Predicting Disease Spread

Xinlong Li

Outline

- Understand the features
- Missing values
- Outliers
- Correlation analysis
- Model
  - Negative binomial regression
  - Random forest
- Regression result

Load the Data

```python
X = pd.read_csv('dengue_features_train.csv')
y = pd.read_csv('dengue_labels_train.csv')
```

```python
X.head()
```

	city	year	weekofyear	week_start_date	ndvi_ne	ndvi_
0	sj	1990	18	1990-04-30	0.122600	0.1037
1	sj	1990	19	1990-05-07	0.169900	0.1421
2	sj	1990	20	1990-05-14	0.032250	0.172
3	sj	1990	21	1990-05-21	0.128633	0.245
4	sj	1990	22	1990-05-28	0.196200	0.262

Description about the features

City and date indicators

city – City abbreviations: sj for San Juan and iq for Iquitos

week_start_date – Date given in yyyy-mm-dd format

NOAA's GHCN daily climate data weather station measurements

station_max_temp_c – Maximum temperature

station_min_temp_c – Minimum temperature

station_avg_temp_c – Average temperature

station_precip_mm – Total precipitation

station_diur_temp_rng_c – Diurnal temperature range

PERSIANN satellite precipitation measurements (0.25x0.25 degree scale)

precipitation_amt_mm – Total precipitation

Description about the features

NOAA's NCEP Climate Forecast System Reanalysis measurements (0.5x0.5 degree scale)

- reanalysis_sat_precip_amt_mm – Total precipitation
- reanalysis_dew_point_temp_k – Mean dew point temperature
- reanalysis_air_temp_k – Mean air temperature
- reanalysis_relative_humidity_percent – Mean relative humidity
- reanalysis_specific_humidity_g_per_kg – Mean specific humidity
- reanalysis_precip_amt_kg_per_m2 – Total precipitation
- reanalysis_max_air_temp_k – Maximum air temperature
- reanalysis_min_air_temp_k – Minimum air temperature
- reanalysis_avg_temp_k – Average air temperature
- reanalysis_tdtr_k – Diurnal temperature range

Satellite vegetation - Normalized difference vegetation index (NDVI) - NOAA's CDR Normalized Difference Vegetation Index (0.5x0.5 degree scale) measurements

- ndvi_se – Pixel southeast of city centroid
- ndvi_sw – Pixel southwest of city centroid
- ndvi_ne – Pixel northeast of city centroid
- ndvi_nw – Pixel northwest of city centroid

Training San Juan and Iquitos Separately

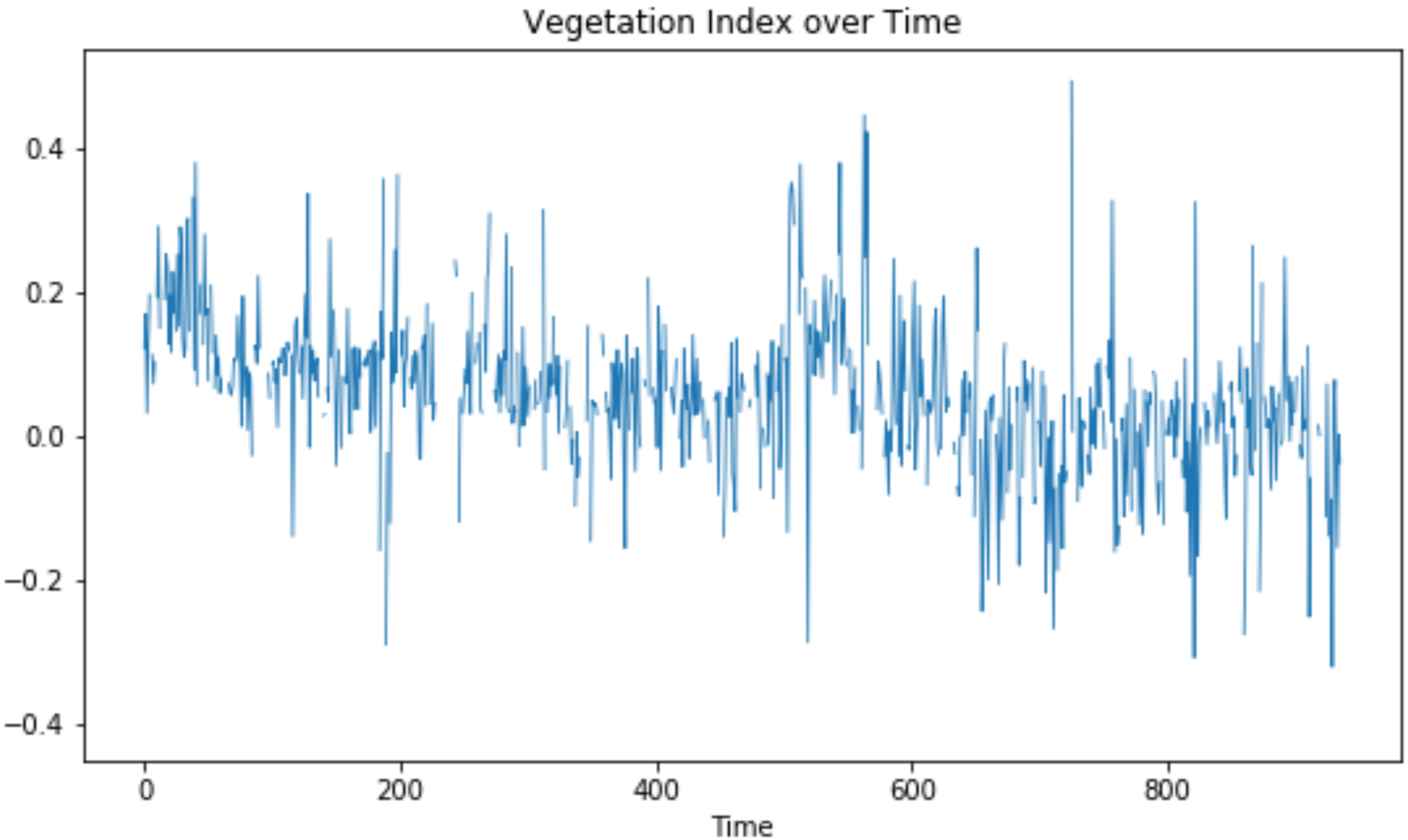
```
X_sj = X.loc[X['city'] =='sj'].copy()
y_sj = y.loc[y['city'] =='sj'].copy()

X_iq = X.loc[X['city'] =='iq'].copy()
y_iq = y.loc[y['city'] =='iq'].copy()
```

Missing Value

```
1  X_sj.isna().sum()
```

```
city                                     0
year                                     0
weekofyear                               0
week_start_date                          0
ndvi_ne                                191
ndvi_nw                                 49
ndvi_se                                 19
ndvi_sw                                 19
precipitation_amt_mm                     9
reanalysis_air_temp_k                    6
reanalysis_avg_temp_k                    6
reanalysis_dew_point_temp_k              6
reanalysis_max_air_temp_k                6
reanalysis_min_air_temp_k                6
reanalysis_precip_amt_kg_per_m2          6
reanalysis_relative_humidity_percent     6
reanalysis_sat_precip_amt_mm             9
reanalysis_specific_humidity_g_per_kg    6
reanalysis_tdtr_k                        6
station_avg_temp_c                       6
station_diur_temp_rng_c                  6
station_max_temp_c                       6
station_min_temp_c                       6
station_precip_mm                        6
dtype: int64
```

