Cyndi Kohashi

Data Analyst Portfolio

About Me

Hello and welcome to my portfolio!

I'm Cyndi Kohashi, a data analyst with a background in visual design and finance.

These opposing backgrounds have led me to the field of data analysis, where I can use both together to discover insights and present results visually.

As a data analyst my goal is to solve problems and answer questions with thorough analysis and visual design.

Table of Contents

GameCo

A fictional video game company requesting global sales analysis to help with developing the marketing budget for 2017.

GameCo's current view is that regional sales have remained the same over time.

Project Overview

Goals:

The analysis is to help inform GameCo's upcoming 2017 marketing budget, and get a better understanding of how their games may fare in the market.

Other key questions include:

- Is our current understanding that regional sales have remained the same over time still correct?
- Are certain types of games more popular than others?
- What other publishers will be main competitors?
- Have any games increased or decreased in popularity over time?

To answer these questions, we will look at sales over time to see if there are any changes in regions, games, publishers, platforms, or genres.

Skills

- **Grouping and summarizing data in Excel**
- **Cleaning data**
- **Performing descriptive analysis**
- **Visualizing and presenting results**

Tools

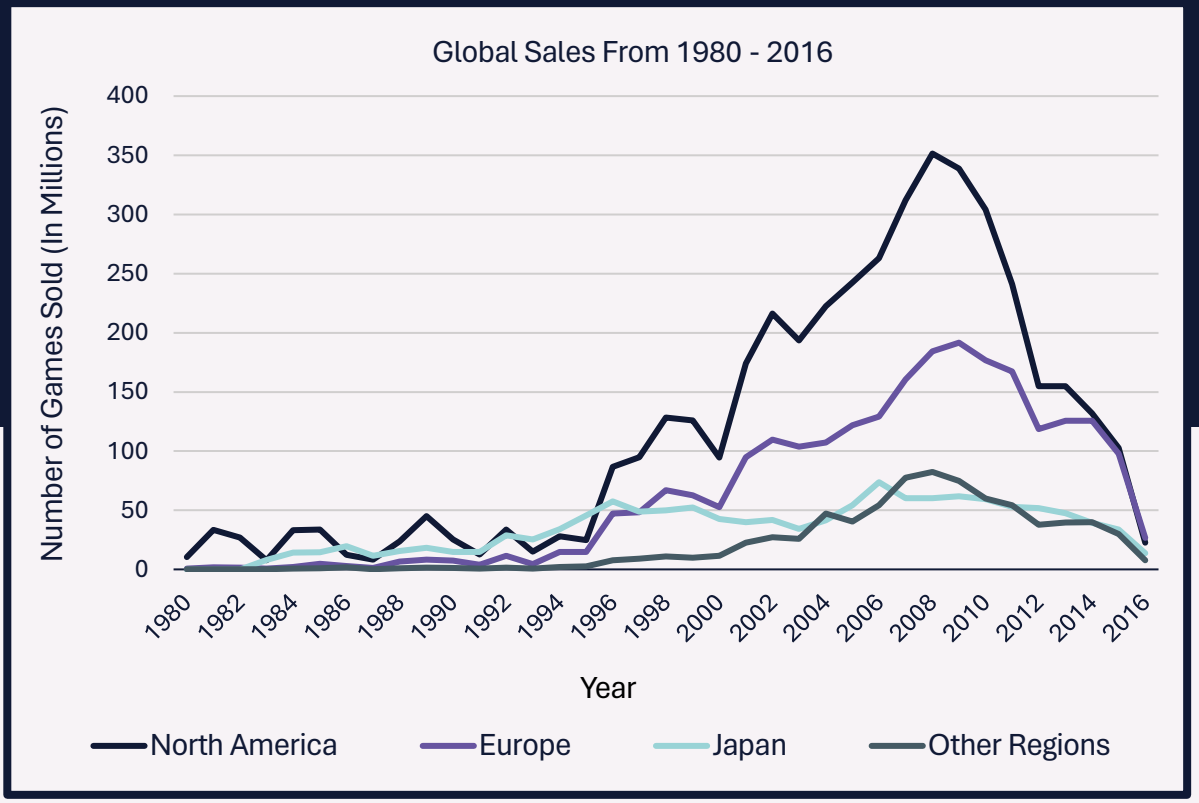
- **Microsoft Excel**
- **Microsoft PowerPoint**

Data

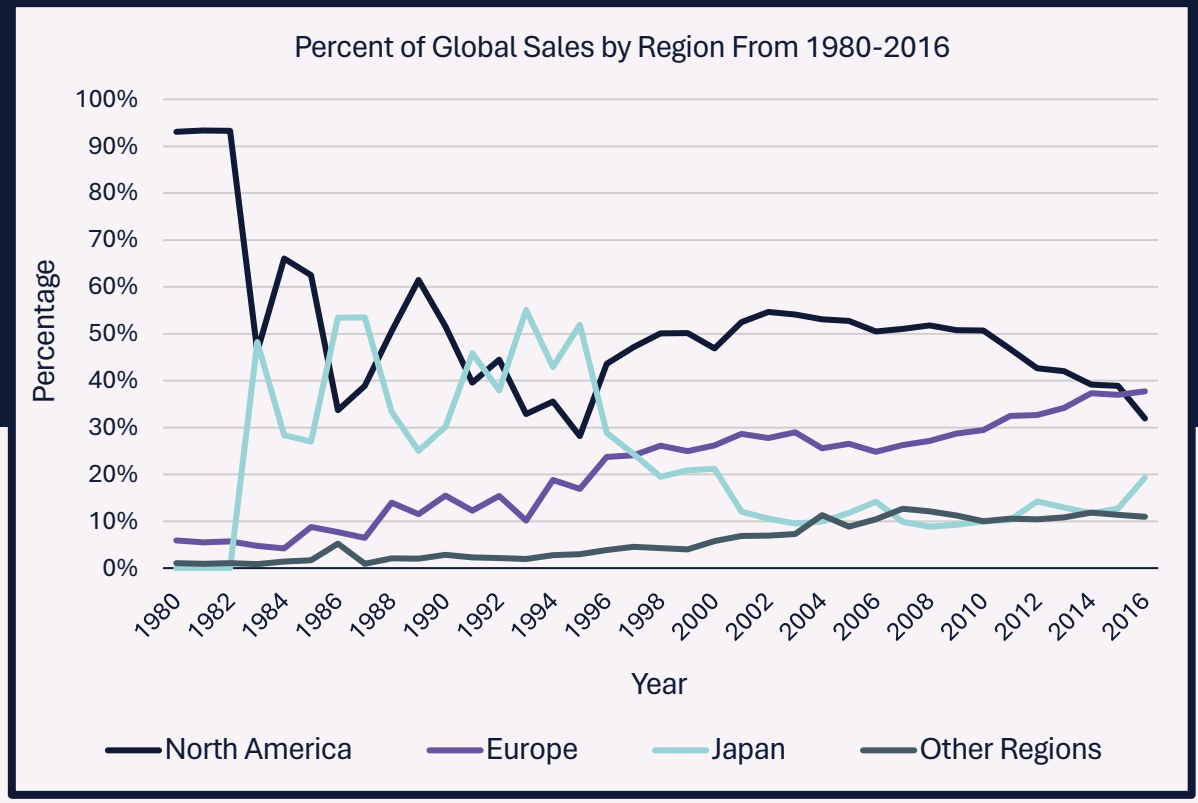
- **Sales data sourced from VGChartz**
- **Includes physical games sold from 1980-2016**
- **Does not contain financial figures**
- **VGChartz's data collection methodology**
- **Project brief**

Regional Sales Have Changed Over Time

Looking at sales over time, we can see that physical video game sales have been declining. Sales in 2016 are **90%** less than in 2008, the peak of game sales.



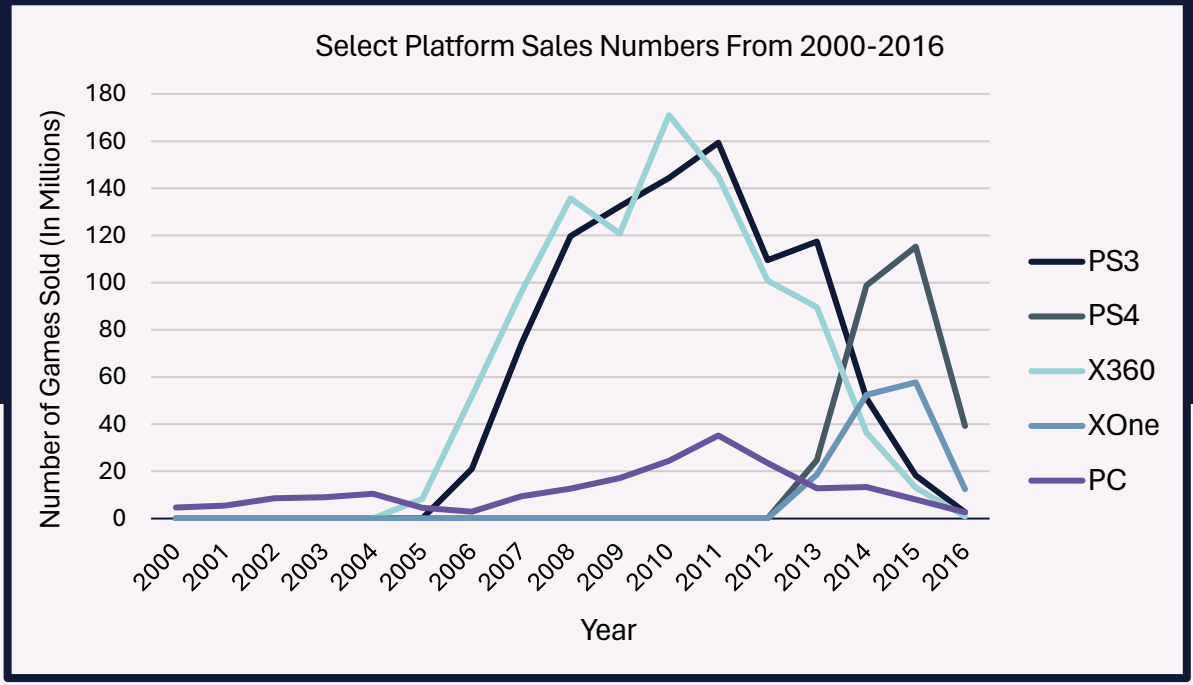
Regional dominance has also changed over time. Generally North America is the largest market, but this has changed recently with **Europe** now making up the most sales at **38%**.



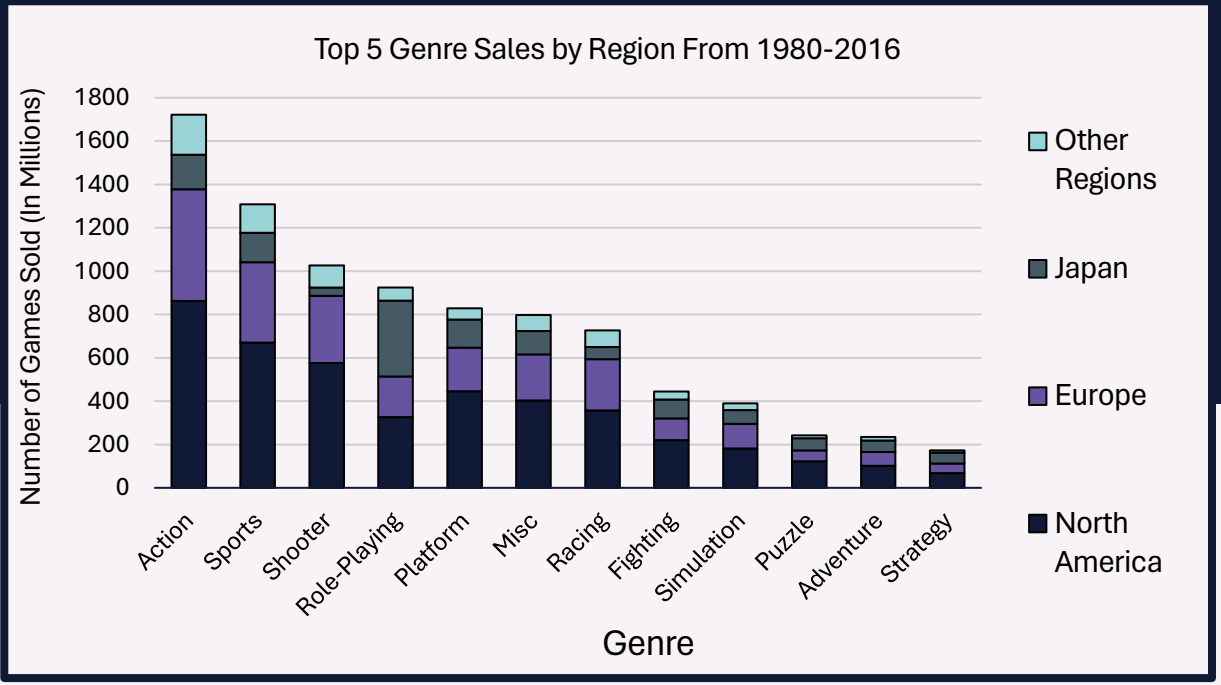
One main factor in the decrease of physical sales is the advancement of technology. Games can now be purchased online, paid for on an ongoing basis via subscription, and so on. This can affect the monetary models GameCo chooses to use moving forward.

Platforms Change, Genres Preserve

Platforms experience high game sales for a short period of time before being replaced with a new model. An example below is of the PlayStation and Xbox consoles. The PC is the longest running platform, but only makes up 2.89% of all sales. We can continue to expect this trend with the release of new consoles in the future.



Recently the most popular genres are **Sports** and **Shooter**, but Action games are still the third highest selling genre in 2016. All three genres are the most popular in Europe and North America.



It would be better to release games based on console generations (PS4, XOne, Wii U) than on console families (PS2, PS3, PS4) unless GameCo has any made any platform exclusive agreements.

GameCo can also research future consoles that may be more relevant when their games finish development.

Action games make up 19% of all global sales from 1980-2016, and 28% of all sales in 2016. It's still a lucrative genre even though it's not the most popular today.

Popular Games and New Publisher Trends

Top 10 Selling Games From 1980-2016:

1. Wii Sports (2006)
2. Super Mario Bros. (1985)
3. Mario Kart Wii (2008)
4. Wii Sports Resort (2009)
5. Pokémon Red/Blue (1996)
6. Tetris (1989)
7. New Super Mario Bros. (2006)
8. Wii Play (2006)
9. New Super Mario Bros. Wii (2009)
10. Duck Hunt (2009)

Top 10 Selling Games of 2016

1. FIFA 17 (PS4)
2. Uncharted 4: A Thief's End (PS4)
3. Tom Clancy's The Division (PS4)
4. Far Cry Primal (PS4)
5. Tom Clancy's The Division (XOne)
6. Overwatch (PS4)
7. No Man's Sky (PS4)
8. Dark Souls III (PS4)
9. FIFA 17 (XOne)
10. Doom 2016 (PS4)

Top 5 Publisher Sales by Genre in 2016	Action	Sports	Shooter	Role - Playing	Fighting
#1 Electronic Arts	--	88%	8%	--	--
#2 Ubisoft	39%	--	61%	--	--
#3 Sony Computer Entertainment	18%	--	64%	--	--
#4 Namco Bandai Games	11%	--	--	40%	42%
#5 Activision	6%	--	86%	8%	--

All these games were published by Nintendo.

5 of these games were for the Wii platform and only 3 were from an Action, Sports, or Shooter genre.

Like regional and genre changes, major publishers have also changed.

Nintendo is the highest selling publisher overall but the 8th highest selling publisher in 2016.

This contrasts with the top 10 games of 2016, which has more publisher diversity.

In 2016 majority of sales from the top 5 publishers were in the top 3 genres.

The outlier is Namco Bandai Games. Their most popular genres were Role-Playing and Fighting.

There are gaps in the publishers' genres that can be capitalized on. For example, Ubisoft, Sony Computer Entertainment, Namco Bandai Games, and Activision didn't sell any Sports games in 2016.

