Name: Niyomwungeri Parmenide ISHIMWE

Andrew-ID: parmenin

DATA, INFERENCE, AND APPLIED MACHINE LEARNING

18-785

ASSIGNMENT 4

24 OCTOBER 2022

I, the undersigned, have read the entire contents of the syllabus for course 18-785 (Data Inference and Applied Machine Learning) and agree with the terms and conditions of participating in this course, including adherence to CMU's AIV policy.

Signature: **Niyomwungeri Parmenide ISHIMWE**
Andrew ID: **parmenin**
Full Name: **Niyomwungeri Parmenide ISHIMWE**

The libraries used:

- import numpy as np
- import pandas as pd
- import matplotlib.pyplot as plt
- from scipy import stats
- from sklearn.linear_model import LinearRegression, LassoLarsIC
- from sklearn.metrics import r2_score, mean_absolute_percentage_error
- import statsmodels.api as sm
- import quandl
- import datetime

QUESTION 1:

It was required to create a regression model including a constant, to calculate the correlation coefficient, to interpret the results, and to give a conclusion about the monthly House Prices in the UK and the FTSE100 index from Yahoo Finance over the period of 01-Jan-1991 to 31-Dec-2016. This is done by first reading the data for both datasets, renaming the unnamed column, filtering data within the specified date range, and sorting them by date.

The next step is calculating the monthly returns for both datasets using the pandas pct_change function, then building a regression model using the **linregress** function from SciPy with the FTSE index as a dependent variable and monthly house prices as the independent variable. That helps to get the slope of 0.09324142754349982, intercept or the constant of 0.0040478376866624555, Pearson correlation coefficient of 0.02655129570190995, a p-value of 0.6409049000031647, and the Standard error of the estimated slope of 0.19970586443555416**.** From that the regression equation can be written as **y = 0.0040478376866624555 + 0.09324142754349982x**.

In addition, the **LinearRegression** function from sklearn.linear_model was used to build the model, then use the model's fit function to train it or fit it with our data, and then the prediction is made using

the predict function of the model to get a prediction of the monthly returns for FTSE using the given monthly returns for house prices. Furthermore, the scatter plot of the actual data is made, and then a line plot of the predicted data.

From that process and the graph below, the correlation coefficient of **0.02655129570190995** was returned. This correlation coefficient shows that there is a very weak positive relationship, which can be said as no relationship as it was very close to zero. This tells that the FTSE100 index and Monthly House Prices cannot be used for linear regression i.e., Monthly House Prices cannot be used to predict future values of the FTSE100 index.

