MS ADS

PORTFOLIO

Jo Vivian | 1 May 2020

Contents

Introduction

Three years ago, I was sitting in a hotel restaurant waiting for breakfast when a man sat at an adjacent table wearing a t-shirt that said, “Data is the new Bacon.” After laughing to myself, I began to seriously think about all the different aspects of my life affected by data. For over 35 years, I have been involved in delivering learning solutions, and I began to wonder, how might applying data science to the area of learning make it more effective? Thus began my foray into the pursuit of a Masters in Data Science.

The Applied Data Science curriculum at Syracuse includes courses on the statistics aspect of data science as well as the programming aspect. The goals of the program don’t just focus on collecting, organizing, and analyzing data, but also includes developing alternative strategies based on the data while not forgetting to incorporate ethical dimensions we should all keep in mind when representing data.

While I found almost all the courses rewarding and I learned from every course, I would like to highlight what I learned from four particular courses:

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

For me, these courses were instrumental in achieving the goals of identifying patterns in data via visualization and demonstrating communication skills regarding data and its analysis. While all my courses contributed to teaching me how to develop a plan

of action to implement the business decisions derived from the analyses, I would like to focus particularly on what I learned from these classes.

MBC638: Data Analysis and Decision Making

PROJECT DESCRIPTION

The primary goal of this project was to follow DMAIC (Define, Measure, Analyze, Improve, Control) steps to improve upon an established issue, defined as a problem statement. It used accessible data that could be collected over the course of ten weeks, including baseline data for comparison and extrapolated future data to measure control.

As a full-time employee, adding the workload of a full-time student would be a challenge in the years ahead. What would slip? Health and fitness? Personal activities? Time management is a process that typically has room for improvement, but under these circumstances, it could become a real problem. With a month of historical work and fitness data for a baseline, I used the next 8 weeks to measure time spent on work, school, fitness, chores, sleep, and personal "fun" activities to determine areas of improvement and extrapolate future data for controls.

A storyboard, seen in Figure 1 captured the entire DMAIC process, providing a high-level overview of the goal and measures of success, and the tools used across the process.

