MARMARA UNIVERSITY
FACULTY OF ENGINEERING
Moving Towards Excellence in Education and Research

CSE4062 2020 SPRING PROJECT

ANALYSIS OF ELECTRIC PRODUCTION OF A WIND TURBINE

Delivery 4: Predictive Analysis

Group: 8

Name	Department	ID Number	Mail Address
Furkan Can Ercan	Industrial Engineering	150316044	furkanercan98@gmail.com
Muhammed Avcı	Mechanical Engineering	150414007	muhammedavci96@gmail.com
Yasin Gök	Computer Engineering	150115058	ygok.96@gmail.com

5) Clustering Analysis

In this part of our project we made analysis about our dataset via clustering techniques. As a technique of clustering we used K means algorithm. We also tried to work with DBSCAN and Agglomerative methods but due to size of our dataset and due to lack of memory and computing power in our computers, we could not run DBSCAN and Agglomerative methods. In our experiments it took about a 1 hour and still the DBSCAN algorithm did not converge. So we decided to stick with K means algorithm.

While using K-means algorithm, to choose a good number of K, we used Elbow method. The elbow method is a very simple and common method; it experiments with different values of k and plots the sum of squared errors. In the plot, there appears a shape similar to elbow. The place where the plot takes an elbow shape shows the place to choose a good K.

Its logic is simple, normally as value of K increases, sum of squared error decreases but if we increase the value of K, we obtain smaller clusters and smaller clusters may not give us the information we wanted. In the plot, as K increases, sum of squared error decreases but this decrease is a drastic decrease at first and as K further increases the decrease in the sum of squared error starts to become smaller and smaller. In the place where the plot takes, the shape of an elbow it describes that the further increase of the value of K will not contribute the decrease of sum of squared error. So we choose a value in the place where plot takes the shape of an elbow. It is of course not guaranteed to have an optimum value of K but it is a good heuristic way to choose a good enough value of K for our experiments.

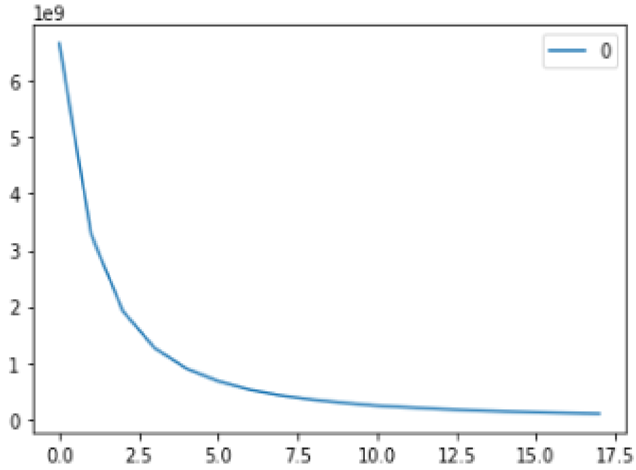


Figure 5.1: K-SSE plot to use elbow method

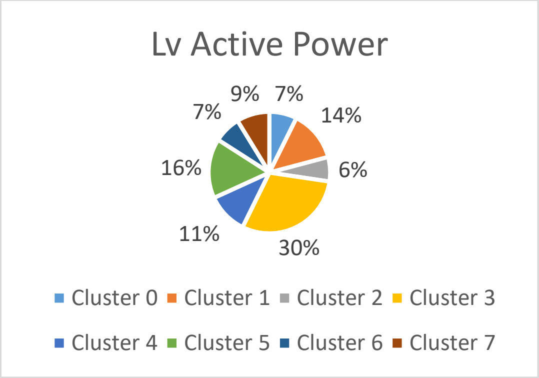
After we decided to use the elbow method we continued with our choices for features to examine with clustering analysis. Since we observed our data in exploratory analysis part, it was easier to choose this time. We discarded the variables such as Date, Time, Day/Night and Month because this variables already described in groups. We also discarded some other variables which has very small standard deviation for example RainMM feature is nearly always 0 and Pressure is nearly always takes 1020 value. For the rest of our variables we first determined a good number of K via elbow method and then made our cluster analysis.

