



DECEMBER 6, 2020

PORTFOLIO MILESTONE

M.S. APPLIED DATA SCIENCE

COURTNEY ZIMMER-BARTELS

SUID: 617132274

[Github.com/czimmerb/Masters_Portfolio](https://github.com/czimmerb/Masters_Portfolio)



Table of Contents

Introduction	2
IST 687 – Introduction to Data Science	2
a. Project Description	2
b. Insights	6
IST 707 – Data Analytics	6
a. Project Description	6
b. Insights	10
IST 722 – Data Warehouse	11
a. Project Description	11
b. Insights	14
IST 719 – Information Visualization	15
a. Project Description	15
b. Insights	17
Conclusion	18
Bibliography	19

Introduction

The Applied Data Science program at Syracuse University's School of Information Studies is based on the foundation of data science principles; data collection, data analysis, strategy/decisions and implementation. These principles are taught throughout the courses and both practiced and assessed through hands on exercises including homework, labs and projects. To show my mastery of these data science principles and the program learning goals, I will highlight coursework that exemplifies each one of the goals listed below.

- Describe a broad overview of the major practice areas in data science.
- Collect and organize data.
- Identify patterns in data via visualization, statistical analysis, and data mining.
- Develop alternative strategies based on the data.
- Develop a plan of action to implement the business decisions derived from the analyses.
- Demonstrate communication skills regarding data and its analysis for managers, IT professionals, programmers, statisticians, and other relevant professionals in their organization.
- Synthesize the ethical dimensions of data science practice.

IST 687 – Introduction to Data Science

a. Project Description

Introduction to Data Science is an introductory course teaching applied examples of data collection, processing, transformation, management, and analysis. Using the statistical software R, our group project was to select a dataset, transform and provide a written analysis of our data. Working with Jonathon Parry and Kevin Vogel, we analyzed Boston crime data, paired with weather data to answer a number of questions including:

- Has there been an increase in crime?
- Does crime occur more frequently at a certain time, day, week, or month?
- Is temperature a predictor of crime?

The initial dataset chosen was from Kaggle.com¹, it contained Boston crime data for 2015-2018 and was merged with weather data from the NOAA² website. A significant amount of data cleaning was needed and our first hurdle to overcome was syncing data types and merging of the two datasets. The logical field to use to merge was the 'date' column, however one dataset used 'POSIXct' format and the other in a 'Date' format. The following code was executed to sync the two 'Date' data types and merge both datasets into the 'Final' data frame.

¹ <https://www.kaggle.com/ankkur13/boston-crime-data>

² <https://www.ncdc.noaa.gov/cdo-web/>

```

Weather$Date <- as.Date(as.POSIXct(Weather$Date,format='%Y-%m-%d %H:%M:%S'))
Crime$OCCURRED_ON_DATE_NOTIME <-
as.Date(as.POSIXct(Crime$OCCURRED_ON_DATE,format='%m/%d/%Y %H:%M:%S %p'))
Final <- merge.data.frame(Crime, Weather, by.x="OCCURRED_ON_DATE_NOTIME", by.y="Date",
all.x = TRUE)

```

After we had the datasets both cleaned and merged, we began analyzing our data starting with some exploratory analysis. To determine if crime has increased from 2015-2018, the following calculations were performed using the DescStat() function. Since the only complete years in our dataset were 2016 and 2017, we chose to utilize the average occurrences per day versus the frequency of crime per year.

Table 1. Average Occurrences of Crime Per Day

Year	Frequency of Crime	Number of Days	Average Occurrences Per Day
2015	53,392	199	268.3
2016	99,134	365	271.6
2017	100,938	365	276.5
2018	74,356	275	270.4

Although the number of crime occurrences per day is high, there is no indication of an increase from year to year. Next, we used the following visualizations to see if crime occurs more during certain time of day, week or month.

