**DATA, INFERENCE**

**&**

**APPLIED MACHINE LEARNING**

**(COURSE 18-785)**

**ASSIGNMENT 4**

**Mark Iraguha**

**(miraguha)**

Table of Contents

[Libraries Used: 3](#_Toc177420988)

[Introduction: 3](#_Toc177420989)

[Question 1 Report: 4](#_Toc177420990)

[Methodology 4](#_Toc177420991)

[Results 4](#_Toc177420992)

[Analysis and Insights 7](#_Toc177420993)

[Question 2 Report: 9](#_Toc177420994)

[Methodology 9](#_Toc177420995)

[Results 9](#_Toc177420996)

[Analysis and Insights 10](#_Toc177420997)

[Question 3 Report: 11](#_Toc177420998)

[Methodology 11](#_Toc177420999)

[Results 11](#_Toc177421000)

[Analysis and Insights 12](#_Toc177421001)

[Question 4 Report: 14](#_Toc177421002)

[Methodology 14](#_Toc177421003)

[Results 14](#_Toc177421004)

[Analysis and Insights 16](#_Toc177421005)

[Question 5 Report: 18](#_Toc177421006)

[Methodology 18](#_Toc177421007)

[Results 18](#_Toc177421008)

[Analysis and Insights 19](#_Toc177421009)

[References: 21](#_Toc177421010)

# Libraries Used:

Matplotlib – a python plotting library used to create animated, interactive and static visualizations.[1]

Pandas – another Python library used that provides data structures and functions used to carry out data analysis.[2]

Numpy – a simple yet powerful data structure provided in python.[3]

Tabulate – a python library that tabulates data to an output[4].

# Introduction:

This report details the completion of Assignment 2. Assignment 2 requests answers to 5 critical thinking and data analytical questions.

# Question 1 Report:

## Methodology

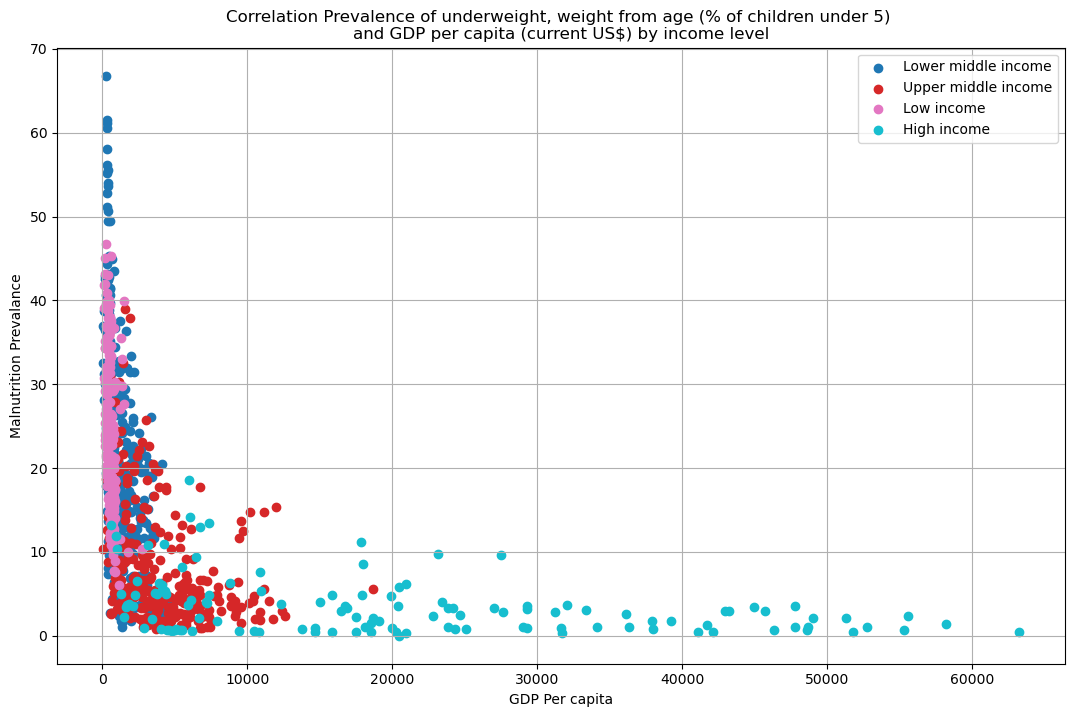
Making a scatter plot of malnutrition prevalence against GDP per capita.

