EMISSIONS IN DEVELOPED AND DEVELOPING COUNTRIES - REPORT

1. Problem Selection. Identify a real-world problem (for example, predicting the number of votes cast for the Democratic and the Republican parties in each county during the 2018 United States Senate elections) and propose a data science solution (for example, building linear regression models). Describe your problem and your solution.

The utilization of energy is tied with the level of development of a given country, given that the vast majority of energy generation produces CO2 as a byproduct, the quantity of CO2 which a country emits should be a good predictor of its level of development, especially when coupled with a countries population, (rate of CO2 production per person). Using data on CO2 production, we aim to classify nations as developed or developing, additional based on historical trends if a country is classified as developing, to determine when this country would reach the status of developed.

In order to accomplish the above, we would incorporate the following task, **Clustering** and **Classification**. Classification. First, we would classify countries into 2 classes developed and developing, then as mentioned above, those countries would have their future levels predicted and would be classified again, this process will repeat until the given country is classified as developed. Clustering. Use different clustering methods to cluster countries into 2 groups developed/developing using variables that would fit the model the best.

2. Data Collection. Identify one or more datasets relevant to your problem. Describe your datasets.

For this project, 2 datasets are collected.

1. The emissions data that contains CO2, N2O, CH4 emissions and emissions for power sector, transportation, buildings, other industry, ratio per gdp and ratio per capita for each country and for each year starting from 1970 to 2012. This dataset is selected from the CORGIS Datasets Project.
https://think.cs.vt.edu/corgis/csv/emissions/

2. As the project tries to predict the countries development status, a true set of data containing the list of countries that are developed and developing was required to test the result of the model. There were no separate dataset that contains the development status of each country. Hence, this dataset was created for this project, by scraping the web. This dataset contains 2 columns, the name of the country and its corresponding development status. The development status is a binary valued column, 1 representing a developed country and 0 representing a developing country.

3. Data Preparation. Detect and correct data quality problems (missing data, noise, outliers, etc.) and transform the data into an appropriate format for data analysis. Describe your data preparation process and report the results obtained.

Some countries had missing data usually for the earlier years and they were not used. Some outliers that were encountered was North Korea had some very high pollution ratios per capita. For example, South

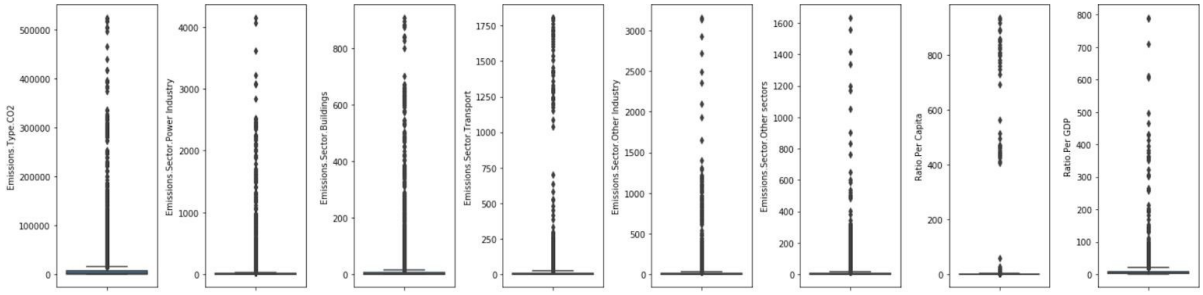
Korea had a pollution ratio per capita of 0.409918 for the year 2012 while North Korea had a ratio per capita of 728.6689. For classification we needed to merge the 2012 data with the development data set by country.

The data was split into 6 groups and each group was modeled using the best classifier and clustering method available.