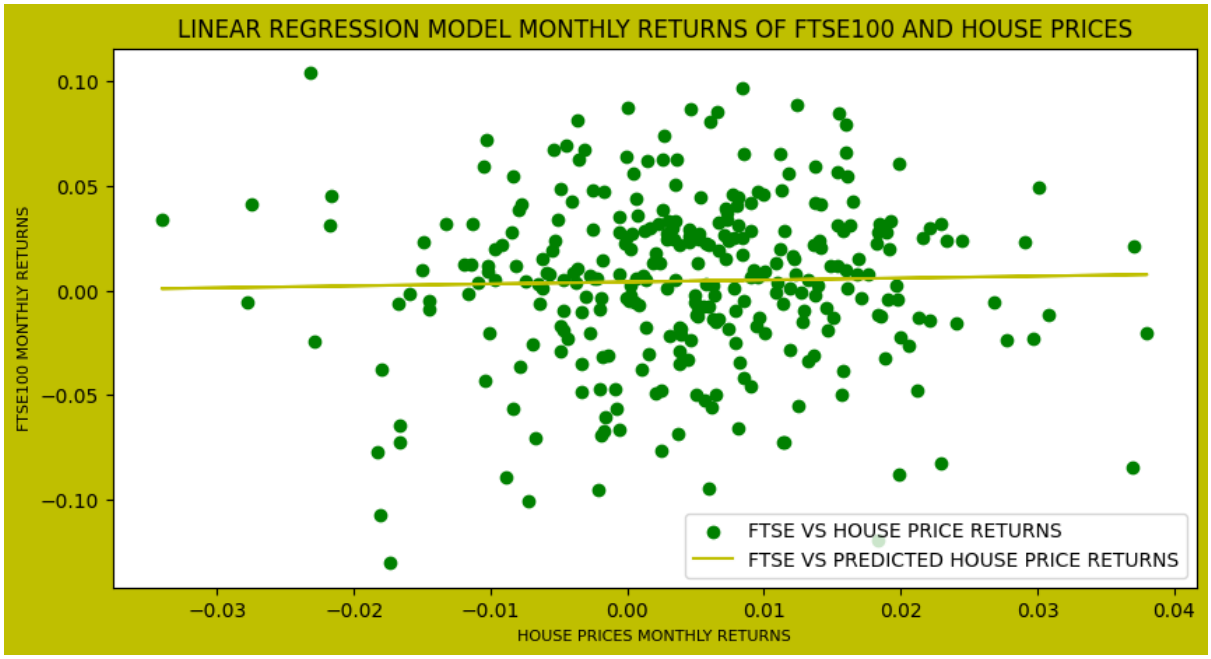

Figure 1: Linear regression model of monthly returns for FTSE100 and house prices

Moreover, a null hypothesis is that "There is no linear relationship between the FTSE100 index and House Prices" and hence the alternative hypothesis is that there is a linear relationship between the FTSE100 index and House Prices. From that, we can infer that a two-tailed hypothesis is an appropriate test because the alternative hypothesis contains the not equal "≠" sign. From the result of the p-value of **0.6409049000031647,** it can be concluded that deviating from the null hypothesis is not statistically significant, and the null hypothesis should not be rejected since the p-value is greater than the significance level of 0.05(p-value > α).

QUESTION 2:

The information about different US colleges and universities including the number of applications received, the number of enrolled students, the number of out-of-state students, the number of admitted students who were in the top 10%, and the number of admitted students who were in the top 25% of their class was provided to predict the graduation rate. This was addressed by first calculating the correlation coefficients between the five variables and the graduation rate and they are depicted in the tables below:

The correlation coefficients between all the variables is:

	Apps	Enroll	Outstate	Top10perc	Top25perc	Grad.Rate
Apps	1.000000	0.846822	0.050159	0.338834	0.351640	0.146755
Enroll	0.846822	1.000000	-0.155477	0.181294	0.226745	-0.022341
Outstate	0.050159	-0.155477	1.000000	0.562331	0.489394	0.571290
Top10perc	0.338834	0.181294	0.562331	1.000000	0.891995	0.494989
Top25perc	0.351640	0.226745	0.489394	0.891995	1.000000	0.477281
Grad.Rate	0.146755	-0.022341	0.571290	0.494989	0.477281	1.000000

Figure 2a: Correlation coefficients between all variables

The correlation coefficients of the 5 variables is:

Apps	0.146755
Enroll	-0.022341
Outstate	0.571290
Top10perc	0.494989
Top25perc	0.477281
Grad.Rate	1.000000

Figure 2b: Correlation coefficients of between 5 variables and the graduation rate

After that, the forward stepwise selection is done to select the best predictors to use with graduation rate as the dependent variable to build a linear regression model. This process involved the use of Ordinary Least Squares (OLS) that compare variables, by comparing their p-values against each other to find the minimum and against the threshold of 0.05, then store it in the inclusion list of the best predictors and continue the process until the conditions are no longer valid. Furthermore, the process

returned 'Outstate', and 'Top25perc' as the only appropriate variables to use within this model to predict the graduation rate.

