Portfolio Milestone Project

Allison R. Deming

SUID: 394248454

Syracuse University

Master of Science in Applied Data Science

Spring 2024

(https://github.com/ardeming/MSADS-Portfolio-Project-2)

(https://youtu.be/2RHq1Klc2Eg)

Portfolio Milestone Project

Table of Contents

Introduction

Data science, an emergent discipline, merges insights from three foundational domains: statistics, computer science, and individual areas of expertise. This expertise can span diverse fields, showcasing the synthesis of statistical methodologies and computational techniques to augment knowledge within that domain. For instance, my background in cognitive experimental psychology, though somewhat unconventional in data science, has been instrumental in shaping my trajectory. After a fulfilling yet challenging tenure as a Psychology Professor spanning two decades, I embarked on a career transition driven by a profound interest in data science. This shift allowed me to reinvigorate the statistical acumen I cultivated during my Ph.D. program while also mastering advanced programming languages beyond my proficiency with Excel and SPSS.

My professional endeavors currently revolve around a Higher Education Resource Services (HRSA) grant and student services, ventures seemingly distant from data science. Nevertheless, data analysis does remain a cornerstone of my responsibilities with the grant, and I firmly believe that integrating data science methodologies into student services can yield substantial benefits. Post-grant conclusion and degree attainment, my aspirations converge on securing a role that seamlessly merges my newfound data science prowess with the pedagogical expertise accrued during my academic tenure, fostering student success through data-informed initiatives.

My journey led me to choose the Syracuse Master of Applied Data Science program offered by Syracuse, drawn by its comprehensive curriculum tailored to equip students with what they need for the workforce. The program's objectives encompass a spectrum of proficiencies, including but not limited to:

- Providing a comprehensive overview of key *practice domains* in data science.

- Proficient *data collection* and organization techniques.

- Discerning patterns through data visualization, *statistical analysis*, and data mining.

- Crafting *alternative strategies* grounded in data insights.

- Formulating actionable plans to actualize *business decisions* derived from analyses.

- Effectively *communicating* data insights to diverse stakeholders within organizations.

- *Ethically* navigating the complexities of data science practice.

This portfolio is a testament to the culmination of skills honed over the past two years within the iSchool's data science program. Spanning diverse projects encompassing programming tasks in R and Python, and exploring ethical dimensions through traditional papers, these efforts encapsulate my unique journey and skillset, representative of my experience within the program.

Project Descriptions

There were many courses that I could have chosen for this portfolio. Narrowing it down was quite hard and was a learning experience in and of itself. I picked five classes I felt would best show what I have learned over the past two years. Focusing on some projects more than others.. They also showcase various means and methods, programming languages, and writing skills. The first course that I picked was IST 618. This course in information policy is more of a non-applied course with papers and presentations, along with one debate that helped me grow my communication skills. I also learned about many ethical dilemmas in data analysis, and how those affect information policy. The first paper was on access and affordability; the second paper was on “Big Tech,” the third paper was on the right to repair one’s technological devices such as cell phones, and the last paper was on cell phone tracking technology along with other GPS tracking technologies and the ethics behind them.

The subsequent course I chose was IST 652, an introductory Python course. One notable aspect of this class was a group project centered on analyzing Airbnb data, where we explored three distinct hypotheses. Unlike other courses this class did not have a formal presentation or write up of the data. Everything was contained in the Jupiter notebook file. Moreover, we incorporated an additional data set sourced from the city of New York (NYC), which provided crime statistics, enabling us to conduct a comprehensive analysis. By joining these datasets, we examined various aspects, including the frequency of terms used in listing descriptions, pricing trends, and the correlation with crime statistics.

(It's worth mentioning that this portfolio features two distinct projects utilizing Airbnb data. The richness and versatility of this dataset allowed for its utilization across multiple projects without redundancy in analysis or any code overlap. Prior clearance from instructors confirmed that the reuse of the same topic was permissible if the datasets differed in the time frame and analytical approach. While this may limit the breadth of showcased variety, the consideration of future course requirements was not immediately apparent during enrollment. Notably, the first project (IST 652) entailed a concise Python analysis and concluded with a comprehensive, multifaceted analysis executed in R (IST 707).)

My next selection for representation from my tenure here was IST 687, Introduction to Data Science. This course delved into data exploration and analysis utilizing the R programming language. For this project the agile method was used along with a Kanban board always to track the project's status. This was a new process for me, which I found very enlightening. Our project focused on the pivotal task of forecasting school placement, homing in on the factors that influence employment prospects for junior, senior high school, and pre-graduate students. Leveraging the wealth of data available within the student database, we embarked on an investigative journey to unearth post-graduation job placement status determinants. This effort facilitated a deeper understanding of data science methodologies and underscored the transformative potential of data-driven decision-making in shaping educational outcomes.

IST 707 presented a formidable challenge for my fourth endeavor compared to its predecessors. In this Data Analytics/Machine Learning course, we explored the Airbnb dataset comprehensively, employing the R programming language as our tool of choice. Unlike previous projects, the objectives here were distinctly business-oriented, exploratory, and predictive elevating our analysis's complexity.

Our project aimed to uncover intricate relationships within the dataset, spanning five key areas of inquiry. Initially, we sought to discern the impact of various amenities on the pricing of Airbnb listings. Subsequently, we analyzed the correlation between review sentiments, proximity to major tourist attractions, and Airbnb pricing, necessitating minor Python coding and integrating Google API for geographical insights.

Furthermore, we investigated the seasonal variability in Airbnb pricing and reviewed frequency in NYC, probing into the influence of time on these metrics. Lastly, we endeavored to identify the geographical distribution of affordable Airbnb listings across the five boroughs of NYC, alongside predicting their availability across different seasons.

This multifaceted project not only stretched our analytical capabilities but also underscored the practical relevance of analytics and machine learning in informing strategic business decisions and forecasting trends within dynamic markets like the hospitality industry. For our analysis, we used decision trees, K-means clustering, Sentiment analysis, and multiple other basic methods for visualization of the big and small data.

For my closing project, I opted for IST 718, delving into Big Data Analytics. Rather than centering my focus solely on the larger group project at the course's conclusion, I was drawn to the first homework assignment, which entailed an intriguing analysis: assessing the potential earnings of the Syracuse Football Coach if they were to coach in alternative conferences, specifically the SEC or all other conferences combined.

This endeavor demanded the union of multiple datasets from various online repositories, facilitating a comprehensive comparison of recent compensation packages offered to college football coaches across different leagues, factoring in similar levels of experience and tenure. Subsequently, the analysis expanded to encompass a broader examination, probing into metrics such as win rates, participation in bowl games, and academic success of student athletes.

By meticulously scrutinizing graduation rates among football program students, we sought to glean insights into the dynamics influencing coaching salaries and performance metrics across diverse collegiate football landscapes. This nuanced exploration underscored the intricacies of data analytics and shed light on the multifaceted considerations shaping the collegiate athletics ecosystem.

Collecting Data

Ensuring meticulousness in data collection is paramount. Merely because a website appears reputable does not guarantee the veracity of its data—a lesson I swiftly learned in my professional journey. While engaging in data science activities, the emphasis lies not solely on sourcing data but also on leveraging it effectively. In my projects, I relied on two primary repositories. The first is Kaggle (IST 687, IST 707, IST 718), a thriving open-source platform (https://www.kaggle.com/), while the second is the Airbnb data science homepage (http://insideairbnb.com/get-the-data/), graciously furnished by the company for exploration and project completion. Given its regular updates, the data undergoes dynamic shifts over time, resulting in nuanced differences between successive projects (IST 652, IST 707).

It should be noted that my SQL class in Database and Database Management was the best example of storing, creating, and organizing data in a database. Unfortunately, the way that class was taught using the remote desktops for all parts of the project including grading there was no way to use it in the portfolio. I just wanted to acknowledge that it would have been a better choice.

