Data Analytics Portfolio

Fariya Asghar



"I believe the best insights happen where business context meets technical skill."

From Banking to Big Data

Hi, I'm Fariya Asghar. For 4 years, I was on the front lines of corporate finance, specializing in credit and risk. My role was to look at a company's financial health and tell a story about its future. I realized the part I loved most was finding that story hidden in the data.

That realization led me here. I'm now building on my practical banking experience by mastering the tools of data analytics. My goal is simple: to bridge the gap between raw data and real-world business strategy.

Business Expertise

- Corporate Credit Analysis
- Financial Risk Assessment
- Stakeholder Communication
- Business Performance Metrics (KPIs)

Technical Skills

- **SQL** (Data Querying)
- **Tableau** (Visualization)
- **Excel** (Advanced Modeling)
- **Python** (Pandas)

Analytical Thinking

- Quantitative Reasoning
- Problem Decomposition
- Data-Driven Storytelling
- Root Cause Analysis

Portfolio Overview

PROJECTS	DESCRIPTION	TOOLS
GameCo. Marketing Strategy	Analysis of Regional Sales Trends & Strategic Recommendations	
Public Health Analysis Influenza Trends	Preparing for Influenza Season	
Rockbuster Stealth LLC	Engineering Rockbuster's Digital Comeback	
Instacart Customer Behavior Analysis	Marketing strategy for an online grocery store	
Pig E. Bank	Predicting Customer Churn	
Airbnb Amsterdam	Unlocking Profitability in the Amsterdam Airbnb Market	
20th Century Geopolitics	Provide actionable recommendations for fleet management and service expansion.	
NYC Citi Bike	Analyze unstructured text and build a network visualization of historical country interrelations.	

GameCo. Marketing Strategy

Analysis of Regional Sales Trends & Strategic Recommendations

Data Source

Public Video Game Sales Dataset (from Kaggle), contextualized for a business case for the client, GameCo.

Methodology

- Grouping Data and Summarizing Data
- Descriptive Analysis
- Visualizing Results
- Presenting Results

Tools

MS Excel (Data Cleaning, Pivoting, Analysis)

MS PowerPoint (Reporting & Visualization)

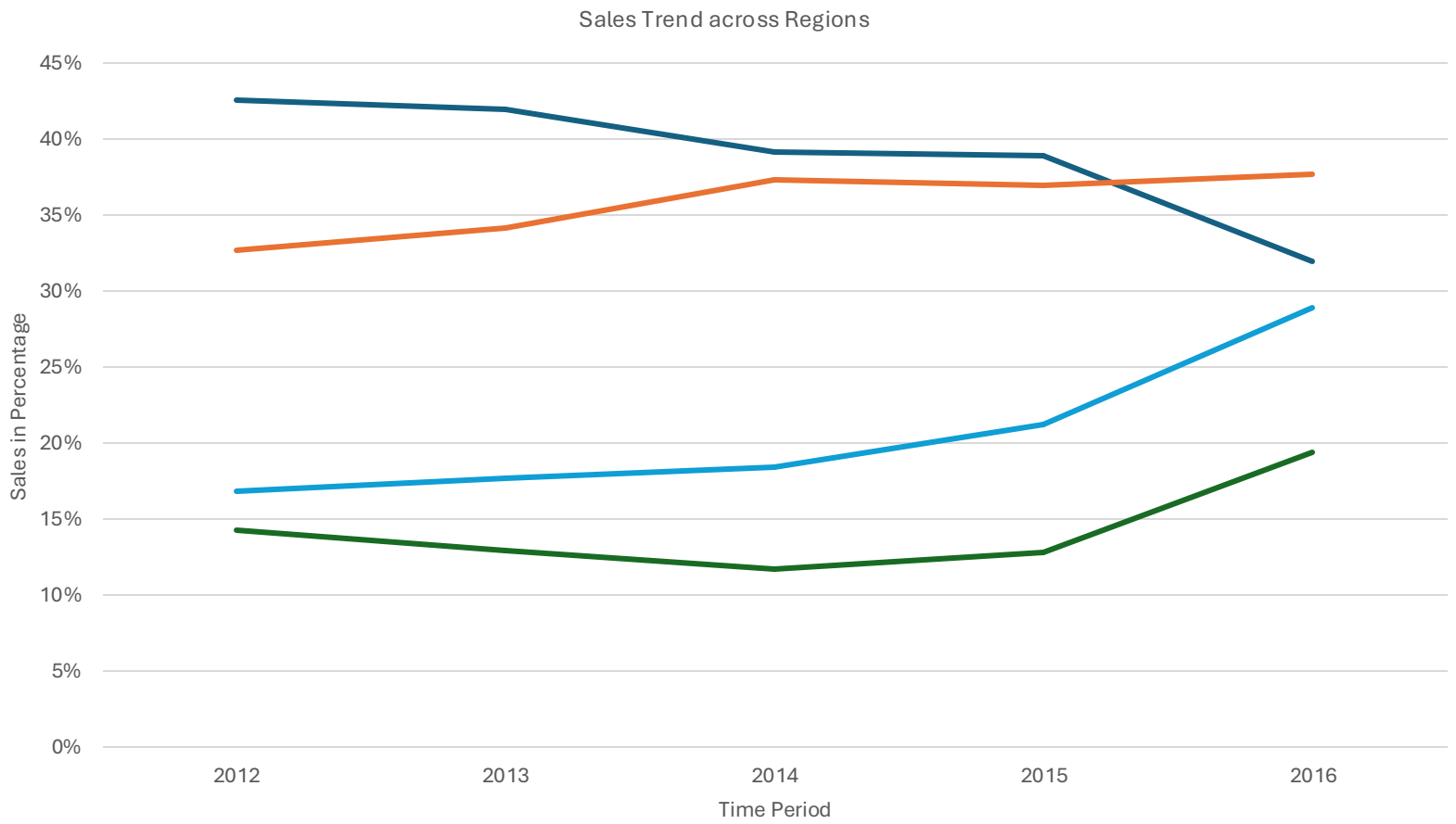
The Business Challenge:

GameCo. executives were planning their 2017 marketing budget, operating under a critical assumption: that their global sales trends were stable, with North America as the unshakeable primary market.

My Role:

To analyze historical sales data to either validate this assumption or provide evidence for a new, data-driven strategy.

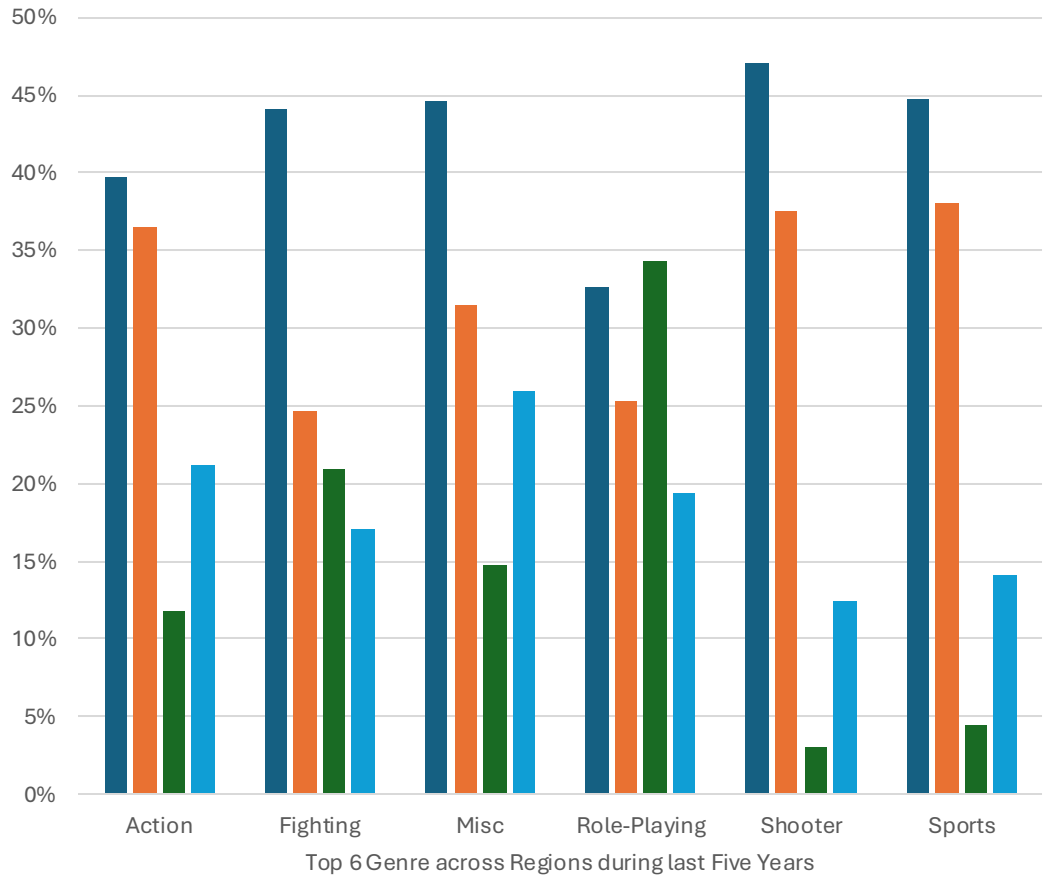
Uncovering a Dramatic Shift in the Market



Analysis revealed that Europe's market share had surpassed North America's, fundamentally challenging the company's core strategy

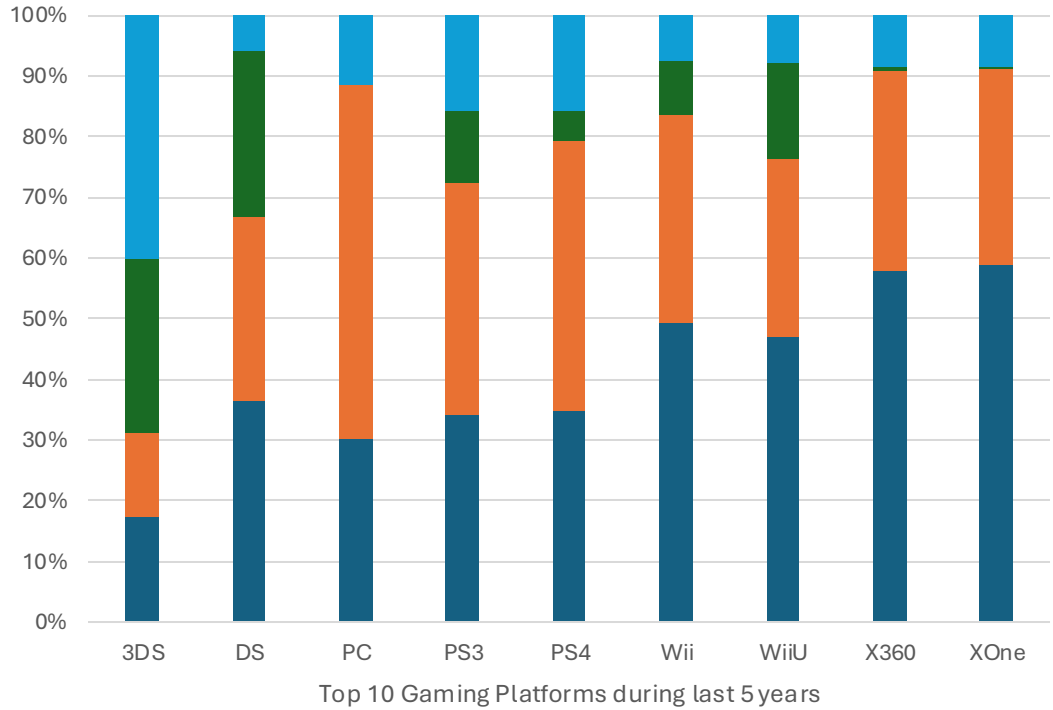
Drilling Down: Genre and Platform Preferences

Genre: NA & EU prefer Shooters; Japan prefers RPGs

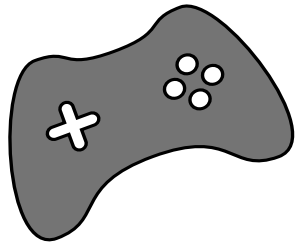


Top 6 Genre across Regions during last Five Years

Platform: Xbox dominates NA; PC is strong in Europe



Top 10 Gaming Platforms during last 5 years



A New, Data-Driven Strategy

Reallocate Budget: Shift focus from the declining NA market to the growing EU and emerging markets.

Localize Content: Align game genres with specific regional platform and taste preferences

Strengthen PC Presence in Europe.

Project Deliverables:
Report
Presentation

Preparing for Influenza Season

Understanding Influenza Burden

Data Source

- CDC Vital Statistics: Influenza deaths by geography
- US Census Bureau: Population data by geography, time, age, and gender
- CDC ILINet: National influenza-like illness visit reports.

Methodology

- Data Integration & Cleaning
- Statistical Analysis (Hypothesis Testing)
- Trend & Seasonal Analysis
- Data Visualization & Storytelling

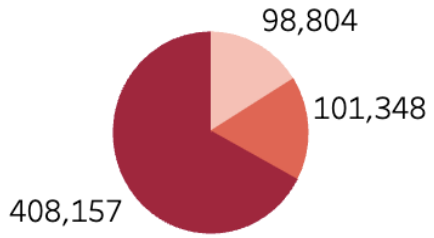
Tools

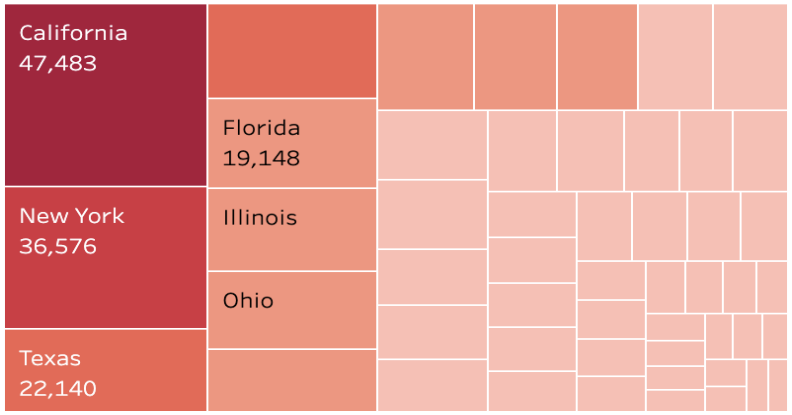
Objective: To identify trends in flu mortality and illness to inform resource allocation and reduce avoidable deaths.

