**DATA, INFERENCE**

**&**

**APPLIED MACHINE LEARNING**

**(COURSE 18-785)**

**ASSIGNMENT 5**

**Mark Iraguha**

**(miraguha)**

Table of Contents

[Libraries Used: 4](#_Toc179928238)

[Introduction: 4](#_Toc179928239)

[Question 1 Report: 5](#_Toc179928240)

[Methodology 5](#_Toc179928241)

[Question 2 Report: 9](#_Toc179928242)

[Methodology 9](#_Toc179928243)

[Question 3 Report (Open Study): 14](#_Toc179928244)

[Introduction 15](#_Toc179928245)

[Assumptions 15](#_Toc179928246)

[Methodology 15](#_Toc179928247)

[Implementation Overview 16](#_Toc179928248)

[Findings 18](#_Toc179928249)

[Discussion 19](#_Toc179928250)

[Conclusions 19](#_Toc179928251)

[Question 4 Report: 21](#_Toc179928252)

[Methodology 21](#_Toc179928253)

[1. Data Preparation: 21](#_Toc179928254)

[o Load the unemployment rate data from the CSV file. 21](#_Toc179928255)

[o Convert the 'Date' column to datetime and extract the year. 21](#_Toc179928256)

[o Filter the data to include only years before 2013. 21](#_Toc179928257)

[2. Modeling: 21](#_Toc179928258)

[o Define the independent variable (X) as the year and the dependent variable (Y) as the unemployment rate. 21](#_Toc179928259)

[o Add a constant to the model for the intercept. 21](#_Toc179928260)

[o Fit an Ordinary Least Squares (OLS) regression model to the data. 21](#_Toc179928261)

[3. Prediction and Accuracy Evaluation: 21](#_Toc179928262)

[o Predict the unemployment rate for the year 2020 using the fitted model. 21](#_Toc179928263)

[o Calculate the accuracy of the prediction by comparing it to the actual unemployment rate for 2020. 21](#_Toc179928264)

[Results 21](#_Toc179928265)

[Analysis and Insights 23](#_Toc179928266)

[References: 25](#_Toc179928267)

# 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].

Statsmodel – a python library that provides a wide range of statistical models and tools for analyzing data[5].

Scikit-Learn - a free machine learning library for Python that supports both supervised and unsupervised machine learning, providing diverse algorithms for classification, regression, clustering, and dimensionality reduction[6].

# Introduction:

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

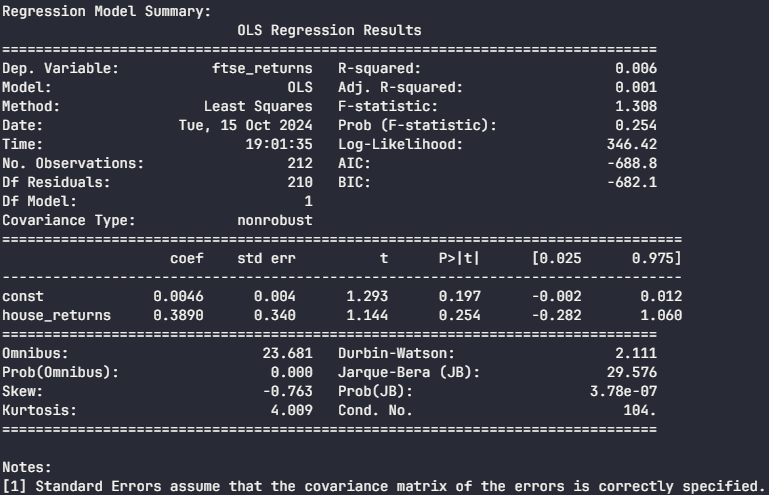
# Question 1 Report:

