Portfolio Milestone Report

Alison Reikher

Syracuse University | Master's in Applied Data Science | Summer 2022

Table of Contents

Introduction

Throughout this program, I had the opportunity to explore various interests of mine. With every course, we were required to present a final project. Each final project had its requirements, but we were always able to pick our own creative dataset. If you know me, you'd know that some of my favorite topics are cats, banned books, and YouTube. It should be no surprise that the corresponding projects were selected for this portfolio.

The projects also corresponded with the three main programming languages in data science: SQL, Python, and R. The projects represent my knowledge and mastery of important concepts in all three languages. Data science isn't just dominated by one language. Data science relies on all three languages to be able to effectively explore the data and answer the target questions.

The projects in this portfolio demonstrate the completion of the seven learning goals. The learning goals aim to provide structure to the program and to help us demonstrate real-world skills. Each project demonstrated several of these goals and combined, all the goals were met. Throughout this portfolio, my three favorite projects will be discussed and connected to these learning goals.

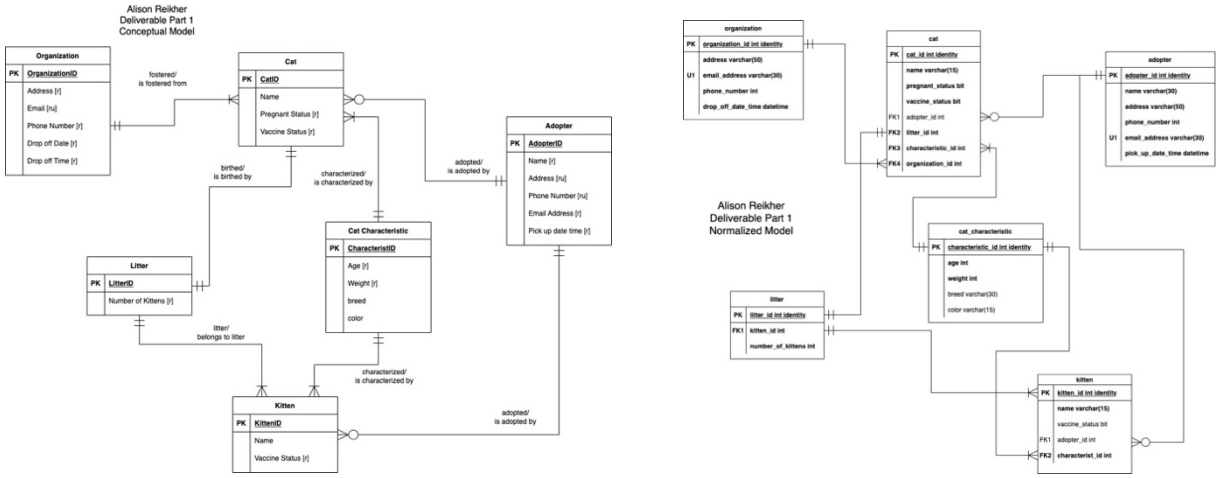
Projects

IST 659: Database Administration

The project assigned for this course required creating a database that would solve a data management problem of my choosing. The problem I chose was inspired by a friend's quarantine hobby. During the pandemic, many people started to foster rescue animals in their spare time. My friend was no exception, she started to foster pregnant cats and their kittens. Along with taking care of them pre- and post-labor, she would also have to find people to adopt these cats. At any given moment, she was unaware of how many cats and kittens were in her house due to the number of animals that she was taking care of. To solve this problem, a

database was created to track how many cats and kittens there were, where they were coming from, and where they were going next.

The first step to creating this database was to create two models: a conceptual and a normalized. These models helped organize the data tables before implementing them in SQL. A lot of thought had to go in before creating the tables in SQL. The models helped establish the relationships before the code was executed.



In addition to these models, the data needed to be created and organized. Most of this data was generated with a random data generator, however certain fields had to logically be restricted against other fields. For example, a cat can't be adopted before it arrived at the house, and a kitten can't be adopted before it was born. The fields had to be tinkered with so that they logically fit a real-world scenario. Excel spreadsheets helped organize all the data before implementation.

