Book

of

Resources

1. Kaggle

- *Description:* A platform for data science competitions that also hosts a vast repository of datasets in various formats.
- *Link:* Kaggle Datasets

2. UCI Machine Learning Repository

- *Description:* A well-known repository with a wide variety of datasets for machine learning and EDA.
- *Link:* UCI Machine Learning Repository

3. Google Dataset Search

- *Description:* A search engine specifically for datasets across the web, including those from various data repositories.
- *Link:* Google Dataset Search

4. AWS Public Datasets

- *Description:* A collection of publicly available datasets hosted on Amazon Web Services for research and development.
- *Link:* AWS Public Datasets

5. Data.gov

- *Description:* The U.S. government's open data portal, providing datasets across various sectors including health, education, and transportation.
- *Link:* Data.gov

6. The World Bank Open Data

- *Description:* Offers free and open access to global development data, including economic, social, and environmental indicators.
- *Link:* The World Bank Open Data

7. Quandl

- *Description:* A platform for financial, economic, and alternative datasets. Many datasets are free, while others require a subscription.
- *Link:* Quandl

8. FiveThirtyEight

- *Description:* Datasets used in FiveThirtyEight's data-driven journalism, covering topics like politics, sports, and culture.
- *Link:* FiveThirtyEight Datasets

9. Stanford Large Network Dataset Collection

- *Description:* A collection of large network datasets from Stanford, suitable for social network analysis and EDA.
- *Link:* Stanford Large Network Dataset Collection

10. OpenStreetMap

- *Description:* Provides free geographic data and mapping to the world, useful for spatial data analysis and EDA.
- *Link:* OpenStreetMap

11. IMDB Datasets

- *Description:* A collection of datasets related to movies, TV shows, and celebrities, useful for EDA in entertainment analytics.
- *Link:* IMDB Datasets

12. European Union Open Data Portal

- *Description:* Offers access to a wide range of data from European Union institutions, agencies, and other bodies.
- *Link:* EU Open Data Portal

13. Global Health Observatory Data Repository

- *Description:* Managed by the World Health Organization, providing datasets on global health indicators.
- *Link:* WHO Global Health Observatory

14. Yelp Open Dataset

- *Description:* A dataset from Yelp, including business details, reviews, and user data, ideal for EDA in business analytics.
- *Link:* Yelp Open Dataset

15. Open Data by Microsoft

- *Description:* A collection of datasets made available by Microsoft for research and analysis.
- *Link:* Microsoft Open Data

16. Azure Open Datasets

- **Link:** Azure Open Datasets
- Microsoft Azure offers a variety of curated datasets that can be used for machine learning and AI models.

17. World Bank Open Data

- **Link:** World Bank Open Data
- The World Bank provides free and open access to global development data, including economic indicators, demographics, and environmental data.

18. Open Data Commons

- **Link:** Open Data Commons
- Open Data Commons offers a collection of datasets licensed under open data licenses, making them freely available for use.

Google BigQuery Public Datasets

- **Link:** Google BigQuery Public Datasets
- Google BigQuery offers a collection of public datasets that can be queried directly using SQL, covering topics like genomics, cryptocurrency, and weather.

Zenodo

- **Link:** Zenodo
- Zenodo is an open-access repository that allows researchers to share datasets, code, and publications. It hosts a wide variety of datasets across multiple disciplines.

OpenML

- **Link:** OpenML
- OpenML is an online platform that provides datasets for machine learning, along with tools for exploring, analyzing, and benchmarking models.

DataHub

- **Link:** DataHub
- DataHub offers a range of datasets across various domains, including economics, healthcare, and geography. The platform also supports data sharing and collaboration.

Harvard Dataverse

- **Link:** Harvard Dataverse
- Harvard Dataverse is a repository where researchers can share, store, and cite research data. It includes a wide variety of datasets across different fields.

Yelp Open Dataset

- **Link:** Yelp Open Dataset
- Yelp provides a public dataset containing data on businesses, reviews, and user information, useful for EDA, sentiment analysis, and recommendation systems.

IMDB Datasets

- **Link:** IMDB Datasets
- IMDB offers datasets that include information on movies, TV shows, and video games, which can be used for analysis and model training.

Open Science Data Cloud (OSDC)

- **Link:** Open Science Data Cloud
- OSDC provides access to a variety of scientific datasets, especially in the fields of biology, genomics, and environmental science.

re3data.org

- **Link:** re3data.org
- re3data.org is a global registry of research data repositories that provides access to a wide range of datasets across various disciplines.

