Click to verify

If you're looking for free datasets for practicing new skills, you're in luck. The number of free, publicly available datasets has only proliferated over time on sites like Google Dataset Search, Kaggle and Data.gov, making for a treasure trove of information data science professionals can practice their skills on. With that in mind, we rounded up free datasets best suited for a variety of competencies, including product purchasing analysis, ad-click prediction, image classification, sentiment analysis and time series analysis. Lending Club Loan Data by Lending Club Instacart Market Basket Analysis by Instacart MNIST Database of Handwritten Digits by Yann LeCun, Corinna Cortes and Christopher J.C. Burges Amazon Product Reviews by Julian McAuley Hourly Energy Consumption by PJM Web Traffic Time Series Forecasting by Google Uber Pickups in New York City by FiveThirtyEight and NYC Taxi and Limousine Commission Individual Household Electric Power Consumption by UC Irvine Tabular Datasets Tabular data is data organized in a table using rows and columns. It's a simple form of data found in spreadsheet and comma-separated values (CSV) files, and often contains mixed data types (having string and numeric values). Tabular data is used to train machine learning models to find relationships between data points and make predictions on new data. 1. Lending Club Loan Data Dataset: Lending Club Loan Data The Lending Club Loan Data set is a great resource for data scientists to practice loan default prediction and expand their finance domain knowledge. It has a massive number of data points, covering all loans made between 2007 and 2015, and it's feature rich, including credit scores, number of finance inquiries and geographical information. It's not always easy to find a finance dataset that checks both boxes. "Sometimes finance data is kind of hard to get," Joe Eddy, data science instructor of the Metis bootcamp in New York City, told Built In. Also consider diving into Lending Club's API, or — as Raja Iqbal, founder of Data Science Dojo, suggested — the UCI Machine Learning Depository's Default of Credit Card Loans dataset, sourced from default payments in Taiwan. 2. Instacart Market Basket Analysis Dataset: Instacart Market Basket Analysis The Instacart Market Basket Analysis set is one of the largest real-world grocery datasets available, making it a go-to for honing product purchasing prediction and analysis. It spans a whopping three million orders placed by 200,000-plus users, with at least four orders per user and some including as many as 100. It also includes the sequence in which users bought products, and the time of day of each purchase. The patterns within the dataset are easily Google-able, but it remains a great resource for sharpening consumer-side predictive work, Eddy said. 3. Avito Context Ad Clicks Dataset: Avito Context Ad Clicks Dataset Avito Context Ad Clicks truly lets data scientists exercise their ad-click prediction and commercial sector data muscles. It holds quite a lot of data and tables, and offers good opportunities to do some feature engineering in a relational data setting. The Avito dataset is similar to a version of Craigslist, as it includes details like item descriptions in ads, geographical details and demand information. 4. Outbrain Click Prediction Dataset: Outbrain Click Prediction The Outbrain Click Prediction dataset deals with predicting what recommended content users will click next. It samples two billion page views, nearly 17 million clicks and a mess of user recommendations that were made across hundreds of publisher sites over the course of two weeks in 2016. (Outbrain is one of the companies that put boxes of sponsored-content articles at the bottom of sites.) "So much of in-practice data science is literally just ad-click predictions," Eddy said. 5. Coffee Reviews Dataset Dataset: Coffee Reviews Dataset This dataset organizes global reviews of coffee between 2017 and 2022 based on factors like blend name, type of roast, price and geographical origin of coffee beans. It is pre-processed and cleaned, and can be used for pandas, data analysis and feature engineering practice. The original version of the dataset comes with 12 features, while the simplified version has nine features. 6. Electric Vehicle Population Data Dataset: Electric Vehicle Population Data Provided by the State of Washington, this dataset displays information about battery electric vehicles (BEVs) and plug-in hybrid electric vehicles (PHEVs) currently registered through the Washington State Department of Licensing. Data is separated into 17 different columns, showing each vehicle's VIN, county and city of registration, make and model, electric type and electric range. Vehicle model years range from 2013 to the current year, with metadata being routinely updated by the Washington government. Image Datasets Image data is data extracted from images or photos, and can include information on pixels and other visual characteristics. This data is found from image files such as JPEGs, PNGs and GIFs, and is used to train machine learning models to recognize and classify certain objects from pictures (leading to abilities like computer vision). 7. ImageNet Dataset: ImageNet ImageNet is a database of over 14 million images, intended for training image recognition, image classification and computer vision models. The database is organized according to the WordNet hierarchy and has over 20,000 synonym sets indexed, with about 1,000 images associated for each of these sets. The ImageNet project has been historically impactful for advancing computer vision and deep learning research. 8. MNIST Database of Handwritten Digits Dataset: MNIST Dataset The MNIST Database of Handwritten Digits is a well-known dataset — consisting of the digits 0 through 9, written in a variety of handwriting styles — remaining as an ideal entry point for image classification newcomers. Basic classification is "pretty much the simplest possible problem for images, but it's a good starting point for anyone who's playing around with neural network image classification from scratch," Eddy said. "One of the hard things about working with neural networks when you're starting is that, sometimes, training and retooling your models is just very time consuming," he added. "So having relatively smaller, simpler datasets dramatically speeds that up." 9. Dogs vs. Cats Dataset: Dogs vs. Cats The Dogs vs. Cats dataset is another recommended starting point for image classification, even being referenced by Keras creator Francois Chollet in his book, Deep Learning With Python. "It's simple enough to be accessible, but complicated enough to allow for meaningful work — my absolute favorite resource to recommend to a beginner in deep learning, in particular those who want to work with Python," Eddy said. "Having a guided approach is extremely helpful." There's a universe of more complex problems waiting beyond these simple classifications, but the core of those problems often involve repeat applications of exactly the kind of work needed to solve simpler ones, "so starting with one or two of those simple datasets will give you a really strong foundation for exploring almost any standard image problem," Eddy said. RelatedA Tech Company's Guide to Deleting Personal Identifying Information Text Mining and Text Analysis Datasets Text mining and text analysis examine and identify patterns in unstructured text data. This data includes any large amount of text that is not traditionally organized or formatted into a table or database. Text mining and analysis can be used for sentiment analysis, topic modeling and named entity recognition, and may apply natural language processing (NLP) to achieve these tasks. 10. Large Movie Review Dataset Dataset: Large Movie Review Dataset The Large Movie Review Dataset, a 2017 cache of IMDB reviews, includes 25,000 reviews for testing and 25,000 more for training, remaining as a popular tool for sharpening sentiment analysis skills. As Towards Data Science noted in a spotlight, be prepared to do a fair amount of cleaning and vectorization before building and training your classifier. But the effort should pay off. "You can [predict sentiment] with traditional NLP techniques or with slightly fancier, modern neural network techniques," Eddy said. "It's a very easy playground for a wide range of different possible techniques." 11. Twitter and Reddit Sentimental Analysis Dataset Dataset: Twitter and Reddit Sentimental Analysis Dataset X (formerly Twitter) and Reddit hold mounds of text conversations and threads, so the Twitter and Reddit Sentimental Analysis Dataset — containing over 160,000 X posts and 37,000 Reddit comments — offers a handy way to pair your sentiment analysis project to some of the biggest social platforms out there. It's also a good opportunity to practice topic modeling, query writing and a bit of text preprocessing, since web users aren't always known for their grammatical precision. 12. Stack Exchange API Dataset: Stack Exchange API Similarly, the Stack Exchange API dataset gives a glimpse into the ecosystem of the Stack Exchange Q&A sites, so you're bound to find some domains of interest — probably with generally cleaner text than the Reddit dataset. Eddy recalled a past project: "I grabbed a bunch of Statistics Stack Exchange questions to analyze what topics were more or less popular, what language was associated with getting more responses to the question. It was really interesting because it was a topic that I had an attachment to." 13. Amazon Product Reviews Dataset: Amazon Product Reviews The Amazon Product Reviews dataset is a sentiment analysis-friendly set that Iqbal points to, particularly for an advanced data scientist who works, or hopes to break into, marketing. It contains 142.8 million reviews, extensive product information and "also viewed" and "also bought" details, culled from user activity between 1996 and 2014. A natural fit for those looking to filter sentiment analysis into building recommender systems. RelatedHow to Do Data Science From Home Without Going Mad Time Series Datasets Time series data is data collected over an interval of time, and can include historical or real-time data points. This data is used in time series analysis and forecasting, which detect patterns and predict when specific changes may occur over time. Time series data helps forecast events like the weather, stock prices or heart rate readings. Eddy stresses two key criteria when picking datasets for time series analysis — especially for newcomers. First, make sure the time interval is fixed. Whether day-to-day, minute-to-minute, hour-to-hour, the key thing is that the data is recorded in a regular, standardized measurement. Second, watch for clear, seasonal patterns that have logical effects. Any of the below time series datasets fit the bill. 14. Hourly Energy Consumption Dataset: Hourly Energy Consumption Hourly Energy Consumption's dataset features over 10 years of hourly energy consumption data in eastern U.S. states in megawatts, provided by PJM Interconnection. This time series set lets data scientists practice how to predict energy consumption on certain times of the day, week, year or special occasions like holidays. 15. International Greenhouse Gas Emissions Dataset: International Greenhouse Gas Emissions This International Greenhouse Gas Emissions dataset covers global greenhouse gas emission levels from 1990 to 2017, provided by the United Nations. The set aims to help forecast emissions trends and possible types present over time, including emission information on carbon dioxide, methane, nitrous oxide and hydrofluorocarbons. 16. Individual Household Electric Power Consumption Dataset: Individual Household Electric Power Consumption Iqbal recommends the Individual Household Electric Power Consumption dataset, from UCI's Machine Learning Depository, for advanced-level time series practice. The set presents measurements of electric power consumption for one household within a one-minute sampling rate over a span of four years. This can help exercise short-term forecasting skills for one subject, being the single home. 17. Web Traffic Time Series Forecasting Dataset: Web Traffic Time Series Forecasting The Web Traffic Time Series Forecasting dataset, provided by Google, contains traffic data to 145,000 Wikipedia articles, with a focus on using said data to predict future web traffic trends. Each time series data point states the name of the Wikipedia article visited and type of traffic represented (desktop, mobile or spider bot traffic). 