**Portfolio Milestone**

Kasirha Goodman

IST 782

10 September, 2023

*Introduction*

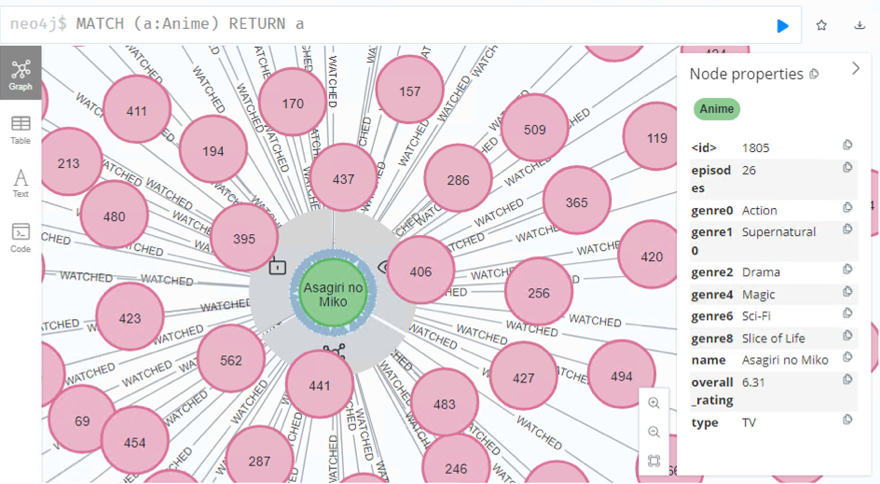
While completing my MS in Applied Data Science at the Syracuse University iSchool, I have had the opportunity to work with a wide variety of languages, datasets, software, and students. My projects have ranged from the smallest scale of building SQL databases for Mike Fudge’s many fictional businesses to the largest scale of scraping metadata from thousands of real videos posted on Tiktok from January 2021 to December 2022. Each of the courses that I have completed has given me both a theoretical background and a practical experience for its topic area. My professors have been there every step of the way to inform and support my learning and my classmates have given me a new appreciation for group work where each member of the group has different strengths and areas of expertise. I have tackled projects both on my own and in a group, finding myself serving the role of designer, coder, presenter, project manager, and editor. In short, my four terms in this program have exceeded my expectations for what I would be able to learn in a single year, and I now feel confident seeking employment in the role of a data scientist.

For this program, I chose the Data Pipelines and Platforms Track. Each term, I took three courses as a full-time student. While all of my courses provided the opportunity for me to do interesting data analysis work, for the purpose of this paper I have chosen four final projects from four of the twelve courses that I took to exemplify the six learning outcomes outlined in the program.

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)

These four projects also show my proficiency in R, Python, and SQL-based query languages, as well as my ability to collect and analyze data in diverse forms and from diverse sources. Three of these projects were completed as a part of a group, in which case I will discuss my contribution to the project. The fourth project was done entirely on my own. Regardless of whether I completed the project solo or as part of a team, my part in each enabled me to gain and demonstrate skills that I intend to utilize in the future. I intend on finding data science work, preferably on the marketing or research and development side of the private sector or the scientific research side of the public sector. I will continue working with the languages and software that I have used during the program, and will expand my work beyond these in the future.

*IST 769: Anime Ratings and Recommendations*

****For my final project in Advanced Big Data Management, I downloaded a large dataset (over 21 million individual entries) from Kaggle that had been scraped from MyAnimeList. MyAnimeList is a website that allows users to keep track of the Japanese and Japanese-style animated shows and films that they have watched, are watching, or want to watch. It is regularly updated, organizes shows by genre, and has a basic recommendation feature. My goal was to create my own recommendations by uploading the data and all its relationships into Neo4j, an open-source graph database management system that uses the SQL-related query language Cypher. I chose a graph database to store the information based on the structure of the data—with users watching and rating multiple shows and shows having multiple attributes attached to them. Such relationships can be easily translated to the kind of node and edge format used by Neo4j. While I did not have to scrape (collect) the data myself, the focus of the course project was to demonstrate our ability to store data in the correct format and to access the data using a query. In order to do this, I had to use one of Syracuse’s virtual machines that had docker and Neo4j installed.