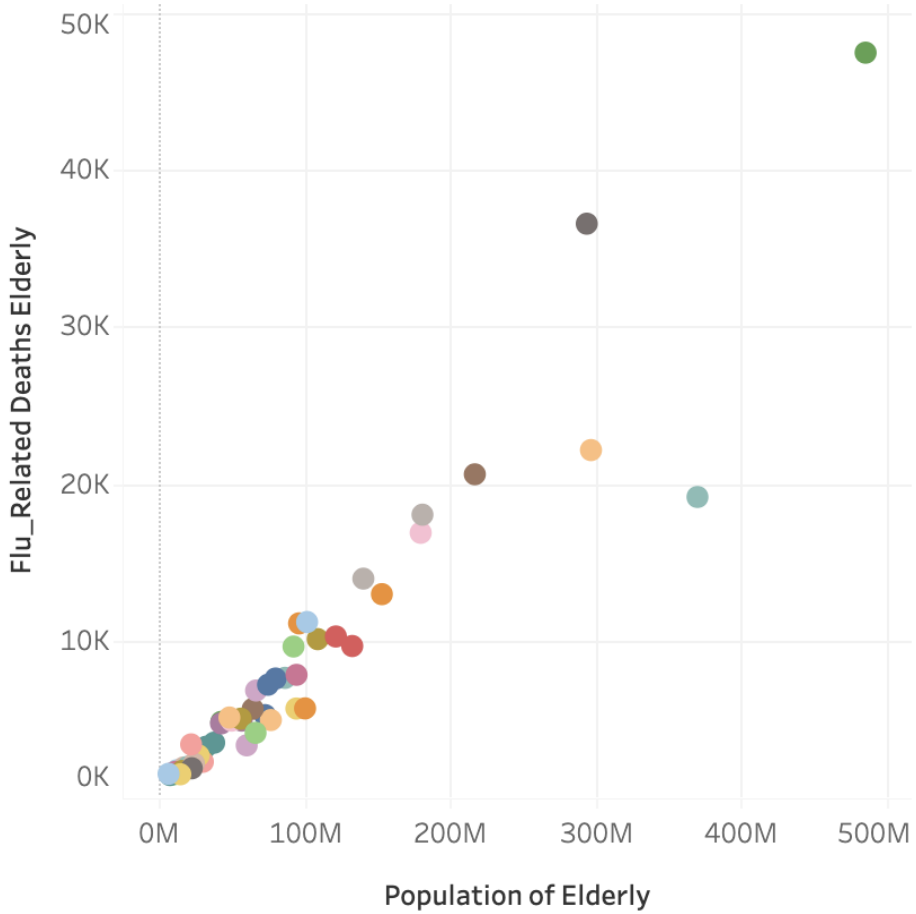
Hypothesis: *If a person's age increases, then their likelihood of mortality from influenza will also increase.*

My Role:
To integrate three disparate public health datasets, identify key trends in mortality and illness, and provide a blueprint for action.

The Data Confirms: Age is the Greatest Risk Factor

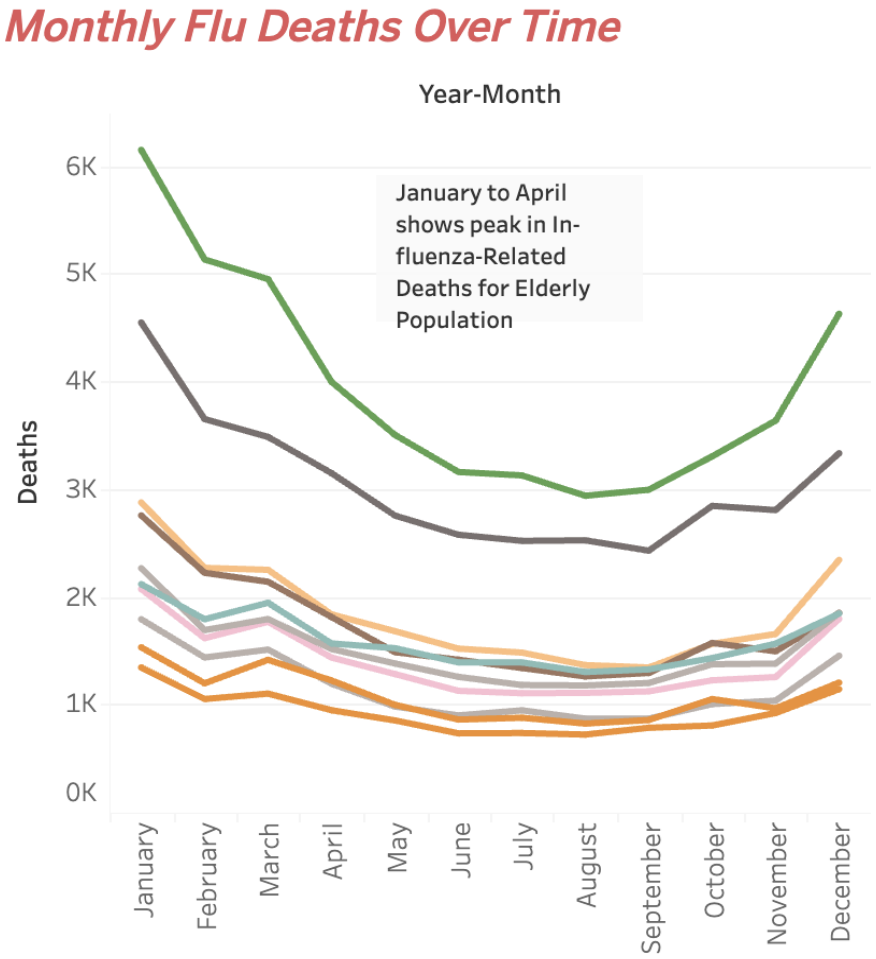


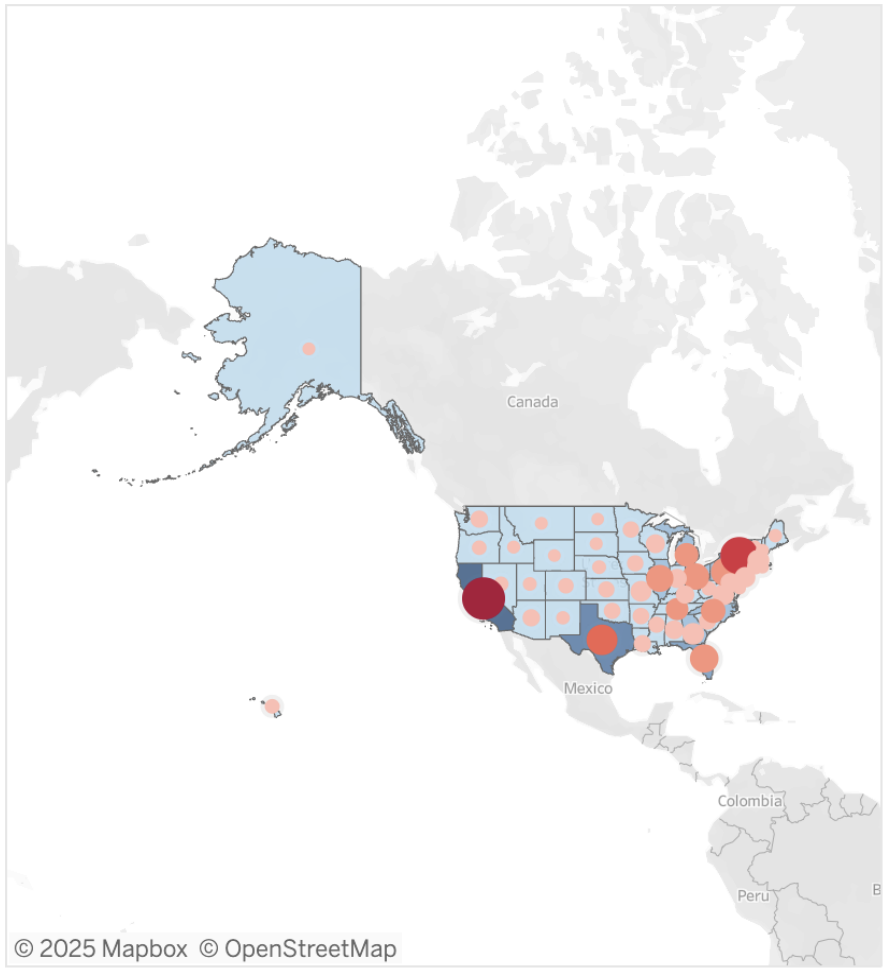




An independent t-test comparing middle-aged vs. elderly mortality rates yielded a p-value of < 0.001. This confirms that the significantly higher mortality rate among the elderly is not due to random chance.

Mapping the Hotspots: Identifying Seasonal and Geographic Peaks





- **Seasonal Peaks:** "The data shows a clear and predictable surge in mortality every year between January and April."
- **Geographic Disparities:** "Resou rces can be targeted to high-mortality states like California, Florida, and Texas, which show consistently higher totals."

A Data-Driven Blueprint for Action

Actionable Recommendations

Prioritize Elderly: Focus vaccination and outreach on populations aged 65+.

Deploy Seasonally: Increase medical staff during peak winter months.

Target Geographically: Allocate resources to regions with historically high mortality.

Project Deliverable:

Rockbuster's Stealth Digital Transformation

Engineering Rockbuster's Digital Comeback

Data Source

Rockbuster' data set including film inventory, customers and payments.

