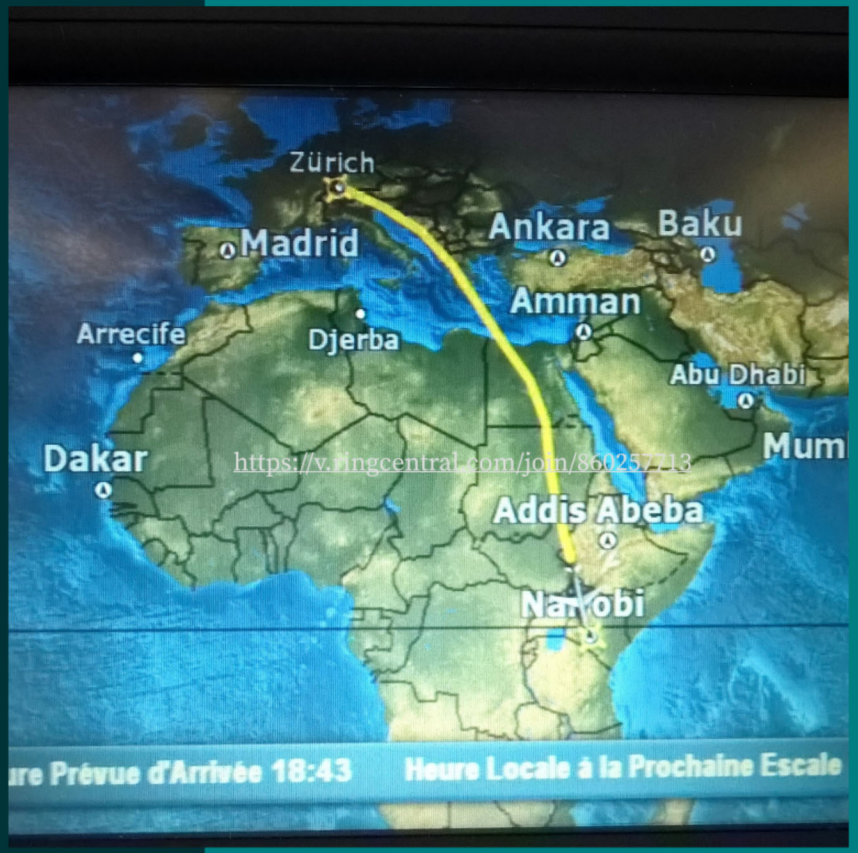
Bias, Bets, and Bytes

A Reflection on Data, Decisions, and Ethics

Introduction to I-School Journey

Data is the New Oil



Effective data transformation is essential for unlocking its true potential, allowing organizations to make informed decisions.

How can we transform raw data into valuable insights?

Data, in its raw form, lacks usability. To derive meaningful conclusions, data must undergo thorough cleaning and processing. The expertise gained from the Master's in Applied Data Science equips individuals with the skills necessary to handle different types of data effectively across various domains such as Retail, Real Estate, Sports, Finance, and Artificial Intelligence.

How can we transform raw data into valuable insights?

Effective data transformation is essential for unlocking its true potential, allowing organizations to make informed decisions.

How can we transform raw data into valuable insights?

Effective data transformation is essential for unlocking its true potential, allowing organizations to make informed decisions.

How can we transform raw data into valuable insights?

Data, in its raw form, lacks usability. To derive meaningful conclusions, data must undergo thorough cleaning and processing. The expertise gained from the Master's in Applied Data Science equips individuals with the skills necessary to handle different types of data effectively across various domains such as Retail, Real Estate, Sports, Finance, and Artificial Intelligence.

Transforming Data Into Actionable Insight

DATA MASTERY

Harness modern practices in data science for effective insights and ethical analysis.

DATA MANAGEMENT

Collect, store, and access data through cutting-edge technologies to facilitate seamless operations and data flow.

INSIGHT GENERATION

Utilize the comprehensive data science lifecycle to generate and communicate valuable insights derived from data analysis.

MODEL BUILDING

Develop and assess predictive models utilizing programming languages such as R and Python for data-driven decision making.

ETHICAL ANALYSIS

Implement ethical principles throughout data analysis and model development to ensure integrity and responsibility in data usage.

DATA MASTERY

Harness modern practices in data science for effective insights and ethical analysis.

DATA MANAGEMENT

Collect, store, and access data through cutting-edge technologies to facilitate seamless operations and data flow.

INSIGHT GENERATION

Utilize the comprehensive data science lifecycle to generate and communicate valuable insights derived from data analysis.

MODEL BUILDING

Develop and assess predictive models utilizing programming languages such as R and Python for data-driven decision making.

ETHICAL ANALYSIS

Implement ethical principles throughout data analysis and model development to ensure integrity and responsibility in data usage.

DATA MANAGEMENT

Collect, store, and access data through cutting-edge technologies to facilitate seamless operations and data flow.

INSIGHT GENERATION

Utilize the comprehensive data science lifecycle to generate and communicate valuable insights derived from data analysis.

MODEL BUILDING

Develop and assess predictive models utilizing programming languages such as R and Python for data-driven decision making.

ETHICAL ANALYSIS

Implement ethical principles throughout data analysis and model development to ensure integrity and responsibility in data usage.

DATA MASTERY

Harness modern practices in data science for effective insights and ethical analysis.

DATA MANAGEMENT

Collect, store, and access data through cutting-edge technologies to facilitate seamless operations and data flow.

INSIGHT GENERATION

Utilize the comprehensive data science lifecycle to generate and communicate valuable insights derived from data analysis.

MODEL BUILDING

Develop and assess predictive models utilizing programming languages such as R and Python for data-driven decision making.

ETHICAL ANALYSIS

Implement ethical principles throughout data analysis and model development to ensure integrity and responsibility in data usage.

Applied Data Science in Action

IST 687

Analyzed market data to identify sales patterns, leading to enhanced strategic initiatives.

IST 687

Analyzed market data to identify sales patterns, leading to enhanced strategic initiatives.

IST 659

Developed a mobile application tailored for finance professionals, delivering actionable insights.

