[Company name]

Clustering

K Means and Hierarchical Clustering

Team TBD

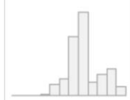
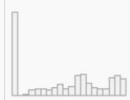
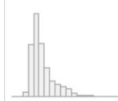
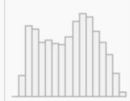
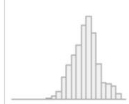
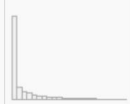
[Date]

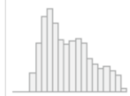
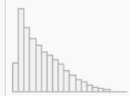
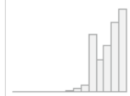
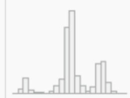
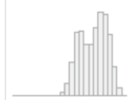
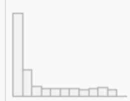
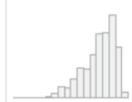
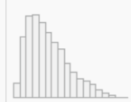
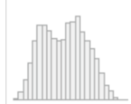
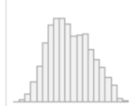
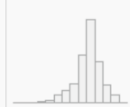
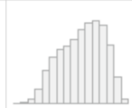
PART I Descriptive Statistics

Dataset Source: https://www.kaggle.com/nicholasjhana/energy-consumption-generation-prices-and-weather (Same as last time)

As an inherited report of multiple regression, we are going to explore the application of logistic regression, K-NN (k nearest neighbor) and classification tree in classifying the #LOAD column. Besides, we will compare these models' performances in the task respectively and asses their alliance of the outcomes.

Below is a chart of the descriptive statistics of variables in pool.

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
1	generation.biomass [numeric]	Mean (sd) : 383.5 (85.3) min < med < max: 0 < 367 < 592 IQR (CV) : 100 (0.2)	435 distinct values		35064 (100%)	0 (0%)
2	generation.fossil.brown.coal.lignite [numeric]	Mean (sd) : 448.1 (354.6) min < med < max: 0 < 509 < 999 IQR (CV) : 757 (0.8)	964 distinct values		35064 (100%)	0 (0%)
3	generation.fossil.gas [numeric]	Mean (sd) : 5622.7 (2201.5) min < med < max: 0 < 4969.5 < 20034 IQR (CV) : 2303 (0.4)	8310 distinct values		35064 (100%)	0 (0%)
4	generation.fossil.hard.coal [numeric]	Mean (sd) : 4256.5 (1962) min < med < max: 0 < 4475 < 8359 IQR (CV) : 3312 (0.5)	7279 distinct values		35064 (100%)	0 (0%)
5	generation.fossil.oil [numeric]	Mean (sd) : 298.3 (52.5) min < med < max: 0 < 300 < 449 IQR (CV) : 67 (0.2)	334 distinct values		35064 (100%)	0 (0%)
6	generation.hydro.pumped.storage.consumption [numeric]	Mean (sd) : 475.6 (792.3) min < med < max: 0 < 68 < 4523 IQR (CV) : 616 (1.7)	3319 distinct values		35064 (100%)	0 (0%)

7	generation.hydro.run.of.river.and.poundage [numeric]	Mean (sd) : 972.2 (400.7) min < med < max: 0 < 906 < 2000 IQR (CV) : 613 (0.4)	1697 distinct values		35064 (100%)	0 (0%)
8	generation.hydro.water.reservoir [numeric]	Mean (sd) : 2605.5 (1835.2) min < med < max: 0 < 2165 < 9728 IQR (CV) : 2680 (0.7)	7040 distinct values		35064 (100%)	0 (0%)
9	generation.nuclear [numeric]	Mean (sd) : 6263.5 (840.3) min < med < max: 0 < 6564 < 7117 IQR (CV) : 1266 (0.1)	2396 distinct values		35064 (100%)	0 (0%)
10	generation.other [numeric]	Mean (sd) : 60.2 (20.2) min < med < max: 0 < 57 < 106 IQR (CV) : 27 (0.3)	112 distinct values		35064 (100%)	0 (0%)
11	generation.other.renewable [numeric]	Mean (sd) : 85.6 (14.1) min < med < max: 0 < 88 < 119 IQR (CV) : 24 (0.2)	87 distinct values		35064 (100%)	0 (0%)
12	generation.solar [numeric]	Mean (sd) : 1432.8 (1680) min < med < max: 0 < 616 < 5792 IQR (CV) : 2508 (1.2)	5344 distinct values		35064 (100%)	0 (0%)
13	generation.waste [numeric]	Mean (sd) : 269.4 (50.2) min < med < max: 0 < 279 < 357 IQR (CV) : 70 (0.2)	268 distinct values		35064 (100%)	0 (0%)
14	generation.wind.onshore [numeric]	Mean (sd) : 5465 (3213.6) min < med < max: 0 < 4849.5 < 17436 IQR (CV) : 4466.5 (0.6)	11477 distinct values		35064 (100%)	0 (0%)
15	total.load.actual [numeric]	Mean (sd) : 28698.3 (4575.8) min < med < max: 18041 < 28902 < 41015 IQR (CV) : 7387.2 (0.2)	15149 distinct values		35064 (100%)	0 (0%)
16	price.actual [numeric]	Mean (sd) : 57.9 (14.2) min < med < max: 9.3 < 58 < 116.8 IQR (CV) : 18.7 (0.2)	6653 distinct values		35064 (100%)	0 (0%)
17	temp [numeric]	Mean (sd) : 289.7 (7.3) min < med < max: 271.9 < 289 < 309.3 IQR (CV) : 11 (0)	19181 distinct values		35064 (100%)	0 (0%)
18	pressure [numeric]	Mean (sd) : 1016.1 (8.2) min < med < max: 974.6 < 1016.8 < 1039.8 IQR (CV) : 8 (0)	695 distinct values		35064 (100%)	0 (0%)
19	humidity [numeric]	Mean (sd) : 68 (14.8) min < med < max: 22.6 < 69.6 < 100 IQR (CV) : 23.4 (0.2)	364 distinct values		35064 (100%)	0 (0%)

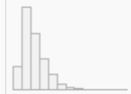
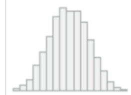
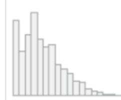
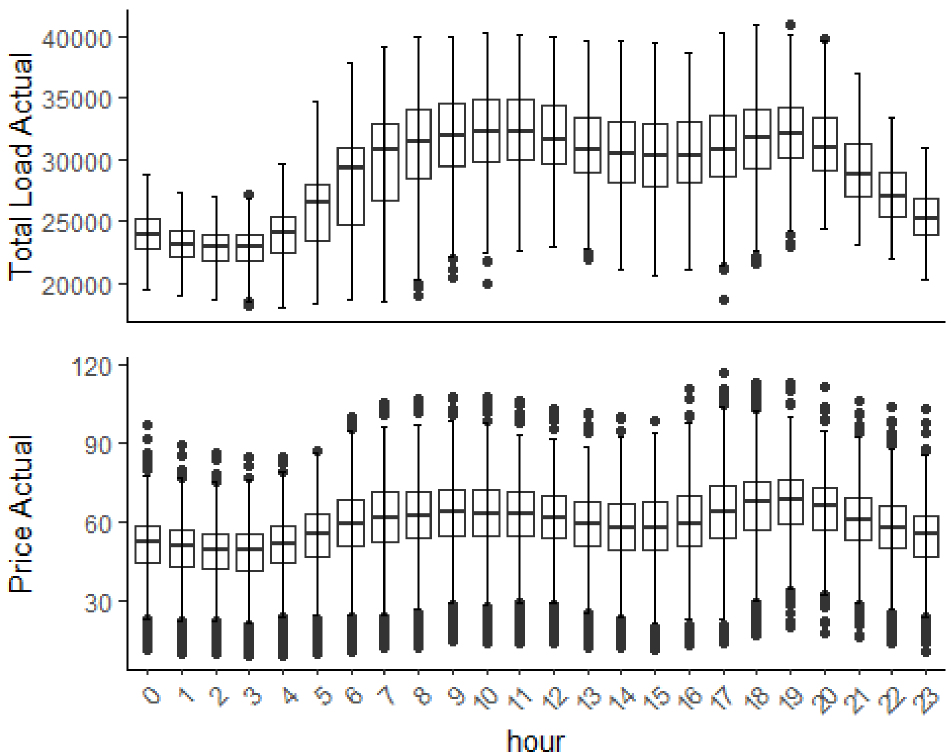
20	wind_speed [numeric]	Mean (sd) : 2.5 (1.3) min < med < max: 0 < 2.2 < 12.8 IQR (CV) : 1.8 (0.5)	60 distinct values		35064 (100%)	0 (0%)
21	wind_deg [numeric]	Mean (sd) : 166.7 (57.2) min < med < max: 0 < 166.2 < 338 IQR (CV) : 80.2 (0.3)	1502 distinct values		35064 (100%)	0 (0%)
22	rain_1h [numeric]	Mean (sd) : 0.1 (0.2) min < med < max: 0 < 0 < 3.2 IQR (CV) : 0.1 (2.8)	40 distinct values		35064 (100%)	0 (0%)
23	rain_3h [numeric]	Mean (sd) : 0 (0) min < med < max: 0 < 0 < 0.5 IQR (CV) : 0 (8.7)	104 distinct values		35064 (100%)	0 (0%)
24	snow_3h [numeric]	Mean (sd) : 0 (0.1) min < med < max: 0 < 0 < 4.3 IQR (CV) : 0 (20.9)	71 distinct values		35064 (100%)	0 (0%)
25	clouds_all [numeric]	Mean (sd) : 24.3 (17) min < med < max: 0 < 22 < 93.6 IQR (CV) : 22.4 (0.7)	434 distinct values		35064 (100%)	0 (0%)