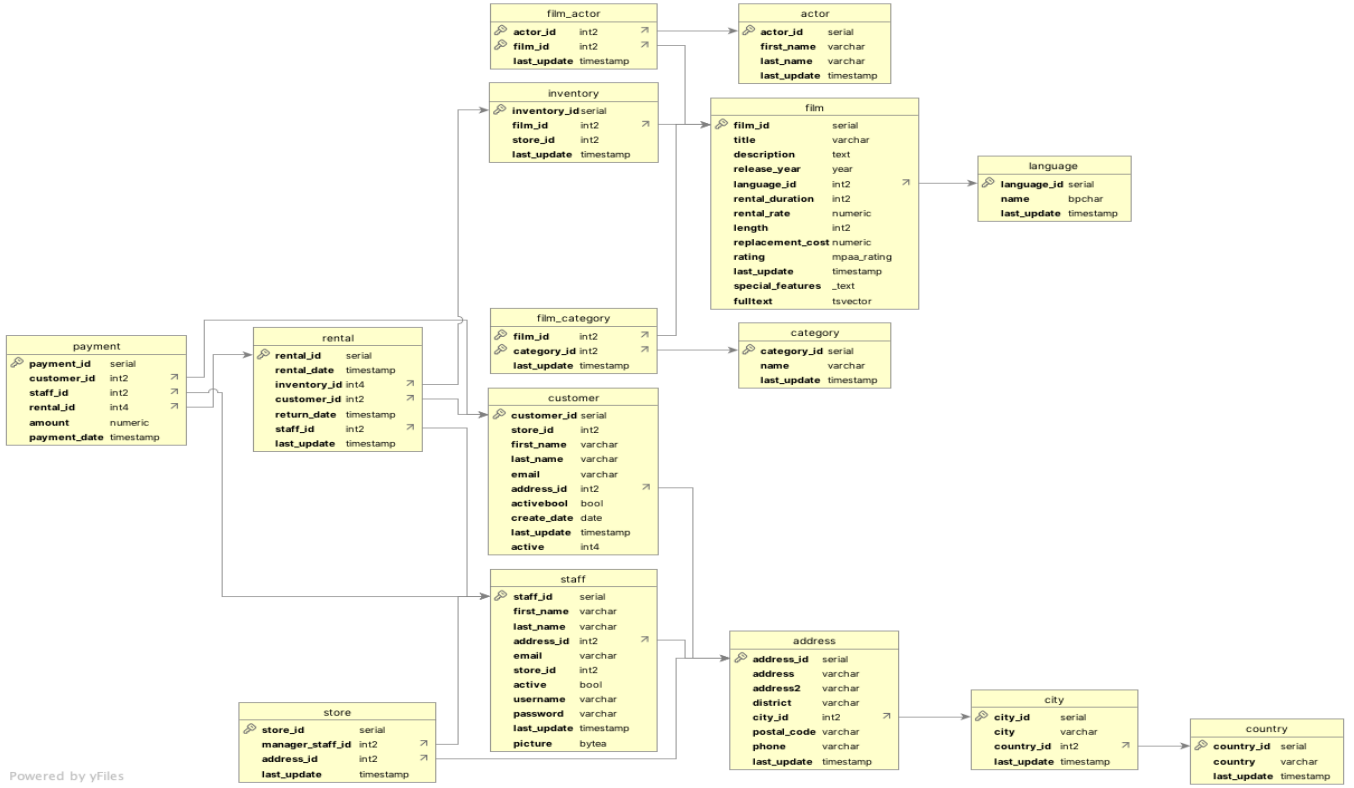
Methodology

1. SQL Database Querying & Integration
2. Customer & Geographic Segmentation
3. Revenue & Content Performance Analysis.

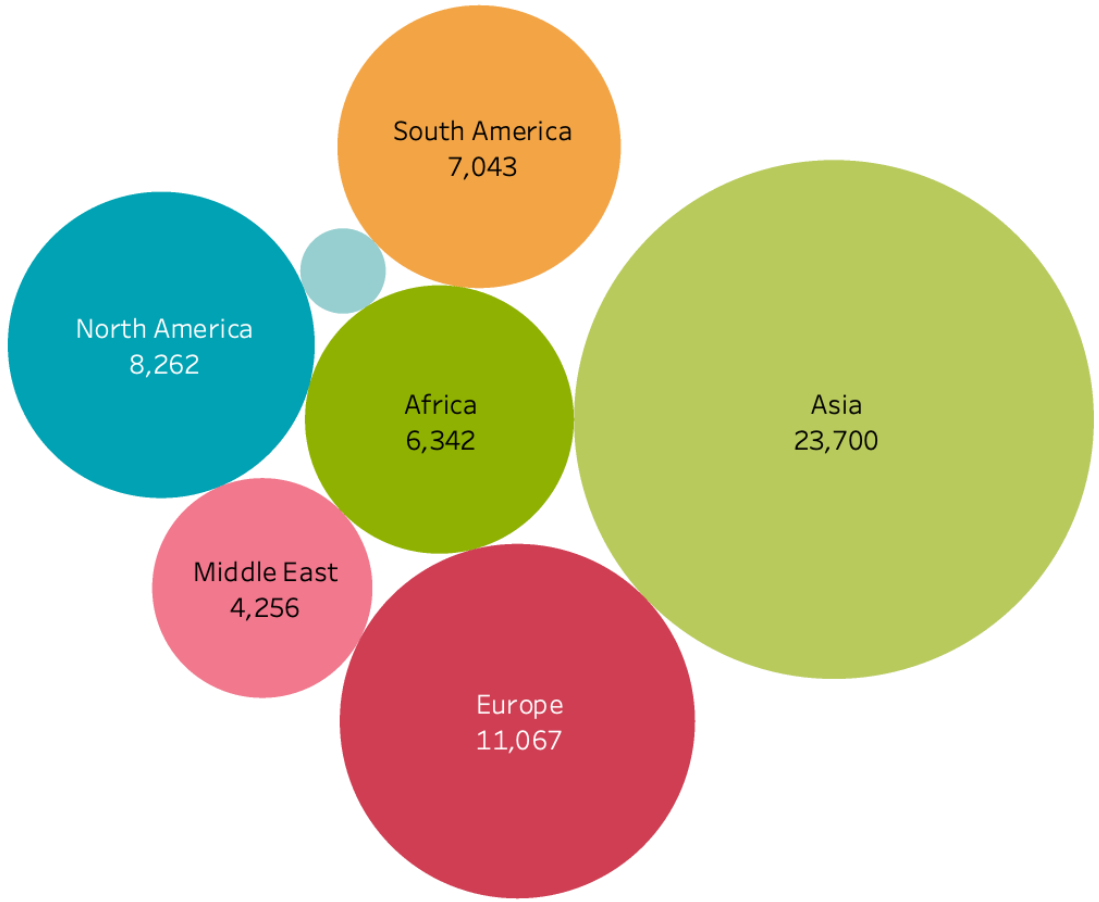
Tools Used

The Challenge: Rockbuster, a global movie rental giant, faced extinction from streaming competitors.

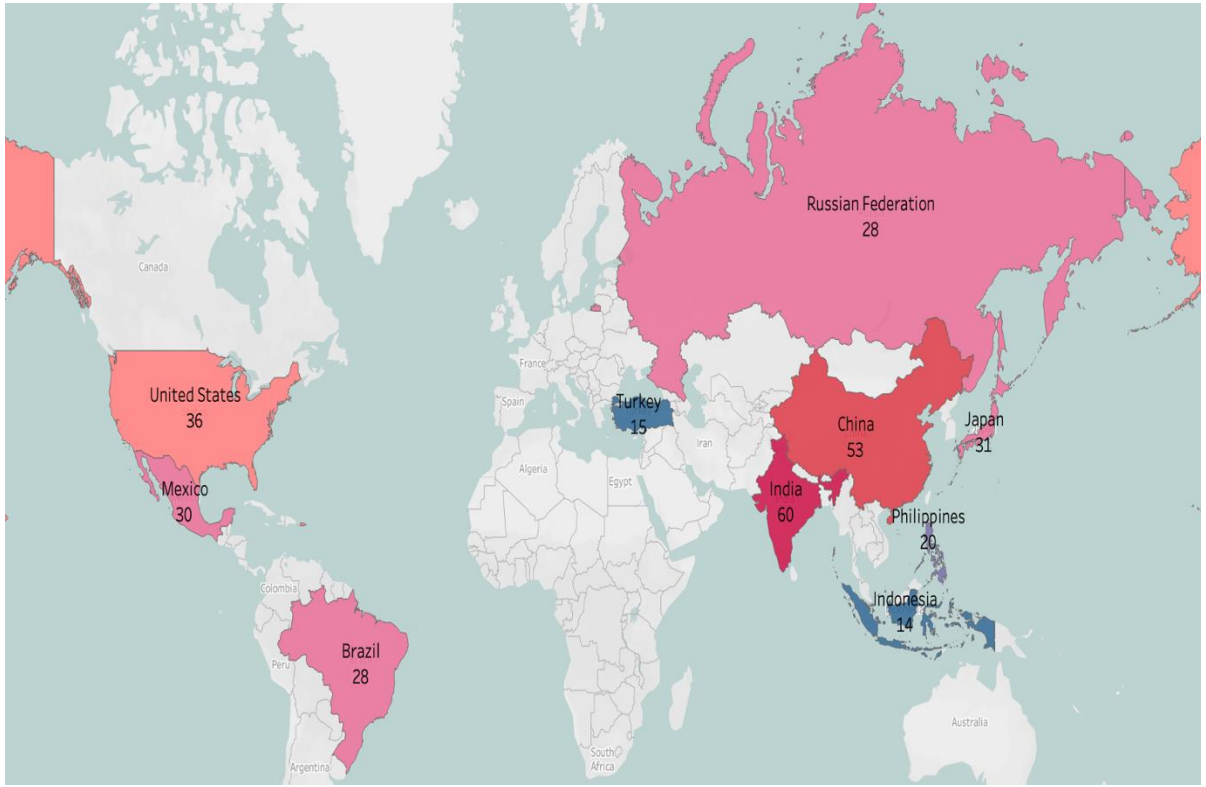
My Role: Analyze their historical customer data to engineer a successful online launch strategy.



The Data Revealed a Hidden Goldmine

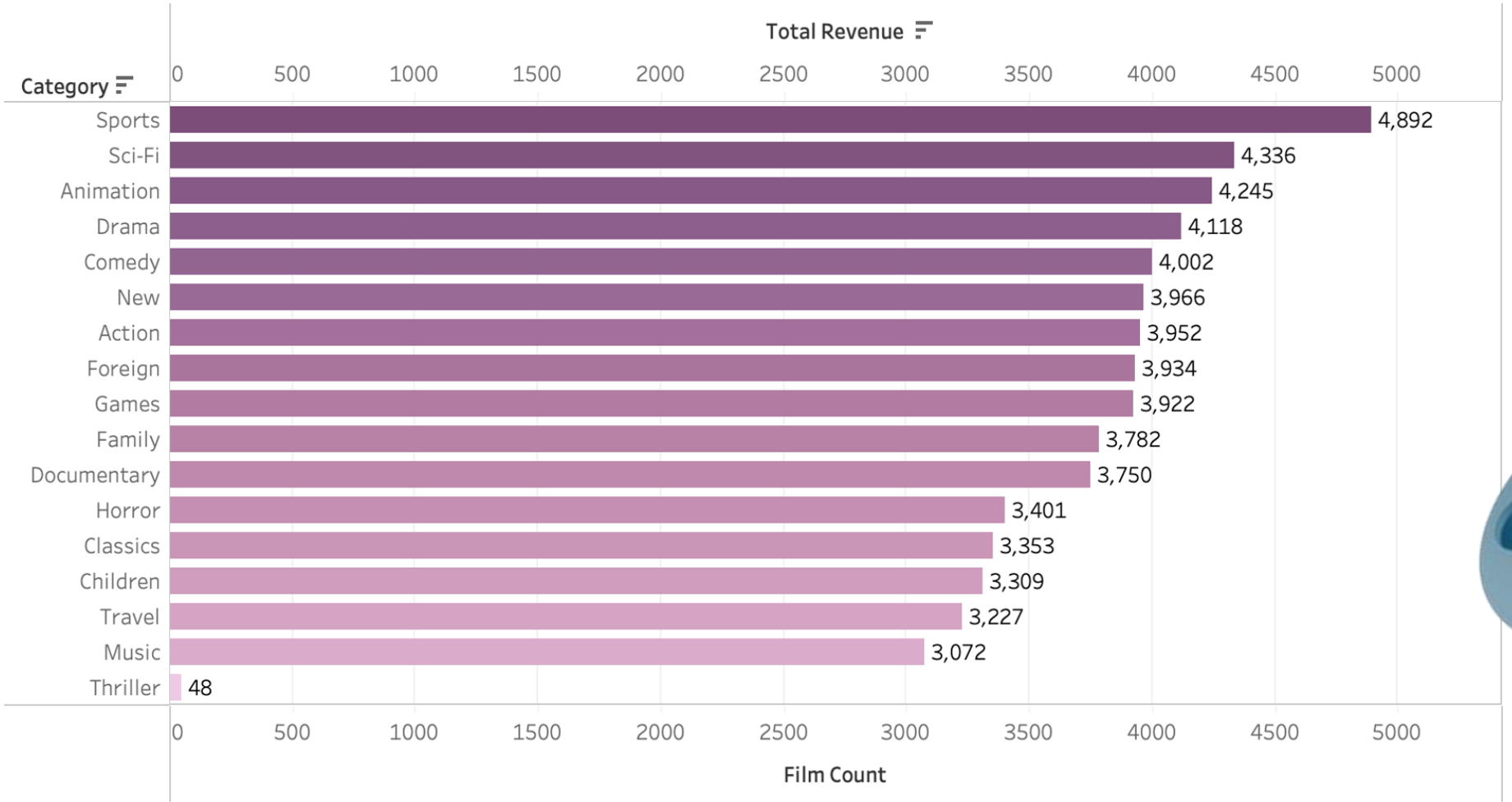


Analysis proved Asia was the dominant market, contributing nearly 40% of all global revenue and revealing the top priority for the online launch.



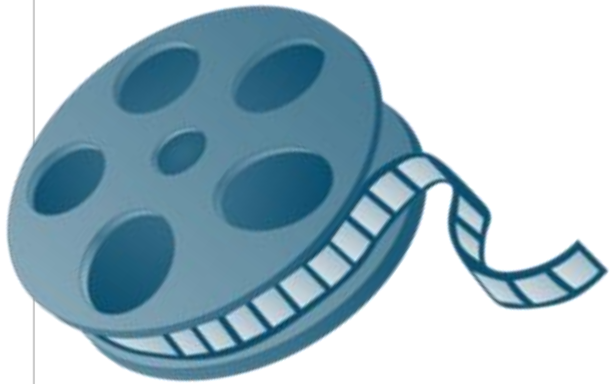
Informing the Strategy: Content & Pricing Insights

Top Content: Sports & Sci-Fi Lead



Pricing Opportunity: The €0.99 Tier

Data showed that while many low-revenue titles were €0.99, other films at the same price point generated 5x more rentals, indicating a clear opportunity for dynamic pricing



A 3-Point Strategy for a Successful Launch

Asia-Focused Growth

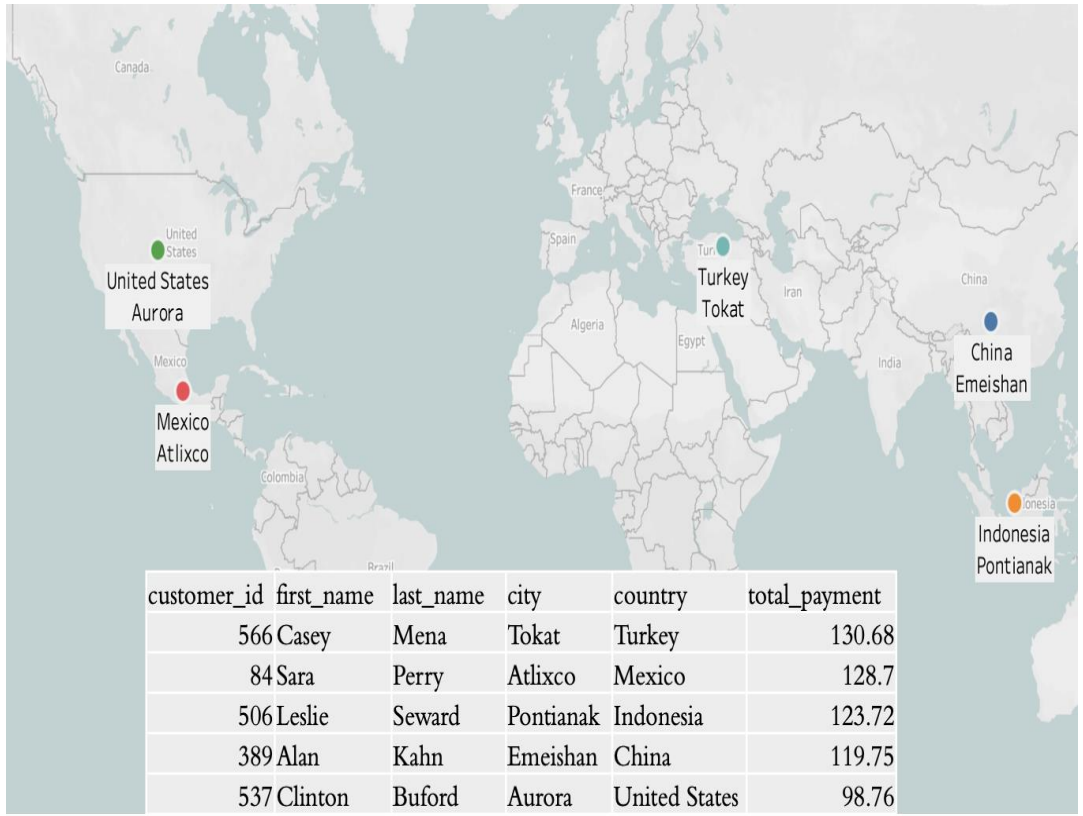
- Localize payments (UPI/Alipay)
- Launch loyalty rewards (free rental after 5 transactions)

