Syracuse University, School of Information
M.S. Applied Data Science

N'Dea Jackson
SUID: 201783916
Email: njacks01@syr.edu
Portfolio Milestone

Table of Contents

1. Introduction

The Applied Data Science program at Syracuse University is offered jointly by the School of Information Studies and the Martin J. Whitman School of Management. This Master of Applied Data Science degree program is designed to be a professional program of study, with a strong emphasis on the applications of data science to enterprise operations and processes, particularly in the areas of data capture, management, analysis, and communication for decision making. The Applied Data Science degree program has seven learning objectives that students are expected to be able to complete including:

1. Describe a broad overview of the major practices 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 their data
5. Develop a plan of actions to implement the business decisions derived from the analyses
6. Demonstrate communication skills regarding data and its analysis for managers, IT professionals, programmers, statisticians, and other relevant professionals in their organization
7. Synthesize the ethical dimensions of data science practice (e.g., privacy)

This portfolio will take a look through four courses, that are a part of the Applied Data Science curriculum, and how the assignments that were completed in these courses align with the learning objectives for the Applied Data Science program.

2. Jackson Memorial Hospital Database

Project Description

The first project selected for this portfolio was from IST 659: Database Administration. Throughout this course, under the direction of Professor Chad Harper, students learned practices that focused on the definition, development, and management of databases for information systems through the use of query language and search specifications. As a final project that would be used to demonstrate the skills learned

throughout the course of the term, a hospital database was created. This fictitious hospital, Jackson Memorial Hospital, enlisted the help of data scientists in order to create a functional database that would allow them to provide the best medical assistance to their patients. The method being used by the hospital at the beginning of the project included physical paper charts and a number of siloed systems that were individual to specific departments. The expected outcome of this project was to create a database that would afford the hospital a better way to maintain all hospital records, including doctor-patient assignments, patient information, patient room assignments, and patient invoice status. This would in turn minimize paperwork and siloed applications as well as increase the level of patient care and the level of staff efficiency.

Data for this project could either be downloaded from an existing database or created by students. For this particular hospital database, 19 fictitious patient records were created for analysis. Both conceptual and logical models (Figure 1 and Figure 2) had to be developed in order to efficiently map out the relationships between all of the moving parts of the hospital, including the patients, the doctors, the hospital bills, the rooms, and the hospital itself. The tables were created using an online platform called draw.io. Query code was created to answer questions about the data such as what patients were assigned to a specific doctor, the insurance provider of a patient given their name, the status of a patient's invoice given the invoice ID, what room a particular patient was assigned to, and the age of a client given their name. These questions were chosen because they span a number of different departments at a hospital as well as would provide answers to questions that would make hospital staff's jobs easier.

