Kennesaw State University
IS 8935 Business Intelligence: Traditional & Big Data Analytics
Dr. Reza Vaezi
Assignment 5
February 12, 2019

Household Power Consumption Clustering Analysis

By Karis Kim

Executive Summary

The goal of this report is to offer the electric utility company innovative power plans based on the analysis and offer business recommendations to increase its revenue. K-Means clustering and cluster distance performance analysis yielded 3 clusters, comprising of high, mid, and low power usage groups. The largest cluster was the low power usage group. The report recommends that geographical data on the highest power usage group be used to market referral incentives in the identified area to attract more customers from high power usage area.

CONTENTS

Contents....2
Business Understanding....3
How the business works....3
Assumptions....3
Data Understanding....3
Attribute Information....4
Attribute Explanation....4
Data Assumptions....5
Data Preparation....9
Modeling....11
k-Means....11
Cluster Distance Performance....13
Evaluation OF Findings....14
Summary of Findings....14
Observations on Usage....15
Busines Recommendations....16
References....16

BUSINESS UNDERSTANDING

HOW THE BUSINESS WORKS

Every state has a Public Utility Commissions (PUC) that regulates how much the power companies can charge customers and what their profit margin or Return on Equity (ROE) can be [2], so that utility companies can be held in check to ensure that the public can affordably and reliably access a basic modern necessity like electric power. So, electric utility companies offer a variety of pricing plans to attract as many customers as possible in a competitive market. Pricing plans include: fixed-rate plans (multi-year/12-month/short term), variable-rate plans, indexed rate plans, flat-rate plans, time-of-use plans, and others [1]. Pricing can be adjusted based on terms of the contract, the local distribution grid, and competitors' offers. In order for power companies to maximize revenue within the regulatory bounds, pricing determinations must be informed with intelligent analytics to balance the demands of both the consumers and the provider.

ASSUMPTIONS

This report is premised on the basic assumption that the electric utility company for which this data is applicable, is not the sole provider of power in this region and is engaged in competition with competitor power providers.

Since power companies are regulated by PUC [2], this report will adopt the assumption that the best model to secure revenue increase is to retain the greatest number of customers in the service region, and to do so with attractive pricing options. It is outside the scope of this report to consider other revenue increasing options such as investments, production cost control and leveraging alternative energy.

The goal of this report is to offer the electric utility company innovative power plans based on the analysis and offer business recommendations to increase its revenue.

DATA UNDERSTANDING

The Household Power Consumption data set contains measurements of household power consumptions over a specific, but unknown time period. There are 8 regular attributes and 2,077 examples (rows), where each example row is a measurement from a single household. All values are numerical data types, but 3 attributes are integer while the rest are real.

While the meta data shows that there are no missing values, a number of attributes such as Sub-metering_1, Sub_metering_2, and Sub_metering_3 show that a significant number of examples have zero as the recorded value (see figure 2). With no time dimension in the data set attributes, it is not possible to determine if the zero values correspond to certain time dimensions in which there was no power usage with the associated equipment, or if the zero values represent households that do not have a submeter installed to measure that particular equipment, or if a zero value is indeed a valid measurement for that attribute. Only 10 of 2,077 examples have a zero value recorded for all 3 submetering attributes. Since cluster analysis is sensitive to outliers, we must determine the course of action for examples with zero values.

A *Filter Examples* operator is employed in RapidMiner to see how many zero values exist (see figure 3 - 6). Sub_metering_3 attribute had 1141 rows with zero values. Sub_metering_2 had 1666 rows with zero values. Sub_metering_1 had 1974 rows with zero values. So, examples with zero values for the sub-metering attributes make up a significant portion of the data set and it does not appear to be a negligible or outlier value.

