CS 539 Machine Learning
Fall 2024
Total marks 100
Due date: 23 September 2024, 11:59 PM
[No submission will be accepted after September 25, 2024, at 11:59pm]
Deliver: Submit via Canvas

Assignment 1

Instructions:
• All parts of question 1 must be answered within a single jupyter notebook or .py file. Submit notebook or .py file for question 1 and pdf document for question 2.
• Follow the file naming conventions: Name your submission file as first_lastname.ipynb or first_lastname.py.
• Use headings to distinguish each question in the notebook.
For any questions, please email rahsan@wpi.edu

QUESTION 1 [55 Marks]

Climate change stands as one of the most urgent challenges confronting our planet today. To effectively comprehend and address this critical issue, access to precise and comprehensive data regarding global temperatures and other climate-related factors is indispensable.

	A	B	C	D	E	F
1	Years	Month	Country	Temperature	Monthly_variation	Anomaly
2	1848	5	Afghanistan	19.573	-0.297	2.037
3	1848	6	Afghanistan	23.894	-0.796	2.136
4	1848	7	Afghanistan	26.507	-0.113	1.937
5	1848	8	Afghanistan	24.498	-0.462	1.937
6	1848	9	Afghanistan	19.068	-1.272	1.865

In this regard, you serve as a data analyst at the National Aeronautics & Space Administration (NASA) and are engaged in researching Earth's climate and temperature. Your work involves utilizing datasets sourced from satellites and ground-based sensors.

You have been entrusted with a dataset “earth_surface_temperatures .csv” encompassing surface temperature data for various countries worldwide, spanning from **Dec 1743** to **December 2020**. Your mission is to conduct an in-depth Exploratory Data Analysis (EDA) on this dataset. This analysis aims to extract insights and answer crucial questions about the data by delving into trends and patterns.

To achieve this, you are expected to:

a. Identify and rectify any missing values in the data using appropriate techniques. **[5 Marks]**
b. Transform the **Years** and **Month** columns into a single column labeled "**Date**" in the **MM-YYYY** format, with a **datetime64[ns]** data type. For example, the year 1848 and month 5 should be unified as a single value, such as 5-1848. **[5 Marks]**
c. Detect and investigate extreme temperature values that might be regarded as outliers. **[5 Marks]**
d. Compute summary statistics for temperature, monthly variation, and anomaly values, including mean, median, standard deviation, and range. **[5 Marks]**
e. Identify the countries included in the dataset and calculate their average temperature values. **[5 Marks]**
f. Determine the overall trend in global temperatures over the years and visualize this trend using a suitable chart. **[5 Marks]**
g. Identify the months with the highest and lowest temperatures for each country and find out whether there are noticeable seasonal patterns in the temperature data. **[5 Marks]**
h. Explore the variation in temperature anomalies on a monthly basis and identify any months with consistently high or low anomalies across the years. **[5 Marks]**
i. Choose five countries and compare the trends in their temperatures over the years, seeking any similar temperature patterns. **[5 Marks]**
j. Explore the potential correlation between temperature and monthly variation *or* anomaly values. Calculate correlation coefficients and create scatterplots to investigate this relationship. **[5 Marks]**
k. Provide an intriguing insight from the dataset by utilizing data visualization techniques such as histograms, box plots, or heatmaps to represent the data's distribution, trends, and relationships. **[5 Marks]**

QUESTION 2 [45 Marks]

As a member of the retail analytics team, you have been contacted by the Category Manager at a retail store, who desires to gain a deeper understanding of the customers who buy chips and their purchasing habits within the region through valuable insights that will eventually be used to inform the store's strategic plan for the chip category in the upcoming six months.

You have received the following e-mail from your manager.

Greetings!

I am following up on our earlier conversation with a few pointers to help you succeed in this task. Here are the key areas you will be working on and what we're looking for in each one:

Firstly, examine the transaction data ("transaction_data" file) and look for inconsistencies, missing data, outliers, correctly identify category items, and numeric data across all tables. If you notice any anomalies, please make the necessary changes in the dataset and save it for further analysis. Having clean data will make it easier for us to conduct an effective analysis.

Secondly, examine the customer data ("purchase_behaviour" file) for similar issues and check for null values. Once you're satisfied with the data, merge the transaction and customer data together for analysis, ensuring that you save your files along the way.

Thirdly, conduct data analysis and identify customer segments. Define the metrics, such as total sales, drivers of sales, and the source of the highest sales. Explore the data, create charts and graphs, and note any interesting trends and insights you find.

Finally, deep dive into customer segments and recommend which segments we should target. Determine if packet sizes are relative and form an overall conclusion based on your analysis.

Here is the task:

Your task is to provide a data-driven strategic recommendation for the upcoming category review. To achieve this, you must first analyze the current purchasing trends and behaviors to understand the customer segments and their chip purchasing behavior. To describe the customers' purchasing behavior, you need to identify relevant metrics. The client has a specific interest in understanding the chip purchasing behavior of different customer segments.
To begin the task, download the comma-separated values (CSV) data files provided to you and conduct preliminary data checks, including:

- Generating and interpreting high-level data summaries.
- Identifying any outliers and, if necessary, removing them (if applicable).
- Verifying the data formats and correcting them, if needed (if applicable).

In addition to the preliminary data checks, it is essential to extract additional features, such as pack size and brand name, from the data. Defining relevant metrics of interest is also crucial to gaining insights into the chip purchasing behavior of different customer segments.
Your ultimate goal is to formulate a strategic recommendation for the Category Manager, based on your findings. Therefore, it is essential that your insights have a commercial application and can be used to inform decision-making.
Lastly, a detailed report on your analysis findings, no longer than 3-4 pages, is required. The report should include any relevant visualizations you have created, as well as your recommendation to the Category Manager, to inform the store's strategic plan for the chip category. Do *not* include any technical aspects of your analysis, such as coding, in the report.
Note: This is an open-ended case study and can be approached in various ways, allowing for flexibility and creativity in the analysis process.
Additional Pointers (column description of purchase behavior):
LIFESTAGE: Customer attribute that determines if they have a family or not, and at what stage of life they are in. For instance, it considers whether their children are in preschool, primary or secondary school.
PREMIUM_CUSTOMER: Customer segmentation approach that distinguishes shoppers based on the price point and product types they purchase. Its purpose is to determine whether customers are willing to pay more for brand or quality or prefer to purchase the most economical options.