Syracuse University

Portfolio Milestone Project

School of Information Studies

Allison R. Hollingsworth Deming

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 whil mastering advanced programming languages beyond my proficiency with Excel and SPSS.

Program Learning Outcomes

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

•Proficient data collection and organization techniques.

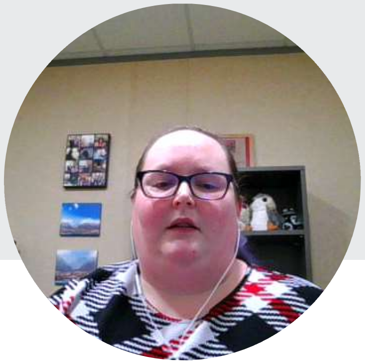
•Discerning patterns through data visualization, statistical analysis, and data mining.

•Crafting alternative strategies grounded in data insights.

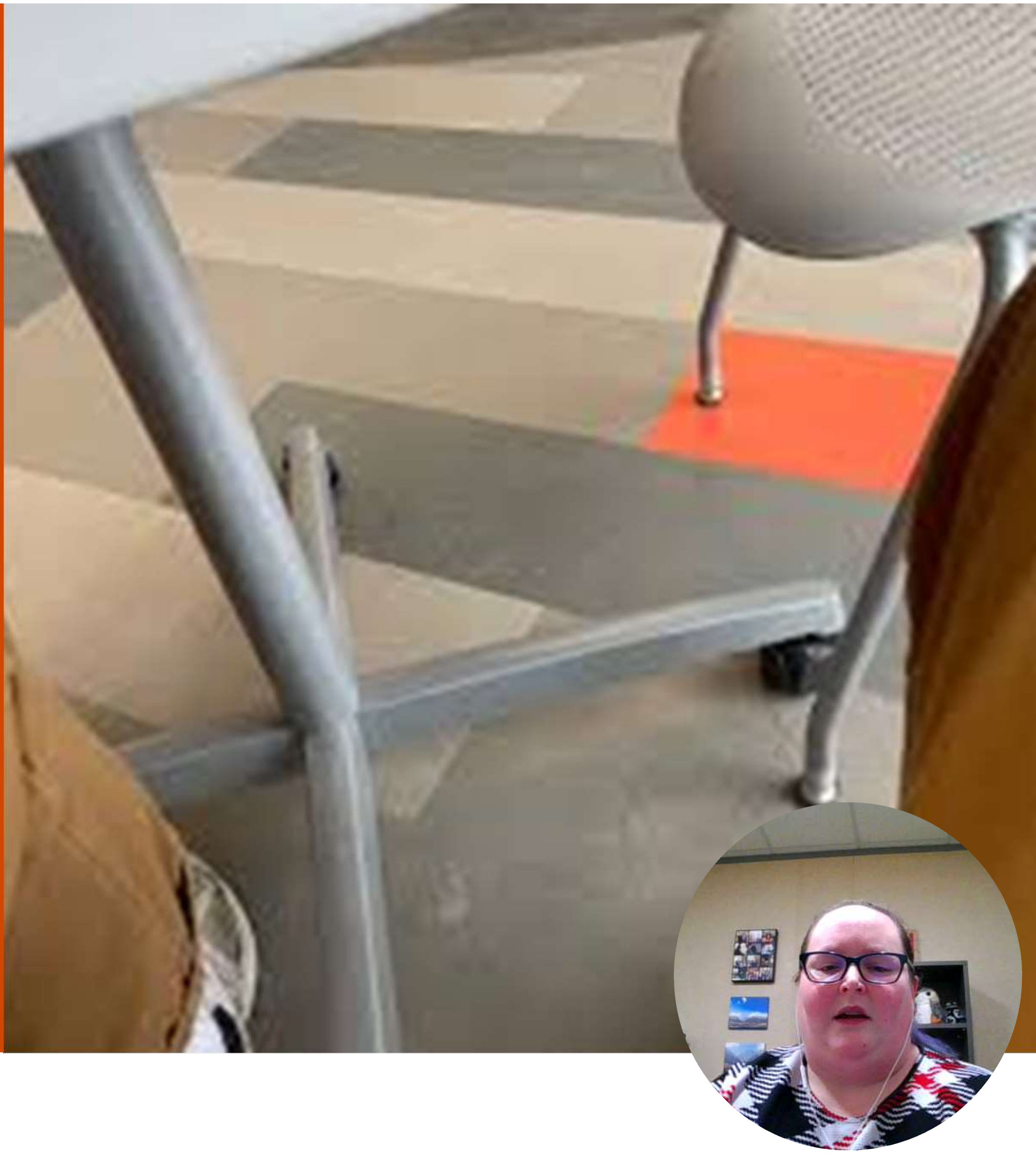
•Formulating actionable plans to actualize business decisions derived from analyses.

•Effectively communicating data insights to diverse stakeholders within organizations.

•Ethically navigating the complexities of data science practice.



Project Descriptions



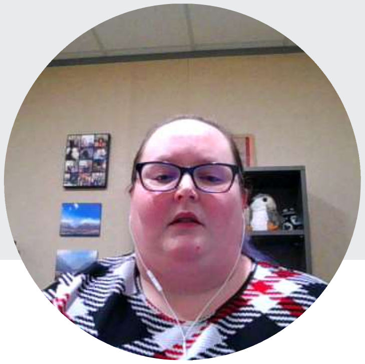
IST 618 – Information Policy

Course Overview:

- Focuses on non-applied concepts with papers, presentations, and a debate to enhance communication skills.
- Explores ethical dilemmas in data analysis and their impact on information policy.

