Master of Science

Applied Data Science

Milestone Portfolio

Introduction

The online MS in Applied Data Science takes an interdisciplinary approach to ensure graduates emerge fully prepared to succeed as data science professionals.

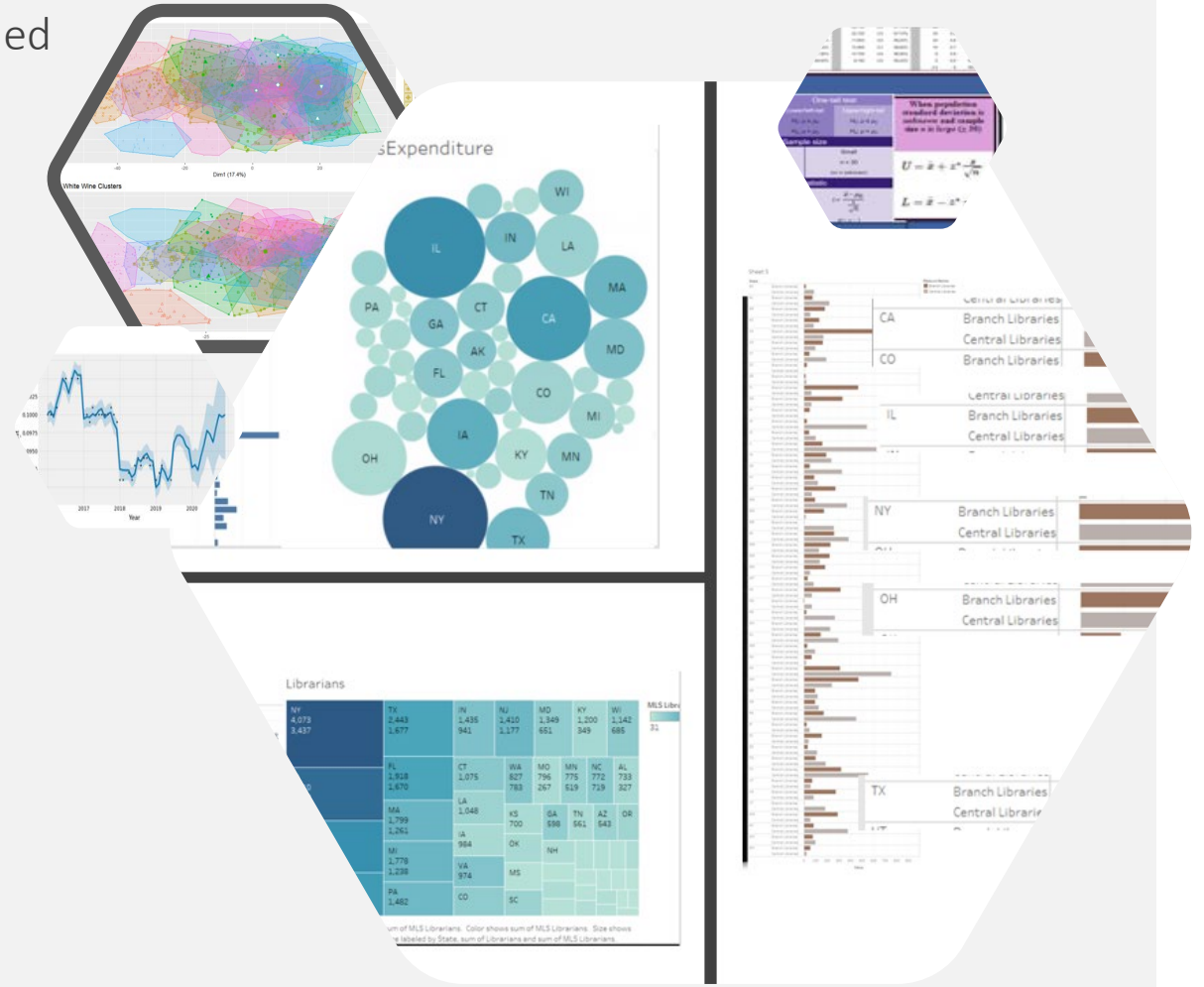
Successful students in the MS in Applied Data Science program will be able to:

- Collect and organize data
- Identify patterns in data via visualization, statistical analysis, and data mining
- Implement various business decisions based on data
- Develop alternative strategies based on the data
- Communicate data findings with managers, IT professionals, and other relevant professionals in their organization
- Recognize the ethical dimensions and privacy concerns of data science practice

Demonstrated Skills

While all courses contributed to the achievement of the defined learning objectives, reports from these four courses provide clear examples of my matriculation through the Syracuse Master of Science for Applied Data Science.

- MBC638: Data Analysis and Decision Making
- MAR653: Marketing Analytics
- IST707: Data Analytics
- IST718: Big Data Analytics



MBC638: Data Analysis and Decision Making

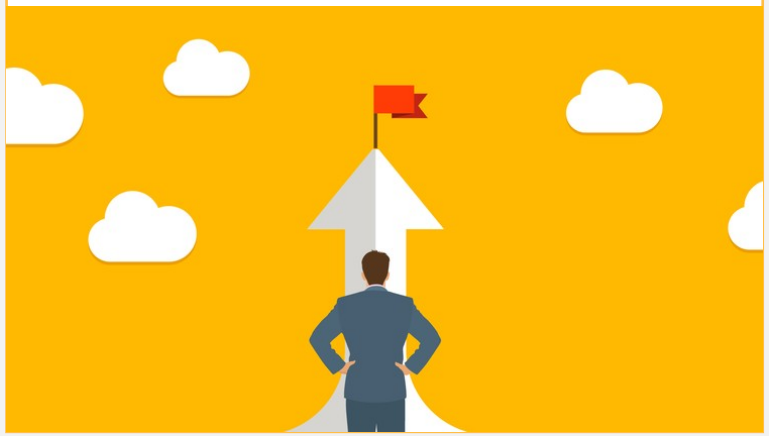
Time Management Process Improvement Project

Time Management Process Improvement Project

Project Description

Goal

- Follow DMAIC (Define, Measure, Analyze, Improve, Control) steps to improve upon an established issue, defined as a problem statement.

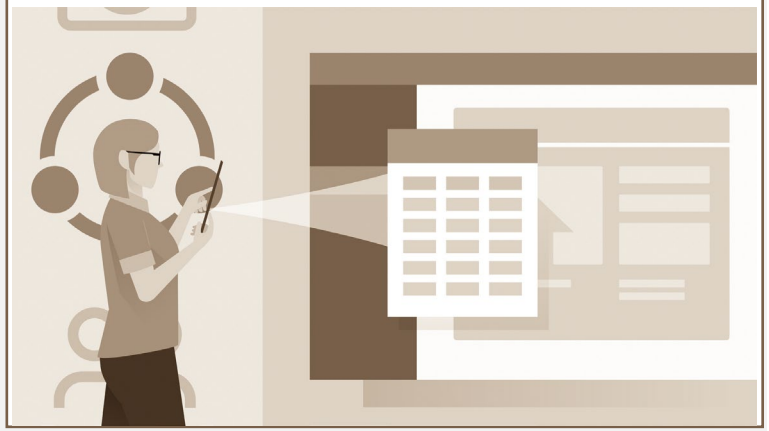


Data

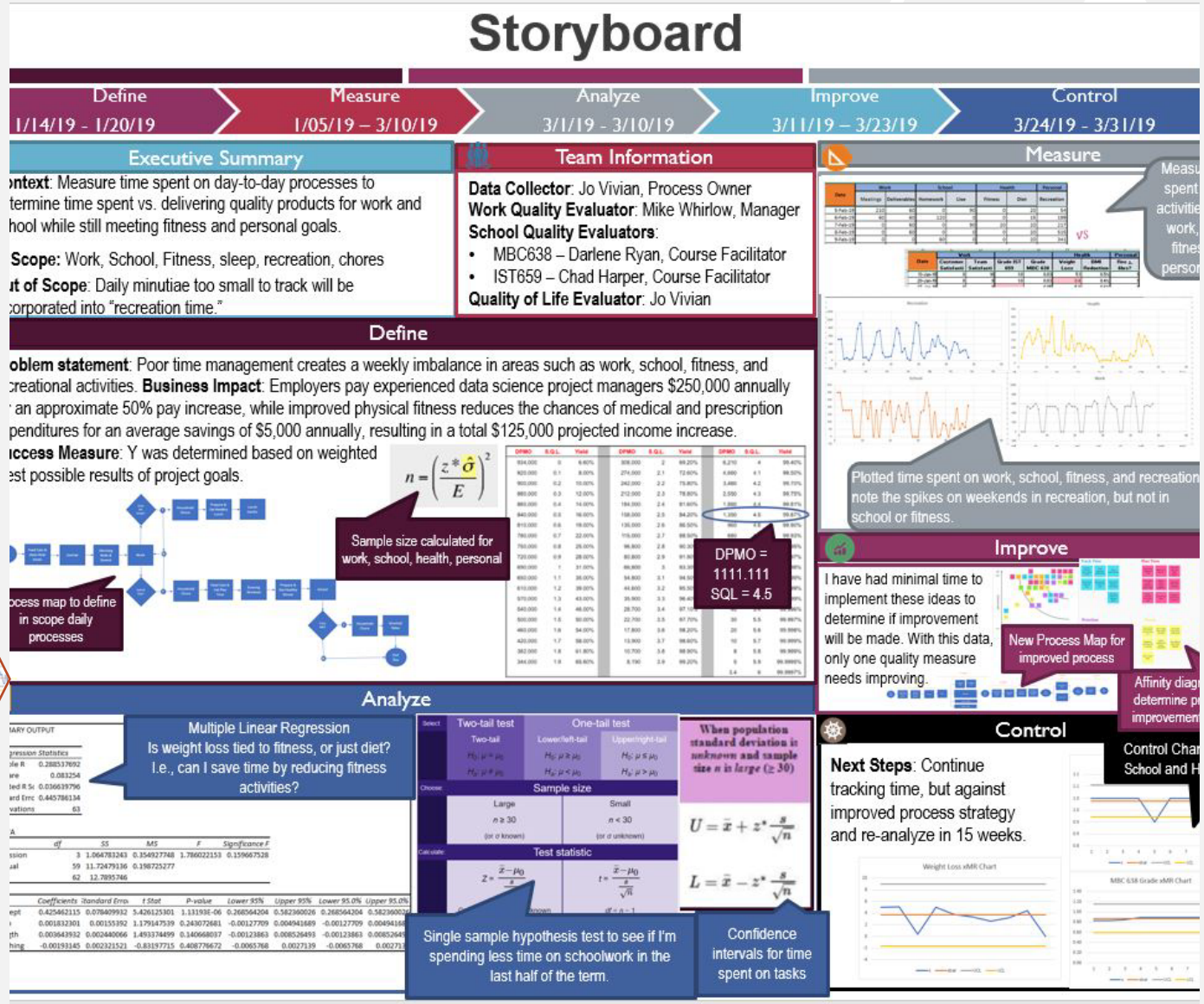
- Accessible data that could be collected over the course of ten weeks, including baseline data for comparison and extrapolated future data to measure control.

Deliverable

- A storyboard and presentation of the DMAIC process used to solve a business problem.

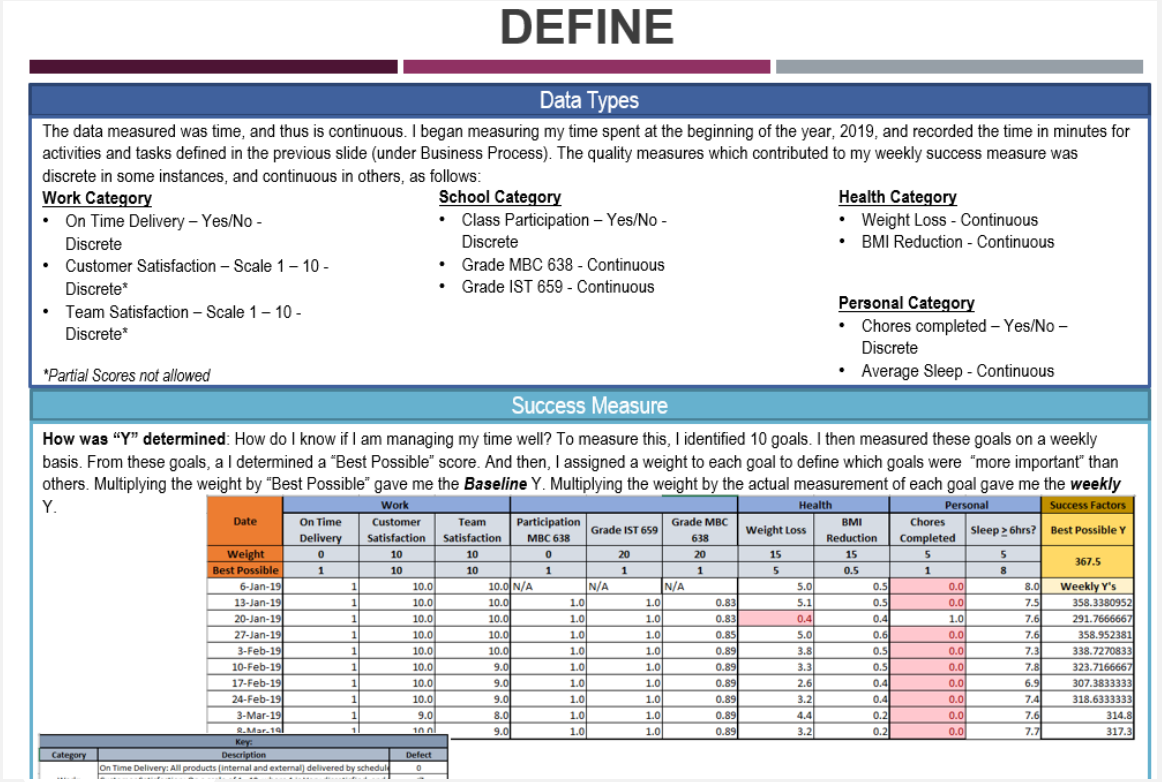


DMAIC Storyboard and Presentation



Define

Success Measure



DEFINE

Data Types

The data measured was time, and thus is continuous. I began measuring my time spent at the beginning of the year, 2019, and recorded the time in minutes for activities and tasks defined in the previous slide (under Business Process). The quality measures which contributed to my weekly success measure was discrete in some instances, and continuous in others, as follows:

Work Category
- On Time Delivery – Yes/No - Discrete
- Customer Satisfaction – Scale 1 – 10 - Discrete*
- Team Satisfaction – Scale 1 – 10 - Discrete*

School Category
- Class Participation – Yes/No - Discrete
- Grade MBC 638 - Continuous
- Grade IST 659 - Continuous

Health Category
- Weight Loss - Continuous
- BMI Reduction - Continuous

Personal Category
- Chores completed – Yes/No – Discrete
- Average Sleep - Continuous