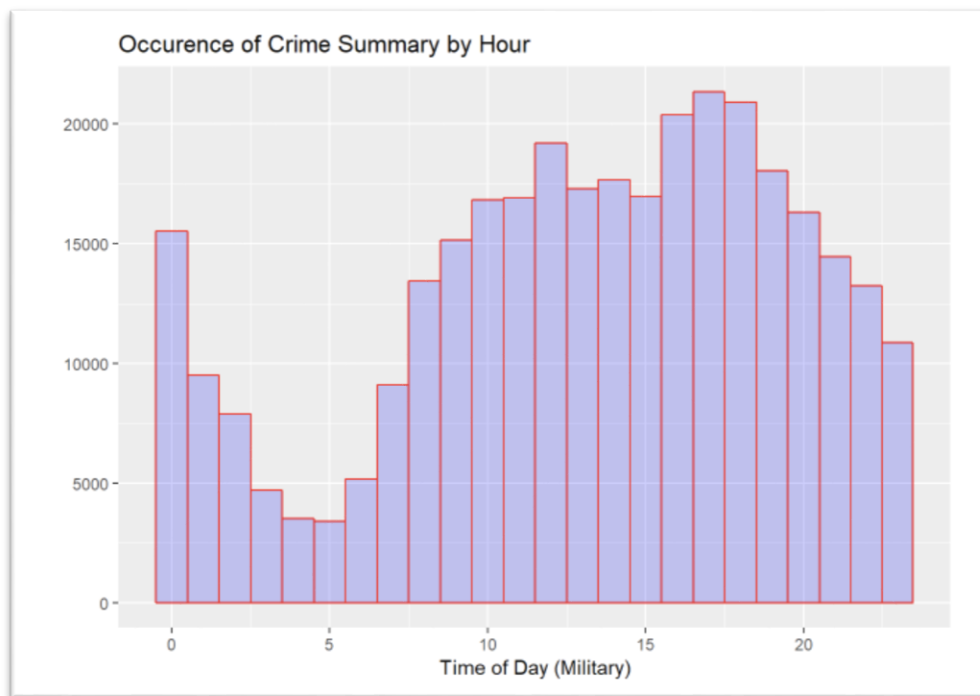


Figure 1. Histogram of Crime Rates by Time of Day

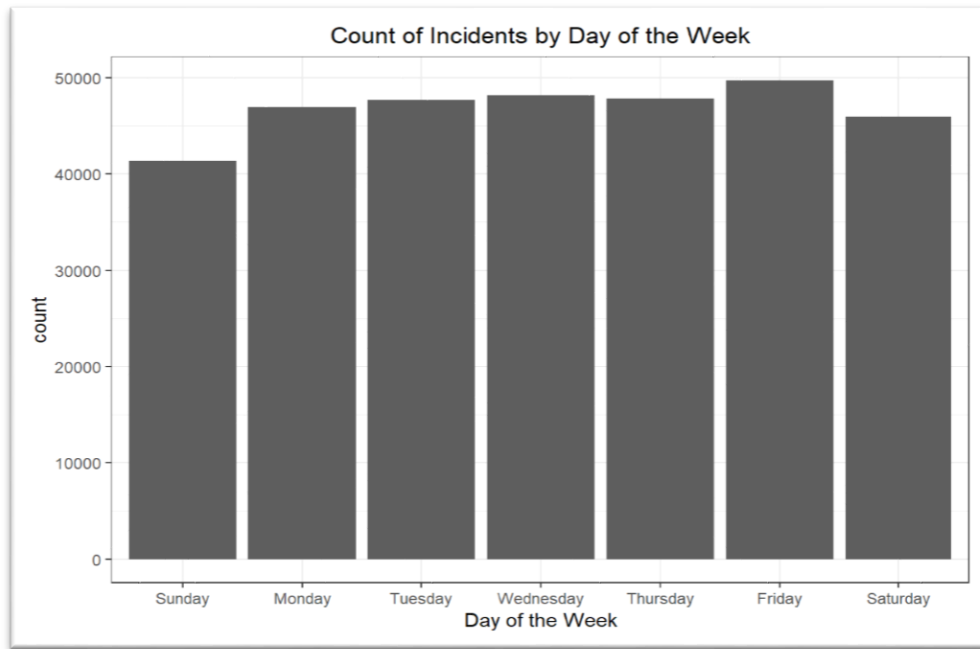


Figure 2. Bar Graph of Crimes by Day of the Week

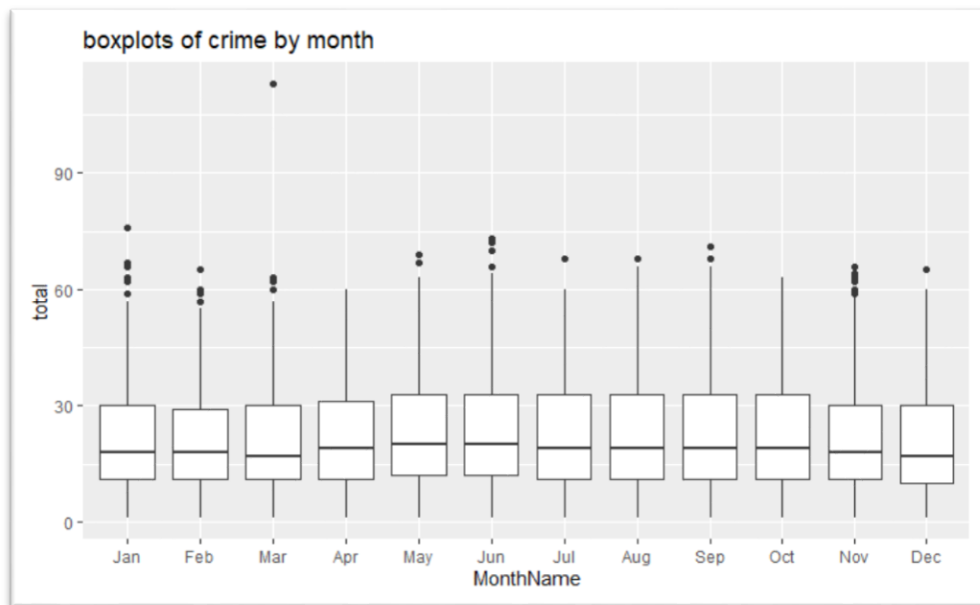


Figure 3. Box Plot of Crimes by Month

From the figures 1-3, we can see that there is very little variation of the number of crimes per month. Whereas, Sunday has a slight dip and Friday has a slight bump in the number of crimes that happen each week. The most variation comes from Figure 1, the number of crimes per hour, where we can clearly see that the number of crimes is very low around 3-6 AM. This makes sense when you factor that there are significantly less people out and about during that time in the morning.

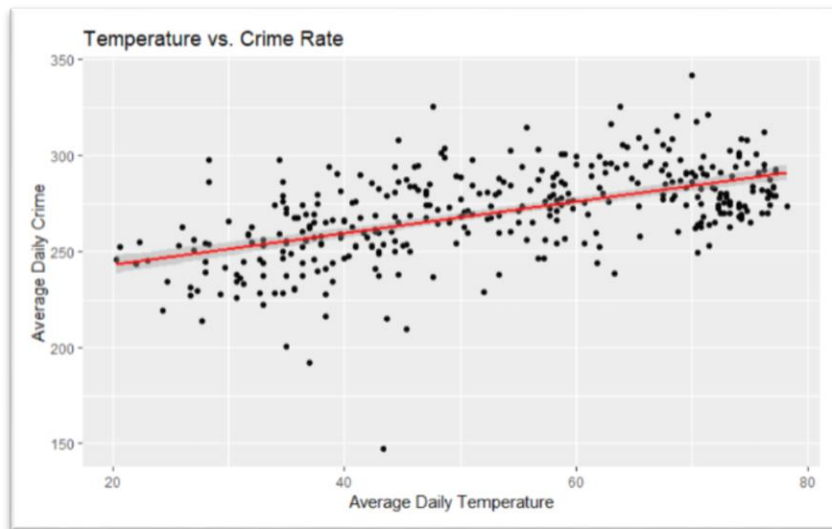


Figure 4. Avg Daily Temperature vs. Avg Daily Crime

When we plot the average daily temperature rate versus the average daily crime, we can see there appears to be a correlation to an increase in temperature and an increase in crime. The last analysis we performed was to see if temperature was a predictor of crime, for this analysis we used linear modeling.

```
Call:
lm(formula = overlapday$Average ~ overlapday$Average_Temp, data = overlapday)

Residuals:
    Min       1Q   Median       3Q      Max
-115.218  -12.046    0.174   11.955   59.212

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    226.85893     3.57241   63.50  <2e-16 ***
overlapday$Average_Temp  0.82368     0.06459   12.75  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.37 on 364 degrees of freedom
Multiple R-squared:  0.3088,    Adjusted R-squared:  0.3069
F-statistic: 162.6 on 1 and 364 DF,  p-value: < 2.2e-16
```

Figure 5. Linear Modeling Results

From our output, we interpret that as temperature increases by 1-degree, daily total of crime occurrences increases by 0.82368. Our p-value is low, which indicates that its statistically significant and that our equation accounts for 30.7% of y values given the adjusted R-squared value.

