APPLIED DATA SCIENCE PORTFOLIO

Syracuse University

Master of Science in Applied Data Science

Spring 2024

ABSTRACT

The Applied Data Science program at Syracuse University's School of Information Studies effectively prepares students with a comprehensive skill set, covering essential areas such as data collection, analysis, and communication. Through hands-on projects and coursework, students gain proficiency in tools like Python, R, SQL, Power BI, and Tableau while emphasizing ethical data practices and effective stakeholder communication, empowering graduates to navigate and excel in data-driven environments.

Amanda Norwood
SU ID 503480562

TABLE OF CONTENTS

INTRODUCTION

Data science is a combination of statistics, programming, and advanced analytics to derive actionable insights that's within an organizations data (IBM). The goal is to assist in making informed decisions. The Applied Data Science program at Syracuse University's School of Information Studies equips students with the skills to collect, store, and analyze data using relevant technology domains, generate and effectively communicate actionable insights through visualizations or reports, and utilize predictive models, all while upholding ethical standards in the model development (Syracuse, 2020).

LEARNING OUTCOMES

Syracuse University's Applied Data Science (Syracuse, 2020) program has six learning outcomes that will be outlined by this portfolio:

1. Collect, store, and access data by identifying and leveraging applicable technologies.
2. Create actionable insight across a range of contexts (e.g. societal, business, political), using data and the full data science life cycle.
3. Apply visualization and predictive models to help generate actionable insight.
4. Use programming languages such as R and Python to support the generation of actionable insight.
5. Communicate insights gained via visualization and analytics to a broad range of audiences (including project sponsors and technical team leads).
6. Apply ethics in the development, use and evaluation of data and predictive models (e.g., fairness, bias, transparency, privacy).

PROJECTS

These five courses that have enabled my development and demonstration to achieve the program's goals by utilizing Python, R, Azure Data Studio, Power BI, and Tableau:

- IST 659 Database Management
- IST 652 Scripting for Data Analysis
- IST 707 Applied Machine Learning
- IST 718 Big Data Analytics
- IST 737 Visual Analytics

IST 659 – DATABASE MANAGEMENT

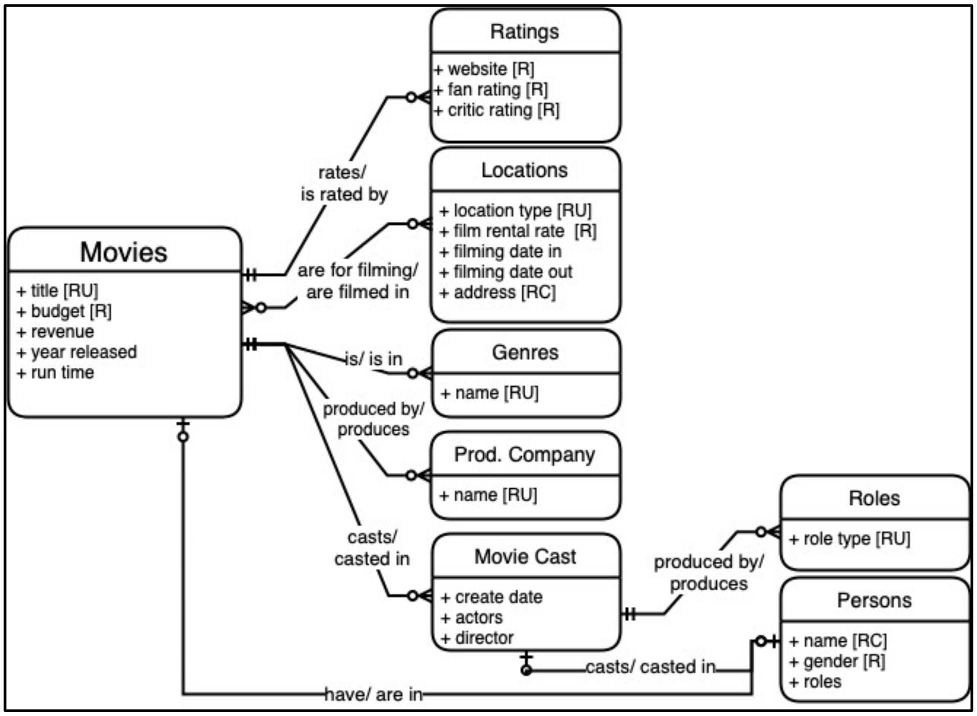
PROJECT GOAL

The project required is the creation of the conceptual data model diagram, logical model diagram, and create and implement the entire database. This included implementing the physical database creation through up and down scripts, creating tables, inserting data, designing views, and executing stored procedures. The project also required us to answer data questions to validate the database functionality. In addition, we were tasked to develop and implement the application.

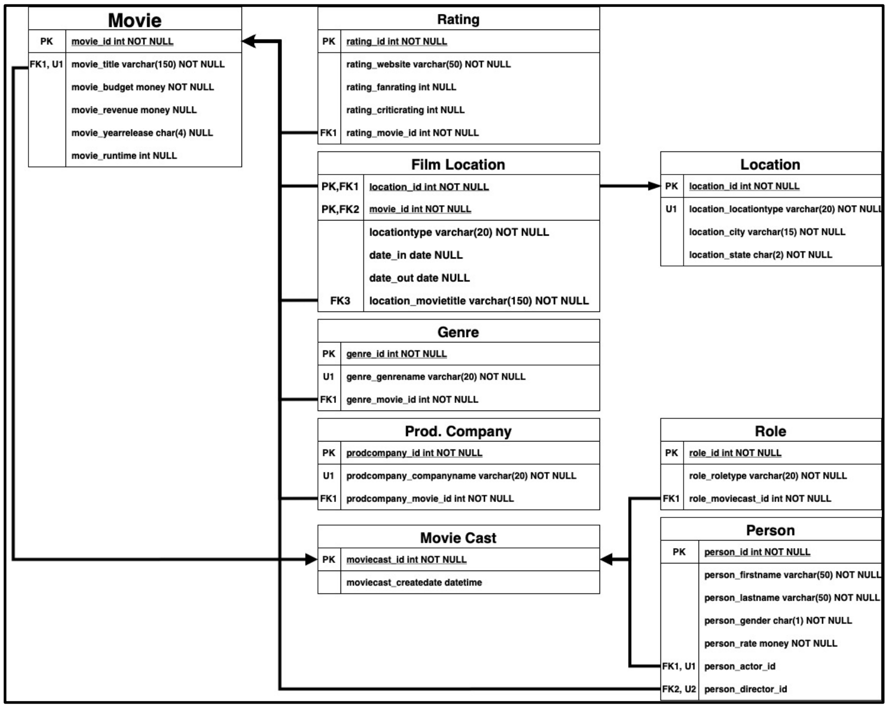
PROJECT DESCRIPTION

Our team created a scenario for a production management company that wants to manage movie production information and track post release performances. The idea is to organize their data, lessen clerical errors, and maintain database integrity and find a solution for their version control. At the time, the production company has been tracking their information in spreadsheets. The mission was to build a database to store the information and provide functionality and ensure data integrity.

