DengAI - Predicting Disease Spread

CS4622 - Machine Learning

Group 16

Viraj Gamage	(140173T)
Gathika Ratnayaka	(140528M)
Thejan Rupasinghe	(140536K)
Menuka Warushavithana	(140650E)

Introduction

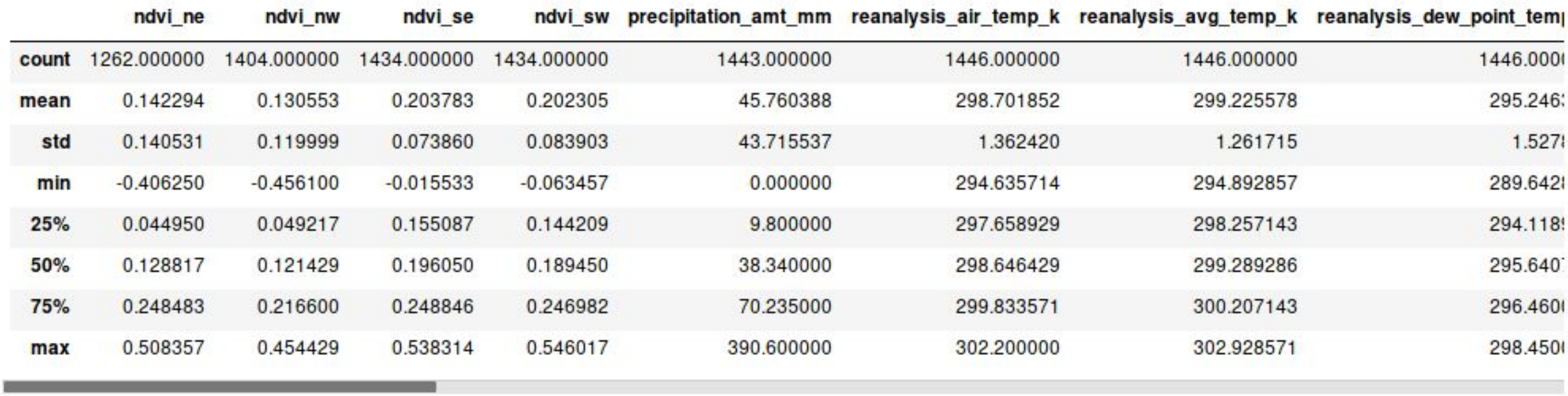
- Competition hosted by DrivenData [drivendata]
- Predicting dengue cases is helpful for health officials
    - To know disease outbreak times beforehand

Methodology

Preprocessing

- Data visualization
- Data cleaning
    - Handling missing values
    - Data integration
- Data normalization