Moreover, using BIC to select the model for predicting the graduation rate, not all variables are useful, 'Apps', 'Enroll', 'Outstate', and 'Top25perc' are the only useful variables. This is obtained using the BIC (Bayesian Information Criterion) model built using the **LassoLarsIC** function from sklearn.linear_model with criterion = 'bic', normalize = False parameters to return the most inappropriate variable that should be removed. This model returned the Top10perc variable as not useful because its coefficient was zero.

To compare the accuracy of these models, predictions of the graduation rate using these models need to be done. These predictions help to evaluate the model accuracy using the actual graduation rate and the predicted graduation rate. After calculating the accuracies by subtracting the MAPE (Mean Absolute Percentage Error calculated using **mean_absolute_percentage_error** function from sklearn.metrics) from 100, it is clear that the model with all five variables is more accurate than any other model with an approximate accuracy score of **80.95%**. The stepwise model with 'Outstate', and 'Top25perc' variables is the last with an approximate accuracy score of **80.79%,** and the BIC model with 'Apps', 'Enroll', 'Outstate', and 'Top25perc' variables is the second with an approximate accuracy score of **80.93%**.

Moreover, the model with all 5 variables has the greatest coefficient of determination (R^2 calculated using **r2_score** function from sklearn.metrics) of **0.3861582005130556,** the second is the BIC model with 4 variables with an R^2 of **0.3856960170430921,** and lastly the stepwise model with an R^2 of **0.37776441749868717.** In addition, by calculating the Bayesian Information Criterion (BIC) of these models, the model with 5 variables has approximately **6283**, the BIC model with 4 variables has **6277**,, and lastly the stepwise model with 2 variables with **6274.** All those three criteria, clearly show that the model with all five variables is the more accurate to predict the graduation rate.

Finally, the model with all five variables is used to predict the graduation rate for Carnegie Mellon University which resulted in **89.20112305346854%.**

QUESTION 3:

As it was required to undertake a study in the domain of transport for one or more countries based on publicly available data by using mathematical facts, then predicting the situation in 2022, and then turning in the report, this study had been undertaken on how the increase in annual Air transport, freight (million ton-km) affects the increase in annual CO2 emission from fuel in Egypt. The methodology for this is to first download the air transport and CO2 emissions by fuel datasets from https://ourworldindata.org/search?q=transport and https://ourworldindata.org/emissions-by-fuel respectively, reading them, filtering the data for Egypt for the period between 1970 and 2020, dropping unnecessary columns and then, one plot showing the two trends is made and is shown next.

The assumption of this study is that there is a linear relationship between air transport as the independent variable and CO2 emissions as the dependent variable. It is assumed that the increase in air transport affects the increase in CO2 emissions. The correlation coefficient of **0.9246357022655756** shows that there is a positive linear relationship between these variables.