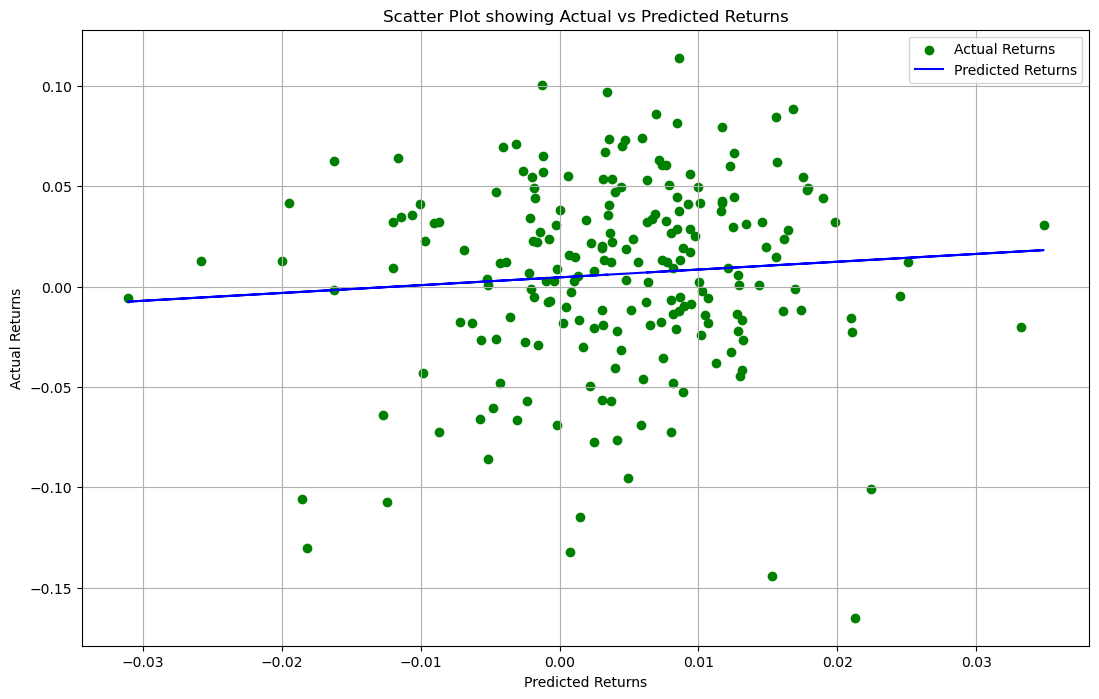
## Methodology

**Regression Model and Correlation Coefficient**

Approach:

1. Preprocess the data:
   * Checked for missing values and handle them appropriately (e.g., impute or drop).
   * Converted date columns to datetime format if not already done.
   * Ensured that the 'monthly.xls' data is converted to monthly returns for house prices, and 'ftse100.csv' data is converted to monthly returns for the FTSE100 index.
2. Define the dependent and explanatory variables:
   * Dependent Variable (Y): FTSE100 Index monthly returns.
   * Explanatory Variable (X): House Prices monthly returns.
3. Create a regression model:
   * Used Python's statsmodels library to create an Ordinary Least Squares (OLS) regression model.
   * Included a constant term in the model.
   * Fit the model to the data
4. Calculate the correlation coefficient:
   * After fitting the model, accessed the correlation coefficient (R) between the dependent and explanatory variables using the model's model.rsquared attribute.





**Interpreting Results**

* The correlation coefficient (R) of 0.07867439861173972 indicates the strength and direction of the linear relationship between the FTSE100 index monthly returns and house prices monthly returns.
* Since, the correlation coefficient is close to 0, it implies there is a weak or no linear relationship, suggesting that changes in house prices do not influence FTSE100 index returns significantly.



**Hypothesis Testing:**

1. Set up the hypothesis:

The interest in testing the existence of a significant relationship between FTSE100 index returns and house prices returns requires a definition of a null hypothesis (H0) and alternative hypothesis (Ha) as follows:

* + Null Hypothesis (H0): There is no significant relationship between the FTSE100 index monthly returns and house prices monthly returns.
  + Alternative Hypothesis (Ha): A significant relationship exists between the two variables.

