Portfolio Milestone
MS Applied Data Science

Jeffrey Kao
SID 999002123

Table of Contents

1. Introduction

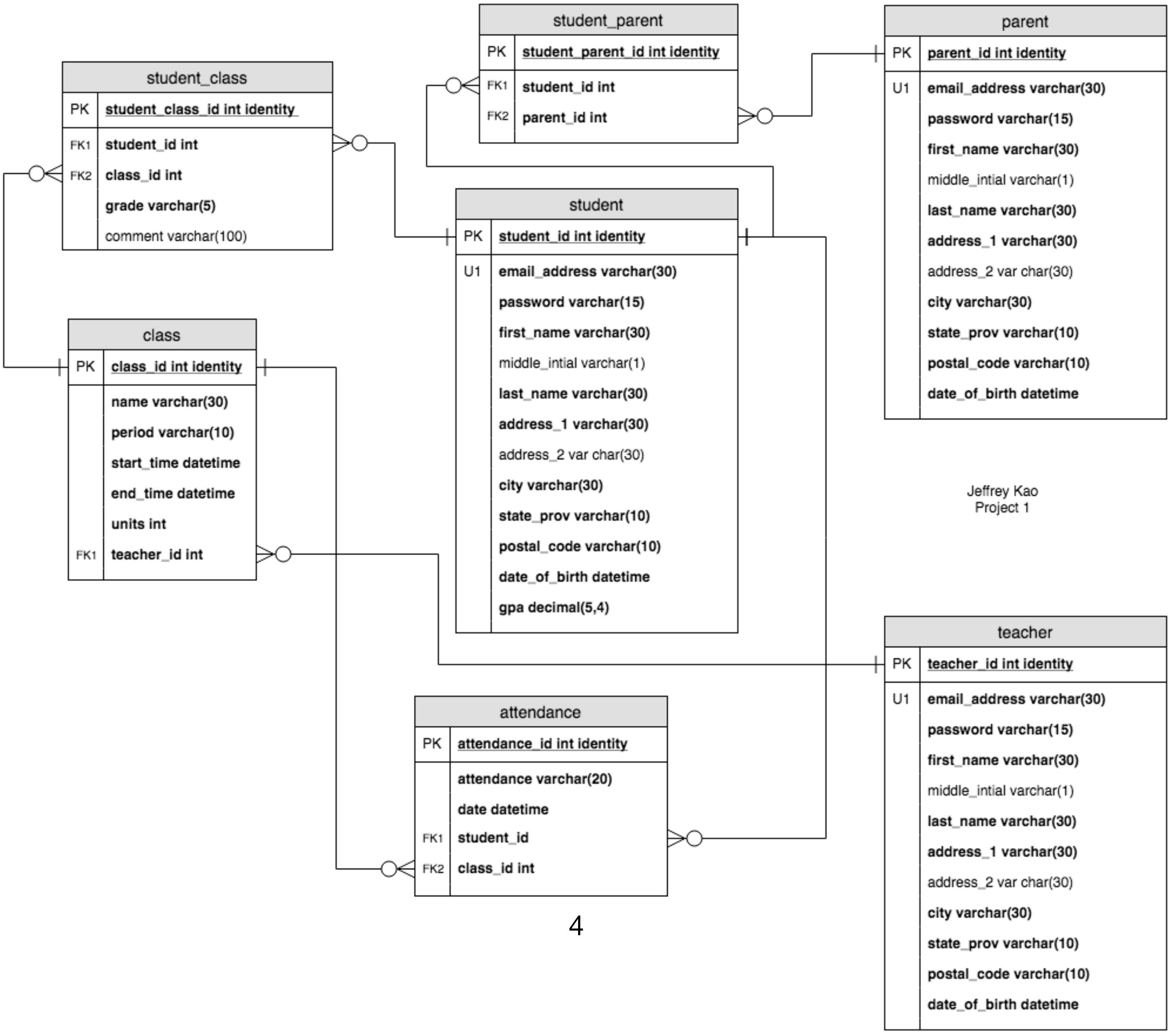
Throughout the MS Applied Data Science Program at Syracuse University, I have achieved the program learning goals with the help of my instructors and fellow students. I am extremely grateful for the opportunity to present my work to the faculty to show my growth as student in the field of data science. In this portfolio I will outline 4 key projects that have exemplified the following learning objectives of this program.