Pricing Optimization

- Test 20% price increase on high-demand €0.99 titles
- Introduce "Binge Pass" (3 films/7 days for €6.99)

Content Expansion

- Boost top genres (Sports/Sci-Fi/Animation)
- Add Asian-localized films (Bollywood/Wuxia)



customer_id	first_name	last_name	city	country	total_payment
566	Casey	Mena	Tokat	Turkey	130.68
84	Sara	Perry	Atlixco	Mexico	128.7
506	Leslie	Seward	Pontianak	Indonesia	123.72
389	Alan	Kahn	Emeishan	China	119.75
537	Clinton	Buford	Aurora	United States	98.76

Project Deliverable:

Instacart Customer Behavior Analysis

Instacart Grocery Basket Analysis

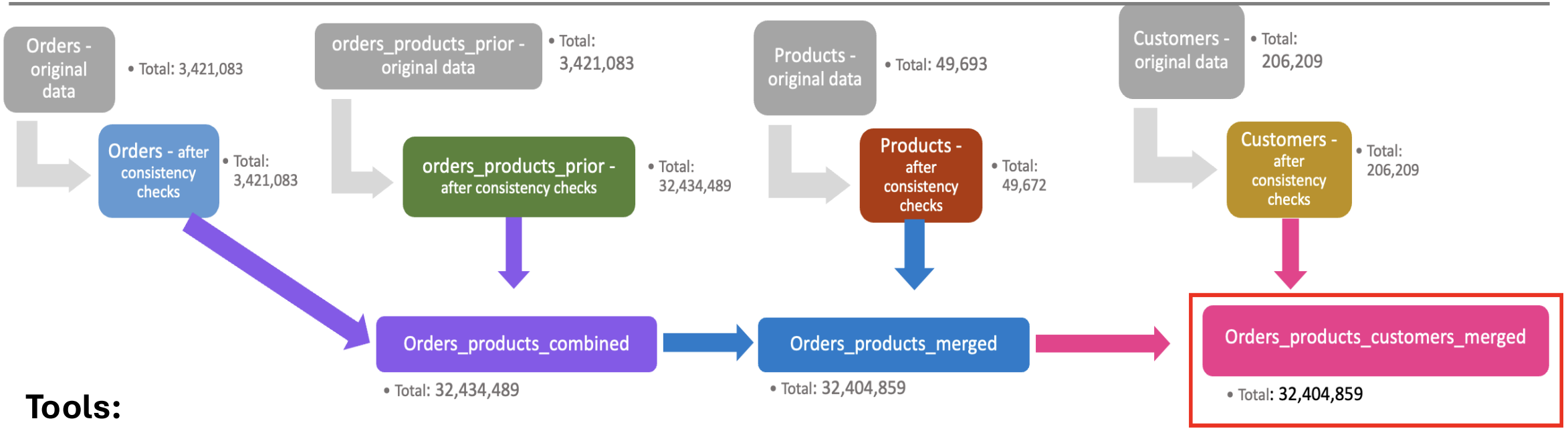
Data Source

- Customer Dataset (CareerFoundry)
- “The Instacart Online Grocery Shopping Dataset 2017” (Kaggle)

Methodology

- Exploratory Analysis
- DanaWrangling and Merging
- Deriving Variables
- Aggregating Data
- Visualization

The Challenge: To analyze over 32 million order lines to uncover actionable customer purchasing patterns for Instacart's marketing and sales teams.

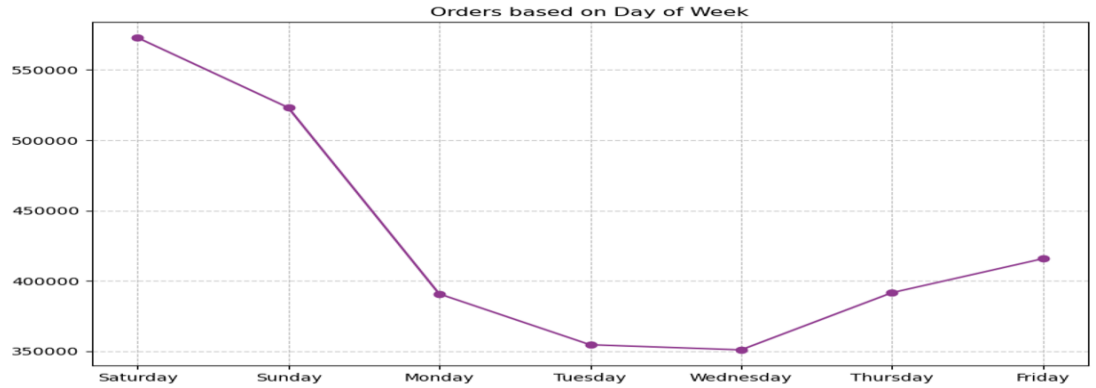


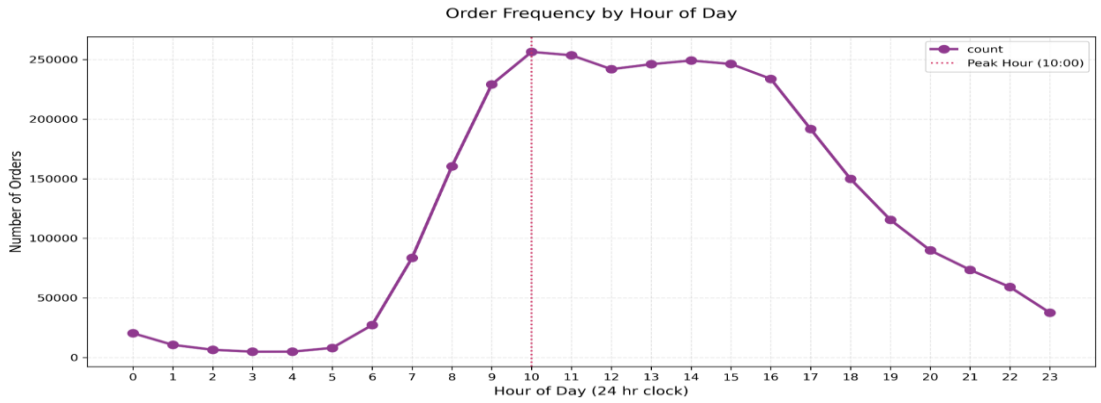
Tools:

My **role** involved rigorous wrangling of multiple datasets, including consistency checks, handling missing values, and deriving new variables for analysis.

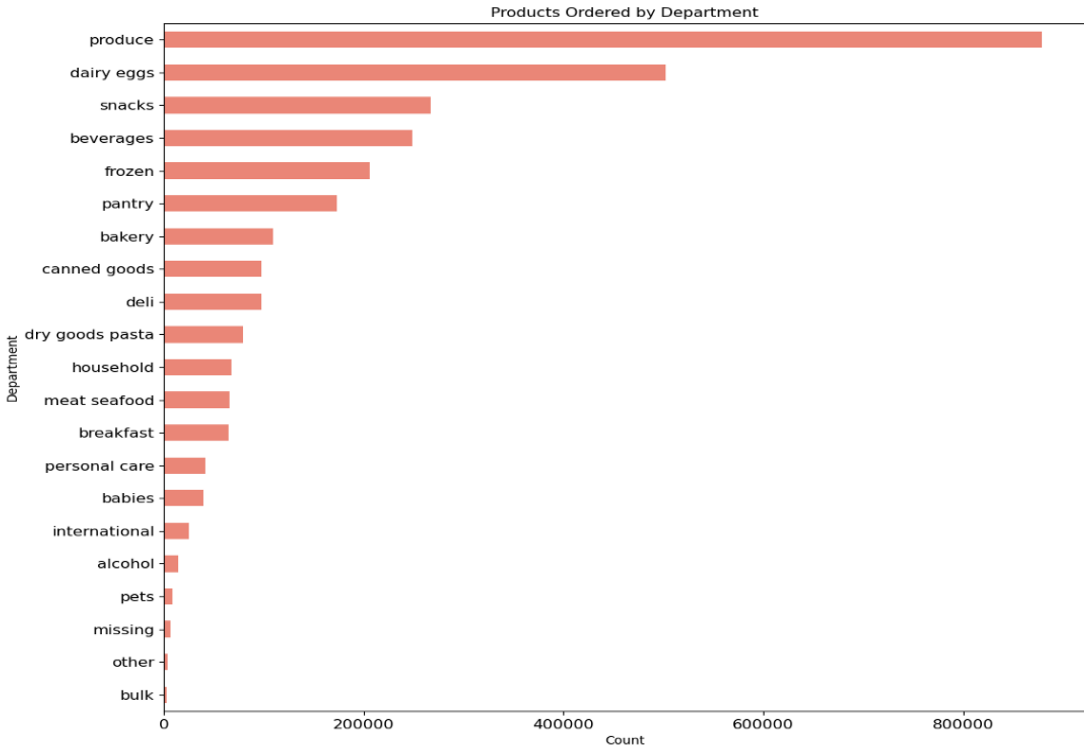
Finding the Pulse: Peak Shopping Times & Top Categories

When They Shop: Weekends and midday (9 AM-4 PM) are the undisputed peak times.



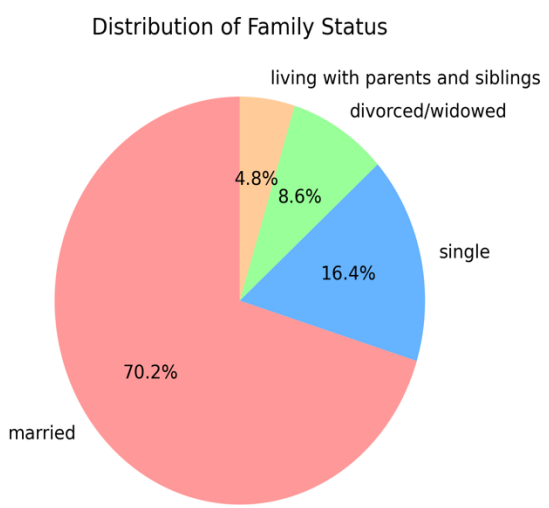


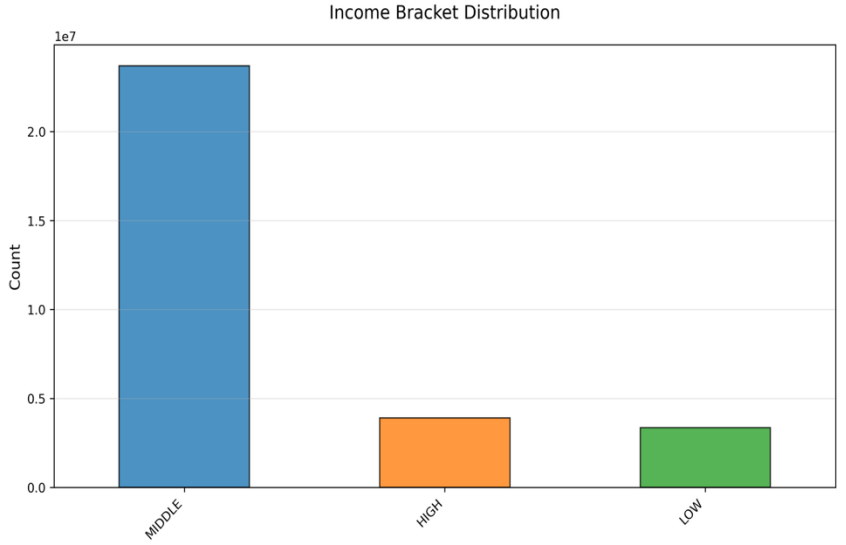
What They Buy: Everyday essentials dominate, with Produce, Dairy/Eggs, and Snacks being the top-selling departments.