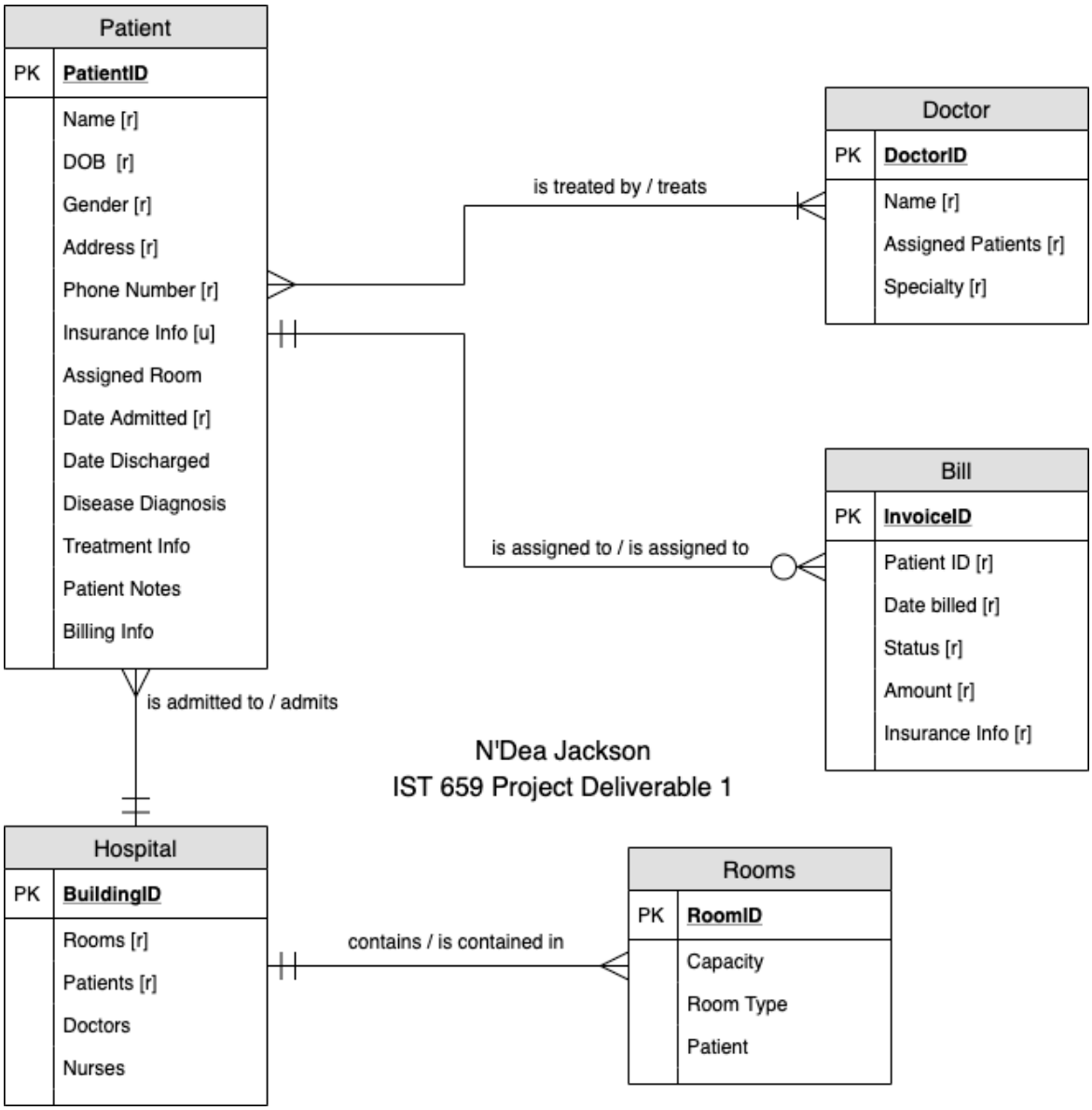


Figure 1 - *Conceptual Model for Jackson Memorial Hospital Database*

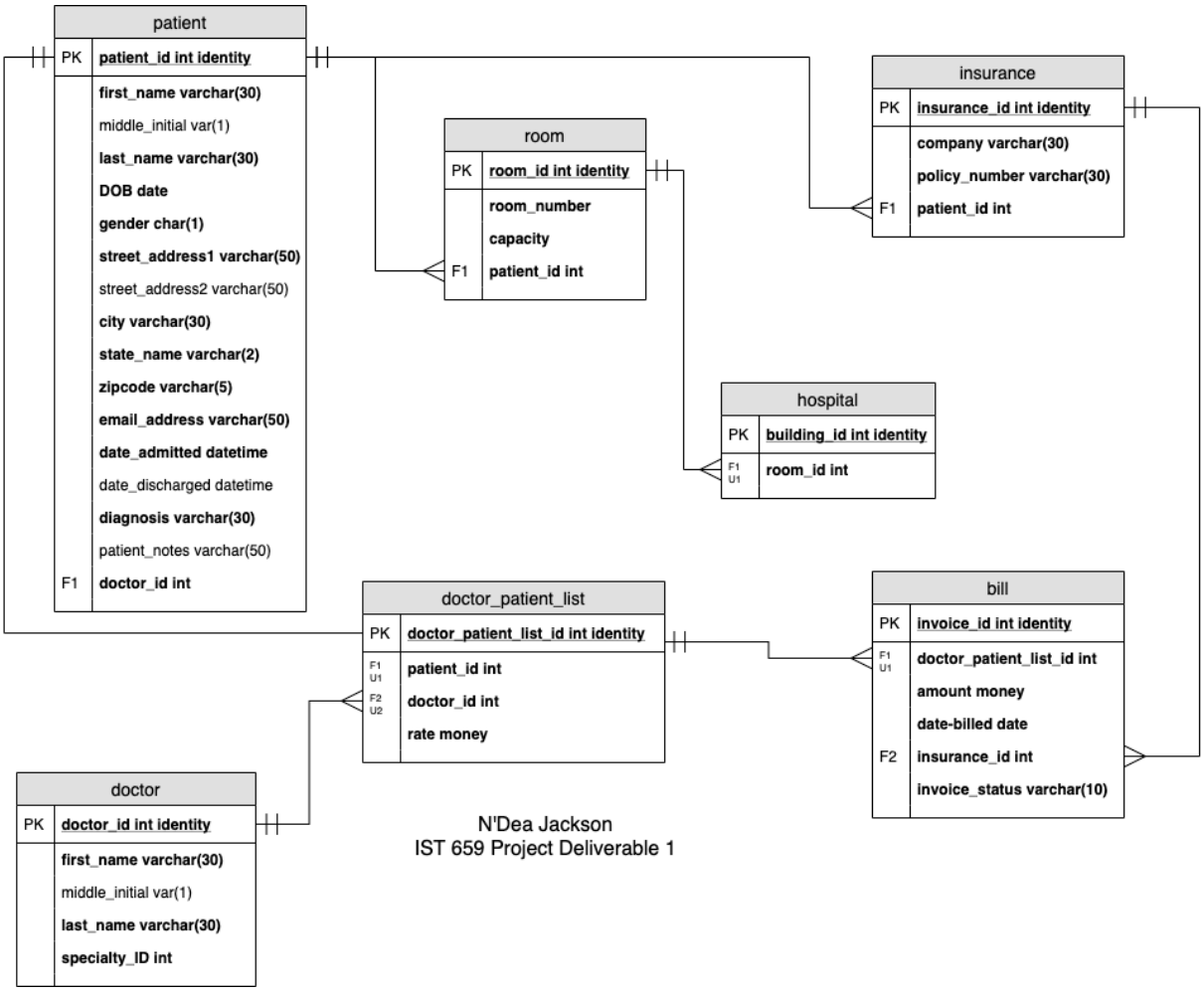


Figure 2 – *Normalized Logical Model for Jackson Memorial Hospital Database*

Question 6: How old is Cody Carrington?

```
SELECT
        patient.patientID,
        patient.first_name AS PatientFirstName,
        patient.middle_initial AS PatientMiddleInitial,
        patient.last_name AS PatientLastName,
        (DateDiff(year, DOB, GETDATE())) AS AGE
FROM patient WHERE patient.last_name = 'Carrington'
```

