Gavin Ellis

Data Analytics Portfolio

Projects & Tools

GameCo

Flu Season

Rockbuster

Instacart

Global Bank

NFL Weather Spread

GameCo

Project Overview

- **GameCo**: Game Co. is an international gaming company with a presence in major markets such as America, Europe, Japan, and other regions.
- **Objective**: Conduct a global descriptive analysis to understand how the development of new games will perform in the market.
- **Key Questions**:
  1. Are certain types of games more popular than others?
  2. What other publishers will likely be the main competitors in certain markets?
  3. Have any games decreased or increased in popularity over time?
  4. How have their sales figures varied between geographic regions over time?
- **Dataset**:

  This dataset includes historical sales of video games (for titles that sold over 10,000 copies) across various platforms, genres, and publishing studios. The data is sourced from the VGChartz website

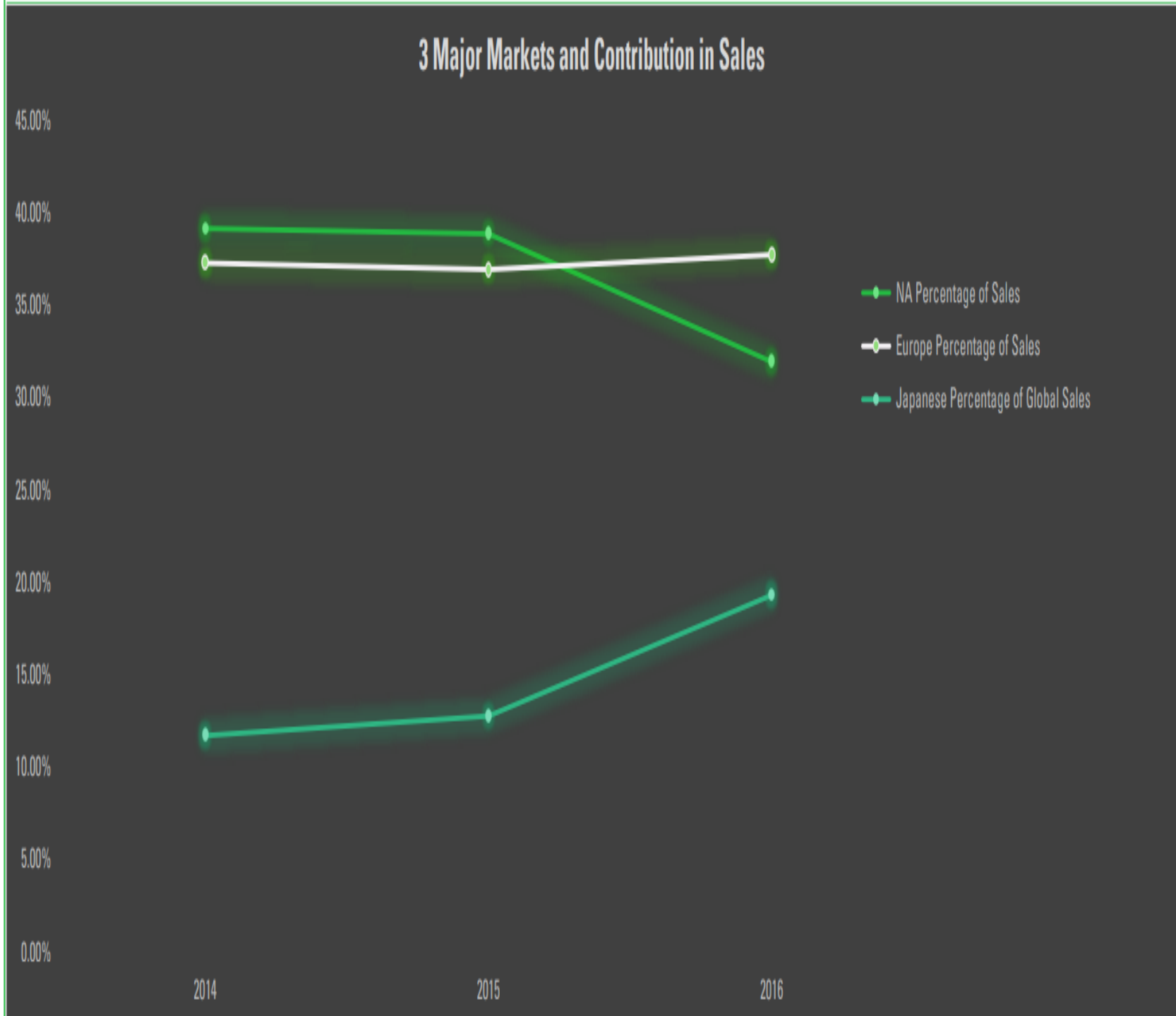
Skills Applied & Challenges

- **Skills Applied:**
  - Improving data quality
  - Data grouping and summarizing
  - Descriptive analysis
  - Pivot table
  - Visualization results in Excel
  - Presenting Results (link)
- **Challenges:**
  - The task involved comprehending business requirements and converting them into analytical insights.

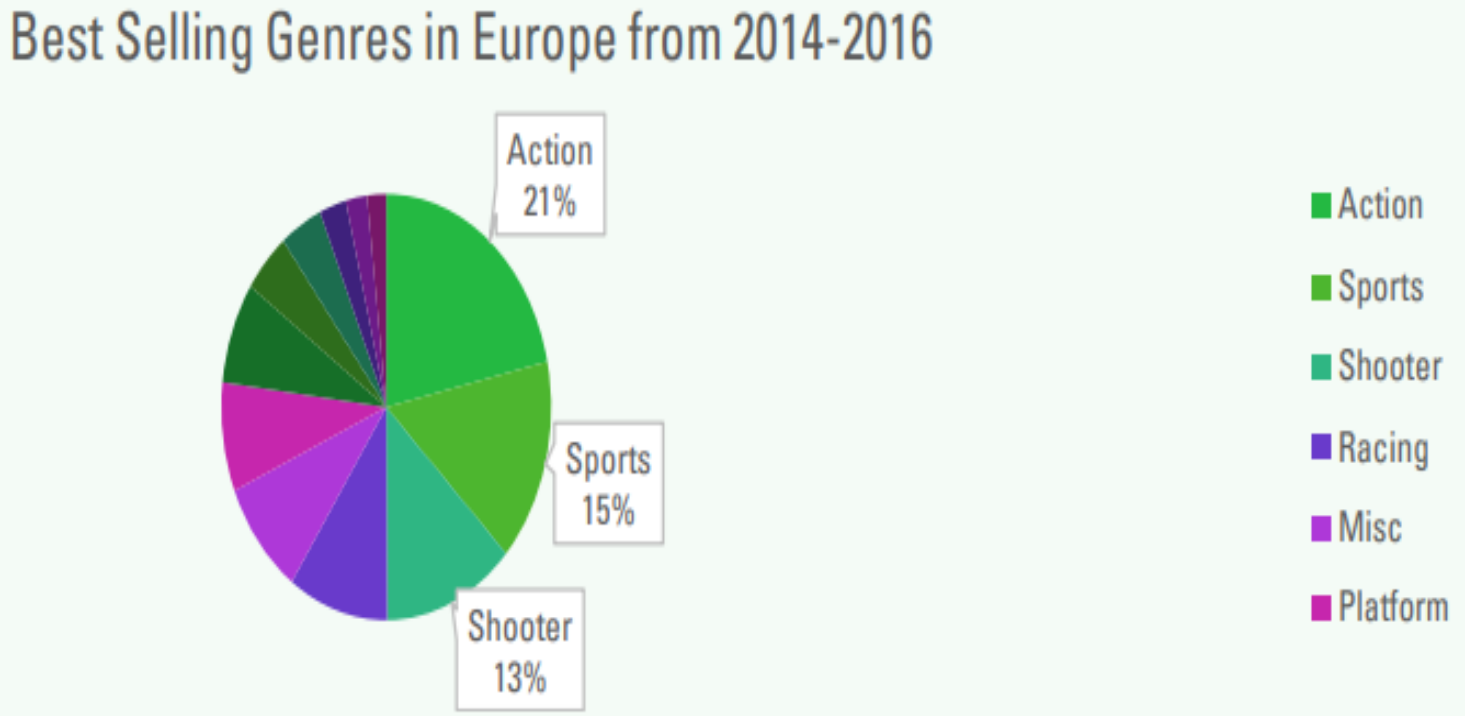
Analysis & Visualizations

- North America consistently leads in video game sales, especially noticeable from the mid-1990s onwards. This suggests a strong market presence and consumer base in the region.
- Japan, while initially strong, shows a relative decline in sales compared to North America and Europe. This could be due to various factors such as market saturation or shifts in consumer preferences.

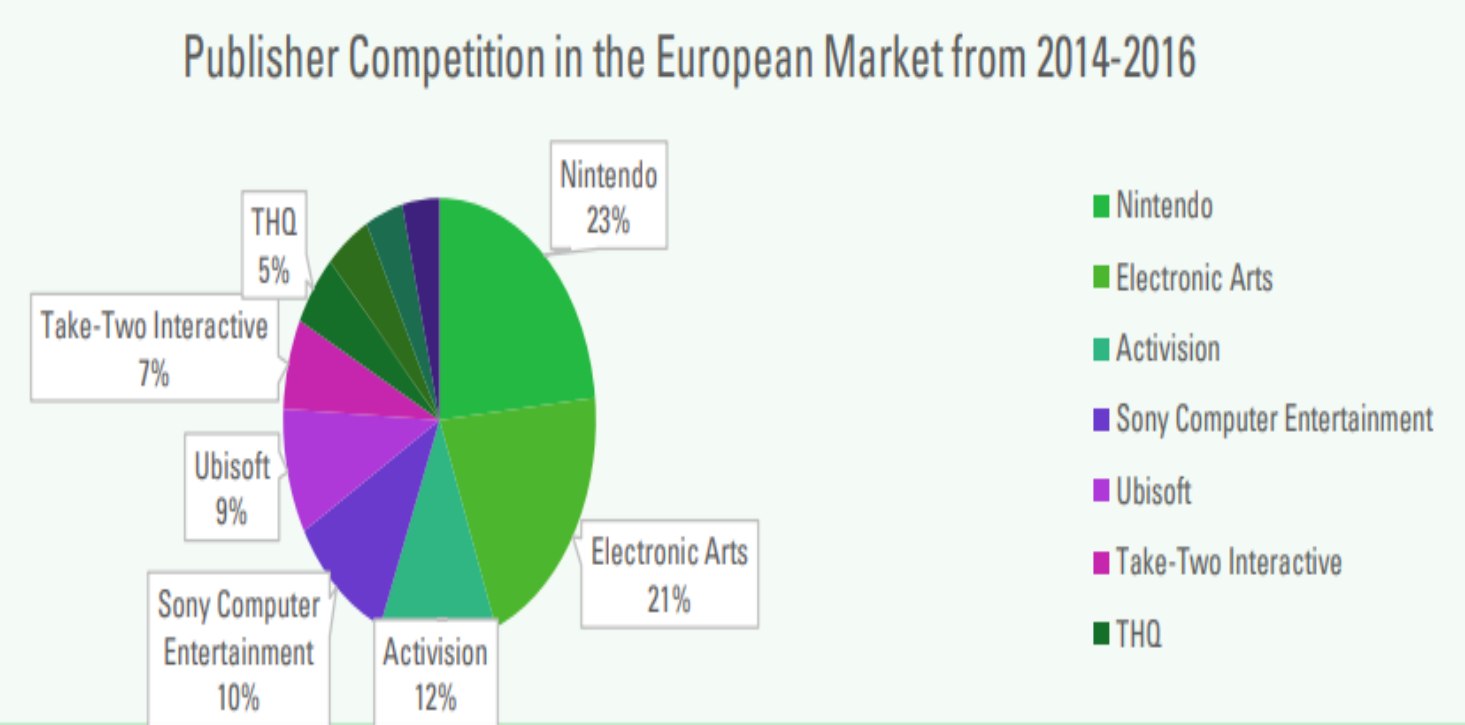
- Europe’s sales contribution starts at around 35%, dips to about 30% in 2015, and then returns to 35% in 2016. This suggests a relatively stable market with minor fluctuations.
- The graph highlights the dynamic nature of the video game market, with North America maintaining a leading position, Europe showing stability, and Japan experiencing significant growth.
- North America consistently contributes the highest percentage of sales, starting just above 40% in 2014, dipping slightly in 2015, and then rising to nearly 50% in 2016. This indicates a strong and growing market presence.



- Action games dominate the market with a 21% share, indicating their widespread appeal and popularity among European gamers.
- The chart highlights the dominant genres that publishers might focus on to capture the European market, with action, sports, and shooter games being the top choices.



- Nintendo holds the largest market share at 23%, indicating its strong presence and popularity in the European market.
- Electronic Arts follows closely with 21%, showing that it is a significant competitor to Nintendo.
- The distribution of market shares among these publishers suggests a competitive landscape where multiple companies are vying for consumer attention and sales.



