# Portfolio Milestone

Joshua A. Gaze

SUID: 411546359

Syracuse University, School of Information Studies

Master of Science Applied Data Science



My name is Joshua Gaze and this is my portfolio milestone demonstrating my proficiency over the material covered throughout the Applied Data Science master’s program at Syracuse University. My first steps into beginning the journey towards a career in data science began during my undergraduate at Kennesaw State University where I wanted to obtain a more refined appreciation of my favorite pastime, albeit from a different perspective, being the game of baseball. Playing the game throughout my childhood provided me great insights about everything going on. When it became evident that I would never progress beyond the high school level, it was time to appreciate the game through another lens in sabermetrics. My studies in mathematics and applied statistics complemented my desire to discern new insights into the game. Throughout these times, an important lesson I’ve taken away, which greatly corresponds with my educational upbringing at Syracuse is that we can learn *some things* which were previously unclear or unknown, and that we can always learn more about a situation. Meaning that we can always expand on our knowledge to form/refine new insights.

**Reviewing the Program Learning Goals:**

As an interdisciplinary program, the master’s in Applied Data Science provides students the opportunity to learn in a broad range of areas related to data science. Students are able to demonstrate that they have successfully comprehended the lessons and material of this data science program through the means of:

1. **Collect, store, and access data by identifying and leveraging applicable technologies**

This in many ways showcases that we ascertain the ability of taking raw data and transforming it into useful information. Using applications and various software to examine patterns in swaths of data, we as data scientists are able to learn more about the end-user’s of our data product to develop effective operational improvement, increase sales, and decrease costs of doing business.

1. **Create actionable insight across a range of contexts (e.g. societal, business, political), using data and the full data science life cycle**

Going through this program, we have mainly focused on retrieving practical insights through the data analytical skills instructed to where the results of those useful insights are utilized to make smart decisions in the environment for which they were appropriated for. I have discerned, throughout the program, that data science is multidisciplinary having lots of different areas of application. Where the practice of data science can be used very broadly across many subject areas while also can be focused within one domain. With the growing need of data to gain that competitive edge in the market these days, it is imperative that companies are asking the right kinds of questions of what they want to achieve. Once the problem is defined and understood, how does one collect data that is relevant to the problem? After the data has been collected, we want to get to a state where we can explore the data and deploy models to gain valuable insights. However, a great deal of preparation must be done before we are able to begin. Data science is about working smart, not hard. Meaning that in order to produce the insights through modeling, we need to thoroughly clean and prepare the data that will be used. Since the actionable insights derived can potentially change the way your team, company, organization does business. After the data cleaning has been completed, we need to summarize the main features of the data set. Allowing the team to discern notable patterns and provide supporting or dissuading context towards assumptions that were held. As the team becomes familiar with the data prepared thus far, they will eventually reach the last step of the data science life cycle being the model development. Where the data is split into the training and test sets that will be used to develop the relevant machine learning models.

1. **Apply visualization and predictive models to help generate actionable insight**

There are several ways in which people are able to comprehend and take in information. So why would we fail ourselves and our audience in not ensuring our product is not properly received? If we are collaborating with several technical analysts, without fearing that the output of our predictive models will not go over their heads. If we were meeting with c-suite stakeholders, the nitty gritty details of our models is of little importance, for the most part, to them. They would instead be preferable towards a big picture takeaway that is immediately comprehendible so as to not waste time. Here we would lean towards data visualizations as it is a tremendous tool in the data scientist’s belt.

1. **Use programming languages such as R and Python to support the generation of actionable insight**

In a market where data is taking a more central component in operations, programming helps us to identify patterns and create practical models of the data structures we will be exposed to. We cannot always rely that the key takeaways from our data is just waiting to be discovered on the surface level of an excel sheet. There will be plenty of manipulation of the data we’ll be exposed to, and depending on the context. The practice of using Excel will not be enough. In order to parse through volumes of data in the most efficient manner, the development of our skills in the languages of Python, SQL, and R will be paramount. Machine learning has been increasing in its relevance for the data science setting. Many business users do not have a comprehensive view of these models that are essentially a blackbox and therefore are very much more likely for individuals to misuse them or even draw an incorrect conclusion from the output. Data Scientists that are well versed in their background of using programming languages will be better able to understand how these advanced tools work in practice.