IST 652

Implemented machine learning algorithms to refine pricing models for short-term rentals.

IST 707

Built predictive models to enhance accuracy in NFL sports betting.

IST 692

Examined ethical considerations in AI, focusing on fairness and bias in advanced models.

Bias, Bets, and Bytes

A Reflection on Data, Decisions, and Ethics

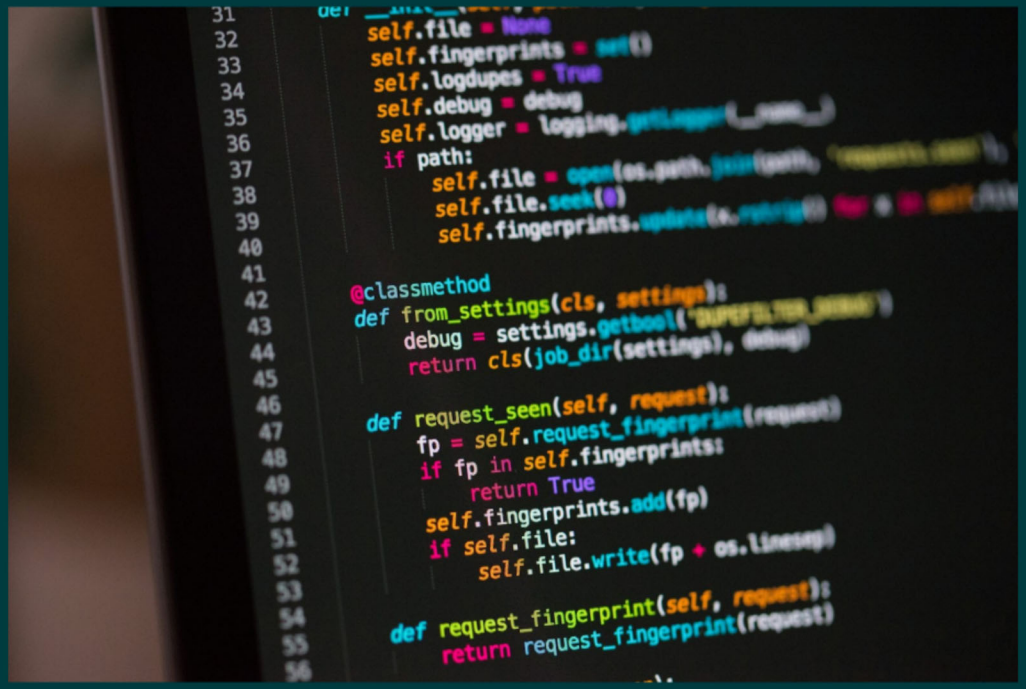

A Journey through the Mind of a Data Scientist
Thank You
Project Highlights
IST 652: Airbnb Data - Seattle
A data-driven tool using Seattle Airbnb data to help both hosts and guests make smarter rental decisions. The project focused on transforming structured and unstructured data from Kaggle into actionable insights through data cleaning, exploratory analysis, and predictive modeling. Using Python and machine learning, the team created pricing and keyword search tools to optimize host strategies and enhance the guest experience.
More Info
IST 687: Amazon Market Analysis
More Info
IST 707: NFL - USA Football Cities
Explored the use of machine learning to predict NFL game outcomes, specifically whether the favored team would win. Using data from the 1960s to 2024, the team analyzed factors like stadium conditions, weather, and game metrics. Models were built using R and Python, applying techniques like clustering, association rules, and supervised learning to uncover betting trends and environmental influences on game outcomes.
More Info
IST 659: TradeGenius
More Info
IST 692: DALL-E & Imagine
More Info

More Info

IST 687: Amazon Market Analysis

Amazon's 2020 sales data from the Indian market was analyzed, focusing on identifying trends in pricing, product performance, and customer satisfaction. The dataset, sourced from Kaggle, included 1,462 cleaned observations and 18 variables. Using the R programming language, the team applied techniques such as data cleaning, regression modeling, and text mining to uncover actionable business insights.

Program Objectives Met



Collected and cleaned real-world data

Used R to model, visualize, and analyze patterns

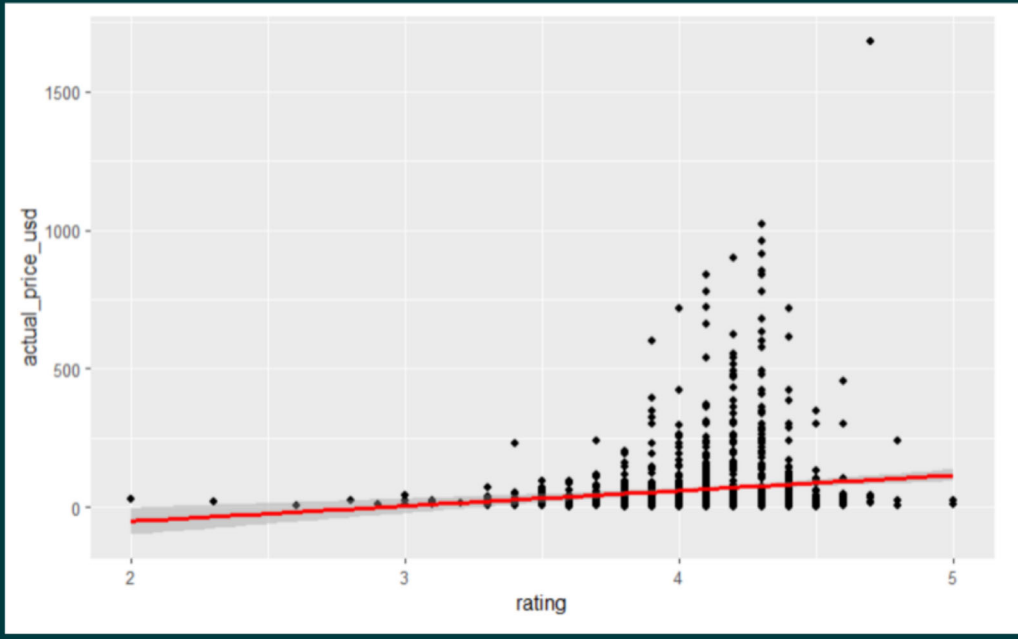
Communicated insights to support business decisions