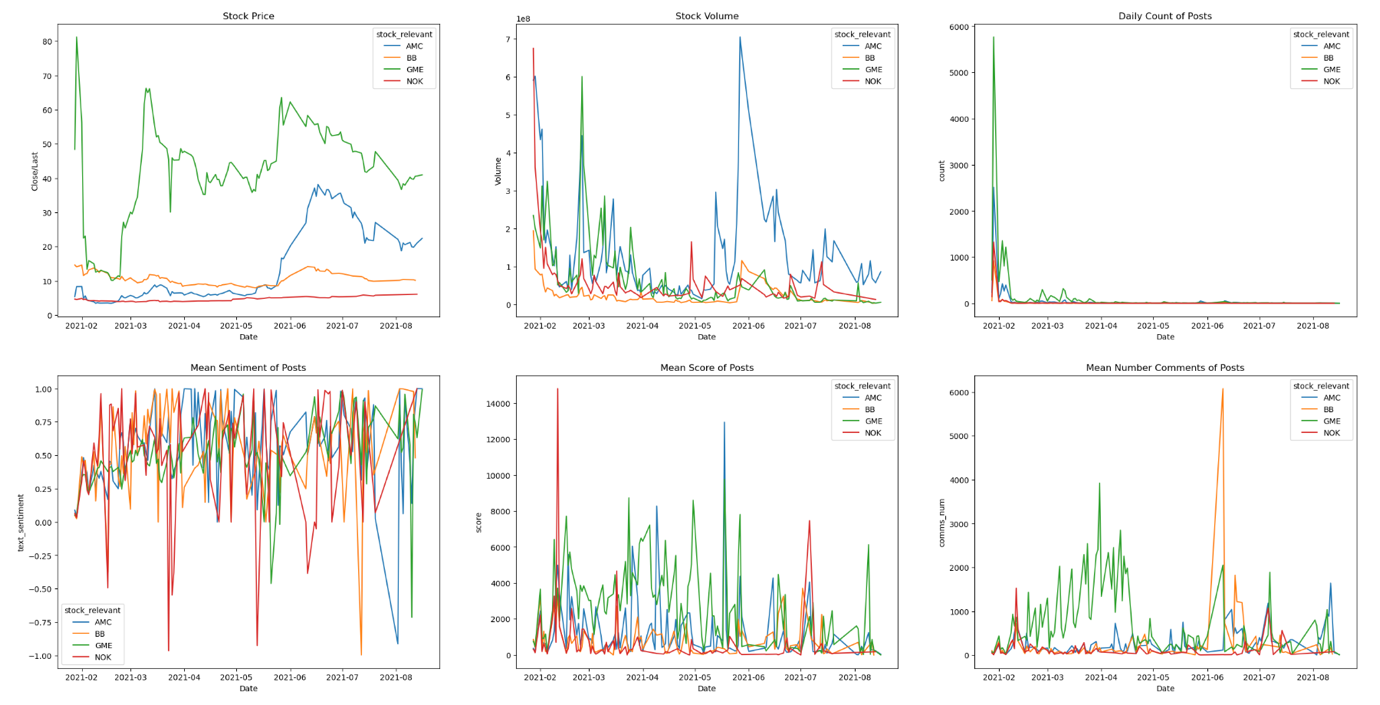
This required me to:

* Launch my virtual machine in Azure
* Download the necessary software to access my virtual machine from my Macbook
* Move the data from Kaggle to the virtual machine in a Python-readable format
* Launch docker to start Spark, Jupyter Notebooks, and Neo4j
* Write Python code that allowed me to use Spark in Jupyter Notebooks
* Load the modified data into Neo4j using a Spark connector
* Save both the Spark and Neo4j databases on the virtual machine
* Query the Neo4j database

This project, which was focused on data ingestion into newly introduced software, clearly connected to the first learning outcome—to collect, store, and access data by identifying and leveraging applicable technologies.

*IST 718: Wall Street Bets: Memes Reflecting and Affecting the Stock Market*

For my final project in Big Data Analytics, I applied Python-based modes of analysis and visualization to a large Kaggle dataset scraped from r/wallstreetbets during the first quarter of 2021. The project also incorporated historical stock market data for key “meme stocks” from this same period. The goal was to determine if factors like use of meme images with text, the use of positive and negative emoji, and the sentiment intensity and polarity of the readable text in posts had any measurable predictive value for the daily change in price of the discussed stocks. This project was completed using Python, and utilized several specialized libraries, including pytesseract, tqdm, the VADER sentiment analyzer, matplotlib, sklearn, scipy, and SMOTE. My partner and I used these libraries to create multiple vizualizations of the data and of our predictive models. These visualizations were further utilized to explain our conclusions—that expressed sentiments regarding stocks on a popular social media platform are not reliable indicators of whether the stock price will go up or down in a day, and should therefore not be used to make investing decisions. My focus on this project was the planning, question and insight generation, writing, data cleaning and preprocessing, and Python code editing.