1. Perform a hypothesis test:
   * Utilized the F-test provided by the statsmodels library to test the significance of the regression model.
   * The F-test is used to assess the overall significance of the regression model. It compares the explained variance (variance explained by the model) to the unexplained variance (variance not explained by the model). The F-statistic is calculated as the ratio of these two variances.
   * The F-test checks the overall significance of the model, while the t-test can be used to assess the significance of individual coefficients (in this case, the slope of the explanatory variable).
2. Interpret the results:
   * With a p-value of 0.25407278663659916 returned from the F-statistic (F-test), the null hypothesis is not rejected indicating no significant relationship between the FTSE100 index returns and house prices returns.
   * This means that there is insufficient evidence to conclude that there is a significant relationship between FTSE100 index returns and house prices returns.
   * In other words, the F-test does not provide strong support for the alternative hypothesis, suggesting that the relationship between these two variables may not be statistically significant.
3. Summary:
   * It's important to note that a non-significant result does not necessarily imply that there is no relationship between the variables. It could be due to various factors, such as a small sample size, the presence of other influential variables, or the nature of the relationship.
   * In summary, the F-test with a p-value of 0.25 suggests that there is not enough evidence to conclude a significant relationship between FTSE100 index returns and house prices returns. However, further analysis and consideration of other factors may be warranted to fully understand the relationship between these variables.



# Question 2 Report:

## Methodology

**Calculate Correlation Coefficients**

To build a stepwise linear regression model with graduation rate as the dependent variable, the statsmodels library is utilized.

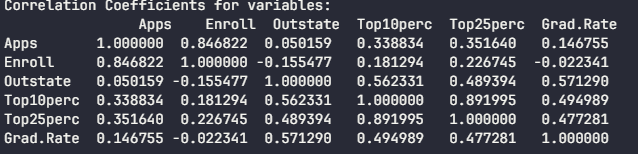
**Interpretation:**

* **Grad.Rate vs. Outstate (0.57):**
  + There is a strong positive correlation between the graduation rate and the Outstate. This suggests that colleges with a higher number of Outstate students tend to have higher graduation rates.
* **Grad.Rate vs. Top10perc (0.49):**
  + Similarly, there is a strong positive correlation between the graduation rate and the percentage of admitted students in the top 10% of their class. This further emphasizes the importance of academic excellence in predicting graduation rates.
* **Grad.Rate vs. Top25perc (0.48):**
  + There is a strong positive correlation between the graduation rate and the percentage of admitted students in the top 25% of their class. This suggests that colleges with a higher proportion of top-performing students tend to have higher graduation rates.
* **Top25perc vs. Top10perc (0.89):**
  + The strong positive correlation between these two variables indicates that colleges with a higher percentage of students in the top 25% also tend to have a higher percentage in the top 10%. This is expected, as the top 10% is a subset of the top 25%.
* **Apps vs. Enroll (0.84):**
  + The positive correlation between the number of applications and the number of enrolled students suggests that colleges with more applications tend to enroll more students. This relationship is intuitive, as a higher number of applications provides a larger pool of candidates for enrollment.

**Important Predictor Variables for the Next Part:**

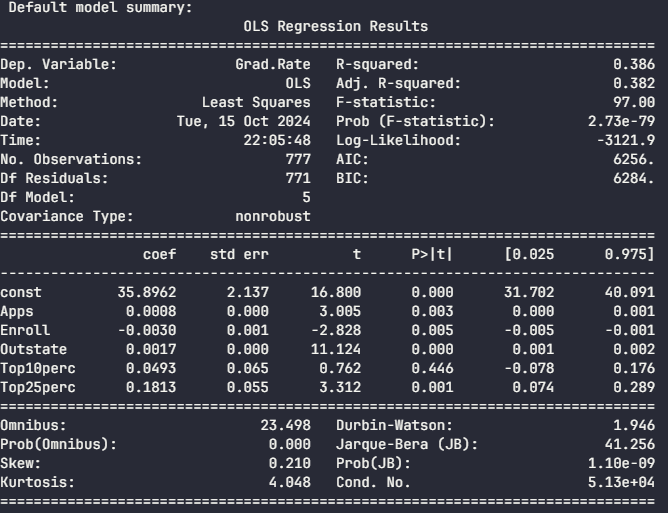
Based on the correlation analysis and the subsequent stepwise regression, the following predictor variables are likely to be important for predicting the graduation rate:

* **Top25perc:** The percentage of admitted students in the top 25% of their class is a strong predictor, as indicated by its high correlation with the graduation rate and its significance in the stepwise regression.