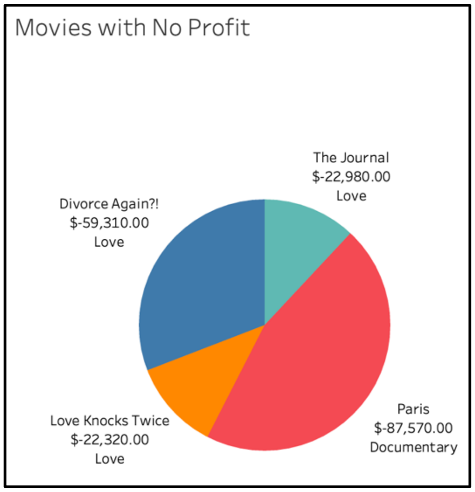
The database included person's name, movie genre, movie title, location, role, budget, rating, and production company. Below is how we planned out the initial build and relationships of the database:

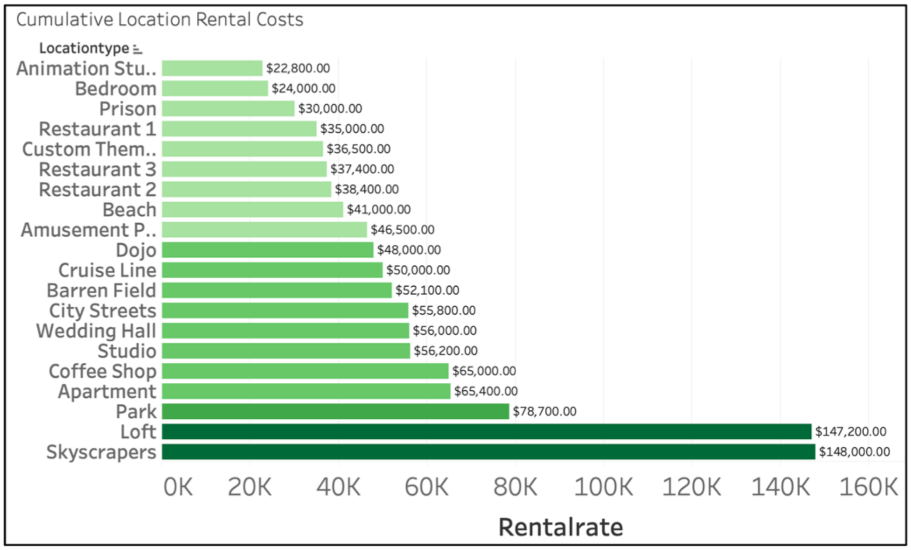


Then we moved on to the logical model to finalize the primary & foreign keys and the properties of specific fields.

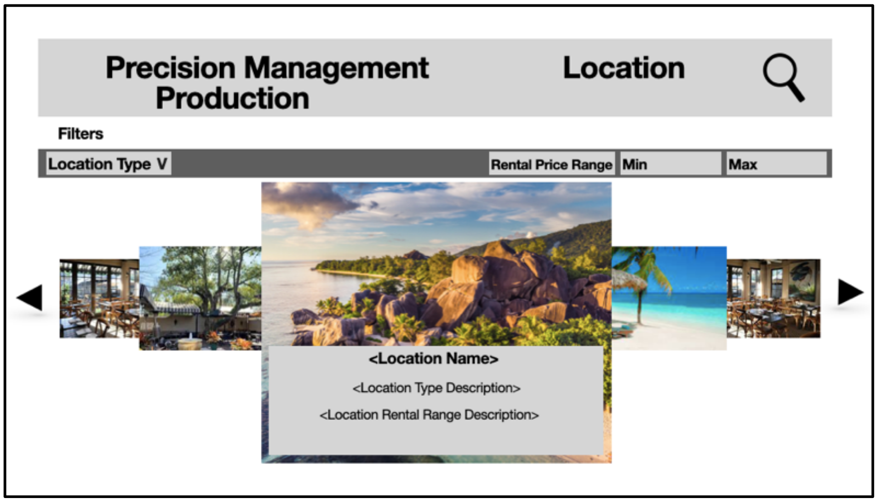

Movie
PK movie_id int NOT NULL
FK1, U1 movie_title varchar(150) NOT NULL
movie_budget money NOT NULL
movie_revenue money NULL
movie_yearrelease char(4) NULL
movie_runtime int NULL
Rating
PK rating_id int NOT NULL
rating_website varchar(50) NOT NULL
rating_fanrating int NULL
rating_criticrating int NULL
FK1 rating_movie_id int NOT NULL
Film Location
PK,FK1 location_id int NOT NULL
PK,FK2 movie_id int NOT NULL
locationtype varchar(20) NOT NULL
date_in date NULL
date_out date NULL
FK3 location_movietitle varchar(150) NOT NULL
Location
PK location_id int NOT NULL
U1 location_locationtype varchar(20) NOT NULL
location_city varchar(15) NOT NULL
location_state char(2) NOT NULL
Genre
PK genre_id int NOT NULL
U1 genre_genrename varchar(20) NOT NULL
FK1 genre_movie_id int NOT NULL
Prod. Company
PK prodcompany_id int NOT NULL
U1 prodcompany_companyname varchar(20) NOT NULL
FK1 prodcompany_movie_id int NOT NULL
Role
PK role_id int NOT NULL
role_roletype varchar(20) NOT NULL
FK1 role_moviecast_id int NOT NULL
Movie Cast
PK moviecast_id int NOT NULL
moviecast_createdate datetime
Person
PK person_id int NOT NULL
person_firstname varchar(50) NOT NULL
person_lastname varchar(50) NOT NULL
person_gender char(1) NOT NULL
person_rate money NOT NULL
FK1, U1 person_actor_id
FK2, U2 person_director_id

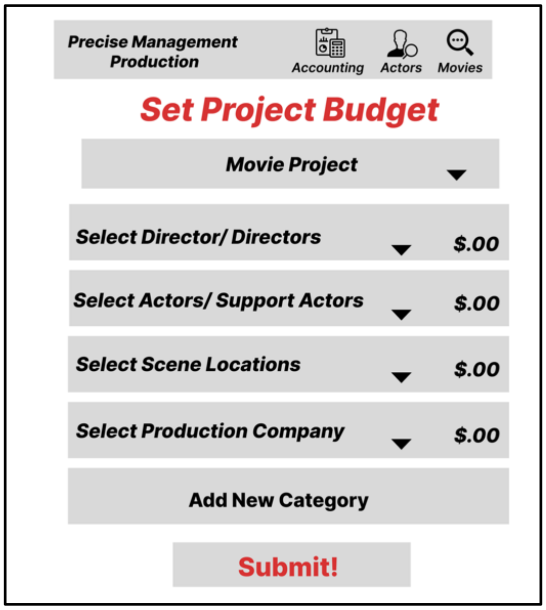
In the end, we were able to help the business answer and visualize their questions through business intelligence tools such as Tableau:





The company aims to leverage the optimized database by integrating it into front-end development processes. The objective is to have an optimized database structure and performance to facilitate efficient data retrieval and manipulation. The screenshots below provide an example of how our data would be utilized in front-end development.


