NOVEC Customer Segmentation Analysis

Anita Ahn Mesele Aytenifsu Bryan Barfield Daniel Kim

Department of Systems Engineering and Operations Research

SYST/OR 699 – Fall 2016-Final Presentation

Agenda

- Introduction
- Problem Statement
- Methodology/ Data Description
- Cluster Analysis
- Applications
- Difficulties/ Lessons Learned
- Conclusion/Recommendation

Introduction – About

- NOVEC: Northern Virginia Electric Cooperative. Locally based electric distribution system
- Services 651 sq miles of area
- 6,880 miles of power lines
- Provides electricity to more than 155,000 home and businesses
- Stretches over multiple Counties:
  - Fairfax, Loudoun, Prince William
  - Stafford, Fauquier

Well-known clients: Potomac Mills Mall, Verizon, AT&T

Introduction – Background on NOVEC's Customers

- NOVEC currently has 3-4 different qualitative consumer segments
    - Residential
    - Small Commercial
    - Large Commercial
    - Church

- These qualitative consumer segments are not homogeneous nor good indicators of consumer's energy usage behavior

Problem Statement

- NOVEC wants to:
  - Segment customers based on their usage of electricity using data already collected for another purpose
  - Determine how these customer segments contribute towards NOVEC's system peak usage
- Why is this Important?

Assumptions & Limitations

- Current data is Stratified Sample, collected for the purpose of rate making
    - Data contains higher population of consumers who use large amounts of electricity (i.e. Large Commercial)
    - Majority of NOVEC's consumers consist of Residential customers

Goals for this Project

Clustering for July data
- Using NOVEC's data on July energy consumption, segment the consumers into groups respective of NOVEC's total peak consumption

Clustering for January Data
- Using the same clustering technique, segment the consumers January usage respective of NOVEC's total peak consumption

Validate Consumer Clusters
- Validate the consumer segments by looking at load profiles

NOVEC Implements
- Using these consumer clusters NOVEC intends to use these customer segments for future forecasting, pricing analysis, and capacity planning

*Project Team will focus on Goals 1-3; Goal 4 will be done by NOVEC

Data Description

Provided Variables	Description
Account	Unique customer identifier
Map Location	Geospatial identifier
Group	Customer Billing Classification (RES, LGCOM, SMCOM, CHRCH)
Usage	Energy expenditure in kilowatt-hour (kWh)
DateTime	MM-DD-YYYY 00:00 (24-hour)

Useful Variables	Description
Account	Unique customer identifier
~~Map Location~~	~~Geospatial identifier~~
~~Group~~	~~Customer Billing Classification (RES, LGCOM, SMCOM, CHRCH)~~
Usage	Energy expenditure in kilowatt-hour (kWh)
DateTime	MM-DD-YYYY 00:00 (24-hour)

Terminology used in Analysis

Consumer's Peak Consumption: Consumer's highest energy usage amount in the time period

Consumer's Average Energy Use: Consumer's average KwH energy usage amount in the time period

Peak System Load: Maximum peak electricity usage in KwH for entire NOVEC's system in time period

Coincident Peak Usage: Consumer's KwH usage at the time NOVEC's system peaked

Worknight/Workday Total Usage: Consumer's total KwH usage during 8am-4pm/ on Monday-Friday for entire month

Weekday/Weekend Total Usage: Consumer's total KwH usage during Monday-Friday/ Saturday-Sunday for entire month

Derived Variables

- Account
- Usage
- DateTime

- Demand Factor

$$\frac{\text{Consumer's Peak Consumption}}{\text{Peak System Load}}$$

- Load Factor

$$\frac{\text{Consumer's Avg Energy Use}}{\text{Consumer'sPeak Consumption}}$$

- Coincident Usage Ratio

$$\frac{\text{Consumer's Coincident Peak Usage}}{\text{Peak System Load}}$$

- Coincident Peak Ratio

$$\frac{\text{Consumer's Coincident Peak Usage}}{\text{Consumer's Peak Consumption}}$$

- Worknight to Workday Usage Ratio

$$\frac{\text{Worknight Total Usage}}{\text{Workday Total Usage}}$$

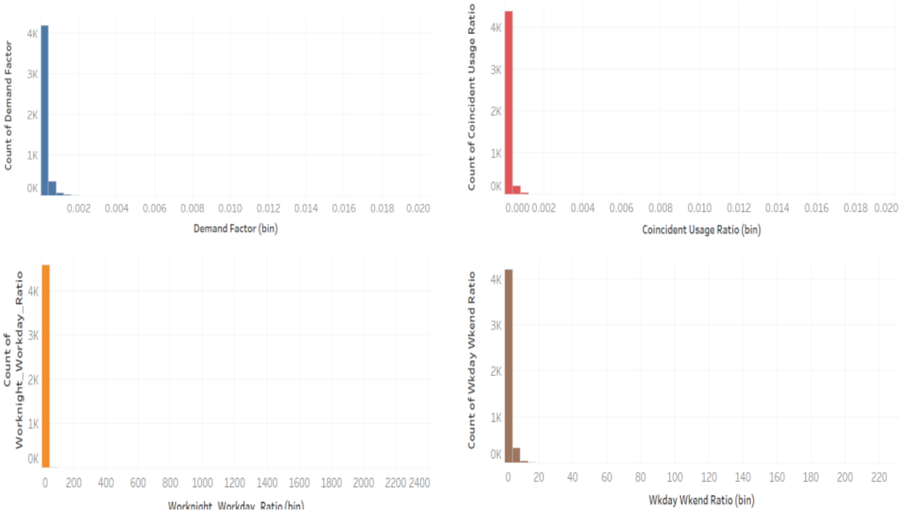
- Weekday to Weekend Usage Ratio

$$\frac{\text{Weekday Total Usage}}{\text{Weekend Total Usage}}$$

Method for Customer Segmentation

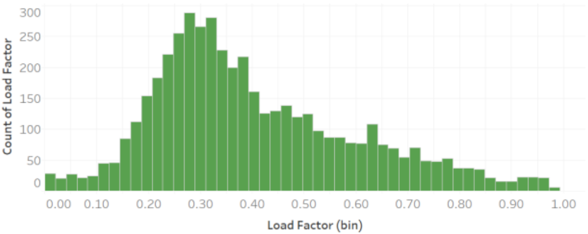
1. Manipulate and transform the data so that it is suitable for the K-means algorithm
2. Determine the optimal number of clusters
3. Run the K-means algorithm
4. Analyze and profile the clusters

