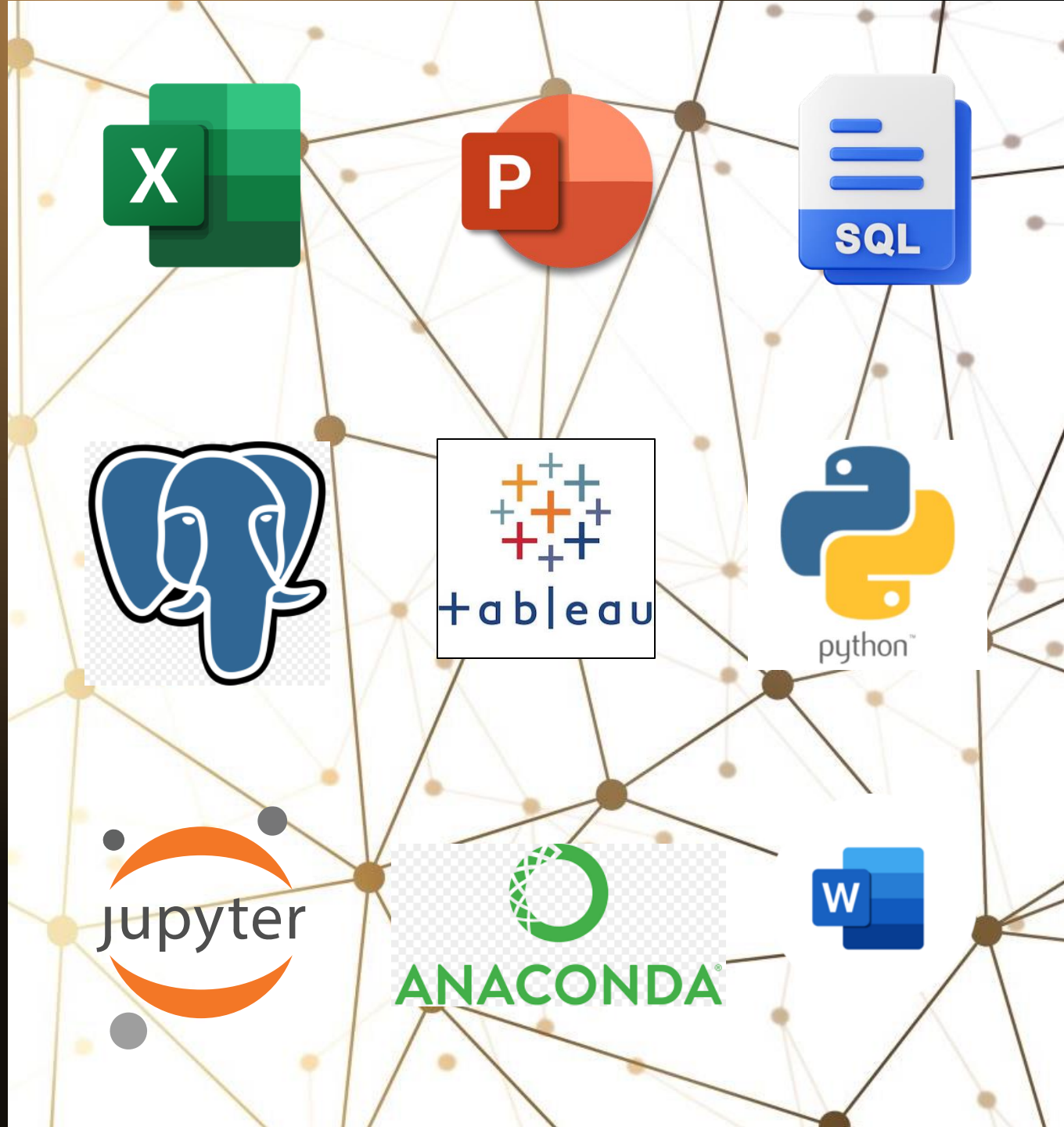
Evan Carr

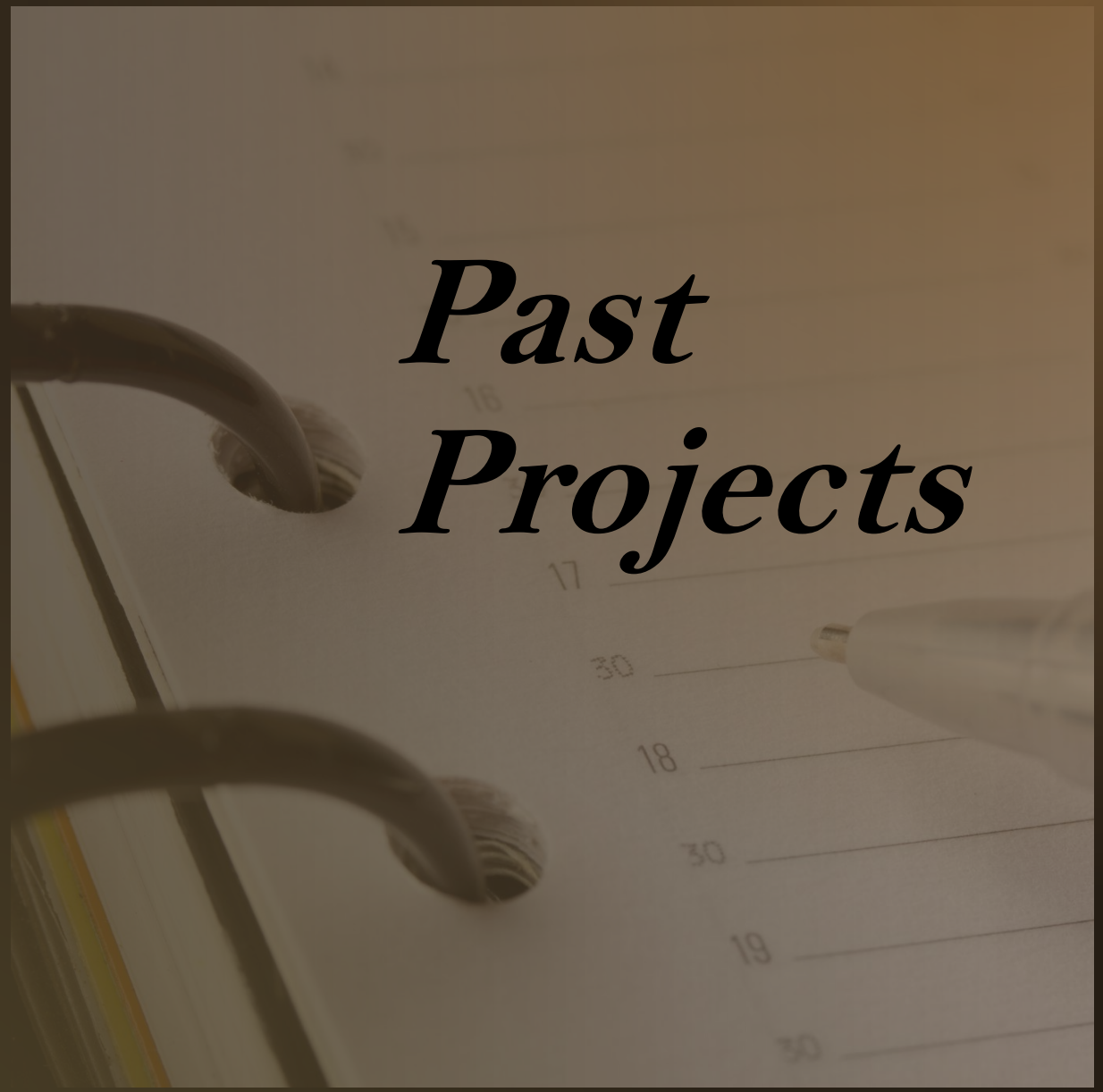
DATA ANALYST PORTFOLIO 2024



About me

Hello!

I’m Evan and I am a data analyst who enjoys the ever-changing world of data! From solving business related problems to deep diving into complex data sets to reveal actionable business insights, I am a quick learning and ambitious analyst who has the necessary expertise in multiple software programs to help clients with transforming and translating raw data across vast industries into clear cut narratives in an organized and transparent manner.



Past Projects

1. GameCo	• Analysis of videogame store
2. Influenza Vaccine	• Preparing medical staff for upcoming influenza season
3. Rockbuster Stealth	• Providing insights regarding video rental stores breakthrough into streaming space
4. Instacart Basket	• Online grocery store's sales patterns through Python coding
5. Pig E. Bank	• Predictive analysis of Customers in a finance space
6. **New York AirBnB Analysis**	• Exploratory analysis of listings in NYC for customer & owner insights

GameCo

Objective:

GameCo wants to use data to inform the development of new games. As such, you've been asked to perform a descriptive analysis of a video game data set to foster a better understanding of how GameCo's new games might fare in the market.

Key Questions

Are certain types of games more popular than others?

What other publishers will likely be the main competitors in certain markets?

Have any games decreased or increased in popularity over time?