Precision Management Production
Location
Filters
Location Type V
Rental Price Range
Min
Max
<Location Name>
<Location Type Description>
<Location Rental Range Description>



LEARNING OUTCOMES

I use SQL daily to write queries and analyze data. However, this course provided a new perspective and appreciation for the role of data engineers in creating and maintaining databases for users. Understanding the structure and planning of the database is crucial for efficient query writing and execution times.

Throughout the course, I developed a deeper understanding of the underlying relationships between entities and constraints, recognizing their importance in maintaining to database integrity. I also gained proficiency in developing, writing, and executing procedures, understanding why they are preferred over update statements due to their reduced potential for human errors.

Furthermore, I also learned the significance of indexing, data denormalization, and optimized query design. Indexing enhances data retrieval speed, while denormalization reduced redundant information, leading to faster query times.

While I was somewhat familiar with common table expressions before the course, I now have a stronger understanding of their use cases and how their application in my daily work.

Reflecting on the project, I realized the integral role of conceptual and logical models in database creations. Spending ample time on planning out the tables and relationships facilitates smoother database implementation.

Overall, this project exemplified the learning outcome of effectively collecting, storing, and accessing data by identifying and leveraging appropriate technologies such as SQL and Tableau (Syracuse).

IST 652 – DATA ANALYSIS

PROJECT GOAL

The goal is to showcase writing scripts from various types of data such as structured and semi structured, define and find patterns within the data, and prepare and transform the data to produce insightful data, patterns, and analyzation. In addition, the main goal is wrangling data to analyze and answer applicable questions.

PROJECT DESCRIPTION

Our project centered around public NFL data from 1997 and onwards, aiming to provide stakeholders, including NFL personnel, with valuable analytics for making informed decisions. The database comprised of comprehensive statistics from each NFL game since 1997, covering passing and rushing yards, quarterback performance, weather conditions, and stadium types.

In addition to the game statistics, we delved into quarterback specific data. We also integrated information on betting odds and game spread to investigate impacts or advantages.

Overall, our project aimed to leverage NFL data to generate actionable insights, aiding organizations in optimizing their strategies and decision-making processes.

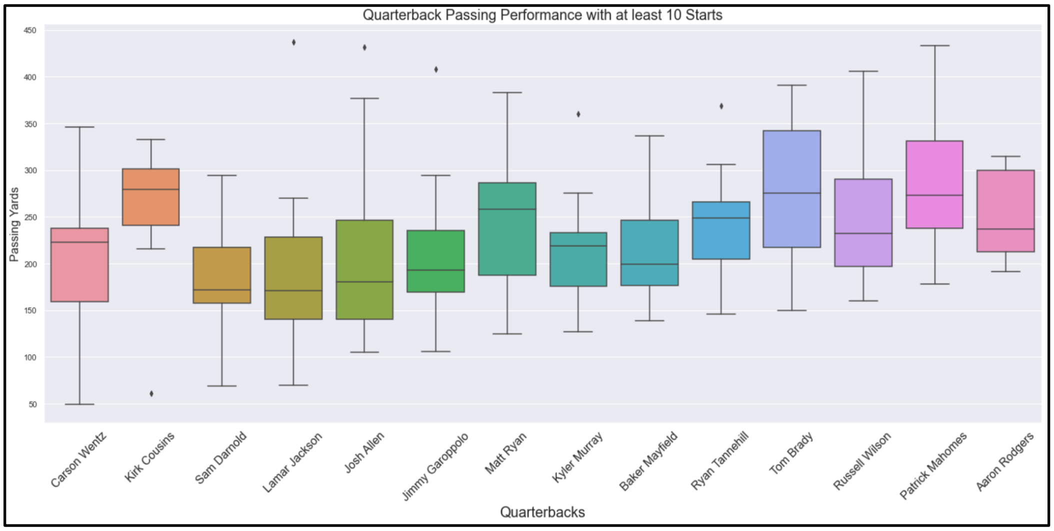
LEARNING OUTCOMES

Prior to taking Scripting for Data Analytics, my experience with Python was limited to basic data cleaning and calculations. However, by the end of the course, I gained valuable insights, particularly in heavy data cleaning, which I found both daunting and rewarding.

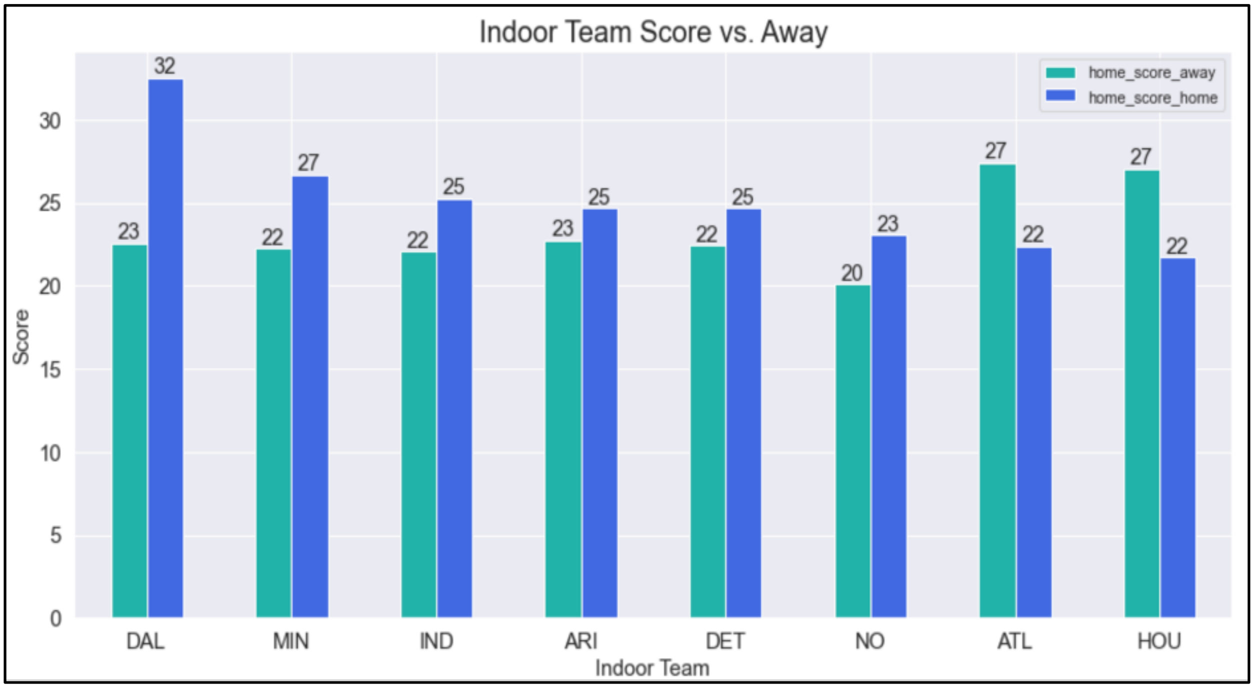
Our team faced the task of joining and merging six datasets, requiring us to think strategically about our approach based on our research questions. This forced us to adopt a holistic and analytical mindset, emphasizing the importance of thorough data cleaning and merging and consistent naming convention across the datasets. These steps were proven to be pivotal in the project's success.

