Internship Report

Investigating the relationship between energy and mining consumption in Nairobi, Kenya and how it affects climate change.

By Mohamed Ziane

In collaboration with Springboard

Fall 2021

Table of Contents

Executive Summary

By using data from varied sources (The World Bank, The Humanitarian Data Exchange, and Meteo Blue), a global understanding of relationships between potential climate change factors that could explain some key variables such as Temperature and Precipitation could be drawn for Africa as a whole and Kenya more specifically.

The project undertook as my capstone project at Springboard will be investigating the relationship between energy and mining consumption in Nairobi, Kenya and how it affects climate change. Analysis will be briefly global (Africa) before focusing specifically on Kenya.

Even though most of the project was centered around Nairobi, two other African cities (Algiers and Cape Town) were added to initially compare whether some trends in Nairobi's temperature could also be correlated with other parts in Africa. However, temperature variation is challenging to rely on to extract an obvious trend as temperature varies throughout seasons.

The top three possible causes and consequences that saw the biggest increase from 1985 to 2019 were then estimated. The biggest cause was the foreign direct investment (by percentage of GDP), followed closely by the C02 emissions, while the precipitation dominated the top possible leading consequence of climate change in Nairobi during this time.

Beyond relationship trends, correlation matrix offered further insights:

- CO2 seems to be more connected with the Total population growth and Methane emissions.
- Precipitation seems to be related to Ores & metal exports, agricultural land, CO2 emissions, foreign direct investments, Methane & Nitrous Oxide emissions, population in urban agglomerations, the total population, the total cloud cover, and the soil moisture.
- For the temperature variation, we could note some positive relationships with the following: Ores & metals exports, CO2 emissions, total population growth, the urban population, the sunshine duration, and the soil temperature

Eventually, using machine learning, a forecast prediction was achieved for the precipitation and temperature for Nairobi from 1985 to 2019. Both supervised-based machine learning models were used: Regression and Random Forest. From the linear regression, the accuracy reached a score of ~ 61% with a mean and median absolute errors of 1.94 and 1.6 degrees F respectively. The metrics, on the other hand, from the Random Forest regressor in combination with cross-validation and hyperparameter search using GridSearch showed scores around 90%.

1. Introduction

In June 2021, I have embarked in an online Data Science bootcamp journey from Springboard that combines programming, mathematics, and technology to analyze data and identify insights.

Data has transformed every industry, and the ability to recognize trends and insights is a highly desirable skill.

The project undertook as my capstone project will be investigating the relationship between energy and mining consumption in Nairobi, Kenya and how it affects climate change. Analysis will be briefly global (Africa) before focusing specifically on Kenya.

The six steps that this project followed are:

- Context & Problem Identification
- Data Wrangling
- Exploratory Data Analysis (EDA)
- Pre-Processing and Training Data Development
- Modeling

2. Context

Climate risks pose serious threats to Kenya's sustainable development goals. With the largest economy in East Africa and a population of 48.5 million, Kenya serves as the regions financial, trade and communications hub.

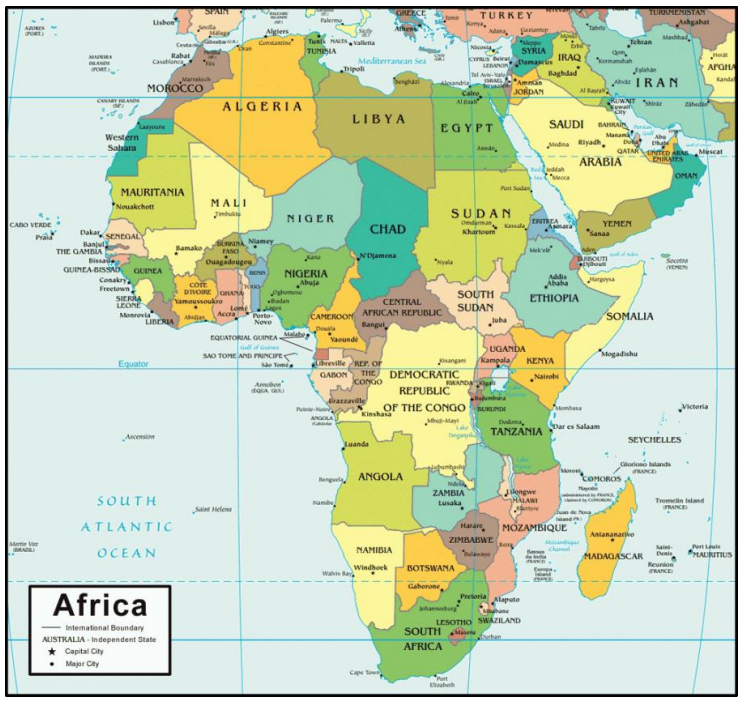
The country's economy is largely dependent on rainfed agriculture and tourism, each susceptible to climate variability and change and extreme weather events. Increasing interseason variability, increasing temperatures, heavy rainfall events and sea level rise lead to severe crop and livestock losses, famine and displacement. High population growth in urban areas is leading to expanding informal settlements, which are at risk from water scarcity, flooding and heat.

Most of the country's coast is low-lying, with coastal plains, islands, beaches, wetlands and estuaries at risk from sea level rise. A sea level rise of 30cm is estimated to threaten 17 percent (4,600 hectares) of Mombasa with inundation. Models estimate that by 2030 climate variability and extremes will lead to losses equivalent to 2.6 percent of GDP annually.

Kenya's geography is dominated by arid and semi-arid plains, with a temperate highland plateau (reaching over 5,000 m) in the center, and a hotter, wetter climate along the coast and the shores of Lake Victoria. Two-thirds of the country receive less than 500 mm (19.6 in) of rainfall per year; coastal and highland areas receive annual averages upwards of 1,100 mm (43 in) and 2,000 mm (78 in), respectively.

Kenya has two rainy seasons: "long rains" from March to June (about 70 percent of total annual rainfall); and "short rains" from October to December. In the west and along the coast, additional significant rainfall occurs outside of these two rainy seasons. Temperatures range from an average of 18°C (64 DegF) in high elevation areas like Nairobi to 26°C (78.8 DegF) in coastal areas such as Mombasa.

Located on the equator, Kenya experiences little seasonal temperature variation.



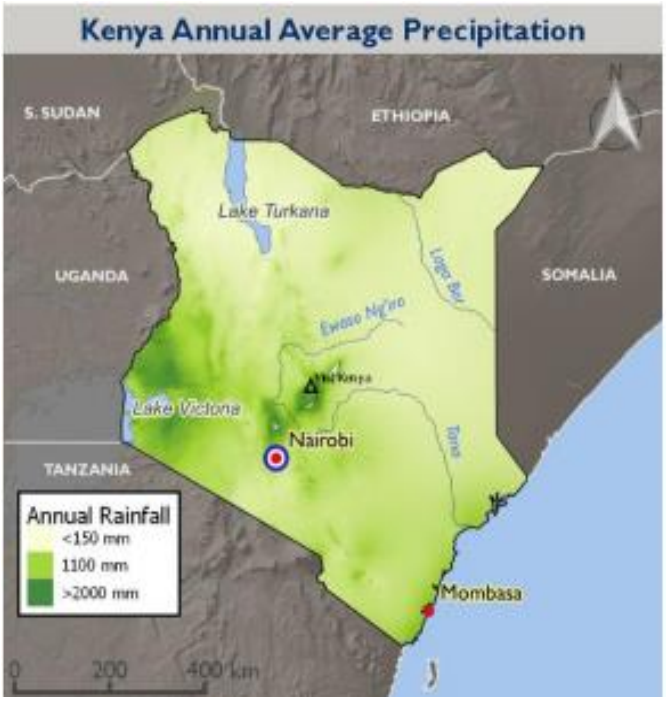


Fig.1 Map of African Countries (Wikipedia), left. Kenya Annual Average Precipitation in 2019 (World Bank)

3. Problem Identification

Stemming from different sources from the World Bank Open Data (***WBOD***), the Humanitarian Data Exchange (***HDE***) and Meteo Blue (meteorological service created at the University of Basel, Switzerland in 2006), several climate-related data were collected to extract insights into possible correlations between causes and consequences of climate change in Kenya. 20+ years of daily data were analyzed and, using machine learning, a prediction of main climate consequences were simulated.