Table 1 *Descriptive Statistics*

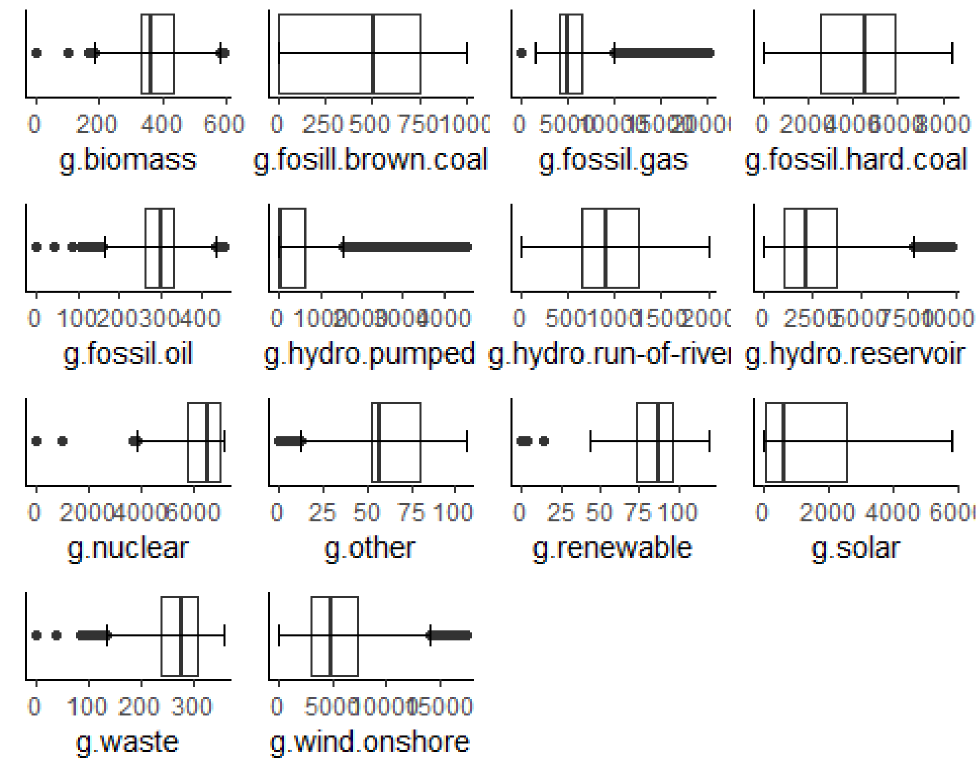
Based on common sense and real-world business, we suspect #LOAD and #PRICE as the target variables. The boxplot shows that #LOAD fluctuates more obviously over time. Meanwhile, power plants wish to know high demand and low demand in the market in order to act accordingly and in time. So, we think #LOAD would be a better choice as a target variable.



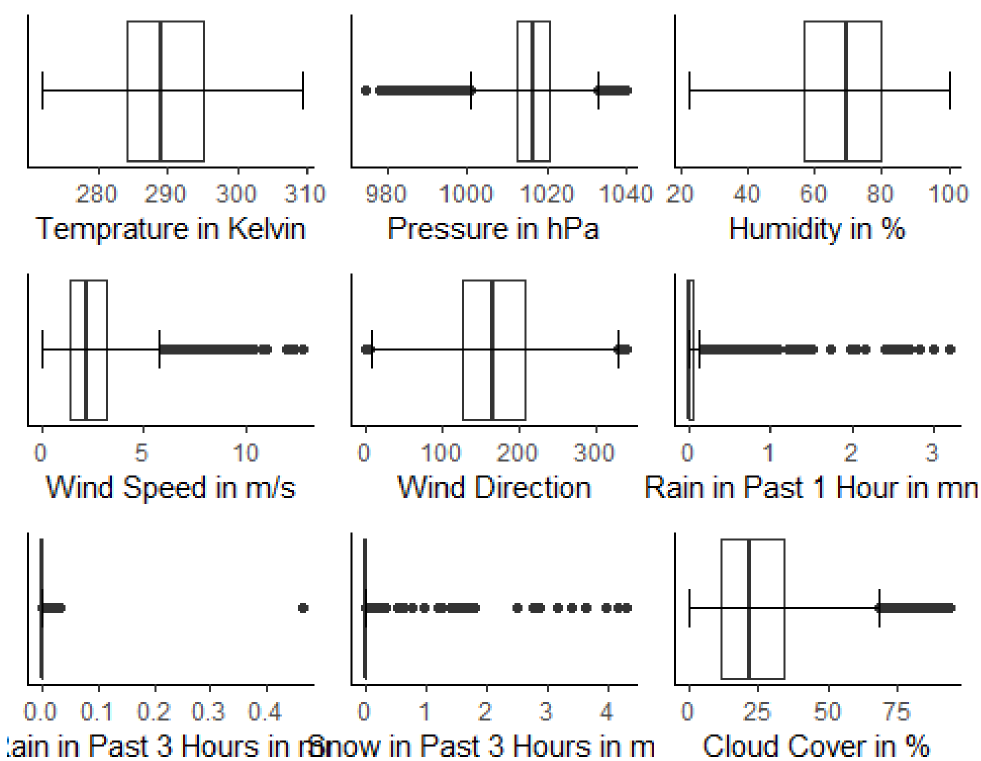
Plot 1 *Demand and Price by Hour*

As last time, we also give the box-and-whisker plots of energy generation methods and weather features. They reflect that these variables are in different scales, suggesting that we need to be cautious when doing the K-NN model. Still, we are not deleting the outliers right now to ensure the integrity of our dataset.