Once all the tables were in SQL, the data was able to be manipulated to answer key questions. Some of the questions include which cats are available, which cats have already been adopted, and which cats still need to be vaccinated. These questions are very important since they help keep all the information organized. It gets more and more difficult to keep track of this information as more and more cats arrive.

Additionally, Microsoft Access forms were created for a more interactive, user-facing platform. These forms include both areas to enter information along with creating checklists to answer specific questions. For example, a form was created for a user to enter in a cat's information when it arrives home rather than coding it into SQL. A checklist was created with all the cats and their vaccination status. This list allows for an easier, user-friendly visual of the cats' medical records.

Reflection and Learning Goals

Collect and organize data

Data was collected for this database by generating data and real-life foster cats and kittens. There are 20 different variables that all needed multiple data points. The data generator helped simplify coming up with certain data such as the information about the cat organizations and adopters. All information that was not generated by the data generator was manually created. Information like cat breed, weight, and age could not be randomly generated since they all dependent on one another. A 6-pound cat couldn't give birth to a 12-pound kitten of a different breed that was older than the mom. The information needed to stay as logical as possible. The research was done to make sure the weights were as close to the average weight for the breed as possible. There also needed to be an agreement about age and adoption day. This data could only be created manually.

Develop a plan of action to implement the business decisions derived from the analyses

Before implementation of the code, stakeholders, business rules, and data questions were established. This helped give the database purpose and structure. Many of these questions were answered through SQL or MS Access via reports. These reports helped inform what action needed to be taken next. The two key actions were finding who needs to be vaccinated and who needs to be adopted. Once this was found, the next plan of action was to either find a home that was right for the cat or kitten or get the cat or kitten vaccinated.

IST 652: Scripting for Data Analysis

For this project, any data source of our choosing was to be selected and analyzed. The data source my partner and I selected was banned book reviews from Amazon and Barnes and Noble. The goal of the analysis was to better understand how readers from Amazon reviewed banned books in comparison to Barnes and Noble. The idea behind this was to better understand how people, in general, viewed these banned books along with understanding how people reviewed books on these two different websites.

The data for this project was manually collected. We agreed on 20 of the most commonly banned books and split the data collection work. We collected information about the author, the year published, the reason for the ban, whether the book was adapted for film or TV, the number of 5 stars through 1-star ratings, and reviews for each star rating. The reviews were pulled from both the Barnes and Nobles and Amazon websites. The comments were chosen through random selection. Longer comments gave us more to work with than smaller ones. Sometimes, there would only be one comment to choose from.

After the data was pulled, we began exploring the data to answer our four data questions. The following are the questions we aimed to answer:

1. Do users rate books differently on different websites?
2. Are there any predominant book genres in our data set?
3. Are there any strong correlations between any variables in either data set?
4. What are the differences between the Amazon reviews and the Barnes and Noble reviews?

<u>Do users rate books differently on different websites?</u>

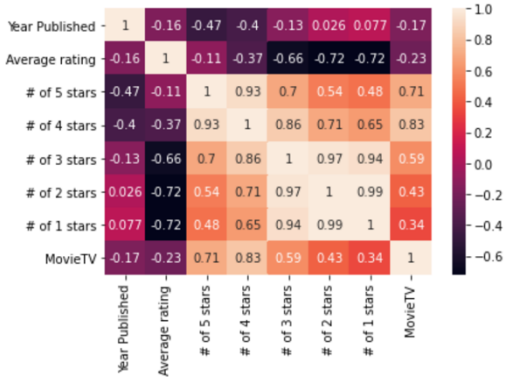
Our findings showed that the average book on Amazon's website is 4.64 out of 5. For Barnes and Noble, it was 4.6 out of 5. These averages are very similar. This shows that users tend to rate banned books similarly on these two websites. While we don't know the global average of all books, based on how similar the star ratings are for each book on each platform, it is fair to assume that the averages are similar and are representative of the true average.

Are there any predominant book genres in our data set?