Analyzing Data

In the ever-evolving landscape of data science, the art of data analysis serves as a cornerstone, empowering professionals to extract valuable insights from vast and complex datasets. Data analysis, a pivotal component within the realm of data science, encompasses a multifaceted approach to examining and interpreting data to uncover meaningful patterns, trends, and relationships. As the foundation upon which informed decision-making and strategic planning are built, data analysis plays a pivotal role in driving innovation, optimizing processes, and unlocking actionable intelligence across various domains. From exploratory data analysis to predictive modeling and beyond, mastering the techniques and methodologies of data analysis equips practitioners with the tools to unravel the mysteries hidden within data, ultimately driving transformative outcomes in the dynamic field of data science.

Amongst all the projects showcasing extensive data analysis, it is my contention that the Airbnb analysis project conducted in R as part of IST 707 stands out as the most comprehensive. This project spans from elementary descriptive analytics to the intricate implementation of diverse machine learning models. Notably, it forecasts the impact of seasonal variations on sentiment and review frequency, while also predicting the availability of specific Airbnb listings throughout the year. Furthermore, through meticulous data cleaning and straightforward analysis, we elucidated the significant contribution of amenities to the pricing dynamics of Airbnb listings in NYC. Delving deeper, we investigated the prevalence of affordable Airbnb options across NYC boroughs and explored how proximity to tourist attractions influences both pricing and sentiment in reviews, thereby enriching our understanding of the Airbnb ecosystem in this bustling metropolis.

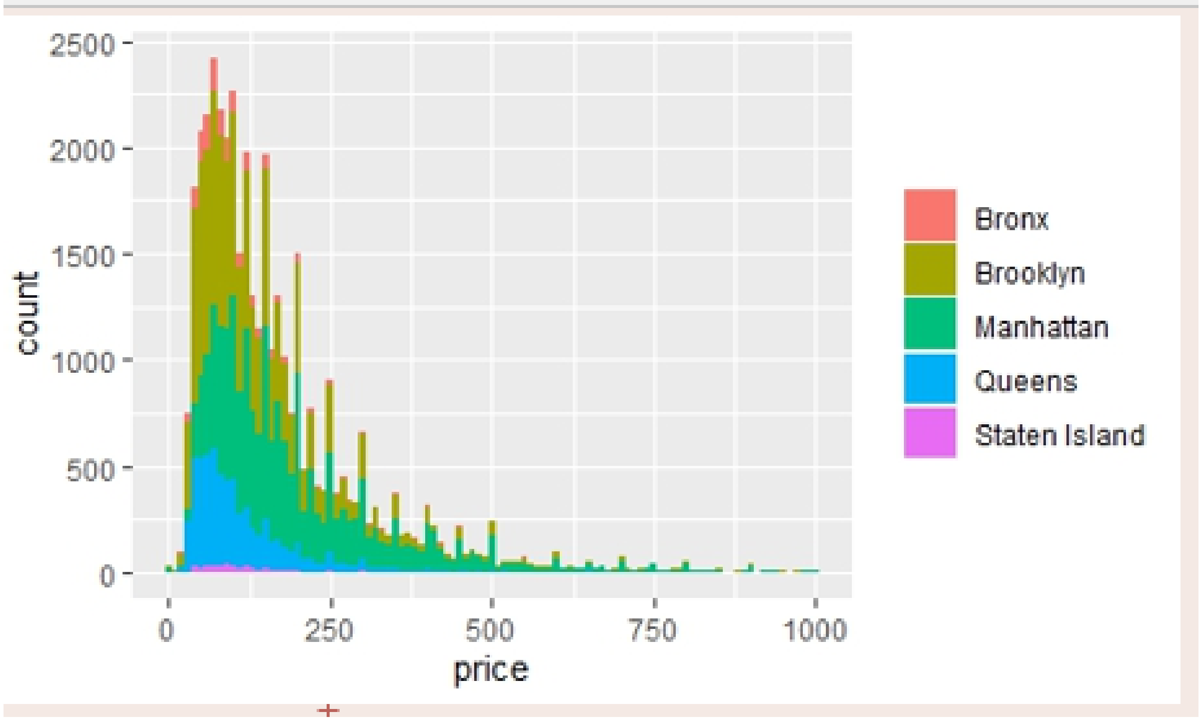


Figure 1: Price Visualizations

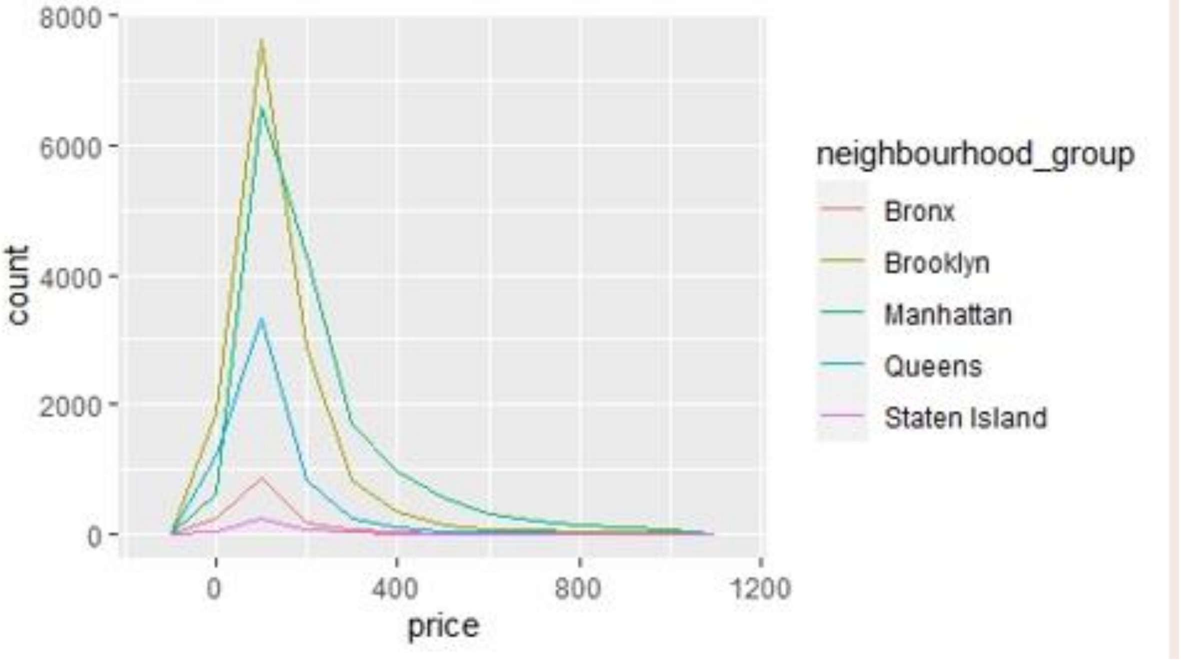


Figure 2: Price Visualizations

Figures 1 and 2 offer straightforward visual representations, illuminating the intricate tapestry of NYC neighborhoods juxtaposed with Airbnb pricing dynamics and the number of available listings. However, Figure 3 delves deeper into the dataset, presenting a nuanced analysis of spatial patterns. Initially, the data underwent discretization into distinct price groupings, ranging from very low to very high. These groupings were then mapped across the latitude and longitude coordinates of each borough, showing compelling insights. Notably, the visualization vividly illustrates a concentration of high and very high-priced listings primarily clustered within Manhattan, particularly prominent in the southern tip, underscoring the borough's premium real estate. Conversely, a notable prevalence of low and very low-priced listings emerges in the Bronx and Staten Island, painting a picture of affordability across the city's landscape.