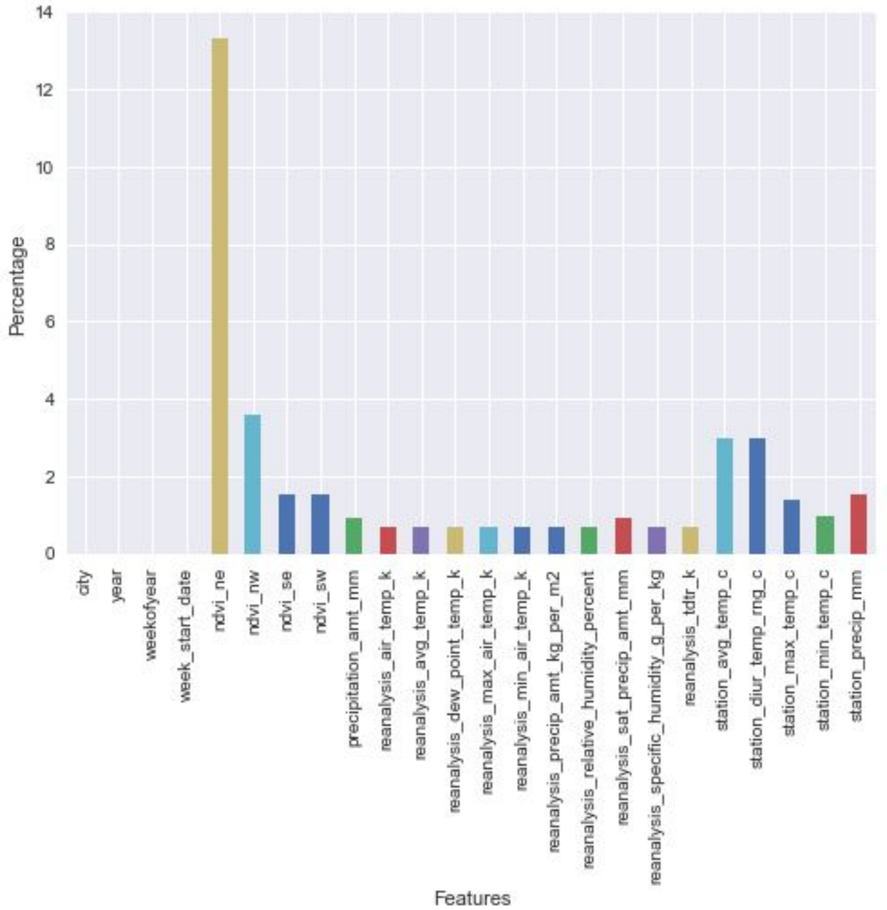
Data Visualization



	ndvi_ne	ndvi_nw	ndvi_se	ndvi_sw	precipitation_amt_mm	reanalysis_air_temp_k	reanalysis_avg_temp_k	reanalysis_dew_point_tem
count	1262.000000	1404.000000	1434.000000	1434.000000	1443.000000	1446.000000	1446.000000	1446.000
mean	0.142294	0.130553	0.203783	0.202305	45.760388	298.701852	299.225578	295.246
std	0.140531	0.119999	0.073860	0.083903	43.715537	1.362420	1.261715	1.527
min	-0.406250	-0.456100	-0.015533	-0.063457	0.000000	294.635714	294.892857	289.642
25%	0.044950	0.049217	0.155087	0.144209	9.800000	297.658929	298.257143	294.118
50%	0.128817	0.121429	0.196050	0.189450	38.340000	298.646429	299.289286	295.640
75%	0.248483	0.216600	0.248846	0.246982	70.235000	299.833571	300.207143	296.460
max	0.508357	0.454429	0.538314	0.546017	390.600000	302.200000	302.928571	298.450

Description of Features

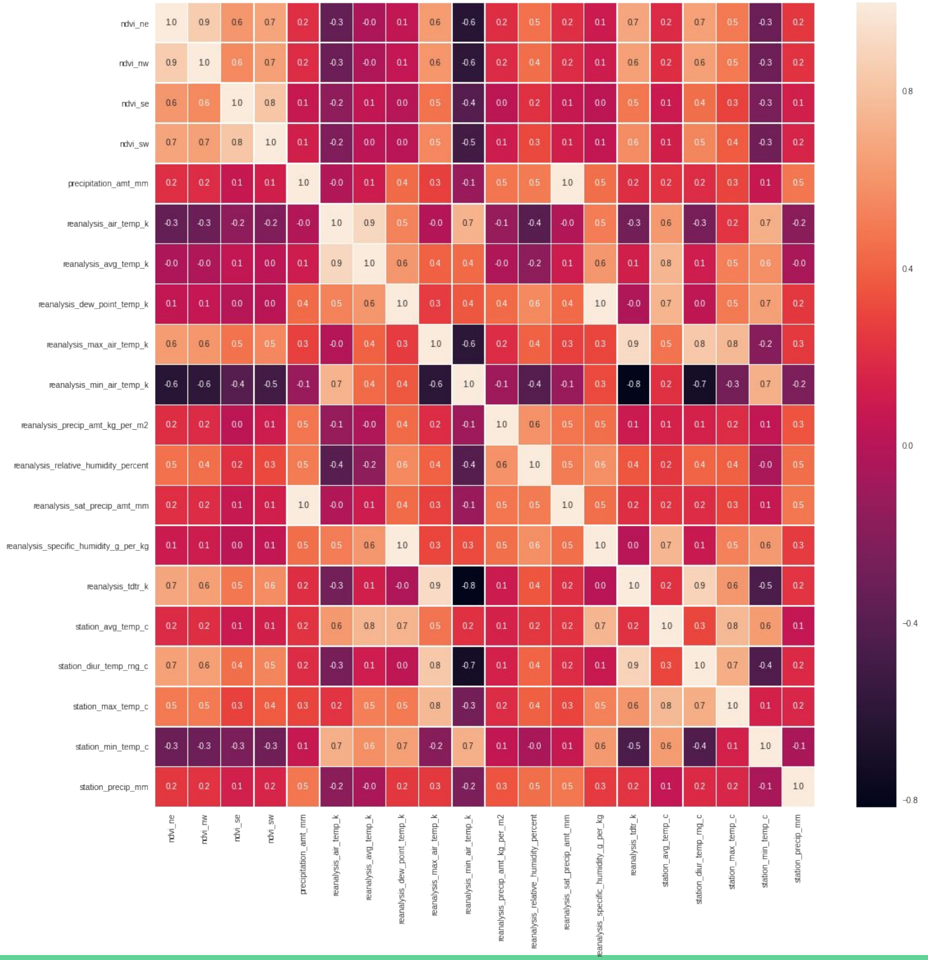
Null Values a Percentage for each Feature