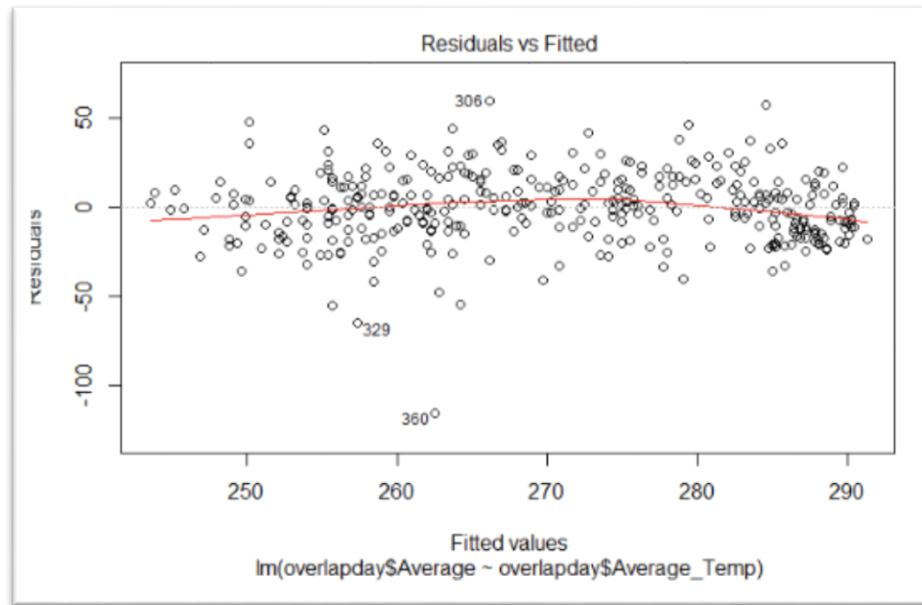


Figure 6. Residuals vs. Fitted for Daily Temperature vs. Daily Occurrences of Crime

Plotting the formula over the data, Figure 6, shows there appears to be a positive linear relationship indicating that as temperature increases, so does the total number of crimes committed daily.

b. Insights

The Applied Data Science course was a fantastic introduction to the world of data science, providing a broad overview and practical examples of data science as well as several of the programs learning goals. From collecting and organizing the data into a usable format. Identifying patterns using visualizations during exploratory analysis as well as statistical analysis in the use of linear modeling. Developing a plan of action to implement business decisions derived from the analysis. Our Analysis showed that temperature does have an impact on the number of crimes committed, this can be translated to summer months having a higher crime activity and that police force employment and scheduling should be adjusted to prevent summer crime spikes. Finally, practiced demonstrating communication skills through both our written paper based on our analysis and findings as well as our in-class presentation of our results.

IST 707 – Data Analytics

a. Project Description

The Data Analytics course was an introduction into data mining methods, allowing for hands on experience to develop solutions to scientific and business problems. The end of course project was a group project to solve a real data mining problem using a dataset of our choosing. My group consisted of David Primrose, Connor Gendron, Rohit Mareddy, Tyler McAfee and myself. Our project was to delve into several NFL statics dataset with the goal of providing insight that could aid in a fantasy football game. Fantasy football is where fans can create their

own roster by competing in a draft among their group of players, you win by amassing the most fantasy points. Our dataset came from Kaggle.com³ and consisted of Basic Stats, Career Stats and Game Logs. Overall, there were 19 different files which we paired down to 9 csv files, these represented football positions that would reward our fantasy football team the most fantasy points.

- Fumbles - Career_Stats_Fumbles.csv
- Passing - Career_Stats_Passing.csv
- Receiving - Career_Stats_Receiving.csv
- Rushing - Career_Stats_Rushing.csv
- Quarterback – Game_Logs_Quarterback.csv
- Running back – Game_Logs_Runningback.csv
- Wide Receiver and Tight End – Game_Logs_Wide_Receiver_and_Tight_End.csv
- Kickers – Game_Logs_Kickers.csv
- Basic Stats – BasicStats.csv

After a significant amount of data cleaning, we began performing some exploratory analysis on our datasets.

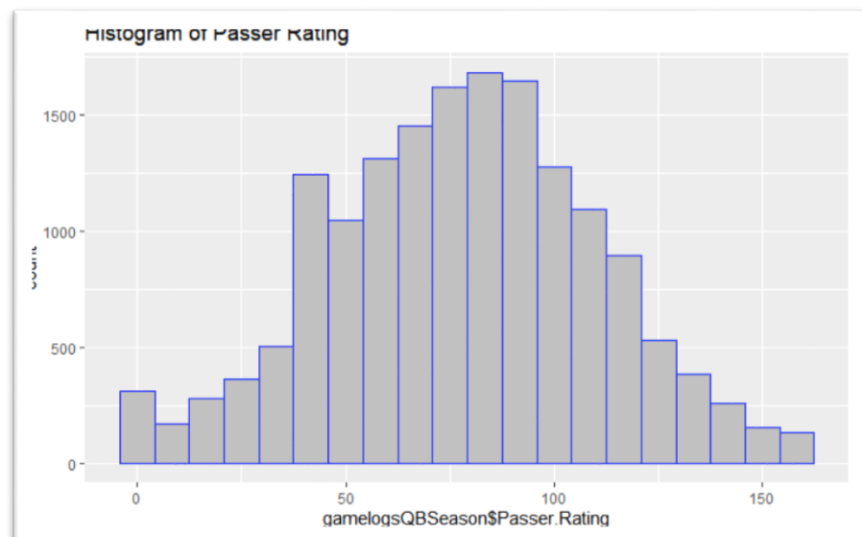


Figure 7. Histogram of Passer Rating from Quarterback Game Logs

Figure 7 is a histogram showing the Passer Rating of quarterbacks (QB), this rating is calculated using a player's passing attempts, completions, yards, touchdowns, and interceptions. It looks to follow a normal distribution with one spike around the 40 mark. Whereas, if we look at Figure 8, histogram showing QB passing yards, you can see that our histogram is skewed towards the left. Passing yards between the 300-425 mark seem to be rarer and a noticeable jump at the 450-500 mark.

