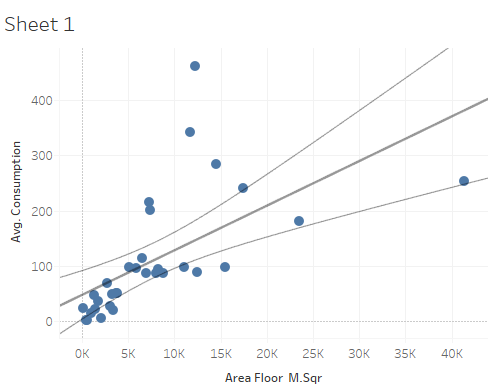
Midterm Part 3

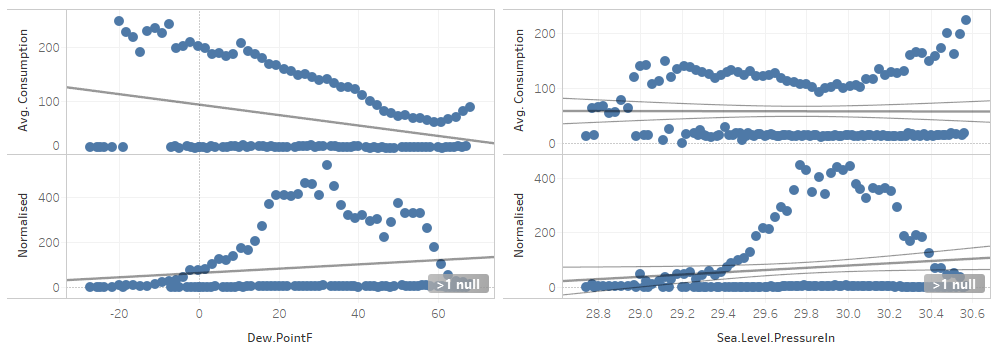
1.Feature Engineering is one of the initial steps for model selection we can see the relation between the independent variable and the dependent variable and do an initial analysis of if we want to add that particular variable to our model

And then see if any independent variables have any relation between each other and pick only the one that would have larger accuracy of prediction.

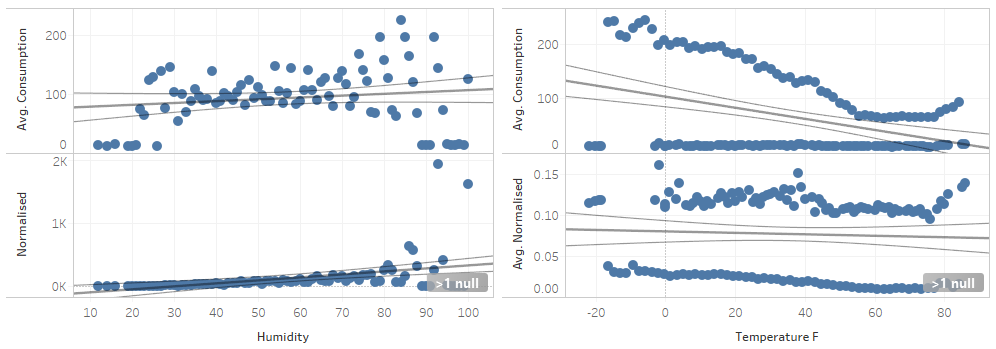
After this we get it checked in R by forward, Backward and exhaustive method of regression but this method will give us an insight in interpreting those results.



Area and the consumption has a positive relationship. We are going to normalize the consumption my getting to consumption per meter square. Then we have plotted each independent variable against consumption and normalized consumption.