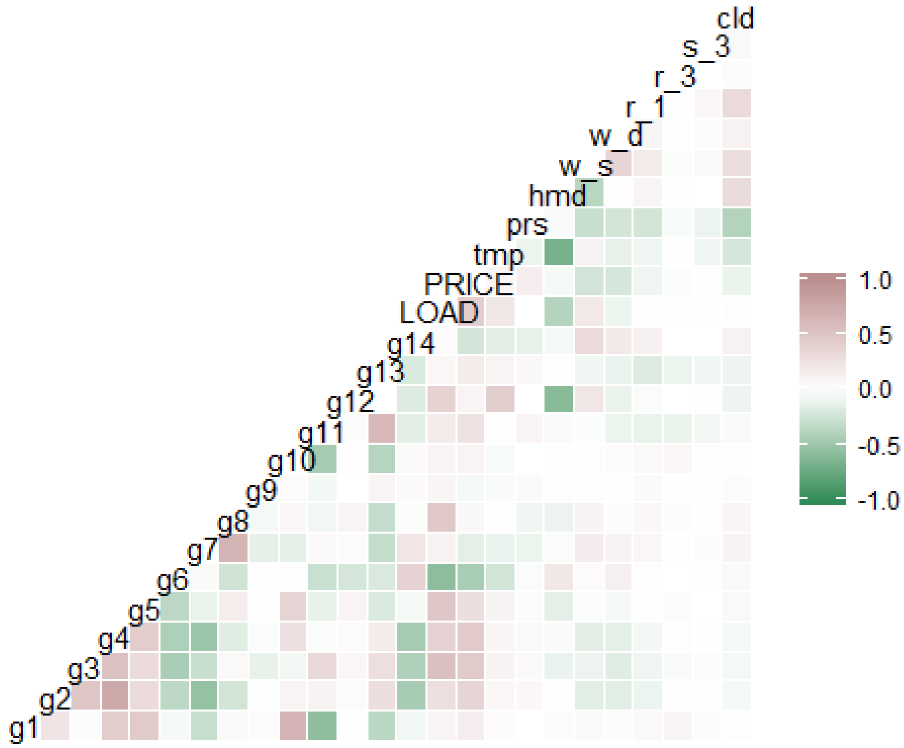
What's more, the correlation matrix and scatterplots are also showed here, which gives us an overall impression of the classification standards. Honestly, if we get familiarized with the distribution and correlation of target variable and potential variables, the model could be judged on an instinctive level of right from wrong, especially true for the classification tree model.



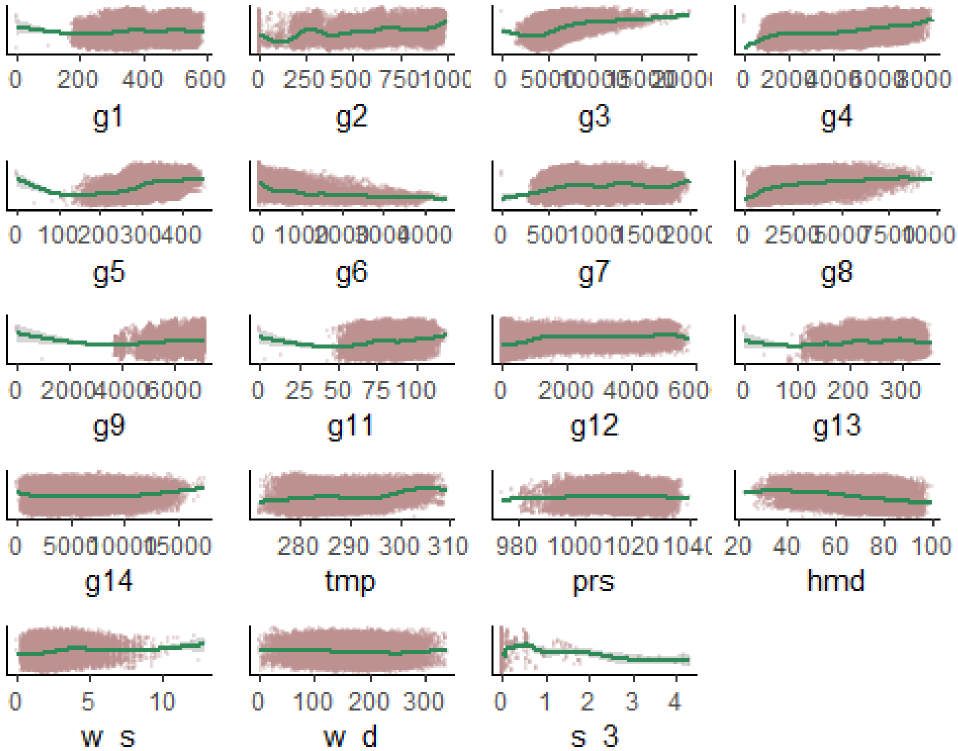
Plot 2 *Energy Generation Boxplot*



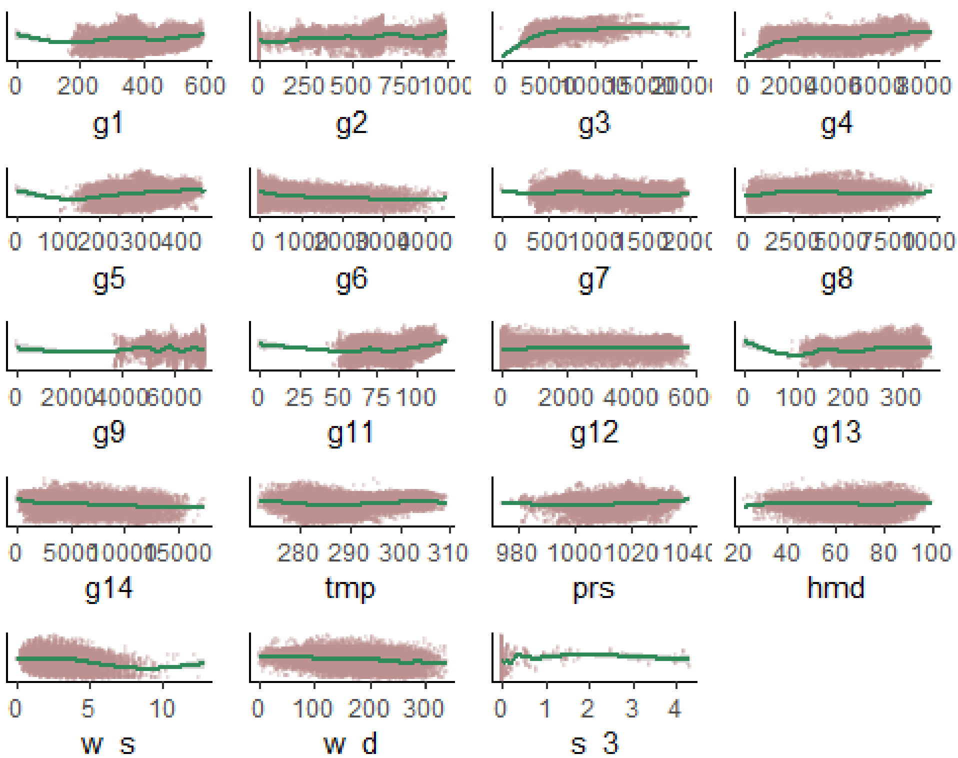
Plot 3 *Weather Feature Boxplot*



Plot 4 *Correlation Matrix*



Plot 5   *Scatterplot of Load vs. Potential Variables*



Plot 6   *Scatterplot of PRICE vs. Potential Variables*

PART II KMeans Clustering

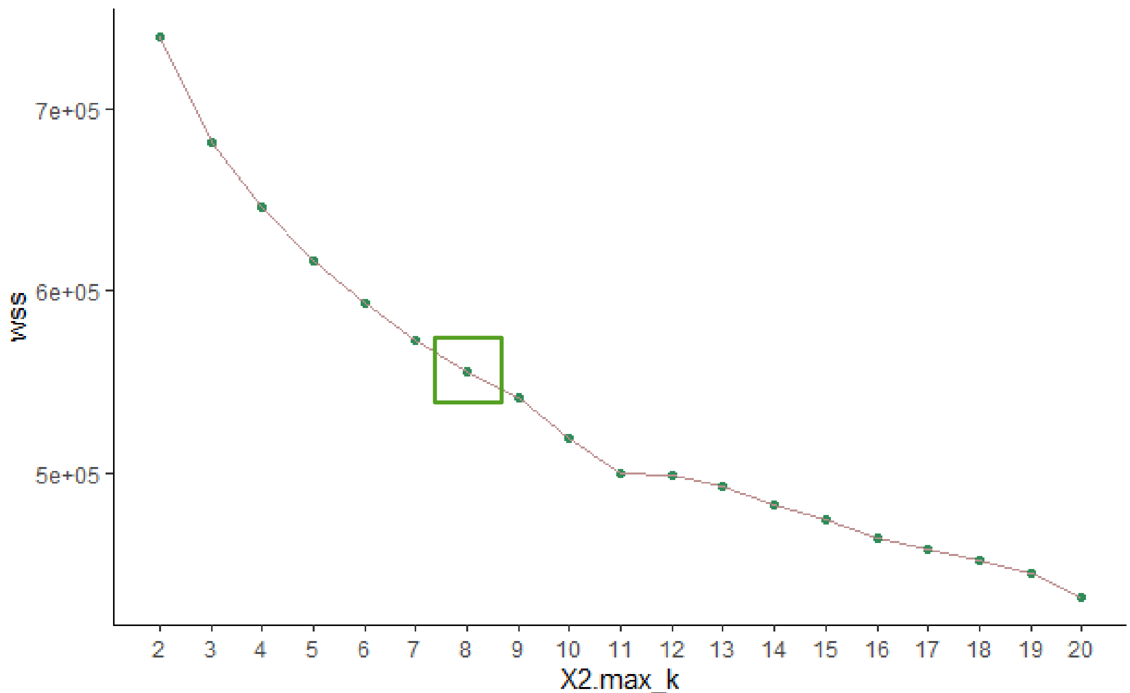
We did not change the dataset for this assignment.

