DengAI: Predicting Disease Spread

Adharsh Rajendran (AXR150830), Mithil Gotarne (MXG180018)

1. Introduction and Problem Definition

Using environmental data from the cities San Juan and Iquitos to predict the number of cases of Dengue fever within a particular time span. As mosquitos thrive in warm and humid climates, countries with these characteristics should have higher cases of dengue fever. Increased amount of precipitation should also contribute to increase of mosquitos and thus cases of dengue fever. Provided with copious amounts of climate data and other factors, we will figure out the large contributors and predict the results of data provided later on.

2. Data Description

1) 3 datasets provided from Driven Data
2) 22 features:
    i. City, year, weekofyear,....
    ii. Climate data is most common, possible to remove correlated features

3. Preprocessing Techniques

a. Process the datasets for the two cities separately, as the geographic distance between the two cities imply the climate data for each is not correlated.
b. Filling in the data not provided:
    i. using median value for feature
    ii. using most frequent category/value of feature

These two methods seem the most useful in capturing data that has not been provided.

Null Values In Dataset

city	0
year	0
weekofyear	0
week_start_date	0
ndvi_ne	194
ndvi_nw	52
ndvi_se	22
ndvi_sw	22
precipitation_amt_mm	13
reanalysis_air_temp_k	10
reanalysis_avg_temp_k	10
reanalysis_dew_point_temp_k	10
reanalysis_max_air_temp_k	10
reanalysis_min_air_temp_k	10
reanalysis_precip_amt_kg_per_m2	10
reanalysis_relative_humidity_percent	10
reanalysis_sat_precip_amt_mm	13
reanalysis_specific_humidity_g_per_kg	10
reanalysis_tdtr_k	10
station_avg_temp_c	43
station_diur_temp_rng_c	43
station_max_temp_c	20
station_min_temp_c	14
station_precip_mm	22

Null Values in Training Data

city	0
year	0
weekofyear	0
week_start_date	0
ndvi_ne	43
ndvi_nw	11
ndvi_se	1
ndvi_sw	1
precipitation_amt_mm	2
reanalysis_air_temp_k	2
reanalysis_avg_temp_k	2
reanalysis_dew_point_temp_k	2
reanalysis_max_air_temp_k	2
reanalysis_min_air_temp_k	2
reanalysis_precip_amt_kg_per_m2	2
reanalysis_relative_humidity_percent	2
reanalysis_sat_precip_amt_mm	2
reanalysis_specific_humidity_g_per_kg	2
reanalysis_tdtr_k	2
station_avg_temp_c	12
station_diur_temp_rng_c	12
station_max_temp_c	3
station_min_temp_c	9
station_precip_mm	5