Missing Value




```
X_sj = X_sj.interpolate()
X_iq = X_iq.interpolate()
```

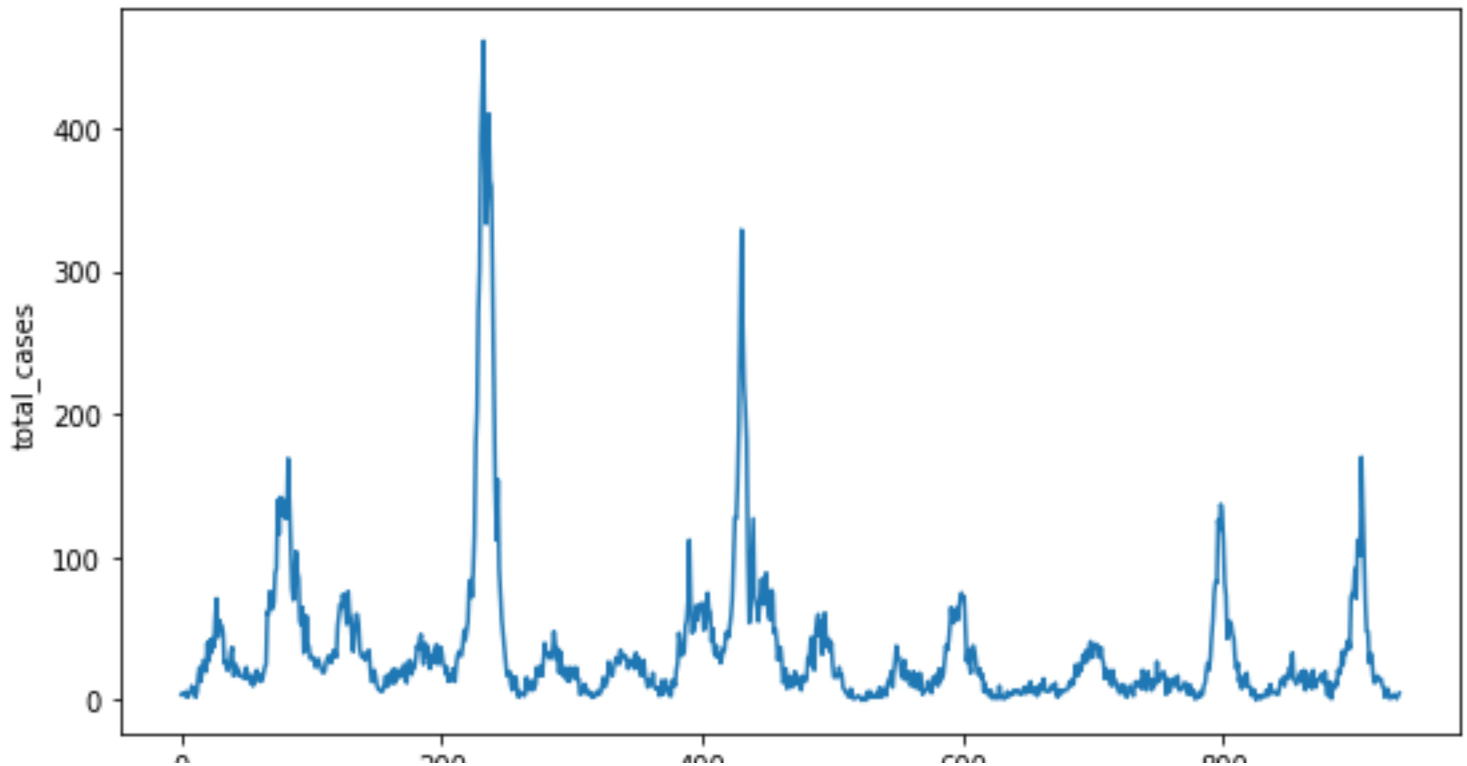
Missing Value

```
1  X_iq.isna().sum()
```

```
city                                      0
year                                      0
weekofyear                                0
week_start_date                           0
ndvi_ne                                   3
ndvi_nw                                   3
ndvi_se                                   3
ndvi_sw                                   3
precipitation_amt_mm                      4
reanalysis_air_temp_k                     4
reanalysis_avg_temp_k                     4
reanalysis_dew_point_temp_k               4
reanalysis_max_air_temp_k                 4
reanalysis_min_air_temp_k                 4
reanalysis_precip_amt_kg_per_m2           4
reanalysis_relative_humidity_percent      4
reanalysis_sat_precip_amt_mm              4
reanalysis_specific_humidity_g_per_kg     4
reanalysis_tdtr_k                         4
station_avg_temp_c                       37
station_diur_temp_rng_c                  37
station_max_temp_c                       14
station_min_temp_c                        8
station_precip_mm                        16
dtype: int64
```