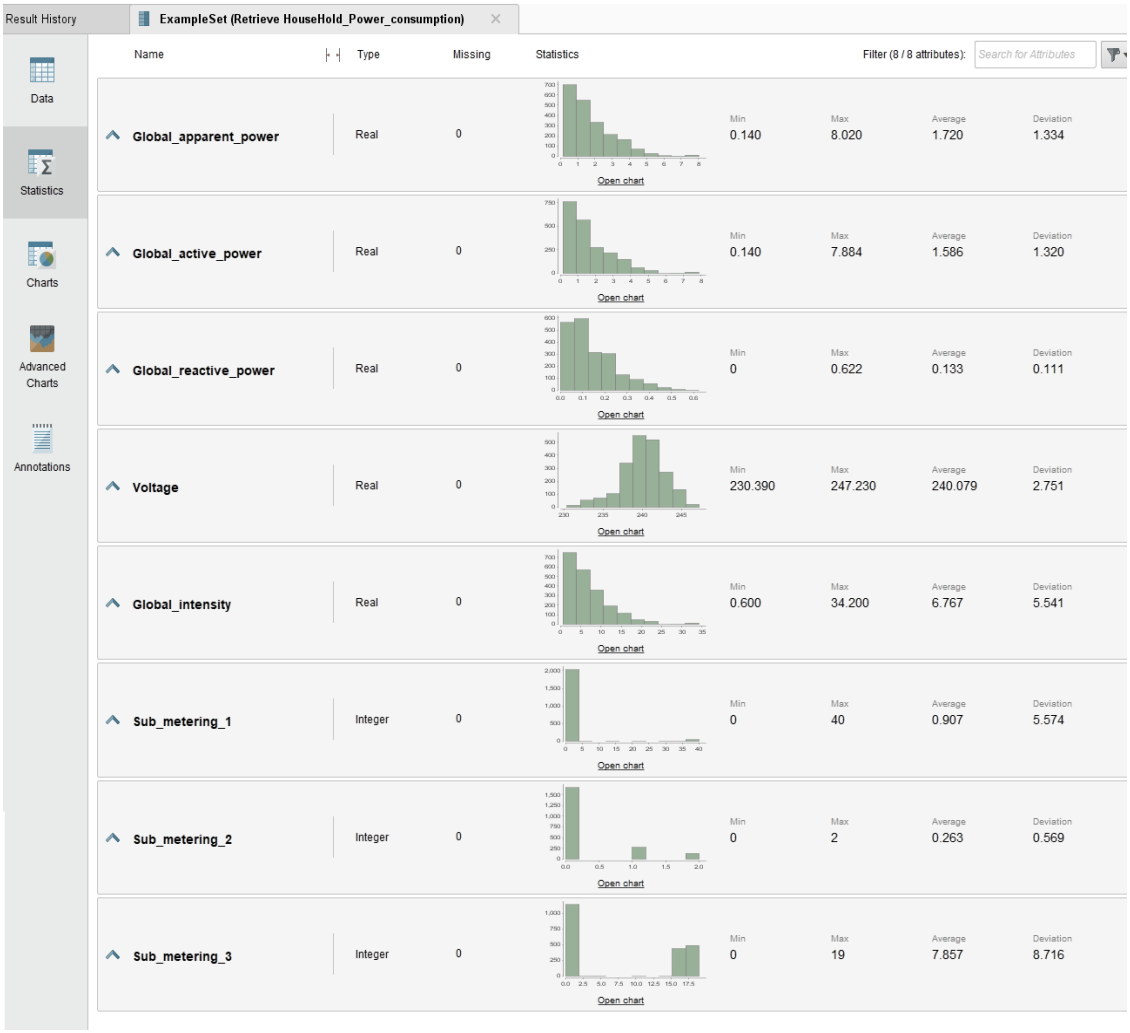
ATTRIBUTE INFORMATION

	Attribute	Description
1	Global_apparent_power	global_active_power + global_reactive_power
2	Global_active_power	household global minute-averaged active power (in kW)
3	Global_reactive_power	household global minute-averaged reactive power (in kW)
4	Voltage	minute-averaged voltage (in volt)
5	Global_intensity	household global minute-averaged current intensity (in amp)
6	Sub_metering_1	energy sub-metering No. 1 (in watt-hour of active energy). It corresponds to the kitchen, containing mainly a dishwasher, an oven and a microwave (hot plates are not electric but gas powered).
7	Sub_metering_2	energy sub-metering No. 2 (in watt-hour of active energy). It corresponds to the laundry room, containing a washing-machine, a tumble-drier, a refrigerator and a light.
8	Sub_metering_2	energy sub-metering No. 3 (in watt-hour of active energy). It corresponds to an electric water-heater and an air-conditioner.

ATTRIBUTE EXPLANATION

1. Apparent power is the sum/combination of active and reactive power, or the total power in an AC circuit [4].
2. Active power is the power actually consumed in an AC circuit and is measured in kilowatts [3].
3. Reactive power is the power that moves back and forth in the circuit and represents the energy that is stored then released in the form of magnetic/electrostatic field [3, 4].
4. Voltage is the measure of the pressure applied to electrons to make them move and measures the strength of the current in a circuit [5].
5. Current intensity is the magnitude of an electric current measured by the quantity of electricity crossing a specified area per unit time [6].
6. Submetering is typically used in multi-tenant residences like apartments, condominiums, townhomes, and student housing. However, submetering can also be installed to enable users to monitor electrical consumption of individual equipment in a building [7]. Submetering for individual equipment may not be found in all unattached residences. Various states have differing regulations on the use of submetering to monitor individual equipment [8].

Figure 1. Descriptive statistics overview of raw data set



DATA ASSUMPTIONS

- Measurements in the data set are collected from regions within the service area of the power company.
- Measurements are in "minute-averaged" units, so measurements are recorded every minute.
- Voltage, Global_intensity, and Global_reactive-power do not reflect on customer's power bills that are charged based on the household's actual consumption.
- Since power bills are based on power consumption measurements in kilowatt hours, the attributes most relevant to this analysis will be those that report kilowatt hours of active power, such as Global_active_power and Sub-metering 1, 2, 3.
- Since zero values make up a significant portion of the sub_metering attributes, those values are not outliers.
- The higher the value of kWh (kilowatt per hour) in the attributes, the greater the revenue for the power company.