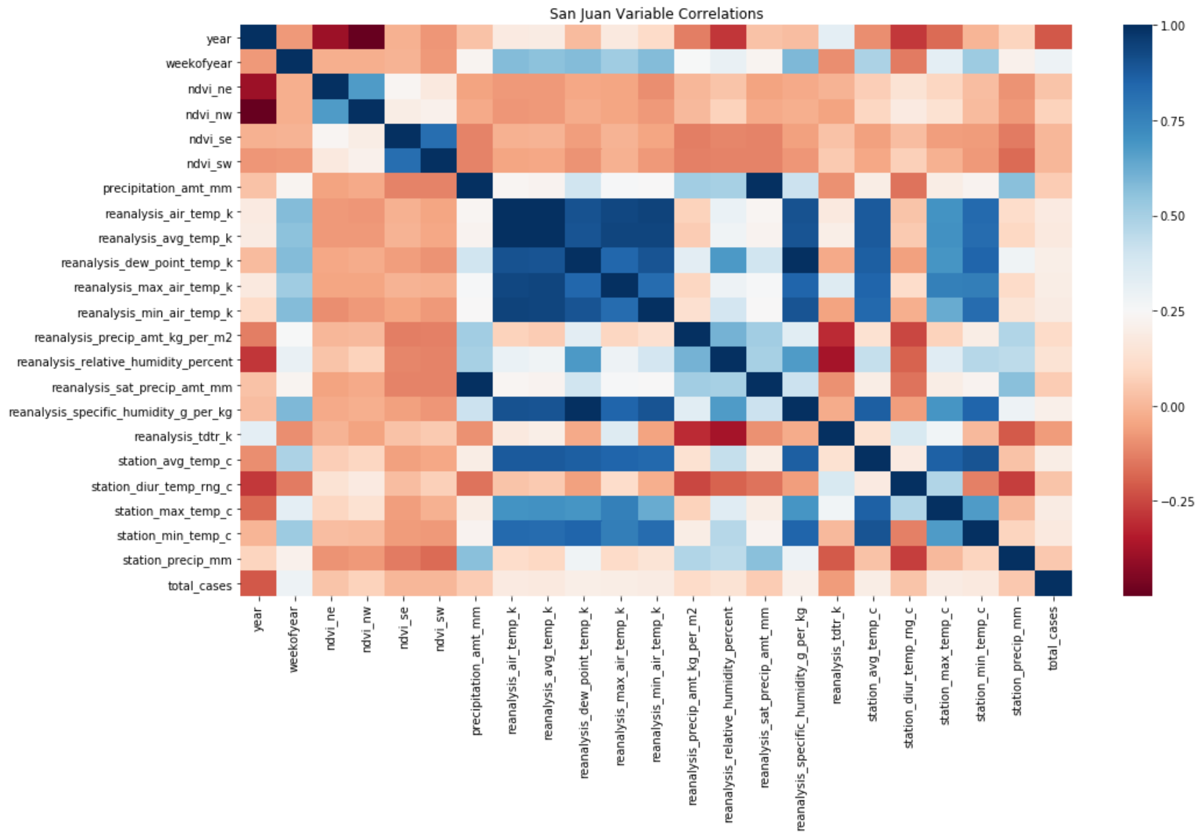
Null Values in Testing Data

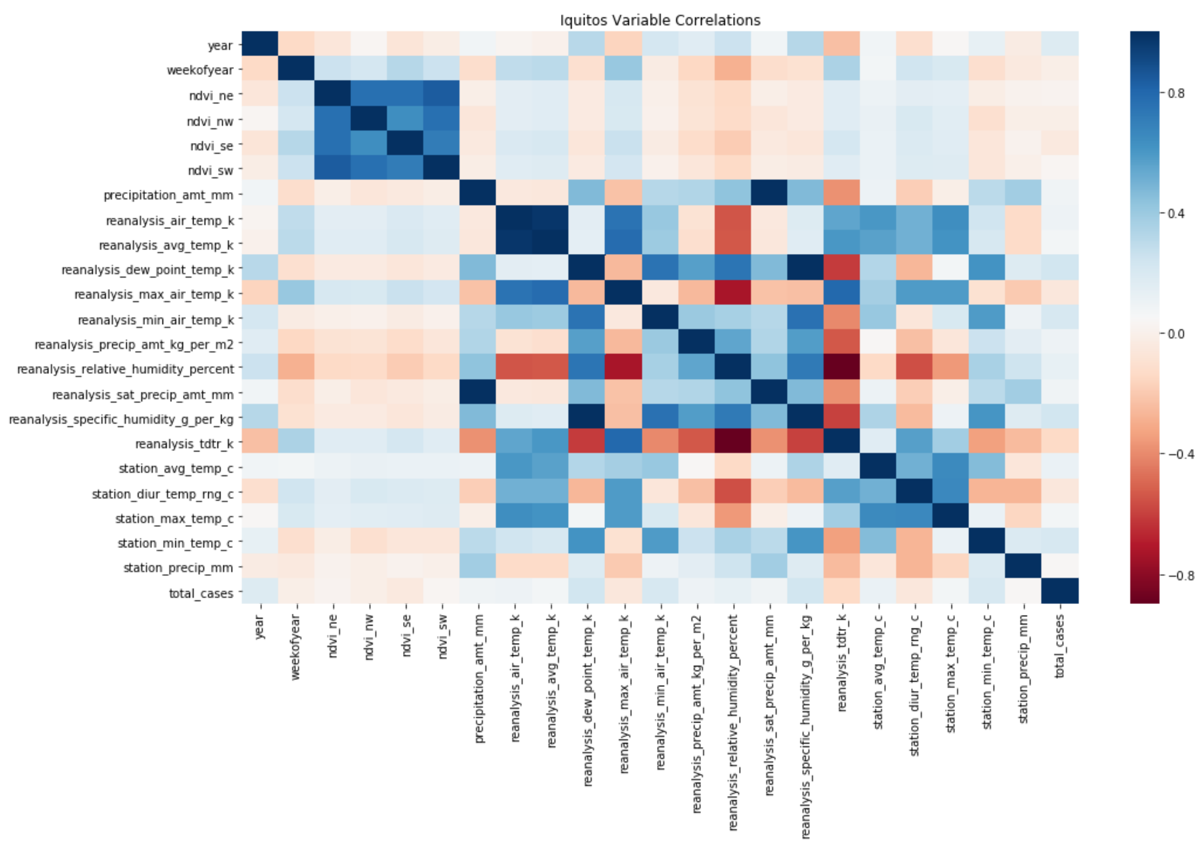
The chart above describes how many of the records within the dataset have null values. By initial observation, the training dataset has far more null values than the test dataset. In order to counteract this issue, we decided on filling in the missing data points rather than eliminating the entire record. The value we chose to replace with is the median value of the column. This is seemed like the most appropriate response as most of the data is climate related. The implication being there is little risk in taking the median of these categories.