³ <https://www.kaggle.com/kendallgillies/nflstatistics>

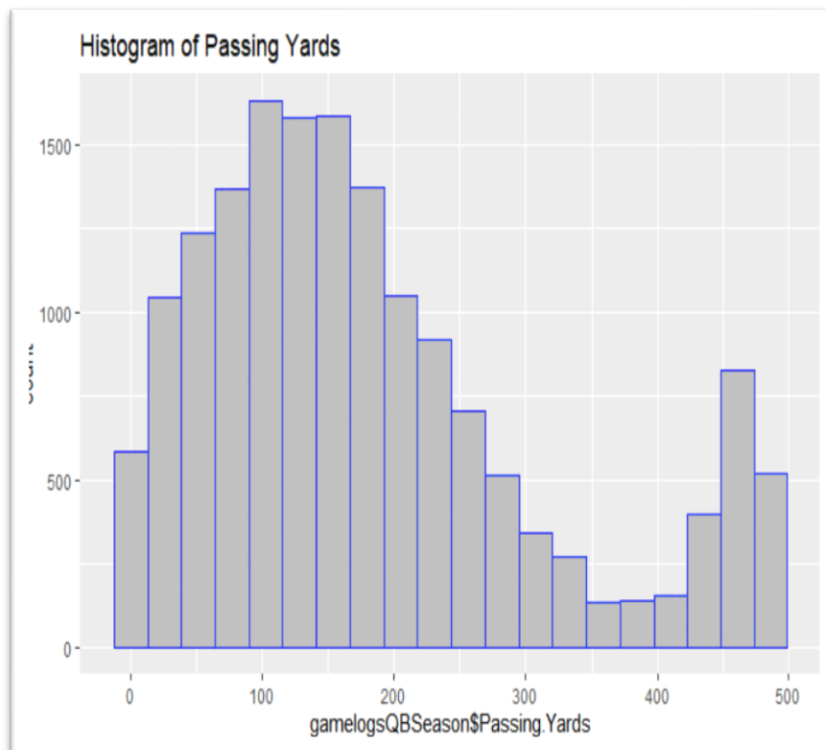


Figure 8. Histogram of Passing Yards from Quarterback Game Logs

After we begin to understand our data better using exploratory analysis, we began utilizing some tools that we learned throughout the course, starting with Association Rule Mining (ARM). ARM was applied to two of our game log datasets, QB and running back/wide receivers (RB/WR). Within our game log datasets, there is a field that records whether the game played was won or lost. We utilized this field by setting the right-hand side of our rule to be Outcome = W, as in outcome of the game is equal to won. We paired down our game log data fields, binned specific fields such as passes complete, attempted and total yards. Sorting our output by lift, Figure 9 shows the top 5 rules in our QB game logs.

##	lhs	rhs	support	confidence	lift	count
## [1]	{Rushing.Attempts=1-10_RushingAttempts,	=> {Outcome=W}	0.09744638	0.8022923	1.606611	2240
##	Rushing.Yards=Negative_RushingYards}					
## [2]	{Rushing.Yards=Negative_RushingYards,	=> {Outcome=W}	0.09579327	0.8007273	1.603477	2202
##	Rushing.TDs=0}					
## [3]	{Rushing.Attempts=1-10_RushingAttempts,	=> {Outcome=W}	0.09579327	0.8007273	1.603477	2202
##	Rushing.Yards=Negative_RushingYards,					
##	Rushing.TDs=0}					
## [4]	{Season=Regular Season,					
##	Home.or.Away=Home,					
##	Ints=0,					
##	Rushing.Attempts=1-10_RushingAttempts}	=> {Outcome=W}	0.09844695	0.7671186	1.536175	2263
## [5]	{Home.or.Away=Home,					
##	Ints=0,					
##	Rushing.Attempts=1-10_RushingAttempts}	=> {Outcome=W}	0.12093792	0.7525717	1.507045	2780

Figure 9. Quarterback Game Log ARM Output

Our top 5 rules have a lift well above 1 and a strong confidence suggesting that there is a good association between the left and right-hand side of our rules. Some noticeable results include rushing attempts between 1-10 appear in 4/5 of our top rules, as well as “Home or

Away” = Home. This indicates that a Home game is most likely to result in a win, and if there is rushing attempts of the QB between 1-10 times during the game, it will also most likely result in a win. Next, we can look at Figure 10, which shows the top 5 rules applied to the RB/WR dataset, using the same right-hand side rule.

	lhs	rhs	support	confidence	lift	count
## [1]	{Longest.Reception=Zero-30_LongestReceptions,					
##	Rushing.Yards=101-150_RushingYards}	=> {Outcome=W}	0.01039412	0.7268314	1.442023	1647
## [2]	{Season=Regular Season,					
##	Rushing.Yards=101-150_RushingYards}	=> {Outcome=W}	0.01046354	0.7227550	1.433936	1658
## [3]	{Home.or.Away=Home,					
##	Receiving.Yards=Zero-50_ReceivingYards,					
##	Rushing.TDs=1}	=> {Outcome=W}	0.01545549	0.7205060	1.429474	2449
## [4]	{Receiving.TDs=0,					
##	Rushing.Yards=101-150_RushingYards}	=> {Outcome=W}	0.01012906	0.7194083	1.427296	1605
## [5]	{Home.or.Away=Home,					
##	Receiving.Yards=Zero-50_ReceivingYards,					
##	Longest.Reception=Zero-30_LongestReceptions,					
##	Rushing.TDs=1}	=> {Outcome=W}	0.01516519	0.7190305	1.426546	2403

Figure 10. Running Back/Wide Receiver Game Log ARM Output

We sorted this outcome by our confidence measure. However, our confidence, lift or support output wasn't to the same level as it was for QB, but we still have some interesting findings. Home games were again a pretty big driver, as was rushing yards in the range of 101-150 yards and longest reception being between 0-30 yards. As part of our data preparation, we needed to calculate fantasy points and use this calculation in some of our predictive modeling. We were able to calculate fantasy points for each game log using a formula derived from the FanDuel⁴ scoring system, the formula used is below.

- QB Fantasy = (PassYards * .04) + (TD * 6) + (Ints * (-1)) + ((RushYards - SackYards) * .1)
- RB/WR/TE Fantasy = (Rec * .5) + (RecYards * .1) + (TDs * 6) + (RushYards * .1)
- Cost = 3500 + (326.086957 * Fantasy)

Using the fantasy points calculated, we attempted to use support vector machine (SVM) to predict whether a player would score more or less than 15 fantasy points in a given game using historical data. The large dataset made using SVM difficult and resulted in us having to use a subset of the QB game logs, by taking a random 20% sample. Our results from the QB game logs resulted in correct predictions approximately 55.47%. Another model we used to predict fantasy football points was decision trees. We subsetted our data this time by recent seasons, which included 2016 and for 2010-2016. We then tested our model using both the QB and RB/WR dataframes. We also focused our model on the variables which will have a greater impact on our fantasy football point calculations, such as, yards gained, interceptions and touchdowns. For our QB data, the models were not very accurate, producing a root mode error of about 45%. Whereas, the running back data set had roughly a 75% accuracy. We can see visual representation of our decision tree models for 2016 in Figures 11 and 12 below. Some things to highlight, our model says that if a player started the game and was not a receiver,

⁴ <https://www.fanduel.com/>

then there is a 71% chance that they will score 10 or more fantasy point, this being supported by 43% of the original data. As for the QB results, a QB has a 68% chance of getting 18 or more fantasy points if passes attempted are greater than 21, sacks are 2 or less and its week 8 or further in the season.