- Papers:

  - Cover topics like access and affordability, "Big Tech," right to repair, and ethics of cell phone tracking technology.
  - Provide a deep dive into critical issues at the intersection of technology and society.



IST 652 – Scripting for Data Analysis

- *Course Highlights:*

  - Conducted a group project analyzing Airbnb data with no formal presentation, utilizing Jupiter notebook files.
  - Incorporated additional data from NYC crime statistics for comprehensive analysis.

- *Project Focus:*

  - Highlighted pricing trends, term frequency in listings, and correlations with crime statistics.
  - Demonstrated practical application of Python programming skills in data analysis.

IST 687 – Introduction to Data Science

- *Course Structure:*

  - Utilized R programming language for data exploration and analysis.
  - Employed agile methodology with a Kanban board for project management.

- *Project Scope:*

  - Investigated factors influencing employment prospects for students, emphasizing data-driven decision-making in education.
  - Applied data science techniques to real-world problems in educational settings.

IST 707 – Applied Machine Learning

- *Course Objectives:*

  - Explored Airbnb dataset extensively using R programming language.
  - Analyzed business-oriented objectives including amenities impact on pricing, review sentiments, and seasonal variability.

- *Analytical Techniques:*

  - Utilized decision trees, K-means clustering, sentiment analysis, and other methods for visualization and prediction.
  - Demonstrated proficiency in advanced analytics and machine learning algorithms.

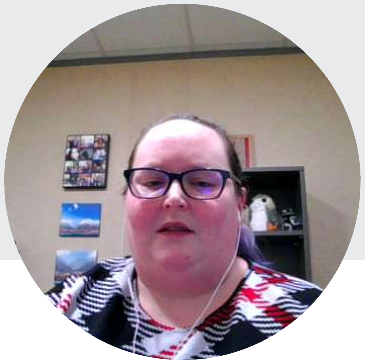
IST 718 – Big Data Analytics

- **Project Focus:**

  - Focused on assessing potential earnings of Syracuse Football Coach in alternative conferences.
  - Integrated multiple datasets for comparison of compensation packages and performance metrics.

- **Exploratory Analysis:**

  - Explored nuances of data analytics in shaping collegiate athletics ecosystem.
  - Highlighted complexities of analyzing large datasets for strategic decision-making.



Collecting Data

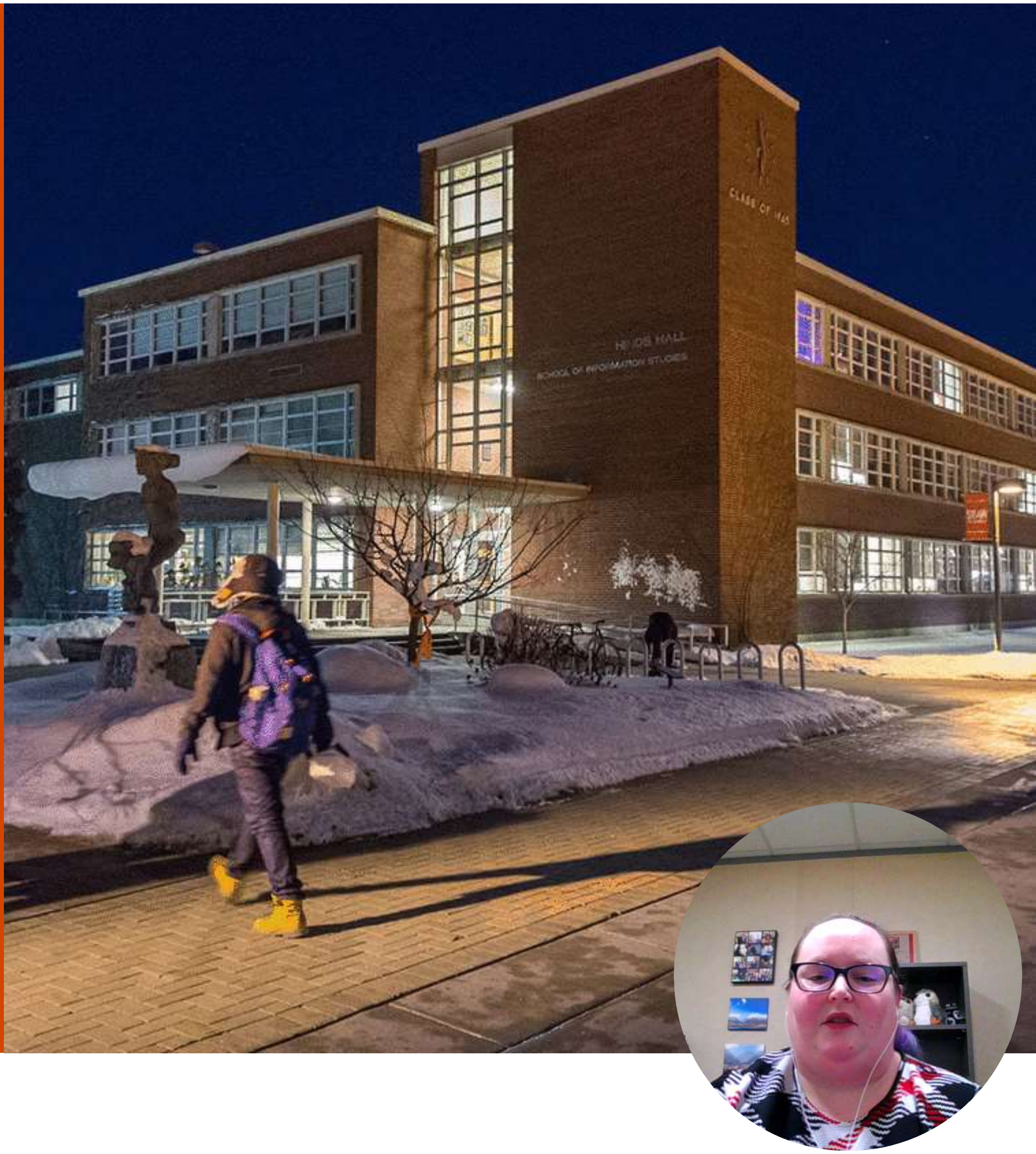
Primary Repositories

- Primary Data Repositories:
  - Kaggle:
    - Used in IST 687, IST 707, IST 718.
    - Open-source platform: https://www.kaggle.com/
  - Airbnb Data Science Homepage:
    - Provided by company for exploration and projects.
    - Source: http://insideairbnb.com/get-the-data/
    - Utilized in IST 652, IST 707.