Figure 2. Histogram showing Sub_metering_1, 2, 3 with many zero values

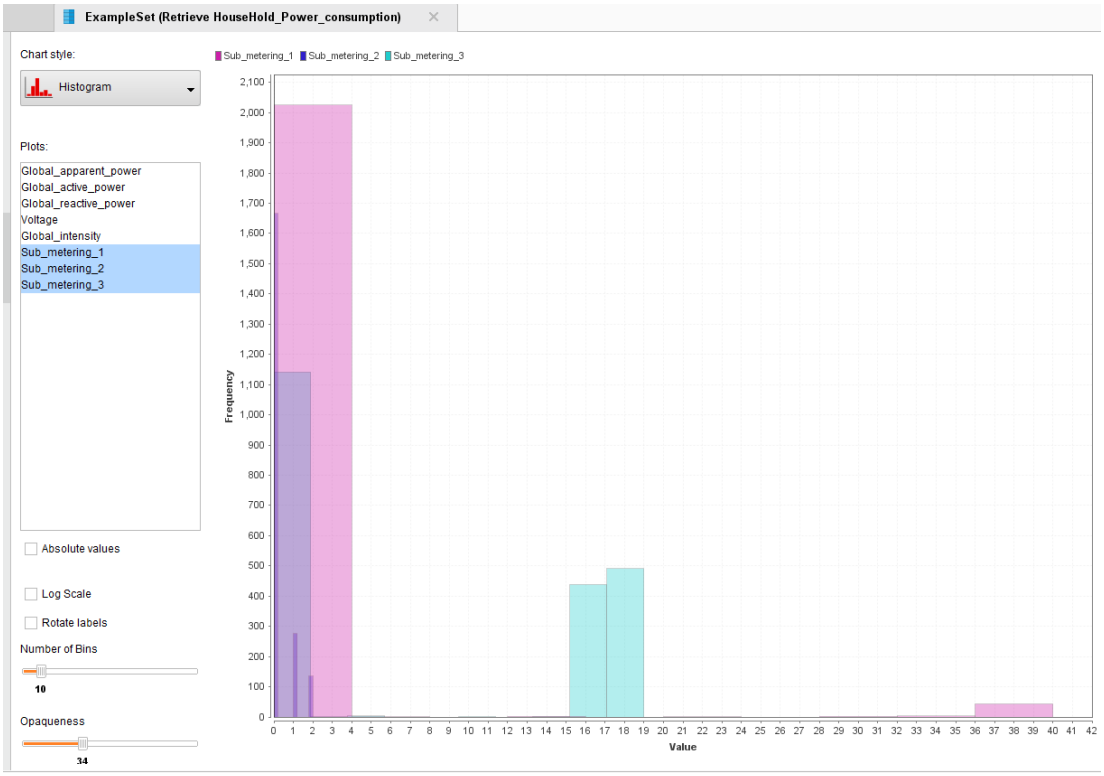


Figure 3. PowerBI histograms of Sub_metering_1, 2, 3 with many zero values

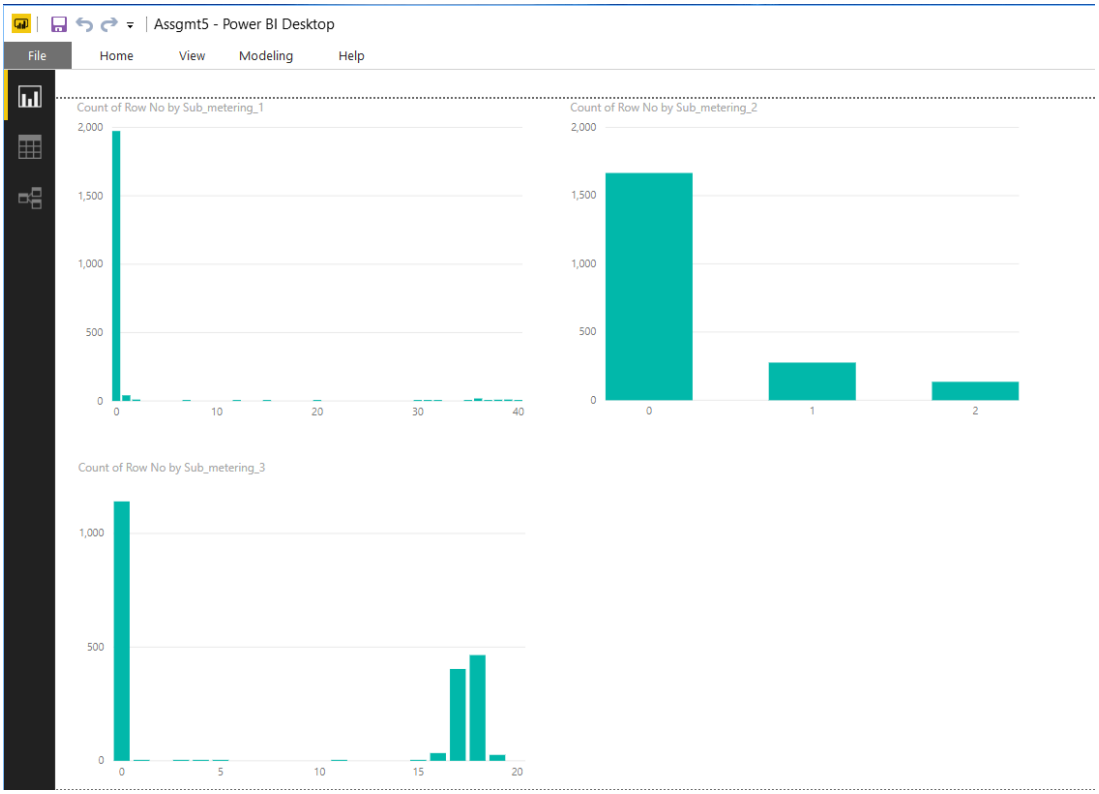
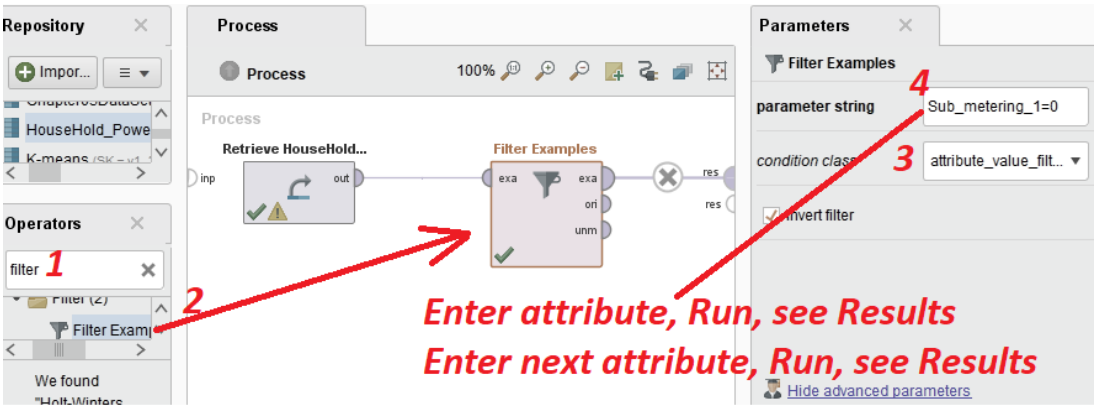
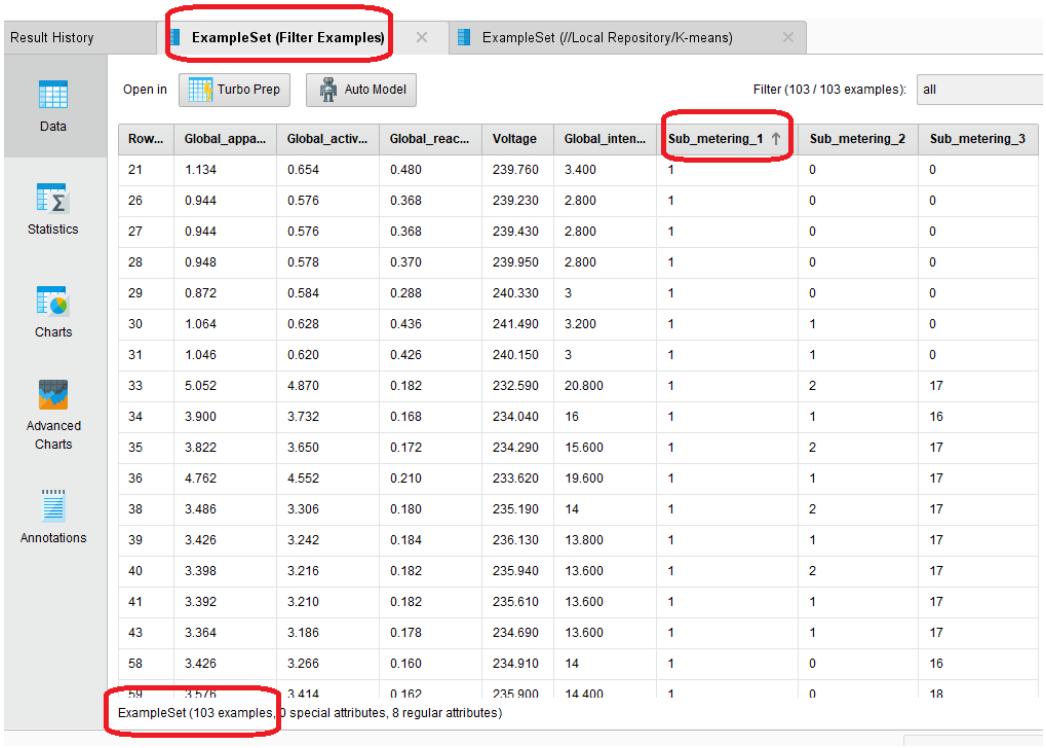


Figure 4. Filter Examples operator to show rows with non-zero values for Sub-metering_1



*Invert filter like above shows rows with non-zero values; to show zero values, unclick invert filter

Figure 5. Filter Example Set shows only 103 examples remaining with non-zero Sub_metering_1, or 1974 zero values



Row...	Global_appa...	Global_activ...	Global_reac...	Voltage	Global_inten...	Sub_metering_1	Sub_metering_2	Sub_metering_3
21	1.134	0.654	0.480	239.760	3.400	1	0	0
26	0.944	0.576	0.368	239.230	2.800	1	0	0
27	0.944	0.576	0.368	239.430	2.800	1	0	0
28	0.948	0.578	0.370	239.950	2.800	1	0	0
29	0.872	0.584	0.288	240.330	3	1	0	0
30	1.064	0.628	0.436	241.490	3.200	1	1	0
31	1.046	0.620	0.426	240.150	3	1	1	0
33	5.052	4.870	0.182	232.590	20.800	1	2	17
34	3.900	3.732	0.168	234.040	16	1	1	16
35	3.822	3.650	0.172	234.290	15.600	1	2	17
36	4.762	4.552	0.210	233.620	19.600	1	1	17
38	3.486	3.306	0.180	235.190	14	1	2	17
39	3.426	3.242	0.184	236.130	13.800	1	1	17
40	3.398	3.216	0.182	235.940	13.600	1	2	17
41	3.392	3.210	0.182	235.610	13.600	1	1	17
43	3.364	3.186	0.178	234.690	13.600	1	1	17
58	3.426	3.266	0.160	234.910	14	1	0	16
59	3.576	3.414	0.162	235.900	14.400	1	0	18