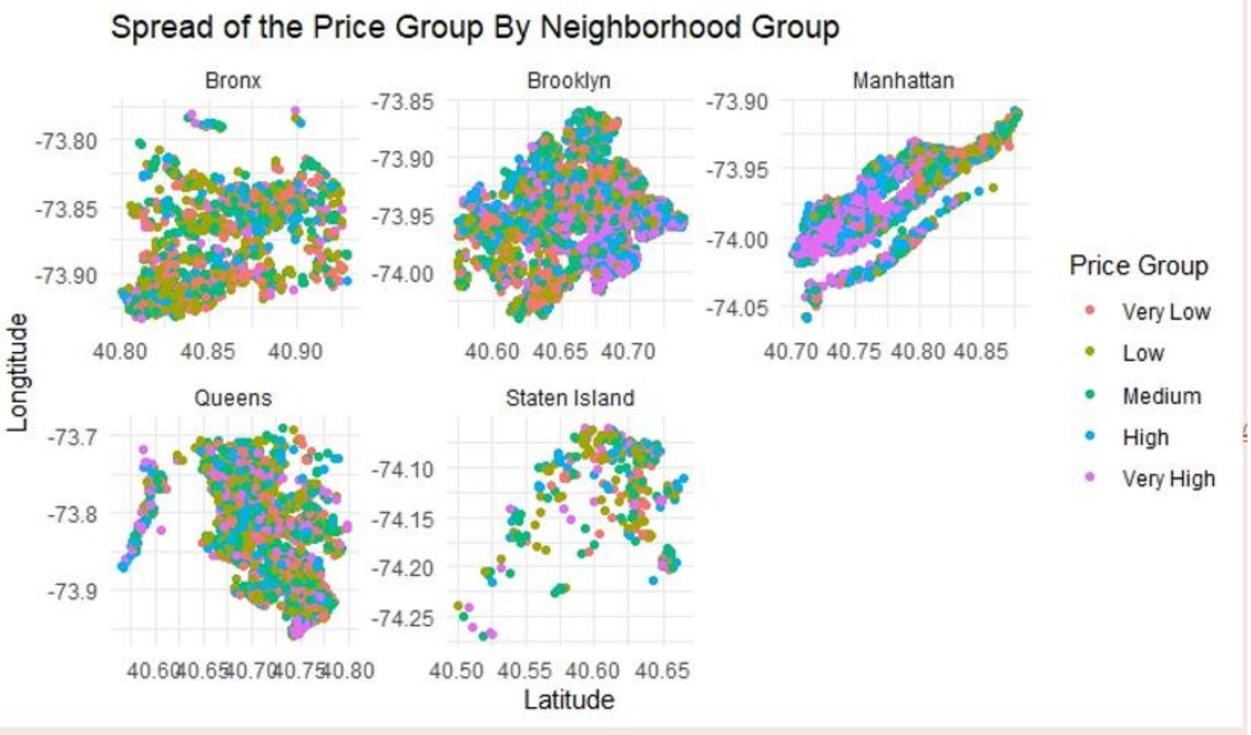


Figure 3: Spread of the Price Grouping by Borough

Next, we move on to our models. The first model was done by a K-Means Analysis. K-Means Analysis stands as a foundational technique in the toolkit of data scientists, offering a powerful method for clustering data points into distinct groups based on similarity. At its core, K-Means aims to partition a dataset into K clusters, where each observation is assigned to the cluster with the nearest mean, or centroid, thereby optimizing intra-cluster similarity while maximizing inter-cluster dissimilarity. With its versatility and scalability, K-Means finds application across various domains within data science, serving purposes ranging from exploratory data analysis and pattern recognition to customer segmentation and anomaly detection. By unraveling hidden structures within datasets and facilitating the identification of coherent groups, K-Means empowers data scientists to gain deeper insights, make informed decisions, and extract actionable intelligence from complex data, thus playing a pivotal role in driving innovation and driving business success.

Through the utilization of K-Means Analysis, we unearthed two significant insights. Firstly, the presence of numerous outliers within the dataset became apparent, necessitating their mitigation. Secondly, our analysis revealed the existence of three distinct clusters within the pricing data: Expensive, Average, and Cheap. This classification neatly delineated the varying price points across the boroughs, as depicted in Figure 4. Furthermore, Figure 5 provides a graphical representation of the distribution of these categories across the boroughs.