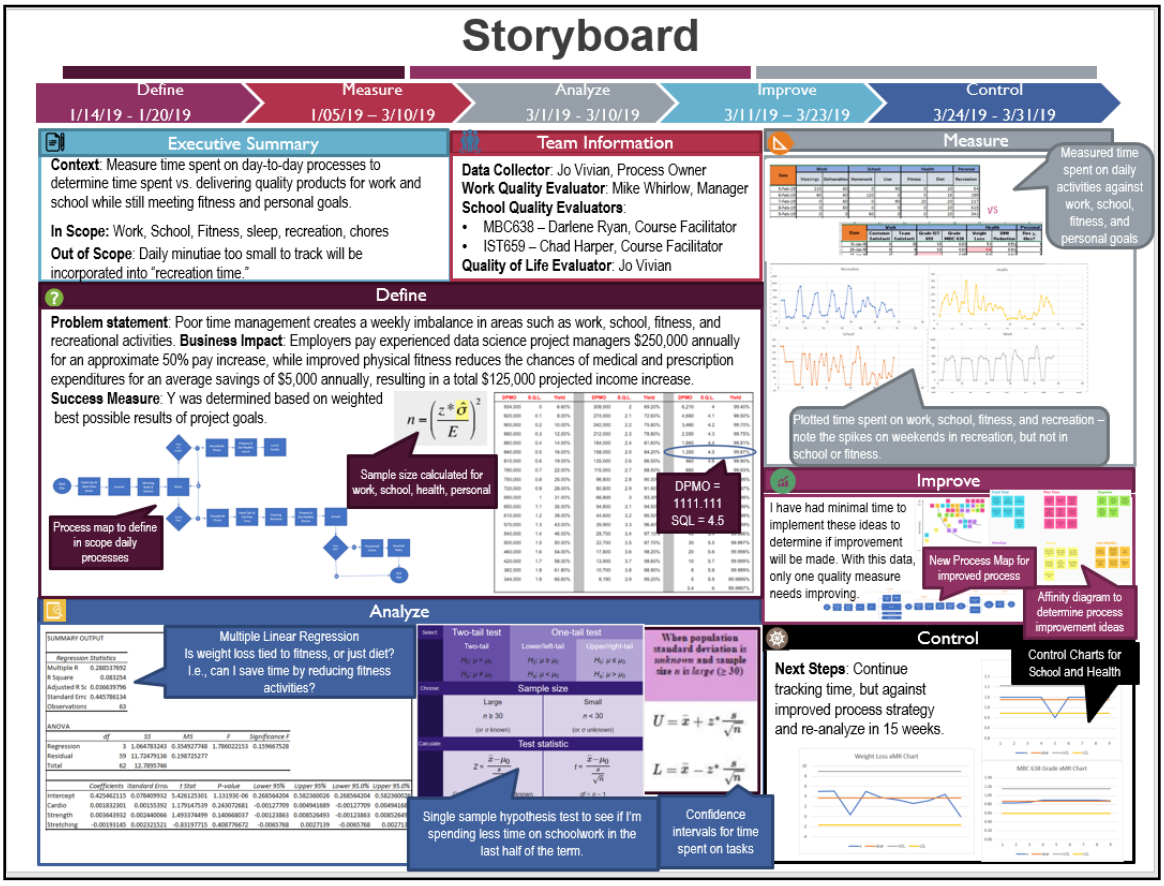

Figure 1 - DMAIC Storyboard

LEARNING GOALS ACHIEVED

DMAIC is a data-driven improvement cycle used for improving, optimizing and stabilizing business processes and designs. This assignment increased my understanding of not only collecting data, but collecting data towards a specific end-goal, and then analyzing the data to detect patterns that would identify processes to be improved. Taken together, this then taught me key skills for developing action plans and strategies based on my data.

MAR653: Marketing Analytics

PROJECT DESCRIPTION

The goal of this project was to address a legitimate business problem for an existing brand or product. To meet this goal, we chose to explore a marketing strategy to improve the number of visits to public libraries. A common misperception today is that public libraries are becoming obsolete due to electronic readers and the availability of information on the internet. Because of this misperception, government funding is decreasing for this invaluable asset. We analyzed data from the 2016 Public Library survey conducted by the Institute of Museum and Library Services. This survey provided information for all public libraries identified by state library administrative agencies in the 50 states and the District of Columbia. It contained 75 columns x 8,813 rows. Figure 2 provides some idea of the data exploration conducted.

Figure 2 - Library Data Exploration

We used K-Means clustering to identify the best grouping of libraries and then conducted extensive data exploration to identify our marketing recommendations. Next, we Used linear regression to find relationships between the number of visits per library and other variables. Figure 3 illustrates the tools and their results.

