Gabriel (Gavi) Book

Professor Shin

Business Intelligence Capstone

**Individual Contributor Report**

# Overview of Contribution

Throughout the semester, my role within the data analytics segment of our project was comprehensive and multifaceted, embodying the responsibilities typically associated with a lead data scientist. My contributions spanned from the meticulous collection of real estate data to the deployment of advanced data analysis techniques, all tailored to address the evolving objectives of our capstone project.

I embarked on the project by sourcing real estate data from an array of platforms, including government databases, real estate websites, and through APIs designed to deliver real-time estate and geospatial data. This task required not only technical expertise in data extraction methodologies but also a deep understanding of the data’s relevance, which I ensured by establishing rigorous accuracy and completeness checks. These efforts set a solid foundation for our data-driven project, ensuring the subsequent analysis was built on reliable and pertinent data.

To further our objectives, I conducted detailed preliminary data extraction from various APIs, which involved writing and executing complex queries to assess data aggregation potential, integrity, and quality. The goal was to determine how well this data could support the overarching project goals, which necessitated a keen eye for data efficacy evaluation and the ability to foresee potential integration challenges.

Recognizing the importance of market dynamics, I also undertook the intricate task of web scraping from leading retail websites, such as Target and Walmart. This process aimed to acquire critical data on baby products, a task that involved crafting sophisticated scraping scripts to parse and collect information that would provide insights into market trends and product availability—a key consideration for young family demographics in our study.

As the project’s direction evolved, I seamlessly adapted my focus to align with the new objectives. This pivot entailed a deep dive into financial data that was particularly relevant to young families, with an emphasis on child-rearing costs and budgeting. The adaptation did not stop there; I extended my analysis to encompass demographic, housing, and other related lifestyle data, which involved aggregating vast datasets to pinpoint the most suitable neighborhoods in NYC for our target demographic. This transition was executed with meticulous attention to detail, ensuring a smooth integration of varied data streams into our project’s revised framework.

With the data repositories at hand, I then located datasets detailing school quality metrics across NYC. The technical challenge here was to develop a Python script capable of data cleansing and aggregation by zip code. This script was not just a tool for data processing; it was a crucial element in our analysis, providing a refined output that depicted average school proficiency—a vital metric for family-friendly neighborhood assessment.

Building on this, I crafted an initial scoring algorithm based on an array of critical factors such as school data, crime rates, and park availability. This algorithm was pivotal in assessing neighborhood family-friendliness. To ensure the validity of our scoring approach, I implemented and tested various machine learning models, a process that involved a careful balance of statistical theory and practical model training. Furthermore, I created a preliminary Tableau visualization to bring these complex datasets to life, offering a visual representation of the family-friendliness of neighborhoods that were both accessible and insightful.

The project also benefited from my development of a script to retrieve specific amenities data from Foursquare’s Places API. This script, which I meticulously integrated with data on commute times, income levels, median rental costs, after-school programs, noise complaints, and existing crime rates and school quality data, enriched our scoring system, providing a multi-dimensional view of each neighborhood.

Continuing my technical endeavors, I refined the neighborhood scoring script to iron out inconsistencies arising from zip code data—a testament to the detail-oriented nature of my work. This refinement was critical in enhancing the accuracy of our model and by extension, the reliability of our project’s conclusions. My contribution extended beyond data analysis to the realm of communication, where I played a key role in developing a presentation PowerPoint that clearly articulated our analytical findings.

A significant development in my contributions was the scripting work to merge zip codes with other geographic delineations such as Neighborhood Tabulation Areas, Census Districts, Community Districts, and PUMAs. This involved the sophisticated handling of geospatial data to amplify the granularity and precision of our analysis—an endeavor that underscored the technical acumen required to manage complex data types effectively.

When integrating economic burden data into our scoring script, I encountered and resolved issues with faulty CSV outputs through meticulous script revisions. This troubleshooting was part of the ongoing process to enhance the Tableau workbook, where I incorporated additional factors to foster greater user interactivity, thereby providing a more dynamic analytical tool for our end-users.