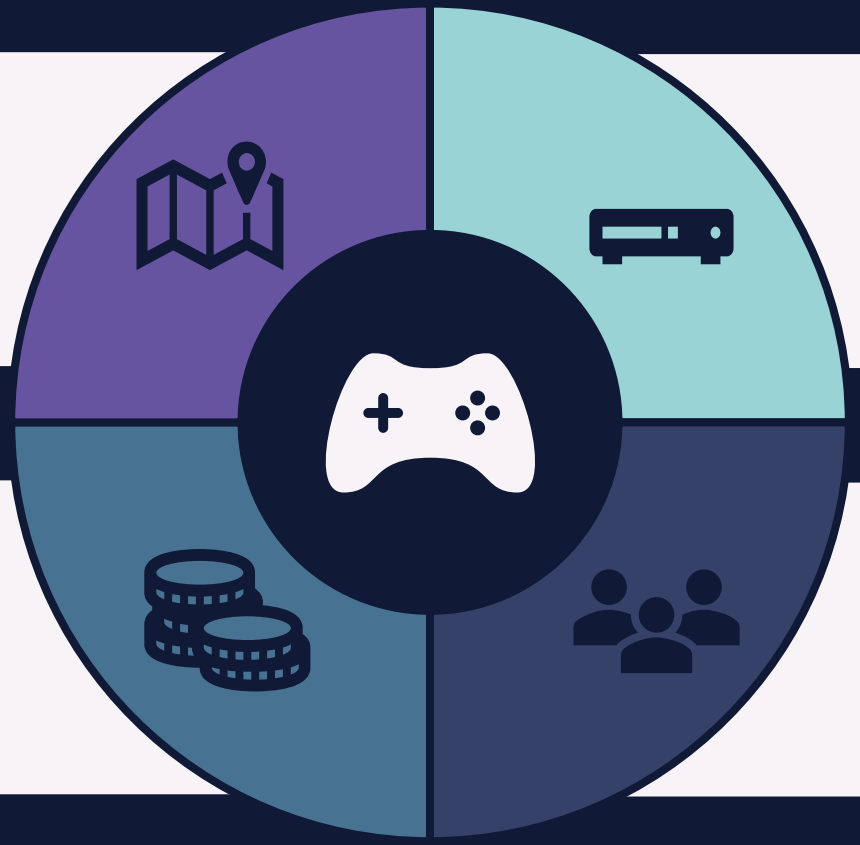
Conclusion and Recommendations

Regions & Genres

Focus on Europe and North America. Prioritize Action, Shooter, and Sports genres. These are the most popular, and there are still gaps in other publisher's genres to take advantage of.

Platforms

Release games on multiple platforms to increase accessibility, but older platforms should be lower priority. These will eventually lose sales as new consoles launch in the future.



Alternate Monetary Models

With the decline of physical sales, consider other income methods like online sales, in-game purchases, subscriptions, etc.

Current Fanbase

Current customers should still be considered and can be surveyed for insights. There may be demand for a sequel or genre that will boost sales.

View : PDF Full Presentation Back to Table of Contents

Influenza Season

Prepare for the next influenza season in the United States by helping a medical staffing agency determine when, where, and how much staff to send to each state.

People part of vulnerable populations are more likely to develop complications and become hospitalized from the flu.

Project Overview

Goals:

The goal is to help a medical staffing agency determine when, where, and how much staff to send to states for the next flu season. The agency sends temporary workers to existing clinics, but there is no budget to hire additional personnel.

People considered part of a vulnerable population can become hospitalized because of the flu. Those hospitals would then need more staff to properly treat those extra patients.

Vulnerable populations are defined by the CDC (Center for Disease Control and Prevention) as adults over age 65, children under 5, pregnant people, individuals with HIV/AIDs, cancer, heart disease, stroke, diabetes, asthma, and children with neurological disorders.

The number of deaths from influenza is an indicator of the severity of flu in that area. Deaths can be prevented with flu shots and adequate staff.

We will be looking at states' population composition, number of deaths, influenza visits, and historical trends to help with staffing and create a priority map of states.

Skills

- **Designing a data research project**
- **Sourcing, cleaning, and profiling data**
- **Data integration and transformation**
- **Statistical hypothesis testing**
- **Visual analysis and forecasting**
- **Visualization in Tableau**

Tools

- **Microsoft Excel**
- **Microsoft Word**
- **Tableau**

Data

- **Influenza deaths in 2009-2017 sourced from CDC**

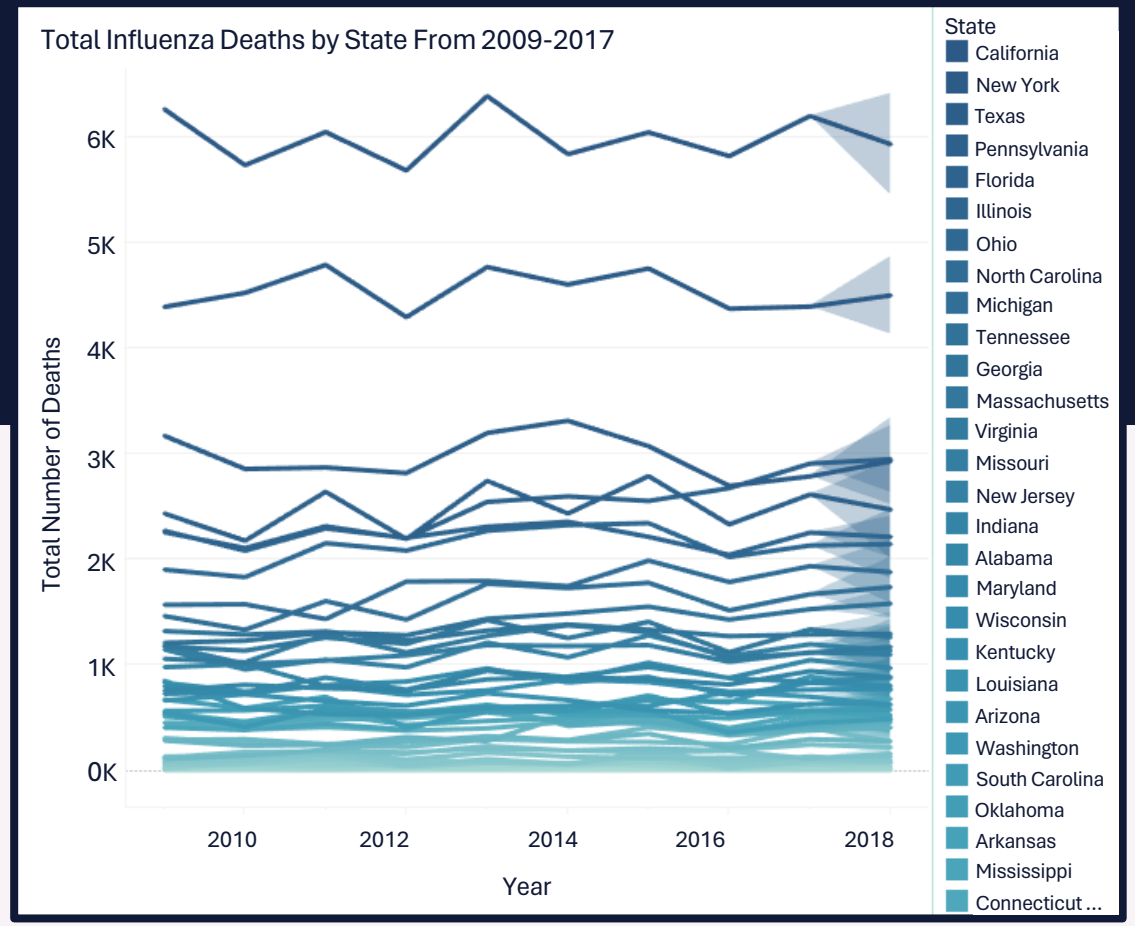
  (Mortality data from certain states were suppressed due to privacy. Death records identify primary cause of death. Influenza-initiated deaths not recorded.)
- **Population data from 2009-2017 sourced from US Census Bureau**

  (Population numbers are estimates.)
- **Influenza visits from 2010-2019 sourced from CDC (Fluview)**

  (Data from Florida suppressed due to privacy. Contains number of medial providers, not individual staff count.)
- **Project brief**

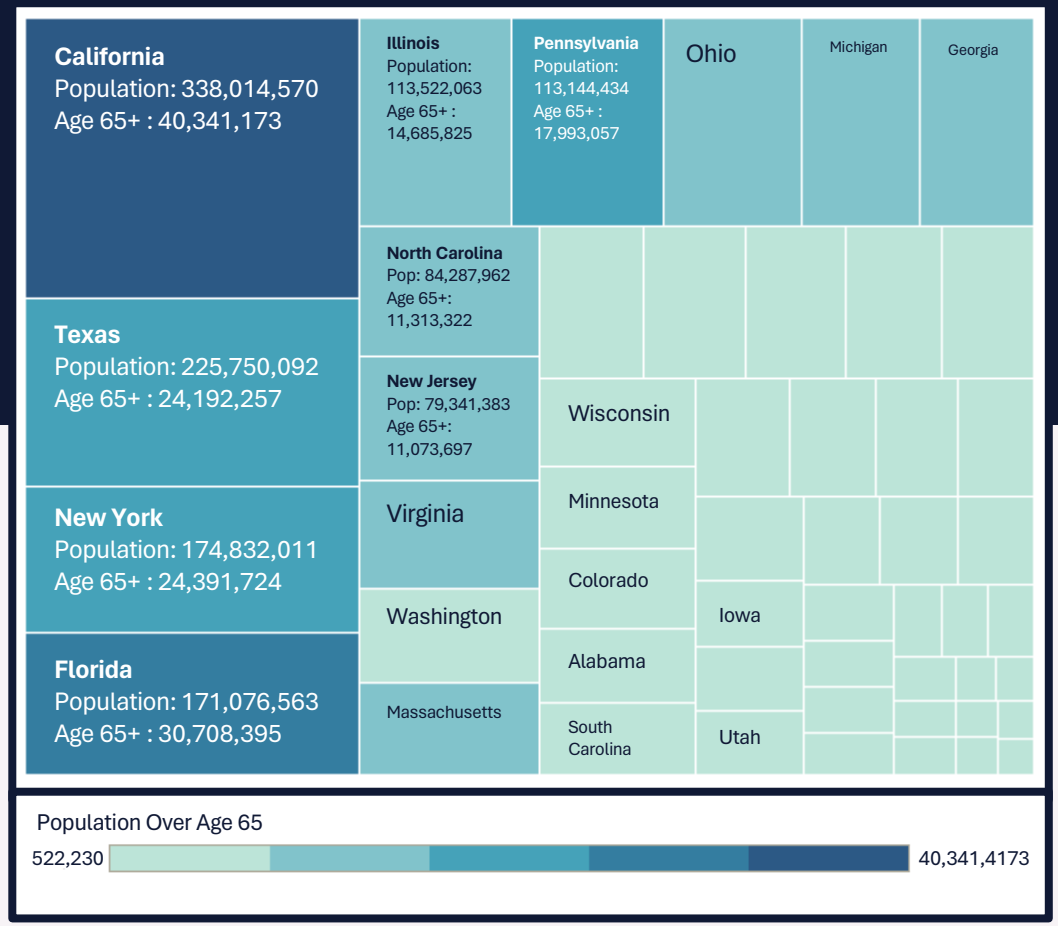
Population Size and Number of Deaths

This graph shows total deaths from influenza by state during 2009-2017. We can see that two states in particular, New York and California, have the most deaths.



The shaded areas are forecast predictions for what may happen in 2018.

Comparing the deaths line graph to population size and vulnerable population (age 65+) size, we can see a relationship start to form.

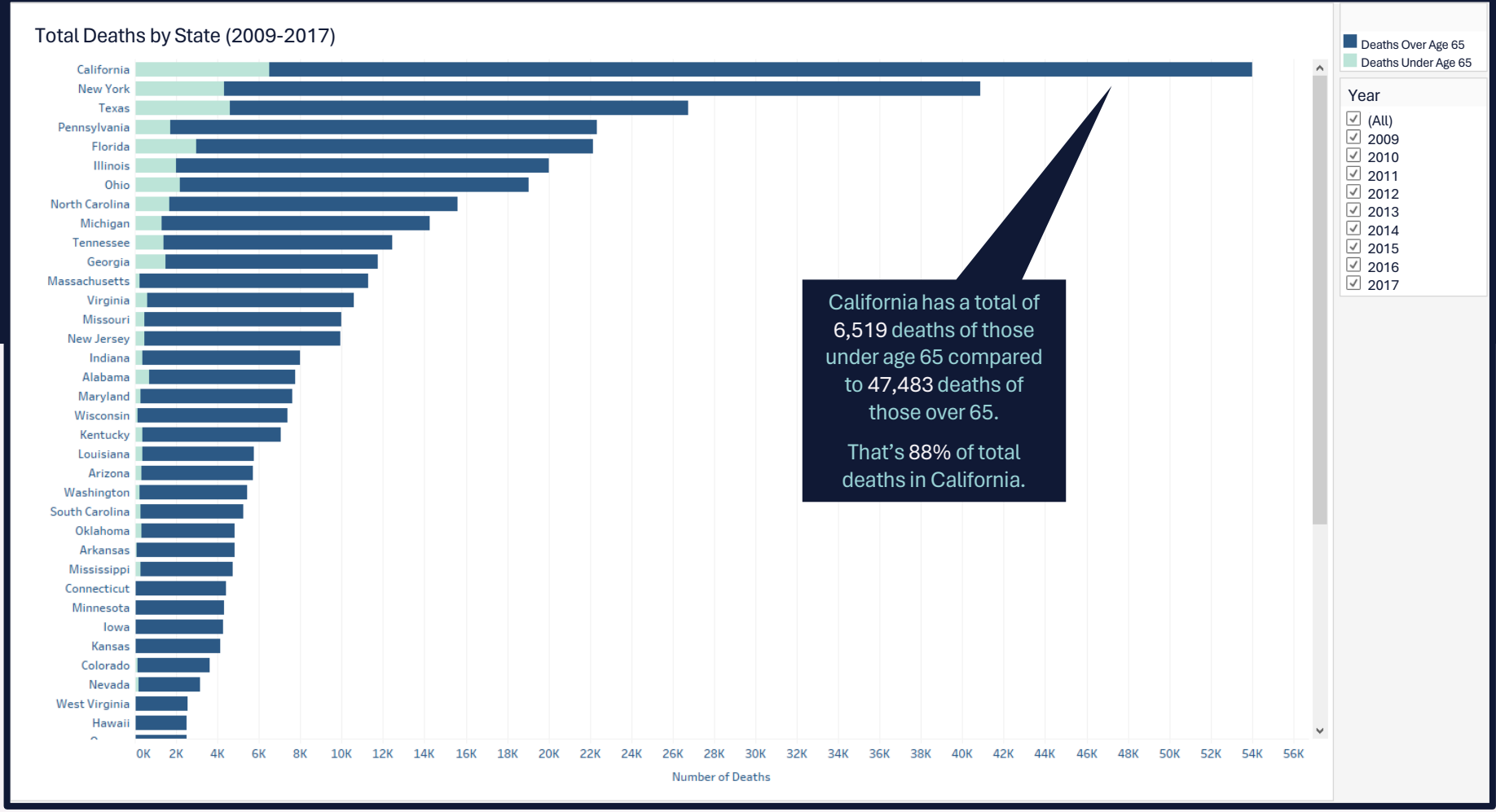


States with larger populations tend to have larger vulnerable populations, resulting in more deaths.