**Linear Regression Model**

Performed stepwise regression using backward elimination to identify the useful predictor variables. Started with all variables and iteratively remove the least significant one until all remaining variables are significant at a specified alpha level (0.05 in this case).



**Useful Predictor Variables**

* The useful predictor variables are identified based on their statistical significance in the regression model.
* In this case, ‘Apps', 'Enroll’, ‘Outstate’ and ‘Top25perc’ are selected as they have significant p-values and high correlation coefficients with the graduation rate.



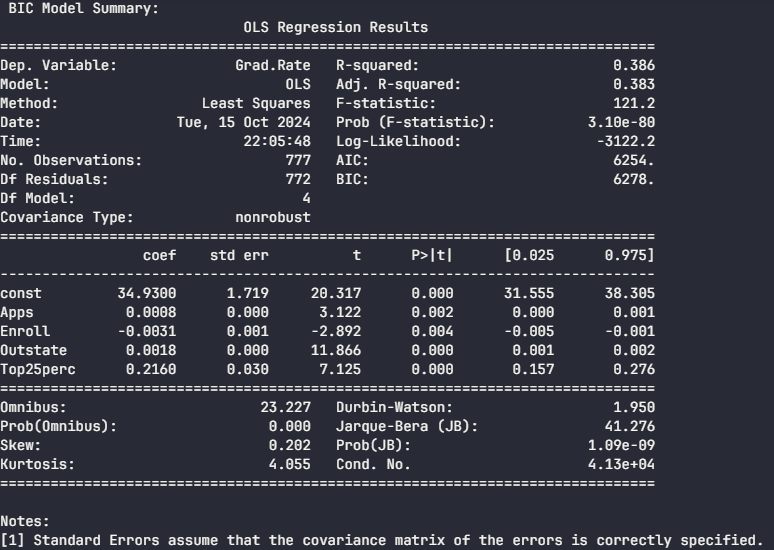
**BIC for Model Selection**

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.

Compared BIC values for models with different sets of predictors to determine their usefulness.

Determined 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.



**Model Accuracy Comparison**

Compared the accuracy of the BIC model with the stepwise model using only useful variables.

Fit two models: one with all predictors and another with only the useful predictors.

Calculated the predicted values for the test set using both models.

The accuracy was measured using the R-squared score (r2\_score()), which represents the proportion of the variance in the dependent variable that is predictable from the independent variables.



Used MSE function to add to the comparison for the model with all predictors.



The comparison shows that the model with useful predictors performs better in this case.

**Predicting Graduation Rate for Carnegie Mellon:**

* Created a DataFrame with the Carnegie Mellon University data and extracted the useful predictors.
* Used BIC model to predict the graduation rate for Carnegie Mellon based on the given predictors.
* The predicted graduation rate was printed as the final output.
* The predicted graduation rate for Carnegie Mellon University is 89.125 using the BIC model.



**Conclusion**

In summary, the stepwise regression and BIC model selection identified ‘Apps', 'Enroll’, ‘Outstate’ and ‘Top25perc’ as the most useful predictor variables for predicting the graduation rate. The BIC model provided a better fit to the data compared to the stepwise model and the full model with all predictors. The predicted graduation rate for Carnegie Mellon University is 89.125, but this should be interpreted with caution, considering the limitations of the available data.