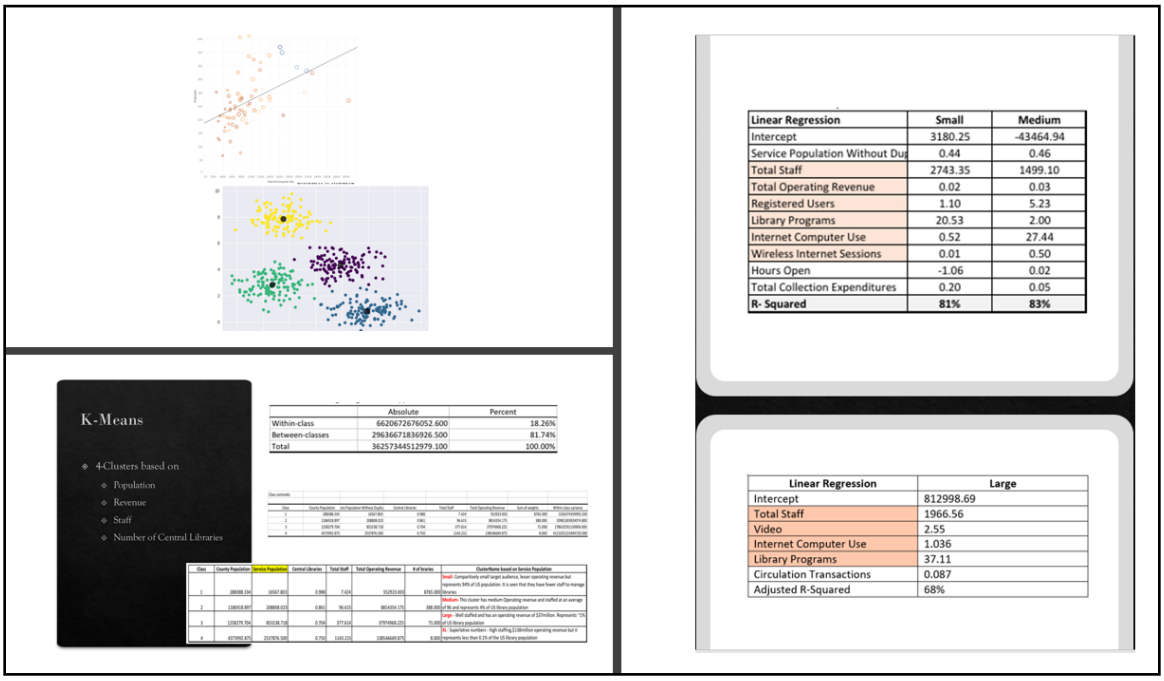

Figure 3 - Tools Used to Analyze Library Visits

Our goal was to help libraries increase visits and we made the following recommendations to this end:

- Increase staff – Adding staff helps in maintenance of library books, visitors get more help during the visits they are more likely to revisit the library again. Every addition of a staff increases ~2000 visits. Understanding that this can't be done without increased funding, we recommend libraries reach out to volunteer communities to increase staff.

- Increase registered users – This can be done by advertising, improving the digital, print collections and increase mobile libraries
- Add Library Programs –For every Library program introduced we see there will be 20 visits increased. Use volunteer data scientists to conduct more studies to identify the ideal target audience for library programs
- Increase computer usage through improved Wi-Fi – Libraries provide a workspace for telecommuters, supply free internet access for people looking for employment opportunities, and offer job and interview training for those in need. According to the ALA, 73% of public libraries assist their patrons with job applications and interviewing skills, and 48% provide access and assistance to entrepreneurs looking to start a business of their own. Improved Wi-Fi increases this type of customer, thus increasing library visits.

LEARNING GOALS ACHIEVED

This project improved my skills in the same areas as the MBC638 project but introduced me to improving a process using marketing analytics as compared to the DMAIC process. Using market response models, product recommendation systems, and resource allocation to improve a business provided a different approach to learning about data collection and analysis and developing alternative strategies. Additionally, this class introduced me to Tableau and the importance of data visualization in communicating the message to the stakeholders.

IST707: Data Analytics

PROJECT DESCRIPTION

The objective of this project was to use predictive analytics to identify a wine varietal based on its description. Using the R programming language, we extracted, cleaned, and analyzed the data and then applied the following 7 models to make predictions:

1. Association Rule Mining
2. K-Means Clustering
3. Decision Trees
4. Naïve Bayes
5. Random Forest
6. Support Vector Machines
7. K-Nearest Neighbor

Association rule mining provided a fundamental insight into common words that were used for each of the wine varietals. The two outputs below provide examples as they were the two most prolific wines. Using K-Means clustering provided 10 lists of words to describe the varieties of wine for both the red and white types of wines. These lists revealed distinct types of wine, because the wines are described in specific ways. Understanding that Decision Tree and Naïve Bayes classifiers were achieving better results with smaller word vector-spaces, revised Decision Tree models were trained using the data from a set of adverbs and adjectives extracted from the description column. As seen in Figure 5, the decision trees and random forests were the most successful models in predicting varietals based on the description of the wine.