18. Uber Pickups in New York City Dataset: Uber Pickups in New York City This dataset supplies date, time and location data for over 20 million Uber and for-hire vehicle trips in the NYC area. The Uber data spans April to September in 2014, while the for-hire vehicles data spans January to June in 2015. Uber's data comes via the New York City Taxi & Limousine Commission. It was released following a 2015 FOIA request by FiveThirtyEight, which delivered much (at the time) eye-opening reporting based on the data. 19. Citi Bike System Data Dataset: Citi Bike System Data The Citi Bike System Data set sheds light on where, when and how far Citi Bike users in New York City ride. This set includes extensive travel information like bike ride ID, start and end times, start and end station IDs and geographical location. Data scientists can utilize this set to determine what days of the week and what times most of these bike rides are taken on. 20. Capital Bike Sharing Dataset: Bike Sharing Bike Sharing is an intermediate-level dataset showing the hourly and daily count of bike rentals in the Capital bikeshare system between 2011 and 2012. Users can practice predicting how many bikes may be rented at certain times based on weather conditions and seasonal factors, and delve into exploratory data analysis, regression modeling or data visualization techniques. Free datasets can be found on websites such as: Google Dataset Search Kaggle Data.gov GitHub Data.world UCI Machine Learning Repository FiveThirtyEight Public datasets can be accessed on websites like Kaggle, Google Dataset Search or GitHub, or through government sources like Data.gov, data.healthcare.gov or data.europa.eu. A dataset is a collection of data often formatted and used for data analysis, research and machine learning projects. A database is a system built for the ongoing management of data, and is often used for storing and accessing data used across a business. Photo by Markus Winkler on Unsplash Curated public datasets are widely used to learn data science and machine learning. But their utility in real-world commercial projects is often overlooked. You must have heard that data scientists spend 80% of their time collecting, cleaning, and preparing the data. Evaluating the viability of an idea will require a decent amount of data. So, do you want to invest time and effort in collecting data only to discover that your idea does not work? It is far cheaper and faster to try your idea on a dataset similar to what you need to collect. First, train a (publically available) model on a similar dataset to see if the results are to your liking. Public datasets can help you rapidly prototype and get clarity about the data you need to collect, saving you time and money. This article gives you a list of 50 public sources for machine learning datasets where you can search and download datasets suitable for your needs. Machine Learning Datasets from Academic Institutes Machine Learning has a much longer history in academic research at universities. So it is not surprising that some of the most versatile open datasets were curated by universities. These are my first go-to place for finding a similar dataset for two reasons. First, these have very diverse datasets for all sorts of machine learning problems. Second, these typically have permissive licenses not prohibiting use for commercial purposes (more about it later). Machine Learning Datasets by Major Cloud Providers Amazon, Microsoft, and Google have rich dataset directories. Some of their datasets are available in cloud data warehouses, which is handy if you are doing machine learning on the cloud. If you are doing machine learning on AWS, Microsoft Azure, or Google Cloud, you should look for a similar dataset here. These are set up to be used easily on the cloud and have permissive licenses. However, these are not as diverse as those in the previous section. Employment and Health Datasets from Governments Before data science became a thing, it was known as statistics. Governments have been collecting all sorts of data for more than half a century. If you are building models in sociology, economic development, education, and health care, a government is quite likely to have a dataset sufficiently similar to what you need for your problem. These are very rich datasets, and it may take you some time to locate the right dataset (but that time will be a fraction of what you will need to collect your own datasets). The licenses are lenient as well to encourage usage. Indian Government's Data and National Data & Analytics Platform (NDAP): Various datasets from the Indian union government and state governments. European Union's Data: European Union's official data source. UK Government Data: Data published by UK's central government, local authorities, and public bodies. US Government's Data: Diverse datasets from the US government. US Government Census Data: US census information like education, employment, health, and housing broken down to zip code level. US Bureau of Labor Statistics: Unemployment, pay & benefits, spending, and other employment data. US Congressional Budget Office: Budget, economic outlook, and projections. US Centers for Disease Control and Prevention: Data for alcohol abuse and various diseases. US Medicare Data: Medicare & Medicaid data Data Commons is an excellent place to explore USA and Euro government data. Socioeconomic Datasets by World Bodies Just like governments, the world bodies like United Nations, WHO, and World Bank have rich socioeconomic datasets. If you are working on models that span across countries, standardizing and stitching together datasets from multiple governments can take quite an effort. Since the goal is to quickly evaluate ML model feasibility before embarking on expensive data collection, cleaning, and labeling, the datasets from world bodies are a better option. United Nations Data (Catalog): Population (and migration), labor market, agriculture, production, price indices, trade, crime, health, environment, tourism, and development data for various countries. UNICEF Data: Childbirth, health, hygiene, nutrition, mortality, education, and development data from United Nations International Children's Emergency Fund. World Health Organization Data: WHO data for health, diseases, pandemics, immunization, pollution, and environmental health data. This is also available at Data Commons for exploration. World Bank Data (Catalog): Economy, growth, agriculture, education, energy, mining, debt, infrastructure, poverty, trade, rural and urban development data. International Monetary Fund (IMF) Data: GDP, trade, price index, exchange rate, monetary and financial data. Asian Development Bank: Similar to World Bank Data but for countries in Asia. Organization for Economic Co-operation and Development (OECD): Similar to World Bank data but mainly for OECD member countries. Financial Datasets by Stock Exchanges and Central Banks Almost all stock exchanges provide historical trading data, and central banks of every nation publish all kinds of financial data. This is your go-to place for trying any time-series models on stock prices, analyzing stock market trends over a period of time, and relations of equities with other asset classes or industrial indicators. Popular Computer Vision Datasets There are a number of image datasets. Over time, these have grown very large and richly annotated. These datasets suffice as a starting point for most of the ML solutions for image-related tasks. Miscellaneous Dataset Sources You may have a problem that does not fall into the broad categories described above. That does not mean that there is no dataset to get you quickly evaluate the feasibility. Google's dataset search engine listed earlier is a good starting point. You can search for the businesses relevant to your task, e.g., yelp for business reviews, Airbnb for rental properties, and YouTube for video. Here are some datasets for sports, news, and other businesses. Do Check the License Please check the dataset's license, especially before using it in a commercial project. A dataset being public does not necessarily imply that you are free to use it as you please. Here are the most common dataset licenses. Open Data Commons (ODC) Community Data License Agreement (CDLA) CDLA Sharing: allowed to use, share and enhance the dataset, but you must give credit and share your data enhancements under the same license. It does not impose any obligations or restrictions on results obtained from the computational use of the data. CDLA Permissive: Same as CDLA Sharing except it gives you the choice to distribute your work under a different license as long as you also include this license for the original dataset. Creative Commons (CC) The PDDL and CC0 are the most permissive licenses. These are rather a renunciation of all rights by the creator. The ODC-By and CC-BY are the next, as they require only attribution acknowledging the use of the dataset. Then comes the CDLA-Permissive and then CDLA-Sharing, as those don't impose any restriction or obligation on the ML model (results obtained from the computational use of the data) you will create. For the rest of the licenses, it depends on whether you intend to use the dataset for commercial purposes. Share-Alike licenses are generally considered "viral", and No-Derivative licenses are the most restrictive. The public datasets are useful while learning Data Science and Machine Learning and rapidly prototyping ideas in commercial settings. Only if an idea is promising, does one need to embark on costly data collection and labeling, With the abundance of free datasets available online, ranging from extensive governmental and economic records to niche areas like Major League Baseball statistics and video game sales, the potential for insightful data science projects is vast. This guide aims to help you navigate these resources, whether you're assembling a portfolio project or honing your SQL and data analysis skills.12 Sources for Free Datasets Anyone Can UseIguazio: Top 22 Free Healthcare Datasets for Machine LearningProvides an overview of 22 open datasets crucial for the development and training of machine learning models in healthcare. These datasets are described as valuable starting points for data scientists and engineers, especially given their open and free nature which can sometimes be challenging to find.Tableau: Free Public Data Sets For AnalysisIt highlights the importance of data in decision-making, emphasizing its role in providing insights and understanding the implications of choices at a granular level. One specific example provided is a COVID-19 data visualization, which serves as a representative of the kinds of visualizations possible with these free data sets.Interview Query: 90+ Free Datasets for Data ScienceProvides a comprehensive overview and categorization of various free datasets useful for data science projects. These datasets cover a broad spectrum of subjects, ranging from governmental and economic data to more specific topics such as Major League Baseball (MLB) statistics and video game sales.Iguazio: Best 10 Free Datasets for Manufacturing10 excellent open manufacturing datasets and data sources for manufacturing data for ML models. This list highlights the significance of open and free datasets for machine learning as crucial for data scientists and engineers working on developing and training ML models for manufacturing, especially given the challenges in accessing manufacturing data365 Data Science: Top 10 Free Dataset Resources for Data Science Projects in 2023The page from 365 Data Science provides a comprehensive list of the top 10 free dataset resources for data science projects in 2023. The article guides readers, especially beginners, through various online resources where they can find free datasets for their projects. These resources include well-known platforms such as Kaggle, Google Dataset Search, GitHub, World Bank Open Data, Data.world, DataHub, Humanitarian Data Exchange, FiveThirtyEight, UCI Machine Learning Repository, and Academic Torrents DataIguazio: Best 13 Free Financial Datasets for Machine LearningProvides a curated list of 13 open financial and economic datasets. These datasets are valuable resources for data scientists and engineers working on developing and training machine learning models in the finance sector.Harvard College: Harvard DataVerseFind datasets across research fields, preview metadata, and download files from Harvard Datavers.Iguazio: 23 Best Free NLP Datasets for Machine LearningThese datasets are categorized into various groups, including Q&A, Reviews and Ratings, Sentiment Analysis, Synonyms, Emails, Long-form Content, and Audio. They are intended for data scientists and professionals to use in training their NLP models for a variety of applications.Column Five: 100+ of the Best Free Data Sources For Your Next ProjectRecognizing the challenge of finding such data, the page offers a curated list of over 100 free data sources from reputable organizations worldwide. These sources are categorized for easy access, aiding users in quickly finding the specific data they need for their projects.Yeshiva University Libraries: Datasets for Computer Science Capstone ProjectsThe page from Yeshiva University Libraries provides a comprehensive list of datasets for Computer Science Capstone Projects. It includes various categories like General Datasets, Subject-Specific Datasets, Datasets Provided by Cloud Providers, and additional options for further exploration.G2: 50 Best Open Data Sources Ready to be Used Right NowThese sources are categorized under various headers including government and global data, financial and economic data, crime and drug data, health and scientific data, academic data, environmental data, business directory data, media and journalism, marketing and social media, and miscellaneous data.Wikipedia: List of Datasets for Machine-learning ResearchThe Wikipedia page provides an extensive compilation of datasets applied in ML research. They include high-quality labeled training datasets for supervised and semi-supervised learning algorithms, which are typically challenging and expensive to produce due to the extensive time required for data labeling.Should You Trust This Data Source?Reputation: A source's credibility can often be gauged by its reputation. Government and academic institutions typically provide trustworthy data.Transparency: Reliable sources are usually transparent about their data collection and update methods. A lack of this information can be a red flag.Updates: Regular updates to a dataset can be a good sign of its reliability and relevance.Can the Dataset Be Inaccurate?Data Collection Methods: Investigate how the data was collected. Biases in the methodology can lead to skewed results.Historical Changes: Remember that data is a snapshot of the past. As conditions change, so does the relevancy of historical data.Cross-Verification: Whenever possible, verify the data with other credible sources to ensure accuracy.Using Free Datasets for ProjectsWhen choosing a dataset for your project, consider its relevance to your goals. Are you trying to demonstrate a specific skill, like data cleaning or complex SQL queries? Pick datasets that allow you to showcase these abilities. Working with a variety of datasets can also broaden your experience and enhance your analytical skills. The data was normalized by A. Bifet and is useful for time-series forecasting and price trend analysis. It contains electricity market data from New South Wales, Australia, collected between May 7, 1996, and December 5, 1998. 