1. Describes a broad overview of the major practice areas in data science.
2. Collect and organize data.
3. Identify patterns in data via visualization, statistical analysis, and data mining.
4. Develop alternative strategies based on the data.
5. Develop a plan of action to implement the business decisions derived form the analyses.
6. Demonstrate communication skills regarding data and its analysis for managers, IT professions, programmers, statisticians, and other relevant professional's in their organization.
7. Synthesize the ethical dimensions of data science practice.

2. IST659: Data Admin Concepts & Database Management - High School Database

   a. Project Description

This project from IST659: Data Administration Concept & Database Management with Professor Harper is a key example that laid down the foundations about data. This was one of the first classes that I took and I gave me a lot of insight about the organization of data. This project focused on designing a database for a school management system for a high school. The goal of this management system was to be able to maintain both attendance and grades for students at a school. In addition, the information of each student's parents was also stored in this database that was created with SQL. Logical models were created in order to map out and organize the relationships between all the necessary data.



Also importantly I had to address the different types of data that where able to be stored in the database software that we were using, Microsoft Access and SQL Server DBMS. Types of variables and other limitation are bases on which database software used. Database was then generated with code that was written based on the model above and data was uploaded into the database. With the creation of this database I was able to readily access information quickly and efficiently. Question that could be answered where: Which class is the largest at the school? How may parents have multiple students at the school? What is the average GPA for a certain grade level? These questions may seem rudimentary, but they are important information that should be readily available in any setting at a school.

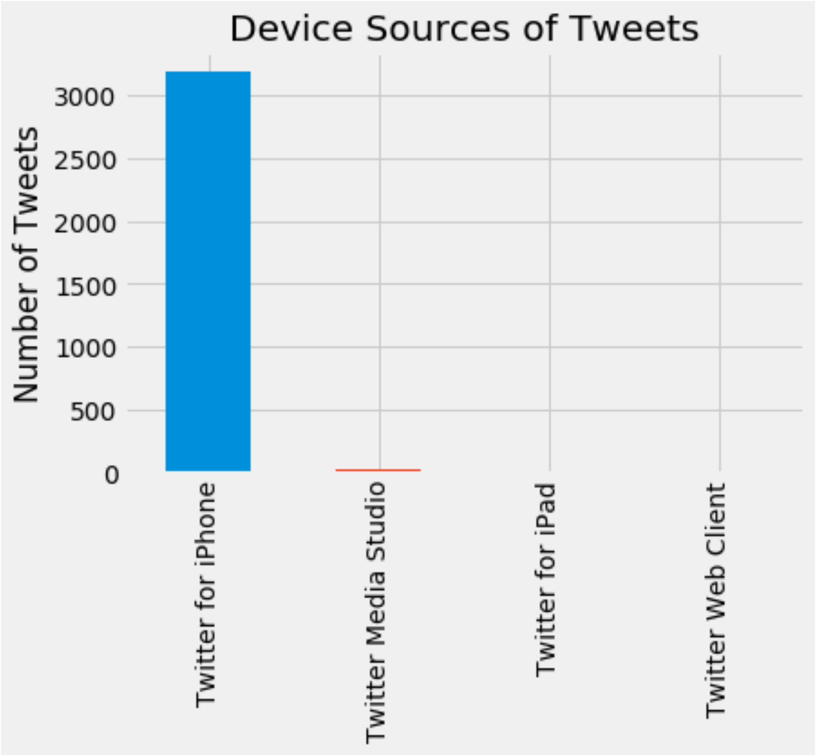
b. Learning Outcome

Even though this was on the first projects that I did for my degree, the impact was everlasting. This was my first foray into using Microsoft Access and SQL Server DBMS. It taught me key fundamentals in how data is created, stored, accessed, and analyzed in a database setting. These skills are key for any data scientist or analyst because almost all data is stored in databases. With this project I was able to get a full understanding of how data is collected and organized in database setting. Furthermore, I was able to derive simple insights that would be necessary for any professional setting.