Results | Messages

	patientID	PatientFirstName	PatientMiddleInitial	PatientLastName	AGE
1	15	Cody	W	Carrington	70

Figure 3 – *Patient's age look-up*

Reflection & Learning Goals

By completing this project, one of the first things that students were able to understand was the importance of mapping the relationships between entities. This step helps to set the tone for the coding that data scientists will have to complete. If there is information that is missing or relationships that are misrepresented, this can dramatically change the effectiveness of the database itself.

This project allowed students to both collect, or create, and organize data within a SQL database, develop some alternative strategies for efficiency based upon the data that was used, develop a plan of action to implement the business decisions that were derived from the database analysis, and to demonstrate communication skills through report writing. The foundational basics that were learned in this Database Management course are transferrable to more advanced courses in the Applied Data Science program such as Advanced Database Management and Data Warehousing, as well as transferred directly into the workforce. This project required students to identify stakeholders who would be concerned with a project such as this based upon the data chosen, create data questions that would later be answered through the analysis performed, create business rules that governed the relationships that existed, as well as create a glossary of relevant terms that would aid in third-party understanding. All of these skills contributed to students' communication skills by giving them practice with tailoring presentations to stakeholders.

3. Boston Crime Data Report

Project Description

The second project chosen to display the learning objectives of the Applied Data Science Program comes from course IST 719: Information Visualization. Throughout this course, under the direction of Professor Gary Krudys, students learned a broad introduction to data visualization for information professionals. Students who took this course were introduced to skills and techniques that are related to information visualization through the use of R programming and Adobe Illustrator. Some of the skills that were acquired by completing this course and that were applied in order to complete the final project using

data cleaning techniques, controlling the R graphics environment, developing custom plots, and discussing issues that are related to the ethics of data visualization. The dataset that was used for this project was an incident crime report dataset that was provided by the Boston Police Department (BPD) to document the initial details surrounding an incident to which BPD officers responded. The dataset contained records from new crime incident reports, which included a reduced set of fields that were focused on capturing the type of incident as well as where and when it occurred. The incidents there recorded in the dataset began in June 2015 and ended in September 2018.

A data story had to be developed for each final project. The data story for this project was that newly elected officials, partnered with the Boston Police Department, enlisted the help of data scientists to help them compile the data on crimes that have occurred around the city in previous years. By gaining insights from this data, the Boston Police Department would be able to get a better understanding of which neighborhoods could benefit from an increased police presence. This study would also give police and officials a better understanding of programs that they could put into place in certain communities in order to curb criminal activity. The data would also then be made publicly available so that potential residents would be able to have a better understanding of their surroundings. Some of the questions that this analysis aimed to answer included the most common crimes that occur in Boston neighborhoods, the different types of crimes that occur, the change in the frequency of crimes across different time periods, the level of time across different years, the districts with the highest crime rates, and the months in which the highest crimes occur. A variety of techniques were used to create the poster including heat maps, histograms, density plots, and exploding pie charts.

After completing analysis, heat maps were created of the top three offenses by frequency in Boston neighborhoods: *Motor Vehicle Accident Response, Larceny (including shoplifting, bicycles, purse snatches), and Medical Assistance (including sick/injured/medical, suicide attempt, sudden death, and death investigation)..* The heat maps can be seen below in figure 4. The epicenter for larceny related events appears to be centered around the Back Bay and North End areas of Boston. When observing the frequency of crimes occurring, histograms and density plots were used to represent the data across months and days of the week, as well as hour of the day. Over the course of the four years being analyzed, 48,496

incidents were recorded on Fridays, making Friday the day of the week with the highest frequency of crime overall. Over the course of a day (24 hours), the trend in crime appears to drop after midnight and rise steadily until it reaches its peak at 5PM. Lastly, on a yearly scale, crime throughout the course of the year is relatively steady until the summer months. Around May, crime begins to rise and steadily increases until August when the summer months come to an end.

