Exercise #05

IT University of Copenhagen (ITU)
Data Mining KSD (DAMIN)
(Autumn 2025)

23 September 2025

Introduction This exercise list will provide hands-on experience in applying data exploration and visualization techniques using real-world datasets. The tasks will introduce you to descriptive statistics, data distributions, and effective data mining and analysis visualization techniques. Each exercise is structured to be practical and approachable for students, offering a deeper understanding of key concepts through the Star Wars dataset. The learning objectives for this exercise encompass:

- Understand and apply descriptive statistics to explore datasets.
- Visualize data distributions and identify patterns.
- Analyze the relationships among numerical and categorical variables.
- Utilize Python libraries such as Pandas, Plotly, Seaborn, and ydata_profiling for efficient data exploration and visualization.
- Interpret visualizations and derive insights from real-world datasets.

Exercise 05.01. *Loading the Dataset* (15-20 minutes) – In this exercise, you will begin by loading the Star Wars dataset into a Pandas `DataFrame`. This is the first step in exploring and analyzing data. Proper loading and inspection of the dataset are critical for understanding its structure, identifying missing values, and preparing for further analysis. Working with real-world datasets like this will help you develop the skills needed for effective data handling in Python.

- **Instructions:**
  - Download and load the dataset.
    - ∗ The Star Wars dataset is provided as a CSV file called `starwars.csv`. Begin by loading the dataset into a Pandas `DataFrame`.
    - ∗ Inspect the basic structure of the data by viewing the first few rows and checking the column names and data types.

  - Inspect the dataset.
    - Display the first 10 rows of the dataset to get an initial view.
    - Print a summary of the dataset's structure, including:
      - The number of rows and columns.
      - The column names and their data types.
      - Information on any missing values.
  - Handle missing data.
    - Identify which columns contain missing values.
    - One alternative is filling missing numerical values with the *mean* of their respective columns and missing categorical values with "*Unknown*." You can also use other methods to handle missing values.
    - Confirm that there are no remaining missing values after this step.

- **Expected Output/Questions:**
  1. What is the size of the dataset (number of rows and columns)?
  2. Which columns contain missing values?
  3. How did you handle the missing values, and what changes were made to the dataset as a result?
  4. Provide a preview of the first and the last 10 rows of the updated dataset after handling missing data.

Exercise 05.02. *Descriptive Statistics* (20-30 minutes) – In this exercise, you will compute descriptive statistics for the Star Wars dataset to summarize its main characteristics. Descriptive statistics help you understand the data's distribution, central tendency, and variability. By analyzing the numeric columns, such as character `height`, `mass`, and `birth_year`, you will gain insights into the dataset and identify any potential outliers. This task will prepare you for more in-depth data analysis and exploration.

- **Instructions:**
  - *Select Numerical Columns*: Identify at least three numerical columns in the dataset that are relevant for statistical analysis, such as:
    - `height`
    - `mass`
    - `birth_year`
  - *Compute Descriptive Statistics*: Use Pandas to calculate the following descriptive statistics for each of the selected columns:
    - Mean
    - Median

    - ∗ Mode
    - ∗ Standard Deviation
    - ∗ Variance

  - – *Interpret the Results*: Analyze the output of the descriptive statistics:
    - ∗ Compare the *mean* and *median* to understand the skewness of the data.
    - ∗ Discuss any large *standard deviations* that might indicate high variability.
    - ∗ Identify outliers in the data by examining the values that deviate significantly from the *mean*.
  - – *Handling Missing Values*: If there are missing values in the selected columns, describe how you handled them in Exercise 05.01 and how it may affect the computed statistics.

- **Expected Output/Questions:**

  1. What are the *mean*, *median*, *mode*, *standard deviation*, and *variance* for the selected columns?
  2. Are there any significant differences between the *mean* and *median*? What does this tell you about the data distribution?
  3. Based on the *standarddeviation* and *variance*, which numerical column has the highest variability?
  4. Are there any potential outliers in the data? If so, which columns and values?

Exercise 05.03. *Visualizing Data Distributions* (20-30 minutes) – In this exercise, you will create visualizations to explore the distribution of numerical variables in the Star Wars dataset. Visualizing data distributions allows you to understand your data's shape, spread, and potential outliers. Histograms, density plots, and box plots are potent tools that help reveal insights about the underlying patterns in numerical data. This task will give you hands-on experience with visualization techniques, preparing you to communicate data insights effectively.

- **Instructions:**
  - – Visualize the distribution of `height`.
    - ∗ Create a histogram to display the distribution of the `height` column. Use an appropriate bin size to represent the data clearly.
    - ∗ Additionally, create a density plot for `height` to visualize the data's smooth distribution.
  - – Visualize the distribution of `mass`.
    - ∗ Generate a box plot for the `mass` column to detect any outliers in the dataset.
    - ∗ Also, create a histogram for `mass` to observe its frequency distribution.

  - Analyze the results.
    - ∗ Interpret the shape of the distributions (e.g., skewed, normal, or bimodal).
    - ∗ Identify any outliers present in the `mass` column based on the box plot.
  - Customize the visualizations.
    - ∗ Ensure each plot has clear labels for the X-axis, Y-axis, and title.
    - ∗ Use appropriate colors and formatting to improve the readability of the plots.