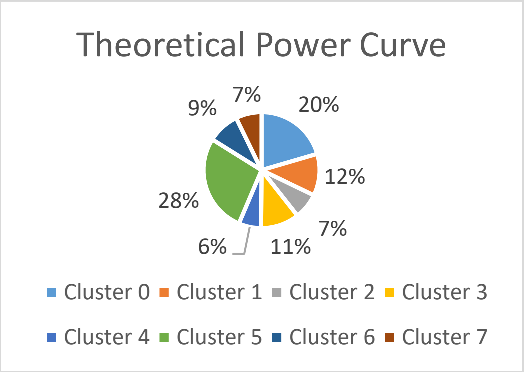
Feature Name	Type	Overall Avg	SSE	Cluster0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
Lv Active Power (kW)	Float64	1367.691704	666440942.4969716	3548	6543	3140	14462	5270	7687	3463	4239
Wind Speed (m/s)	Float64	7.584808	22327.138177423283	8302	7518	2786	8703	5445	1436	7473	6650
Theoretical Power Curve (kW)	Float64	1504.010263	639416324.5652404	9893	5642	3490	5192	3031	13280	4304	3481
Wind Direction	Float64	124.291200	5638038.056385048	8060	10214	2413	7838	3222	9485	2137	4944
Temp	Int 64	16.340095	92838.9379	8561	8291	7601	7227	9301	7317	-	-
Sun Hour	Float64	10.529758	12397.9210	17471	7022	2735	11416	9654	-	-	-
Dew Point	int64	10.836007	30150.3841	5685	4183	6493	9971	6258	8331	1112	5929
Wind Chillc	int64	16.457682	56330.1173	6153	5773	5931	4437	6049	7110	6246	5962
Wind Gust	Int64	14.482996	153801.14222590494	12905	4724	9272	9801	1326	10285	-	-
Humidity	Int64	69.554261	392136.1781263390	12149	8380	8577	4169	8153	6885	-	-
Density	Float64	1.222541	1.5605874423794044	47757	556						

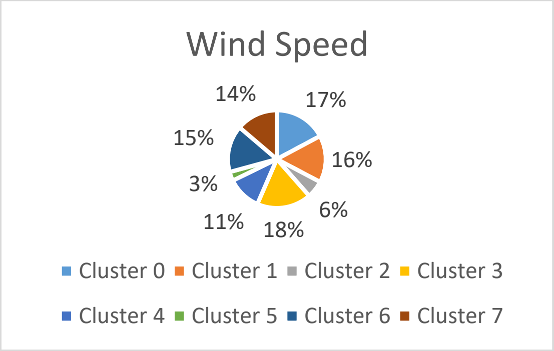
Figure 5.2: Cluster Analysis table of features

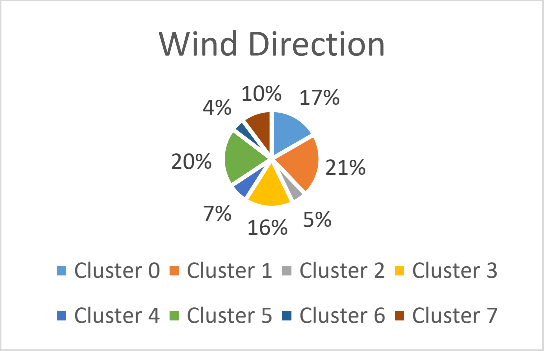
Feature Name	Description
Lv Active Power (kW)	K-means algorithm. Trying different number of K values and using elbow method, select k=8, for Active Power It is the electric production of Wind turbine in kw/h
Wind Speed (m/s)	K-means algorithm. Trying different number of K values and using elbow method, select k=8,for Wind Speed It is the speed of the wind for turbine to turn, in terms of m/s
Theoretical Power Curve (kW)	K-means algorithm. Trying different number of K values and using elbow method, select k=8, for Theoretical Curve It is a calculation done with the values provided by of the manufacturer of the plant. It calculates the theoretical value of the electric production. In terms of kw/h
Wind Direction	K-means algorithm. Trying different number of K values and using elbow method, select k=8, for Wind Direction The direction (angle) of the wind in terms of degrees.
Temp	K-means algorithm. Trying different number of K values and using elbow method, select k=6, for Temp Temperature value in terms of Celsius degrees.
Sun Hour	K-means algorithm. Trying different number of K values and using elbow method, select k=5,for Sun Hour The number of hours from rise of the sun until it sets
Dew Point	K-means algorithm. Trying different number of K values and using elbow method, select k=8, for Dew Point The temperature which allows dews to be created.
Wind Chill	K-means algorithm. Trying different number of K values and using elbow method, select k=8,for Wind Chill The chilling effect which wind causes as in temperature, Celsius degrees.
Wind Gust	K-means algorithm. Trying different number of K values and using elbow method, select k=6,for Wind Gust Sudden increases in the wind speed in terms of m/s
Humidity	K-means algorithm. Trying different number of K values and using elbow method, select k=6,for Humidity Humidity (amount of vaporized water) as percentage.
Density	K-means algorithm. Trying different number of K values and using elbow method, select k=2,for Density Density of the air calculated via pressure and temperature.