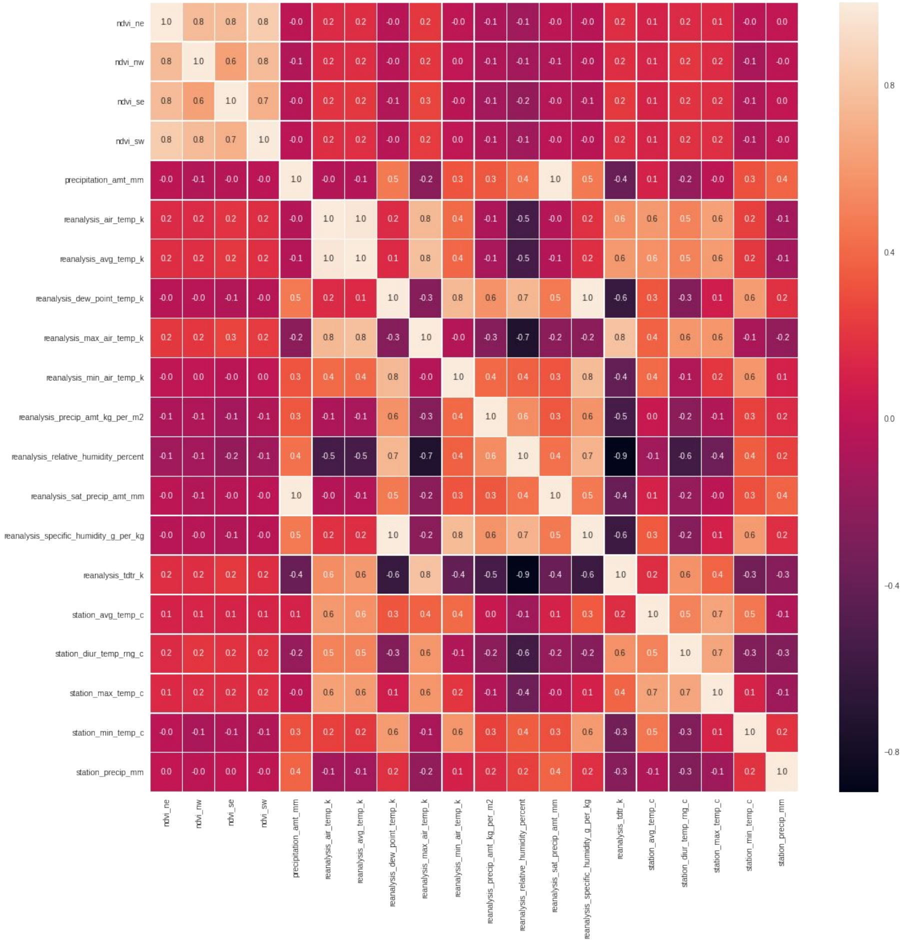


Data Modeling

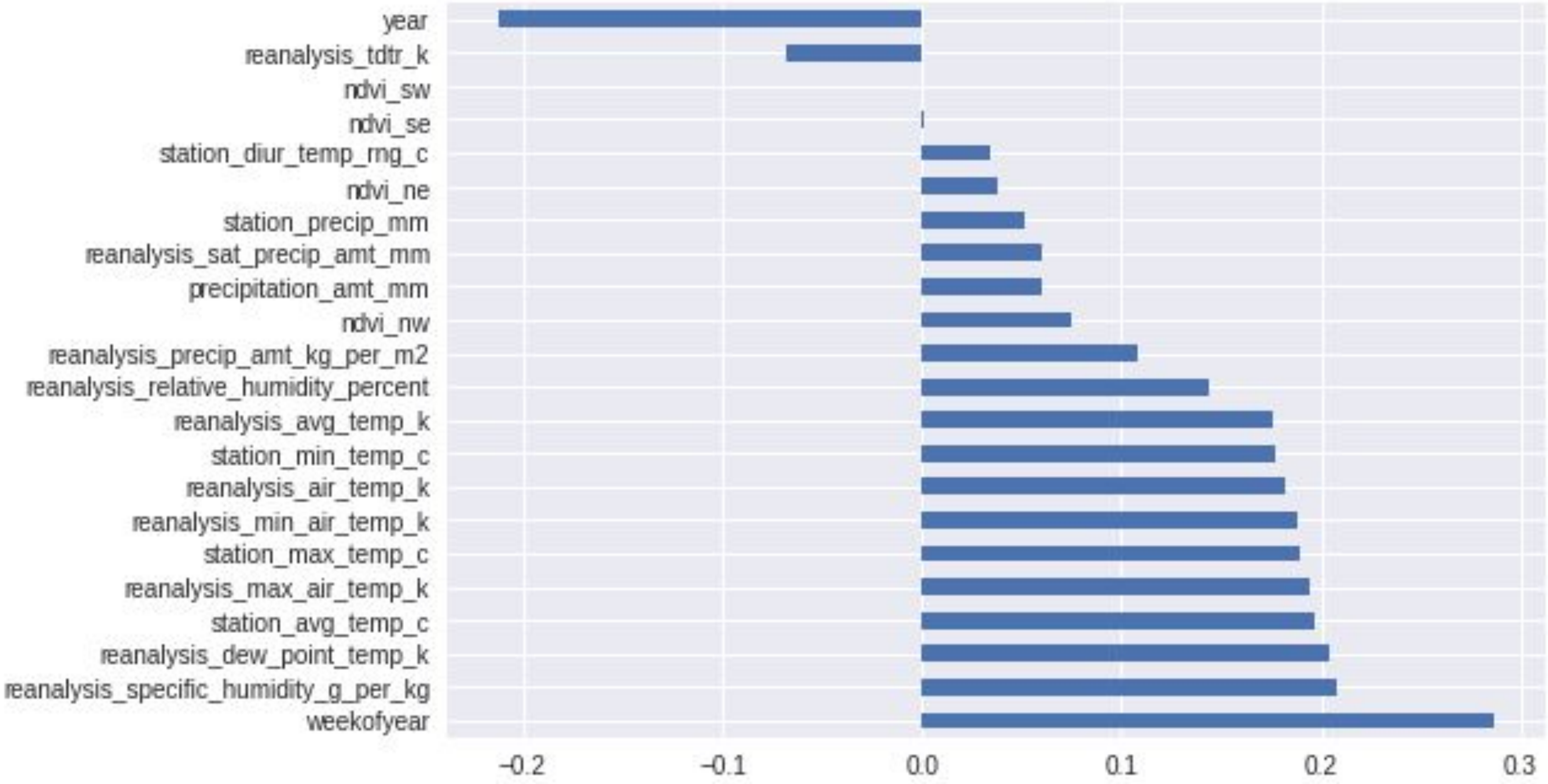
- Correlation between features
- Feature selection
- Engineering new features