- **Expected Output/Questions:**
  1. Based on the histogram and density plot for `height`, what does the distribution look like (e.g., normal, skewed)?
  2. What do you observe from the box plot for `mass`? Are there any significant outliers?
  3. How does the `mass` histogram compare to the box plot in terms of visualizing data spread and variability?
  4. What can you infer about the overall distribution of `height` and `mass` from the visualizations?

Exercise 05.04. *Categorical Data Exploration* (30-35 minutes) – In this exercise, you will explore the categorical variables in the Star Wars dataset, such as `species` and `gender`. Categorical data exploration helps you understand how different categories are distributed within the dataset and whether any notable patterns exist between them. Visualizing the relationship between these variables allows you to uncover exciting trends and better understand the dataset's structure. This task will give you experience in analyzing and visualizing categorical data.

- **Instructions:**
  - Bar chart for `species`.
    - ∗ Create a bar chart that shows the count of characters for each `species` in the dataset.
    - ∗ Ensure that the chart is clearly labeled, with appropriate titles and axis labels.
  - Grouped bar chart for `gender` distribution by `species`.
    - ∗ Create a grouped bar chart that visualizes the distribution of `gender` within each `species`.
    - ∗ Differentiate the groups by color to clearly show the breakdown of `gender` across `species`.
  - Analyze the results.
    - ∗ Identify the most common `species` in the dataset.

        - Explore the `gender` distribution within the most populous `species`.
    - Customize the visualizations.
        - Label each chart with an appropriate title, X-axis, and Y-axis labels.
        - Apply distinct colors to the categories to enhance readability.

- **Expected Output/Questions:**
    1. Which `species` has the most characters in the dataset?
    2. How does the `gender` distribution vary across `species`? Which `species` show an imbalanced `gender` ratio?
    3. Based on the bar charts, are there any `species` or `gender` categories that are underrepresented?

Exercise 05.05. *Correlation and Relationships* (20-30 minutes) – In this exercise, you will explore the relationships between numerical variables in the Star Wars dataset by calculating correlation coefficients and visualizing these relationships. Correlation helps you understand the strength and direction of relationships between variables, which is critical for identifying patterns in data. Additionally, you will visualize these relationships using scatter plots and heatmaps to gain deeper insights into the dataset.

- **Instructions:**
    - Calculate correlation coefficients.
        - Select two numerical columns, such as `height` and `mass`, and compute the *Pearson correlation coefficient* between them.
        - Calculate the correlation between other pairs of numerical columns as well.
    - Visualize the correlation with a scatter plot.
        - Create a scatter plot to visualize the relationship between `height` and `mass`.
        - Add a trend line to the scatter plot to visualize the linear relationship (if any) between the two variables.
    - Create a correlation heatmap.
        - Generate a correlation matrix that shows the correlation coefficients for all the numerical columns in the dataset.
        - Visualize the correlation matrix using a heatmap, with annotations to display the correlation values.
    - Analyze the results.
        - Discuss the strength and direction of the correlation between `height` and `mass`.
        - Identify the strongest positive and negative correlations among the numerical variables based on the heatmap.

- **Expected Output/Questions:**

  1. What is the *Pearson correlation coefficient* between `height` and `mass`? Is it positive or negative?
  2. Based on the scatter plot, is there a clear relationship between `height` and `mass`?
  3. From the heatmap, which numerical columns have the strongest correlation? Which have the weakest correlation?

Exercise 05.06. *Interactive Visualization and Exploratory Data Analysis* (30-45 minutes) In this exercise, you will combine interactive visualization techniques with automated exploratory data analysis (EDA) to gain comprehensive insights into the Star Wars dataset. Using tools like Plotly for interactive visualizations and ydata_profiling for automated profiling, you will explore data patterns, identify correlations, spot potential issues, and interact with the data dynamically. This exercise gives you hands-on experience with advanced data exploration techniques, helping you build intuitive visualizations and gain deep insights into the dataset.

- **Instructions:**

  – *Generate an Exploratory Data Analysis report*: Use the ydata_profiling package to generate an automated EDA report for the Star Wars dataset. This report will summarize the dataset's structure, distribution, correlations, missing values, and potential outliers.

  – *Interactive scatter plot with Plotly*: Create an interactive scatter plot to explore the relationship between `height` and `mass` using Plotly. Interactive plots allow you to hover over data points to view specific details and dynamically zoom in/out.

  – *Interactive Bar Plot for Categorical Variables*: Create an interactive bar plot to explore the distribution of `species` in the dataset. Customize it by showing the `species` count when hovering over each bar.

- **Expected Output/Questions:**

  1. What insights did the automated EDA report provide about the structure and quality of the dataset?
  2. From the interactive scatter plot, do you notice any relationships between `height` and `mass` across `species`?
  3. What do the interactive visualizations reveal about the distribution of categorical variables like `species` and `gender`?
  4. How do the interactive plots enhance your ability to explore the data compared to static plots?