Figure 6. Filter Example Set shows 411 examples remaining with non-zero Sub_metering_2, or 1666 zero values

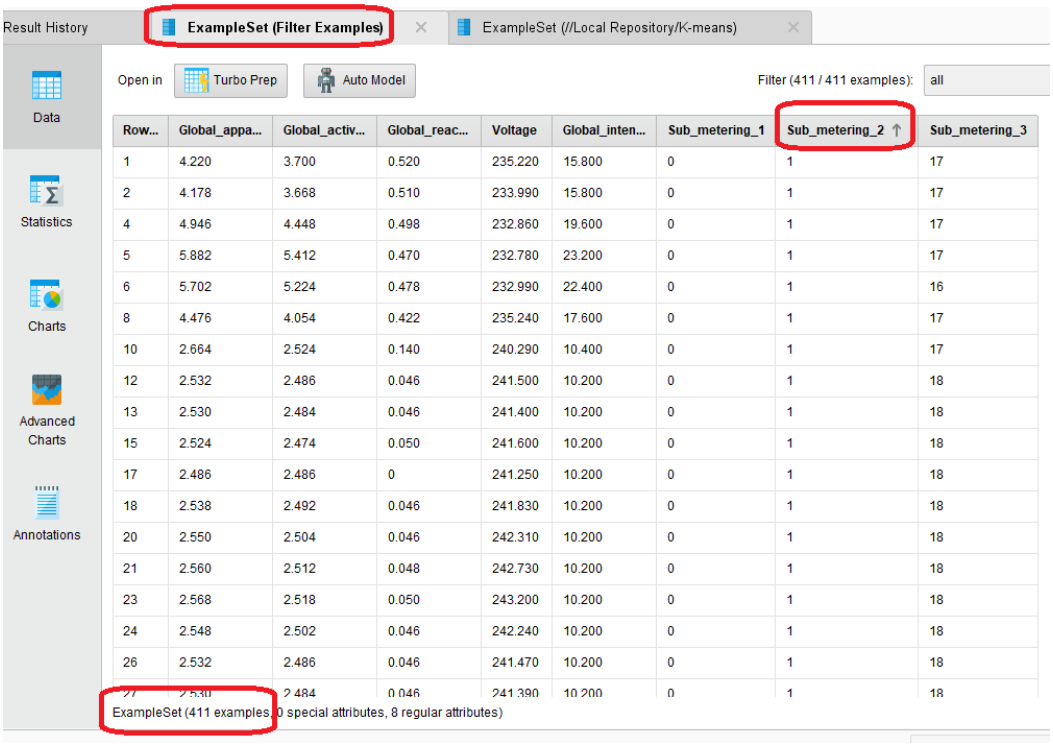


Figure 7. Filter Example Set shows 936 examples remaining with non-zero Sub_metering_3, or 1141 zero values

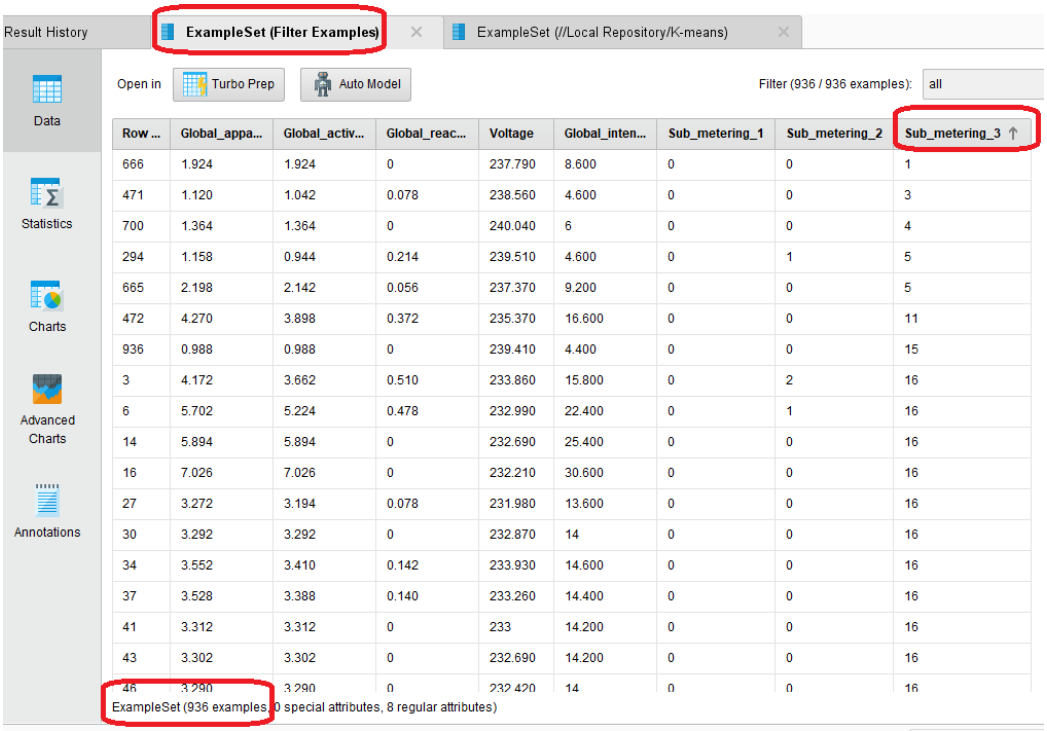
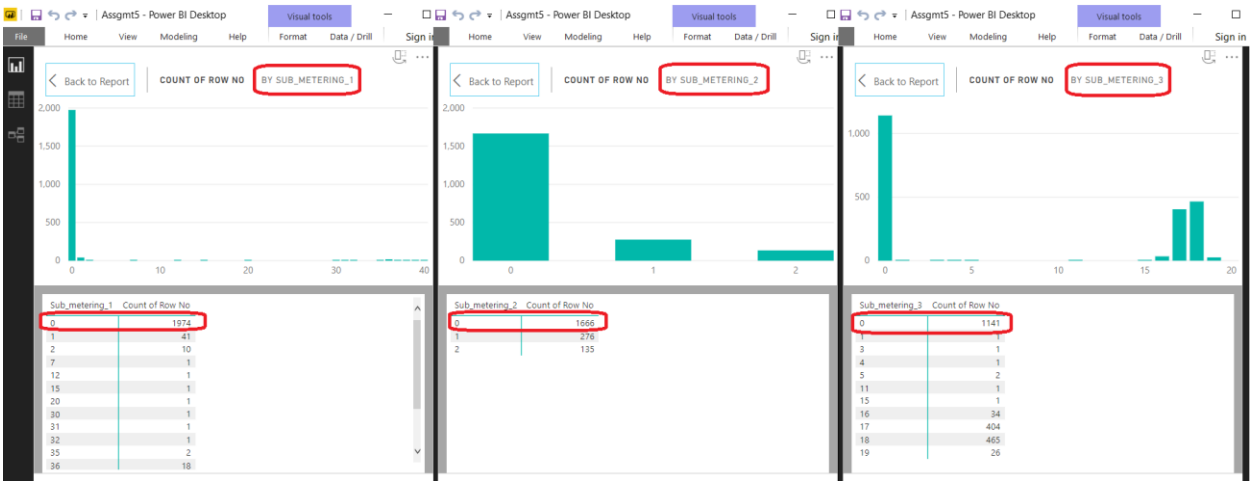


Figure 8. PowerBI counts of zero values for Sub_metering_1, 2, 3



DATA PREPARATION

The k-Means clustering does accept both numeric and polynominal, but distance measures are more effective with numeric. Household Power Consumption data set is already in numerical data type.

