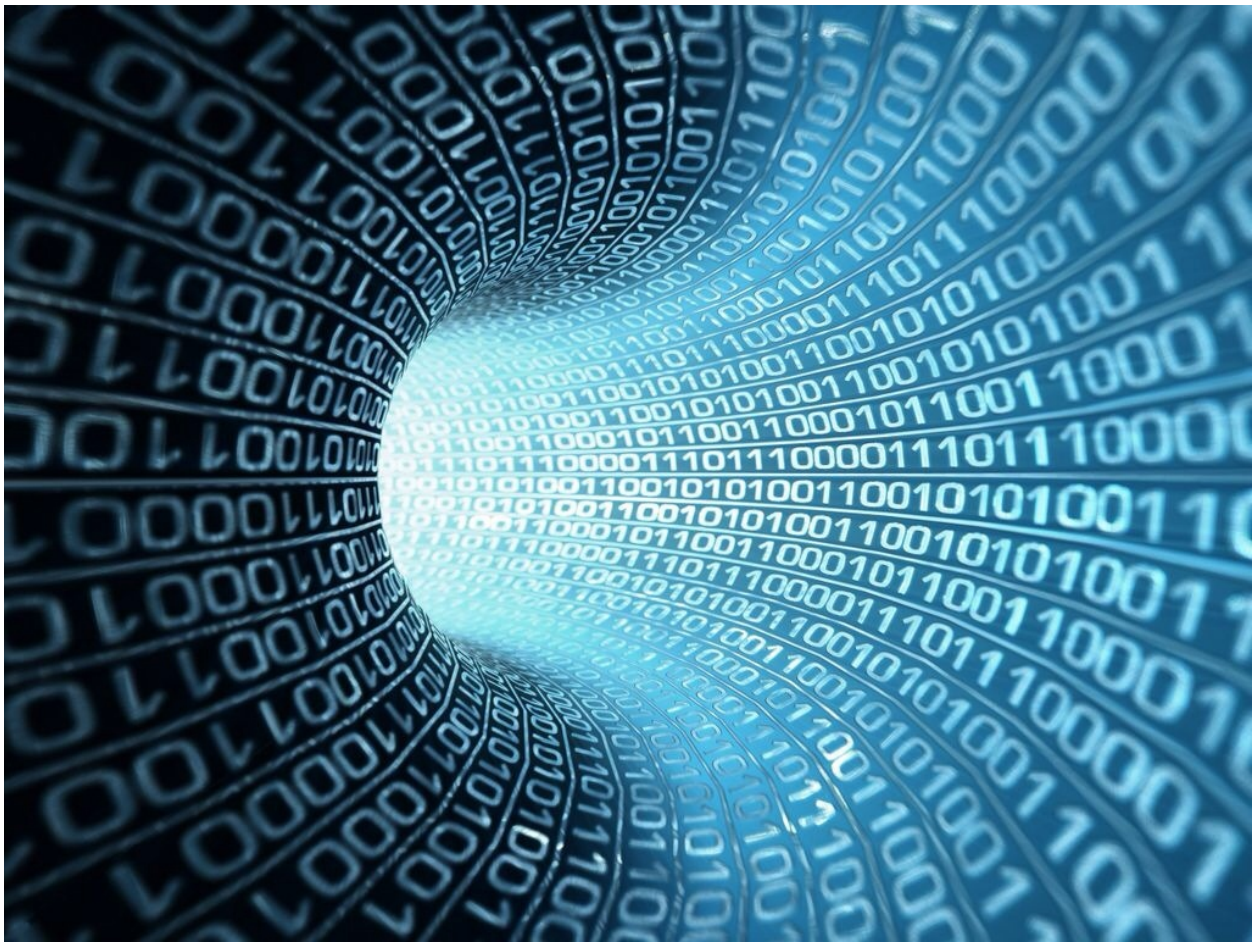
Report: DIAML Assignment 4



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: Tunga Tessema

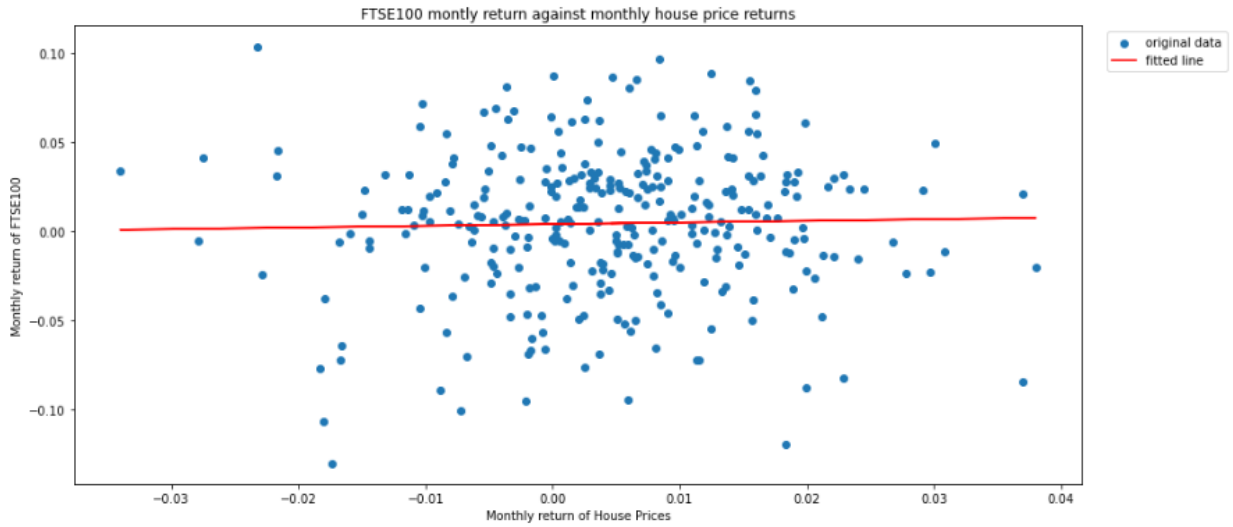
Andrew ID: tchamiss

Full Name: Tunga Tessema

Librairies

- **Pandas:** to read excel and csv files, convert them to dataframes and perform operations on them
- **numpy:** to create an array, calculate square roots and calculate correlation coefficients
- **matplotlib.pyplot:** to plot graphs
- **scipy.stats:** to get a linear regression model and to calculate p-values
- **statsmodels.api**: to calculate linear regression model using least square method
- **quandl**: to download data from quandl
- **Itertools**: to generate all combinations of the independent variables

Question 1




```
LinregressResult(slope=0.09324142754349966, intercept=0.004047837686662456, rvalue=0.026551295701909915, pvalue=0.6409049000031
651, stderr=0.1997058644355541, intercept_stderr=0.002437025309251721)
correlation coefficient is: 0.026551295701909918

H0: slope = 0 (null hypothesis)
H1: slope != 0 (alternative hypothesis)
We us a two-tailed test because the alternative hypothesis is that the slope is not equal to zero
the p-value is: 0.6409049000031651
Since the p-value is greater than or equal to 0.05, we accept the null hypothesis. With 95% confidence level, we can say that t
he slope is zero. This means that there is no significant relationship between the monthly return of houses and stocks
```