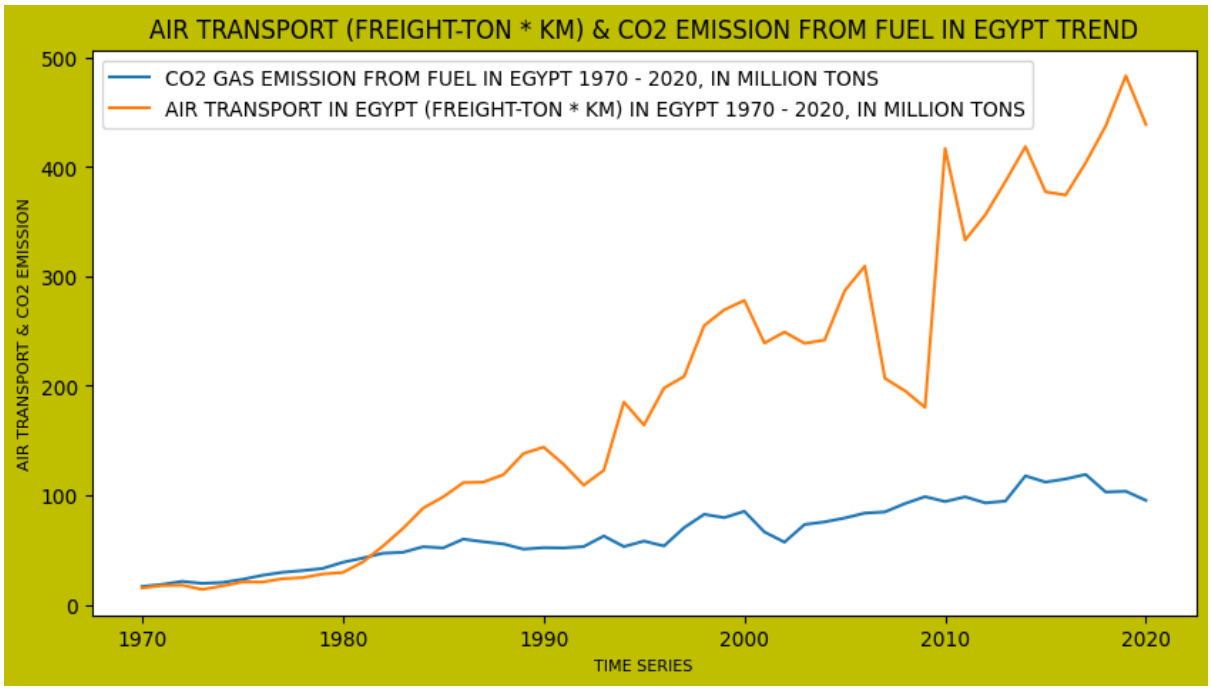

Figure 3: The trend for Air transport and CO2 emission from fuel in Egypt (1970-2020)

After that, a correlation coefficient between the air transport, freight, and CO2 emission from the fuel is calculated as **0.9246357022655756,** which is very close to 1 and it indicates that there is a strong positive relationship between the two variables. Moreover, a model is built using Air transport, and freight (million ton-km) in Egypt as the independent variable and Annual CO2 emissions from oil in Egypt as the dependent variable, then trying to predict the CO2 emissions using the built model and then visualizing actual data on a scatter plot and the model prediction with a line plot on the same plot and it is depicted below. This model has an accuracy score of mean absolute percentage error as low as **17.517510550644687%.**

```
The summary statistic is for air transport is:
 count            51.00
mean      190121457.36
std       141151376.28
min        13899999.62
25%        61599998.47
50%       180309997.56
75%       282507507.32
max       483413879.39
Name: Air transport, freight (million ton-km), dtype: object

The summary statistic is for CO2 emission is:
 count            51.00
mean       64789591.06
std        29106720.68
min        16587787.00
25%        47426816.00
50%        58103712.00
75%        88862133.00
max       119036032.00
Name: Annual CO2 emissions from oil, dtype: object
```

Figure 4: Summary statistics for Air transport and CO2 emission from fuel in Egypt