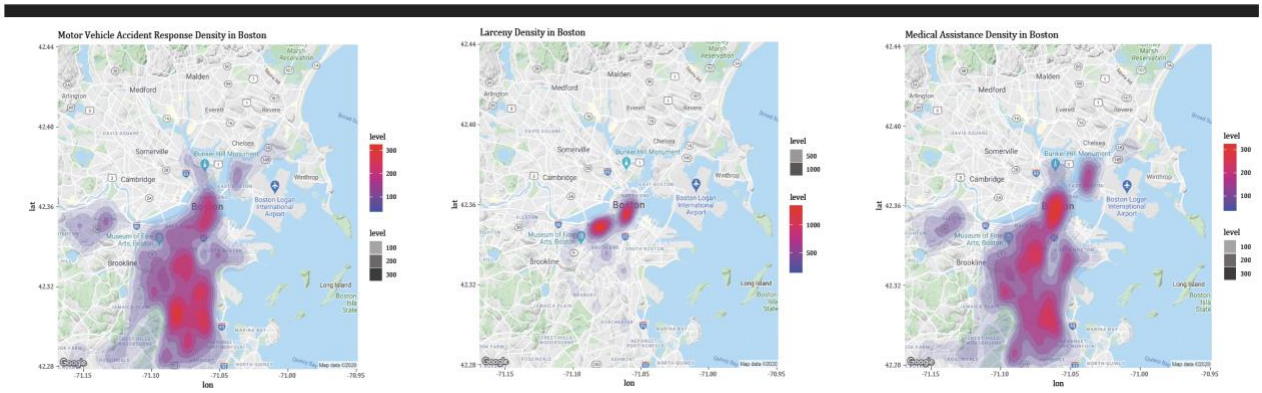


Figure 4 – *Heat Maps of Top Three Offenses Committed in Boston*

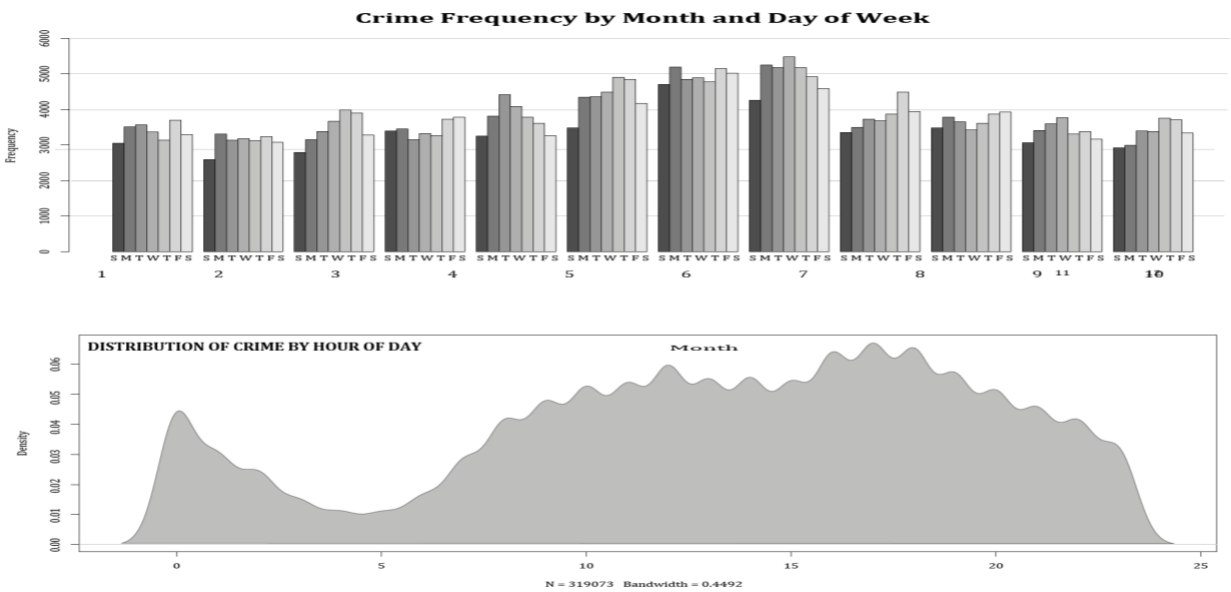


Figure 5 – *Distribution of Crime Frequency by Month, Day of Week, and Hour of Day*

Reflection & Learning Goals

By completing this project, students were able to gain practical knowledge on how to use R to do data cleaning and preparation on a wide range of datasets, identify stories in datasets through various modes of exploration, as well as create rich visual artifacts that aid in communicating data stories. While it is important for data scientists to be able to gain insights from data, it is equally important, if not more important, for them to be able to communicate those findings easily and accurately. All of the skills learned in this course enable students to be able to do just that.

This poster project allowed students to collect and organize data and identify patterns in data via visualization and statistical analysis, develop a plan of actions to implement the business decisions derived from the analyses, as well as demonstrate communication skills regarding the data. Students were expected to select a dataset of their choosing and perform exploratory analysis on it. Throughout the course, students were taught how to effectively visualize data in ways that are not distracting or compromising to the data being presented. The project also helped to sharpen students' communication skills by giving them practice with presenting their analysis to a specified audience. These skills will all be extremely useful in the workforce as data scientists collect and organize data and then need to report on it in a way that is appropriate and relevant for their audience.