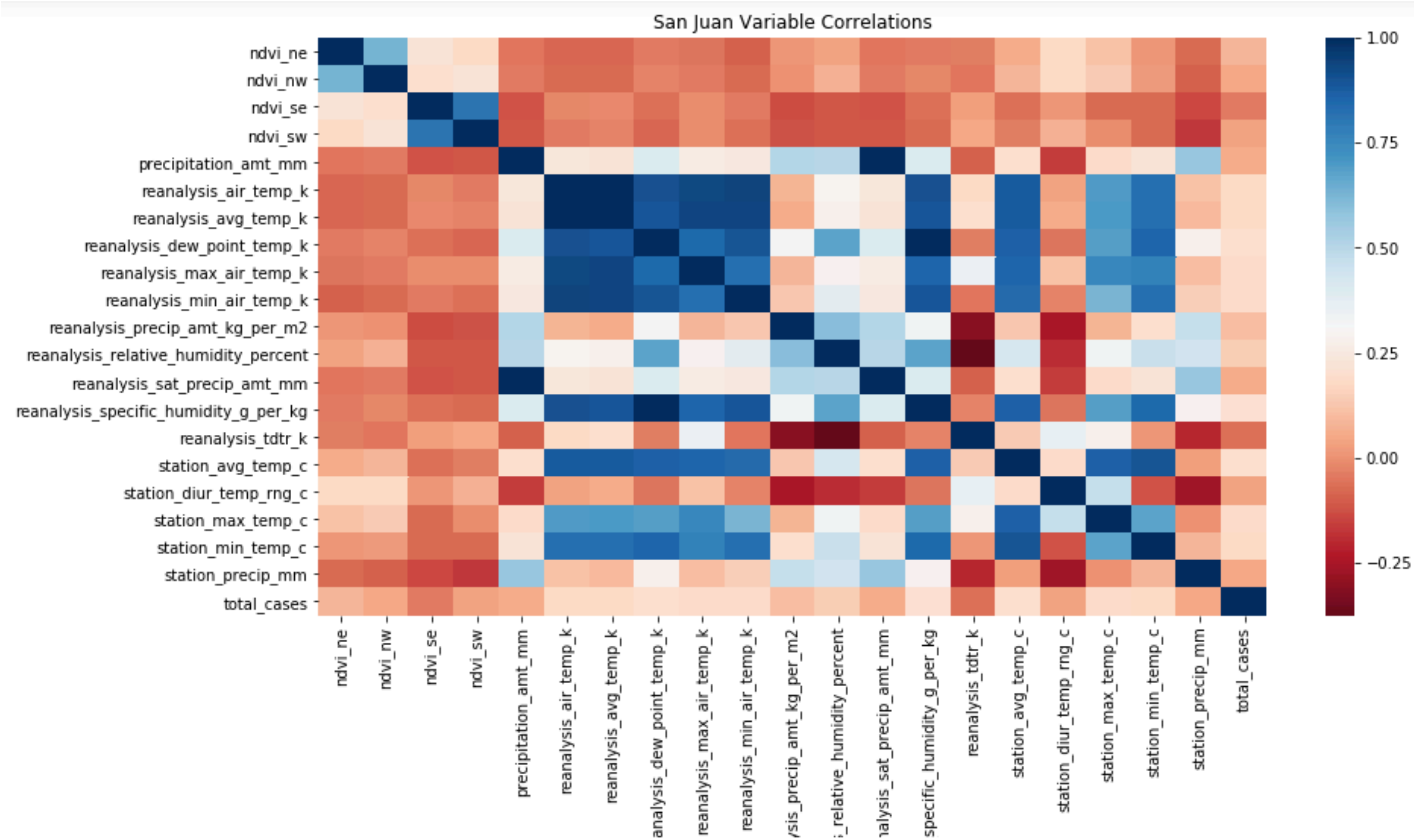
Outliers

Aside from missing values, outliers were also detected, using 3σ as inner outlier limit and 5σ as extreme limit, where σ was the observed standard deviation of the feature. Analysis of these outliers revealed that they were plausible values, and as such, they were not treated for this study



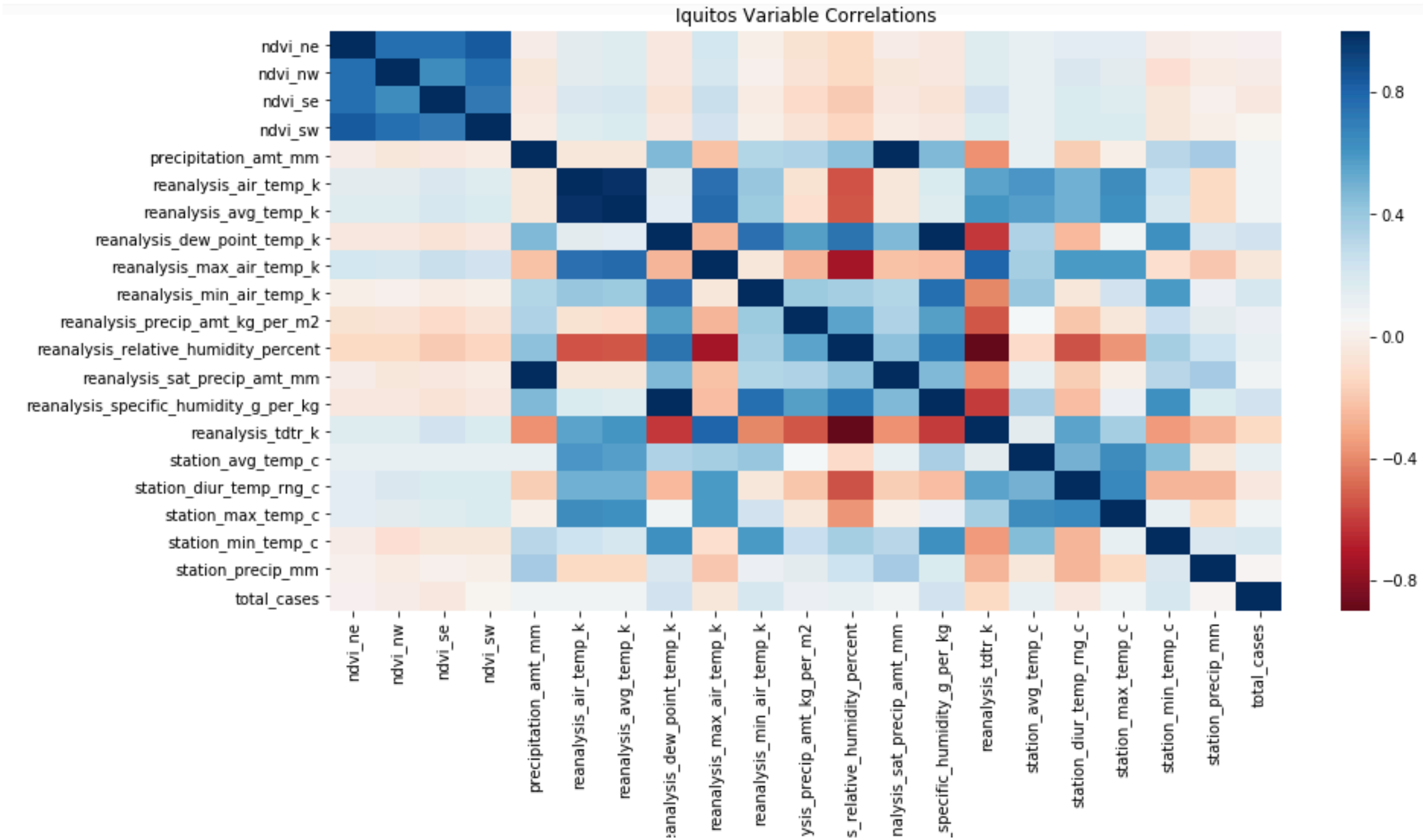
Correlation Analysis

```
sns.heatmap(X_sj.corr(), cmap='RdBu')
plt.title('San Juan Variable Correlations')
```