Visual Insights

```
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
```

Price vs Rating Linear Regression Model

Price vs Rating Linear Regression Plot



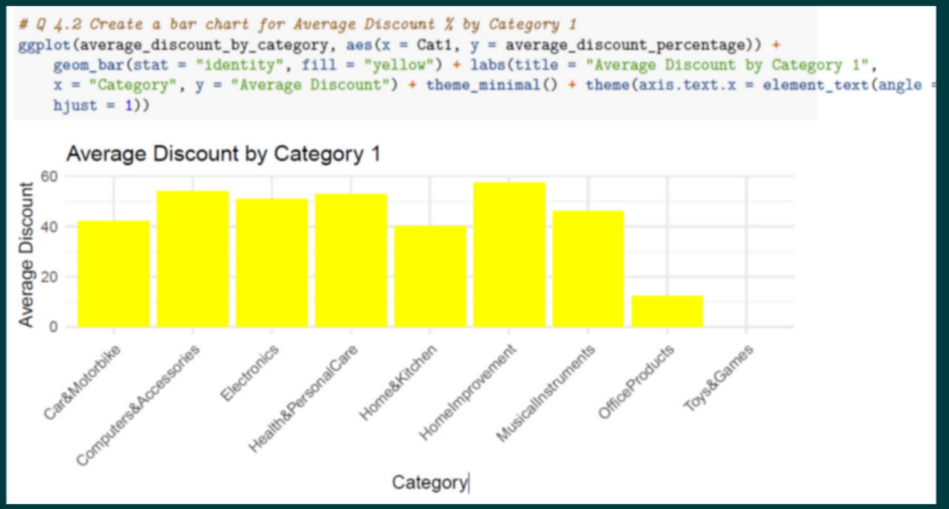
Visual Insights - Cont.

```
## # 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%
```

Average Discount Percentage by Category Table

Average Discount Percentage by Category Bar Chart

```
# 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))
```



Limitations & Observations

Imbalance in product categories (electronics overrepresented)

Few low-rated reviews (only 7 < 3 stars)

Data limited to Indian Amazon market

Reflections & Recommendations

Focus on quality, not just discounting

Promote smart product bundling strategies

Gained full-cycle experience with R and business analytics

Bias, Bets, and Bytes

A Reflection on Data, Decisions, and Ethics

IST 659: TradeGenius

More Info

Trade Genius, a SQL-based financial research platform designed to support smarter investment decisions. Using a relational database schema, the team modeled and queried financial data—such as user profiles, IPO prices, and market tickers—then enforced business logic with triggers and functions. A Figma-based UI mockup showcased how the platform delivers insights to both retail and institutional investors, aiming to reduce technical and cost barriers in financial research.

Program Objectives Met

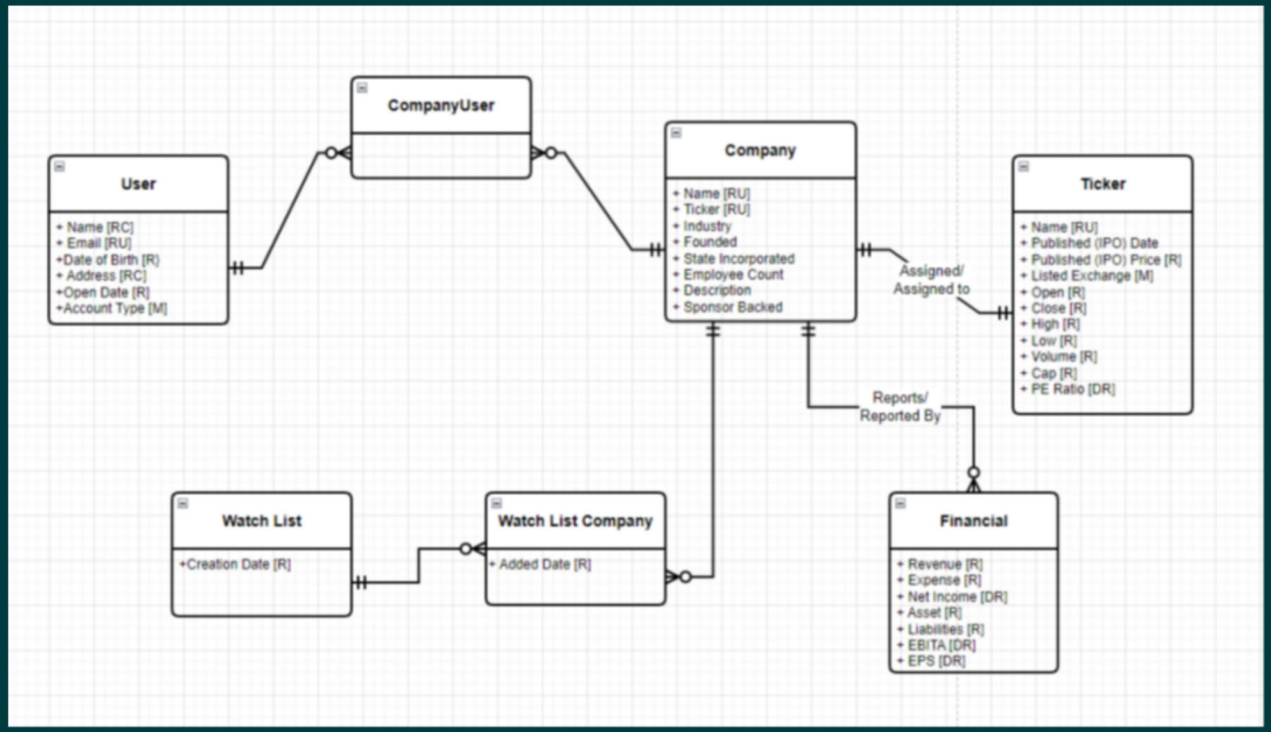
Built a normalized database using conceptual and physical models