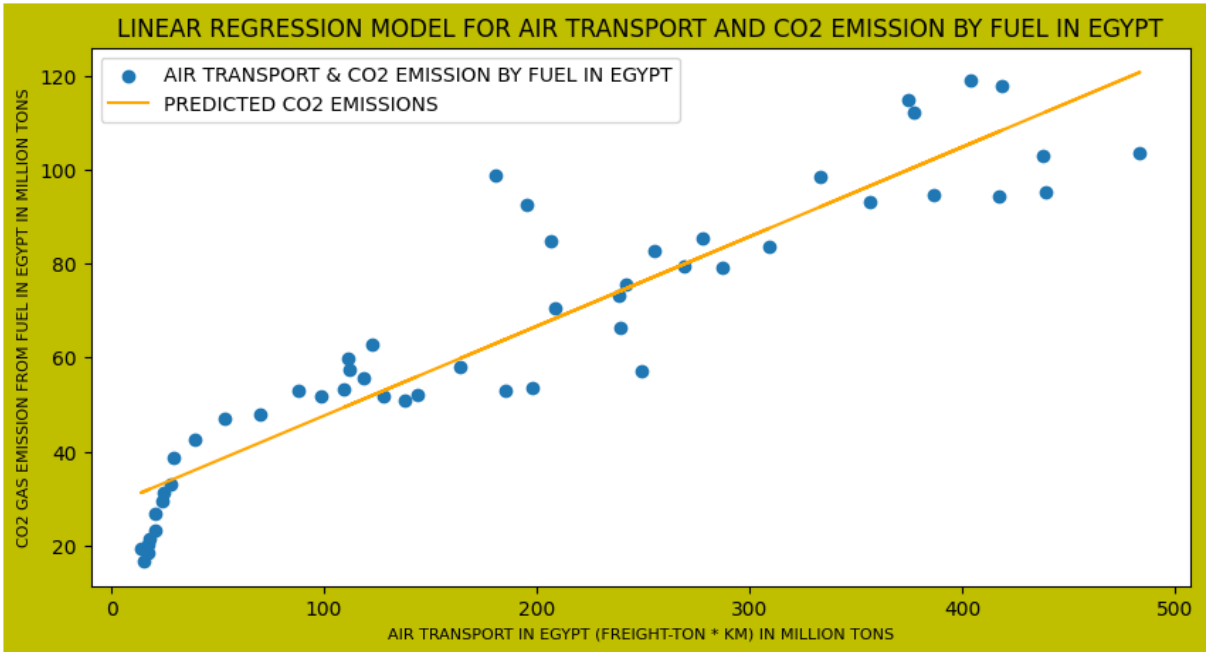

Figure 5: Linear regression model for Air transport and CO2 emission from fuel in Egypt

To predict the CO2 emissions from fuel in 2021, a linear regression model was used with the years range from 1970 to 2020 as an independent variable while the CO2 emissions for that year's range were also used to build and fit the model, and that can predict the CO2 emissions with a low accuracy MAPE score of **9.992150249020963%.** This model is hence used to predict the situation in 2022 and it was forecasted that CO2 emissions in 2021 is **113873565.4094119** and **115761410.57674217** in 2022 in Egypt. In addition, this statistic has a t-statistic of **6.210389082**, and a p-value of **0.000000012** which is very lower compared to the significance level of 0.05 i.e., the null hypothesis is to be rejected.