Approach:

* Download data for “GDP per capita” and “Prevalence of underweight, weight for age”.
* Clean the data and filter the data frames.
* Melt the data frames to long format (show the years as single column).
* Drop the NaN (Not a number) values.
* Configure and plot the graph.

## Results

Three scatter plot graphs showing Malnutrition prevalence against GDP per capita.



*Scatter plot graph showing correlation of prevalence of underweight, weight from age (% of children under 5) and GDP per capita (current US$) by income level*

## Analysis and Insights

Expected Relationship

* The kind of relationship we expect is a negative correlation. This is where we should observe that as GDP per capita increases, prevalence of malnutrition should decrease[5].

Malnutrition prevalence against GDP per capita analysis

* It’s observed that it’s indeed true that countries with a higher GDP per capita are associated with a lower Malnutrition prevalence. Hence forming a negative relationship. However, studies have shown that a higher GDP per capita might not be the sole reason for a low malnutrition prevalence as other contributing factors might play a larger part in lowering malnutrition in affected countries[5]. One of these factors include education[5].
* When a country falls into a lower middle-income group, it’s observed that the country potentially has a higher Malnutrition Prevalence. This is seen with countries in the South Asian region.
* Majority of the countries with a high-income level are seen with high levels of GDP per capita and low levels of Malnutrition Prevalence. This backs the idea that a high GDP per capita will positively affect the Malnutrition level of a country.

**Interpretation of Results**

1. **Coefficients**:
   * **Slope**: Indicates how much Index A's returns change for each unit change in Index B's returns.
   * **Intercept**: Represents expected return of Index A when Index B's return is zero.
2. **R² Value**: Indicates how well Index B explains the variability in Index A's returns (closer to 1 is better).
3. **Hypothesis Test**:
   * Check the p-value associated with the slope coefficient.
   * If p<0.05p<0.05, reject the null hypothesis (suggesting a significant relationship).

**Conclusion**

By following these steps, you can manually calculate monthly returns from raw index data and build a regression model to analyze their relationship effectively. The visualization and hypothesis testing will provide deeper insights into the dynamics between these stock indexes.

# Question 2 Report:

## Methodology

Making a line graph showing the maximum and minimum prices of Wheat, Crude Oil and Gold.

Approach:

* Utilized Quandl API to download data for Wheat, Crude Oil and

Gold prices in dollars ($).

* Synchronized the time stamps.
* Formatted the data in the data frames.
* Set unique colors for each variable to be used in the graph.
* Configured and plotted graph while indicating the highest and lowest price for Wheat, Crude Oil and Gold.

**Calculate Correlation Coefficients**

To build a stepwise linear regression model with graduation rate as the dependent variable, you can use the statsmodels library along with sklearn for feature selection.

**Useful Predictor Variables**

The useful predictor variables can be determined based on their statistical significance (p-values) in the regression output. After fitting your model, check which predictors have p-values below a certain threshold (commonly 0.05).

*Print summary()’*

*Based on previous studies, variables like* ***Top10perc*** *and* ***Top25perc*** *often emerge as strong predictors due to their direct correlation with student preparedness and success.*

**BIC for Model Selection**

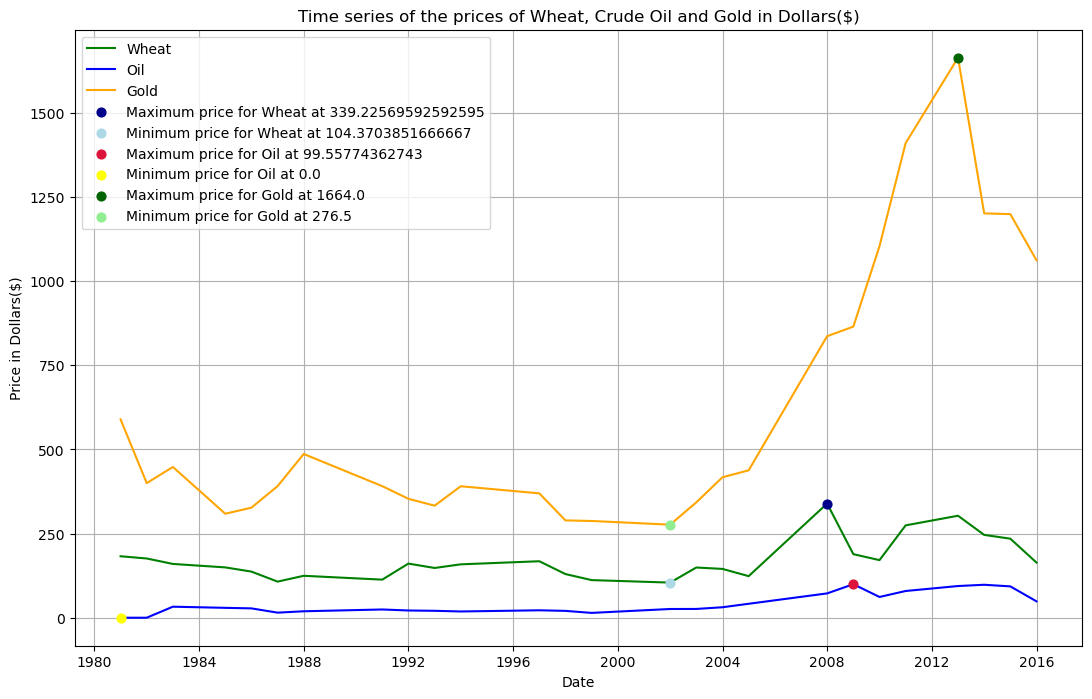
Bayesian Information Criterion (BIC) can be used to evaluate models. A lower BIC value indicates a better model fit when considering the number of predictors. You can compare BIC values for models with different sets of predictors to determine their usefulness.