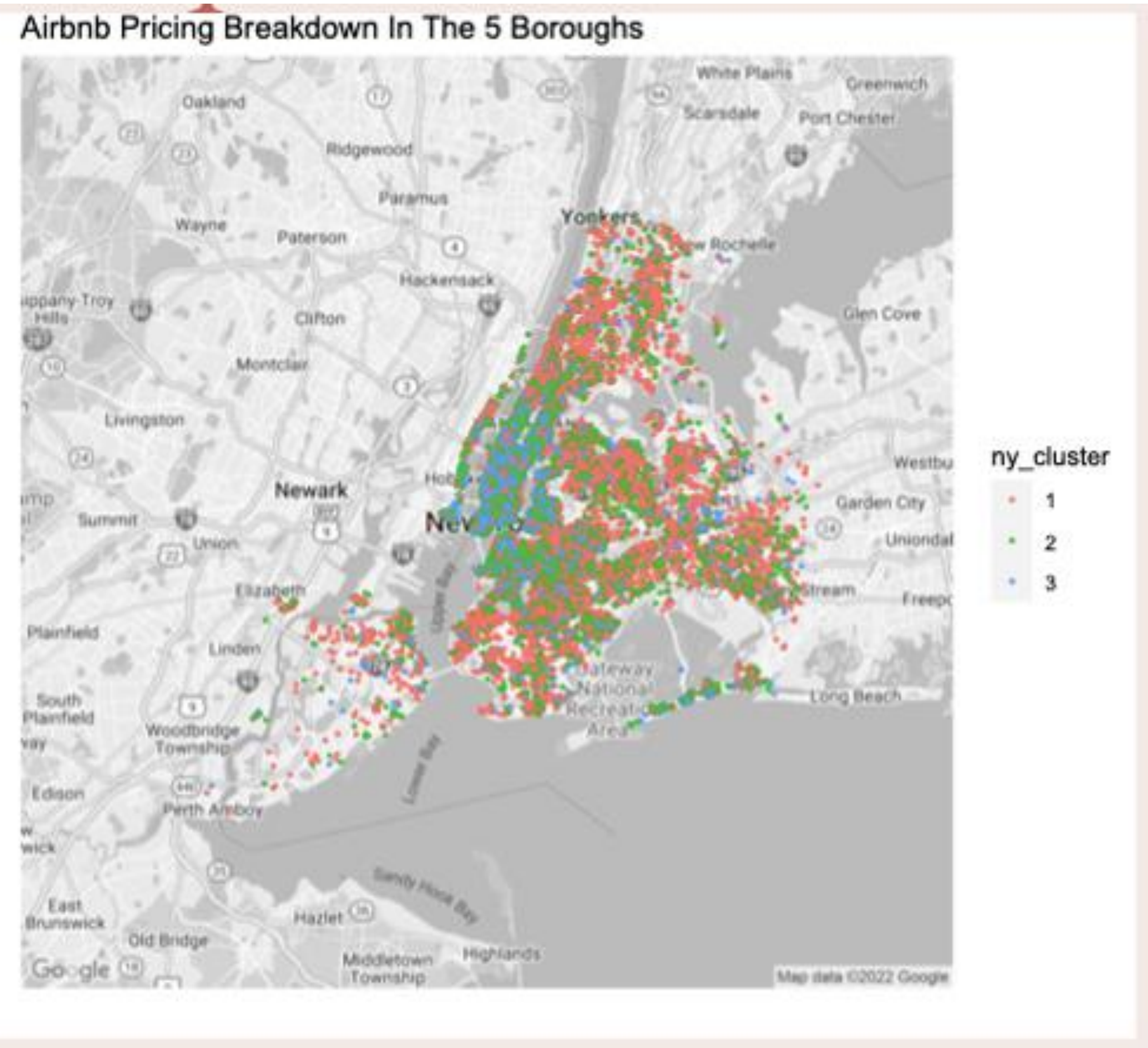


Figure 4: K-Means Cluster Analysis Across the Five Boroughs of Price

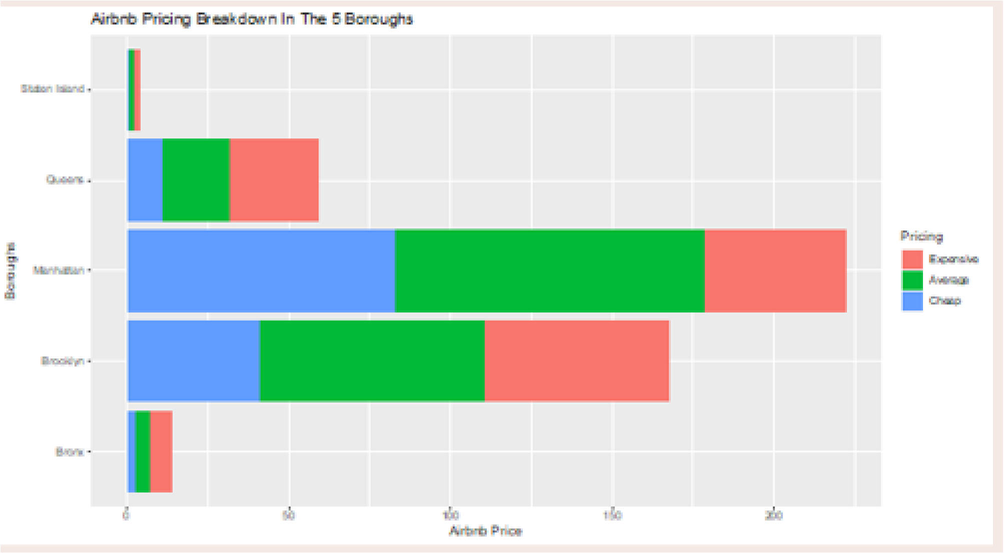


Figure 5: Airbnb Pricing Data in the Five Boroughs

In the realm of data science, the significance of data analysis cannot be overstated, serving as a cornerstone for extracting valuable insights from intricate datasets. From exploratory data analysis to predictive modeling, mastering data analysis techniques empowers professionals to unravel patterns, trends, and relationships, driving innovation and informed decision-making across various domains. One standout project within this landscape is the Airbnb analysis conducted in R as part of IST 707, showcasing a comprehensive exploration ranging from descriptive analytics to sophisticated machine learning models. Through meticulous data cleaning and analysis, the project shed light on the impact of amenities on Airbnb pricing dynamics in NYC and the prevalence of affordable options across boroughs. Additionally, spatial patterns were unveiled, illustrating a concentration of high-priced listings in Manhattan and contrasting with lower-priced options in the Bronx and Staten Island. Furthermore, employing K-Means Analysis revealed critical insights into the dataset, identifying outliers and delineating three distinct pricing clusters—Expensive, Average, and Cheap. These findings, depicted in Figures 4 and 5, underscore the efficacy of K-Means in uncovering natural groupings within data, offering actionable intelligence for strategic decision-making in the field of data science.

