2023 Datathon

Data Analysis with ChatGPT Track Milestones

Setup:

- You will be using various world health dataset found on WHO, download it from ***here*** and set up your project.

Scenario:
You are a Data Analyst at Dasman Diabetes Institute. You have been tasked with analyzing & visualizing WHO data to compare and contrast Kuwait, GCC & Global health metrics.

Milestones

***Level One: Exploring (10 points -** 1 points per task, must complete all tasks to unlock level 2)*

LEVEL 1-A (5pts)

In **Life expectancy and Healthy life expectancy** dataset.

1. Reshape the dataset to look like the following figure, and display the shape of the new dataset. *(1pt)*

Country	Year	Both sexes	Male	Female	Category
Portugal	2019	71.0	69.6	72.2	Healthy life expectancy at birth
Honduras	2000	14.7	15.0	14.6	Healthy life expectancy at age 60
Congo	2019	64.7	63.8	65.6	Life expectancy at birth
Sri Lanka	2010	20.0	18.1	21.8	Life expectancy at age 60

2. What is the total number of unique countries? *(1pt)*
3. What are the years used in the dataset? *(1pt)*
4. What is the average **life expectancy at birth** in Kuwait for **males** over years? *(1pt)*
5. What is the highest **life expectancy at age 60** for **both sexes** over the years? Get the country and year. *(1pt)*

—---------- CALL A JUDGE —----------

LEVEL 1-B (5pts)

6. What is the lowest **healthy life expectancy at age 60** for **females** over the years? Get the country and year. *(1pt)*
7. Which year had the highest **healthy life expectancy at birth** for **males**? *(1pt)*
8. Which top/bottom 5 countries had the highest **life expectancy at birth** for **females** in 2000? *(1pt)*
9. Which top/bottom 5 countries had the lowest **healthy life expectancy at birth** for **males** in 2019? *(1pt)*
10. Which country in the Gulf region had the highest **healthy life expectancy at birth** for **females** in 2019? *(1pt)*

—---------- CALL A JUDGE —----------

Level Two: Analysing (20 points) *(2 points per task,must complete all tasks to unlock level 3)*

NOTE: Make sure you rescale the y-axis so the differences are more apparent

LEVEL 2-A (8 pts - 2 pts per task)

In **Life expectancy and Healthy life expectancy** dataset.

1. Split **healthy life expectancy at birth** in **2019** into the following bins, get the counts and plot this. *(2pts)*
    1) (0,40]
    2) (40,50]
    3) (50,60]
    4) (60,70]
    5) (70,80]
    6) (80,90]
2. Which year has the highest median **for life expectancy at birth** and **healthy life expectancy at birth** for **both sexes**? Visualize this. *(2pts)*
3. Is there a difference in frequency distributions between **2000** and **2019** of the **healthy life expectancy at birth** for **males**(within range(30,90) and bin width=5)? Visualize this. *(2pts)*
4. Draw a plot to display the differences in the **healthy life expectancy at birth** between **Kuwait**, **Japan**, **USA**, and **Germany** for **both sexes** over the years. *(2pts)*

—---------- CALL A JUDGE—----------

LEVEL 2-B (6 pts - 2 pts per task)

5. Visualize the differences in the average **healthy life expectancy at birth** between **males** and **females** over the years? Visualise this. *(2pts)*

6. Which year Kuwait had the best/worst global ranking of **life expectancy at birth** for **males** and **females**? Visualize the **life expectancy at birth** for **males** and **females for years [2000, 2010, 2015, 2019]**. *(2pts)*
7. Which 3 countries had the best / worst progress in the global ranking of **life expectancy at age 60** for **males**? *(2pts)*

—---------- CALL A JUDGE—----------

LEVEL 2-C (6 pts - 2 pts per task)

8. What are the percentages of countries, in the **years 2000, 2010, 2015, 2019,** where the **life expectancy at age 60** for **males** is greater than for females? *(2pts)*
9. Visualize the distributions of **Life expectancy at birth** for **males** and **females,** compute the skewness of **Life expectancy at birth** for **male** and **female**. *(2pts)*
10. Visualize the difference in **life expectancy at birth** for **males** and **females** between 2015 and 2019 for countries in the Gulf region. Which country showed the greatest improvement in **life expectancy at birth?** And which country showed the greatest decline? *(2pts)*