Insights & Recommendations

- **Key Learnings:**

  1. From 1981 to 2008, video game popularity consistently grew each year across North America, Europe, and Japan. However, since 2008, there has been a decline. Interestingly, between 2014 and 2016, Europe performed as well as or better than the U.S. in terms of video game popularity.

  2. From 2014-2016 in the Europe market action, sports, and shooting games are the highest selling genres.

  3. Main competition from 2014-2016 in the Europe market are Nintendo, Electronic Arts, and Activision

- **Recommendations:**
  - GameCo should focus its efforts on developing games for the European market, according to the presented trends.
  - Increase focus on action, sports, and shooting games to compete with other publishers in the European market

- **Links & Deliverables:**
  - Final Presentation

Flu Season

Project Overview

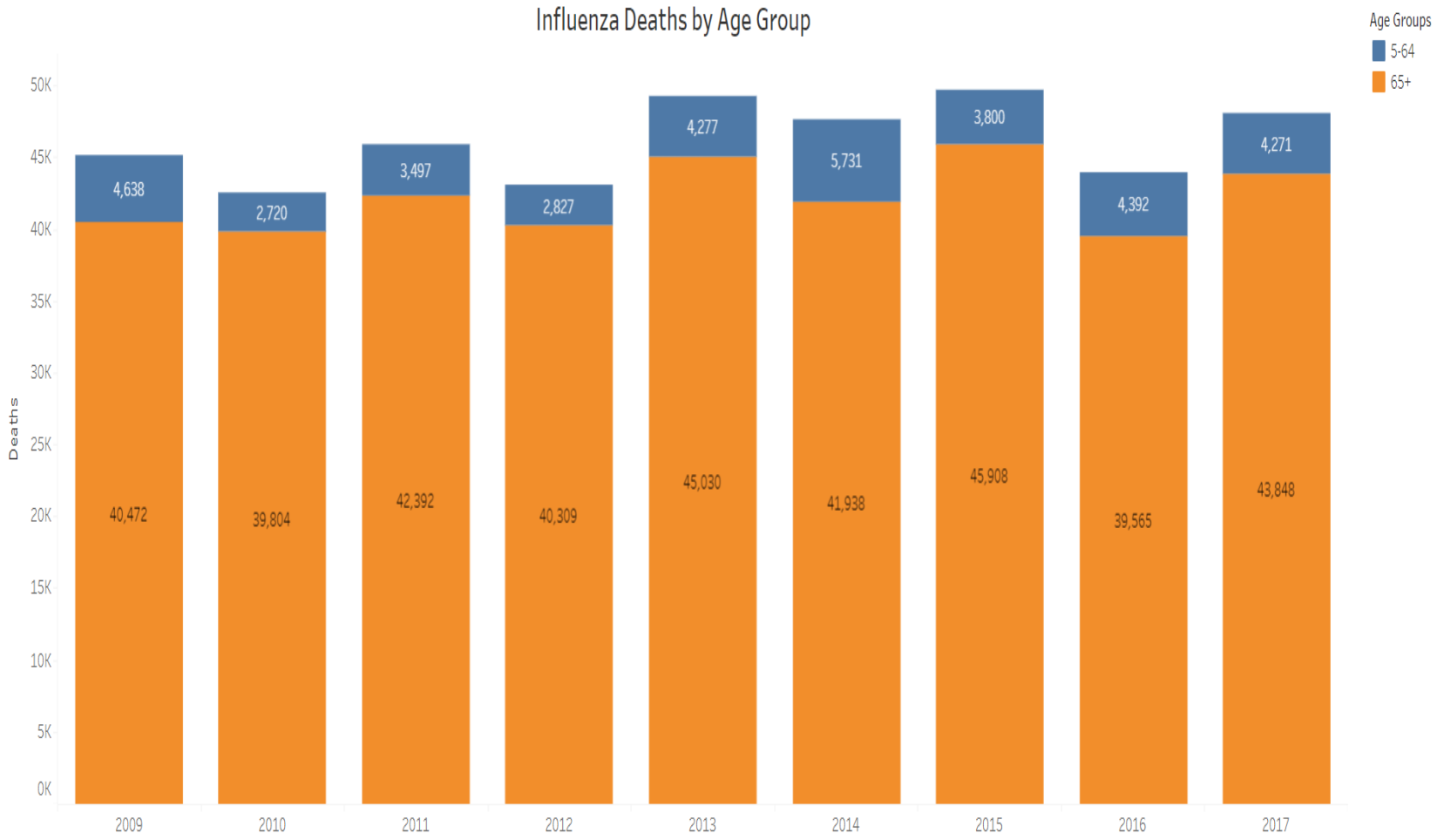
- **Motivation:** During flu season, healthcare facilities in the U.S. frequently experience a substantial rise in patient numbers.
- **Objective:** Determine when to send staff, and how many, to each state.
- **Requirements:**
  - Back up the staffing plan with data on the distribution of medical personnel across the U.S.
  - Examine the seasonal patterns of influenza across different states.
  - Rank states according to the size of their vulnerable populations and classify them as low, medium, or high-need.
  - Pinpoint data limitations that obstruct analysis.
- **Datasets:**
  - Census Population: U.S. population information from 2009-2017 by year and county.
    - Source: US Census Bureau
  - Influenza Deaths: U.S. influenza deaths by age group, state, and year from 2009-2017
    - Source: CDC

Skills Applied & Challenges

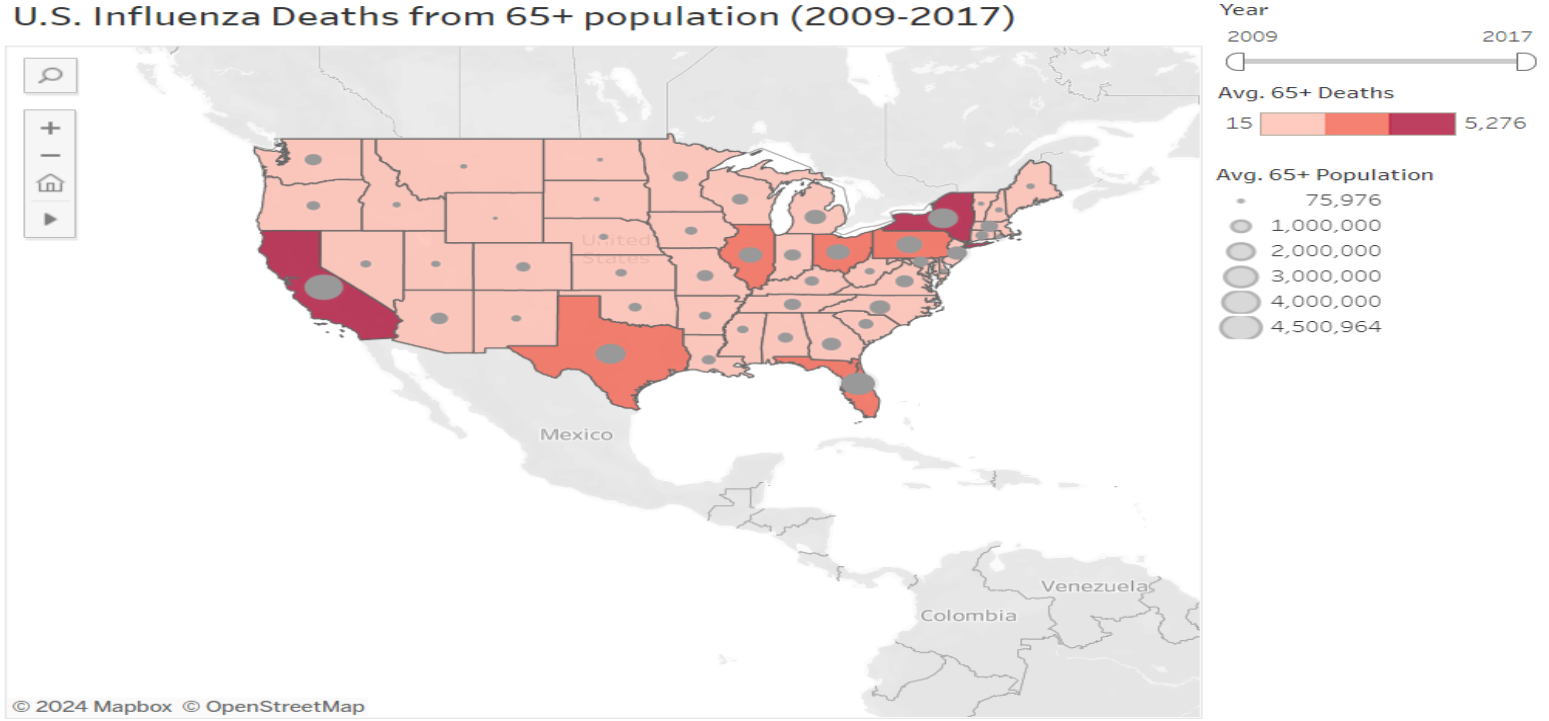
- **Skills Applied:**
  - Converting business requirements into analytical questions
  - Identifying and sourcing relevant datasets
  - Integrating and cleaning data
  - Conducting statistical hypothesis testing
  - Performing visual analysis using Tableau
  - Creating forecasts
  - Crafting narratives with Tableau
  - Presenting results to an audience
- **Challenges:**
  - A major challenge for this project was successfully merging various data sources to achieve a unified analysis.

Analysis & Visualizations

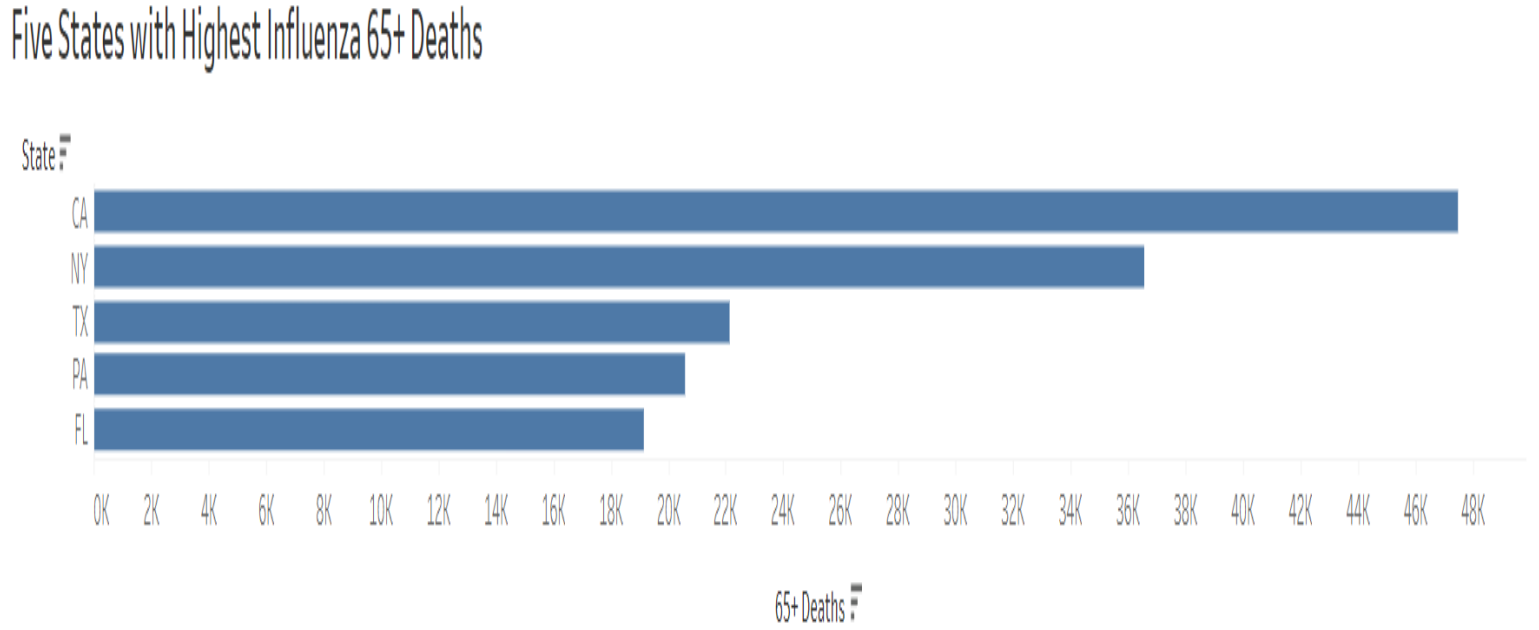
- **Research Hypothesis:**
  - States with higher vulnerable populations (65+) are more likely to die from flu than states with less vulnerable populations
- While the number of deaths fluctuates each year, the overall trend indicates that the 65+ age group remains at a higher risk.
- The data underscores the importance of targeted public health interventions, such as vaccination campaigns and preventive measures, especially for the elderly population.