4. Method

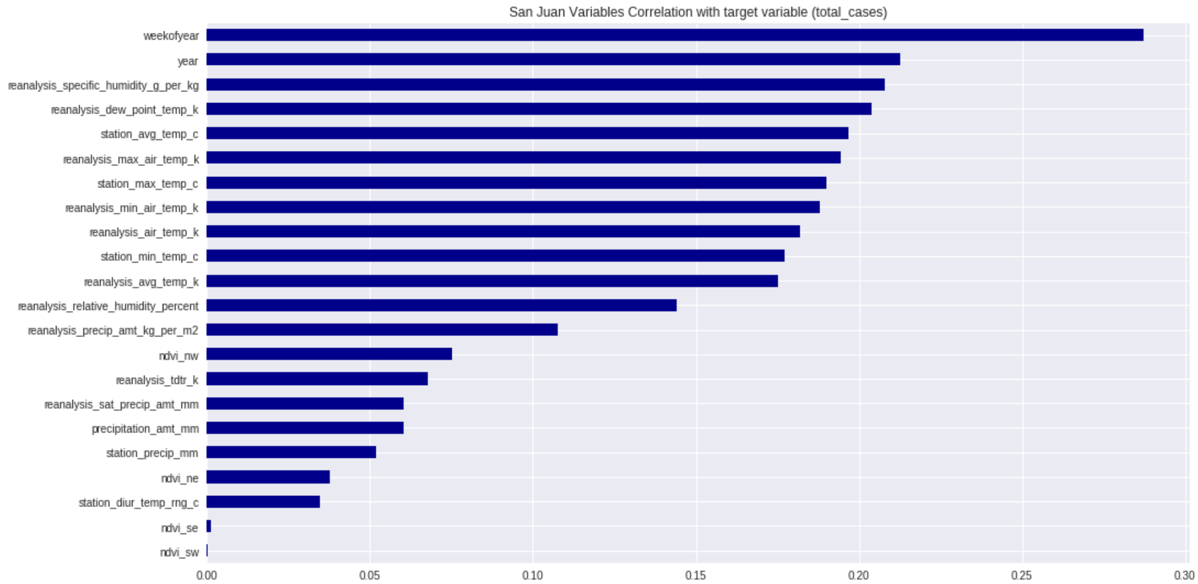
Correlation of Variables in Dataset

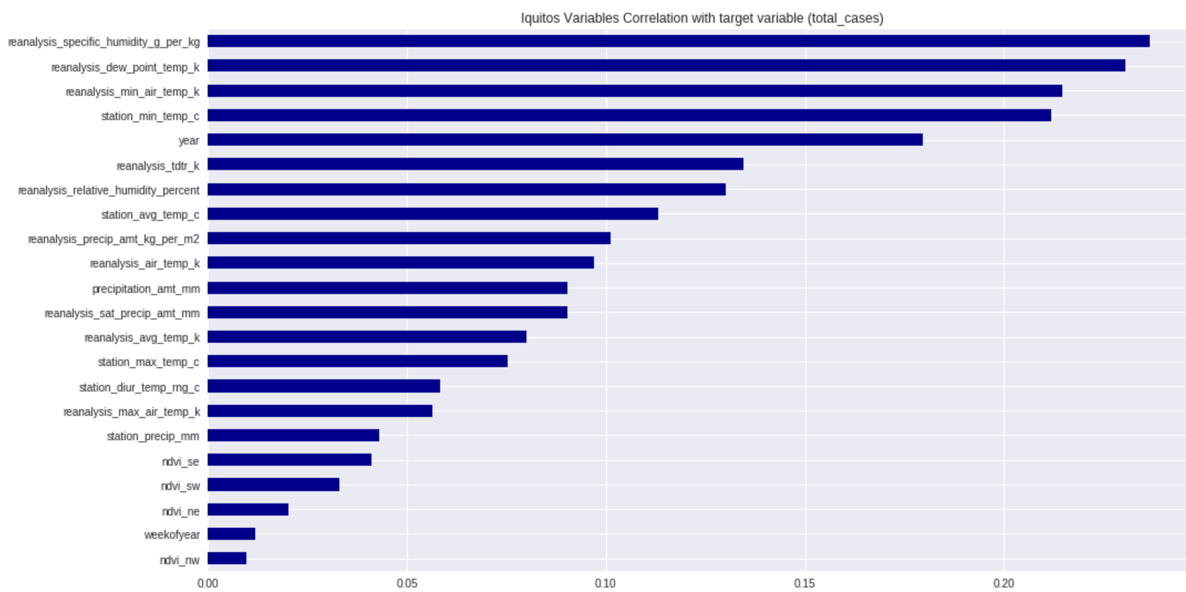
The two heat maps below describe the dataset's internal correlation for both cities. Major correlations of the data involve mostly climate data. For San Juan's data, the prominent examples are temperature related data records. This makes sense as the climate for city will vary very little. Where the data tends to not correlate is with location and humidity. The same results can be said for Iquitos. However the location data for Iquitos is far more correlated but the correlation for the climate data is more concentrated and scattered. The implication for both is some of the climate data can be removed as there is a low variance in it.


San Juan Variable Correlations
year
weekofyear
ndvi_ne
ndvi_nw
ndvi_se
ndvi_sw
precipitation_amt_mm
reanalysis_air_temp_k
reanalysis_avg_temp_k
reanalysis_dew_point_temp_k
reanalysis_max_air_temp_k
reanalysis_min_air_temp_k
reanalysis_precip_amt_kg_per_m2
reanalysis_relative_humidity_percent
reanalysis_sat_precip_amt_mm
reanalysis_specific_humidity_g_per_kg
reanalysis_tdtr_k
station_avg_temp_c
station_diur_temp_rng_c
station_max_temp_c
station_min_temp_c
station_precip_mm
total_cases
1.00
0.75
0.50
0.25
0.00
−0.25


Iquitos Variable Correlations
year
weekofyear
ndvi_ne
ndvi_nw
ndvi_se
ndvi_sw
precipitation_amt_mm
reanalysis_air_temp_k
reanalysis_avg_temp_k
reanalysis_dew_point_temp_k
reanalysis_max_air_temp_k
reanalysis_min_air_temp_k
reanalysis_precip_amt_kg_per_m2
reanalysis_relative_humidity_percent
reanalysis_sat_precip_amt_mm
reanalysis_specific_humidity_g_per_kg
reanalysis_tdtr_k
station_avg_temp_c
station_diur_temp_rng_c
station_max_temp_c
station_min_temp_c
station_precip_mm
total_cases
0.8
0.4
0.0
−0.4
−0.8

Correlation of variables with target variable (total_cases)