Feature-Feature Correlation for San Juan (left) and Iquitos (right)

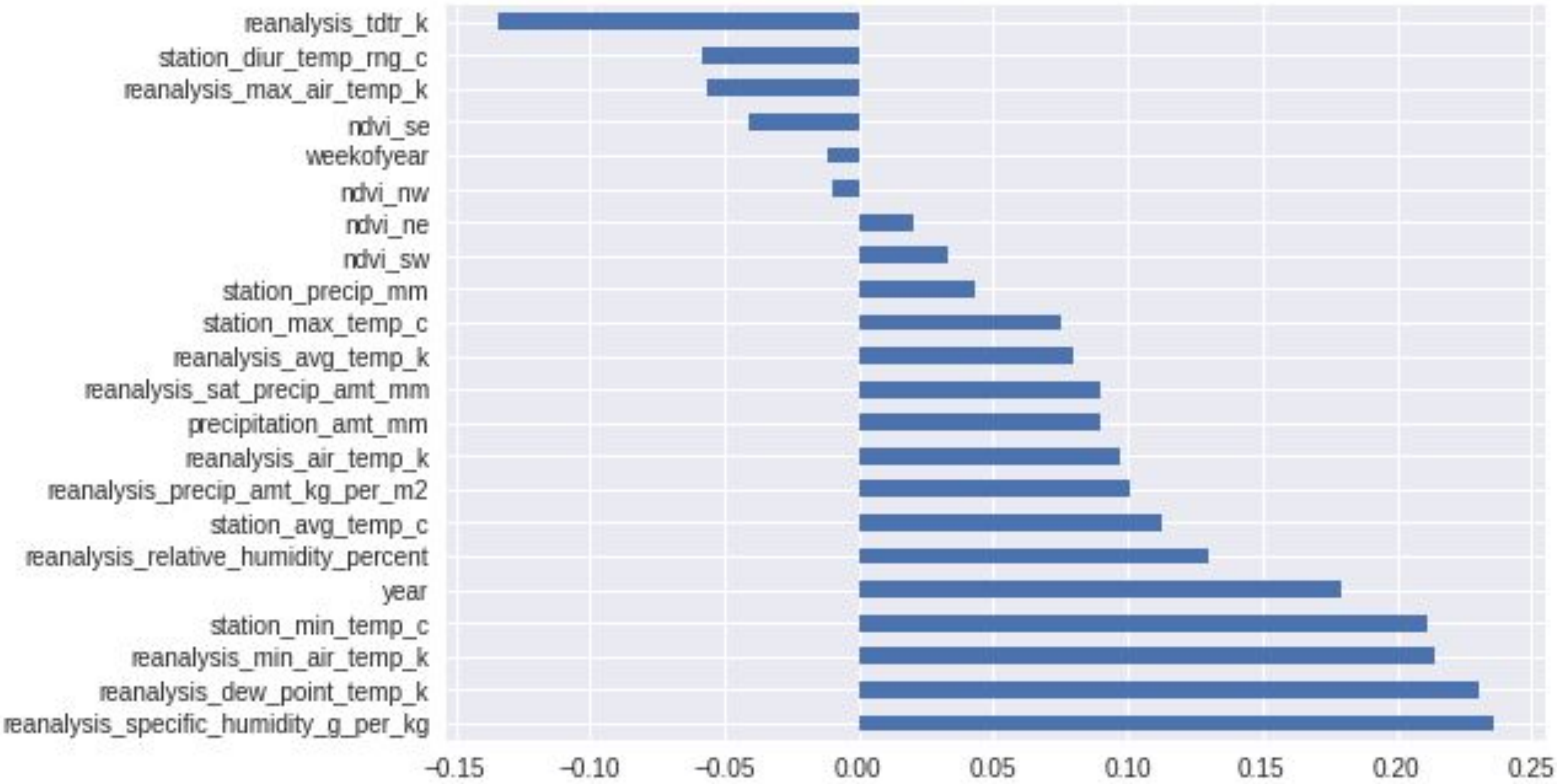