4. Data Wrangling

This step focuses on collecting data, organizing it and making sure it is well defined. Some data cleanings are also carried out in this stage.

A total of 8 different raw tables were sourced. 5 of them, from WBOD and HDE contained possible climate change causes for the entire African region, that ended up being merged into a single table while the three remaining ones, sourced from Meteo Blue, contained possible consequences from climate change in Nairobi (Kenya), but also from Algiers (Algeria) and Cape Town (South Africa).

Even though most of the project was centered around Nairobi, the two other African cities were added to initially compare whether some trends in Nairobi could also be correlated with other parts in Africa.

5. Possible Consequences of Climate Change for Algeria, Kenya and South Africa (from Meteo Blue)

For each of those cities, the 12 following possible related consequences were selected:

- Temperature
- Precipitation
- Snowfall amount
- Relative humidity
- Wind speed
- Wind Dominant direction
- Total cloud covering
- Sunshine duration
- Shortwave radiation
- Mean sea level pressure
- Soil temperature
- Soil Moisture

```
In [28]: # Checking the Nairobi Dataset Info
         nairobi.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 13150 entries, 0 to 13149
Data columns (total 25 columns):
 #   Column                                      Non-Null Count  Dtype
---  ------                                      --------------  -----
 0   timestamp                                   13150 non-null  object
 1   Nairobi Temperature [2 m elevation corrected] 13150 non-null  float64
 2   Nairobi Precipitation Total                 13150 non-null  float64
 3   Nairobi Snowfall Amount                     13150 non-null  int64
 4   Nairobi Relative Humidity [2 m]             13150 non-null  int64
 5   Nairobi Wind Speed [10 m]                   13150 non-null  float64
 6   Nairobi Wind Direction Dominant [10 m]      13150 non-null  float64
 7   Nairobi Cloud Cover Total                   13150 non-null  float64
 8   Nairobi Sunshine Duration                   13150 non-null  float64
 9   Nairobi Shortwave Radiation                 13150 non-null  float64
 10  Nairobi Mean Sea Level Pressure [MSL]       13150 non-null  float64
 11  Nairobi Soil Temperature [0-10 cm down]     13150 non-null  float64
 12  Nairobi Soil Moisture [0-10 cm down]        13150 non-null  float64
 13  N_Temperature_m                             13150 non-null  float64
 14  N_Precipitation_Total                       13150 non-null  float64
 15  N_Snowfall_Amount                           13150 non-null  int64
 16  N_Relative_Humidity_m                       13150 non-null  int64
 17  N_Wind_Speed_10m                            13150 non-null  float64
 18  N_Cloud_Cover_Total                         13150 non-null  float64
 19  N_Sunshine_Duration                         13150 non-null  float64
 20  N_Shortwave_Radiation                       13150 non-null  float64
 21  N_Mean_Sea_Level_Pressure                   13150 non-null  float64
 22  N_Soil_Temperature_10cm                     13150 non-null  float64
 23  N_Soil_Moisture_10cm                        13150 non-null  float64
 24  N_Wind_Direction_Dominant_10m               13150 non-null  float64
dtypes: float64(20), int64(4), object(1)
memory usage: 2.5+ MB
```

Fig.2 Python output of the 12 possible related consequences of climate change in Nairobi (Kenya). The top 12 list shows the old vs. the new after re-naming (bottom 12). The additional column "timestamp" is the time-series column dating from 1985 to 2019. There are 13150 rows of datasets as they are daily records sourced from Meteo Blue.

6. Possible Causes of Climate Change for Algeria, Kenya and South Africa (from WBOD and HDE)

A total of 72 different features were initially sourced as possible causes and 26 were retained as possibly most significant.

```
In [94]: new_merge_edit_final = new_merge_edit_final.rename(columns={"Year":"timestamp"})
         new_merge_edit_final.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 12540 entries, 243 to 12782
Data columns (total 26 columns):
 #   Column                                                                    Non-Null Count  Dtype
---  ------                                                                    --------------  -----
 0   timestamp                                                                 12540 non-null  datetime64[ns]
 1   Year_extract                                                              12540 non-null  int64
 2   CO2 emissions from liquid fuel consumption (kt)_new                       12540 non-null  float64
 3   Energy related methane emissions (% of total)_new                         12540 non-null  float64
 4   Fuel exports (% of merchandise exports)_new                               12540 non-null  float64
 5   Fuel imports (% of merchandise imports)_new                               12540 non-null  float64
 6   Methane emissions in energy sector (thousand metric tons of CO2 equivalent)_new  12540 non-null  float64
 7   Mineral rents (% of GDP)_new                                              12540 non-null  float64
 8   Nitrous oxide emissions in energy sector (thousand metric tons of CO2 equivalent)_new  12540 non-null  float64
 9   Ores and metals exports (% of merchandise exports)_new                    12540 non-null  float64
 10  Total natural resources rents (% of GDP)_new                              12540 non-null  float64
 11  Agricultural land (% of land area)_new                                    12540 non-null  float64
 12  Agricultural land (sq. km)_new                                            12540 non-null  float64
 13  Agriculture, forestry, and fishing, value added (% of GDP)_new            12540 non-null  float64
 14  Arable land (% of land area)_new                                          12540 non-null  float64
 15  CO2 emissions (kt)_new                                                    12540 non-null  float64
 16  CO2 emissions (metric tons per capita)_new                                12540 non-null  float64
 17  Foreign direct investment, net inflows (% of GDP)_new                     12540 non-null  float64
 18  Methane emissions (kt of CO2 equivalent)_new                              12540 non-null  float64
 19  Mortality rate, under-5 (per 1,000 live births)_new                       12540 non-null  float64
 20  Nitrous oxide emissions (thousand metric tons of CO2 equivalent)_new      12540 non-null  float64
 21  Population growth (annual %)_new                                          12540 non-null  float64
 22  Population in urban agglomerations of more than 1 million (% of total population)_new  12540 non-null  float64
 23  Population, total_new                                                     12540 non-null  float64
 24  Total greenhouse gas emissions (kt of CO2 equivalent)_new                 12540 non-null  float64
 25  Urban population_new                                                      12540 non-null  float64
dtypes: datetime64[ns](1), float64(24), int64(1)
memory usage: 2.6 MB

In [95]: table_final = pd.merge(new_merge_edit_final, afr_merge_2, how="left", on='timestamp')
```

Fig.3 Python output of the 26 possible related causes of climate change in Nairobi (Kenya). The additional column "timestamp" is the time-series column dating from 1985 to 2019. There are 12540 rows of datasets as they are daily records sourced from the World Bank and the Humanitarian Data Exchange.