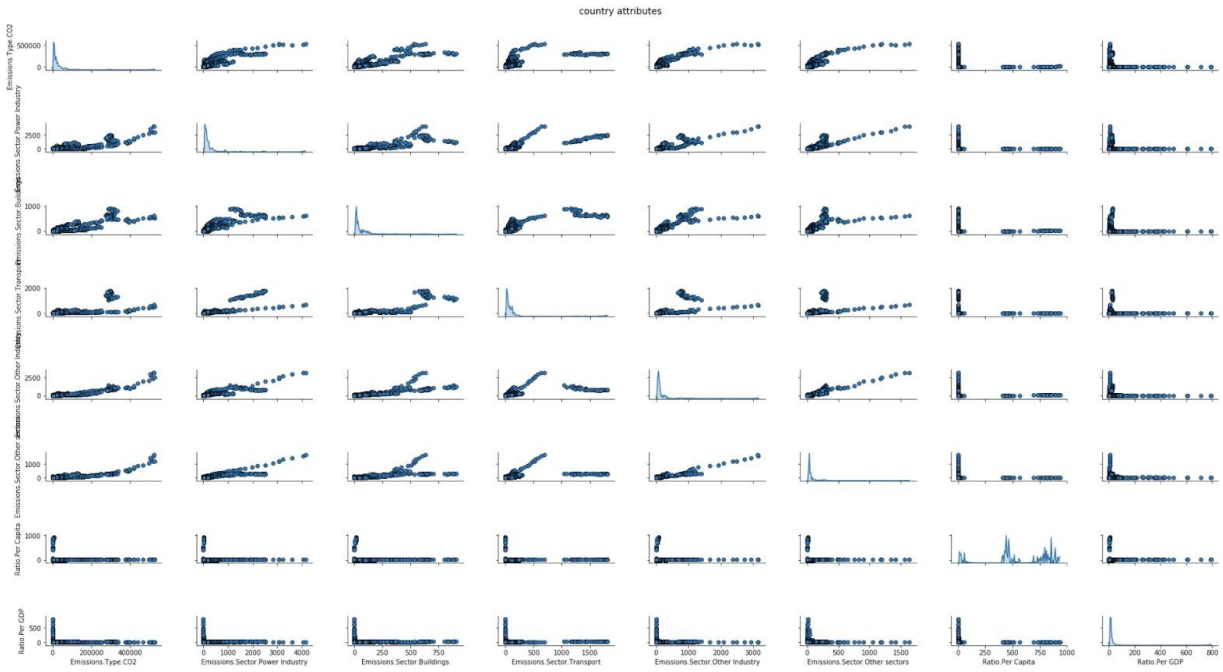
1. The most recent year- 2012
2. Mean of all years
3. Years 1970-1979
4. Years 1980-1989
5. Years 1990-1999
6. Years 2000- 2012

4. Data Exploration. Explore the data using summary statistics and plots and identify the most important variables for data analysis. Describe your data exploration process and report the results obtained.

For the data exploration part, the dataset containing the emissions data is used. As the model needs the predictors that has a higher bias towards the result, some of the columns such as the Emissions.Type.N2O, Emissions.Type.CH4 were not used as a part of the exploration because the model mainly is built to predict the development status based on the CO2 emissions, these two columns didn't contribute to it. The remaining columns were used for the entire building process of the model. The summary statistics of these columns are calculated and visualized using box plots.



Also, to see how the data in each of these columns are a pair plot is created to match each variable with all other chosen predictors.