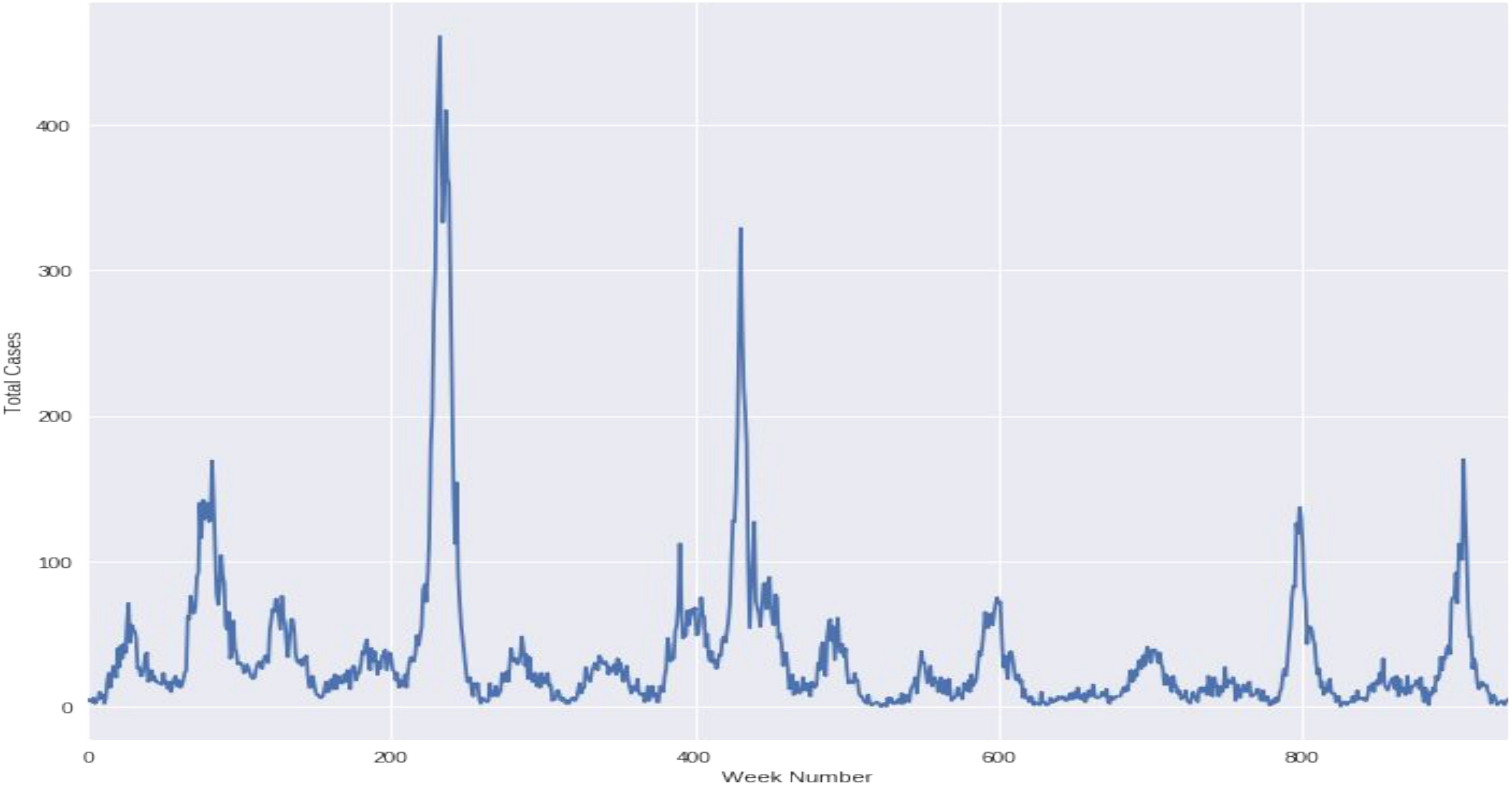
Correlation between *total_cases* and other features for San Juan