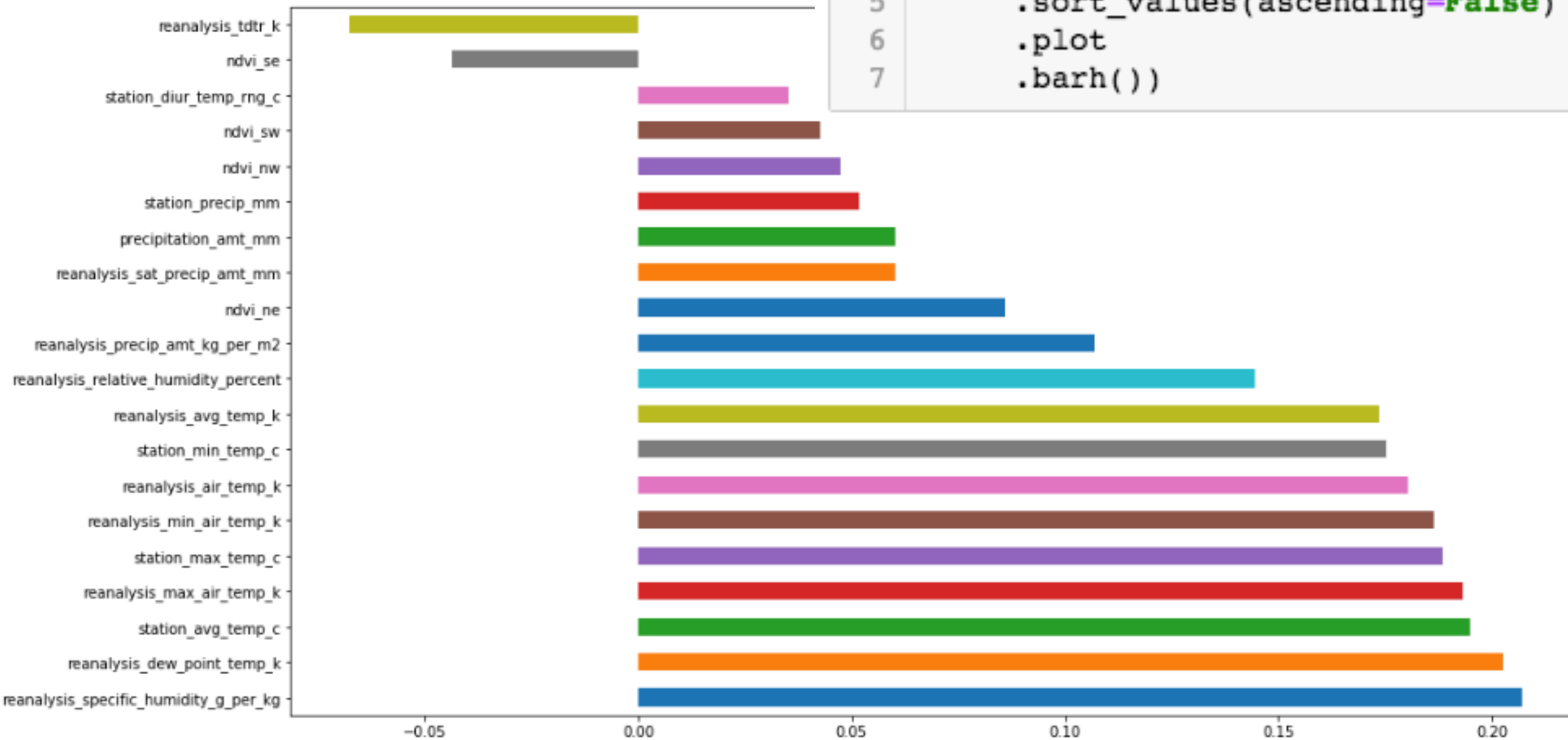
Correlation Analysis

```
3  sj_corr_heat = sns.heatmap(iq_correlations,cmap='RdBu')
4  plt.title('Iquitos Variable Correlations')
```



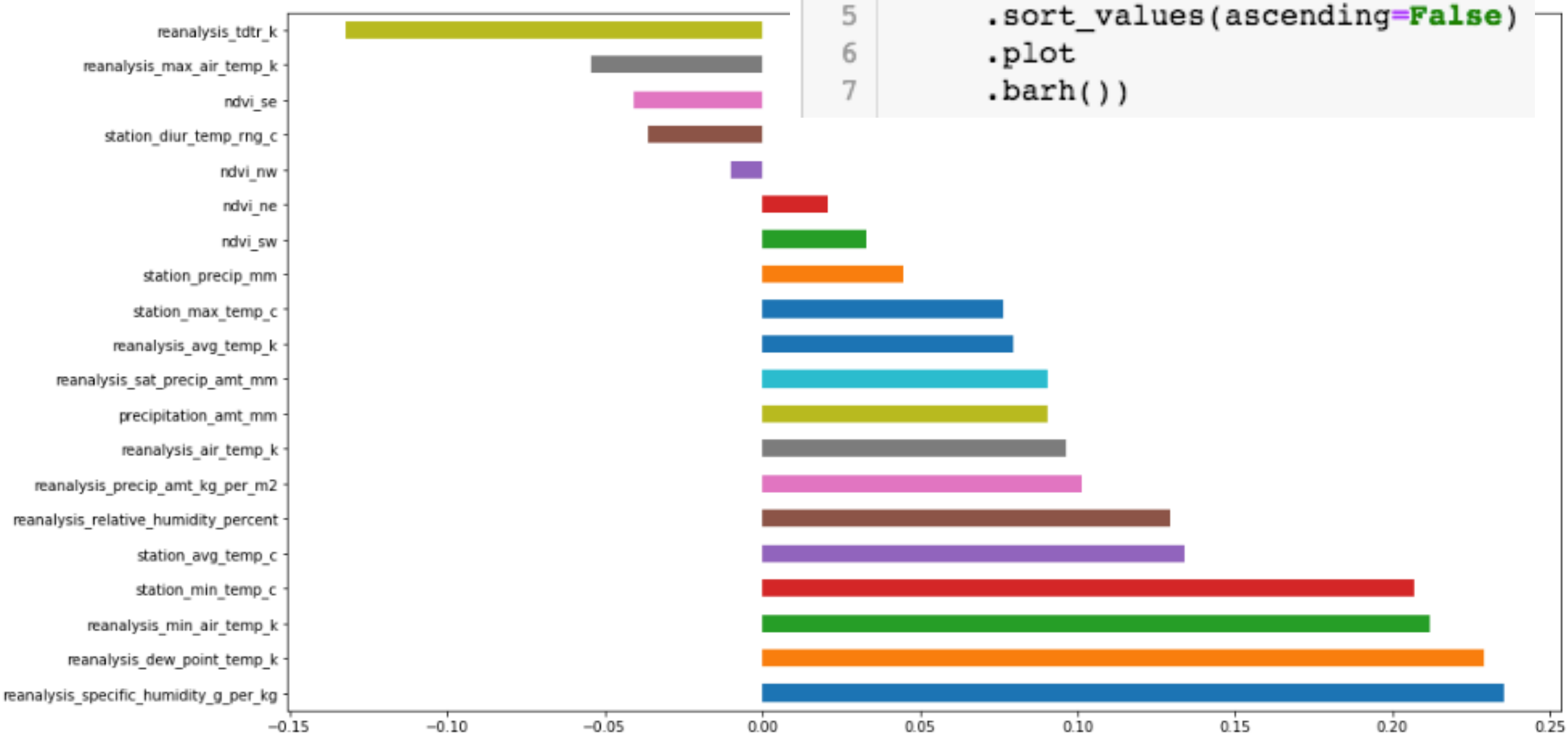
Correlation Analysis

```
(sj_correlations
     .total_cases
     .drop('total_cases')
     .sort_values(ascending=False)
     .plot
     .barh())
```



Correlation Analysis

```
(iq_correlations
     .total_cases
     .drop('total_cases')
     .sort_values(ascending=False)
     .plot
     .barh())
```



Correlation Analysis

-The wetter the better

-The correlation strengths differ for each city, but it looks like reanalysis_specific_humidity_g_per_kg and reanalysis_dew_point_temp_k are the most strongly correlated with total_cases. This makes sense: we know mosquitos thrive wet climates, the wetter the better!