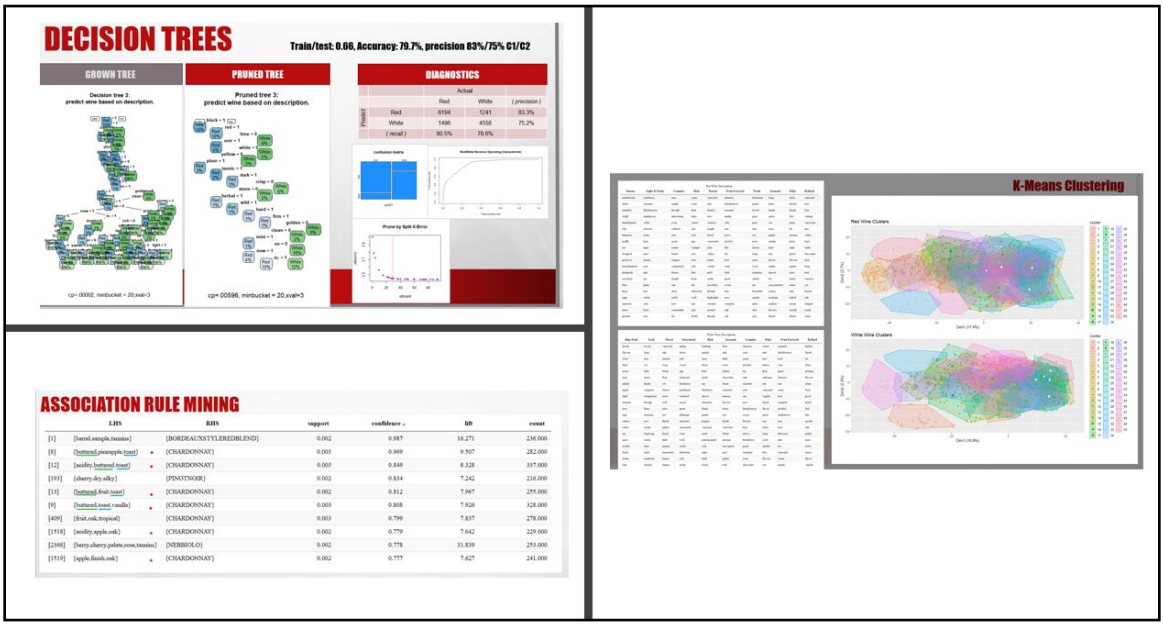

Figure 4 - Models Used to Predict Varietals

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

Figure 5 - Model Accuracy

LEARNING GOALS ACHIEVED

While the project clearly demonstrates learning data collection and analysis, one of the primary focal points of IST 707 is the ability to clearly communicate statistical analysis to stakeholders. While expanding my skills in programming in R, and refining my abilities to make statistical predictions based on data analysis were key components

for this course, a greater emphasis was placed on communicating those results. This was reinforced each week by means of well-documented reports, and then again at term's end when verbally communicating the project results.

IST718: Big Data

PROJECT DESCRIPTION

The purpose of this project was to select an applicable analytical methodology for a real-world problem. Specifically, we analyzed electrical demand data for Schneider Electric's Energy and Sustainability Services (ESS). ESS is an 1,800-employee global team of energy management experts. The team manages an energy spend of more than $30 billion across 500,000 monthly energy invoices at over 430 client sites. They had recently acquired 161 commercial sites and needed predictive analytics to help them manage the electricity for these sites based on data collected from previous energy providers over the past six years. The data used in this analysis is cost (US dollars), usage (kWh), demand (kw), area (square feet), and percent occupied data from 161 commercial real estate buildings across the United States collected over more than 6 years. Data consisted 104 separate files of information from two key clients of Schneider Electric with properties in 64 different zip codes.

A linear regression model was first used to forecast billed cost based on property size (square feet) and occupancy rate (removed outliers prior to running the model) of all the data in aggregate. While some of the early indicators looked promising, it is apparent that there is heteroskedasticity in the data when after exploring the residuals over billed

cost. Linear Regression results and heteroskedasticity evidence can be seen in Figure 6. Because of the Heteroskedasticity, we could not use the regression analysis to predict future costs. Thus we used Facebook Prophet, a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. Figure 7 shows examples of results for a few properties.