1. **Communicate insights gained via visualization and analytics to a broad range of audiences (including project sponsors and technical team leads)**

Tables and data metrics alone cannot provide a comprehensive overview of a dataset. To deliver insights from another perspective, data scientists can use the tool of data visualizations. Similar to building machine learning models, there is a great deal of work that goes into constructing a visualization that conveys a clear story in the data. Every visualization provides a new opportunity of explaining your data to individuals that do not typically digest information in an analytical format. Where the message needs to be immediately apparent upon inspection. This is where the power of visual information comes in. Using various tools such as manipulating the size, color, shape, contrast, background to highlight noteworthy aspects we want our audience to takeaway from the visual.

1. **Apply ethics in the development, use and evaluation of data and predictive models (e.g., fairness, bias, transparency, privacy)**

Through predictive analytics, sensitive information about groups and individuals is predicted that is possibly without the data subjects’ knowledge, from less sensitive or more readily available material via proxy by leveraging the data left by many of other users. This predictive privacy can be violated regardless of whether the prediction was accurate or inaccurate. The challenge that we as data scientists must battle is how to apply a data driven process in a way that measurable reduces over reaching intrusions or judgements but also does not reduce the effectiveness of our data product. Because imagine what could go wrong? Predictive models that reinforce stereotypes in a racial of socioeconomic sense. Healthcare decisions that are incorrectly diagnosed due to the input of data science, having the capability of compromising an individuals health. Among countless of other examples. Going into the data science profession, we must remember that even the most well-intentioned individual can inadvertently commit unethical actions. It is our responsibility throughout the data science life cycle to ensure any present or long term ethical complications stemming from the project are minimized.

**Loyalty Rewards Database**

The course that this course was developed under was IST 659, the Database Administration Concepts and Database Management course. This course was taken under the direction of Dr. Gregory Block in Fall 2021, my first term of the applied data science program of Syracuse University. The goals of this project were to construct and implement a database to solve a data management problem of my choosing. Which I chose to construct a loyalty rewards database that takes credit card holders transaction activity and converts and acts as a ledger of their accrued loyalty points which they could ultimately use for supplementary benefits such as travel, cash back, statement credits, merchandise, etc. The project was divided into two parts of emphasis. The first part being the design specifications detailing the data that is to be tracked and how all of the elements, tables, records work together cohesively. To contextualize the setting in which this database system will be deployed in, we also laid out business rules that dictate how the data is managed. This was represented through the usage of a conceptual and logical model for visualizing the relationships of the database system. Identifying relevant stakeholders, interested parties, and end-users and to detail their involvement and what credentialing would be required for them to access, maintain, or participate within this constructed database system. The second part is the implementation of the design completed in part 1. This included the development of Structured Query Language (SQL) to create the tables and columns to house the loyalty rewards data and to also convey the database constraints that implement the business rules listed previously. Additional goals listed for this project was to include representative statements for the basic Data Manipulation Language (DML) that implement the create, read, update, and delete statements (referred to as CRUD) used in maintaining your data. Actionable insights that were gathered throughout the course of creating this database system for tracking loyalty rewards of credit card transactions was the general scope of how data flows throughout an organization. That things are not as simple as playing around with a couple of excel files, that it can get rather complicated in a quick manner depending on the complexity and scale of your data structures.

Graphical user interface, application

Description automatically generated

This project contributed towards the requirements of the data science program through my development of a database management system which showcased the significance of how data can be stored and accessed. Which is the foundation of what allows data analysts, data engineers, data scientists, and many additional roles to contribute in their work setting. The loyalty rewards database project also refined my skills in developing data systems through the use of the Data Manipulation Language commands of SQL. Throughout the course of the project, one additional method that I developed was to design a method/operation in which card holders would be able to access their transaction/rewards activity. One thing I ensured was paramount when designing this feature was to ensure the API was robust against a SQL injection attack, as well as ensuring that cardholders were unable to access, edit, or manipulate data fields belonging to other accounts. This is an instance in which I was able to apply ethics in the development, use, and evaluation of data through the channels of privacy and security.