- Hot and heavy

-As is known, "cold and humid" is not a thing. So it's not surprising that as minimum temperatures, maximum temperatures, and average temperatures rise, the total_cases of dengue fever tend to rise as well.

- Rain

-Interestingly, the precipitation measurements bear little to no correlation to total_cases, despite strong correlations to the humidity measurements

Model – Negative Binomial Regression

```
print('San Juan')
print('mean: ', y_sj.mean()[0])
print('var :', y_sj.var()[0])

print('\nIquitos')
print('mean: ', y_iq.mean()[0])
print('var :', y_iq.var()[0])
```

```
San Juan
mean:  34.1805555555556
var : 2640.045439691045

Iquitos
mean:  7.565384615384615
var : 115.8955239365642
```

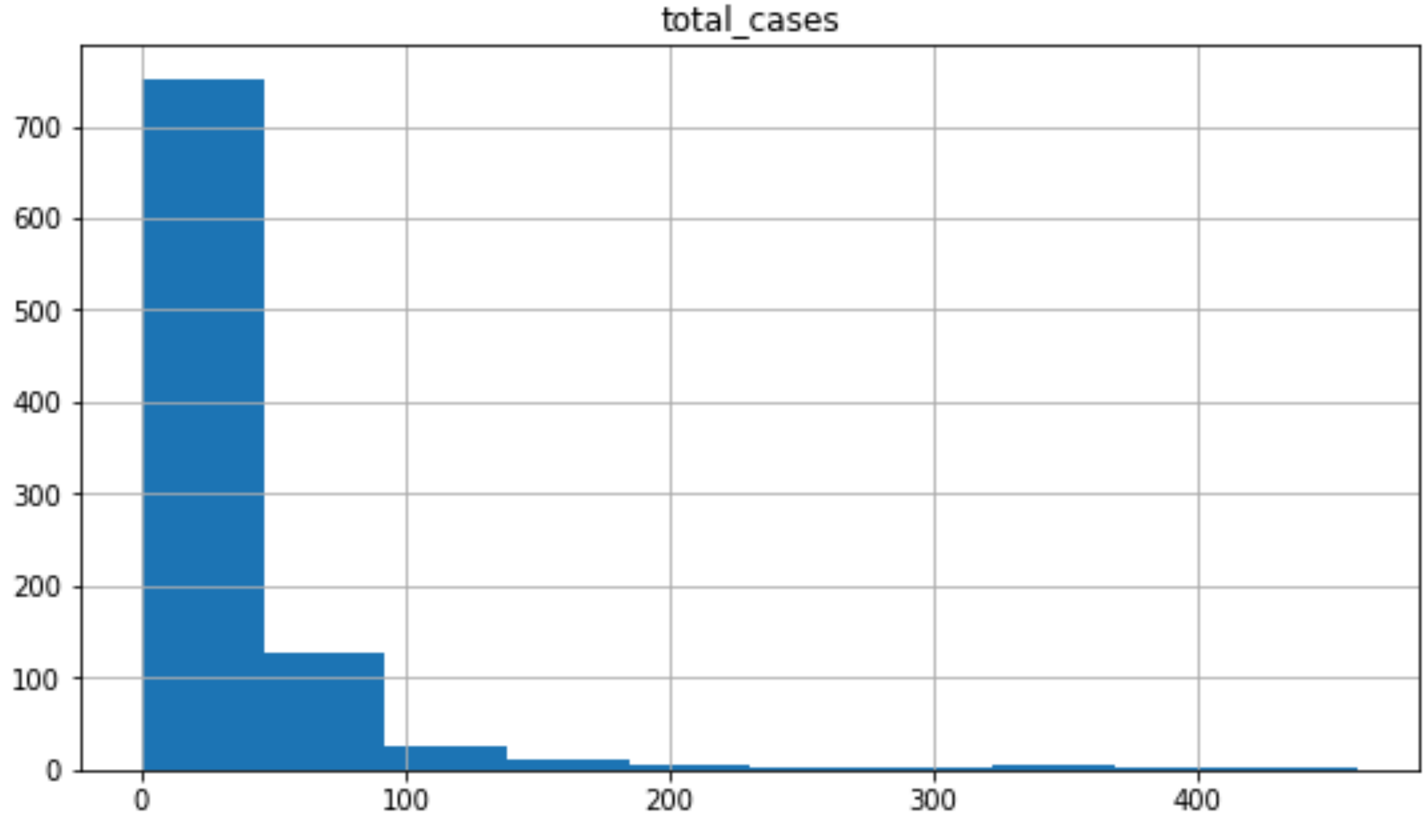
variance >> mean

- suggests total_cases can be described by a negative binomial distribution, so we'll use a negative binomial regression.