*Partial Scores not allowed

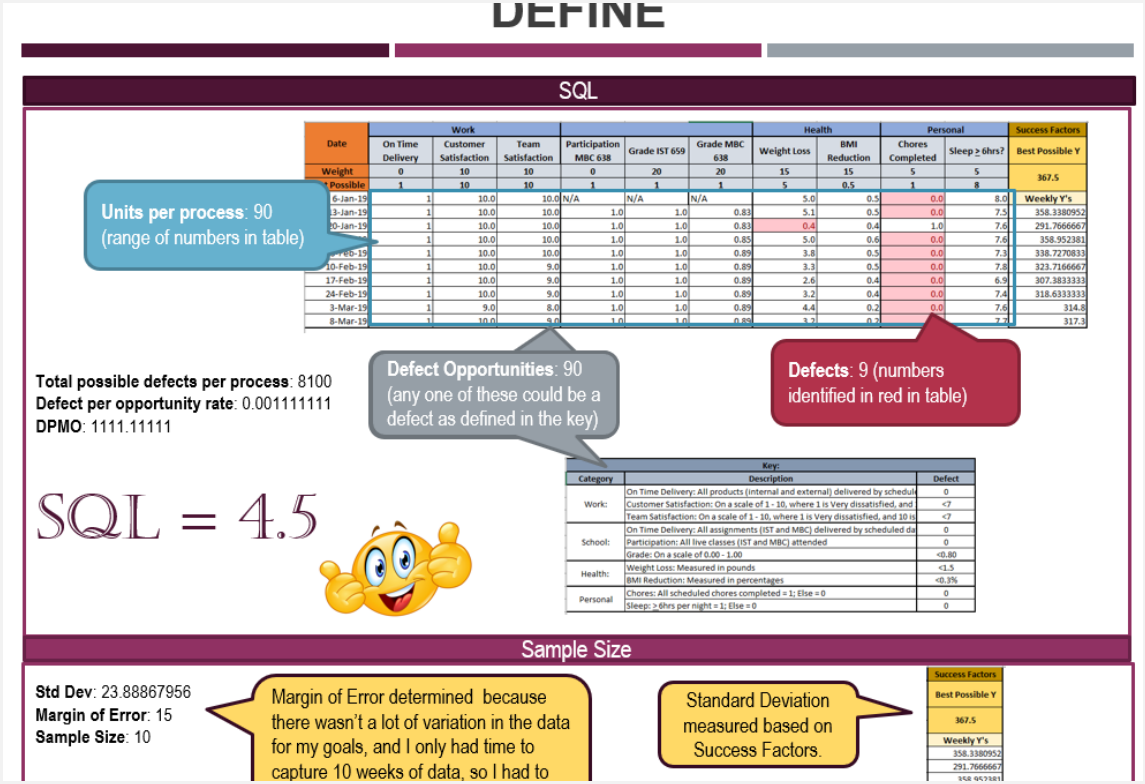
Success Measure

How was "Y" determined: How do I know if I am managing my time well? To measure this, I identified 10 goals. I then measured these goals on a weekly basis. From these goals, a I determined a "Best Possible" score. And then, I assigned a weight to each goal to define which goals were "more important" than others. Multiplying the weight by "Best Possible" gave me the ***Baseline*** Y. Multiplying the weight by the actual measurement of each goal gave me the ***weekly*** Y.

Date	Work						Health		Personal		Success Factors
	On Time Delivery	Customer Satisfaction	Team Satisfaction	Participation MBC 638	Grade IST 659	Grade MBC 638	Weight Loss	BMI Reduction	Chores Completed	Sleep ≥ 6hrs?	Best Possible Y
Weight	0	10	10	0	20	20	15	15	5	5	367.5
Best Possible	1	10	10	1	1	1	5	0.5	1	8	
6-Jan-19	1	10.0	10.0	N/A	N/A	N/A	5.0	0.5	0.0	8.0	Weekly Y's
13-Jan-19	1	10.0	10.0	1.0	1.0	0.83	5.1	0.5	0.0	7.5	358.3380952
20-Jan-19	1	10.0	10.0	1.0	1.0	0.83	0.4	0.4	1.0	7.6	291.7666667
27-Jan-19	1	10.0	10.0	1.0	1.0	0.85	5.0	0.6	0.0	7.6	358.952381
3-Feb-19	1	10.0	10.0	1.0	1.0	0.89	3.8	0.5	0.0	7.3	338.7270833
10-Feb-19	1	10.0	9.0	1.0	1.0	0.89	3.3	0.5	0.0	7.8	323.7166667
17-Feb-19	1	10.0	9.0	1.0	1.0	0.89	2.6	0.4	0.0	6.9	307.3833333
24-Feb-19	1	10.0	9.0	1.0	1.0	0.89	3.2	0.4	0.0	7.4	318.6333333
3-Mar-19	1	9.0	8.0	1.0	1.0	0.89	4.4	0.2	0.0	7.6	314.8
8-Mar-19	1	10.0	9.0	1.0	1.0	0.89	3.2	0.2	0.0	7.7	317.3

Key:		
Category	Description	Defect
Work:	On Time Delivery: All products (internal and external) delivered by schedule	0
	Customer Satisfaction: On a scale of 1 - 10, where 1 is Very dissatisfied, and	<7
	Team Satisfaction: On a scale of 1 - 10, where 1 is Very dissatisfied, and 10 is	<7
School:	On Time Delivery: All assignments (IST and MBC) delivered by scheduled da	0
	Participation: All live classes (IST and MBC) attended	0
	Grade: On a scale of 0.00 - 1.00	<0.80
Health:	Weight Loss: Measured in pounds	<1.5
	BMI Reduction: Measured in percentages	<0.3%
Personal	Chores: All scheduled chores completed = 1; Else = 0	0
	Sleep: ≥ 6hrs per night = 1; Else = 0	0

SQL



DEFINE

SQL

Date	Work						Health		Personal		Success Factors
	On Time Delivery	Customer Satisfaction	Team Satisfaction	Participation MBC 638	Grade IST 659	Grade MBC 638	Weight Loss	BMI Reduction	Chores Completed	Sleep ≥ 6hrs?	Best Possible Y
Weight	0	10	10	0	20	20	15	15	5	5	367.5
Best Possible	1	10	10	1	1	1	5	0.5	1	8	
6-Jan-19	1	10.0	10.0	N/A	N/A	N/A	5.0	0.5	0.0	8.0	Weekly Y's
13-Jan-19	1	10.0	10.0	1.0	1.0	0.83	5.1	0.5	0.0	7.5	358.3380952
20-Jan-19	1	10.0	10.0	1.0	1.0	0.83	0.4	0.4	1.0	7.6	291.7666667
27-Jan-19	1	10.0	10.0	1.0	1.0	0.85	5.0	0.6	0.0	7.6	358.952381
3-Feb-19	1	10.0	10.0	1.0	1.0	0.89	3.8	0.5	0.0	7.3	338.7270833
10-Feb-19	1	10.0	9.0	1.0	1.0	0.89	3.3	0.5	0.0	7.8	323.7166667
17-Feb-19	1	10.0	9.0	1.0	1.0	0.89	2.6	0.4	0.0	6.9	307.3833333
24-Feb-19	1	10.0	9.0	1.0	1.0	0.89	3.2	0.4	0.0	7.4	318.6333333
3-Mar-19	1	9.0	8.0	1.0	1.0	0.89	4.4	0.2	0.0	7.6	314.8
8-Mar-19	1	10.0	9.0	1.0	1.0	0.89	3.2	0.2	0.0	7.7	317.3

Units per process: 90 (range of numbers in table)

Defect Opportunities: 90 (any one of these could be a defect as defined in the key)

Defects: 9 (numbers identified in red in table)

Total possible defects per process: 8100
Defect per opportunity rate: 0.001111111
DPMO: 1111.11111

SQL = 4.5

Key:		
Category	Description	Defect
Work:	On Time Delivery: All products (internal and external) delivered by schedul	0
	Customer Satisfaction: On a scale of 1 - 10, where 1 is Very dissatisfied, and	<7
	Team Satisfaction: On a scale of 1 - 10, where 1 is Very dissatisfied, and 10 is	<7
School:	On Time Delivery: All assignments (IST and MBC) delivered by scheduled da	0
	Participation: All live classes (IST and MBC) attended	0
	Grade: On a scale of 0.00 - 1.00	<0.80
Health:	Weight Loss: Measured in pounds	<1.5
	BMI Reduction: Measured in percentages	<0.3%
Personal	Chores: All scheduled chores completed = 1; Else = 0	0
	Sleep: ≥ 6hrs per night = 1; Else = 0	0

Sample Size

Std Dev: 23.88867956
Margin of Error: 15
Sample Size: 10

Margin of Error determined because there wasn't a lot of variation in the data for my goals, and I only had time to capture 10 weeks of data, so I had to

Standard Deviation measured based on Success Factors.

Success Factors
Best Possible Y
367.5
Weekly Y's
358.3380952
291.7666667

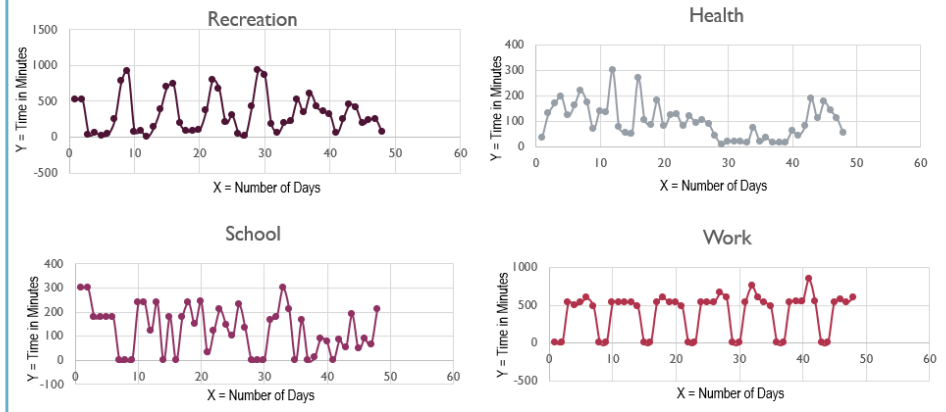
Measure & Analyze

Measure

Measure

Time Series Plots

The seemingly erratic nature of each of these plots makes sense when you consider the time was measured over the course of 10 weeks, including weekends, when naturally, recreation time increases, and work time decreases. Also, because work is constant at 8-9 hours daily during these 10 weeks, there is little to be learned from either the work or recreation charts. What is more informative are the peaks and valleys in school and the long series of minimal time spent on health. Both of these warrant further investigation and indicate areas in need of improvement.



Analyze

Analyze

One Sample One Tail Hypothesis Test

Hypothesis: I have found the best use of my school time, and am spending less time on schoolwork in the last half of the term.

Ho: μ ≥ 131 minutes per day
Ha μ < 131 minutes per day
N = 28 days
Standard Deviation = 110.22 minutes per day
$\bar{x}$ = 127 minutes per day
α = 0.05