2. Phishing Website Detection Due to the lack of a universally agreed upon feature set in phishing detection research, this dataset highlights key attributes and introduces new predictive features. 3. SMS Spam Collection Dataset The SMS Spam Collection is a dataset of 5,574 SMS messages in English, labeled as either ham (legitimate) or spam. This dataset is useful for building and evaluating machine learning models aimed at filtering unwanted messages and improving text classification systems. 4. Students Adaptability Level in Online Education This dataset is designed for predicting students' adaptability levels in online education based on factors like demographics, academic background, internet access, and learning environment. 5. Fruit and Vegetable Disease Detection This dataset contains a comprehensive collection of images categorized to assist in the development of machine learning models for detecting diseases in various fruits and vegetables. 6. Lending Club Loan Data This dataset includes loan data from 2007 to 2015, detailing loan status, payment history, credit scores, finance inquiries, and collections. With 890,000 observations and 75 variables, it is useful for analyzing credit risk, borrower behavior, and loan performance. 7. Startup Success Prediction A startup is a new business aiming for growth, often facing high uncertainty and failure rates. This dataset is designed to predict startup success or failure. 8. Employee's Performance for HR Analytics HR analytics involves using data to make informed decisions about human resources and employee-related matters. It can provide valuable insights to organizations in terms of employee engagement, productivity, and overall organizational effectiveness. 9. Spotify Most Popular Songs Dataset If you're a music enthusiast looking to explore audio features, lyrics, and metadata, this dataset is a solid pick for a data science project. Perfect for music analysis, trend discovery, and building recommendation systems. 10. Census Income With census data on demographics and employment, this dataset is good for predicting whether an individual's income exceeds $50K/year. Useful for classification tasks or for socioeconomic analysis. 11. Car Evaluation With car features like price, capacity, and safety, this dataset is good for evaluating car acceptability. Useful for classification tasks, decision-making models, and recommendation systems. 12. Student Loan Status This dataset is good for analyzing student loan trends, payment behavior, and academic engagement. Useful for predictive modeling, financial analysis, and education research. 13. Average Salary by Job Classification This dataset provides average salaries by job title and grade for full-time regular employees. It offers insights into salary distribution across various roles. 14. Traffic Collision Driver Data Traffic collisions lead to crashes, and understanding driver-related factors can help in accident prevention and risk assessment. This dataset provides details on motor vehicle operators involved in traffic collisions on county and local roadways, capturing key insights into crash incidents and driver behavior. 15. User Demographics and Purchase Behavior Understanding customer trends and improving sales strategies starts with analyzing demographics and purchasing behavior. This dataset includes details like age, gender, and estimated salary, along with whether a user made a purchase. 16. Olympic Games Dataset This dataset provides insights into athlete performance, country participation, and the evolution of the Olympics over time. It is valuable for sports analytics, data visualization, and machine learning applications. 17. Mobile Price Classification In the competitive mobile market, pricing a product correctly is crucial. This dataset captures sales data from various mobile brands to analyze the relationship between device features and price ranges. 18. Student Dropout and Academic Success Prediction Understanding and predicting student outcomes is key to reducing academic dropout rates. This dataset helps identify students at risk by analyzing enrollment data, including academic history, demographics, and socio-economic factors. 19. Animal Condition Classification Assessing the health of animals across different species, this dataset helps build predictive models to determine whether an animal's condition is dangerous based on five key symptoms. 20. Wine Quality Prediction Analyzing wine quality can help winemakers improve production and maintain consistency. This dataset includes key wine attributes, such as acidity and alcohol content, along with quality ratings and wine types. Data science projects often require access to diverse and reliable datasets to build and train models, analyze trends, and derive meaningful insights. While there are numerous sources available, finding high-quality free datasets can be a daunting task. In this article, we will explore 25 reliable sources where you can find free datasets for your data science projects. Each source will be accompanied by a brief explanation and a URL to access the dataset. So, let's dive in! Kaggle: A popular platform for data scientists and machine learning practitioners, Kaggle offers a wide range of free datasets contributed by the community. UC Irvine Machine Learning Repository: A comprehensive repository containing various free datasets suitable for machine learning research. It covers domains like classification, regression, and clustering. Google Dataset Search: Google Dataset Search is a search engine specifically designed to help you find datasets from various sources across the web. Data.gov: The official U.S. government website dedicated to providing open and accessible free datasets from federal agencies. World Bank Open Data: A vast collection of free datasets from the World Bank, covering a wide range of topics related to global development. UNICEF Data: UNICEF Data offers datasets on child well-being, education, health, and other important social indicators. Amazon Web Services (AWS) Public Datasets: AWS provides a collection of public datasets that can be accessed for free, covering domains like biology, climate, and economics. Google Cloud Public Datasets: Google Cloud Public Datasets offer a variety of datasets, including genomics, environmental, and public health data. DataHub: DataHub is a platform that hosts a wide range of free datasets, including social, economic, and scientific data. Data.world: Data.world is a community-driven platform where users can discover, share, and collaborate on free datasets. FiveThirtyEight: FiveThirtyEight provides datasets related to politics, sports, economics, and more. Their datasets are often used for data-driven journalism. OpenML: OpenML is an open science platform that allows users to share datasets and machine learning experiments. GitHub: GitHub hosts numerous repositories containing datasets shared by individuals, organizations, and research institutions. U.S. Census Bureau: The U.S. Census Bureau offers various datasets that provide demographic, economic, and geographic information. European Union Open Data Portal: The European Union Open Data Portal provides access to datasets from EU institutions and member states, covering various domains. Quandl: Quandl is a platform that hosts financial, economic, and alternative datasets suitable for quantitative analysis. Data.gov.uk: The UK government's official data portal, offering a wide range of open datasets. Statista: Statista provides statistical data and charts on various topics, including industries, countries, and consumer behavior. Reddit Datasets: The Reddit community r/datasets is a valuable resource for finding and sharing datasets on a wide range of topics. U.S. Bureau of Labor Statistics: The U.S. Bureau of Labor Statistics offers datasets related to employment, inflation, wages, and more. Data.gov.au: The Australian government's open data portal, providing access to diverse datasets on various subjects. NASA Open Data: NASA Open Data offers datasets related to space exploration, satellite imagery, and climate research. Data.gov.sg: The Singapore government's data portal, offering datasets on topics like demographics, transportation, and health. Open Data Network: The Open Data Network allows you to explore and access datasets from multiple cities and organizations worldwide. Data.gov.hk: The Hong Kong government's data portal, providing datasets on different aspects of the city. Access to high-quality datasets is crucial for successful data science projects. In this article, we have explored 25 reliable sources where you can find free datasets to fuel your data-driven endeavors. Remember to choose the datasets that align with your project requirements and explore the documentation provided by each source for a better understanding of the data. Happy exploring and may your data science journey be filled with valuable insights! Can I use these datasets for commercial purposes?The permissions and licenses for each dataset may vary. It's important to review the terms of use provided by the source before using the datasets for commercial purposes. Are these datasets regularly updated?The update frequency depends on the dataset and the source. Some sources provide real-time data, while others update their datasets periodically. Check the source's documentation for more information. Can I contribute my own datasets to these platforms?Many of the mentioned platforms allow users to contribute datasets. Refer to the respective platform's guidelines for instructions on how to contribute your datasets. What formats are these datasets available in?Datasets can be available in various formats such as CSV, JSON, XML, or Excel. The format availability depends on the source and dataset. Check the documentation or download options provided by the source for the available formats. Are there any restrictions on the use of these datasets?The usage restrictions, if any, will be specified by the dataset source. Make sure to review the terms and conditions or licenses associated with each dataset to ensure compliance. *** Machine Learning books from this Author: Related Articles November 13, 2024 Looking for free datasets for projects? You're in the right place. We've sourced and vetted spectacular datasets for the following: Python R Data science Data visualization Data cleaning Machine learning Probability and statistics Business analysis Excel If you're trying to find free datasets so that you can learn by building projects, we have plenty of options for you. All of our data science courses end with guided projects you can complete using real, high-quality datasets. The projects are designed to accelerate learning and showcase your skills with irresistible portfolio. Identify Customers Likely to Churn: Use an Excel dataset to conduct an exploratory data analysis (EDA) for a telecommunications provider to identify customers who are at risk of churn. Analyze Retail Sales: Work with retail sales data to explore trends and relationships. Build basic models to confirm the statistical significance of your insights. Our Data Analysis with Excel course contains 2 additional projects. Our Data Cleaning with Python path contains 4 other projects. Sign up for free here. Our Data Analysis and Visualization with Python path contains 3 other projects. Sign up for free here. Our R for Data Analysis course contains 2 other projects. Predict House Sale Prices: Use housing data from a city in the United States to build and improve linear regression models. Predict the Stock Market: Use historical data from the S&P 500 Index to make predictions about future prices. Predict Bike Rentals: Use a machine learning dataset of bike rentals and apply decision trees and random forests to predict the number of future bike rentals. Our Machine Learning with Python courses contains 4 other projects. Our Probability and Statistics with Python course contains 9 other projects. Analyze Retail Sales: Work with a retail sales dataset to explore trends and relationships. Build basic models to confirm the statistical significance of your insights. Identify Customers Likely to Churn: Use a training dataset from Kaggle to conduct an Exploratory Data Analysis (EDA) on data from a telecommunications provider to determine customers at risk of churn. Visualize Company Stock Performance: Create a report comprised of data visualizations to answer questions about company stock performance from one of four possible datasets. Our Business Analyst with Power BI career path contains 5 other projects. Sign up for free here. A typical data visualization project might be something along the lines of "I want to make an infographic about how income varies across the different states in the US." There are a few considerations to keep in mind when looking for a good dataset for a data visualization project: It shouldn't be messy because you don't want to spend a lot of time cleaning data. It should be nuanced and interesting enough to make charts about. Ideally, each column should be well-explained so the visualization is accurate. The data set shouldn't have too many rows or columns, so it's easy to work with. Good places to find good datasets for data visualization projects are news sites that release their data publicly. They typically clean the data for you and already have charts you can replicate or improve. FiveThirtyEight is an incredibly popular interactive news and sports site started by Nate Silver. They write interesting data-driven articles, like "Don't blame a skills gap for lack of hiring in manufacturing" and "2022 NFL Predictions." View the FiveThirtyEight Datasets Here are some examples: Airline Safety — public data that contains information on accidents from each airline. US Weather History — a public dataset of historical weather data for the US. Study Drugs — open data on who's taking Adderall in the US. BuzzFeed started