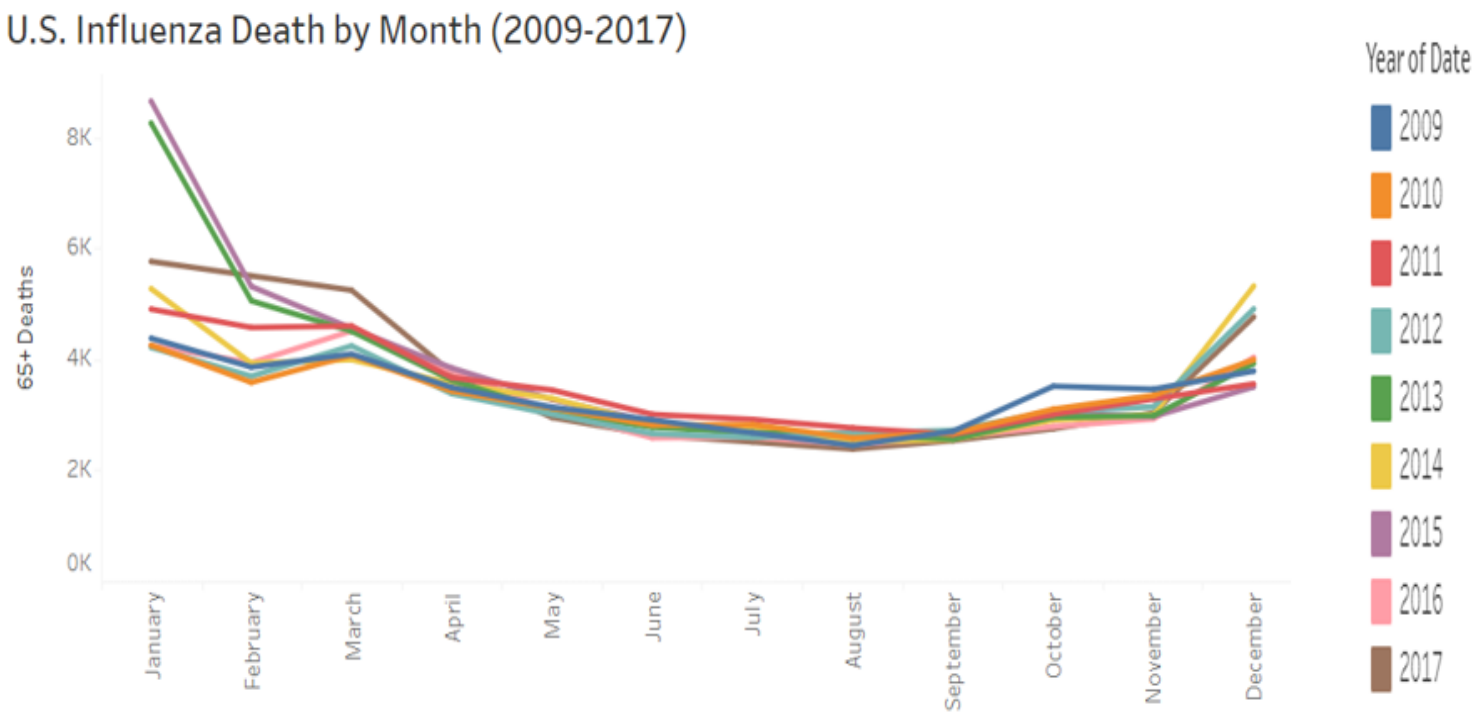
- States like California, Texas, and Florida show darker shades, indicating higher numbers of influenza-related deaths among the elderly.
- States with larger elderly populations, such as California and Florida, also show higher numbers of deaths, suggesting a correlation between population size and influenza mortality.



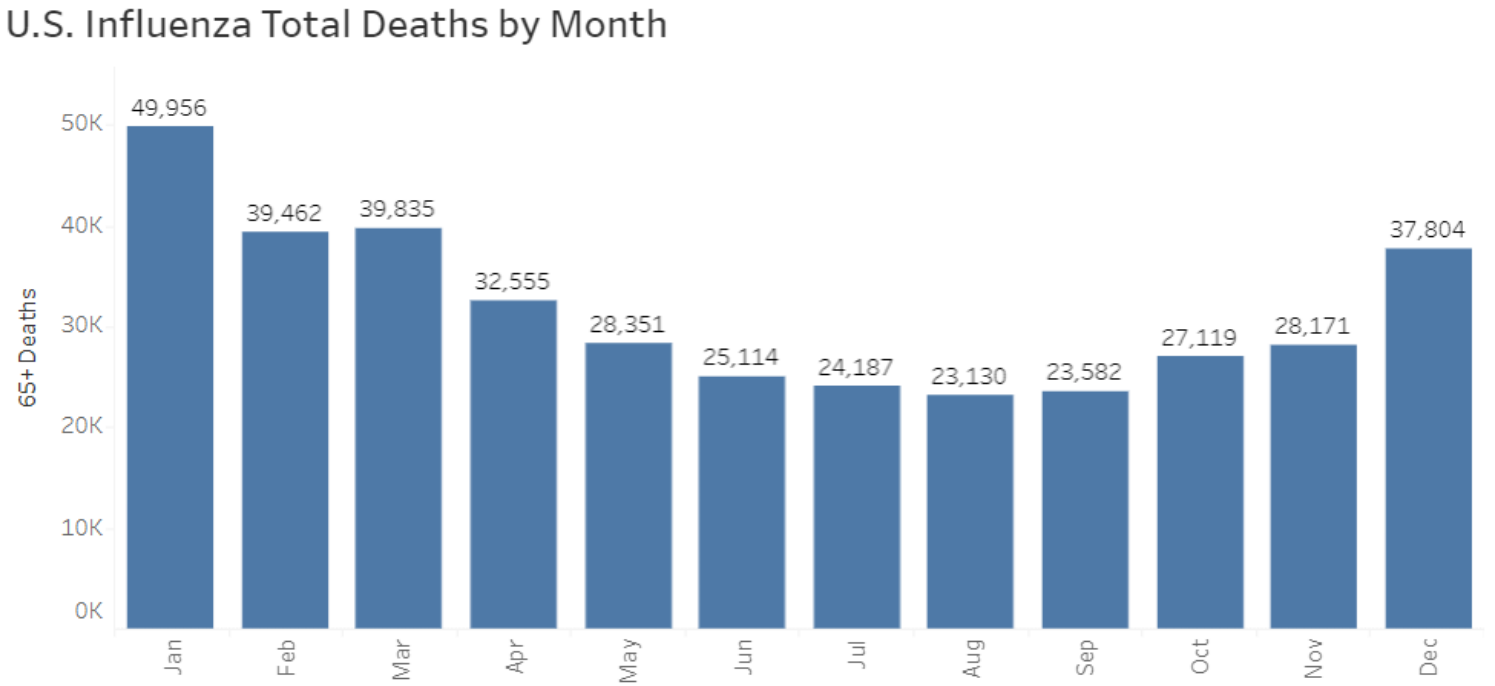
- California (CA) has the highest number of influenza-related deaths among the elderly, indicating a significant impact of influenza in this state.
- New York (NY) and Texas (TX) also show high numbers of deaths, suggesting that these states have substantial elderly populations affected by influenza



- The graph shows clear peaks in influenza deaths during the winter months (December to February) for each year.
- While the overall pattern of winter peaks is consistent, the severity of these peaks varies from year to year. For example, some years like 2014 and 2017 show higher peaks compared to others, suggesting more severe flu seasons.



- The highest number of deaths occurs in January, with approximately 49,962 deaths. This highlights the severe impact of influenza during the winter months.
- The lowest number of deaths is in July, with around 2,147 deaths. This indicates a significant drop in influenza-related mortality during the summer.



Insights & Recommendations

- **Key Learnings:**
  - States with a higher population of individuals aged 65 and older see more influenza-related deaths.
  - Influenza deaths usually rise in December, peak in January, and stay elevated through February and March. This pattern is consistently seen across various states and regions each year.
  - The virus's effect on vulnerable age groups underscores the necessity for targeted interventions and preventative measures.
- **Recommendations:**
  - New York and California are high priority states to send additional staffing
  - To prepare for the influenza season, which lasts from December to March, medical staff should be allocated to each state according to priority.
- **Links & Deliverables:**
  - Tableau Storyboard
  - Interim Report
  - Project Management Plan


HORROR
RENT THEM FOR $1.25 ON THURSDAY
JOY RIDE
the films of GEORGE A. ROMERO
Wes Craven
DARIO ARGENTO
GEORGE ROMERO
WES CRAVEN
Mexican Horror

Rockbuster

Project Overview

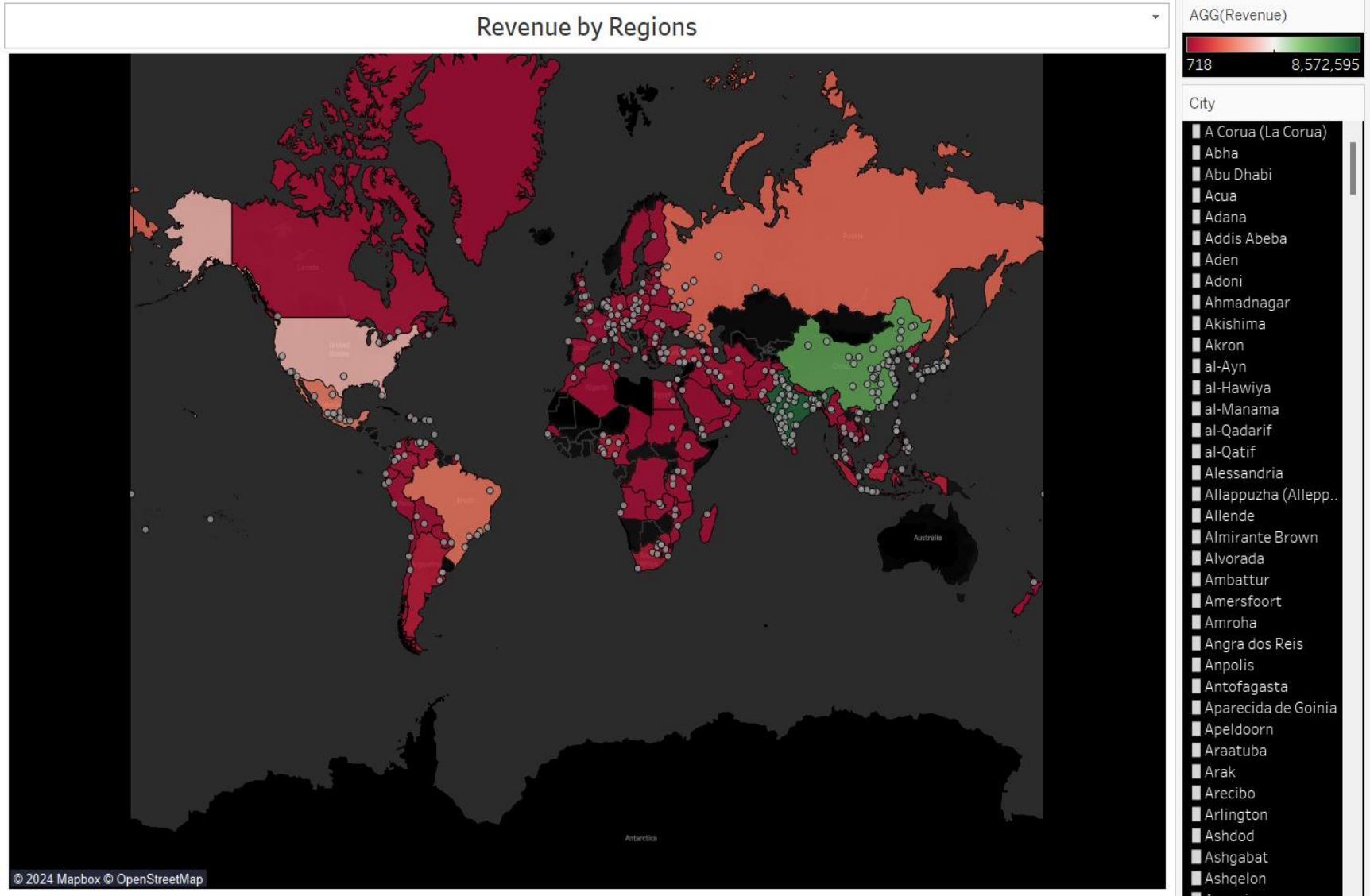
- **Rockbuster:** Rockbuster Stealth LLC, a global movie rental company, once operated stores worldwide. Facing stiff competition from Netflix and Amazon Prime Video they open up an online video service to compete with them.
- **Objective:** Provide business strategy that will help increase revenue for all stores.
- **Key questions:**
  - Which movies contributed the most/least to revenue gain?
  - What was the average rental duration for all videos?
  - Which countries are Rockbuster customers based in?
  - Where are customers with a high lifetime value based?
  - Do sales figures vary between geographic regions?
- Dataset
  - PostgreSQL Tutorial Data Set