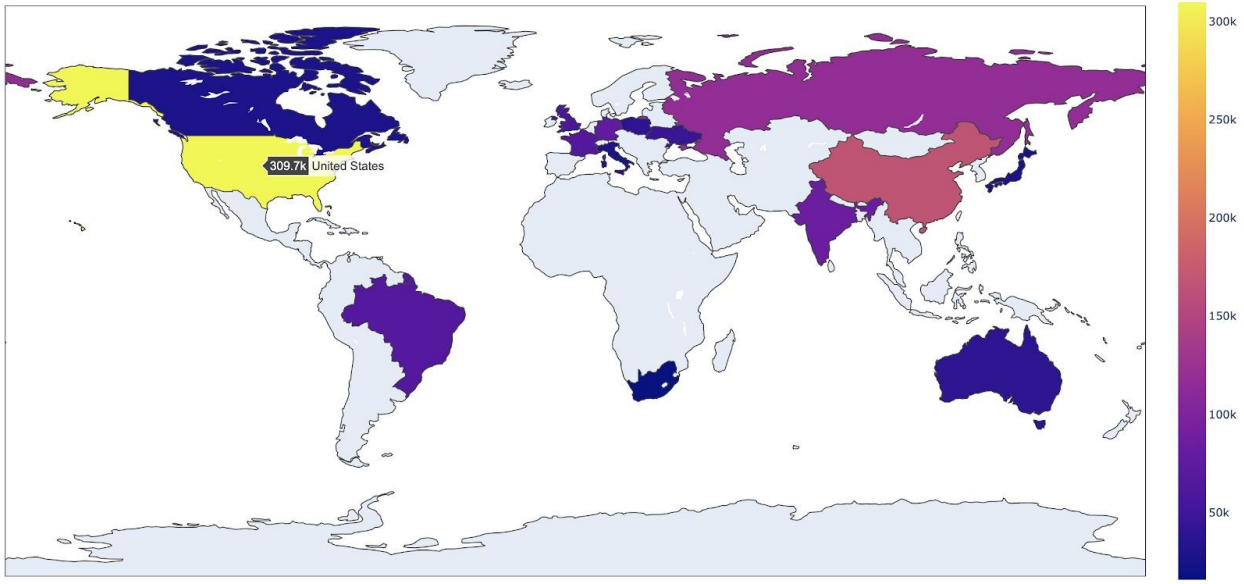


By looking at this visualization, clearly each plot shows how each of the columns are divided into separate clusters based on the distribution of the dataset. Most of the columns contain only 2 clusters, but the Ratio Per Capita column is seen to lie in 3 different clusters.

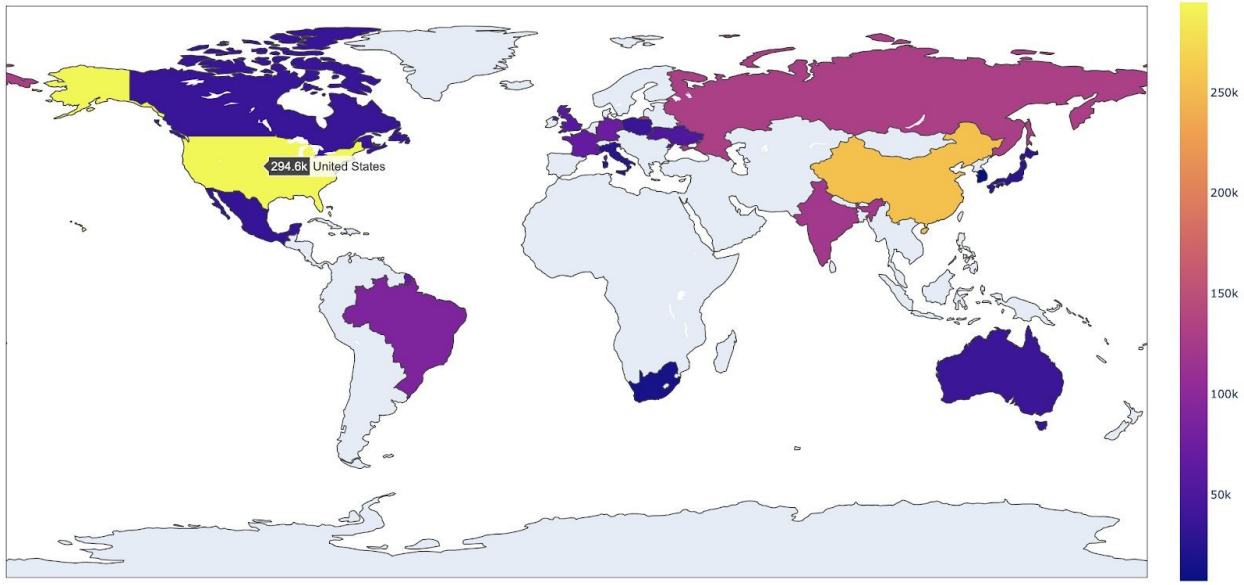
5. Data Modeling. Train and test models using the data. Your data modeling step must include at least two of the following tasks: (1) Regression, (2) Classification, (3) Clustering, and (4) Text analysis. Consider multiple techniques, parameters, and variables. Describe your data modeling process and report the results obtained.