4. National Basketball Association Shot Analysis

Project Description

The third project chosen to display the learning objectives of the Applied Data Science Program comes from course IST 707: Data Analytics. This course, headed by Professor Jeremy Bolton, was designed to give students a general overview in data analytics, as well as familiarity with particular real-world applications, challenges involved in applications, and future directions of the field. This course introduced popular data mining methods for extracting knowledge from the data. One of the main focuses of this course was to teach students how to understand data and how to formulate data mining tasks in order to solve problems using the data. Some of the topics that were covered in this course included data

mining, data preparation, concept description, association rule mining, classification, clustering, evaluation, and analysis. The final objective for the class was a project in which students had to use the skills that were taught throughout the course to solve a real data mining problem. The topic chosen for this project was an NBA dataset that was scraped from the NBA's REST API before it was made publicly unavailable. The NBA uses a 6-camera system in each of the NBA arenas, the AutoSTATS Player-Tracking Technology, that can track 2-dimensional player locations 25 times per second. The data story that was followed for this project was that the NBA needed an analysis tool that would improve league performance by predicting things such as whether or not a game would be won depending on a number of factors, and drive fan satisfaction ratings in the tumultuous 2019-2020 season. This dataset relied on situational variables (dribbles, shot location, and distance from hoop) as well as scenario variables (opponent, venue location).

After all data preparation took place, there were 281 unique player ID's that were represented across 1808 unique games. Visualizations were created that better helped to display the descriptive analytics of the dataset, including stacked bar charts and heat maps of shot attempts on the court itself. The heat map pictured below in figure 6 compares the shot attempts of the 2014-15 Houston Rockets and the Washington Wizards. By looking at this heat map, we can see that the Houston Rockets took quite a bit of three-point shots as well as close-range shots that are a lot more likely to go in. It also appears that they stayed away from the mid-range two-pointers. In comparison, the Wizards also focused on close-range shots but were also heavy on the mid-range shots.

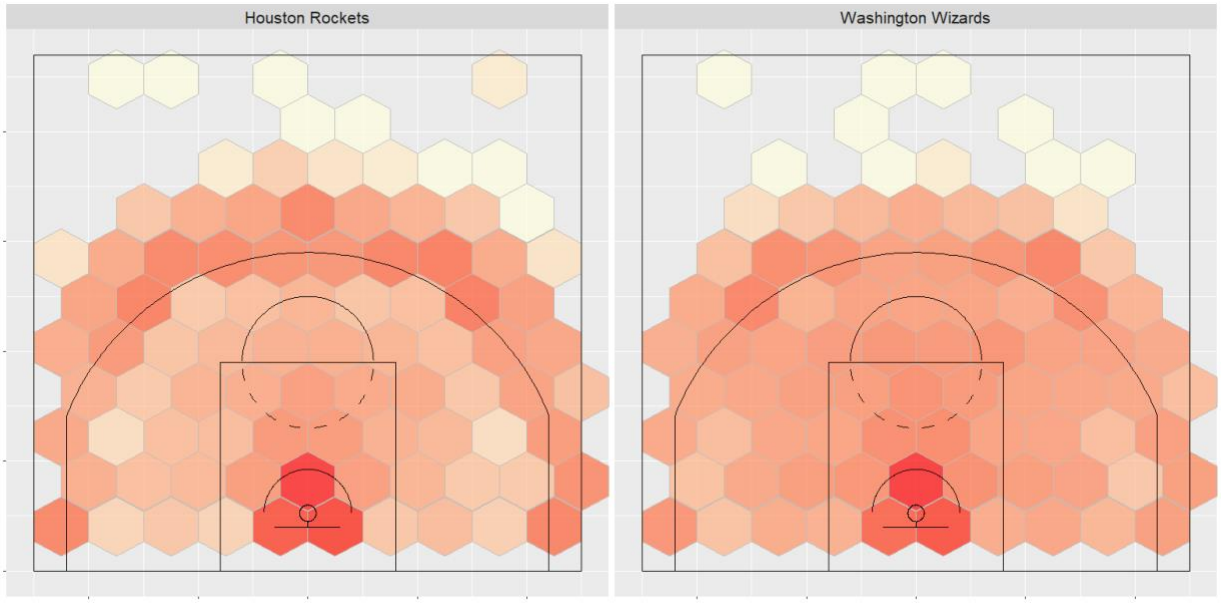
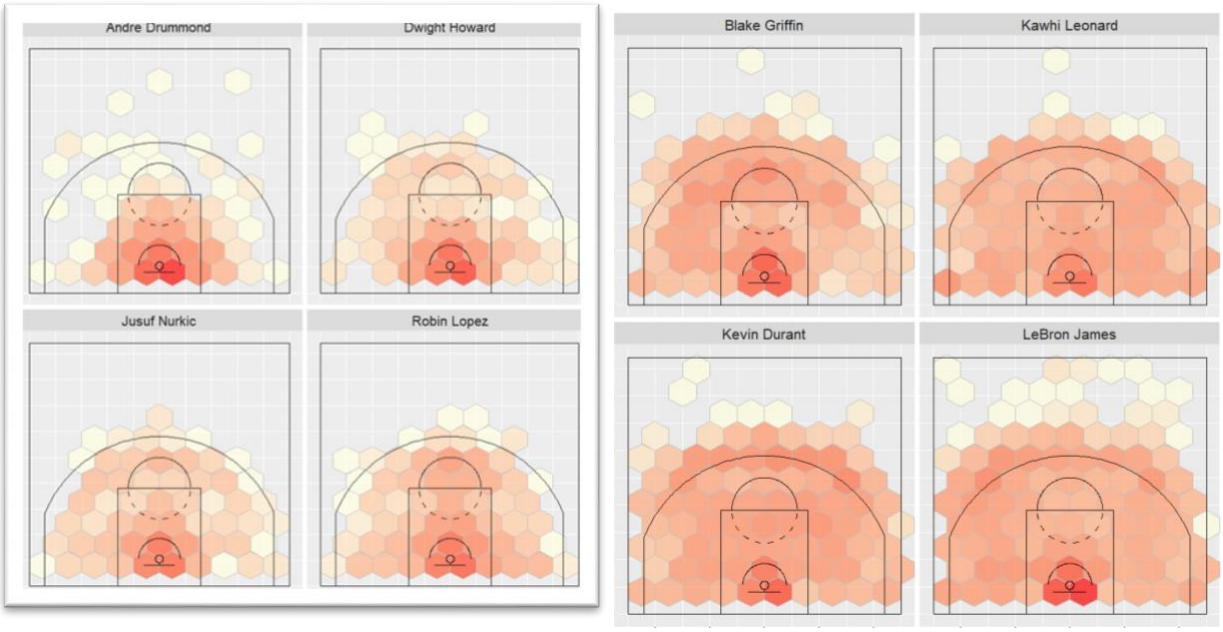


