ANUSHMA SHARMA

DATA ANALYST

ABOUT ME

I’m Anushma Sharma, a data analyst passionate about transforming data into actionable insights. With a background in technical design, I bring strong analytical skills, problem-solving, and attention to detail to data-driven projects. I specialize in trend analysis, data wrangling, and creating impactful visualizations and I am proficient in SQL, Python, Tableau, and Excel. Explore my portfolio to see how I uncover insights and drive informed decision-making.

01
NEW YORK'S CITIBIKE
A comprehensive data dashboard analyzing New York Citi Bike trips for the year 2022

02
ZOMATO
Help Zomato and restaurants enhance satisfaction and growth

03
INSTACART
Strategic marketing plan for the grocery shopping app

04
ROCKBUSTER STEALTH LLC
Insights to optimize content, revenue, and market strategy for online growth.

05
INFLUENZA SEASON
Analysis of staffing requirements for healthcare facilities

06
GAMECO.
Analyze regional sales trends to assess shifts over time.

PROCESS OVERVIEW

Each project begins with identifying the problem and setting clear objectives. Using tools like SQL, Python, Excel, and Tableau, I clean, integrate, and validate data to ensure accuracy for analysis. I draw insights through statistical and descriptive techniques, and I create visualizations to make the findings clear and actionable.

- **Citibike Trip Analysis (2022):** This project examines Citi Bike trip data from 2022 to uncover key usage patterns, identify factors affecting ride volume, and provide strategic recommendations for optimization.
- **Zomato Bengaluru Restaurants:** Analyzed restaurant performance and customer preferences to recommend pricing and marketing improvements.
- **Instacart:** Identified purchasing trends to enhance customer segmentation and improve marketing strategies.
- **Rockbuster Stealth LLC:** Evaluated customer and revenue data to guide an online rental platform's market strategy.
- **Influenza Staffing:** Examined population data to optimize healthcare resource allocation during influenza seasons.
- **GameCo:** Analyzed global sales trends and popular game genres, providing marketing recommendations for regional optimization.

Each project overcame challenges such as incomplete data and balancing computational efficiency. Recommendations were tailored to stakeholder needs, delivering solutions that drive results

TOOLS UTILIZED

EXCEL

TABLEAU

SQL

PYTHON

GITHUB

NewYork's CitiBike

CONTEXT

This project examines Citi Bike trip data from 2022 to uncover key usage patterns, identify factors affecting ride volume, and provide strategic recommendations for optimization.

PROJECT GOAL

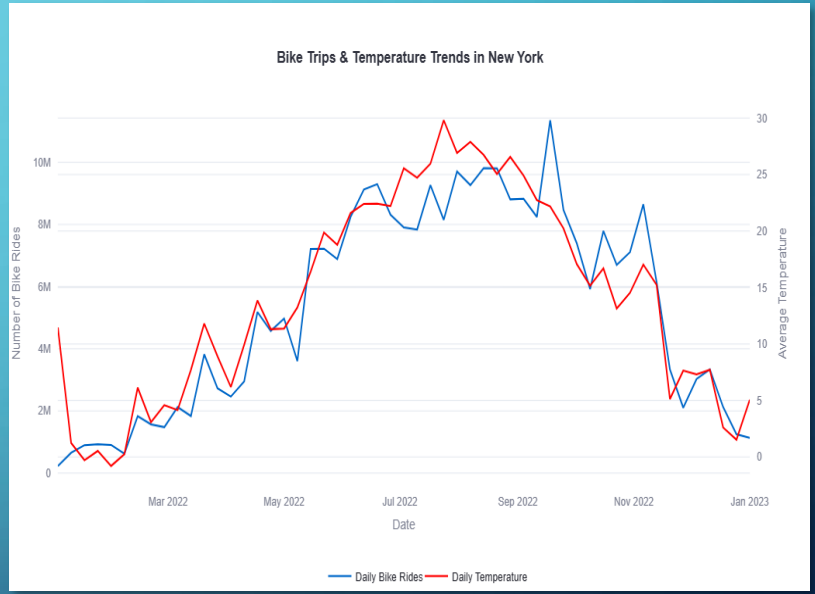
- Analyze how weather conditions impact bike usage.
- Identify the top 20 most popular stations.
- Examine trip patterns and common travel routes.
- Compare subscriber vs. casual user behavior.
- Provide actionable recommendations for system improvements.

SKILLS DEMONSTRATED

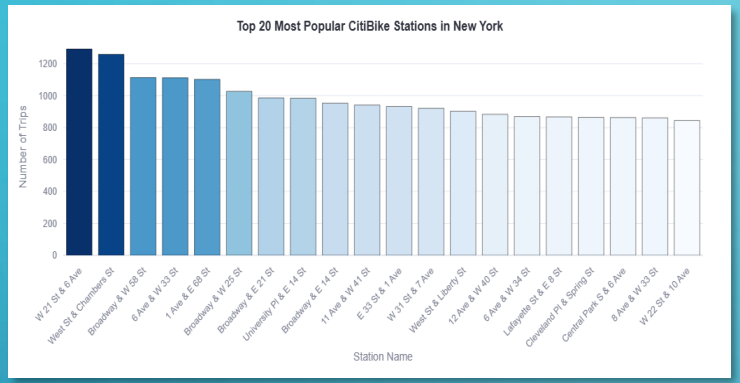
- Processed raw Citi Bike data using Python.
- Identified seasonal trends, peak hours, and ride duration patterns with Exploratory Data Analysis (EDA).
- Created interactive dashboards in Streamlit for stakeholder insights.

ANALYSIS AND VISUALIZATIONS

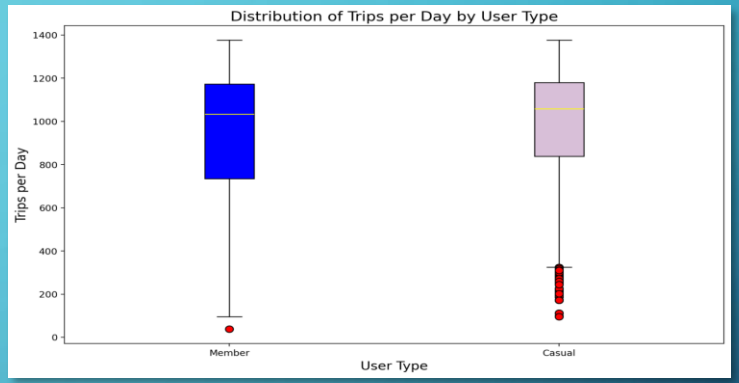
The chart visualizes the relationship between daily bike trips and the average temperature in New York throughout 2022, resampled weekly to highlight trends. The blue line represents bike rides, while the red line shows temperature variations. Warmer weather, especially in spring and summer, correlates with increased ridership, while colder days decline. Temperature spikes often precede a rise in bike usage, emphasizing weather's impact on demand. This insight helps optimize bike distribution and availability based on seasonal patterns.