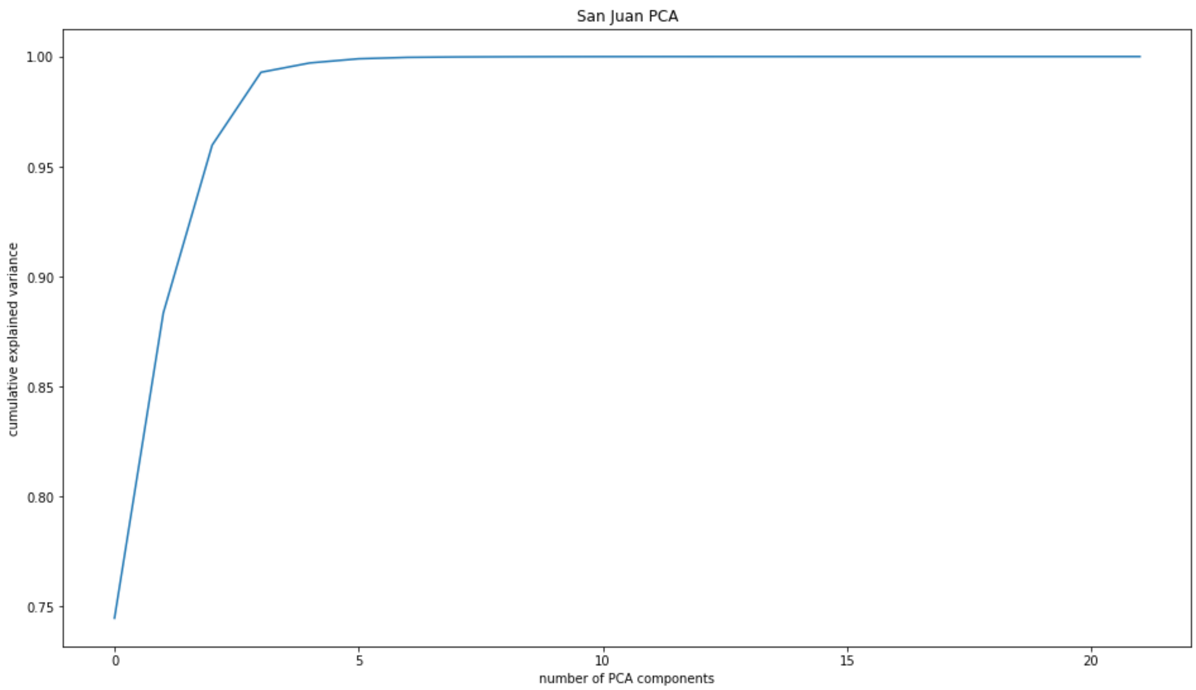


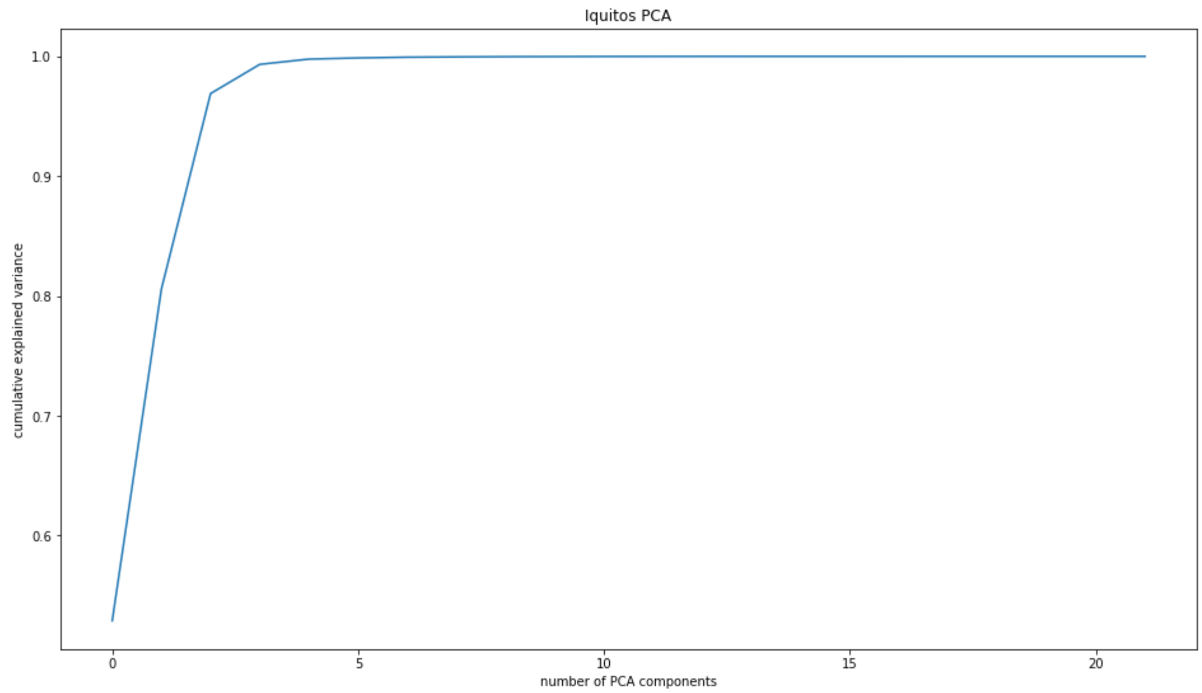


The two graphs above depict the correlation of the different variables with the target variable, the total number of cases of Dengue fever. For San Juan, the variable with the largest correlation is the week of the year. This variable will be one of the main focuses of the data analysis. The other variables with high correlation are the climate data records. As mosquitos thrive in warm and wet climates this makes sense. The week of the year is an interesting fact. The most likely reason this has the highest correlation is because this specific week has optimal conditions for mosquitos. For Iquitos, the climate data, humidity and year dominate the high correlation records. Again, climate is one of the major contributors to the number of Dengue fever cases. Our hypothesis is that due to the fact San Juan is located next to a large body of water, temperature fluctuations will stay predominantly low. So on a weekly basis, the number of cases can increase. As for Iquitos, since this city is located in land, the effects of climate will happen less frequently or on a yearly basis.

