NextHikes IT Solutions Project 2: Data Harmonization and Insights Extraction

Overview

Problem Statement

As a junior data scientist, your role is to address ambiguities and inconsistencies in multiple datasets. The goal is to leverage advanced data-wrangling techniques to produce a clean, unified dataset for future business analysis and modeling.

Key Deliverable

- A well-integrated dataset prepared through systematic wrangling.
- Insightful analysis ready for downstream applications.

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

Tools and Techniques

- **Tools**:
    - Pandas for data manipulation.
    - NumPy for numerical computations.
- **Techniques**:
    - Merging datasets.
    - Imputation of missing values.
    - Outlier detection and handling.
    - Data validation and exploration.

Project Workflow

Phase 1: Data Acquisition

- Load datasets into Python.
- Inspect and explore data structures using .info() and .head().
- Document initial observations about data quality and completeness.

Phase 2: Data Wrangling

- **Task 1 (Week 1)**: Dataset_1 and Dataset_2 Integration
    - Merge the datasets using common keys.

- Handle missing values using appropriate imputation techniques.
- Perform basic statistical analysis to evaluate central tendencies (mean, median, mode).
- Identify and drop unnecessary columns or duplicates.

- **Task 2 (Week 2)**: Dataset_3 Integration
  - Concatenate Dataset_3 with the merged dataset.
  - Perform detailed missing value analysis and outlier detection.
  - Apply statistical methods or capping to handle outliers.

Phase 3: Data Analysis

- **Task 3 (Week 3)**: Advanced Statistical Analysis
  - Analyze skewness of numerical columns and apply transformations as needed.
  - Compute correlations between attributes to identify relationships.
  - Visualize key insights using heatmaps, scatter plots, and boxplots.

Final Submission

The project will culminate in a single, comprehensive submission that includes the following deliverables:

1. **Merged Dataset:**
   - A fully combined dataset from Dataset_1, Dataset_2, and Dataset_3.
   - Includes appropriate handling of missing values and outliers.
2. **Insights and Observations:**
   - Summary of the data wrangling and cleaning processes.
   - Key observations and insights from the final dataset, including trends, distributions, and relationships.
3. **Jupyter Notebook:**
   - A complete notebook with:
     - Data acquisition and loading steps.
     - Data wrangling, merging, and cleaning processes.

        - Skewness and correlation analysis.
        - Visualizations such as boxplots, histograms, scatter plots, and heatmaps.
        - Code outputs and detailed explanations.
    - Uploaded to GitHub for review.

4. **PowerPoint Presentation (PPT):**
    - A concise presentation summarizing:
        - The project overview and objectives.
        - Data wrangling and cleaning steps.
        - Key challenges and how they were addressed.
        - Insights and findings from the dataset.
    - Should include relevant visualizations (e.g., graphs, heatmaps).
5. **GitHub Repository:**
    - The repository must include:
        - The Jupyter Notebook with all code, visualizations, and explanations.
        - Supporting datasets (if necessary for context).
        - The PowerPoint presentation.
        - A README file summarizing the project and instructions for replication.

Learning Outcomes

By completing this project, you will:

- Gain practical experience in dataset cleaning and wrangling.
- Develop an understanding of exploratory data analysis techniques.
- Prepare datasets for advanced modeling and business use cases.

Submission Date: 9th Jan 2025