as a purveyor of low-quality articles but has since evolved and now writes some investigative pieces, like "The court that rules the world" and "The short life of Deonte Hoard." View the BuzzFeed Datasets Here are some examples: NASA is a publicly-funded government organization, and thus all of its data is public. It maintains websites where anyone can download its datasets related to earth science and datasets related to space. You can even sort by format on the earth science site to find all of the available CSV datasets, for example. Sometimes, you just want to work with a large dataset. The end result matters less than the process of reading and analyzing the data. You might use tools like Spark or Hadoop to distribute the processing across multiple nodes. Things to keep in mind when looking for a good data processing dataset: The cleaner the data, the better — cleaning a large dataset can be very time-consuming. The dataset should be interesting. There should be an interesting question that can be answered with the data. Good places to find large public data sets are cloud-hosting providers like Amazon and Google. They have an incentive to host the data sets because they make you analyze them using their infrastructure (and pay them to use it). Amazon makes large datasets available on its Amazon Web Services platform. You can download the data and work with it on your own computer or analyze the data in the cloud using EC2 and Hadoop via EMR. You can read more about how the program works here. Amazon has a page that lists all of the free datasets for you to browse. You'll need an AWS account, although Amazon provides a free access tier for new accounts that will enable you to explore the data without being charged. View AWS Public Datasets Here are some examples: Much like Amazon, Google also has a cloud-hosting service, called Google Cloud Platform. With GCP, you can use a tool called BigQuery to explore large datasets. Google lists all of the data sets on a page. You'll need to sign up for a GCP account, but the first 1TB of queries you make are free. View Google Public Datasets Here are some examples: USA Names — contains all Social Security name applications in the US from 1879 to 2015. Github Activity — contains all public activity on over 2.8 million public Github repositories. Historical Weather — open data from 9000 NOAA weather stations from 1929 to 2016. Wikipedia is a free, online, community-edited encyclopedia. Wikipedia contains an astonishing breadth of knowledge, containing pages on everything from the Ottoman-Habsburg Wars to Leonard Nimoy. As part of Wikipedia's commitment to advancing knowledge, they offer their content for free and regularly generate dumps of all the articles on the site. Additionally, Wikipedia offers edit history and activity, so you can track how a page on a topic evolves over time and who contributes to it. You can find the various ways to download the data on the Wikipedia site. You'll also find scripts to reformat the data in various ways. View Wikipedia Datasets Here are some examples: When you're working on a machine learning project, you want to be able to predict a column from the other columns in a dataset. In order to be able to do this, we need to make sure that: The dataset isn't too messy — if it is, we'll spend all of our time cleaning the data. There's an interesting target column to make predictions for. The other variables have some explanatory power for the target column. There are a few online repositories of datasets that are specifically for machine learning. These datasets are typically cleaned up beforehand, and allow for testing of algorithms very quickly. Kaggle is a data science community that hosts machine learning competitions. There are a variety of externally-contributed, interesting datasets on the site. Kaggle has both live and historical competitions. You can download data for either, but you have to sign up for Kaggle and accept the terms of service for the competition. You can download data from Kaggle by entering a competition. Each competition has its own associated dataset. There are also user-contributed datasets found in the new Kaggle Datasets offering. View Kaggle Datasets Here are some examples: Satellite Photograph Order — a free dataset of satellite photos of Earth — the goal is to predict which photos were taken earlier than others. Manufacturing Process Failures — a dataset of variables that were measured during the manufacturing process. The goal is to predict faults with the manufacturing. Multiple Choice Questions — a dataset of multiple choice questions and the corresponding correct answers. The goal is to predict the answer for any given question. The UCI Machine Learning Repository is one of the oldest sources of datasets on the web. Although the datasets are user-contributed and thus have varying levels of documentation and cleanliness, the vast majority are clean and ready for machine learning to be applied. UCI is a great first stop when looking for interesting datasets. You can download data directly from the UCI Machine Learning repository, without registration. These datasets tend to be fairly small, and don't have a lot of nuance, but are good for machine learning. View UCI Machine Learning Repository Here are some examples: Email Spam — contains emails, along with a label of whether or not they're spam. Wine Classification — contains various attributes of 178 different wines. Solar Flares — attributes of solar flares, useful for predicting characteristics of flares. When looking for a good dataset for a data cleaning project, you want it to: Be spread over multiple files. Have a lot of nuance and many possible angles to take. Require a good amount of research to understand. Be as "real-world" as possible. These types of datasets are typically found on aggregators of datasets. These aggregators tend to have datasets from multiple sources, without much curation. Too much curation gives us overly neat datasets that are hard to do extensive cleaning on. data.world describes itself as 'the social network for data people,' but could be more correctly described as 'GitHub for data.' It's a place where you can search for, copy, analyze, and download datasets. In addition, you can upload your data to data.world and use it to collaborate with others. In a relatively short time it has become one of the 'go to' places to acquire data, with lots of user contributed datasets as well as fantastic datasets through data.world's partnerships with various organizations, including a large amount of data from the US Federal Government. One key differentiator of data.world is they have built tools to make working with data easier – you can write SQL queries within their interface to explore data and join multiple datasets. They also have SDKs for R and Python to make it easier to acquire and work with data in your tool of choice (You might be interested in reading our tutorial on the data.world Python SDK.) View data.world Datasets Data.gov is a relatively new site that's part of a US effort towards open government. Data.gov makes it possible to download data from multiple US government agencies. Data can range from government budgets to school performance scores. Much of the data requires additional research, and it can sometimes be hard to figure out which dataset is the "correct" version. Anyone can download the data, although some datasets require additional hoops to be jumped through, like agreeing to licensing agreements. You can browse the data sets on Data.gov directly without registering. You can browse by topic area or search for a specific dataset. View Data.gov Datasets Here are some examples: Chronic Disease Data — data on chronic disease indicators in areas across the US. The World Bank is a global development organization that offers loans and advice to developing countries. The World Bank regularly funds programs in developing countries, then gathers data to monitor the success of these programs. You can browse World Bank datasets directly, without registering. The datasets have many missing values, and sometimes take several clicks to actually get to data. View World Bank Datasets Here are some examples: World Bank Project Costs — data on World Bank projects and their corresponding costs. Reddit, a popular community discussion site, has a section devoted to sharing interesting datasets. It's called the datasets subreddit, or /r/datasets. The scope of these datasets varies a lot, since they're all user-submitted, but they tend to be very interesting and nuanced. You can browse the subreddit here. You can also see the most highly upvoted datasets here. View Top /r/datasets Posts Here are some examples: New York City Property Tax Data — data about properties and assessed value in New York City. Academic Torrents is a new site that is geared around sharing the datasets from scientific papers. It's a newer site, so it's hard to tell what the most common types of datasets will look like. For now, it has tons of interesting datasets that lack context. You can browse the datasets directly on the site. Since it's a torrent site, all of the datasets can be immediately downloaded, but you'll need a Bittorrent client. Deluge is a good free option. View Academic Torrents Datasets Here are some examples: Enron Emails — a set of many emails from executives at Enron, a company that famously went bankrupt. Student Learning Factors — a set of factors that measure and influence student learning. News Articles — contains news article attributes and a target variable. GitHub has an API that allows you to access repository activity and code. You can get started with the API here. The options are endless — you could build a system to automatically score code quality, or figure out how code evolves over time in large projects. Get started with the Github API Wunderground has an API for weather forecasts that is free up to 500 API calls per day. You could use these calls to build up a set of historical weather data and make predictions about the weather. Get started with the Wunderground API The World Health Organization (WHO) maintains a large dataset on global health at the Global Health Observatory (GHO). The dataset includes all the WHO data on the COVID-19 global pandemic. The GHO offers a diverse range of data on topics such as antimicrobial resistance, dementia, air pollution, and immunization. You can find data on pretty much any health-related topic at the GHO, making it an extremely valuable free dataset resource for data scientists working in the health field. View WHO's datasets. The Pew Research Center is well-known for political and social science research. In the interest of furthering research and public discourse, they make all of their datasets publicly downloadable for secondary analysis, after a set period of time elapses. You can choose from datasets on US politics, journalism and media, internet and tech, science and society, religion and public life, amongst other topics. Climate change is a hot topic at the moment, if you'll pardon the pun. Data scientists who want to crunch the numbers on weather and climate can access large US datasets from the National Centers for Environmental Information (NCEI). OpenML is an online platform that allows users to share and explore datasets for machine learning and deep learning. They hosts a large collection of datasets covering various domains, including image classification, natural language processing, and social sciences. All datasets are contributed by the community and are available for free. OpenML has a strong community component. You can use it to replicate others' experiments, compare your models against existing benchmarks to seek improvement, and contribute to other community members on their projects. View OpenML's datasets. The internet is full of cool datasets you can work with. But for something truly unique, what about analyzing your own personal data? Here are some popular sites that make it possible to download and work with data you've generated. Amazon allows you to download your personal spending data, order history, and more. To access it, click this link (you'll need to be logged in for it to work) or navigate to the Accounts and Lists button in the top right. On the next page, look for the Ordering and Shopping Preferences section, and click on the link under that heading that says "Download order reports."Here is a simple data project tutorial that you could do using your own Amazon data to analyze your spending habits. Facebook also allows you to download your personal activity data. To access it, click this link (you'll need to be logged in for it to work) and select the types of data you'd like to download.Here is an example of a simple data project you could build using your own personal Facebook data. Netflix allows you to request your own data for download, although it will make you jump through a few hoops, and will warn you that the process of collating your data may take 30 days. As of the last time we checked, the data they allow you to download is fairly limited, but it could still be suitable for some types of projects and analysis. OK, so this isn't strictly a dataset – rather a search tool to find relevant datasets. As you already know, Google is a data powerhouse, so it makes sense that their search tool knocks the socks off of other ways to find specific datasets. All you need to do is head over to Google Dataset Search and type a keyword or phrase related to the dataset you're looking for in the search bar. The results will list all the datasets indexed on Google for that particular search term. The datasets are generally from high-quality sources, of which some are free and others available for a fee or subscription. Next steps If you liked this, you might like to read the other posts in our 'Build a Data Science Portfolio' series: Basic concepts 130 Featured In this post we can find free public datasets for Data Science projects. There is a big number of datasets which cover different areas - machine learning, presentation, data analysis and visualization. You can find information for: Data sources - big datasets collections which has curated data and advanced searching Sample datasets - datasets good for data analysis. Starting point for beginners who would like to learn Data Science Datasets resources - useful resources for datasets which can be loaded easily Datasets from Python libraries - load datasets with single line of code from different Python libraries like 'seaborn' NoteThis post will be updated on regular basis so please suggest new ideas and datasets in the comment section below. 