Now from a business and analytical perspective, the data challenged us to think critically about what insights customer truly need to know, beyond what the data alone can provide. Understanding the context in which the data will be used became paramount. Ultimately, the most valuable lesson was extracting actionable insights to facilitate impactful decision-making.

Additionally, I learned that individual statistics alone are not suffice for storytelling. It's crucial to connect the dots to address specific business questions. For instance, merely stating that quarterbacks in dome stadiums score more points overlooks the broader context. It's essential to consider how these quarterbacks perform in outdoor stadiums as well, thereby providing a more comprehensive analysis.



From a coding perspective, this class challenged my ability to showcase data using the most effective visuals. Not all visuals are created equal, and it is true that some visuals display the data better and highlight key insights faster than others. It's best to move beyond simple statistics and consider which visuals best serve the intended purpose. For example, the boxplot displaying passing yards by quarterback provides a comprehensive view, showing the median, high, low, 75 percentile and 25 percentiles. The statistics show a full view on quarterback performance and compare statistics if a team is on the market for a quarterback. Notably, it's evident than Aaron Rodgers and Kick Cousins exhibit the tightest grouping, indicating consistent performance.



In addition to the analysis, the visual above shows that teams that have played 98% of their home games indoors were analyzed. Teams that recently moved into an indoor stadium were not included in this analysis. There are 8 teams with their away and home performance analyzed. 75% of teams with home indoor stadiums hold a strong advantage over their opponent. The Dallas Cowboys perform well over average when at home with a high average of 32 points. However, their away game performance is average compared to other teams.

IST 718 – BIG DATA ANALYTICS

PROJECT GOAL

The goal is to be able to obtain, scrub, explore, model, and interpret the data in a full cycle to analyze real world problems in business.

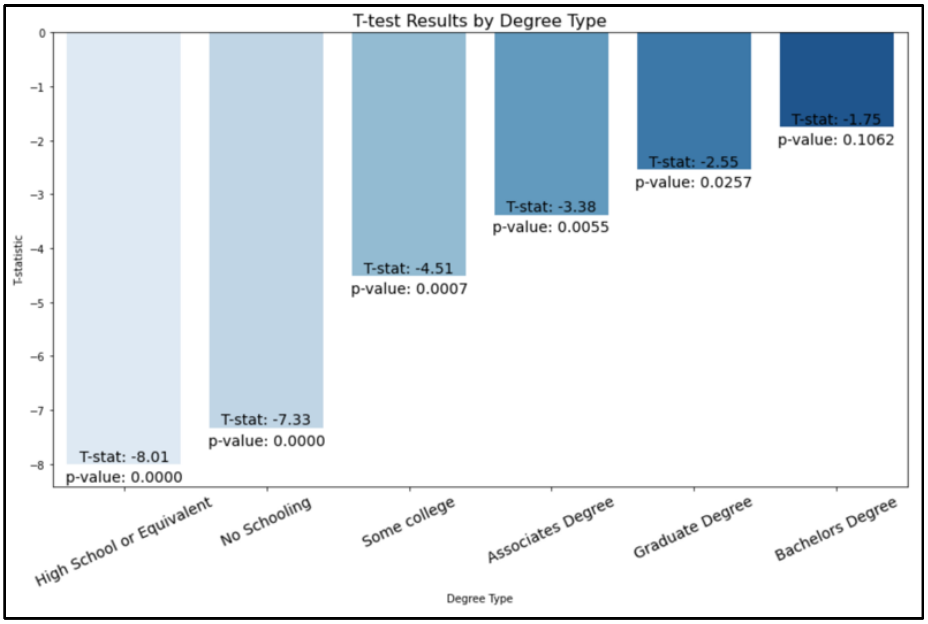
PROJECT DESCRIPTION

The city of Los Angeles, California is looking to address their housing crisis and generate a method to analyze housing affordability based on gender, degree, and profession. The goal is to analyze housing affordability and crime by zip code to assist potential migrants to research pay based on their profession.

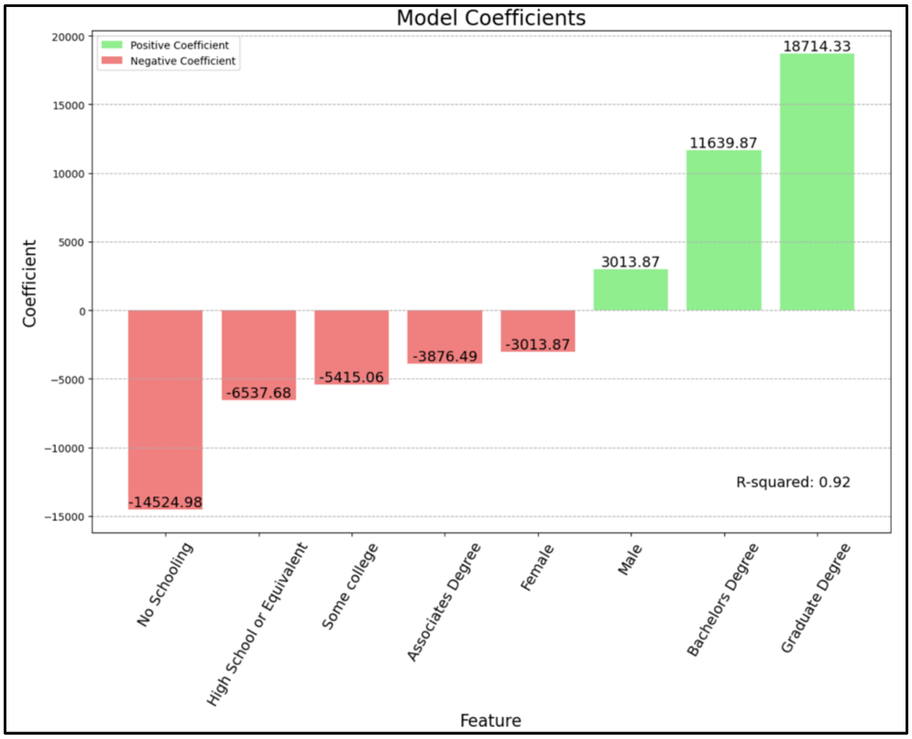
The goal of the analysis is to provide a method for potential L.A. migrants with information regarding the housing affordability and salary based on their gender, profession, or education for them to make informed decisions on their move. We also want to predict zip codes where people can live based on their demographic data.