How have their sales figures varied between geographic regions over time

Tools Used

Excel for analysis

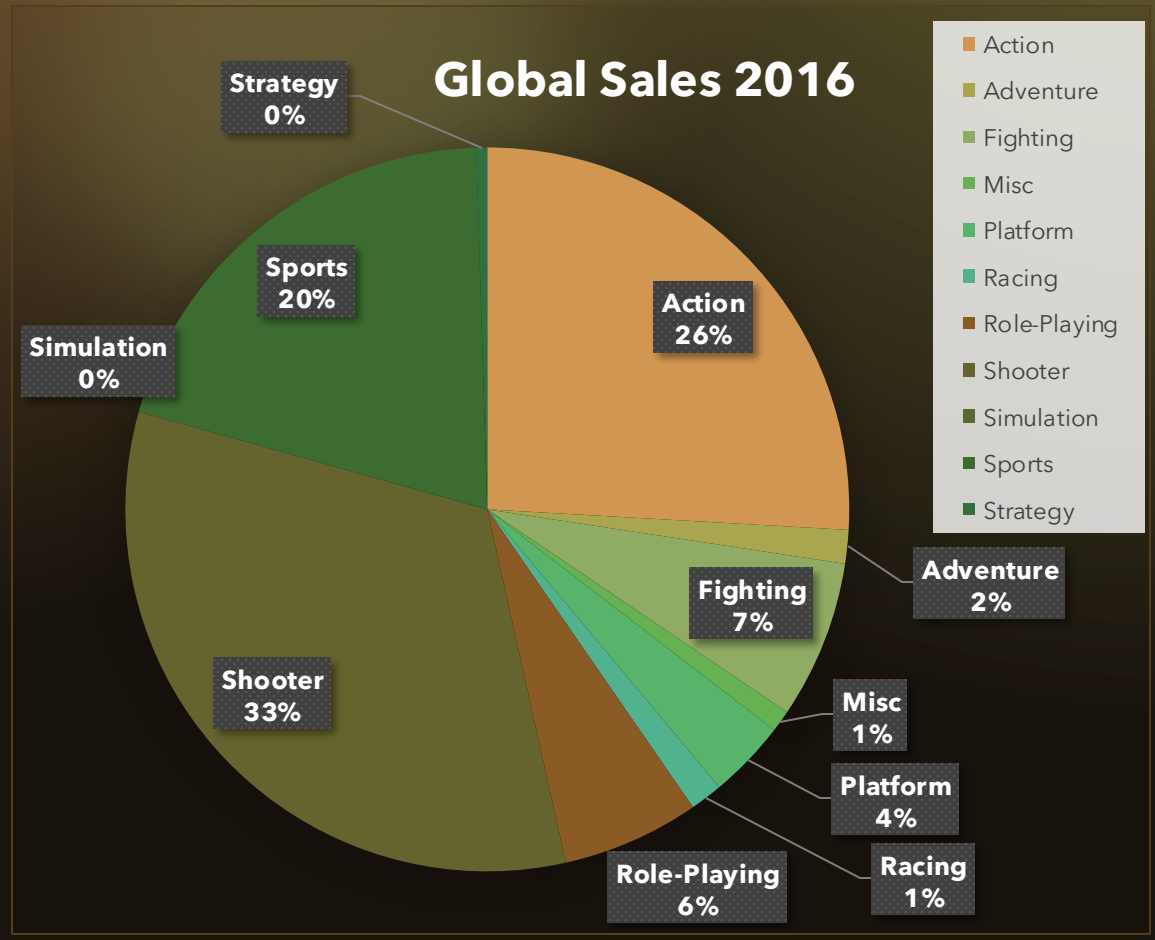
Powerpoint for presentation

Data Overview

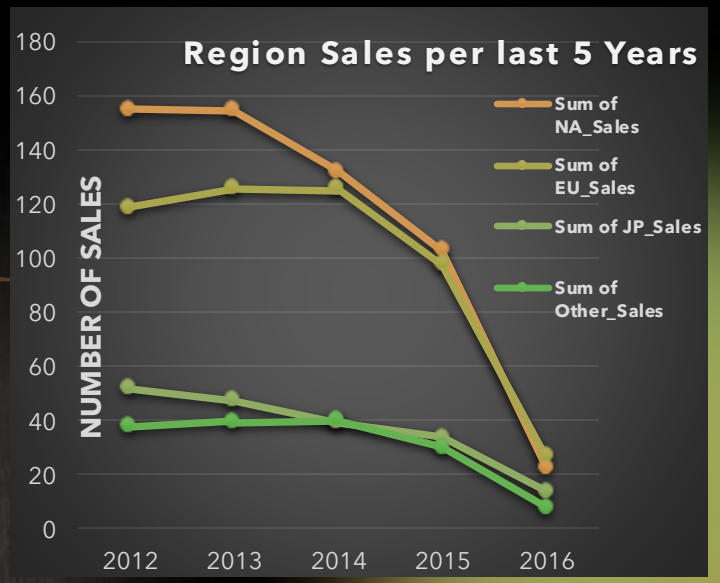
- Dataset was compromised of videogames that sold more than 10,000 copies that span across different genres, platforms, and publishers.
- Data was drawn from VGChartz website
- *Skills Used:*
  - Data Cleaning and Summarizing
  - Pivot Tables
  - Descriptives Analysis and Mapping
  - Visualization Charts

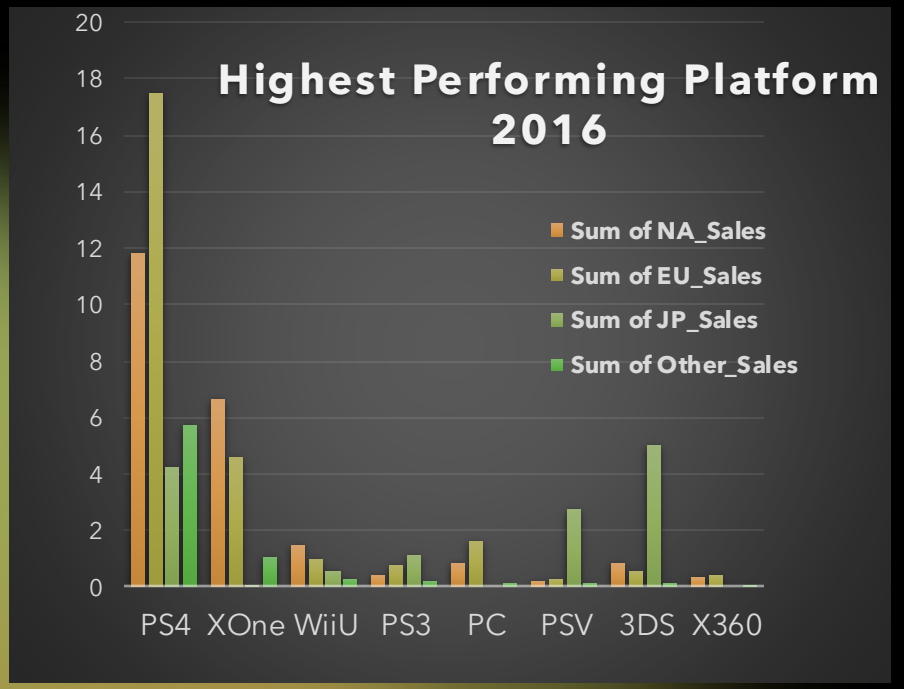
- *Process:*
  - Data was cleaned by removing duplicates, empty rows, columns not useful to analysis, imputing averages, and converting into a uniform format
  - Pivot tables were used to group info into new variables for deeper insights
  - Visualizations were then created based off sales by genre, region, and platform