Data Type Transformation: Using RapidMiner operator *Normalize,* all attributes have been normalized with the resulting standard deviation adjusted to 1. Normalization operator was applied because 3 attributes in particular had standard deviations beyond ±3. Two of those attributes (Sub-metering 1 and 3) are deemed essential to the analysis, and since clustering analysis is sensitive to extreme outlier values, standard deviations like 8.716 (see figure 8) required normalization.

Figure 9. RapidMiner Process to normalize data

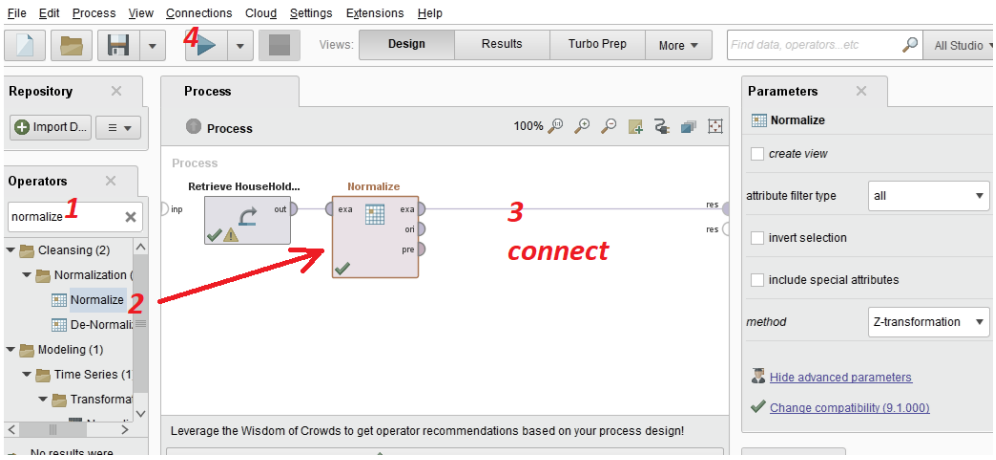


Figure 10. RapidMiner Result with standard deviations beyond ±3 *before* Normalization

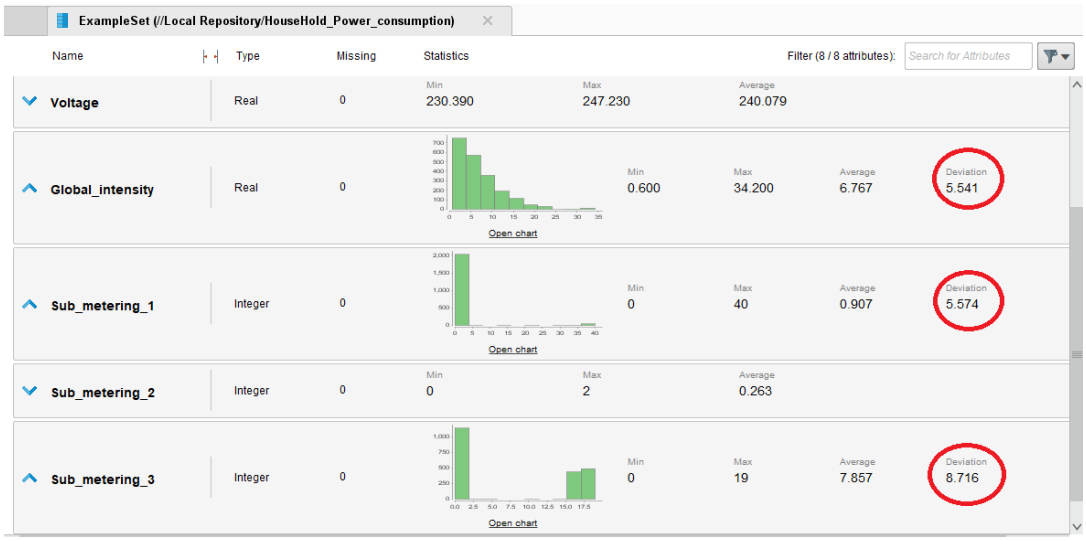
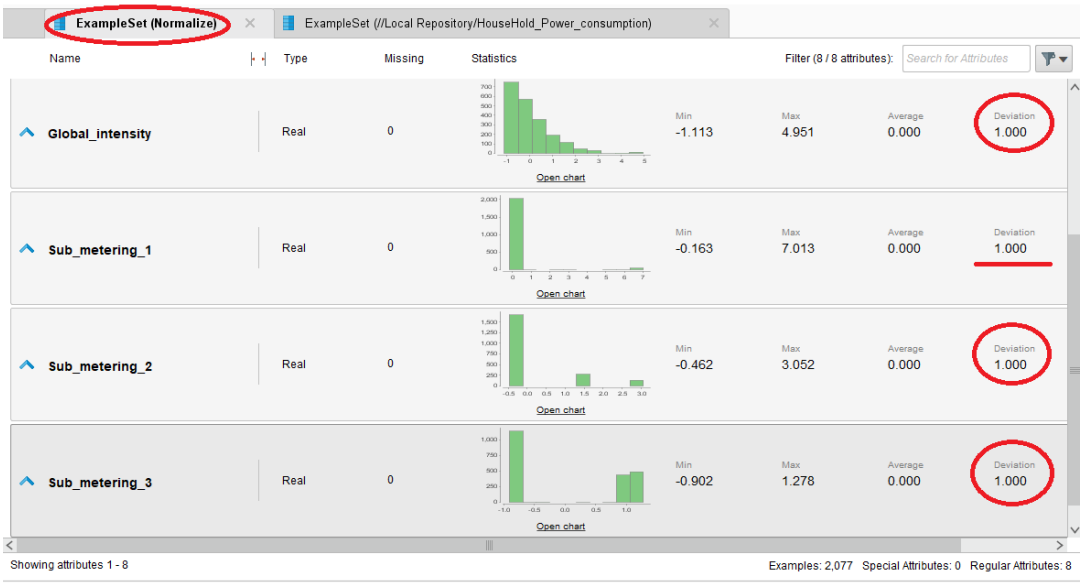


Figure 11. RapidMiner Result with standard deviations after Normalization



Data Preparation of Missing Values: No missing values were found in the data set. Zero values for Sub-metering 1, 2, 3 attributes were left in the data set as is, because the count of zero values were too numerous/significant to justify removing them from the analysis.

Data Preparation of Inconsistent Values: Zero values in Sub_metering 1, 2, 3 are logically inconsistent, since any household that has active power usage would reasonably be expected to also have power usage in the kitchen, refrigerator and water heater measured by Sub_metering 1, 2, 3. However, without more information on the particular data set, this report resumes analysis based on the assumptions aforementioned. These zero values will remain in the data set as per the discussion pertaining to these zero values in the preceding sections.