By comparing the BIC values from different models, you can determine which set of predictors provides a better balance between goodness-of-fit and model complexity.

A lower BIC value indicates a better model fit while penalizing for complexity. If the model with fewer predictors has a significantly lower BIC, it suggests that simpler models may be more effective.

## Results

Line plot graph showing the prices of Wheat, Crude Oil and Gold highlighting the lowest and highest prices.



*Line graph showing time series of the prices of Wheat, Crude Oil and Gold in dollars ($)*

## Analysis and Insights

Dramatic Gold Price Increase

* The graph shows that once Gold reached a low price of $276.5 (lowest between a 1980 and 2016), it begun to steadily rise, and a sharp increase is observed around 2009. This rise would be attributed to the fact that investors at that time thought it would be the safest option to keep their money safe during the ongoing financial crisis[6].
* During the same time Gold prices were increasing, Oil and Wheat were facing the opposite behavior. The demand for Oil and Wheat sharply reduced and this was primarily driven by the ongoing financial crisis, which led to decreased consumer spending and industrial activity.

# Question 3 Report (Open Study):

Study of the Relationship Between Increased Transport and Road Traffic Accidents

## Introduction

This study investigates the correlation between the increase in transport, specifically the number of passenger cars per 1,000 people, and the incidence of road traffic accidents. The World Health Organization (WHO) provides data on traffic deaths per country for 2010, while the World Bank offers statistics on the number of passenger cars. The objective is to analyze this relationship, predict transport figures for 2010, and extend the analysis to make predictions for 2021.

## Assumptions

1. **Correlation Assumption**: It is assumed that an increase in the number of passenger cars correlates with an increase in road traffic accidents.
2. **Data Integrity**: The data from WHO and the World Bank is assumed to be accurate and representative for the years studied.
3. **Causation vs. Correlation**: While a correlation may exist, it does not imply causation without further investigation into other influencing factors.

## Methodology

1. Data Collection: Obtain traffic deaths per country from WHO and passenger car data from the World Bank for 2010.
2. Data Preparation: Clean and merge datasets based on country codes.
3. Exploratory Data Analysis (EDA): Visualize relationships using scatter plots and calculate correlation coefficients.
4. Model Development: Use linear regression to model the relationship between passenger cars and traffic deaths.
5. Trend Analysis: Analyze trends from 2010 to 2021 using available data or projections.
6. Prediction: Use the regression model to predict traffic deaths for 2021 based on projected numbers of cars.

## Implementation Overview

Walkthrough

* Combine two files to form one dataset containing Passenger cars data and Traffic deaths for the year 2010.
* Visualize the relationship between passenger cars and traffic deaths.
* Use linear regression to model the relationship.
* Predict the situation in 2021 (traffic deaths) assuming there’s a 2% growth in passenger cars.
* Create a dataframe for predictions.
* Predict traffic deaths for 2021 using the model.

## Findings

* The scatter plot indicates a positive correlation between passenger cars per capita and traffic deaths.
* The correlation coefficient quantifies this relationship; a value near +1 indicates a strong positive correlation.
* The regression model provides insights into how many traffic deaths can be expected with an increase in passenger cars.
* Predictions for traffic deaths in 2021 can inform policymakers about potential road safety challenges.