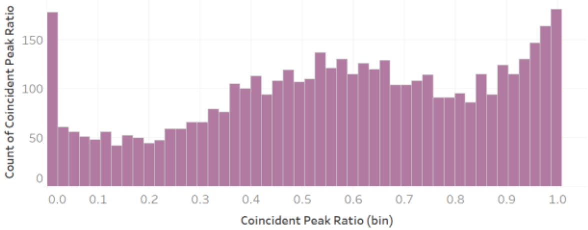
Variable Exploration (Demand Factor, Coincident Usage Ratio, Weekday-Weekend, Worknight-Workday Ratios)



The histograms for these variables show heavily right-skewed distributions. Data will need Log Transformation

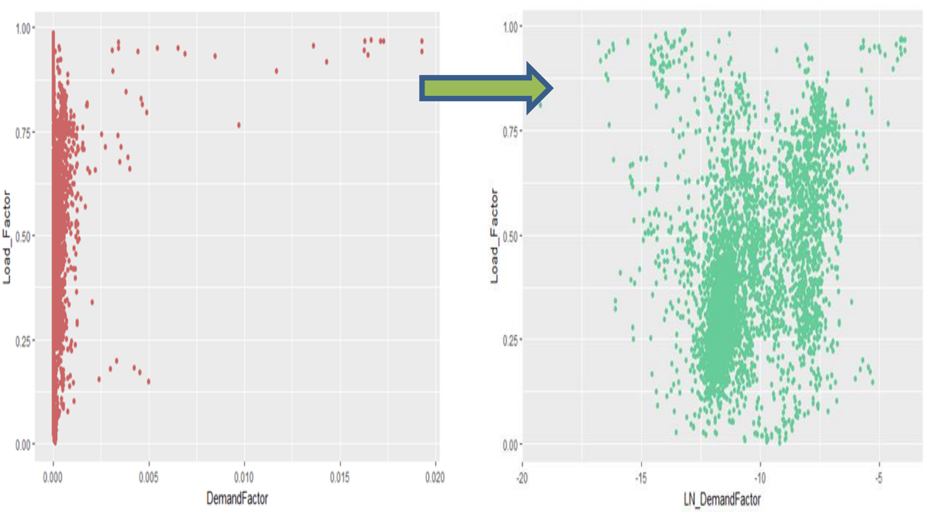
Load Factor and Coincident Peak Ratio Variables





These histograms are not skewed. Okay to use data without Log Transformation

Data Transformation



How do we segment the customers?

Using the **K-Means Clustering Algorithm**!

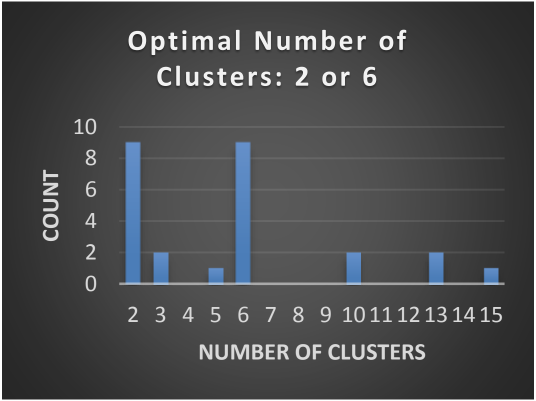
Steps:

1) Choose the number of clusters, ***k.***
2) Generate **k** random points as cluster centroids.
3) Assign each point to the nearest cluster centroid.
4) Recompute the new cluster centroid.
5) Repeat the two previous steps until some convergence criterion is met (usually when assignment of clusters has not changed over multiple iterations).

Requires the user to choose the number of clusters to be generated beforehand.

Finding Optimal Number of Clusters

Index Name	Reference	Number of Clusters
KL	Krzanowski and Lai 1988	6
CH	Calinski and Harabasz 1974	10
Hartigan	Hartigan 1975	5
CCC	Sarle 1983	10
Scott	Scott and Symons 1971	6
Marriot	Marriot 1971	6
TrCovW	Milligan and Cooper 1985	3
TraceW	Milligan and Cooper 1985	6
Friedman	Friedman and Rubin 1967	6
Rubin	Friedman and Rubin 1967	6
Cindex	Hubert and Levin 1976	2
DB	Davies and Bouldin 1979	2
Silhouette	Rousseeuw 1987	2
Duda	Duda and Hart 1973	2
Pseudot2	Duda and Hart 1973	2
Beale	Beale 1969	2
Ratkowsky	Ratkowsky and Lance 1978	6
Ball	Ball and Hall 1965	3
Ptbiserial	Milligan 1980, 1981	3
Frey	Frey and Van Groenewoud 1972	13
McClain	McClain and Rao 1975	2
Dunn	Dunn 1974	2
Hubert	Hubert and Arabie 1985	6
SDindex	Halkidi et al. 2000	13
Dindex	Lebart et al. 2000	6
SDbw	Halkidi and Vazirgiannis 2001	15



Are the clusters really different from each other?

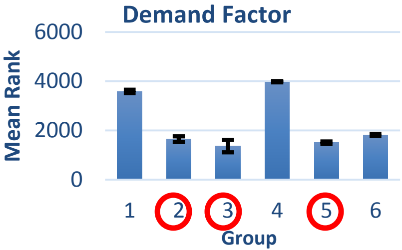
Kruskal-Wallis Test: There are at least two clusters that are statistically different per variable.

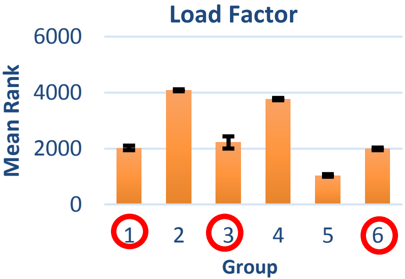
Variable	DF	Chi-Square	P-value
DemandFactor	5	2586.1	< 0.0001
Load_Factor	5	2928.4	< 0.0001
CoincidentUsageRatio	5	3179.2	< 0.0001
Coincident_Peak_Ratio	5	3022.9	< 0.0001
Wknight_wkday_Ratio	5	1504.9	< 0.0001
Wkday_wkend_Ratio	5	1335.8	< 0.0001