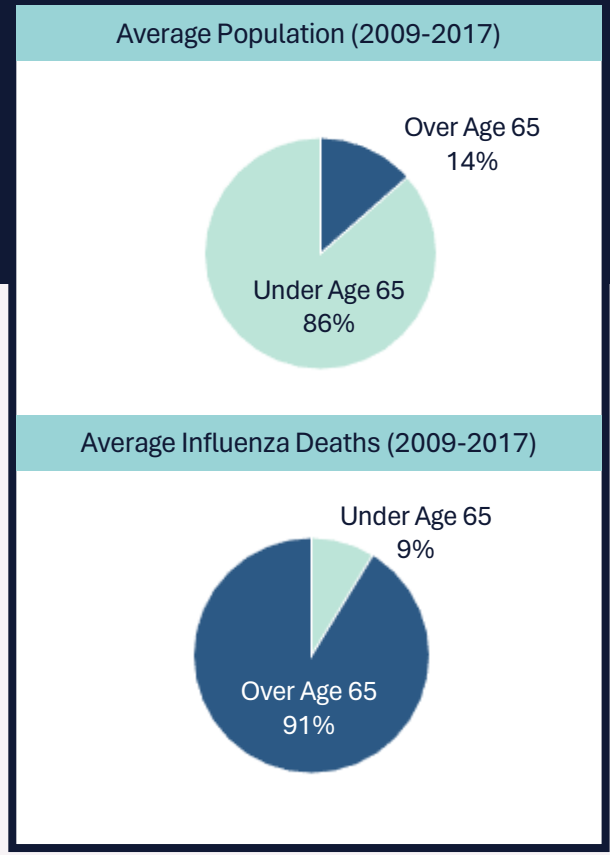
Vulnerable Population Analysis

We do not have data on all members of the vulnerable population (those under age 5 or having certain pre-existing conditions) but we do have data on those age 65 and over.

More people ages 65 and older die from influenza than those under age 65, and this is unfortunately true for all states and every year from the data we have (2009-2017.)
States with the most deaths continue to be those with large populations and large vulnerable populations.

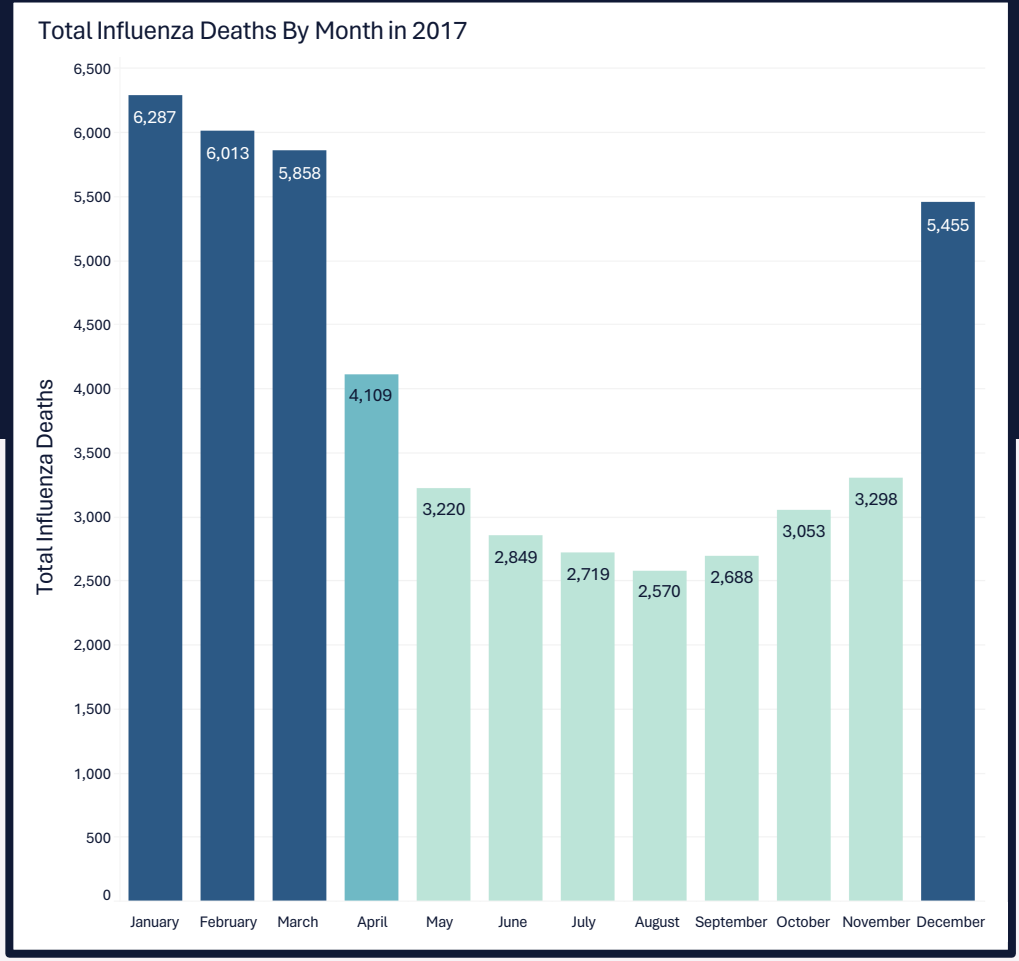


On average, people age 65 and over make up **14%** of the total population, but account for **91%** of all influenza deaths.



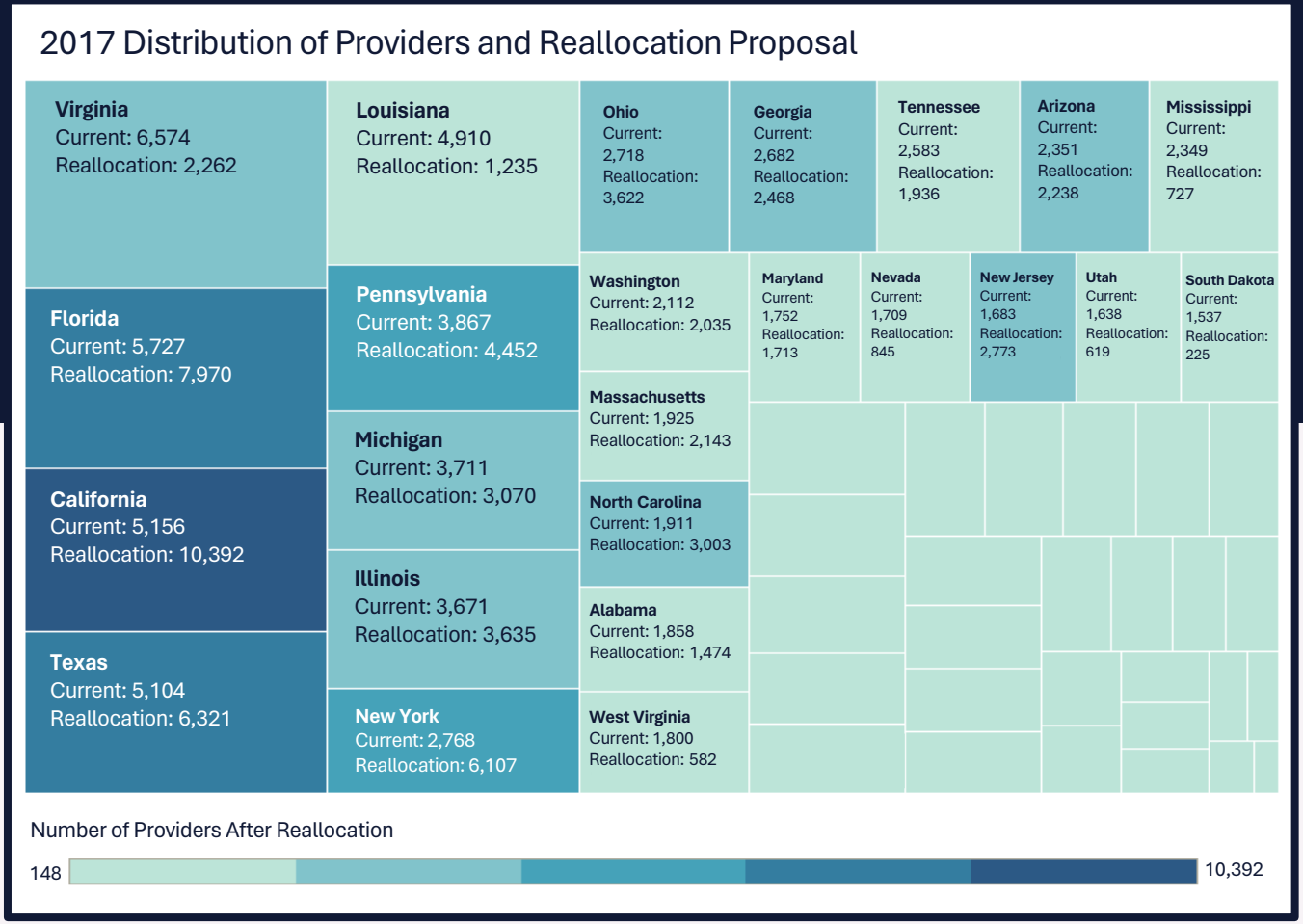
Seasonality and Staff Allocation

Most states have an influenza season from December to March. That is when influenza deaths are at their highest. Some states, such as Florida and New Jersey, show an increase in the months before the season.



16% of states have influenza peaks outside of the season.

This reallocation tree map is based on each state's percentage of the total vulnerable population. That percentage was then applied to the current number of staff in 2017. By using vulnerable population numbers, we can send more staff to areas with more people at risk to prevent deaths.



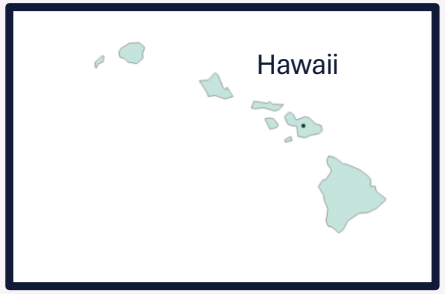
Florida's current staff number was not provided, and is an estimate based on the national average population and average number of patients seen per provider.

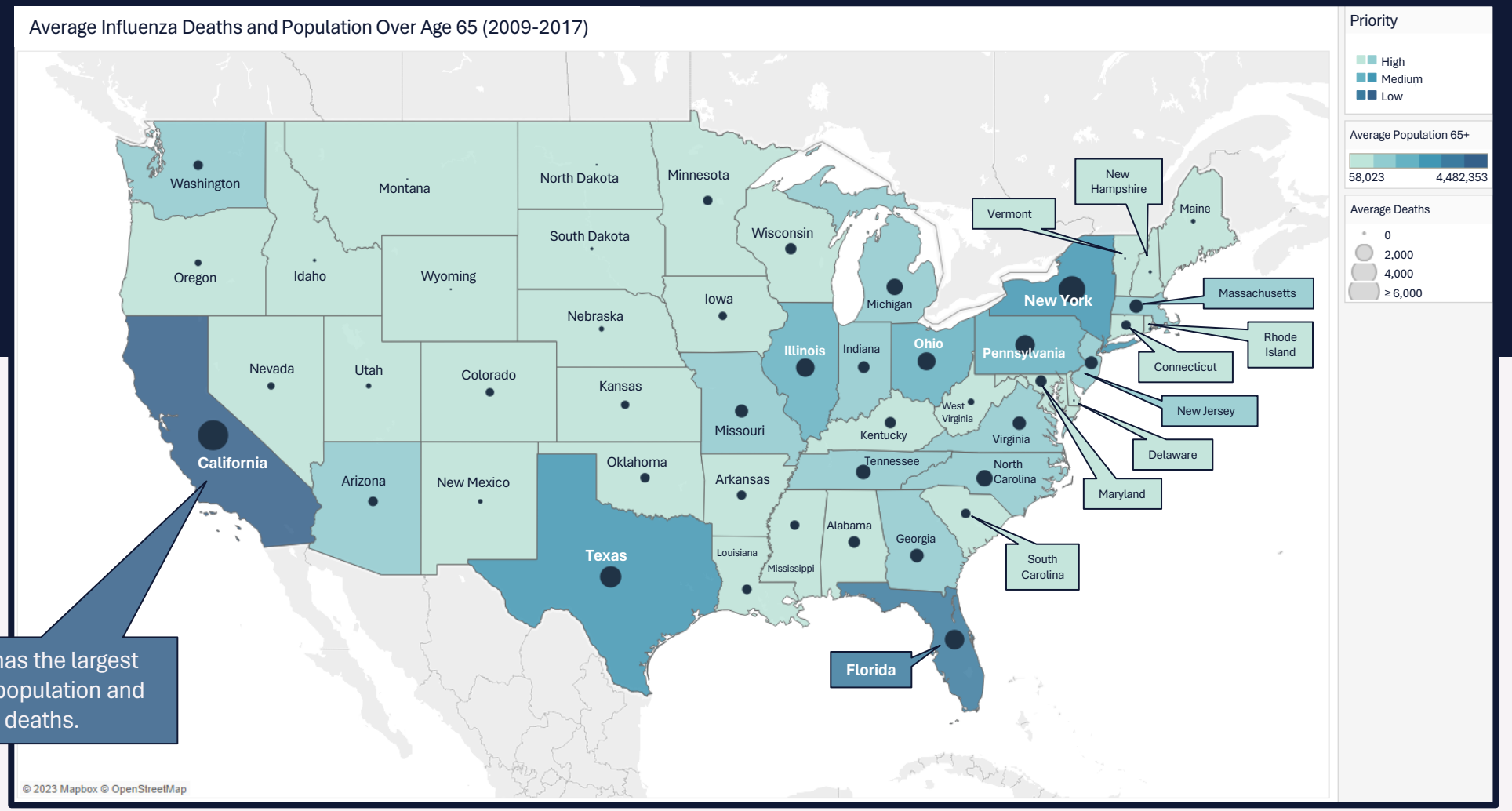
United States Priority Map

This map breaks down states by priority using the average population of those over 65 and the average number of deaths. There are two sections for each level (high, medium, low) to give the categories more specificity.

Our high priority states are California, Florida, New York, Texas, Pennsylvania, Ohio, and Illinois.

Alaska and Hawaii are both low priority states.





California has the largest vulnerable population and most deaths.

Conclusion and Recommendations

Staff Deployment

Relocate staff based on the size of a state's vulnerable population and priority status.

Timing

Staff should mainly be deployed from December to March, but can be sent in November and April to get ahead of the influenza season.

Continued Monitoring

Monitor upcoming influenza season to determine success of analysis. Staff should be surveyed for input and to refine logistics of deployment.

Prevention

Promotion of influenza vaccinations and prevention can help reduce cases across all states.

View : Tableau Visualization Interim Report Back to Table of Contents

Rockbuster Stealth LLC

A fictional movie rental company with physical stores around the globe, Rockbuster Stealth LLC is planning to launch an online video rental service.

The purpose of this analysis is to help the launch strategy by looking at revenue gain, rental statistics, and geographic regions.

Project Overview

Goals:

Rockbuster Stealth LLC plans on launching an online rental platform to compete with other streaming services.

Key questions include:

- Which movies contribute the most/least revenue gain?
- What was the average rental duration for all videos?
- Which countries are customers based in? Where are customers with high lifetime value based?
- Do sales figures vary between geographic regions?

Using SQL, we will retrieve the data we need on customers, location, revenue, and rental duration.