UC Irvine Time Series Data Library

- **Link:** UC Irvine Time Series Data Library
- This library offers time series datasets that are useful for pattern recognition, forecasting, and other time-based analyses.

The Humanitarian Data Exchange (HDX)

- **Link:** Humanitarian Data Exchange (HDX)
- HDX provides datasets related to humanitarian crises, including data on refugees, disasters, and public health.

Global Health Observatory (GHO) Data

- **Link:** Global Health Observatory Data
- The World Health Organization's GHO offers data on global health indicators, including statistics on diseases, health systems, and risk factors.

Tools for Exploratory Data Analysis

---

1. Pandas Profiling

- **Purpose: Automated EDA.**
- **Features: Generates comprehensive reports with insights into data types, missing values, distributions, and correlations.**
- **Link: Pandas Profiling**

2. Sweetviz

- **Purpose: Quick EDA with visualizations.**
- **Features: Produces detailed, interactive HTML reports that visualize various aspects of your dataset.**
- **Link: Sweetviz**

3. D-Tale

- **Purpose: Interactive data analysis and visualization.**
- **Features: Combines the functionality of a Pandas dataframe with interactive, web-based visualization tools.**
- **Link: D-Tale**

4. Plotly

- **Purpose: Interactive visualizations.**
- **Features: Enables the creation of highly customizable and interactive graphs, supporting both Python and JavaScript.**
- **Link: Plotly**

5. Seaborn

- **Purpose: Statistical data visualization.**
- **Features: Built on top of Matplotlib, Seaborn provides high-level functions to easily create complex statistical plots.**
- **Link: Seaborn**

6. Altair

- **Purpose: Declarative visualization.**
- **Features: Designed for statistical visualization with a straightforward API, allowing for the easy creation of complex visualizations.**
- **Link: Altair**

7. Bokeh

- **Purpose: Interactive visualization.**
- **Features: Facilitates the creation of interactive and real-time visualizations directly in web browsers.**
- **Link: Bokeh**

8. Tableau

- **Purpose: Business intelligence and analytics.**
- **Features: Features a drag-and-drop interface, supports numerous data sources, and is ideal for creating dashboards and complex visualizations.**
- **Link: Tableau**

9. Power BI

- **Purpose: Business analytics.**
- **Features: Provides interactive dashboards, integrates with various data sources, and works seamlessly with Microsoft products.**
- **Link: Power BI**

10. Superset

- **Purpose: Data exploration and visualization.**
- **Features: An open-source, web-based tool for visualizing data and creating dashboards.**
- **Link: Apache Superset**

11. Google Data Studio

- **Purpose: Data visualization and reporting.**
- **Features: A free tool that integrates well with other Google services, perfect for building interactive dashboards.**
- **Link: Google Data Studio**

12. Looker

- **Purpose: Data exploration and visualization.**
- **Features: Provides powerful data exploration tools and integrates smoothly with cloud-based data sources.**
- **Link: Looker**

13. Holoviews

- **Purpose: Simplified data visualization.**
- **Features: Built on top of Bokeh and Matplotlib, making it easier to create complex visualizations.**
- **Link: Holoviews**

14. Yellowbrick

- **Purpose: Visual analysis and diagnostic visualization.**
- **Features: Works with Scikit-learn to focus on machine learning model diagnostics through visualizations.**
- **Link: Yellowbrick**

15. Lux

- **Purpose: Smart data discovery.**
- **Features: Extends Pandas to automatically generate visualizations that highlight key aspects of your data.**
- **Link: Lux**

16. Apache Superset

- **Purpose:** Data exploration and visualization.
- **Key Features:**
  - Open-source, customizable, and scalable.
  - Offers a wide range of charts and data exploration options.
  - Supports SQL for advanced querying and data exploration.
- **Use Case:** Scalable data exploration and dashboard creation in an open-source environment.

17. Alteryx

- **Purpose:** Data preparation and advanced analytics.
- **Key Features:**
  - Facilitates data blending, preparation, and advanced analytics without coding.
  - Offers a visual workflow interface for building complex data transformations and predictive models.
  - Integrates with multiple data sources and supports spatial and predictive analytics.

- **Use Case:** End-to-end data preparation, analysis, and predictive modeling.