MODELING

The iterative process of generating clusters using k-Means involves specifying a k number of clusters and assigning data points to the nearest centroid and repeating the process until the Sum of Squared Errors (SSE) is minimized.

K-MEANS

First, the k-Means operator in RapidMiner is used generate the initial random centroid or center point on the normalized data set. Value of k is set to 3 for this iteration. Squared Euclidean Distance is selected for proximity measure of the data points.

Figure 12. RapidMiner k-Means Clustering operator process

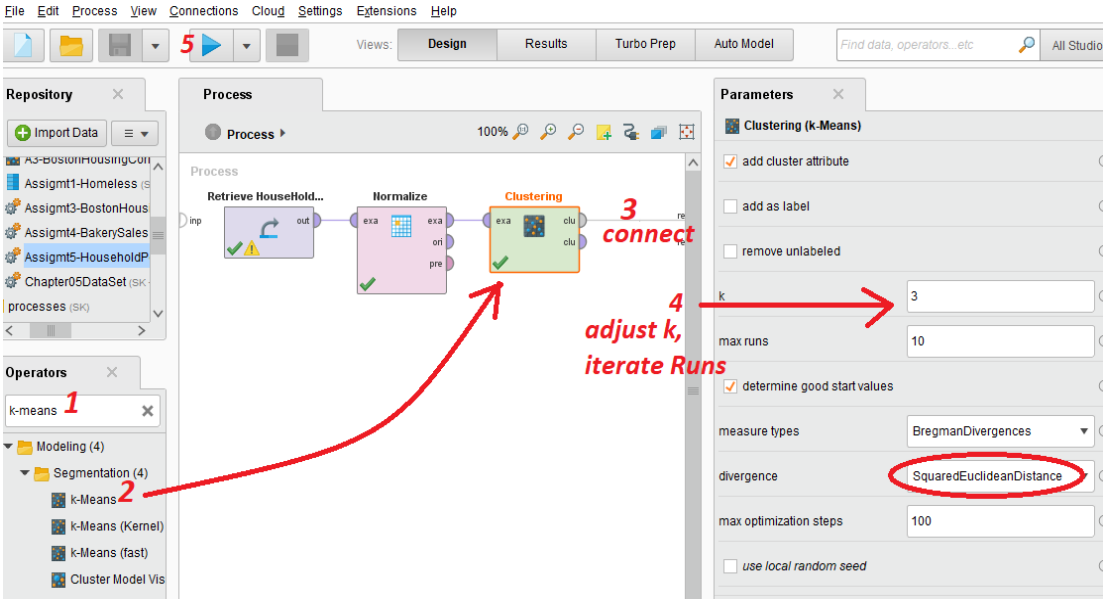


Figure 13. RapidMiner Result cluster description with *k* = 3

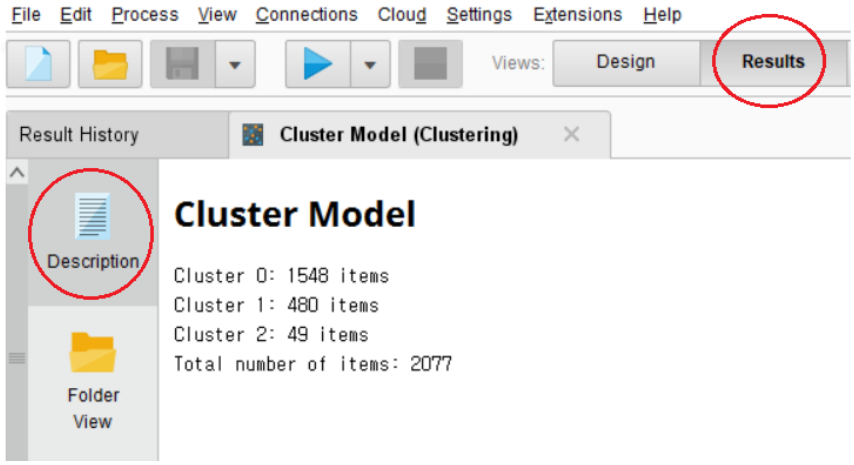


Figure 14. RapidMiner Result cluster graph with *k* = 3

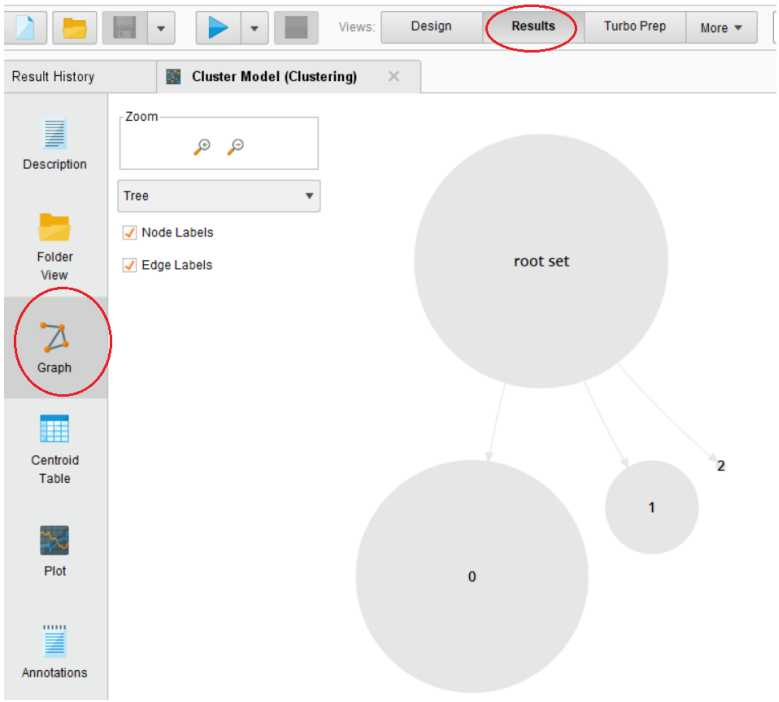
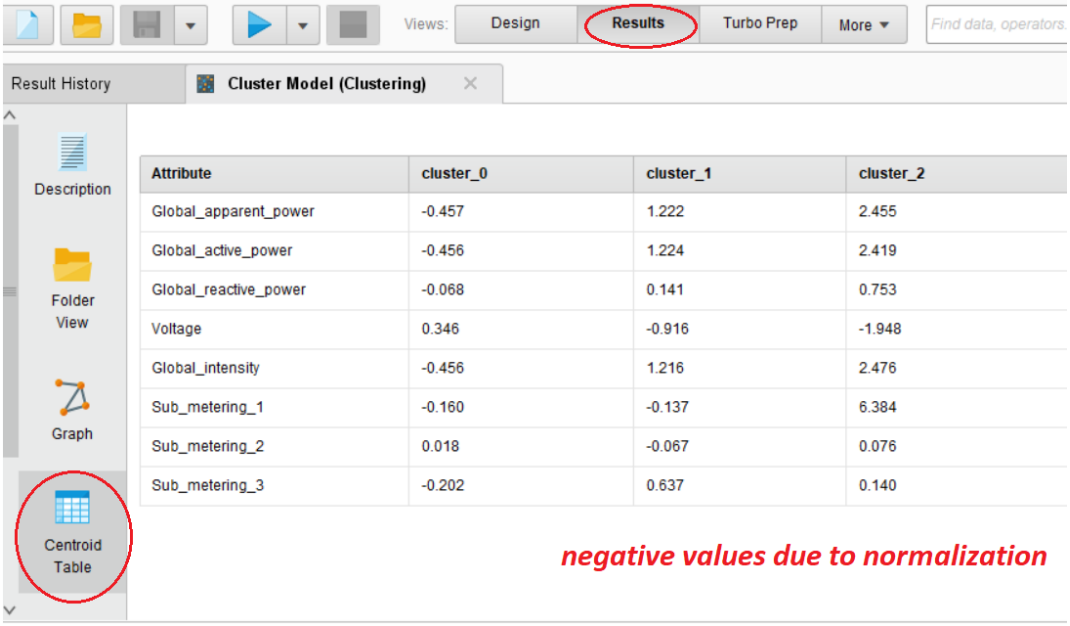


