Bias, Bets, and Bytes

A REFLECTION ON DATA, DECISIONS, AND ETHICS

Cheromaine Smith | 866290307 | May 16, 2025

Table of Contents

Cheromainé Smith

Cortland, NY | 315.746.0536 | csmith27@oswego.edu | Cheromaine Smith | LinkedIn

Data-Driven Analyst | Results-Oriented Specialist | Strategic Project Coordinator

PROFESSIONAL SUMMARY

Detail-oriented Data Analyst with over 5 years of experience in technology, business operations, and higher education administration. Recently completed an M.S. in Applied Data Science, bringing a blend of technical and strategic capabilities to improve workflows, financial oversight, and system performance. As an Operations Specialist and former Functional Business Analyst, I streamline departmental processes, enhance decision-making with data-driven solutions, and coordinate complex projects. Proficient in SQL, Tableau, and Python, with a track record of leading cross-functional teams and improving institutional efficiency.

AREAS OF EXPERTISE

- ✓ Adaptability
- ✓ Business Process Improvement
- ✓ Budget & Financial Oversight
- ✓ Customer Service Excellence
- ✓ Data Analysis and Reporting
- ✓ Data Visualization
- ✓ Departmental Operations
- ✓ ERP System Implementation
- ✓ Event Planning
- ✓ Invoicing and Billing
- ✓ Leadership
- ✓ Personnel Management
- ✓ Process Documentation
- ✓ Project Management
- ✓ Recruitment
- ✓ Stakeholder Communication
- ✓ Strategic Planning
- ✓ Talent Management
- ✓ Team Collaboration
- ✓ Technical Support

CAREER HIGHLIGHTS

Comprehensive Reporting System Development: Collaborated on the creation and maintenance of over 150 reports and data sources, improving workforce visibility and streamlining operations across university departments. Utilized tools including SQL, Tableau, Python, Power BI, Azure, Access, and R Studio.

Transformative ERP Implementation: Successfully led testing and implementation of enterprise systems at Syracuse University, optimizing over 200 business processes and enhancing cross-departmental functionality.

Innovative Recruitment Strategies: Directed Syracuse University's first auxiliary services job fair, resulting in the hiring of 200+ staff. Developed internal dashboards and tools to track hiring metrics, increasing diversity and efficiency.

PROFESSIONAL EXPERIENCE

Operations Specialist I, Department of Mathematics, Syracuse University, Syracuse, NY | October 2024 – Present

Manage departmental operations, budgets, and staff oversight while ensuring compliance with university policies. Provide strategic support to the department Chair and serve as a cross-campus liaison to align academic and administrative priorities.

- Direct daily operations by supervising departmental staff, setting workflow priorities and resolving personnel issues to ensure a smooth-running academic unit.
- Oversee budget planning and reconciliation across multiple funding sources, making recommendations to the Chair to ensure compliance with financial policies.
- Act as hiring lead for faculty, staff, PTI and postdoc positions, managing search processes and maintaining personnel records in line with university procedures.
- Serve as a key liaison to the Dean’s Office, Graduate School, and central administration, aligning departmental initiatives with institutional goals.
- Manage major projects including space planning, graduate orientation, and accreditation efforts, ensuring deadlines and standards were met.
- Monitor and report on faculty start-up budgets, scholarship credits, and financial aid to maintain transparency and accuracy.
- Coordinate maintenance and facilities logistics, submitting work orders to ensure resolution to keep department spaces operational.

Functional Business Analyst I, Human Resources, Syracuse University, Syracuse, NY | April 2023 – October 2024

Serve as a key liaison for testing and implementation of ERP systems, analyzing business requirements to create functional specifications. Manage application maintenance, coordinate security protocols for new employees, and develop data reports to enhance operational efficiency across the university.

- Serve as a primary liaison for ERP system testing, overseeing over 200 processes to ensure successful implementation.
- Analyze business requirements and convert them into detailed functional specifications for development teams, enhancing project clarity.
- Manage application maintenance by diagnosing errors and updating business specifications to optimize system performance.
- Lead security approval processes for over 1,000 new employees, ensuring compliance with HR protocols and policies.
- Develop and maintain a variety of comprehensive data reports to support both operational and ad-hoc reporting needs.
- Coordinate with cross-functional teams to design, test, and implement innovative solutions for business challenges.
- Facilitate training and documentation efforts to ensure seamless user adoption and understanding of new systems.

Talent Management Associate, Human Resources, Syracuse University, Syracuse, NY | August 2021 – April 2023

Managed recruitment reporting and organized talent acquisition events, successfully hiring over 2,000 employees' while enhancing data-driven decision-making through dashboard development and process improvements.

- Developed and analyzed recruitment reports to track metrics on diversity and employee eligibility, enhancing team insights.
- Organized the inaugural auxiliary services job fair, successfully hiring over 200 employees and streamlining the recruitment process.
- Designed and implemented an internal dashboard to measure recruitment methods, enabling data-driven decision-making for talent managers.
- Led the planning and execution of career fairs, contributing to the hiring of over 2,000 temporary employees across the university.
- Created the opening design for Syracuse University's first Virtual Career Fair, collaborating with state partners and attracting over 600 candidates.
- Trained over 50 colleagues on workflow processes and best practices, fostering a more efficient recruitment team.
- Provided technical assistance and support within the HR office, improving operations and resolving data-related issues.

Audit & Invoice Specialist, Robert Half Salaried Professional Services, Syracuse, NY | June 2019 – August 2021

Full-time employee contracted to various companies, where I specialized in auditing and invoicing. Conducted over 125 weekly audits for Protiviti and ensured billing accuracy for Arcadis, collaborating with project teams enhancing operational efficiency.

- Conducted over 125 weekly audits at Protiviti, ensuring the accuracy of internal calculations and claims.
- Collaborated with project managers at Arcadis to process invoices accurately and on time.
- Utilized IBM and BOTs to determine case dispositions, contributing to efficient case management.
- Trained team members on the effective use of auditing systems, enhancing team performance.
- Maintained billing accuracy by adjusting employee rates and approving contracts in Oracle Cloud.
- Compiled and organized receipts for invoicing periods, ensuring documentation integrity and compliance.
- Monitored project setups daily to ensure consistency across client portfolios, supporting operational goals.

Client Processing Representative, Bank of New York Mellon, East Syracuse, NY | June 2016 – June 2019

Collaborating with brokers and clients to enhance operational processes, which significantly improved efficiency and compliance within the trading environment.

- Monitored and amended over 1,000 trades daily, ensuring accurate settlement and minimizing operational crises.
- Collaborated with brokers, clients, and internal departments to address trading issues and improve efficiency.
- Developed and implemented a new procedure for collecting Broker Contact Information, enhancing operational workflow.
- Led the team in migrating multiple accounts to Brussels, overseeing VDS migration and tracking discrepancies.
- Organized and facilitated weekly meetings, creating schedules and distributing minutes to ensure clear communication.
- Analyzed financial models related to trades, contributing to better decision-making and performance evaluation.
- Completed over 76 training courses, including AML/SAR and Data Governance, earning recognition as Employee of the Month.

EDUCATION

Master of Science in Applied Data Science
Syracuse University, Syracuse, NY
May 2023 – May 2025

Bachelor of Science in Business Administration
State University of New York (SUNY) College at Oswego, Oswego, NY
August 2012 – May 2016

CERTIFICATIONS & MEMBERSHIPS

Member of Phi Kappa Phi
Syracuse University, Syracuse, NY

Member of Phi Beta Lambda
State University of New York (SUNY) College at Oswego, Oswego, NY

Director of Communication & Vice President of Public Relations for Enactus
State University of New York (SUNY) College at Oswego, Oswego, NY

Introduction

Data is often called the new oil, not because it is limited but because like oil it must be cleaned, processed and even transformed before it can fuel intelligent and informed decisions. As Clive Humby, the British mathematician and data science pioneer, famously said, “Data is the new oil. It’s valuable, but if unrefined it cannot really be used”. This portfolio represents the culmination of work completed throughout the Master’s in Applied Data Science program, where data was gathered from a range of domains including retail, real estate, sports, finance and artificial intelligence to be analyzed and transformed into actionable insights.

The program is structured to equip students with the technical and analytical skills necessary to become successful data professionals in a world increasingly driven by data. The goal was simple: to not only extract value from data but to make that value practical and applicable. This portfolio is a culmination of those projects that demonstrate the ability to:

- Collect, store, and access data using appropriate technologies.
- Generate and communicate insights using the full data science lifecycle.
- Employ both R and Python to build and evaluate predictive models and visualizations
- Communicate insights gained to a broad range of audiences via visualizations.
- Utilize ethics in the use and evaluation of data systems and predictive models.

In IST 687, the analysis of Amazon sales data focused on extracting actionable business insights from transactional records, leveraging advanced visualization and statistical modeling techniques. In IST 659, a data-driven tool, code-named *Trade Genius*, was developed to support financial decision-makers by delivering high quality market trend research through a mobile platform. In IST 652, the project centered on Airbnb pricing strategies, examining pricing dynamics within the short-term rental market. Through data wrangling and machine learning techniques, patterns were identified, and data-driven pricing recommendations were developed. IST 707 explored the application of predictive analytics and model evaluation to improve decision-making under uncertainty, specifically assessing whether supervised learning could enhance accuracy in NFL Sports Betting. Finally, IST 692 addressed the ethical considerations of data science practice. This included assessing model fairness and investigating algorithmic bias in generative

AI systems, highlighting the importance of accountability and transparency in the development and deployment of data-driven technologies.

The following sections showcase a journey through learning, exploration, and discovery in the field of data science. Each project serves as a case study highlighting how data science techniques were applied to real-world problems and the valuable lessons uncovered along the way.

IST 687: Introduction to Data Science

In the Summer of 2023, Abdullah Awaysheh instructed a class aimed at introducing information professionals to fundamentals about data and the standards, technologies, and methods for organizing, managing, curating, preserving and using data. Through class sessions and projects students would discuss broader issues relating to not only data management but quality control and the publication of data. This project explored Amazon's 2020 sales data with a goal of identifying patterns in product performance, customer satisfaction, and pricing strategies. As Amazon continues to grow into the biggest juggernaut within the retail industry, the aim was to provide actionable insights to executives in the hopes of enhancing performance, identifying weak product categories and uncovering customer sentiment trends.