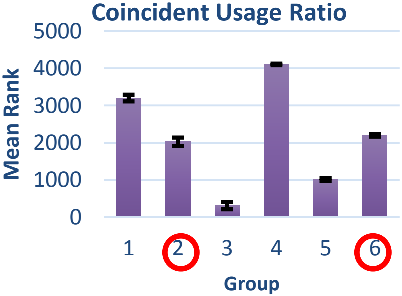
Variable	Description	Variable	Description
Demand Factor	Customer's Peak Consumption/Peak System Load	Coincident Peak Ratio	Customer's Coincident Peak Usage/Customer's Peak Consumption
Load Factor	Customer's Avg Energy Usage/Customer's Peak Consumption	Weekend to Weekday Usage Ratio	Weekday Total Usage/Weekend Total Usage
Coincident Usage Ratio	Customer's Coincident Peak Usage/Peak System Load	Worknight to Workday Usage Ratio	Worknight Total Usage/Workday Total Usage

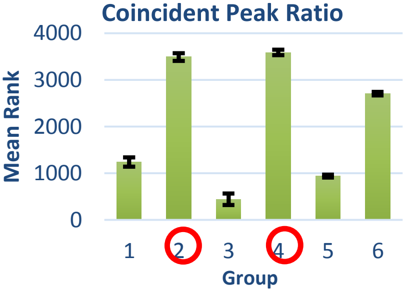
Are the clusters really different from each other?

Post-Hoc Analysis: **Dunn Test** for multiple pairwise comparisons

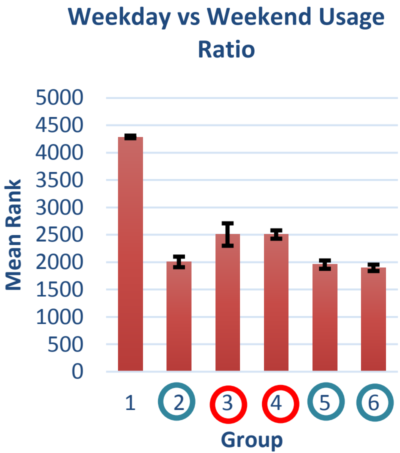


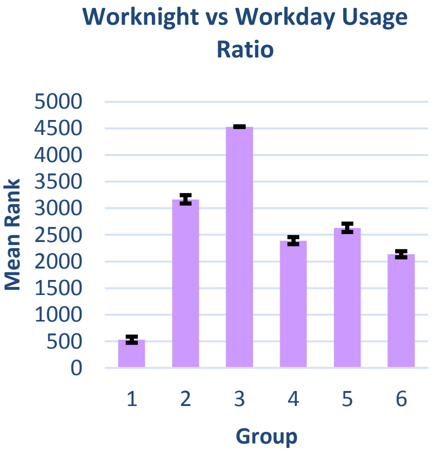




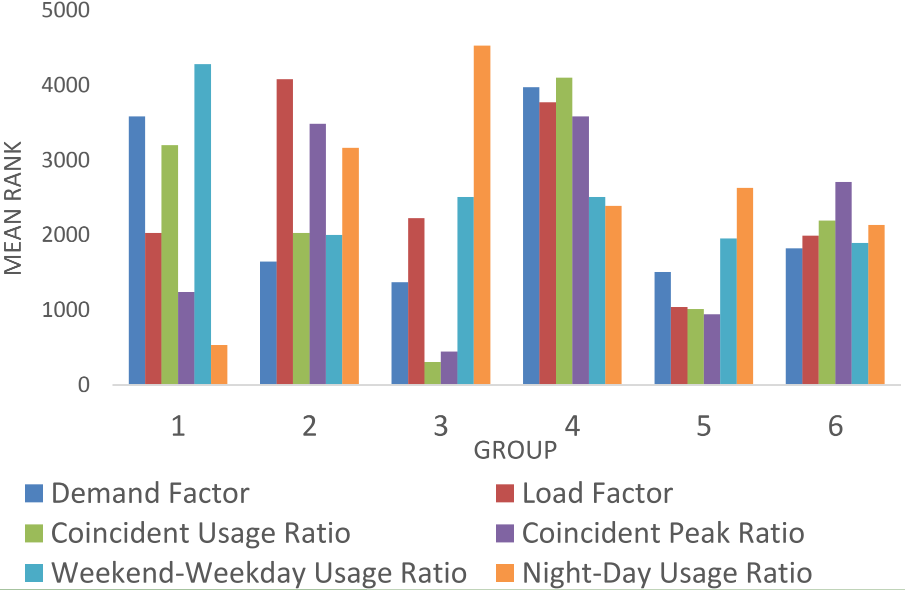


Are the clusters really different from each other?