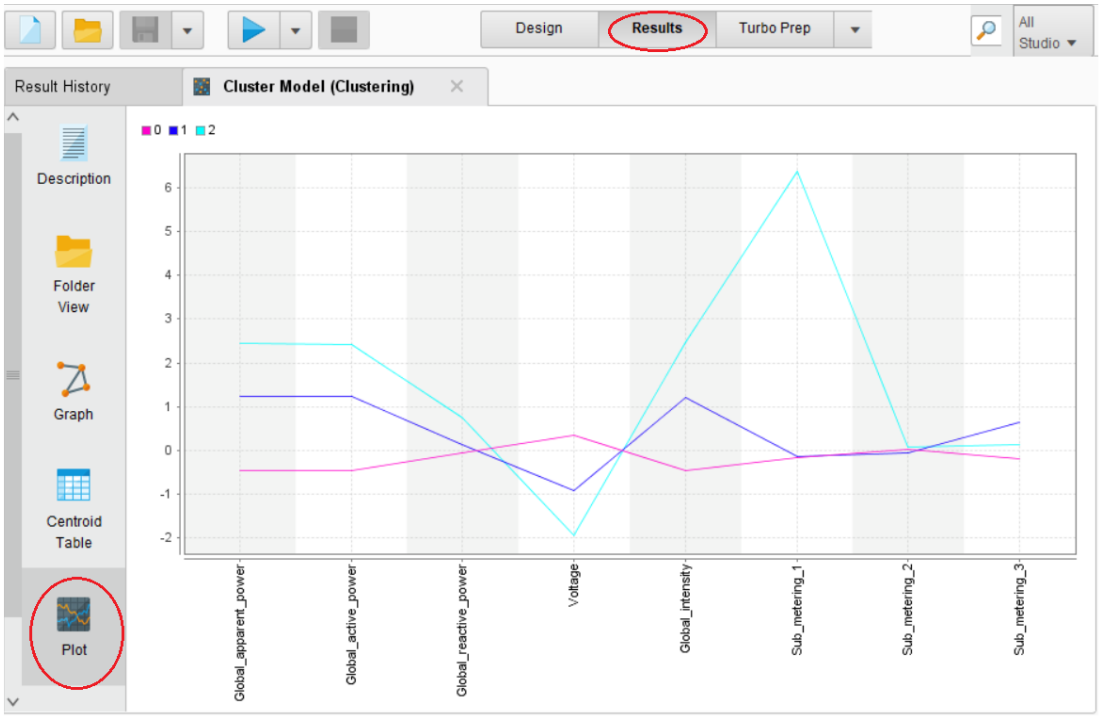
Figure15. RapidMiner Result centroid table with *k* = 3



Attribute	cluster_0	cluster_1	cluster_2
Global_apparent_power	-0.457	1.222	2.455
Global_active_power	-0.456	1.224	2.419
Global_reactive_power	-0.068	0.141	0.753
Voltage	0.346	-0.916	-1.948
Global_intensity	-0.456	1.216	2.476
Sub_metering_1	-0.160	-0.137	6.384
Sub_metering_2	0.018	-0.067	0.076
Sub_metering_3	-0.202	0.637	0.140

negative values due to normalization

Figure 16. RapidMiner Result cluster plot with *k* = 3



CLUSTER DISTANCE PERFORMANCE

Next, the RapidMiner Cluster Distance Performance operator is deployed to evaluate the effectiveness of the clustering groups using SSE and the Davies-Bouldin index. The Performance operator measures the average cluster distance and the Davies-Bouldin index. Large separations between centroids indicates well-separated clusters with the data set data set divided neatly and is desirable. Low Davies-Bouldin index and low average-within-centroid distances indicate better/more cohesive clusters, so low numbers are desirable.

Figure 17. RapidMiner Process with Cluster Distance Performance operator

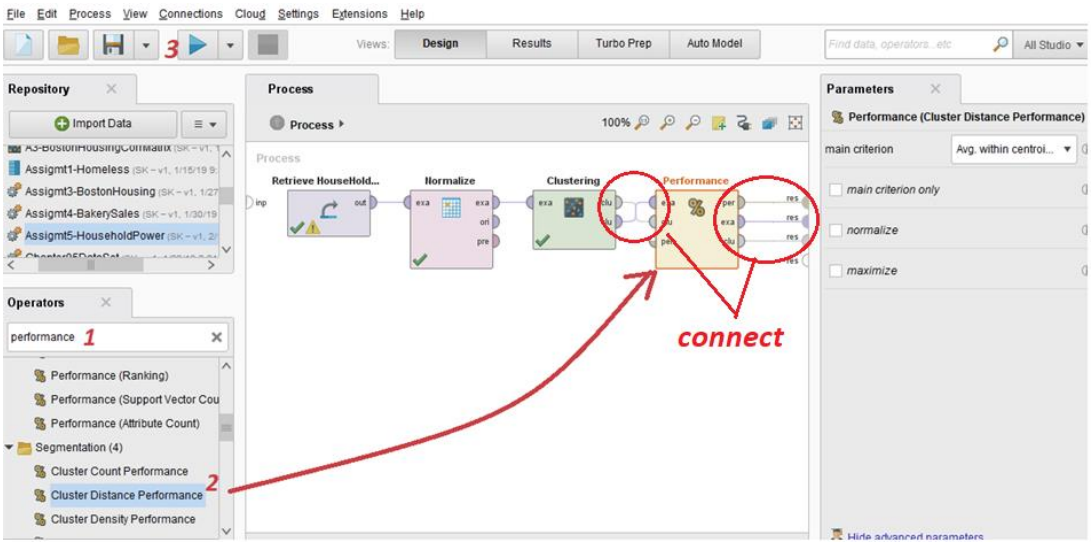
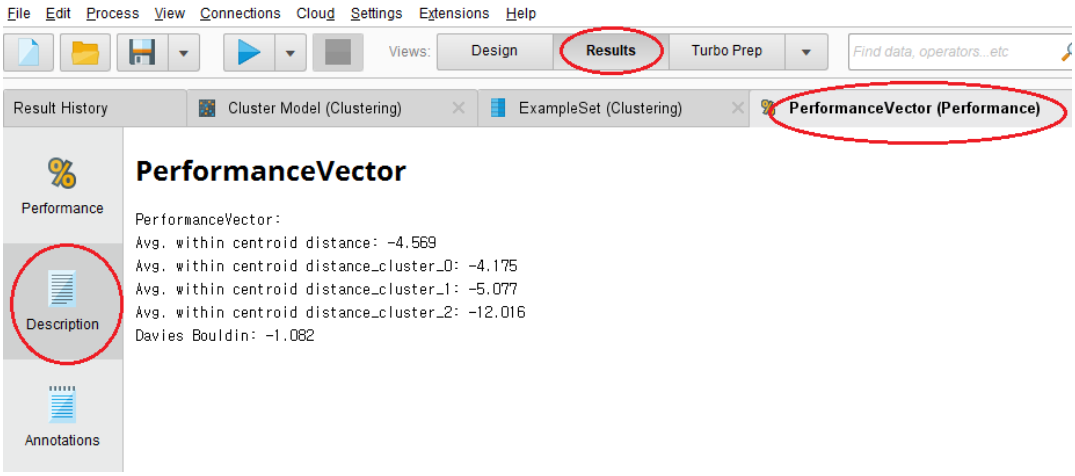