LEARNING OUTCOMES

This course allowed me to learn different types of data tests, machine learning algorithms, and methods of analysis. By the time I took this course, my Python coding and visualization skills had improved vastly.



Among the tools I learned, the t-test stands out as a valuable method for discerning differences between means and quickly identifying variations in data. For instance, when applied to the dataset, any degree type with a p-value less than 0.05 indicates a significant difference in pay. Every degree type except for bachelor's degree yielded a p-value less than 0.05. This suggests that while there are pay discrepancies between genders for most degree types, bachelor's degree holders do not experience such differences.

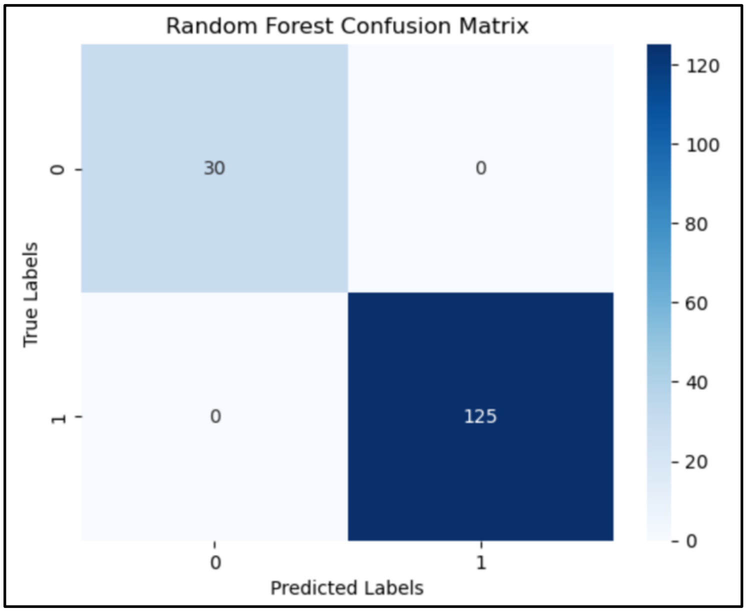


In addition to the t-test, I explored linear regression, a fundamental technique in data science. Through this method, I calculated the model coefficients based on gender, degree type and pay.

Moreover, the visualization below vividly illustrates disparities between gender and degree types. On average, females earn approximately $3000 less than males. Individuals with graduate degrees tend to earn an average of $18.7 more! These insights underscore the significance of education level and gender in determining earning potential in Los Angeles.

In one of my other courses, I had the opportunity to delve into machine learning for the first time and improve it within this course. I utilized both Naïve Bayes and Random Forest algorithms to train my data. Naïve Bayes, known for its probability-based approach, was employed to predict the compatibility of professions based on housing affordability. Impressively, it achieved an accuracy rate of 94.84%.

One of the key takeaways from this project, applicable to real-world scenarios, is the challenge posed by incomplete data. It's often difficult to paint a comprehensive picture when not all the necessary data is available for analysis. However, I've come to understand the importance of finding ways to work around these limitations and informing stakeholders about the constraints within the analysis.



Throughout this project, I've learned that despite our best efforts, incomplete data can potentially skew the narrative. Yet, ultimately, the data speaks for itself. It's crucial for interpreters to recognize and address outliers and restrictions within the data, ensuring that the insights drawn accurately reflect the underlying reality. This experience has underscored the importance of transparency and critical thinking in data analysis, where acknowledging limitations is just as essential as drawing conclusions.

IST 707 – APPLIED MACHINE LEARNING

PROJECT GOAL

The goal of the project is to use the main skills taught in the course to solve real life problems. The learning objectives for the project are apply data mining concepts, algorithms, and evaluation methods to real world problems; employ data storytelling and dive into the data, find useful patterns, and articulate the patterns, how they were found, and why they are valuable and trustworthy.

PROJECT DESCRIPTION

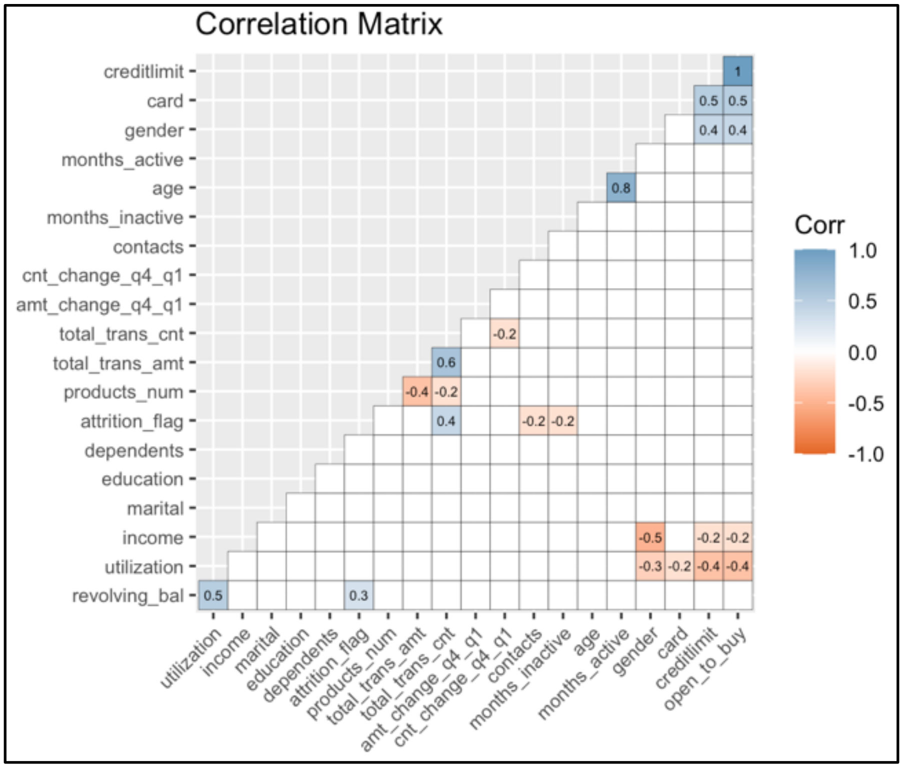
Our data analytics team was recently approached by a bank seeking assistance in understanding why customers were leaving their services. The bank generously provided us with detailed demographic data from their customers.

In this study, we utilize the demographic data provided by the bank to delve into the reasons behind customer attrition. Through a series of analytical models, we meticulously evaluate the data to uncover key insights into customer behavior, particularly identifying indicators signaling their likelihood to discontinue using the bank's services.