$$Z = \frac{127 - 131}{\frac{110.22}{5.29}}$$

Z = -0.19

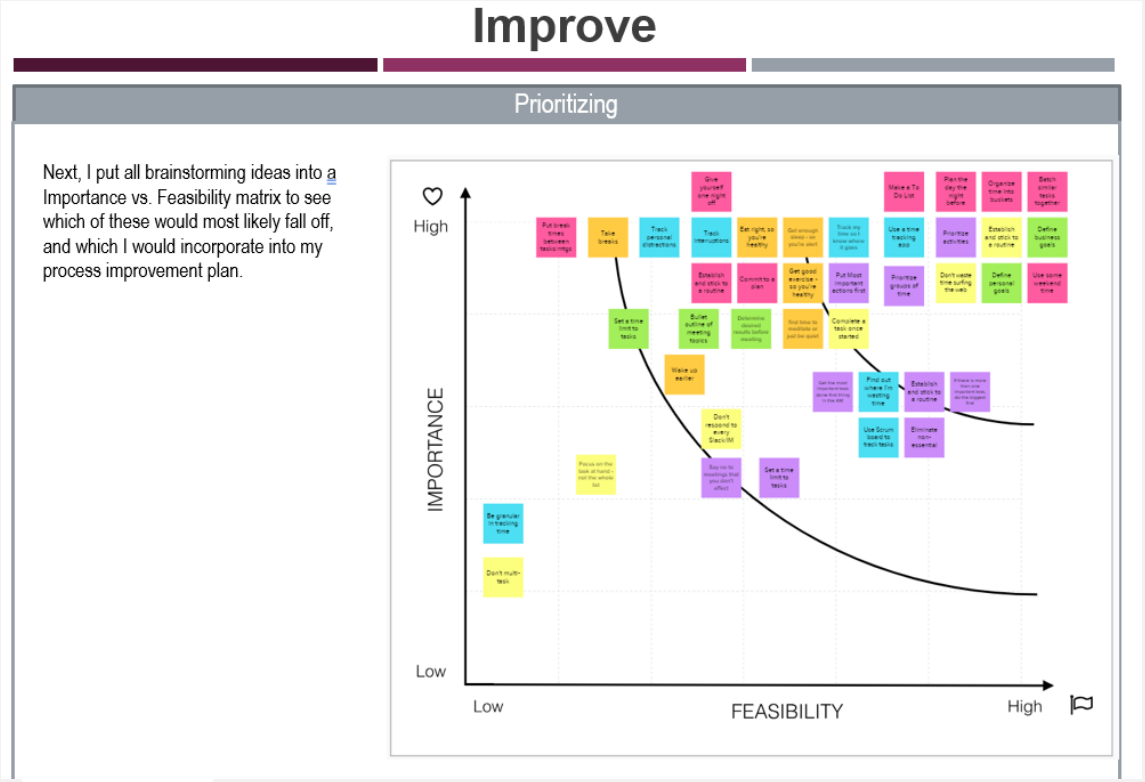
p = 0.4247

Drat.
P isn't low.
Gotta keep Ho

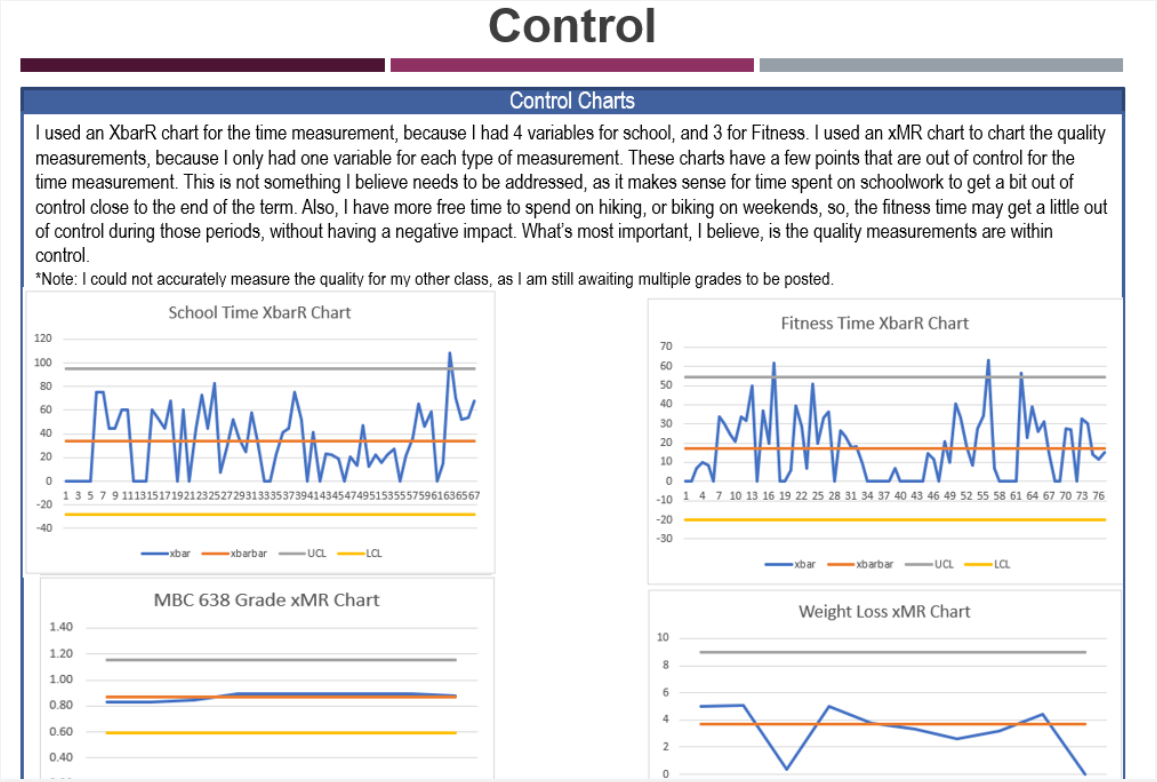
Select:	Two-tail test	One-tail test	
	Two-tail	Lower/left-tail	Upper/right-tail
	H_0: $\mu = \mu_0$	H_0: $\mu \geq \mu_0$	H_0: $\mu \leq \mu_0$
	H_a: $\mu \neq \mu_0$	H_a: $\mu < \mu_0$	H_a: $\mu > \mu_0$
Choose:	Sample size		
	Large	Small	
	$n \geq 30$	$n < 30$	
	(or σ known)	(or σ unknown)	
Calculate:	Test statistic		
	$Z = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$	$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$	
	Can replace s with σ if known	$df = n - 1$	
Identify:	p-value		
	Two-tail	Lower/left-tail	Upper/right-tail
	p = 2 × area past Z or t	p = area left of Z or t	p = area right of Z or t

Improve & Control

Improve



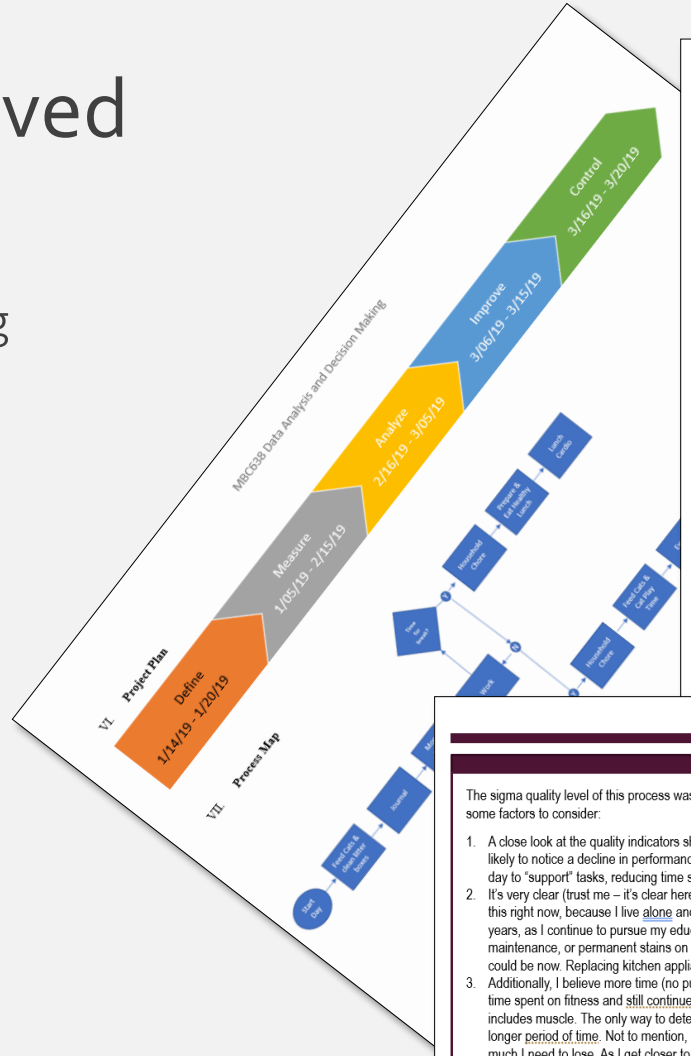
Control



Learning Achieved

Increased understanding

- DMAIC Process
- Collecting Data
- Analyzing Data
- Detecting Patterns
- Developing Strategies



Problem Definition Worksheet – Feedback – Jo Vivian

Please note my comments before moving on to the *Measure* phase.

Content Requirements	Possible Points	Points Earned	Comments
A) **Problem statement**: Is it clear, concise and stated as a problem? Is there evidence this is a problem?	2.0	**2**	Good
B) **Business impact**: Is the business case quantified? Has a measure of success been identified?	2.0	**.5**	This seems more like an objective than a business case. Need to quantify the business case. What is it worth in $ for you to do this project? You would need to make some assumptions. Where is the main measure of success?
C) **Goals**: Clearly stated goal(s)?	1.0	**1**	You have listed many great goals in many areas of your life. Is there a main metric t important to yo
D) **Project scope**: Identified?	1.0	**.8**	This seems to l steps. The first included. Is any
E) **Team**: Identified?	1.0	**1**	Yes
F) **Project plan**: rough timeline or approximate dates/time per each DMAIC step?	1.0	**1**	Good
G) **Process map**: Has clearly identified the start of the current process? Has clearly identified	2.0	**2**	Good job

Process Improvement Project – Feedback – Jo Vivian

Content Requirements	Possible Points	Points Earned	Comments
A) An executive summary is provided in the storyboard format including: Is the storyboard presented in 1 PowerPoint slide? Follows DMAIC? Are tools/graphs/charts used and clearly visible? Do they support findings and conclusions Are arrows, call-out boxes, etc. used to summarize, highlight questions and key learnings? Are expected	5	5.00	1 slide, follows DMAIC, tools used, Callouts used, next steps identified. Very nice job on this! High quality!
B)Is it a cohesive presentation opening with the business process and problem statement? The back-up slides (5-15) detail and support the storyboard content.	2	2.00	Process to be addressed and problem statement are included. Backup slides support the storyboard.
C) Was the success measure clearly identified, operationally defined and baseline identified? (Was the data identified as continuous or discrete, includes SQL?)	3	2.80	Success measure was clearly defined and explained. Baseline was included. It would be good to have averaged your baseline weeks to identify one overall number for your main y. Data was identified as discrete and continuous. SQL is included and work is shown. I am not sure how the total possible defects is 8100 (vs. 100). I see 10 weeks and 10 elements where each one could be a defect and there are 10 red cells within the boxed section of the chart. 10/100 would give a much lower SQL. This resonates with the chart as well. Consider that at a 4.5 sigma level we would have just 1350 defects out of a million
D) Was the data measurement plan or data stratification tree included?	1	1.00	Data Measurement Plan is included
E) Was the data collection method identified?	1	1.00	
F) Was there rationale for the sample size taken? Use of the formula? Is there any reference to measurement error and how to minimize?	1	0.50	Sample size formula was used. Margin of Error (the amount of shift in your main y you want to be able to detect) has to be in the same units as your main Y, so it is points on your weighted averaged of your main Y. An E of 15 is actually quite small when the max is 267 points
G) Are at least 5 different tools and techniques clearly identified? Are the tools linked/ pertinent to the data analysis?	5	**5.00**	Process map Time series plots - nicely labelled Hypothesis test - good showing of your work and good conclusion.I am curious as to why you chose 131 minutes. Multiple linear regression - good conclusions. I wonderif sample size is playing into the results. Confidence intervals - nice work! Affinity diagram - love the ideas. Control charts - Nice! I see 4 weeks where you had a perfect score on the HV. Not sure why that control chart does not show any 10's. Excellent use of tools!
H) Does the data analysis clearly tie to the problem conclusion? Is the "discovery" clear to the reader?	2	**2.00**	Data analysis leads to conclusions. Only opportunity is in the SQL calculation.
Total possible 20 points	20	19.30	Very nice job on this project.

Conclusion

Final Analysis

The sigma quality level of this process was very high, which might seem to indicate it didn't actually need improvement. However, there are some factors to consider:

1. A close look at the quality indicators show a slight downward trend in customer satisfaction, especially internal customers, who are more likely to notice a decline in performance. Perhaps the most obvious, and easiest improvement is to allot a specific amount of time each day to "support" tasks, reducing time spent there, and increasing time spent on work products.
2. It's very clear (trust me – it's clear here at home as well) that household chores are being neglected. It's easy to put less significance on this right now, because I live alone and the only impact is on me (the cats don't care). But, this trend, over the course of the next 1.5 years, as I continue to pursue my education, could potentially result in bigger issues like appliances breaking down due to poor maintenance, or permanent stains on floors or furniture. This wasn't considered in the business impact during the definition, but I believe it could be now. Replacing kitchen appliances and washer/dryer could run as high as $5,000, and carpet cleaning services as high as $500.
3. Additionally, I believe more time (no pun intended) is need to analyze some factors of this process. For example, can I truly cut back on time spent on fitness and still continue to see weight loss? Or, will I encounter issues when the bulk of the weight loss is not fat, but includes muscle. The only way to determine that is to continue working out, but, perhaps at a reduced rate, and track weight loss over a longer period of time. Not to mention, it's very likely I have seen the level of success with weight loss at this stage, because I have so much I need to lose. As I get closer to my target weight, it's very possible diet alone will no longer provide the same degree of success

Next Steps:

1. The first step is to simply spend more time in the Improve phase. I only had about 2 weeks in this phase, which isn't nearly long enough to make any real determinations. Especially given a sample size of 10 weeks – and that's with a margin of error that was too large to begin with, but necessary given the time I had to measure this process. A margin of error of 5 gives me 88 weeks, which is just too long to wait. A margin of 10 gives 22 weeks. Still a pretty long time. I may have to just accept a large margin of error on this first round of improvements, and reassess in another 10 weeks.
2. Since the two hypothesis tests I ran indicated there wasn't a strong relationship between the activities I measured and the quality controls, I will consider other activities to measure, as well as look at other quality control measures – especially try to find some that have more variation and continuous data.

Final Conclusion:

Of the selected quality controls, there is only one (chores) that needs significant improvement. The rest need monitoring. Additional measurement and analysis are needed to determine if there really is a linear relationship between time spent on selected activities and the defined quality control measures.

MAR6653: Marketing Analytics

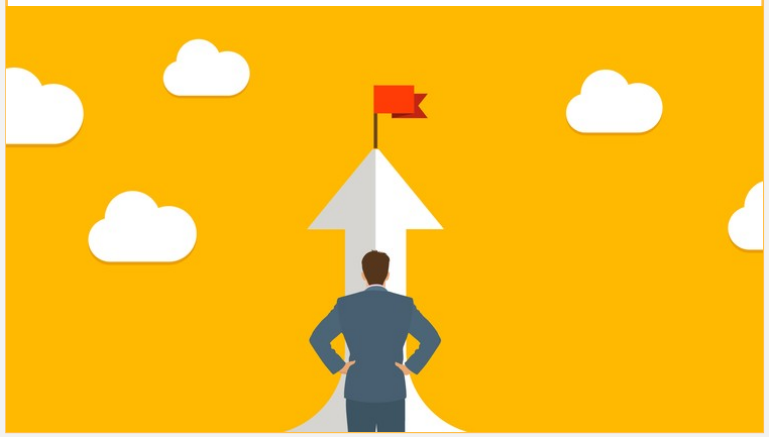
Solving the Business Problem to Increase Visits to Public Libraries

Using Marketing Analytics to Increase Library Visits Project

Project Description

Goal

- Address a legitimate business problem for an existing brand or product.

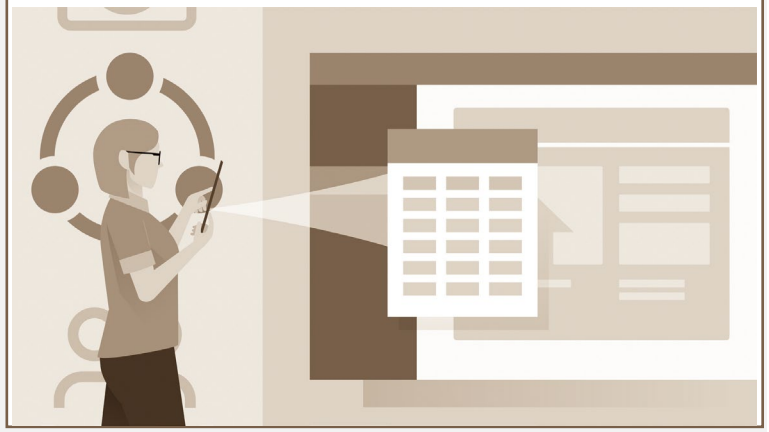


Data

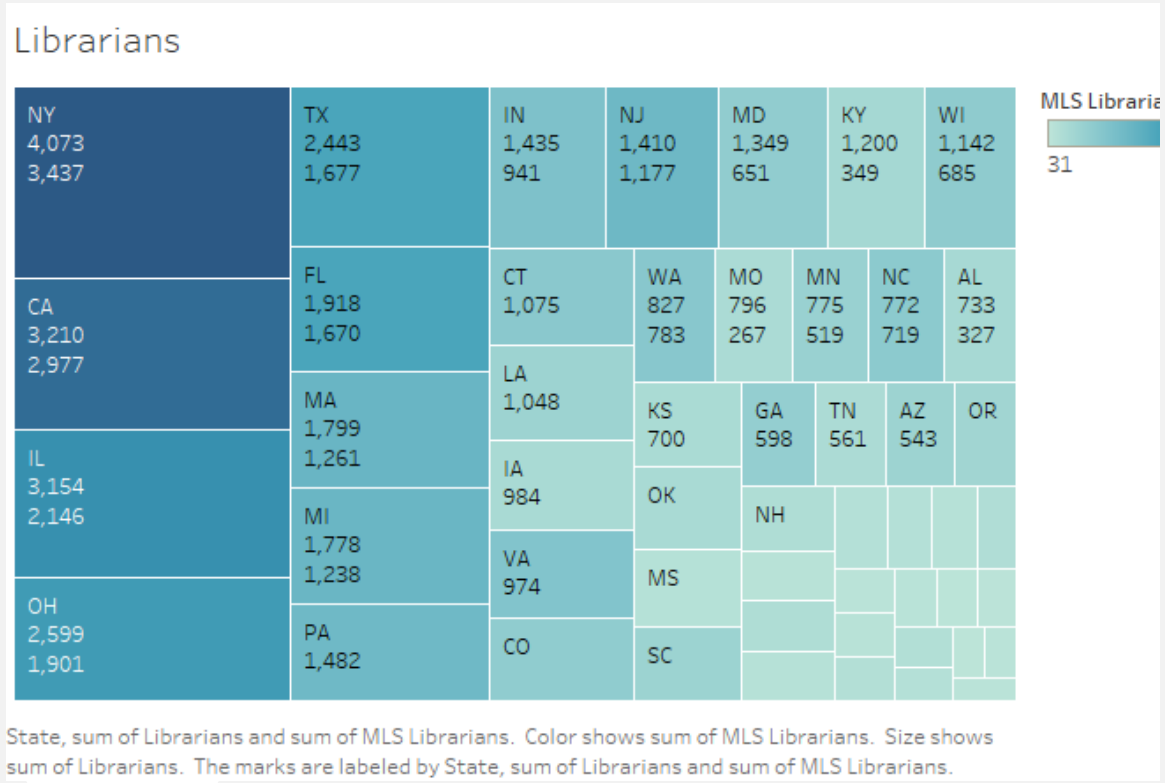
- 2016 Public Library survey conducted by the Institute of Museum and Library Services which includes all public libraries identified by state library administrative agencies in the 50 states and the District of Columbia.