7. Exploratory Data Analysis (EDA)

EDA is an approach for summarizing and visualizing the important characteristics and statistical properties of a dataset. Visualizing the data will help make sense of it to identify emerging themes.

The first step was to load the merged data that contained possible causes of climate change in Africa versus consequences in Kenya (Labeled with "_N" for Nairobi), Algeria (Labeled with "_A" for Algiers) and South Africa (Labeled with "_CT" for Cape Town).

The merged table was also filtered to only contain data ranging from 1985/31/12 until 2019/31/12 on a daily basis and a yearly version of the table was also created using a resampling option.

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 35 entries, 1985-12-31 to 2019-12-31
Freq: A-DEC
Data columns (total 63 columns):
 #   Column                                                                                Non-Null Count  Dtype
---  ------                                                                                --------------  -----
 0   Year_extract                                                                          35 non-null     float64
 1   CO2 emissions from liquid fuel consumption (kt)_new                                   35 non-null     float64
 2   Energy related methane emissions (% of total)_new                                     35 non-null     float64
 3   Fuel exports (% of merchandise exports)_new                                           35 non-null     float64
 4   Fuel imports (% of merchandise imports)_new                                           35 non-null     float64
 5   Methane emissions in energy sector (thousand metric tons of CO2 equivalent)_new       35 non-null     float64
 6   Mineral rents (% of GDP)_new                                                          35 non-null     float64
 7   Nitrous oxide emissions in energy sector (thousand metric tons of CO2 equivalent)_new 35 non-null     float64
 8   Ores and metals exports (% of merchandise exports)_new                                35 non-null     float64
 9   Total natural resources rents (% of GDP)_new                                          35 non-null     float64
 10  Agricultural land (% of land area)_new                                                35 non-null     float64
 11  Agricultural land (sq. km)_new                                                        35 non-null     float64
 12  Agriculture, forestry, and fishing, value added (% of GDP)_new                        35 non-null     float64
 13  Arable land (% of land area)_new                                                      35 non-null     float64
 14  CO2 emissions (kt)_new                                                                35 non-null     float64
 15  CO2 emissions (metric tons per capita)_new                                            35 non-null     float64
 16  Foreign direct investment, net inflows (% of GDP)_new                                 35 non-null     float64
 17  Methane emissions (kt of CO2 equivalent)_new                                          35 non-null     float64
 18  Mortality rate, under-5 (per 1,000 live births)_new                                   35 non-null     float64
 19  Nitrous oxide emissions (thousand metric tons of CO2 equivalent)_new                  35 non-null     float64
 20  Population growth (annual %)_new                                                      35 non-null     float64
 21  Population in urban agglomerations of more than 1 million (% of total population)_new 35 non-null     float64
 22  Population, total_new                                                                 35 non-null     float64
 23  Total greenhouse gas emissions (kt of CO2 equivalent)_new                             35 non-null     float64
 24  Urban population_new                                                                  35 non-null     float64
 25  N_Temperature_m                                                                       35 non-null     float64
 26  N_Precipitation_Total                                                                 35 non-null     float64
 27  N_Snowfall_Amount                                                                     35 non-null     float64
 28  N_Relative_Humidity_m                                                                 35 non-null     float64