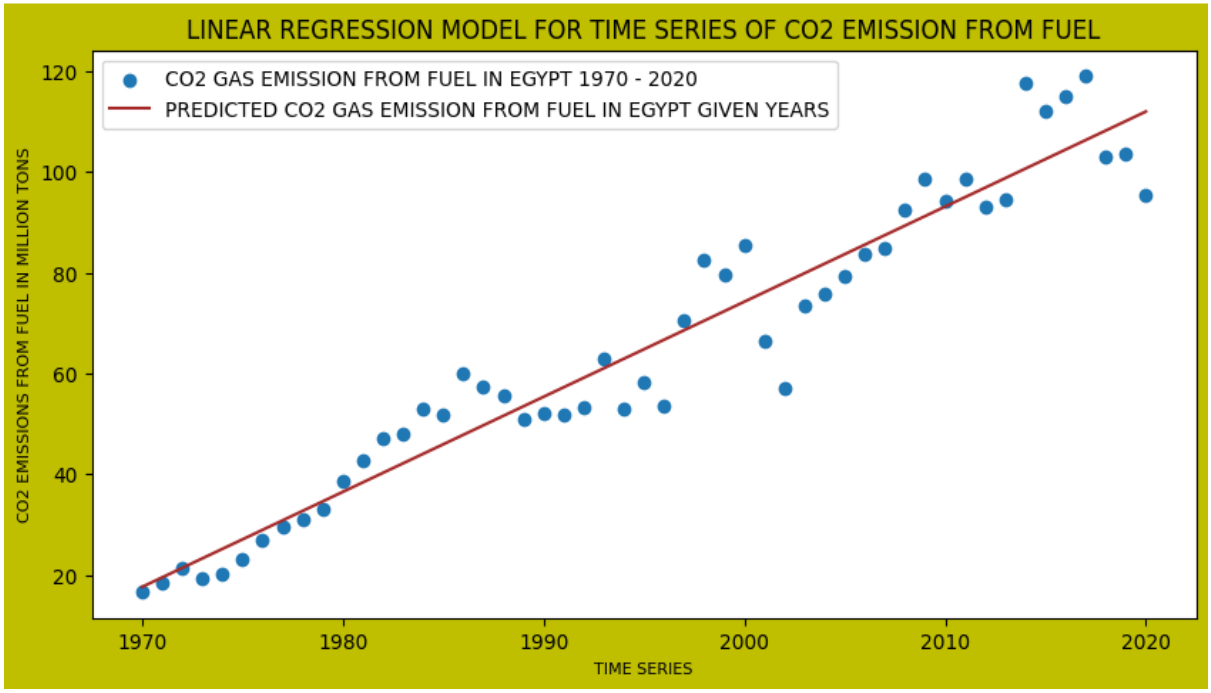

Figure 6: Linear regression for CO2 emission from fuel in Egypt with time series

QUESTION 4:

It was asked to estimate the unemployment rate in Israel from 1980-12-31 to 2013-09-02 by using the data called unemployment rate (per 100 Israeli workforces) from Quandl. This is done by reading the data using the Quandl API with code: ODA/ISR_LUR, filtering the data between the 1980-12-31 and 2013-09-02 period, removing the date as the index, changing the date to their corresponding datums (ordinal representation) as integers, and reshaping them and use them along the with the values for unemployment to fit the built linear regression model with dates as independent variable and the unemployment rate values as the dependent variable. This model is then used to predict the unemployment rate given the dates and then a scatter plot with actual data and a line plot(regression) of the model are illustrated below.