Steps: I first loaded the datafiles of the housing and stock prices using pandas. I renamed the unnamed column in housing data to Date and I converted the date column of the stock prices data from object type to datetime type. I filtered both data frames to get the values from 1991 to 2016 only. I calculated the monthly return of house prices using the pct_change() function. I also calculated the monthly return of stock prices using the pct_change() function. I then used the linregress function from the scipy.stats function to get the regression line coefficients. I plotted the original data and the regression line using matplotlib.pyplot.
I then calculated the correlation coefficient using the corr function of pandas.
The next step was to conduct a hypothesis test to determine whether there is a significant relationship between these two variables. I first defined the null and alternative hypothesis. The null hypothesis is that the slope is equal to zero. The alternative hypothesis is that the slope is not equal to zero. We will use a two-tailed test. I then calculated the t statistic by dividing the slope by the standard error of the slope. I calculated the p-value by using stats.t.sf function. I gave it the t value and the degrees of freedom. The degrees of freedom is the number of rows minus 2.

Insights: I have found the following linear regression formula: y = 0.0932x + 0.0040. I have also found a correlation coefficient of approximately 0.0265. This shows that there is almost no correlation between the monthly return of houses and the monthly return of stocks.
I also found a p-value of approximately 0.64. This is greater than the significance level (0.05). This means that we accept the null hypothesis. We are 95% confident that the slope is zero. Hence, there is no significant relationship between the monthly return of houses and the monthly return of stocks.