IST 659 – Database Management

- Acknowledgement:
  - SQL Class in Database Management.
  - Best example of data storage, creation, and organization.
  - Limitation in portfolio use due to remote desktop setup.
  - Note on potential for better choice.

Data Analysis

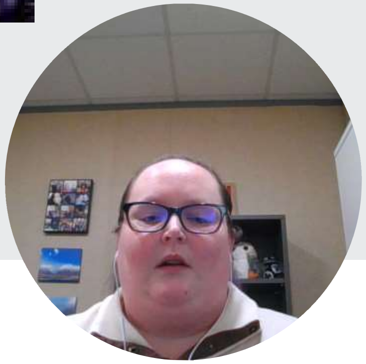


Airbnb Data – IST 707 Overview

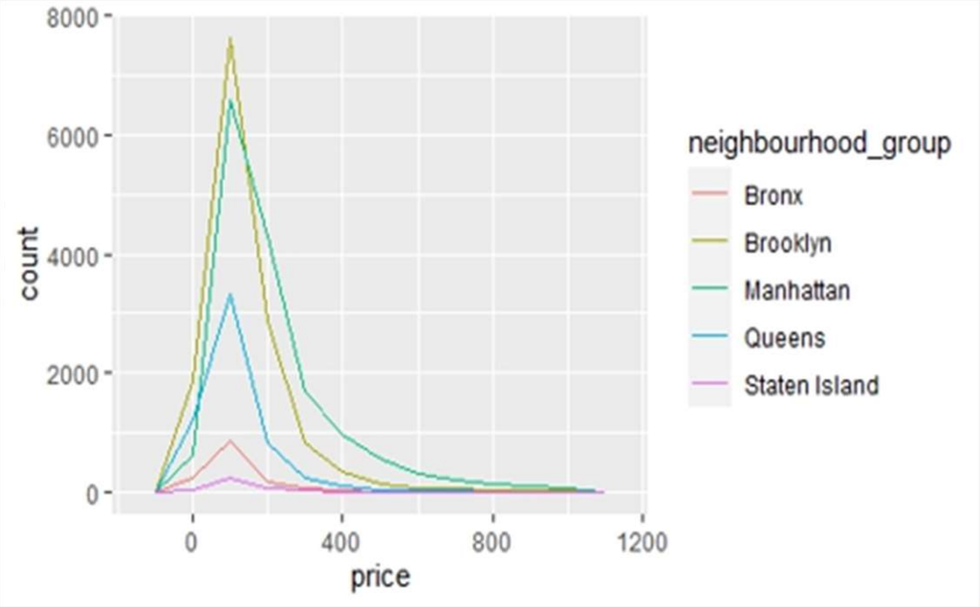
- Comprehensive exploration from elementary descriptive analytics to advanced machine learning models.

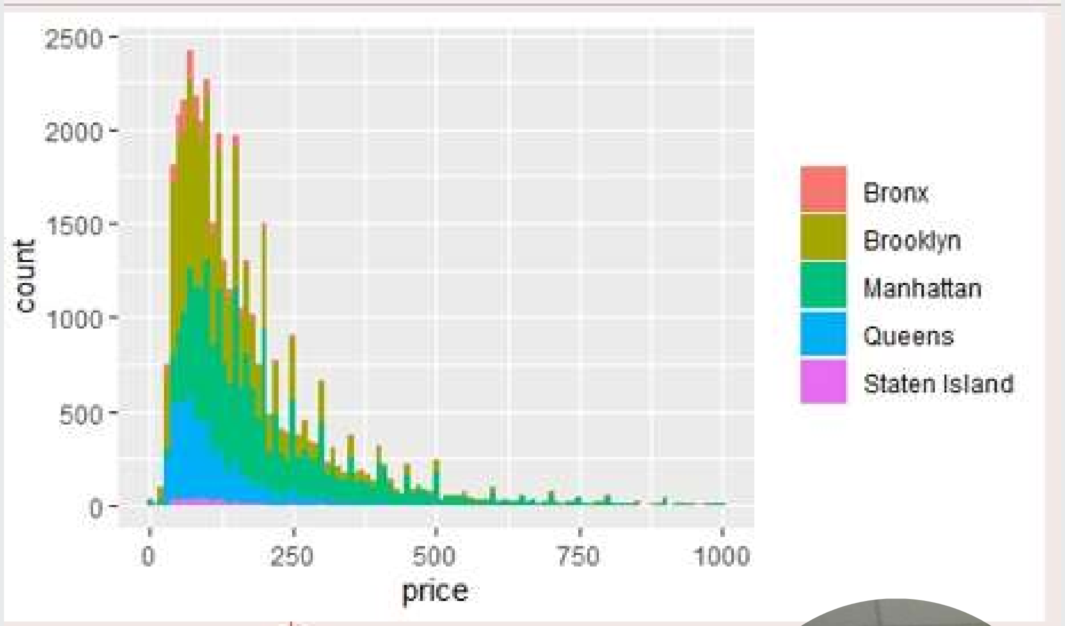
- Focus on forecasting seasonal variations, sentiment analysis, review frequency, and listing availability.