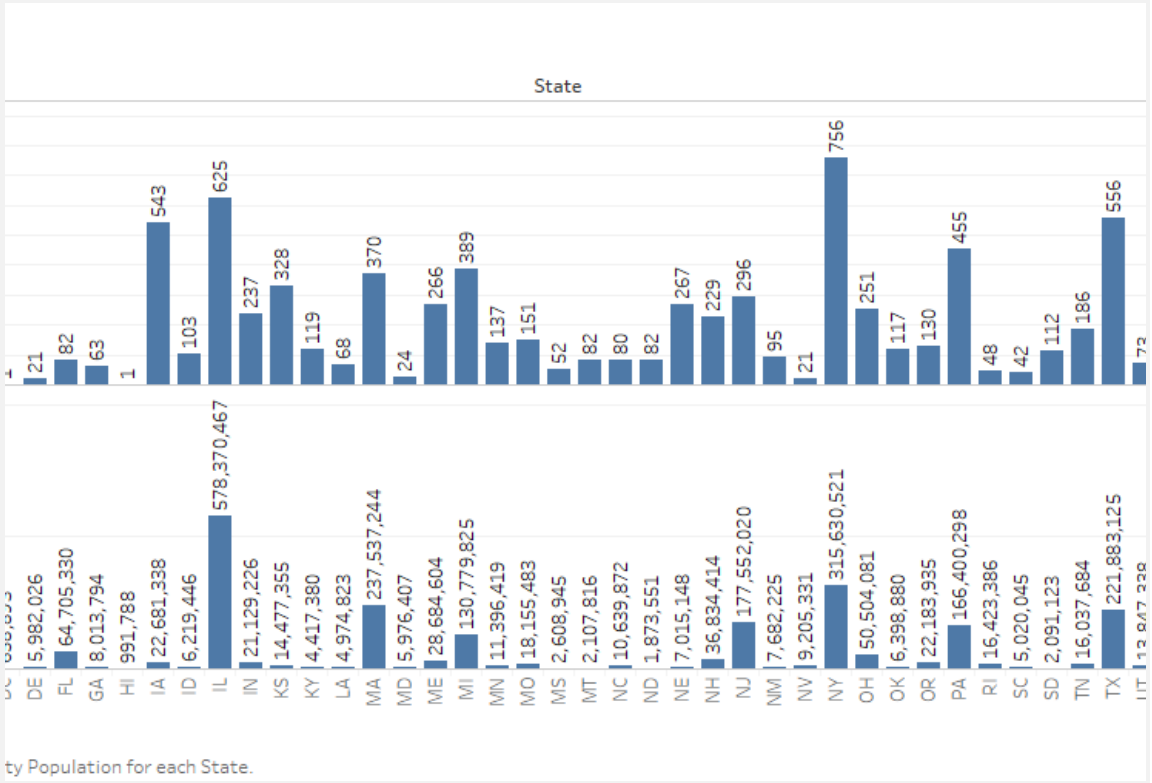
Deliverable

- A presentation of recommendations on how to increase the number of visits to public libraries, thus increasing revenue and funding.

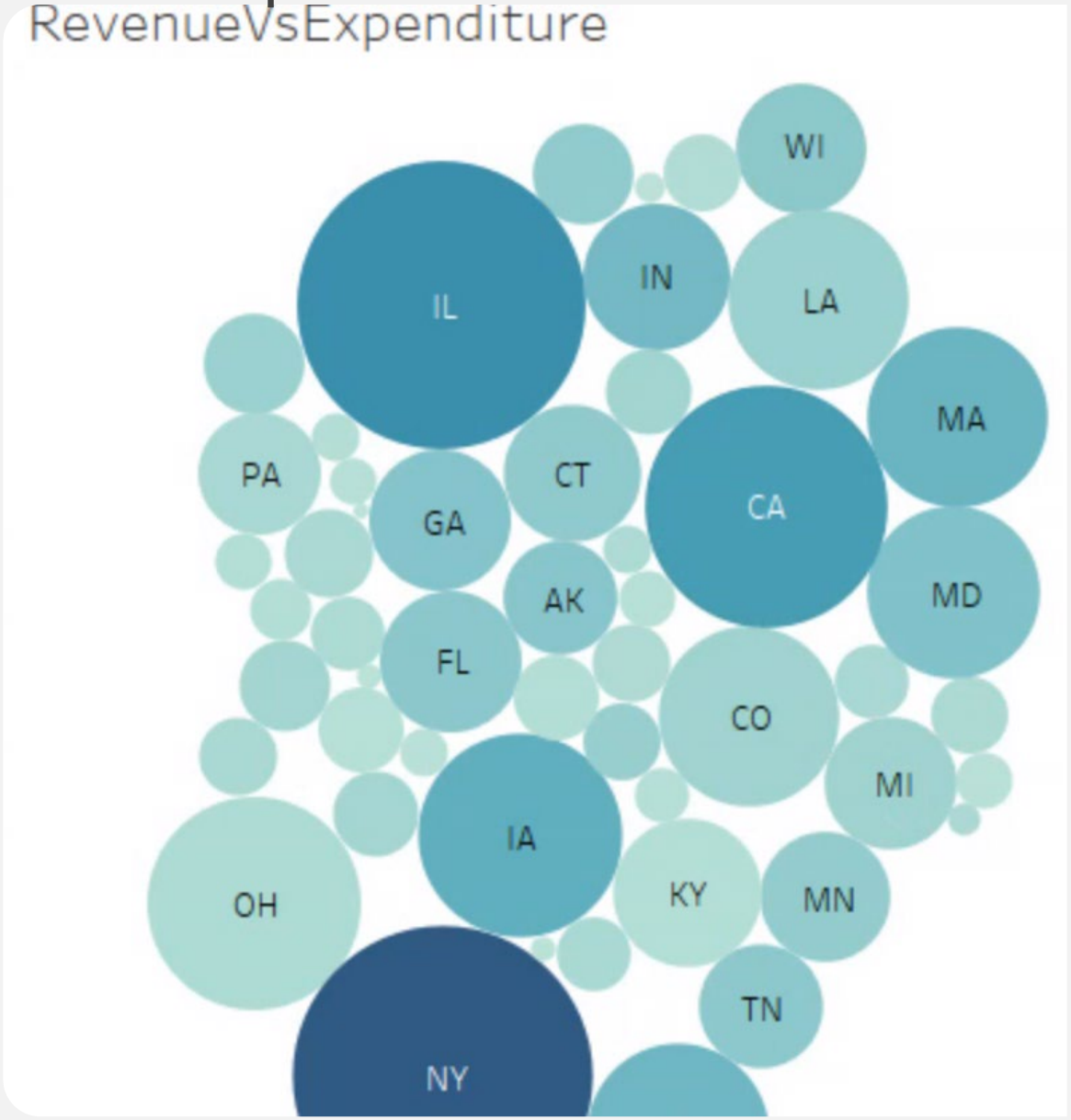


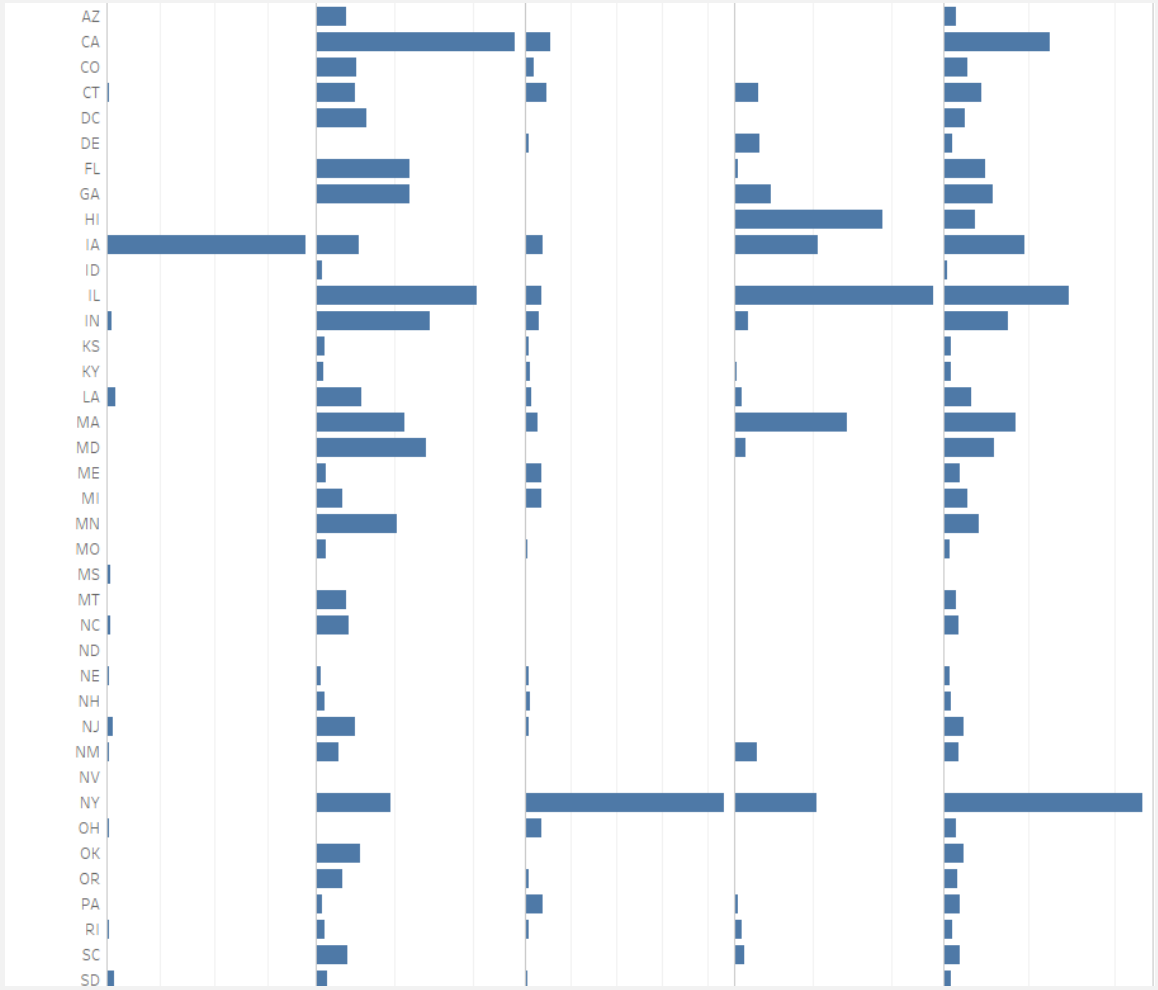
Data Exploration





Data Exploration





K-Means

To cluster Libraries by Size

- 4-Clusters based on
  - Population
  - Revenue
  - Staff
  - Number of Central Libraries

Class centroids:

Class	County Population	ice Population Without Duplic	Central Libraries	Total Staff	Total Operating Revenue	Sum of weights	Within-class variance
1	288388.334	16567.803	0.988	7.424	552923.003	8765.000	1326374359992.020
2	1186918.897	208808.023	0.861	96.615	8814354.175	388.000	20981303924074.800
3	1258279.704	853238.718	0.704	377.614	37974968.225	71.000	178632591150906.000
4	4373992.875	2537876.500	0.750	1143.215	138546669.875	8.000	4121025223494720.000

	Absolute	Percent
Within-class	6620672676052.600	18.26%
Between-classes	29636671836926.500	81.74%
Total	36257344512979.100	100.00%

Class	County Population	Service Population	Central Libraries	Total Staff	Total Operating Revenue	# of braries	ClusterName based on Service Population
1	288388.334	16567.803	0.988	7.424	552923.003	8765.000	Small- Comparitively small target audience, lesser operating revenue but represents 94% of US population. It is seen that they have fewer staff to manage libraries
2	1186918.897	208808.023	0.861	96.615	8814354.175	388.000	Medium- This cluster has medium Operating revenue and staffed at an average of 96 and represents 4% of US library population
3	1258279.704	853238.718	0.704	377.614	37974968.225	71.000	Large - Well staffed and has an operating revenue of $37million. Represents ~1% of US library population
4	4373992.875	2537876.500	0.750	1143.215	138546669.875	8.000	XL - Superlative numbers - high staffing,$138million operating revenue but it represents less than 0.1% of the US library population