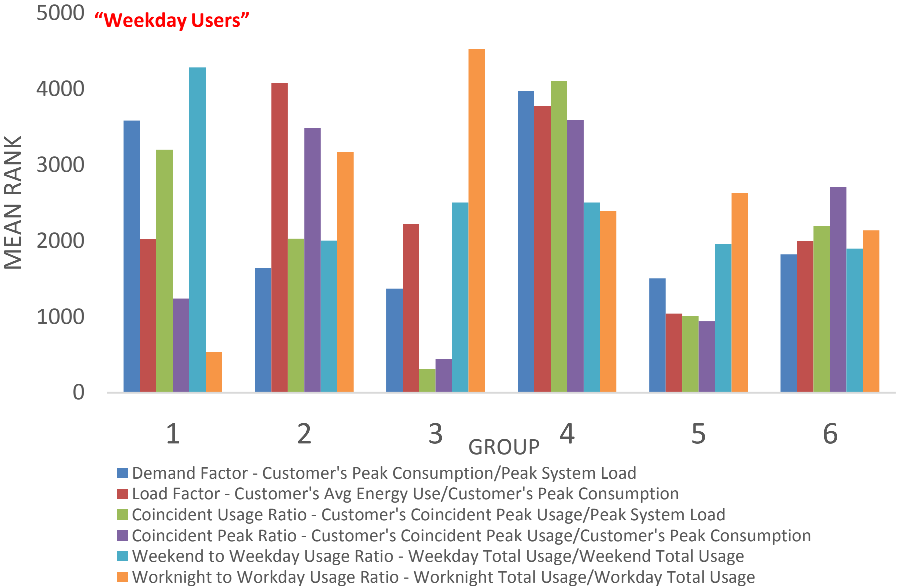




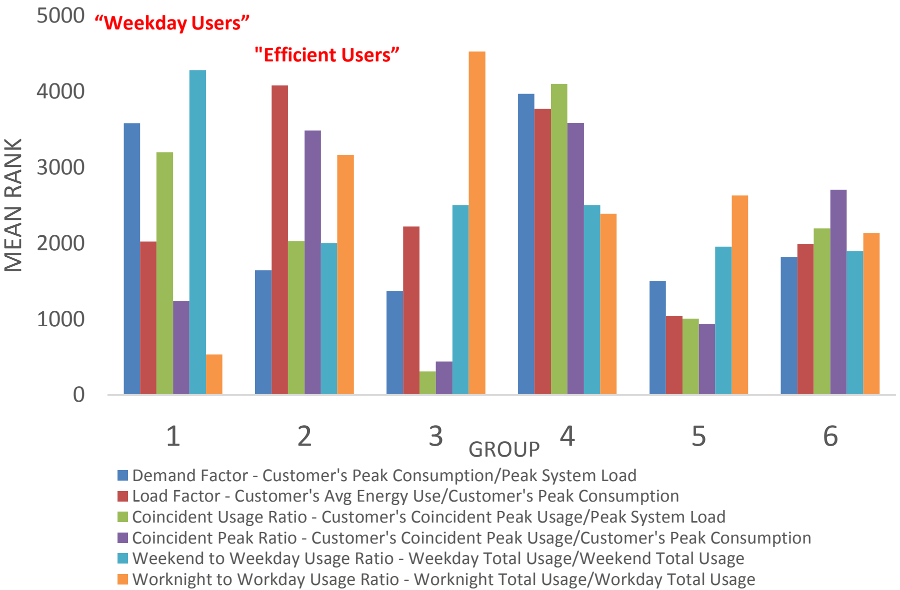
Customer Profiling



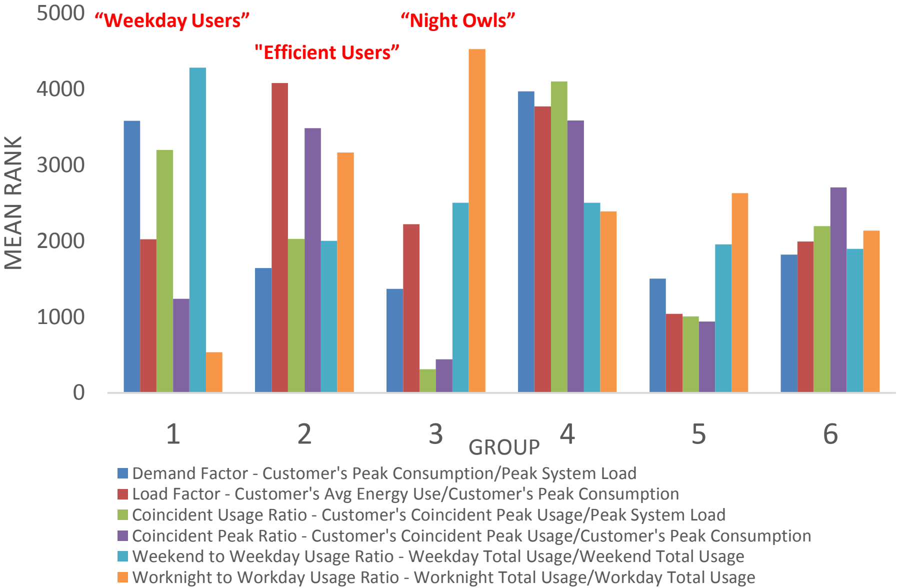
Customer Profiling



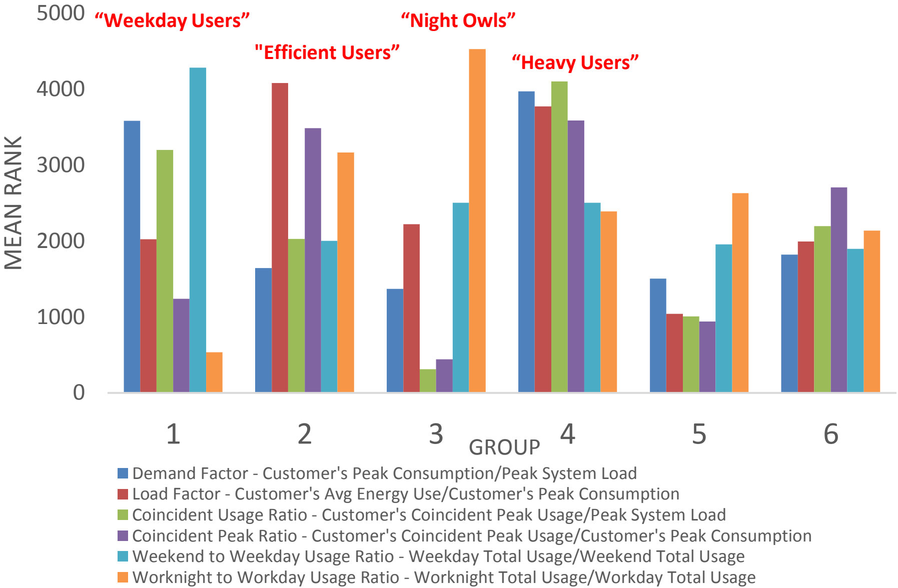
Customer Profiling



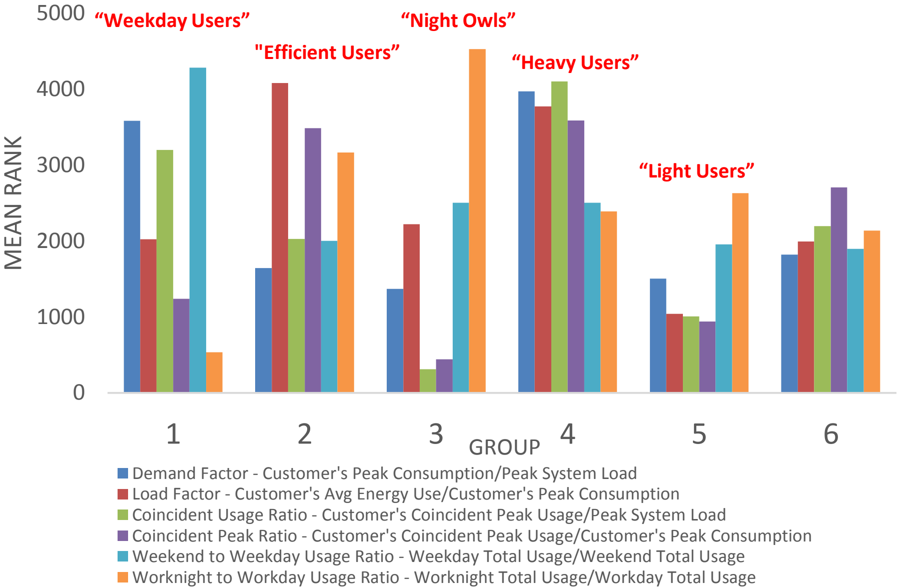
Customer Profiling



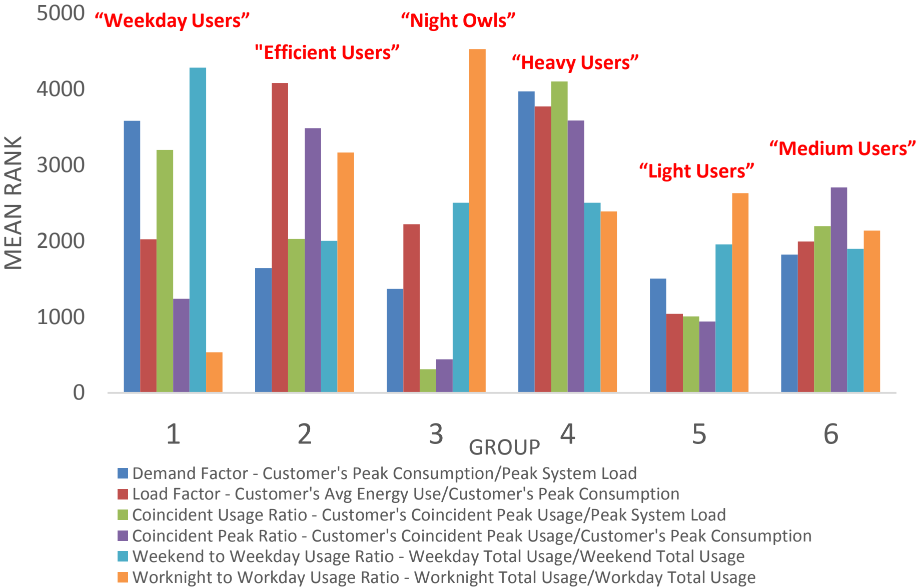
Customer Profiling



Customer Profiling

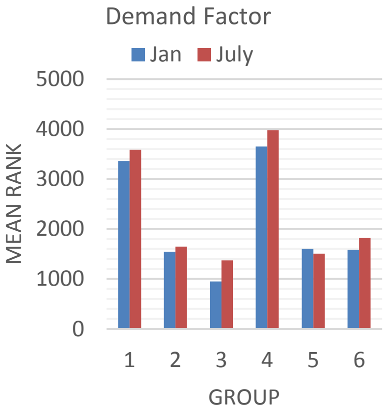


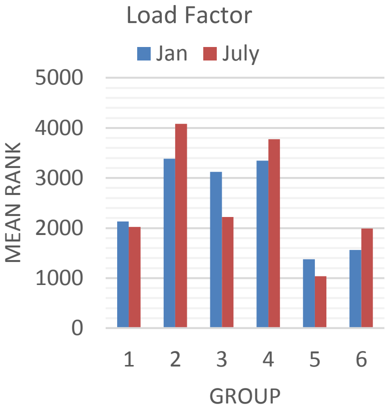
Customer Profiling



Jan vs July Usage Behavior

Group	User Type	Group	User Type
1	Weekday Users	4	Heavy Users
2	Efficient Users	5	Light Users
3	Night Owls	6	Medium Users



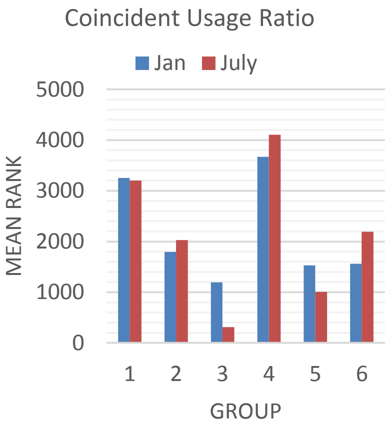


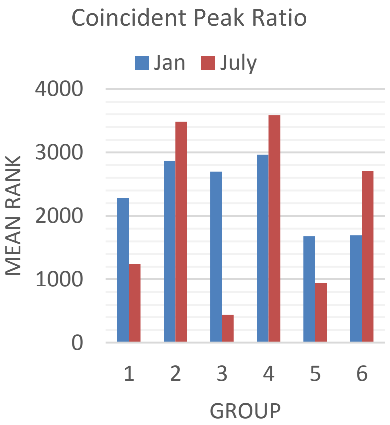