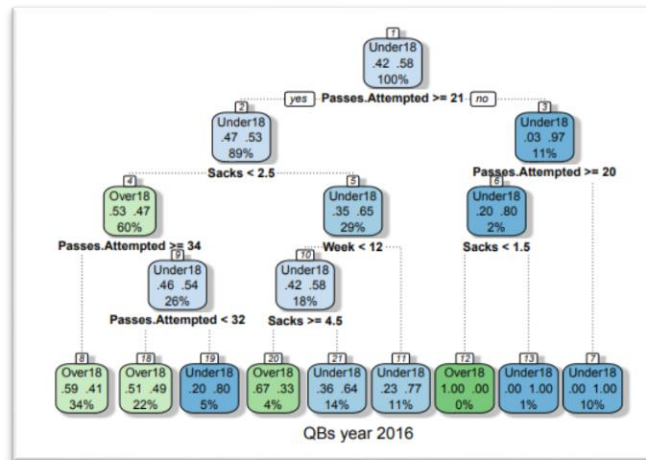


Figure 11. Quarterback Game Log Decision Tree Output for 2016

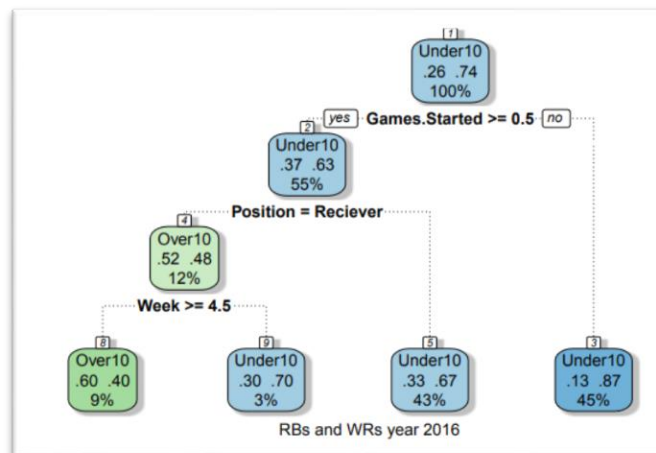


Figure 12. Running back/Wide Receiver Game Log Decision Tree Output for 2016

b. Insights

This project was a lot of fun and allowed us to use many of the skills taught in the data analytics class as well as prior classes. Due to our group having so many members, we had the resources to apply multiple different data mining techniques on our data. Also, having such an enormous dataset, it required us to create a research plan and perform a lot of exploratory analysis to better understand which data would be the most useful in answering our real-world problem. We also needed to do a significant amount of data cleaning, such as replacing NA's with zeros, binning specific data fields, merging datasets, such as the running back and wide receiver game logs. Finally, we began applying predictive modeling to our data to see if we could truly use historical football statistics to predict fantasy football points. These techniques included Clustering, Association Rule Mining, Decision Trees, Naïve Bayes and Support Vector Management. As discussed above, some of these yielded interesting reports, others proved to

just not work effectively such as Naïve Bayes, which only yielded roughly 50% accuracy in our model. As well as our SVM models, which yielded around 50% accuracy and took an enormous amount of time processing, even on a sample dataset, this limited our ability to rerun the models as much as we would have liked. Even though we didn't yield the results we had hoped for when starting this project, we got an enormous amount of real world experiencing using multiple different models.

IST 722 – Data Warehouse

a. Project Description

IST 722, Data Warehouse was an introduction to business intelligence and the techniques of building a BI solution. In this course, we had hands on experience using several different software including, Microsoft Excel, SQL Server Management Studio, Visual Studio and Power BI. For our group project, we were to use two databases belonging to a fictitious online retailer, Fudgemart and a fictitious online DVD by mail and video service, Fudgeflix. Our task was to build a data warehouse and Business Intelligence solution that will merge the two databases in order to facilitate future business decisions. Working alongside my team members, Kathi Fox, Aaron Talley and Dane Lyons, we decided to build a solution to provide Customer Satisfaction Reporting, more specifically understand which products and/or titles received the highest customer reviews. Our project plan consisted of the following steps and milestones;

- Design and Create the Data Warehouse utilizing the Star Schema design.
- Develop a Staging Environment between Source Databases and the Data Warehouse.
- Generate Data Flow Processes to move and transform data from Source Databases to the Data Warehouse.
- Analyze the data to gain valuable business insights towards the principle objective of the project.
- Present our Findings utilizing graphical visualizations.

We started by creating a high-level dimension and fact table worksheet, focusing on which attributes and metrics we wanted to collect from our existing databases, how to merge them into new dimension tables and any new derived columns that would be needed to answer business decisions. In Figure 13, you can see our completed star schema design, showing our dimension and fact tables that will be populated by both the Fudgemart and Fudgeflix databases.

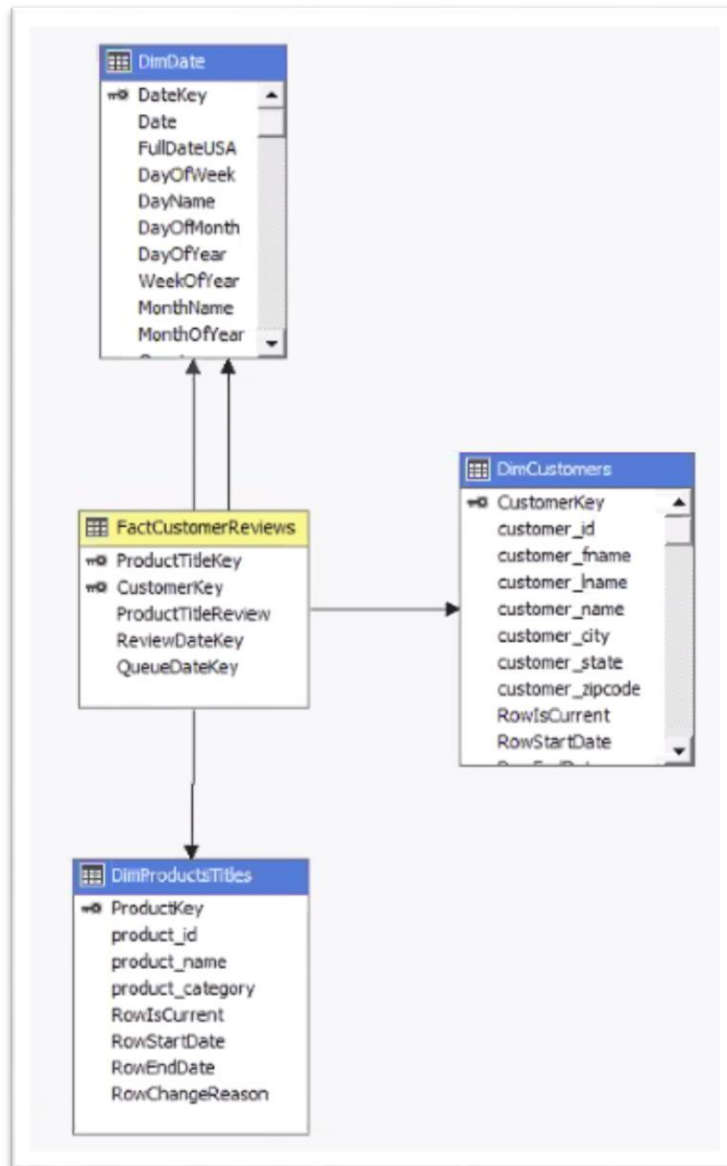


Figure 13. Star Schema

Our next step was creating the ETL processes that would feed data from our existing databases into our newly created Data Warehouse.