Model – Negative Binomial Regression

```
2 y_sj.hist()
```

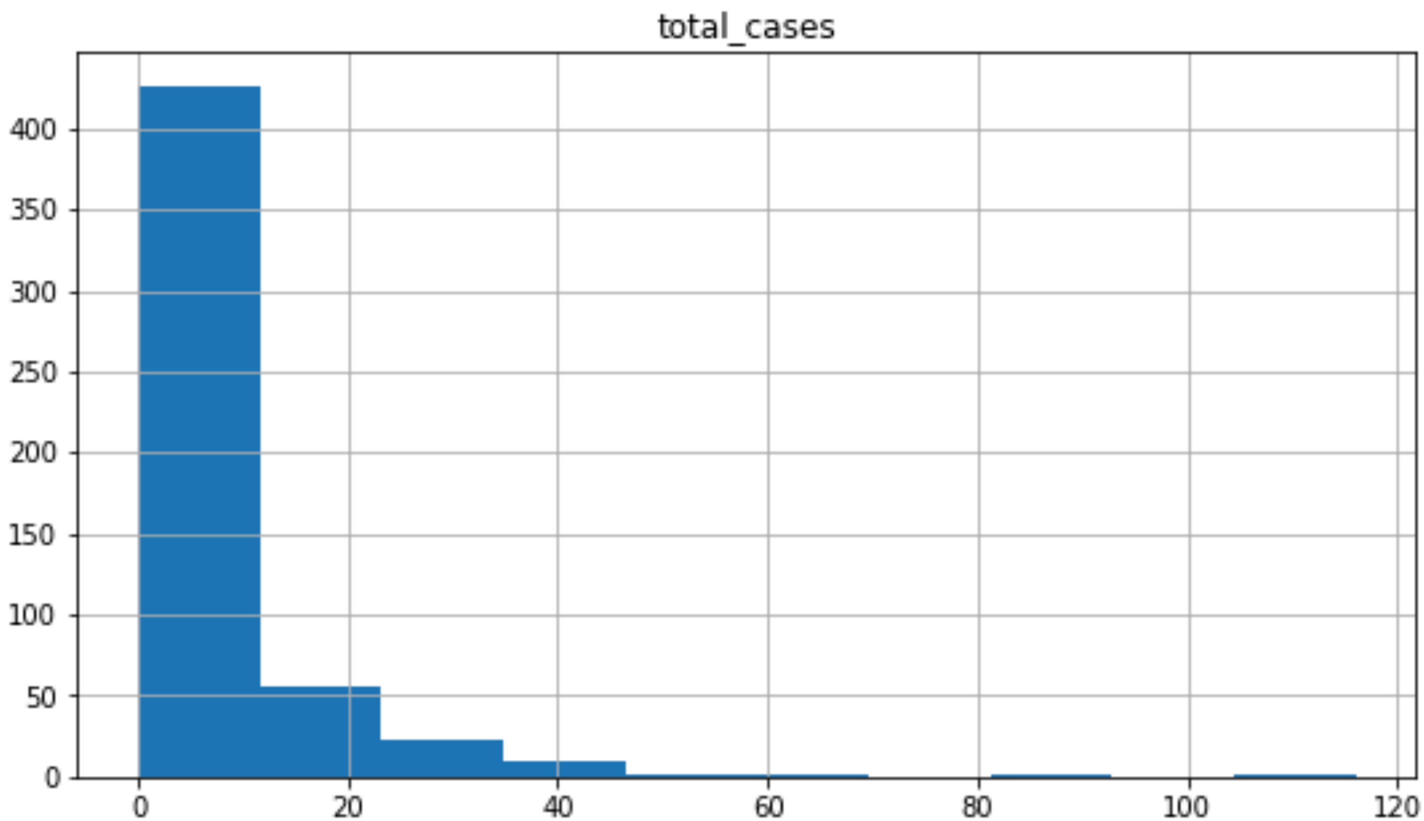
```
array([[<matplotlib.axes._subplots.AxesSubplot object at 0x10cd3243
      dtype=object)
```



Model – Negative Binomial Regression

```
2  y_iq.hist()
```

```
array([[<matplotlib.axes._subplots.AxesSubplot object at 0x10d9f5
       dtype=object)
```



Model – Negative Binomial Regression

```
sj_train = X_sj.head(800)
sj_test = X_sj.tail(X_sj.shape[0] - 800)

iq_train = X_iq.head(400)
iq_test = X_iq.tail(X_iq.shape[0] - 400)
```

Model – Negative Binomial Regression

```
def get_best_model(train, test):
    # Step 1: specify the form of the model
    model_formula = 'total_cases ~ 1 + ' \
                    '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'

    grid = 10 ** np.arange(-8, -3, dtype=np.float64)

    best_alpha = []
    best_score = 1000
```

Model – Negative Binomial Regression

```python
# Step 2: Find the best hyper parameter, alpha
for alpha in grid:
    model = smf.glm(formula=model_formula,
                    data=train,
                    family=sm.families.NegativeBinomial(alpha=alpha))

    results = model.fit()
    predictions = results.predict(test).astype(int)
    score = eval_measures.meanabs(predictions, test.total_cases)

    if score < best_score:
        best_alpha = alpha
        best_score = score
print('best alpha = ', best_alpha)
print('best score = ', best_score)
```

Model – Negative Binomial Regression

```
# Step 3: refit on entire dataset
full_dataset = pd.concat([train, test])
model = smf.glm(formula=model_formula,
                data=full_dataset,
                family=sm.families.NegativeBinomial(alpha=best_alpha))

fitted_model = model.fit()
return fitted_model
```

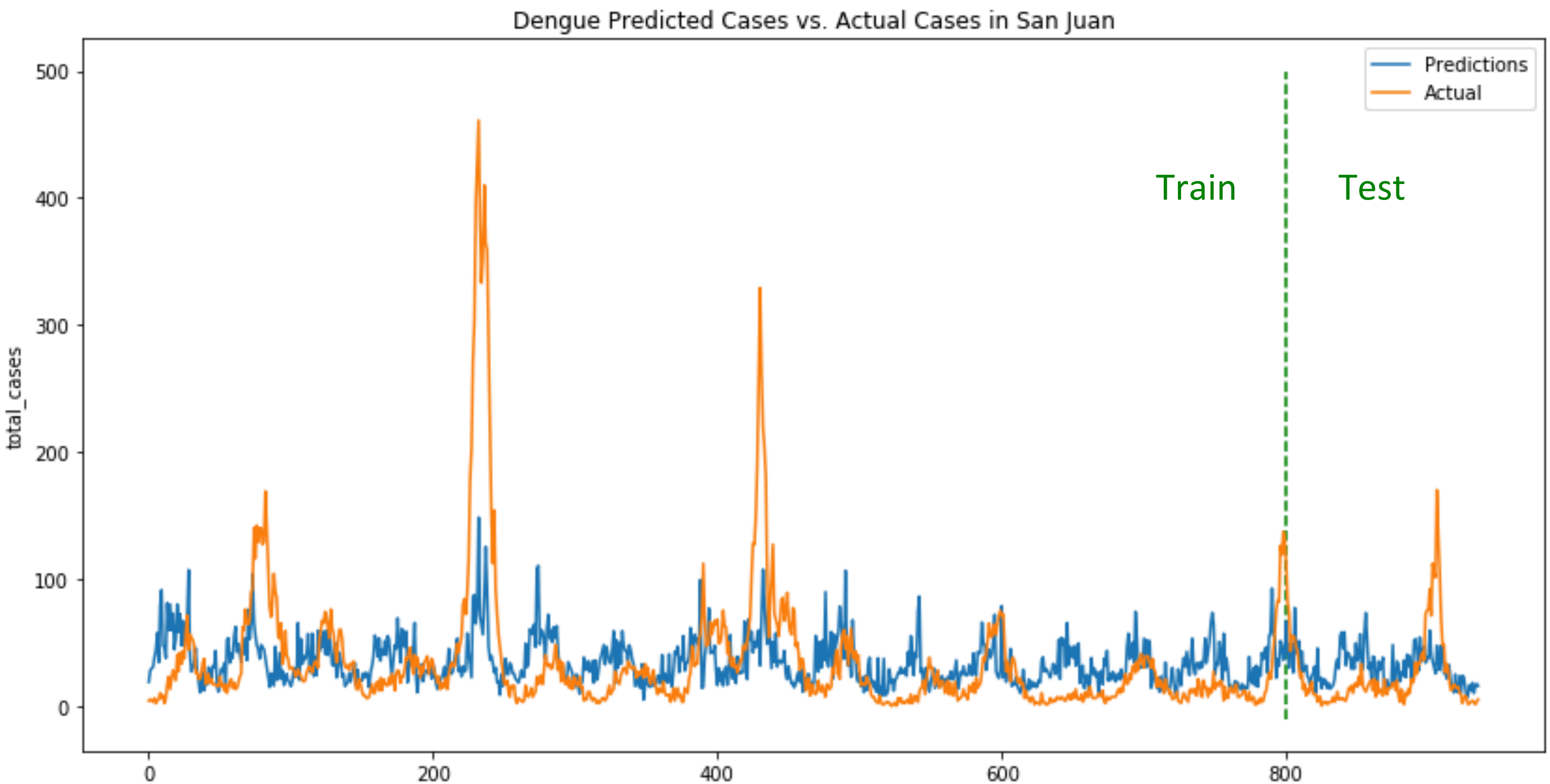
Model – Negative Binomial Regression

```
# Step 3: refit on entire dataset
full_dataset = pd.concat([train, test])
model = smf.glm(formula=model_formula,
                data=full_dataset,
                family=sm.families.NegativeBinomial(alpha=best_alpha))

fitted_model = model.fit()
return fitted_model
```