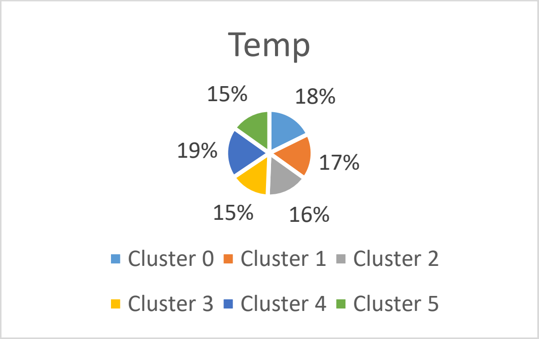
Figure 5.3: Description of analysed features and used methods

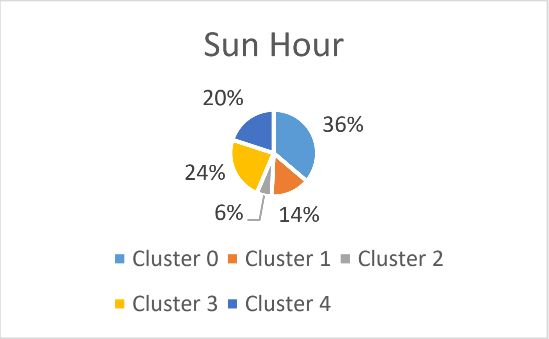

Lv Active Power
9%
7%
7%
14%
6%
16%
11%
30%
■ Cluster 0 ■ Cluster 1 ■ Cluster 2 ■ Cluster 3
■ Cluster 4 ■ Cluster 5 ■ Cluster 6 ■ Cluster 7

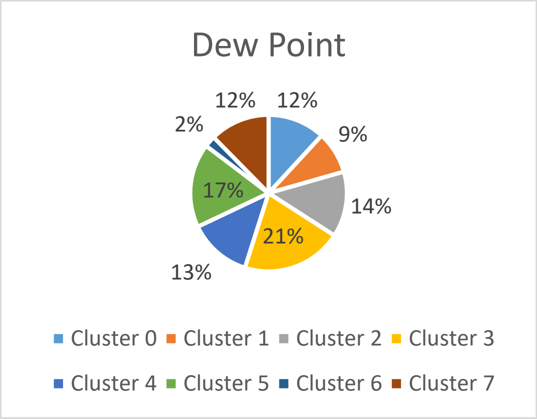

Theoretical Power Curve
7%
9%
20%
12%
28%
7%
6%
11%
■ Cluster 0 ■ Cluster 1 ■ Cluster 2 ■ Cluster 3
■ Cluster 4 ■ Cluster 5 ■ Cluster 6 ■ Cluster 7

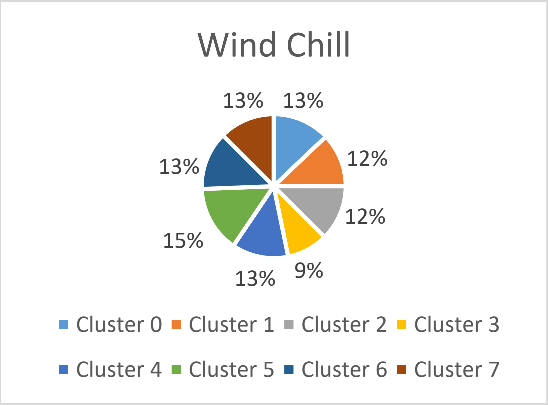

Wind Speed
14%
17%
15%
16%
3%
6%
11%
18%
■ Cluster 0 ■ Cluster 1 ■ Cluster 2 ■ Cluster 3
■ Cluster 4 ■ Cluster 5 ■ Cluster 6 ■ Cluster 7