As k-mean is very sensitive to the first choice, and unless the number of observations and groups are small, it is almost impossible to get the same clustering. The discussion is taken out on the charts below.

Before moving into the stage of K Means Clustering, we need to do some pre-process to the original dataset. We remove the category variables and target variable to assure the practicability of K Means clustering. Then, we need to standardize our data since K Means, in essence, is computing the distance between observations and centers, so we want to minimize the scale bias.

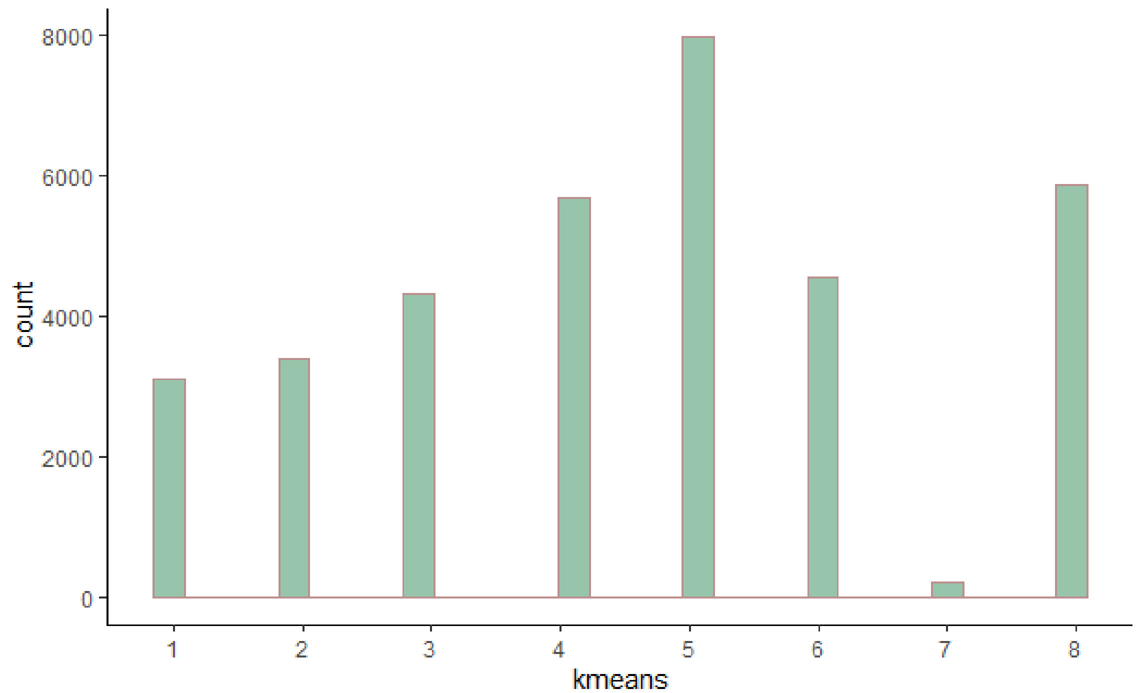
We also need to keep in mind that the number of clusters(k) depends on the nature of the data set, the industry, business and so on. Although there is a rule of thumb that $k = \sqrt{\frac{n}{2}}\ (n\ is\ the\ volume\ of\ observations)$, it will not work for our dataset of more than 35,000 observations.[1] Meanwhile, due to the limitation of our laptop's RAM, we choose elbow method instead of Silhouette Coefficient.

[1] https://www.quora.com/How-can-we-choose-a-good-K-for-K-means-clustering



Plot 1 *Elbow of WSS (Within Sum of Squares)*

As K Means somehow is a random procedure, we set the trial times to 20 and iteration to 100 during each time in order to minimize the bias. From Elbow plot, we can tell the optimal k is 8, where the curve is starting to have a diminishing return.



Plot 2 *Distribution of clusters (k=8)*

Because we are including 24 variables in the clustering, we are not able to show the scatterplot here. The principal is that, the size of each cluster should be as even as possible. Except for cluster 7, which is very susceptible to be outlier, all other clusters are relevantly ideal. It might be good to have homogeneity between clusters, if not, a thinner data preparation might be required. Then, what is the feature of each cluster?

Plot 3 *Heatmap of K Means (k=8)*

Overall speaking, the color code indicates that the greener the block is, the lower the value it has, similarly, the pinker the block is, the higher the value it has. Take cluster 8 for an example, it ranks highest in the feature of #g6 *(#generation hydro pumped storage consumption)* and a relatively low in the feature of actual #price, g2 *(#generation fossil brown coal/lignite)*, g3 *(#generation fossil gas)*, g4 *(#generation fossil hard coal)*, and g5 *(#generation fossil oil)*. We can learn others by analogy and are not going to repeat here.

PART III Hierarchical Clustering

We should do the same data pre-process as K Means.

Hierarchical clustering is an alternative approach to k-means clustering for identifying groups in the dataset. It does not require us to pre-specify the number of clusters to be generated as is required by the k-means approach. Furthermore, hierarchical clustering has an added advantage over K-means clustering in that it results in an attractive tree-based representation of the observations, called a dendrogram.

After precise permutation, we decide to use Euclidean to calculate distance within and between clusters and Ward's minimum variance method to realize linkage.

The distance calculated by Euclidean between these two points is quantified based on the Pythagoras Theorem. It's sensible and convenient to use the Euclidean norm, because this is the only norm up (up to linear transformation\choice of basis) that has an associated an inner product, which is essentially a generalization of the notion of an angle, namely, rotation invariant, which makes Euclidean the most appropriate as our topic of energy is also rotation invariant.[2]

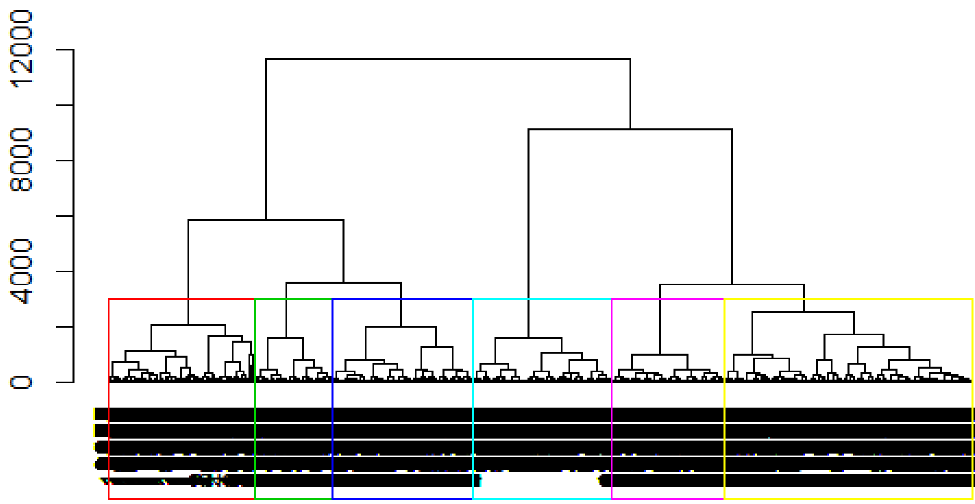
There is no absolute right from wrong in choosing the distance method. So is the linkage method. Euclidean distance is the basis of Ward's method, which implies a specific archetype of a cluster compared to other linkages of our dataset.[3] Our principal is to find the method that presents most evenly