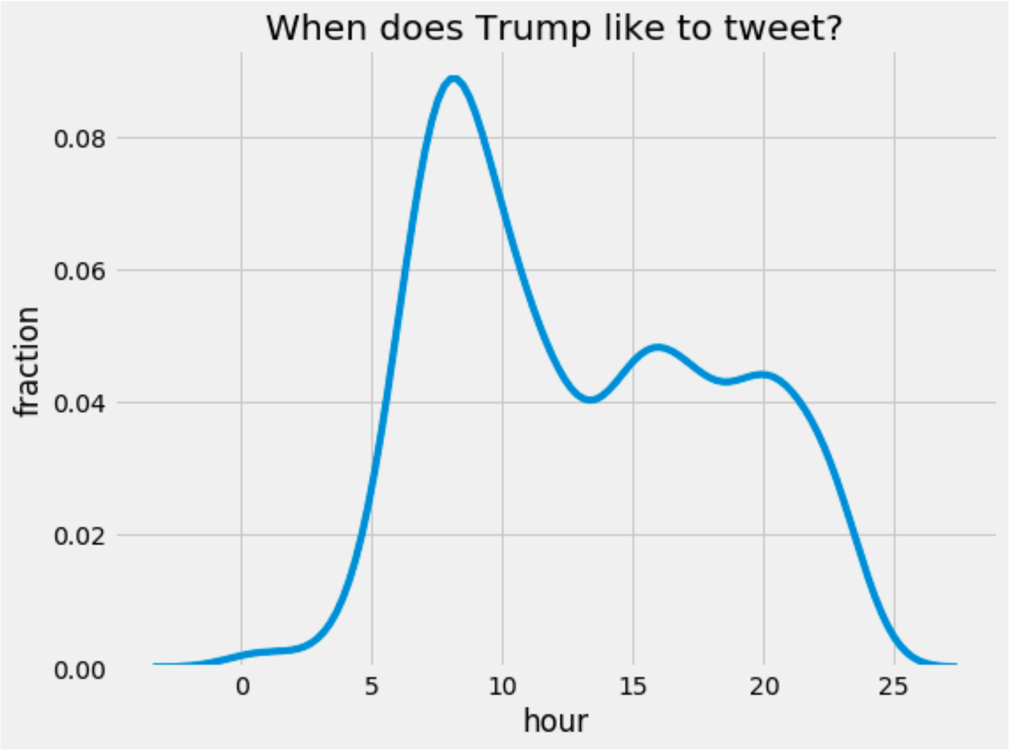
3. IST652: Scripting for Data Analysis - Trump's Tweets
    a. Project Description

This was one of the most interesting projects that I did during the course of program. IST652: Scripting for Data Analysis with Professor Debbie Landowski taught me different ways to use Python to collect, organize, and analyze data. This project was the culmination of the course and was interesting and insightful. I first used signed up for a Twitter developer account and was able to use the Twitter API to aggregate a collection of Presidents Trump's tweets. Because of his prolific Twitter usage there was a lot of data. I was able to collect and clean up the data to be more readable for both the user and the computer for analysis using techniques learned in class. Examples of the was removing unnecessary HTML or normalizing the time zone. Furthermore after I cleaned up the data I performed sentiment analysis on the tweets. Here I was able to identify some patterns in data and visualized them. Some of the examples of analysis and visualization are shown below.