ANALYSIS AND VISUALIZATIONS



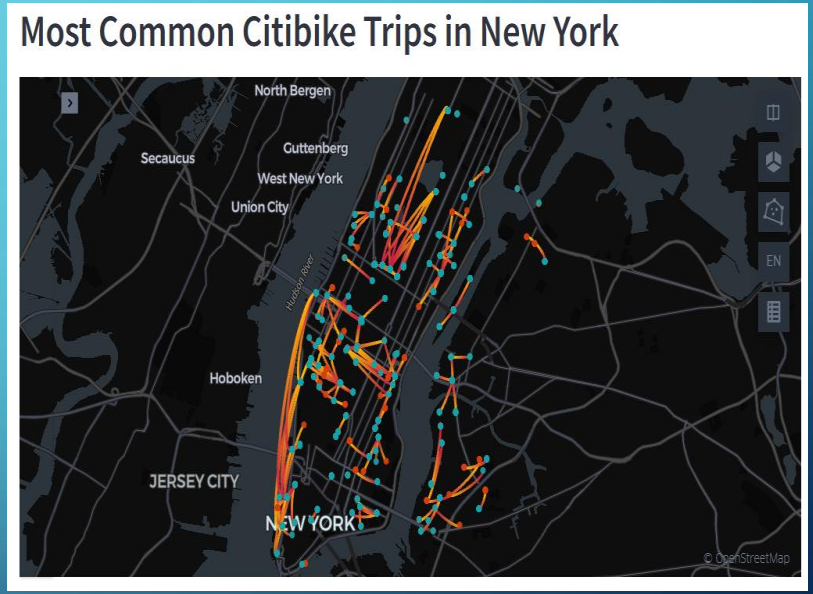
This bar chart showcases the top 20 most popular Citi Bike stations in New York for 2022 based on trip counts. Stations like W 21 St & 6 Ave, West St & Chambers St, and Broadway & W 58 St saw the highest demand, exceeding 1,000 trips. High-ranking stations are concentrated in business districts, transit hubs, and tourist areas, highlighting the need for efficient bike distribution and additional docking spaces to meet demand.



This box plot visualizes the distribution of daily Citi Bike trips in 2022 by user type. Members show greater variability in trips, with a wider range and higher fluctuations, while casual riders have a narrower distribution with occasional low-trip outliers. Members' frequent and varied usage may contribute to supply-demand imbalances, especially during peak periods. Understanding these patterns helps optimize bike availability and station management.

ANALYSIS AND VISUALIZATIONS

This map visualizes the most common Citi Bike trips in New York during 2022, with thicker lines indicating higher trip frequencies. Central Park S & 6 Ave is the busiest station (135 trips), followed by Roosevelt Island Tramway, Soissons Landing, and 5 Ave & E 72 St. These locations, near tourist attractions and parks, suggest high demand for short, leisurely rides rather than long-distance commuting. Understanding these patterns helps optimize bike distribution in high-traffic areas.



KEY FINDINGS AND RECOMMENDATIONS

Citi Bike's 2022 analysis shows supply-demand imbalances, with peak-hour congestion at Central Park S & 6 Ave and Roosevelt Island Tramway. Ridership fluctuates with weather, casual riders dominate tourist areas, and members contribute to trip variability, highlighting the need for better bike distribution and dock expansion.

Citi Bike should use real-time redistribution, expand docks in high-demand areas, and adjust supply based on weather. Targeted membership incentives and a real-time dashboard can further optimize operations, ensuring a smoother riding experience across NYC.

Dashboard Link

Zomato Bengaluru Restaurant Trends

CONTEXT

The dataset explores restaurant performance in Bengaluru, focusing on customer ratings, votes, price range, cuisine popularity, and location trends. It aims to uncover patterns in customer preferences and engagement.

PROJECT GOAL

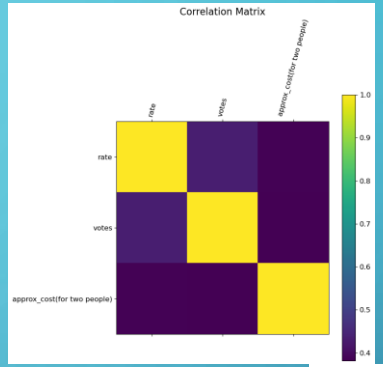
To provide actionable insights for Zomato and restaurant owners, helping them enhance customer satisfaction, optimize marketing strategies, and improve business outcomes.

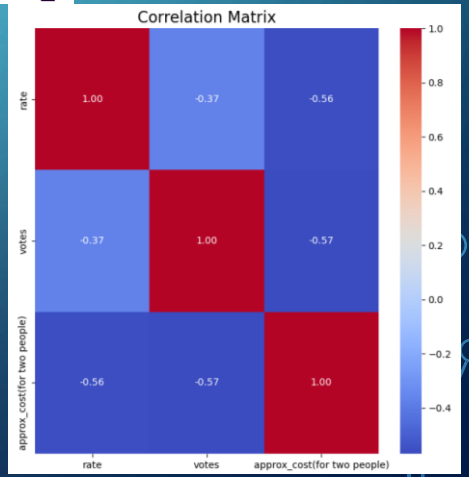
SKILLS DEMONSTRATED

- Sourced and prepared open data for analysis
- Conducted exploratory data analysis using Python
- Created visualizations and dashboards with Python and Tableau
- Applied supervised machine learning regression for prediction and performed unsupervised machine learning clustering to identify groups

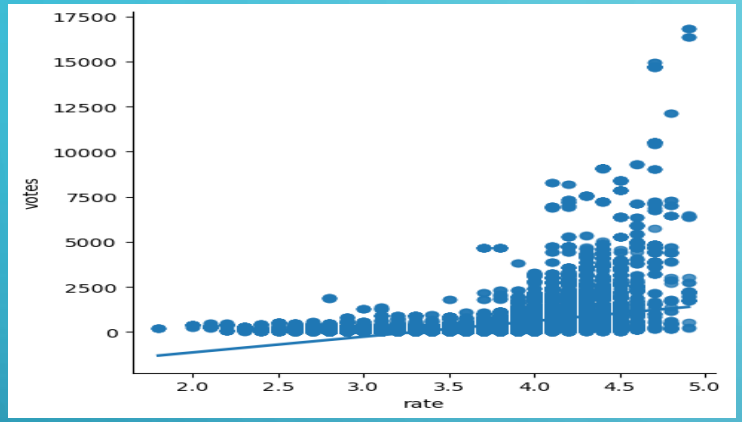
ANALYSIS AND VISUALIZATIONS

The correlation matrix reveals key relationships between ‘rate’, ‘votes’, and ‘approx_cost (for two people)’. A moderate negative correlation between rate and cost suggests that higher costs often result in lower ratings, likely due to unmet expectations. Similarly, the negative correlation between votes and cost indicates that pricier restaurants attract fewer customers, leading to fewer votes. The weak negative correlation between rate and votes suggests that customer engagement (votes) has little impact on ratings. Overall, affordability strongly influences both customer satisfaction and engagement.

