Data Analytics and Visualizations - Summer 2024
Herald College
University of Wolverhampton

Home-Work-1
Data Wrangling with Python and Pandas.

July 9, 2024

About this Home-Work:
This home-work has the collection of the exercises and are expected to be completed individually. Before you begin the exercises please read the instructions carefully.

Contents

1 Home-Work Overview

In this work, you will perform a extensive data wrangling with provided small dataset using what you've learned so far. Tools and Python Package which can be used for this assignments, listed but are not limited to:

1. **Pandas library.**
2. **Numpy library.**
3. **Matplotlib library.**

1.1 Learning Objectives

On successful completion of this module, the learner must be able to:

1. Apply data cleansing and statistical operations on datasets to address the issues of data quality.
2. Select and apply suitable methods and analyses techniques for data of various structure and content and present summary statistics.
3. Extract various information from a given dataset using statistical techniques.

2 Exercise -1: Data Cleaning and Pre-processing

2.1 Dataset:

Dataset provided is `foodgrainsproduction_fiscalyear.csv`
This dataset contains a data on total production of various grains in Nepal during various fiscal year. Using the dataset complete the following tasks:

2.2 Tasks:

- **Task1: Handling Special Characters and Missing Values**
  1. Load the dataset into a Pandas Data-frame.
  2. Replace special characters and non-numeric values with "NaN".
  3. Fill missing values with appropriate methods(forward fill, median replacement or drop).
- **Task2: Checking and Dropping Duplicate Values**
  1. Check for duplicate rows if found drop any duplicate rows.
  2. Verify that all duplicates have been removed.
- **Task3: Checking for Outliers**
  1. Identify outliers in the numerical columns using the IQR method.
  2. Remove rows if identified as outliers.
  3. Convert columns to numeric values if not already numeric and ensure all numerical data is within realistic ranges.

3 Exercise -2: Data Wrangling with Pandas

3.1 Dataset:

To complete this exercise you will need to have following collections of dataset:

1. Dataset1: `customer_data.csv`
2. Dataset2: `product_data.csv`
3. Dataset3: `sales_data.csv`

3.2 Tasks:

For completion of the task follow the instructions:

- **Task1: Data Loading and Cleaning**
  1. Load all three datasets into Pandas Data-frames.
  2. Check for missing values and handle them appropriately.
  3. Convert the "OrderDate" and "CustomerSince" columns to datetime format.
  4. Check and remove any duplicate if present.
- **Task2: Sub-setting and Filtering**
  1. Subset the "sales_data" Data-frame to include only orders placed in the last year.
  2. Subset the "customer_data" Data-frame to include only customers who have made a purchase within last year.
  3. Filter the "product_data" to include only products that belong to specified category(e.g."Electronics")
- **Task3: Group Analysis**
  1. Calculate the total revenue for each region.(Revenue = Quantity*Price)
  2. Determine the average and median order value per customer.
  3. Find the top 5 customers in terms of the number of orders placed.
- **Task4: Merging and Aggregation**
  1. Merge the "sales_data" with "customer_data" on "CustomerID".
  2. Merge the resulting Data-frame with "product_data" on "ProductID"
  3. Calculate the total revenue per product category.
  4. Determine the average revenue per order for each region.
- **Task5: Advanced Analysis**
  1. Identify trends in sales over time (e.g., monthly revenue).
  2. Determine the churn rate (percentage of customers who did not make a purchase in the last year).
  3. Perform cohort analysis to understand the customer retention (e.g. grouping customers by the month they made their first purchase analyzing their purchasing behavior overtime).
- **Task6: Reporting and Interpretation**
  1. Summarize the key findings from your analysis in a report.

2. Provide actionable insights based on the data(e.g., recommendations fro improving sales, targeting specific customer segments, etc.)
3. Highlight any limitations of your analysis and suggest potential improvements or further analysis that could be conducted.