Lesson 1.5

1. Load Titanic Dataset.csv from <https://github.com/sanadv/MLCourse>
2. Handel messing data
3. Remove duplicates
4. Correct inconsistencies
5. Detect outliers using IQR and Z-Score
6. Perform Normalization and Standardization
7. Plot the data before and after cleaning

Interview Questions

* **Scenario**: You are working on a dataset that includes patient records in a healthcare database. After initial analysis, you identify that 15% of the patient age data is missing. The dataset is fairly large, with over 100,000 records. The age data is important for your analysis to track disease prevalence across age groups.
* **Question 1**: Describe how you would handle these missing values. Discuss the techniques you would consider and explain your choice. How would your approach change if missing values were identified in a more critical feature, such as diagnosis information?
* **Scenario**: You are tasked with cleaning a financial dataset used for forecasting stock prices. The dataset contains some apparent outliers in the volume of trades, which could potentially skew the predictive models. Preliminary analysis shows that these outliers represent days with significant market news.
* **Question 2**: What methods would you use to detect these outliers, and how would you decide whether to remove or adjust these data points in the dataset? Outline the potential impacts of your decision on the forecast model's performance.
* **Scenario**: Imagine you are preparing a dataset for a machine learning model that predicts real estate prices. The dataset features have varying scales and distributions, including property size in square feet and local crime rate per 1,000 residents.
* **Question 4**: Would you choose to normalize or standardize these features, and why? Provide a detailed explanation of how each process would affect the data and the model's learning process. What might be the implications of choosing one method over the other in terms of model performance and accuracy?