Lab 1: Real-World Data Cleaning, Transformation, and Visualization
17/06/2023

Objective: This exercise aims to provide hands-on experience with real-world data preprocessing and visualization. The goal is to understand the challenges of cleaning, transforming, and visualizing data to derive valuable insights.

Instructions:

Dataset: Use the Titanic dataset from the Seaborn library (seaborn.load_dataset('titanic')). This dataset has numerous missing and inconsistent data entries.

Data Cleaning and Transformation:

- First, identify missing data in the dataset using the isnull() function in Pandas.
- Then, choose an appropriate strategy to handle the missing data. This can be deletion, imputation, or any other method that you think is appropriate.
- Conduct necessary transformations such as scaling and normalization of data. For instance, you can convert the 'fare' column to log scale to deal with its skewness.
- Conduct feature engineering if needed. For example, create a new column 'Family_Size' by adding 'SibSp' and 'Parch' columns.

Data Visualization:

- Once the data is cleaned and transformed, proceed with visualization using Matplotlib and Seaborn libraries.
- Plot a histogram of the age distribution of passengers.
- Using a box plot, compare the fares paid by passengers who survived and who did not.
- Create a heatmap to visualize the correlation among different numeric variables in the dataset.
- Use a pairplot to observe the pairwise relationship between the different classes of the 'survived' column.

By the end of this exercise, you should have a cleaned and transformed dataset and several visualizations that highlight different aspects of the data.

Note: There's no one-size-fits-all solution in data preprocessing and visualization, and different approaches might be appropriate for different situations. Feel free to try different strategies and visualizations that you think would be appropriate for this dataset. This is an exploratory exercise to encourage you to think critically about how to handle real-world data.

Upload your notebook with the solutions to your GitHub, and update Real-World Data Cleaning, Transformation, and Visualization.docx with the link by Tuesday, 12 am GMT.