 29  N_Wind_Speed_10m                                                                      35 non-null     float64
 30  N_Cloud_Cover_Total                                                                   35 non-null     float64
 31  N_Sunshine_Duration                                                                   35 non-null     float64
 32  N_Shortwave_Radiation                                                                 35 non-null     float64
 33  N_Mean_Sea_Level_Pressure                                                             35 non-null     float64
 34  N_Soil_Temperature_10cm                                                               35 non-null     float64
 35  N_Soil_Moisture_10cm                                                                  35 non-null     float64
 36  N_Wind_Direction_Dominant_10m                                                         35 non-null     float64
 37  Year_extract_x                                                                        35 non-null     float64
 38  A_Temperature_m                                                                       35 non-null     float64
 39  A_Precipitation_Total                                                                 35 non-null     float64
 40  A_Snowfall_Amount                                                                     35 non-null     float64
 41  A_Relative_Humidity_m                                                                 35 non-null     float64
 42  A_Wind_Speed_10m                                                                      35 non-null     float64
 43  A_Cloud_Cover_Total                                                                   35 non-null     float64
 44  A_Sunshine_Duration                                                                   35 non-null     float64
 45  A_Shortwave_Radiation                                                                 35 non-null     float64
 46  A_Mean_Sea_Level_Pressure                                                             35 non-null     float64
 47  A_Soil_Temperature_10cm                                                               35 non-null     float64
 48  A_Soil_Moisture_10cm                                                                  35 non-null     float64
 49  A_Wind_Direction_Dominant_10m                                                         35 non-null     float64
 50  CT_Temperature_m                                                                      35 non-null     float64
 51  CT_Precipitation_Total                                                                35 non-null     float64
 52  CT_Snowfall_Amount                                                                    35 non-null     float64
 53  CT_Relative_Humidity_m                                                                35 non-null     float64
 54  CT_Wind_Speed_10m                                                                     35 non-null     float64
 55  CT_Cloud_Cover_Total                                                                  35 non-null     float64
 56  CT_Sunshine_Duration                                                                  35 non-null     float64
 57  CT_Shortwave_Radiation                                                                35 non-null     float64
 58  CT_Mean_Sea_Level_Pressure                                                            35 non-null     float64
 59  CT_Soil_Temperature_10cm                                                              35 non-null     float64
 60  CT_Soil_Moisture_10cm                                                                 35 non-null     float64
 61  CT_Wind_Direction_Dominant_10m                                                        35 non-null     float64
 62  Year_extract_y                                                                        35 non-null     float64
dtypes: float64(63)
memory usage: 17.5 KB
```

Fig.4 Python output of the final merged table combining both possible causes and consequences of climate-related changes in Kenya (“N_”), Algiers(“A_”) and Cape Town (“CT_”). The additional column “Year_extract” is the time-series column dating from 1985 to 2019 on a yearly basis. There are 35 rows of datasets as they are daily records transformed into yearly ones. Sourced from the World Bank, Humanitarian Data Exchange and Meteo Blue