$$\text{Demand Factor} = \frac{\text{Consumer's Peak Consumption}}{\text{Peak System Load}}$$

$$\text{Load Factor} = \frac{\text{Consumer's Avg Energy Use}}{\text{Consumer's Peak Consumption}}$$

Jan vs July Usage Behavior

Group	User Type	Group	User Type
1	Weekday Users	4	Heavy Users
2	Efficient Users	5	Light Users
3	Night Owls	6	Medium Users



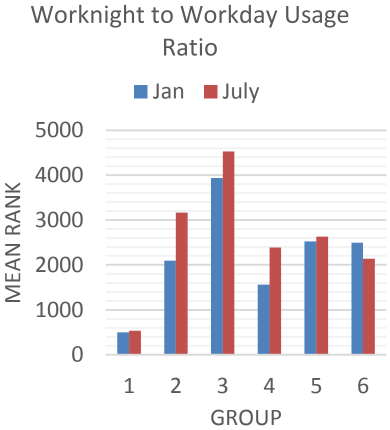


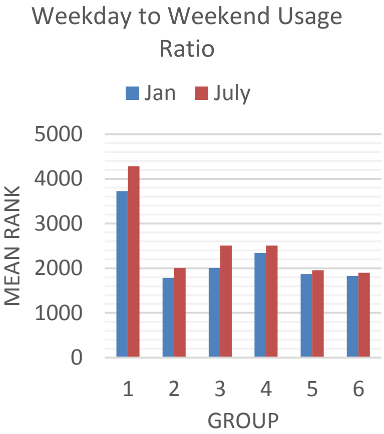
$$\text{Coincident Usage Ratio} = \frac{\text{Consumer's Coincident Peak Usage}}{\text{Peak System Load}}$$

$$\text{Coincident Peak Ratio} = \frac{\text{Consumer's Coincident Peak Usage}}{\text{Consumer's Peak Consumption}}$$

Jan vs July Usage Behavior

Group	User Type	Group	User Type
1	Weekday Users	4	Heavy Users
2	Efficient Users	5	Light Users
3	Night Owls	6	Medium Users