Figure 6 – *Houston Rockets vs. Washington Wizards Heat Map*



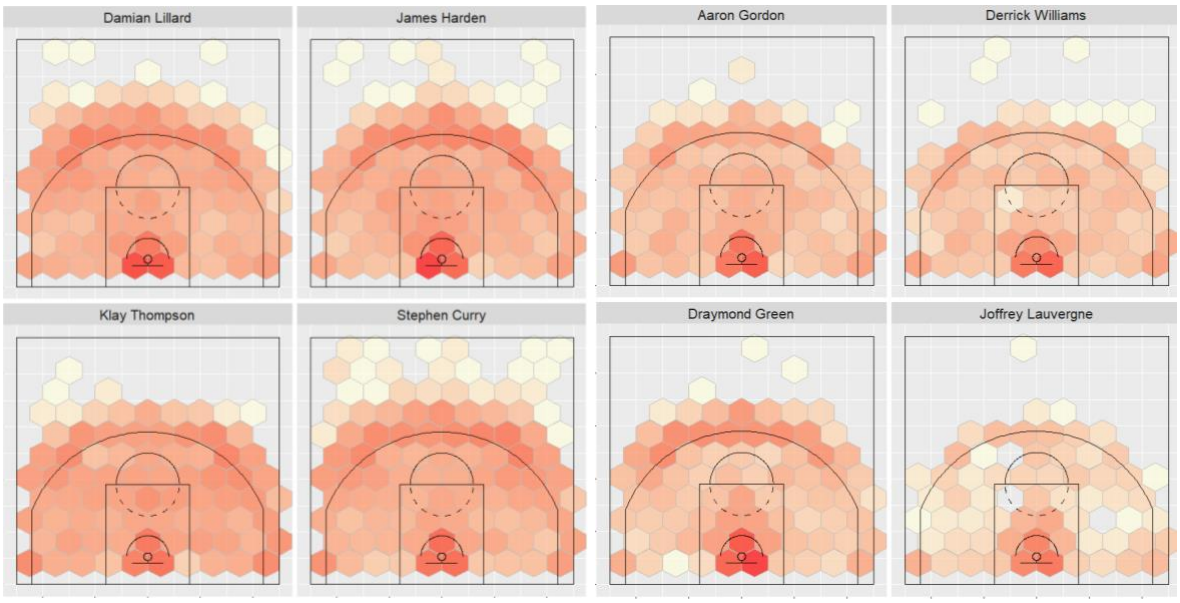


Figure 7 – *Four Clustered Player Profiles*

Other models that were used in the completion of this project included decision trees, association rule mining, and support vector machines. Of all of the models that were used, the SVM ended up being the most effective in predicting whether or not a shot would be made or not with an accuracy of 79.3%, which was substantially higher than the other models. In predicting whether a game would be a win or a loss, decision trees proved to be the most effective with an accuracy of 64.54%. Clustering was used in this project to determine the player types. These player types could be used by coaches and general managers to get a better understanding of how their players are matched as well as just get a better understanding of the types of players that they have on their roster. This could help with decision-making that could ultimately determine the result of a play, a game, or even a season. Lastly, association rule mining was used to develop tens of thousands of rules that gave data scientists an idea of some of the top factors that influence games. Some of the major factors that were most useful in determining whether or not a game would be won included the *home team, location,* and *teamID*. When determining if a shot would be missed or not, the major factors were *seconds remaining, period of the game, the type of shot, time left on the shot clock, shot distance,* and *defender distance*. Lastly, when determining whether or not a shot would be made, the major contributing factors were *type of shot, plater, game clock*, and

defender distance. These were the major attributes that data scientists recommended that the NBA consider when aiming to improve league performance.

Reflection & Learning Goals

This project further contributed to the application of the seven learning objectives as outlined by the Applied Data Science program. Students were required to collect and organize data, identify patterns in the data via visualizations and data mining, develop alternative strategies based on the data used, implement business decisions derived from the analyses, and demonstrate communication skills by reporting on findings with recommendations.