**Netflix Catalogue Poster**

The course that this course was developed under was IST 719, the Information Visualization course. This course was taken under the direction of Dr. Gary Krudys in Spring 2022, my third term of the applied data science program of Syracuse University. The goals of the project were to identify stories within a data source and attempt to portray unique insights about the data through a visual lens. The data source I decided to comb through and tell a story through the use of visualizations was with the media catalogue carried by Netflix for its subscribers. While there is a tremendous amount of valuable data within the confines of the Netflix operation, I gained the valuable perspective from this course that sometimes the same data can tell a grander tale when relayed through the "proper" medium. Data visualizations possess the capability to portray information in a different perspective which can make it easier to comprehend by expressing it visually and demonstrating relationships between data points/features. The utilization of labels, colors, and attention-grabbing images allows for the human brain to process and retain gained insights. Which leads the intended audience towards a deeper understand of the story the data source is wishing to reveal. While data measures existing within the confines of a table or the output of a model often have informative takeaways, it's purely one dimensional and not more can be expressed. Take the world map on the final poster of the Netflix Catalogue for instance. A table can show us that the United States dominates Netflix landscape of countries that produced the content for its media catalogue. But the map visual also showcases that there's another dimension of insights at play with proximity. Where we see the North American continent by far comprising a majority of the catalogue, where many parts of the world are producing minimal numbers than what some would expect. Another perspective that was derived was when we performed a textual analysis of the diction choice of the media project's descriptions. This can be viewed towards the bottom of the poster with the "Netflix" logo where I was able to super impose a word cloud plot onto the company's logo. Here I was able to ascertain that it is very common to see descriptions having word choice displaying the density of words including 'world', 'girl ', 'love', 'Christmas', among others. There were a number of technologies that were utilized throughout the course of making this poster communicating Netflix's catalogue. For the general data cleaning and preparation of Netflix's data of its media catalogue, I used the programming services of RStudio with an assortment of the typical packages. In addition to data cleaning, I used R to perform my exploratory data analysis to prepare myself with the story this project wanted me to communicate about Netflix's catalogue. R lastly being used to generate the plotly map of the world showcasing the number of media projects being generated in each country. To generate the word cloud within the confines of the Netflix logo, I switched languages over to Python where I used the packages of the Python Imaging Library (PIL), matplotlib-pyplot, and the wordcloud package. The last technology utilized in this project was the software package of Adobe Illustrator. The main purpose of using this tool was for the general organization and layout of designing the actual poster. This was my first exposure to a photo editing software tool so while there was a learning curve to it, I found the experience incredibly enlightening and of great benefit should I ever need to use this going forward.

Timeline

Description automatically generated

This project contributed towards my refinement and competence within Syracuse's applied data science program through my development in understanding the nuances of conveying information through the channels of visualizations. This poster conveys actionable insight as to what words we would typically have in the description of a media project. Where this visual insight was delivered through the usage of programming languages such as R and Python.

# Credit Card Clustering

  This project was completed in the Winter term of 2022 which was my second term in the applied data science program. The course this project is associated with was the IST 707 - Applied Machine Learning/Data Analytics course under the direction of Professor Ying Lin. The goals of the project were to use skills developed throughout the course towards a self-selected dataset that reflects the overall goals of the course. Those goals included that I (1) document, analyze, and translate data mining needs into technical designs and solutions. (2) Apply data mining concepts, algorithms, and evaluation methods to real-world problems. And (3) employ data storytelling and dive into the data, find useful patterns, and articulate what patterns have been found, how they are found, and why they are valuable and trustworthy. The dataset selected for this project pertained with a banking institutions population of active credit card holders that consisted of 8,950 across 18 attributes that was sourced from Kaggle. The goals we wanted to tackle within this project was to construct a means in which to segment the credit card holders of the banking institution such that we could use the business insights, gathered through predictive classification models, to improve upon their operations and thusly their overall satisfaction with the financial services that the bank provides.

There were two main technologies utilized to develop the research behind clustering the banking institution's cardholders. For general exploratory data analysis and initial model constructing, the R programming language was serviced. Parts that were vital to the project that were generated using R was the transformation our quantitative attributes. As this was the distribution amongst our dataset's attributes:

Chart

Description automatically generated with medium confidence

Examining the skewness of our data attributes gave insight as to how we would proceed with outliers and/or missing values. With a skewed distribution such as what is shown above, we thought it to be prudent to perform imputation with the median over the mean as the measure of central tendency when dealing with NA’s. Due to the number of records in the dataset, we decided to keep outliers in the dataset as we didn’t want to diminish the sizes of our training and testing datasets when generating models. The other roadblock that the packages within R helped overcome was the varying scales of our quantitative attributes. It was important that we'd be able to relatively compare quantitative attributes with different scaling systems as if they were under a uniform scale. For instance, the data attribute *PAYMENTS* ranged in values from 0 to 50721.5 and *TENURE* ranged in values from 6 to 12. Therefore, before we could even consider using and machine learning algorithms to produce anything that could be useful, we thought to transform these irregularly scaled variables such that they were all on the same scale through normalizing all our quantitative variables to have a mean of 0 and a standard deviation of 1.