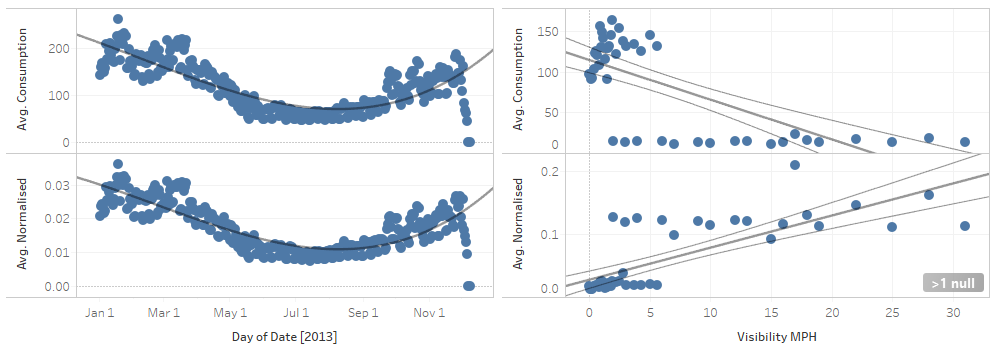


Dew Point & Sea. Level Pressure: the correlation is not clear only visual analysis would not be sufficient.



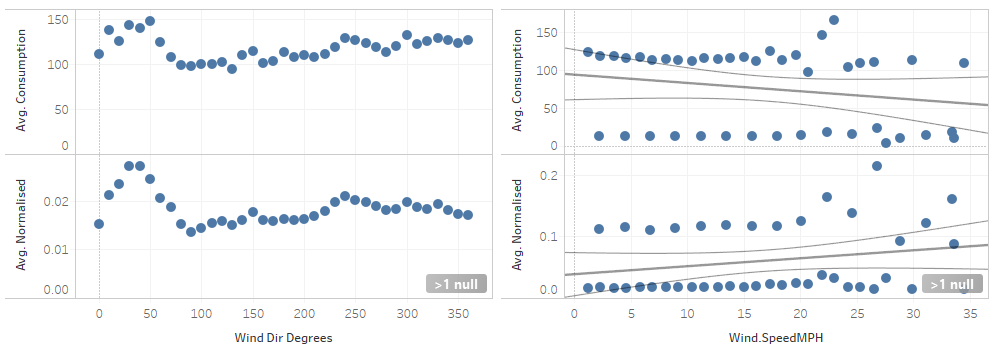
Humidity: has a positive relationship.

Temperature: has a negative relationship which is much more clear when plotted against consumption.

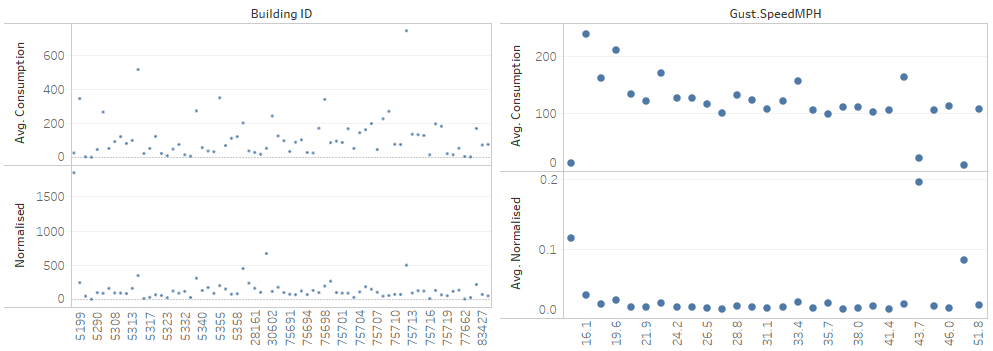


Date has a polynomial relationship.

Visibility has a positive relationship with Normalized consumption and negative relationship with avg. consumption.

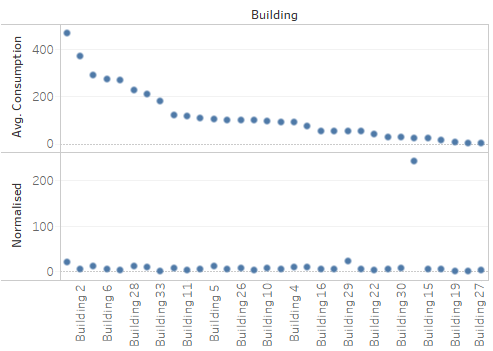
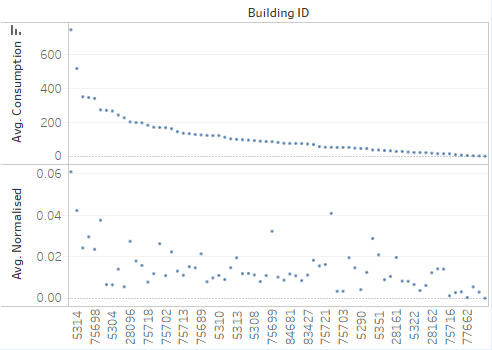


Wind direction does not have any correlation but wind direction has a positive correlation.