The temperature variation above 1985 to 2019 is consistent throughout the three cities (North to South). However, temperature variation is very hard to extract an obvious trend as temperature varies throughout seasons. Can we identify which other factors may have had a more significant increase from 1985 to 2019? The remaining of the project fosuced only on Nairobi, Kenya.

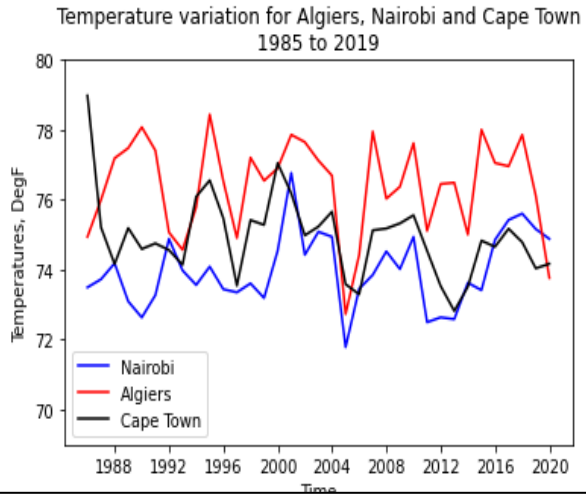


Fig.5 Temperature variation (in Fahrenheit) vs. time (from 1985 to 2019) for Algiers (Red line), Cape Town (Black line) and Nairobi (Blue line). Curves were extracted using Python from the final merged table.

8. Top 10 Climate-related causes and consequences in Nairobi. Kenya from 1985 to 2019

The yearly final merged table was used as the input after removing datasets from South Africa and Algeria. Then, a simple ratio of values belonging to 2019 was divided by the 1985 ones to evaluate the

Let's start by calculating the ratio of the temperature

In [321]:
```
def create_temp_ratio(d):
    y1985 = float(d['value'][d['Year'] == 1985])
    y2019 = float(d['value'][d['Year'] == 2019])
    ratio = [y2019/y1985]
    return ratio
```

In [322]:
```
create_temp_ratio(df_melt[df_melt['Parameters'] == 'N_Temperature_m'])
```

Out[322]:
```
[1.0188636500852428]
```

In [347]:
```
# Creating a loop to calculate the ratios on other variables
final={}
for b in df_melt['Parameters']:
    par = df_melt[df_melt['Parameters']==b]
    final[b]=create_temp_ratio(par)
print(final)
```

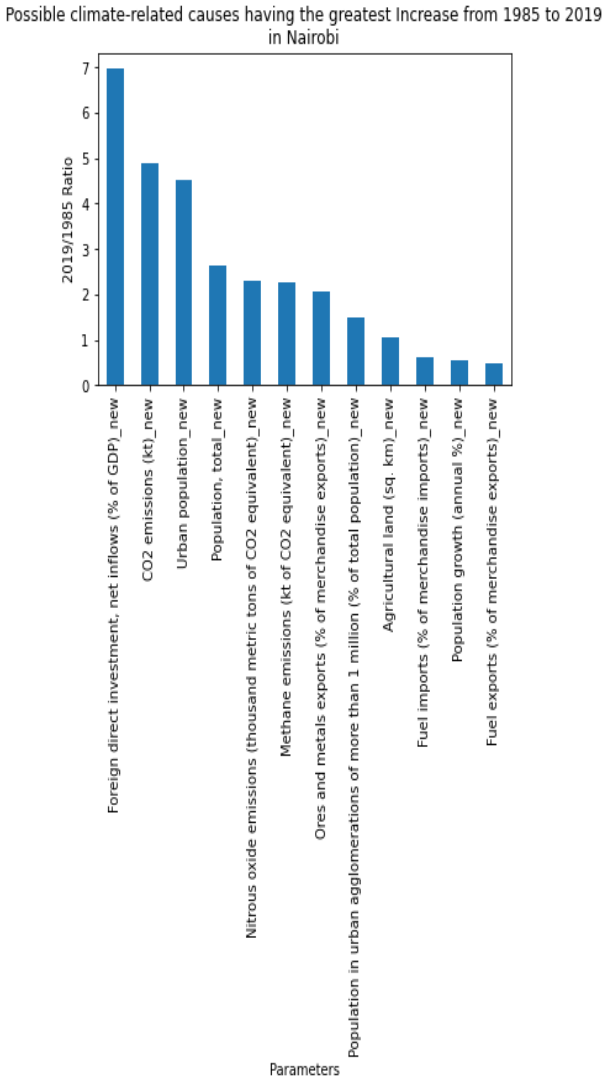
```
{'Year_extract': [1.0171284634760704], 'Fuel exports (% of merchandise exports)_new': [0.4726518872755196], 'Fuel imports (% of merchandise imports)_new': [0.6120599924955565], 'Mineral rents (% of GDP)_new': [-57.68436391171622], 'Ores and metals exports (% of merchandise exports)_new': [2.0492771050517673], 'Agricultural land (sq. km)_new': [1.0484217391409845], 'CO2 emissions (kt)_new': [4.88161667891709
7], 'Foreign direct investment, net inflows (% of GDP)_new': [6.97112479529732], 'Methane emissions (kt of CO2 equivalent)_new': [2.27978
1924205935], 'Mortality rate, under-5 (per 1,000 live births)_new': [0.44411589937190527], 'Nitrous oxide emissions (thousand metric tons
of CO2 equivalent)_new': [2.303785669968179], 'Population growth (annual %)_new': [0.5653222498187518], 'Population in urban agglomeratio
ns of more than 1 million (% of total population)_new': [1.491785482646817], 'Population, total_new': [2.6449544730193884], 'Urban popula
tion_new': [4.52483196259996], 'N_Temperature_m': [1.0188636500852428], 'N_Precipitation_Total': [6.684931026776905], 'N_Relative_Humidit
y_m': [0.9878956974134134], 'N_Wind_Speed_10m': [0.8916294349510168], 'N_Cloud_Cover_Total': [1.0467007960301582], 'N_Sunshine_Duration':
[0.87180751080150 5], 'N_Shortwave_Radiation': [0.8489636510809363], 'N_Mean_Sea_Level_Pressure': [1.0001023389861163], 'N_Soil_Temperatur
e_10cm': [1.0021401124965323], 'N_Soil_Moisture_10cm': [1.2316621558438265], 'N_Wind_Direction_Dominant_10m': [0.9996273373214357]}
```

Fig.6 Ratios of Temperature only (Top) vs. Ratios of the remaining possible causes and consequences variables in climate change in Nairobi, Kenya from 1985 to 2019 (Bottom).

biggest increase in that time period. An initial test was performed on one variable only (Temperature) before creating a loop over all the other variables (causes and consequences).

The top three possible causes and consequences that saw the biggest increase from 1985 to 2019 are displayed in the two histograms below. The biggest cause was the foreign direct investment (by percentage of GDP), followed closely by the C02 emissions, while the precipitation dominated the top possible leading consequence of climate change in Nairobi during this time period. Quick relationship plots show that:

- The temperature is still very hard to link with C02 emissions even though a direct relationship trend can be noticeable
- The precipitation seems to go hand in hand with the CO2 emissions and foreign direct investments



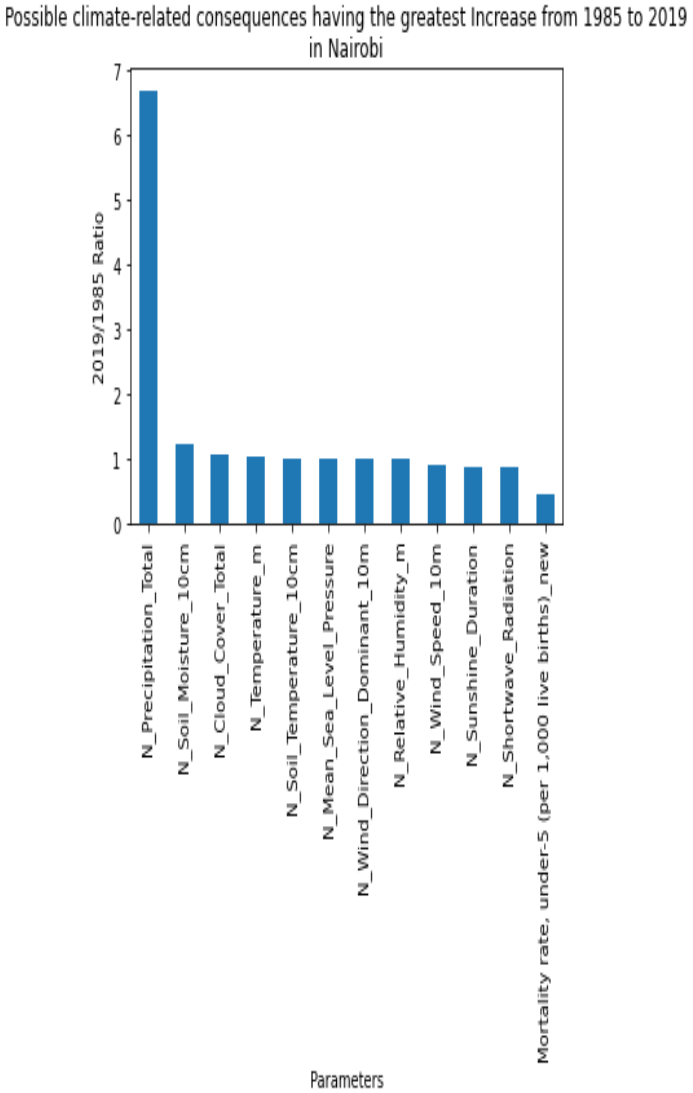


Fig.7 Top 10 of the greatest increase of possible climate-related causes (left) and consequences (right) from 1985 to 2019 in Nairobi, Kenya.

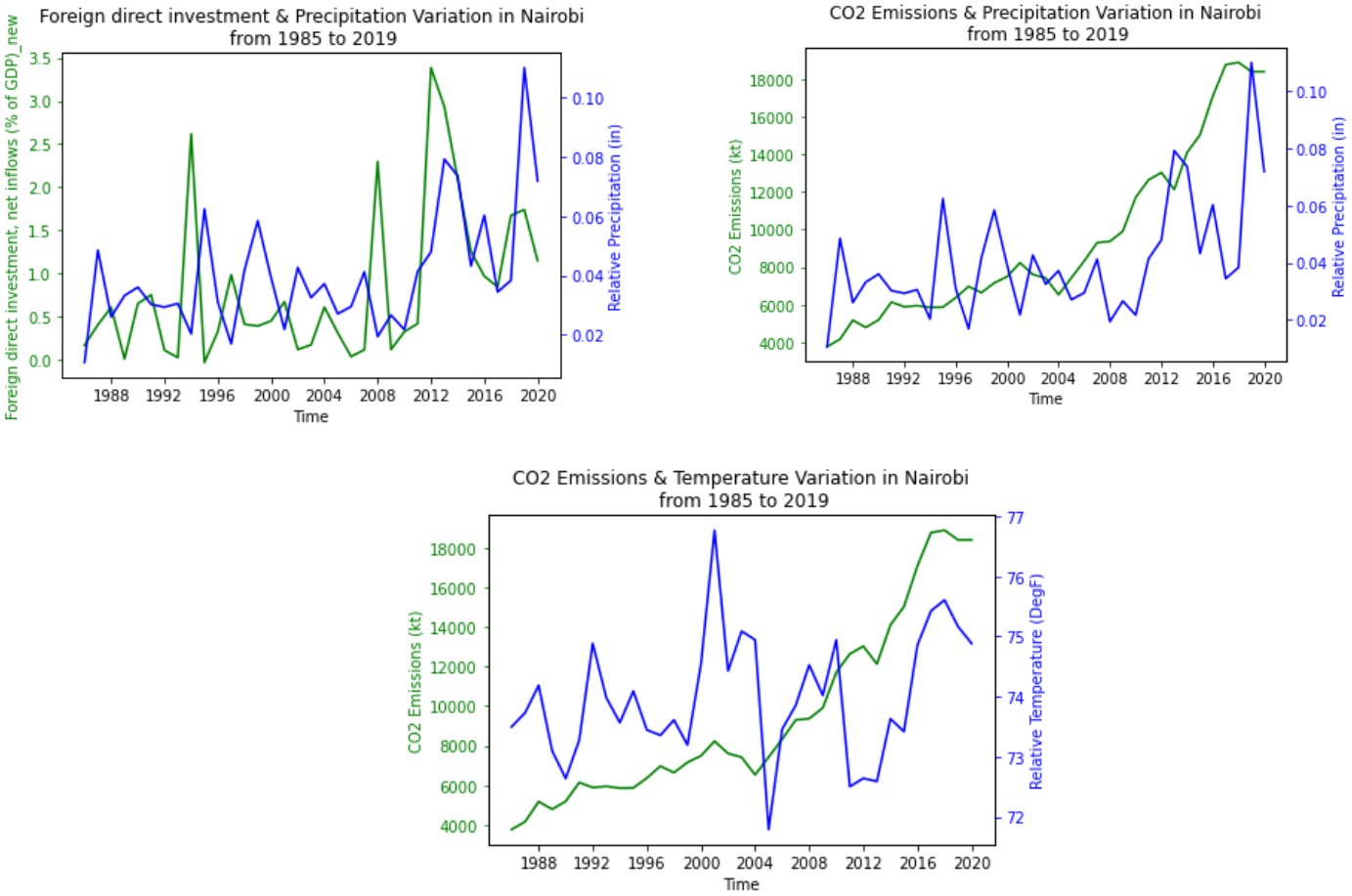