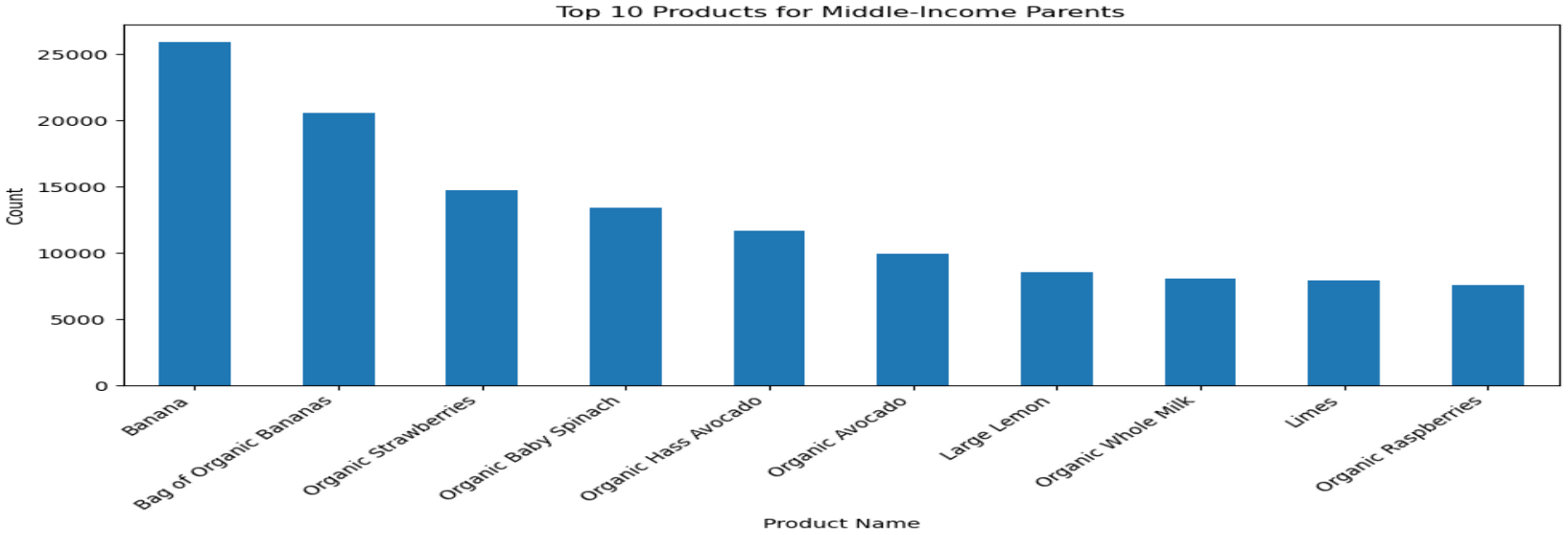


This data provides a clear roadmap for ad scheduling, shopper staffing, and inventory management.

Beyond the Numbers: Building the Core Customer Persona







Persona: The "Savvy Parent"

Demographics: Married, middle-income, with dependents (forms the largest and most loyal customer segment).

Shopping Behavior: Prioritizes fresh, healthy, and organic options for their families, with items like bananas, spinach, and strawberries being top purchases.

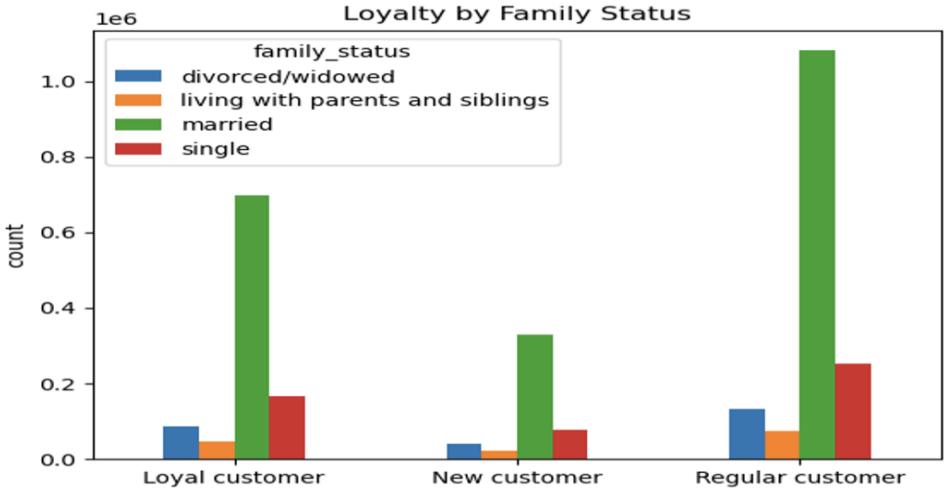
The Strategic Value: This persona is Instacart's core engine. All marketing, promotions, and product recommendations should be tailored to their needs.

A Data-Driven Blueprint for Marketing & Sales

Target the "Savvy Parent" Persona: "Launch 'Weekly Family Meal' bundles and 'Healthy Kids Lunchbox' campaigns featuring top organic products to resonate with the core customer."

Optimize Ad Spend & Promotions: "Focus ad budget on weekends (9 AM-4 PM). Launch targeted 'Late-Night Convenience' promotions to capture high-spending, low-volume nighttime shoppers."

Increase Average Order Value: "Implement a 'Goes well with' recommendation engine to cross-sell items (e.g., suggest 'Organic Berries' when a user adds 'Yogurt')."

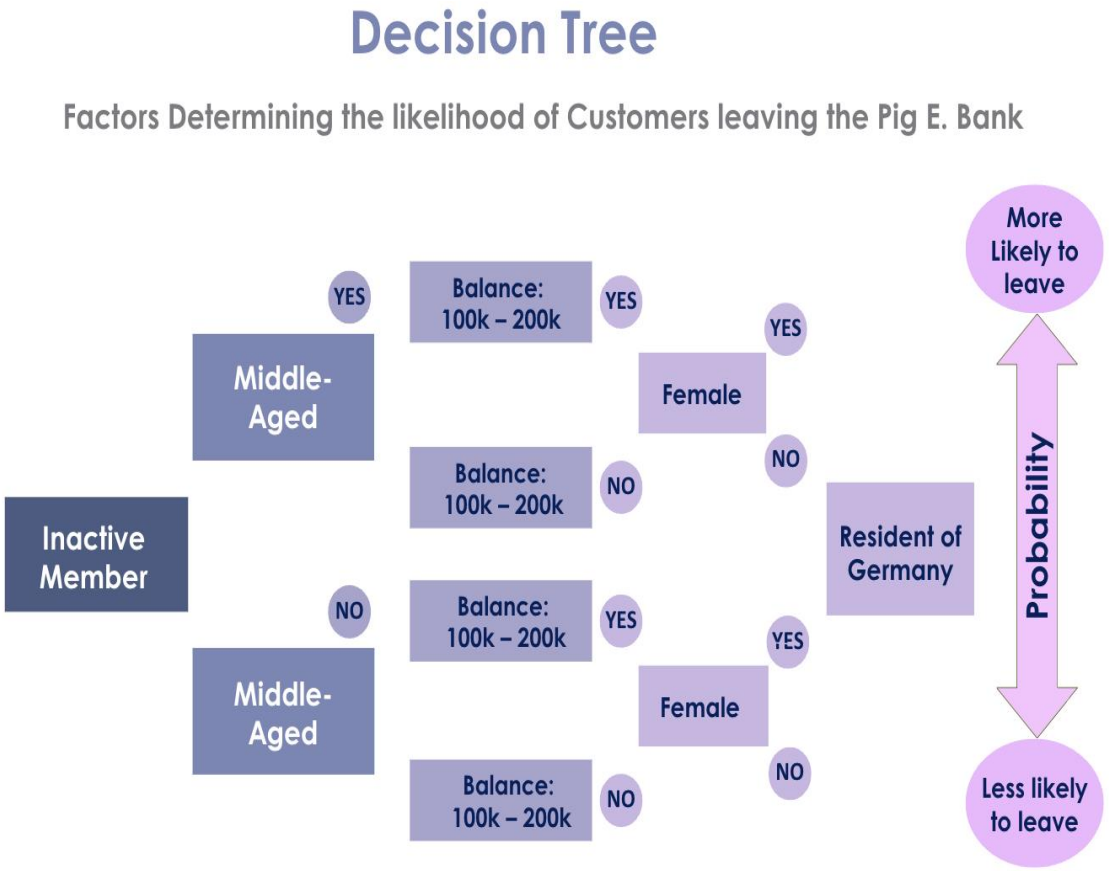


Project Deliverable:
Full Project Report, SQL Queries & Visualizations

Pig E. Bank


banking
banks
financial
capital
reserve
customers
deposits
business
bank
lending
credit
loans
system
intermediary
fractional
goldsmiths
known
many
centuries
receipts

Predicting Bank Customer Churn



The Challenge:

Pig E. Bank was reactively losing customers without understanding the root causes. They needed a data-driven way to identify high-risk customers *before* they leave.

Tools

My **goal** was to move beyond simple analytics to build a predictive model that uncovers the specific factors driving churn.

Methodology

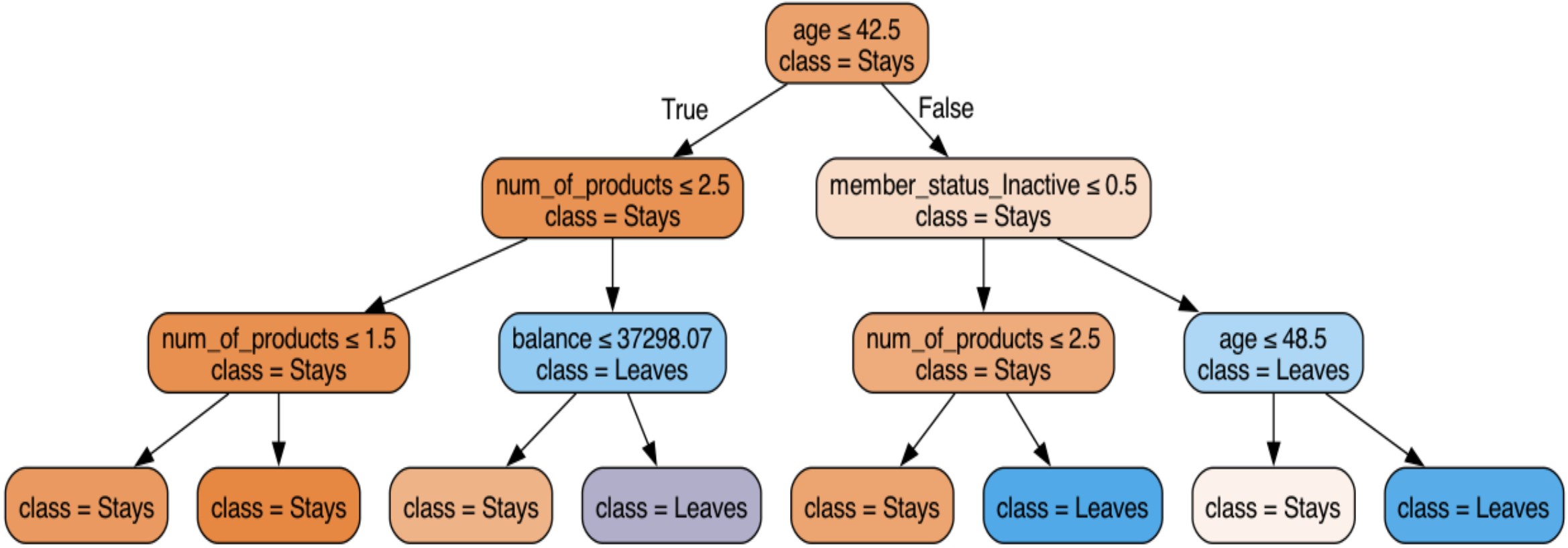
- Data Cleaning
- Descriptive Statistics
- Decision Tree

From Data to Prediction: The Churn Decision Tree

I developed a Decision Tree classification model using scikit-learn to map the customer journey towards churn.

The model immediately identified the single most important predictor: **being an 'Inactive Member' dramatically increases the likelihood of leaving.**

Further splits reveal that factors like **balance**, **age**, and **number of products** are also significant predictors.



Building the 'At-Risk' Customer Persona

Status: Inactive Member. (70% of churned customers are inactive)

Demographics: Middle-Aged (40-60), often **Female**, and a resident of **Germany**.

Financials: Holds a **High Balance** (typically $100k - $200k).

Key Finding: Churn is not driven by new or low-value customers. The greatest risk comes from established, high-balance members who have become disengaged with the bank's services.

Row Labels	Count of member_status
Active	61
Inactive	143
Grand Total	**204**

Row Labels	Count of country
France	77
Germany	75
Spain	52
Grand Total	**204**

Row Labels	Count of age
Adults	**60**
Middle- Aged	**123**
Elderly	**21**
Grand Total	**204**

Row Labels	Count of gender
Female	**121**
France	51
Germany	39
Spain	31
Male	**83**
France	26
Germany	36
Spain	21
Grand Total	**204**

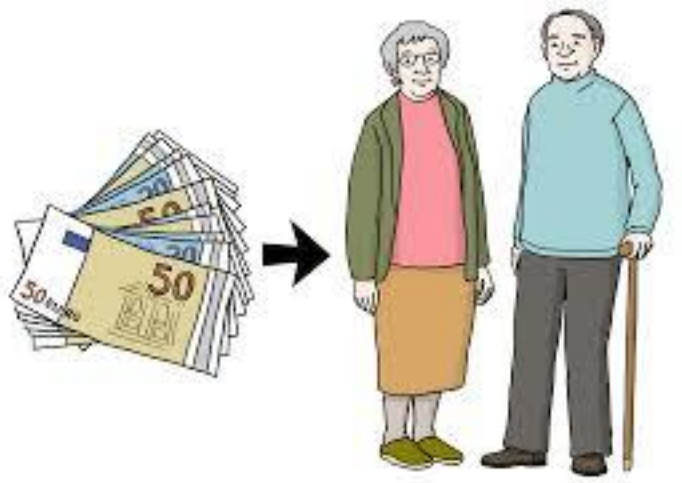
Row Labels	Count of balance
0	**56**
Low Balance	**3**
Medium Balance	**25**
High Balance	**117**
Above 200,000	**3**
Grand Total	**204**

A Proactive Strategy to Reduce Churn

Launch Re-Engagement Campaigns: Proactively target **inactive, high-balance members** with personalized offers, new feature tutorials, or consultations with financial advisors to demonstrate value.