Linear Regression

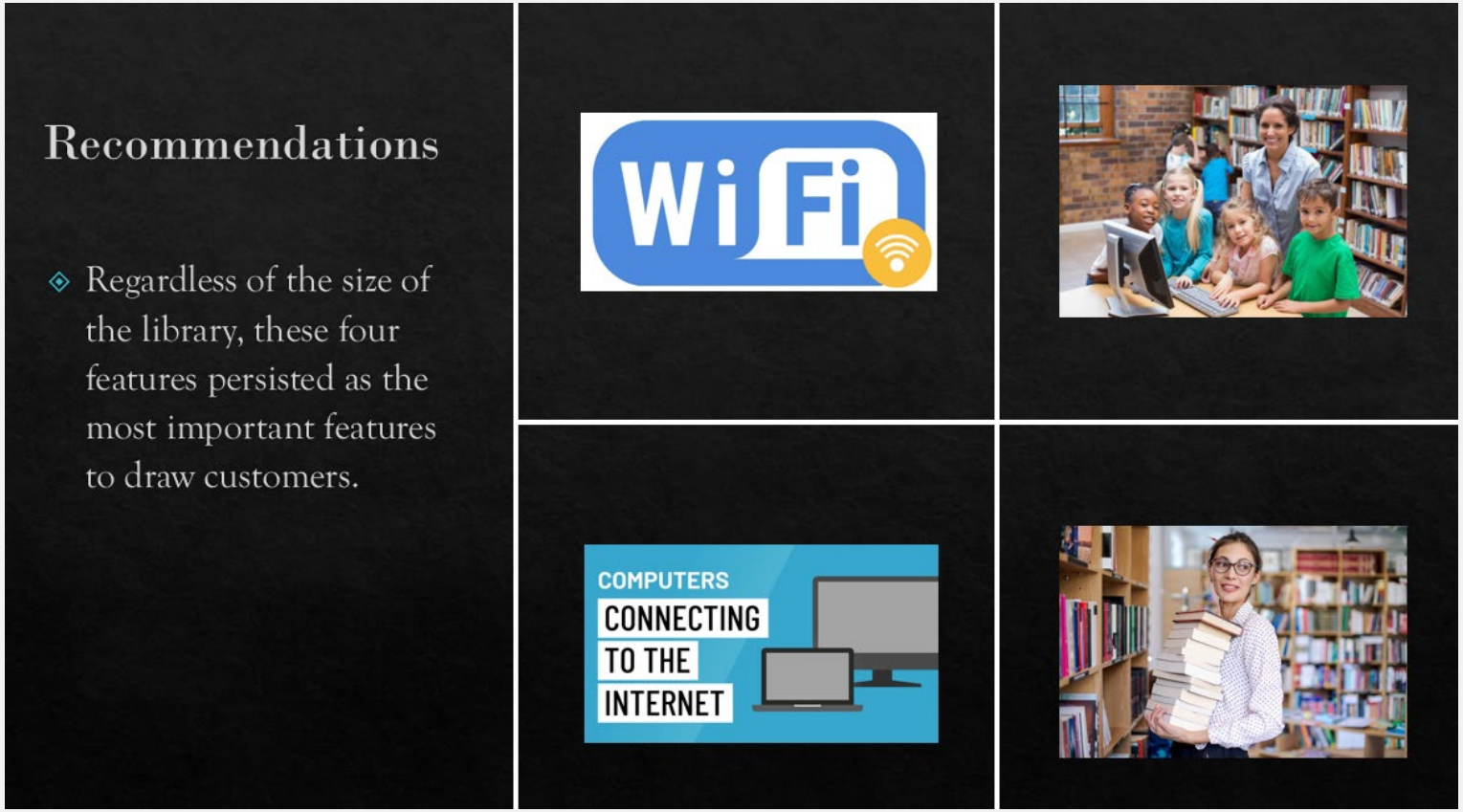
Linear Regression	Large
Intercept	812998.69
Total Staff	1966.56
Video	2.55
Internet Computer Use	1.036
Library Programs	37.11
Circulation Transactions	0.087
Adjusted R-Squared	68%

Linear Regression	Small	Medium
Intercept	3180.25	-43464.94
Service Population Without Dup	0.44	0.46
Total Staff	2743.35	1499.10
Total Operating Revenue	0.02	0.03
Registered Users	1.10	5.23
Library Programs	20.53	2.00
Internet Computer Use	0.52	27.44
Wireless Internet Sessions	0.01	0.50
Hours Open	-1.06	0.02
Total Collection Expenditures	0.20	0.05
R- Squared	**81%**	**83%**

Learning Achieved

Increased understanding

- Solving business problems with Marketing Analytics using
  - Market response models
  - Product recommendation systems
  - Resource allocation to improve a business
- Using data visualization in Stakeholder communications



IST707: Data Analytics

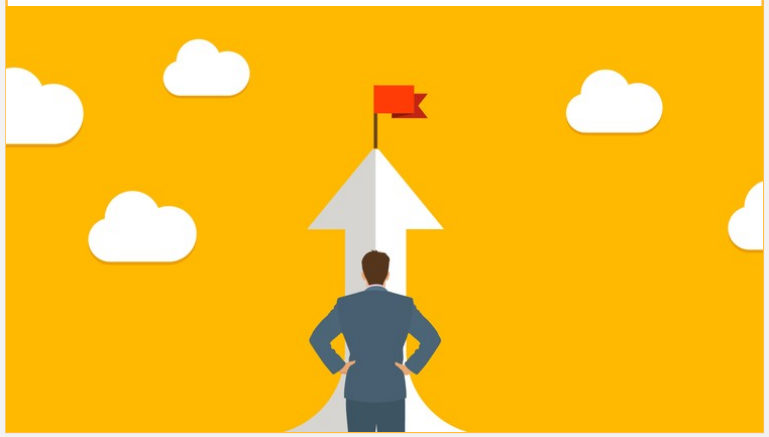
Virtual Sommelier Project

Virtual Sommelier Project

Project Description

Goal

- Use predictive analytics to identify a wine varietal based on its description.

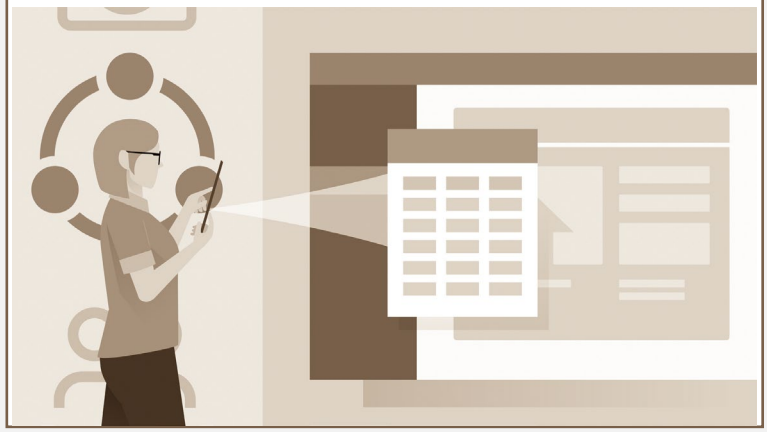


Data

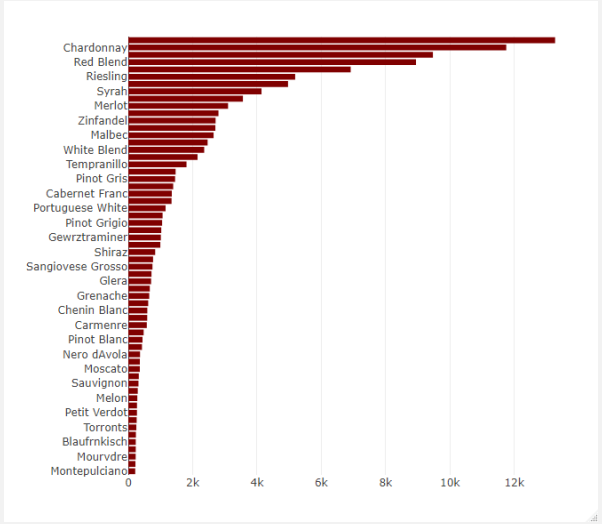
- Obtained through Kaggle, includes approximately 130,000 wine reviews with fourteen variables.

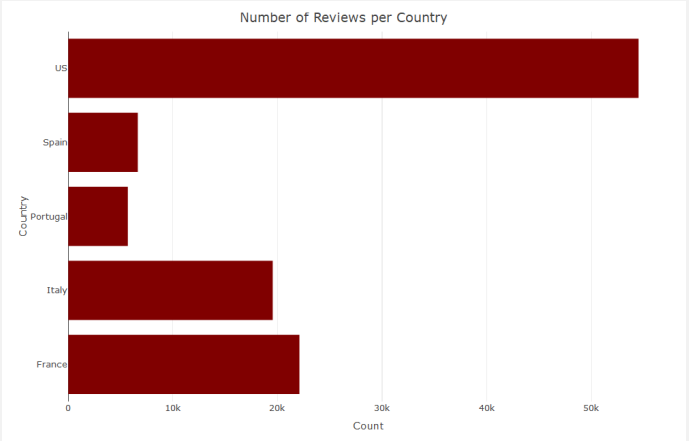
Deliverable

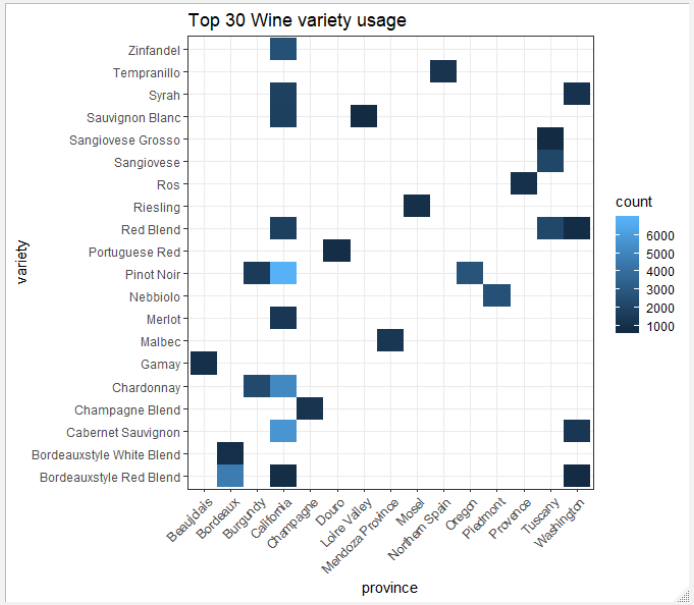
- R-Code, report and presentation explaining the data exploration and modeling to determine if we could predict a wine based on a taster's description.