Alternative Strategies with Data

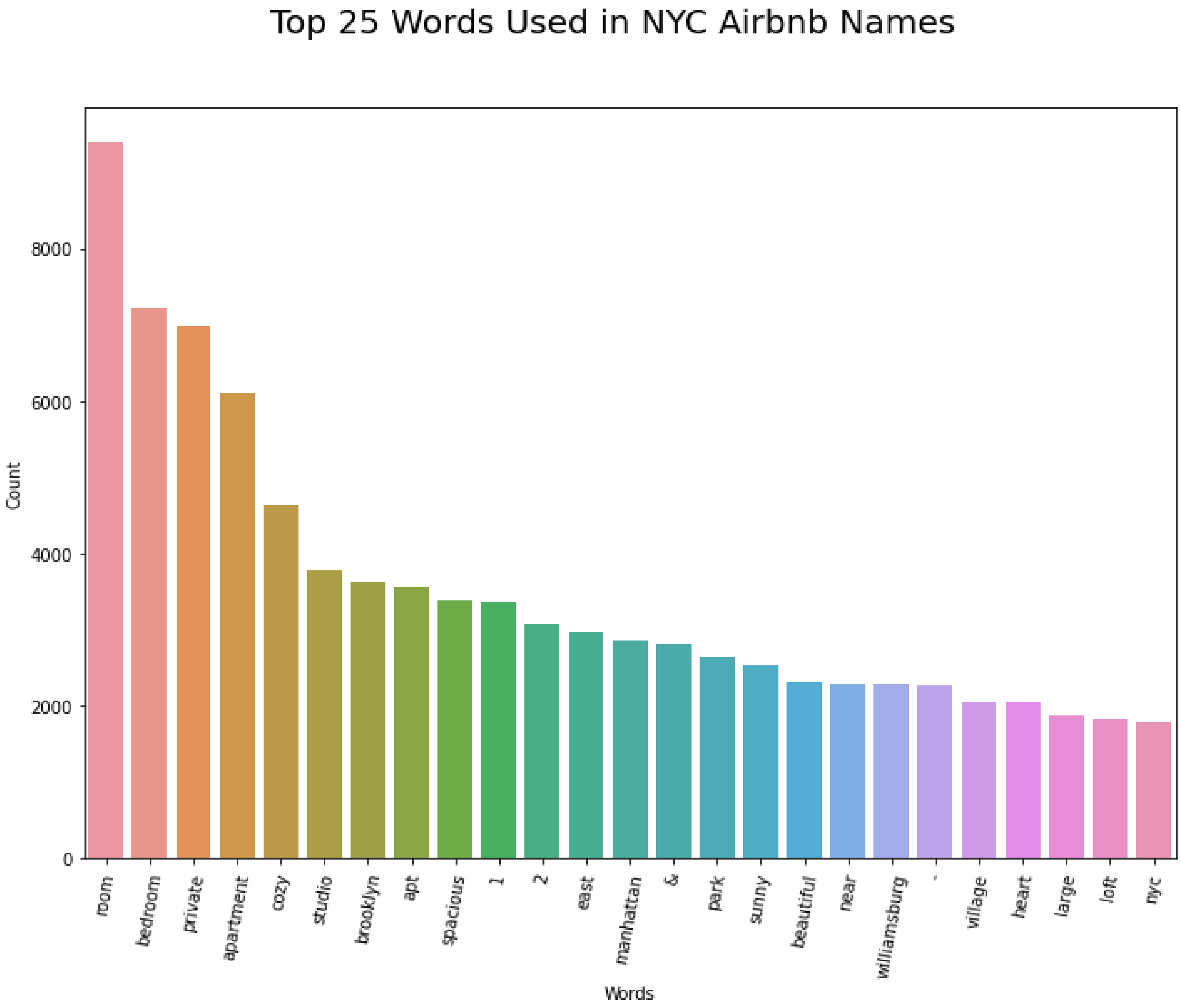


Figure 6: Airbnb Top 25 Words Used in Names

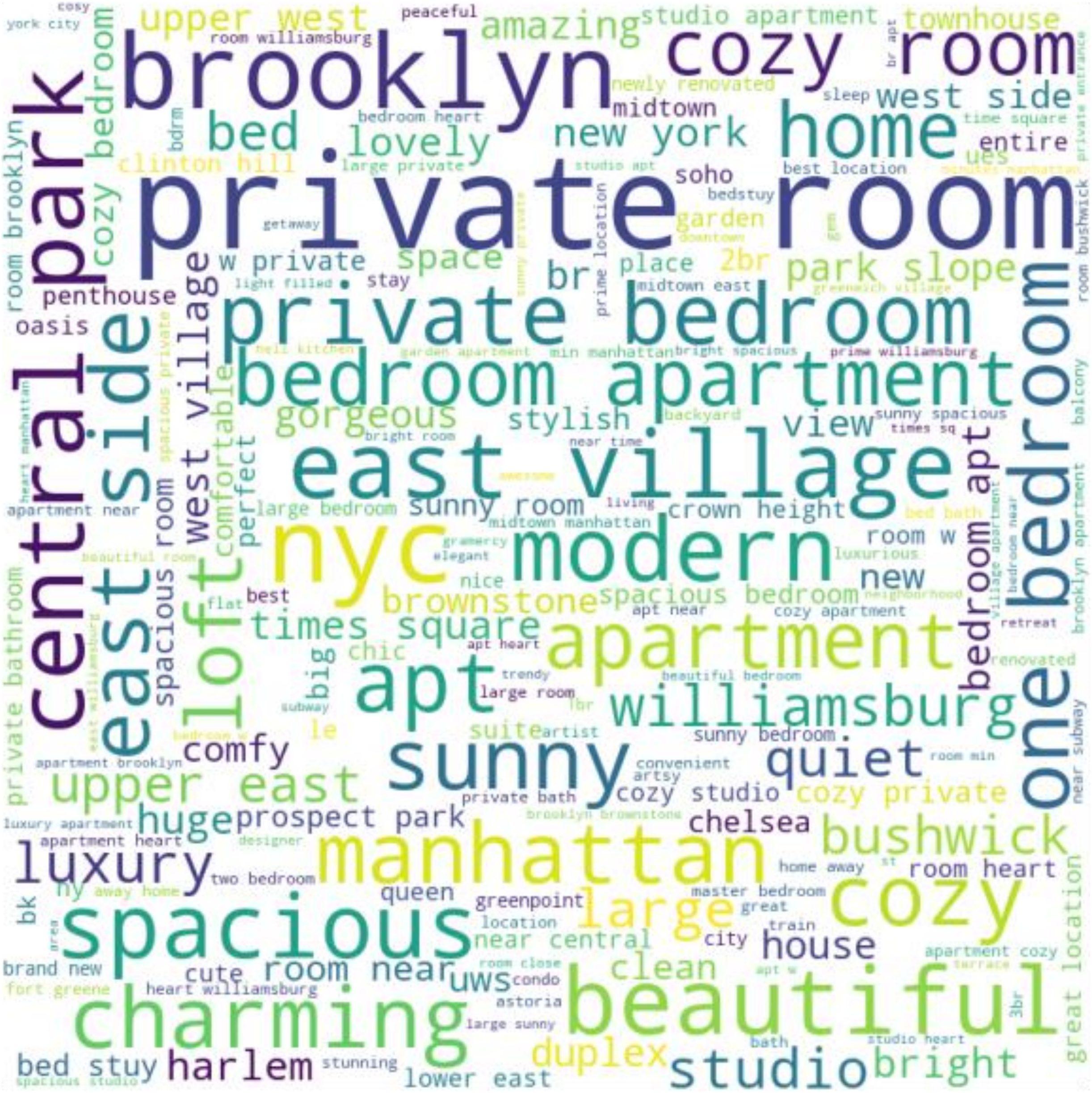


Figure 7: Word Cloud Representing Reviews of the Properties on Airbnb

Above are two visual representations, each showcasing insights derived from the Airbnb Dataset utilized in Python (IST 652). Our exploration delved into the intricacies of handling unconventional data, specifically textual information encompassing reviews and Airbnb location names. Initially, we opted for a traditional bar graph presentation. However, we quickly recognized its limitations in effectively conveying the nuances of the data. Bar graphs, with their linear X-axis, struggled to establish meaningful associations between the data points and their corresponding labels.

In response, we ventured into a less conventional yet more insightful avenue: employing a word cloud for our subsequent analysis. Figure 2 vividly captures the essence of each review by highlighting the prevalence and significance of individual words. Unlike the bar graph, where the prominence of words is represented by their frequency, the word cloud offers a visually compelling depiction. Each word's size reflects its frequency, offering a more intuitive and artistic interpretation of the data.

This innovative approach not only enhances comprehension but also elevates the aesthetic appeal of the presentation. While I have explored various alternative strategies in my analytical endeavors, the utilization of word clouds stands out as a remarkably effective and visually captivating method.

Business Insights

In today's data-driven business landscape, the role of data analysis in shaping strategic decisions has become paramount. Data analysis serves as the compass guiding organizations through the sea of information, providing invaluable insights that drive informed decision-making and steer businesses towards success. By harnessing the power of advanced analytical techniques and tools, businesses can sift through vast volumes of data to uncover trends, patterns, and correlations that would otherwise remain obscured. From market trends and consumer behavior to operational efficiencies and risk management, data analysis offers a lens through which executives and stakeholders can gain a deeper understanding of their business landscape. Armed with these insights, organizations can make evidence-based decisions, mitigate risks, capitalize on opportunities, and ultimately gain a competitive edge in today's dynamic marketplace. Thus, data analysis not only serves as a catalyst for innovation but also as a corner stone in the strategic decision-making process, driving businesses towards sustainable growth and prosperity.

The project I wish to highlight was completed using R for my original data science course, IST 687. Unlike previous analyses, this project navigated complexities primarily through a multiple regression problem and simpler analytics. Leveraging my extensive background in statistics, coupled with my experience in academia, rendered this project particularly meaningful. The dataset used hailed from an Indian school system and encompassed job placement records, boasting substantial volume (over 21,000 student observations). Interestingly, despite the magnitude of data, statistical significance was attained, albeit devoid of true variance accounted for—a caveat often encountered when grappling with large datasets. This project's essence lay in scrutinizing the student database of organizations to discern the myriad factors influencing post-graduation job placement. This pursuit resonated deeply, given its relevance to the current job in student services: predicting the success trajectories of diverse student cohorts. From anticipating collegiate achievements based on backgrounds to forecasting post-collegiate employability, this analysis delved into pivotal questions of education and career trajectories.