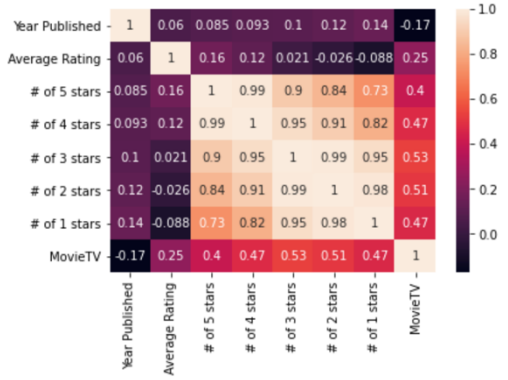
Three genres appeared multiple times in our data set. Young Adult Fiction appeared 6 times, Children's Literature appeared 3 times and Graphic Novel appeared 3 times. From these genres, it is clear that the books that are most often banned are aimed at children, teens, and young adults.

Are there any strong correlation between any variables in either data set?

To answer this question, two correlation matrices were created. The following are these two matrices.



Amazon Matrix



Barnes and Noble Matrix

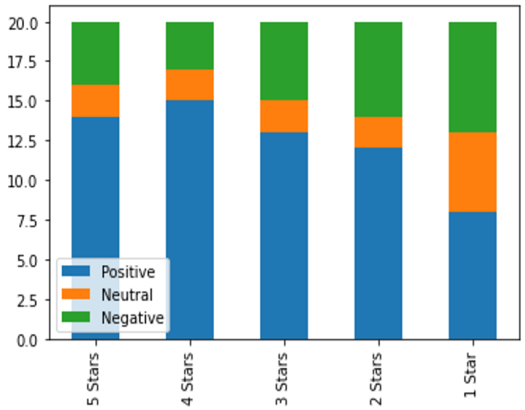
Looking first at the Amazon Matrix, the only noteworthy correlations exist are with MovieTV. All other correlations are either trivial or insignificant. For instance, it is trivial that the average is correlated with the number of stars the review gets. Also, for the Year Published correlations, they are all either close to zero or less than 50%. In the MovieTV correlations, we see the strongest correlations for the number of 5 stars and the number of 4 stars. With these two correlations being significant, this shows that a user is more likely to rate a banned book highly if it had been adapted into a movie or TV show. For Amazon, this makes sense. A user could go watch a movie directly on the website and then go purchase the book immediately after.

The Barnes and Noble Matrix only has trivial correlations. All significant correlations are in relation to the adjacent star ratings. It is trivial that if there are a high number of 5-stars, there is most likely a high number of 4-stars. The same can be said for 1- and 2-star ratings. The one notable detail of this matrix is the low significance it has for MovieTV. Unlike Amazon, the fact that the book was adapted to a movie or TV had little to no bearing on how a user rated the book. If anything, the highest correlation is with the lower star ratings. This could suggest that users are more inclined to give a lower rating if the book at been adapted.

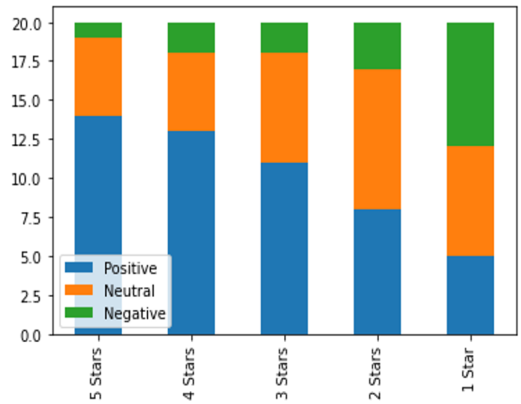
What are the differences between the Amazon reviews and the Barnes and Noble reviews?

The answer to this question was explored in two ways. The first way to understand the differences between these reviews is to look at the individual words used in the reviews. The approach for this required us to create word frequency lists. After removing the stopwords from both sets, we found the top ten most frequently used words for each rating for each data set. It turned out that the most common across all ratings and both datasets is the word 'book'.