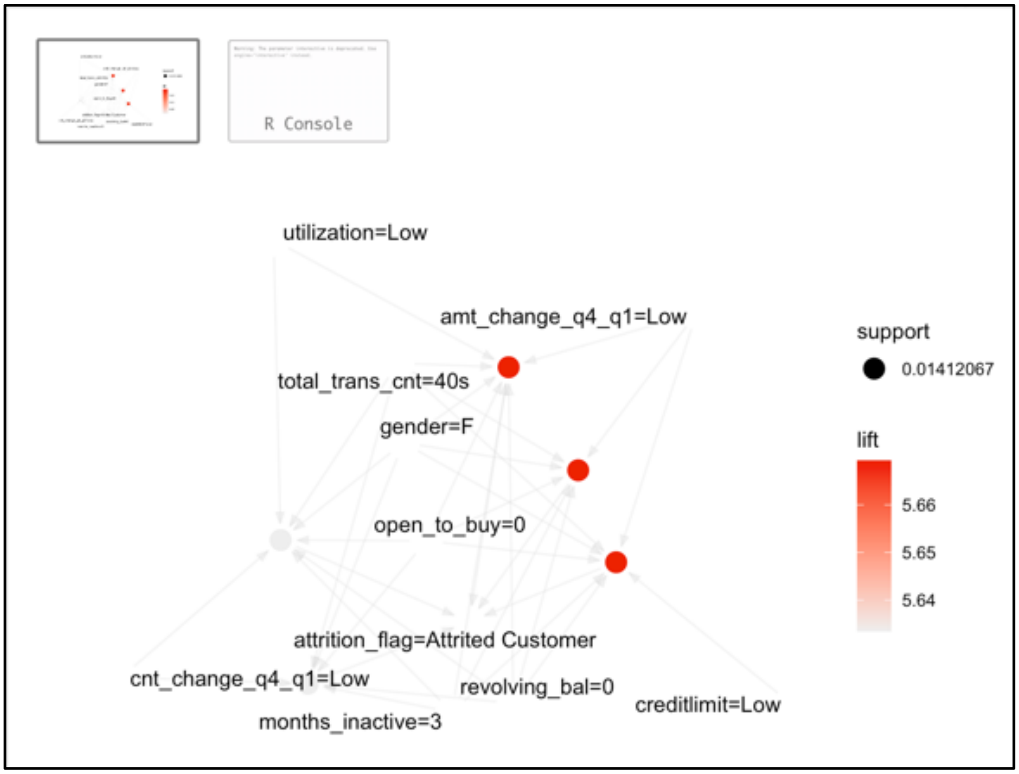
These findings can be invaluable to the bank's marketing team, enabling them to reach out to customers with personalized recommendations and refine their products and services to better meet customer needs. By leveraging these insights, the bank can proactively address customer churn and enhance overall customer satisfaction.

LEARNING OUTCOMES

This course proved to be the most challenging for me, primarily because my prior experience was limited to IST 718 with some exposure to machine learning. This course significantly heightened the importance of machine learning in our project. We used multiple machine learning algorithms and gained insights into how overfitting can compromise the integrity of the system. Overfitting is when the programming learns the specifics of the data instead of the general trends of the data.

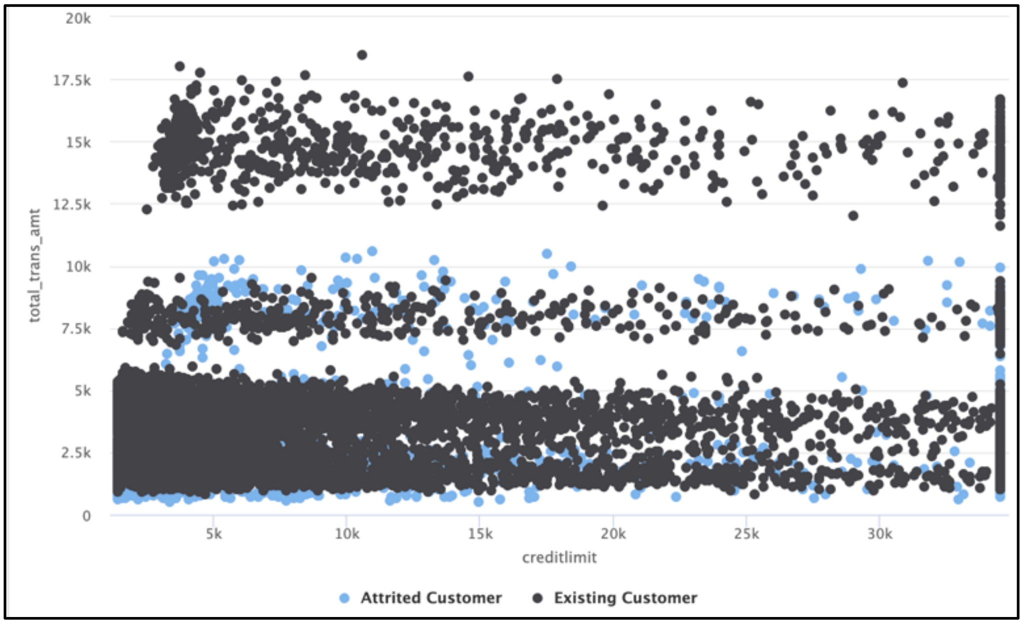


While previous courses predominantly used Python, I opted for R in this project as a challenge to learn R. This decision allowed me to compare the two programming languages. While certain tasks and coding language may be easier to accomplish in Python, I found R's visualizations to be far more appealing and interactive than Python.



Throughout the course, I learned effective methods for visualizing data and discovered how to create interactive plots, adding depth to our analyses. From a machine learning perspective, I found exploring the concepts like rules, lift, and support particularly intriguing because it can be manipulated slightly to spit out different results. Understanding how each setting impacts data results was eye-opening, and I enjoyed experimenting with different configurations to shape the narrative of the data.

It later became evident that meticulous scrutiny is essential to ensure the reliability of both the data and results. When presenting findings to stakeholders, clear definitions are paramount. By defining these terms, stakeholders can grasp the truth behind the data which facilitates informed decision making.



IST 737 – VISUAL ANALYTICS

PROJECT GOAL

The goal is to create an interactive dashboard in Tableau to answer business questions that can clearly be answered and navigated through the dashboards. The visualization focuses on a specific message with a good flow from high level into details.