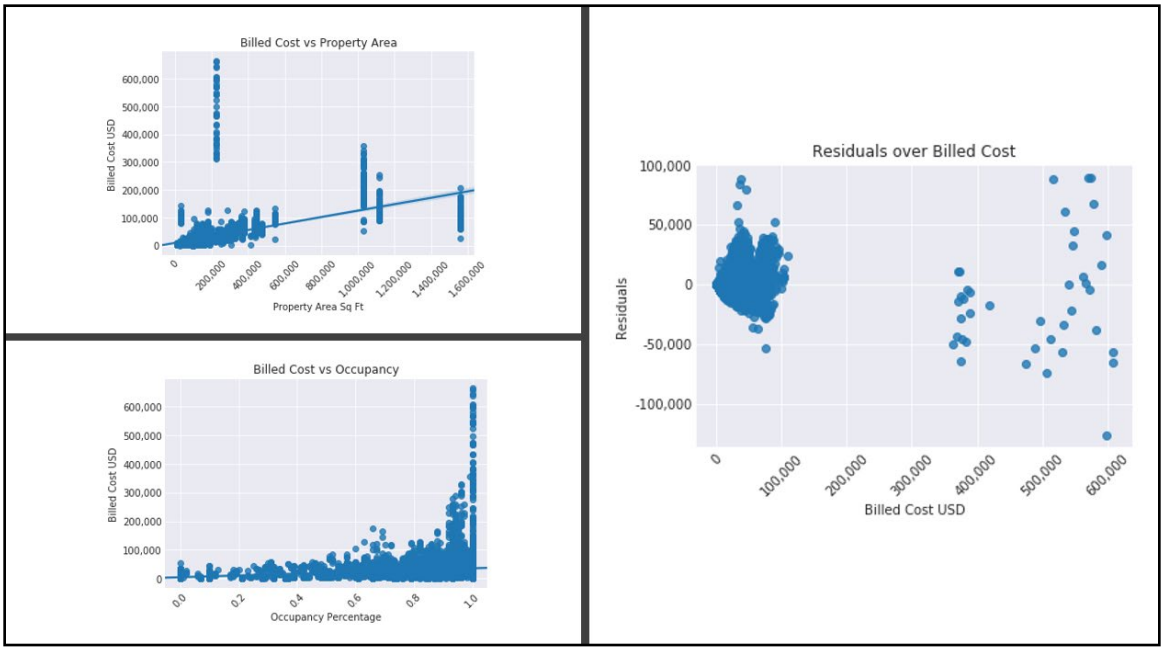

Figure 6 - Linear Regression and Heteroskedasticity

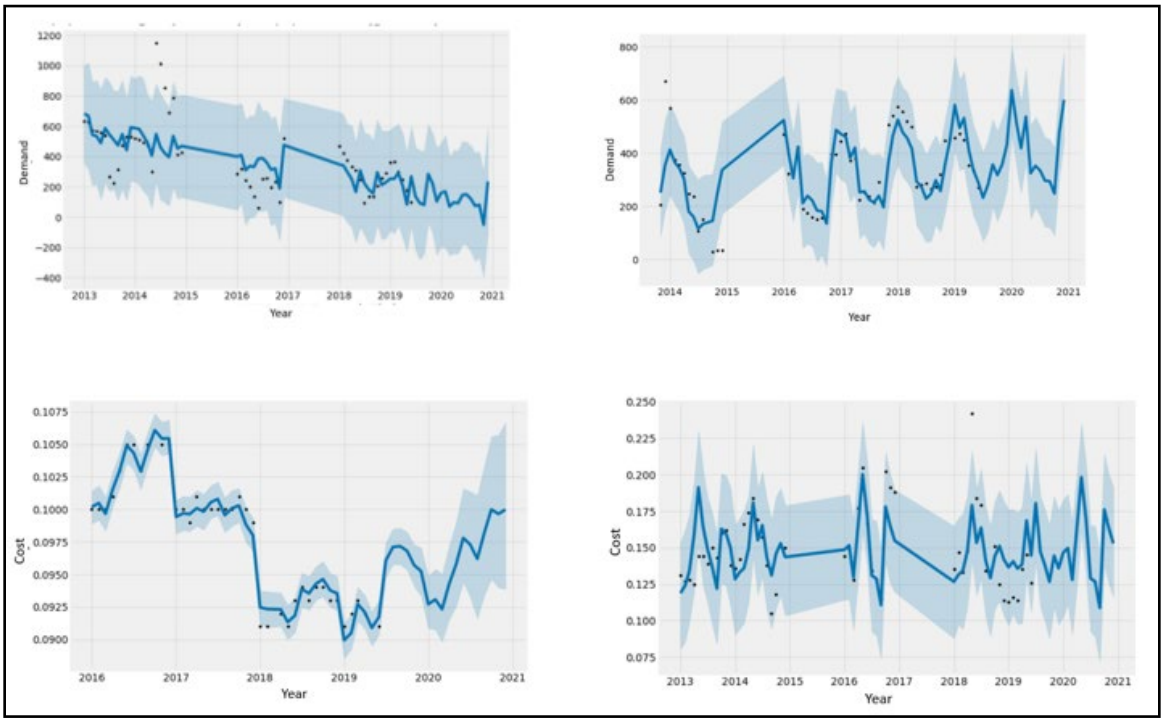

Figure 7 - Facebook Prophet Predictions

The challenge that comes with interpreting the results from the Prophet is that each property is analyzed separately. Thus, a general conclusion is difficult to make in aggregate of all the properties. In order to explore the results for each property, it became necessary to create a dynamic dashboard that allows the end user to filter through each property as needed. Figure 8 is a screenshot of the dashboard showing one random property from the data set. This interactive tool was created using python and the bokeh library to plot the results of the time series estimates by property that were created with Prophet.

