NextHikes

Overview

Problem Statement

You are a junior data scientist assigned a new task to perform data wrangling on a set of datasets. The datasets have many ambiguities. You must identify those and apply different data-wrangling techniques to get a dataset for further usage.

Dataset

- Download Dataset_1[dataset_1 - Google Sheets], Dataset_2[dataset_2.xlsx - Google Sheets], and Dataset_3 [dataset_3 - Google Sheets] and upload the datasets for your given analysis.

Attribute information:

- date = date of the ride
- season - 1 = spring, 2 = summer, 3 = fall, 4 = winter
- holiday - whether the day is considered a holiday
- working day - whether the day is neither a weekend nor a holiday
- weather:-

  1: Clear, Few clouds, Partly cloudy, 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 - "feels like" temperature in Celsius
- humidity - relative humidity
- windspeed - wind speed

- casual - number of non-registered user rentals initiated
- registered - number of registered user rentals initiated
- count - number of total rentals

Important libraries to be used

- Pandas is a high-level data manipulation tool
- NumPy is used for working with multidimensional arrays

The following concepts are to be used with the help of a business use case:

- Data acquisition
- Different methods for data wrangling:
  1. **Merge datasets**
  2. **Identify unique values**
  3. **Drop unnecessary columns**
  4. **Check the dimensions of the dataset**
  5. **Check the datatype of the dataset**
  6. **Check datatype summary**
  7. **Treat missing values**
  8. **Validate the correctness of the data at the primary level if applicable**

Tasks:

Task 1: Data Acquisition and Data Wrangling on dataset 1 and dataset 2

[All methods to be applied for data wrangling concerning the listed above and combine_data from dataset1 and dataset2, Work on central tendency.].- Week 1

Task 2: Data Acquisition and Wrangling on Dataset 3.

Concatenate combine_ data with Dataset_3. To be Worked on missing values and outliers.-Week 2

Check the skewness and correlation of the data. -Week 3

Learning Outcomes

Technical Skills:

1. Data Acquisition
2. Data Wrangling

Badges

Each week, one user will be awarded one of the badges below for the best performance in the category below.

In addition to being the badge holder for that badge, each badge winner will get +20 points to the overall score.

This approach aims to support and reward expertise in different parts of the Core Python and Basic Data Science.

There will also be a mark that will be added to the most innovative approach.

Interim Submissions

To be categorized into the different weeks till 2nd week

- Your employer wants a quick meeting after you've done a first quick pass of the data and wants to know whether further investigation is useful. To achieve this, summarize your findings from Task 1 and Task 2
- Link to your GitHub code that includes your Jupyter Notebook].

Feedback

You may not receive detailed comments on your interim submission but will receive a grade.

Final Submission **(15-10-2024)**

- Work on the Jupyter Notebook, which will be saved in the Project 2 folder.
- Link to your GitHub Account that should include your required code.

Feedback

You will receive comments/feedback in addition to a grade.