Since this project used Python for most of its heavy lifting, it is a good demonstration of the fourth learning outcome of this program—to use programming languages to support the generation of actionable insight. In addition, an important part of this project was visualizing the stock market data, the activity on r/wallstreetbets, and their relationship to one another—which ties into the third learning outcome—to apply visualization and predictive models to help generate actionable insight.

*IST 707: Predicting Tiktok Virality*

For the final project in Applied Machine Learning, my partner and I generated our own data and analyzed it using R. We used Apify to scrape 981 distinct metadata attributes from around 8000 videos posted on Tiktok during the 2021-2022 period. We chose the videos based on the hashtags that were used to index it on the app. We separated these hashtags into three categories: top ten popular hashtags in 2021, top ten popular hashtags in 2022, and top ten popular hashtags of all time (referred to as “evergreen” hashtags). We found the lists of these popular hashtags in articles discussing social media trends of the selected time periods. While some data cleaning (choosing which of the 981 attributes to keep and which to discard) was done by hand in Excel, the bulk of the preprocessing was done using R. All the data was converted into formats readable by R, and then discretized so that association rules could be mined from the data. The target attribute in this project was playcount—the number of views that a tiktok video has. The playcount determines whether or not a video is “viral”—which we defined as meaning that it performed above the 3rd quartile of its peers. Ultimately, none of the available variables (hashtags, creator popularity, video likes, creation time, duration, number of shares, number of comments, and TikTok verification) were reliable predictors of whether or not a video would go “viral”. This reflects the findings of other social media research, but contradicts the claims of paid sites and services that guarantee celebrity to their prospective influencer clients.

My part in this project was the data acquisition, data cleaning and preprocessing, data analysis, and R coding. Since this was early in the program for me, I felt more comfortable letting my partner take charge of the big picture things like generating insight from the computational findings. With his guidance, I was able to achieve the second learning outcome—to create actionable insight across a range of contexts using data and the full data science life cycle. Our findings also led us to further questions regarding fairness, bias, and ethical business practices in the social media sphere. Thus this project also contributed to the sixth learning outcome—to apply ethics in the development, use and evaluation of data and predictive models.

*687: Analyzing and Predicting US Natural Gas Production*

The last project that I will mention in this portfolio was also my first project in the program. I worked on a team of five and let my more experienced classmates direct me in my role. As a result, the content of this project does not reflect my later interests, but it did give me valuable experience working with a team to create a coherent product. For Introduction to Data Science, the final project analyzed natural gas production data from the United States over the past two decades. The data was sourced from the U.S. Energy Information Administration’s API, and documents gas production, number of drill rigs, and futures prices. We explored the data to visualize the pattern of growth across regions and built a predictive model to determine the next possible “boom” region for natural gas. My main responsibilities in this project were visualizing and analyzing the historical data. Our professor also wanted each member of the team to contribute equally to the class presentation. This meant that I needed to understand not only my small part of the project, but also its place in the larger whole. I also needed to be able to communicate to a class of laypeople the insights generated by our team’s resident natural gas expert—who chose this topic as an extension of work he was already doing as part of his job. This experience gave me the fifth learning outcome—to communicate insights gained via visualization and analytics to a broad range of audiences.

*Conclusions*

This program has been a journey of discovery from start to finish. While I often bit off more than I could comfortably chew, my struggles ultimately led to greater learning. Besides the stated learning outcomes of this program, I also honed my time management and group work skills. In the past I struggled to work as a part of a group—usually taking the lead and shouldering most of the burden of work myself. In this program, I found that I could trust my classmates to be just as passionate and dedicated to the work as I was, which allowed me to give them the creative freedom to change the project from something I made into something *we* made. Learning to trust my group also helped me to become more confident in my own solo work. If my work was good enough for other, more experienced, people to accept, then I should be willing to accept it as well. As my master’s degree comes to a close, I find myself ready and able to enter the workplace as a competent data scientist. I still have much to learn, but if I can do a semester’s worth of work in 12 week sprints, then I can learn on the job.

Link to video presentation: <https://zoom.us/clips/share/A2F3MSAMt_uogJh8qmD86DRT4uKqM02Hs4Krl1rFG0MQs6otIQ>