Skills

- **Database querying**
- **Filtering, cleaning, and summarizing data**
- **Joining tables**
- **Performing subqueries**
- **Using CTEs (common table expressions)**
- **Creating a data dictionary**
- **Creating an ERD (entity relationship diagram)**

Data

- **Rockbuster Stealth LLC data provided by CareerFoundry**
- **Project brief**

Tools

- **Microsoft Excel**
- **Microsoft Word**
- **Microsoft PowerPoint**
- **PostgreSQL**
- **DbVisualizer**
- **Tableau**

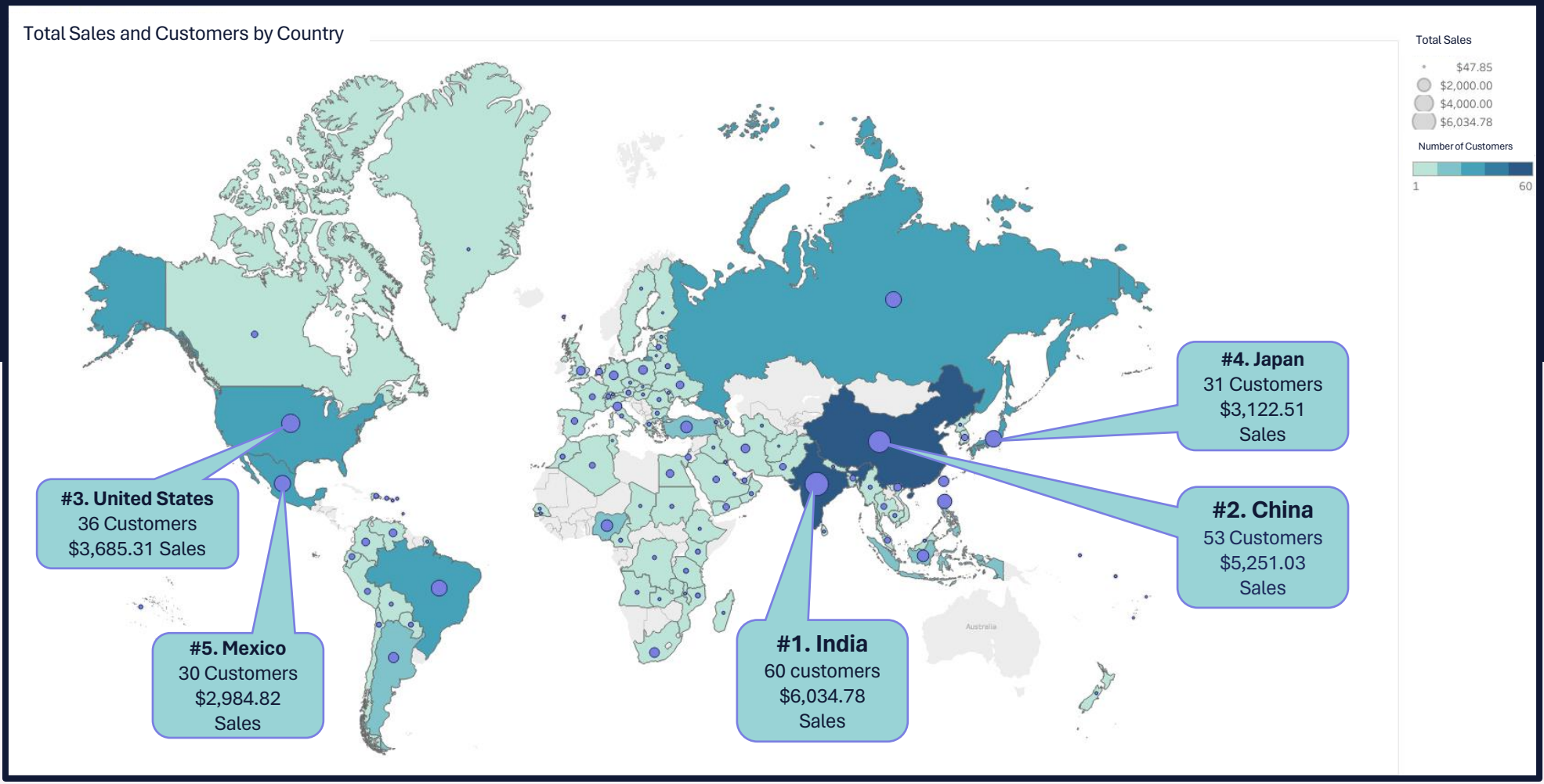
Rockbuster Sales and Customers Around the World

Regions with higher sales and customer numbers include Asia, North America, and South America.

1,000
Number of films Rockbuster has

5 Days
Average rental duration

PG-13
Most common film rating



$4.99
Maximum rental rate

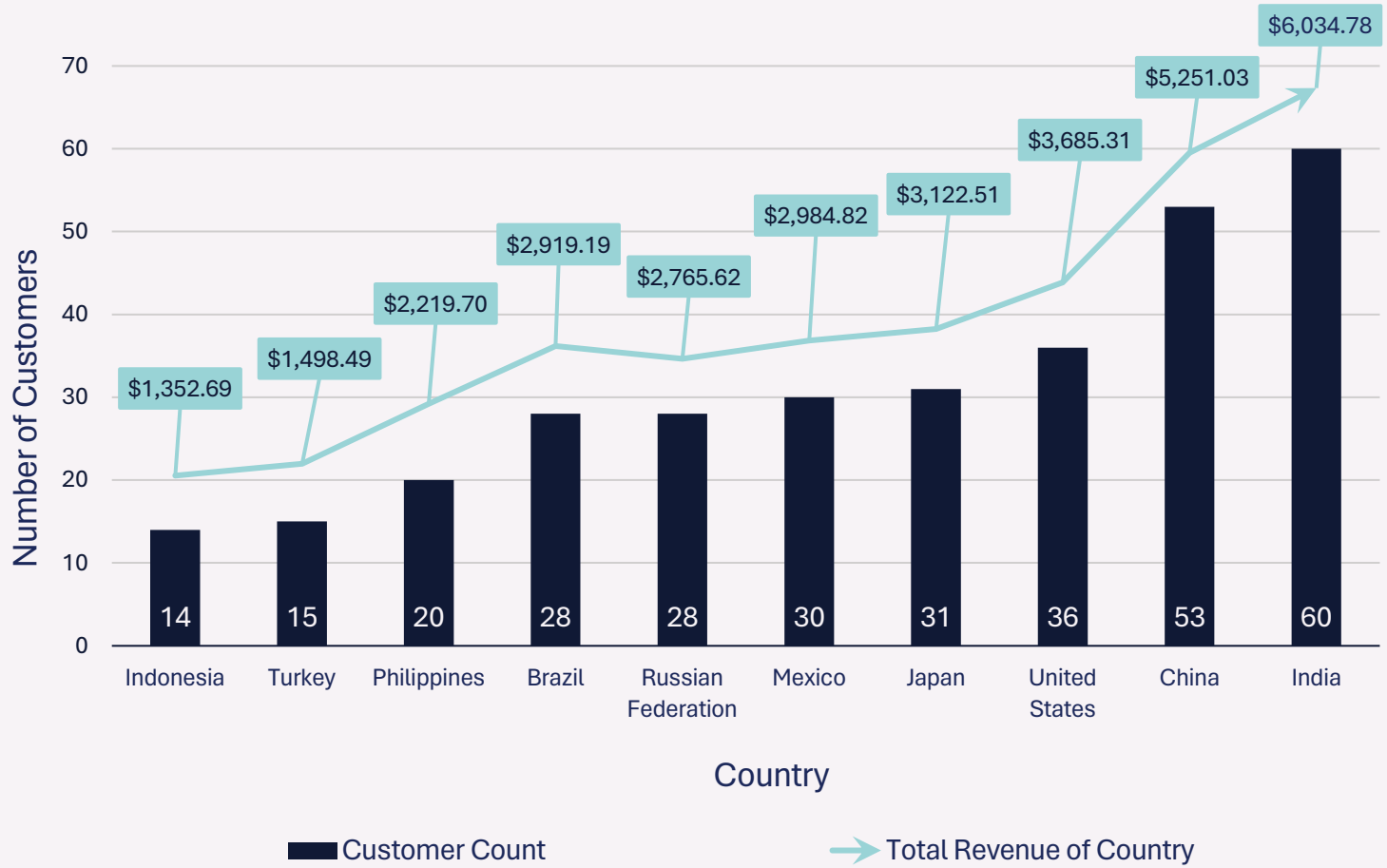
599
Number of current customers

$211.55
Most revenue generated by one customer

Top 10 Countries With the Most Customers and Sales

The countries with the most customers were also the countries with the largest sales.

These top countries make up **52.59%** of total customers and **51.92%** of total global revenue.



5 of the top 10 revenue generating customers were from the top 10 countries. This shows us that Rockbuster lifetime customers are diverse, but the top countries are still strong revenue generators.

Top 10 Revenue Generating Customers Globally

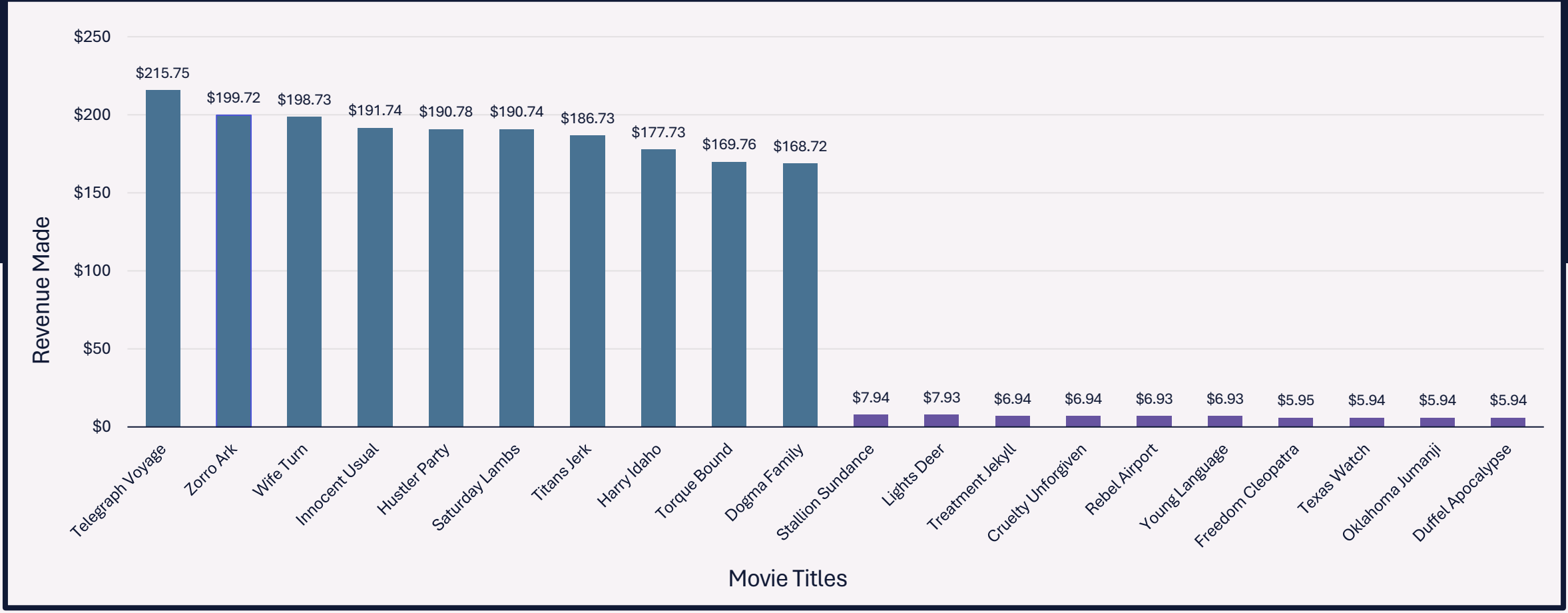
Name	City	Country	Total Revenue
Eleanor Hunt	Saint-Denis	Réunion	$211.55
Karl Seal	**Cape Coral**	**United States**	**$208.58**
Marion Snyder	**Santa Bárbara d'Oeste**	**Brazil**	**$194.61**
Rhonda Kennedy	Apeldoorn	Netherlands	$191.62
Clara Shaw	Molodechno	Belarus	$189.60
Tommy Collazo	Qomsheh	Iran	$183.63
Ana Bradley	**Memphis**	**United States**	**$167.67**
Curtis Irby	Richmond Hill	Canada	$167.62
Marcia Dean	**Tanza**	**Philippines**	**$166.61**
Mike Way	**Valparai**	**India**	**$162.67**

Top 10 Movies Generating the Most and Least Revenue

The top 10 revenue generating movies all have the maximum rental rate of $4.99, but their average rental duration is short at 4 days.

The bottom 10 revenue generating movies all have the minimum rental rate of $0.99, and a longer average rental duration of 6 days.

Rockbuster can increase revenue by increasing movie prices, but also by increasing the rental duration of the best-selling movies.



Conclusion and Recommendations

Customer Priority

Survey customers to get their input about the online launch. Consider starting a loyalty reward program for lifetime value customers.

Increase Revenue Options

Offer a monthly subscription alongside one-time rentals to reach more customers. Price movies over $4.99. Newer movies, for example, could cost more as their newness will be the incentive to rent them.

Increase Rental Duration

Launch promotions or deals to increase rental duration (e.g., rent for 4 days, get the 5th day free.)

Top Countries

Focus market on the top 10 countries, as they have the most customers and generate the most revenue.

View : Full Presentation | Data Dictionary | SQL Code | Tableau Visualization | Technical Report | Back to Table of Contents

Instacart

An online grocery delivery or pick-up service.

The goal of this project is to uncover customer purchasing patterns, which can be used for a targeted marketing strategy.

Project Overview

Goals:

By looking at customer, order, and product data with exploratory analysis we can find customer purchasing patterns for marketing.

Key questions include:

- Which days and hours have the most orders, so that ads can be run during slower times?
- When is the most money spent during the day?
- Which departments have the most orders?
- What does the brand loyalty of the customer base look like?
- Do ordering habits differ based on loyalty?
- Is there a regional difference in order habits?
- Is there a connection between age and family status?
- What are the differences between customer demographics and between customer profiles?

Because of the large size of datasets, we will use Python to retrieve the information we need, analyze it, and create visuals.

Skills

- **Data cleaning, wrangling, subsetting, and merging**
- **Data consistency checks**
- **Deriving new variables**
- **Grouping and aggregating data with Python**
- **Visualization with Python**
- **Creating population flow and reporting in Excel**

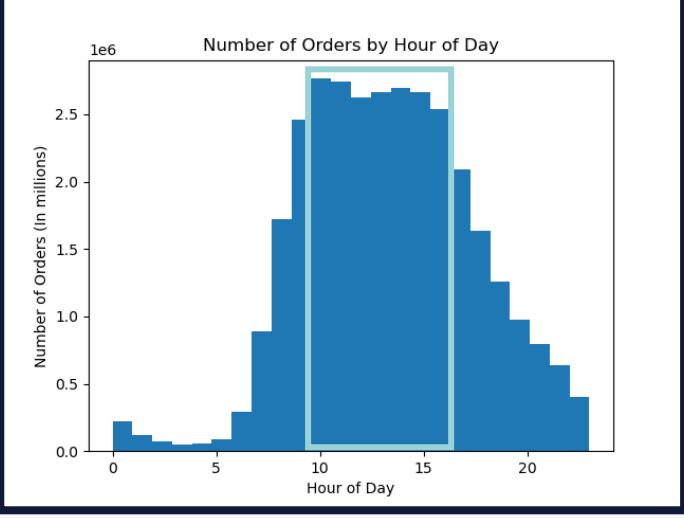
Data

- **Data is the Instacart Online Grocery Shopping Dataset 2017, accessed from https://www.instacart.com/datasets/grocery-shopping-2017 via Kaggle on Nov 15th, 2023.**
- **Customer data and prices were created for educational purposes**
- **Data dictionary was provided**
- **Project brief**

Tools

- **Microsoft Excel**
- **Anaconda**
- **Jupyter Notebook**
- **Python**
- **Python libraries Pandas, NumPy, Seaborn, Matplotlib, and SciPy**

Ordering Habits

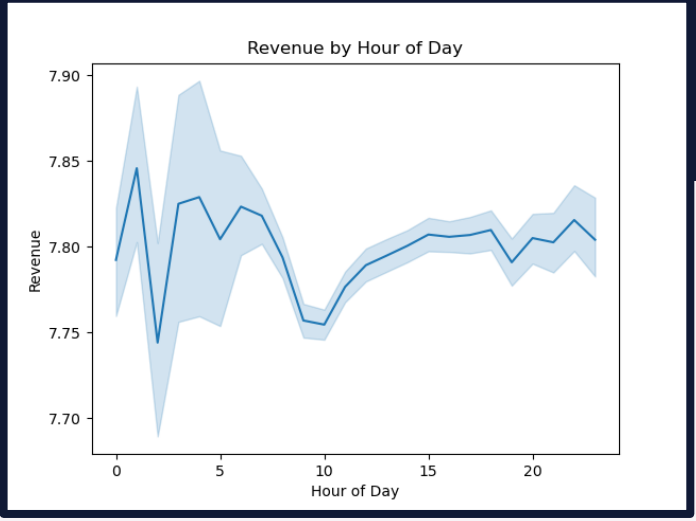


The busiest days of the week are **Saturday** (0.0) and **Sunday** (1.0).