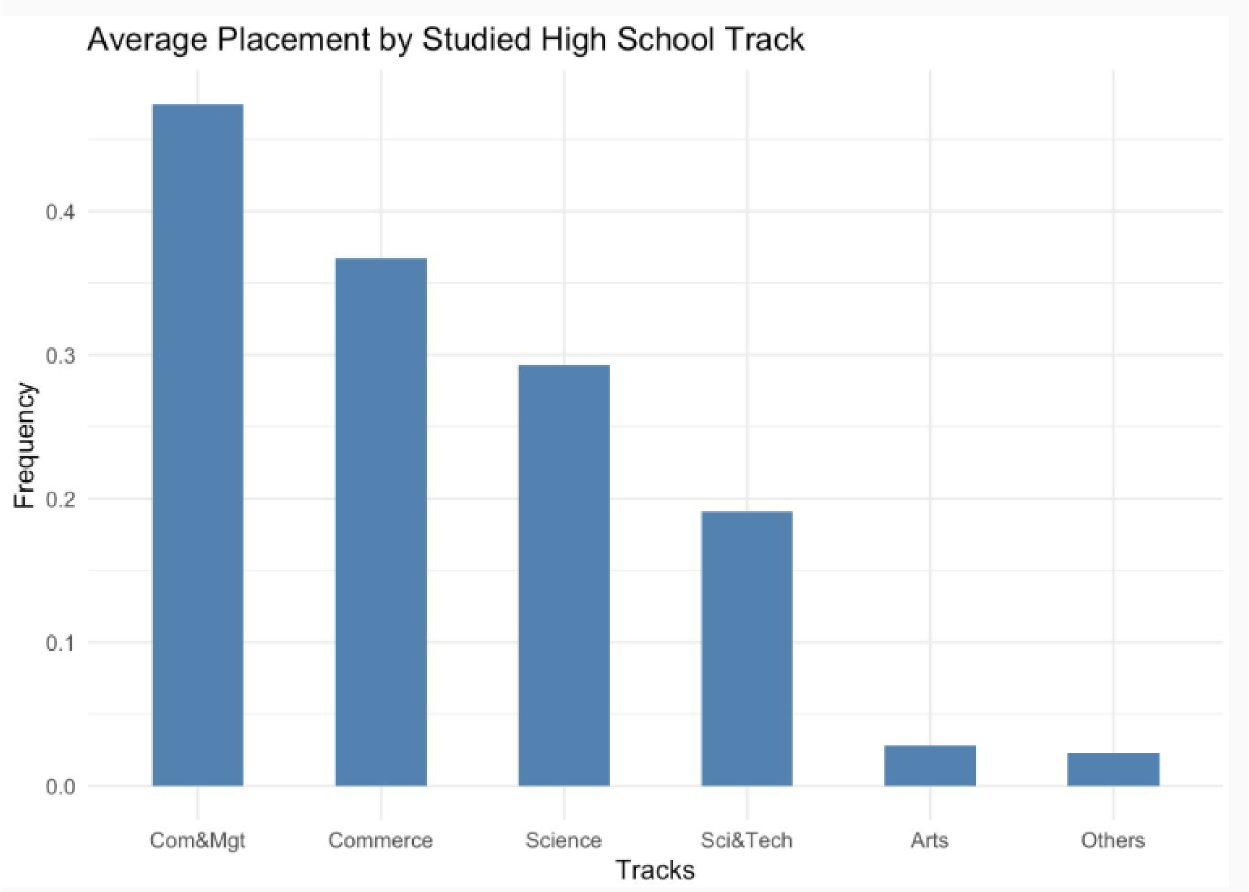


Figure 8: Average Placement by Studied High School Track

Similar to high schools in the United States, the educational landscape in India offers distinct tracks catering to diverse student interests and aspirations. These tracks encompass Communication and Management, Commerce, Science, Science and Technology, Arts, and Others. Among these, Communication and Management attract the largest student cohort, followed by Commerce and Science, with Science and Technology trailing closely behind. Arts represents the smallest percentage of the student population, followed by Others, both constituting minor segments within the educational framework. Figure 8 visually encapsulates the distribution of students across these tracks within the dataset, providing a clear insight into the relative prevalence of each educational pathway.