Figure 18. RapidMiner Result of Cluster Distance Performance with *k* = 3



EVALUATION OF FINDINGS

SUMMARY OF FINDINGS

Best *k* is 3

After multiple iterations of clustering with k values from 2 to 6, and an examination of their respective Performance vectors, this report finds that 3 clusters can best represent different groups of customers from this Household Power Consumption data set. Of the various k values, the average-within-centroid distance and Davies Bouldin index was the lowest with 3 clusters. Moreover, 3 clusters is a logical division to categorize customers into 1) high power usage group, 2) mid power usage group, and 3) low power usage group.

Cluster_0: Low power usage group (1548 of 2077 items)
Of the 3 clusters, cluster_0 has the lowest Global_apparent_power, Global_active_power, Global_reactive_power, Global_intensity, Sub_metering_1 and Sub-metering_3. However, custer_0 has the highest for Voltage and is in the middle for Sub_metering_2.

Cluster_1: Mid power usage group (480 of 2077 items)
Of the 3 clusters, cluster_1 holds the middle in Global_apparent_power, Global_active_power, Global_reactive_power, Voltage, Global_intensity, and Sub_metering_1. However, cluster_1 has the highest usage in Sub_metering_3, which measures electric water heater and air conditioner. Cluster_1 has the lowest usage in Sub_metering_2, which measures laundry room, refrigerator and light.

Cluster_2: High power usage group (49 of 2077 items)
Of the 3 clusters, cluster_2 has the highest Global_apparent_power, Global_active_power, Global_reactive_power, Global_intensity, Sub_metering_1, and Sub_metering_2. However, cluster_2 has the lowest Voltage and is in the middle for Sub_metering_3.

Figure 19. Centroid table with lowest and highest power usage

Attribute	cluster_0	cluster_1	cluster_2
Global_apparent_power	-0.457	1.222	2.455
Global_active_power	-0.456	1.224	2.419
Global_reactive_power	-0.068	0.141	0.753
Voltage	0.346	-0.916	-1.948
Global_intensity	-0.456	1.216	2.476
Sub_metering_1	-0.160	-0.137	6.384
Sub_metering_2	0.018	-0.067	0.076
Sub_metering_3	-0.202	0.637	0.140

Lowest power user **Highest power user**

OBSERVATIONS ON USAGE

- The lowest usage group, cluster_0, and the highest usage group, cluster_2 has an interesting inverse relationship in Voltage. That is, the highest voltage has the lowest usage in most of the attribute categories, while the lowest voltage has the highest usage in most of the attributes (see figure 19).
- When Sub_metering_1 (which measures the kitchen, containing mainly a dishwasher, an oven and a microwave) shows low usage, then overall power usage also appears to be low. Inversely, when Sub_metering_1 had high usage, the overall power usage also appeared to be high (see figure 19).
- The lowest power user group had the lowest usage of Sub-metering_1 (kitchen) and Sub_metering_3 (water heater, air conditioner), but not the lowest Sub_metering_2, which measures laundry room, containing a washing-machine, a tumble-drier, a refrigerator and a light (see figure 20).
- The highest power user group had the highest usage in Sub-metering_1 (kitchen) and Sub_metering_2 (laundry room, refrigerator, light), but not the highest in Sub_metering_3, which measures electric water heater and air conditioner (see figure 20).

Figure 20. Sub_metering usage summary by cluster group

Power User Group	Sub_metering_1 **Kitchen**	Sub_metering_2 **Laundry Rm, Refrigerator**	Sub_metering_3 **Water Heater, AC**
Low user cluster_0	Lowest		Lowest
Mid user cluster-1		Lowest	Highest
High user cluster_2	Highest	Highest	

- The largest number of examples or items fell into the low power usage cluster, with 1548 of 2077 items in cluster_0.
- The smallest number of examples or items fell into the high power usage cluster, with 49 items in cluster_2.

Note: Above bullet points are mere observations on the cluster groups and not intended to draw correlations or associations, which would require respective data modeling and analysis.

BUSINES RECOMMENDATIONS

Since power companies are regulated by PUC with regard to the charge rate and profit margins, the business assumption on which this analysis was built is that attracting more customers is the best means to create more revenue. In order to attract more customers, additional information is needed to further augment this analysis to yield better informed recommendations.

For instance, for the highest user group (cluster_2), this report recommends that the geographical information for those users be considered together to identify a region with the greatest density of high power users and offer referral incentives to attract more customers from the region with the greatest density of high power users.

For the low power user group (cluster_0), this report recommends that a time-of-usage attribute be considered to identify both time of day (short term) and also peak power usage seasons (long term), and offer prepaid plans prior to peak power usage seasons or offer multiple time-of-day variable plans to attract more customers.

Also, this report recommends the company to obtain information on competitor power companies rates and plans, to offer price matching incentives to attract new customers.

REFERENCES

1 Make an Informed Decision: How to Find the Right Energy Plan for You. (2018, May 29). Retrieved from https://www.saveonenergy.com/learning-center/post/picking-a-plan/

2 Girouard, C. (2015, April 23). Advanced Energy Perspectives: How Do Electric Utilities Make Money? Retrieved from https://blog.aee.net/how-do-electric-utilities-make-money

3 What is Active, Reactive and Apparent Power - definition and explanation. (2017, May 17). Retrieved from https://circuitglobe.com/what-is-active-reactive-and-apparent-power.html

4 Mishra, A. (2017, July 11). *Quora*[What is real, reactive, and apparent power?]. https://www.quora.com/What-is-real-reactive-and-apparent-power.

5 The NEED Project. (2017). Measuring Electricity [PDF file]. Retrieved from https://www.need.org/Files/curriculum/infobook/Elec3S.pdf

6 Current intensity. (n.d.). In *Merriam-Webster*. Retrieved from https://www.merriam-webster.com/dictionary/current intensity

7 Utility submeter. (2018, September 07). Retrieved from https://en.wikipedia.org/wiki/Utility_submeter

8 Utility Submetering. (2016, January 15). Retrieved from http://www.ncsl.org/research/energy/utility-submetering.aspx