The dataset, sourced from Kaggle, featured 16 variables and 1,465 observations, which were then ultimately refined to 18 variables and 1,462 cleaned observations after a thorough round of preprocessing. Key fields included product pricing, rating, review content, and category information, with values originally in Indian rupees that were later converted to USD. The cleaning process involved removing NA values, parsing and splitting complex hierarchical category fields, and formatting monetary values. Next, descriptive statistics and visualizations were used to understand patterns. Finally, the team modeled the data and performed text analysis to evaluate relationships, predict values and correlate language to customer satisfaction.

Through the project, with the use of R programming language, linear regression, SVM, data visualization techniques, text mining and sentiment analysis, the top potential biases identified were:

1. **Category Imbalance** – There were far more observations in the electronics category than under toys.

2. **Negative Reviews** – There was not a lot of representation of negative reviews, in a stack of 1455 reviews only 7 came in with ratings under a 3.

3. **Geographic limitations** – This data was using the Indian Amazon market which is not representative of the global market.

PROGRAM OBJECTIVES

The following objectives were met during the completion of this project:

1. Collect, store, and access data by identifying and leveraging applicable technologies
    a. The project sourced a dataset from Kaggle, which needed to meet requirements of more than 10 variables and greater than 1,000 observations. Tools like R Studio and packages such as read_csv() were used to access and manipulate the data efficiently.
2. Create actionable insight across a range of contexts (e.g., societal, business, political), using data and the full data science life cycle
    a. Insights were extracted from Amazon sales data to provide business recommendations, including marketing strategies, pricing patterns, and customer satisfaction pulls that answer real-world questions relevant to decision making in the retail market.
3. Apply visualization and predictive models to help generate actionable insight
    a. A diverse range of visual tools (e.g., histograms, word clouds, bar charts) and models (e.g., linear regression, multiple regression, and support vector machines) were used to reveal patterns in ratings, pricing, and customer sentiment
4. Use programming languages such as R and Python to support generation of actionable insight
    a. This was a central objective of this project as we used R exclusively for data munging, cleaning, analysis and visualization, demonstrating technical fluency across the full analysis pipeline.
5. Communicate insights gained via visualization and analytics to a broad range of audiences
    a. Through visual storytelling and interpretation (e.g., identifying top revenue categories, average rating by price, discount patterns), the team was able to communicate the findings to a non-technical audience, suitable for decision-making.

PROJECT VISUALIZATION

The following three visuals illustrate the main findings and insights from the project:

1. **Price vs Rating Linear Regression Plot**

```
Call:
lm(formula = actual_price_usd ~ rating, data = AmazonData4)

Residuals:
    Min      1Q  Median      3Q     Max
-103.35  -58.43  -42.58   -1.65 1580.03

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -160.92      48.13  -3.344 0.000848 ***
rating         55.25      11.72   4.715 2.65e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 129.7 on 1460 degrees of freedom
Multiple R-squared:  0.015,     Adjusted R-squared:  0.01432
F-statistic: 22.23 on 1 and 1460 DF,  p-value: 2.649e-06
```

Fig 1.1 Price vs Rating Linear Regression Model

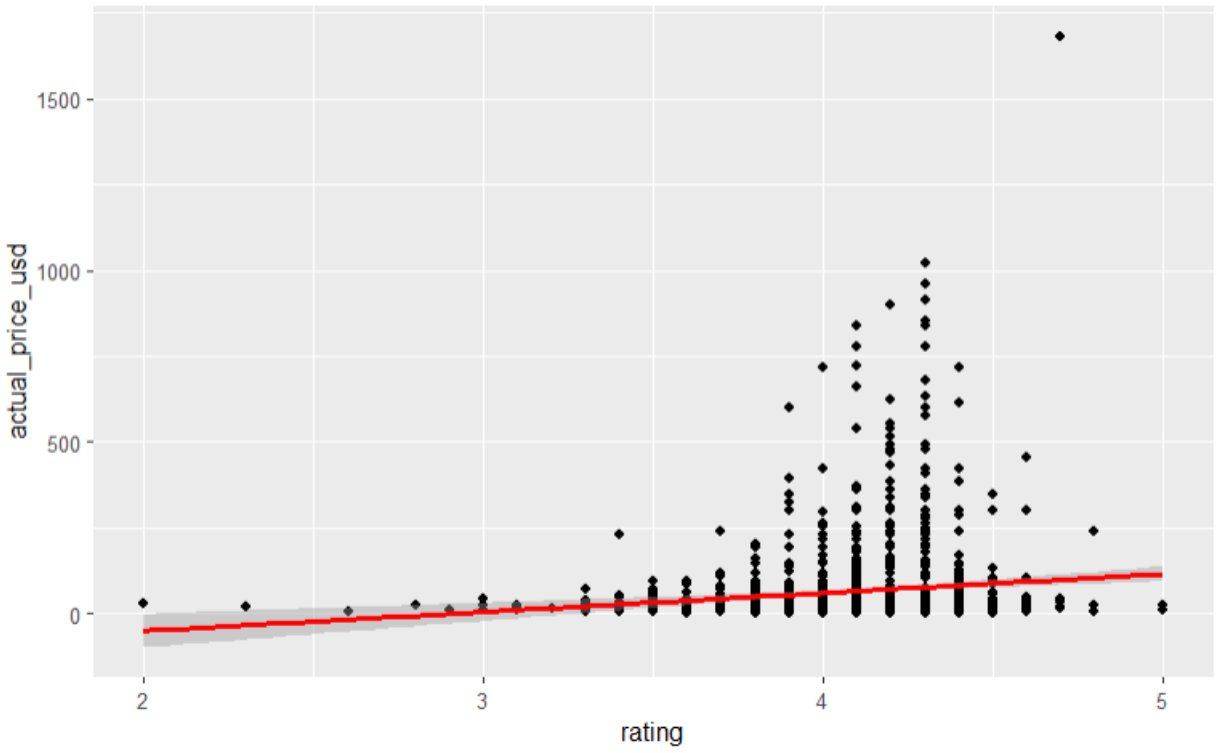


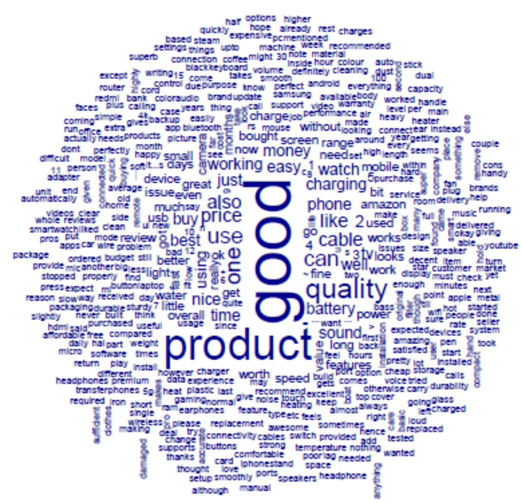
Fig 1.2 Price vs Rating Linear Regression Plot
This plot demonstrates a slight positive correlation between product price and user rating, suggesting that more expensive products tend to receive higher ratings.

2. **Word Cloud of Product Reviews**

Q 5.1: Review Title Wordcloud

```
# Wordcloud based on Review Title, words used at least twice Top
# words: good product, nice, quality, money, price
```

Fig 2.1 Word cloud of review titles



Q 5.2: Review Content Wordcloud

```
# Wordcloud based on Review Content, words used at least 3 times
# Larger word cloud, but similar top words: good, product, quality,
# price, easy, phone, batter
```

Fig 2.2 Word cloud of product reviews

These word clouds show the most frequently used words in product reviews. Words like “good”, “quality”, and “value” dominate, indicating positive customer sentiment in the dataset.

3. Average Discount by Category Bar Chart

```
## # A tibble: 9 x 2
##   Cat1                  average_discount_percentage
##   <chr>                                       <dbl>
## 1 HomeImprovement                              57.5
## 2 Computers&Accessories                        53.9
## 3 Health&PersonalCare                          53
## 4 Electronics                                  50.8
## 5 MusicalInstruments                           46
## 6 Car&Motorbike                                42
## 7 Home&Kitchen                                 40.2
## 8 OfficeProducts                               12.4
## 9 Toys&Games                                    0
```

```
# Home Improvement has the highest discount percentage with 57.5%
```

Fig 3.1 Table of Average Discount Percentage by Category

```
# Q 4.2 Create a bar chart for Average Discount % by Category 1
ggplot(average_discount_by_category, aes(x = Cat1, y = average_discount_percentage)) +
    geom_bar(stat = "identity", fill = "yellow") + labs(title = "Average Discount by Category 1",
    x = "Category", y = "Average Discount") + theme_minimal() + theme(axis.text.x = element_text(angle =
    hjust = 1))
```

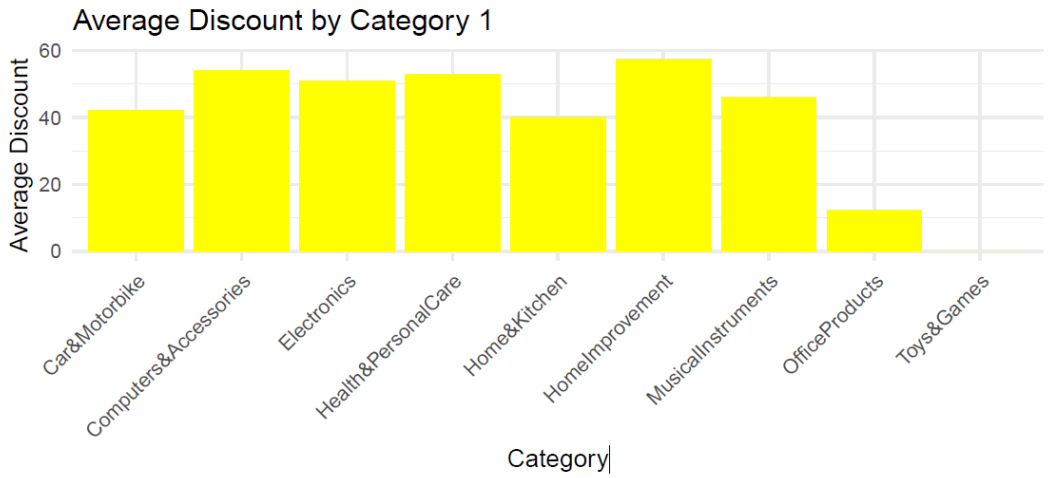


Fig 3.2 Bar Chart of Average Discount Percentage by Category

This bar chart compares average discount percentages across Tier 1 product categories, revealing which departments rely heavily on discounting to drive sales.

