Click to prove you're human

Example data sets

Access to quality data is a vital component of informed decision-making across various sectors like business, academia, healthcare, and government. Yet, gathering reliable datasets can be time-consuming and costly. Fortunately, there are numerous free public datasets available that can aid analysis and offer valuable insights. Let's explore some notable sources categorized by theme: Climate and Environmental Datasets. The National Oceanic and Atmospheric Administration (NOAA) provides extensive climate data through NOAA Climate Data Online at . Similarly, NASA Earth Science offers datasets related to Earth sciences including climate and environmental information at . Next, we have Government and Public Datasets that cover a wide range of topics from demographics to public health. The U.S. government's data portal, Data.gov (, houses over 200,000 datasets across various subjects. The European Union's Open Data Portal at also offers valuable insights into government operations and societal trends. Economic and Financial Datasets provide critical information on global economic indicators, financial markets, and development statistics. World Bank Open Data (is a comprehensive resource for global development data, while the International Monetary Fund's IMF Data (offers insights into lending, exchange rates, and other economic and financial indicators. Health-related datasets contain crucial information on public health including disease prevalence, healthcare infrastructure, and health outcomes. The Global Health Observatory at provides data from the World Health Organization on global health statistics. Additionally, the Centers for Disease Control and Prevention (CDC) offers datasets on public health through its data portal at . Lastly, astronomy and space-related datasets offer insights into celestial objects, planetary systems, and cosmic phenomena. These resources facilitate research in various fields of study. NASA Exoplanet Archive: A Hub for Exoplanet Research The NASA Exoplanet Archive provides valuable insights into the discovery and study of exoplanets, contributing significantly to our understanding of the cosmos. Given article text here Looking at the list of provided datasets, it includes product purchasing analysis tools such as Lending Club Loan Data, Instacart Market Basket Analysis, Avito Context Ad Clicks Dataset and Outbrain Click Prediction Dataset. These tools offer valuable resources for data scientists to analyze various financial transactions like loan defaults, consumer purchases and ad clicks. The datasets themselves cover a wide range of domains including finance, consumer behavior and marketing. Some notable examples include Lending Club's massive dataset on loan defaults, Instacart's large grocery dataset which allows for analysis of consumer purchasing patterns, and Avito Context Ad Clicks Dataset, which provides a rich source of data for ad click prediction. Another useful resource is the 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. Across numerous publisher sites over a two-week period in 2016, data was collected, including from companies like Outbrain that feature sponsored content articles at the bottom of websites. According to Eddy, much of practical data science involves predicting ad clicks. The Coffee Reviews Dataset is another valuable resource, compiling global coffee reviews from 2017 to 2022 based on factors such as blend name, roast type, price, and geographical origin, with both original and simplified versions available for analysis. Additionally, the Electric Vehicle Population Data dataset provided by the State of Washington offers insights into registered battery electric vehicles and plug-in hybrid electric vehicles, including details like vehicle identification number, county and city of registration, make and model, and electric range. Image datasets, such as ImageNet with over 14 million images, are crucial for training machine learning models in image recognition and classification, organized by the WordNet hierarchy with thousands of synonym sets. The MNIST Database of Handwritten Digits is a foundational dataset for image classification, featuring digits 0 through 9 in various handwriting styles, ideal for beginners. Eddy notes that basic classification is a simple yet effective starting point for neural network image classification. The Dogs vs. Cats dataset is also recommended for image classification, striking a balance between simplicity and complexity, making it a favorite among data scientists. These datasets are essential for advancing research in computer vision and deep learning, providing valuable resources for training and testing machine learning models. Looking for a beginner's guide in deep