---

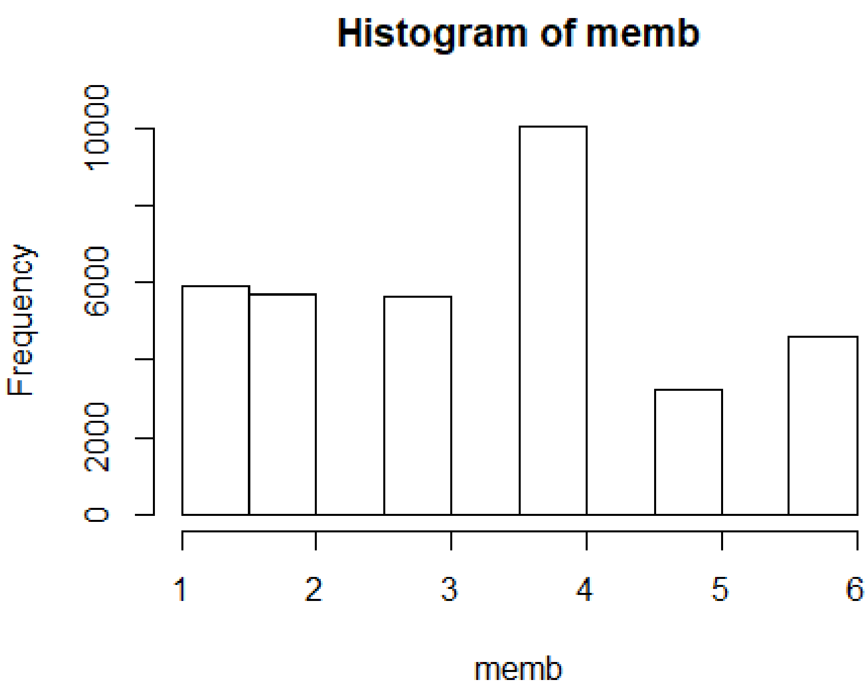
[2] https://www.quora.com/Why-do-people-use-Euclidean-distance-instead-of-Manhattan-distance-or-some-other-norm-metric

[3] https://stats.stackexchange.com/questions/195446/choosing-the-right-linkage-method-for-hierarchical-clustering

distributed clusters and long enough branch lines given that agglomerative cluster is a bottom-to-top process.

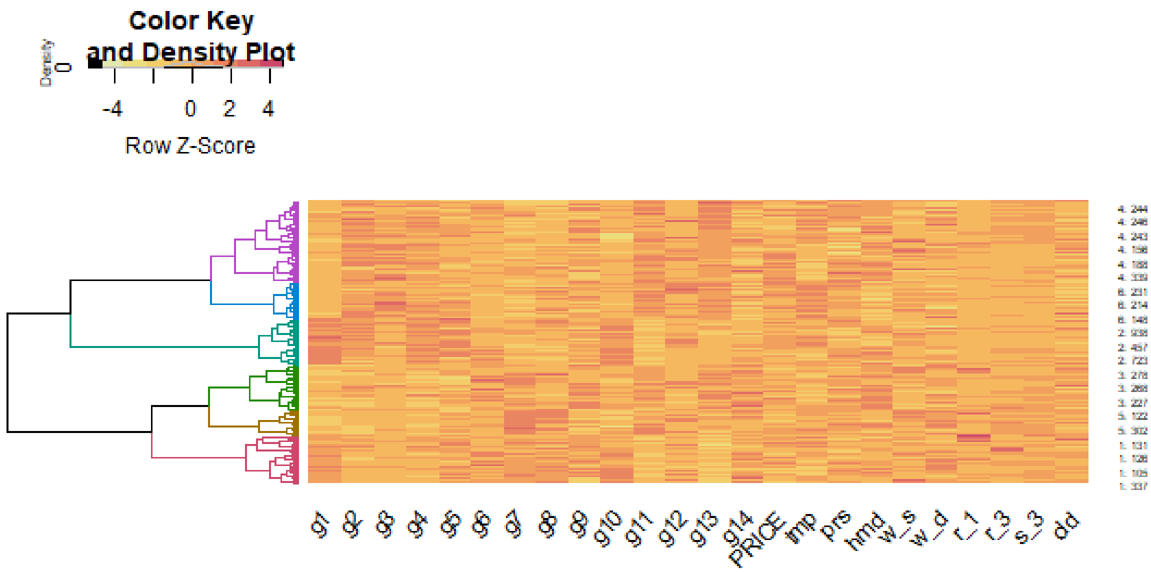


Plot 4 *Dendrogram of Hierarchical Clustering (k=6)*



Plot 5 *Distribution of Hierarchical Clustering (k=6)*

We must admit that the optimal number of clusters is somehow subjective and depends on the method used for measuring similarities and the parameters used for partitioning. From observations and tries of k ranging from 4 to 6, we think k=6 is the best choice.



Plot 6 *Heat map of Hierarchical Clustering (k=6)*

We can interpret this plot in the same way as K Means heat map. The redder of the color code, the higher the value and the yellower, the lower. Let's pick the pink group for an example, we can see an obviously redder region in features of g7 *(#generation hydro run-of-river and poundage*), g8 *(#generation hydro water reservoir*), and g10 *(#generation other)* and yellower region in features of g2 (#generation fossil brown coal/lignite) and tmp (#Temperature). Please understand that we are including too many variables, it is impossible to induce each cluster within single two-dimensional standard. However, we can put it into a multiple regression to do some feature engineering to reflect the clusters' reliability.

PART IV Feature Engineering with K Means

We paste K Means Clustering and Hierarchical Clustering Methods to original dataset for further rationalization

First, let's review the original multiple regression. Several important indices we should remember: Adjusted R^2 0.912 – the multiple regression can explain 91.2% of the observations; $RMSE$ 1355 - the differences between values (sample or population values) predicted by a model or an estimator and the values observed (Reasonable as our predictor is beyond the level of 20,000); vif <5 – there is no perfect multicollinearity in the multiple regression; p <0.05 – all variables are statistically significant.

```
Residual standard error: 1360 on 28030 degrees of freedom
Multiple R-squared:  0.912,     Adjusted R-squared:  0.912
F-statistic: 1.46e+04 on 20 and 28030 DF,  p-value: <0.0000000000000002

                      ME RMSE  MAE   MPE MAPE
Test set 0.0000000000001 1355 1058 -0.23  3.7
         ME RMSE  MAE   MPE MAPE
Test set 14 1288 1021 -0.17  3.6
```

Table 1 *Multiple Regression Statistics*

```
 g1  g2  g3  g4  g5  g6  g7  g8  g9 g11 g12 g13 g14 tmp prs hmd w_s w_d
2.5 2.9 2.2 3.9 1.9 1.8 3.1 2.6 1.1 2.7 2.5 2.2 1.9 2.3 1.2 3.1 1.7 1.3
s_3  t2
1.0 2.7
```