1. Dataset Sources Below we can find a table of dataset collections. Most of them have advanced searching by: file type size number of rows tags Other soruces: 2. Datasets samples There are several listed below which are used in this site for demonstration of data science basics: 3. Datasets resources 4. Read Kaggle Datasets To read Kaggle datasets we can use the Python library kaggle. Downloading dataset from kaggle with Python code is available from method: dataset_download_file: import kaggle kaggle.api.authenticate() kaggle.api.dataset_download_file('dorianlazar/medium-articles-dataset', file_name='medium_data.csv', path='data/') For more information and examples refer to: How to Search and Download Kaggle Dataset to Pandas DataFrame 5. Load Datasets by Python libraries In this section we can find several useful datasets for different purposes like: machine learning visualization testing creating own datasets with fake data 5.1 datasets - machine learning Python library datasets offers a huge number of free and easy to use datasets. It can be installed by: pip install datasets To list all available datasets we can use method: datasets.list_datasets(): from datasets import list_datasets, load_dataset print(list_datasets()) It will return more than 7000 datasets. To load dataset we can use method: datasets.load_dataset(dataset_name, **kwargs): squad_dataset = load_dataset('squad') squad_dataset This give us two datasets: training dataset validation dataset DatasetDict({ train: Dataset({ features: ['id', 'title', 'context', 'question', 'answers'], num_rows: 87599 }) validation: Dataset({ features: ['id', 'title', 'context', 'question', 'answers'], num_rows: 10570 }) }) To access the dataset for training we can use: squad_dataset['train']. Finally we can loaded as Pandas DataFrame by: import pandas as pd pd.DataFrame(squad_dataset['train']) 5.2 pandas - test datasets Pandas offers multiple ways to download datasets with a single line of code. Let's cover few of them starting with the test data in Pandas github: Pandas data files - csv, xml, html scrape wiki tables Next we can load data from Pandas by scraping wikipedia: pd.read_html(')[2] create datasets We can create random or fake datasets with Pandas by: 5.3 seaborn - visualization datasets Seaborn offers free tests which are good for visualization. With single line of code we can get DataFrame good for data wrangling and visualization: import seaborn as sns df = sns.load_dataset('flights') All datasets available from seaborn library: seaborn-data. sklearn-learn - machine learning We can get sample datasets from sklearn-learn by methods like: load_iris from sklearn.datasets import load_iris iris = load_iris() To find more sample datasets from sklearn we can use the next code: from sklearn import datasets dir(datasets) This will list all available options like: 'load_sample_images', 'load_svmlight_file', 'load_svmlight_files', 'load_wine', 'make_biclusters', 'make_blobs', 'make_checkerboard', 'make_circles', You can find more about sklearn-learn datasets on this link: sklearn.datasets: Datasets. 5.4 dataprep - data analysis To load dataset we can use method: load_dataset from dataprep.datasets import load_dataset df = load_dataset("titanic") to list datasets we can use: from dataprep.datasets import get_dataset_names get_dataset_names() which results into several datasets like: ['waste_hauler', 'wine-quality-red', 'countries', 'house_prices_train', 'iris', 'adult', 'covid19', 'titanic', 'patient_info', 'house_prices_test'] More information about dataprep datasets: Datasets DataPrep 6. Top 10 sites with interesting datasets The table below is based on this Kaggle list: Top 10 sites with interesting datasets Here's a concise Markdown table: | Dataset Source | Description | |--------------------------------|-------------| | **UCI Machine Learning Repository** | Collection of datasets for research and education. | | **Google Dataset Search** | Search engine for datasets across various domains. | | **Data.gov** | U.S. government's open data portal. | | **World Bank Open Data** | Global development datasets on education, health, and more. | | **Reddit's r/datasets** | Community-driven subreddit for sharing datasets. | | **AWS Public Datasets** | Large public datasets hosted by AWS. | | **FiveThirtyEight** | Data journalism site with political, sports, and economic datasets. | | **Data.gov.uk** | UK government's open data portal. | | **DataHub.io** | Platform for open datasets across multiple fields. | | **Data.world** | Collaborative platform for dataset discovery and sharing. | This keeps it clear and to the point. 7. Conclusion In

this article, we covered free datasets sources and discussed common ways to download dataset from them. Through practical examples, we learned how to download and use those datasets in Python and Pandas. We covered different Python libraries which offer public datasets for learning. Finally, we covered how to create test datasets with fake data. Those datasets and ideas should be sufficient for practicing and learning data science. By using DataScientYst - Data Science Simplified, you agree to our Cookie Policy. Whether you're a student exploring new concepts or a seasoned professional, there's an undeniable truth we all stumble upon: the importance of quality data. We've all heard the saying, "garbage in, garbage out," and it's a reminder that our projects are only as good as the data we feed them. Seeking robust and relevant datasets is key to the success of any endeavor. And let's face it, there's no better way to excel and make an impact than by rolling up your sleeves and diving into projects, headfirst. This is especially true when it comes to data-centric work, where each new dataset presents a unique challenge and an opportunity to sharpen your skills. I'm here to let you in on a secret: there's a treasure trove of fascinating and accessible datasets just waiting for your discovery. No more struggling to find that perfect data source or feeling limited by data constraints. If you're eager to take your projects to the next level and dive into some exciting resources, you've come to the right place. In this article, we will guide you through a carefully curated list of websites and sources that offer an abundance of datasets, perfect for any idea you want to bring to life. Get ready to bookmark these go-to websites and embark on a journey of endless project possibilities, where you can focus on honing your craft, gaining valuable experience, and maybe even creating something that will revolutionize the industry. So, without further ado, let's unlock the door to a world of data-driven exploration! Table of contents Kaggle is a prime platform for accessing datasets due to its vast repository covering diverse topics like astronomy, diabetes, and more. With user-friendly features for filtering by license type and topic, Kaggle ensures easy access to high-quality datasets, making it ideal for both beginners and experienced data scientists seeking valuable resources for their projects. Find Kaggle Datasets here. To master Kaggle, read this article. AWS Data Exchange is a robust platform for data exchange, offering a wide range of datasets from various providers, including government agencies and companies. It is particularly useful for projects as it provides a one-stop shop for diverse data needs. With data from multiple sources, it saves time and offers consistency. The platform's reliability and the ability to find unique datasets make it a valuable resource for any data-related project. Click here to checkout this website to find dataset. Data.world is a fantastic resource for datasets due to its community-driven nature. It offers a vast collection of user-uploaded datasets, ensuring variety and specialty. The platform fosters collaboration and data exchange, making it ideal for finding unique and specific datasets. The website's search functionality and categorization make it easy to navigate and discover relevant data quickly. Click here to explore Data.world! GitHub, while primarily known as a code repository, has evolved into a valuable resource for datasets. Its vast community of users often shares datasets alongside their code, providing a unique perspective. The website is ideal for projects as it offers a one-stop shop for code and data, with a simple interface. GitHub's search functionality and filtering options make finding relevant datasets efficient and straightforward. Click here to checkout this website to find dataset. OpenDataSoft is a reliable data-sharing platform, offering a comprehensive directory of open datasets. Its strength lies in its focus on data sharing and collaboration. The website is ideal for projects as it provides a one-stop source for diverse data needs, with a user-friendly interface. The platform's commitment to open data and its global reach make it a valuable tool for anyone looking for transparent and accessible datasets. Click here to checkout Open Data Soft. DataHub is an excellent platform for accessing free and open datasets on various topics. Its strength lies in its comprehensive collection of data from different sources, making it a one-stop shop for projects. The website is user-friendly, well-organized, and efficient for finding relevant data quickly. DataHub's commitment to open data and its partnership with the data analytics company Qlik ensure a reliable and valuable resource for any data-driven project. Click here to checkout this website to find dataset. The Google Public Data Explorer is a unique tool that provides access to a vast array of public data from international organizations and academic institutions. While it is not a direct dataset source, the platform offers a user-friendly interface to explore and visualize data. This makes it ideal for gaining insights and understanding trends. The tool's strength lies in its ability to make complex data accessible and its suitability for projects requiring dynamic data representation. Click here to explore Google Public Data Explorer! Data.gov is an extensive open data platform provided by the US government. It centralizes data from a wide range of federal, state, and local agencies, covering diverse topics. The website is well-structured, allowing users to easily search and filter data by format, topic, and agency. For projects, Data.gov offers a reliable and authoritative source of data, ensuring consistency and authenticity. Its regular updates and commitment to open data make it an invaluable resource for anyone seeking US-specific datasets. Click here to explore this website to find datasets. Data.gov.uk is the UK government's comprehensive open data portal. It provides access to a wide range of data from UK public bodies and agencies, ensuring transparency and accessibility. The website is well-organized and user-friendly, making it easy to navigate and find relevant data. For projects, it offers a reliable source of UK-specific data, covering various topics such as economics, health, and education. The platform's commitment to open data and its regular updates make it a valuable resource for data-driven projects. Click here to explore data.gov. The Census Data of India website offers a rich collection of demographic, economic, and social data about the country. It provides valuable insights into India's population, including various statistics and indicators. The website is essential for projects focusing on India, offering detailed information at the national and regional levels. Click here to explore datasets at Census Data of India. Open Government Data Platform India: This is a centralized platform for open data from the Indian government, covering various sectors. India Data Portal: It provides open data on various themes, including agriculture, education, and energy. National Data Portal: This portal offers a comprehensive collection of datasets, publications, and maps from the Indian government. Ministry of Statistics and Program Implementation: The official website of the ministry provides data on economic indicators, demographics, and more. Open Data Initiative India: This initiative aims to provide open data on governance, infrastructure, and other key areas. The World Bank Open Data platform is a comprehensive source of global development data. It offers extensive datasets covering a wide range of indicators, including finance, health, education, and more. The website is well-organized, allowing users to easily search and filter data by country, topic, and indicator. For projects with a global focus, this platform is invaluable, providing authoritative and up-to-date information. The World Bank's commitment to data transparency and its extensive coverage of economic and social development data make it a go-to resource for researchers, policymakers, and data analysts alike. Click here to explore this website to find more datasets. UN Data, maintained by the United Nations, is a rich repository of global data covering a wide range of topics. The platform offers data on areas such as population, environment, trade, and human development, providing valuable insights into global trends and issues. UN Data is well-structured and user-friendly, allowing users to search and filter data by country, region, and theme. For projects with a global perspective, this website is essential, offering reliable and authoritative data. The UN's commitment to data transparency and its comprehensive coverage of