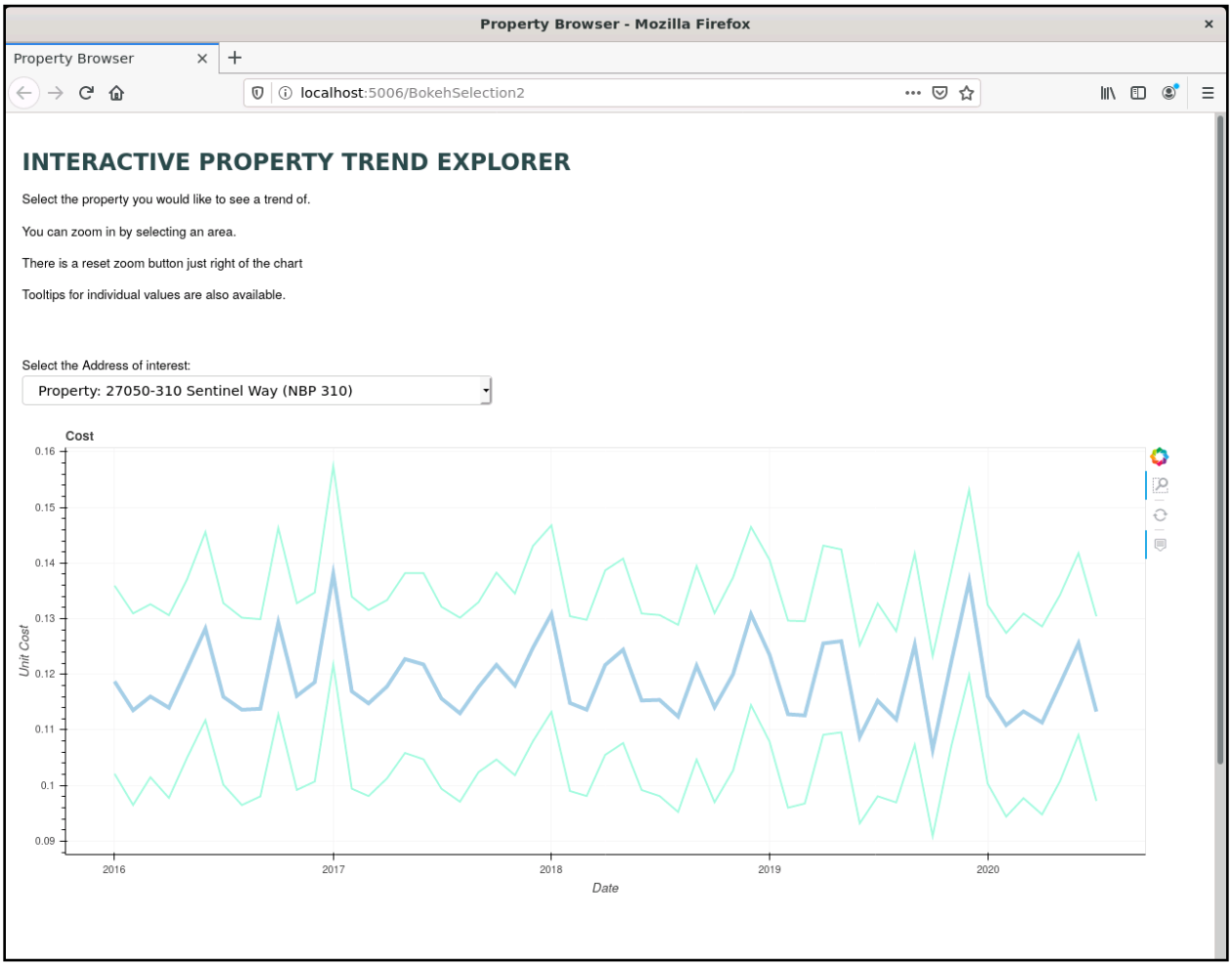

Figure 8 - Time Series Dashboard

LEARNING GOALS ACHIEVED

The primary differentiator between IST718 and IST707, aside for programming in Python instead of R was the size of the data. The analysis of big data requires more time spent both on collecting the data as well as cleaning and exploring it. Big Data is not typically collected from a single source, but from an aggregate of sources, and ensuring the data integrity is a crucial learning objective in this course. This project and course were effective in teaching me the fundamentals of finding the data and then selecting the appropriate analytical methods to the chosen problem; interpreting the data, models,

analysis, and findings and drawing the appropriate conclusions; and then presenting the results in a meaningful way.

Conclusion

It was very difficult to select only four projects to highlight in this portfolio as I genuinely enjoyed almost every project I worked on, and I felt like each of them were instrumental in teaching me the objectives of the major practice areas in data science. However, these four projects demonstrate a thorough understanding of collecting and managing data. The projects in this portfolio also outline my understanding on data analysis using statistical methods and data mining techniques for tasks such as regression, classification, or clustering. Communicating data results can best be done with data visualizations to identify patterns, and from this, actionable