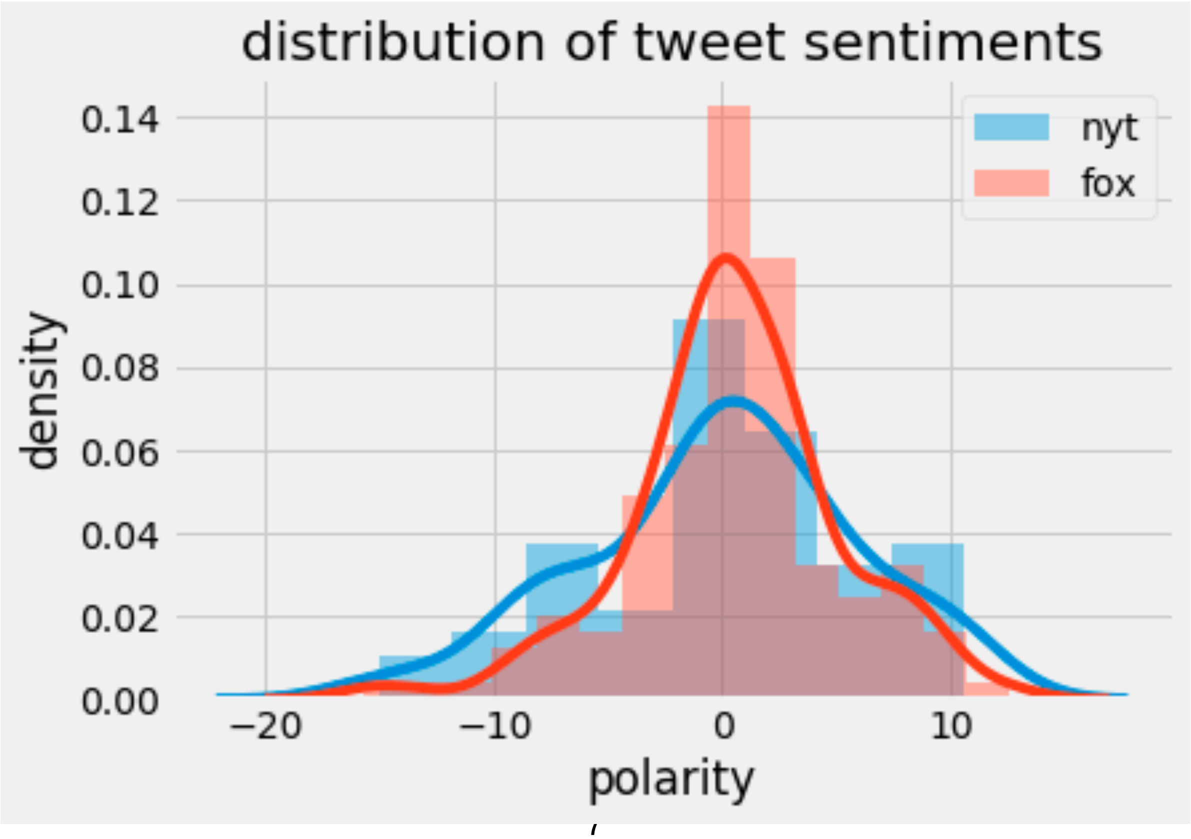
What is the source of Trump's tweets?



When does Trump like to tweet?