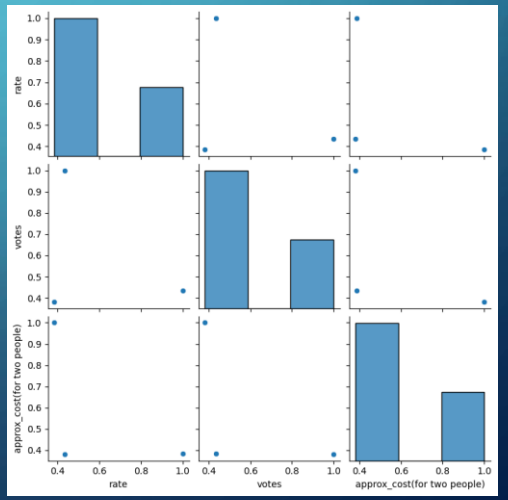




ANALYSIS AND VISUALIZATIONS



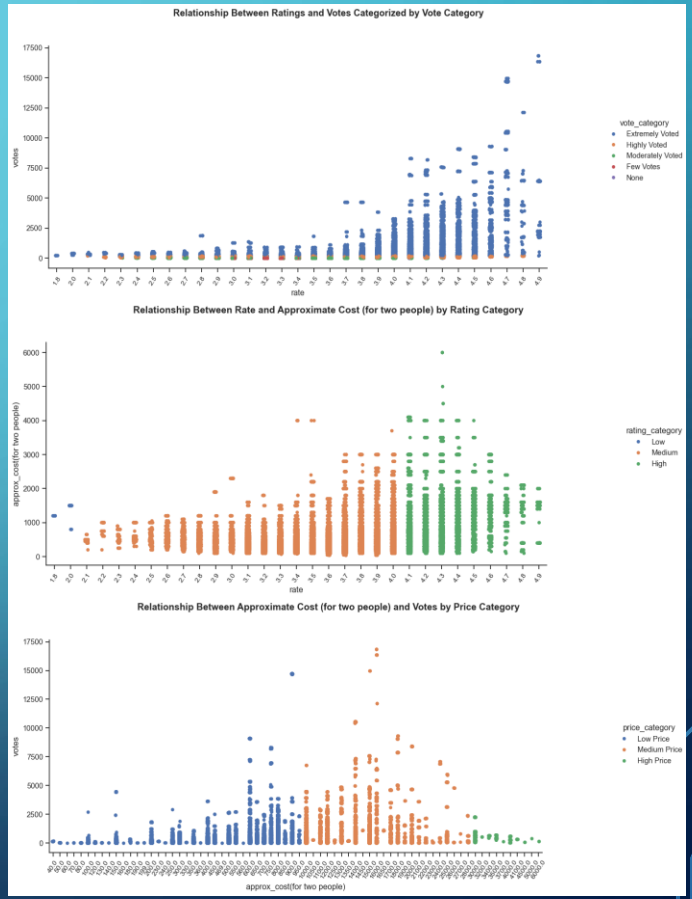
The scatter plot shows a positive correlation between restaurant ratings and the number of votes. As ratings increase from 2.0 to 5.0, the number of votes generally rises, with a noticeable concentration of highly rated restaurants (above 4.0) receiving significantly more votes. This suggests that higher-rated restaurants are more popular and attract greater customer engagement.



The pair plot shows that each variable (rate, votes, and approx_cost) has a skewed distribution, with a high concentration of data points at lower values. There are weak positive trends between 'rate' and 'votes' and between 'approx_cost' and 'rate', suggesting slight associations between rating, popularity, and cost. Further exploration of 'rate' and 'votes' could reveal insights into customer preferences while examining 'approx_cost' and 'rate' might show if higher-rated restaurants are generally more expensive.

ANALYSIS AND VISUALIZATIONS

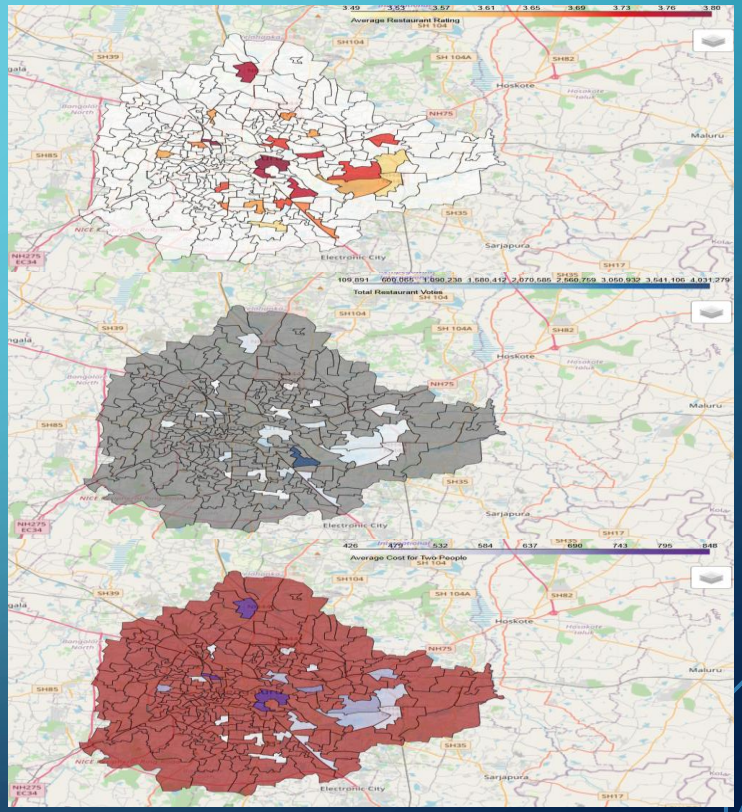
Higher-rated restaurants (4.0 and above) receive significantly more customer engagement, with exceptional ratings (4.5–4.9) driving strong popularity. Moderate pricing (₹500–₹1500 for two) attracts the most votes, reflecting customer preference for this range, while high-priced venues engage a niche audience. High-rated restaurants span premium and affordable prices, whereas lower-rated establishments (below 3.0) are generally cheaper and receive fewer votes. Maintaining high ratings and offering value through moderate pricing can maximize customer interaction and popularity.



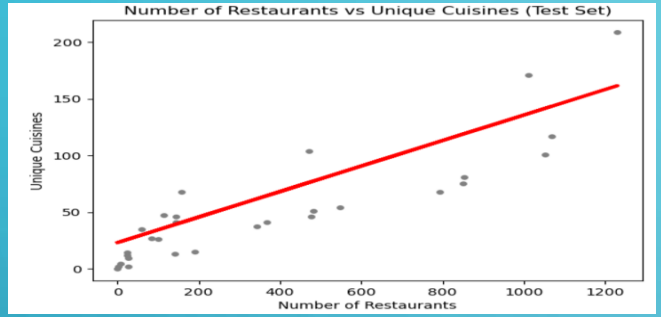
ANALYSIS AND VISUALIZATIONS

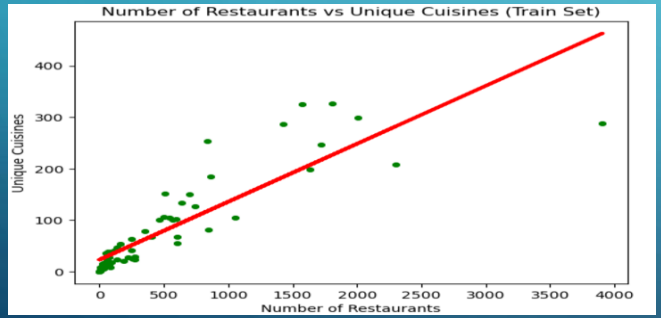
The analysis of Bengaluru's wards reveals distinct patterns in food quality, customer engagement, and dining costs. High-rated wards like Koramangala, BTM Layout, and Shantala Nagar showcase superior food quality (above 3.7) and significant customer engagement, reflected in high vote counts and premium dining options. Moderate-rated areas such as HSR Layout and Rajaji Nagar balance affordability with good quality (3.6–3.7) and steady customer participation. In contrast, lower-rated wards like Dodda Nekkundi and Arakere (below 3.6) tend to offer budget-friendly dining but receive fewer votes, indicating limited customer engagement. These insights emphasize the diverse dining landscape and customer preferences across the city.