socio-economic and environmental data make it a trusted source for researchers, policymakers, and anyone working on international development initiatives. Click here to explore the datasets at UN Data. Eurostat is the European Union's statistical office, providing comprehensive data on EU member states. It offers a wide range of data covering economic, social, and agricultural topics, among others. The website is user-friendly, allowing easy navigation and data exploration by country, indicator, and theme. Eurostat is particularly valuable for projects focused on Europe, offering reliable and up-to-date information. The office's commitment to data transparency and its extensive coverage of EU-specific data make it an essential resource for understanding the economic and social dynamics of the region. Eurostat also provides data visualization tools and analytical reports, further enhancing its usefulness for researchers and policymakers alike. Click here to explore the websites to find datasets. FRED Economic Data, hosted by the Federal Reserve Bank of St. Louis, is an extensive database of US and international economic data. It offers thousands of economic time series, covering various indicators such as inflation, employment, interest rates, and more. The platform is user-friendly, providing powerful search and filtering tools to navigate the vast dataset. FRED Economic Data is ideal for projects requiring economic analysis, offering a one-stop shop for historical and current economic indicators. The Federal Reserve Bank's commitment to data transparency and its regular updates make this platform a trusted source for researchers, economists, and anyone interested in economic trends and forecasting. Click here to checkout the FRED Economic Data here. The UCI Machine Learning Repository is a well-known and trusted source for machine learning and artificial intelligence research and education. Maintained by the University of California, Irvine, it offers a diverse collection of datasets specifically curated for ML and AI applications. The repository is user-friendly, providing a comprehensive dataset search and filtering system. It is valuable for projects as it covers various data types, from text and images to time series and sensor data. The platform's commitment to supporting ML research and education, along with its regular updates, makes it an indispensable resource for students, researchers, and practitioners in the field of machine learning and artificial intelligence. Click here to checkout datasets in UCI ML Repository. OpenML is a collaborative platform designed specifically for the machine learning community. It offers a unique approach to data sharing by providing not just datasets but also machine-learning tasks and flows. This platform allows researchers and practitioners to share and reproduce experiments easily. OpenML is well-organized, with a user-friendly interface, making it simple to search and explore datasets, tasks, and flows. The platform fosters reproducibility and transparency in ML research, making it ideal for projects requiring a more comprehensive approach to data and experimentation. OpenML's commitment to openness and its active community make it a valuable resource for advancing machine-learning practices. Click here to explore this website to find datasets. CMU StatLib is a renowned statistical database provided by Carnegie Mellon University. It offers a rich collection of datasets specifically curated for statistical and machine learning research and education. The database is well-organized, providing a comprehensive search and browsing experience. CMU StatLib is valuable for projects requiring statistical analysis, offering a diverse range of data types and topics. The platform's association with a leading university ensures the reliability and quality of the datasets. CMU StatLib's regular updates and commitment to supporting statistical research and education make it an indispensable resource for students, researchers, and practitioners in statistics and machine learning. Click here to explore this website to find datasets. Google Dataset Search is a specialized search engine by Google, launched in 2018, designed to help researchers find freely available online data. It allows filtering by data type and is based on schema.org metadata standards. The service complements Google Scholar and offers a user-friendly interface accessible on mobile devices. Click here to checkout Google Datasets Search! Open Data Monitor is a unique website that aggregates open data portals from around the world. It serves as a discovery platform, making it easier for users to find datasets from different countries and regions. The website is well-designed, providing a centralized search engine for exploring global open data initiatives. Open Data Monitor is valuable for projects requiring international data, offering a diverse and comprehensive collection of sources. The platform promotes transparency and accessibility, ensuring that users can quickly locate relevant data portals and gain insights into global open data practices. Its continuous updates and expansion make it a dynamic resource for anyone working with international data. Click here to explore this website to find datasets. DataPortals.org is a comprehensive global registry specifically designed to help users discover open data portals from cities, regions, and countries around the world. It serves as a centralized platform, providing easy access to a wide range of datasets offered by governments and organizations. DataPortals.org is valuable for projects and research that require diverse and localized data. The website promotes transparency and open data practices, ensuring users can quickly locate and utilize datasets that align with their specific needs. With regular updates and a growing community, DataPortals.org has established itself as a dynamic and trusted resource for anyone seeking open data sources, offering a unique perspective on global data initiatives. Click here to explore datasets at DataPortals.org. Data Is Plural is a unique initiative that curates interesting and diverse datasets from various sources on the web. It takes the form of both a newsletter and an archive, providing a regular stream of data-related content. Data Is Plural offers a broad range of topics, covering areas that are often overlooked by traditional data platforms. This makes it ideal for projects requiring unique and specialized data. The newsletter format provides a convenient way to discover new datasets, while the archive ensures a growing collection of valuable resources. Data Is Plural's commitment to exploring the "plural" nature of data and its focus on lesser-known datasets make it a dynamic and intriguing resource for data enthusiasts, researchers, and anyone seeking fresh perspectives in their projects. Click here to explore this website to find datasets. Nasdaq is a renowned global electronic marketplace for buying and selling securities, particularly known for its focus on technology stocks. While primarily a stock exchange, Nasdaq also offers a wealth of data and analytics tools on its website. This includes real-time market data, company profiles, financial news, and investment analysis. For projects involving stock market analysis or financial research, Nasdaq is a valuable resource, providing authoritative data and insights. The platform offers various data products, APIs, and solutions tailored to different user needs. Nasdaq's reputation, combined with its commitment to innovation and data transparency, makes it a trusted source of financial information for investors, traders, and researchers worldwide. Click here to access Nasdaq datasets. Yelp is a well-known crowd-sourced review platform that periodically releases large datasets of its business and review data. The Yelp Dataset is valuable for academic research and data science competitions, offering insights into consumer behavior and preferences. It provides rich information, including business details, user reviews, and ratings, allowing for a wide range of analytical projects. The dataset is unique due to its scale and real-world applicability. Yelp's commitment to data transparency and its impact on local businesses make this dataset a valuable resource for researchers and data scientists, offering a window into consumer trends and behavior patterns. Checkout datasets at yelp here. The Pew Research Center is a non-profit think tank that conducts surveys and research on a wide range of topics, including social issues, media usage, and political attitudes. The center is known for its commitment to providing unbiased and reliable data to the public. Its website offers easy access to a wealth of datasets, making it a valuable resource for projects requiring public opinion and demographic information. The Pew Research Center's data covers a diverse range of subjects, such as technology adoption, social trends, and global attitudes. The platform provides user-friendly data exploration tools and detailed methodology explanations, ensuring transparency and understanding. Researchers, journalists, and anyone interested in societal insights will find the center's data and analyses invaluable, offering a window into the beliefs and behaviors of diverse populations. Click here to checkout this website to find datasets. NASA Open Data is a fascinating portal that provides access to a wide range of scientific data from NASA's various missions and research. It offers a unique opportunity to explore space science, Earth science, and aerospace research data. The platform is user-friendly, allowing easy discovery and download of datasets, images, and even software. NASA Open Data is ideal for projects requiring scientific and space-related information, providing authoritative and detailed insights. The platform's commitment to data transparency and its continuous updates ensure that researchers, students, and enthusiasts can access the latest findings and contribute to further exploration. With data from NASA's renowned missions, this portal offers a window into the universe, inspiring innovation and discovery. Click here to explore NASA Open Data. Figshare is a trusted repository designed for hosting and sharing scientific research outputs, including datasets, code, and other research artifacts. It provides a platform for researchers to make their work openly accessible and citable. Figshare is valuable for projects requiring scientific data, offering a wide range of disciplines, such as life sciences, social sciences, and computer sciences. The platform ensures proper credit and attribution to researchers, promoting open science practices. Figshare's user-friendly interface allows easy search and download of datasets, fostering collaboration and reproducibility. With a commitment to long-term data preservation and an expanding community, it has become an indispensable resource for researchers, institutions, and anyone seeking open scientific data and resources. Click here to explore this website to find datasets. BuzzFeed News Data is a unique initiative by the BuzzFeed data journalism team, where they release the datasets used in their investigative journalism and articles. The platform offers a range of datasets covering topics like politics, social issues, and media. BuzzFeed News Data provides valuable insights into the data-driven stories that shape our world. The datasets are often accompanied by explanatory articles, providing context and understanding. This initiative promotes data transparency and accountability, allowing researchers and the public to explore and analyze the information themselves. BuzzFeed News Data is ideal for projects requiring real-world, contemporary datasets with a focus on current affairs. It bridges the gap between data and storytelling, offering a dynamic resource for data journalists and researchers alike. Checkout the following links to find the datasets: Reddit's "/r/datasets" community, or "subreddit," is a vibrant and unique collection of datasets shared and discussed by its members. It offers a diverse range of datasets covering various topics, from science and technology to social sciences and hobby projects. The community-driven nature of "/r/datasets" provides a dynamic and engaging space for data enthusiasts to collaborate and explore. The subreddit is valuable for finding specialized and niche datasets that may not be easily accessible elsewhere. It fosters a culture of data sharing and discussion, with members offering insights, feedback, and suggestions. For projects requiring unique or specific data, "/r/datasets" is a valuable resource, providing a combination of crowd-sourced data and expert advice. It bridges the gap between data enthusiasts and experts, creating a collaborative environment for data exploration and discovery. Click here to explore datasets. I hope that this list of resources would prove extremely useful for people looking out for doing pet projects or side projects. For the starters, this is definitely a gold mine. Make sure you pick a few side projects and continue to work on them. If you can think of any application of these datasets or know of any popular resources which I have missed, please feel free to share them with me in the comments below. Looking forward to hearing from you. Tired of looking for the ultimate guide to sourcing the perfect datasets for data science projects? Look no further! This blog streamlines your search, guiding you toward the ideal datasets and accelerating your data science journey with ProjectPro's innovative hands-on projects. Langchain Project for Customer Support App in Python Downloadable solution code | Explanatory videos | Tech Support Start Project The State of Data Science 2020 report states that most data scientists spend around 70% of their time analyzing datasets, making data preparation and understanding one of the most significant and time-consuming tasks during a data science project lifecycle. Just as an architect relies on sturdy blueprints