PROJECT DESCRIPTION

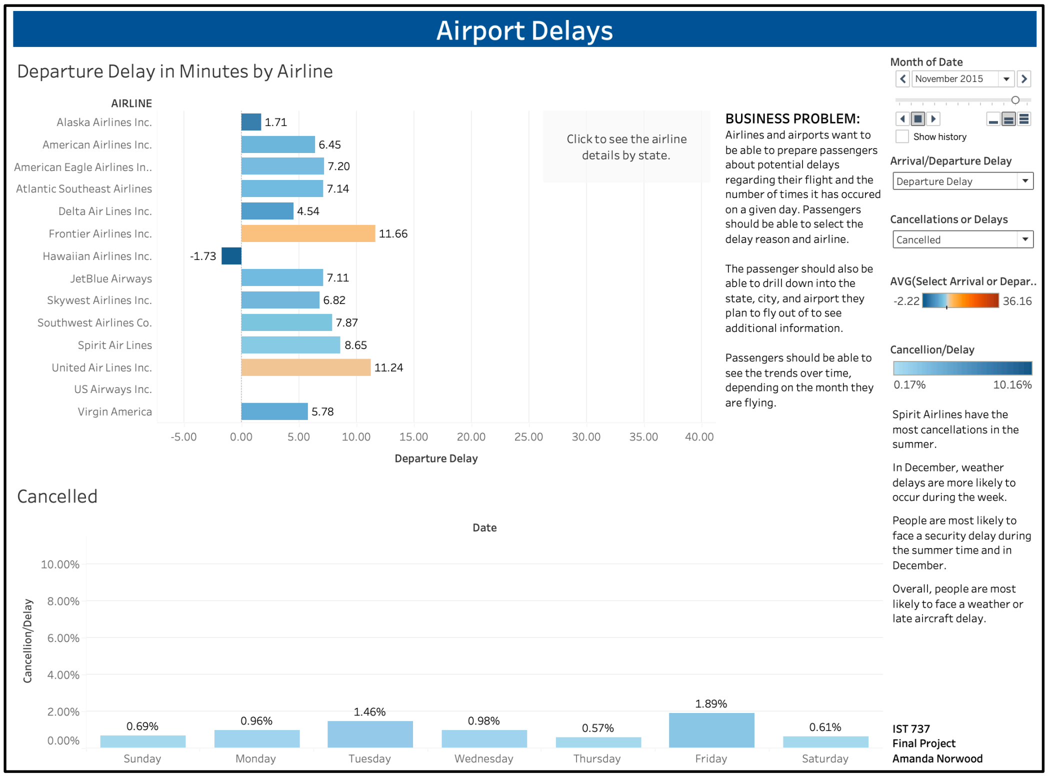
Airports and airlines want to track and be able to provide passengers with information about potential delays or cancellations on a given day based on the airline, airport, day, or state. For functionality purposes, users should be able to see data at a high level and be able to drill down to see the details of the data they select.

LEARNING OUTCOMES

I have experience with Power BI and the hands-on experience with Tableau through IST 737 is valuable. Less is more in terms of visualization. Most users do not enjoy clutter and that translates to dashboard pages. Dashboards should be clear and concise without overwhelming a user. And they should have the ability to drill down on the data if they more granulate information.

Again, as I have learned from previous course, a single data point is not enough to tell a story or trend. It is the holistic assessment of data points that tie together the narrative we are trying to tell.

For example, with the visualization of airport delay's below, we can see that for the month of November 2025, Frontier and United Airlines had an average departure delay of roughly 11 minutes. That is great to see as an airline, airport, or passenger. But how does it impact them if they are flying out of different airports? That is when the functionality to drill down is essential. A user can click on the airline and see more details.



BUSINESS PROBLEM:
Airlines and airports want to be able to prepare passengers about potential delays regarding their flight and the number of times it has occured on a given day. Passengers should be able to select the delay reason and airline.

The passenger should also be able to drill down into the state, city, and airport they plan to fly out of to see additional information.

Passengers should be able to see the trends over time, depending on the month they are flying.

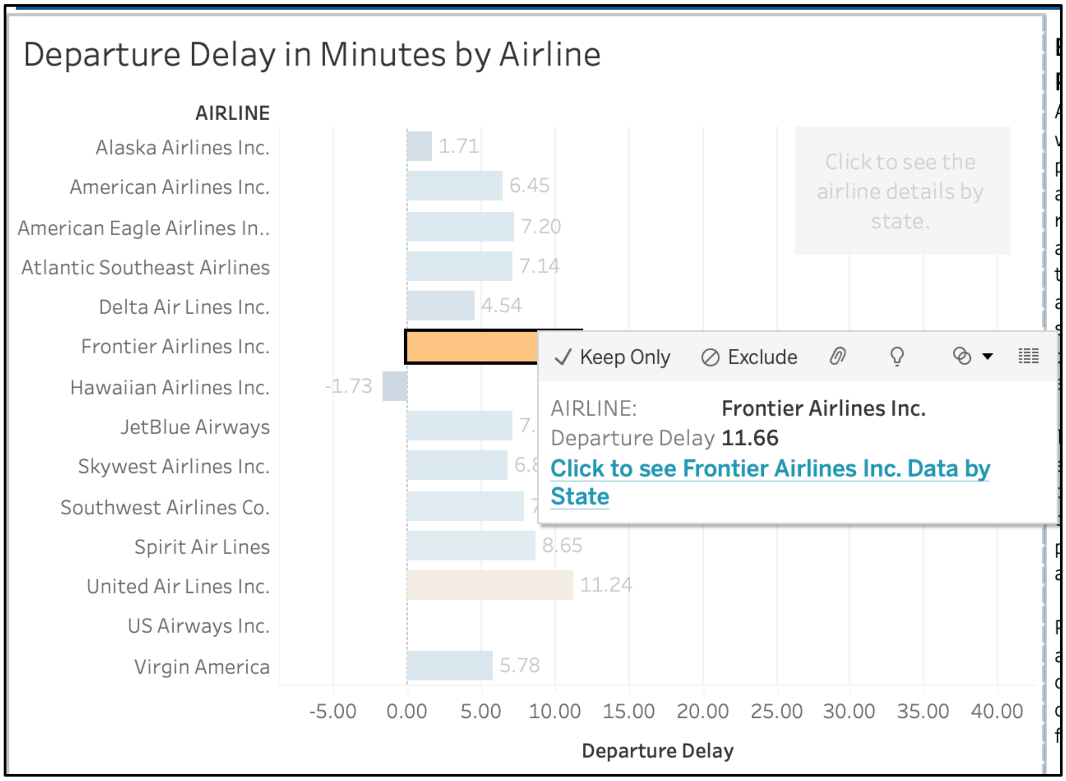
Spirit Airlines have the most cancellations in the summer.

In December, weather delays are more likely to occur during the week.

People are most likely to face a security delay during the summer time and in December.