REFLECTION

The project successfully met its objectives by uncovering key insights about pricing, customer satisfaction, and product category performance on Amazon. The team found that Office Products had the highest average ratings, though this may be influenced by a smaller sample size or the geographic location of the Amazon sales data. A noteworthy finding was that higher actual prices tend to be positively correlated with higher ratings, reinforcing the idea that customers associate price with perceived quality. On the contrary, deep discounts may trigger skepticism and lead to lower ratings. The team recommended that Amazon needs to focus on product quality over large discounts, especially in high-revenue categories like Smart TVs and home electronics. Marketing strategies should promote bundling quality items like a Smart TV with a sound bar rather than relying solely on pricing tactics. Through the completion of this project, the team gained the skills to clean data, process natural language, conduct advanced visualizations and models, and use R for business analysis. Most importantly, this project demonstrated how structured data analysis can turn raw product and customer review data into strategic insights that inform business decisions.

IST 659: Data Administration Concepts & Database Management

In the Fall of 2023, Gregory Zink instructed a class aimed at definition, development, and management of databases for information systems. Through class sessions and projects students would learn data analysis techniques, data modeling, schema design, query languages, search specifications, data administration concepts and an overview of file organization for databases. In the rapidly evolving world of financial services, individual investors and small firms often face a significant disadvantage due to limited access to high-quality research tools. The IST 659 Trade Genius project was developed to address this gap by designing a comprehensive financial research platform that standardizes data access and insights for retail investors, hedge funds, venture capitalists, and journalists. The platform's core purpose is to empower users with data-driven

research tools to make informed investment decisions without the traditional barriers of cost or technical complexity.

The data utilized in this project was structured in a relational database schema and included user profiles, company financials, market tickers, IPO prices, account types, and platform privileges. The team began by using both physical and conceptual models to design the database. The data was then prepared and queried using SQL in Azure Data Studio. From this point the team was able to build the views, triggers, and functions then optimize it with logic constraints. Finally, using Figma the team was able to create a mockup of the UI/UX to visualize the user experience. However, with the demonstration of the UI presented some potential biases like uneven industry representation or overly restrictive assumptions about account limitations which may not scale well in real world scenarios.

PROGRAM OBJECTIVES

The following objectives were met during the completion of this project:

1. Collect, store, and access data by identifying and leveraging applicable technologies
    a. The team created several conceptual and physical data models, defining the database architecture. Using SQL for indexing, query optimizations, stored procedures, and triggers to demonstrate understanding of backend systems.
2. Create actionable insight across a range of contexts (e.g., societal, business, political), using data and the full data science life cycle
    a. The team discussed analyzing IPO trends, ticker categories, and industry performance to drive equity insights and trading strategies.
3. Communicate insights gained via visualization and analytics to a broad range of audiences (including project sponsors and technical team leads)
    a. This was a central objective of this project as the team created detailed dashboards mockups, UI/UX flows, and real-world use case scenarios tailored for institutional and retail audiences.

PROJECT VISUALIZATION

The following three visuals illustrate the main findings and insights from the project:

1. **Conceptual Database Model**

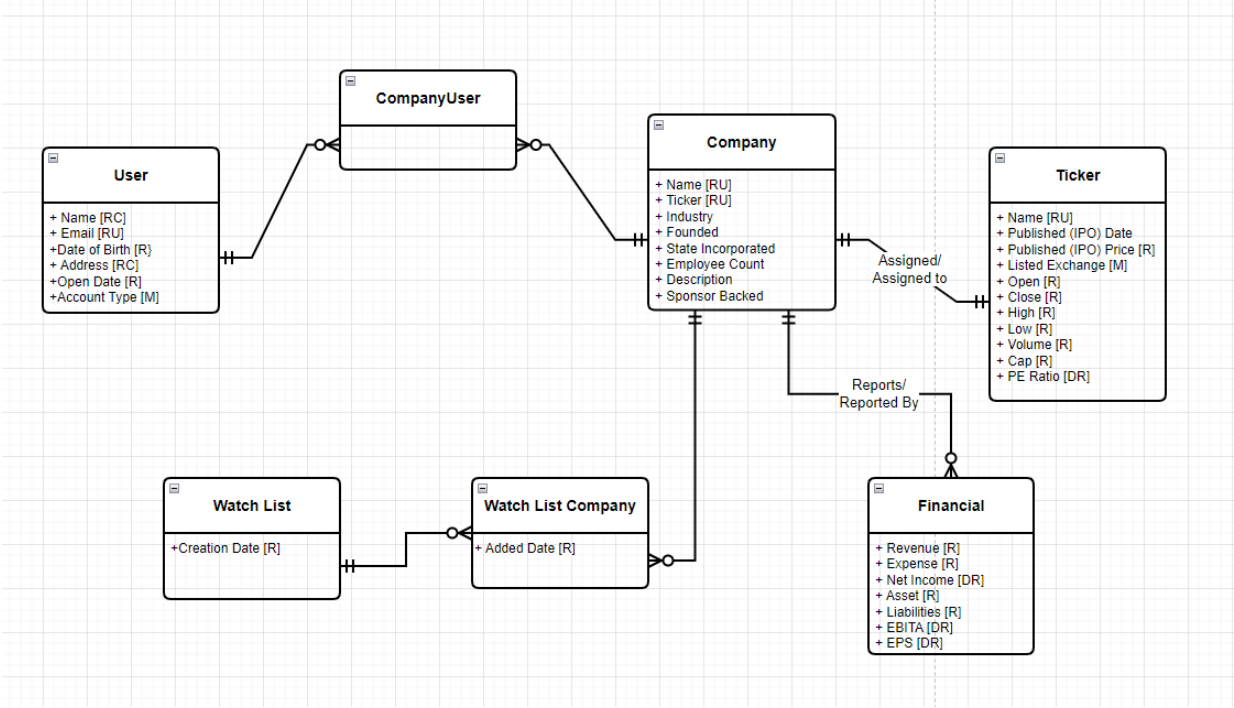


Fig4.1 Trade Genius Conceptual Model

This diagram showcases the high-level relationships between entities such as users, companies, and tickers ensuring normalized data structure aligned with business logic.

2. Company IPO Price Analysis

```
570    -- Q3
571
572    SELECT c.company_name, c.company_industry, t.ticker_ipo_price,
573        avg(ticker_ipo_price) OVER (partition by company_industry) as avg_price,
574        min(ticker_ipo_price) OVER (partition by company_industry) as min_price,
575        max(ticker_ipo_price) OVER (partition by company_industry) as max_price
576    from Companies C
577    JOIN tickers t on t.ticker_name = c.company_ticker
578    where company_industry IN ('Financial Services','Technology')
579    order by company_industry
580
```

Results | Messages

	company_name	company_industry	ticker_ipo_price	avg_price	min_price	max_price
1	Apple Inc.	Technology	22.0000	21.500000	21.0000	22.0000
2	Microsoft Corporation	Technology	21.0000	21.500000	21.0000	22.0000

Fig5.1 SQL Code and Preview of Price Analysis

This result screen highlights minimum, maximum, and average IPO prices across the “Technology” and “Financial Services” industries, helping analysts compare current performance against historical benchmarks.

3. Dashboard Mock-Up

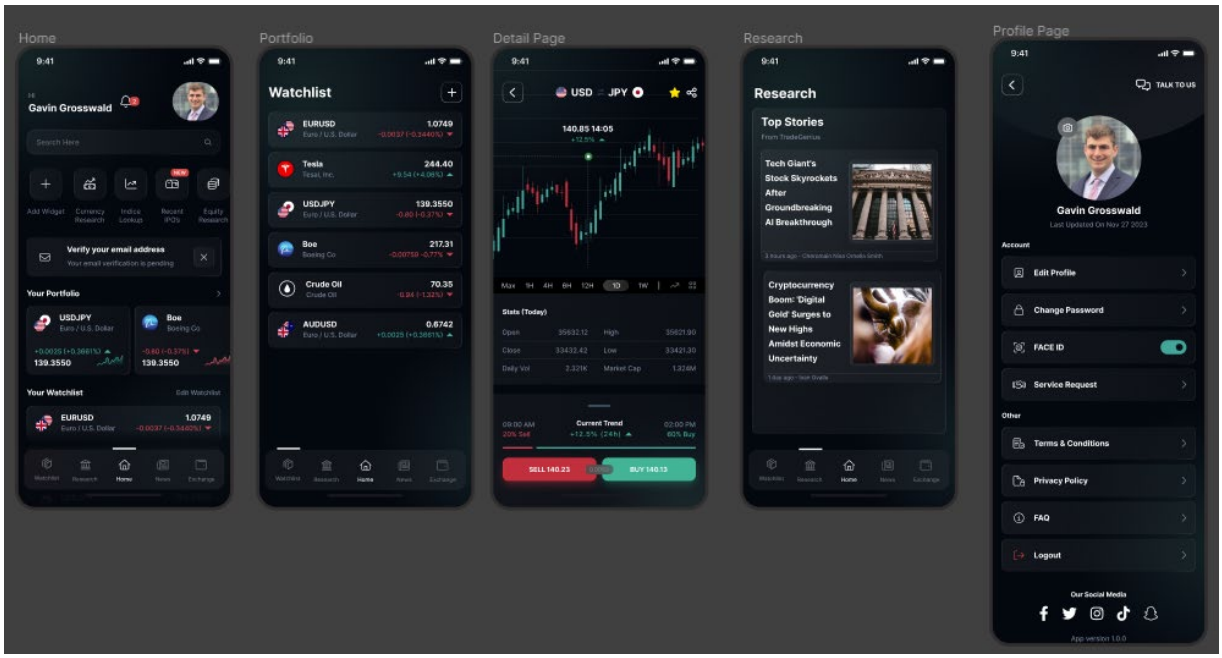


Fig 6.1 Mock-Up of the Mobile Platform

This visual demonstrated the Trade Genius portfolio page UI, showing real-time company insights, equity scores, and watchlist features designed for investor decision-making.