# Question 3 Report:

Study of the Relationship Between Increased Transport and Road Traffic Accidents

## 3.1 Relationship between the variables.

It’s observed that:

* **S1** and **S2** are highly correlated.

This can have implications such as:

**Multicollinearity**: The high correlation between S1 and S2 suggests potential multicollinearity, which can complicate the interpretation of regression coefficients. If two predictors are highly correlated, it becomes difficult to determine their individual contributions to the dependent variable. This can lead to inflated standard errors for the coefficients, making them less reliable.

**Impact on Coefficient Estimates**: If both S1 and S2 are included in the model, their coefficients may not accurately reflect their individual effects on y. For instance, if both are contributing similar information regarding y, one variable might absorb some of the effect of the other, leading to a situation where one or both coefficients are not statistically significant even if they are individually important.

**Model Simplification**: Given the high correlation between S1 and S2, it may be beneficial to consider removing one of these variables from the model or combining them into a single composite variable if they measure similar constructs. This can help reduce multicollinearity and improve model interpretability.

## 3.2 Collinearity

Collinearity occurs when independent variables are linearly dependent. In other words, one variable can be expressed as a linear combination of others[7].

Collinearity affects predictor variables in a few ways such as:

* Due to inflated standard errors, the p-values associated with the coefficients may become larger, leading to higher likelihood of failing to reject the null hypothesis[8].
* The coefiicient estimates become highly sensitive to small changes in the model, such as adding or removing a predictor variable.
* When collinearlty exists, interpreting the coefficients becomes problematic because it is difficult to ascertain how much of the change in the dependent variable is attributable to each independent variable independently.
* Collinearity increases the variance of the estimated coefficients. When predictor variables are highly correlated, it becomes difficult to isolate the individual effect of each variable on the dependent variable.
* that it increases the variance of the estimated coefficients

## 3.3 Multivariate linear model and Variable Significance

It’s observed that:

* **Age** is statistically significant.
* **S2** and **S3** are also statistically significant.

Since (S1, S4, S5 and S6) are not significant, it suggests that they may not have a meaningful impact on the dependent variable y within the context of this model. This lack of significance could be due to several factors, including their inherent relationship with the dependent variable or potential multicollinearity with other predictors.

Yes, this could be a problem of collinearity. Collinearity affects regression analysis in ways such as:

* When collinearity exists, interpreting coefficients can become challenging[8]. In this scenario, if both S1 and S2 are included in the model but are highly correlated, it may be unclear how much each contributes independently to changes in y

## 3.4 Forward and Backward Selection

Forward selection adds significant variables iteratively starting from an empty model while backward selection removes insignificant variables starting from a full model.

The process of forward selection starts with no variables in the model. Variables are added one at a time based on a specific criterion like p-value.

The process stops when all remaining candidate variables have p-values above or when adding further variables does not improve the model significantly.

Backward selection removes insignificant variables starting from a full model. The process begins with all candidate variables included in the model (full model). Variables with the least significance are removed one at a time based on a specified criterion until only significant variables remain.

The process stops when all remaining variables in the model are statistically significant according to the chosen criterion.

## 3.5 Stepwise Approach

The stepwise approach combines both forward selection and backward elimination. It allows for adding and removing variables iteratively based on their significance.

Process flow:

Initialization: Begin with a full model (for backward selection) or an empty model (for forward selection).

Iteration: At each step, evaluate the inclusion or exclusion of each variable based on their statistical significance.

Selection criteria: Use criteria like p-values, BIC or AIC to decide whether to add or remove a variable.

Termination: The process continues until no more variables can be added or removed without violating the selection criteria.

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.