learning with Python? Eddy suggests starting with a guided approach to build a strong foundation. Exploring simple image classification problems can help develop skills for tackling more complex issues, which often involve similar work. Beginning with one or two simpler datasets can be beneficial for exploring various standard image problem types. Text mining and text analysis are used to examine patterns in unstructured data, including sentiment analysis, topic modeling, named entity recognition, and natural language processing (NLP). The Large Movie Review Dataset is a 2017 cache of IMDB reviews with 50,000 training samples and 25,000 testing samples. It's useful for sharpening sentiment analysis skills. A Twitter and Reddit Sentimental Analysis Dataset offers an opportunity to pair your project with big social platforms like X and Reddit. This dataset contains over 160,000 posts and 37,000 comments. The Stack Exchange API Dataset gives a glimpse into the ecosystem of Q&A sites, providing opportunities for topic modeling, query writing, and text preprocessing. Amazon Product Reviews Dataset is a sentiment analysis-friendly set that Iqbal recommends, particularly for advanced data scientists in marketing. It contains over 142.8 million reviews. The dataset contains details from user activity between 1996 and 2014, making it an ideal fit for those looking to incorporate sentiment analysis into recommender systems. Time series data refers to information collected over a specific interval of time, including historical or real-time data points. This type of data is used in time series analysis and forecasting, which detect patterns and predict when specific changes may occur over time. For newcomers to time series analysis, Eddy stresses two key criteria for selecting datasets: ensure the time interval is fixed and watch for clear seasonal patterns with logical effects. Some suitable time series datasets that fit these criteria include: 14. Hourly Energy Consumption Dataset: A dataset featuring over 10 years of hourly energy consumption data in eastern U.S. states, provided by PJM Interconnection. 15. International Greenhouse Gas Emissions Dataset: A dataset covering global greenhouse gas emission levels from 1990 to 2017, provided by the United Nations. 16. Individual Household Electric Power Consumption Dataset: Measurements of electric power consumption for one household over a four-year period, with one-minute sampling rates. 17. Web Traffic Time Series Forecasting Dataset: Data containing traffic information for 145,000 Wikipedia articles, focusing on predicting future web traffic trends. 18. Uber Pickups in New York City Dataset: Date, time and location data for over 20 million Uber and for-hire vehicle trips in the NYC area from April to September 2014. Citi Bike System Data Dataset: Insights into Bike Rides in NYC The Citi Bike System Data set provides a comprehensive view of where, when, and how far Citi Bike users ride in New York City. It includes detailed travel information such as bike ride ID, start and end times, start and end station IDs, and geographical location. Here is the list of datasets and projects, reorganized for better readability: **Biology:** * Expression Omnibus (GEO) * Biology Gene Ontology (GO) * Global Biotic Interactions (GloBI) * Harvard Medical School (HMS) LINCS Project * Human Genome Diversity Project – Stanford * Human Microbiome Project (HMP) * ICOS PSP Benchmark * Imported bats * Invasive species * Journal of Cell Biology DataViewer * KEGG * Mammal data * MIT Cancer Genomics Data * National Centers for Environmental Information * National Oceanic and Atmospheric Administration Fisheries * NCBI Proteins * NCBI Taxonomy * NCI Genomic Data Commons * OpenSNP genotypes data * Palmer Penguins * Pathguid Protein Interactions Catalog * Penguins * Protein Data Bank * Psychiatric Genomics Consortium * PubChem Project * Rfam * Sanger Catalogue of Somatic Mutations in Cancer (COSMIC) * Sanger Genomics of Drug Sensitivity in Cancer Project (GDSC) * Sequence Read Archive (SRA) * Shrinking salmon * Stowers Institute Original Data Repository * Systems Science of Biological Dynamics (SSBD) Database * The Cancer Genome Atlas (TCGA) * The Catalogue of Life * The Personal Genome Project * UCSC Public Data * UniGene * Universal Protein Resource (UnitProt) * Zoo animal lifespans **Climate and Weather:** * Actuaries Climate Index * Australian Weather * Aviation Weather Center * Canadian Meteorological Centre * Charting The Global Climate Change News Narrative 2009-2020 * Climate Data from UEA * Dutch Weather * European Climate Assessment & Dataset * German Climate Data Center * Global Climate Data Since 1929 * NASA Global Imagery Browse Services * NOAA Bering Sea Climate * NOAA Climate Datasets * NOAA