LINK 1
LINK 2
LINK 3



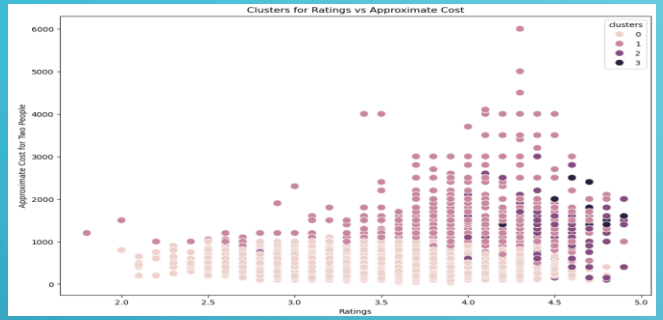
ANALYSIS AND VISUALIZATIONS





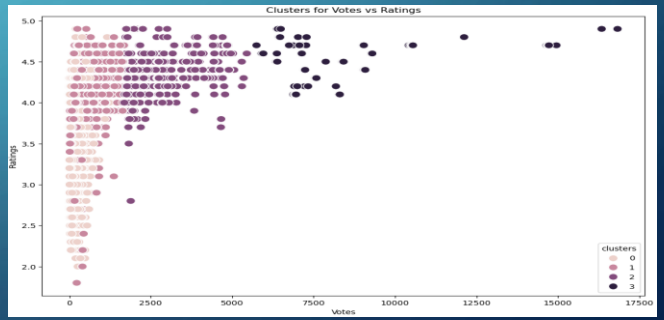
Both the test and train sets show a strong positive correlation between the number of restaurants and unique cuisines, with identical slopes (0.1127), indicating consistency. The R^2 scores (test: 0.7038, train: 0.7616) highlight good model performance, though slightly better on the training set, suggesting potential overfitting. Higher MSE in the training set (1895.42 vs. 714.63 for the test set) may be due to more data or outliers. While the model captures trends well, predictions at higher restaurant counts are less reliable, pointing to opportunities for fine-tuning.

ANALYSIS AND VISUALIZATIONS



The clusters show expected patterns: low-cost, low-rated restaurants in Cluster 0, mid-cost, mid-rated in Cluster 1, and high-cost, highly-rated in Clusters 2 and 3. Higher costs generally align with better ratings, though overlaps in mid-cost ranges and sparse high-cost representation suggest refinement opportunities. Outliers with high costs and low ratings highlight anomalies for further investigation. Adding variables like votes, cuisine type, or location could improve clustering and insights.

The clustering of Votes vs. Ratings effectively distinguishes restaurants by popularity and quality. Cluster 0 represents low-visibility restaurants, Cluster 1 includes moderately popular ones, and Clusters 2 and 3 capture highly popular, well-rated establishments, with Cluster 3 highlighting top-tier performers. A positive trend between votes and ratings is evident. Overlap in mid-range votes and a dense Cluster 0 suggest potential data imbalance. Adding features like cost or cuisine and refining the algorithm could enhance segmentation. Outliers in Cluster 3 warrant further analysis to identify success factors.



KEY FINDINGS AND RECOMMENDATIONS

The analysis shows a positive correlation between restaurant ratings and customer engagement, with moderately priced restaurants (₹500–₹1500 for two) attracting the most votes. High-rated, high-cost establishments receive niche engagement, while budget-friendly options often have lower ratings. Geographically, areas like Koramangala and Shantala Nagar show high participation, while Dodda Nekkundi reflects lower engagement. Clustering reveals clear patterns between cost, ratings, and popularity, though data imbalances and outliers suggest areas for improvement.

Recommendations include maintaining high ratings through quality service, targeting marketing for moderately priced restaurants, and enhancing Zomato's recommendations by factoring in cuisine type and location. Further analysis of outliers and refining clustering models can provide better insights and segmentation.

GITHUB REPOSITORY
TABLEAU LINK

Instacart

CONTEXT

This project involves a comprehensive analysis of Instacart's sales data to identify patterns in customer purchasing behavior. Stakeholders seek insights to inform a targeted marketing strategy aimed at various customer segments.

PROJECT GOAL

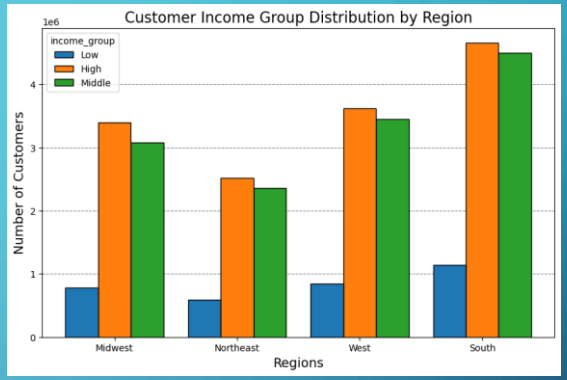
The main objective is to generate actionable insights and recommend strategies to improve customer segmentation, enhancing the effectiveness of marketing efforts.

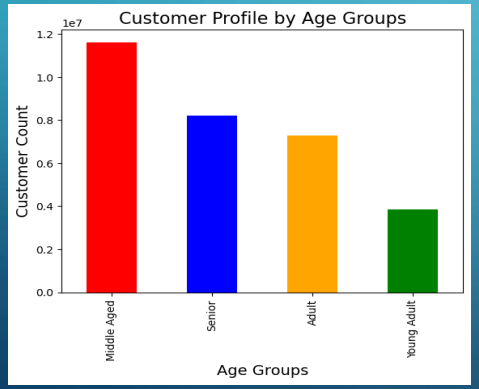
SKILLS DEMONSTRATED

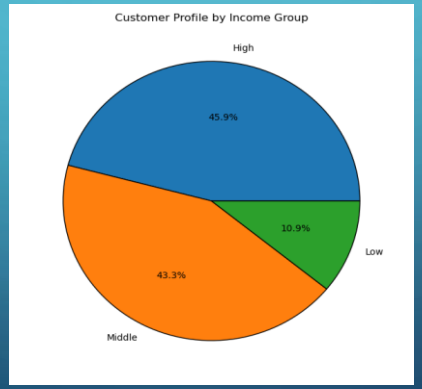
- Python and Jupyter Notebook proficiency
- Data wrangling and data subsetting
- Performing consistency checks
- Merging and integrating data from multiple dataframes
- Creating new variables for deeper analysis
- Grouping and aggregating data to identify trends
- Visualizing data using Python libraries

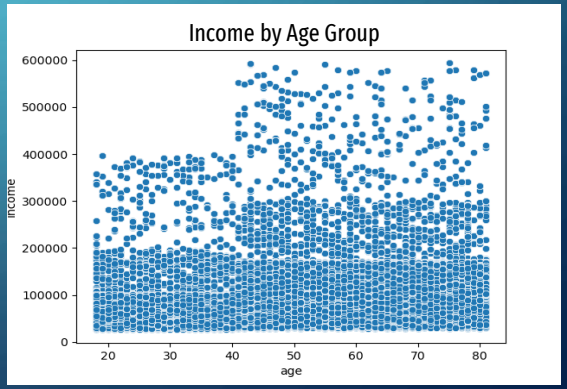
ANALYSIS AND VISUALIZATIONS

Key patterns were uncovered, including peak shopping times and popular product categories, along with an in-depth analysis of customer purchasing behaviors. Outliers were also removed to improve dataset reliability. This thorough analysis yielded valuable insights to guide targeted marketing campaigns.