5. San Francisco Airport Data: Pandemic Recovery Survey

Project Description

The fourth and final project chosen to display the learning objectives of the Applied Data Science Program comes from the course MAR 653: Marketing Analytics. This course, headed by Professor Shaam Ramamurthy, was designed to have students focus on developing marketing strategies and resource allocation decisions driven by quantitative analysis. Students who took this course learned about topics such as market segmentation, marker response models, customer profitability, product recommendation systems, churn predictions, media attribution models, and resource allocation. Like the other courses within this portfolio, a final project was required of students to prove that they had retained the information from the course and could apply the skills to real life problems. The dataset that was used for this particular project focused on the San Francisco airport (SFO). San Francisco airport conducts extensive research on customers that come through their facilities on topics such as satisfaction levels, preferences, and demographics throughout the course of the year. By collecting this data, the airport is able to refine their practices as well as strategically team up with partners to ensure that the San Francisco airport experience remains above satisfactory. Some decisions that are often impacted by this data include transit issues, concessions, and any other key areas of concern that arise from the survey.

Some of the major variables that were targeted during this analysis included questions on how concerned travelers were about flying during the current COVID-19 situation, travelers opinions on the steps that SFO was taking to keep travelers safe during this time, and what else travelers feel like SFO could do to help travelers feel like their health is being protected. Some of the other variables that were represented within this data set included demographics on traveler age, income, and country of residence, as well as information on travelers' awareness of services that are offered at SFO. By conducting this research, it was my hope to be able to identify a group of travelers that may need a little more encouragement in believing that it is safe to travel during this COVID-19 period. After identifying this group, we hope to make recommendations to SFO on things that they could do to help sway these customers to travel from SFO airport.

The surveys that SFO conducted were filled out on paper by travelers who passed through the airport. There were 1,086 participants who participated in the survey or 16 different questions. K-Means clustering was the main method that was used for this particular analysis. The elbow plot method was used within Excel to determine how many clusters would work best for the data. It was determined that three clusters would work best. The three customer segmentations can be seen below in Figure 8.

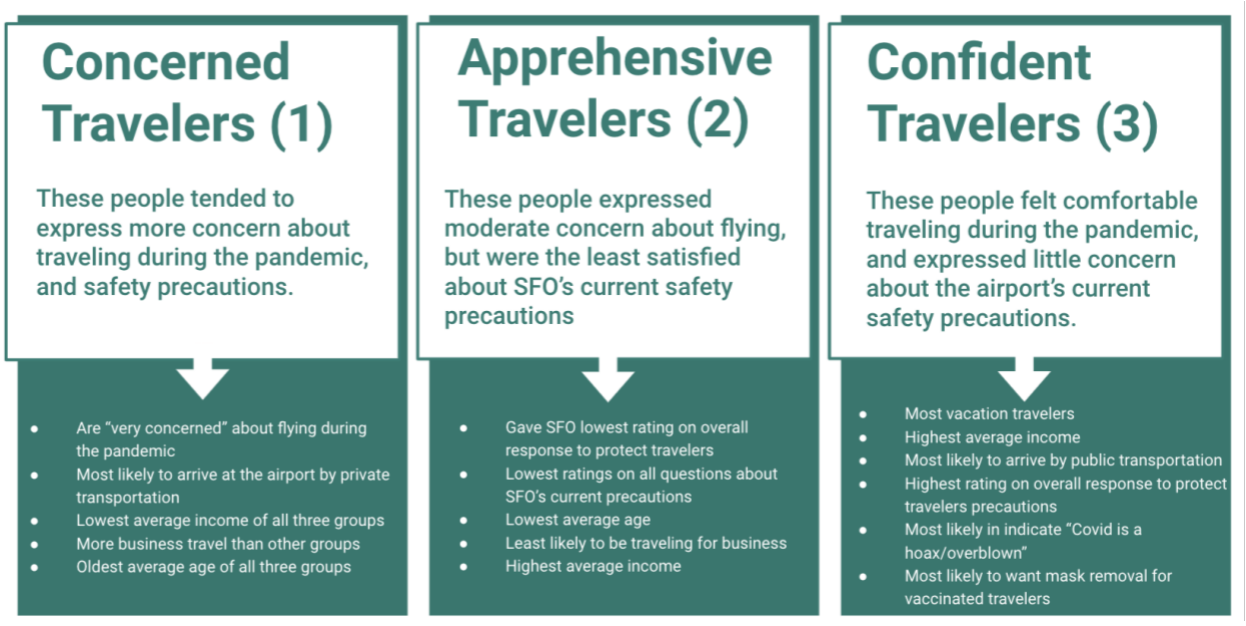


Figure 8 – Customer segmentation

The overall goal of the analysis was to make a recommendation to SFO on which customers to target and on what things to change. The recommendation was for SFO to target the customers who were in the "Concerned Travelers" cluster. These customers were older people who were less likely to use public transportation and who gave low ratings about the current airport policies. Targeting these customers would alleviate tensions associated with resuming post pandemic transportation. Another recommendation was for SFO to take extra precautions to meet the needs of these most vulnerable groups. The recommendations were proposed in an 8 month plan that would first focus on the very concerned travelers and then would transition at the end to focus on the customers who were going to resume traveling post-pandemic. The 8-month plan can be seen below in Figure 9.