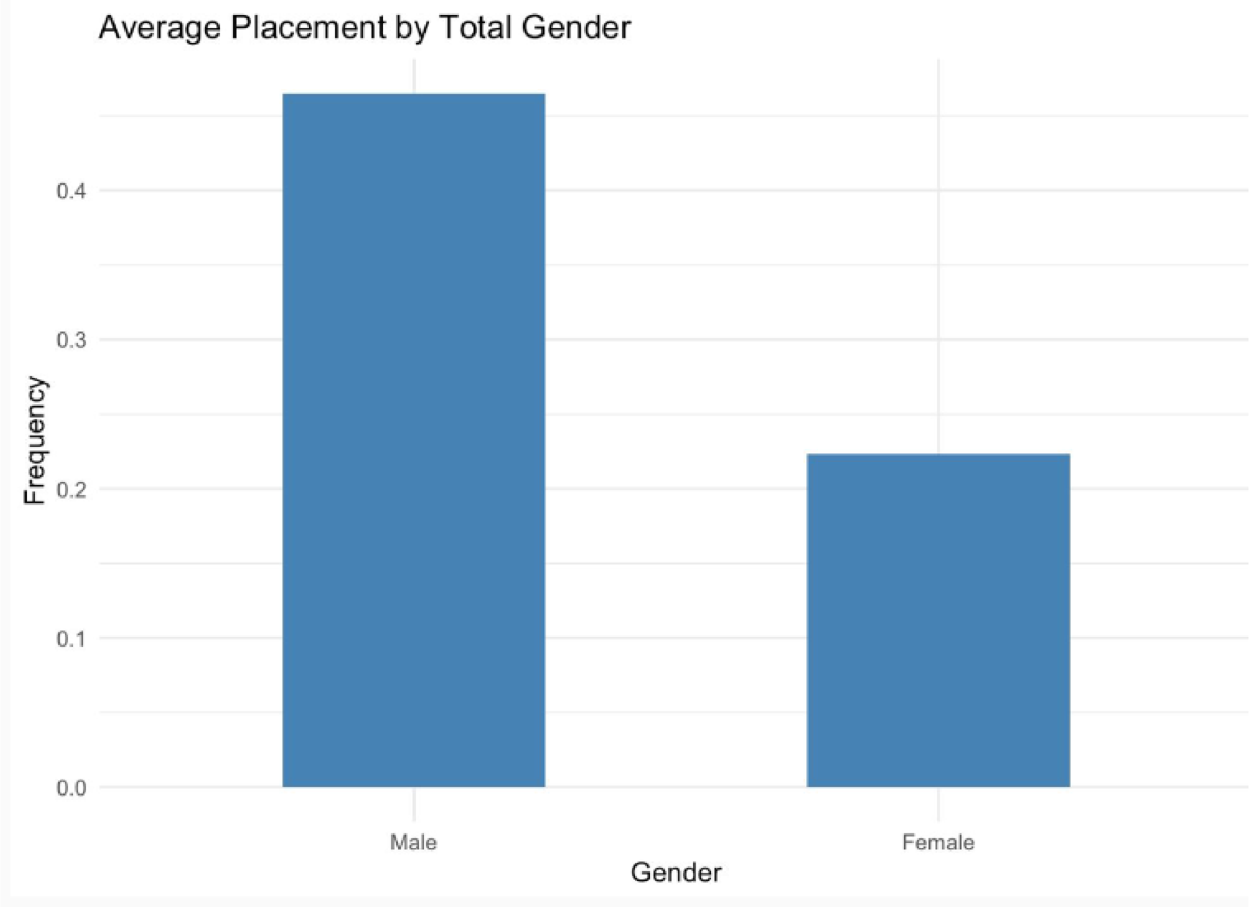


Figure 9: Average Placement by Total Gender

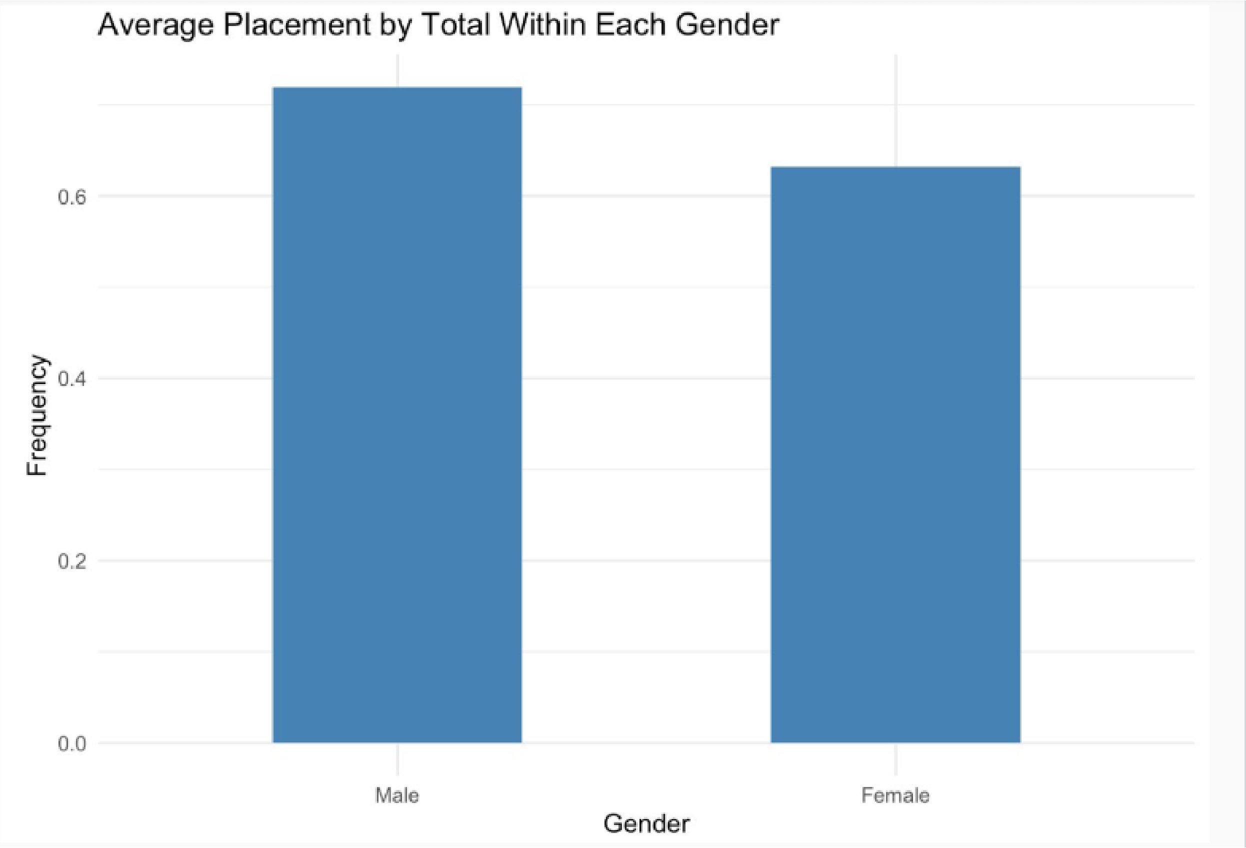


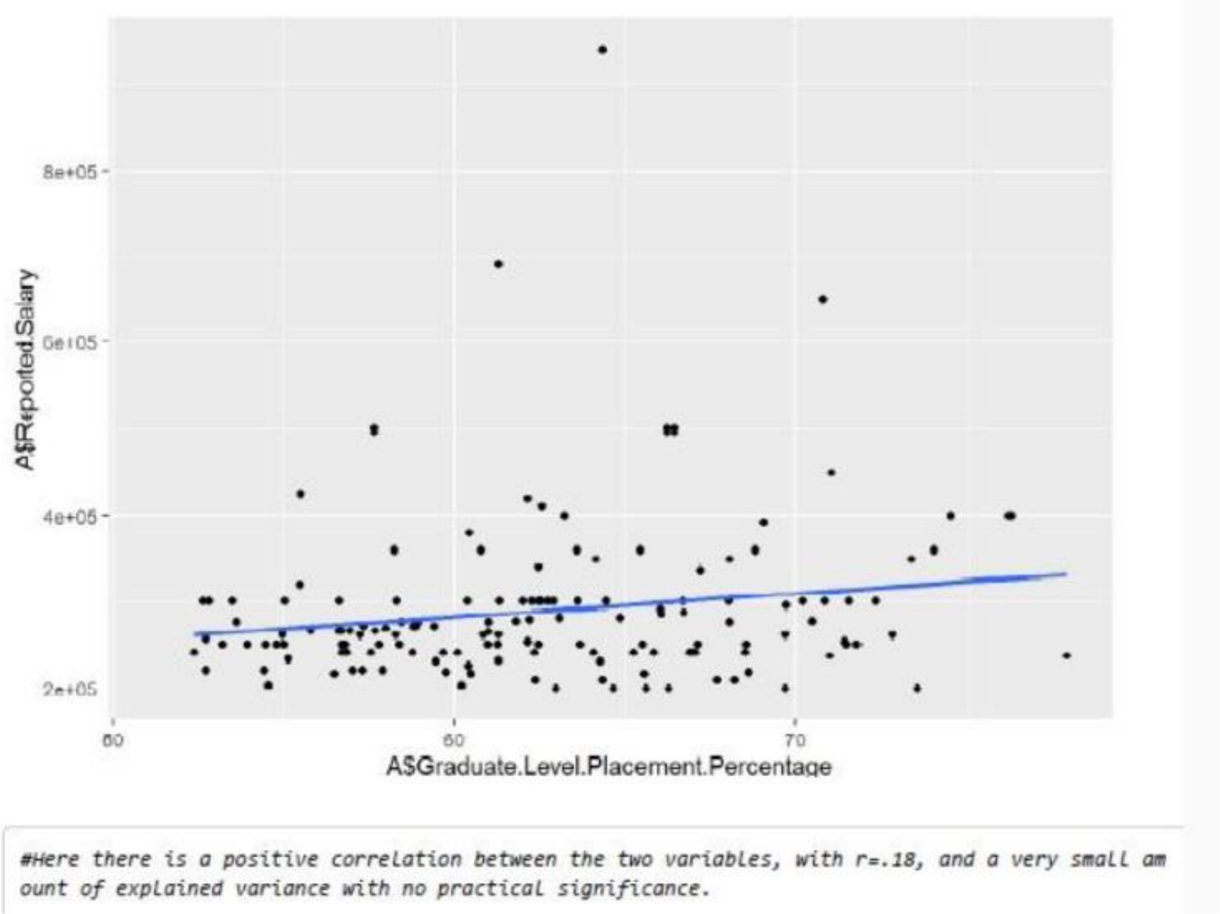
Figure 10: Average Placement by Total Within Each Gender (Proportionate)

Initially, the dataset was segregated by gender, revealing an apparent bias that raised concerns regarding its representativeness. Upon initial examination of the raw data, it became evident that the distribution of job placements was skewed, with only 22% of female students securing placements compared to 47% of male students. This glaring disparity prompted a deeper dive into the dataset, uncovering a substantial discrepancy in the number of female (approximately 7,500) versus male (approximately 13,500) students. Consequently, the dataset was heavily influenced by these imbalanced numbers. In response, efforts were made to rectify this imbalance, leading to the creation of Figure 10, which illustrates the average placement proportionate to the total number of students within each gender category. Upon closer inspection, it became apparent that when viewed proportionately, both female and male students achieved job placements at nearly equal rates, with males enjoying a slight advantage.

This observation hints at the possibility of enduring cultural biases, which might persistently influence the disparity between the number of working female and male students in contemporary society.

In reflecting on this dataset from an educational institution within the United States, it would likely prompt recommendations to increase the recruitment of female students. Notably, our analysis reveals that female students possess equal potential to secure jobs compared to their male counterparts, highlighting an untapped demographic within India. A cursory examination unveils parallels with historical data from the United States, where similar trends emerged following the integration of women into colleges. Despite this milestone, disparities persisted, with women facing hurdles in job placement and encountering pay differentials compared to men. Regrettably, these inequities endure to this day in certain sectors, underscoring the enduring struggle for gender equality. Given this historical trajectory in the United States, it is plausible that India may witness a similar evolution, necessitating proactive measures to address entrenched biases and promote gender parity in educational and professional spheres.

Like the United States, India administers placement tests to all graduates, and this data forms a crucial component of our dataset. Utilizing this information, we sought to predict various outcomes, such as reported salaries, as depicted in Figure 11. Surprisingly, we observed that the graduate-level placement percentage, akin to the GRE or Graduate Record Examination, exhibited minimal correlation with reported salaries. Despite initial expectations of a strong association, the correlation was merely positive, with an R value of 0.18 and negligible variance explained, indicating a lack of practical significance. Consequently, relying solely on graduate examination scores to forecast job placement success and salary prospects may not be advisable. However, other paired tests revealed positive correlations, which is unsurprising given that general knowledge or intelligence is a common trait among all students. Figures 12 and 13 offer insights into these correlations, underscoring the nature of predictors influencing job placement and salary outcomes.



```
#Here there is a positive correlation between the two variables, with r=.18, and a very small am
ount of explained variance with no practical significance.
```