Question 2

```
apps correlation: 0.14675459955109238
enroll correlation: -0.022341038639948456
outState correlation: 0.571289928248201
top10 correlation: 0.4949892348013402
top25 correlation: 0.47728116437578355

the chosen independent variables (forward regression): ['Outstate', 'Top25perc']

the model with the least BIC: ['Outstate', 'Top25perc']
its BIC: 6274.3329824422635

R-squared with all variables: 0.3861582005130556
R-squared with 2 variables: 0.37776441749868717

prediction by model with all variables: [89.20112305]
prediction by model with 2 variables: [87.09352366]
actual garduation rate: 74
```

	Collar	BIC
0	[Apps]	6619.397697
1	[Enroll]	6635.926794
2	[Outstate]	6329.339476
3	[Top10perc]	6417.933788
4	[Top25perc]	6435.453828
5	[Apps, Enroll]	6563.239399
6	[Apps, Outstate]	6319.696848
7	[Apps, Top10perc]	6424.078072
8	[Apps, Top25perc]	6441.599055
9	[Enroll, Outstate]	6330.752936
10	[Enroll, Top10perc]	6411.105482
11	[Enroll, Top25perc]	6423.813497
12	[Outstate, Top10perc]	6283.333458
13	[Outstate, Top25perc]	6274.332962
14	[Top10perc, Top25perc]	6418.124429
15	[Apps, Enroll, Outstate]	6320.507268
16	[Apps, Enroll, Top10perc]	6396.704915
17	[Apps, Enroll, Top25perc]	6401.190886
18	[Apps, Outstate, Top10perc]	6287.773696
19	[Apps, Outstate, Top25perc]	6279.392965
20	[Apps, Top10perc, Top25perc]	6423.737387
21	[Enroll, Outstate, Top10perc]	6289.984326
22	[Enroll, Outstate, Top25perc]	6280.769562
23	[Enroll, Top10perc, Top25perc]	6407.885029
24	[Outstate, Top10perc, Top25perc]	6279.766569
25	[Apps, Enroll, Outstate, Top10perc]	6288.068853
26	[Apps, Enroll, Outstate, Top25perc]	6277.675824
27	[Apps, Enroll, Top10perc, Top25perc]	6392.747969
28	[Apps, Outstate, Top10perc, Top25perc]	6285.110958
29	[Enroll, Outstate, Top10perc, Top25perc]	6286.137256
30	[Apps, Enroll, Outstate, Top10perc, Top25perc]	6283.746453

Steps: I first read the college file using pandas. I then calculated the graduation rate's correlation with the five independent variables. I then used forward regression to select the best model. The forward regression function adds one variable at a time, builds the model and stores the p-value corresponding to every variable in a pandas series. The variable who has the least p-value is chosen and if its value is less than the significance level, it is added to the least of variables that we will include in our best model.

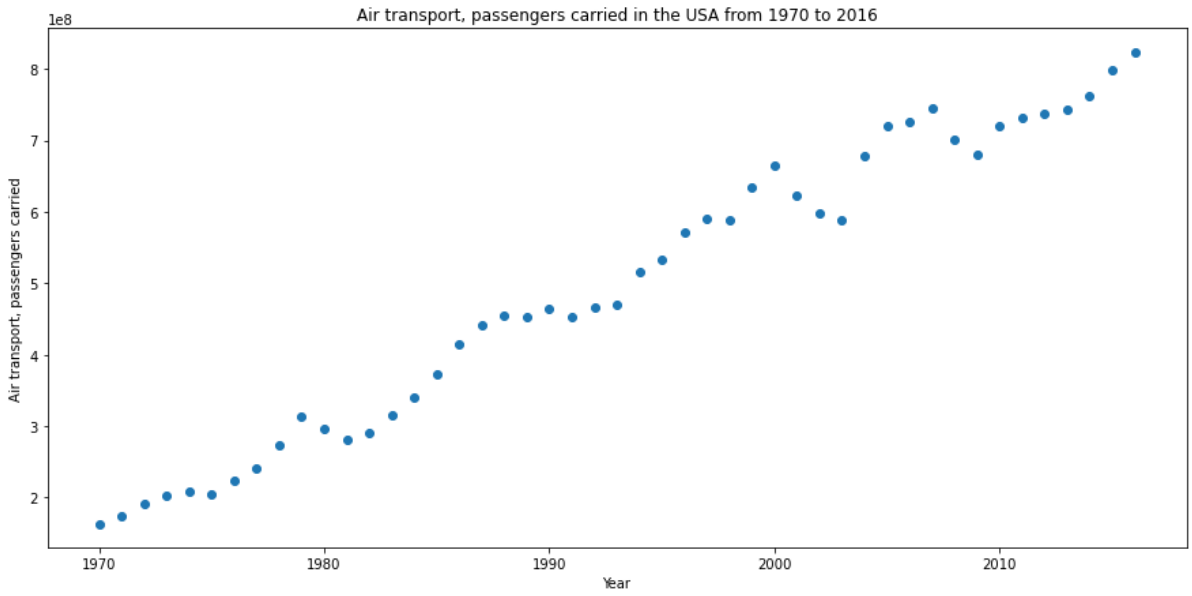
I then used the best subset method to select the best model based on the least BIC value. I first used the itertools.combinations library to create all possible combinations of the 5 independent variables. I got 31 combinations. I then built a model for each one of them and stored their bic in a dataframe. I then retrieve the model with the minimum BIC.