Skills Applied & Challenges

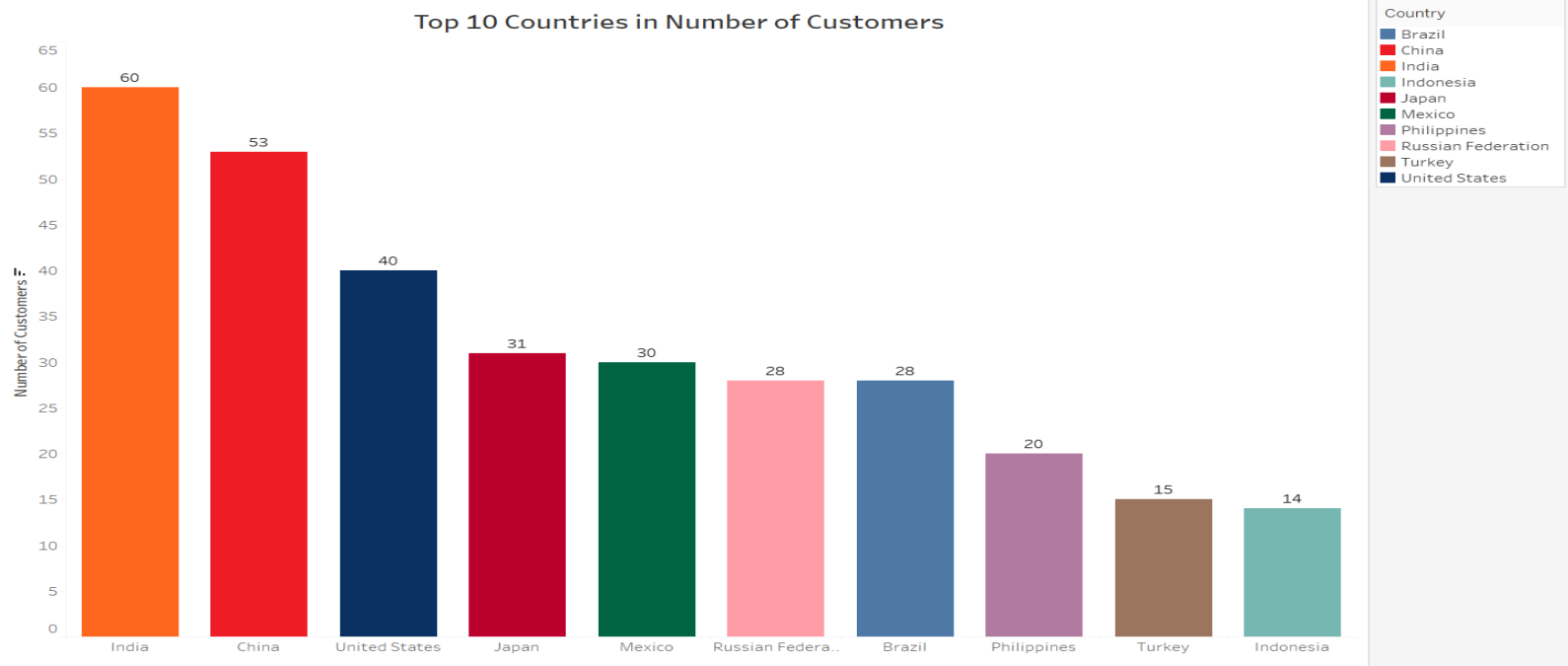
- **Skills Applied:**
  - Relational databases
  - SQL
  - Creating a data dictionary
  - Database querying
  - Data filtering
  - Data cleaning and summarizing
  - Joining tables
  - Subqueries
  - Common table expressions
  - Presentation
- **Challenges:**
  - Handling complex queries and ensuring data accuracy.

Analysis & Visualizations

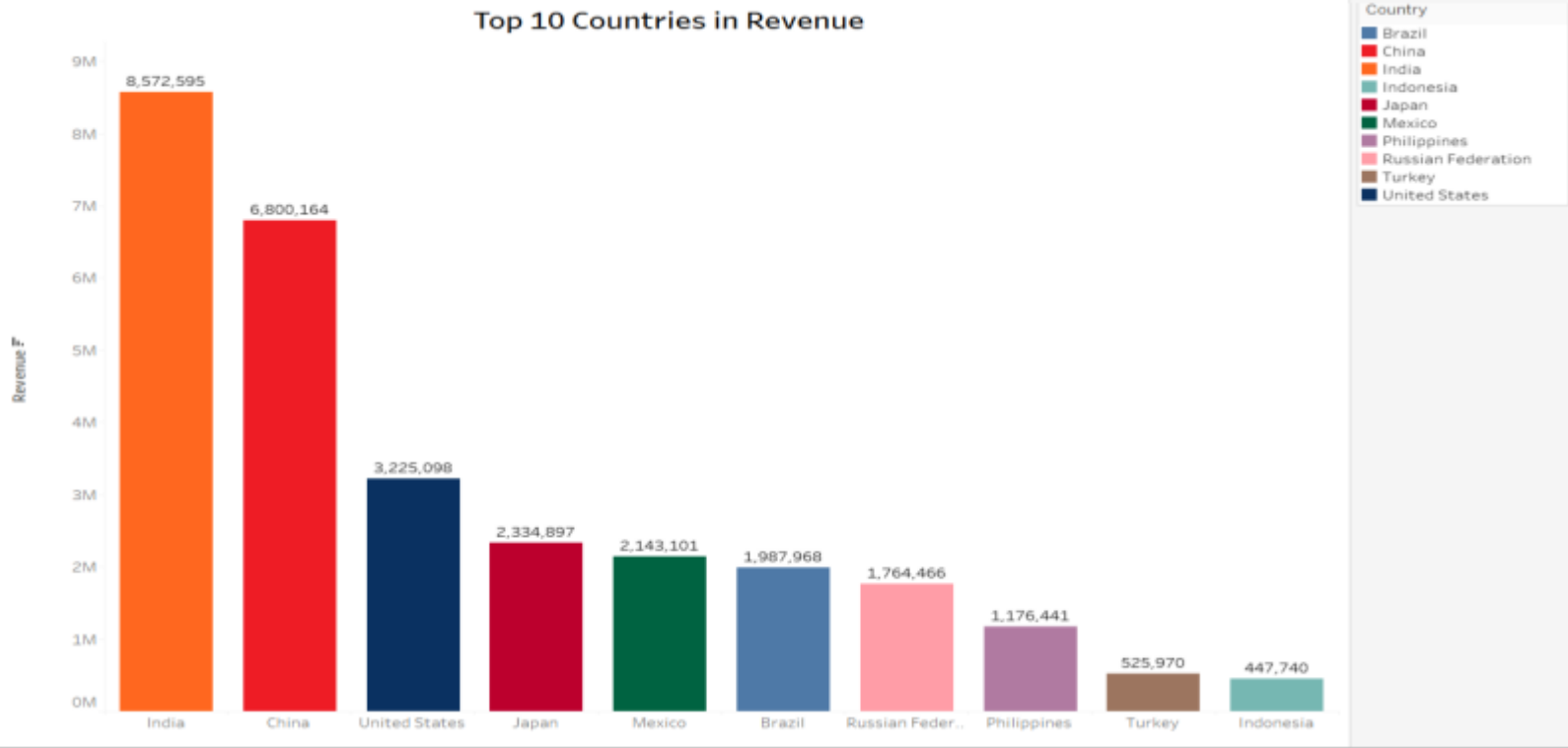
- Based on the map we can identify that majority of sales are coming from the eastern hemisphere.
- China and India are the top revenue generators, reflecting significant customer engagement.
- The Australian market remains untapped and could be vital for boosting Rockbuster's revenue.



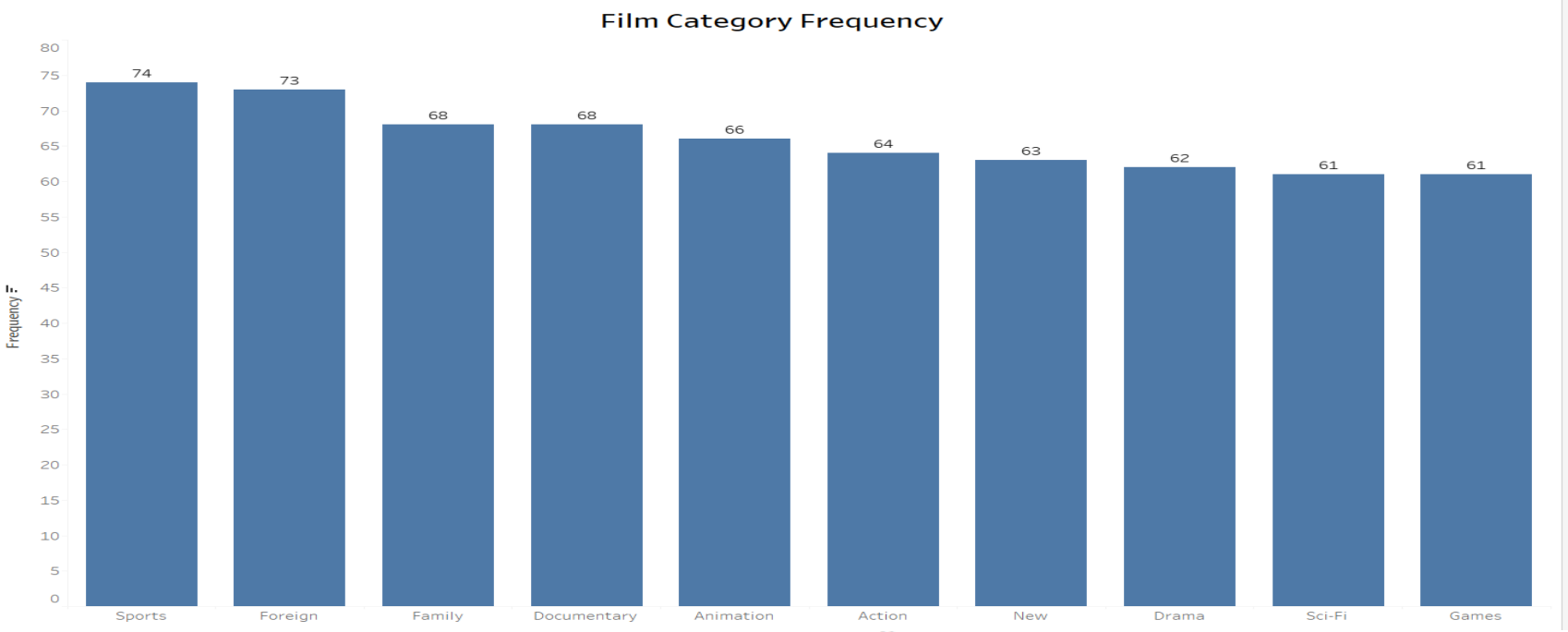
- Three of the top five countries with the highest number of customers are in Asia.
- India has almost double the number of customers compared to the U.S.



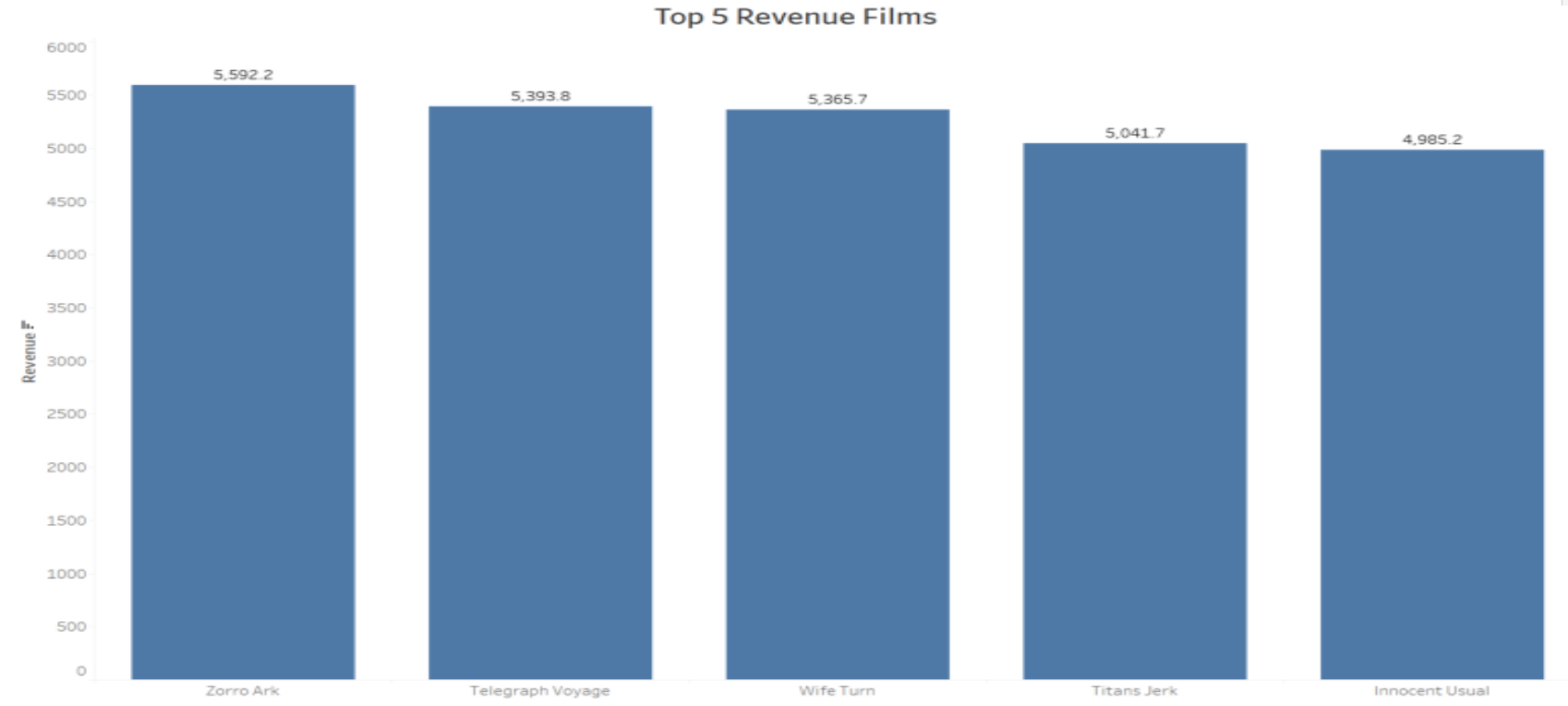
- India and China generate two to three times more revenue than all U.S. stores combined.
- Three of the top five countries in revenue reside in Asia