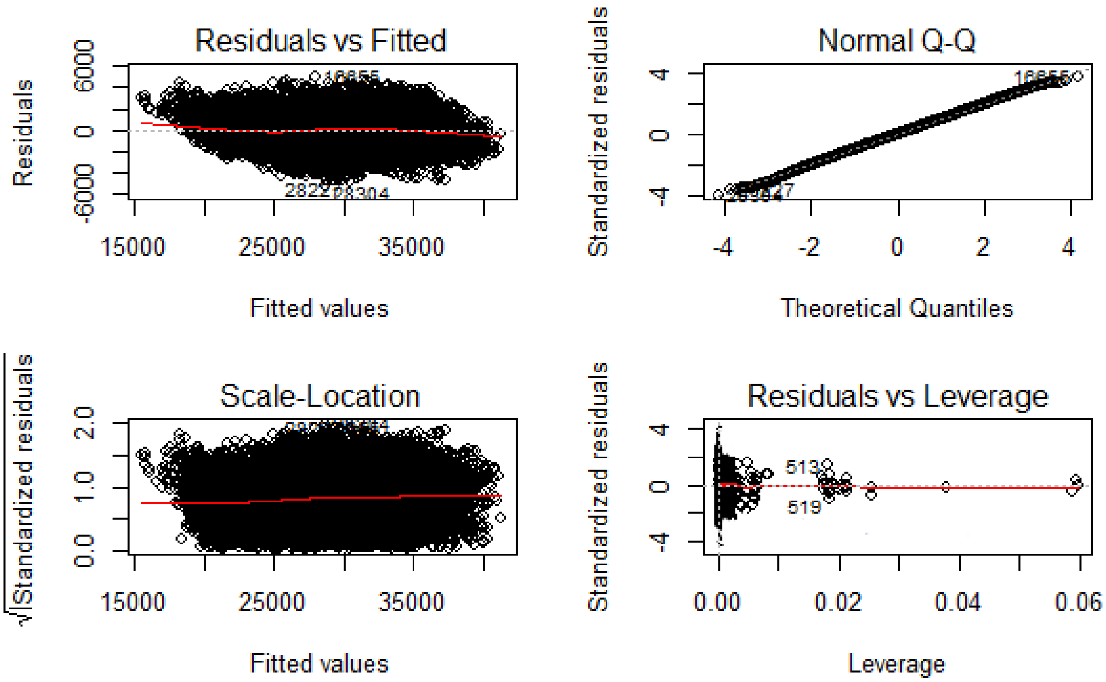
Table 2 *VIF coefficients*

```
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.361e+04  1.344e+03 -17.568  < 2e-16 ***
## g1           2.020e+00  1.463e-01  13.801  < 2e-16 ***
## g2           8.580e-01  3.784e-02  22.676  < 2e-16 ***
## g3           7.510e-01  5.270e-03 142.509  < 2e-16 ***
## g4           6.965e-01  7.924e-03  87.899  < 2e-16 ***
## g5           7.845e+00  2.049e-01  38.286  < 2e-16 ***
```

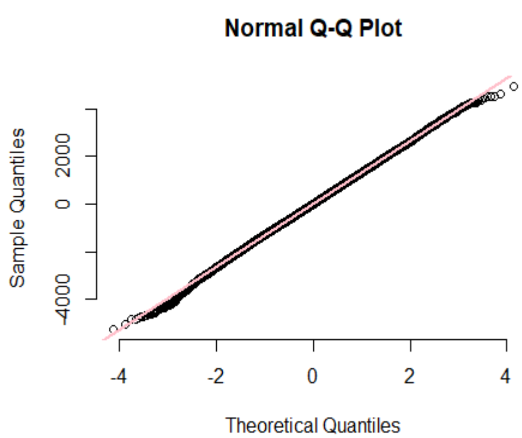
```
## g6          -1.316e+00  1.329e-02 -99.032  < 2e-16 ***
## g7           8.848e-01  3.470e-02  25.499  < 2e-16 ***
## g8           1.010e+00  6.987e-03 144.608  < 2e-16 ***
## g9           7.709e-01  9.915e-03  77.745  < 2e-16 ***
## g11          2.569e+01  9.204e-01  27.906  < 2e-16 ***
## g12          8.103e-01  7.418e-03 109.225  < 2e-16 ***
## g13          5.704e+00  2.312e-01  24.675  < 2e-16 ***
## g14          7.223e-01  3.426e-03 210.804  < 2e-16 ***
## tmp          1.909e+01  1.636e+00  11.666  < 2e-16 ***
## prs          1.895e+01  1.072e+00  17.670  < 2e-16 ***
## hmd          9.137e+00  9.325e-01   9.798  < 2e-16 ***
## w_s          3.393e+01  7.603e+00   4.463 8.11e-06 ***
## w_d         -8.298e-01  1.551e-01  -5.350 8.88e-08 ***
## s_3          1.009e+03  2.302e+02   4.384 1.17e-05 ***
## t2          -5.559e+02  2.591e+01 -21.456  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

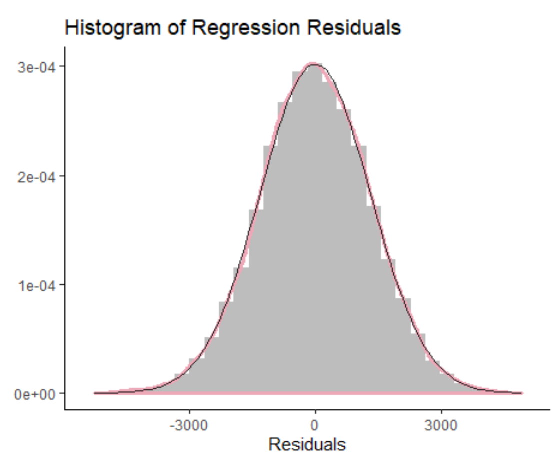
Table 3 *Multiple Regression Coefficients Significance Level*

Then, look at the plots of out prediction model. The Residuals vs. Fitted plot shows an almost flat line overlapping with $y = 0$ and dots are distributed randomly, meaning that the model fits observations quite well; The Normal Q-Q plot fits $y = x$, meaning that two probability distributions(predicted and observed) align with each other; The Residuals vs. Leverage plot has a fitted line overlapping with $y = 0$ and no cook's distance in it, meaning that there is no influential cases, that is, we have removed outliers to an appropriate and acceptable level; The Residual Histogram is distributed almost perfect in accordance with normal distribution line, which is an outcome we are very happy to see.



Plot 7 *Residual Combo of Original Model (Without Outliers)*





Left_Plot 8 *QQ Plot of Original Model (Without Outliers)*

Right_Plot 9 *Residual histogram vs. Normal Distribution Line (Without outliers)*

Then, how is our new multiple regression with clusters? We will look at the multiple regression with K Means Clustering first. Adjusted R^2 is increased to 0.922 – not so meaningful as we already have too many variables; $RMSE$ decreased to 1265 – the model predicts/fits better; vif <5 – there is no perfect multicollinearity in the multiple regression; p <0.05 – K Means does help in improving our multiple regression model and they are statistically significant.

How to interpret the categories, or K Means, then? So, the categorical variable is automatically transferred into $n - 1$ variables and cluster 1 (kmeans1) is set as the basis. Take kmeans2 for an example, if the observation belonged to cluster 2, the demand would be around 864.5 lower than cluster 1. This is especially useful when we are doing market segmentation.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2.209e+04  1.314e+03 -16.814  < 2e-16 ***
g1           4.001e+00  1.665e-01  24.028  < 2e-16 ***
g2           9.087e-01  3.766e-02  24.131  < 2e-16 ***
g3           6.909e-01  6.105e-03 113.173  < 2e-16 ***
g4           7.259e-01  8.027e-03  90.440  < 2e-16 ***
g5           1.015e+01  2.115e-01  47.997  < 2e-16 ***
g6          -1.437e+00  1.660e-02 -86.548  < 2e-16 ***
g7           1.088e+00  3.486e-02  31.197  < 2e-16 ***
g8           1.017e+00  7.024e-03 144.756  < 2e-16 ***
g9           6.974e-01  9.970e-03  69.949  < 2e-16 ***
g11          1.504e+01  9.750e-01  15.423  < 2e-16 ***
g12          7.537e-01  7.406e-03 101.771  < 2e-16 ***
g13          5.110e+00  2.466e-01  20.721  < 2e-16 ***
g14          7.035e-01  3.414e-03 206.044  < 2e-16 ***
tmp          1.262e+01  1.622e+00   7.780 7.50e-15 ***
prs          1.933e+01  1.045e+00  18.497  < 2e-16 ***
hmd          1.142e+01  9.231e-01  12.372  < 2e-16 ***
w_s          3.276e+01  7.446e+00   4.400 1.09e-05 ***
w_d         -9.809e-01  1.508e-01  -6.504 7.96e-11 ***
s_3          6.786e+02  2.075e+02   3.270  0.00108 ** 
p_h          6.294e+02  2.523e+01  24.947  < 2e-16 ***
kmeans2     -8.645e+02  5.017e+01 -17.232  < 2e-16 ***
kmeans3     -5.395e+02  5.182e+01 -10.411  < 2e-16 ***
kmeans4     -1.422e+03  5.597e+01 -25.415  < 2e-16 ***
kmeans5     -2.895e+02  4.945e+01  -5.855 4.81e-09 ***
kmeans6     -2.400e+02  5.877e+01  -4.085 4.43e-05 ***
kmeans7     -1.036e+03  1.285e+02  -8.061 7.89e-16 ***
kmeans8      2.492e+02  5.514e+01   4.519 6.24e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 1282 on 27981 degrees of freedom
Multiple R-squared:  0.9217,	Adjusted R-squared:  0.9216 
F-statistic: 1.22e+04 on 27 and 27981 DF,  p-value: < 2.2e-16
```