Clustering

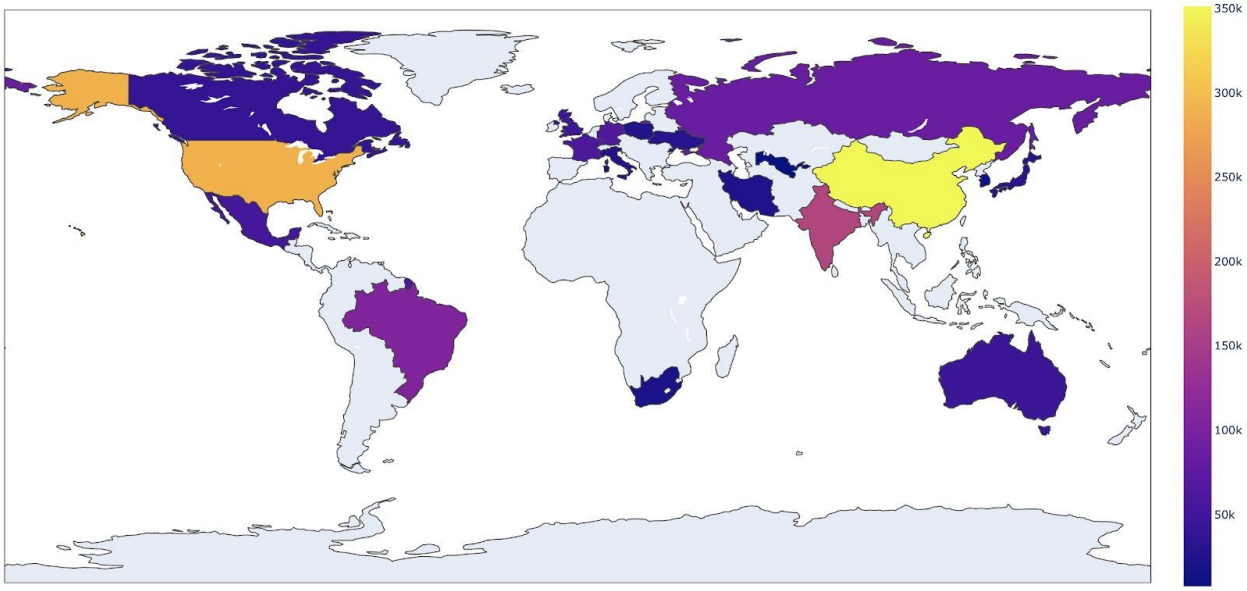
Clustering is used to cluster the countries into two separate clusters, one for developed countries and another for developing countries. The technique used for clustering is DBSCAN. This technique was picked because this technique will do a good job in selecting areas which has high density of observations both for the developed and developing countries. The clustering is done on many subsets of data. It is done on all the six groups of data that was split before. The results of the clustering for each group of data is shown below,