- I found the most popular genres across the globe in 2016 and then narrowed my search regionally

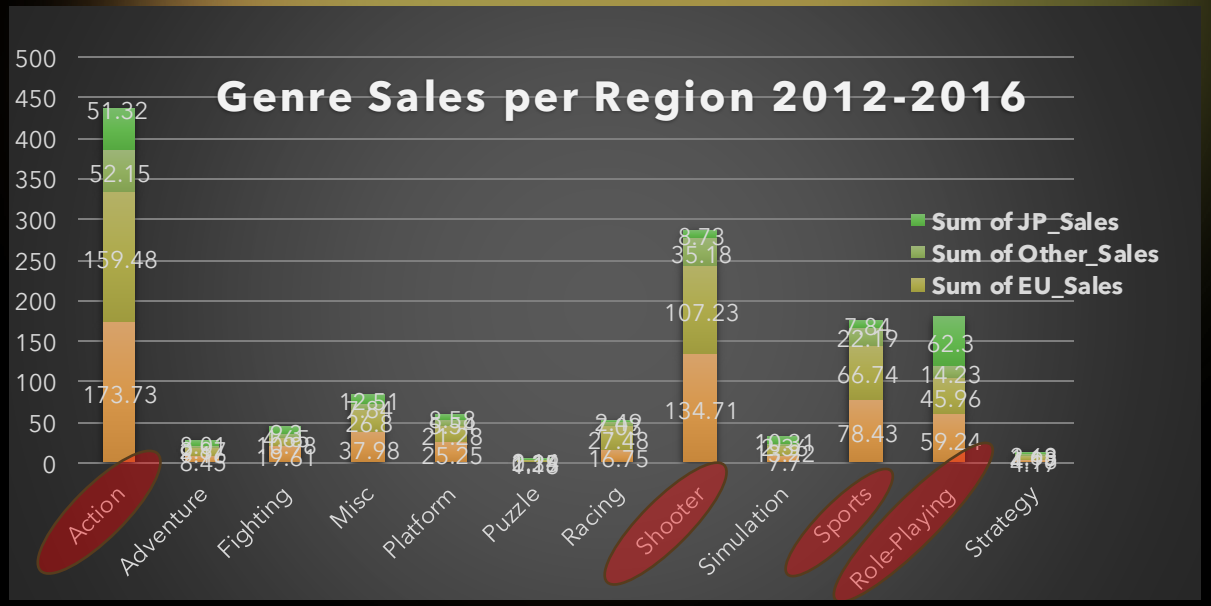


Data Results





- After inspection of GameCo's history I found that data consisted of sales across 4 regions during the years 2007-2016
  - Japan, Europe, North America, and Other Regions
- I then performed a deeper dive into what genres generated the most revenue and what gaming platform was most popular in each region
  - PS4, Xbox One, 3DS