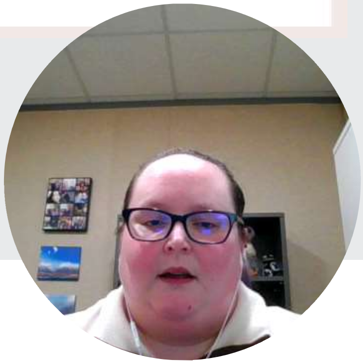
- Highlights the impact of amenities on pricing dynamics and the prevalence of affordable options across NYC boroughs.

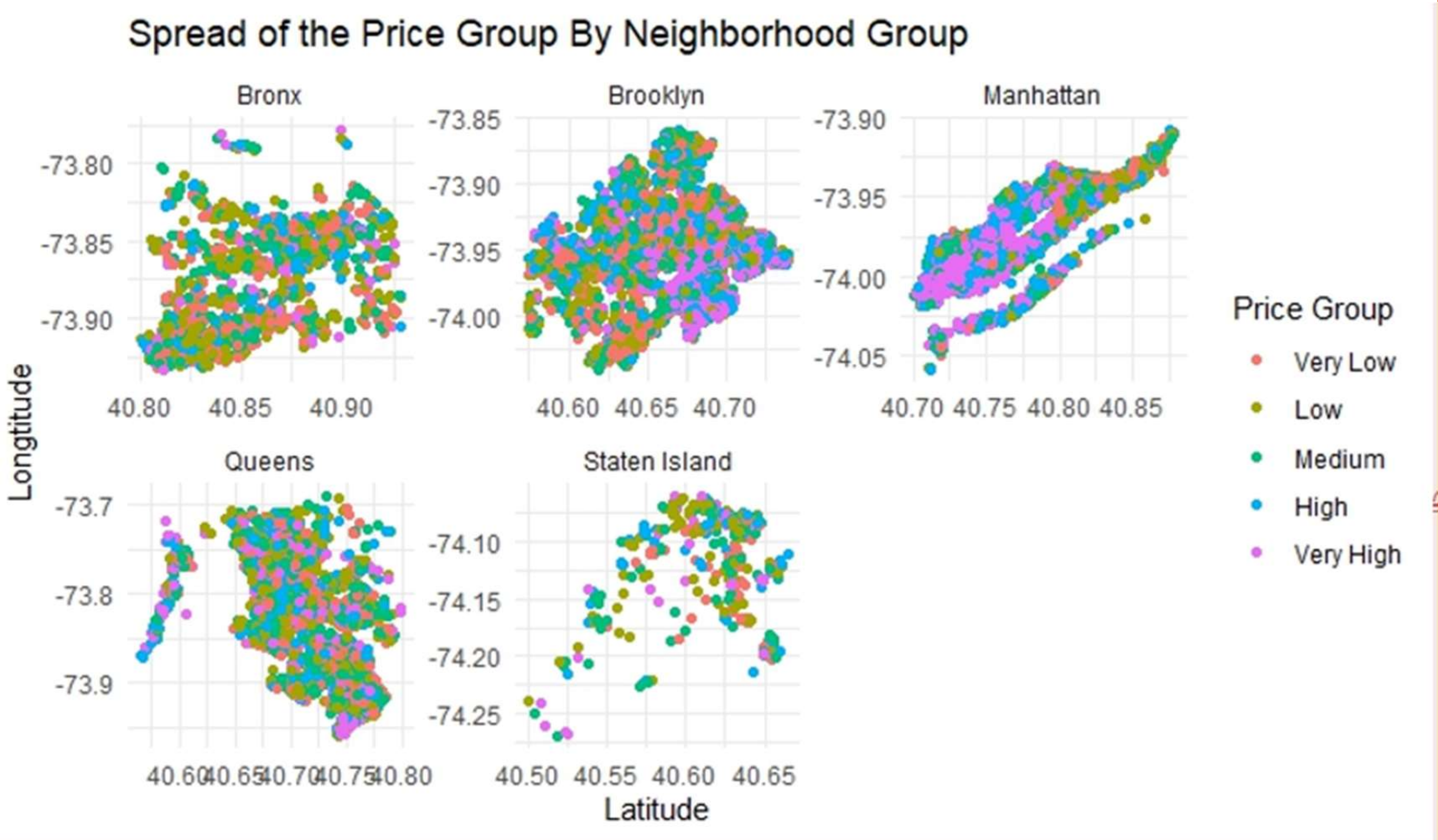


Analyzing Data – Airbnb









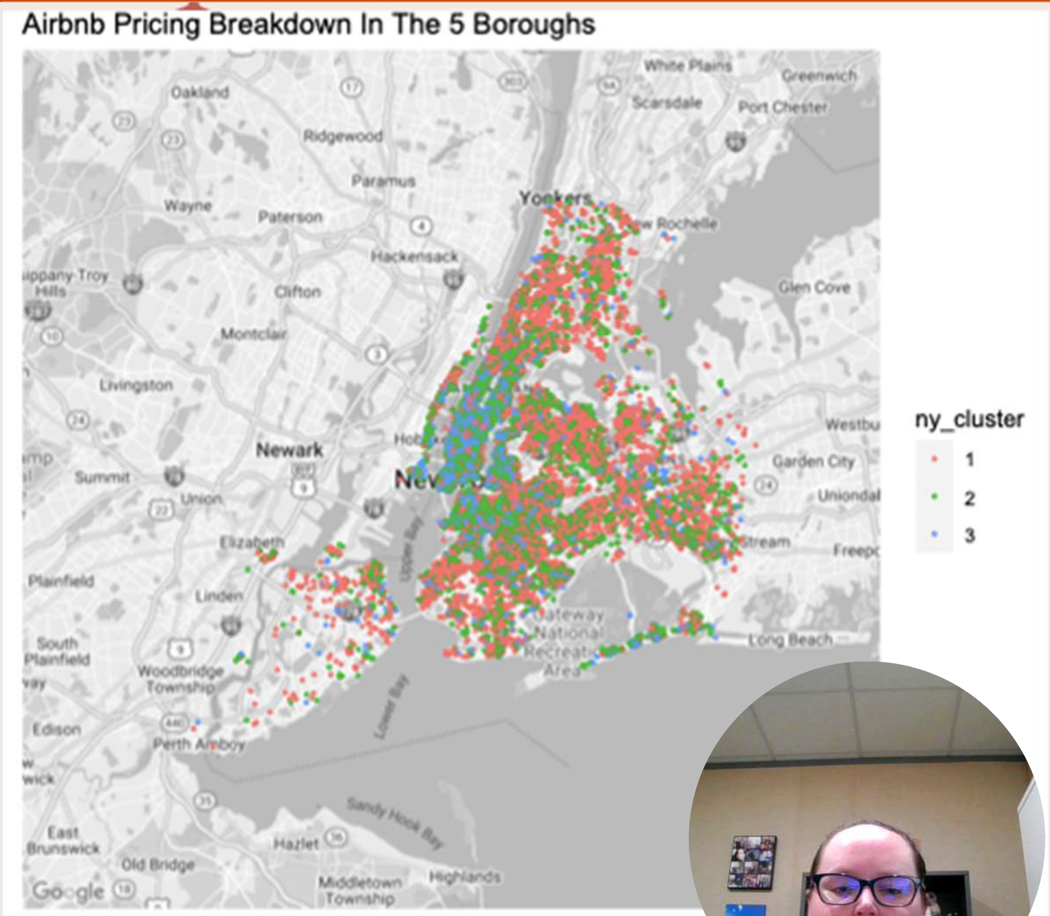
Analyzing Data – Airbnb in R

Clustering discretized Data in R.