Queried, optimized, and managed data using SQL

Analyzed IPO trends and industry performance for trading insights

Designed a UI/UX prototype tailored for financial audiences

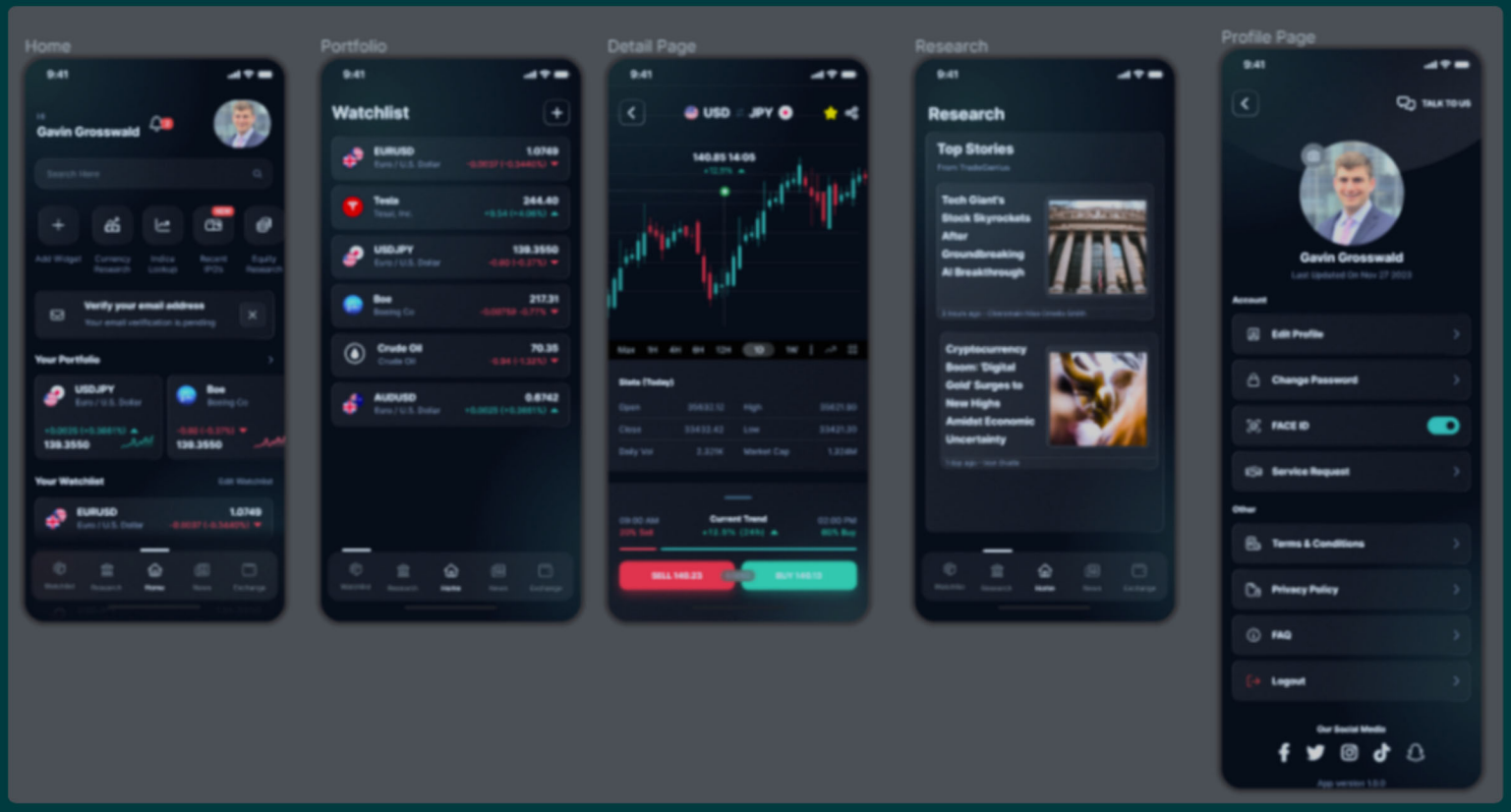
Visual Highlights



Conceptual Model

Defined entity relationships for scalable storage

Visual Highlights - Cont.



Previewed investor dashboards with real-time features

Limitations in the Platform Design

Industry data imbalance (tech and finance overrepresented)