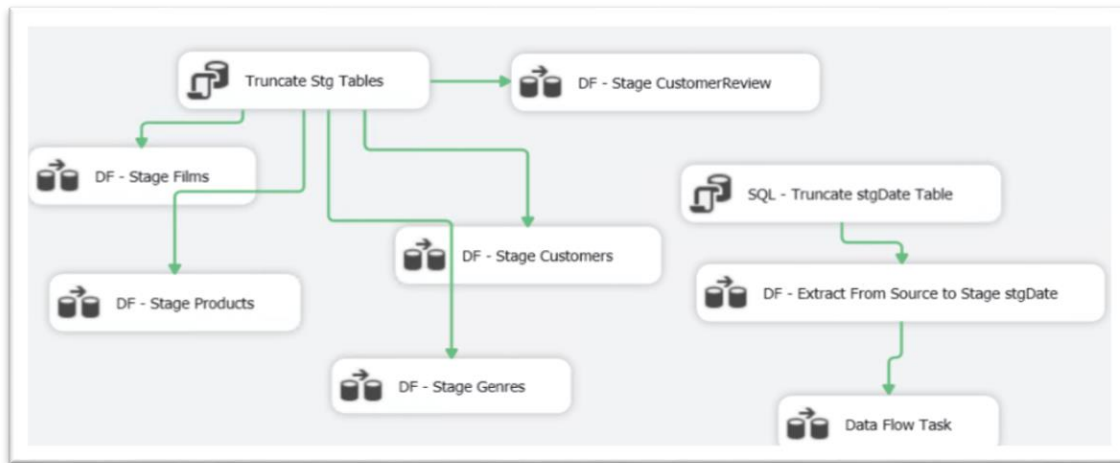


Figure 14. Dimension ETL Process



Figure 15. Fact Table ETL Process

Figure 14 depicts our creation of ETL feeds that processed data from Fudgemart & Fudgeflix to a staging environment and eventually to our new created dimension tables, merging both databases data into a single dimension for product & customers. Figure 15 depicts the ETL process of creating the Customer Review fact table metrics that are outlined in the star schema in Figure 13. After we completed our data warehouse and populated our tables with data, we connected our warehouse to Power BI and created a dashboard we hoped would provide more insight into our customers and their product reviews. Figure 16 displays screenshots of our two-page dashboard analyzing our customer review data. Some highlights we glean from the dashboard, the average rating for all products is 2.64. We also learned that 435 reviews have no rating, which means that we need to implement some rules into our data collection to prevent no rating reviews from being collected in the future. Movies have the majority of reviews, but they also have the most products. TV seasons have the highest average rating, whereas clothing has the lowest average rating. Most of our reviewing customers live in the coastal US and reviews peak at the beginning of the year, which makes sense following the end of year holiday gift giving season.