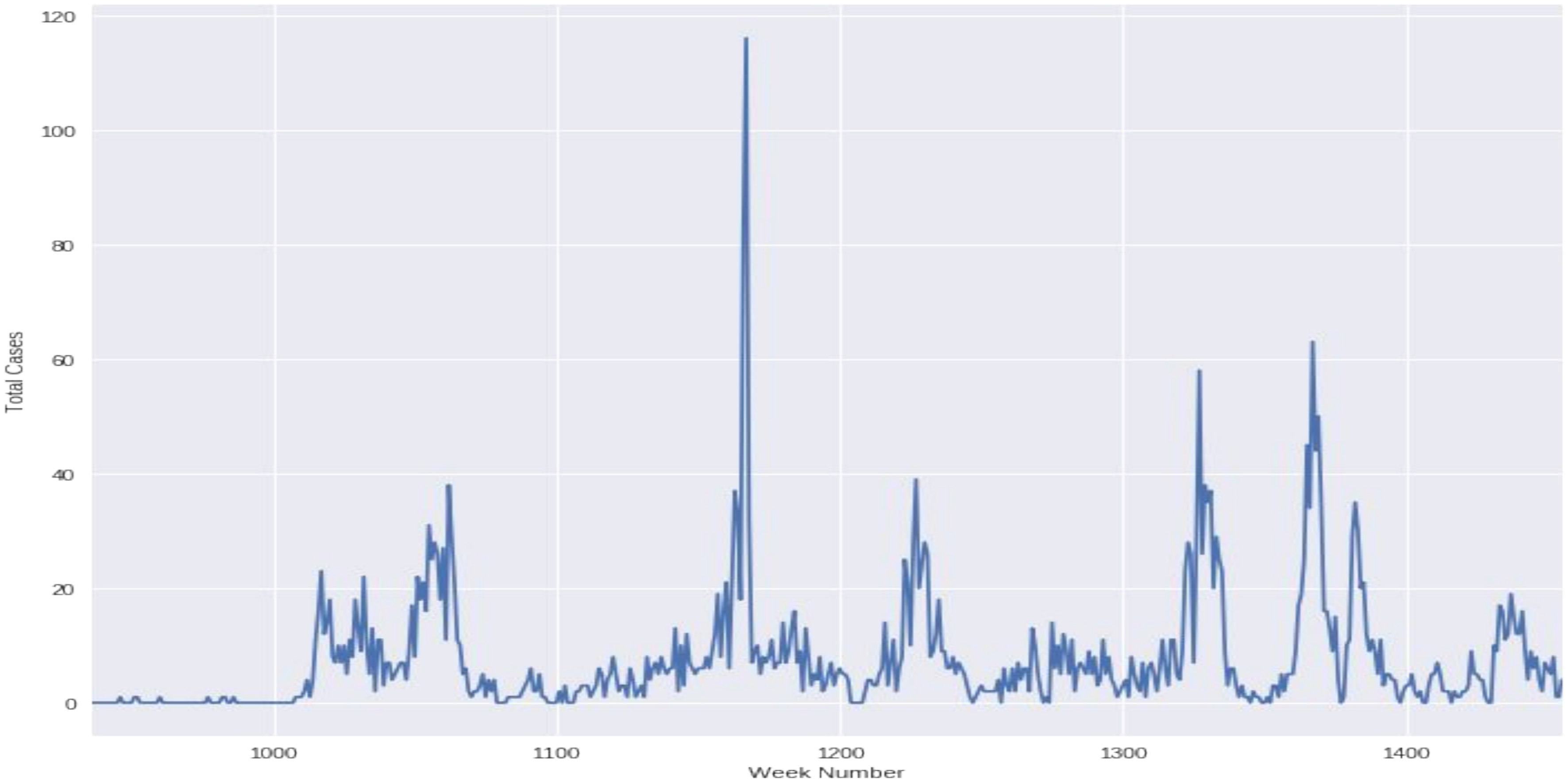
Correlation between *total_cases* and other features for Iquitos