My involvement with the project report was extensive, focusing on the methodology to ensure clarity and comprehensiveness. As the project neared completion, I further refined the Tableau integration to maximize data visualization and interactivity. This continual revision of both the Tableau workbook and the project report ensured that our findings were not only accurately quantified but also presented in a way that was both engaging and informative for our intended audience.

In conclusion, my technical contributions, as outlined, were integral to the project’s success, reflecting a proactive, leadership-oriented approach to data analysis and visualization. Each step of the way, I ensured that the work performed was in lockstep with our project’s goals, employing a range of technical skills that spanned the entire data science spectrum.

# Updates/Changes/Difficulties

The project demanded a dynamic approach as it evolved, requiring updates to the initial workplan. Initially, I was tasked with extracting financial data for the Young Couples Finance App. However, when the project’s focus expanded to include a comprehensive analysis of neighborhoods suitable for young families, it necessitated a shift in data acquisition and analysis strategies. This transition brought forth challenges that tested my technical adaptability and problem-solving skills.

One of the more significant difficulties I encountered was with data discrepancies, particularly in the geospatial information associated with zip codes. The integrity of the data was paramount for our scoring algorithm, and the inaccuracies posed a risk to the credibility of our analysis. In response, I developed and iterated on Python scripts, enhancing their error-handling capabilities. Through this process, I established more robust data validation protocols, which included the implementation of regular expression checks and the incorporation of geospatial data cross-referencing. These measures greatly improved the consistency of the data and reinforced the integrity of our outputs.

Another issue arose from the integrity of CSV outputs, which are often prone to encoding errors and formatting inconsistencies that can corrupt data processing. To tackle this, I refined the data ingestion and cleansing scripts, incorporating exception handling and encoding normalization. I conducted rigorous testing and validation processes, utilizing both unit tests and integration tests, to ensure that our scripts could handle a variety of data anomalies. These enhancements not only resolved the immediate CSV-related issues but also fortified our data pipeline against potential future discrepancies.

The resolution of these challenges was not a straightforward path; it required a thoughtful application of both established and innovative data-handling techniques. The iterative improvements to our scripts and validation processes ultimately led to a more reliable and accurate dataset, which was crucial for the success of our project. These technical resolutions did not merely address the symptoms of the challenges we faced but also strengthened the foundation of our data processing framework, ensuring the final deliverable was of the highest quality.

# Lessons Learned

Throughout the course of this project, several lessons have been imprinted on my professional outlook. If given the opportunity to revisit the project, I would place an even greater emphasis on the flexibility of data structures from the onset, preparing for potential shifts in project scope. Learning to anticipate and plan for such changes would allow for smoother transitions and less time spent reworking existing frameworks.

A pivotal lesson was the power of visualization in conveying complex data narratives. Prior to this project, I had no experience with Tableau. It was both exciting and enlightening to discover the capabilities of this tool, which allowed me to bring our data to life in a compelling and easily digestible format. The ability to visualize our findings added a new dimension to our analysis and facilitated a more impactful presentation of our results.

Working with geospatial data was another fresh and enriching experience. Understanding the nuances of this data type and learning to integrate it effectively into our analysis provided me with a new perspective on spatial analysis and its potential applications.

One of the more nuanced lessons this project imparted was the complexity of team dynamics in a data-driven environment. I learned firsthand the challenges and intricacies of collaborating within a diverse team, where aligning individual efforts to a unified analytical goal required clear communication, patience, and a willingness to adapt to each other's working styles and expertise. This aspect of the project was a profound exercise in teamwork and collaboration, which is as critical to the success of a project as the technical skills each member brings to the table.