Develop a Targeted Regional Strategy: Since German customers have a higher churn rate, create marketing campaigns and product bundles specifically tailored to the German market.

Innovate for the Core Persona: Design new products and services that cater to the financial goals of middle-aged customers (e.g., advanced investment tools, retirement planning services).



Project Deliverables:
Jupyter Notebook with scikit-learn model: [Available Upon Request]
Full Analysis in Excel.

Airbnb Amsterdam

Strategic Market Analysis of Amsterdam Airbnb

Objective: To move beyond simple averages and identify data-driven, actionable strategies for new investors in the complex and competitive Amsterdam short-term rental market.

Key Questions:

1. What are the true Drivers of Price?

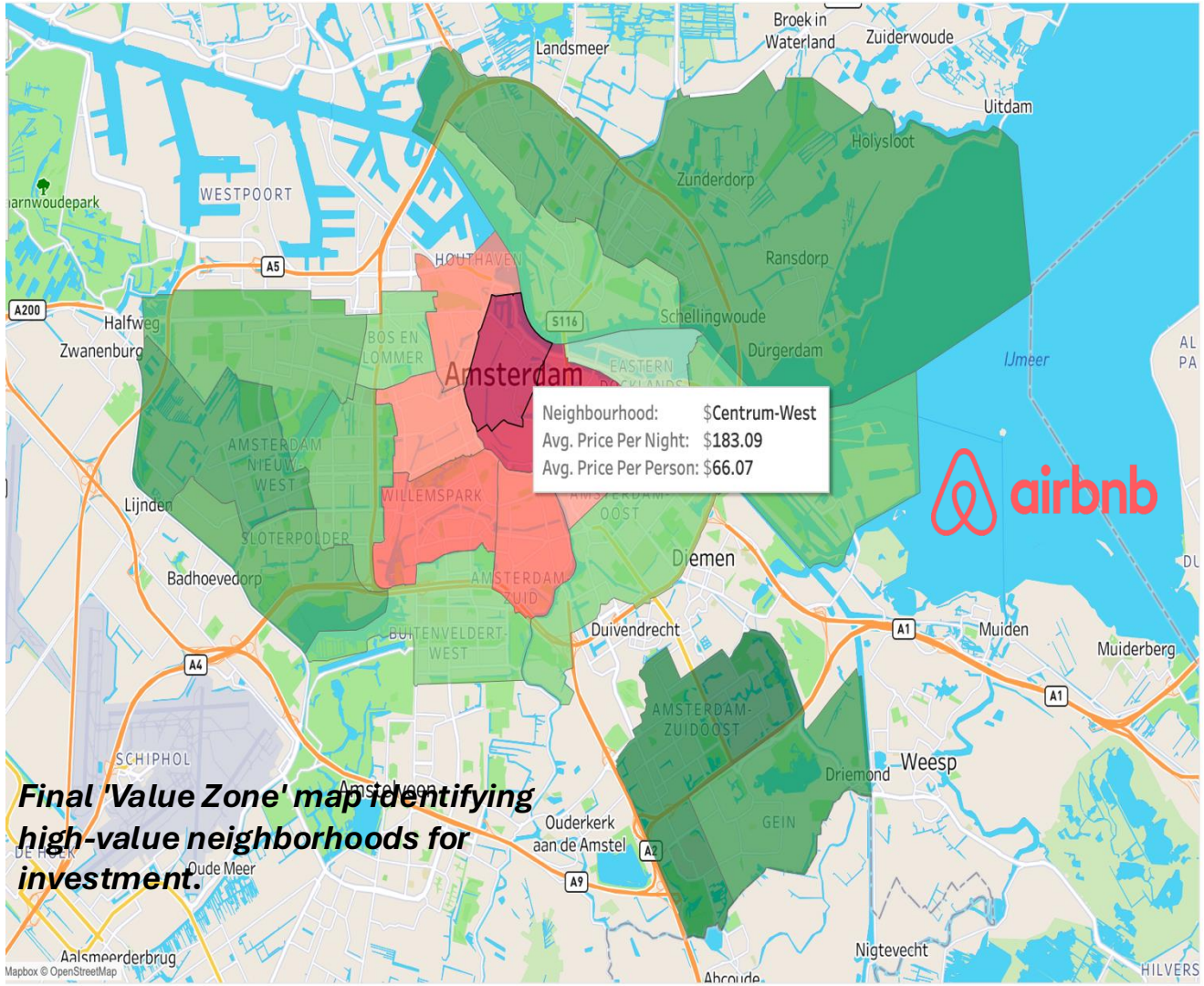
2. How do successful Listings compete and generate revenue??

3. Where are the hidden opportunities for new investors?

Data: Inside Airbnb (Dec 2018), City of Amsterdam

Methodology:

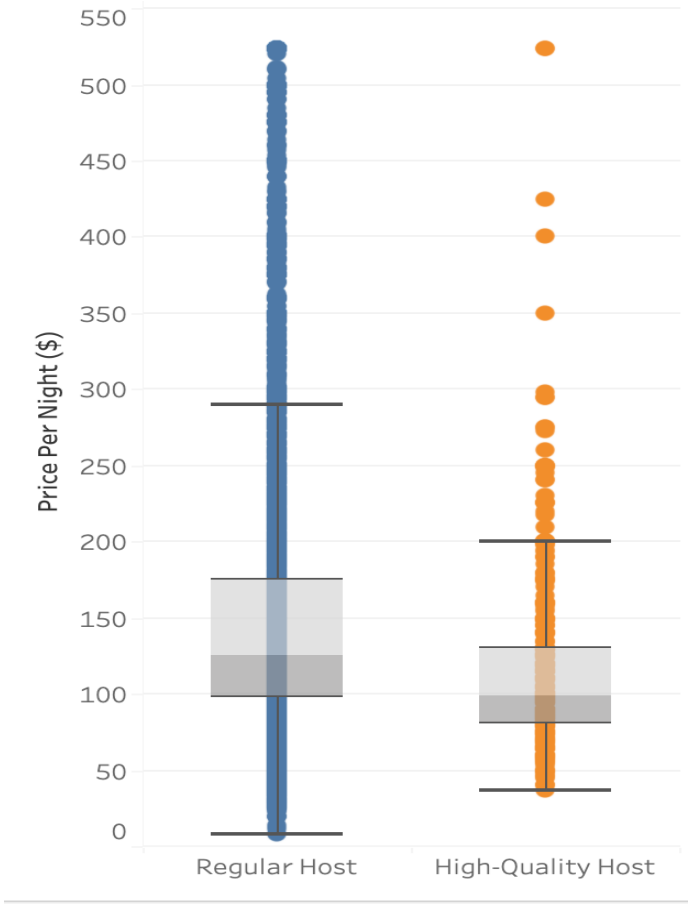
- Data Wrangling & EDA (Python, Pandas)
- Supervised & Unsupervised Machine Learning (Scikit-learn)
- Time-Series Analysis & Forecasting (Statsmodels)
- Interactive Dashboarding (Tableau)



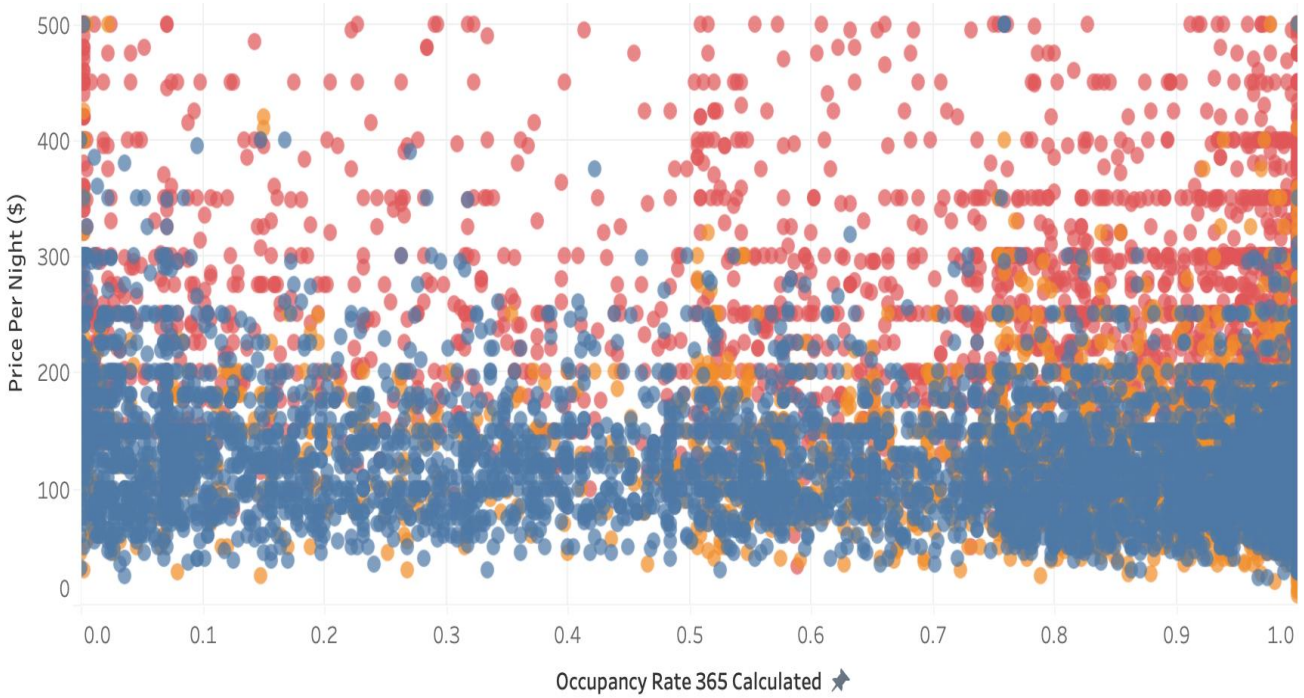
Final 'Value Zone' map identifying high-value neighborhoods for investment.

Tools Used:

From Pricing Myths to Data-Driven Market Segments



Initial finding: Top-rated hosts don't charge a price premium, suggesting a value-based market.



K-means clustering revealed three distinct market personas based on performance.

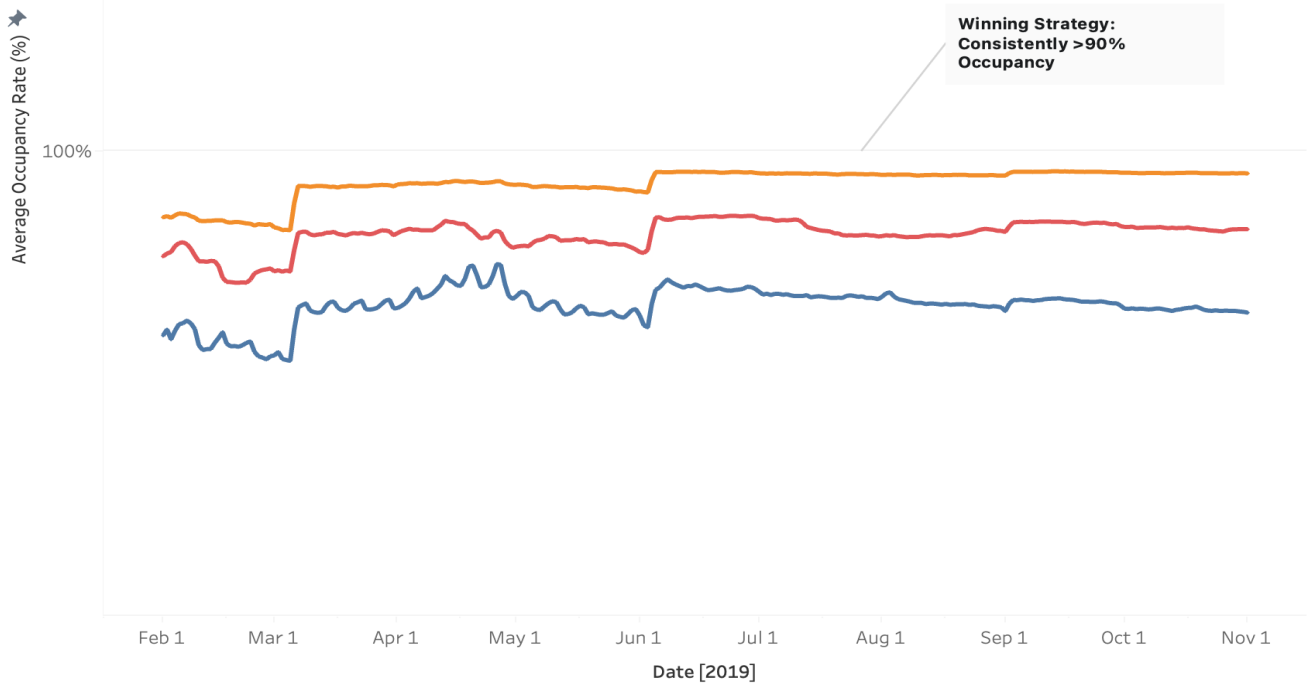
Analytical Journey: My analysis began by disproving the common myth that better hosts charge more. This led me to investigate the true drivers of success. Using k-means clustering, I segmented the entire market into three data-driven personas:

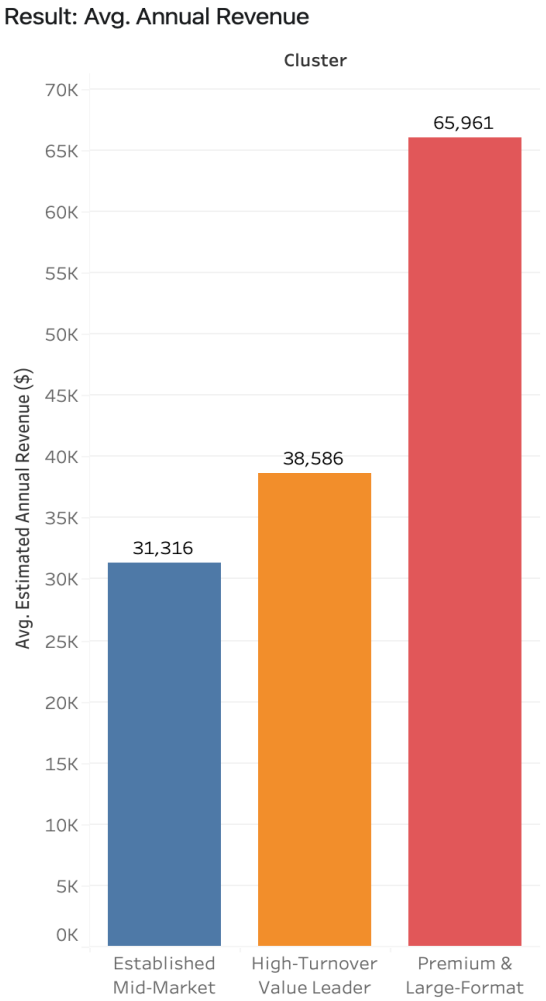
- **Premium & Large-Format:** High-price, high-margin listings.
- **High-Turnover Value Leaders:** Competitively-priced listings with near-constant occupancy.
- **Established Mid-Market:** Older listings with moderate performance.