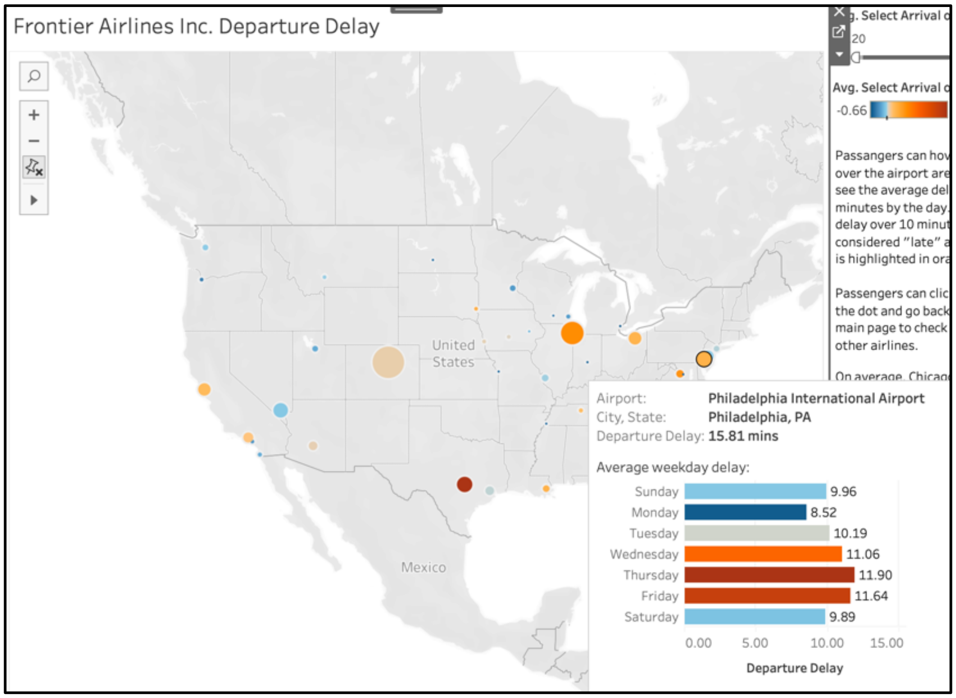
Overall, people are most likely to face a weather or late aircraft delay.

IST 737
Final Project
Amanda Norwood



AIRLINE: Frontier Airlines Inc.
Departure Delay 11.66
Click to see Frontier Airlines Inc. Data by State

They can further drill down to their state and select their airport. And use the tooltips to see the details without being overwhelming by the data. In this scenario, a person can look at Philadelphia's airport and see that Thursday and Fridays are the busiest days to travel and Frontier airlines is likely to be delayed by an average of 11 minutes. This means that if a user is looking at a connecting flight through there, they may want to add extra buffer time.



Frontier Airlines Inc. Departure Delay

g. Select Arrival d
20

Avg. Select Arrival d
-0.66

Passangers can hov
over the airport are
see the average del
minutes by the day.
delay over 10 minut
considered "late" a
is highlighted in ora

Passangers can clic
the dot and go back
main page to check
other airlines.

On average, Chicag

Airport: Philadelphia International Airport
City, State: Philadelphia, PA
Departure Delay: 15.81 mins

Average weekday delay:

Also, color theory and consistency are important through a single dashboard. Colors must mean the same thing across the report or users will be confused. A user should be able to look at the report and say that blue clearly means good while red means not good.

Legends and slicers are also important because it provides information and versatility. The user can slice and dice the data to their choosing.

Overall, the dashboard should follow color theory, have a flow, and tie in all the data points together to paint a whole picture.

CONCLUSION

Based on the outlined learning outcomes and the detailed project descriptions, it's evident that the Applied Data Science program at Syracuse University's School of Information Studies successfully equipped students with a comprehensive and well-rounded skill set to tackle real-world data challenges.

Throughout the program, I've gained proficiency in various essential areas of data science, including collection, storage, and access by utilizing modern technologies such as Python, R, SQL, Power BI, and Tableau. The emphasis on the OSEMN (Hotz, 2023) method has been ingrained in me as a systematic approach in data analysis, ensuring accuracy and reliability in the results that I am able to produce.

The projects I've had the joy of working on serve as tangible evidence of my ability to apply these skills across diverse technology domains. Whether it was database management, data analysis, big data analytics, machine learning, or visual analytics, each project presented unique challenges that I was able to tackle effectively. By addressing business issues, I've learned the importance of understanding stakeholders needs and translating the data insights into actionable recommendations.

Moreover, these projects have highlighted the interdisciplinary nature of data science, requiring not only technical expertise but also strong communication skills to convey findings to audiences with varying knowledge levels.

Additionally, the emphasis on ethics in data science has underscored the programs commitment to responsible and transparent data practices. Nowadays, data can be skewed to fit a specific narrative and it's important as the person analyzing the data to communicate the data's limitations up front and how it was derived.

From a workplace standpoint since I joined the program, I have:

- **Refined SQL Writing Skills**: through IST 772 homed in my ability to write SQL queries adept at handling large datasets efficiently, including queries for performance end ensuring data integrity and accuracy.
- **Implemented Automation Python Scripts**: through IST 652, I have learned how to leverage Python scripting to automate the generation of reports directly from SQL queries to export and format to Excel or CSV, streamlining the reporting process and reducing manual efforts.
- **Automated Analysis with Python Scripts**: through various courses at Syracuse, I have learned to develop Python scripts to automate repetitive analysis tasks typically performed and calculated in Excel. This automation has saved time and ensures consistently and accuracy.
- **Python Integration with Power BI:** through IST 718, I have extended my skills to integrate Python scripts with Power BI for enhanced analysis capabilities. By importing Python-generated analysis results into Power BI, I've facilitated more comprehensive and dynamic data visualization and exploration.
- **Enhanced Power BI Semantic Models:** through IST 772, I have contributed to improving and redesigning Power BI semantic models to optimize data relationships. The optimization resulted in faster retrieval times, minimized cache time, and improved overall query performance.
- **Built Loan Default Machine Learning Model**: through IST 707, I have developed a machine learning model for predicting loan defaults, utilizing algorithms and techniques to analyze historical data and identify patterns indicative of potential default risks.
- **Data Storytelling**: through IST 737, I learned the significance of integrating color theory into dashboards for effective data visualization. Most importantly, I have learned that a single data point is not sufficient to convey the complete picture; it's the aggregation of data points that paints the full picture. The flow of the data storytelling also contributes to capturing the stakeholder's attention by guiding them through the narrative and fostering a deeper understand of the insights being presented.
- **Stakeholder Communication**: throughout all the I have grasped the paramount importance of transparency in communicating about data, including its limitations, strengths, and the methodologies used for analysis. Upholding data integrity is crucial for presenting actionable insights because stakeholders must have confidence in the data underlining the decisions.

In conclusion, the Applied Data Science program at Syracuse University has equipped me with a robust toolkit to navigate the complexities of the data landscape. I am confident in my ability to make meaningful contributions to organizations by leveraging data-driven insights to inform decision-making and drive innovation. Also, I have learned the importance of communication to stakeholders regarding the data to ensure transparency and fairness.

SOURCES

Hotz, N. (2023, January 19). OSEMN Data Science Life Cycle. Data Science Process Alliance. https://www.datascience-pm.com/osemn/

IBM. (n.d.). What is Data Science? https://www.ibm.com/topics/data-science

Syracuse University. (2020, June 17). Applied Data Science Master's Degree - ischool: Syracuse University. Syracuse School of Information. https://ischool.syr.edu/academics/applied-data-science-masters-degree/