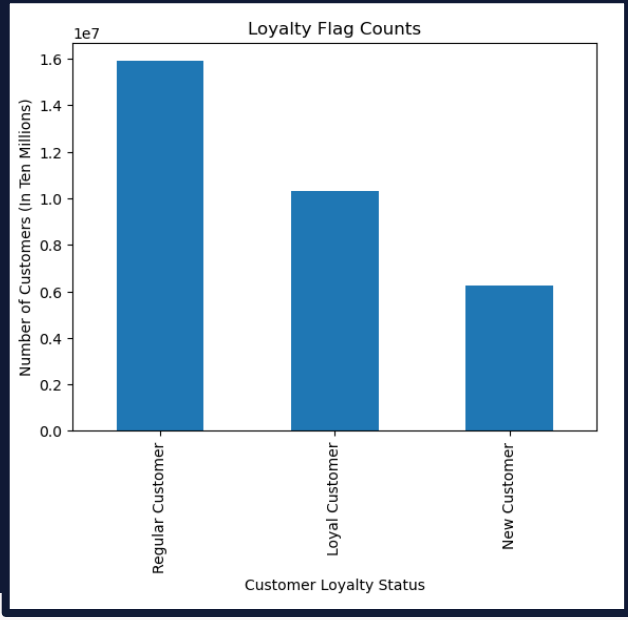
The busiest hours are 10-16 (**10AM-4PM**).

Ads should be run on weekdays before 10AM or after 4PM.

The most money spent on orders during the day is around **1AM** and **4AM**.



The shaded area of this line chart shows the possibility of values. Because of the large size of the dataset, the graph was created using a random sample.



The loyalty flag is based on max orders.

New Customer:
10 or less max orders

Regular Customer:
10 to 40 max orders

Loyal Customer:
Over 40 max orders

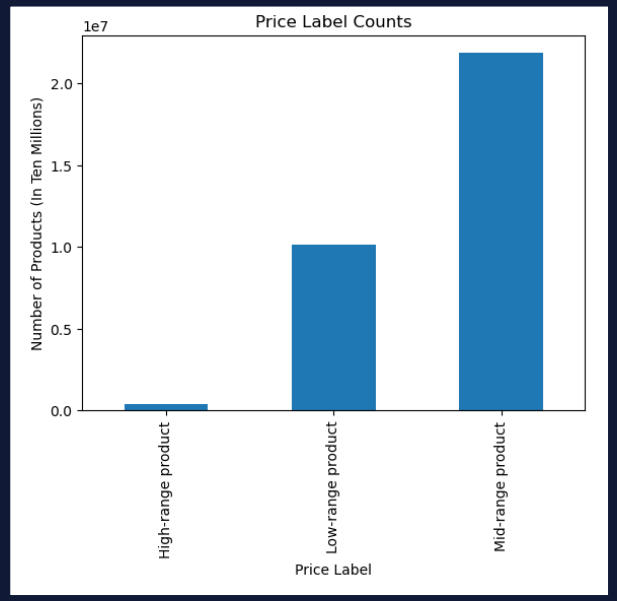
49% of all customers are Regular Customers.

Customer frequency is based on the average time since a last order (Frequent: 10 days or less, Average: 10 to 20 days, Non-Frequent: over 20 days).

Order Frequency Based on Customer Loyalty			
Loyalty Flag	**Frequent Customer**	**Non-Frequent Customer**	**Average Customer**
Loyal Customer	**699,445**	5,336,152	4,258,430
New Customer	205,506	4,038,836	563,613
Regular Customer	334,246	14,073,920	1,483,341

Loyal Customers make up **56%** of all Frequent Customers, but Regular Customers make up **60%** of Non-Frequent Customers.

Products, Regions, and Departments



The Price Label variable groups products on price.

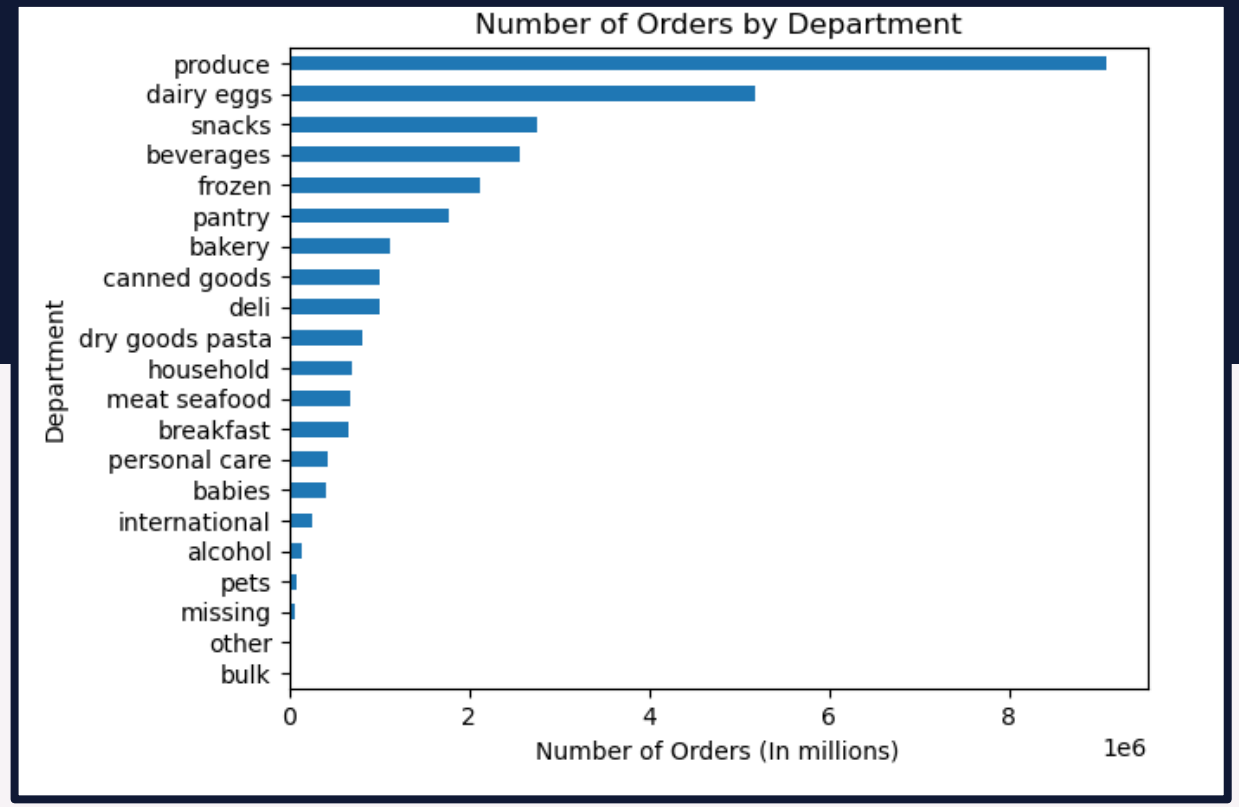
High-Range:
Greater than $15

Mid-Range:
Over $5 to $15

Low-Range:
$5 or less

Most products are considered Mid-Range.

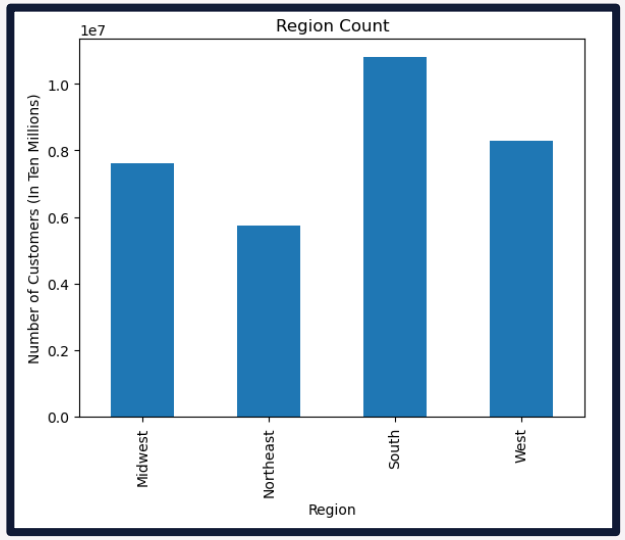
The departments with the most orders are Produce, Dairy & Eggs, Snacks, Beverages, and Frozen. Produce makes up **29%** of all orders.



33% of all customers are in the **South region**, followed by West (26%), Midwest (23%), and Northeast (18%).

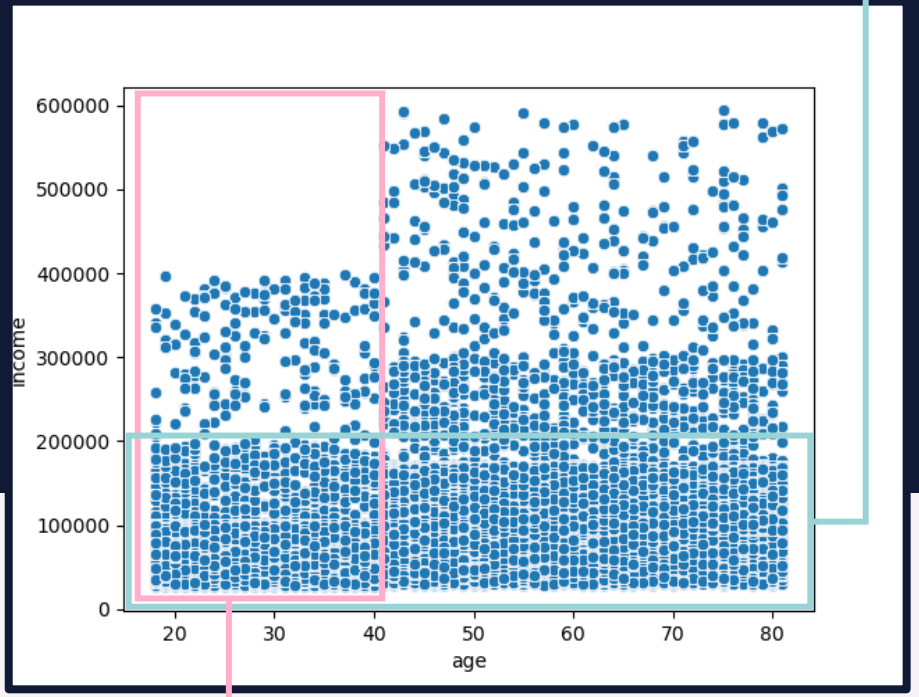
While the South region has the largest customer base, spending habits appear to be the same among all regions.

All regions had around 32% of Loyal Customers, 19% of New Customers, and 49% of Regular Customers.



Age, Family Status, And Income

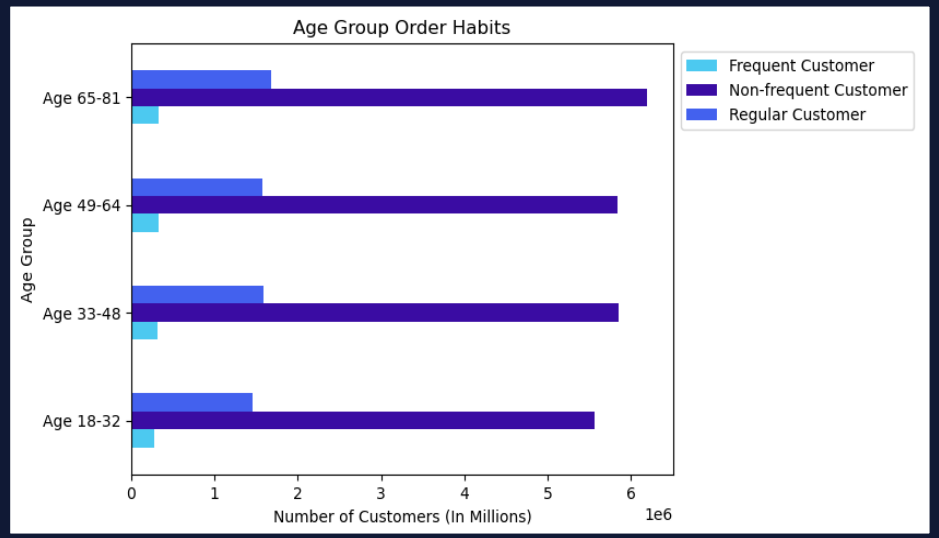
Most people across all ages have an income of less than $200,000.



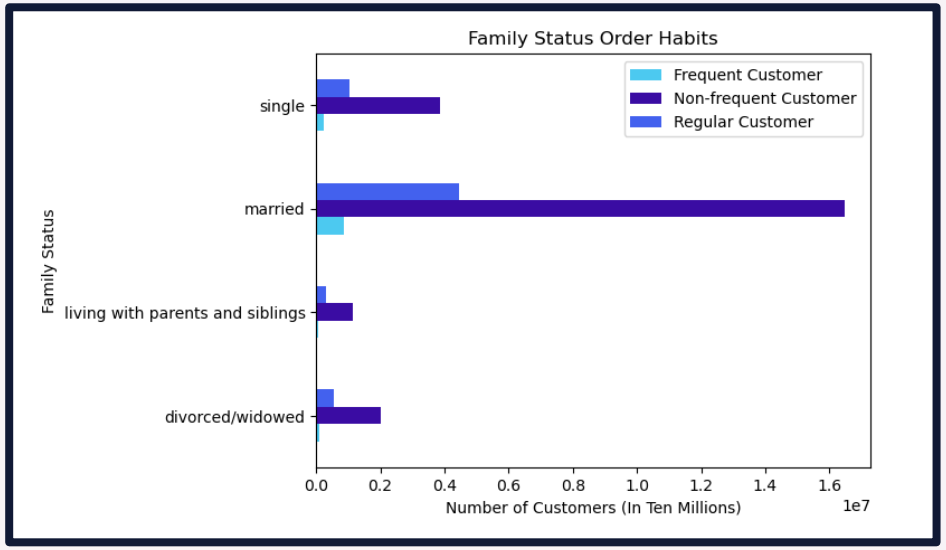
No one under age 40 is earning more than $400,000.

Those 40 years and older have more people earning an income of $300,000 than those under 40.

$200,000 and $300,000 seem like high income figures, and while there are some high income points it doesn't seem like the numbers are skewed by a few outliers.



There didn't seem to be a relationship between age, family status, and ordering habits.
The number of customers in each frequency category seem evenly distributed amongst the age groups.

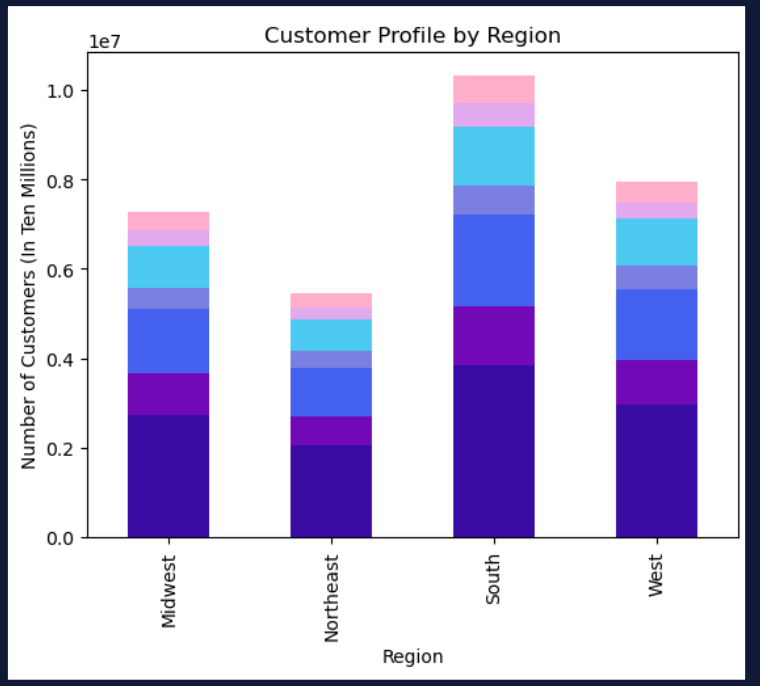


When looking at family status, customers that are married make up the most customers in each frequency category.
Married customers have a wide age range though, from 18-81.