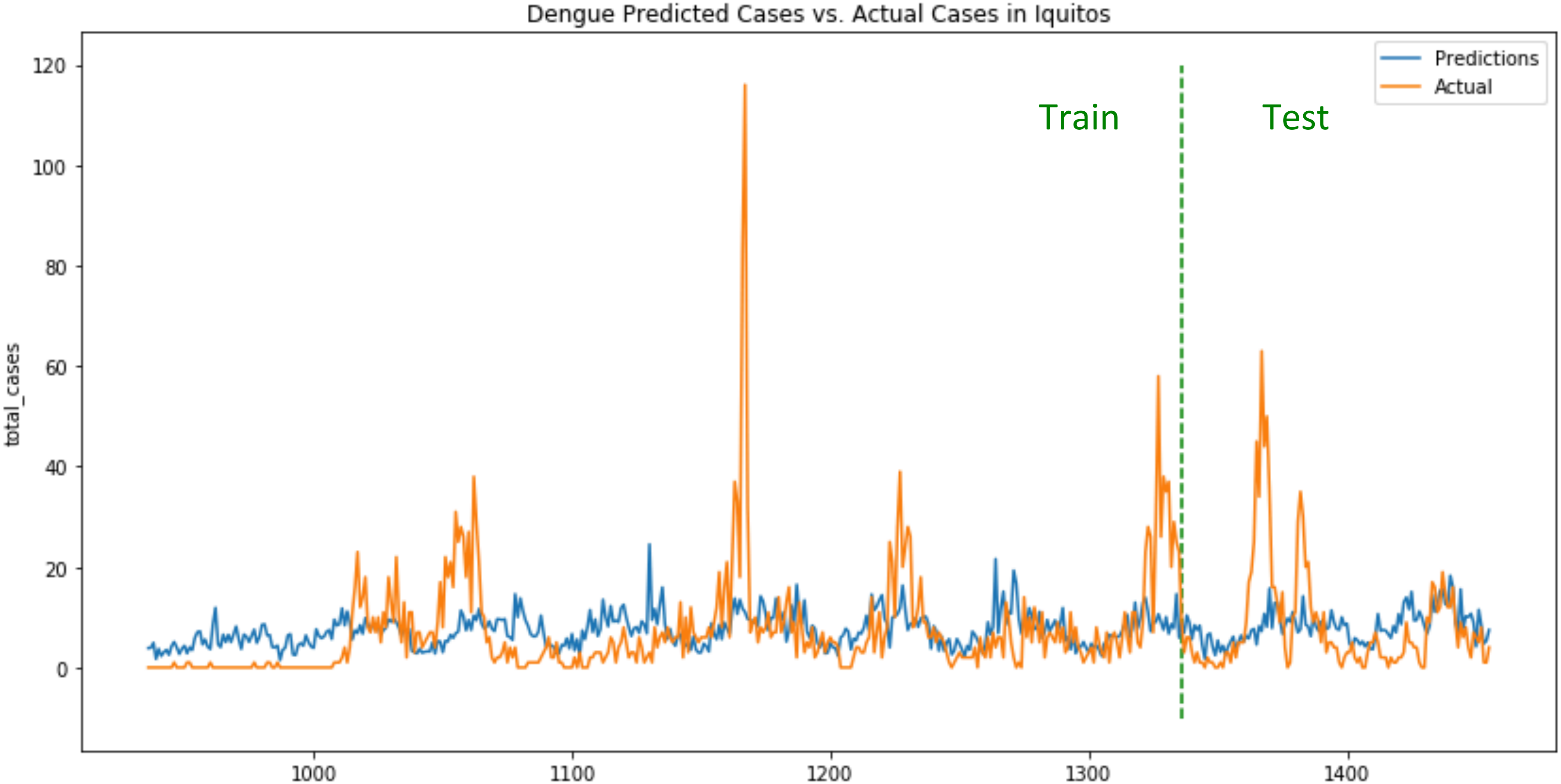
- Best score for San Juan is 23.3
- Best score for iquitos is 7.0