Fig.8 Foreign direct investment vs precipitation variation in Nairobi (Top left), CO2 emissions vs. precipitation variation in Nairobi (Top right) and CO2 emissions vs. Temperature variation in Nairobi (Bottom). From 1985 to 2019.

9. Correlation Matrix to identify global relationship trends between all those parameters

One way to identify relationships between all those variables is to use the Pearson coefficient which is a measure of linear association between variables. It has a value between -1 and 1 where: -1 indicates a perfectly negative linear correlation while +1 would reflect a perfectly positive correlation instead.

From the matrix, the following main observations could be inferred:

- C02 seems to be more correlated with the following: the Total & Urban Population & Methane emissions. Interestingly, it is also related to the mortality rate (for children under 5 years old). The latter may be surprising, but it may be explained by the fact that infants do not spend as much time as do adults or older kids?

- Nitrous oxide emissions appear to have more correlations than CO2 does: Agricultural land, Methane emissions, populations in urban agglomerations and the total population. Like CO2, it is also inversely related with the mortality rate for infants under five years old.

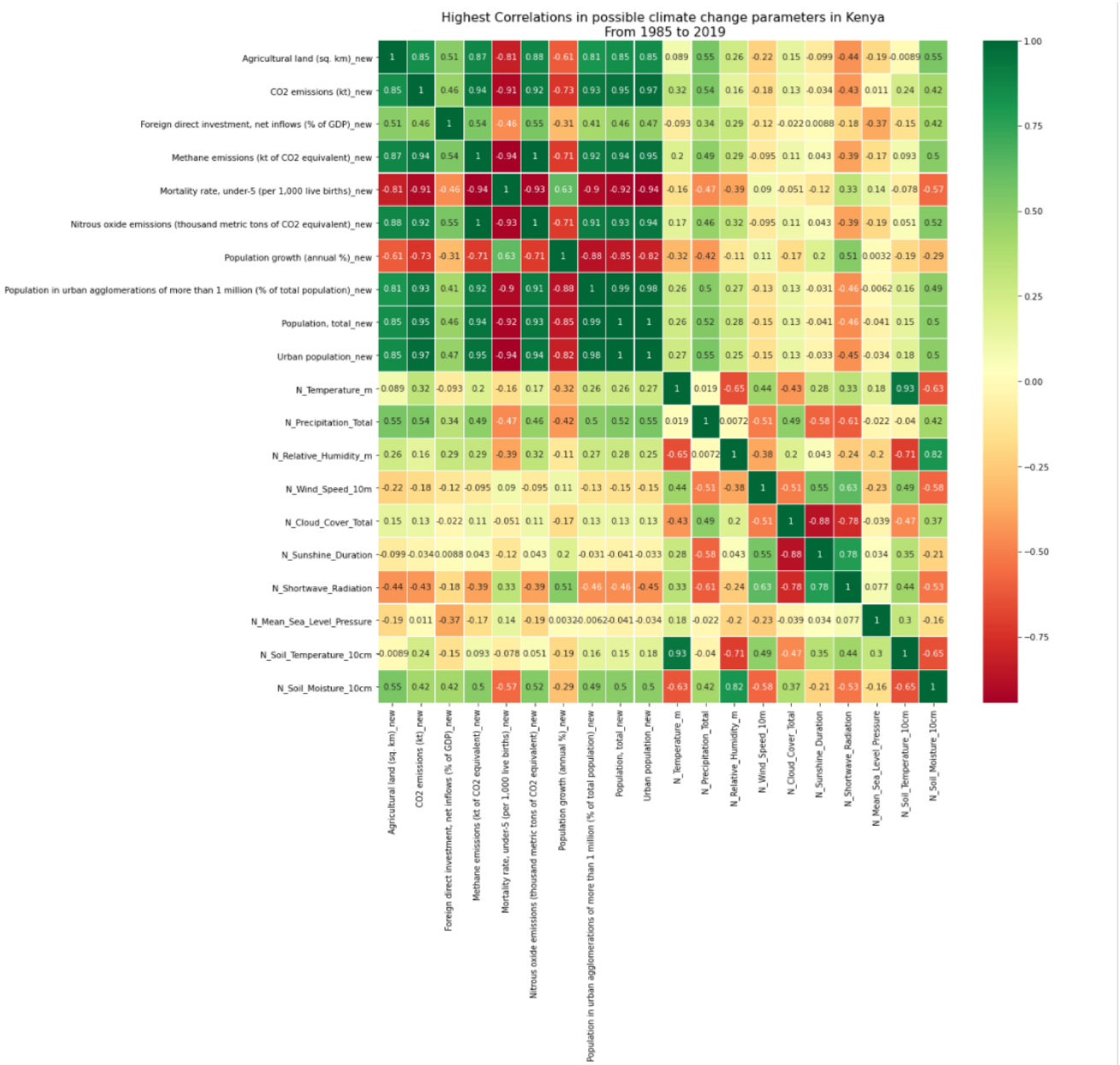


Fig.9 Highest Pearson-based correlations in possible climate change parameters in Kenya from 1985 to 2019.

- Precipitation seems to be related to Ores & metal exports, agricultural land, CO2 emissions, foreign direct investments, Methane & Nitrous Oxide emissions, population in urban agglomerations, the total population, the total cloud cover and the soil moisture.

The temperature variation, as expected and seen earlier, is hard to use to assess obvious relationships with other variables. However, we could note some positive relationships with the following: Ores & metals exports, CO2 emissions, total population growth, the urban population, the sunshine duration and the soil temperature.

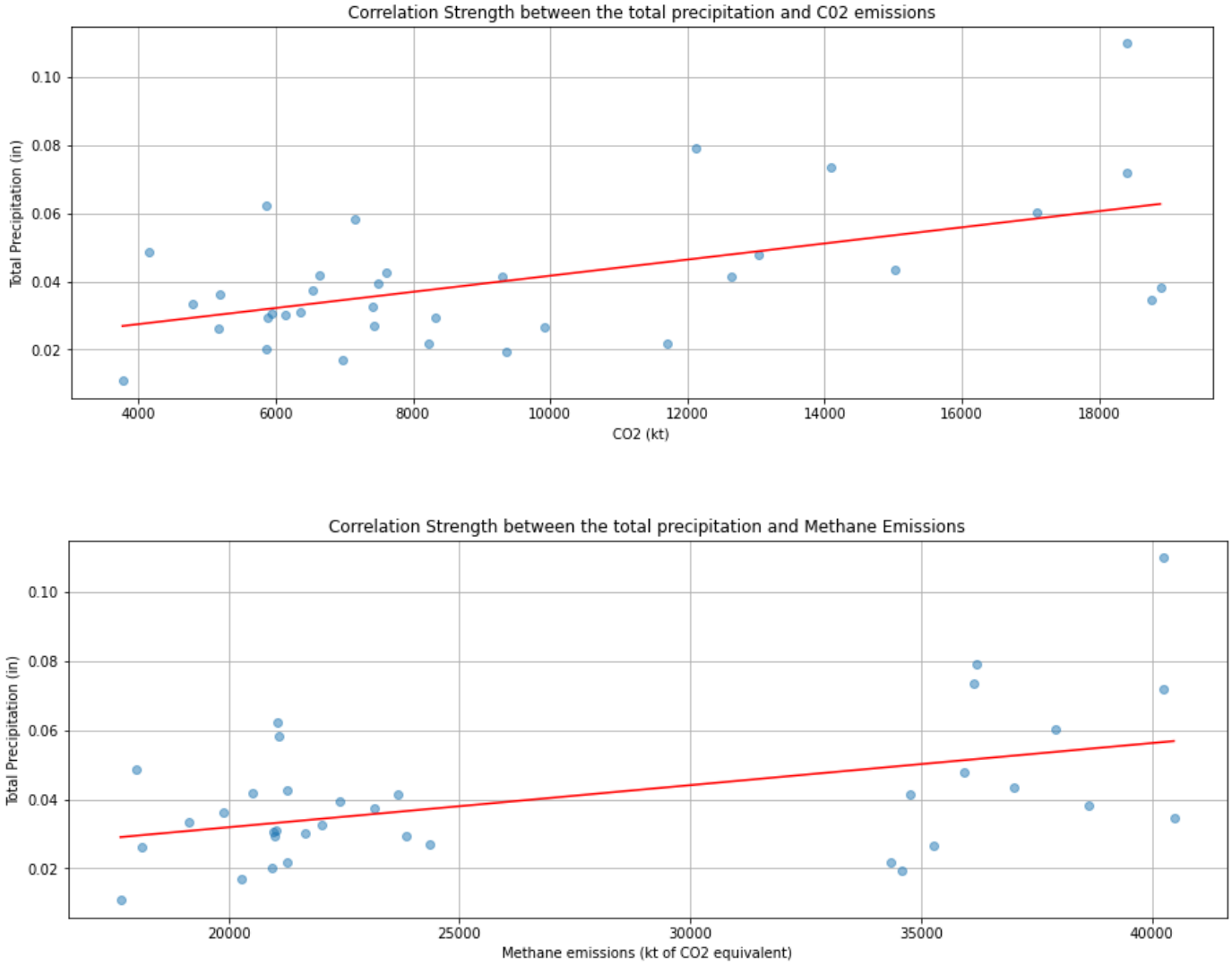


Fig.10 Correlation strength between the total precipitation and CO2 emissions (Top) and methan emissions (bottom)

10. Pre-Processing, Training Data Development and Modeling

This last part focuses on leveraging the cleaned and processed dataset to make predictive insights. Machine learning is used to assess the biggest positive correlations between causes and consequences for Temperature and precipitation changes for the last 30 years in Kenya.

Machine Learning Background

With the field of machine learning, there are two main types of tasks: supervised and unsupervised. The main difference between the two types is that supervised learning hails from having prior knowledge of what the output values for our samples should be. Therefore, the goal of supervised learning is to learn a function that, given a sample of data and desired outputs, best approximates the relationship between input and output observable in the data. Unsupervised learning, on the other hand, does not have labeled outputs, so its goal is to infer the natural structure present within a set of data points. Two supervised-based models were used: Ordinary Least Squares Regression and Random Forest.

11. Data Input for Predictions using Ordinary Least Squares (OLS)

A total of eight different features were selected to predict temperature and precipitation from 1985 to 2019. While, both of those two variables will be assessed independently, the features remained the same. Those 8 features were selected based on the Pearson correlation matrix seen in the previous part and are:

- CO2 emissions
- Methane emissions
- Nitrous oxide emissions
- Total population
- Urban population
- Relative Humidity
- Mean sea level pressure
- Soil moisture

The OLS, is a type of linear least squares method for estimating the unknown parameters in a linear regression model. OLS chooses the parameters of a linear function of a set of explanatory variables by the principle of least squares: minimizing the sum of the differences between the observed features and those predicted by the linear function of the two independent variables (Precipitation and Temperature).

12. Results of predictions of Temperature and Precipitation.

For a very simple model such as OLS, the accuracy results does not seem too bad, but after several iterations, the score seemed to still hover around a 61% accuracy.