Revealing Two Paths to Profitability: A Trade-off Analysis

Price vs. Volume Trade-off

Success is a strategic choice between maximizing occupancy and maximizing price.





Key Insight: The analysis revealed that the highest occupancy does not necessarily lead to the highest revenue. This uncovered two distinct, successful strategies in the market:

A **High-Volume** model (driven by the "High-Turnover" cluster) which ensures a consistent revenue stream.

A **High-Margin** model (driven by the "Premium" cluster) which yields the highest potential annual revenue.

Conclusion & Actionable Recommendations

Strategic Recommendations

Path 1: The High. Turnover Strategy (Lower Risk)

What: Focus on smaller (1-2 bedroom) 'Entire home/apt' listings.

Where: Target the **"Value Zone" neighbourhoods** (e.g., Westerpark, Oud-Oost).

How: Price competitively to maximize occupancy (>90%)

Path 2: The Premium Margin Strategy (Higher Risk)

What: Invest in larger, high-end properties (3+ bedrooms)

Where: Target the **premium central districts** (Centrum-West, Zuid).

How: Price for high margins to achieve higher potential annual revenue.

Study Limitations

Analysis based on a 2018-2019 data snapshot. Further analysis would be needed to account for recent market changes.

Project Deliverables:

View the Interactive Storyboard

Explore the Full Code and Analysis

Visualizing 20th Century Geopolitical Networks

Visualizing 20th Century Geopolitics

Objective: To transform unstructured historical text from Wikipedia into a dynamic network graph, uncovering hidden patterns of interrelation between world powers throughout the 20th century.

Key Questions:

- Can we create a structured dataset of country interactions from raw text?
- Who were the most central and influential nations?
- Can data reveal the major geopolitical blocs of the era?

Methodology:

- Web Scraping (BeautifulSoup)
- Natural Language Processing (spaCy for NER)
- Relationship Extraction & Graph Modeling (NetworkX)
- Community Detection & Interactive Visualization (Pyvis)

```
G = nx.from_pandas_edgelist(relationships_df,
                source = "source",
                target = "target",
                edge_attr =  "value",
                create_using = nx.Graph())

print(f"NetworkX graph created successfully.")
print(f"Number of nodes (countries): {len(G.nodes())}")
print(f"Number of edges (relationships): {len(G.edges())}")
```

```
NetworkX graph created successfully.
Number of nodes (countries): 45
Number of edges (relationships): 154
```

Tools Used:

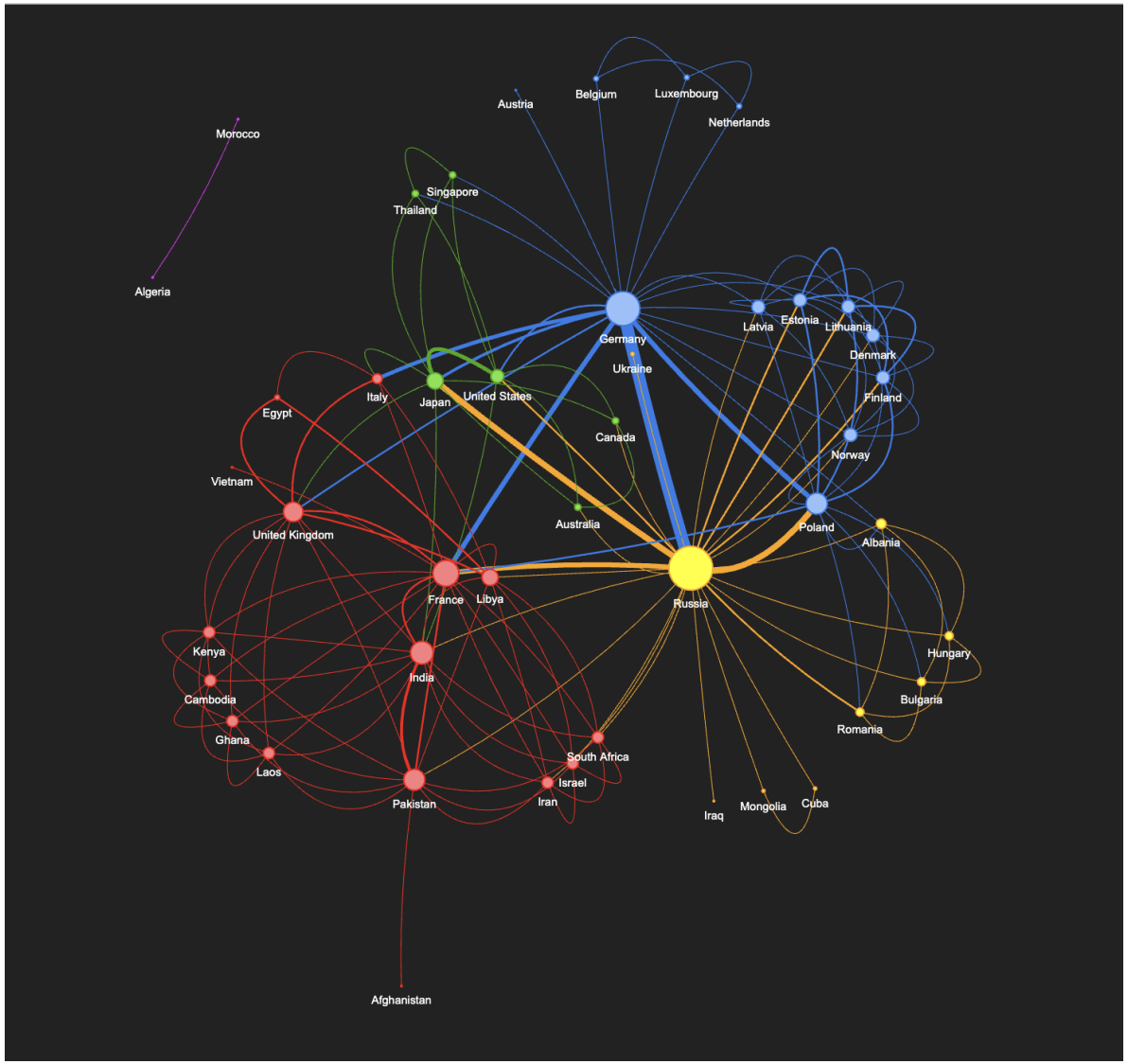
spaCy

NLP Automatically Identifies Geopolitical Blocs

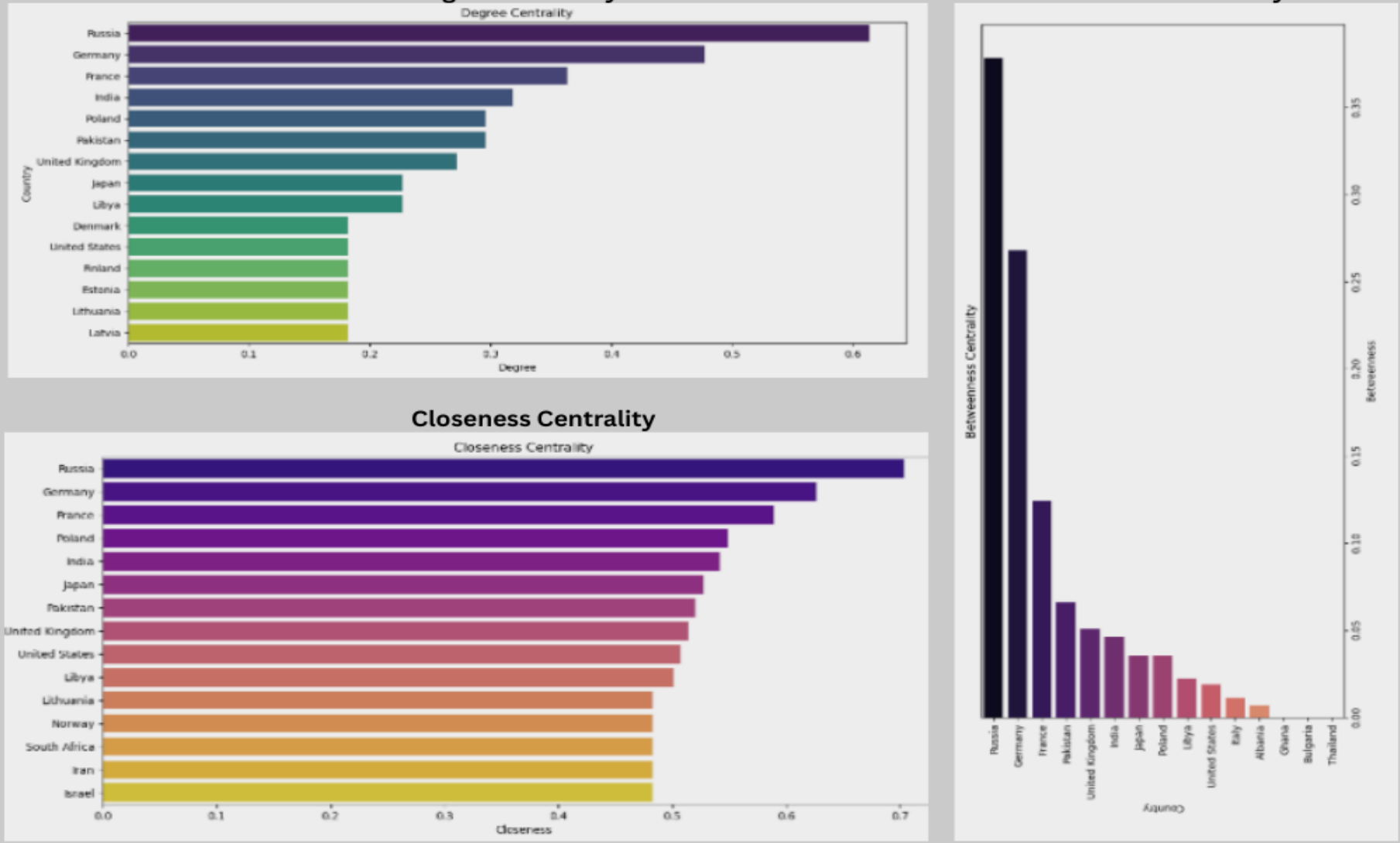
Key Insight:
The most powerful finding was that an unsupervised community detection algorithm, with no prior historical knowledge, successfully partitioned the network into historically coherent blocs. The data itself revealed the structure of the World Wars and the Cold War.

Key Blocs Identified:

- Central European/WWII Bloc (Blue)
- Soviet/Eastern Bloc (Yellow)
- Post-Colonial/UK-Legacy Bloc (Red)



A Deeper Dive: Quantifying Historical Influence



Key Insights:

- **Most Connected (Degree):**

Russia, Germany, and France were the primary "hubs," having direct relationships with the most other nations.

- **Most Efficient (Closeness):**

These same powers could also spread influence most efficiently through the network, as shown by their top Closeness scores.

- **Most Critical "Bridges" (Betweenness):**

Russia and Germany had overwhelmingly high scores, confirming their roles as the essential connectors between the East, West, and other global spheres.

Conclusion & Project Reflection

Conclusion:
This project successfully demonstrated that advanced data science techniques can create valuable, structured insights from a source as seemingly chaotic as a historical article. It proves that measurable patterns of historical relationships are embedded within unstructured text.

Challenges & Lessons Learned:
The greatest challenge was the extensive data wrangling required to standardize historical names (e.g., "Soviet Union," "Prussia"). The project was a powerful lesson in the importance of a robust data cleaning pipeline and the creative application of NLP to problems where no clean dataset exists.

