Data Preprocessing in Pandas

This assignment is designed to introduce you to tools that will be useful in future practical assignments.

You will learn to:

- work with data using Python and the Pandas package
- perform data preprocessing
- find simple patterns in data

Introduction Currently, Python is one of the most widely used programming languages. One of its advantages is the large number of packages available to solve a wide range of tasks. In our course, we recommend using the Pandas, NumPy, and SciPy libraries, which simplify reading, storing, and processing data. In future work, you will also become familiar with the Scikit-Learn package, which implements many machine learning algorithms.

Getting Started To start working with data, you must first load it from a file. In this assignment, we will work with data in CSV format, which is intended for storing tabular data: columns are separated by commas, and the first row contains the column names.

Example of loading data into Pandas:

```
python
Copy code
import pandas
data = pandas.read_csv('titanic.csv', index_col='PassengerId')
```

The data will be loaded as a DataFrame, which allows you to work with it conveniently. In this case, the parameter `index_col='PassengerId'` specifies that the `PassengerId` column is used as the row numbering for this DataFrame.

To view the data, you can use several methods:

- The more common Python approach (if only one index is specified, row selection is performed):

```
python
Copy code
data[:10]
```

- Or you can use a DataFrame method:

```
python
Copy code
data.head()
```

One way to access DataFrame columns is by using square brackets and the column name:

```
python
Copy code
data['Pclass']
```

To calculate some statistics (count, mean, maximum, minimum), you can also use DataFrame methods:

```
python
```

```
Copy code
data['Pclass'].value_counts()
```

For a more detailed list of DataFrame methods, refer to the documentation.

Materials The dataset is sourced from the Kaggle website: *Titanic: Machine Learning from Disaster*.

Assignment Instructions Load the `titanic.csv` dataset and, using the methods described above, answer the following questions:

1. How many men and women were on the ship? Provide two numbers separated by a space.
2. What proportion of passengers survived? Calculate the percentage of survivors (no percentage sign needed).
3. What percentage of the passengers were in first class? Provide the answer as a percentage (no percentage sign needed).
4. What was the age of the passengers? Calculate the mean and median age. Provide two numbers separated by a space.
5. Do the number of siblings/spouses correlate with the number of parents/children? Calculate the Pearson correlation between the `SibSp` and `Parch` features.
6. What was the most common female name on the ship? Extract the first name of each passenger (from the `Name` column). This task is a typical example of what data analysts face. The data is very diverse and noisy, but you need to extract the necessary information. Try manually reviewing a few values in the `Name` column and develop a rule for extracting first names and distinguishing between male and female names.

Round answers to two decimal places if necessary.

Each answer should be in a text file with the answer on the first line. Note that the files you submit should not contain a newline at the end, as this is a limitation of the Coursera platform. We are working on removing this limitation.