K-Means Cluster Analysis Map

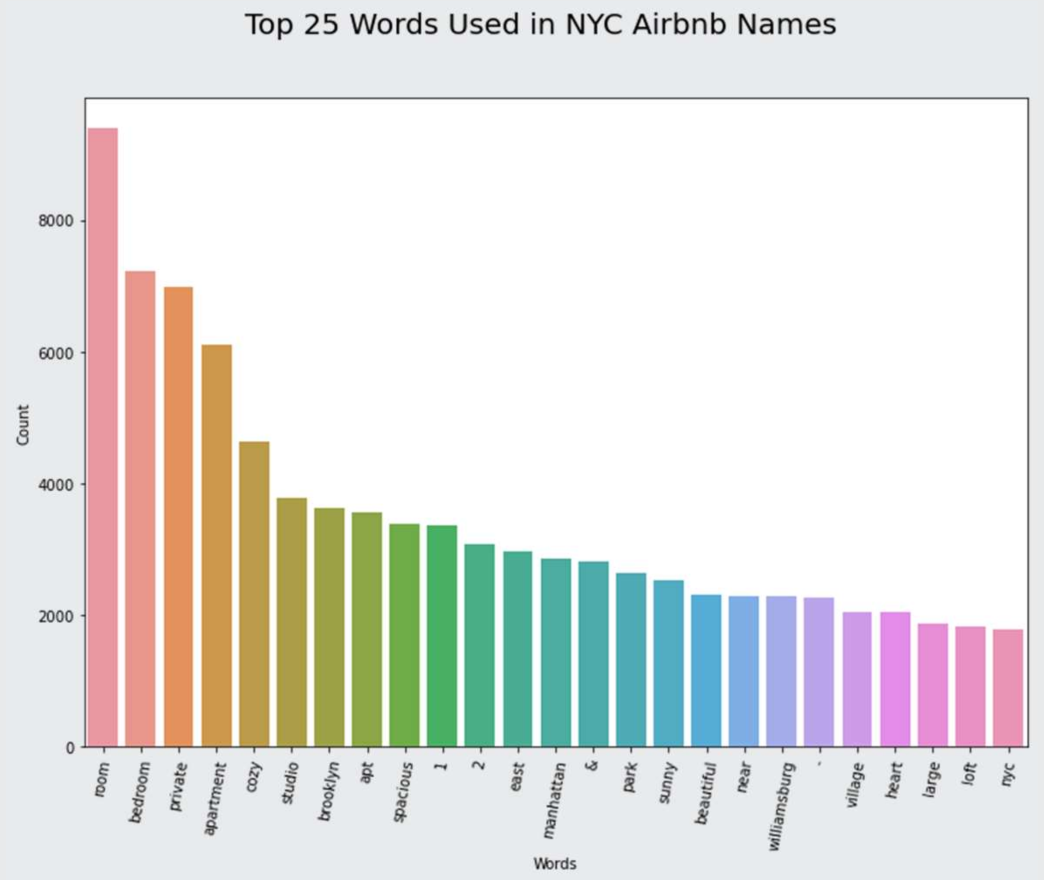
K-Means Analysis

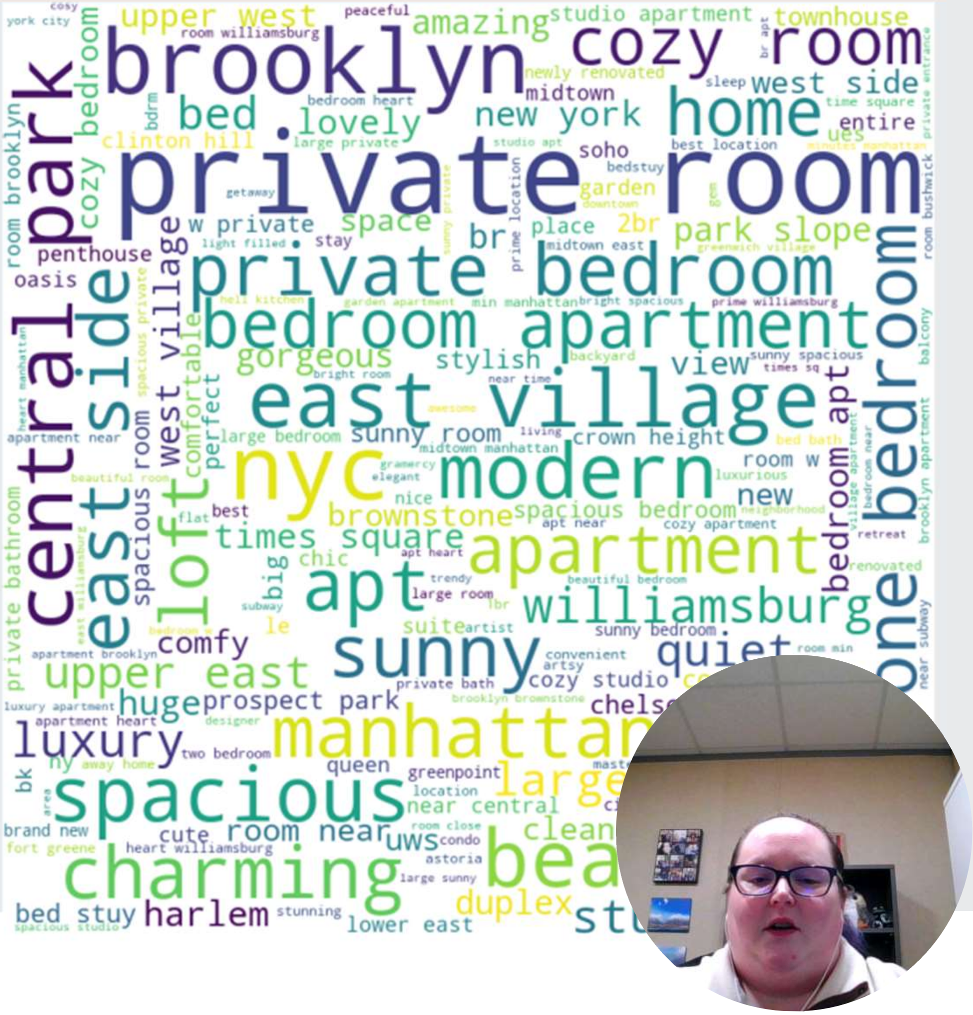
It was found through a clustering analysis that there were naturally three groups within the pricing data data. While we had initially divided it into five this is more accurate due to the statistical analysis backing it up. 1 is the cheapest cluster, 2 is moderate, and 3 is the expensive cluster.



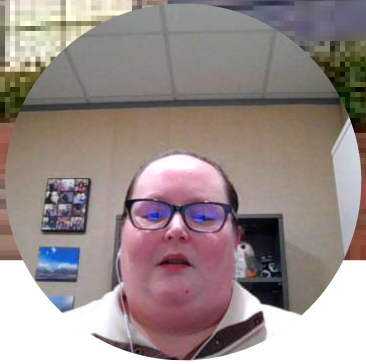
Alternative Strategies with Data - Python