Worknight to Workday

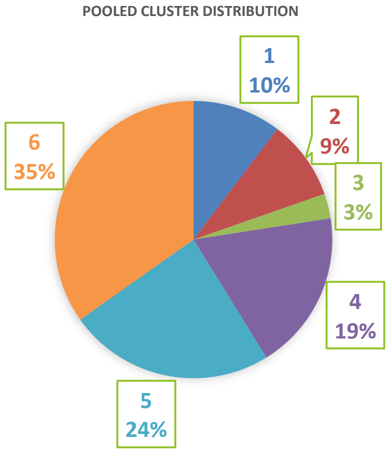
$$\frac{\text{Worknight Total Usage}}{\text{Workday Total Usage}}$$

Weekday to Weekend

$$\frac{\text{Weekday Total Usage}}{\text{Weekend Total Usage}}$$

Cluster Distribution

Group	User Type	Group	User Type
1	Weekday Users	4	Heavy Users
2	Off-Peak Users	5	Light Users
3	Night Owls	6	Medium Users



Groups /Year	2011	2012	2013	2014	2015	Average
1	5%	11%	11%	11%	15%	10%
2	11%	11%	9%	8%	8%	9%
3	2%	3%	3%	3%	3%	3%
4	19%	17%	19%	19%	18%	19%
5	23%	22%	22%	29%	24%	24%
6	39%	36%	36%	30%	32%	35%

Example: Impact on System Peak by Different Customer Types

Effect of adding **1** “Heavy User” is equivalent to…..

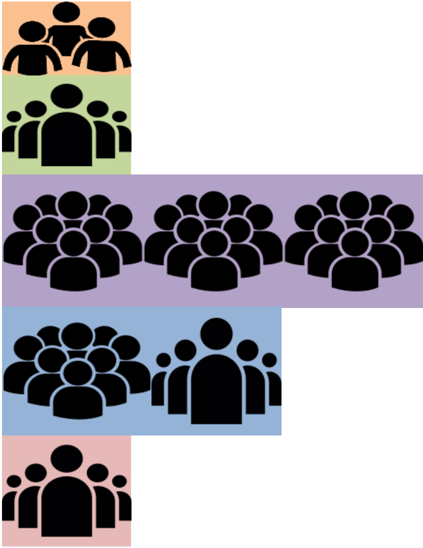
7 “Weekday Users”

58 "Efficient Users”

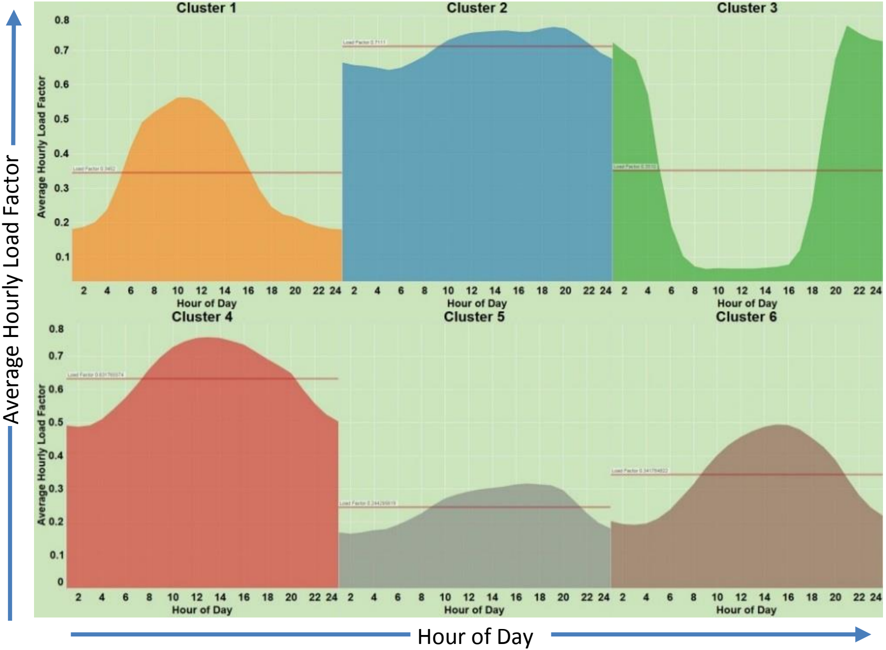
361 “Night Owls”

162 “Light Users”

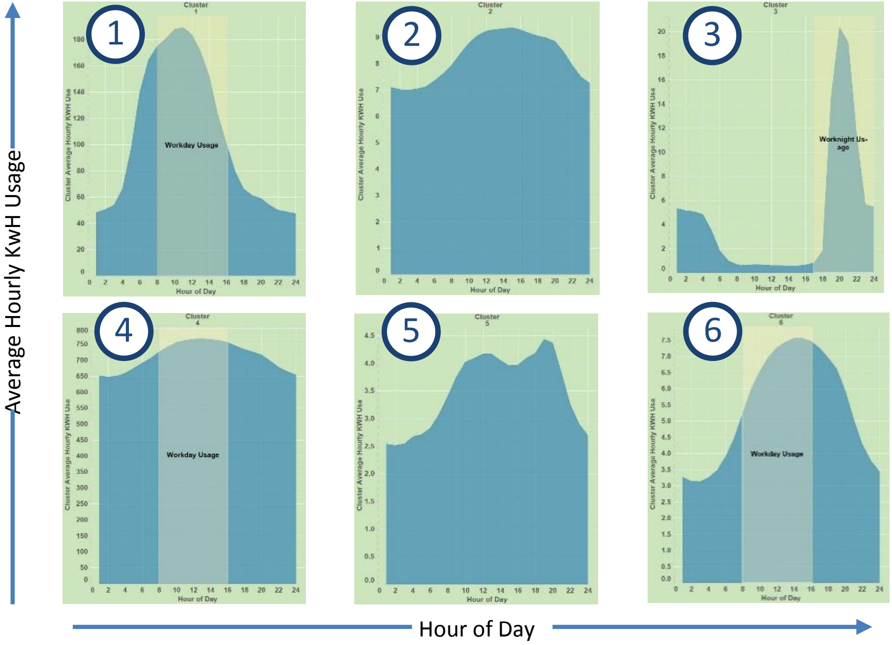
58 “Medium Users”



Cluster Load Factor Profiles



Workday vs. Worknight Cluster Profiles



Application: Estimating System Peak Usage

NOVEC can use our cluster analysis to identify future customers and their predicted impact on peak system load.

$$\text{Coincident Usage Ratio} = \frac{\text{Consumer's Energy Usage}}{\text{Peak System Load}}$$

Group	Coincident Usage Ratio	Lower 95% CI	Upper 95% CI
1	8.54E-05	6.81E-05	1.03E-04
2	1.10E-05	9.31E-06	1.26E-05
3	1.75E-06	2.03E-08	3.48E-06
4	100*6.31E-04 = 0.0631 6.31%	100*3.11E-04 = 0.031 3.1%	100*9.52E-04 = 0.095 9.5%
5	3.90E-06	2.60E-06	5.20E-06
6	1.10E-05	8.52E-06	1.34E-05

Application: Group with Minimal Impact to Peak

If NOVEC saw an increase of 1000 customers in group 3 (Night Owls), the peak system load would increase according to table.

