Jeremy Dang

ID: 131177953

GitHub: https://github.com/danceswithme/nba_analysis

Final Journal Entry

**Introduction**

This research project was both a challenge and also a good learning experience in regards to both higher levels of research for a possible future career and also doing something that I have an interest in. This project broadened my horizons and got me more comfortable with python, Jupyter, and also got my feet wet with Git, GitHub, API pulls,  and MongoDB. I started initially with attempting to analyze tweets with the word “rigged” or the hashtag #rigged, although I was later informed that this was a common practice called sentiment analysis. Additionally the API changes to Twitter, now called X, by Elon Musk, made it much more costly to do what I wanted to do which meant I had to find a different medium for my data collection. I decided on basketball since I like its statistics and history, and chose to use basketball-reference.com as it had the data I needed and was known to be welcoming for data analysts and scientists to use its data, responsibly, as needed. I wanted to see if players who chose new teams had similar motivations to regular people when it comes to moving to new cities for a new job or a change of scenery. I pulled the cities and states and made API requests for the average city temperature, and state tax and per capita income to compare between.

**Week 1: Initial Research and Project Redirection**

During the first week, I embarked on initial research to determine the focus of my project. My original plan was to collect Tweet data to train a classifier to identify the subjects of tweets containing "rigged" or "#rigged." However, I discovered that accessing Twitter's search

API was prohibitively expensive, prompting a change in direction. My new project idea stemmed from a coworker's suggestion about NBA player movements. I also set up a GitHub repository to ensure version control and transparency. This week taught me the importance of adaptability and the reality that data can be costly.

**Week 2: Data Collection Strategy and Manual Data Entry**

In the second week, I delved into collecting data from Basketball-Reference.com. Initially, I considered using Stathead, but it did not meet my specific needs, so I manually collected data from various pages. This process involved copying and pasting player information into Excel, removing duplicates, and identifying players who moved to new teams. The labor-intensive nature of this task highlighted that data analysis often involves significant manual effort. I also researched historical player movement trends and identified three comparison categories: cost of living, state tax, and annual average temperature.

**Week 3: Data Compilation and API Research**

By the third week, I continued to fill out the Excel sheet, which included 250 players, and researched API sources for future data collection. An internet outage highlighted the dependency on technology and the importance of having backups. I completed the Excel sheet and set up a GitHub repository to practice version control and proper documentation. This week reinforced the value of persistence and the necessity of technology in data analysis.

**Week 4: Preparing Data for API Lookups**

During the fourth week, I finalized the Excel sheet, ensuring each player's movement was recorded in separate rows. I discovered a weather website with a generous API allowance and

wrote code to convert city names to latitude and longitude for API lookups. Troubleshooting the Python programs for uploading data to MongoDB was a valuable learning experience in debugging and the interaction between different programs. This week emphasized the importance of attention to detail and the iterative nature of programming.

**Week 5: API Lookups and Data Challenges**

In the fifth week, I began API lookups for annual temperature, state tax rates, and city consumer price indexes. The US Bureau of Labor Statistics (BLS) API presented challenges, leading me to switch to a more accessible dataset, the per capita income (PCI) project data from the United States Reap. This week underscored the importance of flexibility and the need to investigate datasets thoroughly. I also learned the power of virtual environments for managing dependencies.

**Week 6: Combining and Cleaning Data**

The sixth week focused on combining and cleaning data from various sources. I downloaded state-specific PCI data and wrote a Python script to add relevant rows to MongoDB. Balancing this project with other academic commitments taught me time management skills. Interactions with peers provided an outsider perspective, reinforcing the complexity and interest of my research. This week highlighted the necessity of thorough data preparation and peer feedback.

**Week 7: Standardizing Tax Data**

During the seventh week, I tackled the challenge of standardizing state tax data. Given the variability of tax data across states, I decided to categorize tax rates into tiers (none, low,

moderate, high) to facilitate analysis. This simplification made the data more manageable and comparable. I also separated entries into different collections in MongoDB to future-proof the database. This week demonstrated the importance of simplifying complex data and planning for future use.

**Week 8: Data Analysis and Future Proofing**

In the eighth week, I finalized the tax data and uploaded it to MongoDB. I began data analysis using Jupyter Notebook, performing significance testing and bar chart plotting. This period required shaking off the rust in data analytics and preparing for geo mapping in QGIS. I learned the importance of maintaining skills and the value of interactive visualizations for data analysis. This week also emphasized the need for consistency and future-proofing data.

**Week 9: Trend Graphing and QGIS Mapping**

The ninth week involved trend graphing and mapping in QGIS. Despite finding QGIS less beginner-friendly than ArcGIS, I persevered and created maps to visualize player movements. I also created frequency bar graphs to identify popular destination cities over time. Sharing my research with coworkers provided valuable insights and recommendations. This week reinforced the significance of data visualization and the benefits of collaboration.

**Week 10: Finalizing Data and Mapping**

In the tenth week, I focused on finalizing data analysis and mapping in QGIS. The resource-intensive nature of QGIS and occasional crashes highlighted the importance of using appropriate software and having backups. I updated trend notebooks to ensure consistency and

explored alternative visualization tools like Colab. This week emphasized the importance of consistency, appropriate tool selection, and the potential of interactive visualizations.

**Week 11: Wrapping Up and Reflecting**

In the final week, I concentrated on wrapping up the project, conducting final tests, and preparing the report. I reflected on the challenges and successes of the project, recognizing the value of the skills acquired, including version control with Git and GitHub. Planning the final journal and academic-style paper provided an opportunity to summarize and share my research. This week highlighted the importance of thorough documentation, reflection, and sharing findings with the broader community.

**Conclusion**

Going through this project was time-consuming but did show me the benefits of challenging oneself and being placed in a situation where one feels “dumb” in. This doesn’t mean that the research is above your head and you don’t get anything out of it, it means you are delving deeper into something that you don’t know well but would like to know more of and it can help you grow as you then can’t always rely on an outside source. Additionally, I came out of this wanting to grow my skills more in the data analytics sphere, along with trying to improve my project in the future by collecting data from more than just star players and also cleaning up my data and making sure everything is good and would be easily readable and usable by another person.