Total cases vs. Week of the year for **San Juan**



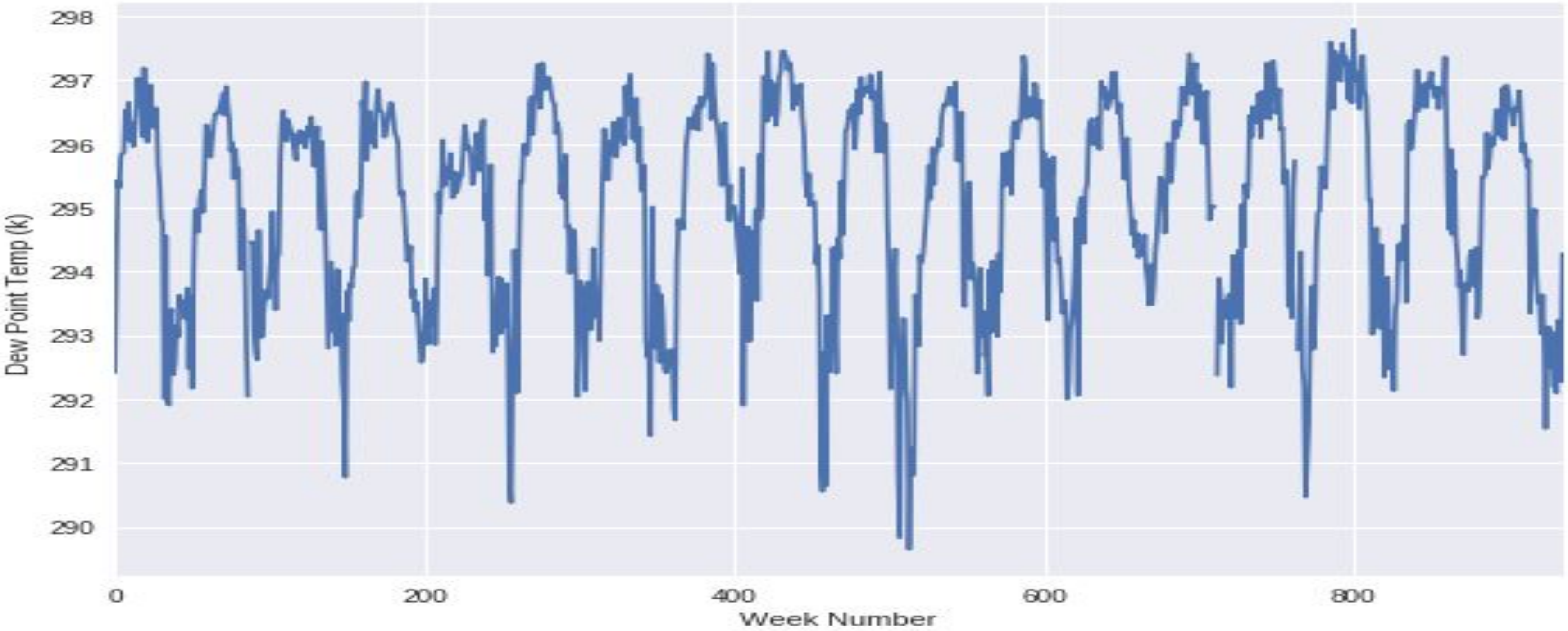
Total cases vs. Week of the year for **Iquitos**



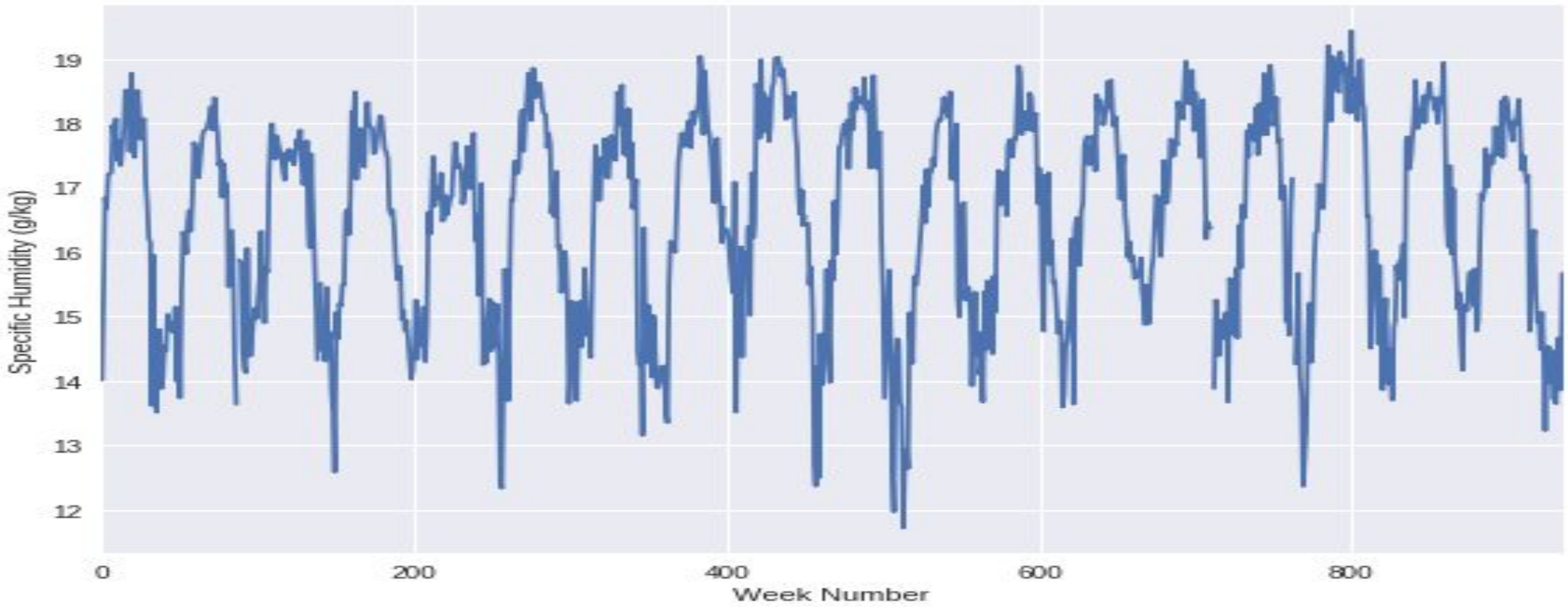
Feature Selection

- For San Juan , ‘week of the year’ has the highest correlation with the total number of cases.
- It must be due to climatic changes that happens in relation to the period of the year (week number)
- We plotted graphs of climatic factors against the week number.
- A pattern existing between these climatic factors and total number of cases in relation to week number could be identfied.