The buildingID does not have any relation between themselves and no correlation between the ID and consumption and Normalized.

Gust Speed does not have any correlation with either consumption or Normalized.



On plotting the Building ID and Building in descending order the behavior of consumption and Normalized is different as the Building ID and Building number are attributes which do not have any relationship between each other.

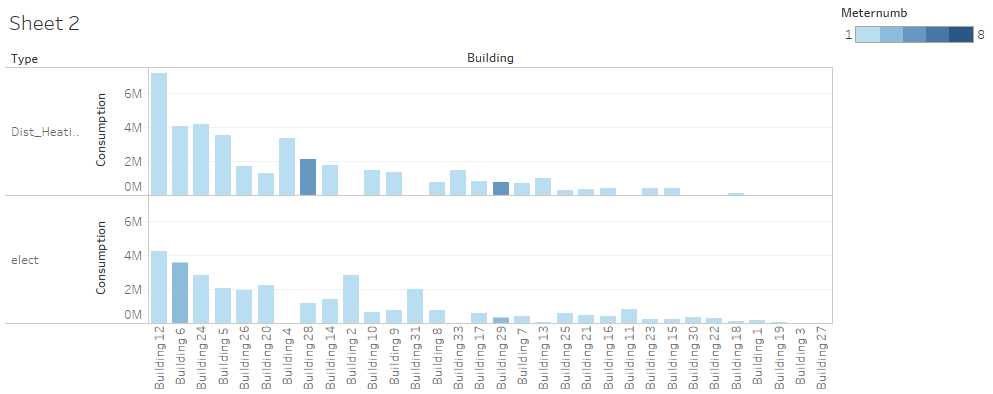
The rest of the plots on the Dashboard are against factors so we will select them only on R modelling.

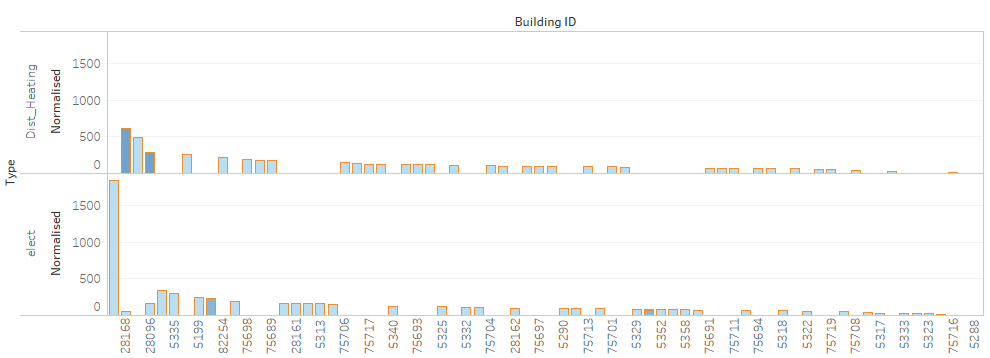
3. From the Dash board you can see that all the building IDs have either electricity or heating (except 28096, 28168).

Building ID can be used to understand the model better.

You can see in the picture below.