Table 4 *Multiple Regression with K Means Coefficients Significance Level*

	GVIF	Df	GVIF^(1/(2*Df))
g1	3.413024	1	1.847437
g2	3.044261	1	1.744781
g3	3.078002	1	1.754424
g4	4.224684	1	2.055404
g5	2.097715	1	1.448349
g6	2.954178	1	1.718772
g7	3.327145	1	1.824046
g8	2.838813	1	1.684878
g9	1.190358	1	1.091035
g11	3.199804	1	1.788800
g12	2.640186	1	1.624865
g13	2.598250	1	1.611909
g14	2.057634	1	1.434446
tmp	2.357123	1	1.535292
prs	1.242419	1	1.114638
hmd	3.176456	1	1.782262
w_s	1.702097	1	1.304644
w_d	1.270968	1	1.127372
s_3	1.335560	1	1.155665
p_h	2.713964	1	1.647411
kmeans	184.105567	7	1.451411

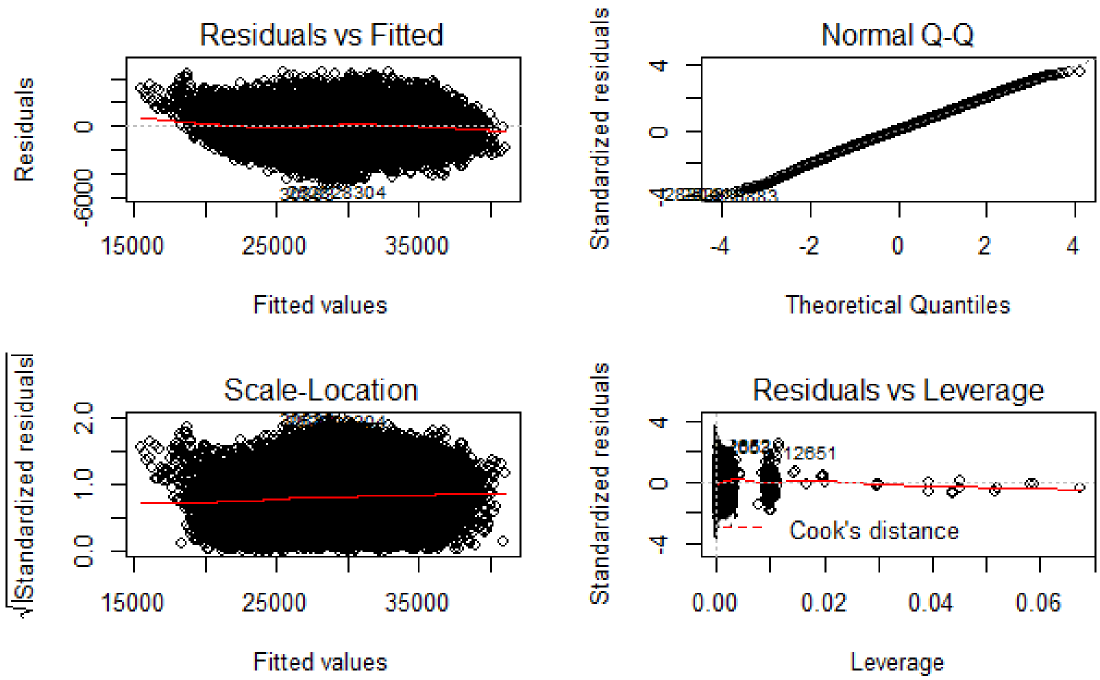
Table 5 *Multiple Regression with K Means VIF Coefficients*

	ME	RMSE	MAE	MPE	MAPE
Test set	1.850572e-14	1280.886	1012.375	-0.2016624	3.592419

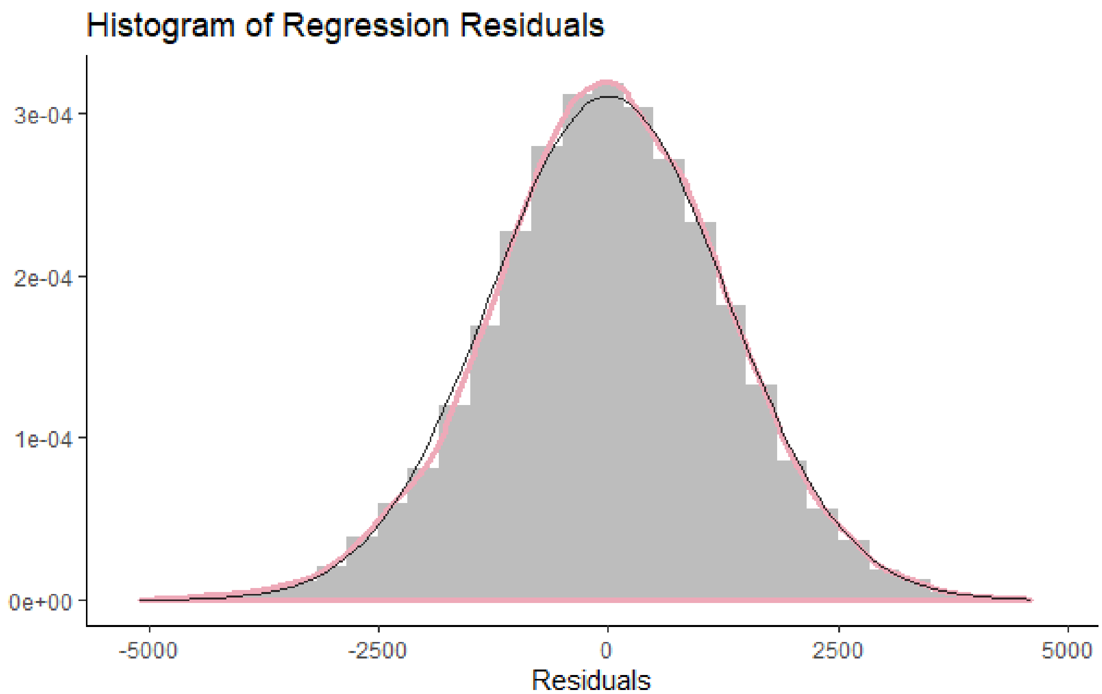
	ME	RMSE	MAE	MPE	MAPE
Test set	-10.8252	1265.276	994.9398	-0.2463268	3.53089

Table 6 *Multiple Regression with K Means Accuracy*

We are not going to repeat the plots as we don’t see any unusual signs.



Plot 10 *Residual Combo of Multiple Regression with K Means*



Plot 11 *Residual histogram vs. Normal Distribution Line (Multiple Regression with K Means)*

PART V Feature Engineering with Hierarchical Clustering

Then, we will look at the multiple regression with Hierarchical Clustering. Adjusted R^2 is increased to 0.920 – not so meaningful as we already have too many variables; $RMSE$ decreased to 1277 – the model predicts/fits slightly better; vif <5 – there is no perfect multicollinearity in the multiple regression; p <0.05 – Hierarchical Clustering does help in improving our multiple regression model and they are statistically significant. Still, there is no overfitting issue. The coefficient interpretation is the same as is stated above.

