OpenData and Exploratory Data Analysis

Introduction to Data Science

Author: Eng. Carlos Andrés Sierra, M.Sc.
carlos.andres.sierra.v@gmail.com

Lecturer
Computer Engineer
School of Engineering
Universidad Distrital Francisco José de Caldas

2024-II

Outline

1. OpenData and Data Sources

2. Exploratory Data Analysis

Outline

1 OpenData and Data Sources

2 Exploratory Data Analysis

Data Sources and Formats

- **Data** is everywhere. It is generated by people, machines, and devices.
- **Data** is stored in databases, files, and APIs.
- **Data** is structured, semi-structured, and unstructured.
- **Data** is processed and analyzed to extract knowledge and insights.

Data Sources and Formats

- **Data** is everywhere. It is generated by people, machines, and devices.
- **Data** is stored in databases, files, and APIs.
- **Data** is structured, semi-structured, and unstructured.
- **Data** is processed and analyzed to extract knowledge and insights.

Data Sources and Formats

- **Data** is everywhere. It is generated by people, machines, and devices.
- **Data** is stored in databases, files, and APIs.
- **Data** is structured, semi-structured, and unstructured.
- **Data** is processed and analyzed to extract knowledge and insights.

Data Sources and Formats

- **Data** is everywhere. It is generated by people, machines, and devices.
- **Data** is stored in databases, files, and APIs.
- **Data** is structured, semi-structured, and unstructured.
- **Data** is processed and analyzed to extract knowledge and insights.

Third-party APIs

A **third-party API** is an API that is provided by a third-party company or organization.

Propietaty Data Sources

Propietary Data Sources are data sources that are owned by a company or organization. Access to proprietary data is restricted to authorized users.

Open Data

Open data is data that is freely available to everyone to use and republish as they wish, without restrictions from copyright, patents, or other mechanisms of control.

Data Streaming

Data streaming is the process of transferring data from a source to a destination in a continuous and real-time manner.

Outline

1 OpenData and Data Sources

2 Exploratory Data Analysis

Data Profiling and Cleaning

- **Data profiling** is the process of analyzing data to understand its structure, quality, and content.
- **Data cleaning** is the process of detecting and correcting errors and inconsistencies in data.
- **Data profiling** and **data cleaning** are important steps in the data analysis process.

Data Profiling and Cleaning

- **Data profiling** is the process of analyzing data to understand its structure, quality, and content.
- **Data cleaning** is the process of detecting and correcting errors and inconsistencies in data.
- **Data profiling** and **data cleaning** are important steps in the data analysis process.

Data Profiling and Cleaning

- **Data profiling** is the process of analyzing data to understand its structure, quality, and content.
- **Data cleaning** is the process of detecting and correcting errors and inconsistencies in data.
- **Data profiling** and **data cleaning** are important steps in the data analysis process.

Exploratory Data Analysis

- **Exploratory data analysis** is the process of analyzing data to summarize its main characteristics.
- **Exploratory data analysis** is the first step in the data analysis process.
- **Exploratory data analysis** is the process of generating insights and hypotheses from data.

Exploratory Data Analysis

- **Exploratory data analysis** is the process of analyzing data to summarize its main characteristics.
- **Exploratory data analysis** is the first step in the data analysis process.
- **Exploratory data analysis** is the process of generating insights and hypotheses from data.

Exploratory Data Analysis

- **Exploratory data analysis** is the process of analyzing data to summarize its main characteristics.
- **Exploratory data analysis** is the first step in the data analysis process.
- **Exploratory data analysis** is the process of generating insights and hypotheses from data.

Outliers Detection

- An outlier is an observation that is significantly different from other observations in a dataset.
- **Outliers** can skew the results of data analysis.
- **Outliers** can be detected using statistical methods and visualization techniques.

Outliers Detection

- An outlier is an observation that is significantly different from other observations in a dataset.
- **Outliers** can skew the results of data analysis.
- **Outliers** can be detected using statistical methods and visualization techniques.

Outliers Detection

- An outlier is an observation that is significantly different from other observations in a dataset.
- **Outliers** can skew the results of data analysis.
- **Outliers** can be detected using statistical methods and visualization techniques.

Extract-Tranform-Load Pipelines

- An extract-transform-load (**ETL**) pipeline is a process that extracts data from sources, transforms the data into a format that is suitable for analysis, and loads the data into a destination.
- An ETL pipeline is a key component of the data analysis process.
- An ETL pipeline is a repetitive process that requires automation.

Extract-Tranform-Load Pipelines

- An extract-transform-load (**ETL**) pipeline is a process that extracts data from sources, transforms the data into a format that is suitable for analysis, and loads the data into a destination.
- An ETL pipeline is a key component of the data analysis process.
- An ETL pipeline is a repetitive process that requires automation.

Extract-Tranform-Load Pipelines

- An extract-transform-load (**ETL**) pipeline is a process that extracts data from sources, transforms the data into a format that is suitable for analysis, and loads the data into a destination.
- An ETL pipeline is a key component of the data analysis process.
- An ETL pipeline is a repetitive process that requires automation.

Outline

1 OpenData and Data Sources

2 Exploratory Data Analysis

Thanks!

Questions?

Repo: *https://github.com/EngAndres/ud-public/tree/main/courses/data-science-introduction*