Account limitations may be too rigid for real-world scaling

No real-time market integration in prototype phase

Reflections & Recommendations

Mastered SQL logic, query optimization, and data integrity controls

Developed backend-first design for financial systems

Built tools that promote equitable access to financial intelligence

Highlighted how strong database design powers scalable insights

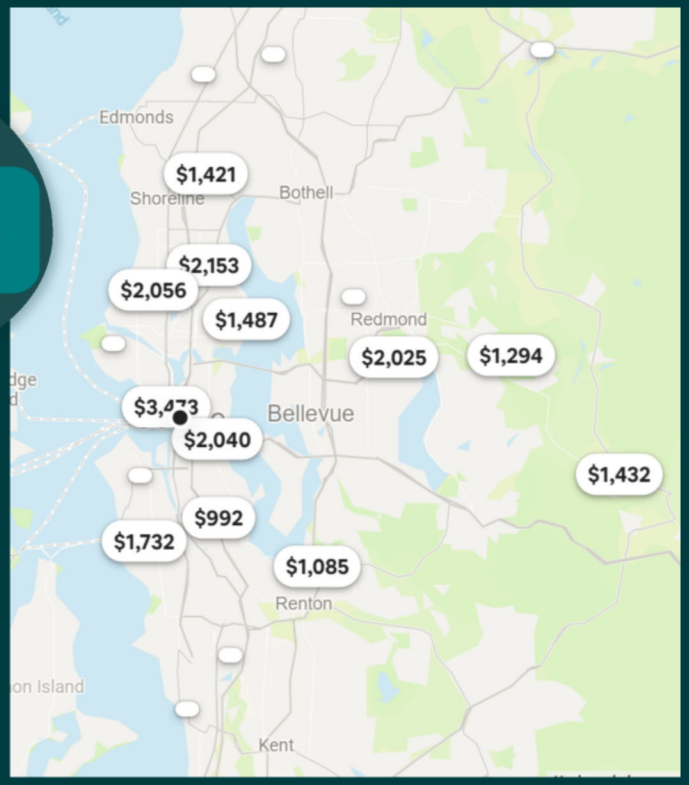
Bias, Bets, and Bytes

A Reflection on Data, Decisions, and Ethics

IST 652: Airbnb Data - Seattle

A data-driven tool using Seattle Airbnb data to help both hosts and guests make smarter rental decisions. The project focused on transforming structured and unstructured data from Kaggle into actionable insights through data cleaning, exploratory analysis, and predictive modeling. Using Python and machine learning, the team created pricing and keyword search tools to optimize host strategies and enhance the guest experience.

More Info



Program Objectives Met

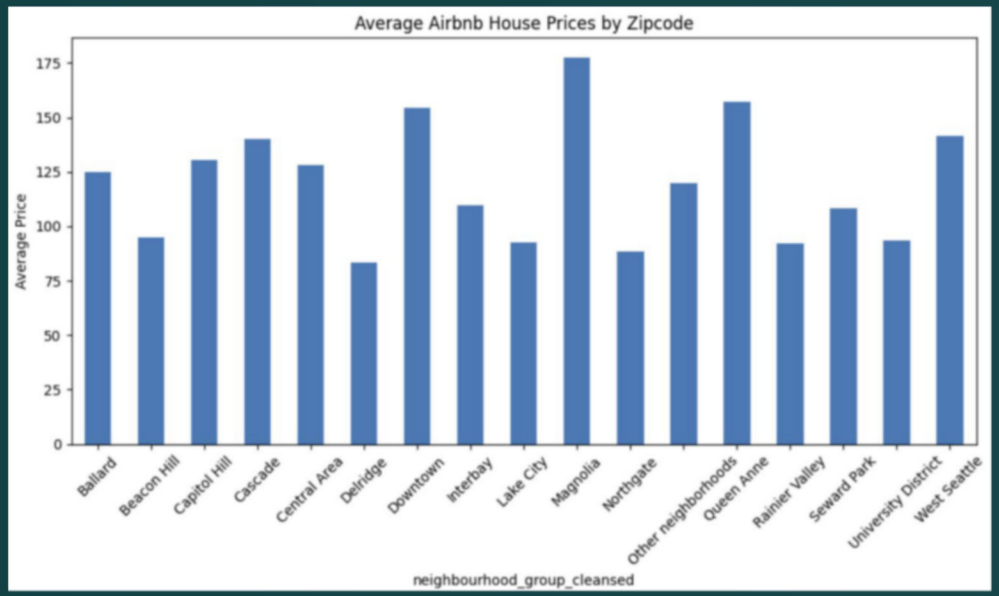
Sourced and transformed three Airbnb datasets from Kaggle

Built pricing models and search tools using Python and ML

Created visualizations for technical and non-technical users

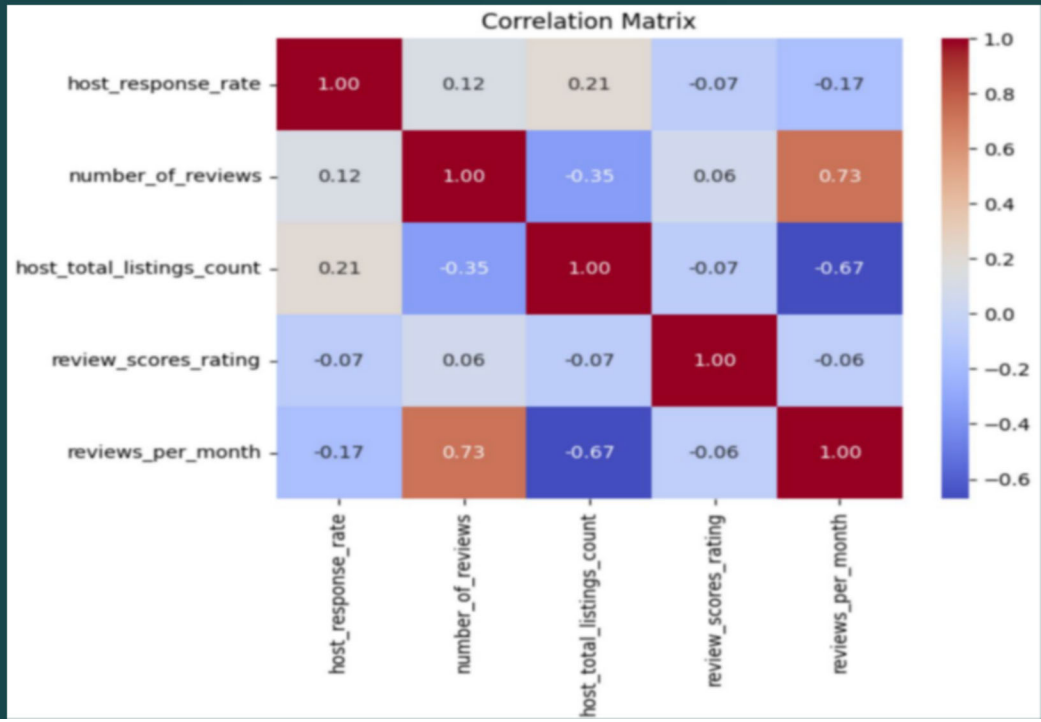
Addressed ethical concerns like fairness and model bias

Visual Insights



Rental Price By Neighborhoods

Host Review Correlation Matrix



Visual Insights - Cont.

```
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
```

Keyword/Phrase Search Tool

Limitations & Ethical Considerations

Low R^2 values in early pricing models

Lacked seasonal and macroeconomic factors

Data limited to one city (Seattle)

Need for fairness across income brackets

Reflections & Recommendations

Built custom tools for hosts and guests

Learned Python scripting, EDA, and ML modeling

Discovered pricing trends and listing quality signals

Demonstrated real-world value in travel tech analytics

Bias, Bets, and Bytes

A Reflection on Data, Decisions, and Ethics

IST 707: NFL - USA Football Cities

Explored the use of machine learning to predict NFL game outcomes, specifically whether the favored team would win. Using data from the 1960s to 2024, the team analyzed factors like stadium conditions, weather, and game metrics. Models were built using R and Python, applying techniques like clustering, association rules, and supervised learning to uncover betting trends and environmental influences on game outcomes.