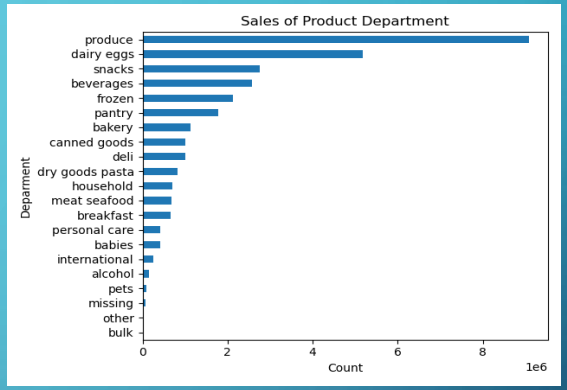


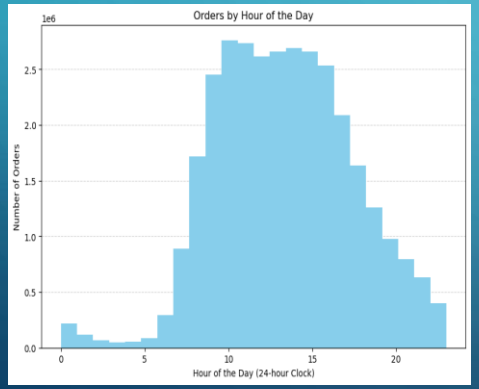


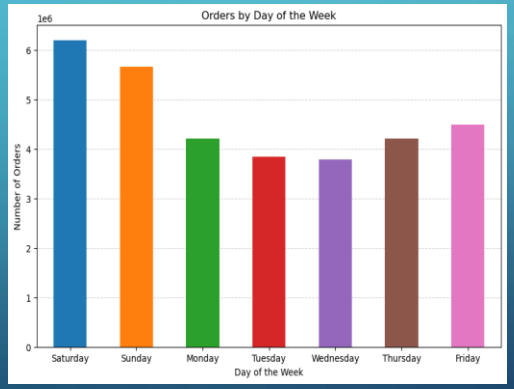
ANALYSIS AND VISUALIZATIONS

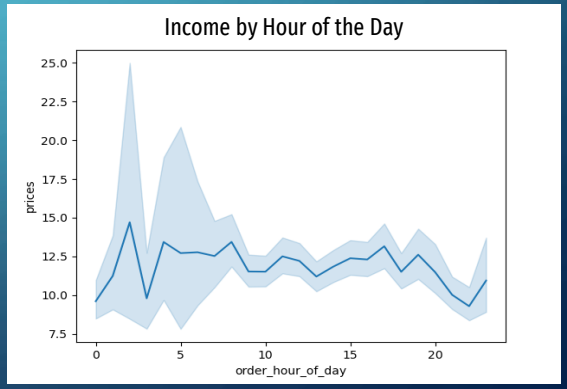
Saturday and Sunday are the busiest days of the week, while Tuesday and Wednesday experience the least activity. Most orders are placed from late morning through early afternoon, peaking between 9 AM and 4 PM, with fewer orders occurring in the early morning and late evening.

The top-selling departments are produce, dairy and eggs, snacks, and beverages, a trend consistent across different age groups and regional analyses.









KEY FINDINGS AND RECOMMENDATIONS

Saturday and Sunday are the busiest days, peaking mid-morning, while Tuesday and Wednesday are the slowest. Midweek discounts could help boost sales.

Top categories include produce, dairy, snacks, and beverages. Targeted promotions on popular and less-sold items may increase orders.

With a primary demographic of middle-aged, high-income parents, Instacart could focus on family-oriented discounts. To compete with in-store shopping, partnerships with stores offering exclusive delivery discounts may attract more users.

GITHUB REPOSITORY

Rockbuster Stealth llc

CONTEXT

Rockbuster Stealth LLC, once a leader in physical movie rentals, is transitioning to an online rental service to compete with streaming platforms like Netflix and Amazon Prime.

PROJECT GOAL

This presentation provides data-driven insights to optimize the online platform for content offerings, pricing strategies, and regional market focus.

DESCRIPTIVE STATISTICS

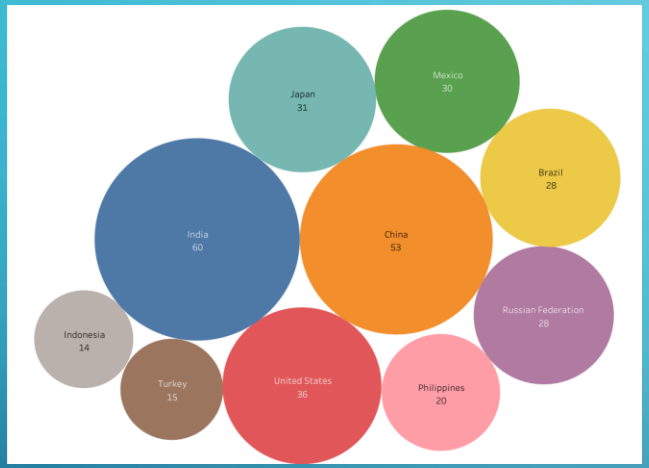
LENGTH: Min - 46 min, Max - 185 min, Avg - 115 min
RENTAL RATE: Min - $0.99, Max - $4.99, Avg - $2.98
RENTAL DURATION: Min – 3 days, Max – 7 days, Avg – 4.98 days
REPLACEMENT COST: Min - $9.99, Max - $29.99, Avg - $19.98
MOST FREQUENT:
RATING: PG-13
LANGUAGE: English
RELEASE YEAR: 2006

SKILLS DEMONSTRATED

- Gather customer, movie, and payment data using SQL.
- Calculate metrics like average rental duration, revenue per movie, and customer count per country.
- Create data visualizations in Tableau to illustrate key insights such as top revenue-generating movies and geographic revenue distribution.
- Develop strategic recommendations based on the data analysis to improve Rockbuster's market positioning and online expansion.

ANALYSIS AND VISUALIZATIONS

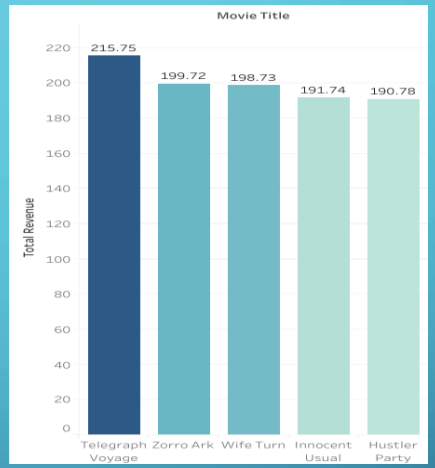
TOP 10 COUNTRIES BY CUSTOMER NUMBERS



- Top countries in Rockbuster's customer base: **India** and **China**.
- Next largest markets: **U.S.** and **Japan.**
- Growth potential markets: **Mexico, Brazil,** and **Russia**.
- Expansion opportunities in smaller markets: **Philippines, Turkey,** and **Indonesia.**