Principal Component Analysis

As the majority of the data is climate data, there will be various similarities with some of the features. Dropping the columns with the least variance and features with heavy commonalities is the method we are choosing to increase accuracy.





The two graphs above depict the Cumulative Explained Variance and PCA components relationship. Variance is at its peak just around 5 PCA components and plateaus just after that. As the variance does not increase after a certain point. This is likely due to the fact that most of the data is climate data. Implications of this representation are the number of components we can use is between 4 to ~9. This will be optimal and efficient.

Techniques

1. Random Forest

   We decided on this technique to learn more about decision trees. Sklearn provides an excellent library to learn about random foresting. It is also an useful tool to learn about feature analysis and comparison.

2. XGBoost

   This tool is relatively new and is known to be a powerful tool in Machine Learning. Execution speed and model performance are the key characteristics that make this tool worth exploring.

3. Neural Networks

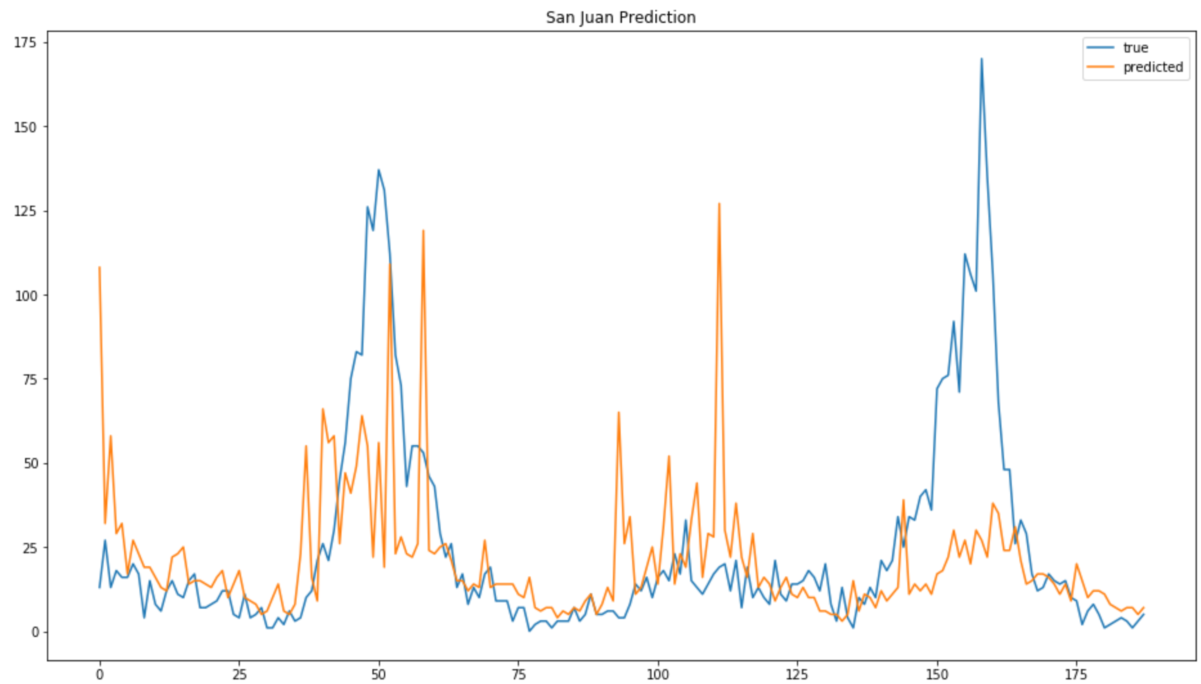
   A Neural Network is one of the first tools we learned to use. This basic tool uses classification and clustering tools. As this tool is a developed pattern recognition device, we decided practicing with this would provide a great learning experience.