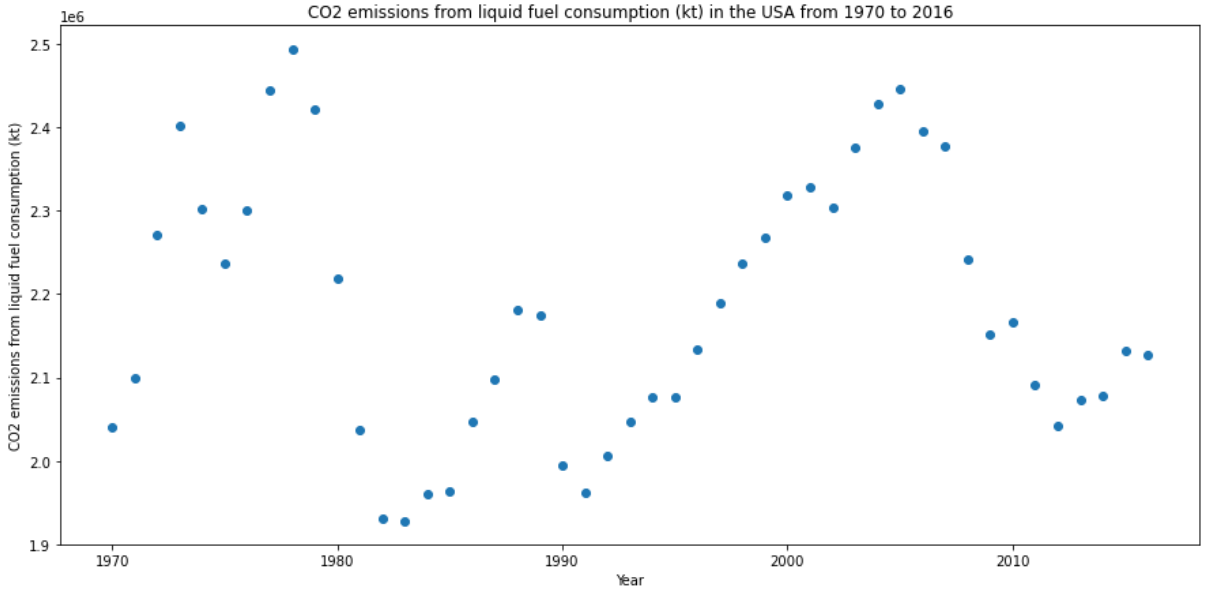
I then used the r-squared value to compare the accuracy of the model chosen with the model that includes all variables.

Next, I used both models to predict the graduation rate of CMU.