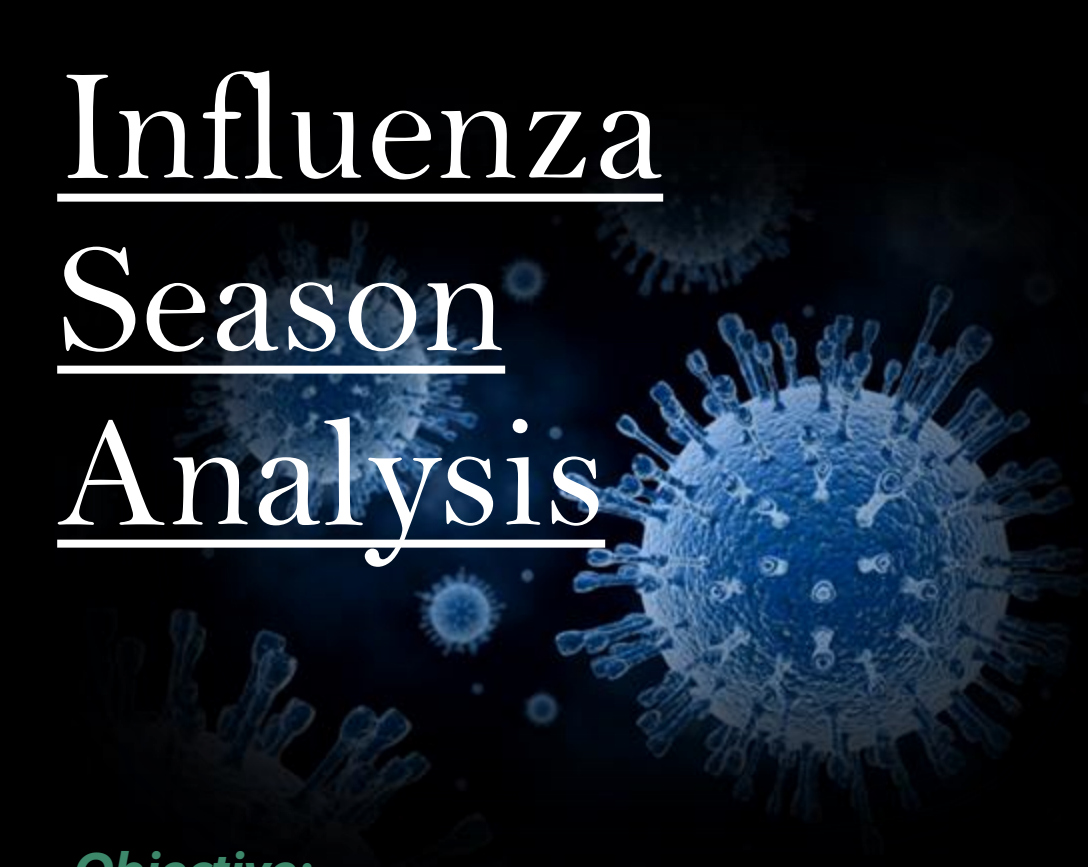


Split	Split Budget Based off Region Simplicity and Allocate specialized approach to maximize each region •50% to North America and Europe •25% to Japan •25% to "Other" regions
Focus on	Focus on top 3 grossing genres •Action, Shooter, Sports get majority of marketing budget (75%) for NA, EU, and Other regions •Japan receives other 25% of budget for Roll-Playing and Action genres
Narrow down	Narrow down which ad space to best utilize •PS4 and XOne are most popular and newest game consoles out •Both platforms have large digital market à access to wide variety of AD space

Results

- ✓ Percentages were allocated to each region as a matter of my own opinion and should be adjusted in further analysis per business requirements
- ✓ Once focused on the top grossing genres I recommend that GameCo continue to allocate resources to Action, Shooter, and Sports games in EU and NA
- ✓ By focusing on the platforms that prove to generate the most revenue GameCo's profit should continually grow by utilizing ad space

Influenza Season Analysis



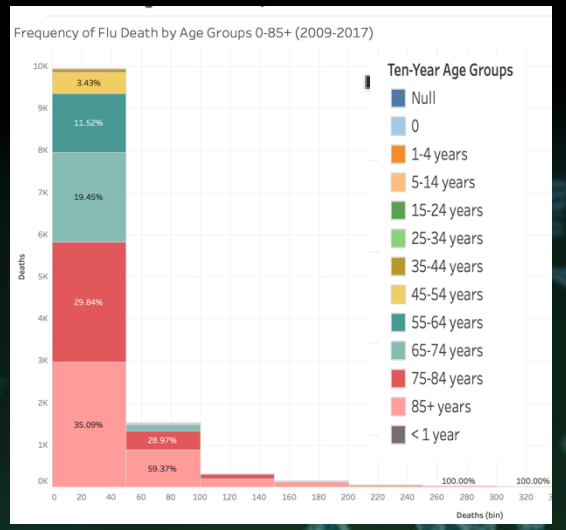
Objective:

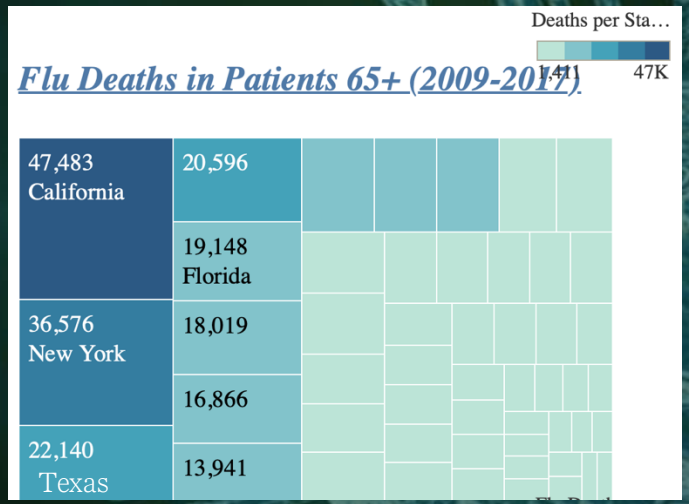
Devise a plan that utilizes additional (but limited) staffing agency resources to states most in need during the influenza season cold weather spike.

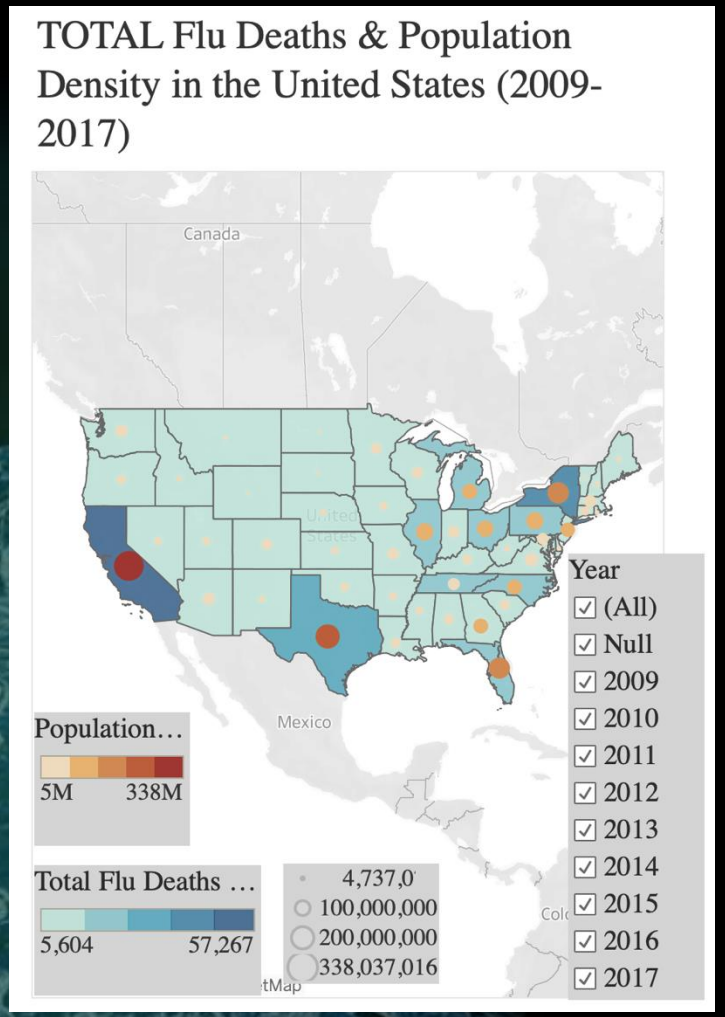
- **Data Overview**
  - Flu deaths by geography provided by CDC trusted website
  - Population data drawn from US Census Bureau
    - Consists of data by age, time, and gender
- **Skills**
  1. Data cleaning, transforming, and integration
  2. Statistical hypothesis testing
  3. Visual Analysis
  4. Forecasting
  5. Presentation expertise
- **Tools**
  1. Excel
  2. Word
  3. Tableau