- Sports and Foreign films are the most frequent, with percentages of 74% and 73% respectively. This indicates a high prevalence of these genres in the dataset.
- Sci-Fi films have the lowest frequency at 61%, suggesting they are less common compared to other genres.



- "Zorro Ark" leads with the highest revenue of approximately $5,922 million, indicating its significant financial success.
- "Titans Jerk" and "Innocent Teasal" have slightly lower revenues, at approximately $5,042 million and $4,985 million respectively. While not as high as the top three, these figures still represent substantial earnings.



Insights & Recommendations

- **Key Learnings:**
  - Zorro Ark was the highest grossing film
  - The average rental duration for all videos is 5 days, reflecting steady consumer engagement.
  - Sports, foreign, and family genres are the most popular.
  - India, China, and the United States dominate in both customer numbers and revenue share, significantly influencing the market.
  - The three highest customers in revenue reside in Asia
- **Recommendations:**
  - Encourage discounts for films so customers rent movies for longer periods to retain activity and loyalty
  - Focus on a market campaign for Australia since it has larger population, which could lead to increase in revenue.
  - Limit the least popular film categories to make room for the genres that are top selling in number of purchases.
- **Project Deliverables & Links:** GitHub Repository, Final Presentation, Data Dictionary

Instacart

Project Overview

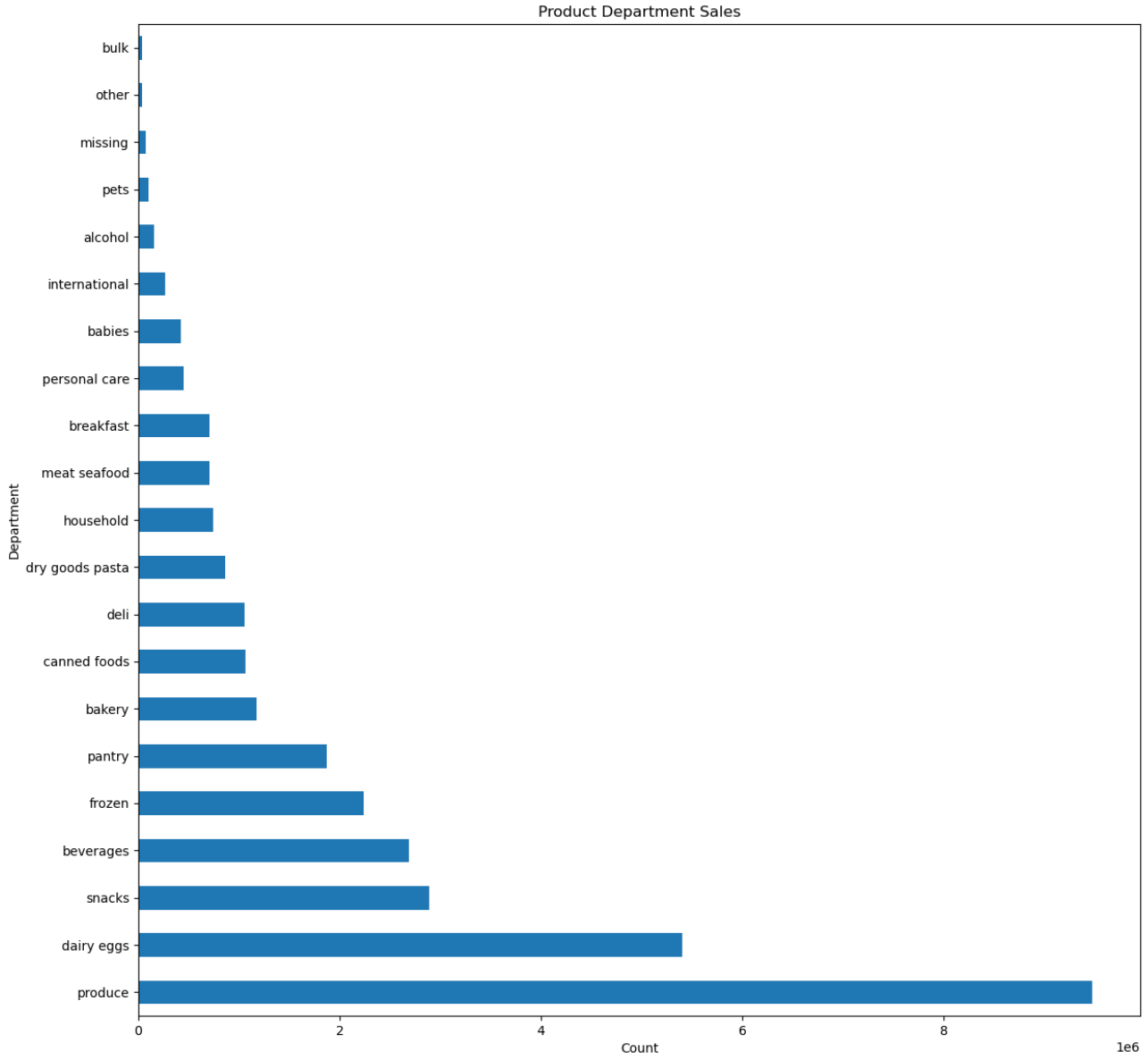
- **Instacart:** Instacart, a top online grocery store, allows customers to order groceries via an app. To boost sales, the company aimed to adopt a more strategic approach to customer targeting and improve segmentation by analyzing historical data.
- **Objective:** Conduct an initial data and exploratory analysis to derive insights and recommend strategies for improved segmentation based on the given criteria.
- **Key Questions:**
  - What are the busiest days of the week and hours of the day?
  - At what times of the day do people tend to spend the most money?
  - How can simple price range groupings be used to optimize marketing and sales efforts?
  - Which types of products are most popular?
  - What are the characteristics and spending habits of different customer profiles
- Datasets & Citations:
  - Datasets: Customers, Orders, Products, Departments
  - Citations:
    - "The Instacart Online Grocery Shopping Dataset 2017", Accessed via Kaggle.
    - "Customers Data Set", Provided by CareerFoundry.

Skills Applied & Challenges

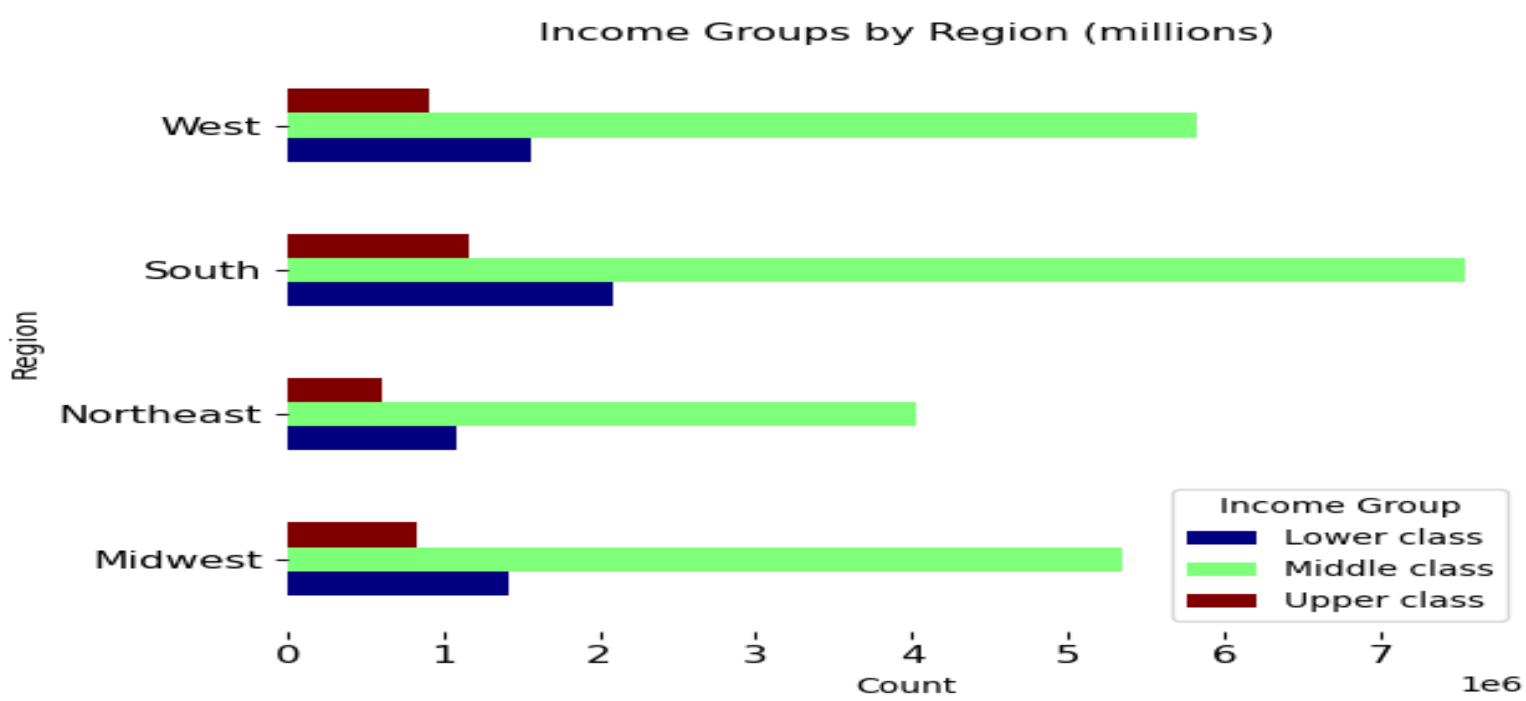
- **Skills Applied:**
  - Python Programming:
    - Conducted all coding and analysis tasks within Jupyter Notebook.
    - Performed data wrangling, subsetting, filtering, and summarizing using Pandas.
    - Merged datasets and conducted consistency checks to ensure data accuracy.
    - Derived new variables and executed complex data transformations.
    - Grouped and aggregated data to extract detailed insights.
  - Data Visualization:
    - Developed advanced visualizations with Matplotlib and Seaborn.
    - Presented analytical results through clear and impactful charts.
- **Challenges:**
  - Managing the dataset's 30 million records presented significant challenges in cleaning, standardizing, and formatting for analysis.
  - Memory shortages were addressed through efficient coding practices, ensuring meaningful data transformation.