REFLECTION

The Trade Genius project achieved its goal of building a foundational database system and research tool capable of supporting both institutional and retail trading decisions. The mobile platform successfully integrates complex data relationships and business rules like using triggers to limit the different account types while enabling user-defined search functionality through table-valued functions. Some key results included:

- A structured SQL database architecture.
- Optimized queries for retrieving financial performance data.
- Built-in logic for data integrity and platform-specific constraints.
- A user-friendly, modern dashboard prototype.

Some recommendations for the future development of this platform would be to integrate real-time market data via APIs and expand industry datasets for balance. Through the completion of this project, the team gained valuable skills in SQL development, database modeling, query optimization, interface prototyping and agile task management. Most importantly, this project demonstrates the power of data engineering and structured database design to normalize complex financial

research. It highlights how thoughtful backend architecture and SQL driven logic can empower smarter investment decisions and transform how financial intelligence is distributed.

IST 652: Scripting for Data Analysis

In the Spring of 2024, Ying Lin instructed a class aimed at scripting for the data analysis pipeline. Through class sessions and projects, students would learn how to acquire, access and transform data into the forms of structured, semi-structured, and unstructured data. In the US, the average person is paying approximately $1,986 per week to go on vacation. These rising vacation costs highlights a growing need for smarter, budget-conscious choices in travel planning. The objective of this project was to explore how historical Airbnb data from Seattle can be transformed into a data-driven tool that helps both hosts and guests make more informed decisions. The goal was to design a tool that could assist hosts with competitive pricing, feature optimization, and location strategy, while also helping guests search for accommodations with tailored amenities that meet their needs beyond the standard filters.

The core data came from Kaggle and consisted of three interlinked datasets: calendar.csv, listings.csv, and reviews.csv. These included detailed records on listing prices, availability, location, amenities, host characteristics, and guest reviews. Together these categories offered a 360- degree view of the short-term rental market in Seattle. This project followed a multi-step process:

1. **Data Cleaning and Preprocessing**
    a. Each dataset underwent extensive formatting and null handling. Dates were standardized, price values were stripped of symbols and converted to numeric, and text fields were cleansed of null values to ensure compatibility with machine learning models. Non-useful fields like URLs were removed, and newly calculated fields like price per guest were created.
2. **Exploratory Data Analysis (EDA)**
    a. The EDA focused on understanding price distribution, identifying optimal pricing ranges, and detecting relationships between variables using plots, bar charts, and correlation matrices. This phase uncovered useful insights like the average price range of $100-$150, price differences by neighborhood, and amenity prevalence.
3. **Predictive Modeling**

    a. Using random forest regression and decision tree models, the team attempted to predict listing prices based on variables such as bedrooms, bathrooms, location, and review scores. Models were evaluated using metrics like MSE, RMSE, and R-squared.
4. **Tool Development**
    a. **Price Suggestion Tool:** for host and takes input like bedrooms, bathrooms, and zip code to output and estimated competitive price.
    b. **Keyword Search Tool:** for guest to find listings that match specific needs like "kid friendly" or "Jacuzzi" without having to open each listing individually.
    c. **Top Listings Report:** based on aggregated review scores to help users find the best-rated stays in Seattle.

PROGRAM OBJECTIVES

The following objectives were met during the completion of this project:

1. Collect, store, and access data by identifying and leveraging applicable technologies
    a. This objective was central to this project as it sourced data from Kaggle.com in three CSV files (calendar.csv, listings.csv, and reviews.csv).
2. Create actionable insight across a range of contexts (e.g., societal, business, political), using data and the full data science life cycle
    a. Using the data, the team suggested Business insights for Airbnb hosts (e.g., optimal pricing, best neighborhoods) and societal value in terms of improving access to family-friendly, affordable rentals.
3. Apply visualization and predictive models to help generate actionable insight
    a. The project uses bar plots, correlation matrix, and graphs to visualize price ranges, host performance, and review scores. The team then applied random forest and decision tree models for price prediction and trend identification.
4. Use programming languages such as R and Python to support generation of actionable insight
    a. This objective was central to the project as the team used a combination of Python, Pandas, and Jupyter Notebooks for data handling and analysis. Python was imperative in the data preprocessing, model building, keyword searching and visualizations completed.

5. Communication insights gained via visualization and analytics to a broad range of audiences (including project sponsors and technical team leads)
    a. Insights were clearly communicated using graphs that showcased for example price by neighborhood, tables and correlation matrix for both non-technical and technical audiences. The report included a detailed explanation of the models and business values that were accessible to a variety of stakeholders.
6. Apply ethics in the development, use and evaluation of data and predictive models (e.g., fairness, bias, transparency, privacy)
    a. Some ethical considerations that the team noted needing further investigation are the challenges of modeling in an unequal economy, fairness for low-income families, model bias and transparency in pricing tools.

PROJECT VISUALIZATION

The following four visuals illustrate the main findings and insights from the project:

1. **Price Distribution of Airbnb Listings in Seattle**

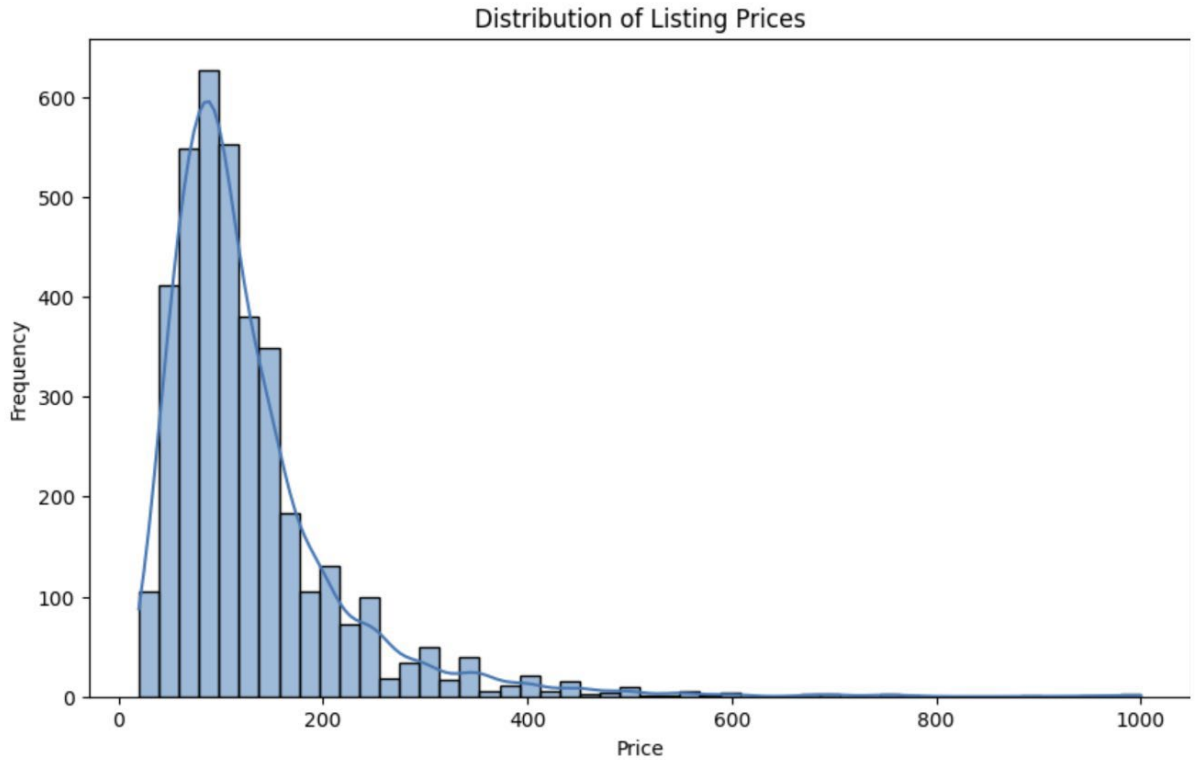


Fig 7.1 Price range of Airbnb Listings in Seattle

This histogram shows that most listings fall within the $100 - $150 range and taper off by early $200, informing strategic pricing decisions for hosts.