```
[22]: # Instantiate the regressor class
      regressor = LinearRegression()

      # Fit and build the model by fitting the regressor to the training data
      regressor.fit(X_train, y_train)

      # Make a prediction set using the test set
      prediction = regressor.predict(X_test)

      # Evaluate the prediction accuracy of the model
      print("Accuracy of Linear Regression: %.2f" % regressor.score(X_test, y_test))
      print("The Mean Absolute Error: %.2f degrees fahrenheit" % mean_absolute_error(y_test, prediction))
      print("The Median Absolute Error: %.2f degrees fahrenheit" % median_absolute_error(y_test, prediction))

[22]: LinearRegression()
      Accuracy of Linear Regression: 0.61
      The Mean Absolute Error: 1.94 degrees fahrenheit
      The Median Absolute Error: 1.60 degrees fahrenheit
```

Fig.11 Metrics from the Linear Regression after 50+ iterations.

13. Random Regressor with Random Forest

A model that can work very well in a lot of cases is the Random Forest. For regression, this is provided by Sklearn's RandomForestRegressor class.

A pipeline was designed to assess the performance using cross-validation. The latter was performing the fitting as part of the process. I first used the default settings for the random forest and then went on to investigate some different hyperparameters.

The cross validation mean absolute error for the Temperature and the Precipitation were: 0.84 Deg F and 0.03 in respectively.

Similarly, the mean absolute error for both variables were lower from the Random Forest than from the Linear Regression results.

Final Model Selection for Temperature

Linear regression model performance

[228]:
```
# 'neg_mean_absolute_error' uses the (negative of) the mean absolute error
lr_neg_mae = cross_validate(lr_grid_cv.best_estimator_, X_train, y_train,
                            scoring='neg_mean_absolute_error', cv=5, n_jobs=-1)
```

[229]:
```
lr_mae_mean = np.mean(-1 * lr_neg_mae['test_score'])
lr_mae_std = np.std(-1 * lr_neg_mae['test_score'])
lr_mae_mean, lr_mae_std
```

[229]: (0.9705988486490866, 0.009298283213350815)

[230]:
```
mean_absolute_error(y_test, lr_grid_cv.best_estimator_.predict(X_test))
```

[230]: 0.9582399477626851

Random forest regression model performance

[231]:
```
rf_neg_mae = cross_validate(rf_grid_cv.best_estimator_, X_train, y_train,
                            scoring='neg_mean_absolute_error', cv=5, n_jobs=-1)
```

[232]:
```
rf_mae_mean = np.mean(-1 * rf_neg_mae['test_score'])
rf_mae_std = np.std(-1 * rf_neg_mae['test_score'])
rf_mae_mean, rf_mae_std
```

[232]: (0.8506507086401445, 0.007173446937250973)

[237]:
```
mean_absolute_error(y_test, rf_grid_cv.best_estimator_.predict(X_test))
```

[237]: 0.8414370581985658

The random forest model has a lower cross-validation mean absolute error. It also exhibits less variability. Verifying performance on the test set produces performance consistent with the cross-validation results.

Final Model Selection for Precipitation

Linear regression model performance

[137]:
```
# 'neg_mean_absolute_error' uses the (negative of) the mean absolute error
lr_neg_mae = cross_validate(lr_grid_cv.best_estimator_, X_train, y_train,
                            scoring='neg_mean_absolute_error', cv=5, n_jobs=-1)
```

[138]:
```
lr_mae_mean = np.mean(-1 * lr_neg_mae['test_score'])
lr_mae_std = np.std(-1 * lr_neg_mae['test_score'])
lr_mae_mean, lr_mae_std
```

[138]: (0.0505739255377571, 0.0019226454034038647)

[139]:
```
mean_absolute_error(y_test, lr_grid_cv.best_estimator_.predict(X_test))
```

[139]: 0.055810620582645236