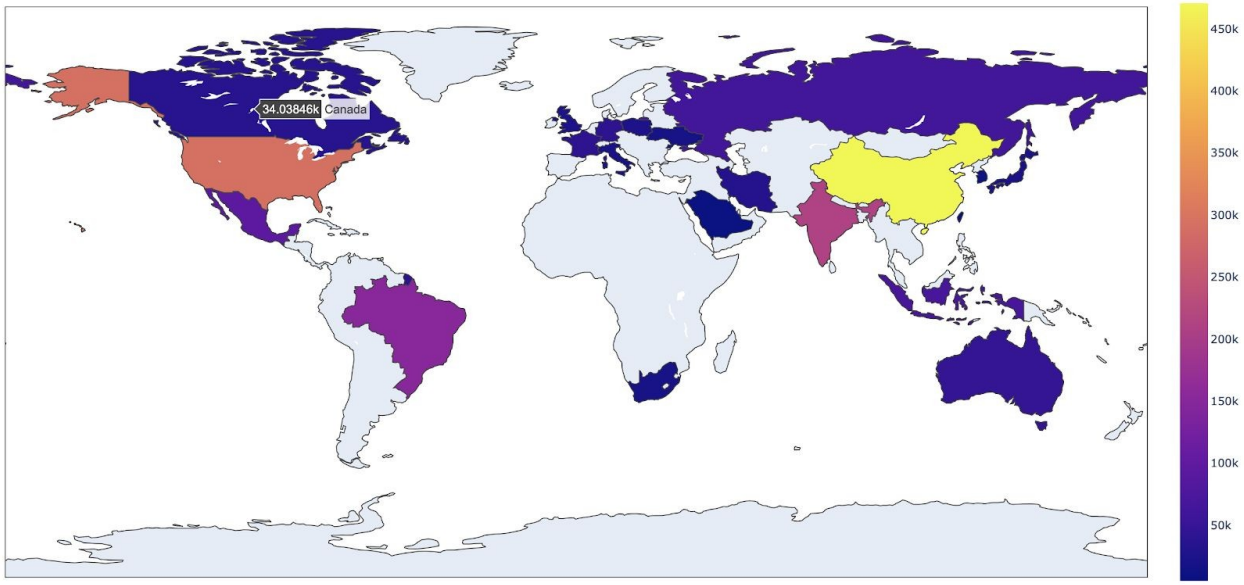
1970 - 1979



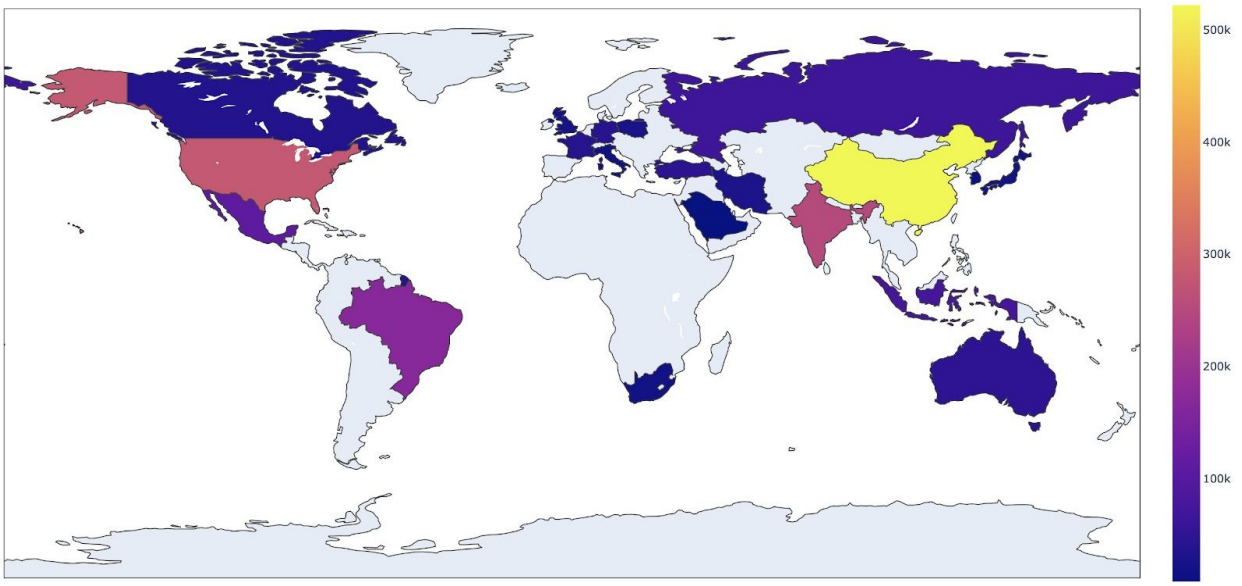
1980 - 1989



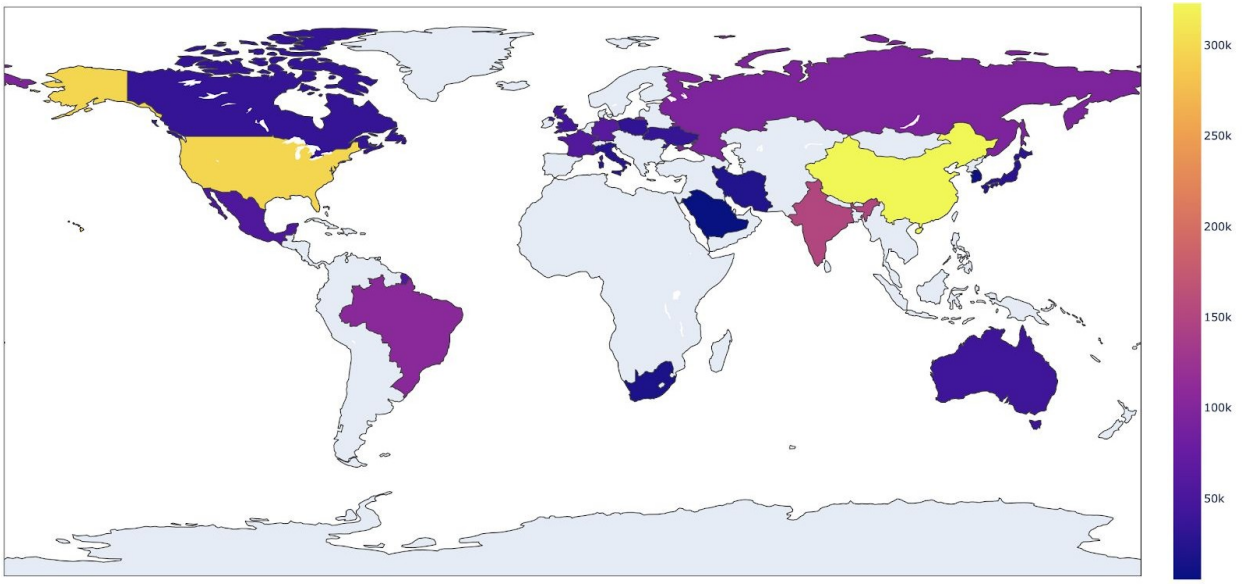
1990 - 1999



2000 - 2012



2012



All years (1970 - 2012)

On observing all these different maps, it is evident that if a country emits **50 - 100 kilotons of CO2** by all major sectors operating in a country, it can be labelled as a developed country. Also, the western countries such as the countries in North America and South America and also Australia remain constant in all the different subsets of data, whereas the small countries in Africa, Europe and Asia tends to get changed into the status of developed countries in the increase in the number of years of emitting CO2.