Distribution of sentiments of tweets about Fox and the NYT



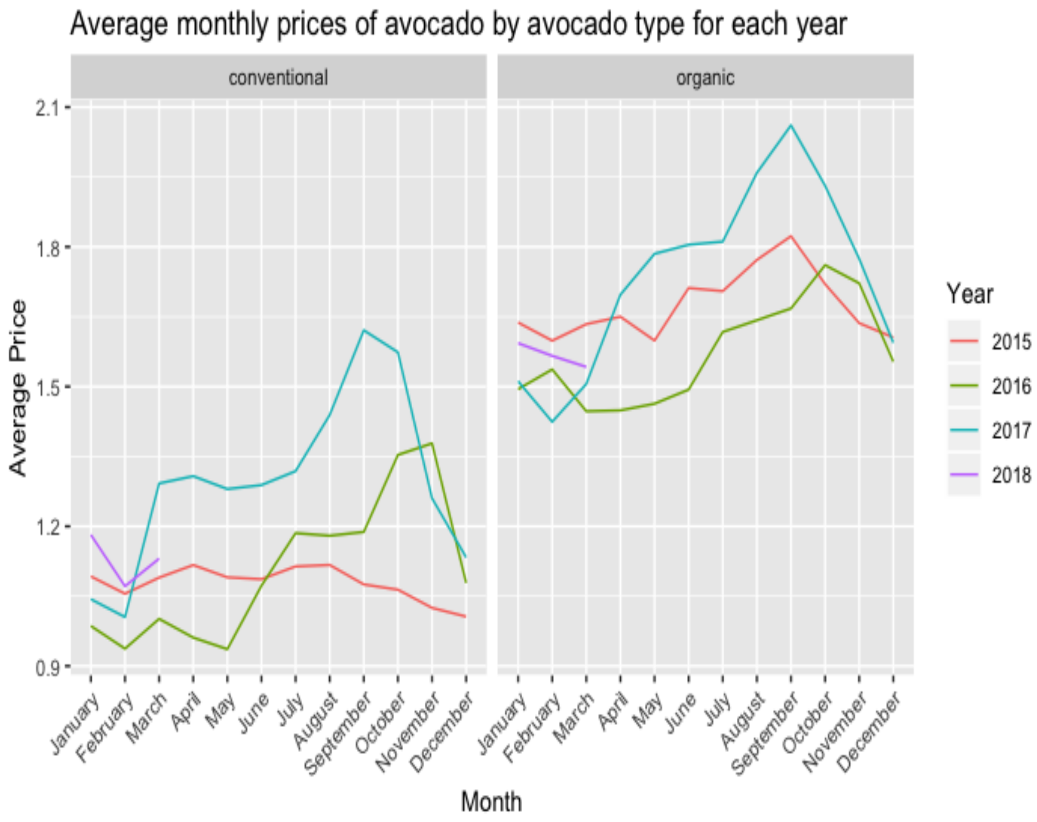
b. Learning Outcome

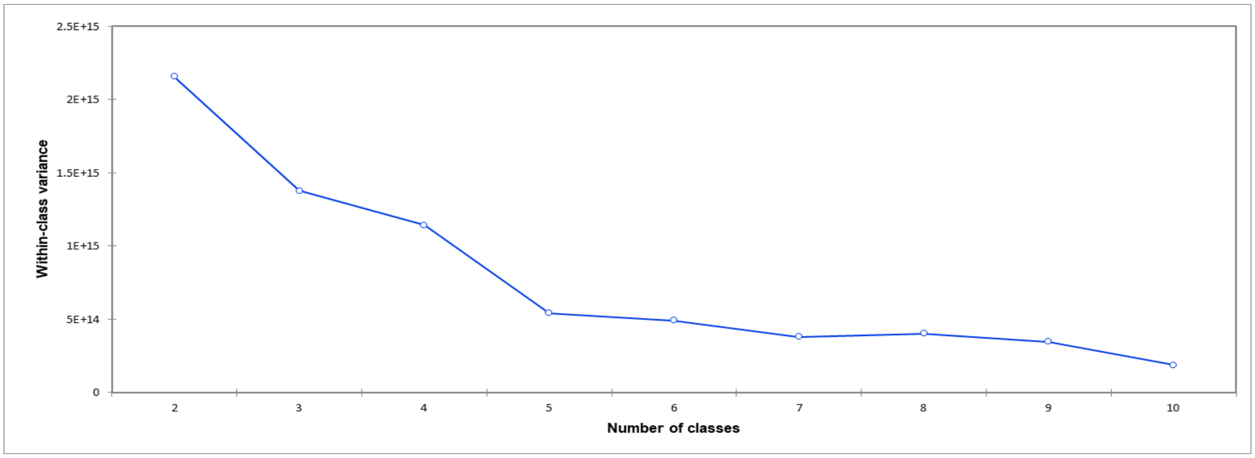
This project tackled a lot of learning objectives that were key to the applied data science program. Using an API is common way to obtain data from large companies that have a plethora of data like Twitter and Facebook. Because of the popularity of these applications the data is immense and many insights can be derived from it. The practice of using APIs is important not only in data science, but in all software development so it was very beneficial to me. I was able to access the Twitter API to collect the data and organize it. Visualizations were key for this project because the data was very text heavy and there was little you could get from the data looking at in chart format. Furthermore, I was also able to communicate my insights and findings as a presentation for my class, which showed my communication skills about data to relevant other data-minded people in my class. I also tackled the ethical dimensions of data science in this project by disclaiming that while sentiment analysis is improving all the time it is not necessarily always correct. Sentiment analysis must always be taken with a grain of salt and I echoed in my presentation in class that decisions should not always be made with out additional data to back it up. In addition, I brought up the perils of essentially tweeting too much because of the inability to delete what you tweet. Once you tweet it out it is most likely stored in a database in perpetuity.

4. MAR653: Marketing Analytics - Analysis of Avocado Pricing and Distribution

a. Project Description

In MAR654: Marketing Analytics taught by Professor Andrew Petersen, I did a group project that was an overview of the avocado market in the US. Data was collected online that had sales records weekly of avocados in the US. We were able to look analyze sales trends and look at the different habits of avocado buyers. Also we were able to segment the customers into clustered groups and recommend a plan of action to boost avocado sales. Analysis was done using R and Excel. Some examples of visualizations that we derived from the data are shown below.





K-Means clustering reached 5 segments of customers.