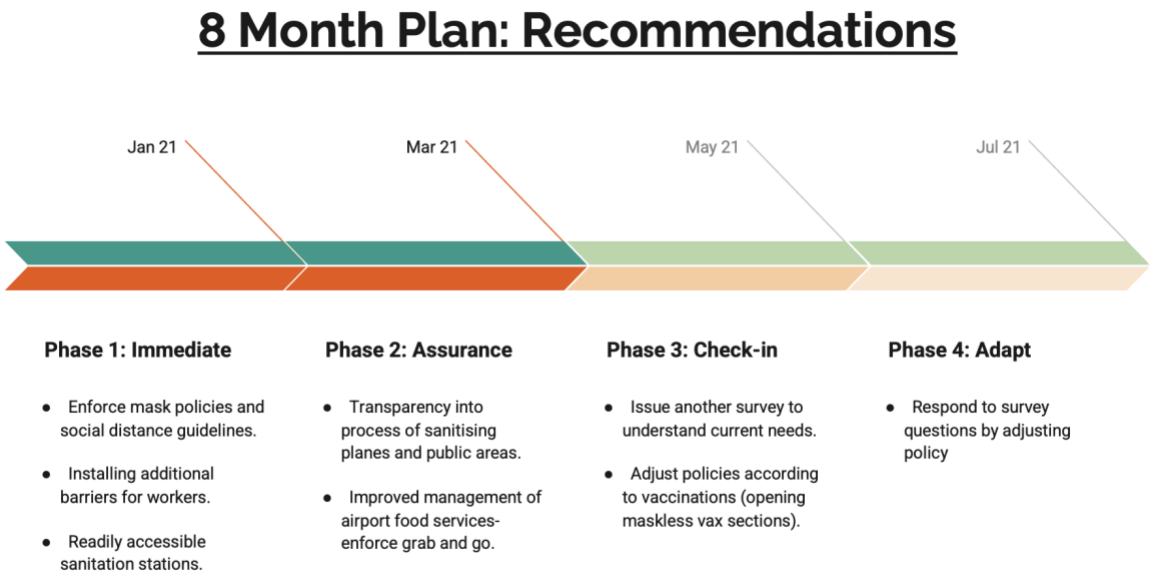


Figure 9 – Recommendations for SFO

Reflection & Learning Goals

By completing this project, students were allowed to collect and organize data of their choosing to target a specific customer set and optimize some marketing goal. They were also taught to create various strategies to lead to the same outcome, improving client lifetime value (CLV). Customer segmentation helps students to learn their audiences in hopes of

providing more value to their customers. This project also allowed students to practice their communication skills by presenting findings through a series of group presentations. This project also allowed students the ability to implement business decisions derived from the analyses that were completed.

6. Conclusion

The assignments selected for this portfolio successfully demonstrated the implementation of the program learning objectives as detailed by the Applied Data Science program course of study. These assignments allowed students the ability to collect and organize data through a number of different avenues, using a number of different applications to program. Many statistical models were used such as association rule mining, clustering, regression, and classification. On top of these models, visualization techniques learned by students throughout these courses were also applied to help students describe their data as well as give them a way to identify patterns that may not be as easily seen by the above analysis. After completing all of these analyses, students worked on their communication skills by report writing, presentations, or poster creation. By creating data stories, students were able to demonstrate their ability to tailor their presentations to specific stakeholders, showing that they understood their audience and what was important to them. Students got to test their hand at understanding ethical boundaries to data such as security and access requirements that would ensure that people only had access to that data that was relevant to them. These projects, successfully completed, allow students the ability to showcase their understanding of the learning objectives.

The Applied Data Science program at Syracuse University is designed to be a professional program of study, with a strong emphasis on with a strong emphasis on the applications of data science to enterprise operations and processes, particularly in the areas of data capture, management, analysis, and communication for decision making. The program outlines seven core learning objectives that students should get out of their courses. The skills learned in this degree program align directly with these objectives, and if completed successfully, equip students with practical skills that they need to translate their academic learnings directly into the workforce. Students leaving this degree program are armed with the knowledge that they need to tackle a wide range of data science problems,

the resources to explain their findings to a variety of audiences, as well as business knowledge to relate it back to.

References

JACKSON, N. J. (2021). *njacks01/AppliedDataSciencePortfolio*. N’DEA JACKSON. https://github.com/njacks01/AppliedDataSciencePortfolio/tree/main/IST%20659

JACKSON, N. J. (2021). *njacks01/AppliedDataSciencePortfolio*. N’DEA JACKSON. https://github.com/njacks01/AppliedDataSciencePortfolio/tree/main/IST%20707

JACKSON, N. J.(2021). *njacks01/AppliedDataSciencePortfolio*. GitHub. https://github.com/njacks01/AppliedDataSciencePortfolio/tree/main/IST%20719

JACKSON, N. J. (2021). *njacks01/AppliedDataSciencePortfolio*. GitHub. https://github.com/njacks01/AppliedDataSciencePortfolio/tree/main/MAR%20653