Wind Direction
10%
4%
17%
20%
21%
7%
16%
5%
■ Cluster 0 ■ Cluster 1 ■ Cluster 2 ■ Cluster 3
■ Cluster 4 ■ Cluster 5 ■ Cluster 6 ■ Cluster 7


Temp
15%
18%
19%
17%
15%
16%
■ Cluster 0 ■ Cluster 1 ■ Cluster 2
■ Cluster 3 ■ Cluster 4 ■ Cluster 5


Sun Hour
20%
36%
24%
6%
14%
■ Cluster 0 ■ Cluster 1 ■ Cluster 2
■ Cluster 3 ■ Cluster 4


Dew Point
12%
12%
2%
9%
17%
14%
21%
13%
■ Cluster 0 ■ Cluster 1 ■ Cluster 2 ■ Cluster 3
■ Cluster 4 ■ Cluster 5 ■ Cluster 6 ■ Cluster 7


Wind Chill
13%
13%
12%
13%
12%
15%
13%
9%
■ Cluster 0 ■ Cluster 1 ■ Cluster 2 ■ Cluster 3
■ Cluster 4 ■ Cluster 5 ■ Cluster 6 ■ Cluster 7

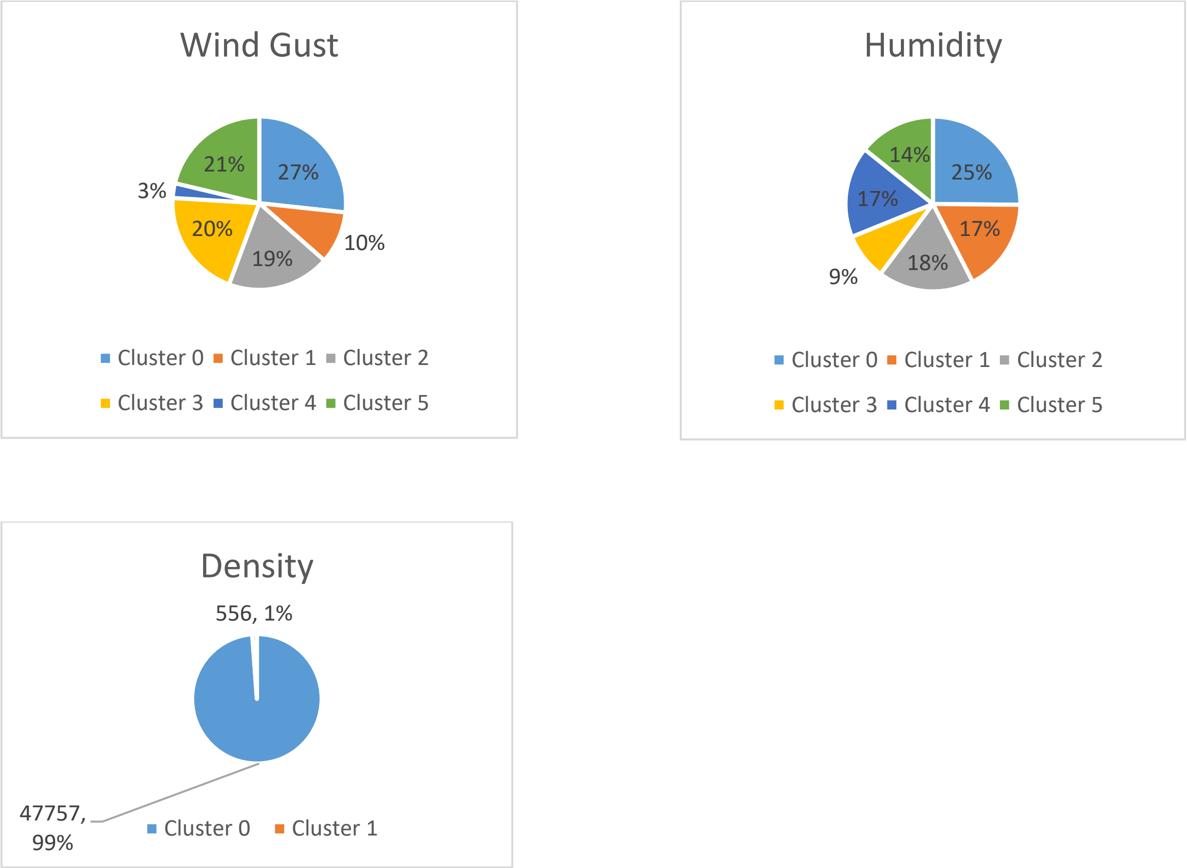
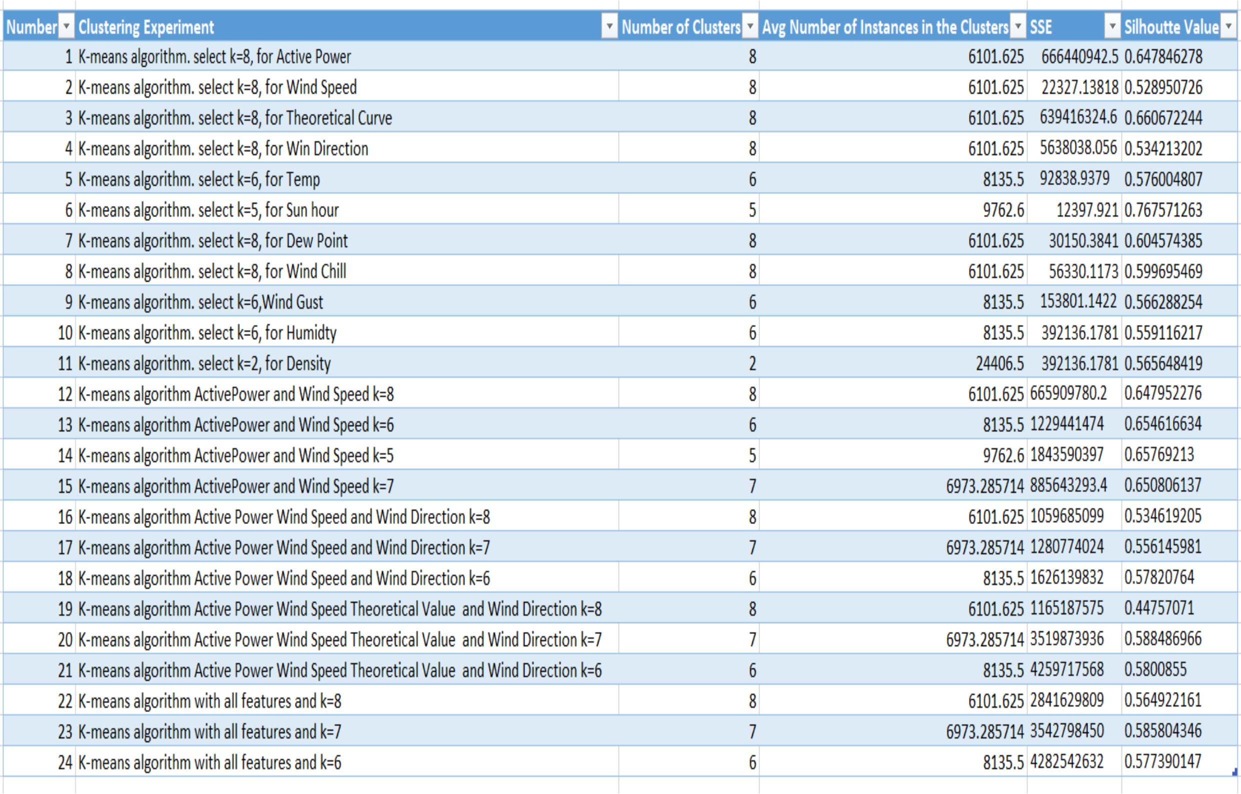


Figure 5.4 : Pie charts for all of the features

By looking at pie charts we realized that for most of our features there are 2-3 clusters that includes 15% or more of the data and rest of the clusters are sharing the smaller parts. This is even more present in the examples with lesser K values, on those ones, there is one or two clusters that includes 20% or more of the data and again rest of the clusters shares the remaining part. But in these examples and in pie charts we did not looked at the silhouette score. It is another scores which evaluates whether the data is in the right cluster. It measures the inter-cluster and intra-cluster differences. It looks for a data point that if there are another cluster nearby and if there is whether this data point should be in its cluster or should be in that close cluster. It is important for clusters to minimize the intra-cluster distances and maximize the inter-cluster distances. While SSE (Sum of Squared Error) measures the coherence in one cluster, which means intra-cluster distances, silhouette score also account for inter cluster distance.