5. Experimental Results and Analysis

a. XGBRegressor Results

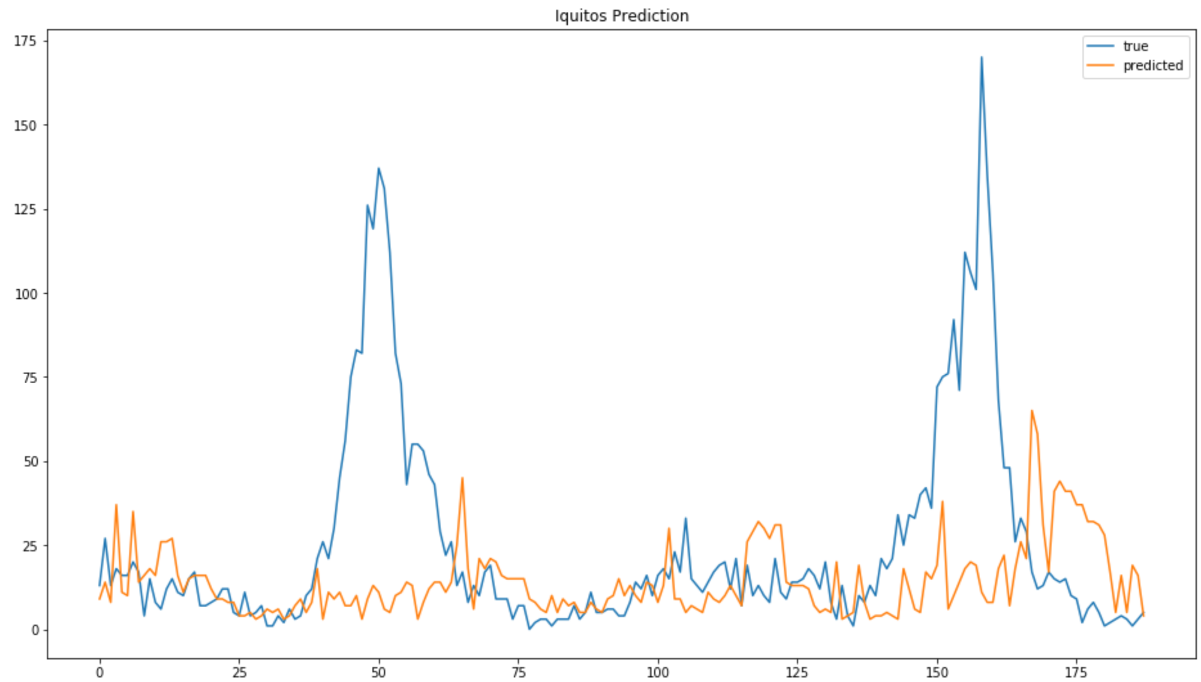
```
clf = XGBRegressor(max_depth=5, n_estimators=100)
clf.fit(X_train, y_train)
y_true, y_pred = y_test, clf.predict(X_test).astype(int)
print("Mean Absolute Error(MAE): %f" %MAE(y_true, y_pred))
```

```
Mean Absolute Error(MAE): 17.430851
```



```
clf = XGBRegressor(max_depth=3, n_estimators=100)
clf.fit(X_train, y_train)
y_true, y_pred = y_test, clf.predict(X_test).astype(int)
print("Mean Absolute Error(MAE): %f" %MAE(y_true, y_pred))
```

```
Mean Absolute Error(MAE): 20.718085
```