2. **Average Prices by Neighborhood**

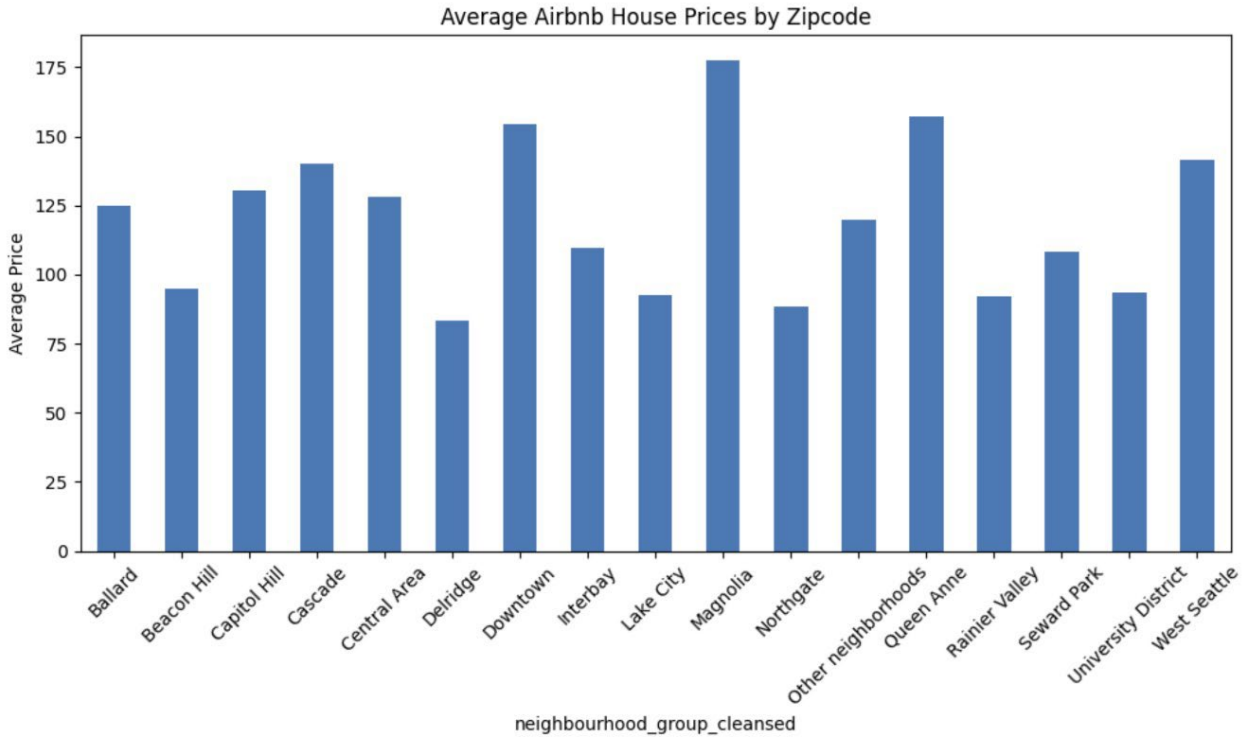


Fig 8.1 Average Rental Price in Neighborhoods in Seattle
A bar chart showcasing the price disparity across neighborhoods. Magnolia emerged as the most expensive and Delridge as the most affordable.

3. **Correlation Matrix of Review and Host Variables**

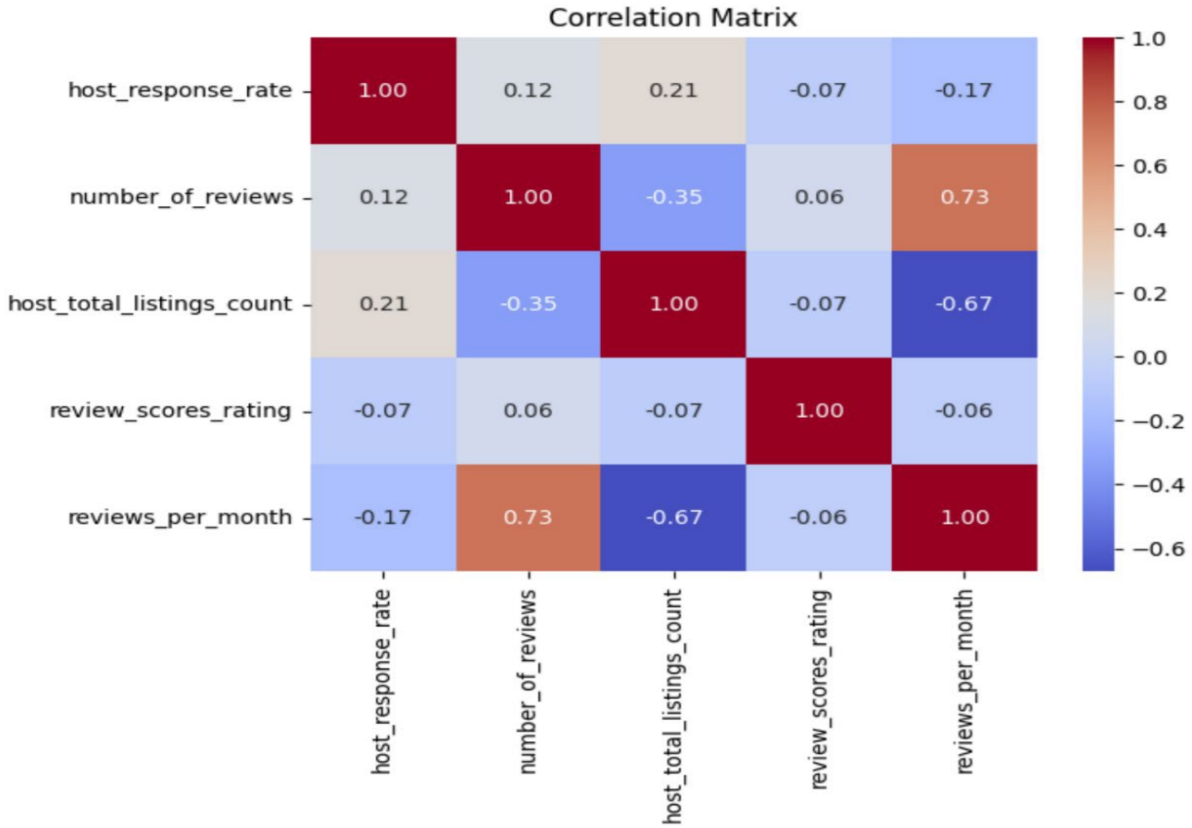


Fig 9.1 Host Review Matrix

This matrix explores the relationship between host behavior like response rate and guest reviews. It revealed a surprising negative correlation between reviews per month and the number of listings a host manages.

4. **Keyword Search Tool**

```
Enter a keyword to search for listings: family/Kid Friendly
Listings containing the keyword 'family/Kid Friendly':

Listing ID: 227636
Name: Large Ballard/Fremont apartment
Host Response Time: within an hour
Bedrooms: 1.0
Bathrooms: 1.0
Beds: 2.0
Price: 120.0
Neighbourhood Cleansed: West Woodland
Minimum Nights: 2
Guests Included: 2
Number of Reviews: 131
Review Scores Rating: 97.0
Listing Link: https://www.airbnb.com/rooms/227636
--------------------------------------------------

Listing ID: 566435
Name: "THE 5-STAR HOUSE"
Host Response Time: within an hour
Bedrooms: 1.0
Bathrooms: 3.5
Beds: 1.0
Price: 65.0
Neighbourhood Cleansed: Crown Hill
Minimum Nights: 2
Guests Included: 2
Number of Reviews: 37
Review Scores Rating: 95.0
Listing Link: https://www.airbnb.com/rooms/566435
--------------------------------------------------

Listing ID: 2197982
Name: Private Studio in Seattle
Host Response Time: within an hour
Bedrooms: 0.0
Bathrooms: 1.0
Beds: 1.0
Price: 80.0
Neighbourhood Cleansed: Maple Leaf
```

Fig 10.1 Guest Keyword/Phrase Search Tool

These listings provide an example of how the keyword tool works in terms of a user provides a must have for their rental and the tool returns a filtered list that fits that metric.

REFLECTION

The Seattle Airbnb project demonstrates how well-organized, cleaned, and analyzed data can provide tangible value in the real estate and tourism industries. While the initial predictive models faced accuracy challenges, as reflected by low R2 values, they laid a solid foundation for future iterations by establishing necessary features, variables, and machine learning techniques. However, it became evident that economic instability and untracked variables like seasonality or macroeconomic factors could impact predictive accuracy. This may be a significant tool improvement needed for future models.

Some recommendations from the team include incorporating real-time market dynamics, expanding the dataset to include more cities for broader insights, and integrating sentiment analysis from review comments. Some key skills learned through the completion of this project included data cleaning in Python, EDA with pandas and matplotlib, model training using scikit-learn, and the application of random forest and decision tree algorithms. Most importantly, this project demonstrated how data-driven tools can be designed to offer custom insights to two very different user types, Airbnb hosts and guests, empowering better decision-making in a competitive travel market.

IST 707: Applied Machine Learning

During that same semester of Spring 2024, Jeremy Bolton taught a class aimed at providing a general overview of industry standard machine learning techniques and algorithms. Through course sessions and projects students were able to focus on building machine learning models, optimization, real-world applications and future directions in the field while gathering hands-on experience with modern data science packages. This project investigates factors influencing NFL game outcomes, specifically whether the favored team wins. Given the unpredictable nature of sports, this study focuses on historical NFL data to analyze correlations between geographic location elevations, weather, stadium type, and game metrics and their effects on outcomes. The overarching goal is to support better decision-making in betting scenarios by identifying conditions that make favorites more or less likely to win.

To answer this, we analyzed data spanning from the 1960s to 2024, combining score data with stadium and weather information. The primary dataset includes historical NFL games scores merged with stadium data and environmental

conditions collected from Kaggle. Secondary datasets included team identifiers, stadium attributes, and betting lines. Various models were applied that aimed to find meaningful patterns across win rates, betting spreads, and environmental features, offering valuable insights for strategic game prediction and betting. Some key steps in the process of investigating the data were:

1. **Data Cleaning & Preparation:** cleaning and merging multiple datasets, formatting due to inconsistencies in team names, and weather data.
2. **Exploratory Data Analysis (EDA):** Generating heat maps, correlation matrices and word cloud to uncover initial trends and relationships
3. **Modeling:** Exploring win percentages and Super Bowl predictions, studying team behavior across different environmental conditions and locations.

This project used R and the Apriori algorithm to focus on the supervised & unsupervised learning, correlation analysis and predictive modeling concepts. However, this investigation was met with a lot of biases and challenges. A couple challenges the team faced were:

1. **Data Sparsity:** Earlier decades had fewer records and missing fields.
2. **Model Limitations:** Many models had a modest predictive accuracy of 50% to 57%.
3. **Unquantifiable Factors:** Player injuries, team morale, and last-minute lineup changes were not measurable data that was included.
4. **Frequency Bias:** Heavily reported teams may be overrepresented in the word clouds analysis.