Customer Profiles

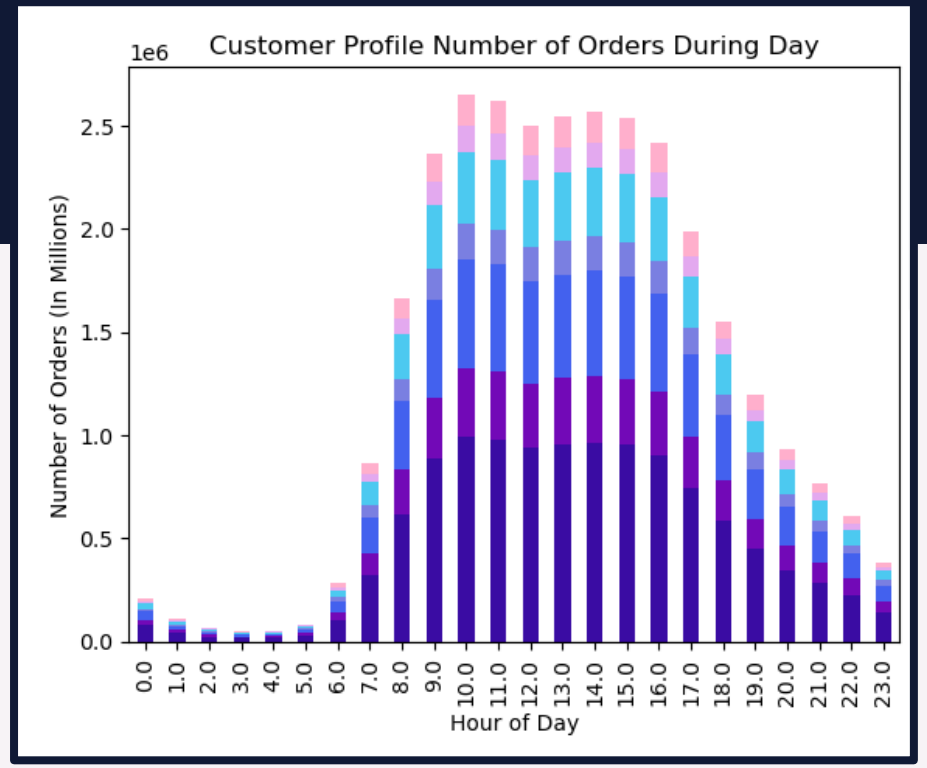
Based on age, family status, and number of dependants we can classify the customers into different demographic categories:

- **Middle married dependants** - Age 33-64, married with dependants
- **Middle single no dependants** - Age 33-64, single with no dependants
- **Older married dependants** - Age 65-81, married with dependants
- **Older single no dependants** - Age 65-81, single (single/divorced/widowed) with no dependants
- **Younger married dependants** - Age 18-32, married with dependants
- **Younger single dependants** - Age 18-32, living with parents and siblings, no dependants
- **Younger single no dependants** - Age 18-32, single no dependants



All regions have a similar amount of the largest demographic, the middle married dependants profile.

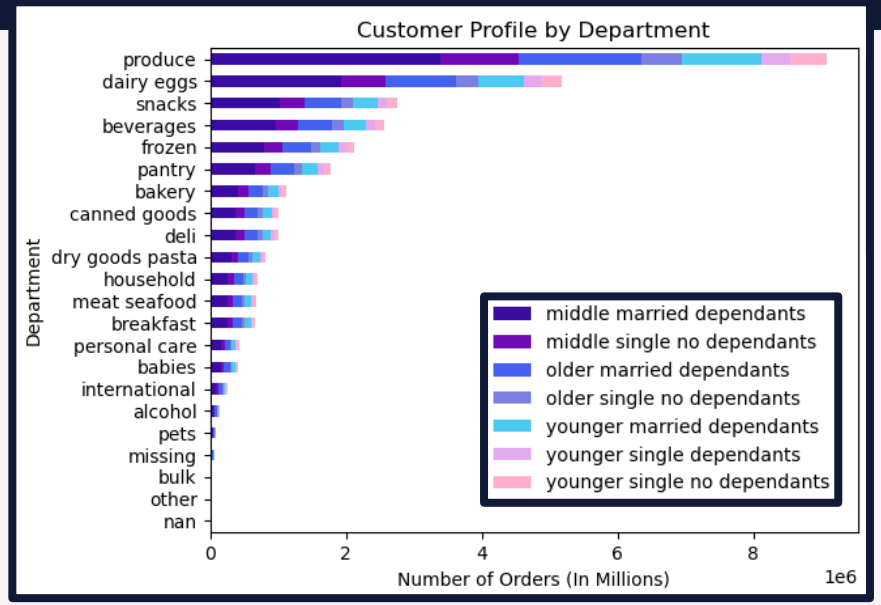
The South Region is not only the largest region, but it also has the most customers in every customer profile.



All profiles follow the general trend of ordering the most around hour 10, but also have a significant number of orders placed at 9 to 16 (9AM - 4PM).

The middle age married dependants group continue to make up the bulk of all orders. There are also a fair amount of produce orders from older and younger married customers with dependants.

All groups make up a similar amount purchases from the Babies department, regardless of dependants.



Conclusion And Recommendations

Time of Day/Week

The busiest times are weekends from 10AM-4PM. If marketing wants to run ads during slow periods, it should be during the weekdays before 10AM and after 4PM.

Regions & Family Status

Focus advertising on the largest part of the customer base, the South Region and those married of any age and dependants.

Promotions & Deals

Starting a price match program for in-store prices can work with coupons and deals to increase customer frequency. The current loyalty program is based on delivery drivers, not customers. A customer based one can also increase loyalty.

Departments & Products

Promote produce, dairy/eggs, snacks, beverages, and frozen department products. These are the most popular departments.

View : Full Report Python Code Back to Table of Contents

Pig E. Bank

A fictional global bank looking to increase customer retention.

Customer data is analyzed to identify factors that contribute to client loss. These factors are then modeled in a decision tree.

Project Overview

Goals:

The goal of this project is to assist the sales team of Pig E. Bank increase customer retention, by identifying factors that would cause a customer to leave.

To do this we will use data mining techniques to assess the quality of our data, clean it, and then generate some descriptive statistics.

From there I looked at demographics first (age, gender, country) and then usage (activity, number of products, tenure, balances, salary, credit card status, credit score) to see what might be top factors in leaving.

The top exit factors were then modeled in a decision tree as part of a final report.

Skills

- **Big data and data ethics**
- **Data mining**
- **Data quality assessment and cleaning**
- **Descriptive statistics**
- **Predictive analysis**
- **Time series analysis and forecasting**

Data

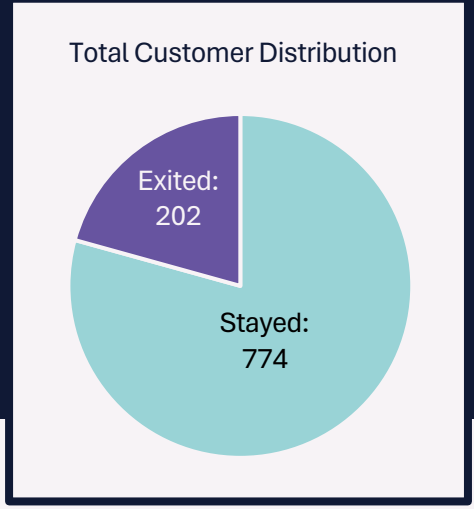
- **Customer data provided by CareerFoundry**
- **Project brief**

Tools

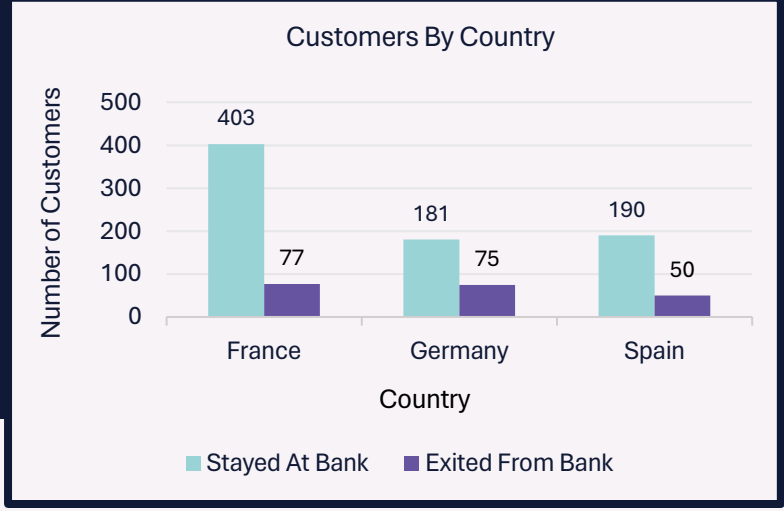
- **Microsoft Excel**
- **Microsoft Word**
- **Microsoft PowerPoint**

Customer Demographics

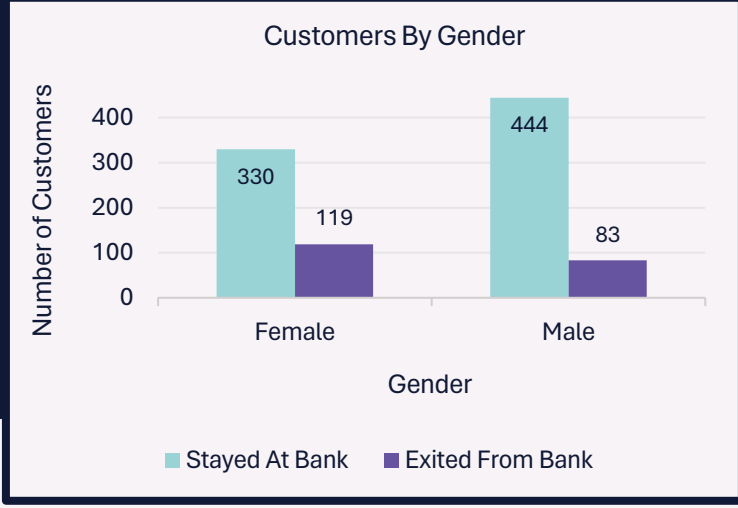
20.70% of customers have left the bank. **79.30%** remain.



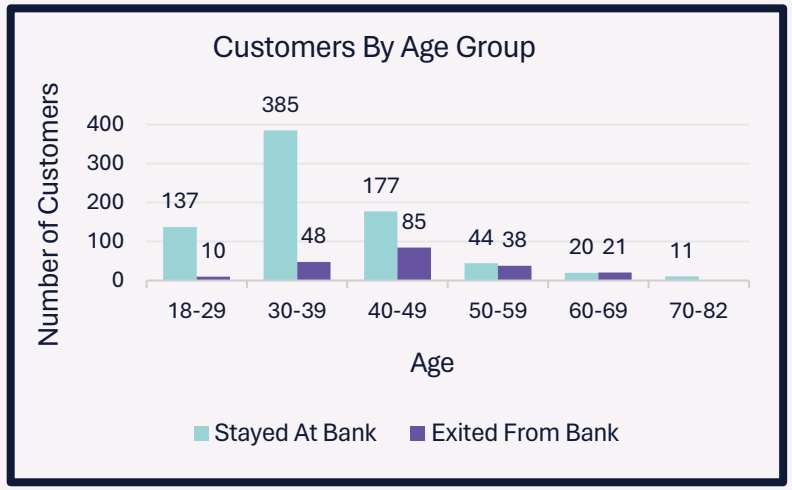
Germany has **29.30%** of customers leave compared to France's **16.04%**, even though more people from France left.

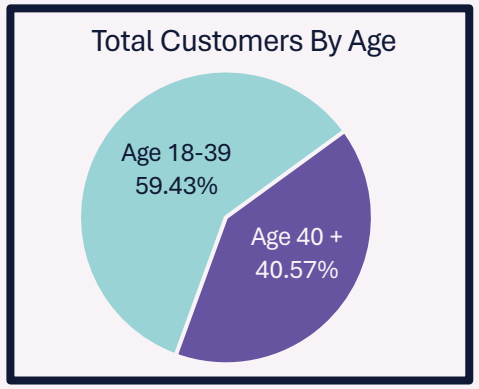


26.50% of all female customers left, compared to **15.75%** of all male customers.



Customers that left were, on average, older than those that stayed. If we combine the age groups, we can see that **10%** of customers under age 40 left, but **36.36% of those over 40 left the bank.**

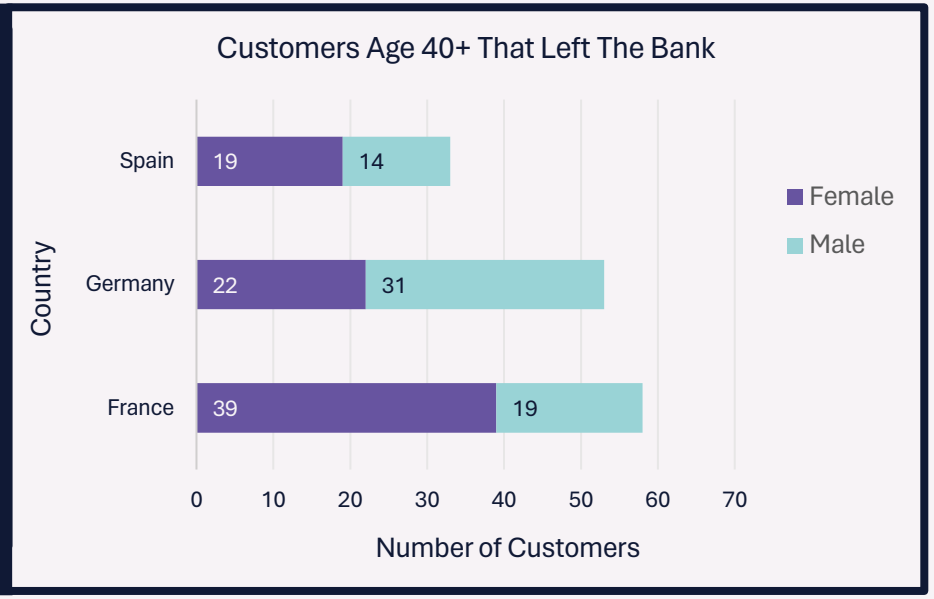




Most customers are under 40 even though that is a small age range.

Out of all these characteristics, I would consider age a factor in leaving.