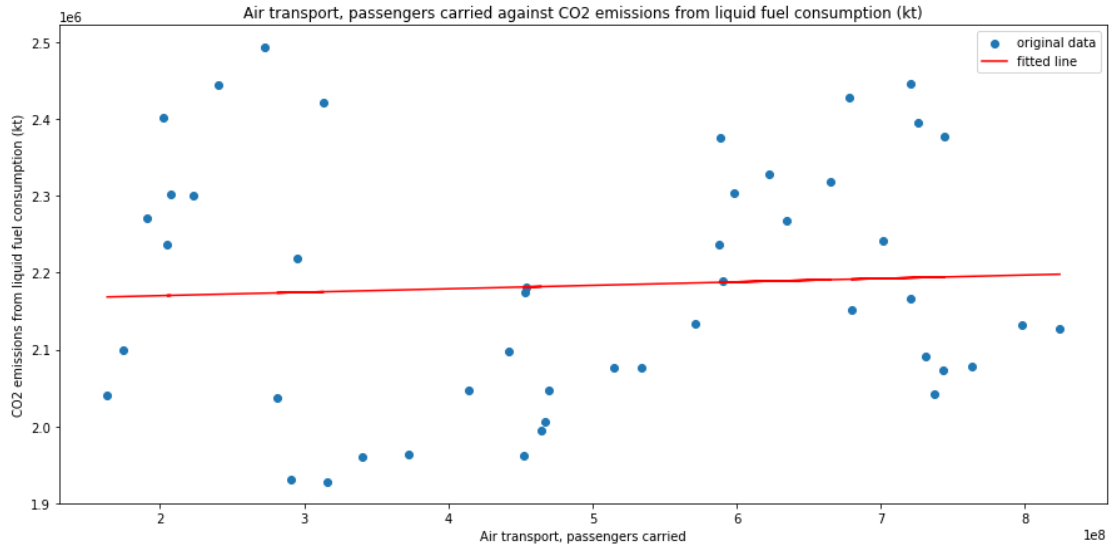
Insights: Calculating the correlation coefficient between the graduation rate and every variable gives us the type of linear relationship between a specific variable and the graduation rate. A correlation coefficient is a value between -1 and 1. The value 1, tells us there is a perfect positive linear relationship, -1 tells us there is a negative linear relationship and 0 tells us there is no linear relationship. Any value between -1 and 1, tells us how strong or weak the relationship is. With the correlation coefficients that I have found, I see that 'outstate' has the most relationship (0.57), followed by top10 (0.49) and top25 (0.47). We can see that applications received has a very low correlation (0.15) and number of enrollments has an even lower and negative correlation (-0.02). This gives us an idea of which variables might contribute greatly to the value of the graduation rate. However, this doesn't tell us which variables to choose. The correlation coefficients correspond to the sample given. To determine if this sample correlation represents the whole population, we have to conduct hypothesis testing.
The forward regression function that I used does that. It calculates the p-value for every variable added in the model. It specifies a null hypothesis, saying that the coefficient for that variable is zero and the alternative hypothesis saying that the coefficient for that variable is not equal to zero. If the p-value is less than 0.05 (significance level), we reject the null hypothesis. This means that the variable has significantly contributed to the model since its slope is significantly different from zero. By this method, I have found that 'outstate' and 'top25' are the best independent variables for the model.
I then used the best subset method to select the model with the least BIC value. The Bayesian Information Criterion is a metric that is used to compare the goodness of fit of different regression models. I have found that the model with the 'outstate' and 'top25' variables has the least BIC. So I have found the same set of predicting variables using the forward regression and best subset method using the BIC. This could be explained by the fact that BIC measures the goodness of fit but also penalizes the use of many parameters to combat overfitting.
To compare the accuracy of the chosen model compared to the accuracy of the model with all variables, I have used their r-squared value. R-squared (Coefficient of determination) represents how well the predicted values fit compared to the original values. The value from 0 to 1 is interpreted as percentages. The higher the value is, the better the model is. We can see that the model with all variables has an r-squared value of 0.386 while the chosen model has an r-squared value of 0.377. The model with all variables is slightly more accurate. Its value is bigger by only 0.009. The complexity of adding 3 more predictor variables for such a small increase is not worth it.
For Carnegie Mellon University, the model with all variables predicts 89.20 while the model with 2 variables (chosen one) predicts 87.09.