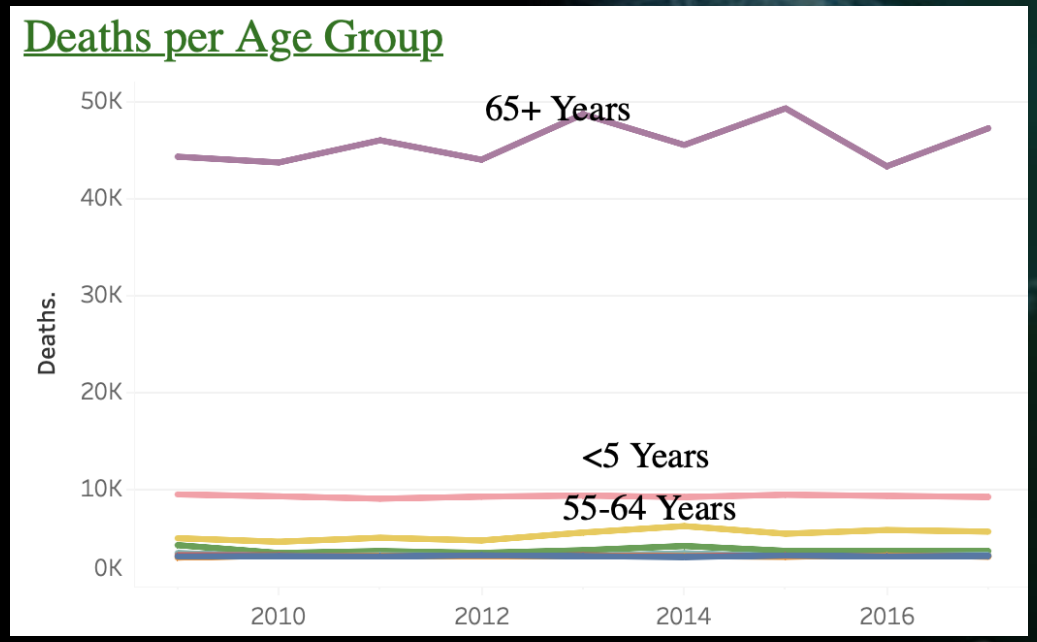
Analysis

- I integrated and cleaned data from two separate sources then conducted statistical analysis to find the most at risk age groups (65+)