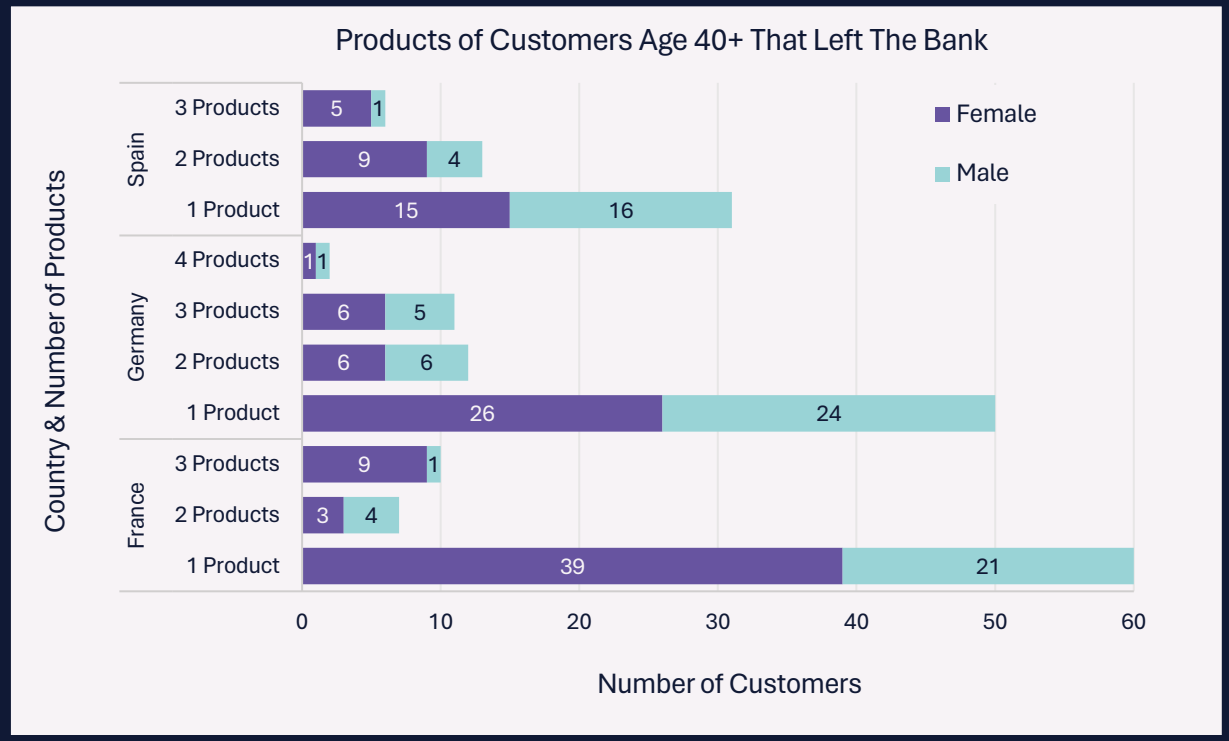
It makes up the largest percentage and is also part of the country and gender demographics.



Customer Usage

Aspects that I would **not** consider a factor for leaving are:

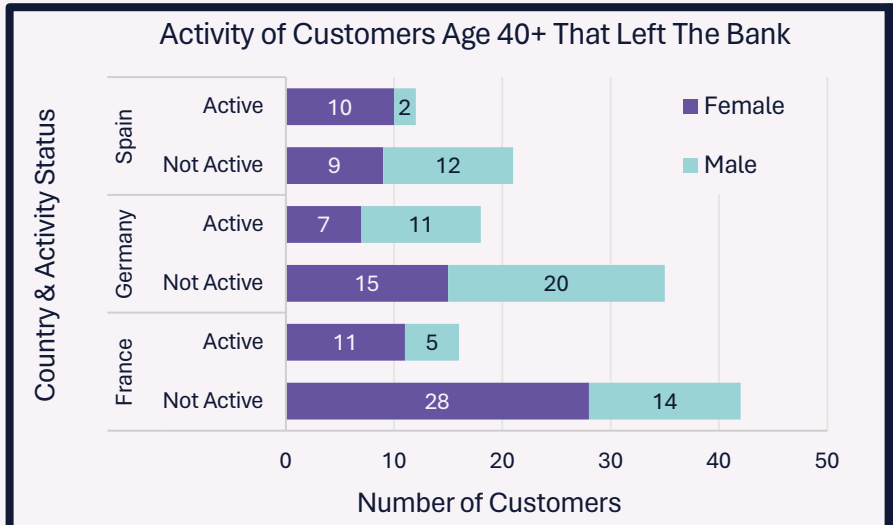
- **Credit Card Status**
  20.61% of those with a credit card left the bank, and 20.91% of those without one stayed. These ratios are similar.
- **Credit Score**
  The scores of those that left were like those that stayed and followed the same pattern.
- **Tenure**
  Customers with both short and long terms left and were relatively proportional to those that stayed.
- **Estimated Salary**
  For all salary ranges, only 0.31%-1.54% of all customers left the bank. Those that left were proportional to those that stayed.



27.87% of customers that left had only one product. This is much larger than the 12.98% of those that left with two or more products.

Across our demographics, majority of those that left also only had one product.

I would consider this an exit factor.



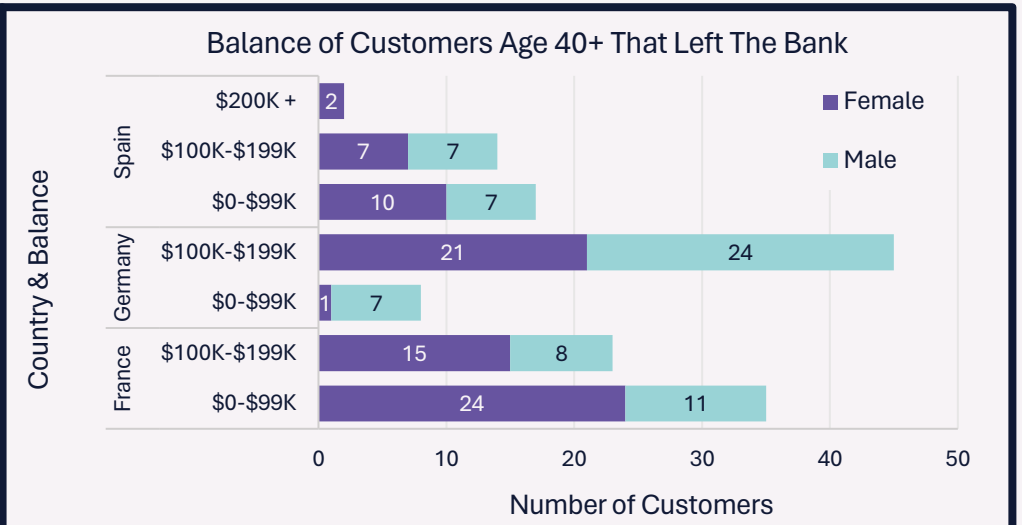
Activity status and balances were also factors in exiting the bank.

29.56% of all inactive customers left, compared to the 12.22% of active customers that left the bank.

This is 68.06% of those that left over age 40.

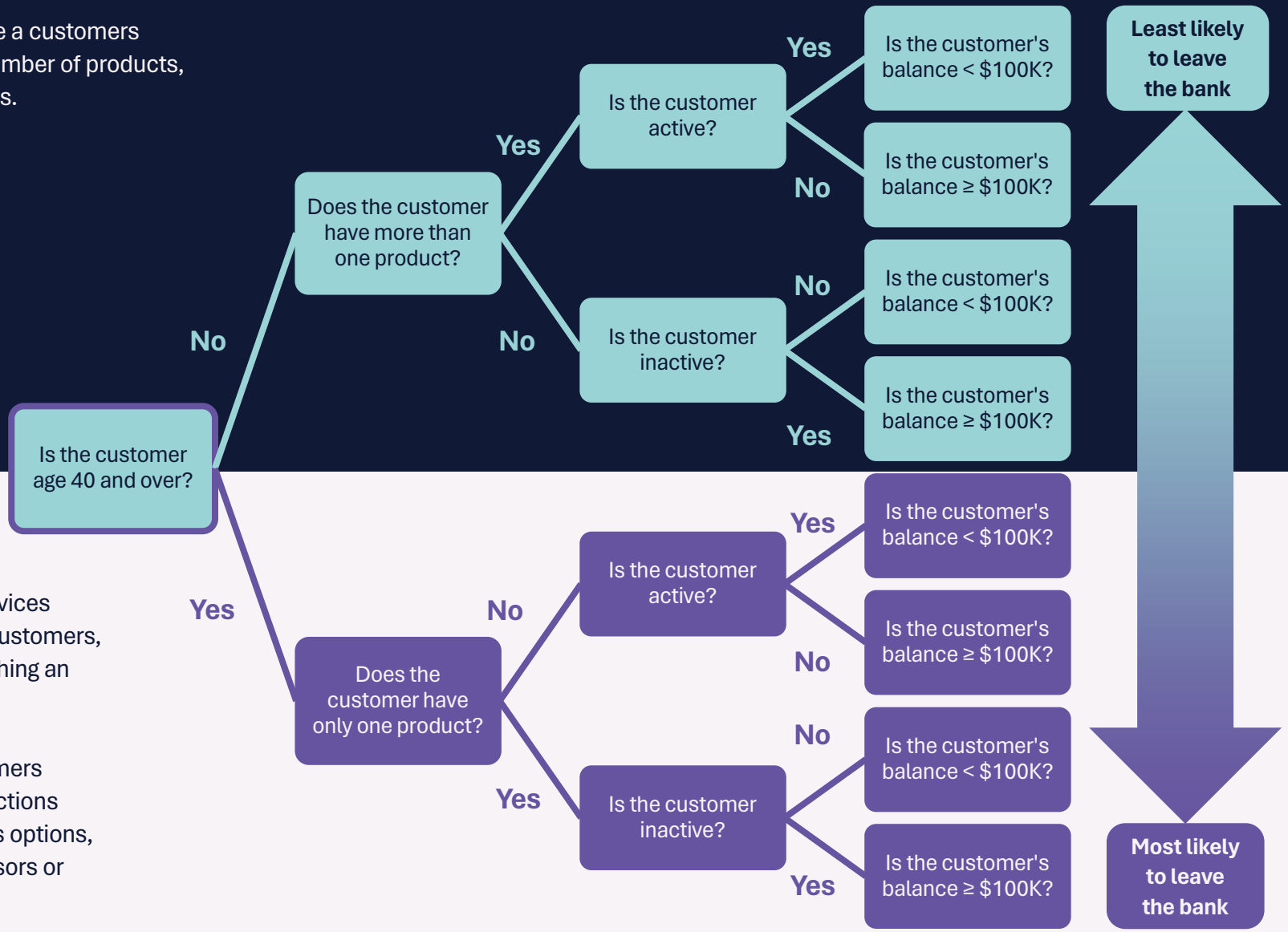
25.52% of customers with balances over $100K left, compared to the 15.99% of those that left with balances less than $100K.

58.33% of customers over the age of 40 that left had balances over $100K.



Decision Tree: Will A Customer Leave the Bank?

The top factors that indicate a customers leaving the bank are age, number of products, activity status, and balances.



Pig E. Bank's products and services may appeal more to younger customers, or advertising may not be reaching an older customer base.

In my experience, older customers tend to prefer in-person interactions at branches, safer investments options, and may need retirement advisors or long-term care planning.

Conclusion and Recommendations

Expand Investments & Savings

To retain customers with higher balances we can increase investment, savings options, and rates.

Survey Customers

Customers leaving the bank should be surveyed as to why, and in what ways Pig E. Bank can meet their needs.

Long-term Investments

Mortgages, installment loans, and even direct deposits will keep customers active. Promote long-term products.

Incentivize Cross-selling

Rewards and contests can motivate employees to cross-sell. Rate specials can increase products per customer.

View : Final Report Back to Table of Contents

UFC Historical Analysis

The UFC (Ultimate Fighting Championship) is a mixed martial arts organization that was founded in 1993.

As someone with no knowledge of the sport, I chose to analyze match data to learn about the UFC and gain insights into what contributes to a win or loss.

Project Overview

Goals:

The project was led by my own key questions, with the overall goal to learn more about the UFC and gain an understanding of what contributes to wins or losses.

Key questions I created included:

- Who has the most wins?
- What is the most common way to win?
- What factors contribute to winning?
- Why do the fighters in the Red corner have more wins than those in the Blue corner?
- Does age, height, weight, or reach create a difference in winning or losing?

To find out the answers I started with a visual analysis, then picked out strongly correlated variables to explore further.

What soon became apparent was that the scope of the project was very narrow, and that winning is more complicated than focusing on characteristics of fighters.

Skills

- **Sourcing open data**
- **Data cleaning, wrangling, and subsetting**
- **Performing exploratory visual analysis**
- **Creating geographical visualizations**
- **Supervised machine learning with linear regression**
- **Unsupervised machine learning with k-means clustering**
- **Sourcing and analyzing time-series data**
- **Creating a data dashboard**

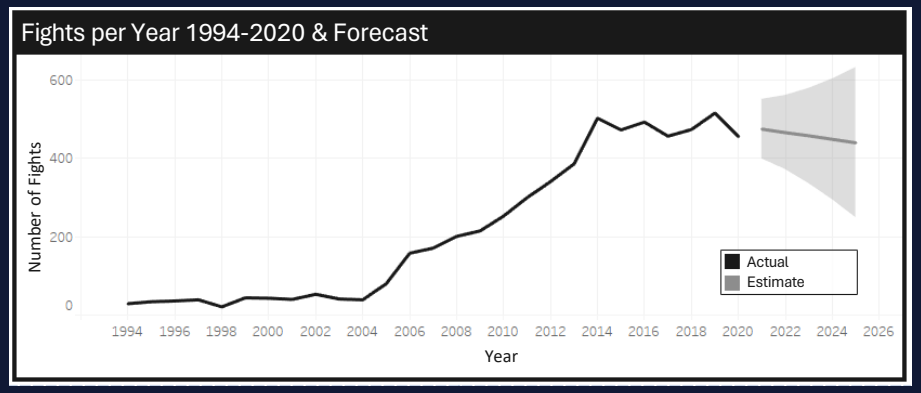
Data

- **UFC-Historical Data from 1993-2021 by Rajeev Warrier via Kaggle**
- **Data includes fights from 1994-2021 and has fighter statistics**
- **World-countries.json by Kostya via Kaggle**
- **Project brief**

Tools

- **Microsoft Excel**
- **Anaconda**
- **Jupyter Notebook**
- **Python**
- **Python libraries Pandas, NumPy, Seaborn, Matplotlib, Folium, and scikit-learn**
- **Tableau**

Methodology



My first instinct was to remove fights before 2000 during cleaning, to keep the data more current.

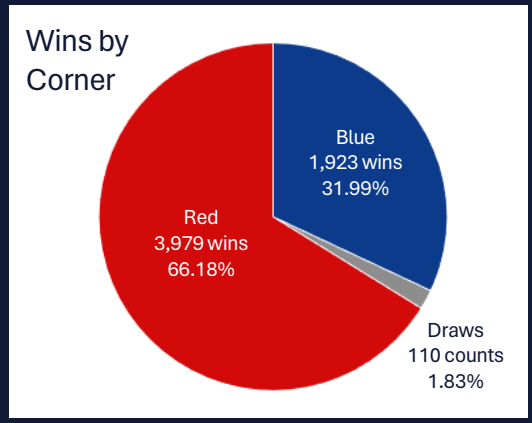
Near of the end the project I included the older data back in to show the overall growth of the UFC in terms of matches per year.

There was a large difference between Red corner wins and Blue corner wins, which led to one of my key questions.

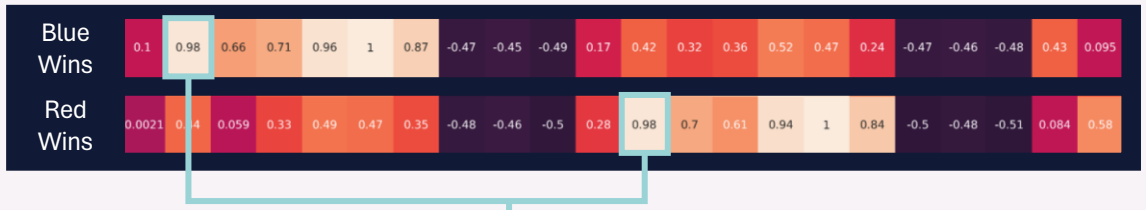
The "favored" fighter is usually in the Red corner to enter the cage second and increase entertainment of the match.

I thought this might've been a contributing factor to wins, but being "favored" usually means being experienced.

A fighter can be affected by this psychologically, but there are outside sources that influence their mental state. We don't have data on the mental state of fighters in this set.



I removed much of the knockdown, strike attempt, ground control, etc. columns because I didn't think they would be needed. Then I created a heatmap of the remaining variables.



The highest correlation with wins, at 0.98, was total rounds fought. This is what I would continue forward with and test with machine learning.

Age, height, weight, and reach did not have strong correlations with wins, ranging from 0.884 to -0.51.