With our analysis, we were able to paint a picture of the avocado market in the USA. Key takeaways that we saw were: the further you are from Mexico the more expensive the avocado is, pricing is steady in places with low demand, the words "X-Large" have a negative effect on sales, and organic avocados have a high correlation to price.

b. Learning Outcome

In this group project, I was able to to communicate with other team members and work to together to gather insights about avocados. We also explored different marketing strategies and business directions for avocados and presented them to our peers. By identifying sales patterns via visualizations, statistical analysis, and data mining we were able to make marketing strategies and develop a plan of action for the avocado market. Examples of such strategies and plans based on the data and further analyses: elimination of the "X-Large" avocado category and just grouping them with large and pricing segmentation to our different clusters that we identified. Our communication skills were further tested with the form of a presentation to out peers and instructor in class.

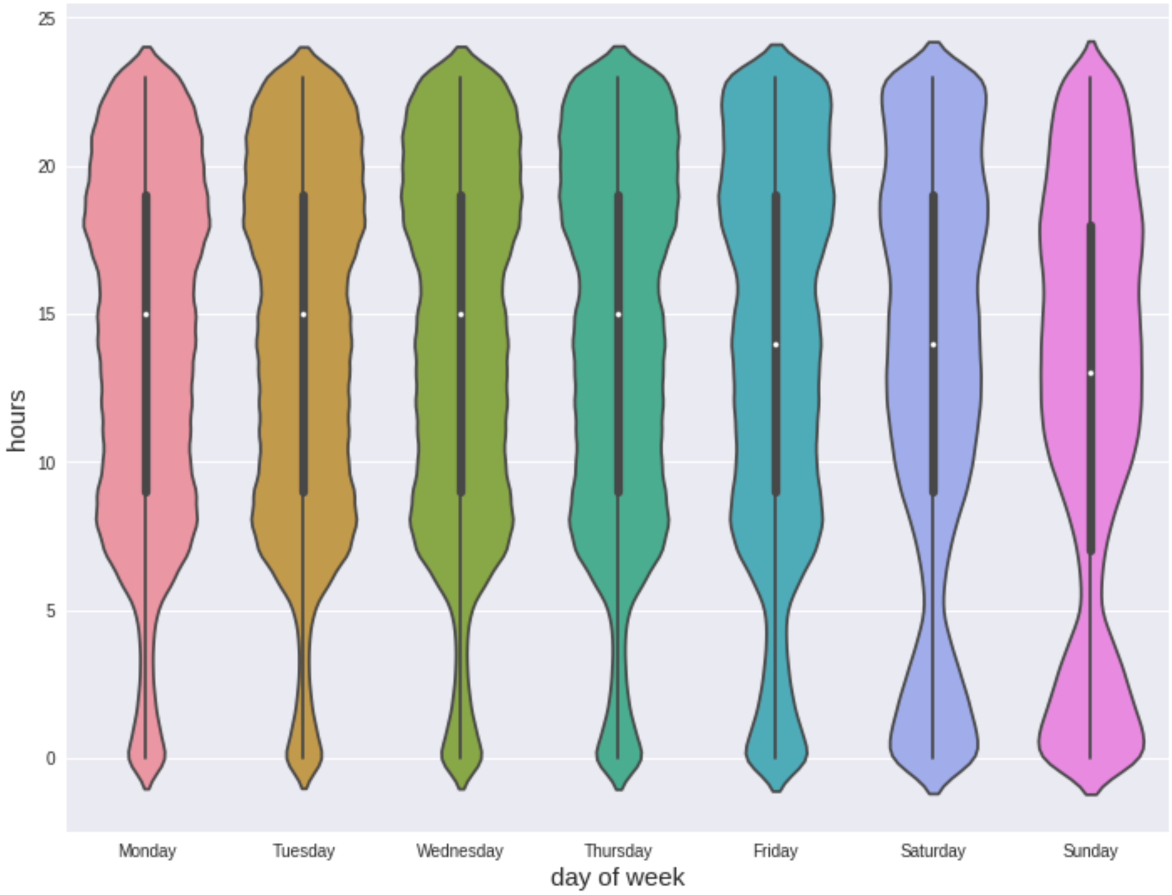
5. IST718: Big Data Analytics - NYC Taxi Duration

