ANDREW ID: maniraha

Names: Albert MANIRAHO

Kigali, January 31, 2022

DATA ANALYTICS ASSIGNMENT 1 REPORT

All the libraries used:

```
from scipy import stats
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from statsmodels.graphics.tsaplots import plot_acf
import math
import seaborn as sns
import statsmodels.api as sm
import warnings
warnings.filterwarnings('ignore')
```

QUESTION 1.

- The historical daily weather data for France was downloaded and loaded in the environment for use.
- The encoding was done to make our data best for use.
- High Gust Wind\xa0(km/h) and Events columns were dropped since they are unwanted in the dataset.
- By using .info(), primary information of the dataset was obtained.
- Since there was some empty or missing data, a proper interpolation was made to fill in the missing values.

QUESTION 2.

The correlation matrix that shows correlation coefficients between variables is calculated. It is used to investigate the dependence between multiple variables at the same time.

From the correlation, the following is the heatmap of the correlation matrix:

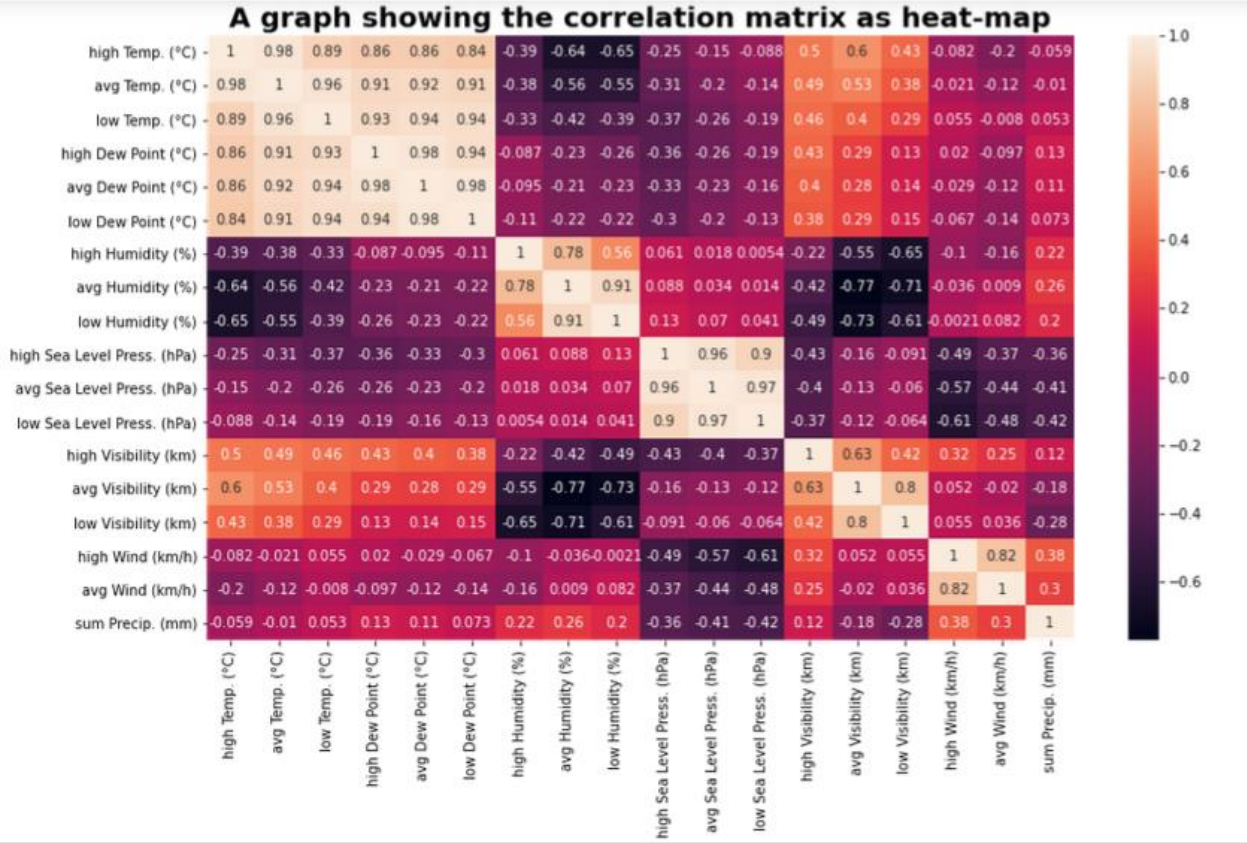

Figure 1: The heatmap for the correlation matrix.