More Info

Program Objectives Met

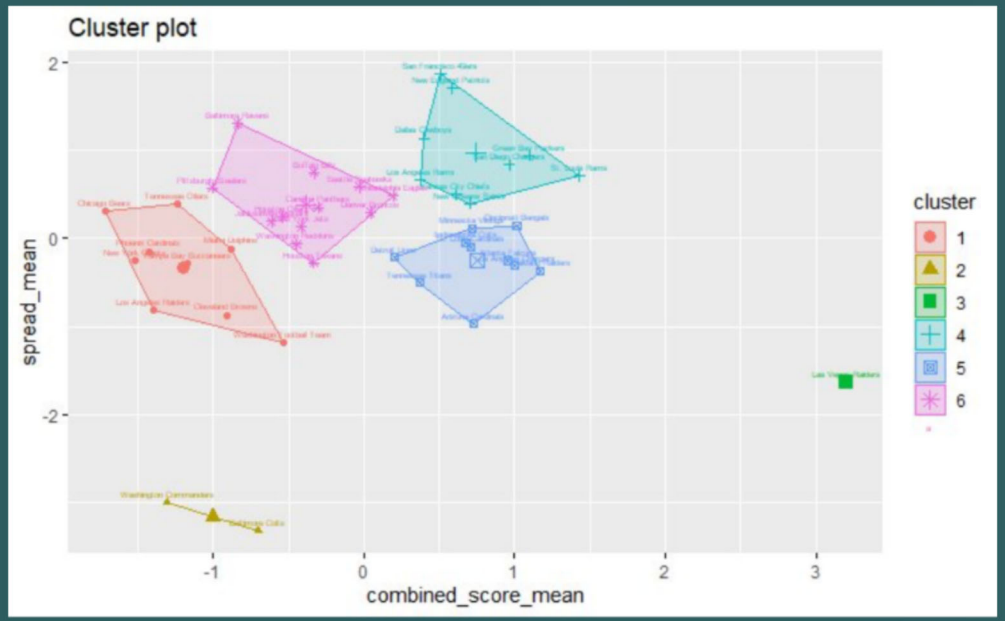
Collected and cleaned multi-decade NFL game, stadium, and weather data

Applied machine learning models (SVM, clustering, random forest)

Visualized patterns in game outcomes and betting behaviors

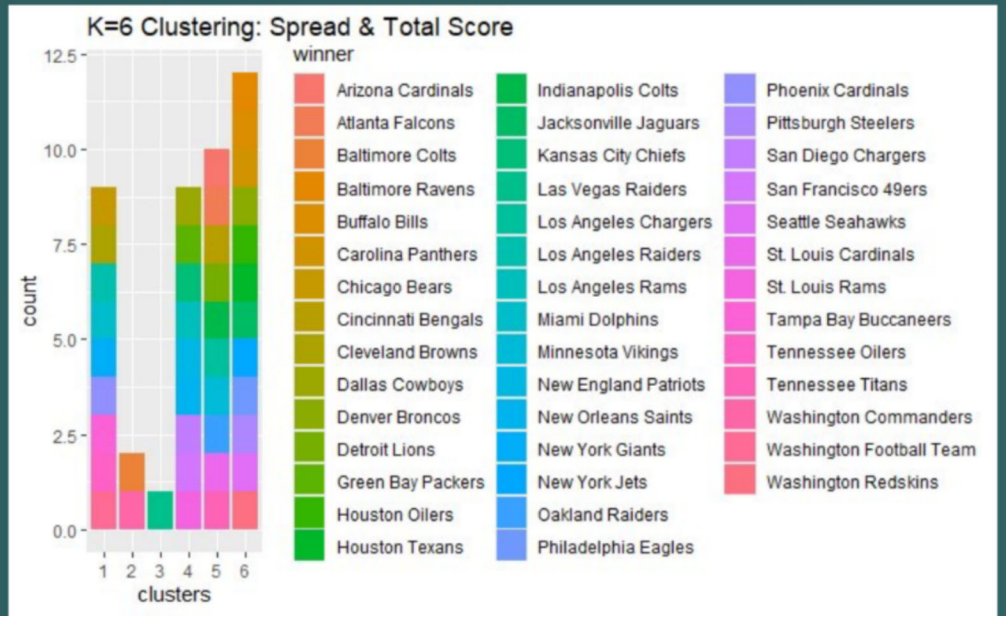
Addressed fairness, prediction limitations, and responsible gambling

Visual Insights



Scores By Team Cluster

Cluster Graph with Total Score



Visual Insights - Cont.