a. Project Description

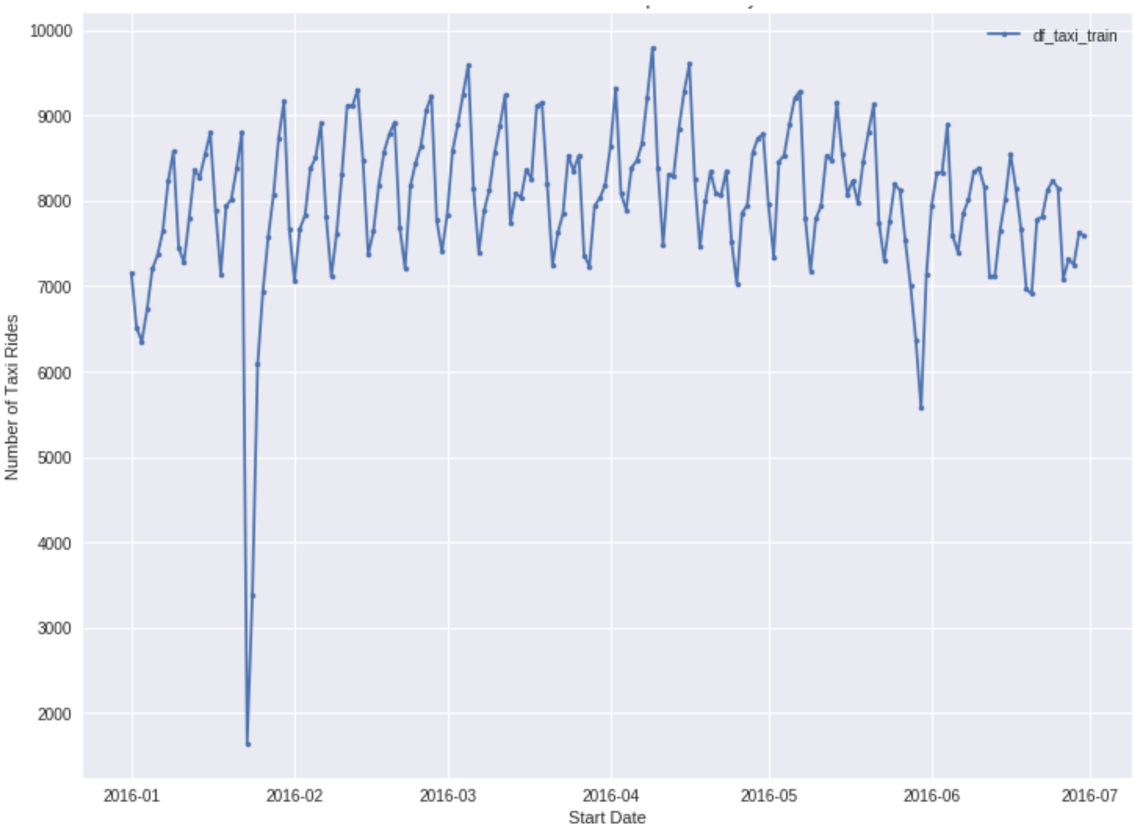
In IST718: Big Data Analytics with Professor John Fox, I did a group project that was by far the most challenging and time consuming of anything in this program. Using challenge data that was provided on Kaggle we sought to predict trip duration rides in NYC based a a variety of variables including pick-up and drop-off time, pick-up longitude and latitude, the number of passengers, trip duration, collision data, and traffic data. The goal was to accurately predict ride duration so that we could find the best time to take a taxi thus optimizing route, time, and happiness. As a group we cleaned and optimized the data for analysis and we also brought in other datasets such as weather and collision data in order to try to more accurately predict the duration of the taxi ride. Using different prediction models, linear model and random forest, we were able to to predict the with approximately 80% accuracy the duration of a taxi ride.

Wrangling such large datasets and combining them with other data was one of the biggest challenges in this project. Some examples of visualization graphics are displayed below.