—---------- CALL A JUDGE—----------

Level Three: Using ChatGPT (*5 points per task, must complete all tasks to unlock level 4)*

LEVEL 3-A (10 pts - 5 pts per task)

1. Create an app that does exploratory data analysis of this dataset from the World Health Organization. Make sure we show various types of plots, with summary and descriptions when prompted (links/buttons). Use a Large Language Model (such as ChatGPT) to create the code, you are only allowed to make minimal changes to the code that is generated. Show the Judge your working app as well as the prompt(s) used and code generated. *(5pts)*
2. Write a function that takes in a **single argument** (a text string) and **outputs** either a **-1 (negative), 0 (neutral), or 1 (positive) sentiment**. The function should call a **Large language model** (eg ChatGPT) on the given text to carry out the sentiment analysis. *(5pts)*

—---------- CALL A JUDGE—----------

LEVEL3-B (10 pts - 5 pts per task)

3. Extract the reviews from a **Google maps business review** for Dasman Diabetes Institute in Kuwait. Store the data in **csv format** where the columns are the **Author's name**, the **star review**, and **text review**. *(5pts)*

   Hint: Use web scraping or API calls to automate the process rather than doing it manually.

4. Write a function that takes in the **csv file name**, and outputs a dataframe where the columns are:
Author name, **star review**, **text review**, **sentiment** (-1, 0 or 1) *(5pts)*

—---------- CALL A JUDGE—----------

Level Four: Modeling *(5 points per task)*

LEVEL 4-A (15 pts - 5 pts per task)

Modelling:

1. Create a new dataset to look like the following figure containing (**Country, Year, Gender**, **mean total cholesterol, and GNI**) by merging relevant datasets on **Year** and **Country**(**how='inner'** is required to merge all datasets), and display the shape of the new dataset. *(5pts)*

Country	Year	Gender	mean total cholestrol	GNI
Haiti	2016	Male	4.5	1370.0
Malaysia	2016	Male	5.0	9880.0
Turkmenistan	2018	Female	4.6	6500.0
Liberia	2011	Male	3.9	500.0

2. Is there a relationship (linear relationship) between **GNI** and **mean total cholesterol?** Get the type and value of the relationship. *(5pts)*
3. Which relationship is stronger? **Mean total cholesterol** of **males** or **females with GNI** (Round the relationship value to 2 decimal places), visualise relationships. *(5pts)*

—---------- CALL A JUDGE—----------

LEVEL 4-B (15 pts - 5 pts per task)

4. Create a new dataset to look like the following figure containing (**Country, Year**, **NCD mortality, Gender**, and **GNI**) by merging relevant datasets on **Year** and **Country**(**how='inner'** is required to merge all datasets), and display the shape of the new dataset. *(5pts)*

Country	Year	NCD Mortality	Gender	GNI
Uganda	2013	18.6	Female	810.0
Timor-Leste	2006	18.2	Male	1340.0
Georgia	2017	35.4	Male	4040.0
Panama	2009	13.7	Male	6940.0
Iraq	2000	33.5	Male	1520.0

5. Create 2 plots displaying the correlation between **GNI** and **NCD mortality**
   5.1. **Before** 2010 (year < 2010)
   5.2. **After** 2010 (year > = 2010)
   5.3. In which period was the relationship the strongest? *(5pts)*

6. Which 3 countries in the Gulf region had the:
   6.1. Strongest relationship between **GNI** and **NCD mortality**
   6.2. Weakest relationship between **GNI** and **NCD mortality** *(5pts)*

—---------- CALL A JUDGE—----------

LEVEL 4-C (15 pts)

7. Create a new dataset to look like the following figure containing (**Country, Year**, **Gender**, **GNI**, and **Overweight adult's percentage**) by merging relevant datasets on **Year** and **Country** (**how='inner'** is required to merge all datasets), and display the shape of the new dataset. *(5pts)*