Question 3


Air transport, passengers carried in the USA from 1970 to 2016
1e8
Air transport, passengers carried
Year
1970
1980
1990
2000
2010
2
3
4
5
6
7
8


CO2 emissions from liquid fuel consumption (kt) in the USA from 1970 to 2016
1e6
CO2 emissions from liquid fuel consumption (kt)
Year
1970
1980
1990
2000
2010
1.9
2.0
2.1
2.2
2.3
2.4
2.5

```
correlation coefficient is: 0.05727996729562196
```




```
LinregressResult(slope=4.4767526782506634e-05, intercept=2161245.0513842343, rvalue=0.05727996729562196, pvalue=0.7021411998479
979, stderr=0.00011631626643718978, intercept_stderr=61898.28686219432)
```

Steps: The trend I am studying is the relationship between the number of passengers using air transport and the amount of co2 emissions from liquid fuel consumption. I used the following data sources:

- Air transport, passengers carried: [Air transport, passengers carried | Data (worldbank.org)]
- CO2 emissions from liquid fuel consumption (kt): [CO2 emissions from liquid fuel consumption (kt) | Data (worldbank.org)]

I assume that as the number of passengers using air transport increases the amount of co2 emission increases also.
To study the relationship I have used the following methodology. I first saw loaded both files into pandas dataframes. I removed headers that were not needed and renamed the first row as the column. I selected the row corresponding to the United states. I also looked at the dates available and chose a range that both datas would be available. I chose the dates from 1970 to 2016. I then plotted a graph of the passengers carried over the years to get a sense of what the trend was. As you can see on the first graph, the number of passengers using air transport is increasing over the years. At some points, the number of passengers decreases slightly. For example, from 1979 to 1981, the number of passengers decreased. But overall, the number of passengers using air transportation has been increasing over the years.
I also plotted the amount of co2 emissions over the years. As you can see on the second graph, the amount of co2 emissions goes from 1.9 billion kt to 2.5 billion kt. It seems to increase until it reaches a peak then decreases sharply. However, it never goes below 1.9 billion kt.
After getting an idea of both datasets individually, I merged the data frames using their years as a criteria. I then calculated their correlation coefficient. Their correlation coefficient is 0.057. This is a very low number close to 0, it shows us that the two datas have almost no correlation.

I built a linear regression model for the datasets. I used the number of air passengers as an independent predictor and the amount of co2 emissions as the dependent variable.
I then plotted the original data and the fitted line. As you can see on the third graph the data points are far from the fitted line. Using the linear regression, I obtained a p-value of 0.70. Using this value I can conduct a hypothesis test. The null hypothesis is that the slope is equal to zero and the alternative hypothesis is that the slope is not equal to zero. 0.70 is greater than 0.05 (our significance level). This means that we accept the null hypothesis. We are 95% confident that the number of passengers using air transport are not significantly related to the amount of co2 emission from liquid fuel consumption.
This might be explained by the following points:

- The number of passengers is not equal to the number of flights. The same amount of co2 is emitted whether the plane is full or not. So even though the number of passengers increase, we don't know if the increase is significant enough to increase the number of flights
- Air transportation is not the main liquid fuel consumer

Hence, since these two variables are not significantly related, I cannot predict the amount of co2 emission in 2021 given the number of passengers using air transport.

Question 4

```
LinregressResult(slope=0.13973362299465242, intercept=-270.18337210338683, rvalue=0.5407348863387454, pvalue=0.0011585109247269
068, stderr=0.03904193819898755, intercept_stderr=77.92859536418155)
```


```
2020 prediction: 12.078546345811048
MAPE in %: 21.992601540272005
```

Steps: I first downloaded the data from quandl and chose the values from 1980 to 2013. I then built a linear regression model using the linregress method. I used the year as the independent variable and the unemployment rate as the dependent variable. I found the following equation y = 0.1397x - 270.1834. Y is the unemployment rate and x is the year. I also found a p-value of 0.001. The p-value is less than 0.05 (significance level). This means that we reject the null hypothesis (slope is equal to zero). Hence, we are 95% confident that the year has a significant relationship with the unemployment rate in Israel.
I used the above equation to predict the unemployment rate in 2020. I found an unemployment rate of approximately 12.08.
To estimate the accuracy of the model I used the mean absolute percentage error. MAPE is calculated using the following formula:

MAPE = (1/n) * Σ(|actual – forecast| / |actual|) * 100

where:

- **Σ** – A symbol that means “sum”
- **n** – Sample size
- **actual** – The actual data value
- **forecast** – The forecasted data value

I wrote a function to calculate MAPE by using the above formula. I got a MAPE value of 21.99%. This tells us that the average difference between the forecasted value and the actual value is 21.99%.