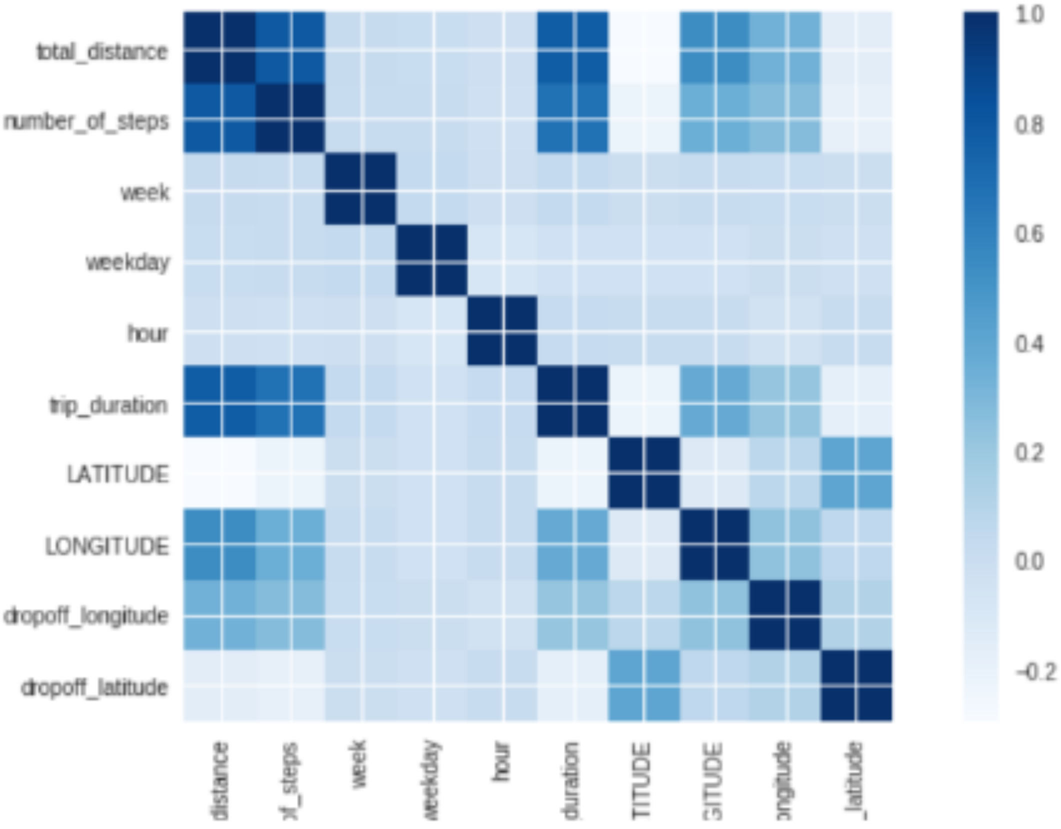
Counts or Rides per Day by Hour



Number of Taxi Rides Per Day



Correlation Analysis



b. Learning Outcome

I learned the most from this project out of anything in the entire applied data science program. It was tough and challenging task to wrangle and clean so much data and make accurate predictions and assessments with it. Collecting, combining and organizing different datasets was a good experience. Using predictive models are a major practice area in data science that are increasingly important. With such a large amount of data, visualizations were key in order to fully understand the data. A lot of statical analysis and data mining went in to figure out the best predictive models to work with. Based on the data we had to come up with alternative strategies for instance removing collision data from our predictive modeling because it was not helping. And finally communications skills of all this data and analysis was presented to our peers and instructor.

6. Conclusion

All the projects above have demonstrated the achievement of the learning objectives of the applied data science program. Throughout the program I have tackled the major practice areas of data science. Data was collected and organized in many shapes and forms like from databases, APIs, and data scraped from the internet. Visualization were done in several different languages and programs in order to more clearly show the data quickly and efficiently. Statical analysis and data mining techniques were done at every junction like correlation, regression, and clustering in order to make accurate assessments of the data. After assessing and analyzing data, I was able to make alternative strategies to either make more accurate predictions or better informed marketing and create a plan of action for business decisions. Communication skills were developed and driven by class discussions and presentation to other data-minded peers. And the ethical dimensions were tackled by investigating privacy and using only relevant data needed to make informed decisions and outcomes.

All in all, the skills that I have acquired from the applied data science program at Syracuse University have better prepared me for the wide range of problems related to data in the real world. I will be able to collect, manage, and analyze data with the various techniques that I have learned from this program. I thank all the faculty and staff that have helped me along the way.

References

Kao, Jeffrey (2019) https://github.com/jeffthestampede/PortfolioMilestone