Some scripts from initial project plans were ultimately not used because those project trajectories were not pursued. Additionally, certain exploratory data analysis (EDA) scripts, which included heatmaps, correlation matrices, and other matplotlib functionalities, were set aside. These scripts, while informative, offered a limited view of the factors we were considering and did not align with our revised methodology, which focused on a scoring system for neighborhoods.

# Technical Self-Assessment

The project has significantly advanced my technical skill set. I have developed a deeper proficiency in Python, particularly in web scraping and data cleansing, which I foresee being a cornerstone in my future endeavors as a data professional.

Learning Tableau was a highlight, as it expanded my toolkit for data visualization. The integration of geospatial data analysis techniques has also been an invaluable addition to my skill set. These skills not only enriched this project but will also be of tremendous utility in my future career.

Another realization was the intricacy involved in data collection and cleansing. In contrast to previous experiences with pre-curated datasets, this project required sourcing data from multiple origins, each with its own set of challenges. The process was far more demanding than anticipated, involving a complex and time-consuming effort to ensure data quality and usability. This aspect of data work was both surprising and challenging, underscoring the importance of meticulous data preparation in the analytics workflow.

Moreover, the project has expanded my understanding of the data analytics field. It has highlighted the importance of contextual data analysis and the role of a data analyst in bridging the gap between data and decision-making. The course has reinforced my belief in the power of data and analytics, and I am surprised by the extent to which data can influence strategic decisions in real-world scenarios.

The field of data analytics, as experienced through this course, lived up to and, in many ways, exceeded the surrounding hype. The practical challenges of data collection and cleansing brought to light the reality that data work is as much about perseverance and attention to detail as it is about technical prowess. This experience has not only fortified my technical capabilities but also my resolve to pursue a data-centric career path with a newfound appreciation for the foundational work that enables powerful analytics.

# Code and Final Report

The technical codebase I have built and refined throughout the course is a testament to the analytical rigor and technical acumen that I brought to the project. All scripts, including those for data extraction, cleansing, scoring, and visualization, have been documented and are available in the project’s [Google Drive folder](https://drive.google.com/drive/folders/1giNu8mRjvwORIAflQ0CZso82A4LoNPHU?usp=sharing). In this separate Google Drive folder, I will include every script I wrote, even the unused ones, alongside all datasets, CSV outputs generated from my scripts, and the Tableau workbooks I developed, to thoroughly document and highlight the extensiveness of my individual work on this project.

My contributions, particularly the neighborhood scoring system and the dynamic Tableau dashboards, were pivotal in the final report. These tools not only provided the team with a robust analytical framework but also equipped the end-users with interactive and informative insights into the family-friendliness of neighborhoods across NYC.

The Tableau dashboards, in particular, became a highlight of our project, enabling stakeholders to engage with the data in a more meaningful way. As such, my work was integral to the final report, reflecting the depth and application of the technical skills I have honed throughout the course.

# Conclusion

In closing, the capstone project was a profound learning experience that extended beyond technical skill enhancement to include valuable insights into the intricacies of collaborative data analysis. Navigating the challenges of data collection, cleansing, and visualization—particularly with tools and data types new to me—has not only equipped me with practical skills but also a deeper understanding of the analytical process. The lessons learned from the complexities of teamwork will be particularly invaluable as I move forward in my career. With a reinforced appreciation for the discipline of data science, I am eager to apply these competencies and insights to future endeavors, confident in the knowledge and experience gained from this comprehensive project.

As this capstone project comes to a close, I extend my deepest appreciation to our professor, whose guidance was pivotal throughout this journey. Her expertise pushed us to challenge ourselves beyond the comfort zones of our pre-existing skill sets, encouraging exploration and innovation in equal measure. It was through her encouragement that we were able to navigate the complexities of real-world data analysis with confidence. This experience has been invaluable, not only in the technical skills and collaborative approaches it has instilled but also in fostering an analytical mindset that will undoubtedly be an asset in all future endeavors. For this, and for the opportunity to undertake such a comprehensive project, I am profoundly grateful.