Real-time Weather Models * NOAA SURFRAD Meteorology and Radiation Datasets * UEA Climatic Research Unit * Wahington Post Climate Change * WorldClim – Global Climate Data * WU Historical Weather Worldwide **Complex Networks:** * AMiner Citation Network Dataset * Community Resource for Archiving Wireless Data * CrossRef DOI URLs * DBLP Citation dataset * DIMACS Road Networks Collection * NBER Patent Citations * Network Repository * Small Network Data * Stanford GraphBase * Stanford Large Network Dataset Collection * The Laboratory for Web Algorithmics (UNIMI) * UCI Network Data Repository * UFL sparse matrix collection * 3.5B Web Pages from CommonCrawl 2012 * Computer Networks * CAIDA Internet Datasets * ClueWeb09 – 1B web pages * ClueWeb12 – 733M web pages Global Network and Data Repository Overview #### Web data is a crucial component of computer networks, encompassing various aspects like internet-wide scan data repository and OONI: Open Observatory of Network Interference. This dataset is utilized by companies such as Rapid7 Sonar for internet scans. Furthermore, the Peer-to-Peer Trace Archive offers insights into computer networks. Additionally, 38-Cloud is a platform that detects cloud-based phenomena. African agricultural survey datasets are used to analyze air quality and other environmental factors like Earth's climate system. The Alabama Real-Time Coastal Observing System facilitates monitoring of coastal areas. UCSD Network Telescope provides data on IPv4 /8 net while the Complete Plants Checklist offers insights into global vegetation. NASA's EOSDIS – Earth observing system data, Global dataset of historical crop yields, and Global Wind Atlas are other notable projects focused on Earth sciences. Other notable datasets include IceCube - South Pole Neutrino Observatory, Integrated Marine Observing System (IMOS), Ligo Open Science Center (LOSC), Marinexplore – Open Oceanographic Data, NASA Earth Science, National Estuarine Research Reserves System-Wide Monitoring Program and NSSDC (NASA) data of 550 space spacecraft. Economically relevant datasets are the Optimized Soil Adjusted Vegetation Index, Sloan Digital Sky Survey (SDSS), Smithsonian Institution Global Volcano and Eruption Database, US forest tree distributions, USGS Earthquake Archives, World glaciers. Additionally, there are datasets related to academia like Academic parental leave policies Economics, Aggregated smartphone movements Economics. Here is the rewritten text: Rooftop water tanks, grants for science and economics, sidewalk grates, space dollars, and more - a vast array of datasets related to economics, including The Atlas of Economic Complexity, The Big Mac Index, The Center for International Data Economics, and many others. This collection also includes data on education, such as College Scorecard Data, Education data, unified, and Wikipedia revisions for editors. Additionally, it contains energy-related datasets like BLUEd, COMBED, DBFC, DEL, ECO, EIA, Electricity prices, Electricity utilities, and more. The dataset also features entertainment data, including Anthony Bourdain's travels, BFI film industry statistics, Billboard music hits and lyrics, Boy bands, Breaking Bad data, Classical music data, Crossword data, Drama, Friends TV show analysis, Hunger Games survival, Indian movie theatres, Movie dialog, Movie scripts and genders, Music artists, NYC film and TV permits, Pinball, Ramen ratings, Rotten Tomatoes reviews, Star Trek data, Studio Ghibli data, Talk radio transcripts, Tarantino movie data, Billionaire list Finance. Geospatial and government data from various sources including GitHub, GeoNames, GADM, and more, covering topics such as cities, countries, national parks, weather service data, openaddresses, and other geographical features. Also includes datasets on administrative areas, natural environment, climate change, urban infrastructure, noise pollution, timezones, and population demographics in regions like USA, UK, Canada, Australia, Europe, and others. Government agencies worldwide, including those in Finland, Hong Kong, India, Indonesia, and several countries in North America, have reported various data and policy initiatives. These include international labor treaties, licensed firearms dealerships, and local municipalities' governance structures such as Istanbul's government. Government bodies also focus on public health policies, pension plans, and open data portals like Indonesia's Data Portal Government. The text appears to be a list of various government and health-related datasets, reports, and databases from the United States and other countries. These resources cover topics such as US government data, census information, healthcare statistics, and COVID-19 research. Some examples include: * Government reports from