  After our data cleaning and transformation was completed in R, we utilized the Python programming language for when developing our machine learning models such as K-Means Clustering and Hierarchical Clustering.

Chart, line chart

Description automatically generated

Examining the totality of the dataset and using the elbow plot, shown above, that the optimal number of clusters to partition the banking institution's cardholders was 4 segments. The reason this value was chosen is due to the diminishing returns that become apparent in the chart above. Additionally, we felt that segmenting the customer base of the bank will become increasingly difficult with the additional number of clusters to proceed with, so it was determined that this inflection point occurs at the number of clusters being at a value of 4.

Chart, scatter chart

Description automatically generated

Graphical user interface

Description automatically generated

Diagram

Description automatically generated

Using the K-Means clustering unsupervised machine learning algorithm on k=4 groups, we used the plots provided above to generate the following interpretations about the 4 clusters. This was one of the goals we intended on delivering to our "banking client" so that they can tailor their marketing and operational attention accordingly. We found that the cardholders of *cluster 1* are frequent users of their credit card with a moderate tier of income that was primarily spending their credit on lower priced goods. While also confidently paying off a decent majority of their balance at the end of their cycle period. The cardholders within *cluster 2* are high frequency users of the credit card with a higher tier of income spending their money on higher priced consumer products. While also confidently paying off a good majority of their balance at the end of their cycle period. The cardholders of cluster 3 are infrequent users of the credit card that take out a high number of items against their line of credit. Meaning that cardholders falling into this category should closely be examined to ensure they are not falling behind on their balance payoffs. the cardholders within cluster 4 are infrequent users of the credit card, with a relatively lower income tier spending their money on lower priced consumer products.

This project provided an opportunity for examining and repurposing a dataset that was not provided to us by the instructor, where we were solely responsible for identifying patterns both within and between the established clusters of credit card holders at this bank. The collection of the dataset required special attention and respect towards cardholder privacy, where any identifying data attributes that were not useful for meaningful analysis were ultimately removed. Which is a vital aspect to take into account because consider the ethical ramifications if this project were utilized in a different setting than a master's program and were instead in the private sector with an actual banking institution. We would now be ascribing attributes onto specific individuals regardless as to the accuracy of those predictions.

## **Summary of All Program Learning Outcomes**

This portfolio has demonstrated the successful implementation of the laid-out learning objectives and the significant practice areas in the Applied Data Science program of Syracuse University. Data was collected and managed in a database environment where it could then allow retrieval for accomplishing further data science tasks using statistical methods and data mining for classification, regression, or clustering. Data visualizations were utilized to convey ideas and insights that aren't best delivered through the medium of model output and metrics in our Information Visualization (IST-719) and Applied Machine Learning/Data Analytics (IST-707) courses. Where we were, respectively, able to identify textual patterns and country distribution of the media items in Netflix's catalogue and able to see the usage behaviors behind credit card holders segmented into clusters in a banking institution. The many domains of media, banking, and loyalty rewards showcases that I can apply the data science techniques I’ve learned throughout this program to generate actionable insights. Especially in the projects relating to finance, to develop safeguards to ensure the end-users are safe from exploitation in a negative aspect to ensure that no ethical boundaries have been breached. Being able to perform all of these achievements through utilizing the programming languages and software packages to enable myself to deliver the insights my data was able to convey. Through this and many more aspects, showcases my achievement of all the applied data science program learning outcomes.

## **Reflections about the Program**

My time throughout this master’s program has been challenging, yet not in the negative sense. I have had to learn new ways to analyze information and formulate questions that truly aim at the core of what’s going on in my data landscape. Learning new ways of communicating information from both a technical purview but more so in the human connection. Since I am not the end-user of my data product, those that I am aiming to support and how I deliver my data science solutions is vital to the overall success of the project. Communicating through data visualizations to convey patterns in the data that could otherwise go unsaid and bring zero value. My instructors have instilled in me an analytical approach to derive insight and seek patterns but always from purview of skepticism. Showing that we as data scientists can learn *some* things that are unclear or previously unknown and that we can always learn more about our data. The skills of applied data science are useless when they exist in the confines of a black box where they cannot be interpreted and communicated in a meaningful way. That there’s plenty more to consider about data than simply the output results of a model. It has to be interpreted and appreciated for the message it is trying to convey. It must be comprehendible to those that are not familiar with it. All this so that we can expand our knowledge to deliver the insights we produce as data scientists from Syracuse University.