$$MAD = \frac{1}{n}\sum_{i=1}^{n}\left|A_i - \hat{A}_i\right|$$

Model – Negative Binomial Regression



Model – Negative Binomial Regression



Model – Random Forest

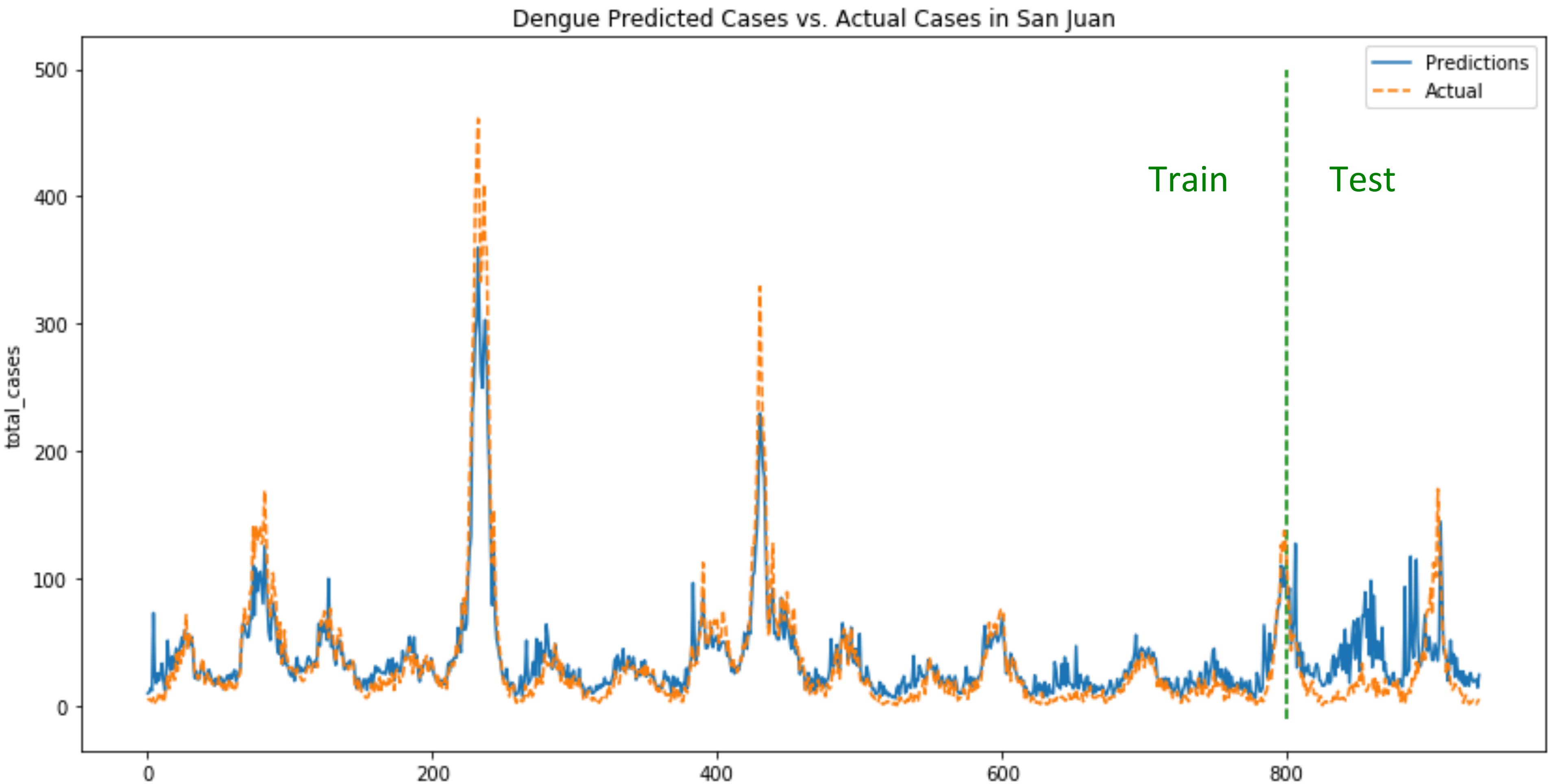
```
from sklearn.ensemble import RandomForestRegressor
model_sj = RandomForestRegressor(100)
model_iq = RandomForestRegressor(100)

model_sj.fit(sj_train.drop(["total_cases"],axis=1), sj_train['total_cases'])
model_iq.fit(iq_train.drop(["total_cases"],axis=1), iq_train['total_cases'])
```

```
ypred_sj = model_sj.predict(X_sj.drop(["total_cases","fitted"],axis=1))
ypred_iq = model_iq.predict(X_iq.drop(["total_cases","fitted"],axis=1))
```

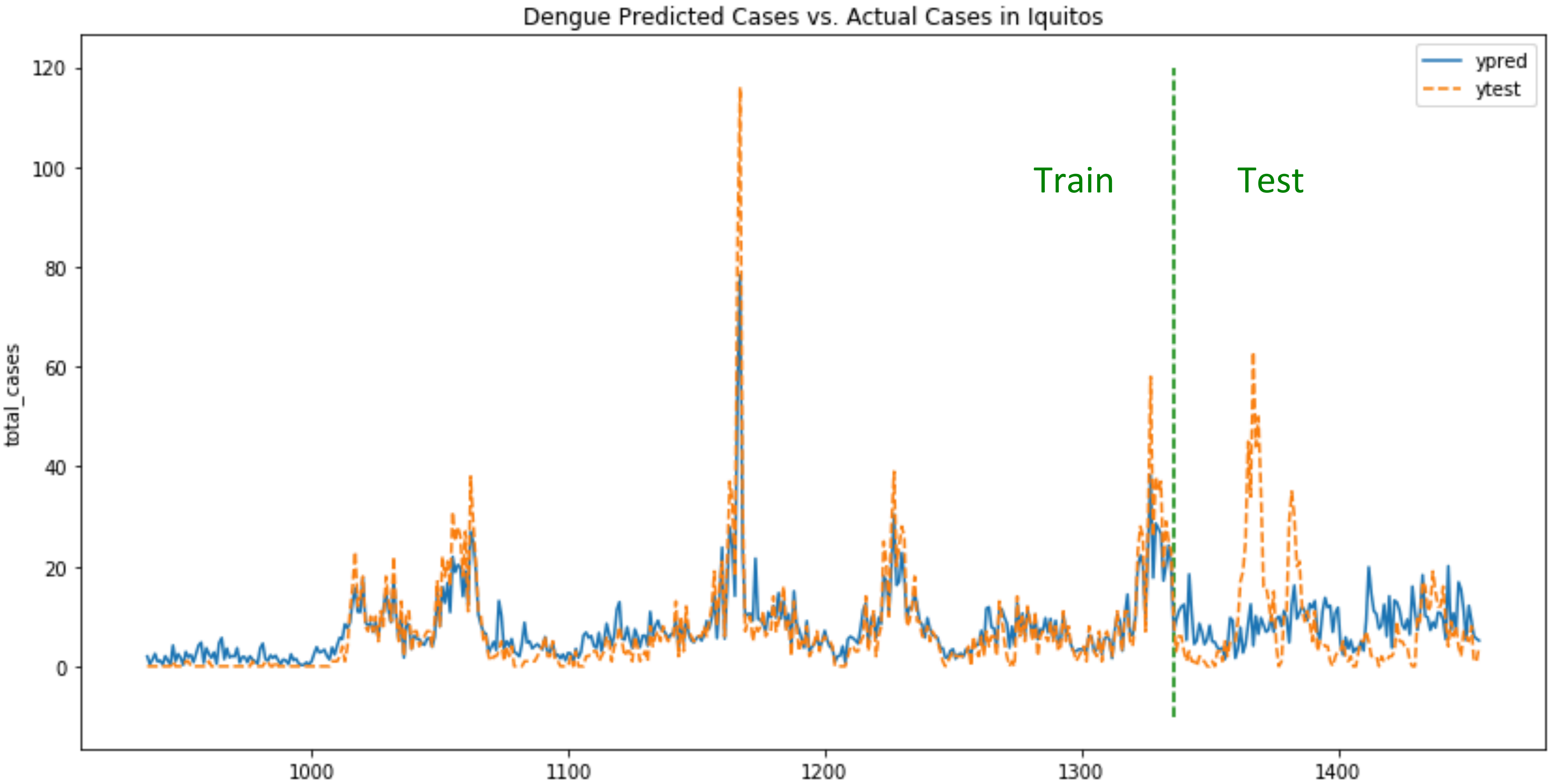
- Best score for San Juan is 12.8
- Best score for iquitos is 3.7

$$MAD = \frac{1}{n}\sum_{i=1}^{n}\left|A_i - \hat{A}_i\right|$$

Model – Random Forest



Model – Random Forest



Thanks

Correlation Analysis

- The reanalysis specific humidity and reanalysis dew point temperature were the most strongly correlated with total cases. This supported the assumption that mosquitoes thrive in wet climates, which could lead to more dengue cases.

- Temperature and total dengue cases showed positive correlation, indicating higher cases of dengue during warm weather.

- In general, the precipitation measurements had weak correlation to total cases.

Variable Rescaling

The wide variation in the value ranges resulted from the use of different scale, and necessitated rescaling to avoid biasing the data models. All fields were brought to comparable scales, such as °C for temperature and mm for precipitation.