the US House of Representatives and state governments * Healthcare datasets from the CDC, Census Bureau, and National Center for Education Statistics * COVID-19 data repositories from Johns Hopkins CSSE Health and other organizations * Healthcare research papers on topics like coronavirus, cancer, and medications * Demographic databases from Gapminder World and census information These resources are likely intended to support research, policy-making, or public health initiatives. The list spans various domains, including government, healthcare, education, and demographics. Given text: paraphrase this text: Severe workplace injuries Health Social assistance programs Health Study Forrest Health Subnational COVID-19 case counts Health The Cancer Genome Atlas project (TCGA) Health The Cancer Imaging Archive (TCIA) Health The COVID Tracking Project Health Two decades of tobacco (and e-cigarette) laws Health US dairy information Health US emergency room visits Health US vaccination rates in adults Health US workplace safety Health World Health Organization Global Health Observatory Health Yahoo Knowledge Graph COVID-19 Datasets Health Zika virus data Health Canada Science and Technology Museums Corporation’s Open Data Museums Cooper-Hewitt’s Collection Database Museums Metropolitan Museum of Art Collection API Museums Minneapolis Institute of Arts Museums Museum of Modern Art collection Museums Natural History Museum (London) Data Portal Museums Rijksmuseum Historical Art Collection Museums Tate Collection metadata Museums The Getty vocabularies Museums Automatic Keyphrase Extraction Natural Language Blizzard Challenge Speech Natural Language DBpedia Natural Language Dirty Words Natural Language Flickr Personal Taxonomies Natural Language German Political Speeches Corpus Natural Language Google Books Ngrams (2.2TB) Natural Language Google MC-AFP Natural Language Google Web 5gram (1TB 2006) Natural Language Hansards text chunks of Canadian Parliament Natural Language LJ Speech Natural Language Machine Comprehension Test (MCTest) of text from Microsoft Research Natural Language Machine Translation of European languages Natural Language Making Sense of Microposts 2016 Natural Language Microsoft MAchine Reading COmprehension Dataset (or MS MARCO) Natural Language Multi-Domain Sentiment Dataset (version 2.0) Natural Language Noisy speech database for training speech enhancement algorithms and TTS Natural Language Open Multilingual Wordnet Natural Language SaudiNewsNet Collection of Saudi Newspaper Articles (Arabic 30K articles) Natural Language SMS Spam Collection in English Natural Language Stanford Question Answering Dataset (SQuAD) Natural Language The Big Bad NLP Database Natural Language Universal Dependencies Natural Language USENET postings corpus of 2005~2011 Natural Language Webhose – News/Blogs in multiple languages Natural Language Wikidata – Wikipedia databases Natural Language A range of miscellaneous data Other Adult height over time Other Advice sought Other Broadband access in the US Other Catholic women Other Confidence Other Dating survey Other Emotion and word associations Other Graphic Design data Other Happy moments Other Jeans pockets Other Joke database Other Makeup shades Other Photographer biographies Other Random card choices Other Various datasets from public sources, including Statista, Washington Post, SOCR, UFO Reports, Wikileaks, Yahoo Webscope, Zenodo, and others. These datasets cover a wide range of topics such as social networks, email communications, online reputation management, software development projects, government-sponsored cyberattacks, and more. Additionally, there are datasets related to sports, including baseball ballparks, college football songs, cricket matches, golfing discs, and soccer/football play-by-play data. Time series datasets include hard drive failure rates, heart rate, MIT dataset, and UC Riverside dataset. Commercial Vehicle Safety: Transport Data Collection - A Comprehensive List This article provides an extensive compilation of transportation-related datasets from various sources, including government agencies, research institutions, and private companies. The data sets cover a range of topics, such as: * Commercial vehicle safety * Traffic information * GPS trajectories * Autonomous driving * Parking tickets * Taxi trip data * Flight schedules * Urban traffic * Public transportation * Vehicle management The list also includes datasets from international organizations and countries, providing insights into global transportation trends. The article concludes by inviting readers to contribute to the list or suggest improvements. Grab your essential SQL reference materials now and discover key functionalities from various database providers. (ADD SPELLING ERRORS - SE)