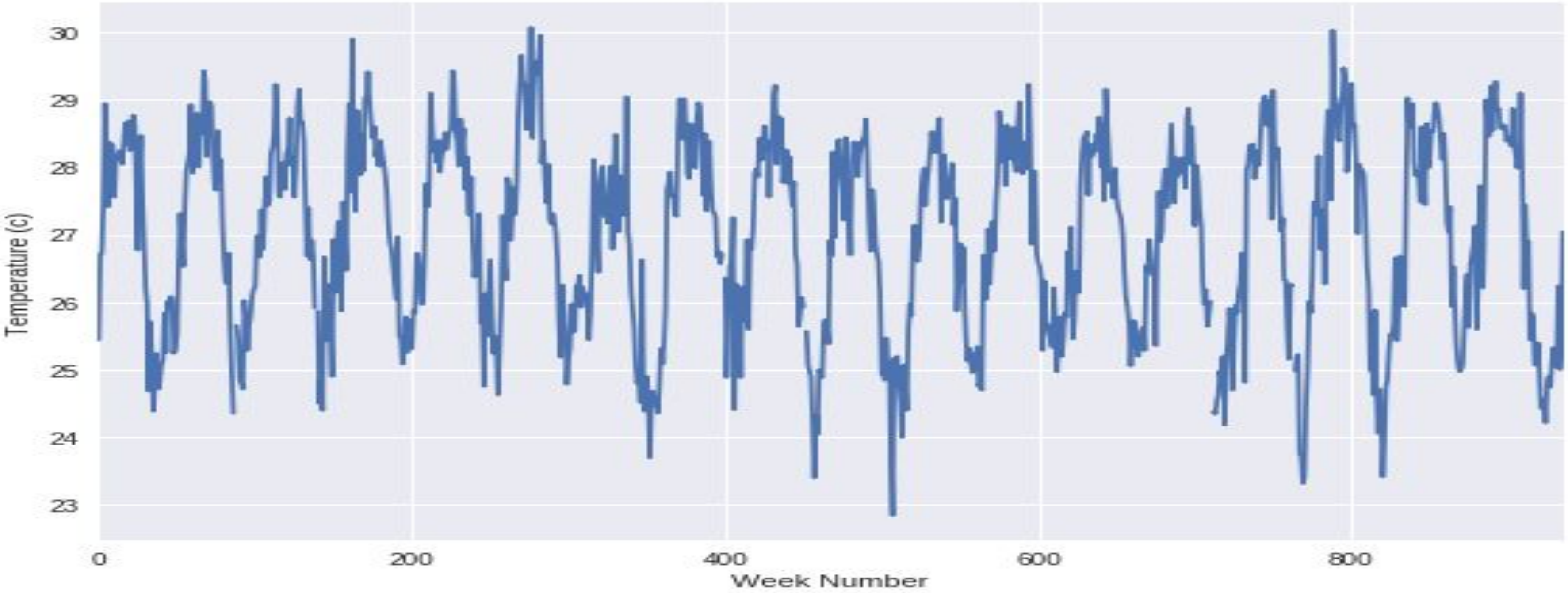
Dew Point Temp vs Week Number



Specific Humidity vs Week Number



Temperature vs Week Number



Engineering New Features

- reanalysis_dew_point_temp_k
- reanalysis_specific_humidity_g_per_kg
- reanalysis_precip_amt_kg_per_m2

Results and Analysis

Comparing Different Models

Model	MAE for San Juan	MAE for Iquitos
Linear Regression	26.987	6.539
Support Vector Regression (kernel=’linear’)	22.792	5.686
Support Vector Regression (kernel=’rbf’)	21.810	5.617
Gradient Boosting	19.491	5.726
K-Nearest Neighbour Regression	26.482	6.521
Random Forest regression	19.800	6.385

Comparison of Models Contd...

- Gradient Boosting model and Random Forests models show the best results for the data set of San Juan
- Gradient Boosting and Support Vector Regression Model with linear kernel and rbf (radial basis function) kernel outperforms all other models for the data set of Iquitos
- KNN regression model’s score and linear regression model’s score is around that of baseline model’s
- Further processing with the selected models

Tuning Hyper-parameters with Grid Search

- Using sklearn.model_selection.GridSearchCV

Model	**MAE for San Juan**
Gradient Boosting	16.101
Random Forests	16.728

Model	**MAE for Iquitos**
Gradient Boosting	5.623
SVR (kernel=’linear’)	4.872
SVR (kernel=rbf’)	5.252

Conclusion

Conclusion

- Importance of Preprocessing
    - Filling missing values
- Feature Engineering
    - Using plotted graphs for features
    - Using the trend of a feature
- Model Selection
    - Decision Trees are the best model at all

Thank You!