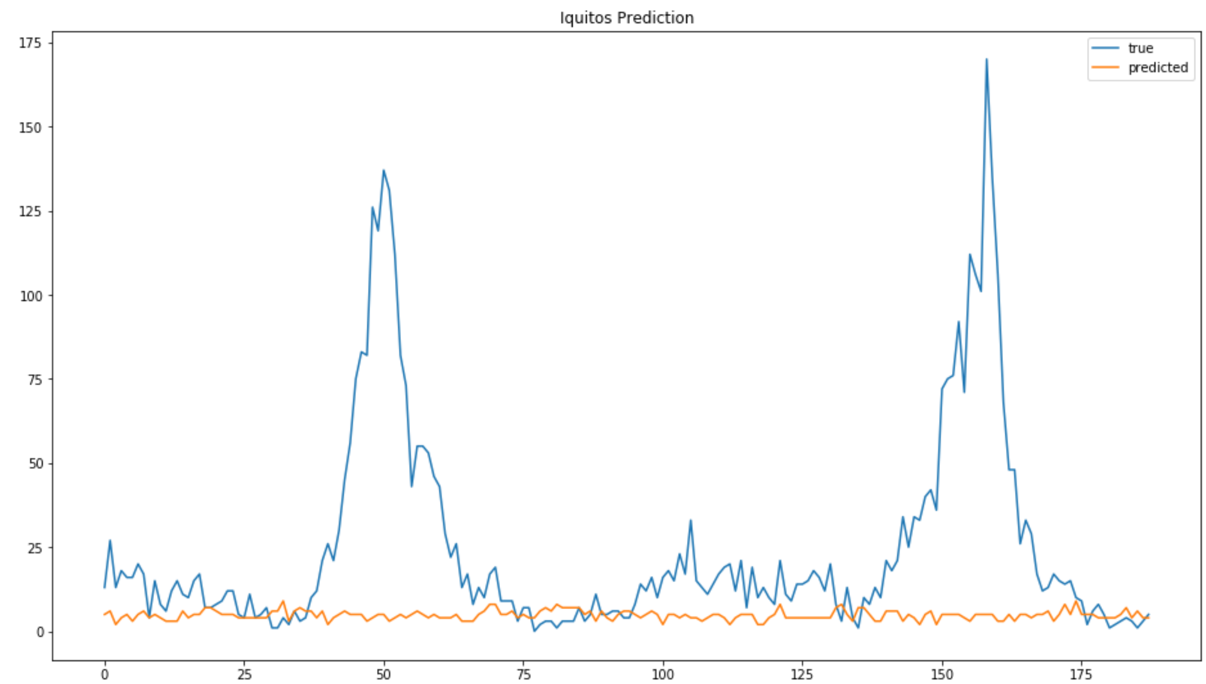
b. MLPRegressor Results

```
clf = MLPRegressor(max_iter=10000, hidden_layer_sizes=(100,))
clf.fit(X_train, y_train)
y_true, y_pred = y_test, clf.predict(X_test).astype(int)
print("Mean Absolute Error(MAE): %f" %MAE(y_true, y_pred))
```

```
Mean Absolute Error(MAE): 26.484043
```

```python
clf = MLPRegressor(max_iter=10000, hidden_layer_sizes=(13,13,13))
clf.fit(X_train, y_train)
y_true, y_pred = y_test, clf.predict(X_test).astype(int)
print("Mean Absolute Error(MAE): %f" %MAE(y_true, y_pred))
```

```
Mean Absolute Error(MAE): 20.925532
```



Submission Results

Technique	Results w/o PCA	Results w/ PCA	Results w/ Data Separation	Results w/ Data Separation and PCA
XGBoost	26.6274	26.8558	25.8173	27.6563
Random Forest	26.6779	26.4928		
Neural Networks			29.1611	32.1875

The results are Mean Absolute Error which is the required metric for the submission on the Driven Data competition.

6. Conclusion

Each of the models had very similar scores ranging from 25 to 32. The only major difference created was separating the cities' datasets during the preprocess. Our initial assumption was XGBoost would overwhelm the other two methods chosen. However, it only slightly out performed.

7. Contributions

Adharsh Rajendran: Random Foresting Code, Report
Mithil Gotarne: Data analysis, XGBoost code, Neural Network Code

8. References

- Donges, Niklas. "The Random Forest Algorithm – Towards Data Science." *Towards Data Science*, Towards Data Science, 22 Feb. 2018, towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd.
- "A Gentle Introduction to XGBoost for Applied Machine Learning." *Machine Learning Mastery*, 21 Sept. 2016, machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/.
- "A Beginner's Guide to Neural Networks and Deep Learning." *Skymind*, skymind.ai/wiki/neural-network.
- "Thanks, Global Warming--Mosquito-Borne Diseases Are on the Uptick." *Scientific American*, www.scientificamerican.com/article/mosquito-borne-diseases-on-the-uptick-thanks-to-global-warming/.
- "Predictive Modeling of Dengue Fever Epidemics: A Neural Network Approach." *ResearchGate*, www.researchgate.net/publication/324150583_Predictive_Modeling_of_Dengue_Fever_Epidemics_A_Neural_Network_Approach.
- "DengAI: Predicting Disease Spread - Benchmark - DrivenData Labs." *DrivenData*, drivendata.co/blog/dengue-benchmark/.