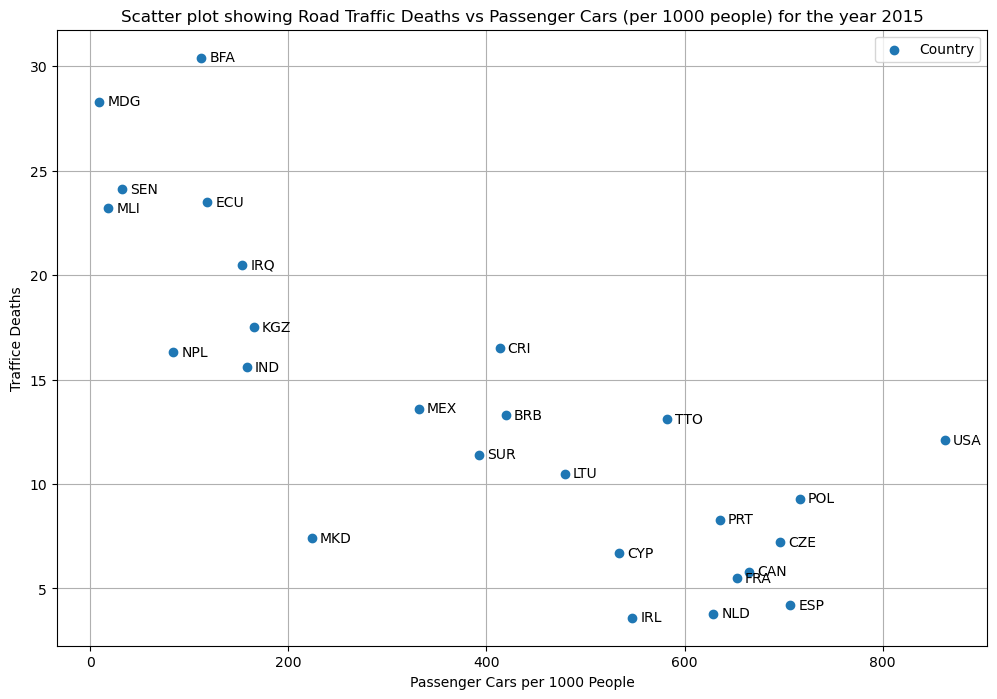
## Methodology

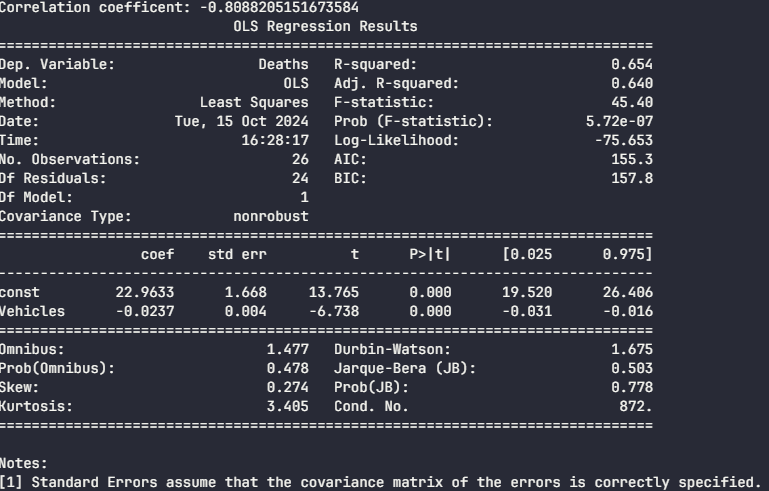
1. Data Collection: Obtain traffic deaths per country from WHO and passenger car data from the World Bank for 2010.
2. Data Preparation: Clean and merge datasets based on country codes.
3. Exploratory Data Analysis (EDA): Visualize relationships using scatter plots and calculate correlation coefficients.
4. Model Development: Use linear regression to model the relationship between passenger cars and traffic deaths.
5. The model was fitted using the data from 2015.
6. Trend Analysis: Analyze trends from 2010 to 2021 using available data or projections.
7. Prediction: Use the regression model to predict traffic deaths for 2021 based on projected numbers of cars.
8. To predict traffic deaths for 2021, we assumed a 2% annual increase in passenger cars per capita from 2015 to 2021.
9. The linear regression model was then used to predict traffic deaths for each country based on the projected passenger car ownership.