With the use of Wordclouds, we discovered more about the language used to describe the books for each star rating. The higher ratings tended to be more positive than lower ratings. As the rating decreased, we noticed that the Amazon reviews reflected more on the reasons why the book was banned while the Barns and Noble reviews reflected more about the content of the book itself. Words in the Amazon review suggest a criticism of being inappropriate for age groups. Words in the Barnes and Noble review criticize the writing and don't factor in the reason for the ban as much.



Amazon Sentiment Bar chart



Barnes and Noble Bar chart

The second approach to analyzing the text was through sentiment analysis. The bar charts above represent the number of each sentiment score by star rating. For the Amazon reviews, they were all mostly positive. Even at the 1-star level, there were considered more positive reviews than negative. There were very few neutral reviews. It was unexpected for there to be such a high number of positive reviews across all star ratings. It was also unexpected that there would be more positive reviews for 4-stars than 5-stars. This could be a reflection as to how users tend to rate books in general. However, the sample size is too small to make an informed conclusion.

On the Barnes and Noble side, the results were much more expected. The number of negative reviews was low for 5-stars and gradually increased as the star rating went down. The number of positive reviews was high for 5-stars and gradually decreased as the star rating went down. Unlike Amazon, Barnes and Noble had a lot of neutral comments. This is partially because there wasn't a review for every single book. The sentiment analyzer will give blank values a 0 score if there is nothing to analyze. A 0 sentiment would be considered neutral. However, there were still a significant number of positive and negative reviews.

Reflection and Learning Goals

Describe a broad overview of the major practice areas in data science

This project incorporated several different major areas of data science practice. The major areas include statistics, machine learning, and natural language processing. Summary

statistics were used to get a better overview of the data. It helped with understanding how the two sites rated the books. Machine learning was used in the creation of the correlation matrices. They were used to find any meaningful correlations within both datasets. Natural language processing was used to analyze the sentiment of the reviews. This told us how many of the reviews were positive, negative, and neutral.

Collect and organize data

Both datasets used in this analysis were manually collected and the work was split between my partner and me. We collected information about the book and the relevant information regarding the Amazon and Barnes and Noble reviews. The data was then organized into two dataframes for each dataset.

Identify patterns in data via visualization, statistical analysis, and data mining

Patterns were identified using visualization, statistical analysis, and data mining. There were several different visualizations used for this project and all served to answer different questions. Two correlation matrices were created where the darker colors reflected a more negative correlation, and the lighter colors reflected a more positive correlation. Word clouds were created to visually see the most frequently used words. Bar charts were created to help visually see the number of positive, negative, and neutral reviews for each star rating for each dataset. Statistical analysis was used when exploring the averages between the two sites' ratings. Data mining was done when observing the frequently used words for the two review datasets. The text was analyzed to determine the overall sentiment and how the review relates to a book being banned.

Demonstrate communication skills regarding data and its analysis for managers, IT professionals, programmers, statisticians, and other relevant professionals in their organization

A presentation was done for this project. I worked with my partner to create a slide deck based on the report we had written together. The focus of our presentation was to discuss the datasets we collected and answer the questions we had set out to answer. All the visualizations we created were used in the presentation to better demonstrate our findings.

Synthesize the ethical dimensions of data science practice

The project does pose an ethical dilemma in data science. The reviews were collected from public websites. However, we did not have consent from the reviewer to analyze their review. Since the purpose of our analyses was to understand how users feel about banned books, we are potentially judging these reviewers on their morals. While there is no harm done with our analysis, there is a larger question of whether users should have the right to not have their thoughts analyzed.

IST 707: Applied Machine Learning

For this project, we were asked to pick any dataset of our choosing and apply at least 2 machine learning algorithms to this dataset. The dataset I chose was the YouTube Dislikes Dataset from Kaggle. This dataset consists of data from the YouTube explore page before YouTube removed the display of the dislike count.

There were three algorithms I chose for this assignment. The first was Associative Rule Mining. The goal of this was to find any interesting rules between the variables. The most common rule found was that as the variables like likes, dislikes, views, and comment count increase, the other variables also increase. This rule is expected. When a video is getting more attention, it will receive more feedback in the form of likes, dislikes, and comments.