PROGRAM OBJECTIVES

The following objectives were met during the completion of this project:

1. Collect, store, and access data by identifying and leveraging applicable technologies
    a. For this project, the team obtained and worked with three datasets from Kaggle: spreadspoke_scores.csv, NFL Teams and NFL Stadiums.
2. Create actionable insight across a range of contexts (e.g., societal, business, political), using data and the full data science life cycle
    a. The project itself focused on sports betting, which while fast-growing has varied societal impacts like risk of addiction, economic effects and legality. The full data science pipeline was executed through this project from data collection, to cleaning, exploratory data analysis, modeling then communicating the findings.

3. Apply visualization and predictive models to help generate actionable insight
    a. The project utilized bar plots, heatmaps, and word clouds to show performance under different weather conditions, geographic patterns, and win/loss distributions. The team used several different machine learning models to investigate the efficiency of predicting game outcomes based on historical trends with the hopes of improving betting predictions.
4. Use programming languages such as R and Python to support generation of actionable insight
    a. Programming languages such as Python and R, with libraries of pandas, scikit and ggplot2 were central to analyzing, cleaning and structuring the data obtained from multiple sources. Python assisted with modeling and preprocessing, while R was used for visualizations like heatmaps and bar plots. The dual use of these languages throughout the project demonstrates fluency across both languages for different stages of analysis.
5. Communication insights gained via visualization and analytics to a broad range of audiences (including project sponsors and technical team leads)
    a. The team used different visualizations to clearly communicate win/loss trends, team performance and text analysis to tell a visual story. The team communicated findings for both sports fans and technical analysts, with visual and written explanations tying trends back to betting behavior.
6. Apply ethics in the development, use and evaluation of data and predictive models (e.g., fairness, bias, transparency, privacy)
    a. The team concluded by addressing ethical issues in betting like the potential for addiction, responsible gambling and the overall fairness of predictions. The project noted a lot of cleaning needed as the team had to exclude games without full scoring information or stadium information.

PROJECT VISUALIZATION

The following three visuals illustrate the main findings and insights from the project:

1. **Favorite Win Rate by Stadium (US)**

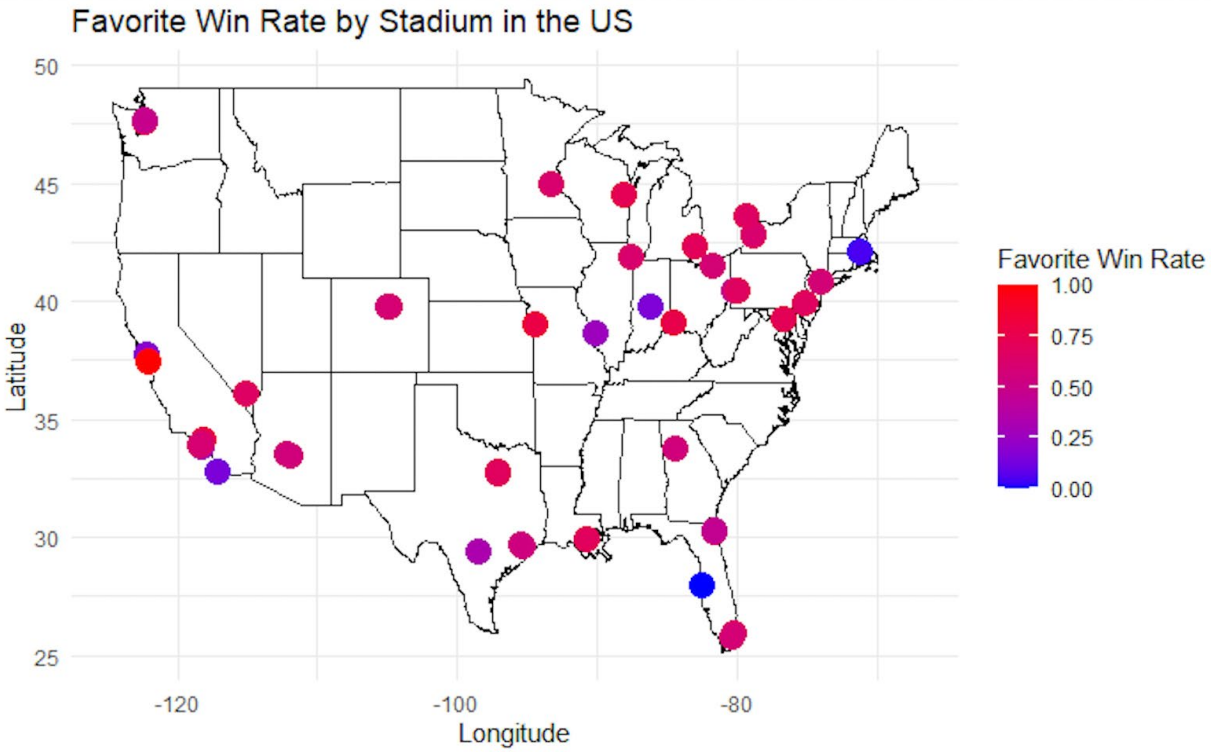


Fig 11.1 Stadium Win Rate Heatmap

This heatmap shows the variance in favorite team win percentages across NFL stadiums, revealing geographic and climate-based patterns. For example, higher win rates in colder climates like Green Bay.

2. **K-Means Cluster Plot – Score & Spread**

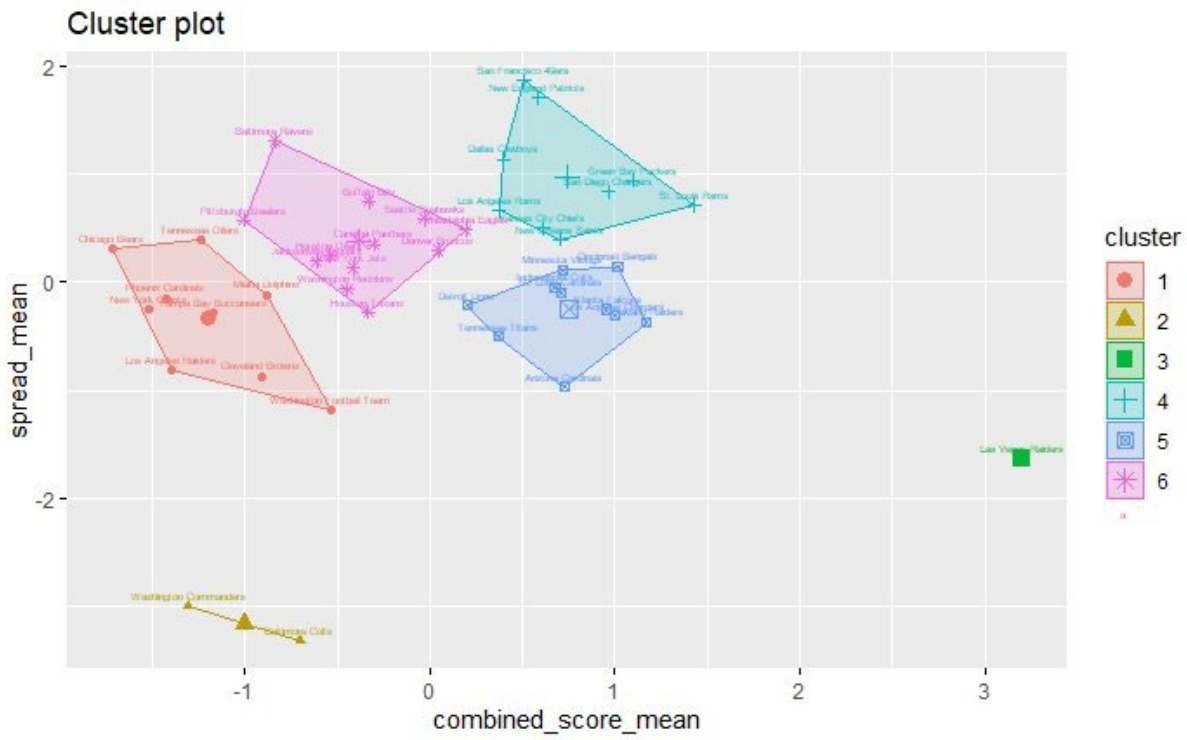


Fig 12.1 Score By Teams Cluster

Visualizes six clusters of teams grouped by average spread and total points scored, highlighting performance tiers. For example, Cluster 4 includes historically dominant teams like the 49ers and Packers.