```
##### Random forest regression model performance
```

[140]:
```
rf_neg_mae = cross_validate(rf_grid_cv.best_estimator_, X_train, y_train,
                            scoring='neg_mean_absolute_error', cv=5, n_jobs=-1)
```

[141]:
```
rf_mae_mean = np.mean(-1 * rf_neg_mae['test_score'])
rf_mae_std = np.std(-1 * rf_neg_mae['test_score'])
rf_mae_mean, rf_mae_std
```

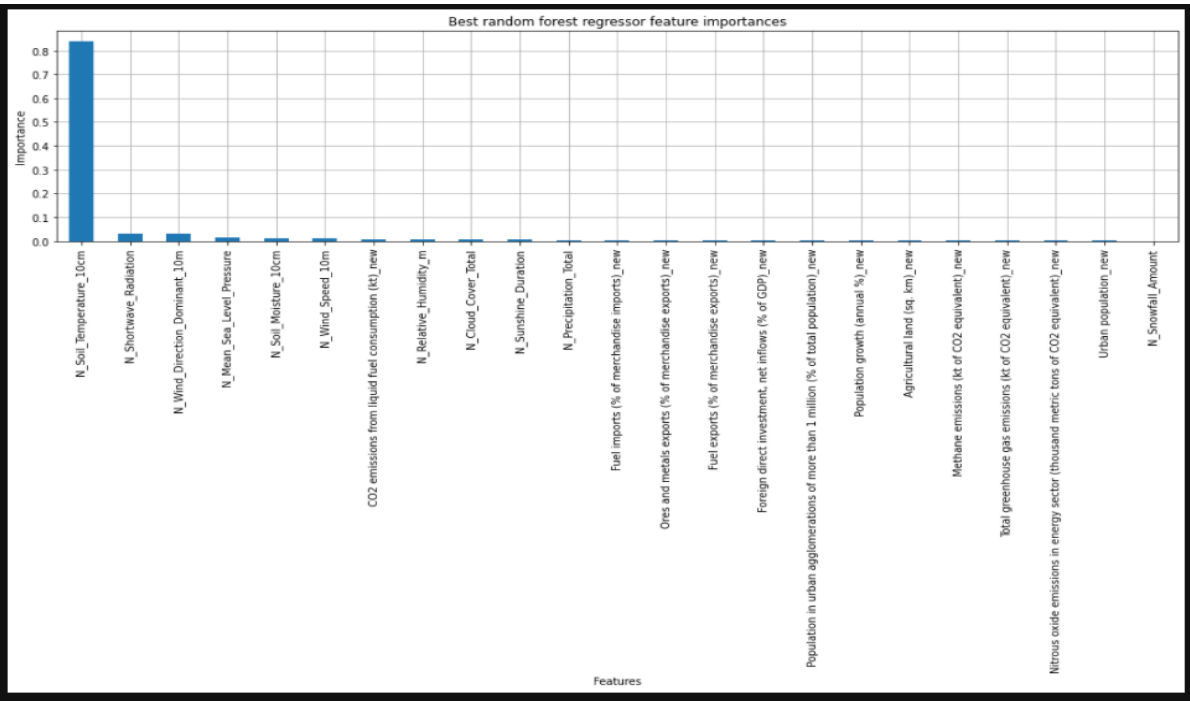
[141]: (0.02744622065722029, 0.0016177603704547372)

[142]:
```
mean_absolute_error(y_test, rf_grid_cv.best_estimator_.predict(X_test))
```

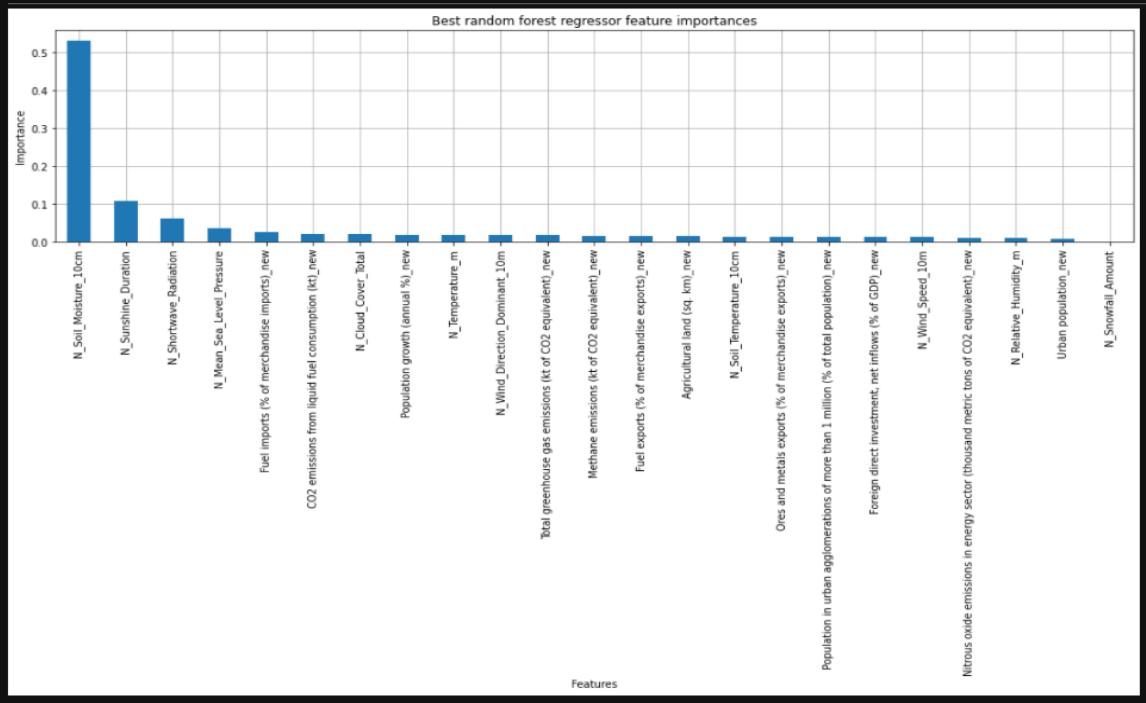
[142]: 0.03048765922513634

The random forest model has a lower cross-validation mean absolute error. It also exhibits less variability. Verifying performance on the test set produces performance consistent with the cross-validation results.

Fig.12 Metrics from the Linear Regression after 50+ iterations compared with Random Forest metrics results




```
These results suggest that Soil Temperature is the biggest positive features (in terms of consequences).
Similarly, the first positive features in terms of possible causes are: CO2 emissions, but this is very small, so conclusion cannot be drawn for the temperature alone.
This makes intuitive sense and is consistent with what we saw during the EDA work.
Let' shave a look at the precipitation instead to see if we can isolate some more prominent causes.
```




```
These results suggest that Soil Moisture, Sunshine Duration, Shortwave Radiation, Mean Sea Level Pressure are among the biggest positive features (in terms of consequences).
Similarly, the biggest positive feature sin terms of possible causes are: Fuel Imports & CO2 emissions!
This makes intuitive sense and is consistent with what we saw during the EDA work.
```

Fig.13 Best Random Forest Regressor feature importance from the Temperature (Top) and Precipitation (Bottom) predictions

14. Conclusion

This project, far from being exhaustive, used limited information and used the assumption that data from capitals (in that case Nairobi) could be extrapolated to a whole country. As such, more work should be done to gather more data from other cities for a single country and enhance the machine learning by increasing the number of training dataset.

The first step in developing a model was to examine performance by using the mean of temperature and precipitation individually. This proved to be helpful in establishing a baseline for comparison, however it was not as useful or as accurate as a linear model or random forest model. If I predicted those variables by using the mean, on average we would be off by about 3 Deg F and 0.27 in for the Temperature and Precipitation respectively.

In the process of building the linear model, missing values were imputed with the median and mean values. However, the initial linear model was overfitting and needed to be adjusted by the number of features. Through cross-validation, the value of k was set to focus on: Soil Temperature and CO2 emissions for the Temperature. Similarly, the value of k was set to focus on: Soil Moisture, Sunshine Duration, Shortwave radiation, Mean sea level pressure, Fuel Imports and CO2 emissions. These features fit our initial assumptions from the EDA.

After testing both the linear model and random forest model, the best performance was from using the forest regression model. Comparison of the two demonstrated that performance on the test set was consistent cross-validation results. Additionally, the cross-validation mean absolute error was lower using the random forest regressor.

Despite the limitations in data availability, the Data Science Method has proven to be a very useful tool to extract key information that could be used to gain quick insights and lead the ground for future more detailed research work.