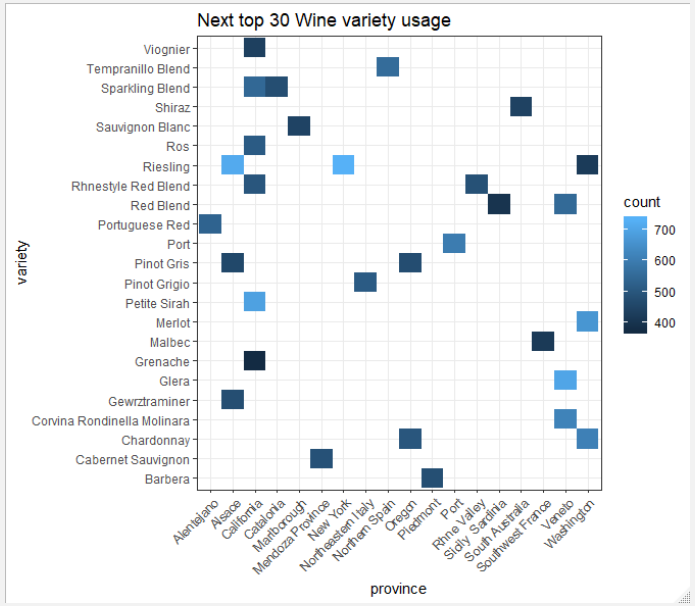


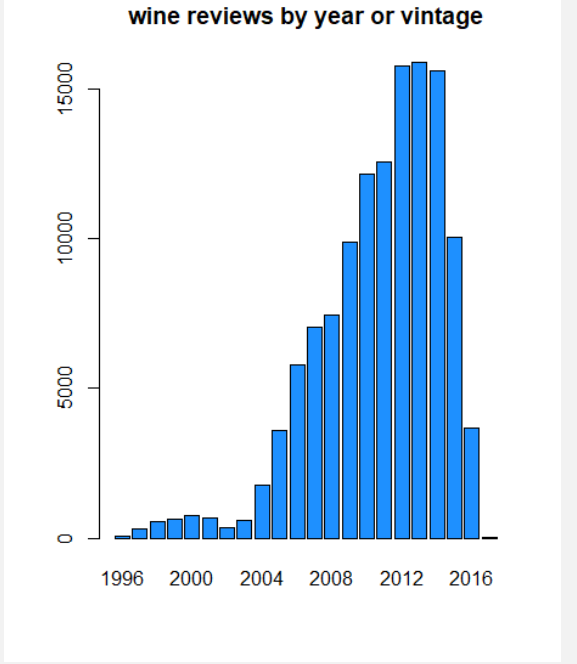
Data Exploration

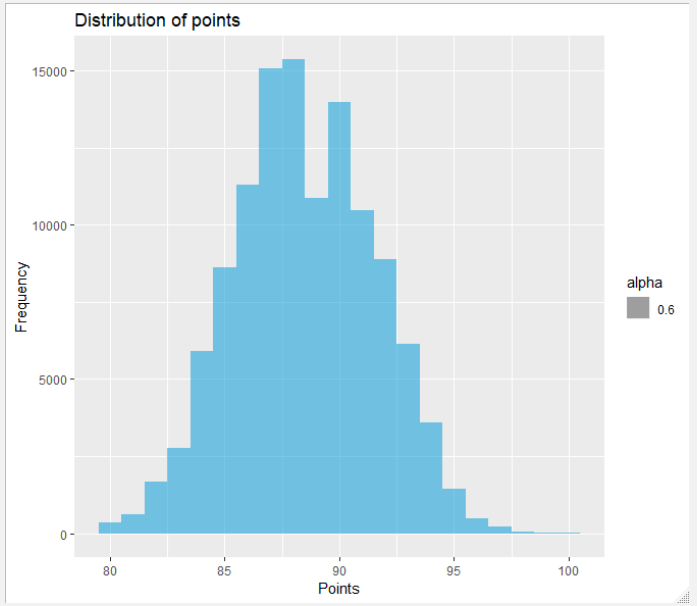




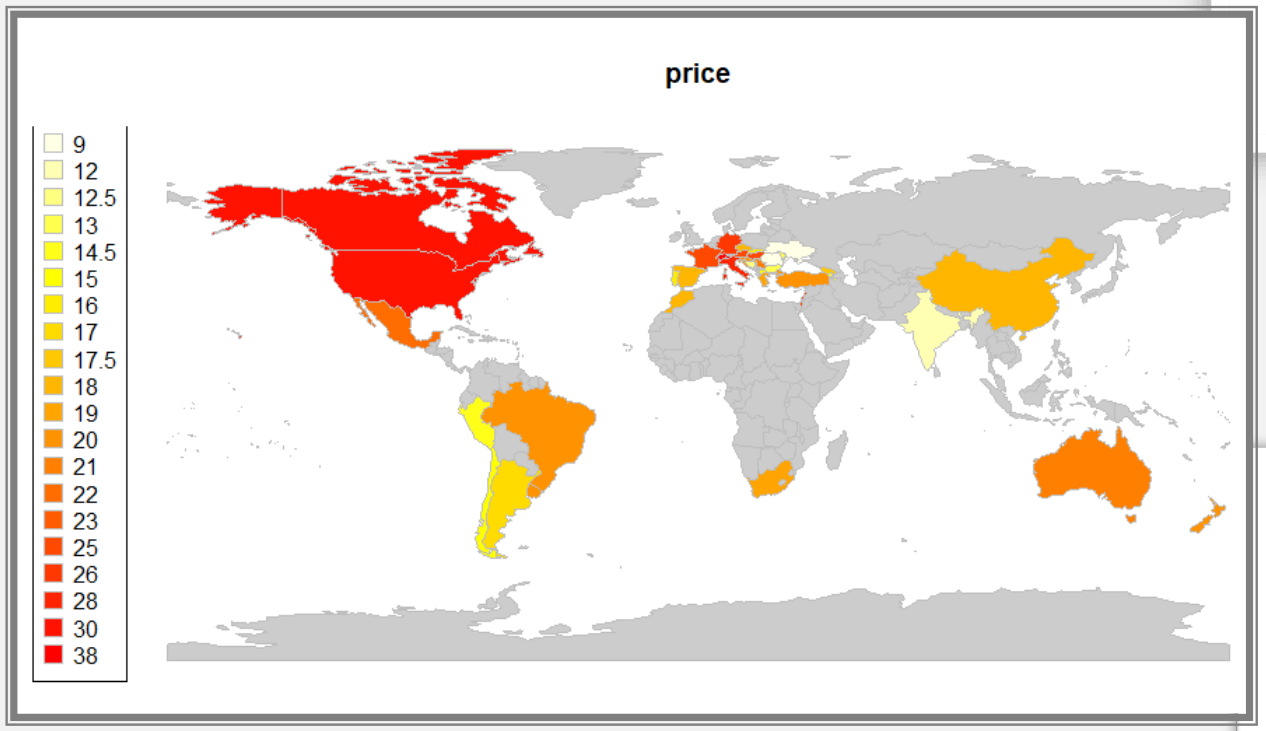




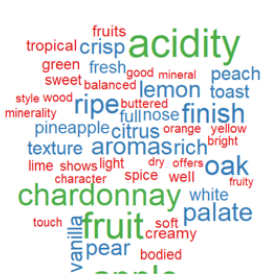




Data Exploration



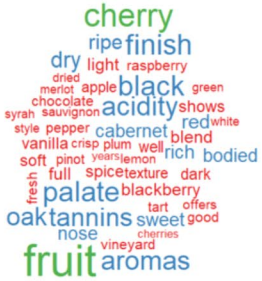
Chardonnay



Red Blend

Riesling

United States



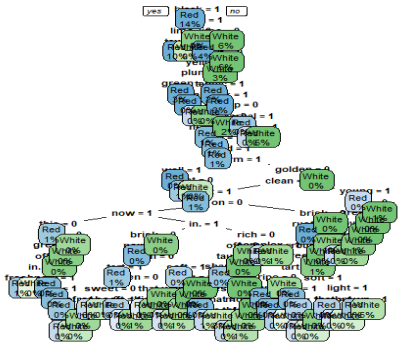
France

Non-US Countries

	LHS	RHS	support	confidence	lift	count
[1]	{barrel,sample,tannins}	{BORDEAUXSTYLEREDBLEND}	0.002	0.987	16.271	236.000
[8]	{buttered,pineapple,toast}	{CHARDONNAY}	0.003	0.969	9.507	282.000
[12]	{acidity,buttered,toast}	{CHARDONNAY}	0.003	0.849	8.328	337.000
[193]	{cherry,dry,silky}	{PINOTNOIR}	0.002	0.834	7.242	216.000
[13]	{buttered,fruit,toast}	{CHARDONNAY}	0.002	0.812	7.967	255.000
[9]	{buttered,toast,vanilla}	{CHARDONNAY}	0.003	0.808	7.926	328.000
[409]	{fruit,oak,tropical}	{CHARDONNAY}	0.003	0.799	7.837	278.000
[1518]	{acidity,apple,oak}	{CHARDONNAY}	0.002	0.779	7.642	229.000
[2368]	{berry,cherry,palate,rose,tannins}	{NEBBIOLO}	0.002	0.778	31.839	253.000
[1519]	{apple,finish,oak}	{CHARDONNAY}	0.002	0.777	7.627	241.000

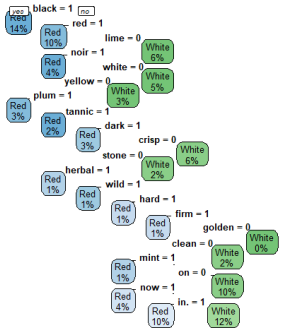
GROWN TREE

Decision tree 3:
predict wine based on description.

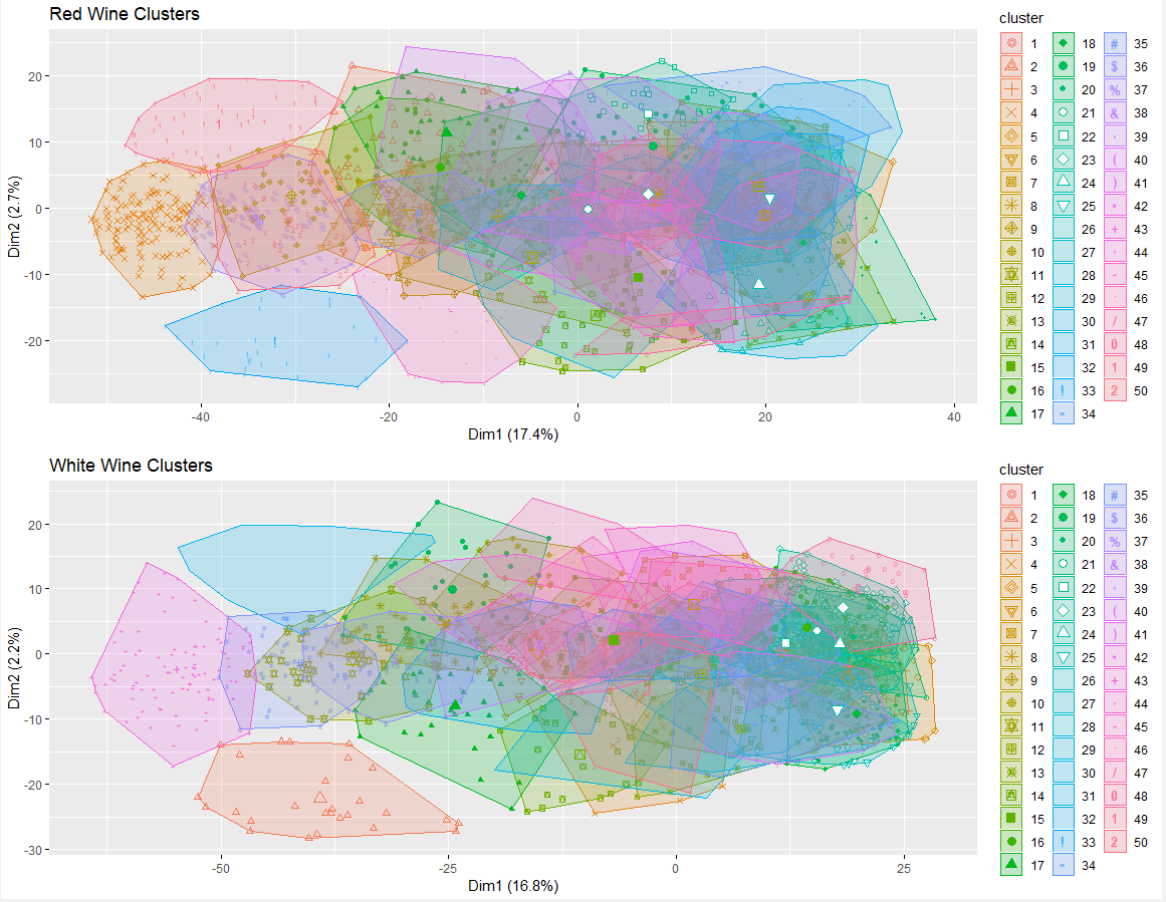


cp=.00002, minbucket = 20,xval=3

Pruned tree 3:
predict wine based on description.



cp=.00596, minbucket = 20,xval=3



Learning Achieved

Increased understanding

- Data collection and analysis
- Data modelling
- Predictive analytics
- Communicating analytical results

Model	Accuracy	
	White	Red
Decision Trees	79%	81%
Naïve Bayes	56%	1.4%
SVM	72%	72%
Random Forest	72%	81%
KNN	64%	76%

CONCLUSION

- IDENTIFYING WINE BY A REVIEW IS FEASIBLE
- WINE DISTRIBUTORS CAN USE FOR MARKETING AND TARGETING SPECIFIC CONSUMER TASTES
- SAVE THE SOMMELIER SALARY $150K/YEAR

IST718: Big Data

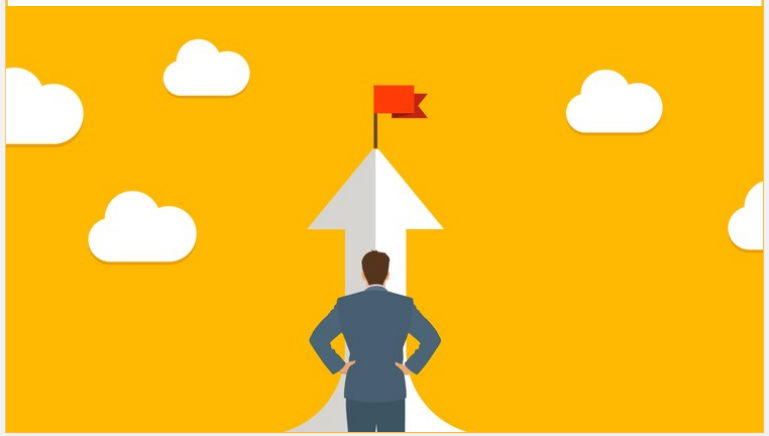
Turn on the Lights Project

Turn on the Lights Project

Project Description

Goal

- Select an applicable analytical methodology for a real-world problem

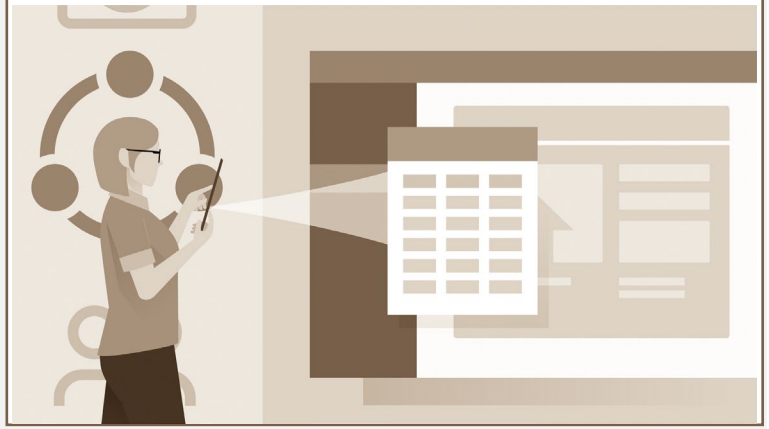


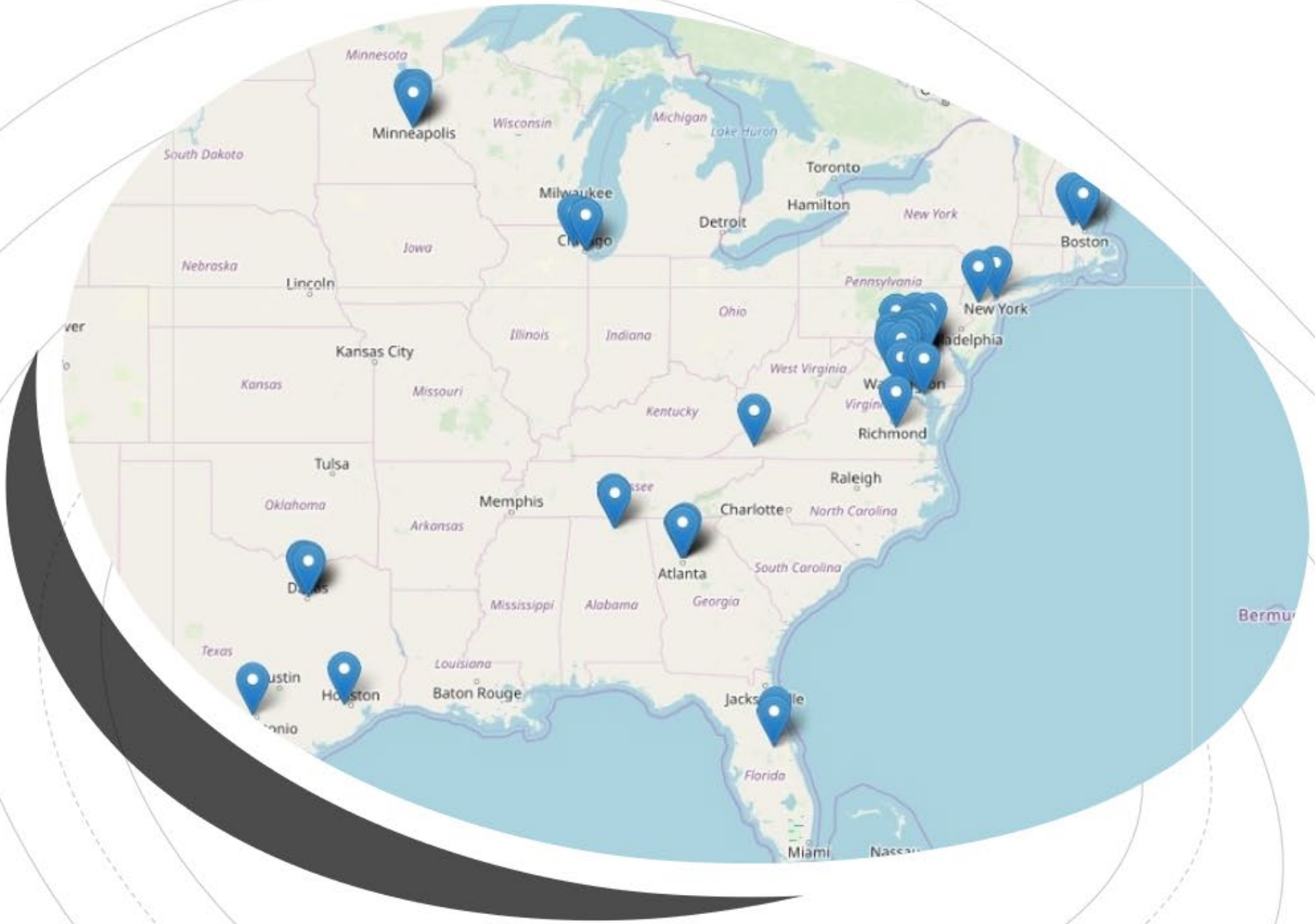
Data

- Electrical demand data for Schneider Electric's Energy and Sustainability Services entailing cost, usage, demand, area, and percent occupied data from 161 commercial properties over 6 years.