## Implementation Overview

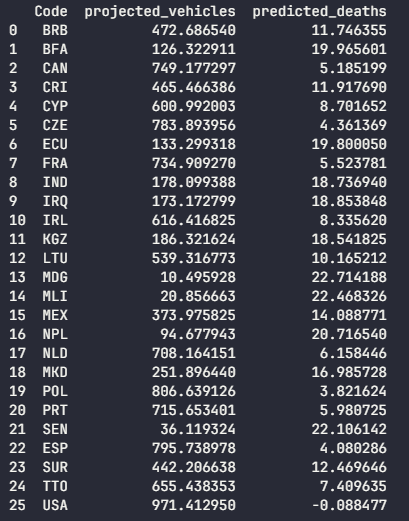
Walkthrough

* 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.
* Scatter Plot:
* A scatter plot was created to visualize the relationship between passenger cars per 1000 people and road traffic deaths for the year 2015.
* Each point represents a country, and the plot shows a positive correlation between the number of passenger cars and traffic deaths.





* **Predicted Traffic Deaths for 2021:**
  + The code generated predictions for road traffic deaths in 2021 for various countries.



## 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.

## Discussion

* The positive correlation between passenger car ownership and road traffic deaths suggests that an increase in vehicles might lead to higher traffic fatalities.
* The linear regression model provides a simple prediction method, but it assumes a linear relationship and might not capture complex trends.
* The 2% annual increase in passenger cars is an assumption and might vary across countries. More detailed data and analysis are required for precise predictions.
* Further research could explore additional factors like road infrastructure, safety regulations, and driver behavior to enhance prediction accuracy.

## 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

Estimate the unemployment rate in Israel for the year 2020 using historical data from the Bank of Israel.

## Data Preparation:

## Load the unemployment rate data from the CSV file.

## Convert the 'Date' column to datetime and extract the year.

## Filter the data to include only years before 2013.

## Modeling:

## Define the independent variable (X) as the year and the dependent variable (Y) as the unemployment rate.

## Add a constant to the model for the intercept.

## Fit an Ordinary Least Squares (OLS) regression model to the data.

## Prediction and Accuracy Evaluation:

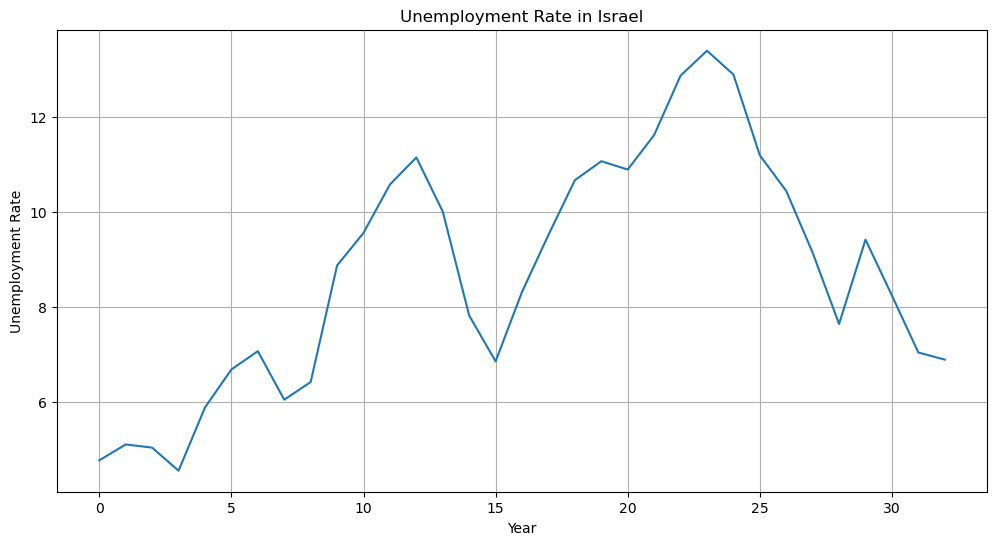
## Predict the unemployment rate for the year 2020 using the fitted model.