to design an extraordinary building, data scientists depend on high-quality datasets to build robust data science solutions.From discovering unique and relevant datasets for visual storytelling to exploring popular ones for streaming data, this blog is your go-to resource to finding the best datasets for your data science projects. Table of Contents Data is the lifeblood of data science projects. Without high-quality data, it's impossible to build accurate and reliable models. Therefore, finding suitable data sets for data science projects is crucial for successful data analysis and insights. So, where can you find good datasets for your data science projects? There are various sources where you can discover diverse and reliable datasets- Vipul Patel, Chief AI Officer at Elevate AI, shares his opinion on finding suitable datasets for your projects in online data repositories- A few online repositories like Google Dataset Search, Kaggle, and the UCI Machine Learning Repository offer datasets to analyze for projects around data science, ML, data visualization, etc. Source- Wikipedia With over 16 million users across 190 countries, Kaggle is a renowned online community for data scientists and ML enthusiasts. It hosts over 273k public datasets, 915k public notebooks, and 2,000 pre-trained, ready-to-deploy ML models, ranging from healthcare and finance to social sciences and natural language processing. Kaggle also hosts data science and ML competitions, providing a platform for data scientists to test their skills and collaborate with others. Source- Google Founded in 1987, the UCI Machine Learning Repository is another famous repository for ML data sets. It includes a carefully curated collection of more than 653 public datasets that have been well-documented and categorized for various research and educational uses. As a foundational resource for ML research, the UCI Machine Learning Repository offers sample data sets for algorithm development and evaluation. Abhishek Kumar Annamraju, Product Manager and Technical Consultant at Flutura (a Part of Accenture), lists Google Datasets as a valuable dataset for ML projects- Source- Google Dataset Search Google Dataset Search is a widely used data set search engine that lets users find datasets by topic, keyword, or license type. It indexes over 25 million datasets from various sources, including government websites, academic institutions, and public repositories. By offering a centralized platform for exploration and discovery, Google Dataset Search streamlines the process of locating datasets. Many governments worldwide are committed to open data. This means they are making their data publicly available for free. This can be a great source of data sets for data science projects. Here are a few examples of government databases and open data initiatives- One of the largest sources of development data is the World Bank Open Data website, which spans a wide range of topics like infrastructure, poverty, education, and health. The World Bank database is an all-inclusive resource for studying global development trends and patterns. The World Bank Open Data website gives users access to over 200,000 publicly available documents and reports, 2,000 databases, and over 14 million indicators. Source- Wikipedia The US Census Bureau is the primary data source about the people and economy of the United States. It covers a dozen topics across 130+ surveys and programs, such as the decennial census, the American Community Survey, and other surveys, to collect more than 2.5 million tabular data, including raw public data, maps, profiles, and more on a wide range of topics, such as population, housing, income, and education. Source- EU Data Portal The European Union Open Data Portal provides access to public data from the European Union and its member states. It hosts over a million datasets, 181 catalogs, 1478 news pieces, and 209 data stories from 36 countries, covering various topics, including agriculture, environment, transportation, and public affairs. Let us dive into the list of free data sets you can use for data visualization, data analysis and processing, data cleaning, machine learning, and other data science tasks. 20 Free Datasets For Data Science Projects This section will walk you through a curated collection of 25 free data sets that will serve as your data science compass. These data sets, carefully selected from various domains, will equip you with the raw materials to develop compelling data visualizations, streamline your data processing tasks, tackle data cleaning challenges, build robust ML models, and leverage the power of real-time data streams. I think that they are fantastic. I attended Yale and Stanford and have worked at Honeywell,Oracle, and Arthur Andersen(Accenture) in the US. I have taken Big Data and Hadoop,NoSQL, Spark, Hadoop Admin, Hadoop projects. I have been happy with every project. They have really brought me into the... Ray han Tech Leader | Stanford / Yale University Having worked in the field of Data Science, I wanted to explore how I can implement projects in other domains, So I thought of connecting with ProjectPro. A project that helped me absorb this topic was "Credit Risk Modelling". To understand other domains, it is important to wear a thinking cap and... Gautam Vermani Data Consultant at Confidential Not sure what you are looking for? View All Projects This section compiles a selection of data sets that will spark your imagination for data visualization and help you create engaging stories with data. These datasets will inspire you to produce data visualizations that effectively communicate the necessary insights- whether your goal is to visualize global trends, analyze social media trends, or dig into the complexities of financial markets. Source- Wikipedia FiveThirtyEight (now known as 538) is a data journalism website that publishes stories on various topics, including sports, politics, and science. They also maintain a repository of open-source data sets on GitHub that they use for their analyses and are available for others to use. These data sets provide valuable insights across several domains, making them ideal resources for data visualization projects. Data science experts often use these interesting data sets to visualize data, such as their election data sets covering historical voting trends, candidate performance, etc. Their sports datasets include player statistics, game outcomes, and team performances and can help create engaging visualizations showcasing trends within the sports world. Source- FiveThirtyEight Datasets Below are some valuable datasets offered by FiveThirtyEight- Source- HDX The Humanitarian Data Exchange (HDX) platform provides access to open-source datasets related to humanitarian crises and development. It offers a vast collection of datasets covering various topics, including natural disasters, food security, health, and displacement. The platform hosts 20,340 datasets from 254 locations combined from 1947 sources. These data sets can be used for several data visualization projects, such as creating maps to show the spatial distribution of humanitarian crises, and creating interactive visualizations to inform humanitarian decision-making. Source- Humanitarian Data Exchange Datasets Sample Datasets by The HDX Below are some valuable datasets offered by the Humanitarian Data Exchange- Source- WHO The World Health Organization (WHO) is a specialized United Nations agency providing global public health leadership. They maintain a vast repository of healthcare data and data on several health-related topics, including data on diseases, health systems, and health outcomes. These data sets can be used for various data visualization projects, such as creating maps to show the geographical distribution of health indicators, visualizing trends in health indicators over time, and creating interactive visualizations to inform public health policies and interventions. Source- World Heath Organization Datasets Sample Datasets by WHO Below are some valuable datasets offered by WHO- Global Health Estimates (Offer the latest available data on causes of mortality and disability worldwide by WHO region and country, age, sex, and income group). Immunization Data and Statistics (Provides data on immunization coverage and trends for various diseases worldwide) Source- NASA The National Aeronautics and Space Administration, or NASA, offers data on space exploration and Earth science. NASA's Open Data Portal provides access to vast data sets covering various topics, including astronomy, planetary science, Earth observations, and aeronautics. You can use these NASA data sets for any typical data visualization project, such as creating maps to show the geographical distribution of data, visualizing trends in climate data over time, and comparing different regions or entities based on their data. Source- NASA Databases Sample Datasets by NASA Below are some valuable datasets offered by NASA- Global Surface Water Datasets (Provide global coverage of surface water extent from 1993 to present, derived from satellite observations) Air Quality Data (Provides air quality data for various locations worldwide, including measurements of PM2.5, ozone, and nitrogen dioxide). Get confident to build end-to-end projects Access to a curated library of 250+ end-to-end industry projects with solution code, videos and tech support. Request a demo Did you know that the top five happiest countries in 2023 were Finland, Denmark, Iceland, Israel, and the Netherlands? Source- World Happiness Report 2023 The World Happiness Report is a landmark survey of the global state of happiness. It is published annually by the Sustainable Development Solutions Network (SDSN) and is based on a study of over 150 countries. The report primarily uses the Gallup World Poll data and ranks countries based on their citizens' reported happiness levels, using factors such as life expectancy, social support, and income. You can use the World Happiness Report data for various data visualization projects, such as creating maps to show the geographical distribution of happiness scores, visualizing trends in happiness scores over time, and comparing different countries or regions based on their happiness scores. Source- World Happiness Report Data Best Datasets For Data Processing Projects Data processing is the backbone of data science, transforming raw data into a clean and organized state ready for analysis. This section showcases a collection of carefully selected data sets that will test your data processing skills. From handling missing values and dealing with outliers to normalizing data and feature engineering, these datasets provide ample opportunities to hone your data-wrangling skills. Source- Amazon Amazon Web Services (AWS) provides a comprehensive collection of publicly accessible datasets through its Open Data Registry. The datasets are available in CSV, JSON, and parquet formats, making them suitable for various data processing tasks. AWS Open Data Registry project datasets are ideal for various data processing projects, including land cover classification, and also support scientific research in fields like climate change and renewable energy development. Source- AWS Open Data Registry Datasets Sample Datasets by AWS Open Data Registry Below are some valuable datasets offered by AWS Open Data Registry- Unlock the ProjectPro Learning Experience for FREE Google Cloud Public Datasets is a repository of over 200 freely accessible datasets from various industries, curated to simplify data processing and analysis. These datasets are hosted on Google Cloud Storage, and you can easily access and process them using Google Cloud Platform (GCP) services like BigQuery and Dataproc. Data science enthusiasts can easily use the Google Cloud public datasets to process data for tasks like image classification and anomaly detection and to extract insights and patterns related to various topics, such as traffic patterns and image content. Source- Google Cloud Public Datasets Sample Datasets by Google Cloud Platform You can download data offered by GCP, such as OpenStreetMap Dataset (Provides worldwide OpenStreetMap data, facilitating geospatial analysis and mapping applications) Taxi Trip Dataset (Provides trip information for taxi rides in various cities, enabling research on urban mobility patterns) The MovieLens dataset is a vast collection of movie ratings from over 1 million users. It includes information on the movies, such as their genre, release year, as well as the ratings and timestamps for each user's ratings. The dataset is a valuable resource for building recommendation systems as it helps train ML models to recommend movies to users based on their past ratings. The MovieLens dataset can be used for several data processing projects, including training recommendation systems to recommend movies to users based on their past ratings, and developing new recommendation algorithms that are more accurate or efficient than existing algorithms. Source- MovieLens Datasets Sample Data by MovieLens Datasets Here are a few sample datasets you can use from MovieLens Datasets- Stack Exchange Data Explorer is a powerful tool that lets you explore and analyze data from across the Stack Exchange network, a collection of over 180 Q&A websites covering various topics. The Data Explorer provides access to a vast repository of rich textual data, including questions, answers, comments, and user information. Stack Exchange datasets are highly useful for various purposes in data processing projects, such as training NLP models for question answering or sentiment analysis, analyzing user behavior, data science community dynamics, and identifying patterns and correlations in data related to programming languages, software development, and other topics. Source- Stack Exchange Data Explorer Datasets Sample Data by Stack Exchange Datasets Here are a few sample datasets you can use from Stack Exchange datasets- Stack Overflow Posts (Provides over 25 million questions, 