QUESTION 3.

- The historical daily electricity consumption data for France was downloaded and loaded in the environment for use.
- The encoding was done to make our data best for use.
- The unwanted columns were dropped from the dataset.
- Since there was some empty or missing data, a proper interpolation was made to fill in the missing values.

QUESTION 4.

- The electricity consumption data date column was previously an object, so we needed to change its format into datetime format for the ease of synchronization.
- The weather condition data too, its date column format was changed from object into datetime format.
- Now that the two datasets have common column of date, then we used the date column to merge the two datasets.
- The scatter plot of energy consumption against mean temperature is as follows:

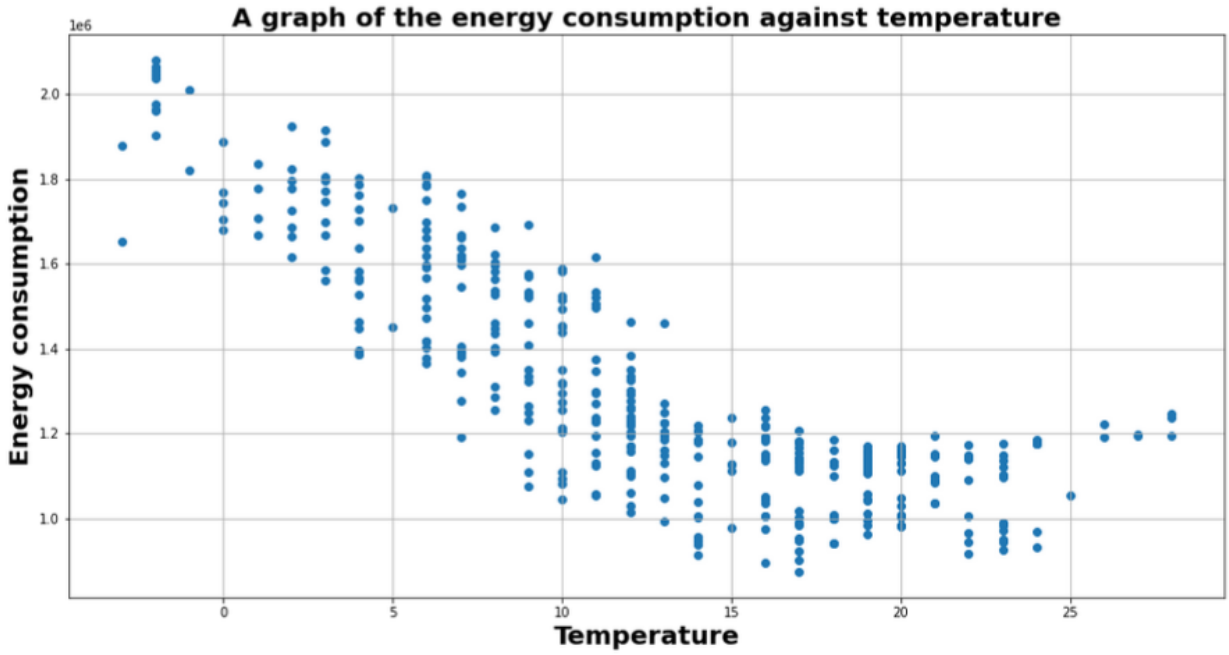

Figure 2: The graph of the energy consumption against temperature in France.

From the graph, at the starting it is seen that when the temperature is low, there is an increased/high energy consumption since people need to use the heating systems and heating materials.