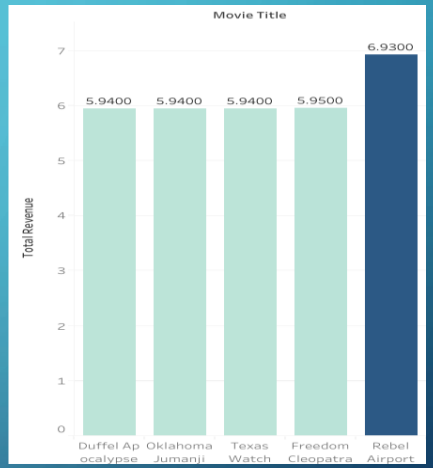
TABLEAU LINK

TOP 5 & LEAST 5 REVENUE CONTRIBUTING MOVIES



Top 5 Movies:
1. Telegraph Voyage
2. Zorro Ark
3. Wife Turn
4. Innocent Usual
5. Hustler Party

TABLEAU LINK



Least 5 Movies:
1. Oklahoma Jumanji
2. Duffel Apocalypse
3. Texas Watch
4. Freedom Cleopatra
5. Rebel Airport

TABLEAU LINK

ANALYSIS AND VISUALIZATIONS

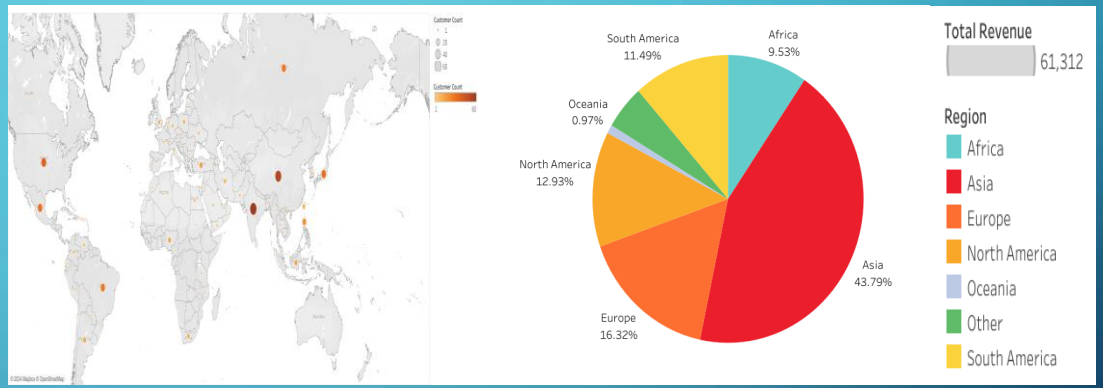
TOP MOVIE GENRES

Top Movie Genres:

1. Sports
2. Sci-Fi
3. Animation
4. Drama
5. Comedy

TABLEAU LINK

CUSTOMER & SALES DISTRIBUTION BY GEOGRAPHIC REGIONS



- **Asia** leads revenue with 43.79%, followed by **Europe** (16.32%) and **North America** (12.93%).
- **South America** and **Africa** contribute moderately at 11.49% and 9.53%.
- **Oceania,** with the lowest share (0.97%), suggests potential for growth focus.

TABLEAU LINK: CUSTOMER DISTRIBUTION
SALES DISTRIBUTION

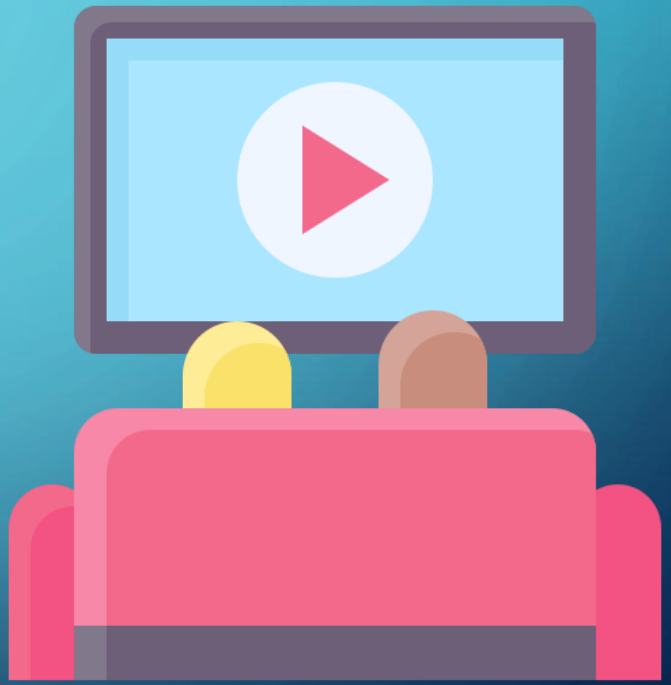
KEY FINDINGS AND RECOMMENDATIONS

Rockbuster should expand high-demand genres like Sports, Sci-Fi, and Animation, tailor content for top markets like Asia, and diversify with niche genres such as Foreign Films to attract wider audiences.

Regional efforts should focus on high-revenue Asia, with additional targeted expansions in Oceania and Africa. Collaborating with local creators in emerging markets like South America and Africa can also boost growth.

Marketing should prioritize PG-13 and NC-17 movies for high returns and promote PG and G-rated films to build family audience loyalty. This approach will enhance global reach, audience engagement, and revenue.

GITHUB REPOSITORY



Project Plan For Upcoming Influenza Season

CONTEXT

This project aims to support effective resource allocation and staffing for influenza season by understanding the impact of influenza on both vulnerable and non-vulnerable populations across U.S. states. Vulnerable groups include individuals below 5 years and above 65 years, who face higher risks of severe outcomes from influenza.

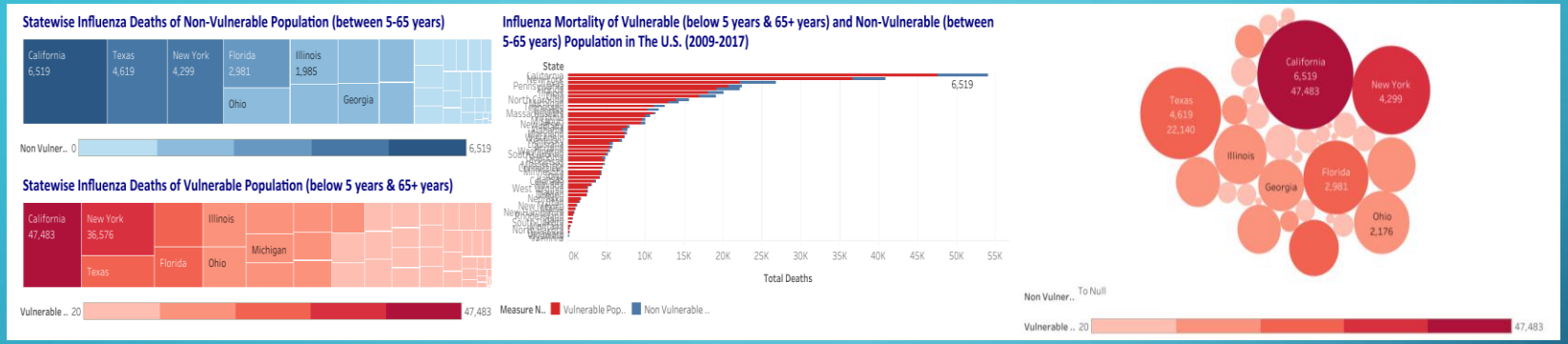
PROJECT GOAL

The primary goal is to analyze and compare influenza-related staffing needs by examining differences in death rates between vulnerable and non-vulnerable populations (those aged 5 to 65). By identifying patterns and trends in influenza death rates among these groups, the project seeks to inform resource distribution and support effective management strategies.