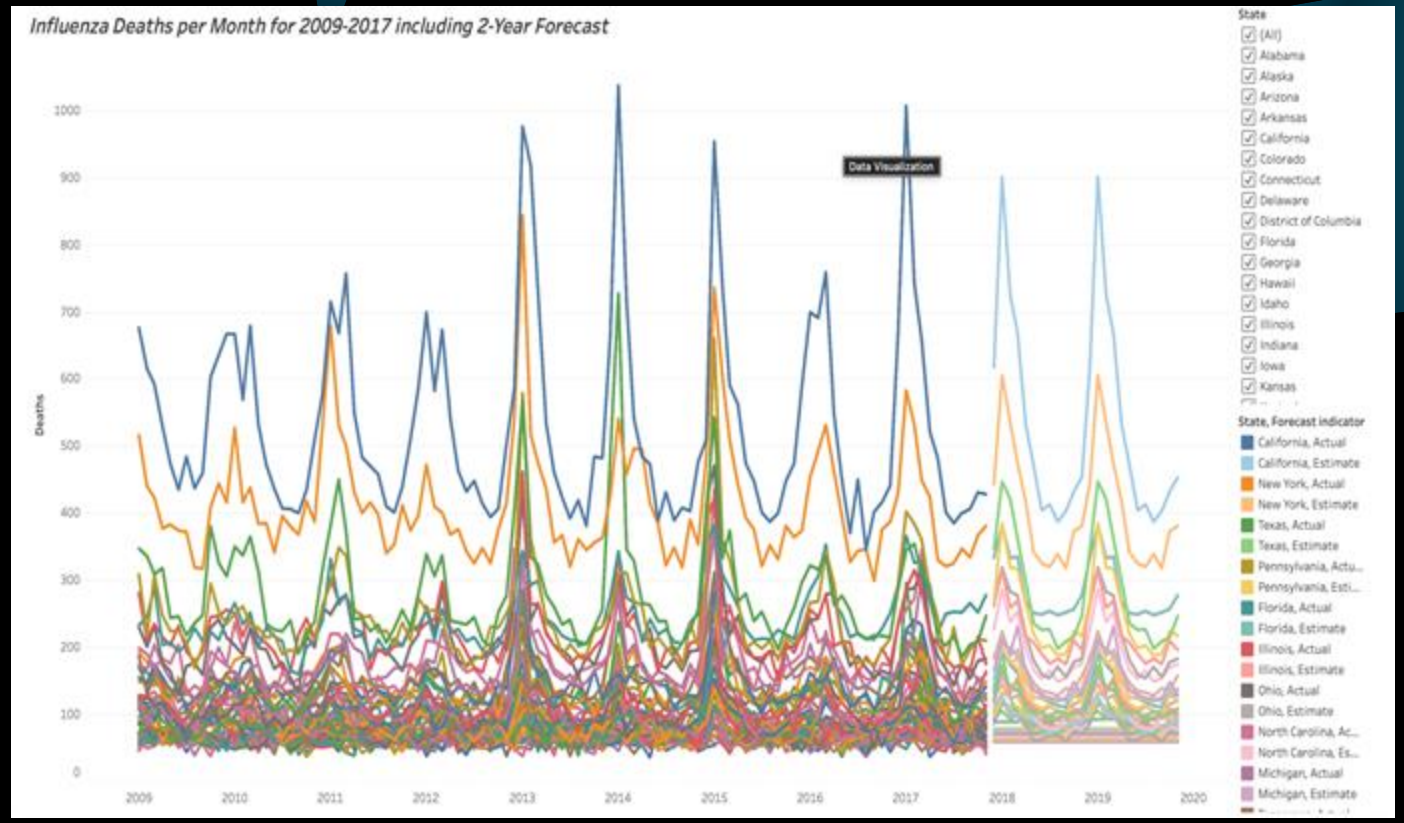


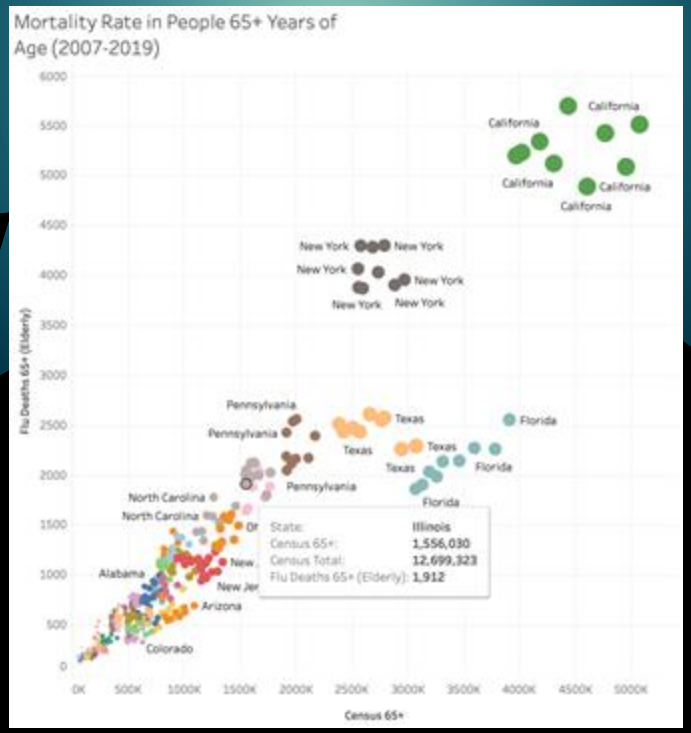


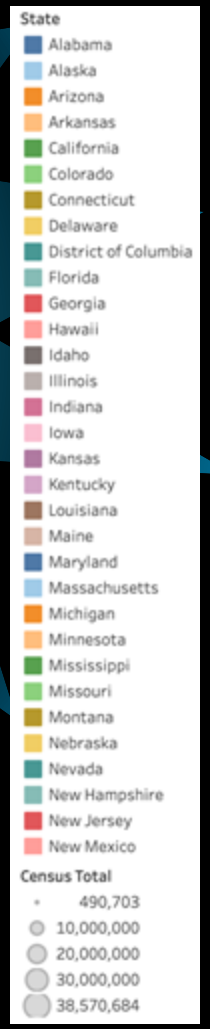
- I then determined which states these deaths were occurring in the most
  - California, New York and Texas

Analysis Continued

- After determining which demographic and area was most affected by the flu it was determined that the more dense the population the higher the risk of spread
- Seasonality played major role when breaking down flu deaths by month and I found that spread was more likely to occur in the colder months
- 2-Year forecast was implemented to determine how staff should prepare for future Influenza season







Results & Recommendations

Project Deliverables:

- Interim reports to keep stakeholders updated on analysis and results
- Tableau Storyboard
- Screen-casted presentation of Tableau Storyboard

People 65+ years of age are the highest risk of dying from the Flu virus

The most densely populated states tend to propagate high deaths and general spread of Flu virus