And when the temperature increases a bit, the energy consumption is decreased a bit to a certain point that the temperature is high and then the energy consumption increases again because when the temperature is high, people will need to turn on cooling systems and equipments.

QUESTION 5.

The quadratic model was fitted using the using the average temperature and the energy consumption using the python code np.polyfit.

First, I plotted the scatter plot of the energy consumption against temperature.

Then, I plotted the line fitting the as much points as possible.

The following is the graph of the energy consumption against temperature polynomial fitting:

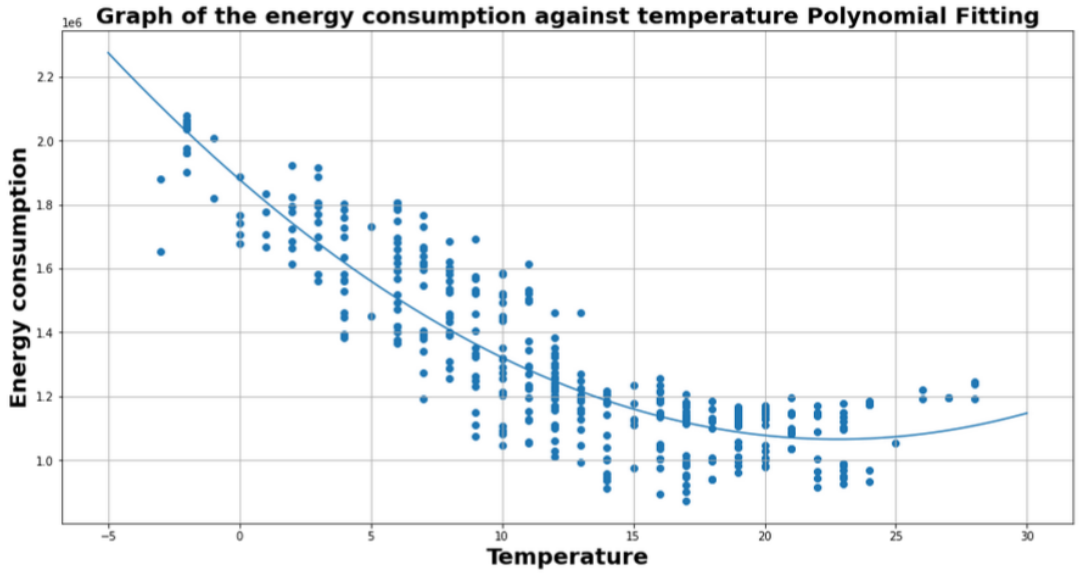

Figure 3: The graph of energy consumption against temperature polynomial fitting.

QUESTION 6.

The minimum energy is: 1065866.1724185245
The optimal temperature is: 22.78846153846154