Analysis & Visualizations

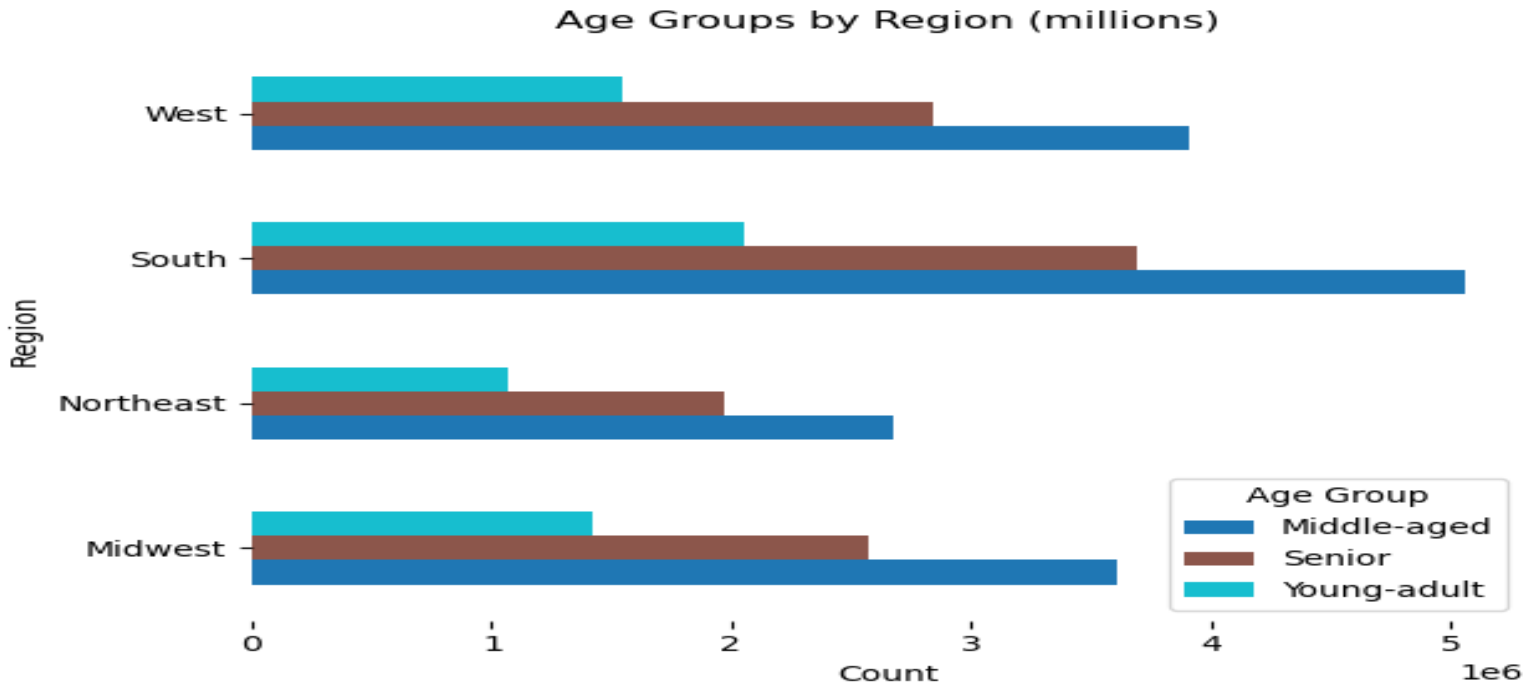
- It’s evident that certain product categories sell better than others in millions.
- The top 10 categories are produce, dairy and eggs, snacks, beverages, frozen foods, pantry staples, baked goods, canned items, deli products, and dry goods and pasta.
- Notably, produce products dominate sales, outperforming all other departments.



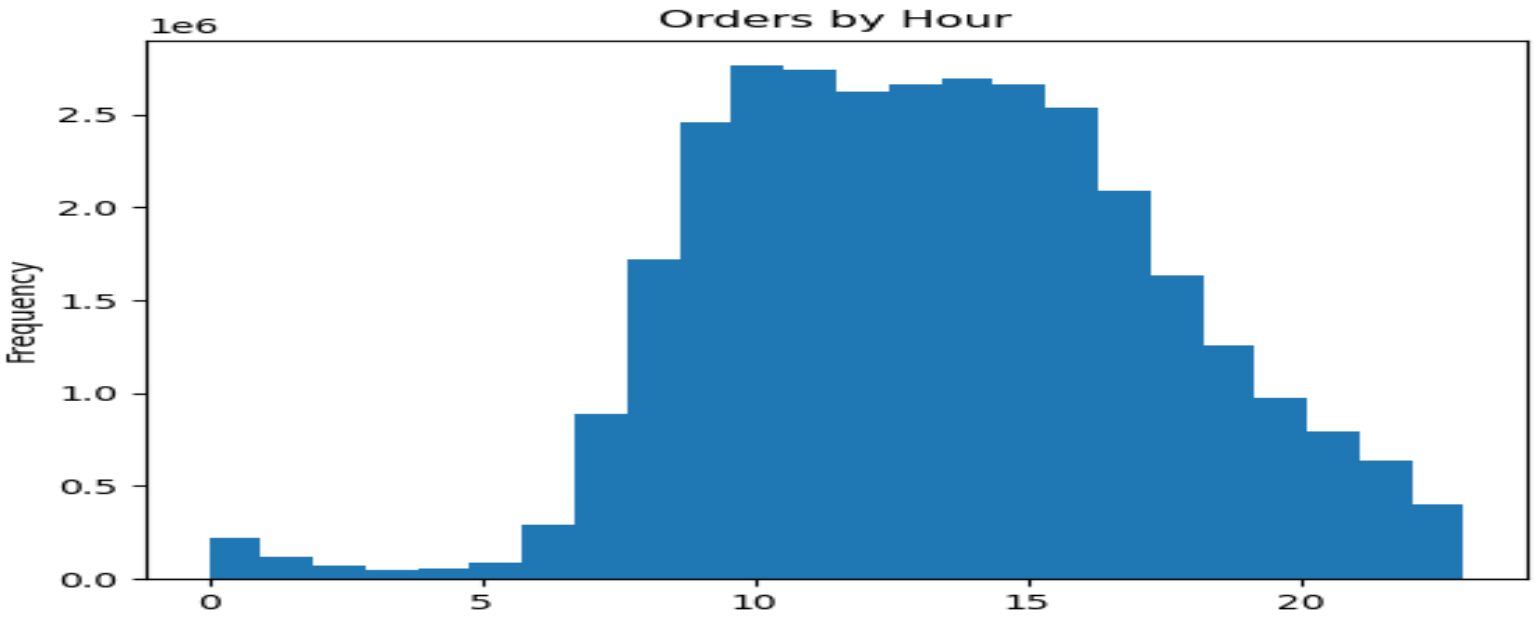
- In all regions, the majority of customers belong to the middle-income group, with the South region having the highest proportion and the Northeast the lowest.
- The high-income and low-income groups are relatively similar in size across all regions.



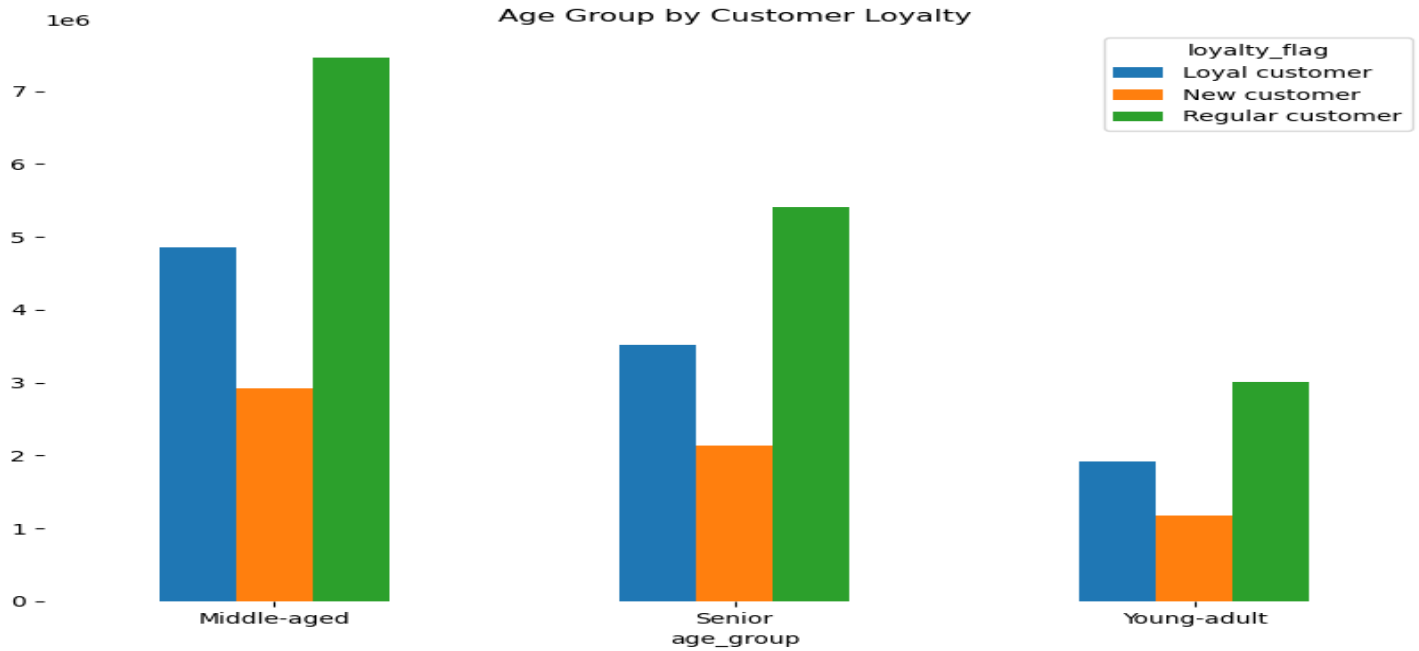
- Across all regions, middle-aged adults are the most frequent Instacart customers, followed by seniors and then young adults.
- The South region has the highest number of middle-aged customers, while the Northeast has the fewest.



- Throughout the day, the highest volume of orders occurs between 10 am and 3 pm, while the lowest volume is seen in the early morning hours from midnight to 5 am.
- After 5 am, order activity gradually increases, peaking from 10 am to 3 pm, and then starts to decline after 4 pm.



- The middle-age group also holds the highest number of loyalty flag customers with regular being first and loyal customers being second highest.



Insights & Recommendations

- **Key Learnings:**
  - Saturdays and Sundays are the busiest days of the week, with the highest order volumes occurring between 10 AM and 3 PM
  - People tend to order more expensive items before 6-7 AM.
  - Most of the products people order are mid-range (5-10 USD) products.
  - The most popular food products are produce, dairy eggs, and snacks.
  - The most notable difference in shopping behavior among customer groups is related to income with the middle-class contributing to 70% of sales.
- **Recommendations:**
  - Given that some customers have very high incomes, consider designing premium product lines and luxury items to attract more high-income customers to Instacart.
  - Schedule advertisements during off hours and the least busiest days of the week. Therefore, ads are recommended to run during 12am-5am, while increasing ads on Tuesday and Wednesday.
  - Consider adding incentives to poorly performing departments to possible entice purchases.
- **Project Deliverables & Links:** Github Repository, Project Report

Global Bank

Project Overview

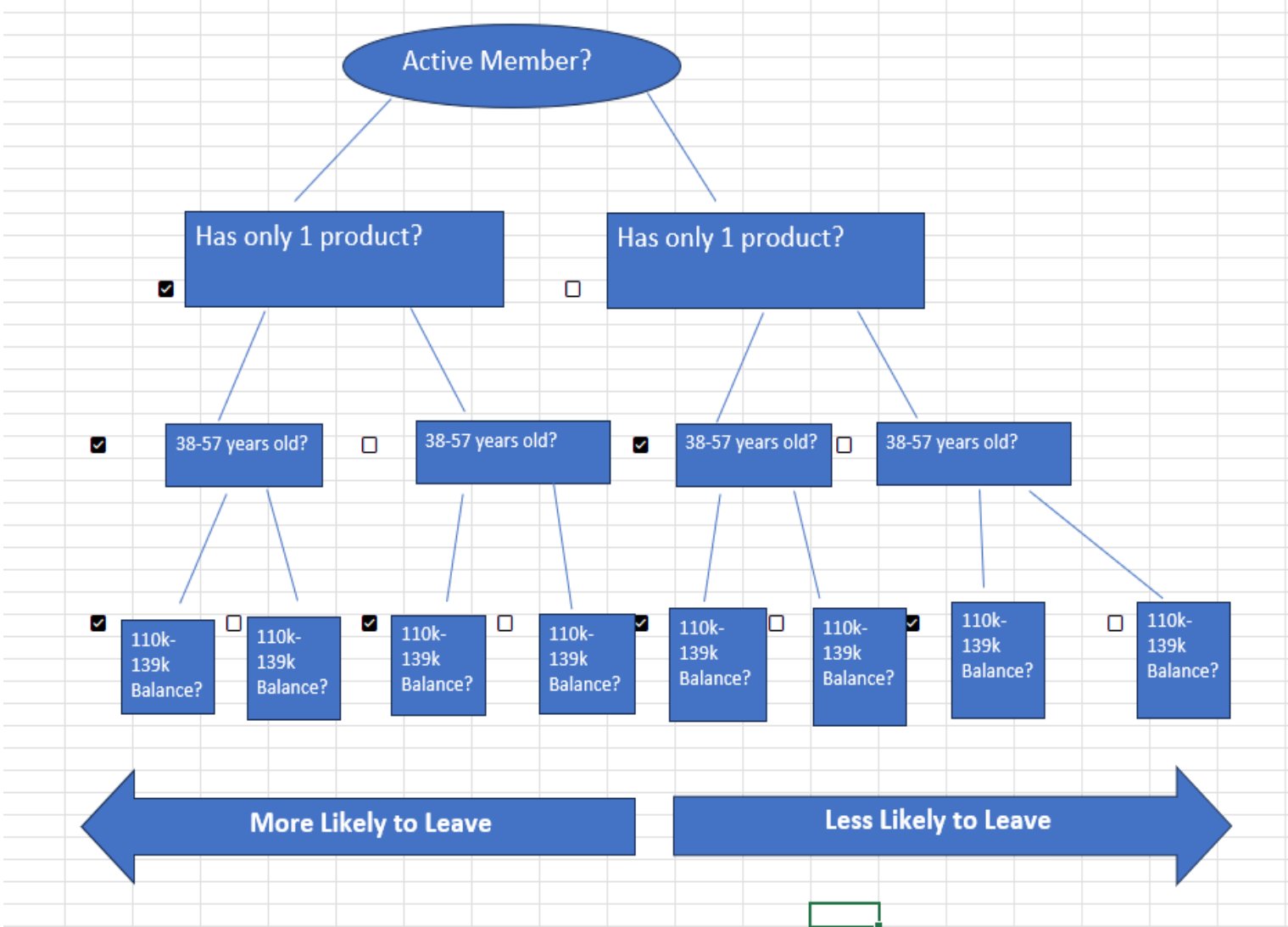
- **Global Bank:** Pig E. Bank is a global institution committed to delivering outstanding financial services.
- **Objective:** The goal of this project was to conduct a comprehensive analysis of customer satisfaction data at Pig E. Bank. The focus was on identifying factors contributing to customer attrition, with the ultimate aim of developing effective strategies to enhance customer retention
- **Key Question:**
  - What are the key risk-factors in identifying customers who are most likely to churn?
- **Data source:** Career Foundry

