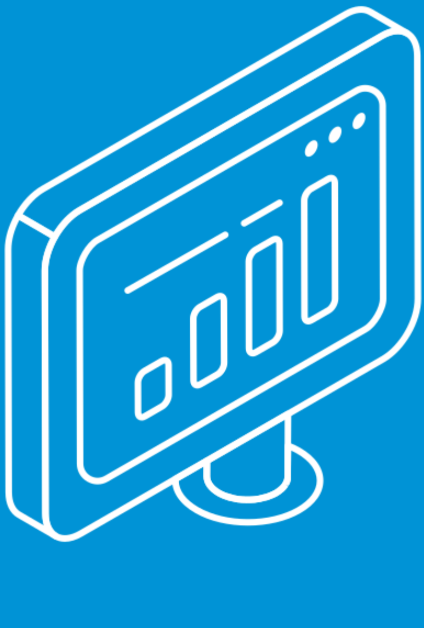
TARDIS - BOOTSTRAP

CLEANING AND VISUALIZING TRAIN DELAY DATA



TARDIS - BOOTSTRAP

Welcome to the Tardis project bootstrap!

This guided bootstrap is designed to help you get started with data cleaning and visualization, essential skills for the main project.

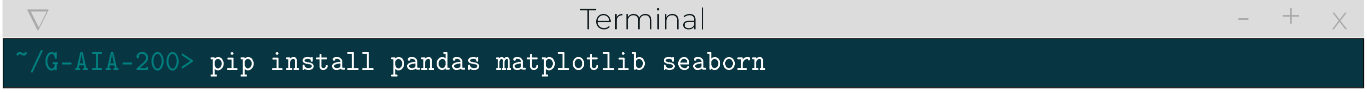
By the end of this bootstrap, you will:
- Explore a dataset.
- Identify and address common data issues (missing values, duplicates, inconsistent formats).
- Generate basic visualizations to understand trends.
- Prepare the dataset for analysis in the main project.

Dataset

We will use a simplified train delay dataset for this exercise. The dataset contains:
- `date`: The date of the train departure.
- `departure_station`: The station where the train departs.
- `arrival_station`: The station where the train arrives.
- `scheduled_time`: The scheduled departure time.
- `actual_time`: The actual departure time.
- `delay_minutes`: The delay in minutes (negative values indicate early departure).

Step 1: Loading the Dataset

First, install the required libraries:



```
~/G-AIA-200> pip install pandas matplotlib seaborn
```

Now, in a Jupyter Notebook (`touch tardis_eda.ipynb`), use pandas functions to load the dataset.

- ✓ Which pandas function can load a CSV file?
- ✓ How can you preview the first few rows of the dataset?
- ✓ What columns do you see?
- ✓ How many rows are present in your dataset?

Step 2: Exploring Data

Use pandas methods to explore the dataset's structure:

- ✓ Which method helps you see column names, data types, and missing values?
- ✓ Are all the data types appropriate?
- ✓ What issues might arise based on this exploration?

Step 3: Identifying Data Issues

Find missing values and duplicates using pandas functions:

- ✓ Which pandas function helps detect missing values?
- ✓ How do you check for duplicate entries?
- ✓ How would you handle these issues if you found them?

Step 4: Initial Cleaning

Decide how you will handle missing values and duplicates:

- ✓ Will you remove or replace missing values? What factors influence your decision?
- ✓ Which pandas methods would help you remove missing values and duplicates?
- ✓ After cleaning, verify your dataset again. What improvements do you notice?

> Maybe df.dropna(), df.drop_duplicates() ?

Step 5: Adjusting Data Types

Convert date and time columns to appropriate formats:

- ✓ Which pandas functions convert columns to datetime?
- ✓ Why is it important to have correctly formatted dates and times?

> Look at panda.to_* functions.

Step 6: Feature Engineering

Create new columns to enhance your analysis:

- ✓ How can you extract the day of the week from a date?
- ✓ Define your own function to categorize delays into groups (e.g., on time, slight delay, major delay).
- ✓ Why might categorizing delays be useful for analysis?

Step 7: Simple visualization

Use Matplotlib or Seaborn to visualize your cleaned data:
- Which functions would help create a histogram or boxplot of delays?
- What initial trends can you identify through these visualizations?

Step 8: Ideas for Further Visualizations

Here are some additional visualization ideas to explore later:

✓ Countplot: Explore the frequency of delay categories or stations.
✓ Heatmap: Visualize correlations between different variables.
✓ Lineplot: Analyze delays over time (daily or weekly trends).
✓ Which visualization might help uncover additional insights?

Step 9: Clean Data

Save your cleaned dataset to a CSV file:

✓ Which pandas method exports your DataFrame to a CSV?

Conclusion

Congratulations! You have successfully:

✓ Cleaned missing and inconsistent data.
✓ Engineered useful features for analysis.
✓ Created basic visualizations to explore trends.

These skills will be critical for your Tardis project. Now, you are ready to tackle the full dataset and build your predictive model!

v 1.0.0

{EPITECH}