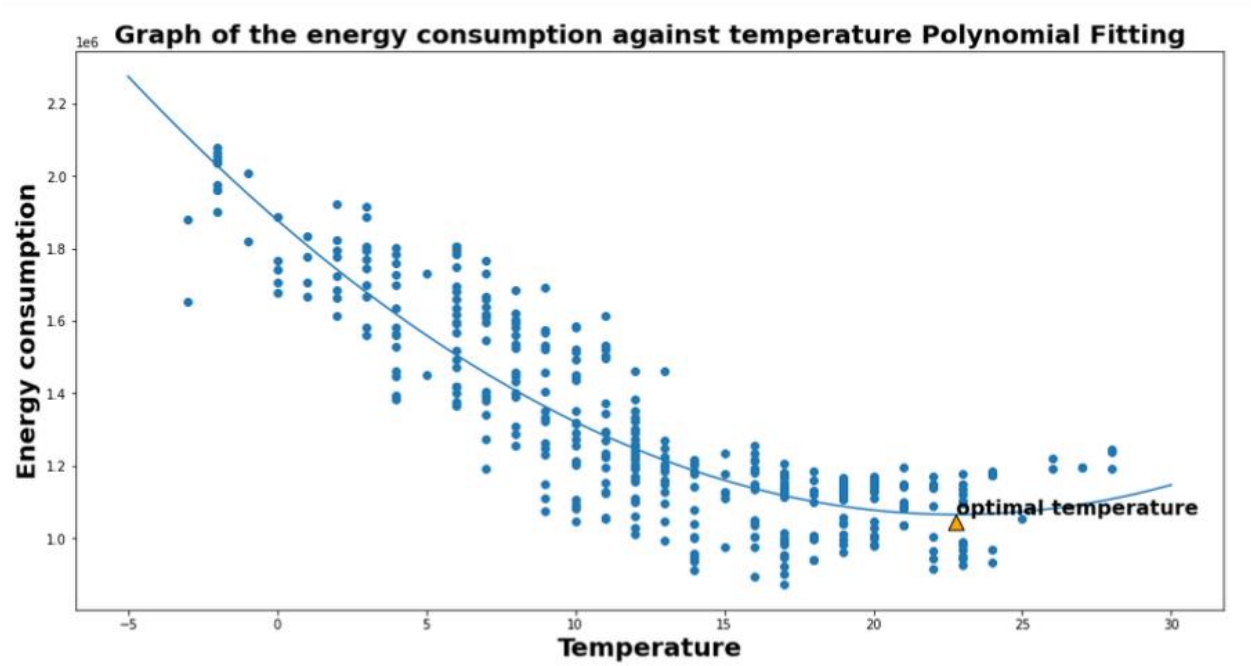

Figure 4: The graph that shows the point where coincide the optimal temperature and minimal energy consumption.

QUESTION 7.

The selected variables are:

['avg Dew Point\xa0(°C)',
 'high Humidity\xa0(%)',
 'low Humidity\xa0(%)',
 'low Sea Level Press.\xa0(hPa)',
 'high Visibility\xa0(km)']

The coefficient of determination R_Squared is: 0.750

R-squared:	0.750

Carnegie Mellon University Africa

QUESTION 8.

By increasing the explanatory variables through squaring them.

The selected variables are now:

['high Temp.\xa0(°C)*',
 'low Temp.\xa0(°C)*',
 'low Humidity\xa0(%)*',
 'high Visibility\xa0(km)*',
 'high Temp.\xa0(°C)',
 'low Temp.\xa0(°C)',
 'high Dew Point\xa0(°C)',
 'high Visibility\xa0(km)',
 'high Wind\xa0(km/h)']

The coefficient of determination R_Squared is: 0.826

R-squared:	0.826

Yes, there is an improvement in the accuracy of the model since the coefficient of determination R_Squared has moved from 0.750 to 0.826 and this a result of increasing the explanatory variables.

QUESTION 9.

The days of the week that are selected are:

'Monday',
 'Tuesday',
 'Wednesday',
 'Thursday',
 'Friday',
 'Saturday',
 'Sunday']

The coefficient of determination R_Squared is: 0.904

R-squared:	0.904

Yes, there is an improvement in the accuracy of the model since the coefficient of determination R_Squared has moved from 0.826 to 0.904 and this a result of including dummy variables for the day of the week in the multivariate regression.

QUESTION 10.

The modelling approach will lead to Overfitting since we keep on increasing explanatory variables this will help improve the accuracy of the model. But keeping doing this will lead to overfitting.

Two approaches that could be used to prevent overfitting

Cross-validation

One of the most effective methods to avoid overfitting is cross validation. This method is different from what we do usually. We use to divide the data in two, cross validation divides the training data into several sets. The idea is to train the model on all sets except one at each step. If we have k sets, we will train the model k times with a new testing set at each step. This cross-validation technique is called k-fold.

Regularization methods

Regularization methods are techniques that reduce the overall complexity of a machine learning model. They reduce variance and thus reduce the risk of overfitting. The regularization methods allow the variance of the model to be considerably reduced without increasing the bias. We will return to the bias/variance dilemma in the last section.

Many regularization techniques exist:

1. L1
2. Ridge
3. L2
4. Lasso