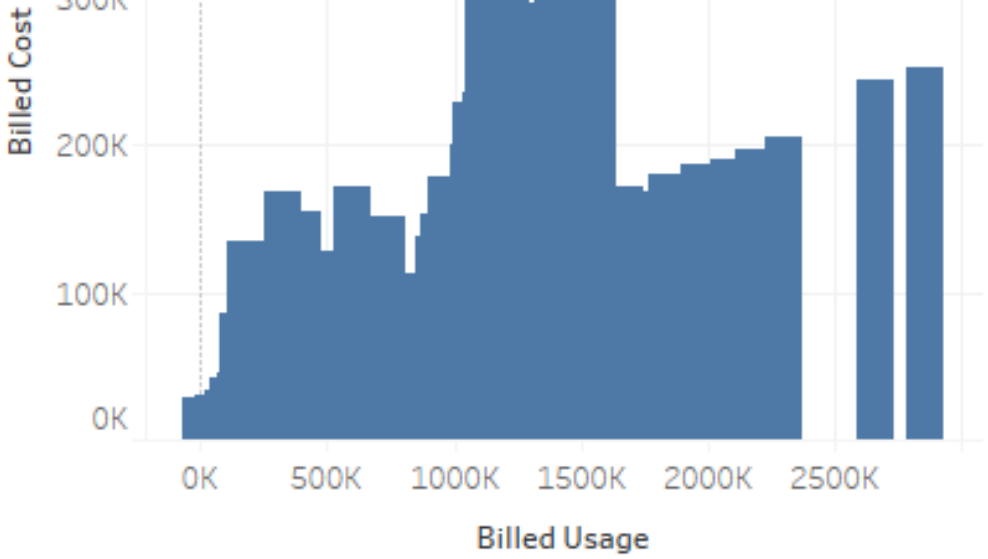
Deliverable

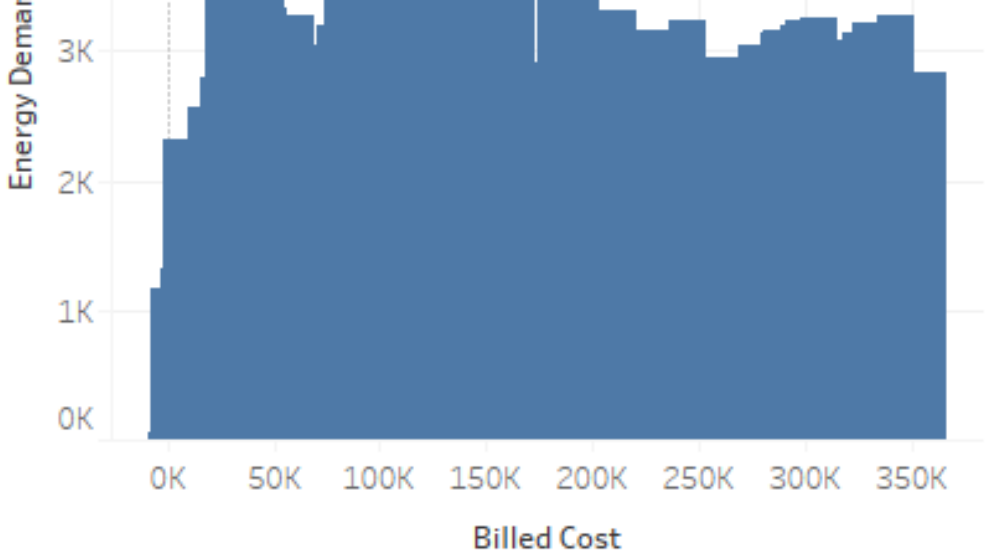
- Python Code, report and presentation and dashboard application predicting electrical demand based on property location.



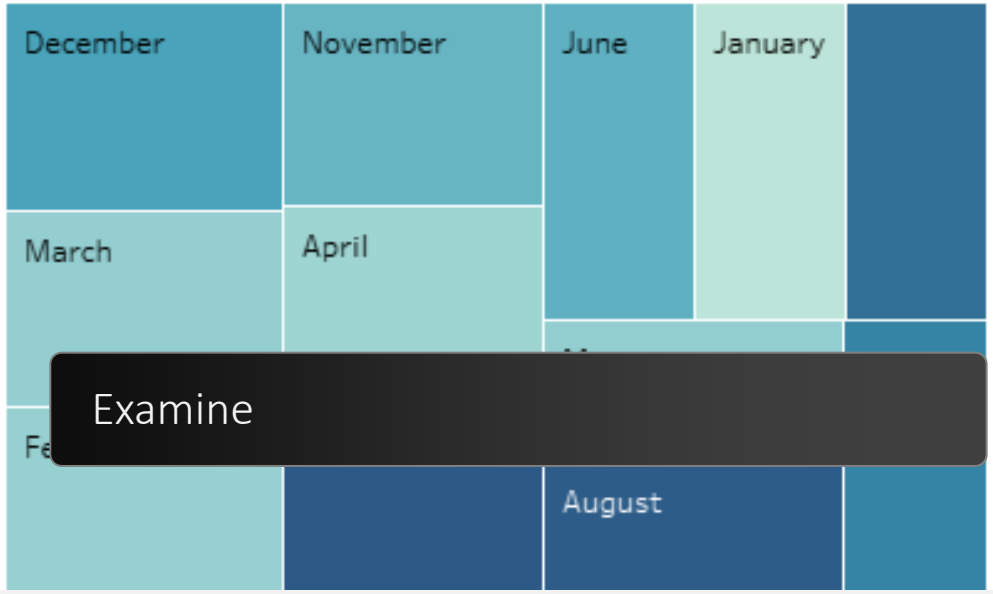

Minnesota
Minneapolis
Wisconsin
Michigan
Lake Huron
South Dakota
Toronto
Hamilton
Milwaukee
Detroit
New York
Boston
Iowa
Nebraska
Lincoln
Pennsylvania
New York
Ohio
Illinois
Indiana
Kansas City
West Virginia
Kansas
Missouri
Kentucky
Richmond
Tulsa
Raleigh
Oklahoma
Memphis
Charlotte
North Carolina
Arkansas
Atlanta
South Carolina
Mississippi
Alabama
Georgia
Texas
Louisiana
Baton Rouge
Florida
Miami