	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

Rules where favorites lost

Limitations & Ethical Risks

Low predictive accuracy (best model ~57%)

Sparse or incomplete data in early decades

Unmodeled variables: injuries, morale, real-time changes

Frequency bias toward heavily covered teams

Reflection & Recommendations

Demonstrated machine learning's limits in dynamic sports contexts

Learned R/Python for EDA, modeling, and visual storytelling

Built models for clustering, association, and outcome prediction

Suggested future enhancements using real-time and player-level data

Bias, Bets, and Bytes

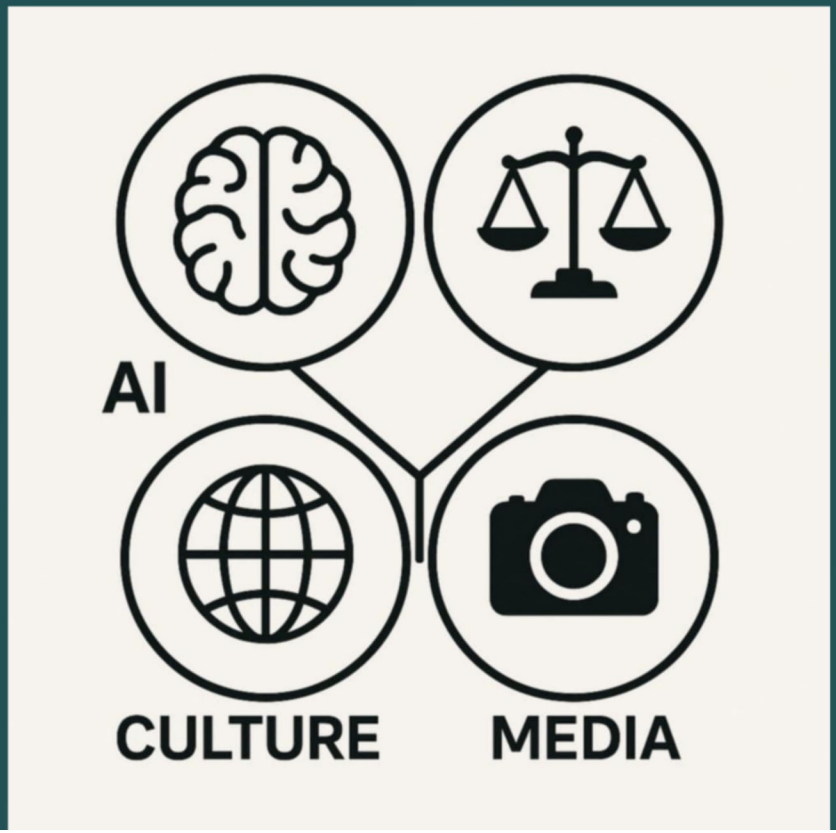
A Reflection on Data, Decisions, and Ethics

IST 692: DALL-E & Imagine

More Info

Examine fairness in AI-generated media by analyzing outputs from OpenAI's DALL·E and Meta's Imagine. The project evaluated how these tools portray race, gender, and culture in response to culturally specific prompts like "A Traditional Indian Woman" or "A Rural African Village." Findings revealed how AI models often reinforce stereotypes, exoticize non-Western subjects, and marginalize modern diversity. This critical project combined ethical analysis, prompt engineering, and image-based comparison to assess representational harms and propose responsible AI improvements.

Program Objectives Met



Used emerging AI tools to generate culturally contextual image datasets

Analyzed AI behavior across sociopolitical and geographic contexts

Created actionable insights for developers, researchers, and ethicists

Centered ethical evaluation around fairness, equity, and representation

Visual Insights

Prompt 1: A modern Indian Family

<- DALL-E

Imagine with Meta AI ->

Visual Insights - Cont.

