CONCLUSIONS FOR DENGAI ANALYSIS

# #INITIAL & EXPLORATORY ANALYSIS

### ###GRAPH: Frequency histogram of all variables in training set (both cities together)

Various variables are normally distributed but others are not

Total cases are skewed right

### ###GRAPH: Frequency histogram of all variables in training set for \*\*SJ\*\*

Same as above

### ###GRAPH: Frequency histogram of all variables in training set for \*\*IQ\*\*

Same as above

### ###GRAPH: Climate variables by time for \*\*SJ\*\*

(feature by week\_start\_date)

Seasonal effects can be seen by most variables but some have no pattern

### ###GRAPH: Climate variables by time for \*\*IQ\*\*

Same as above

### ###GRAPH: Climate variables by week for \*\*SJ\*\*

Flat boxplot include variables:

ndvi\_ne,

ndvi\_nw,

ndvi\_sw,

ndvi\_se,

precipitation\_amt\_mm,

reanalysis\_precip\_amt\_kg\_per\_m2, reanalysis\_relative\_humidity\_percent,

reanalysis\_sat\_precip\_amt\_mm,

reanalysis\_tdlr\_k,

station\_diur\_temp\_rng\_c,

station\_precip\_mm

The rest of the variables so an increase starting at week 17, peaking at week 41 then coming back down again.

### ###GRAPH: Climate variables by week for \*\*IQ\*\*

Same as above

### ###GRAPH: Total\_cases by time for \*\*SJ\*\*, \*\*IQ\*\* and together

Bar graph shows various peaks but there does appear to be seasonality to the cases

### ###GRAPH: Average Total\_cases by week for \*\*SJ\*\*, \*\*IQ\*\*

The rest of the variables so an increase starting at week 17, peaking at week 41 then coming back down again for SJ.

If IQ, the lowest point is at week 30 then it goes up until about week 50. Week 51 and 52 appeak to have a large drop possible due to Christmas.

### ###GRAPH: Total\_cases by climate variables (both cities together)

No pattern emerges

### ###GRAPH: Total\_cases by climate variables for \*\*SJ\*\*

Same as above

### ###GRAPH: Total\_cases by climate variables for \*\*IQ\*\*

Same as above

### ###Compare the means between same variables in different cities

The means of the two cities are significantly different for each feature.

## ##Compare similar variable values within the dataset

### ###Difference in max air temp

"station\_max\_temp\_c"" and "reanalysis\_max\_air\_temp\_k" (scaled to Celcius) are different to each other by 2-4 degrees C. Do not drop either variable at this point.

### ###Difference in min air temp

"station\_min\_temp\_c"" and "reanalysis\_min\_air\_temp\_k" (scaled to Celcius) are different to each other by minus 1-2 degrees C. Do not drop either variable at this point.

### ###Difference in average air temp

"station\_avg\_temp\_c"" and "reanalysis\_avg\_temp\_k" (scaled to Celcius) are different to each other by 0-2 degrees C. Do not drop either variable at this point.

### ###Difference in total precipitation

"station\_precip\_mm", "precipitation\_amt\_mm", "reanalysis\_sat\_precip\_amt\_mm", "reanalysis\_precip\_amt\_kg\_per\_m2" -> when looking at each one of these in combination, most have a median difference around zero but there are many outliers.

The only attributes which had exactly the same values are precipitation\_amt \_mm and reanalyisi\_sat\_amt\_mm. One of these two variables can be removed.

## ##Review of climate factors independently (SJ ONLY)

10-fold cross fitting of Decision tree with only total\_cases and listed variables

Overfitting is possible in all options.

### ###Decision Tree with vegetation data

form <- "total\_cases ~ ndvi\_se + ndvi\_sw + ndvi\_ne + ndvi\_nw"

"average error using k-fold cross validation and decision tree algorithm: 96.582 percent"

### ###Decision Tree with temperature data

form <- "total\_cases ~ station\_max\_temp\_c + station\_min\_temp\_c + station\_avg\_temp\_c + station\_diur\_temp\_rng\_c + reanalysis\_dew\_point\_temp\_k + reanalysis\_air\_temp\_k + reanalysis\_max\_air\_temp\_k + reanalysis\_min\_air\_temp\_k + reanalysis\_avg\_temp\_k + reanalysis\_tdtr\_k"

"average error using k-fold cross validation and decision tree algorithm: 97.331 percent"

### ###Decision Tree with precipitation data

form <- "total\_cases ~ station\_precip\_mm + precipitation\_amt\_mm + reanalysis\_sat\_precip\_amt\_mm + reanalysis\_precip\_amt\_kg\_per\_m2"

"average error using k-fold cross validation and decision tree algorithm: 97.439 percent"

### ###Decision Tree with humidity data

form <- "total\_cases ~ reanalysis\_relative\_humidity\_percent + reanalysis\_specific\_humidity\_g\_per\_kg"

"average error using k-fold cross validation and decision tree algorithm: 97.546 percent"

# #ANALYSIS OF OUTLIERS

### ###GRAPH: Boxplot of climate variables (test and train)

Many outliers exist for all variables but they will be ignored for now

# #MISSING VALUES

### ###Check for missing values (is.na)

There are many missing values for both SJ and IQ

### ###Missing values: Remove all rows with an NA in it

Removed rows using na.omit

### ###Missing values: Using last non-NA value

Filled in the last no-NA value using na.locf

# #REMOVE week\_start\_date AND city FROM DATASET

Created a new dataframe to keep all the original variables (iq/sj\_train\_labels.startweek) and removed city and week start date from the lastna dataframe.