18. RapidMiner

- **Purpose:** Data science and machine learning platform.
- **Key Features:**
    - Provides a visual workflow designer for data preparation, modeling, and deployment.
    - Supports automated machine learning and feature engineering.
    - Offers detailed analysis and explanations of model results.
- **Use Case:** Building and deploying machine learning models with a focus on interpretability.

19. H2O.ai

- **Purpose:** Scalable machine learning and AI.
- **Key Features:**
    - An open-source platform for building and deploying machine learning models.
    - Supports automated machine learning (AutoML) for rapid model development.
    - Integrates with Python, R, and other languages for seamless data science workflows.
- **Use Case:** High-performance machine learning and AI for large-scale data analysis.

20. Kedro

- **Purpose:** Pipeline development and EDA.
- **Key Features:**
    - Helps build reproducible, maintainable data science code with a focus on pipeline development.
    - Integrates seamlessly with Jupyter notebooks for conducting EDA.
    - Allows version control and tracking of data analysis processes.
- **Use Case:** Organizing complex EDA and data science projects into modular, maintainable pipelines.

21. Great Expectations

- **Purpose:** Data validation and profiling.
- **Key Features:**
    - Provides tools to validate, document, and profile your data in a systematic way.
    - Automatically creates "expectations" (assertions about data) to check for data quality.
    - Generates reports that can be integrated into your EDA process.
- **Use Case:** Ensuring data quality and consistency during EDA.

22. Dataprep

- **Purpose:** Simple, fast EDA and data preparation.
- **Key Features:**
    - Enables quick data cleaning, summarization, and visualization.
    - Offers a user-friendly API that is highly compatible with Pandas.
    - Automatically generates summary statistics and visualizations for your data.
- **Use Case:** Streamlining the process of EDA and data cleaning with minimal code.

23. Visidata

- **Purpose:** Terminal-based data exploration.
- **Key Features:**
    - Provides a fast, terminal-based interface for exploring and analyzing datasets.
    - Supports a variety of data formats, including CSV, Excel, and SQL databases.
    - Allows for quick slicing, dicing, and summarizing of data directly from the terminal.
- **Use Case:** Efficient data exploration for those who prefer command-line interfaces.

24. PandasGUI

- **Purpose:** Interactive data analysis with Pandas.
- **Key Features:**
    - Provides a graphical user interface (GUI) for Pandas dataframes.
    - Offers tools for filtering, sorting, and visualizing data interactively.
    - Supports real-time interaction with data without needing to write code.
- **Use Case:** Simplifying EDA for those who prefer a GUI approach to working with Pandas.

25. Autoviz

- **Purpose:** Automated visualization and EDA.
- **Key Features:**
    - Automatically generates visualizations and summary statistics for your data.
    - Detects relationships, correlations, and distributions without manual intervention.
    - Supports a wide range of data formats, including CSV and Excel.
- **Use Case:** Quickly generating insightful visualizations with minimal manual effort.

26. Exploratory

- **Purpose:** Data analysis and visualization.
- **Key Features:**
    - Offers a GUI-based tool for performing EDA, visualization, and statistical modeling.
    - Supports a wide range of statistical and machine learning methods.
    - Integrates with R and provides features for reproducible analysis.
- **Use Case:** Comprehensive EDA and statistical analysis for users who prefer a GUI-based environment.

27. Databricks Notebooks

- **Purpose:** Collaborative data analysis and EDA.
- **Key Features:**
    - Offers a cloud-based notebook environment with strong support for big data and Apache Spark.
    - Provides collaborative features for real-time teamwork on EDA and data analysis.
    - Supports rich visualizations and integrates well with machine learning pipelines.
- **Use Case:** EDA in collaborative, big data environments, especially with Spark.

28. Orange

- **Purpose:** Data mining and machine learning.
- **Key Features:**
    - Provides a visual programming interface for data mining and EDA.
    - Allows drag-and-drop workflows for data analysis, visualization, and modeling.
    - Includes tools for clustering, classification, and data visualization.
- **Use Case:** Visual EDA and data mining, particularly for those who prefer low-code environments.

29. ExplainaBoard

- **Purpose:** EDA for natural language processing (NLP) tasks.
- **Key Features:**
    - Offers tools for analyzing and understanding the performance of NLP models.
    - Provides visualizations and insights tailored to text-based data.
    - Allows comparison of model performance across datasets and tasks.
- **Use Case:** EDA and performance analysis for NLP projects.