Prompt 2: A rural African Village

<- DALL-E

Imagine with Meta AI ->

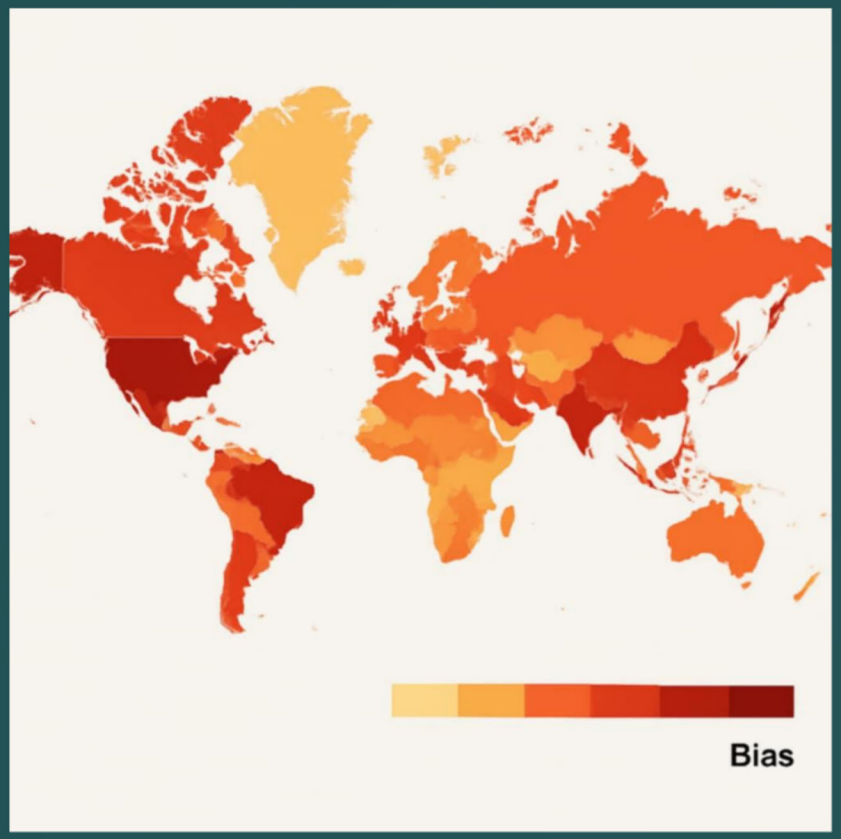
Visual Insights - Part 3

Prompt 3: A Black American CEO

<- DALL-E

Imagine with Meta AI ->

Ethical Risks & Observed Bias



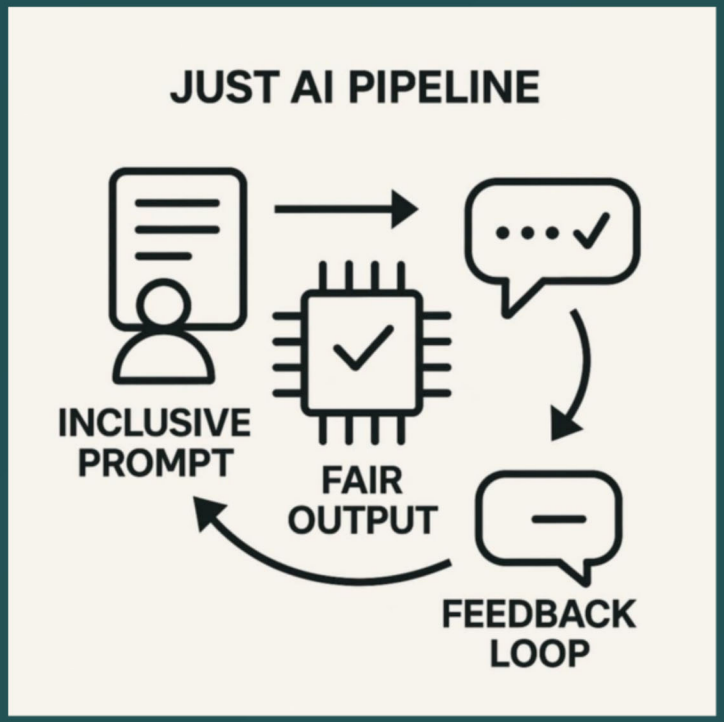
Cultural misrepresentation and stereotyping common across models

Eurocentric defaults in attire, roles, and appearance

Underrepresentation of modern, diverse imagery in training data

Gender roles depicted rigidly and unequally

Reflections & Recommendations



Developed visual literacy and critical analysis of AI-generated media

Applied sociotechnical frameworks to assess fairness in model outputs

Proposed ethical AI improvements: dataset diversity, feedback loops, and identity-agnostic prompts

Gained practical experience evaluating generative tools for representational justice

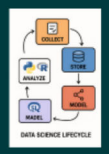
Bias, Bets, and Bytes

A Reflection on Data, Decisions, and Ethics

Conclusion and Reflections

Practice makes Perfect

I gained comprehensive experience designing and executing full data workflows—from data collection and storage to model building and insight communication. Each project sharpened my ability to use tools like Python, R, SQL, and machine learning libraries to develop actionable insights across business, societal, and technical domains. I learned to translate raw data into impactful solutions that support real-world decision-making.



Ethics in Data Usage

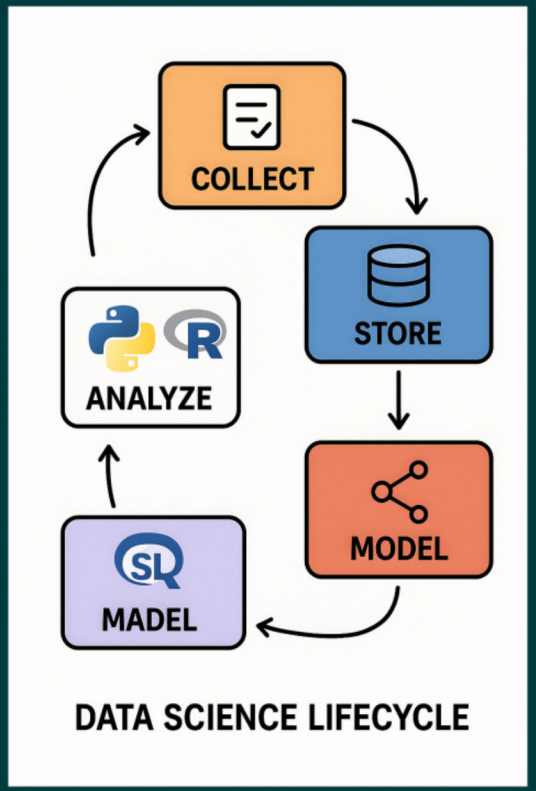
Beyond technical execution, these projects taught me to consider the broader implications of data use. From detecting bias in AI-generated images to addressing fairness in pricing tools, I learned to critically assess the societal and ethical impacts of data systems. Responsible data science requires transparency, fairness, and inclusion—especially when decisions affect diverse populations.

Future Directions in Data-Driven Decisions

As data continues to shape industries and influence public policy, the ability to combine technical skills with ethical reasoning becomes critical. I am prepared to contribute as a data professional who not only builds intelligent systems but also asks critical questions about their impact. My future work will focus on designing data tools that inform decision-making while prioritizing user trust, equity, and transparency.

Practice makes Perfect

I gained comprehensive experience designing and executing full data workflows—from data collection and storage to model building and insight communication. Each project sharpened my ability to use tools like Python, R, SQL, and machine learning libraries to develop actionable insights across business, societal, and technical domains. I learned to translate raw data into impactful solutions that support real-world decision-making.

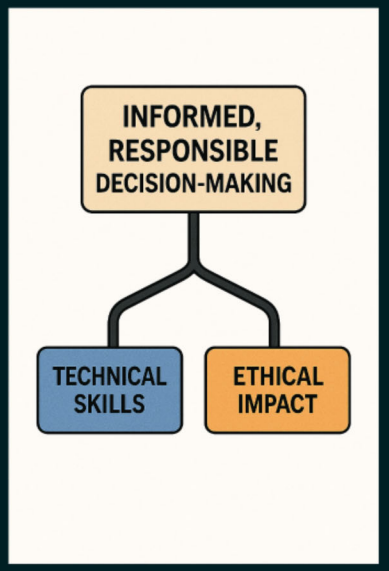


Ethics in Data Usage

More than just technical execution, these projects taught me to consider the broader implications of data use. From detecting bias in AI-generated images to addressing fairness in pricing tools, I learned to critically assess the societal and ethical impacts of data systems. Responsible data science requires transparency, fairness, and inclusion—especially when decisions affect diverse populations.

Future Directions in Data-Driven Decisions

As data continues to shape industries and influence public policy, the ability to combine technical skills with ethical reasoning becomes critical. I am prepared to contribute as a data professional who not only builds intelligent systems but also asks critical questions about their impact. My future work will focus on designing data tools that inform decision-making while prioritizing user trust, equity, and transparency.



Bias, Bets, and Bytes

A Reflection on Data, Decisions, and Ethics

A Journey through the Mind of a Data Scientist

Thank You