Introduction to ML
Assignment 2

September 9, 2025

1 Introduction

This assignment is a hands-on exercise designed to help you master fundamental data science skills using Python. Through a series of tasks, you will learn to explore, clean, and analyze a real-world dataset. The goal is to build your proficiency with key libraries such as **pandas**, **scikit-learn**, and **matplotlib**. Using **seaborn** for visualization is optional, but recommended for creating professional-looking plots.

Data Description

The dataset contains hourly bike rental counts from a bike-sharing system. The features include:

- **instant**: Record index
- **dteday**: Date
- **season**: Season (1:springer, 2:summer, 3:fall, 4:winter)
- **yr**: Year (0: 2011, 1:2012)
- **mnth**: Month (1 to 12)
- **hr**: Hour (0 to 23)
- **holiday**: Whether the day is a holiday or not
- **weekday**: Day of the week
- **workingday**: 1 if the day is neither a weekend nor a holiday, 0 otherwise.
- **weathersit**: Weather situation:
  - 1: Clear, Few clouds, Partly cloudy
  - 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist

  - 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
  - 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

- **temp**: Normalized temperature in Celsius.
- **atemp**: Normalized feeling temperature in Celsius.
- **hum**: Normalized humidity. The values are divided to 100 (max)
- **windspeed**: Normalized wind speed. The values are divided to 67 (max)
- **casual**: Count of casual users
- **registered**: Count of registered users
- **cnt**: Count of total rental bikes including both casual and registered

(IF YOU FIND ANY MISTAKE IN THIS DESCRIPTION, PLEASE LET ME KNOW)

---

2 General Exploration & Data Cleaning

First, you will need to import the necessary libraries and load your dataset. If you use any other libraries to solve other tasks, don't forget to import them at the beginning of your notebook. Questions should be answered in a Jupyter notebook. You can use markdowns instead of comments.

Data Import

```python
import pandas as pd
import matplotlib.pyplot as plt

from sklearn.preprocessing import OneHotEncoder


# Read Bikeshare dataset
df_bike = pd.read_csv("Bike.csv")

# Rename the ’cnt’ column to ’cnt_rental_bike’ for clarity
df_bike.rename(columns={’cnt’: ’cnt_rental_bike’}, inplace=True)
```

Listing 1: Data Import

1. What are the first few rows of your loaded DataFrame? How many rows and columns does it have? What are the data types for each column?

2. Are there missing values in your dataset? If so, which columns are affected?

3. Are there duplicate rows?

4. The dataset has several columns related to date and time. Create a new column, `datetime`, that combines the date and time information. What is the appropriate data type for this new column?

5. Use the function `describe()` from pandas to analyze basic statistics of your data set (no need to use `print()`, pandas prints a structured table automatically).

6. What is the total number of bike rentals? What are the average, minimum, and maximum rental counts?

7. Visualize the distribution of bike rentals between different seasons. Which season has the highest average rental count?

8. Compare the rental numbers for casual users versus registered users. How do their average daily usage patterns differ?

9. What are the unique values of `weathersit` column? Replace them by descriptive labels:

   - 1: 'clear'
   - 2: 'cloudy'
   - 3: 'light_rain'
   - 4: 'heavy_rain'

10. What is the data type of the `weathersit` column?

11. Visualize the relationship between weather conditions (`weathersit`) and the number of rentals. Which weather situation corresponds to the highest number of rentals on average? Does it make sense?

12. Use the function `OneHotEncoder` from `Scikitlearn` to encode your column.

13. Consider the advantages and disadvantages of using one-hot encoding on a categorical variable. Does your answer depend on whether the variable is nominal (no order, like weather) or ordinal (has a natural order, like a rating scale)?

14. What can you say about the linear dependence of the columns produced by one-hot encoding? What solution exists to avoid this problem? Change your `OneHotEncoder` function accordingly.

15. Repeat the process of encoding using ordinal encoding for a suitable categorical column. Why can we consider that this method is more appropriate than one-hot encoding?

16. Choose a numerical feature and normalize it to a range of $[0, 1]$.

17. Standardize the same feature using standardization.

18. Which features would you consider normalizing, and which would you standardize? Justify your choices based on their distributions.

---

3 Conceptual Questions

1. A column with values 'male', 'female', 'NaN' is an example of what type of data?

   - (A) Ordinal
   - (B) Nominal
   - (C) Continuous
   - (D) Numerical

2. Explain the difference between nominal and ordinal data types. Provide an example for each that is not from this dataset.

3. If two features are perfectly linearly dependent, it means they contain redundant information for a linear model.

   - (A) True
   - (B) False

4. Consider a one-hot encoded feature with N categories. If you do not drop any of the one-hot columns, what is the relationship between the sum of the values in these N columns and the number of rows?

5. Why is it a common practice to drop one of the columns after one-hot encoding a feature for use in a linear regression model? What is the statistical term for this issue?

6. You have a dataset with a column representing a country's rank in a competition (1st, 2nd, 3rd, etc.). Would you recommend using one-hot encoding or ordinal encoding for this feature? Justify your choice.

7. Describe a scenario where using one-hot encoding could lead to a significant increase in the dimensionality of your dataset.

8. Which of the following scaling methods transforms the data to have a mean of 0 and a standard deviation of 1?

- (A) Normalization (Min-Max Scaling)
- (B) Standardization
- (C) Log Transformation
- (D) Robust scaling

9. Normalization is generally preferred when the data contains significant outliers, as it is more robust to them than standardization.

   - (A) True
   - (B) False

10. A feature has values $\{10, 22, 27, 53\}$. Calculate the normalized values for this feature using Min-Max scaling.

11. A feature has a mean (μ) of 50 and a standard deviation (σ) of 10. A specific data point has a value of 75. Calculate its standardized value (Z-score).

12. What is the main purpose of scaling features before training a machine learning model? Name at least two types of models that are particularly sensitive to the scale of input features.