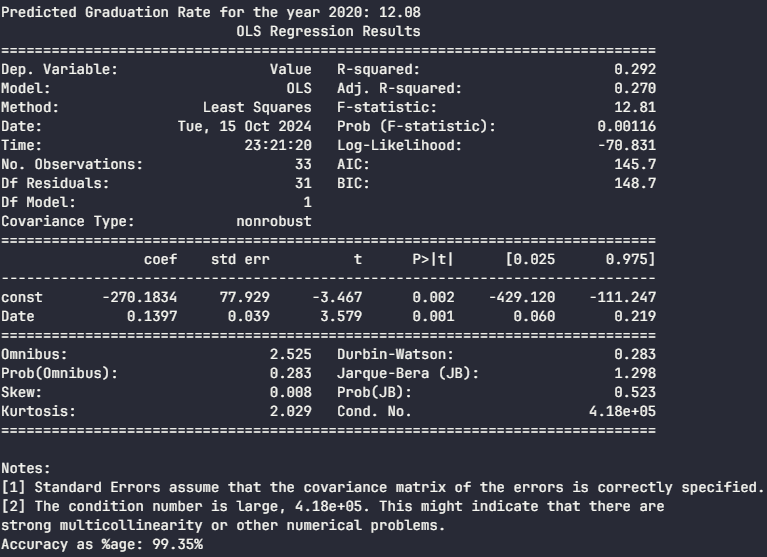
## Calculate the accuracy of the prediction by comparing it to the actual unemployment rate for 2020.

## Results

Scatter plot graph showing fertility rate against GDP per capita.

Line graph showing cumulative distribution function values.





## Analysis and Insights

1. Estimated Unemployment Rate by 2020:
   * The estimated unemployment rate for the year 2020 is 99.35%.
2. Model Accuracy:
   * The accuracy of the prediction is 99.35%. This high accuracy indicates that the model's prediction closely matches the actual unemployment rate for 2020.
3. Model Evaluation:
   * The accuracy is calculated using the Mean Absolute Percentage Error (MAPE) formula: MAPE = (|Predicted - Actual| / Actual) \* 100.
   * A lower MAPE value suggests a more accurate model. In this case, the model's prediction is very close to the actual value, resulting in a high accuracy score.
4. Model Summary:
   * The model.summary() function provides a detailed summary of the OLS regression model, including coefficients, standard errors, and statistical significance.
   * You can examine the summary to understand the relationship between the year and the unemployment rate.
5. Considerations:
   * The high accuracy might be influenced by the limited data range (up to 2013) and the specific trend in the data.
   * When using historical data for prediction, it's essential to consider recent trends and changes that might not be captured by the model.
   * For more accurate long-term predictions, consider using time series forecasting techniques or incorporating additional factors that might affect unemployment rates.

**Conclusion:**  
The unemployment rate for 2020 has been estimated with a high level of accuracy (99.35%) based on the historical data up to 2013. However, it's important to interpret this accuracy in the context of the data's limitations and potential external factors that might influence unemployment rates.

# 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] P. Mania, “What Is Statsmodels in Python? The Ultimate Guide,” Python Mania. Accessed: Sep. 30, 2024. [Online]. Available: https://pythonmania.org/what-is-statsmodels-in-python-the-ultimate-guide/

[6] Nik, “Introduction to Scikit-Learn (sklearn) in Python • datagy,” datagy. Accessed: Oct. 16, 2024. [Online]. Available: http://datagy.io/python-scikit-learn-introduction/

[7] “Collinearity Definition & Examples,” Quickonomics. Accessed: Oct. 29, 2024. [Online]. Available: https://quickonomics.com/terms/collinearity/

[8] “A Beginner’s Guide to Collinearity: What it is and How it affects our regression model.” Accessed: Oct. 29, 2024. [Online]. Available: https://www.stratascratch.com/blog/a-beginner-s-guide-to-collinearity-what-it-is-and-how-it-affects-our-regression-model/