Skills Applied

- **Skills Applied:**
  - Big data
  - Data ethics
  - Data mining
  - Predictive analysis
  - Time series analysis and forecasting

Analysis & Visualizations

- **Client Segmentation:** Categorized the data into two groups for analysis: former clients and current clients.
- **Trend Analysis:** Thoroughly analyzed the former clients' data to identify common behavioral patterns.
- **Model Development:** Developed a decision tree to systematically understand and predict the factors contributing to customer churn.



- Majority of customers are aged between 38-57 years
- Majority of customers reside in France. However, the impact French customers of leaving is not very much with the rest of the regions.
- Majority of members of have a remaining balance between $110,000-$140,000.

Exited From Bank Table

Row Labels	Count of ExitedFromBank?
0	79.41%
1	20.59%
Grand Total	100.00%

Credit Score Table (ExitedFromBank 1)

Credit Sc	Count of Credit Score
Poor	28.92%
Fair	32.35%
Good	22.55%
Very Good	11.27%
Exceptiona	4.90%
Grand Total	100.00%

Country Table (ExitedFromBank (All))

Estimated S	Count of Country
France	48.44%
Germany	25.93%
Spain	25.63%
Grand Total	100.00%

Tenure Table (ExitedFromBank (All))

Tenure	Count of Tenure
0	3.83%
1	10.39%
2	11.10%
3	10.19%
4	8.68%
5	10.09%
6	9.49%
7	9.38%
8	10.90%
9	10.90%
10	5.05%
Grand Total	100.00%

Gender Table (ExitedFromBank? (All))

Estimated Sal	Count of Gender
Female	46.62%
Male	53.28%
Unknown	0.10%
Grand Total	100.00%

Age Table (ExitedFromBank 1)

Age	Count of Age
18-27	1.47%
28-37	17.16%
38-47	43.14%
48-57	24.51%
58-67	13.24%
68+	0.49%
Grand Total	100.00%

Active Member Table (ExitedFromBank (All))

ActiveMemb	Count of IsActiveMember
0	49.24%
1	50.76%
Grand Total	100.00%

Number of Products Table

Number of P	Count of NumOfProducts
1	51.46%
2	45.01%
3	3.33%
4	0.20%
Grand Total	100.00%

Balance Table (ExitedFromBank 0)

Balance	Count of Balance
0	37.23%
30,000-39,9	0.25%
40,000-49,9	0.51%
50,000-59,9	1.40%
60,000-69,9	1.91%
70,000-79,9	3.30%
80,000-89,9	4.32%
90,000-99,9	4.83%
100,000-109,	6.99%
110,000-119,9	7.50%
120,000-129,	9.02%
130,000-139,	7.37%
140,000-149,	4.96%
150,000-159,	3.81%
160,000-169,	2.92%
170,000-179,	2.16%
180,000-189,	1.02%
190,000-199,	0.51%
Grand Total	100.00%

Credit Card Table (ExitedFromBank? (All))

Has Credit Ca	Count of HasCrCard?
0	29.36%
1	70.64%
Grand Total	100.00%

Expected Salary Table (ExitedFromBa (All))

Expected	Count of Estimated Salary
0-9,999	5.45%
10,000-19,9	5.95%
20,000-29,9	4.44%
30,000-39,9	4.84%
40,000-49,9	4.54%
50,000-59,9	4.24%
60,000-69,9	4.34%
70,000-79,9	6.86%
80,000-89,9	4.84%
90,000-99,9	5.45%
100,000-109	4.44%
110,000-119	5.95%
120,000-129	5.75%
130,000-139	4.44%
140,000-149	5.65%
150,000-159	4.64%
160,000-169	4.24%
170,000-179	4.64%
180,000-189	4.44%
190,000-199	4.84%
Grand Total	100.00%

Insights & Recommendations

- **Key learnings:**
  - Inactivity appears to be a significant factor contributing to customers leaving the bank.
  - Majority of former members are aged between 38–47-year-olds.
  - A significant proportion of former customers had higher account balances, specifically between $100k-$140k.
  - Majority of the former customers held only one product.
  - A higher proportion of former customers are female, while current customers are predominantly male.
- **Recommendations:**
  - Boost customer engagement with loyalty programs, personalized offers, and consistent communication.
  - Create customized financial products and services designed to meet the specific needs of individuals between 38-47-year-olds.
  - Enhance personalized financial advisory services to cater specifically to the financial needs of individuals with higher account balances
  - Promote product diversification among customers.
  - Gain insights into the unique needs and preferences of female customers.
- **Project Deliverables & Links:** Project Report

NFL Weather Spread

Project Overview

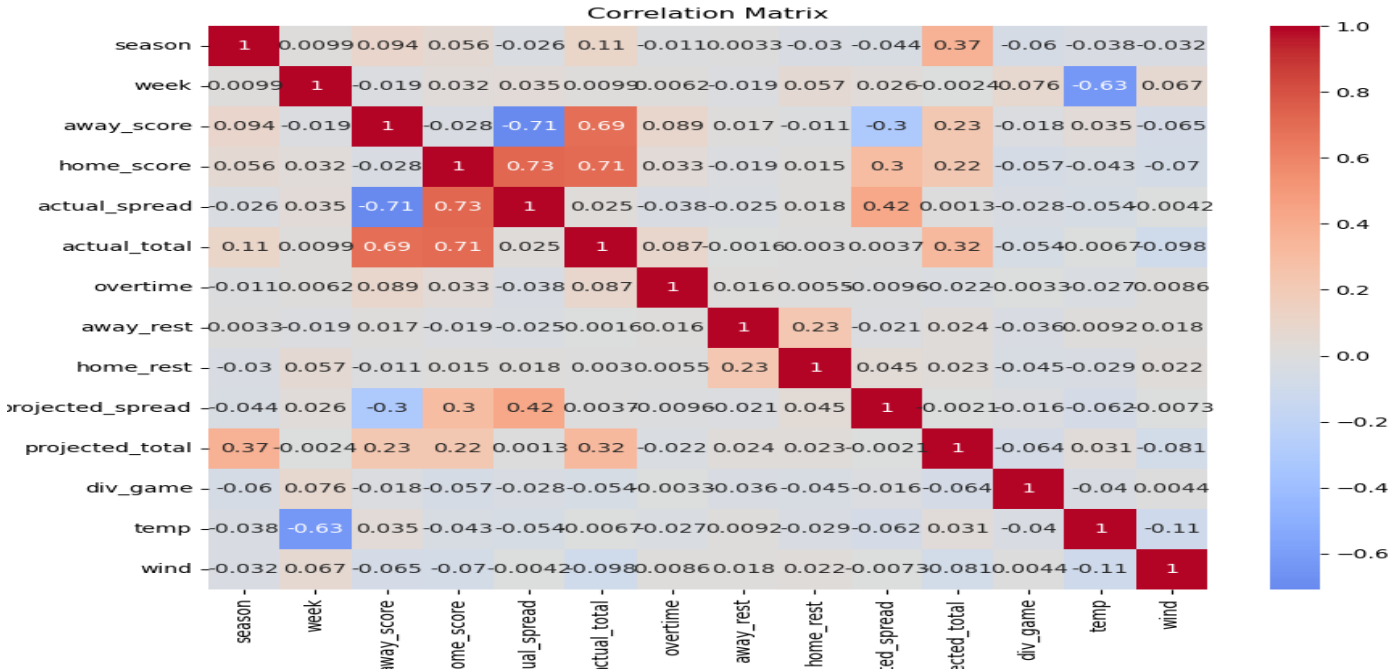
- **Objective:** Analyze how various geospatial and environmental factors, such as game location, stadium type, and weather conditions, influence the outcomes of NFL games. By examining these variables, we aim to identify patterns and correlations that can provide insights into team performance, game outcomes, and the impact of environmental conditions on the game.
- **Key Questions:**
  - Is there a correlation between the spread line and the actual score difference?
  - How do different teams perform against the spread?
  - How do temperature, wind speed, and other weather conditions affect game outcomes and scores?
- **Dataset:**
  - Data originates from the GitHub nflverse community using ProFootballReference which contains teams, scores, temperature, wind, spread, and over & under.
  - Link: nfldata/DATASETS.md at master · nflverse/nfldata (github.com)

Skills Applied

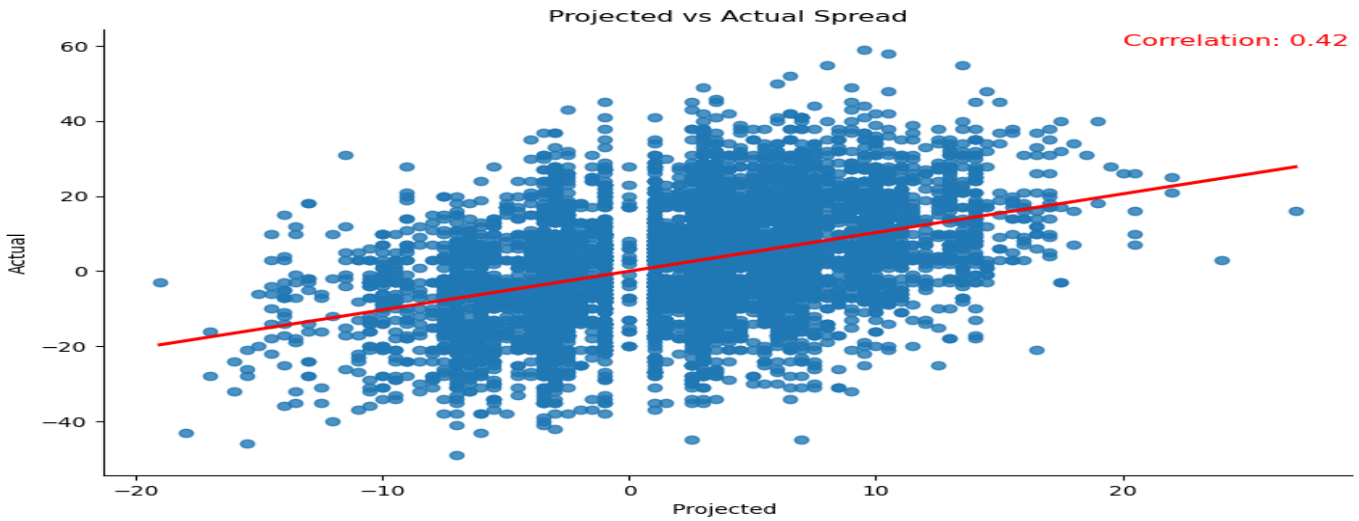
- **Python Programming:**
  - Data cleaning
  - Correlation charts
  - Geographic visualization
  - Regression
  - Machine Learning
  - Time-series analysis
- **Tableau:**
  - Geographic Visualization
  - Scatterplots