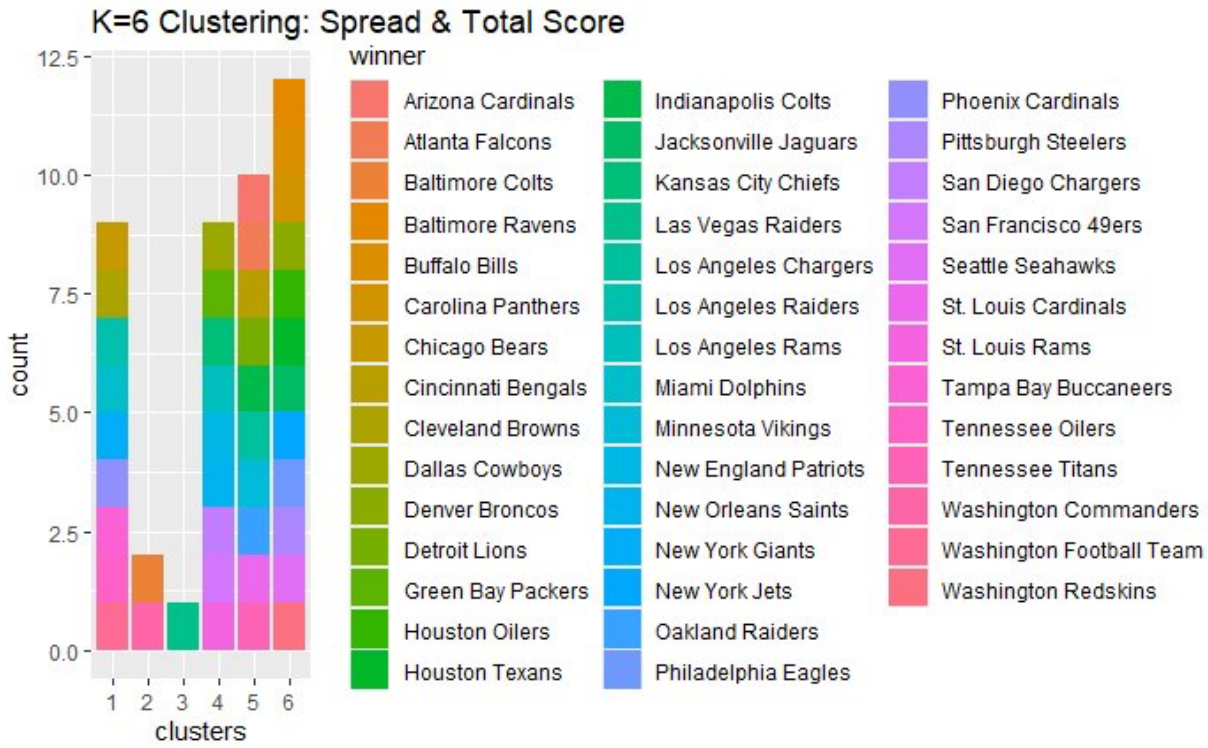


Fig 12.2 Graph identifying the clusters and Teams

3. Apriori Association Rules – Favorite Lost

	lhs		rhs	support	confidence	coverage	lift	count
[1]	{stadium=Gillette Stadium}	=>	{favorite_won=FALSE}	0.01681416	0.9500000	0.01769912	2.055726	190
[2]	{winner=Oakland Raiders}	=>	{favorite_won=FALSE}	0.01699115	0.9846154	0.01725664	2.130631	192
[3]	{team_home=Tennessee Titans}	=>	{favorite_won=FALSE}	0.01557522	0.8461538	0.01840708	1.831011	176
[4]	{winner=Tennessee Titans}	=>	{favorite_won=FALSE}	0.01911504	0.9953917	0.01920354	2.153950	216
[5]	{winner=San Diego Chargers}	=>	{favorite_won=FALSE}	0.02619469	0.9833887	0.02663717	2.127976	296
[6]	{team_home=San Diego Chargers}	=>	{favorite_won=FALSE}	0.02256637	0.8225806	0.02743363	1.780000	255
[7]	{stadium=Qualcomm Stadium}	=>	{favorite_won=FALSE}	0.02274336	0.8263666	0.02752212	1.788193	257
[8]	{team_home=Indianapolis Colts}	=>	{favorite_won=FALSE}	0.02398230	0.8089552	0.02964602	1.750516	271
[9]	{winner=Washington Redskins}	=>	{favorite_won=FALSE}	0.02876106	0.9759760	0.02946903	2.111936	325
[10]	{winner=Indianapolis Colts}	=>	{favorite_won=FALSE}	0.03053097	0.9942363	0.03070796	2.151450	345
[11]	{team_home=New England Patriots}	=>	{favorite_won=FALSE}	0.02991150	0.8733850	0.03424779	1.889937	338
[12]	{winner=New England Patriots}	=>	{favorite_won=FALSE}	0.03973451	0.9868132	0.04026549	2.135387	449
[13]	{team_home=New England Patriots, stadium=Gillette Stadium}	=>	{favorite_won=FALSE}	0.01681416	0.9500000	0.01769912	2.055726	190
[14]	{team_home=San Diego Chargers, winner=San Diego Chargers}	=>	{favorite_won=FALSE}	0.01513274	0.9884393	0.01530973	2.138905	171
[15]	{stadium=Qualcomm Stadium, winner=San Diego Chargers}	=>	{favorite_won=FALSE}	0.01513274	0.9884393	0.01530973	2.138905	171
[16]	{weather_wind_mph=[4,10), winner=San Diego Chargers}	=>	{favorite_won=FALSE}	0.01690265	0.9896373	0.01707965	2.141498	191
[17]	{team_home=San Diego Chargers, stadium=Qualcomm Stadium}	=>	{favorite_won=FALSE}	0.02256637	0.8252427	0.02734513	1.785761	255
[18]	{team_home=San Diego Chargers, weather_temperature=[55,72)}	=>	{favorite_won=FALSE}	0.01654867	0.8165939	0.02026549	1.767045	187
[19]	{team_home=San Diego Chargers, weather_wind_mph=[4,10)}	=>	{favorite_won=FALSE}	0.01778761	0.8170732	0.02176991	1.768083	201
[20]	{stadium=Qualcomm Stadium, weather_temperature=[55,72)}	=>	{favorite_won=FALSE}	0.01663717	0.8173913	0.02035398	1.768771	188
[21]	{stadium=Qualcomm Stadium, weather_wind_mph=[4,10)}	=>	{favorite_won=FALSE}	0.01778761	0.8170732	0.02176991	1.768083	201
[22]	{team_home=Indianapolis Colts, winner=Indianapolis Colts}	=>	{favorite_won=FALSE}	0.01699115	0.9948187	0.01707965	2.152710	192
[23]	{team_home=Indianapolis Colts, weather_wind_mph=[0,4)}	=>	{favorite_won=FALSE}	0.02380531	0.8078078	0.02946903	1.748033	269
[24]	{team_home=Indianapolis Colts, weather_temperature=[72,97]}	=>	{favorite_won=FALSE}	0.02380531	0.8078078	0.02946903	1.748033	269
[25]	{team_home=Washington Redskins, winner=Washington Redskins}	=>	{favorite_won=FALSE}	0.01637168	0.9840426	0.01663717	2.129391	185
[26]	{weather_wind_mph=[0,4), winner=Indianapolis Colts}	=>	{favorite_won=FALSE}	0.02000000	0.9955947	0.02008850	2.154389	226
[27]	{weather_temperature=[72,97], winner=Indianapolis Colts}	=>	{favorite_won=FALSE}	0.02159292	0.9959184	0.02168142	2.155090	244
[28]	{over_under_line=[44,63.5], winner=Indianapolis Colts}	=>	{favorite_won=FALSE}	0.01707965	1.0000000	0.01707965	2.163922	193
[29]	{team_away=New England Patriots, winner=New England Patriots}	=>	{favorite_won=FALSE}	0.01725664	0.9948980	0.01734513	2.152881	195
[30]	{team_home=New England Patriots, winner=New England Patriots}	=>	{favorite_won=FALSE}	0.02247788	0.9806950	0.02292035	2.122147	254
[31]	{team_home=New England Patriots, weather_temperature=[-6,55)}	=>	{favorite_won=FALSE}	0.01575221	0.8989899	0.01752212	1.945344	178
[32]	{team_home=New England Patriots, weather_wind_mph=[10,40]}	=>	{favorite_won=FALSE}	0.01690265	0.8488889	0.01991150	1.836929	191
[33]	{weather_temperature=[-6,55), winner=New England Patriots}	=>	{favorite_won=FALSE}	0.01761062	0.9754902	0.01805310	2.110885	199
[34]	{spread_favorite=[-26.5,-6.5), winner=New England Patriots}	=>	{favorite_won=FALSE}	0.01716814	1.0000000	0.01716814	2.163922	194
[35]	{weather_wind_mph=[10,40], winner=New England Patriots}	=>	{favorite_won=FALSE}	0.01796460	0.9712919	0.01849558	2.101800	203
[36]	{schedule_season=[2.01e+03,2.02e+03], winner=New England Patriots}	=>	{favorite_won=FALSE}	0.01504425	1.0000000	0.01504425	2.163922	170
[37]	{schedule_season=[2e+03,2.01e+03), winner=New England Patriots}	=>	{favorite_won=FALSE}	0.01504425	0.9826590	0.01530973	2.126397	170
[38]	{score_away=[24,59], winner=New England Patriots}	=>	{favorite_won=FALSE}	0.01522124	0.9942197	0.01530973	2.151414	172
[39]	{score_home=[27,70], winner=New England Patriots}	=>	{favorite_won=FALSE}	0.01628319	0.9839572	0.01654867	2.129207	184
[40]	{over_under_line=[44,63.5], winner=New England Patriots}	=>	{favorite_won=FALSE}	0.01823009	0.9951691	0.01831858	2.153468	206
[41]	{team_home=San Diego Chargers, stadium=Qualcomm Stadium, winner=San Diego Chargers}	=>	{favorite_won=FALSE}	0.01513274	0.9884393	0.01530973	2.138905	171
[42]	{team_home=San Diego Chargers, stadium=Qualcomm Stadium, weather_temperature=[55,72)}	=>	{favorite_won=FALSE}	0.01654867	0.8165939	0.02026549	1.767045	187
[43]	{team_home=San Diego Chargers, stadium=Qualcomm Stadium, weather_wind_mph=[4,10)}	=>	{favorite_won=FALSE}	0.01778761	0.8170732	0.02176991	1.768083	201
[44]	{team_home=Indianapolis Colts, weather_wind_mph=[0,4), winner=Indianapolis Colts}	=>	{favorite_won=FALSE}	0.01690265	0.9947917	0.01699115	2.152651	191
[45]	{team_home=Indianapolis Colts, weather_temperature=[72,97], winner=Indianapolis Colts}	=>	{favorite_won=FALSE}	0.01690265	0.9947917	0.01699115	2.152651	191
[46]	{team_home=Indianapolis Colts, weather_temperature=[72,97], weather_wind_mph=[0,4)}	=>	{favorite_won=FALSE}	0.02380531	0.8078078	0.02946903	1.748033	269
[47]	{weather_temperature=[72,97], weather_wind_mph=[0,4), winner=Indianapolis Colts}	=>	{favorite_won=FALSE}	0.01955752	0.9954955	0.01964602	2.154174	221
[48]	{team_home=Indianapolis Colts, weather_temperature=[72,97], weather_wind_mph=[0,4), winner=Indianapolis Colts}	=>	{favorite_won=FALSE}	0.01690265	0.9947917	0.01699115	2.152651	191

Fig 13.1 Rules for Favorite Lost

Demonstrates strong rules for when favorites tend to lose to aid bettors in identifying potential upsets.