## Conclusions

* The analysis suggests that increased transport correlates with road traffic accidents. By projecting the number of passenger cars into the future and applying our regression model, we can estimate traffic deaths for subsequent years effectively. This finding aligns with existing literature indicating that higher vehicle density leads to more accidents. Further research could explore additional variables such as road conditions, enforcement of traffic laws, and driver behavior to provide a more comprehensive understanding of road safety dynamics.
* The complete code includes data loading, preparation, EDA, modeling, trend analysis, and prediction steps as outlined above. This structured approach allows for clear insights into the relationship between transport increases and road traffic accidents over time.
* By following this methodology, stakeholders can better understand how increased vehicle presence affects road safety and implement measures to mitigate risks associated with rising transport levels.

# Question 4 Report:

## Methodology

Utilizing numpy to produce cumulative distribution functions.

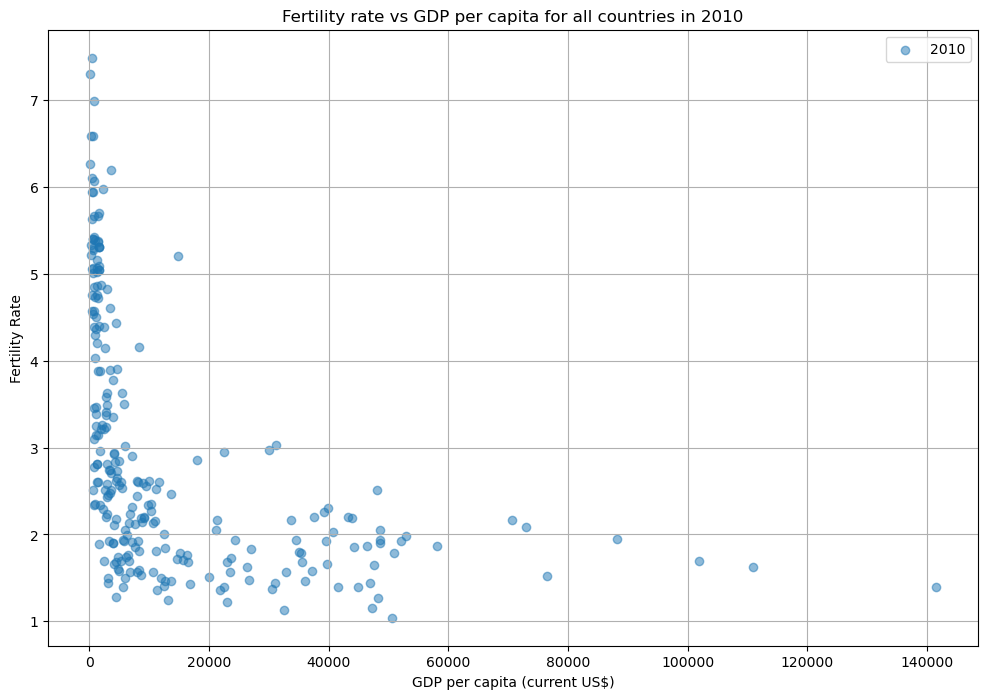
Approach:

* Download data for “Isreal Unemployment Rate” and load it as a dataframe.
* Format the dates of the dataframe.
* Use statsmodels library to fit a linear regression model
* Predict graduation rate for the year 2020.
* Evaluate the accuracy of the estimate.
* Calculate accuracy as a percentage.