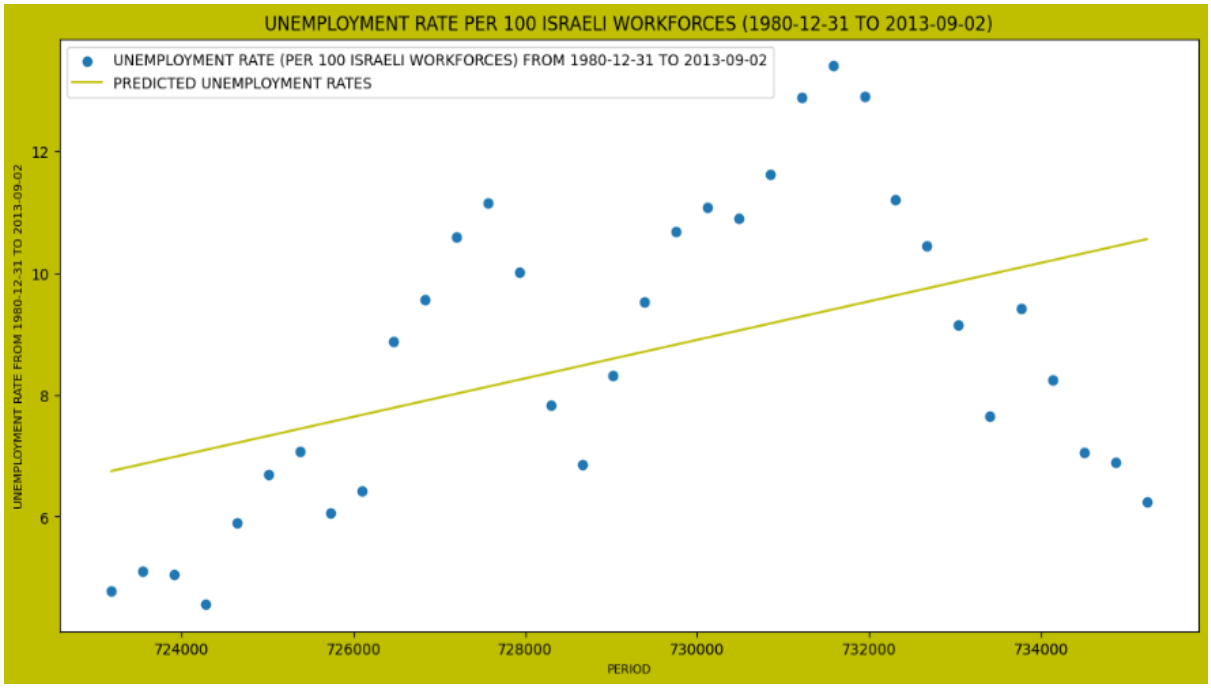

Figure 6: The linear regression for the unemployment rate in Israel (1980-2013)

This model estimated that the unemployment rate in 2020 will be **11.36127564243202** per 100 Israeli workforces. The accuracy of this estimate can be evaluated using the Mean Absolute Percentage Error (MAPE) which measures and defines the machine learning model accuracy using the absolute difference between the actual and predicted values, divided by the actual value to obtain the error of a given model. This means that the lower the MAPE, the better the model. Moreover, the MAPE for this model is **23.710851338953358%** which seems to be low, and this implies that the accuracy is **76.28914866104664 %**.

REFERENCES

1. Miko and Charles, “Miko,” Real Statistics Using Excel. [Online]. Available: https://www.real-statistics.com/descriptive-statistics/assumptions-statistical-test/. [Accessed: 23-Oct-2022].
2. Real Python, “Linear regression in python,” Real Python, 24-May-2022. [Online]. Available: https://realpython.com/linear-regression-in-python/. [Accessed: 23-Oct-2022].
3. G. Toth, “Feature selection methods with python,” DataSklr, 23-Dec-2020. [Online]. Available: https://www.datasklr.com/ols-least-squares-regression/variable-selection. [Accessed: 23-Oct-2022].
4. AakkashVijayakumar, “AakkashVijayakumar/stepwise-regression: Project uploaded to Python Community,” GitHub. [Online]. Available: https://github.com/AakkashVijayakumar/stepwise-regression. [Accessed: 23-Oct-2022].
5. nlahrinlahri 13111 gold badge11 silver badge33 bronze badges and David DaleDavid Dale 1, “How to do stepwise regression using sklearn?,” Data Science Stack Exchange, 01-Feb-1965. [Online]. Available: https://datascience.stackexchange.com/questions/24405/how-to-do-stepwise-regression-using-sklearn/24447#24447. [Accessed: 23-Oct-2022].
6. “Lasso model selection: AIC-BIC / Cross-validation,” scikit. [Online]. Available: https://scikit-learn.org/stable/auto_examples/linear_model/plot_lasso_model_selection.html. [Accessed: 23-Oct-2022].
7. “Search,” Our World in Data. [Online]. Available: https://ourworldindata.org/search?q=transport. [Accessed: 23-Oct-2022].
8. H. Ritchie, M. Roser, and P. Rosado, “CO2 emissions by Fuel,” Our World in Data, 11-May-2020. [Online]. Available: https://ourworldindata.org/emissions-by-fuel. [Accessed: 23-Oct-2022].
9. “Scipy.stats.ttest_ind#,” scipy.stats.ttest_ind - SciPy v1.9.3 Manual. [Online]. Available: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_ind.html. [Accessed: 23-Oct-2022].
10. “MAPE - mean absolute percentage error in Python,” AskPython, 16-Oct-2020. [Online]. Available: https://www.askpython.com/python/examples/mape-mean-absolute-percentage-error#:~:text=What%20is%20MAPE%3F,termed%20by%20the%20model%20evaluation. [Accessed: 23-Oct-2022].