REFLECTION

This project applied a wide range of machine learning techniques to extract insights from historical NFL game data. Models like decision trees and random forests showed that cumulative wins are a strong predictor of win percentage, while clustering methods revealed that teams from specific geographic areas perform consistently well in certain weather conditions. Teams like the Miami Dolphins and Tampa Bay Buccaneers excelled in high temperature and humidity, while teams like the Chiefs showed better scoring performance in windier environments. Association rule mining uncovered strong rules that link specific team-stadium combinations to higher or lower likelihoods of the favorite winning. However, these models also highlighted the inherent unpredictability of sports in that the random forest predicted a 49ers Super Bowl win in 2024 based on historical factors, but the real-world outcome deviated.

Support vector machines yielded the best predictive accuracy at approximately 56.7%, suggesting a nonlinear relationship in the data. Even then this performance illustrates the ceiling of current models using only historical team, stadium, and weather date. To improve the model, future work should incorporate player-level metrics, injury reports, and real-time updates to improve prediction performance. Overall, the project provided meaningful insights into the multifactorial nature that is NFL outcomes. It also demonstrated the limits of historical data in predicting inherently dynamic and human-driven events, offering a balanced perspective on the power and boundaries of machine learning in sports analytics.

IST 692: Responsible AI

In the Fall semester of 2024, Jasmina Tacheva taught a class aimed at providing students with the critical skills necessary to discuss and evaluate more just and equitable AI models. Through class sessions and the final project, students leveraged Python or R packages to build these models. This project, Unveiling the Bias, was created to analyze and address representational harms in images generated by OpenAI's DALL-E and MetaAI' s Imagine. As generative AI becomes more prevalent in content creation, concerns about how these tools depict race, ethnicity, gender, and culture have grown. The core problem investigated in this project was the tendency of these models to reinforce stereotypes, exoticize non-Western identities, and erase cultural nuances, thereby contributing to social and ethical harm.

The dataset was generated by prompting both AI systems with culturally loaded terms such as “A Traditional Indian Woman” or “A Rural African Village”. These controlled prompts were intentionally selected to examine how AI interprets different identities across cultural, socioeconomic, and geographical contexts. The project followed a structured approach:

1. **Data Generation**: Using prompt-based query requests in AI Image generators.
2. **Output Analysis**: Searching visually and contextually for harmful patterns such as cultural misrepresentation, stereotypes, and bias.
3. **Comparison**: Compare both platforms to assess variance in bias and inclusivity.
4. **Conclusion**: Documenting harm then proposing solutions and ethical guardrails.

Concepts like exoticism, cultural erasure, and stereotype reinforcement were central to the analytical framework. Through the completion of this project, biases were found in both the training data and output logic of the models. DALL-E frequently depicted non-Western subjects through outdated or overly stylized visuals, and reinforced cliches. On the other hand, Imagine sometimes defaulted to Eurocentric standards in terms of lightening skin tones and depicting Western attire regardless of prompt context. This project also emphasized the role of missing or skewed data in model training, revealing how underrepresentation in datasets leads to inaccurate, reductive portrayals. While no programming was used directly to build models, the team used critical analysis and visual content review as the primary methodology that allowed for an interpretive, interdisciplinary review of AI’s sociocultural impact.

PROGRAM OBJECTIVES

The following objectives were met during the completion of this project:

1. Collect, store, and access data by identifying and leveraging applicable technologies
    a. This project leveraged DALL-E by OpenAI and Imagine by Meta AI to collect visual outputs based on controlled, prompt-based queries. These prompts served as experimental inputs used to demonstrate thoughtful use of emerging technology to generate and analyze image-based data.
2. Create actionable insight across a range of contexts (e.g., societal, business, political), using data and the full data science life cycle

    a. This project explored societal impacts, especially around representation fairness, and inclusion in AI-generated media. Some actionable insights included identifying specific patterns of misrepresentation, exoticism, and stereotype propagation.
3. Communication insights gained via visualization and analytics to a broad range of audiences (including project sponsors and technical team leads)
    a. The project narrative clearly communicates bias issues in AI systems using relatable examples and image-based outputs. It focuses on breaking down complex ethical issues in an accessible but structured way which is ideal for developers, content creators and social scientists alike. With the use of side-by-side comparisons, analysis by prompt and plain language interpretation, this report is inclusive and easily digestible by all readers alike.
4. Apply ethics in the development, use and evaluation of data and predictive models (e.g., fairness, bias, transparency, privacy)
    a. This program objective was not only fully met but central to the project. The project tackled ethical AI challenges including representational harm, cultural misappropriation, stereotype reinforcement, gender bias, racial bias and erasure. The team ended with offers of detailed strategies for ethical improvements within AI including data diversification, user feedback loops, blind testing and fairness safeguards.

PROJECT VISUALIZATION

The following three visuals illustrate the main findings and insights from the project:

1. **DALL-E vs Imagine: A modern Indian Family**

Fig 14.1 DALL-E Modern Indian Family

Fig 14.2 Imagine with Meta AI Modern Indian Family

Visual Summary: While the systems offered different depictions of family structures, the common thread was consistent: women were portrayed in tradition Indian attire, while men, family size, and backgrounds varied more significantly.

Insight: This highlights a gendered cultural bias in both AI systems, where female identities are more rigidly tied to traditional markers of ethnicity, reinforcing the stereotype that women must visually embody "authentic" culture.

2. DALL-E vs Imagine: A rural African Village

Fig 15.1 DALL-E Rural African Village

Fig 15.2 Imagine with MetaAI Rural African Village

Visual Summary: DALL-E focused on huts and rural poverty, while Imagine romanticized village life, both ignoring the diversity and modern development across African communities.

Insight: This reflects how AI systems reduce African life to outdated, monolithic images, perpetuating colonial-era stereotypes of underdevelopment and simplicity.

3. **DALL-E vs Imagine: A Black American CEO**

Fig 16.1 DALL-E Black American CEO

Fig 16.2 Imagine with MetaAI Back American CEO

Visual Summary: DALL-E provided a professional, neutral image, while Imagine showed mostly male CEOs in high-status settings and a single female CEO in a less prestigious office, styled in Eurocentric ways.

Insight: This illustrates intersectional bias, where race and gender combine to produce limited portrayals of leadership, reinforcing norms about who "looks" like a CEO. In these systems portrayal a Black American CEO must be a man or a woman with flat ironed hair in a suit.

REFLECTION

The project revealed that both DALL-E and Imagine carry inherent biases in how they depict race, gender, and culture. These systems often rely on narrow, stereotypical visual cues such as associating Indian women strictly with traditional dress or portraying African villages as impoverished ad static. While Imagine tended to show slightly more diverse outputs than DALL-E, both models exhibited Eurocentric and gender biases, especially in professional or cultural contexts. Some key recommendations for improvement include expanding and diversifying training datasets to include regional, socioeconomic, and gender variation. Technical solutions like removing race or gender identifiers from prompt influence, ethical flagging systems, and user-feedback loops could enhance both fairness and adaptability. From an educational standpoint, this project deepened the understanding of AI bias, critical media analysis, and the ethical responsibilities tied to machine learning systems. The skills developed particularly in visual literacy, ethical critique, and sociotechnical evaluation are essential for designing future AI systems that are as inclusive and representative as they are innovative.

Conclusion

Across the sequence of courses in the Applied Data Science Program, I gained practical, end-to-end experience in collecting, storing, analyzing, and communicating data-driven insights. Combined, these projects demonstrate my capability to:

- Select and utilize appropriate data storage and access technologies.
- Extract and communicate actionable insights in business and societal contexts.

- Build predictive models and visualizations that support decision-making.
- Leverage both R and Python effectively in data workflows.
- Articulate findings to both technical and non-technical stakeholders
- Apply ethical reasoning to the design and evaluation of data-driven systems

Most importantly, I developed an awareness of the social implications of data science, learning how to apply ethical principles in evaluating both models and their outcomes. As data increasingly shapes every facet of our lives, the responsibility of data scientists extends beyond technical skills. It demands critical thinking, domain understanding, understanding of human behavior, and ethical awareness. This portfolio reflects not only my technical proficiency, but my preparedness to contribute meaningfully to the evolving landscape that is data science.

References

Smith, C. (2023a). *MSADS-Portfolio/IST 659 - data admin concepts & database management* . GitHub. https://github.com/CherSmith39/MSADS-Portfolio/tree/a521d654f9f33402c0eb13a53c69e82cd1cf4c64/IST%20659%20-%20Data%20Admin%20Concepts%20%26%20Database%20Management

Smith, C. (2023b). *MSADS-Portfolio/IST 687 - intro to data science* . GitHub. https://github.com/CherSmith39/MSADS-Portfolio/tree/a521d654f9f33402c0eb13a53c69e82cd1cf4c64/IST%20687%20-%20Intro%20to%20Data%20Science

Smith, C. (2024a). *MSADS-Portfolio/IST 652 - scripting for data analysis*. GitHub. https://github.com/CherSmith39/MSADS-Portfolio/tree/a521d654f9f33402c0eb13a53c69e82cd1cf4c64/IST%20652%20-%20Scripting%20for%20Data%20Analysis

Smith, C. (2024b). *MSADS-Portfolio/IST 692 - responsible AI* . GitHub. https://github.com/CherSmith39/MSADS-Portfolio/tree/a521d654f9f33402c0eb13a53c69e82cd1cf4c64/IST%20692%20-%20Responsible%20AI

Smith, C. (2024c). *MSADS-Portfolio/IST 707 - applied machine learning* . GitHub. https://github.com/CherSmith39/MSADS-Portfolio/tree/a521d654f9f33402c0eb13a53c69e82cd1cf4c64/IST%20707%20-%20Applied%20Machine%20Learning

Talagala, N. (2022, October 12). *Data as the new oil is not enough: Four principles for avoiding data fires*. Forbes. https://www.forbes.com/sites/nishatalagala/2022/03/02/data-as-the-new-oil-is-not-enough-four-principles-for-avoiding-data-fires/