Coincident Usage Ratio

$$\frac{\text{Consumer's Energy Usage}}{\text{Peak System Load}}$$

Group	Coincident Usage Ratio	Lower 95% CI	Upper 95% CI
1	8.54E-05	6.81E-05	1.03E-04
2	1.10E-05	9.31E-06	1.26E-05
3	1000*1.75E-06=1.75E-03 .175%	1000*2.03E-08=2.03E-05 .00203%	1000*3.48E-06=3.48E03 .348%
4	6.31E-04	3.11E-04	9.52E-04
5	3.90E-06	2.60E-06	5.20E-06
6	1.10E-05	8.52E-06	1.34E-05

Challenges

- Data available from NOVEC is not clean
    - Inconsistency in qualitative customer characteristics
    - Missing data points
- Stratified sampling of data over represented the population of heavy users
- Lessons Learned
    - Sampled roughly 500 of customer's data on housing properties to find only 30% correlation between house size and energy usage.
    - Customer clustering will not be consistent over different months because customer's energy usage behavior changes due to seasonality such as weather, vacation and holidays.

Conclusions

- 6 Different Customer Segmentation from January & July data based on consumer's pattern of energy consumption

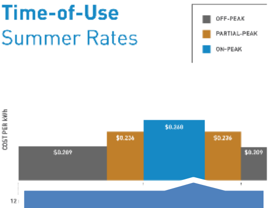
- Verified and validated the clusters using different techniques.

- Clustering of consumers will benefit NOVEC in:
    - New customer identification
    - Gaining knowledge of when certain consumers use more/less electricity

- Limitation in Analysis: Future improvements in technology, change in family dimension and new energy sources

Conclusions

- Can be implemented in the following NOVEC's system



Time-of-Use Pricing

Reduce customer's expenses by shifting your energy use to partial-peak or off-peak hours of the day

Load Management Program

Reduce peak electric demand by installing load management switches to AC and water heater and hold down power cost

Capacity Planning

Planning and building the capacity of electric distribution network to support its customer base and potential growth.

Recommendation for Future Work

- Add additional metrics to cluster customers
    - seasonality effects
    - holiday effects
    - Different time of day effects

- Should NOVEC pursue to use the existing data for segmentation, We recommend to apply importance sampling technique for detailed analysis of the stratified survey data.

- Suggest to perform a survey with fair representation of all types of customers with all levels of usage and a direct analysis of the survey results can be made.

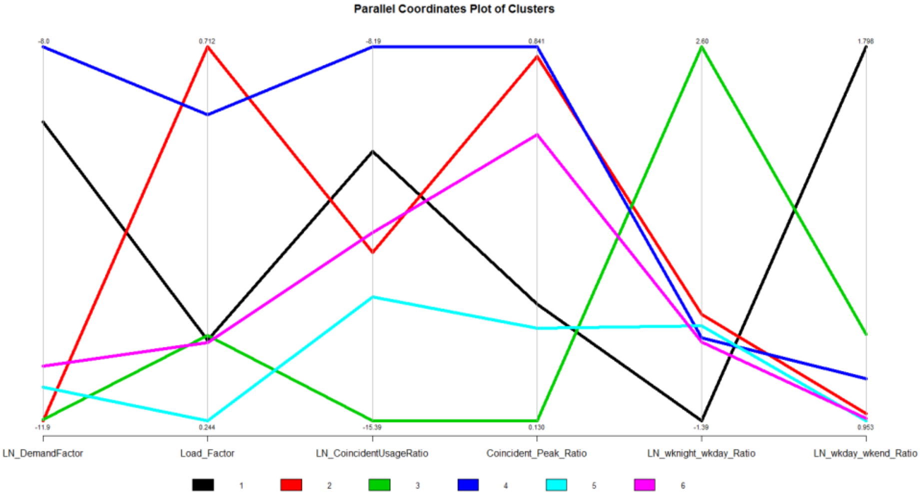
Questions ?

Special Thanks to…

OR/SYST 699 Project Class
Professor Hoffman
Professor Xu
NOVEC
GMU's Faculty

Backup slides

Characterizing the Clusters

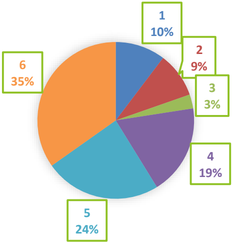


Characterizing the Clusters

Variable	Description
Demand Factor	July Peak / Peak System Load
Load Factor	July Avg / July Peak
Coincident Usage Ratio	Coincident Usage / Peak System Load
Coincident Peak Ratio	Coincident Usage / July Peak

Group	Demand Factor	Load Factor	Coincident Usage Ratio	Coincident Peak Ratio	Worknight vs Workday Usage Ratio	Weekday vs Weekend Usage Ratio
1	3.07E-04	0.34	8.54E-05	0.35	0.31	7.36
2	1.31E-05	0.71	1.10E-05	0.82	0.84	2.66
3	4.10E-05	0.35	1.75E-06	0.13	55.36	3.84
4	7.12E-04	0.63	6.31E-04	0.84	0.63	2.90
5	1.85E-05	0.24	3.90E-06	0.31	0.78	2.68
6	1.71E-05	0.34	1.10E-05	0.67	0.61	2.66

CLUSTER DISTRIBUTION



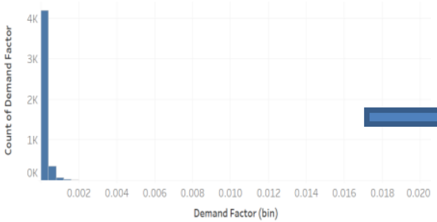
- Cluster 1: highest Weekday to Weekend Usage ratio. "Weekday Users" (i.e. Weekday Businesses)
- Cluster 2: highest Load Factor & High Coincident Peak Ratio. "Consistent System-aligned Users"
- Cluster 3: highest Weeknight to Weekday Usage Ratio. "Night time users" (i.e. Night-owls)
- Cluster 4: highest Demand, Coincident Usage, and Coincident Peak. "Heavy System-aligned users"
- Cluster 5: lowest Load Factor & lowest Weekday to Weekend Usage. "Light Inconsistent Weekend Users"
- Cluster 6 has medium Coincident Usage and medium Coincident Peak Usage. "Medium Weekend users"