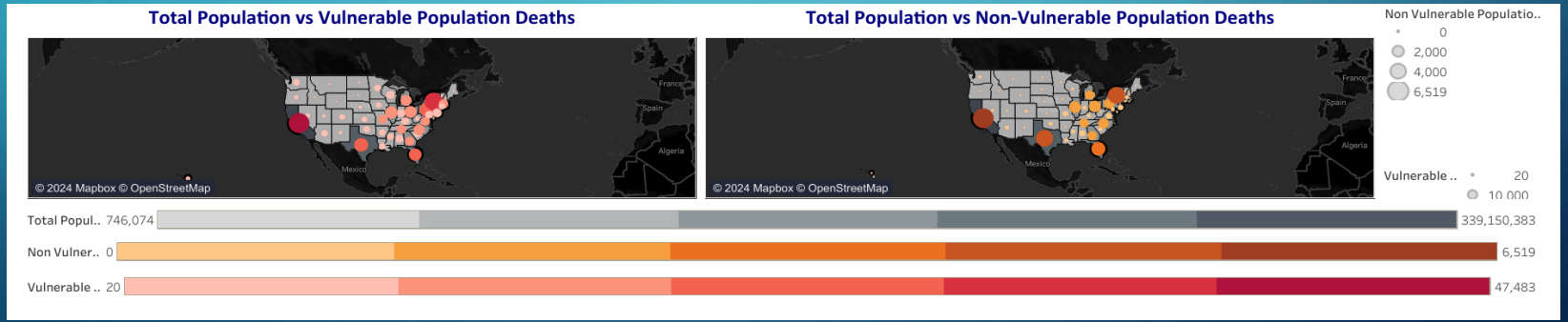
SKILLS DEMONSTRATED

- Statistical analysis for identifying death rate patterns across populations.
- Data cleaning and integration, working with CDC influenza death data and US Census Bureau population data (2009-2017).
- Data visualization to illustrate differences between vulnerable and non-vulnerable populations.
- Insightful interpretation to inform decisions on staffing needs and resource allocation during influenza season.

ANALYSIS AND VISUALIZATIONS

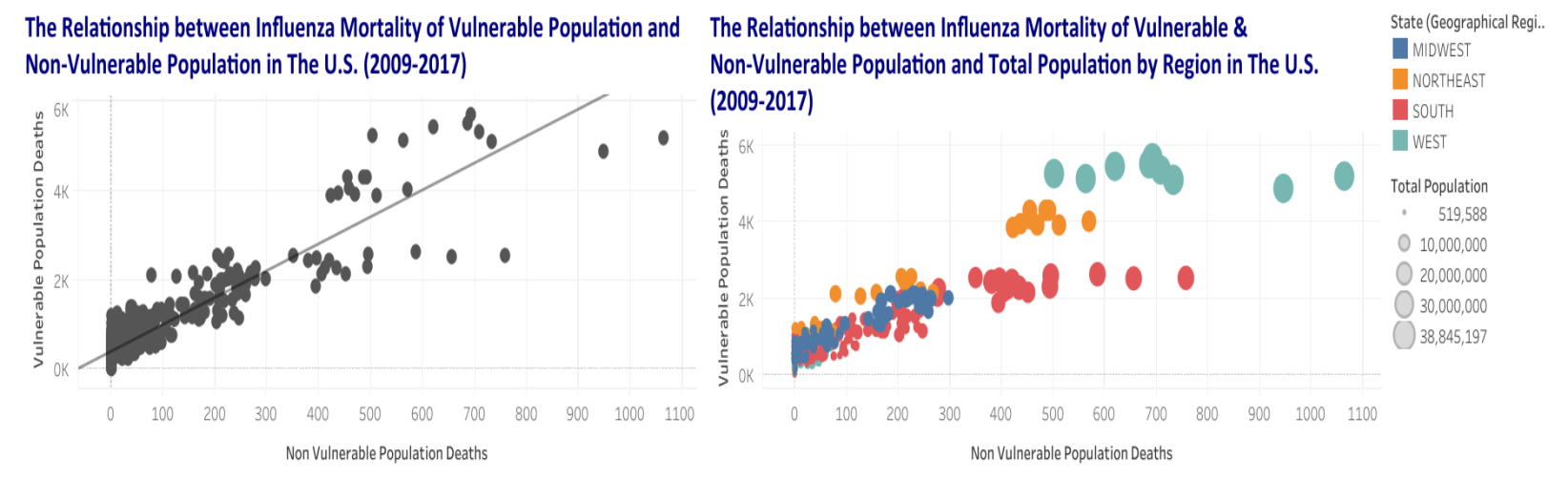


This visualization clearly shows which states have higher death counts in vulnerable populations, providing insight into regional disparities and potential areas for targeted interventions.



This visualization highlights both overall population distribution and specific impacts on vulnerable groups, pinpointing states with the greatest need for support.

ANALYSIS AND VISUALIZATIONS



A scatterplot with a trend line shows a strong positive correlation (R^2 = 0.84) between deaths in vulnerable and non-vulnerable populations, indicating that as deaths in vulnerable groups rise, deaths among non-vulnerable populations also increase significantly.

KEY FINDINGS AND RECOMMENDATIONS

States with high death counts significantly impact vulnerable groups, especially those under 5 and over 65, with a strong correlation (R^2 = 0.84) between deaths in vulnerable and non-vulnerable populations. Midwest and Southeast states are particularly affected.

Recommendations include prioritizing high-death states, focusing on vulnerable populations, deploying additional resources during peak seasons, and boosting preventive efforts. Continuous monitoring is advised to support regions with rising deaths among vulnerable groups.

TABLEAU LINK

GameCo

CONTEXT

GameCo, an emerging video game company, is focused on using data insights to drive the development of new games and to optimize its marketing investments.

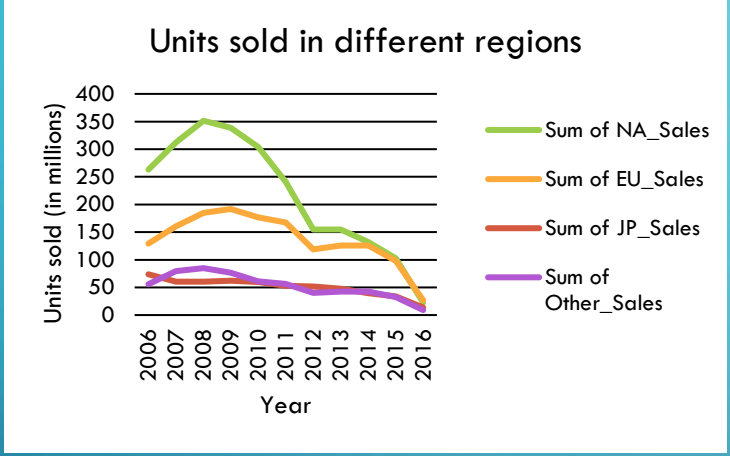
PROJECT GOAL

- Examine global sales trends to gain insights into market patterns.
- Identify popular game genres to focus on high-demand areas.
- Analyze leading publishers to better understand the competitive landscape.