After making analysis about features individually, we made experiments with different combination of features to gain a more general insight about our data. We made these experiments with our 4 most important features and with all features included. This time we had the chance to see the alteration on the silhouette score and SSE as K changes. The results are on the table below.



Number	Clustering Experiment	Number of Clusters	Avg Number of Instances in the Clusters	SSE	Silhoutte Value
1	K-means algorithm. select k=8, for Active Power	8	6101.625	666440942.5	0.647846278
2	K-means algorithm. select k=8, for Wind Speed	8	6101.625	22327.13818	0.528950726
3	K-means algorithm. select k=8, for Theoretical Curve	8	6101.625	639416324.6	0.660672244
4	K-means algorithm. select k=8, for Win Direction	8	6101.625	5638038.056	0.534213202
5	K-means algorithm. select k=6, for Temp	6	8135.5	92838.9379	0.576004807
6	K-means algorithm. select k=5, for Sun hour	5	9762.6	12397.921	0.767571263
7	K-means algorithm. select k=8, for Dew Point	8	6101.625	30150.3841	0.604574385
8	K-means algorithm. select k=8, for Wind Chill	8	6101.625	56330.1173	0.599695469
9	K-means algorithm. select k=6,Wind Gust	6	8135.5	153801.1422	0.566288254
10	K-means algorithm. select k=6, for Humidty	6	8135.5	392136.1781	0.559116217
11	K-means algorithm. select k=2, for Density	2	24406.5	392136.1781	0.565648419
12	K-means algorithm ActivePower and Wind Speed k=8	8	6101.625	665909780.2	0.647952276
13	K-means algorithm ActivePower and Wind Speed k=6	6	8135.5	1229441474	0.654616634
14	K-means algorithm ActivePower and Wind Speed k=5	5	9762.6	1843590397	0.65769213
15	K-means algorithm ActivePower and Wind Speed k=7	7	6973.285714	885643293.4	0.650806137
16	K-means algorithm Active Power Wind Speed and Wind Direction k=8	8	6101.625	1059685099	0.534619205
17	K-means algorithm Active Power Wind Speed and Wind Direction k=7	7	6973.285714	1280774024	0.556145981
18	K-means algorithm Active Power Wind Speed and Wind Direction k=6	6	8135.5	1626139832	0.57820764
19	K-means algorithm Active Power Wind Speed Theoretical Value and Wind Direction k=8	8	6101.625	1165187575	0.44757071
20	K-means algorithm Active Power Wind Speed Theoretical Value and Wind Direction k=7	7	6973.285714	3519873936	0.588486966
21	K-means algorithm Active Power Wind Speed Theoretical Value and Wind Direction k=6	6	8135.5	4259717568	0.5800855
22	K-means algorithm with all features and k=8	8	6101.625	2841629809	0.564922161
23	K-means algorithm with all features and k=7	7	6973.285714	3542798450	0.585804346
24	K-means algorithm with all features and k=6	6	8135.5	4282542632	0.577390147

Figure5.5: Table for experiments

After we made all these experiments we had some insights about clustering and about our dataset. As we mentioned before while making these experiments we also calculated silhouette score. While making these experiments at first we chose our K values with elbow method, which means we only looked for sum of squared errors while choosing our K value. After calculating silhouette score we saw that our scores are not bad but also not the best. So to get better scores we played with number of K(s) and saw that when we decrease the value of K generally we increased our silhouette scores. So these experiments taught us that we should not focus only for one measure. We learned from this experiments that we should balance for both lesser SSE values and higher silhouette scores. Because silhouette scores takes the inter cluster measures into account. This is also important for having better clusters. We also discovered that as we increase the number of features the sum of squared error is also increases but silhouette scores does not always increase or decrease with the number of features. Special to our dataset and similar datasets, since we have big numbers as features our sum of squared errors tend to be huge numbers and it gets hard to evaluate. So it may be wiser to scale the data for situations similar to this. It is also harder to visualize when we have features more than 3. So we should decrease our features, we should pick the most important of them. We may do this via using principal component analysis or some other methods. Especially for K means to decrease the number of features will have an impactful effect on the sum of squared error.