Introduction to Data Science
Fall 2023
Total marks 100
Due date: 23 October, 2023, 11:59 PM

Assignment 2

Instructions:

- All questions must be answered within a single notebook or .py file.
- Follow the file naming conventions: Name your submission file as RollNo.ipynb or RollNo.py (e.g., i22_xxxx.ipynb, where xxxx is your Roll Number).
- Use headings to distinguish each question in the notebook.
- Late submissions will not be accepted and will be given a zero.
- Any form of plagiarism will result in a zero for both parties involved.
- AI-generated content is prohibited. Detection of such content will lead to a zero score.

QUESTION 1 [40 Marks] [PANDAS DATA MANIPULATION]: Assume you are an analyst working on World Happiness Report (https://www.kaggle.com/datasets/unsdsn/world-happiness) and trying to find answers to key questions. The World Happiness Report is a landmark survey of the state of global happiness. The first report was published in 2012, the second in 2013, the third in 2015, and the fourth in the 2016 Update. The World Happiness 2017, which ranks 155 countries by their happiness levels, was released at the United Nations at an event celebrating International Day of Happiness on March 20th. The report continues to gain global recognition as governments, organizations and civil society increasingly use happiness indicators to inform their policy-making decisions. Download the datasets from the link above and they have been uploaded in the submission files as well. It contains dataset for the years 2015-2019. Do the following on the dataset.

a. Load the five datasets in pandas dataframes

b. Retrieve the list of countries in 2015 that have Health (Life Expectancy) value between 0.5-1.

c. Retrieve the top 10 countries with highest Happiness Score for the year 2016.

d. Display the mean Happiness Score of the countries grouped by the Region for any year.

e. Identify if there are any outliers in your datasets (for all datasets 2015-2019)

f. Analyze if there is any relationship between the Happiness Score and Economy (GDP per Capita) of the country for any year.

g. Combine the datasets from 2015-2019 to construct 1 dataframe that contains the following columns. Country, Region, Year (coming from the respective dataset) and Happiness Score for each year as a new column. Your dataframe should look below

Country	Region	2015	2016	2017	…
Singapore	Southeastern Asia	6.798	6.739	….	
….	…	…	…		

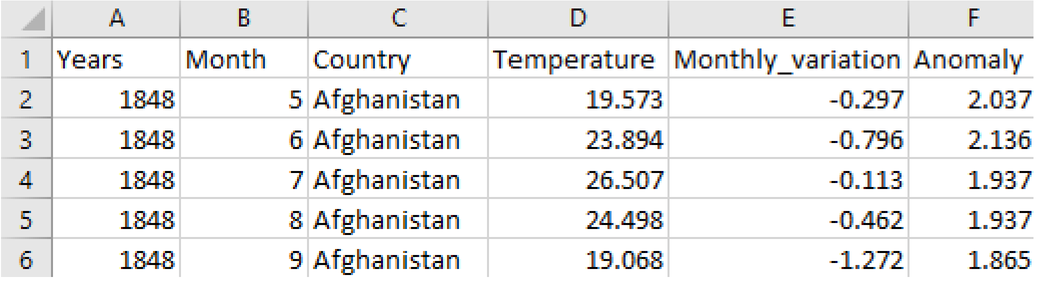
h. For any 2 countries of your choice draw a line chart (x-axis contains the years from the dataframe constructed in part g and y-axis contains the Happiness Score values for the respective years). Identify the trend.

i. For the year 2015, identify the regions that have an Economy (GDP per Capita) value less than 0.5.

j. Identify which 2 variables have the highest correlations across all the countries for the years 2015 and 2016.

QUESTION 2 [60 Marks] [EXPLORATORY DATA ANALYSIS (EDA)]:

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 encompassing surface temperature data for various countries worldwide, spanning from **May 1848** 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. Ensure that the data types of all columns are consistent with their values and make conversions where necessary. **[5 Marks]**
c. 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]**
d. Detect and investigate extreme temperature values that might be regarded as outliers. **[5 Marks]**

e. Compute summary statistics for temperature, monthly variation, and anomaly values, including mean, median, standard deviation, and range. **[5 Marks]**
f. Identify the countries included in the dataset and calculate their average temperature values. **[5 Marks]**
g. Determine the overall trend in global temperatures over the years and visualize this trend using a suitable chart. **[5 Marks]**
h. 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]**
i. 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]**
j. Choose five countries and compare the trends in their temperatures over the years, seeking any similar temperature patterns. **[10 Marks]**
k. Explore the potential correlation between temperature and monthly variation *or* anomaly values. Calculate correlation coefficients and create scatterplots to investigate this relationship. **[5 Marks]**

As a **bonus**, you have the opportunity to provide an intriguing insight from the dataset by utilizing data visualization techniques such as histograms, box plots, and heatmaps to represent the data's distribution, trends, and relationships. Your creativity and accuracy in this aspect will also be taken into account when assessing your work.