The only other rule I was able to find outside this scope was through the channel called Sky Sports Football. I found a rule that says that this channel is most frequently associated with receiving between 22 and 20,000 likes and receiving between 3 and 404 dislikes. This channel is also the most frequently appearing channel on the explore page, so this explains why it is the only one that came up.

The second algorithm used was Sentiment Analysis. Sentiment Analysis was used to analyze the titles, comments, tags, and descriptions for all the videos that had more dislikes than likes. The purpose of this analysis was to understand how a video could be popular enough to reach the trending page and yet be considered mostly disliked. The comments came up mostly positive. 78% of the comments were classified as positive. The reason for this could be

because of censorship of comments. The channel owner or the YouTube algorithm can remove comments that could be violations of the terms of service. The descriptions are the second most positive. 55% of the descriptions were classified as positive. Since the descriptions are written by the creator, it makes sense for this to be somewhat more positive. However, it is important to keep in mind that certain topics will be classified as positive or negative. In this case, it is unavoidable for the description to be negative. Tags are the second most negative. It is 33% negative and 35% neutral. Tags were the most descriptive because they are essentially descriptions without the stopwords. These tags being neutral and negative suggest the topic of these videos could be the cause for the negative sentiment. And if the topic itself is negative or controversial, that can explain the high dislike ratio. Titles had the most negative and neutral sentiment with 32% negative and 56% neutral. The titles being so neutral and negative makes sense. Since the title is what makes the viewer click, the title must relate to the content. It wouldn't make sense for a serious or negative topic to have an overly positive title.

The third algorithm used is correlation matrices. Two matrices were created. The first one was for the entire dataframe. The most significant correlation coefficient was 0.78 between view counts and likes. This implies that as views increase, the number of likes tends to also increase. The second matrix is for the dataframe with just the videos with more dislikes than likes. A strong correlation of 0.84 was found between dislikes and comment count. This implies that as the number of dislikes increases, the number of comments left on the video also increases.

Reflection and Learning Goals

Describe a broad overview of the major practice areas in data science

The major practice area in data science demonstrated in this project was the machine learning algorithms used. Machine learning is becoming more and more common in data science practice. This project explored 3 different algorithms which all contributed to a better understanding of the dataset.

Develop alternative strategies based on the data

There were three strategies developed to better understand the data. The Apriori Method was used to understand any rules between the text and numerical data. Sentiment analysis was used the better understand the text data for the target videos. Correlation matrices were used to understand how the numerical data corresponded to one another. All these strategies have merits of their own. Some even overlap. By developing these alternative strategies to approach the dataset, a broader understanding of the data was able to be made.

Demonstrate communication skills regarding data and its analysis for managers, IT professionals, programmers, statisticians, and other relevant professionals in their organization

For this project, a final presentation was done to show the various approaches to exploring the data through machine learning. Since I worked on this project by myself, I created the slide deck on my own based on the final report I had written. I included background information, summary statistics, and a walkthrough of the results from the analysis. Visuals were also used to help demonstrate the findings.

Conclusion

I'm incredibly proud of all three projects I discussed in this report. These projects show a broader overview of each topic through the lens of various different data science concepts. I used this opportunity to explore topics such as foster kitten adoption, banned book reviews, and YouTube explore page data. Having the freedom to choose leads to more interesting questions being asked and a better understanding of the data science concepts.

It is important to be able to apply data science concepts in a variety of languages. SQL allows for databases to be easily created and organized. Databases can help bring something as chaotic as foster cat adoption to order. Python has great modules like Pandas that allow for dataframes to be made from datasets. It also allows for sentiment analysis and correlation analysis. R allows for easy statistical exploration. It allows for a variety of machine learning algorithms such as associative rule mining, sentiment analysis, and correlation analysis.

All three projects demonstrate completion of all seven learning goals. These learning goals help apply the work done in the project to the real world. It is important to be able to use

the skills learned in these courses outside of class. It is much easier to learn about data science when it is done through the lens of interesting subject areas such as cats, banned books, and YouTube data. Data science is still a new and growing field, so it is important as data scientists that we continue to learn and explore different topics.