recommendations are developed to reflect tangible business decisions. This portfolio demonstrates multiple data visualization techniques used for communication, as does the descriptive text associated with the visualizations. Finally, the projects in this portfolio prove that the ethical dimensions of data science practice was applied, especially the project for Big Data Analytics. Only relevant and timely data was used in that project while considering data privacy when analyzing the company's proprietary information.

Each course in this graduate program has built on previous courses, which taken altogether, contributed to a high-level of understanding and well-developed skills that I plan to apply to in my career. Specifically, I intend to use this knowledge to improve the field of learning by providing analytics to predict which learning objects a student needs to focus on to improve his or her chances of success in a chosen profession.

References

Venkatesan, R., Farris P., Wilcox R. (2015). Cutting Edge Marketing Analytics: Real World Cases and Data Sets for Hands on Learning. NY, NY. Pearson.

Larose, D. (2016). Discovering Statistics. (3rd ed). NY, NY. W.H. Freeman and Company

Forecasting at Scale, Facebook Open Source. Retrieved from: https://facebook.github.io/prophet/

Tan P., Steinbach, M., Karpatne, A., Kumar V. (2019). Introduction to Data Mining. (2nd ed). NY, NY. Pearson

Miller, T. (2015). Modeling Techniques in Predictive Analytics with Python and R. NY, NY. Pearson