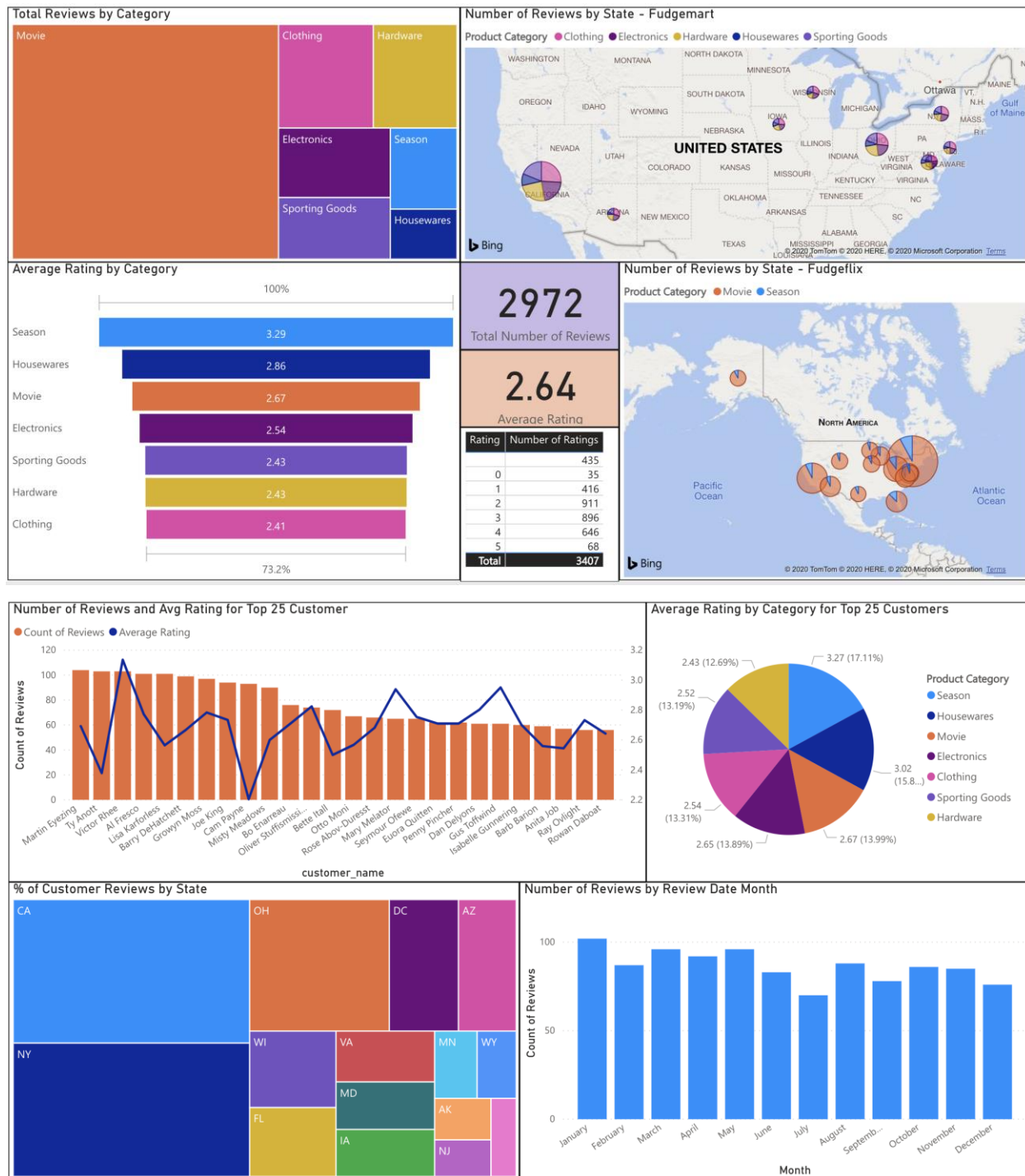


Figure 16. Power BI Dashboard

b. Insights

The topics covered in IST 722 – Data Warehouse was unlike any other course I took in this program and was equals part challenging and rewarding. This course taught me a new way to approach data collection, it really made me think about the data at my disposal and what business case I felt could be tackled using that data. We also had to really understand the specifics of the data being collected, what data type was the data being stored, how many

characters and the similarities and differences between our existing databases. Developing alternative strategies based on the data was a big lesson learned here. The existing data in our databases was all that we had to work with, which means that we really had to think about what insights could be gleaned from that data and how that data could be merged and structured in our data warehouse. There were several problem-solving meetings within the group to address staging table issues as well as rethinking our dimension attributes. For example, we wanted to incorporate genre into our dimension, specifically for our movie/season products. However, we discovered that genre was a many to one relationship with our titles, meaning that for every movie title there could be multiple genres. This resulted in our movie titles being duplicated, having each row be the movie title with a different genre. After several failed attempts to work through this issue, we made the business decision to remove the genre field and move along without it. In addition, this course was a great practical course on the ethical dimensions of data science practice. Part of data science means collecting and storing of sensitive and private data, whether it be company financials or private customer information. This course was an excellent introduction in best practices of storing sensitive data and how you can segregate, create securities as well as maintain access levels to all different data.

IST 719 – Information Visualization

a. Project Description

Information Visualization is a course dedicated to teaching the skills and techniques through R and Adobe Illustrator to effectively create and present analytical visualizations. Our solo project in this course was to take the skills learned regarding design concepts, storytelling through visualizations and effective communication to create a poster of a topic chosen by us. I began this process by selecting data, I chose to use weekly retail sales data of avocados in the United States. This data was collected by the Hass Avocado board⁵ and was provided for download on their website. I was able to download every year available, which was 2017 – July 2020. Looking at the raw data, I planned to answer 3 business questions;

- Is there a specific area where organic avocados outsell conventional ones?
- Pricewise, what is the best time for a consumer to buy avocados?
- Has the rise in popularity of avocados translated to an increase in sales, specifically in a new region?

Using R Studio, I began by uploading my 4 datasets and merging them into one dataframe. After performing some exploratory analysis, I discovered that my dataframe contained Total U.S. data, Regional data as well as individual areas. I decided to segment my dataframe into three separate dataframes, one containing only the Total U.S. weekly avocado sales data, another

⁵ <https://hassavocadoboard.com/>

containing the regional data and the third containing the individual cities. After the data cleaning process, I began creating some plots to answer my business questions.