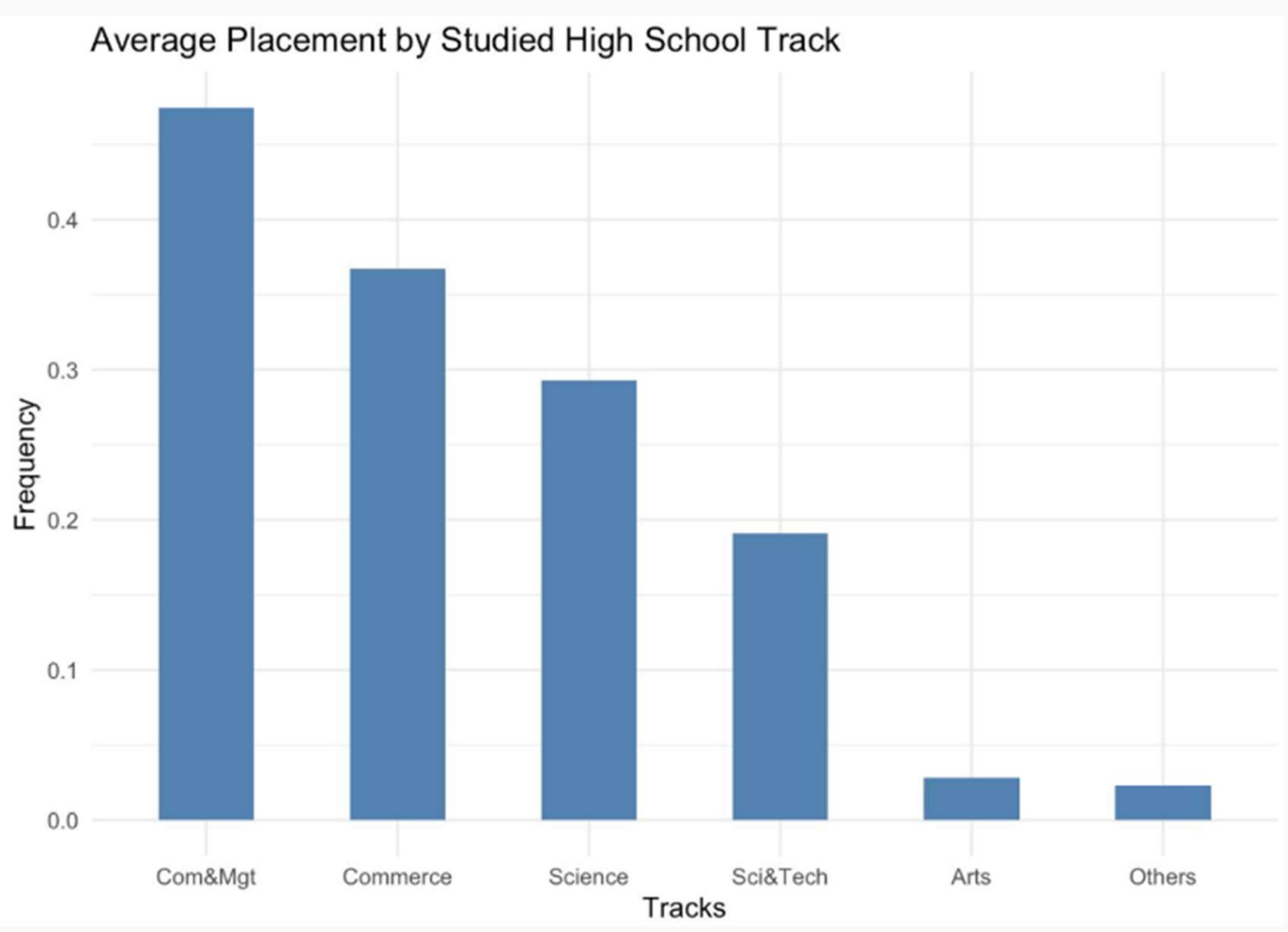
IST 652 - Python


Top 25 Words Used in NYC Airbnb Names



Business Insights





Placement by School Track

Communication and Management

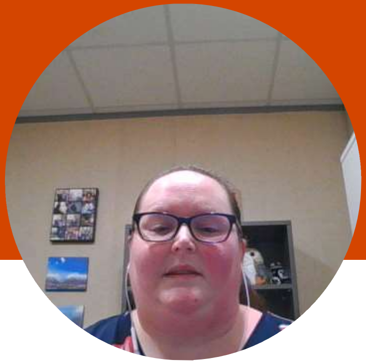
Commerce

Science Only

Science and Technology

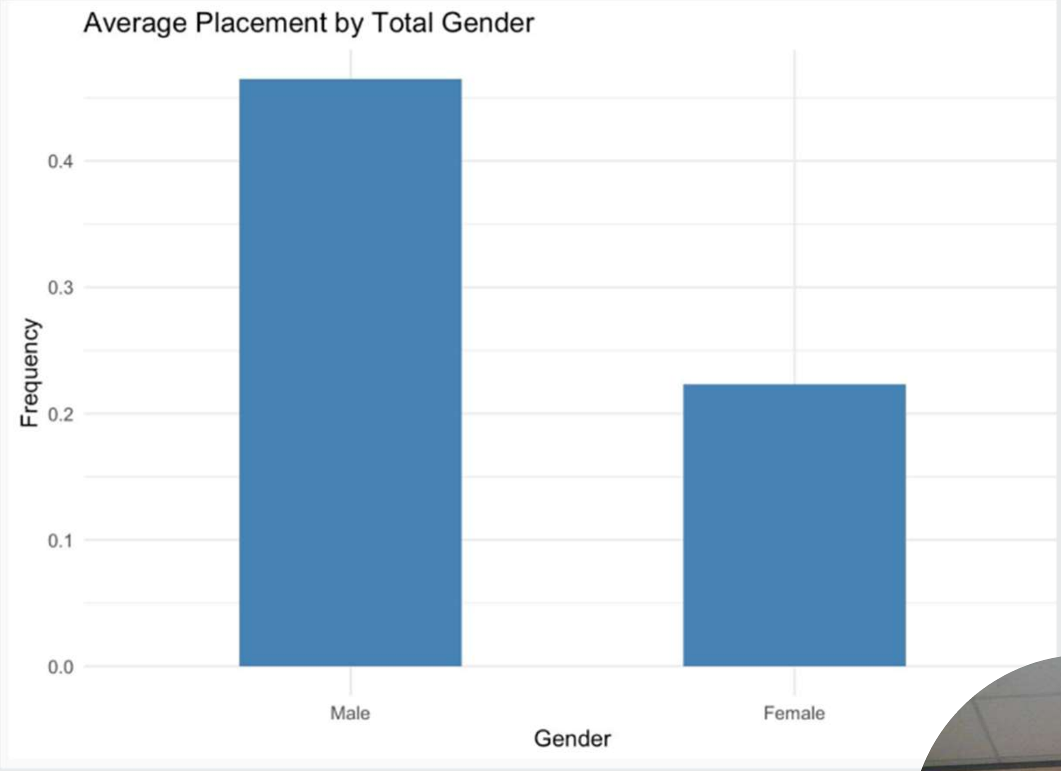
Arts

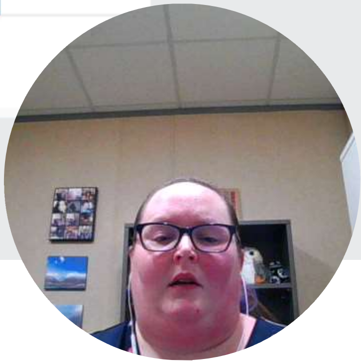
Others



Average Placement by Total Gender

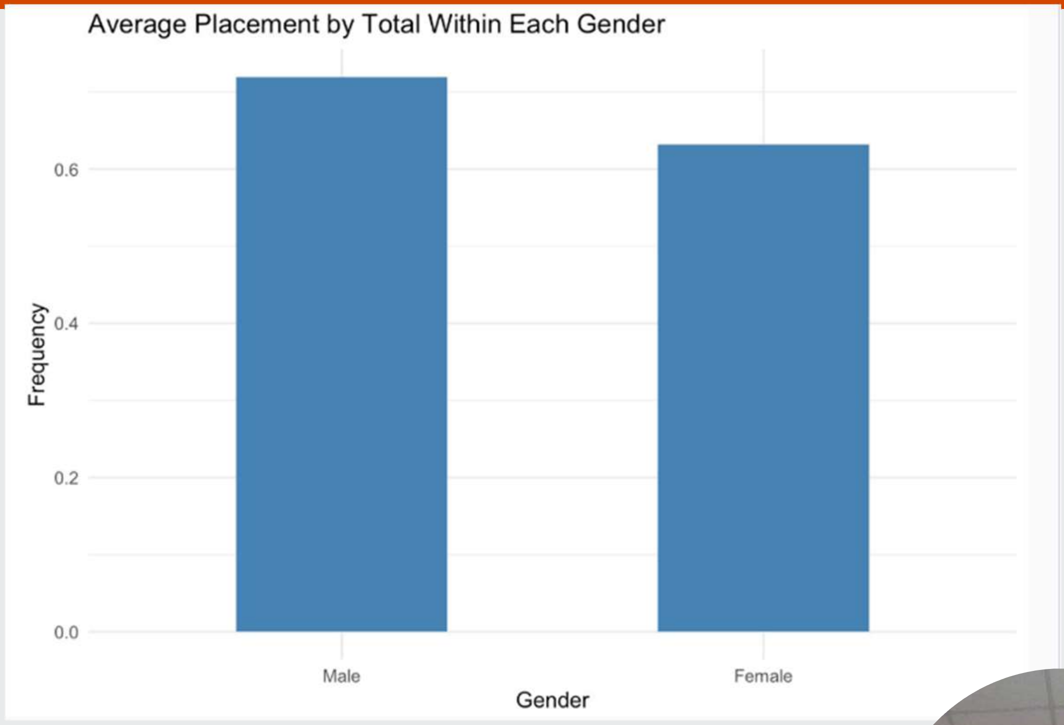
As you can see there is about double the number of males compared to females in this school system. While the data is a little bit old it is likely that cultural norms influenced the representation of males and females in the data.

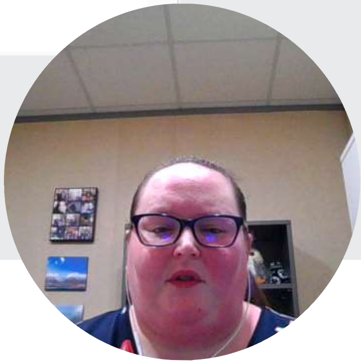




Average Placement by Total Within Each Gender

Taking out the bias that is inherent in the data we can see that proportionally, the success rate of being placed following training and exams is almost equal with males and females performing similarly, although the males are still slightly outperforming females.