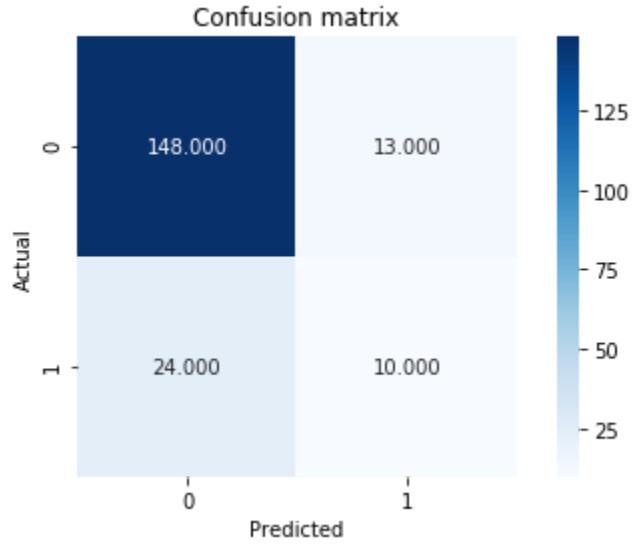
Classification

The first step was to split the data into training and test using a 75/25% split respectively using “Emissions.sector.power.industry”,”Emissions.sector.buildings”, ”Emissions.sector.transport” as the variables for x and data with “development status” as the variable for y. Then we standardized the data and fitted it to a Linear SVM model. K nearest neighbors was also tried but it was a little too strict in and would classify almost every country as developing which didn’t sound right. Then we used the predict function with our data to get prediction results for all the countries. After the prediction we built a confusion matrix to see how countries moved categories and also got the classification report. For our result we mostly saw that more than half of developed countries are being classified as developing. We also saw that for every developed country being classified as developed there was also a developing country being classified as developed. The US is classified as developing with this model and canada and the UK for example are classified as developed. Out of all the 6 year groups the data model seems to work best with the group with years 2000-2012.

Scores

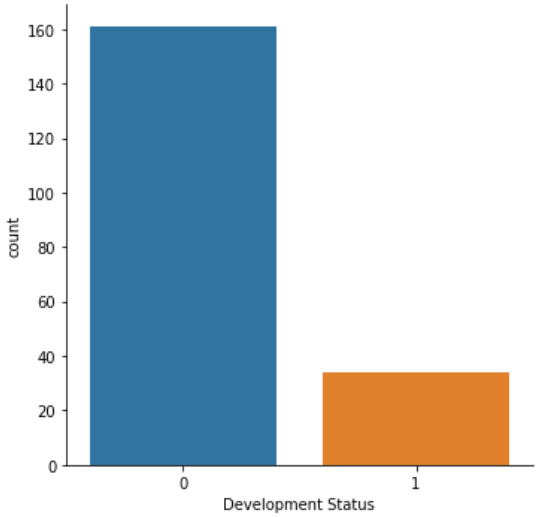
	precision	recall	f1-score	support
0	0.86	0.92	0.89	161
1	0.43	0.29	0.35	34
accuracy			0.81	195
macro avg	0.65	0.61	0.62	195
weighted avg	0.79	0.81	0.80	195

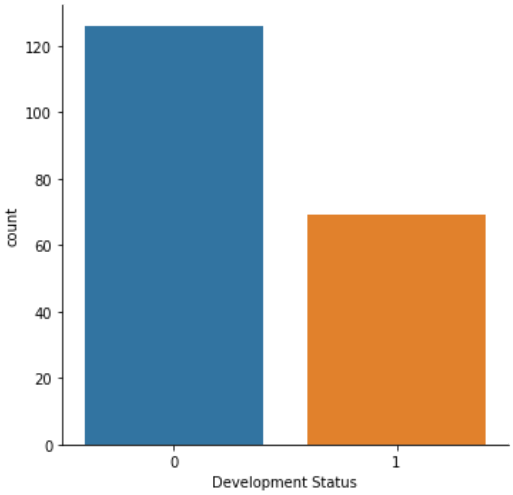
Confusion matrix of the results



Before

After


160
140
120
100
80
60
40
20
0
count
0
1
Development Status


120
100
80
60
40
20
0
count
0
1
Development Status