Country	Year	GNI	Gender	Overweight adult's percentage
Malawi	2010	440.0	Female	27.3
New Zealand	2006	26400.0	Female	55.7
Netherlands	2001	26840.0	Both sexes	49.6
Cabo Verde	1983	420.0	Male	8.5
Algeria	1980	2040.0	Both sexes	33.3

8. Build a linear regression model to predict the **Overweight adult's percentage** from **GNI (one feature).**
   8.1. Data cleaning: You must drop missing values.
   8.2. Splitting the dataset into training and test dataset:
      - **Test dataset:** Years [2016, 2015]. The test dataset size must be **960** rows.
      - **Training dataset:** All years except [2016, 2015].
   8.3. Fitting the model on a training dataset to predict **Overweight adult's percentage**.
   8.4. Evaluating the model on the test dataset by using the mean absolute error (**MAE**).The **MAE** should be **less than 12**. *(10pts)*

—---------- CALL A JUDGE—----------

LEVEL 4-D (20 pts - 5 pts per task)

Unsupervised learning:

1. Create a new dataset to look like the following figure containing (**Country, Year**, **NCD Mortality, Gender**, **BMI mean**, and **GNI**) by merging relevant datasets on **Year**, **Country** and **Gender** (**how='inner'** is required to merge all datasets), and display the shape of the new dataset. *(5pts)*

Country	Year	NCD Mortality	Gender	BMI mean	GNI
Afghanistan	2016	35.1	Male	22.3	570.0
Afghanistan	2016	36.0	Female	23.7	570.0
Afghanistan	2015	35.4	Male	22.3	610.0
Afghanistan	2015	35.7	Female	23.6	610.0
Afghanistan	2014	35.7	Male	22.2	650.0

2. Find the optimal number of clusters for clustering similar data together. *(5pts)*
3. Use an unsupervised learning algorithm to cluster data. *(5pts)*
4. Use visualisation to gain some insights from the new clusters. *(5pts)*

—---------- CALL A JUDGE—----------

LEVEL 4-E (40 pts - 5 pts per task)

Supervised learning:

1. Create a new dataset to look like the following figure containing (**Country, Year**, **Gender**, **Prevalence Hypertension**, **Mean Total Cholesterol**, **Overweight adult's percentage**, **BMI mean**, **GNI**, **Income Group**, and **NCD Mortality**) by merging relevant datasets on **Year**, **Country**, and **Gender** (**how='inner'** is required to merge all datasets), and display the shape of the new dataset. *(5pts)*

Country	Year	NCD Mortality	Gender	hypertension Prevalence	mean total cholestrol	Overweight adult percentage	BMI mean	GNI	IncomeGroup
Lesotho	2010	36.8	Female	46.5	3.9	48.8	26.2	1260.0	Lower middle income
Paraguay	2012	18.2	Male	61.1	4.6	50.7	25.9	5160.0	Upper middle income
Armenia	2011	32.0	Male	46.3	4.4	51.4	25.1	3470.0	Upper middle income
Bahamas	2012	16.9	Female	42.5	4.8	66.5	28.6	NaN	NaN
Bhutan	2016	19.4	Male	42.5	4.1	25.2	23.4	2600.0	Lower middle income

2. Build a model to predict **NCD Mortality**. *(35pts)*
   - 2.1. Data cleaning:You must **NOT** drop missing values.
   - 2.2. Feature Engineering.
   - 2.3. Selecting the features and target.
   - 2.4. Splitting the dataset into training and test dataset:
     - **Training dataset:** Years [2014, 2013, 2012, 2011, 2010, 2009].
     - **Test dataset:** Years [2016, 2015]. The test dataset size must be **620** rows.
   - 2.5. Fitting the model on the training dataset to predict **NCD Mortality**.

2.6. Tuning the model's parameters to improve the model performance.
2.7. Evaluating the model on the test dataset by using the mean absolute error (**MAE**).The **MAE** should be **less than 1.6**

—---------- CALL A JUDGE—----------