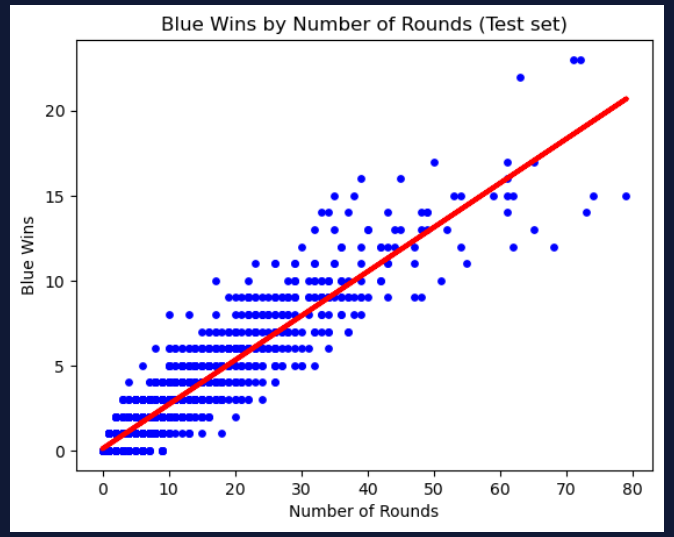
I later created a heat map with the variables I did remove to try and find another strong correlation with wins, but there were none.

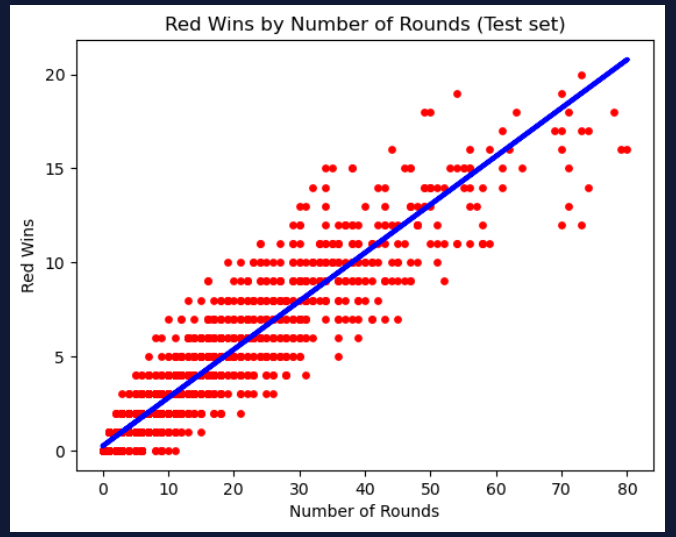
The idea that fighting more meant a person won more seemed obvious to me, because the fighter was gaining more experience and having more chances to win.

It was after testing that I started thinking of expanding the project.

The data did not have viewer or fan metrics, so I continued to answer my other questions and started looking at individual careers instead.

Linear Regression Analysis





Based on the strong correlation between wins and total rounds fought, I formed the following hypothesis:

If a fighter is in more rounds, they will have more wins.

This was tested with a supervised machine learning technique, linear regression. The r-squared value is around 0.87, showing it to be a good fit for the data. **87%** of the variances in wins can be explained by the number of rounds fought.

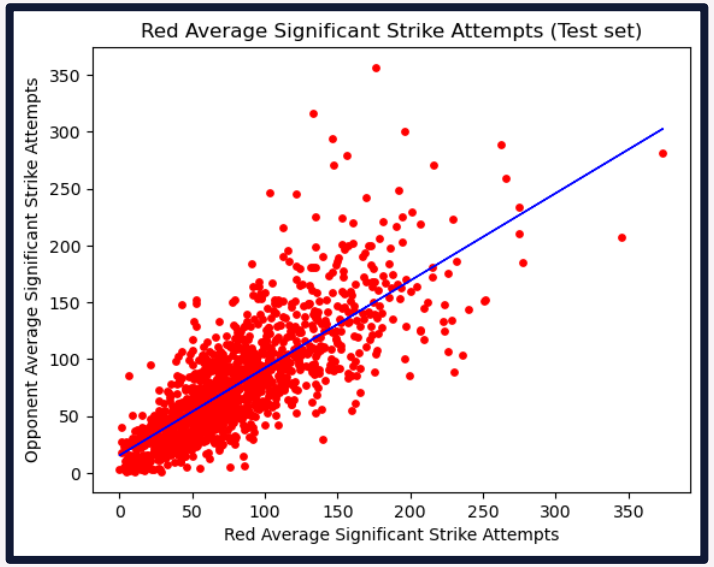
The root mean squared error was 1.41, a small number indicating that this analysis was a good fit for the data. It is also of relative size to the number of data points.

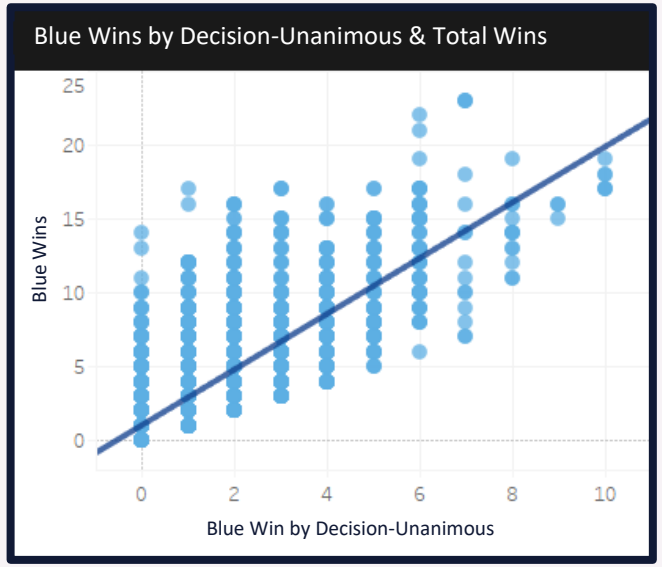
I also ran the same test on significant strike attempts and opponent significant strike attempts to look at other aspects.

These were strong correlated, and my hypothesis was that if a fighter is making more attempts, their opponent will respond in turn.

The root mean squared error was quite large though, and Red strikes had a negative R-squared score of **-0.35**.

This indicated that a horizontal line would be a better fit for the data.





Even looking at types of wins, like **Decision – Unanimous**, resulted in a moderate **60%** of explained variances.

This is an example of the Blue chart, and the Red chart has a similar shape and variance percentage.

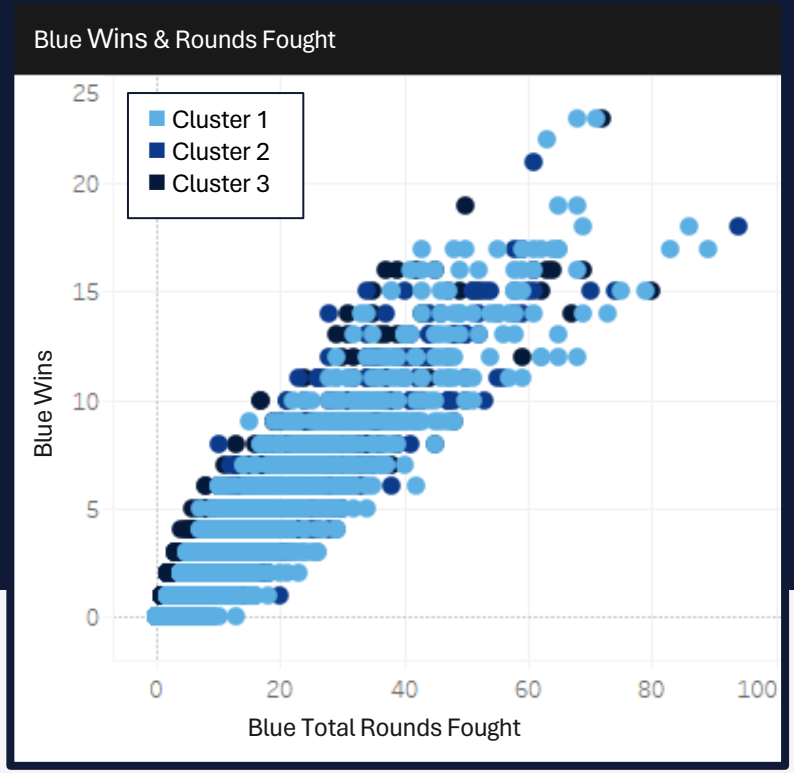
Decision – Unanimous wins are the most common and imply that fighters fought the full rounds of the match.

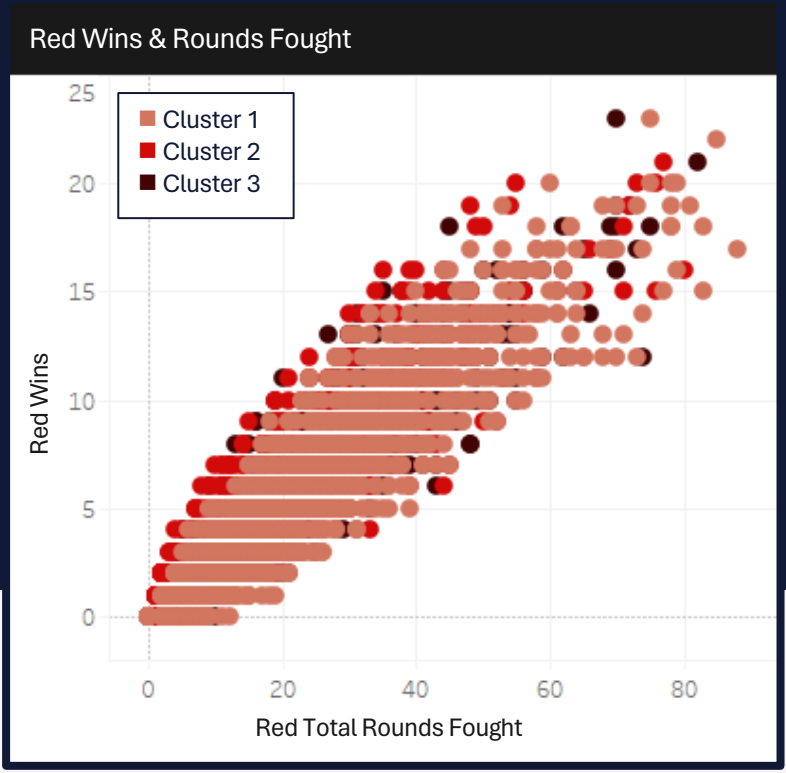
K-means Clustering Analysis

The next test was a k-means clustering, continuing to look at wins and total rounds fought.

The clusters did not show up as divided vertically or horizontally, but instead show as stacked on top of each other.

This may be because the data is arranged by matches, so every fighter's total wins and rounds fought are recorded at each match .





There didn't seem to be one factor that influenced the clusters. This is better illustrated by the clusters' descriptive statistics.

Cluster 1 in both Blue and Red corners has highest average rounds, but Cluster 3 has the lowest rounds, highest average wins, and age.

This suggests that the group with lower rounds has more wins, contrasting our previous hypothesis.

Blue Clusters	Average Total Rounds Fought	Average Wins	Average Age
Cluster 1	***10.21***	2.63	29.13
Cluster 2	9.68	2.47	29.37
Cluster 3	8.58	***2.88***	***29.60***

Red Clusters	Average Total Rounds Fought	Average Wins	Average Age
Cluster 1	***15.03***	3.77	29.44
Cluster 2	11.87	3.74	29.79
Cluster 3	14.65	***4.12***	***29.84***

UFC Fighters and Matches

While fighting in more rounds may contribute to more wins, this project has shown me that wins are multifaceted and there's no way to guarantee a prediction. Humans are unpredictable and the UFC fighters should be celebrated as the sport continues to grow and thrive.

Top records from our data include:

Donald Cerrone
Highest total wins at **23,** and the **most matches at 36**

Charles Oliveira
Most submission wins at **14,** and **most Blue corner wins** at **11**

Derrick Lewis
Most KO/TKO wins at **12,** 75% of total wins

Anderson Silva
Longest win streak with **16 wins**

Ron van Clief
Oldest fighter at **51,** with 1 match in 1994

Jessica Andrade
11 wins, most of all the women's weight classes

Randy Couture
Second oldest fighter at **47,** with 16 wins.

Jim Miller
Most matches at **36,** with 21 wins

Location of Fights From 1994-2021

Our data ends in March 2021. Many of these top records could've been passed since then.

A good example of this is **Jim Miller**, who now has the **most fights at 46** and the **most wins at 26** (as of January 2024).

I was not a UFC, or even much of a sports fan before this analysis project, but I'm glad to have learned about something new and surprisingly complex.

While the results of the analysis may have been predictable, I was able to use machine learning techniques that were new to me. I also gained experience picking my own data and developing my own questions and project objectives.

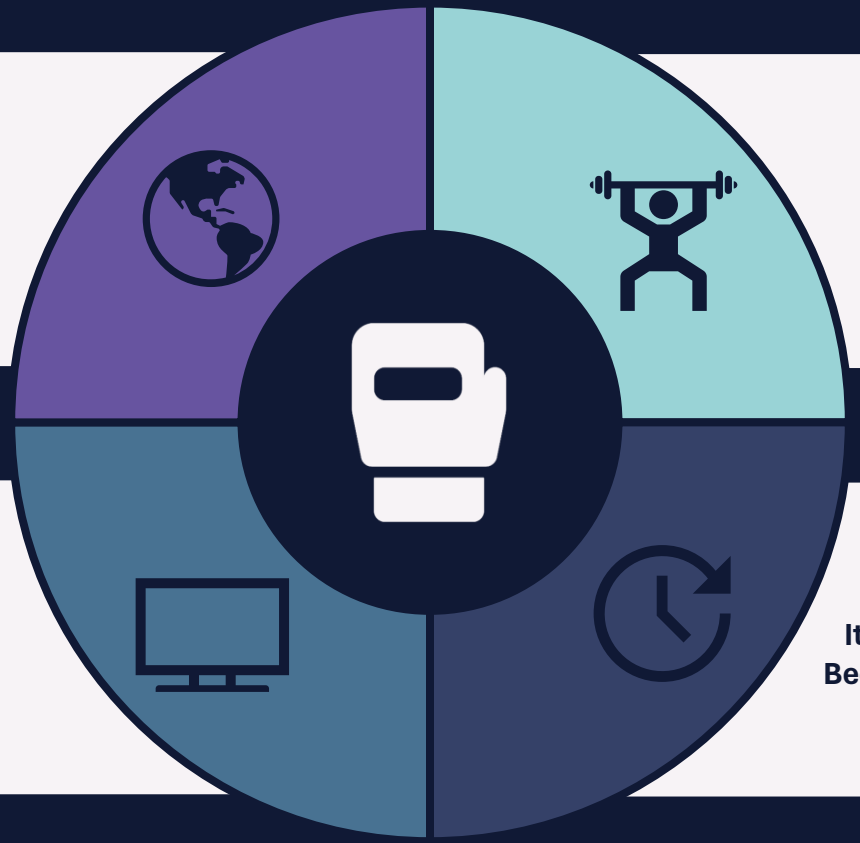
Recommendations and Next Steps

Expanding Globally

The UFC should continue to host fights internationally, to grow their fanbase and recruit new fighters.

Care for Injuries

Because of the nature of the sport, injuries can be severe. To increase career length and fighter wellbeing, the full effects of injuries should be researched.



Viewership Exploration

This data didn't have viewership or fan metrics, but I would be interested in analyzing that data to look at increasing the fanbase and interactivity.

Future Data

It would be interesting to see fight data after 2020. Because of the structure of the data, it may be more suited to a different machine learning model.

View : Tableau Visualization Python Code Back to Table of Contents

Thank You!

Back to Top

Contact

Credits : Three Data illustrations, disease illustration, health illustration, people illustrations, two business illustrations, and money illustration by Storyset. LinkedIn icon, save icon, and link icon created by Freepik – Flaticon. MMA icon created by Radhe Icon – Flaticon.

Credits: GiHub icon, spreadsheet icon, doc file icon, pdf icon, programming language icon, and business class icon created by Pixel perfect – Flaticon. Email icon created by Unicornlabs – Flaticon. UI icon created by Iyahicon – Flaticon.