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -2.373e+04  1.326e+03 -17.896  < 2e-16 ***
g1             4.173e+00  1.764e-01  23.650  < 2e-16 ***
g2             8.355e-01  3.775e-02  22.130  < 2e-16 ***
g3             7.039e-01  5.427e-03 129.697  < 2e-16 ***
g4             7.299e-01  7.985e-03  91.400  < 2e-16 ***
g5             9.412e+00  2.089e-01  45.062  < 2e-16 ***
g6            -1.294e+00  1.333e-02 -97.095  < 2e-16 ***
g7             1.102e+00  3.588e-02  30.722  < 2e-16 ***
g8             1.015e+00  6.879e-03 147.515  < 2e-16 ***
g9             7.013e-01  1.026e-02  68.314  < 2e-16 ***
g11            1.418e+01  9.883e-01  14.347  < 2e-16 ***
g12            7.743e-01  7.322e-03 105.757  < 2e-16 ***
g13            3.748e+00  2.620e-01  14.304  < 2e-16 ***
g14            6.995e-01  3.459e-03 202.243  < 2e-16 ***
tmp            1.557e+01  1.641e+00   9.488  < 2e-16 ***
prs            1.974e+01  1.054e+00  18.732  < 2e-16 ***
hmd            9.952e+00  9.199e-01  10.818  < 2e-16 ***
w_s            4.543e+01  7.445e+00   6.102 1.06e-09 ***
w_d           -7.286e-01  1.518e-01  -4.799 1.60e-06 ***
s_3            5.128e+02  1.847e+02   2.776 0.005499 ** 
p_h            5.958e+02  2.541e+01  23.447  < 2e-16 ***
hierarchical2 -5.177e+02  3.615e+01 -14.320  < 2e-16 ***
hierarchical3  7.284e+02  4.626e+01  15.747  < 2e-16 ***
hierarchical4  7.161e+02  4.643e+01  15.422  < 2e-16 ***
hierarchical5  1.909e+02  5.026e+01   3.799 0.000145 ***
hierarchical6  1.111e+03  4.965e+01  22.382  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1292 on 27983 degrees of freedom
Multiple R-squared:  0.9205,	Adjusted R-squared:  0.9204 
F-statistic: 1.296e+04 on 25 and 27983 DF,  p-value: < 2.2e-16
```

Table 7 *Multiple Regression with Hierarchical Clustering Significance Level*

	GVIF	Df	GVIF^(1/(2*Df))
g1	3.772139	1	1.942200
g2	3.011975	1	1.735504
g3	2.394460	1	1.547404
g4	4.115757	1	2.028733
g5	2.014468	1	1.419319
g6	1.874945	1	1.369286
g7	3.468285	1	1.862333
g8	2.680266	1	1.637152
g9	1.242018	1	1.114459
g11	3.236090	1	1.798914
g12	2.540064	1	1.593758
g13	2.887429	1	1.699244
g14	2.078648	1	1.441752
tmp	2.376471	1	1.541581
prs	1.243160	1	1.114971
hmd	3.104964	1	1.762091
w_s	1.675258	1	1.294318
w_d	1.267809	1	1.125970
s_3	1.041578	1	1.020577
p_h	2.710266	1	1.646289
hierarchical	39.921605	5	1.445842

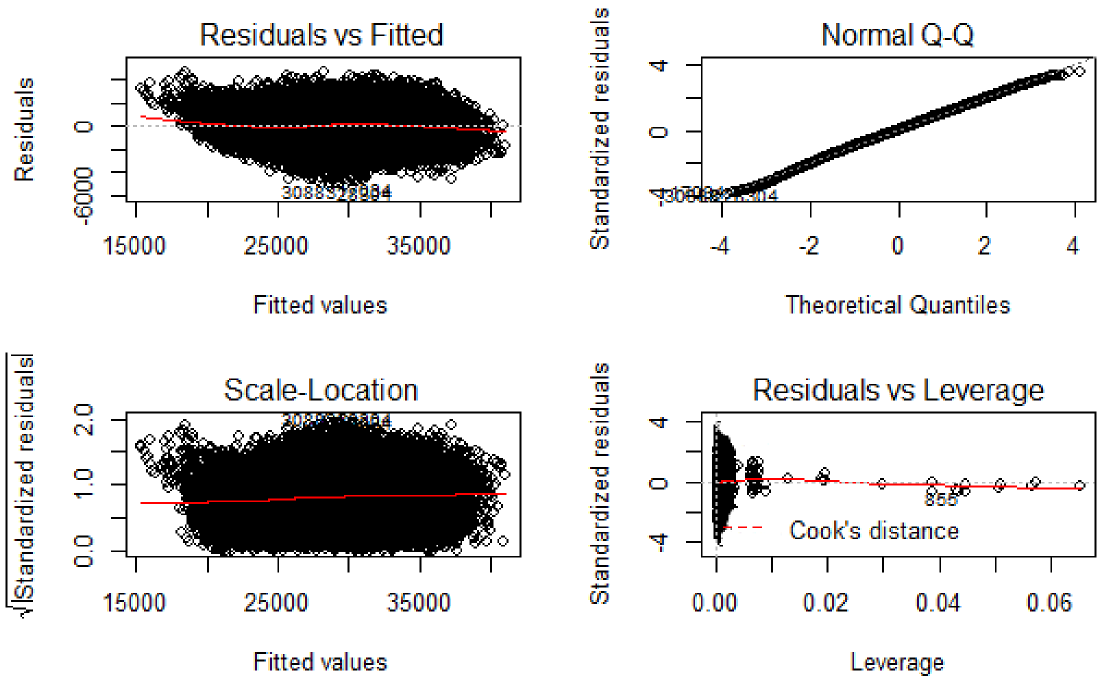
Table 8 *Multiple Regression with Hierarchical Clustering VIF Coefficients*

	ME	RMSE	MAE	MPE	MAPE
Test set	-1.249953e-14	1291.092	1025.62	-0.2045925	3.639147

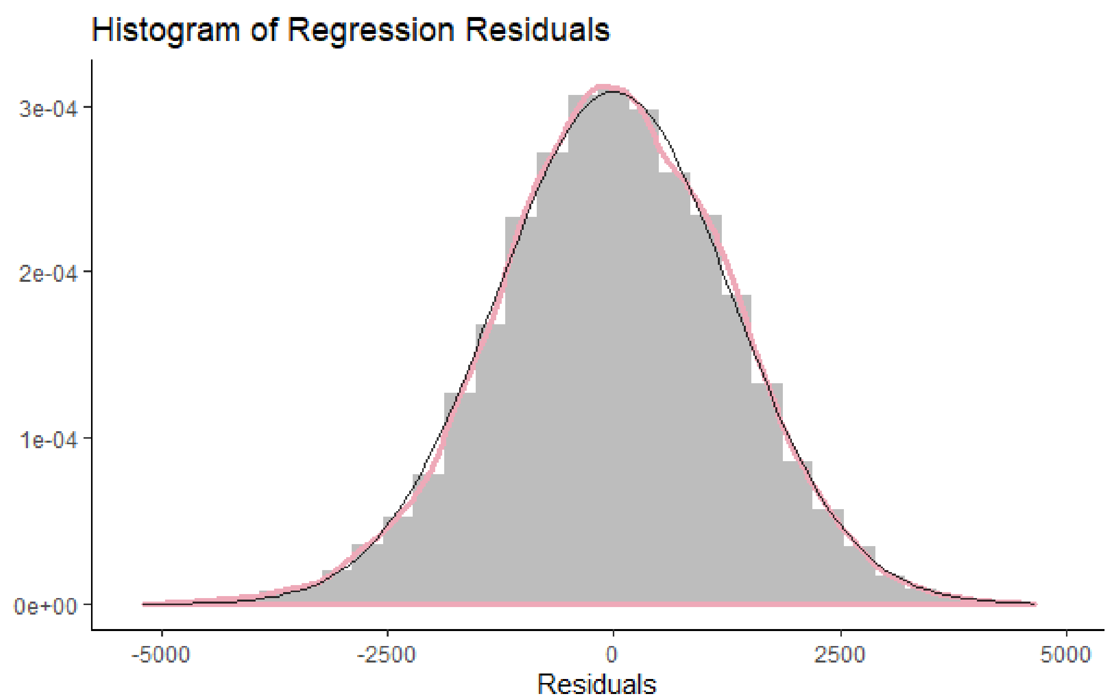
	ME	RMSE	MAE	MPE	MAPE
Test set	-12.10338	1277.22	1007.101	-0.252777	3.573953

Table 9 *Multiple Regression with Hierarchical Clustering Accuracy*

There is a tiny slight issue with Residuals vs. Leverage plot. We see some of the values are beyond cook's distance as typically, cook's distance ranges from -3 to 3. We think it might be because we should try another k in hierarchical clustering, like 4 or 3. Also, perhaps we should remove more data in this case, or we should consider a non-linear regression.



Plot 12 *Residual Combo of Multiple Regression with Hierarchical Clustering*



Plot 11 *Residual histogram vs. Normal Distribution Line (Multiple Regression with Hierarchical Clustering)*