Project Deliverables:

Explore the Full Code and Analysis

NYC Citi Bike Strategic Dashboard

NYC Citi Bike: A Strategic Analysis

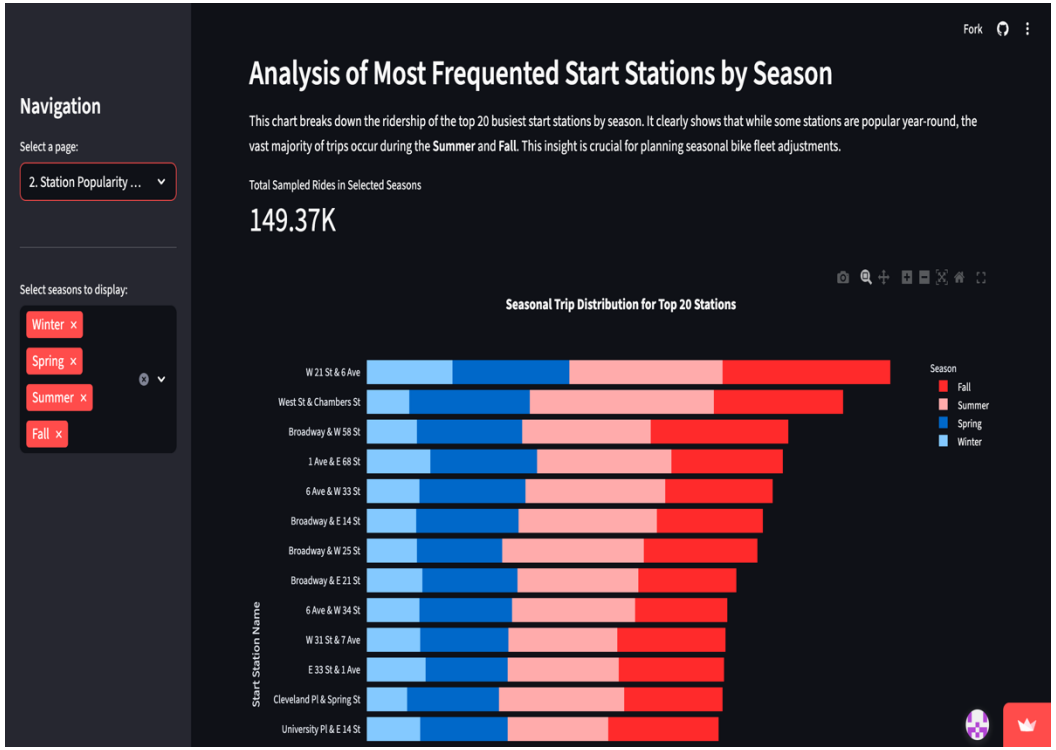
Objective: To analyze over 30 million trip records to develop a data-driven strategy for optimizing bike distribution, managing seasonal demand, and identifying expansion opportunities.

Key Questions:

1. What are the most popular stations and routes?
2. How does ridership change with the seasons?
3. Where are the primary "bike highways" of New York City?

Methodology:

- Data Sourcing & Enrichment (API)
- Large-Scale Data Aggregation (Pandas)
- Interactive Dashboard Development (Streamlit)
- Geospatial & Time-Series Visualization (Plotly, Kepler.gl)



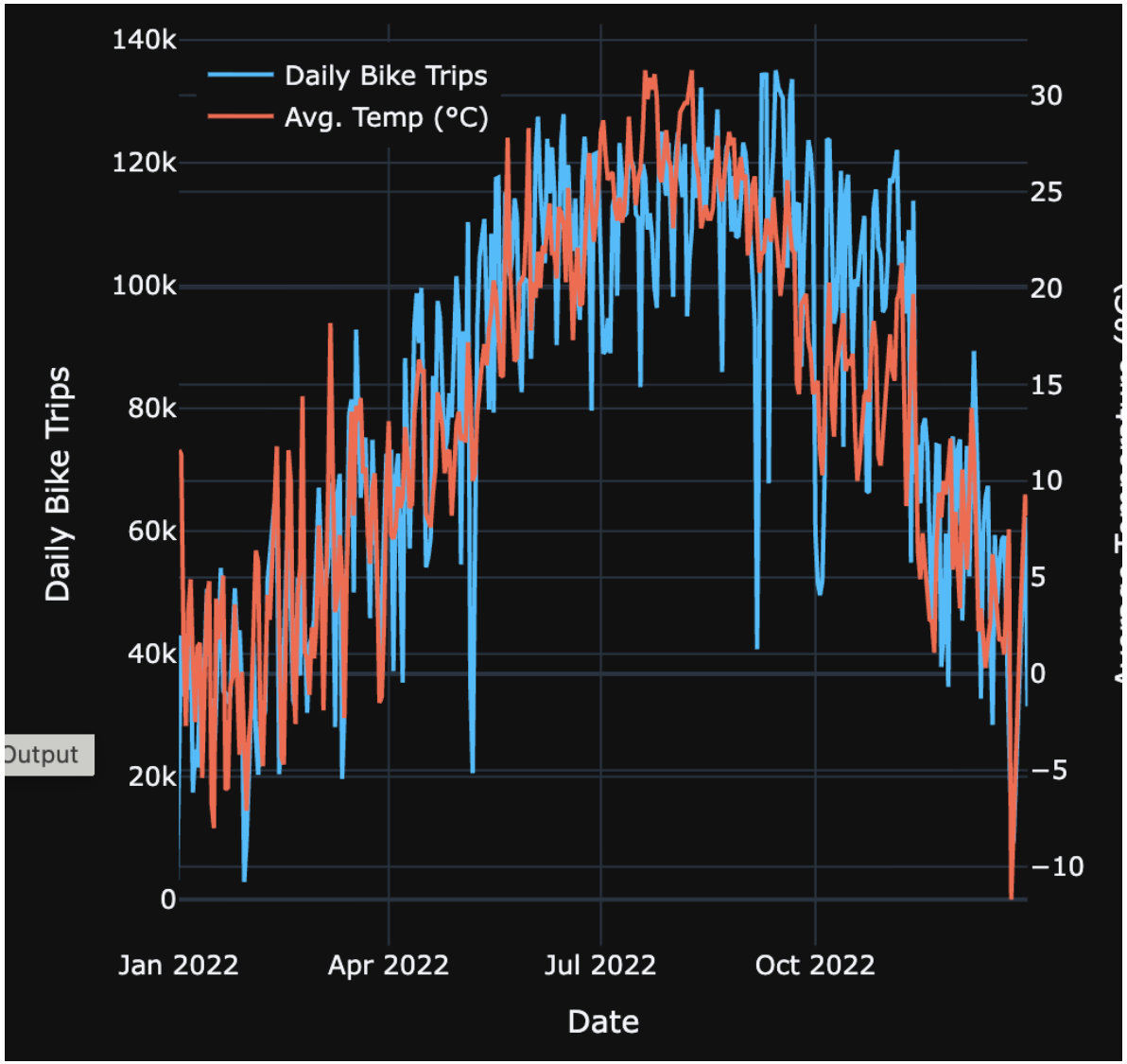
Tools Used:

Python, plotly, kepler.gl, Streamlit

Uncovering the Power of Seasonality

Ridership is Driven by Temperature:

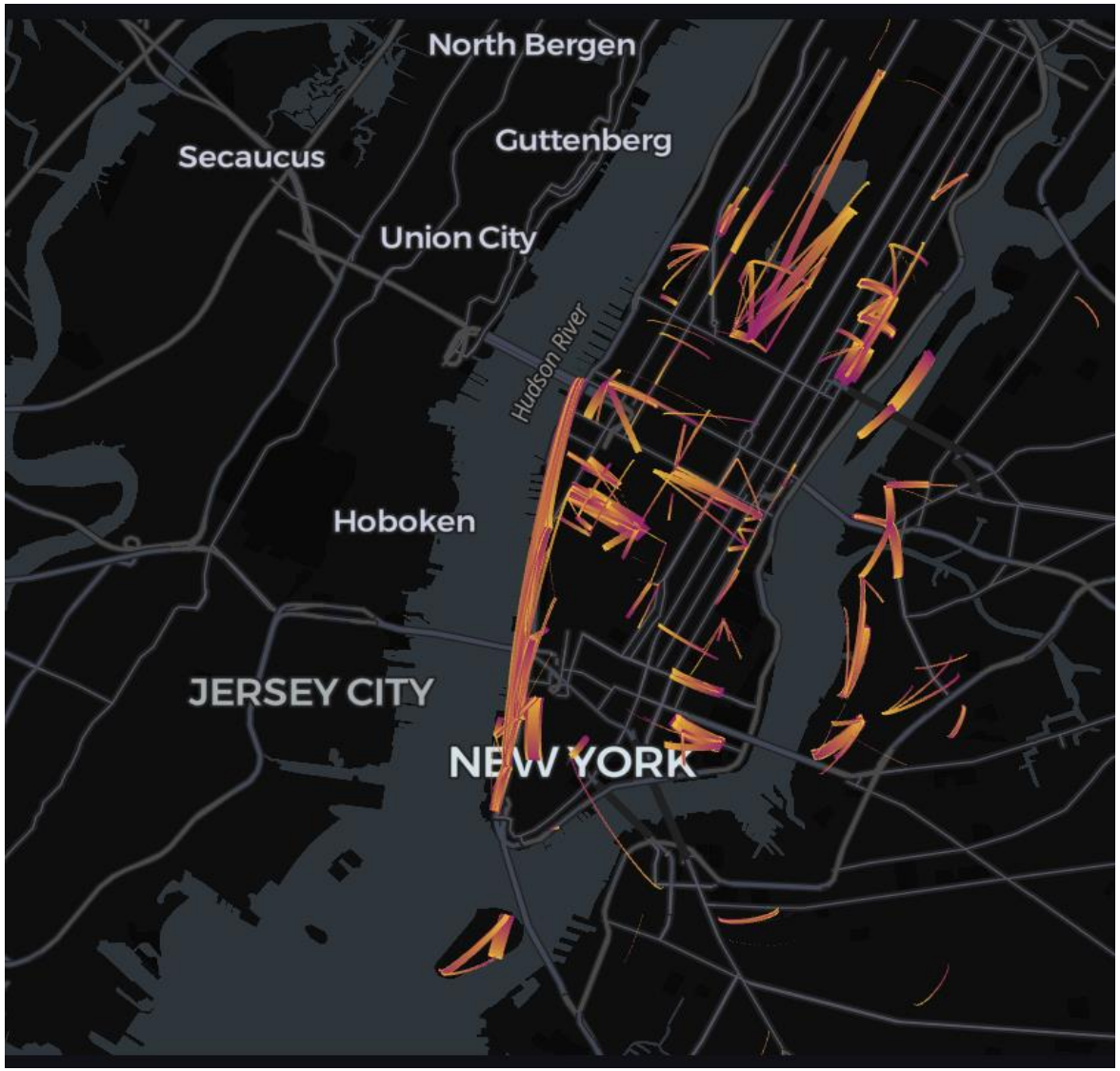
- The analysis revealed an undeniable and powerful correlation between temperature and bike usage. Ridership soars in the summer and plummets in the winter, providing a predictable pattern for managing the entire fleet.
- A "one-size-fits-all" year-round operational strategy is inefficient and costly.



A Deeper Dive: Identifying High-Demand Corridors

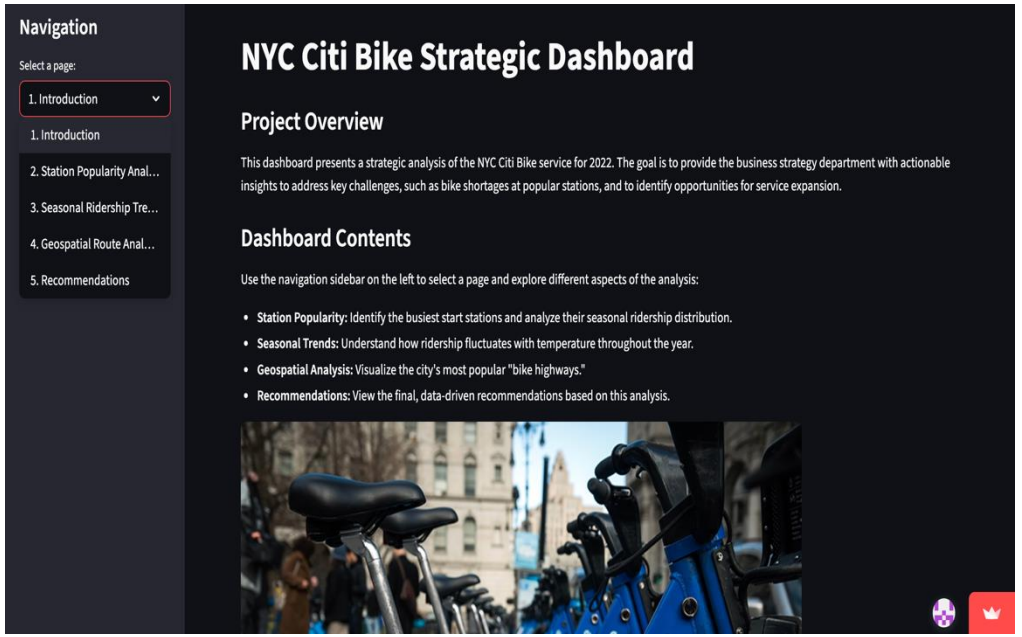
Key Insights:

- **Manhattan is the Epicenter:** The vast majority of high-traffic routes are concentrated in Midtown and Lower Manhattan.
- **Bridges are Critical Arteries:** The Williamsburg and Brooklyn Bridges are the most important commuter corridors connecting the boroughs.
- **Waterfronts are Key:** Routes along the Hudson and East Rivers are extremely popular for both commuting and recreation.



Conclusion & Actionable Recommendations

- **Strategic Recommendations):**
  - **Path 1: The Seasonal Strategy (Lower Cost):** Scale back the active fleet by 40-50% in the winter months (Nov-Mar) to dramatically reduce operational costs.
  - **Path 2: The Rebalancing Strategy (Higher Efficiency):** Focus logistical efforts on keeping the high-demand "bike highways" and commuter hubs stocked during peak hours.
  - **Path 3: The Expansion Strategy (Future Growth):** Use the geospatial data to prioritize new station installations in underserved areas that lie between popular start and end points.



Project Deliverables:

View the Live IDashboard Streamlit

Explore the Full Code

Next Project

Dataset:

https://www.kaggle.com/datasets/computingvictor/transactions-fraud-datasets