35 million answers, 90 million comments, and 25K tags from Stack Overflow) Ask Ubuntu Posts (Provides over 412K questions, 520K answers, 1.5 million comments, and 3.2K tags from Ask Ubuntu) Data Processing Projects For Practice Here are some projects you can add to your data science portfolio that use data processing techniques- This section will prepare you to tidy up your data like a pro with our roundup of popular data sets for data-cleaning projects. Let us explore a few widely used datasets specifically chosen to help you practice data-cleaning techniques, such as exploratory data analysis (EDA), ensuring your datasets are error-free and ready for analysis. Source- Academic Torrents Academic Torrents is a platform that provides a curated collection of over 127.15TB of research data from scientific papers across various disciplines. These exciting data sets are available through the BitTorrent protocol, enabling efficient and distributed access to large datasets. The platform is a valuable resource for researchers and data scientists working on various data-cleaning projects. Leveraging these datasets to build a data cleaning project can involve cleaning datasets by performing exploratory data analysis to identify incomplete or inaccurate entries, standardizing data formats to ensure consistency for further analysis, and enhancing the dataset by adding relevant information from external sources. Source- Academic Torrents Sample Data by Academic Torrents Datasets Here are a few sample datasets you can use from Academic Torrents datasets- YASP 3.5 Million Data Dump (Provides an extensive collection of over 3.5 million interesting data-driven articles) OpenAlex Snapshot (Offers metadata for 209M works (journal articles, books, etc.); 2013M disambiguated authors; 124k venues (journals and online repositories); 109k institutions; and 65k Wikidata concepts) Source- Reddit With approximately 1.1 billion users worldwide, Reddit is a popular community-driven online forum and discussion platform. It offers a separate section called the datasets subreddit, or /r/datasets, a valuable resource for data cleaning projects due to its vast collection of real-world data, diverse topics, and active community involvement. The subreddit provides various datasets from various domains and sources, suitable for multiple data-cleaning needs. The datasets are typically open-access and freely available, making them accessible to many users. This further encourages discussion and collaboration among users, facilitating the exchange of data-cleaning techniques and best practices. Source- Reddit Datasets Sample Data by Reddit Datasets Here are a few sample datasets you can use from Reddit datasets- Source- Data.world Data.world (or the 'social network for data people') is a platform that provides access to a vast repository of open data (nearly 129K datasets) from various sources, including multiple US government agencies, academic institutions, and non-profit organizations. The platform offers a user-friendly interface for searching, exploring, and downloading datasets, and it also provides tools for data analysis and cleaning. The datasets are curated by experts and undergo rigorous quality checks, ensuring the reliability of the data. The datasets are freely accessible and maintained by a community of researchers, and include detailed metadata and documentation, providing context and facilitating data understanding. Source- Data.world Datasets Sample Data by Data.world Datasets You can acquire data sets from Data.world platform such as Climate Change Data (Offers data from World Development Indicators and Climate Change Knowledge Portal on climate systems, greenhouse gas emissions, and energy use) UN Human Development Index (Provides data on the United Nations Human Development Index (HDI) for various countries over time for 2022) Another useful dataset is the Billion Word Language Modeling Benchmark Dataset, which can be found on Google Research and is a valuable resource for data cleaning projects in natural language processing (NLP). It provides a massive collection of text data from various sources, including books and news articles. Various use cases of the Billion Word Language Modeling Benchmark Dataset for data cleaning projects include developing data cleaning algorithms on a large corpus of real-world text data, and assessing the impact of data cleaning on downstream NLP tasks, such as sentiment analysis and text summarization. Source- Billion World Language Modeling Benchmark Dataset Data Cleaning Projects For Practice Here are some projects you can add to your data science portfolio that use data-cleaning methods, such as exploratory data analysis and others- Interesting Datasets For Machine Learning Projects In machine learning, datasets are the secret tools that power algorithms, enabling them to discover hidden patterns and make predictions. This section curates a collection of intriguing datasets that will challenge your machine-learning skills and inspire you to implement ML in real-world scenarios. Whether you want to predict customer churn, analyze medical trends, or optimize marketing campaigns, these datasets will provide the perfect platform to hone your machine-learning expertise. Source- Nasdaq Data Link Trusted by over 800K professionals, Nasdaq Data Link is a platform that provides access to over 250 datasets covering economic and financial data, such as historical and real-time stock prices, economic data, and company financials. The platform offers data sets in various formats, including CSV, XML, and JSON, and provides a user-friendly API for accessing data programmatically. You can use the Nasdaq Data Link free datasets for various machine-learning tasks. For instance, you use these datasets to analyze stock price data and indices, predict economic indicators to identify trends and patterns, and develop trading algorithms based on technical analysis and machine learning. Source- Nasdaq Data Link Datasets Sample Data by Nasdaq Data Link Datasets You can find free datasets on the Nasdaq Data Link platform, such as US Federal Reserve Data Releases (Provides official US figures on money supply, government finances, bank assets and debt, exchange rates, etc.) World Bank Data (Provides data from hundreds of countries and regions worldwide, from multiple categories such as finance, economy, climate change, government expenditures, etc.) With a global community of 262K quants, researchers, data scientists, and engineers, QuantConnect provides a platform for developing and deploying trading algorithms and access to a comprehensive collection of financial datasets. These datasets cover various asset classes, including stocks, futures, options, and foreign exchange, ideal for various ML applications in the financial domain. Data scientists can employ QuantConnect datasets for various ML tasks in finance to train ML models to identify profitable trading signals and predict market movements, evaluate trading algorithms using ML techniques, and optimize portfolio allocation using ML models Source- QuantConnect Datasets Sample Data by QuantConnect Datasets You can find free datasets on the Nasdaq Data Link platform, such as The Statlog Shuttle Landing Discovery Dataset is a challenging and valuable dataset for ML projects offered by the UCI Machine Learning Repository, which offers a collection of shuttle flight-related measurements, including altitude, velocity, acceleration, and sensor readings. Data enthusiasts can leverage this dataset to identify anomalous patterns in sensor data to predict potential shuttle landing failures, build predictive models to predict equipment failures and evaluate various ML algorithms for classifying shuttle flights based on their sensor readings. Source- Statlog Shuttle Landing Discovery Dataset Sample Data by Statlog Shuttle Landing Discovery Dataset You can pick sample data from the Statlog Shuttle Landing Discovery dataset, such as Shuttle Test Set (A subset of the data that can be used to evaluate the performance of trained ML models) Shuttle Landing Failure Reason Codes (A table of codes that indicate the specific reasons for shuttle landing failures) The House Prices: Advanced Regression Techniques dataset, hosted on Kaggle, offers detailed information about houses sold in Ames, Iowa, along with their sale prices. This dataset aims to train ML models to accurately predict a house's sale price based on various features, such as its size, location, amenities, and other characteristics. You can use this dataset to develop ML models to predict the houses' sale price of houses based on their features, analyze factors that impact real estate prices in different regions, and build recommendation systems that recommend suitable houses to potential buyers as per their preferences and budget. Source- House Prices Advanced Regression Techniques Dataset Machine Learning Projects For Practice Here are some fascinating projects for you that use ML methods- Bonus: Public Datasets For Data Science Projects As data science projects grow in complexity, the need for real-time data sources has become increasingly important. While traditional data science projects rely on pre-existing datasets, an increasing amount of valuable data online services generate in real-time, such as social media feeds, financial market data, and sensor readings. Here are a few bonus streaming data sources that can be used for your next streaming data project- Source- GitHub GitHub is a valuable resource for real-time data science projects, offering a curated collection of open-source datasets available in streaming or real-time formats. These datasets provide a continuous data stream, enabling researchers and developers to analyze and respond to real-time events. You can use streaming datasets from GitHub to analyze social media streams to gain insights into public opinion on trending topics or any organization, develop real-time models to predict stock prices based on market data or news sentiment, and several other innovative data science projects. Source- GitHub Picture this: a constant flow of live tweets at your fingertips. That's what Twitter's real-time streaming API offers! Twitter offers access to a vast real-time data stream through its Twitter Streaming API. It provides a valuable resource for data science projects that require the latest insights into public sentiment, trending topics, and breaking news. These real-time datasets offer a window into millions of users' collective thoughts and opinions worldwide. You can use real-time Twitter data for various data science tasks, such as analyzing real-time tweets to understand public opinion on trending topics or political discussions or to develop real-time models to predict market trends, identify potential crises, or detect unusual patterns in public sentiment. You can even use this data to analyze real-time user behavior and sentiment to optimize social media campaigns, target advertising, and personalize user experiences. Source- Twitter API Ready to go on a data expedition with Pew Research Center's real-time streaming datasets? From its Political Typology survey data to the News Interest Index data, the Pew Research Center provides access to a collection of real-time datasets through its American Trends Panel (ATP), offering insights into public opinion, political attitudes, and social trends in real-time. These fascinating datasets allow you to analyze sentiments during significant events, track sudden shifts in public opinion, analyze emerging trends in social attitudes, or even forecast cultural movements. With Pew's real-time datasets, you are not just observing trends; you are navigating the currents of societal shifts in real time! Source- Pew Research Center Datasets Bonus Projects For Your Data Science Portfolio Series Here are some exciting projects you can add to your data science portfolio- Explore Interesting Datasets For Data Science Projects With ProjectPro In data science, datasets are the foundation for significant discoveries and innovative solutions. But how can you leverage the power of these datasets to solve real-world business problems? The key lies in understanding how to use these datasets effectively, transforming raw data into actionable insights. This is where ProjectPro comes in, offering a vast collection of data science projects designed to provide hands-on experience in implementing datasets and building effective data solutions. Through these projects, aspiring data scientists can explore data in real-world scenarios, learn how to tackle challenges, and gain valuable insights from various datasets. Whether analyzing customer behavior, predicting market trends, or optimizing operational efficiency, the ProjectPro repository provides the perfect platform to hone data science skills and gain invaluable experience. So, jumpstart your data science journey with ProjectPro, where datasets are not just numbers but tools to transform businesses and shape the future. FAQs on Datasets For Data Science Projects What are some factors to consider when choosing a dataset for data science projects? When choosing a data set for a data science project, it is essential to consider the following factors- Relevance- The data set should be relevant to the problem you are trying to solve. Quality- The data set should be of high quality, with minimal errors and missing values. Size- The data set should be large enough to train your ML models or perform your analysis. Accessibility- The data set should be easy to access and download. What are the different types of datasets for data science projects? Data science projects employ various types of datasets, including- Structured data- Organized data stored in tables, spreadsheets, or databases. Unstructured data- Text, images, audio, and video that lack a predefined format. Time-series data- Data collected over time, such as stock prices or sensor readings. Real-time data- Data that is continuously generated and streamed, such as social media feeds or financial market data.