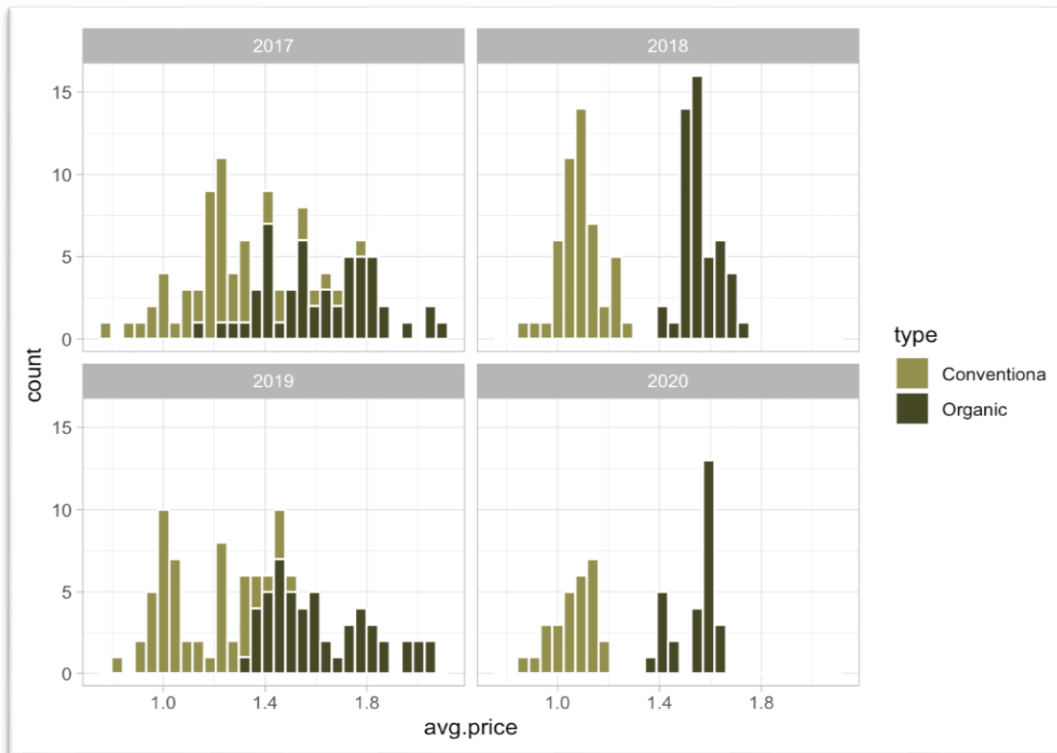


Figure 17. Average Price per Avocado Histogram Facet Plot

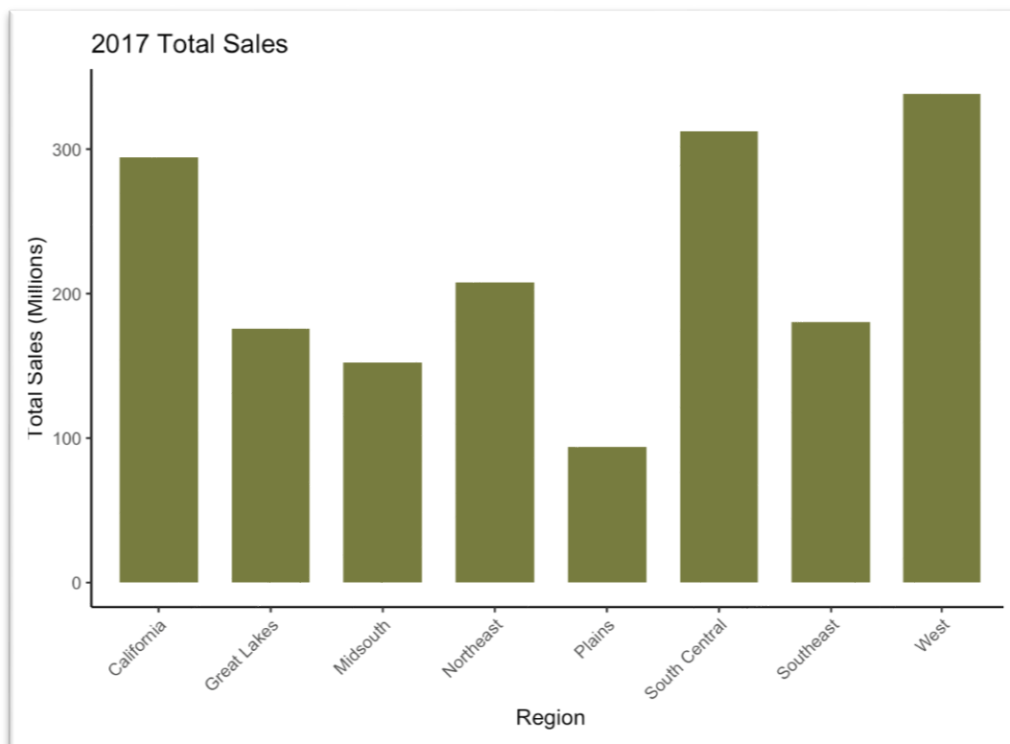


Figure 18. 2017 Total Sales by Region

Figure 17 and 18 are examples of plots created directly in R, before being edited in Adobe Illustrator and placed into my poster. The next step in my process was to select a color palette and consider how I was going to use color to tell my story. Figure 19 consist of the colors I used in my poster. Very Light-Yellow Green was used as the background, I used Very Light Avocado Green for any plot representing conventional avocados and Very Dark Avocado Green for organic avocados. I used the Dark Brown for text, including headers and axis on my graphs.






Dark Brown	Very Light Yellow Green	Very Light Avocado Green	Medium Yellow Green	Very Dark Avocado Green
				
HEX #5c2d04	HEX #dbdbb2	HEX #908c48	HEX #73793b	HEX #42421c
RGB 92, 45, 4	RGB 219,219,178	RGB 144,140,72	RGB 115,121,59	RGB 66,66,28
CMYK 0,51,96,64	CMYK 0,0,19,14	CMYK 0,3,50,44	CMYK 5,0,51,53	CMYK 0,0,58,74

Figure 19. Color Palette

The next design element I focused on was choosing a font. I decided to only utilize one font throughout the poster, bolding for headers and regular for the rest of the text, this font was ‘**Georgia**’. The last design element was choosing how to layout out my poster, the overall poster is broken out into four sections. The top, having the title, author’s name and a brief introduction to the topic. The largest section is the middle, which is broken out into two equal parts, the left has the focal visualization which is a long bar graph representing total volume of avocados sold by cities and answers the first business question. To the right of the focal visualization are three graphs representing the seasonality of avocado buying and price. The bottom section represents regional reporting of volumes sold and answers the last business question. The poster in its entirety can be seen on my GitHub repository, github.com/czimmerb/Masters_Portfolio/tree/main/IST719_InformationVisualization

b. Insights

Information Visualization really made me think about the story and what information I wanted to tell with my graphs and charts. There is much more to communicating data then just putting together graphs. From understanding which graphs to choose, to colors, fonts and most importantly the layout of your presentation. I found this class to be incredibly helpful, in my current job as a Business Analyst at Popular Bank, I create a lot of reporting and dashboards

and never really thought past the data and graphs. Being able to successfully communicate is a very important skill for data scientist, having the ability to understand who your audience is and being able to adjust your reporting, visualizations and even words to fit the audience you are trying to reach is a skill this class has taught me.

Conclusion

My portfolio demonstrates that I have learned, practiced and understood the programs learning goals and have provided numerous examples of how I've mastered them. From understanding the major practice areas of data science in majority of my courses, but specifically Introduction to Data Science and Data Analytics. Using descriptive analytics to get a sense of the current state of things and to better understand the data. Modeling for predictive and forecasting purposes, as well as using the analysis to make and recommend further actions. Collecting and organizing data, all projects required either collecting data or organizing, such as Information Visualization where the data collected had to be first merged and then segmented into 3 dataframes based on total US, region and area. Identifying patterns in data via visualization, statistical analysis and data mining, all four of my projects submitted in my portfolio demonstrates some form of this learning goal. Every project requires exploratory analysis, it allows you to get a sense of your data, identify issues, such as missing values, and allows you to develop an analysis plan from there. As well as creating dashboards in Power BI, information poster and the use of statistical models and accuracy graphs to make predictive analysis. Develop alternative strategies based on the data was exemplified in my project for Data Warehouse where we were given two databases with the goal of merging data into one cohesive data warehouse. We quickly learned as we progressed that some of our initial plan would not work due to the structure of our existing data. Develop a plan of action to implement the business decisions derived from the analyses, our Boston crime data analysis spotlighted some interesting trends and could be extremely beneficial for staffing and scheduling in order to keep crime at a steady or lower rate going forward. Demonstrating communication skills, Information visualization was instrumental in learning this program goal as it really focused on how you are presenting your data, your data story and ensuring that you are communicating effectively to your audience. Lastly, synthesize the ethical dimensions of data science practice, Data Warehouse really demonstrated this topic as it highlighted the importance of instituting the proper data procedures to ensure that all data is protected.

Bibliography

References

(n.d.). Retrieved September 01, 2020, from <https://streeteasy.com/>

AnkurJain. (2018, October 04). Crimes in Boston. Retrieved September 13, 2019, from <https://www.kaggle.com/ankkur13/boston-crime-data>

Gillies. (2017, June 9). NFL-Statistics-Scrape. Kaggle.Com.
<https://www.kaggle.com/kendallgillies/nflstatistics>

Hass Avocado Board. (n.d.). Retrieved September 5, 2020, from <https://hassavocadoboard.com/>

Inside Airbnb. Adding data to the debate. (n.d.). Retrieved September 5, 2020, from <http://insideairbnb.com/index.html>

National Centers for Environmental Information (NCEI). (n.d.). Climate Data Online. Retrieved September 13, 2019, from <https://www.ncdc.noaa.gov/cdo-web/>