



Datasets for Python EDA Project

Topic 1: Ground Water - Water Quality Regulatory Information in California USA

Source:

<https://catalog.data.gov/dataset/ground-water-water-quality-regulatory-information>

Dataset here:

<https://drive.google.com/drive/folders/19n6FbRtCkYHdaSIFm6q-HkxX1yWiM-8x?usp=sharing>

Approx. Size: 75.000 rows

Topic 2: U.S. Chronic Disease Indicators (CDI), 2023 Release

Source:

https://data.cdc.gov/Chronic-Disease-Indicators/U-S-Chronic-Disease-Indicators-CDI-2023-Release/g4ie-h725/about_data

Approx. Size: 1185676 rows

Topic 3: U.S. National Outbreak Reporting System (NORS)

Source:

https://data.cdc.gov/Foodborne-Waterborne-and-Related-Diseases/NORS/5xkq-dg7x/about_data

Approx. Size: 62415 rows

Topic 4: Real drinking-water quality in Germany

Source:

Data: <https://zenodo.org/records/3884398>

Data dictionary and additional context:

<https://www.spotseven.de/wp-content/uploads/2018/03/rulesGeccolc2018.pdf>

Example of questions:

Is there a time of the day when the quality of the water is the worst?

Is there a day of the week when the quality of the water is the worst?

Is the water quality better during the weekends or during the week?

Approx. Size: 140.000 rows (Time series with secondly data for 4 months)



Datasets for Python EDA Project

Topic 5: Chemical in Cosmetics in the state of California

Source: <https://data.ca.gov/dataset/chemicals-in-cosmetics>

Dataset:

<https://drive.google.com/drive/folders/1SZFPh-YpraoJRAKclFPUYmNTL450e8Y2?usp=sharing>

Example of questions:

Which category contains most of the products that have been discontinued?

Which chemicals are more present per product category?

Which companies have discontinued products and which ones haven't at all?

Approx. Size: 114.000 rows with product and ingredients information

Topic 6: YouTube Trending Videos

- Example Questions:
 - What types of content trend most often?
 - Which categories receive the most engagement?
- [Link to Dataset](#)

Topic 7: 2021 Kaggle Machine Learning & Data Science Survey

Overview: The data comes from a survey conducted by Kaggle, covering demographic questions (e.g., gender, education level) and inquiries about programming languages, tools, and machine learning algorithms used.

- Example Questions:
 - What is the most popular programming language in 2021?
 - What are the most commonly used cloud platforms in 2021?
 - What is the most common framework in 2021?
- [Link to Dataset](#)

Topic 8: Medical Appointment No Shows

Overview: This dataset collects information from 100,000 medical appointments in Brazil, focusing on whether patients show up for their appointments.

- Example Questions:
 - What features are important for predicting if a patient will attend their scheduled appointment?
- [Link to Dataset](#)



Datasets for Python EDA Project

Topic 10: Nice Ride Minnesota 2017

Overview: This dataset contains data from Nice Ride MN for the year 2017, including daily weather data from NOAA.

- Example Questions:
 - Is there a preference for taking rides in particular weather conditions?
- [Link to Dataset](#)

Topic 11: Historical Rainfall data of Bangladesh

Overview: Historical Rainfall data of Bangladesh from 1970-2016

- Example Questions:

What are the best locations to open a store for rainproof products?
Which is the location with the lowest average yearly rainfall?
Which quarter has the lowest rainfall in that location?
- Links to Dataset:
 - <https://www.openml.org/api/v1/json/data/41539>
 - https://api.openml.org/data/v1/download/21230844/rainfall_bangladesh.arff

Topic 12:

Overview: Flight Prices

Example Questions:

Does price vary with Airlines?

Does ticket price change based on the departure time and arrival time?

- [Flight Price](#)

Topic 13:

Overview: Real Estate

- [World's Real Estate Data\(147k\)](#)

Topic 14:

Overview: Cars

- [Used Cars Dataset](#)



Datasets for Python EDA Project

- [Car Features and MSRP](#)

Topic 15:

Overview: Job Reviews

- [Glassdoor Job Reviews](#)

Topic 16:

Overview: Developer Survey

- [Stack Overflow Insights - Developer Hiring, Marketing, and User Research](#)

Topic 17:

Overview: MOOC

- [MOOC Dataset](#)

Topic 18:

Overview: Real Estate

- [Saudi Arabia Real Estate \(AQAR\)](#)