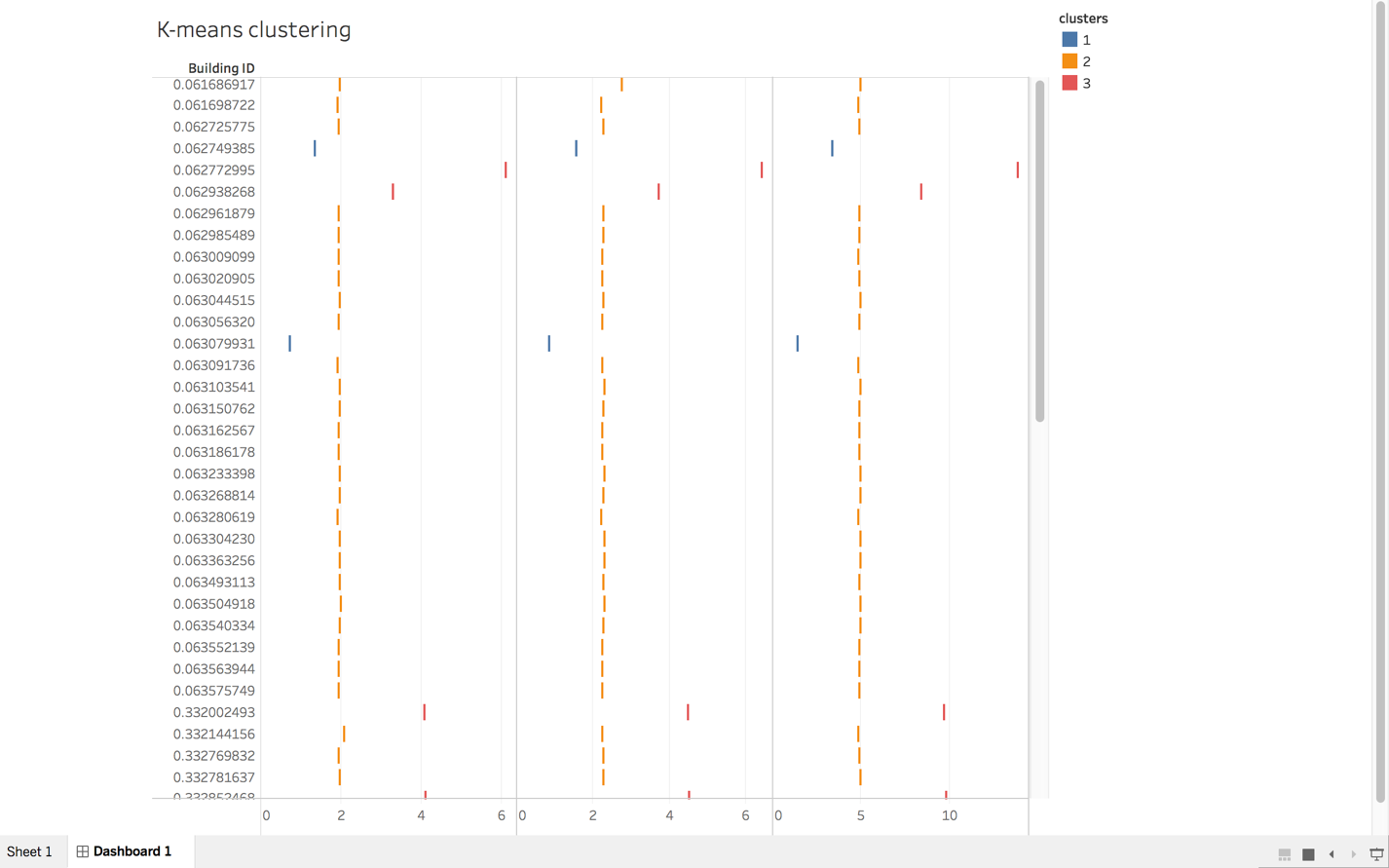
There are more unique combination in second graph.





R- Tableau Integration

K-means clustering



For Prediction, Classification

We had used library(h2o) to run our models and get the best possible model.

SCRIPT\_REAL('

library(h2o)

y.dep <- .arg1

system.time( rforest.model <- h2o.randomForest(y=.arg1, x=.arg2, training\_frame =

train.h2o, ntrees = 1000, mtries = 3, max\_depth = 4, seed = 1122)

For both prediction & classification our best model was random forest. But I have attached a linear regression model that we tried invoking in tableau.

I followed the tutorials that was specified in this link

<https://github.com/h2oai/h2o-world-2014-training/blob/master/tutorials/extab/tableau.md>

To invoke my model into Tableau but it kept throwing Error:Base Parse [213221] and it didn’t execute my R code in the Calculated field column.