Figure 12: Test Scores by Reported Salary

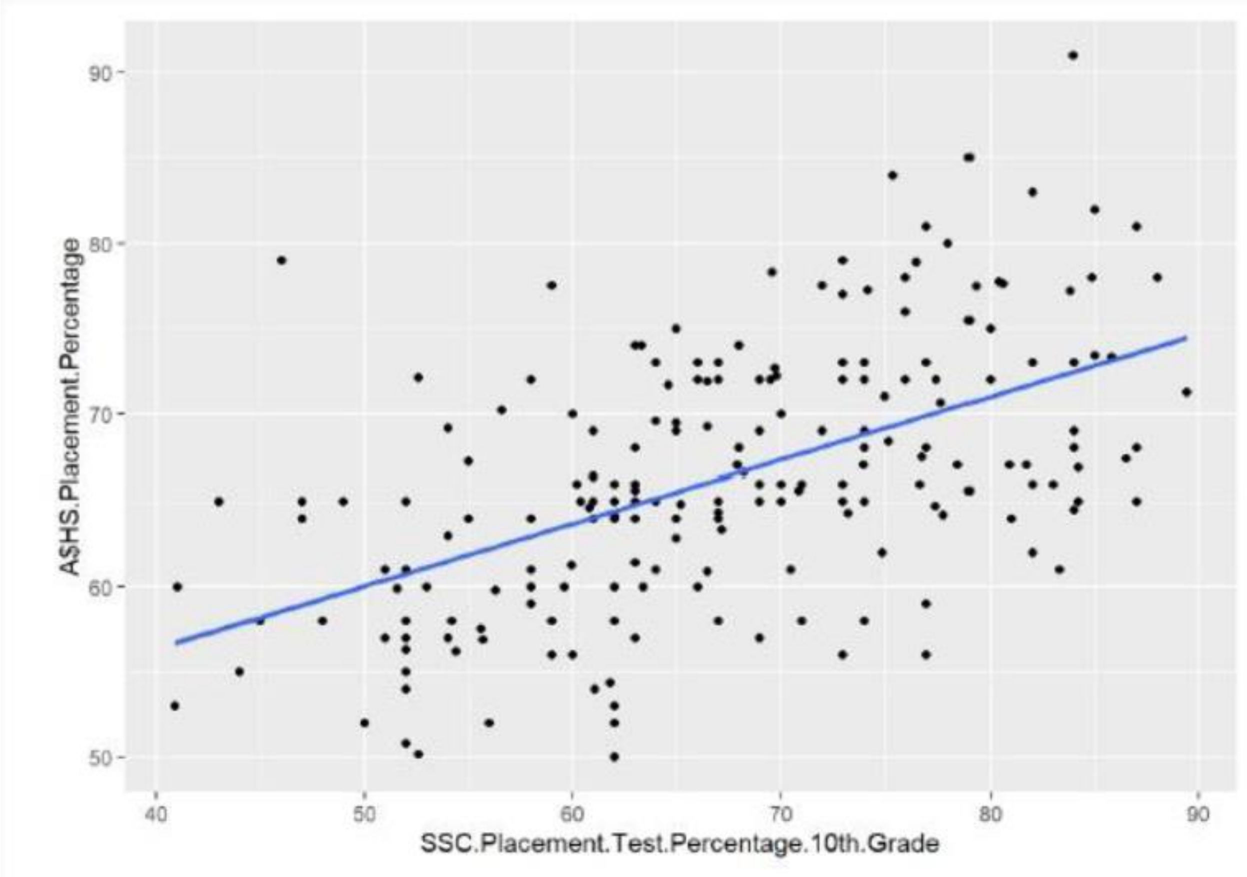


Figure 13: Test Scores by Grade Level

The data analysis conducted in this study provides valuable insights that can inform strategic decision-making across various aspects of business operations. Firstly, the analysis underscores the pivotal role of data-driven approaches in shaping business strategies, highlighting the importance of leveraging advanced analytical techniques to uncover trends, patterns, and correlations within large datasets. Such insights enable organizations to make evidence-based decisions, mitigate risks, and capitalize on emerging opportunities in the marketplace.

Specifically, the project undertaken in IST 687 demonstrates the potential of data analysis to inform educational policies and practices. By examining factors influencing post-graduation job placement, such as educational tracks and gender disparities, businesses can identify areas for improvement and implement targeted interventions to enhance outcomes for diverse student cohorts. For instance, the observation of gender disparities in job placement rates underscores the importance of promoting gender equality and diversity within educational institutions and the workforce, which can have long-term implications for organizational performance and competitiveness.

Moreover, the analysis of graduate-level placement tests and their correlation with reported salaries provides valuable insights into the efficacy of existing assessment measures in predicting job placement success. This highlights the need for organizations to critically evaluate the relevance and effectiveness of assessment tools and consider alternative predictors when making hiring decisions.

Overall, the findings of this study underscore the transformative potential of data analysis in driving informed decision-making and strategic planning across various domains. By harnessing the power of data, organizations can gain a deeper understanding of their business landscape, identify areas for improvement, and ultimately achieve sustainable growth and prosperity.

Communication

For every project culminating at the end of each of these project-based courses, the customary deliverables encompassed three essential components: the methodically crafted code underpinning the analysis, a comprehensive written report elucidating methodologies, discoveries, images, and an articulate presentation delivered to peers and the professor, encapsulating the project's key facets (IST 618, IST 707, IST 718). The commenting of code not only facilitates constructive feedback on the analytical approach but also furnishes stakeholders with a coherent written roadmap to trace the trajectory of the conclusions drawn. Across all projects constituting my portfolio, a concerted effort was invested in ensuring that my code remained extensively commented, elucidating the rationale behind every code block. Evidently, from the first endeavor examining what tracts make Indian school children likely to be placed in jobs (IST 687) to the completion focusing on the Machine Learning analysis of Airbnb's data (IST 707), discernible strides have been made in enhancing the clarity and depth of commentary within my codebase.

While code undeniably holds significance, its practical utility in the workplace is limited by the scarcity of individuals equipped to decipher it, even when meticulously annotated. Hence, mastering alternative modes of conveying findings emerges as a linchpin for success in data science.

As previously outlined, within the framework of the four coding-centric project courses, two additional deliverables emerged: the comprehensive report and the stakeholder presentation. The former, crafted using Microsoft Word initially, transitions seamlessly to Adobe Acrobat for a polished presentation of findings alongside captivating visualizations. On the other hand, the latter is meticulously prepared using PowerPoint, facilitating convenient dissemination through Adobe Acrobat.

However, this merely scratches the surface. The crux of the matter lies in the art of translation—distilling the intricate realms of data science, replete with hypotheses, inquiries, relational nuances, and weighty data analyses, into bite-sized, comprehensible nuggets that render the findings accessible to colleagues less versed in the intricacies of the field.

Ethical Dimensions

In data science, ensuring privacy remained a paramount priority throughout each project. Specifically, adherence to the General Data Protection Regulation (GDPR) was a primary concern for our teams (IST 618, IST 687), particularly given my role in handling sensitive user data within my current position at a university. Throughout the program, I deliberately avoided analyses or datasets containing sensitive data (IST 652, IST 687, IST 707, IST 718), such as medical information, personally identifiable data, or financials. This proactive approach was implemented at the outset of each project to preemptively sidestep the necessity of scrubbing sensitive information from datasets. This resulted in a lack of instances wherein such actions were required. It is crucial to acknowledge that while this approach may not always be feasible in real-world scenarios, it underscores the necessity of vigilance in safeguarding the privacy and integrity of individuals potentially impacted by our research endeavors.

Conclusion

The compilation presented herein attests to the proficient attainment of Syracuse University's Applied Data Science program's seven learning objectives while also showcasing a nuanced comprehension of key practice domains within the field of data science, namely business analysis, computer science, and data analysis (IST 718). Data procurement entailed the utilization of diverse sources, including Kaggle and open-source databases, subsequently structured and processed utilizing either R or Python (IST 652, IST 687, IST 707, IST 718). Through machine learning algorithms, visual representations, and statistical modeling techniques, discernible patterns within the dataset were elucidated (IST 718, IST 707). Subsequent analyses culminated in formulating alternative strategies and actionable plans, with a. emphasis on the business implications and option of running an Airbnb in New York City (IST 718). Syracuse University's School of Information Studies empowers students to cultivate essential competencies in data analysis, business analysis, computer science, and communication, equipping them to furnish actionable insights to diverse audiences.

The Applied Data Science program places paramount importance on bridging the divide between business stakeholders and IT experts, a pivotal aspect in tackling data-related challenges, addressing business needs, and enhancing operational effectiveness. Furthermore, the program fosters a reflective consideration of the ethical dimensions inherent in data management and analysis, recognizing the volume of data in our world. It underscores the imperative for data scientists to exercise vigilance in safeguarding the privacy of individuals and organizations while also ensuring that the data utilized in analysis remains unbiased and representative.

I want to thank the faculty, student services staff, and technology support employees for their help throughout this program. While this may seem odd, I would also like to thank my fellow students whose insight, knowledge, and willingness to help someone from outside the technology field was essential to all learning. I know that without their help, I would not have made it to the point that I am at where I can have the chance to showcase my work and learning.