Examine



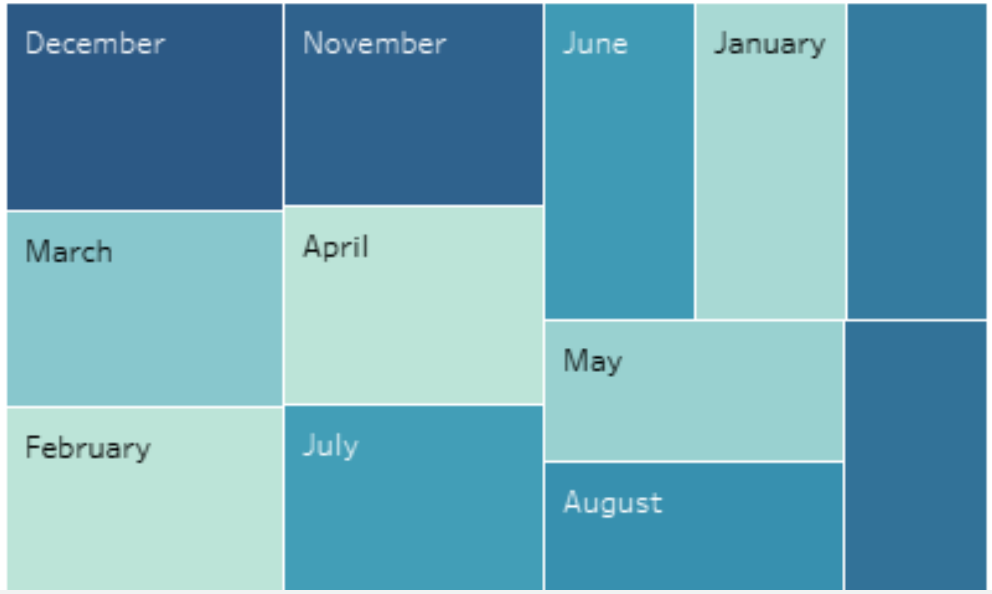


Demand by Month & Area

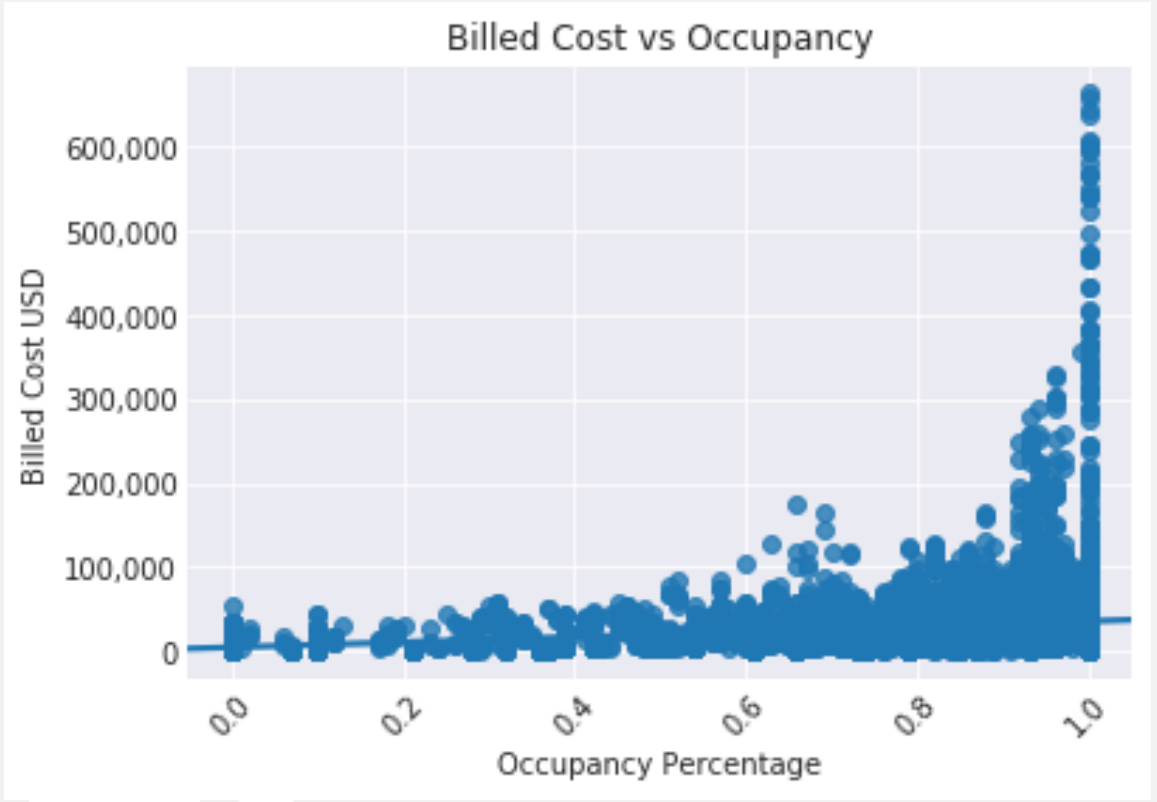


Examine

Demand by Month & Occupancy



Interpret – Linear Regression

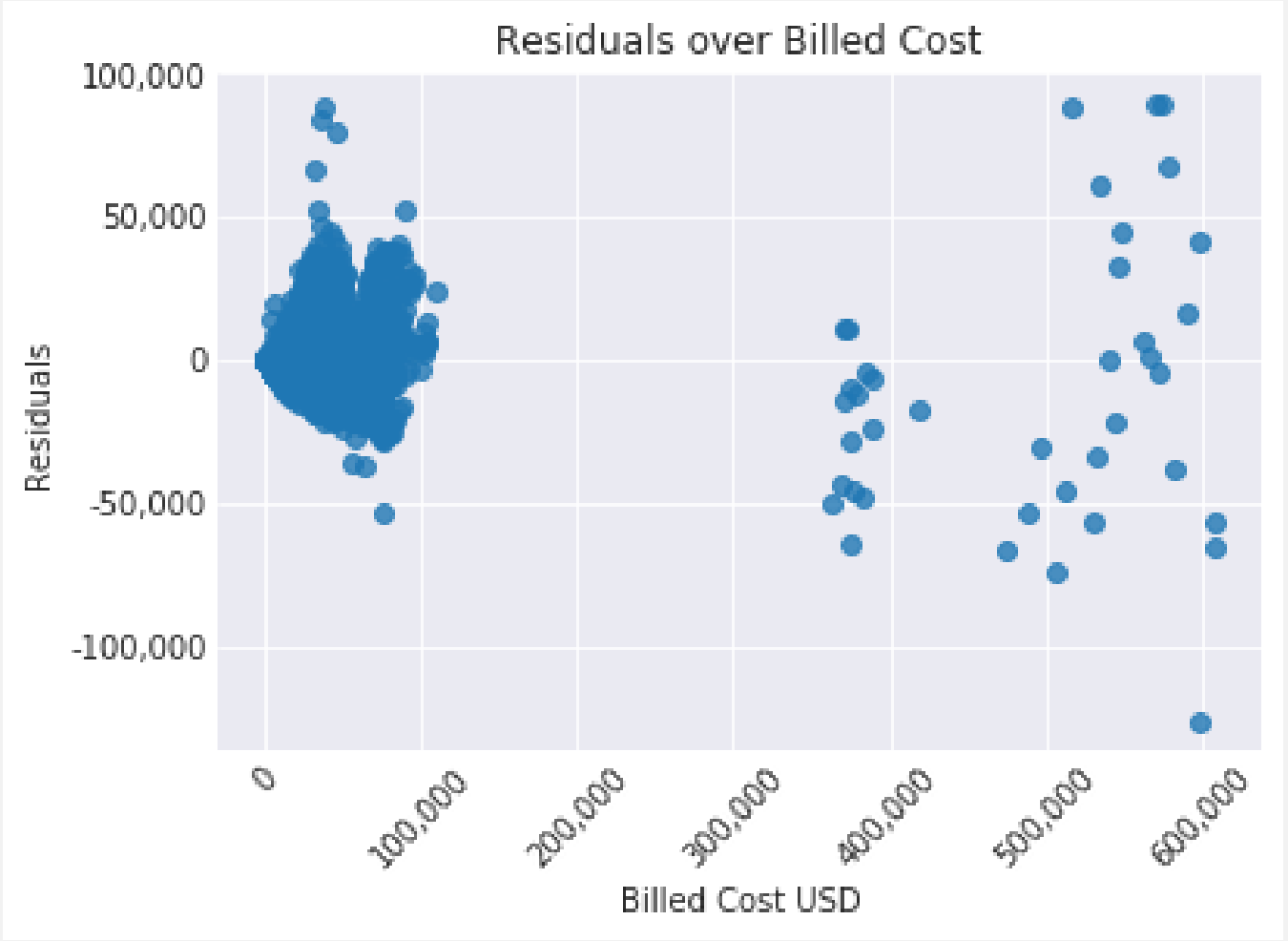


Regression Results

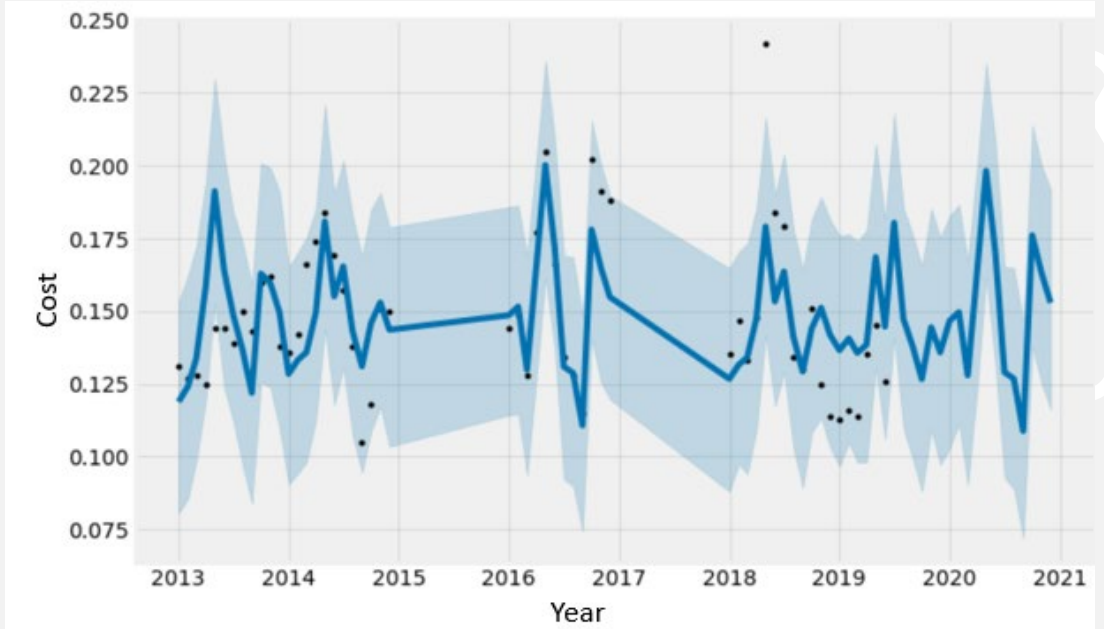
- Positive Results
- 94% R-Squared value

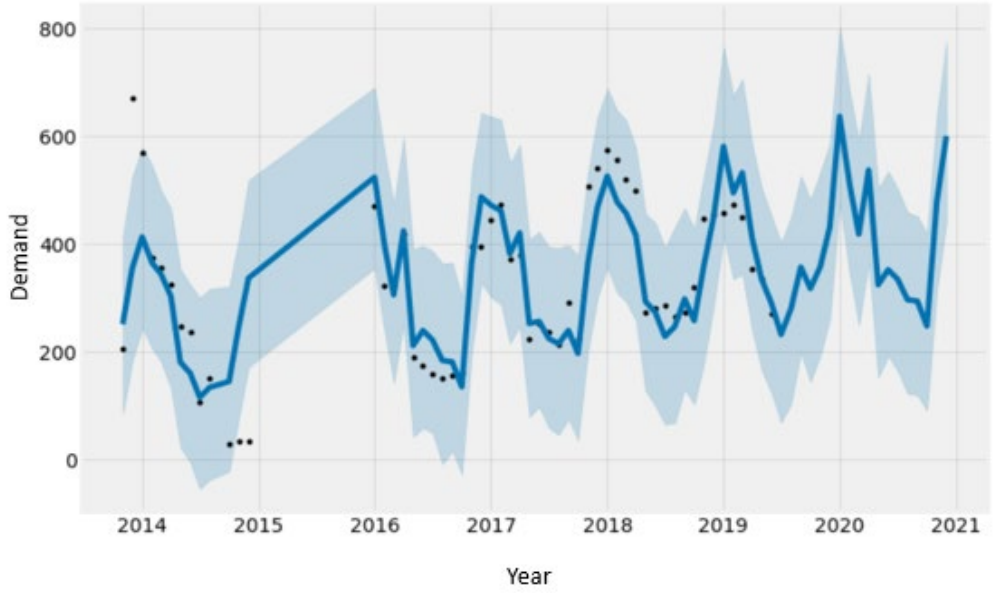
```
                            OLS Regression Results
==============================================================================
Dep. Variable:             BilledCost   R-squared:                       0.949
Model:                            OLS   Adj. R-squared:                  0.949
Method:                 Least Squares   F-statistic:                 2.183e+04
Date:                Sun, 01 Dec 2019   Prob (F-statistic):               0.00
Time:                        15:48:32   Log-Likelihood:                -49970.
No. Observations:                4726   AIC:                          9.995e+04
Df Residuals:                    4721   BIC:                          9.998e+04
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   -747.4694    515.681     -1.449      0.147   -1758.444     263.506
Occ         3392.4267    576.342      5.886      0.000    2262.528    4522.325
Energy        14.0526      0.648     21.692      0.000      12.783      15.323
Area           0.0294      0.002     16.782      0.000       0.026       0.033
BilledUsage    0.0345      0.001     35.612      0.000       0.033       0.036
==============================================================================
Omnibus:                     1678.400   Durbin-Watson:                   0.585
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            67917.230
Skew:                           0.989   Prob(JB):                         0.00
Kurtosis:                      21.466   Cond. No.                     4.13e+06
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.13e+06. This might indicate that there are
strong multicollinearity or other numerical problems.

Proportion of Training Set Variance Accounted for:  0.949

Proportion of Test Set Variance Accounted for:  0.943
```

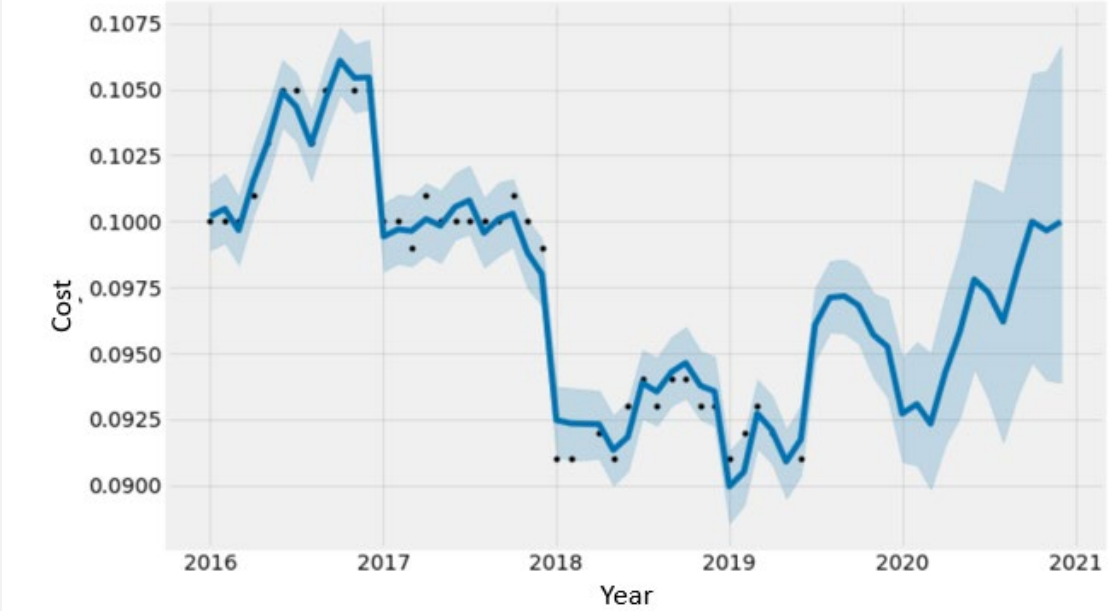
Heteroskedasticity




Demand
Year


Cost
Year


Demand
Year


Cost
Year

Interpret: Prophet Dashboard

Property Browser - Mozilla Firefox

Property Browser

localhost:5006/BokehSelection2

INTERACTIVE PROPERTY TREND EXPLORER

Select the property you would like to see a trend of.

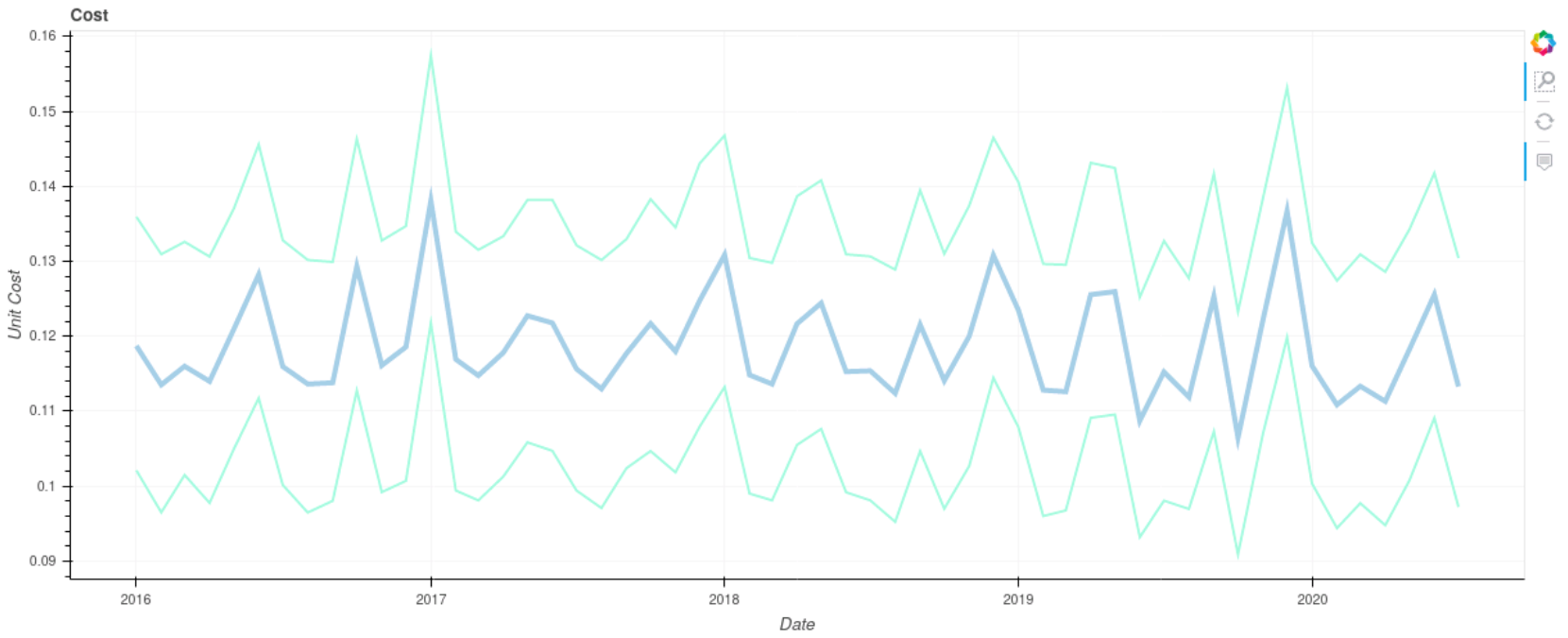
You can zoom in by selecting an area.

There is a reset zoom button just right of the chart

Tooltips for individual values are also available.

Select the Address of interest:

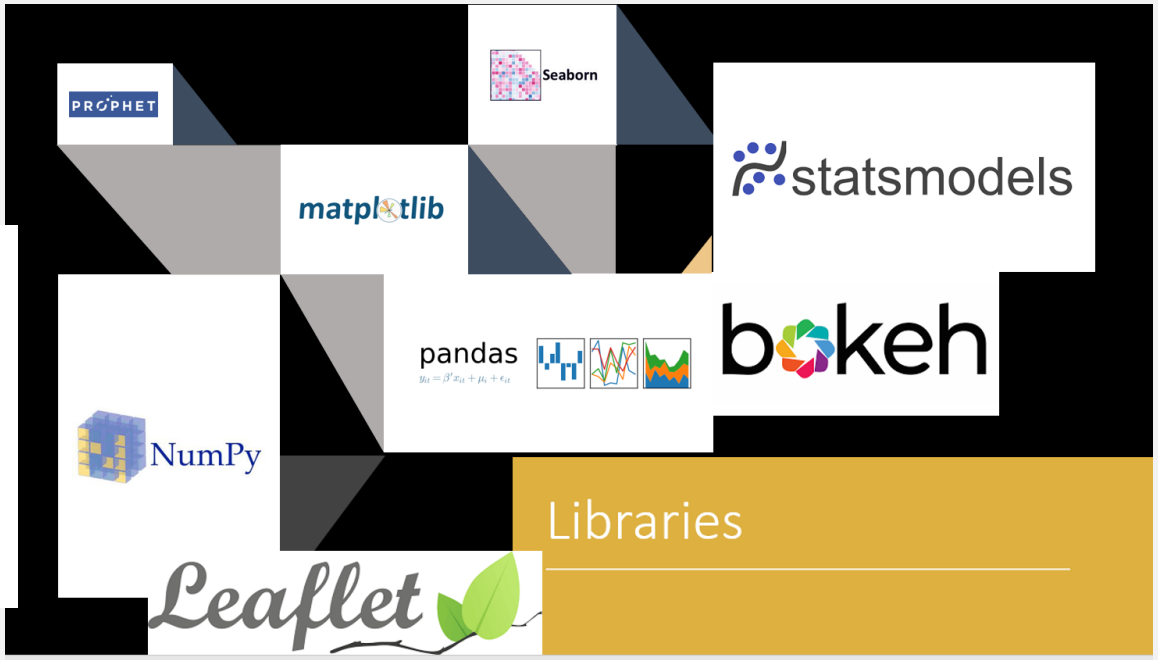
Property: 27050-310 Sentinel Way (NBP 310)



Learning Achieved

Increased understanding

- Big Data Analytics
- OSEMIN Process
- Aggregating Data
- Solving Real-World Problems



ThankYou

Jo Vivian

SUID: 360829521

cvivian@syr.edu

https://github.ibm.com/jo-vivian/2SU_MS-ADS_Portfolio