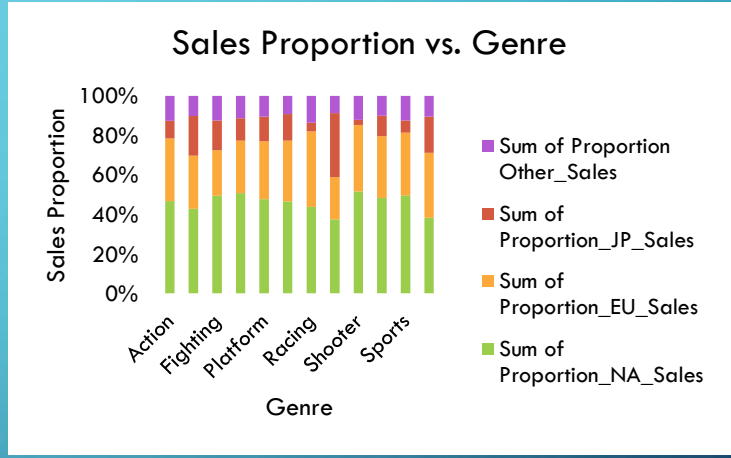
SKILLS DEMONSTRATED

- Organizing, sorting, and filtering data in Excel for streamlined analysis.
- Performing descriptive analysis to uncover trends in game sales.
- Cleaning data to ensure precision and consistency in findings.
- Creating visualizations to present insights in an engaging manner.
- Utilizing data storytelling techniques to inform strategic planning.

ANALYSIS AND VISUALIZATIONS

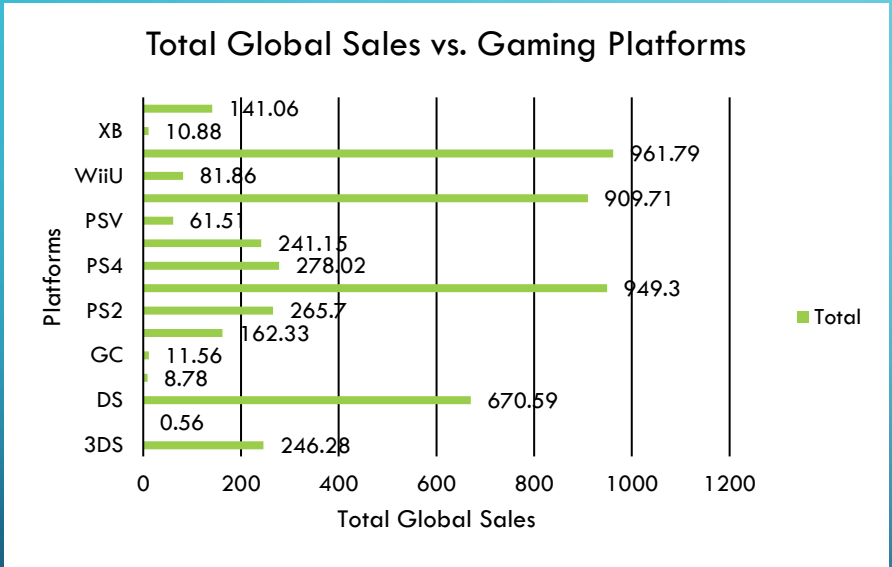


Over the past decade, **North American** sales peaked in 2006-2009, then declined post-2010. **Europe** grew until 2009, leveled off, and dropped in 2016. **Japan** saw steady declines, with larger drops after 2012. **Other regions** mirrored Europe, declining after 2010.



Over the past decade, **shooter games** saw the strongest sales in North America (52%) and Europe (33%), while **role-playing games** were most popular in Japan (32%). **Racing games** performed best in Europe (38%), and **adventure and strategy games** had balanced sales across regions.

ANALYSIS AND VISUALIZATIONS



Over the past decade, **PS3, X360, Wii**, and **DS** have consistently been top performers, achieving high cumulative global sales and dominating the gaming market. More recent data, however, suggests that newer consoles like **PS4, XOne**, and **3DS** are emerging with significant growth potential, indicating strong consumer demand and an upward trend in their market presence.

ACTIONS AND RECOMMENDATIONS

To optimize the 2017 marketing budget, North America should receive a larger share due to high sales, focusing on targeted ads and influencer partnerships. Europe also merits increased efforts with locally tailored campaigns.

In Japan, prioritize Role-Playing games and collaborate with local influencers and events. For emerging markets, allocate funds for localized marketing and research. Regular analytics monitoring will enable timely budget adjustments to maximize ROI.

GITHUB

Pig E Bank

CONTEXT

Pig E. Bank, an international financial institution, aims to strengthen its anti-money-laundering compliance efforts by better understanding client behavior.

PROJECT GOAL

The goal is to analyze client data to identify key factors that lead clients to leave the bank, providing insights to reduce attrition.

SKILLS DEMONSTRATED

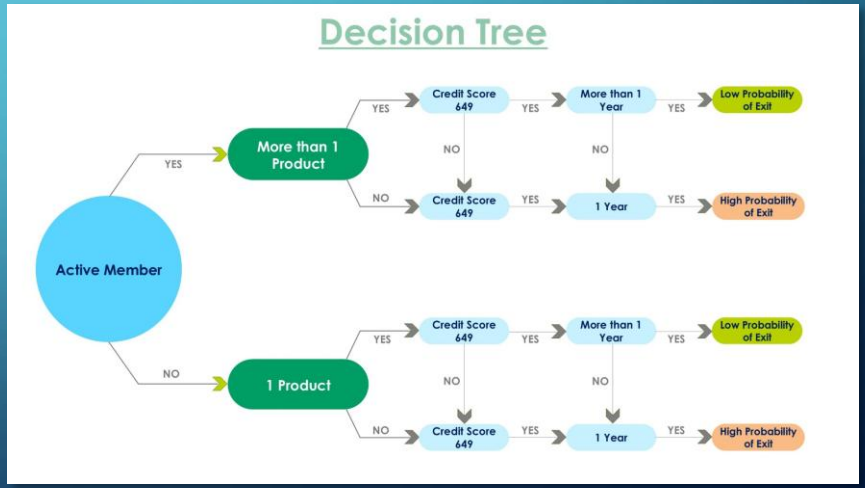
- Conducted exploratory data analysis
- Applied data mining techniques
- Built data models for analysis
- Performed predictive analysis to identify potential attrition factors

ANALYSIS AND VISUALIZATIONS

Column Name	All Clients			Exited Clients			Retained Clients		
	Min	Max	Mean	Min	Max	Mean	Min	Max	Mean
Credit Score	376	850	649	376	850	637	411	850	652
Age	18	82	39	22	69	45	18	82	38
Tenure	0	10	5	0	10	5	0	10	5
Balance	$0.00	$2,13,146.20	$78,002.72	$0.00	$2,13,146.20	$90,239.22	$0.00	$1,97,041.80	$74,830.87
Number of Products	1	4	2	1	4	1	1	3	2
Estimated Salary	$371.05	$1,99,725.39	$98,574.54	$417.41	$1,99,725.39	$97,155.20	$371.05	$1,99,661.50	$98,942.45

Descriptive statistics for All Clients, Exited Clients and Retained Clients.

A decision tree model identifying the key factors that most strongly predict a customer's likelihood of leaving the bank



KEY FINDINGS AND RECOMMENDATIONS

The analysis shows that clients with higher credit scores, longer tenures, and larger account balances have better retention, while those with Poor credit, shorter tenures, and $0 balances are more likely to exit. Female clients and those aged 46-55 also show higher churn rates.

To improve retention, the bank should offer tailored services like credit-building resources for lower-credit clients, premium offerings for high-credit clients, and personalized support for female clients and older age groups. Additionally, encouraging product diversification, active membership, and providing targeted engagement strategies could help reduce churn across various client segments.

THANK YOU

CONNECT WITH ME