Analysis & Visualizations

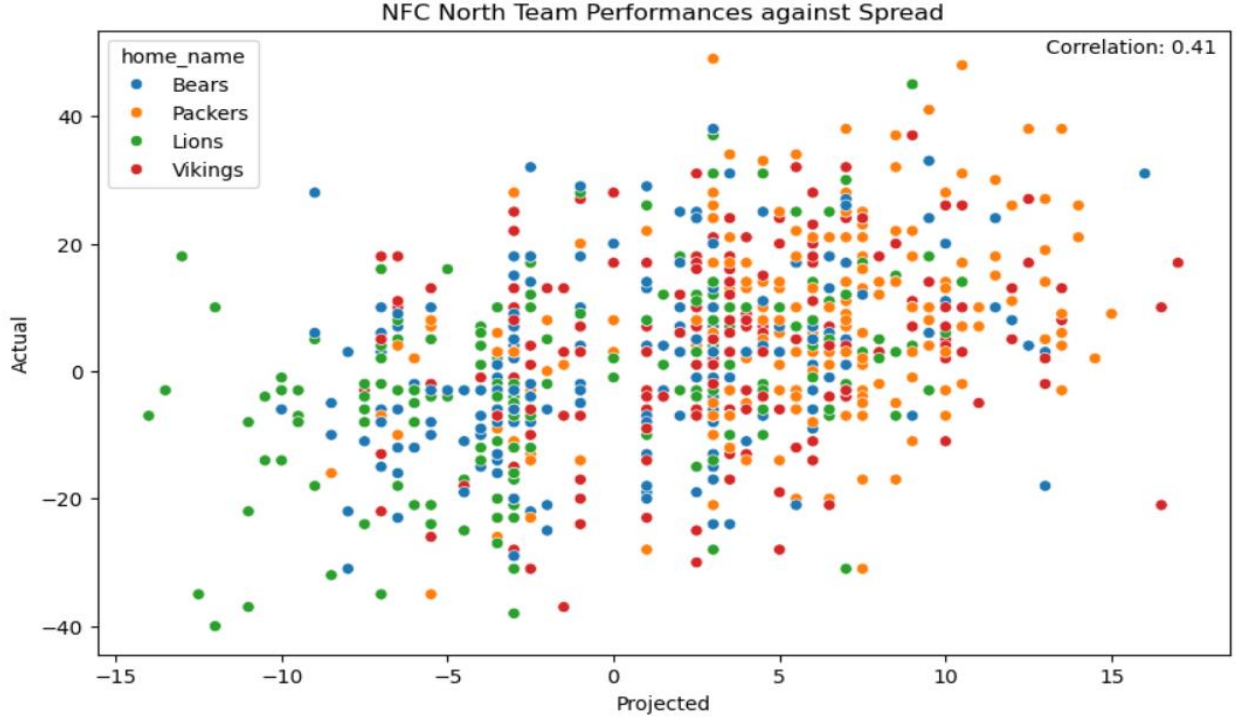
- Variables within the dataset with the highest correlations based on the matrix would be the projected and actual spreads along with totals at 0.42 and 0.32 respectively.
- There appears to be a negative relationship between actual spread and temp at -0.05.



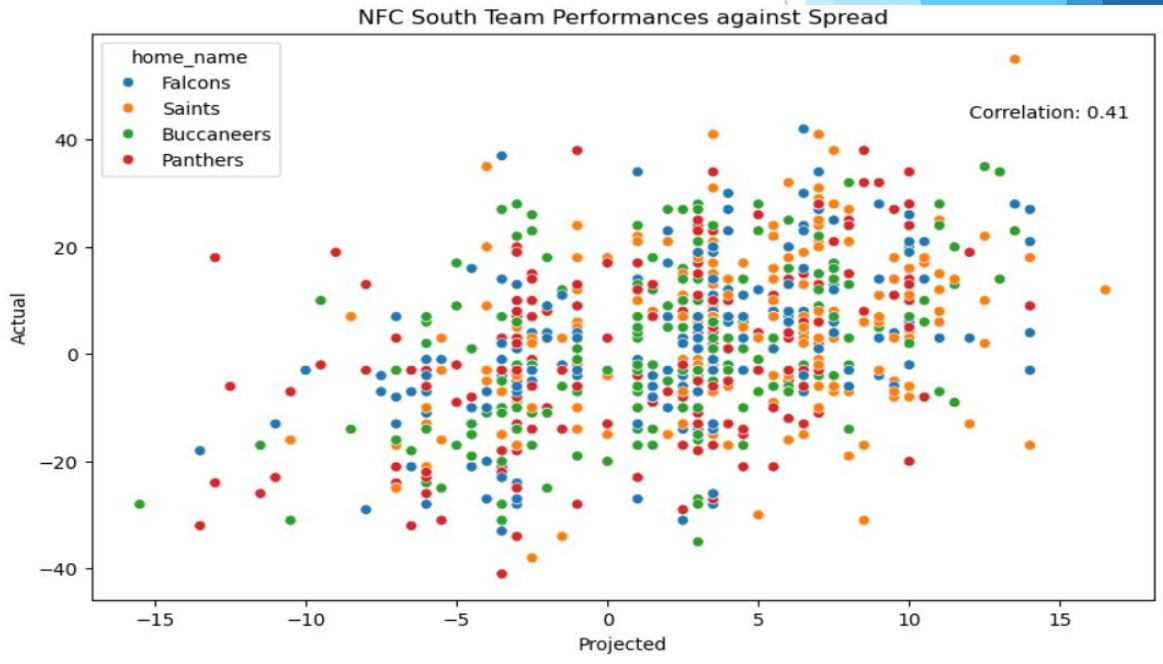
- Projected vs actual spread scatterplot indicate a moderate relationship at 0.42 with plenty of variation.
- This also indicates that sportsbooks projected spreads show tons of variation.



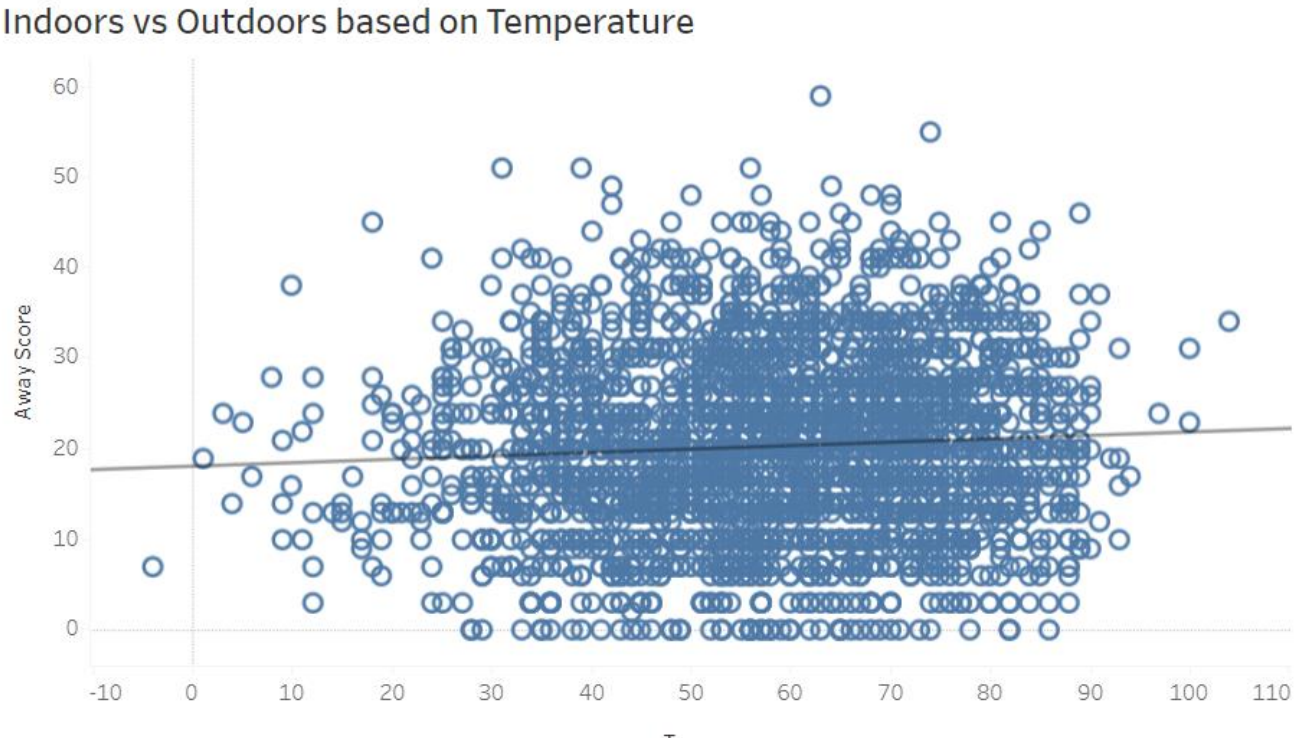
- NFC North team performances against the spread vary as the Lions and Packers were unfavored(e.g.; -5) and favored(e.g.; 5) to win their matchups, respectively.
- The relationship with these teams are moderate with a correlation of 0.41.



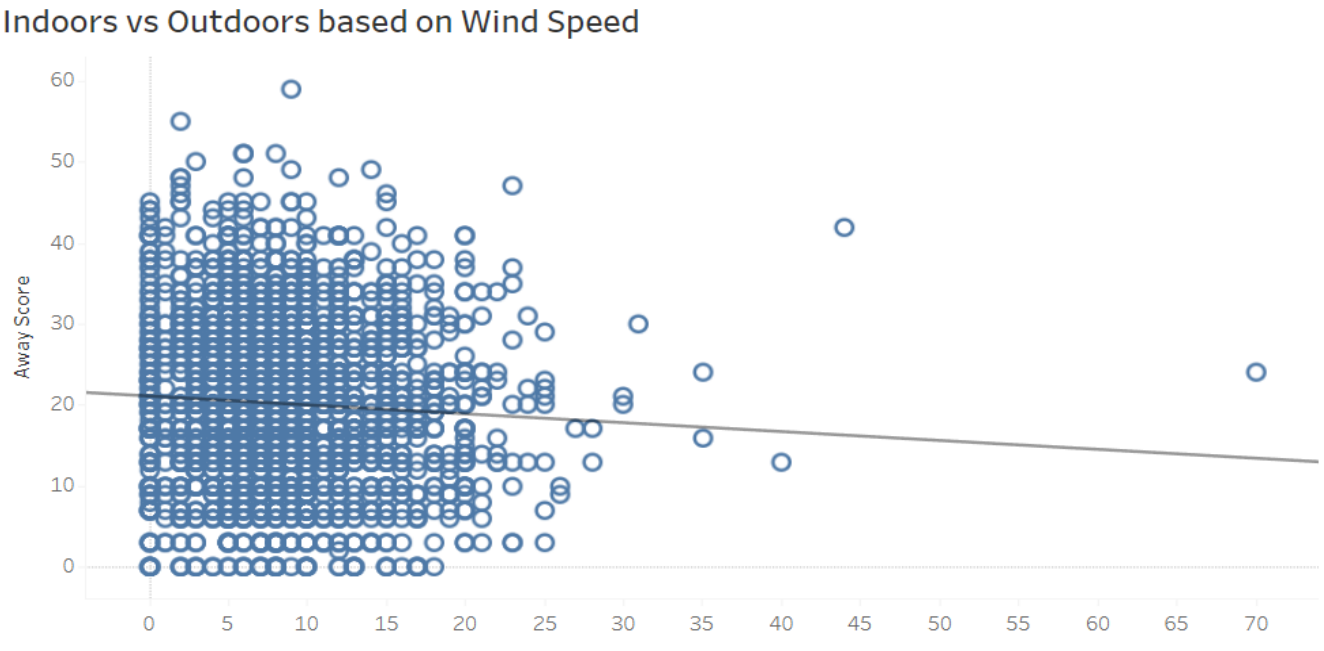
- NFC South team performances against the spread vary as the Panthers and Saints were unfavored(e.g. -10) and favored(e.g. 10) to win their matchups, respectively.
- The relationship with these teams are moderate with a correlation of 0.41.



- The relationship between visiting indoor teams/warm climate teams score and temperature is virtually none with a correlation of 0.004.
- The visiting teams scores vary a ton regardless of temperature.



- The relationship between visiting indoor teams/warm climate teams score and wind speed is virtually none with a correlation of 0.003.
- The visiting teams scores vary a ton regardless of wind speed.



Key Learnings & Recommendations

- **Key Learnings:**
  - The projected spreads from sportsbooks and the actual spread of the NFL games had some correlation for prediction.
  - NFL game outcomes for away teams that are indoor or in warmer climate cities are hardly affected by the temperature and wind speed.
  - NFL away score clusters vary regardless of weather elements.
  - Total points scored by both NFL teams tend to remain consistent throughout the years and months.
  - Northern city home NFL teams have a little bit higher spreads than other regions.
- **Recommendations:**
  - Include additional datasets that have more statistics such as passing, receiving, and rushing, turnovers; etc.
  - Include travel distance data for every game to see how different time zones affect outcomes.
  - Include detailed sportsbooks datasets that include more details and information different kinds of bets to compare with weather and etc.
- **Project Deliverables & Links:**
  - GitHub Repository

Thank You

Click the images below to connect