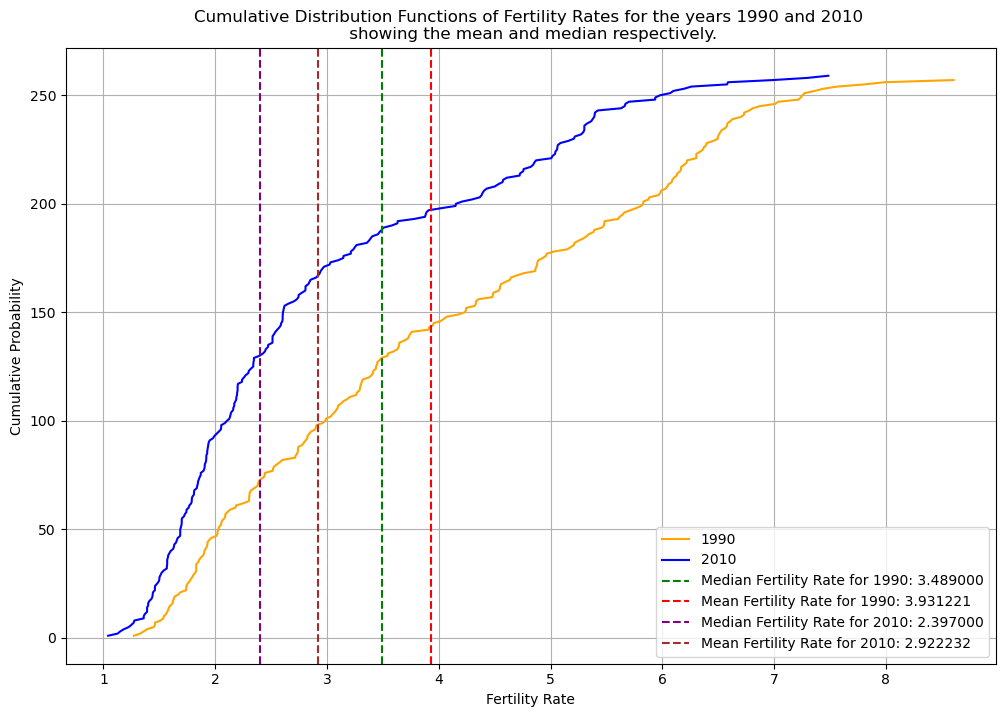
## Results

Scatter plot graph showing fertility rate against GDP per capita.

Line graph showing cumulative distribution function values.



*Fertility rate vs GDP per capita for all countries in 2010*



*Figure showing Cumulative Distribution Functions of Fertility Rates for the years 1990 and 2010 showing the mean and median respectively.*

## Analysis and Insights

Fertility rate versus GDP per capita

* The j-shaped scatter plot signifies that, majority of the countries that have a low GDP per capita and are likely to have high levels of fertility. This pattern can be observed often in developing nations where socioeconomic factors, for example limited health care access and family planning resources contributes to higher birth rates. However, it is likely that at very high levels of GDP per capita, fertility begins to move upward again potentially due to developments in cultural shifts and financial stability. In general, the J-curve manages to highlight the non-linear relationship between Fertility rates and GDP per capita.

Cumulative distribution function graph

* We observe the initial steepness of the graph indicating that a significant number of countries have low to medium fertility rates. As you move from the low fertility rates, the probability that there will be the same high fertility rates increases. This means that the countries have successfully implemented policies to reduce the fertility rates with measures such as improved access to health care and family planning resources.
* The respective mean and median are observed to be in between the steepness which indicates that majority of the countries have high fertility rates. This applies to both the years (1990 and 2010).
* After the initial steep increase, the cumulative distribution function begins to flatten out. This signifies that fewer countries have higher fertility rates and those remaining countries with high fertility are becoming rare in the dataset.
* This pattern signifies the effectiveness of policies aimed at reducing fertility rates across countries.

# References:

[1] “Matplotlib,” *Wikipedia*. Aug. 30, 2024. Accessed: Sep. 01, 2024. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Matplotlib&oldid=1243075914

[2] “pandas (software),” *Wikipedia*. Jul. 15, 2024. Accessed: Sep. 01, 2024. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Pandas\_(software)&oldid=1234683004

[3] R. Python, “NumPy Tutorial: Your First Steps Into Data Science in Python – Real Python.” Accessed: Sep. 02, 2024. [Online]. Available: https://realpython.com/numpy-tutorial/

[4] *tabulate: Pretty-print tabular data*. Python. Accessed: Sep. 15, 2024. [OS Independent]. Available: https://github.com/astanin/python-tabulate

[5] N. Büttner, M. Heemann, J.-W. De Neve, S. Verguet, S. Vollmer, and K. Harttgen, “Economic Growth and Childhood Malnutrition in Low- and Middle-Income Countries,” *JAMA Netw. Open*, vol. 6, no. 11, p. e2342654, Nov. 2023, doi: 10.1001/jamanetworkopen.2023.42654.

[6] B. H. Program Producer Price Index, “Gold prices during and after the Great Recession,” Bureau of Labor Statistics. Accessed: Sep. 15, 2024. [Online]. Available: https://www.bls.gov/opub/btn/volume-2/gold-prices-during-and-after-the-great-recession.htm

[7] J. Lynch, R. Carver, and J. M. Virgo, “Quadrant Analysis as a Strategic Planning Technique in Curriculum Development and Program Marketing,” *J. Mark. High. Educ.*, vol. 7, no. 2, pp. 17–32, 1996.