Communication

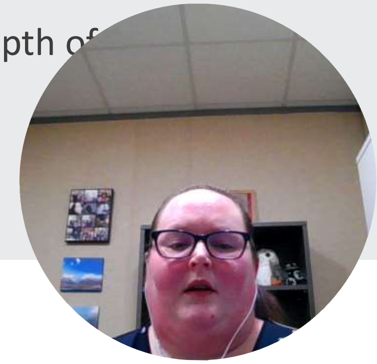
Overview of Project Deliverables

- For every project culminating at the end of each project-based course (IST 618, IST 707, IST 718), three essential components were delivered:

1. Methodically crafted code underpinning the analysis.
2. Comprehensive written report elucidating methodologies, discoveries, images, and insights.
3. Articulate presentation delivered to peers and the professor encapsulating key project facets.

- Commenting of code facilitates constructive feedback, providing stakeholders with a coherent roadmap to understand analytical approaches and conclusions.

- Effort invested in ensuring extensively commented code across all projects, enhancing clarity and depth of commentary over time.



Importance of Conveying Findings

- While code holds significance, its practical utility in the workplace is limited by the scarcity of individuals equipped to decipher it, even with meticulous annotations.

- Mastery of alternative modes of conveying findings emerges as a linchpin for success in data science.

- Translation of complex data science realms into comprehensible nuggets is essential for rendering findings accessible to colleagues less versed in the field.

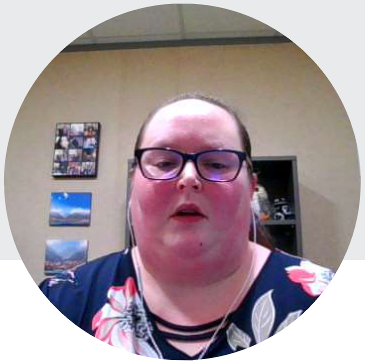
Ethical Dimensions

Ethical Dimensions Privacy in Data Science

Privacy is a top priority in data science projects. Adherence to GDPR guidelines was crucial in projects like IST 618 and IST 687, and in all analyses in my given my role in handling sensitive data at the university.

Throughout the program, I avoided datasets containing sensitive information (IST 652, IST 687, IST 707, IST 718) to proactively protect privacy. This approach minimized the need for data scrubbing.

While not always feasible, this highlights the importance of vigilance in safeguarding individuals' privacy and integrity in research.



Syracuse University

Masters of Science in Applied Data Science Portfolio

Conclusion

Allison R. Deming (Hollingsworth)

iSchool