# #CORRELATION ANALYSIS

## ##Correlation analysls before missing values are addressed

Correlation graph not easy to visualize within R

## ##Comparison of correlations with the other non-na dataframes

### ###Correlation with na.omit

Correlation graph not easy to visualize within R

### ###Correlation with lastna

Correlation graph not easy to visualize within R

## ##Conclusion about correlation USE LASTNA

Different methods of imputing missing values had no impact on correlation. Will stick with lastna as the final version.

# #FEATURE SELECTION and DIMENTIONALITY REDUCTION

We will use various methods to see if we can find any features that need to be eliminated

## ##Feature selection via CaretR (Remove redundant features)

### ###CaretR (Remove redundant features) for SJ

The following is the output of the CaretR correlation. findCorrelation function: This function searches through a correlation matrix and returns a vector of integers corresponding to columns to remove to reduce pair-wise correlations.

[1] 16 10 8 9 18 11 12 7 5

[1] "reanalysis\_specific\_humidity\_g\_per\_kg"

[2] "reanalysis\_dew\_point\_temp\_k"

[3] "reanalysis\_air\_temp\_k"

[4] "reanalysis\_avg\_temp\_k"

[5] "station\_avg\_temp\_c"

[6] "reanalysis\_max\_air\_temp\_k"

[7] "reanalysis\_min\_air\_temp\_k"

[8] "precipitation\_amt\_mm"

[9] "ndvi\_se"

### ###CaretR (Remove redundant features) for IQ

The following is the output of the CaretR correlation. findCorrelation function: This function searches through a correlation matrix and returns a vector of integers corresponding to columns to remove to reduce pair-wise correlations.

[1] 17 11 16 10 9 7 5 3 4

[1] "reanalysis\_tdtr\_k"

[2] "reanalysis\_max\_air\_temp\_k"

[3] "reanalysis\_specific\_humidity\_g\_per\_kg"

[4] "reanalysis\_dew\_point\_temp\_k"

[5] "reanalysis\_avg\_temp\_k"

[6] "precipitation\_amt\_mm"

[7] "ndvi\_se"

[8] "ndvi\_ne"

[9] "ndvi\_nw"

## ##Feature selection via Boruta

### ###Boruta for SJ

21 attributes confirmed important: ndvi\_ne, ndvi\_nw, ndvi\_se, ndvi\_sw,

reanalysis\_air\_temp\_k and 16 more;

1 attributes confirmed unimportant: precipitation\_amt\_mm;

### ###Boruta for IQ

15 attributes confirmed important: precipitation\_amt\_mm,

reanalysis\_air\_temp\_k, reanalysis\_avg\_temp\_k, reanalysis\_dew\_point\_temp\_k,

reanalysis\_min\_air\_temp\_k and 10 more;

7 attributes confirmed unimportant: ndvi\_ne, ndvi\_nw, ndvi\_se, ndvi\_sw,

reanalysis\_max\_air\_temp\_k and 2 more;

## ##Feature Selection using Random Forest

### ###Random Forest for SJ

| **% Inc MSE** | |
| --- | --- |
|  |  |
| **ndvi\_se** | 37.442900 |
| **ndvi\_sw** | 23.897640 |
| **ndvi\_nw** | 14.704934 |
| **reanalysis\_specific\_humidity\_g\_per\_kg** | 13.107286 |
| **reanalysis\_dew\_point\_temp\_k** | 12.932920 |
| **ndvi\_ne** | 12.373046 |
| **reanalysis\_relative\_humidity\_percent** | 12.324789 |
| **reanalysis\_min\_air\_temp\_k** | 12.194454 |
| **station\_precip\_mm** | 12.182469 |
| **precipitation\_amt\_mm** | 10.185727 |
| **reanalysis\_precip\_amt\_kg\_per\_m2** | 9.870209 |
| **reanalysis\_sat\_precip\_amt\_mm** | 8.691110 |
| **reanalysis\_max\_air\_temp\_k** | 8.588030 |
| **reanalysis\_air\_temp\_k** | 7.381811 |
| **station\_avg\_temp\_c** | 7.368847 |
| **reanalysis\_avg\_temp\_k** | 7.064424 |
| **station\_diur\_temp\_rng\_c** | 6.760298 |
| **reanalysis\_tdtr\_k** | 6.710697 |
| **station\_min\_temp\_c** | 6.343531 |
| **station\_max\_temp\_c** | 5.451985 |

### ###Random Forest for IQ

| **% Inc MSE** | |
| --- | --- |
|  |  |
| **reanalysis\_specific\_humidity\_g\_per\_kg** | 13.760845 |
| **station\_avg\_temp\_c** | 13.281377 |
| **reanalysis\_dew\_point\_temp\_k** | 11.658470 |
| **reanalysis\_precip\_amt\_kg\_per\_m2** | 11.108627 |
| **station\_max\_temp\_c** | 10.650739 |
| **station\_diur\_temp\_rng\_c** | 10.135905 |
| **reanalysis\_relative\_humidity\_percent** | 10.052518 |
| **reanalysis\_air\_temp\_k** | 9.346375 |
| **reanalysis\_tdtr\_k** | 8.533772 |
| **reanalysis\_min\_air\_temp\_k** | 8.188367 |
| **reanalysis\_max\_air\_temp\_k** | 6.869346 |
| **reanalysis\_avg\_temp\_k** | 6.032521 |
| **ndvi\_sw** | 5.744110 |
| **ndvi\_ne** | 5.433112 |
| **ndvi\_se** | 4.197998 |
| **station\_precip\_mm** | 3.702077 |
| **reanalysis\_sat\_precip\_amt\_mm** | 3.651217 |
| **precipitation\_amt\_mm** | 3.532636 |
| **ndvi\_nw** | 1.696376 |
| **station\_min\_temp\_c** | 1.574023 |