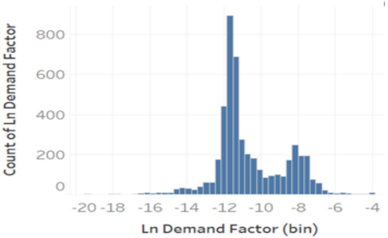
Characterizing Clusters - Jan vs. July Usage Behavior

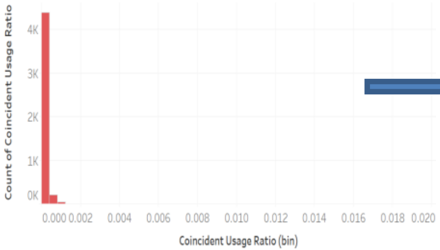
Number of clusters are still the same

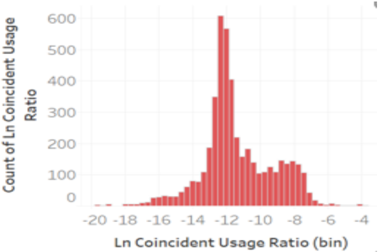
	Load Factor		Demand Factor		Coincident Usage Ratio		Coincident Peak Ratio		Weekday vs weekend Usage Ratio		Worknight vs Workday Usage Ratio	
Group	**Jan**	**July**	**Jan**	**July**	**Jan**	**July**	**Jan**	**July**	**Jan**	**July**	**Jan**	**July**
1	0.39	0.34	3.85E-04	3.07E-04	2.22E-04	8.54E-05	0.59	0.352	5.27	7.36	0.41	0.31
2	0.66	0.71	2.00E-05	1.31E-05	1.36E-05	1.10E-05	0.73	0.823	2.60	2.66	0.91	0.84
3	0.58	0.35	2.23E-05	4.10E-05	7.49E-06	1.75E-06	0.68	0.13	2.90	3.84	5.99	55.36
4	0.62	0.63	1.04E-03	7.12E-04	8.94E-04	6.31E-04	0.75	0.841	2.97	2.9	0.71	0.63
5	0.30	0.24	2.49E-05	1.85E-05	1.21E-05	3.90E-06	0.45	0.306	2.72	2.68	1.04	0.78
6	0.32	0.34	2.20E-05	1.71E-05	1.21E-05	1.10E-05	0.46	0.674	2.65	2.66	0.93	0.61

Demand Factor and Coincident Usage Ratio Variables



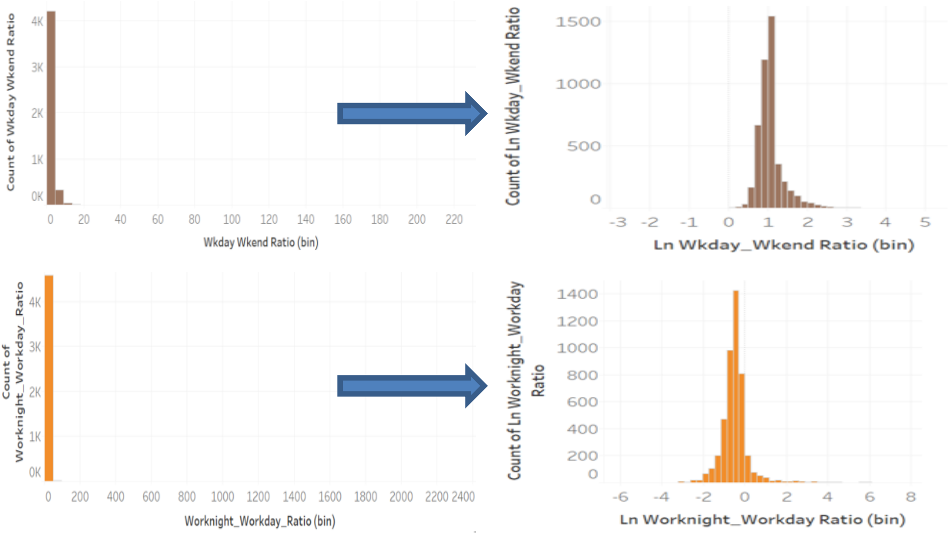






The histograms for these variables show heavily right-skewed distributions.
Data will need Log Transformation

Weekday-Weekend and Worknight-Workday Ratio Variables



The histograms for these variables show heavily right-skewed distributions.
Data will need Log Transformation

Correlation between Variables

	LN_wknight_wkday_Ratio	LN_wkday_wkend_Ratio	LN_DemandFactor	LN_CoincidentUsageRatio	Load_Factor	Coincident_Peak_Ratio
LN_wknight_wkday_Ratio	1	-0.33	-0.22	-0.33	0.06	-0.12
LN_wkday_wkend_Ratio		1	0.3	0.14	-0.06	-0.21
LN_DemandFactor			1	0.89	0.31	0.18
LN_CoincidentUsageRatio				1	0.47	0.53
Load_Factor					1	0.61
Coincident_Peak_Ratio						1

Final Dataset

4210 accounts from 2011 - 2015

First 10 rows of final dataset……

Account	Year	LN_DemandFactor	Load_Factor	LN_CoincidentUsageRatio	Coincident_Peak_Ratio	LN_wknight_wkday_Ratio	LN_wkday_wkend_Ratio
10082850	2012	-11.7535	0.199261	-12.6783	0.396606	-0.89576	0.859554
10082850	2015	-12.2443	0.25726	-12.7265	0.617367	-0.71967	0.857048
10527647	2012	-12.3759	0.207939	-15.5041	0.043799	-2.39967	2.733209
10666330	2012	-11.2132	0.194807	-12.7386	0.21752	-0.70438	0.871958
10723900	2014	-11.6977	0.36114	-12.4543	0.469283	-0.27415	1.04557
10733030	2013	-11.788	0.375203	-11.9948	0.813121	-0.90507	1.03771
10754460	2012	-7.74963	0.836301	-7.81337	0.938248	-0.29766	0.922956
10754464	2015	-6.98338	0.799851	-7.05504	0.930845	-0.25573	1.089342
10830230	2012	-16.7197	0.915939	-16.7557	0.964706	-0.28745	0.894123
1083080	2013	-12.1423	0.210786	-13.0914	0.387112	-0.33889	1.008229
………	…….	…….	…….	………	………..	………..	………..
………	…….	…….	…….	………	………..	………..	………..