- California, New York, Texas

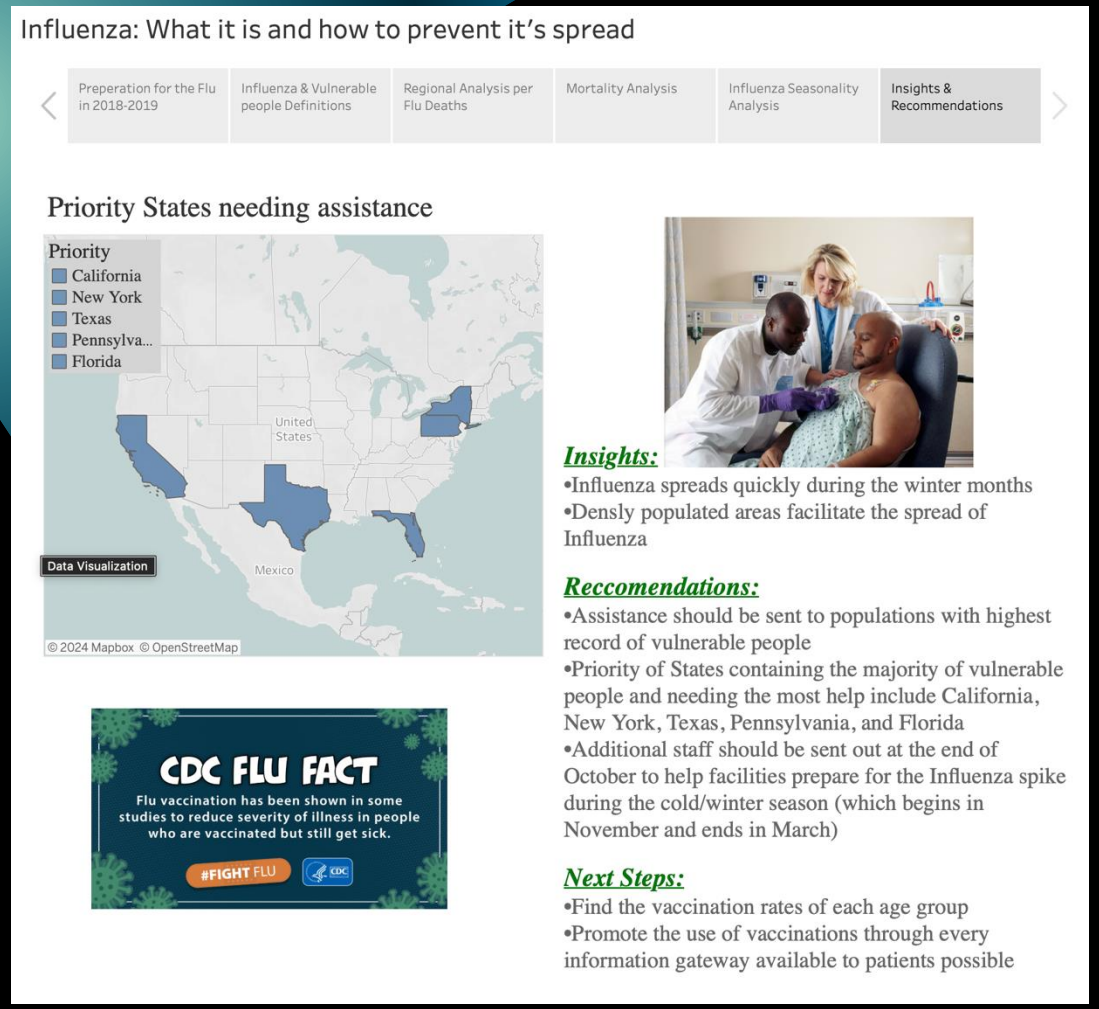
Staff should be sent out ahead of the years cold weather spike to cull the mortality rate

- Send staff out to states in need at end of October in preparation for Dec-Mar up-ticks

Utilize every information gateway possible to make people aware of the dangers of influenza virus

- Ad space, commercials, etc…

Stress importance of influenza vaccines for elderly people 65 or older



03 Rockbuster Stealth

Intro:

Rockbuster Stealth is a movie rental company trying to break into the online rental space. Using their existing data will help gain a better understanding on how to form a business strategy to break into a fiercely competitive sector.

Key Questions & Objectives

- 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

- 'Dataset is compromised of film inventory, customers, and payments amongst other things'
- Rockbuster dataset was loaded into PostgreSQL database for analysis
- Data was collected over 3 month span

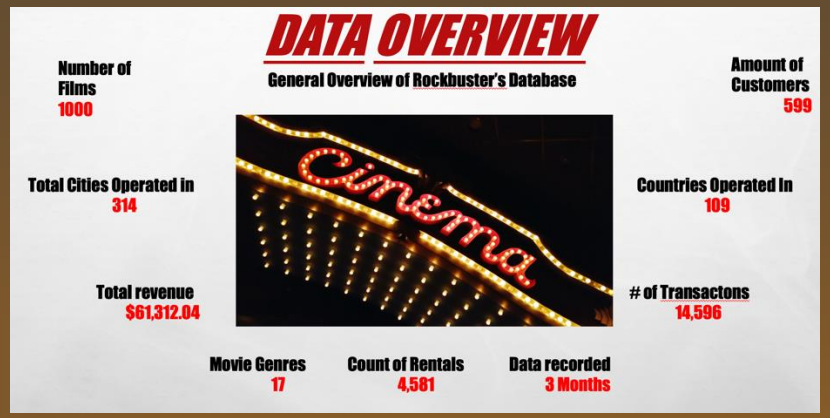
Tools

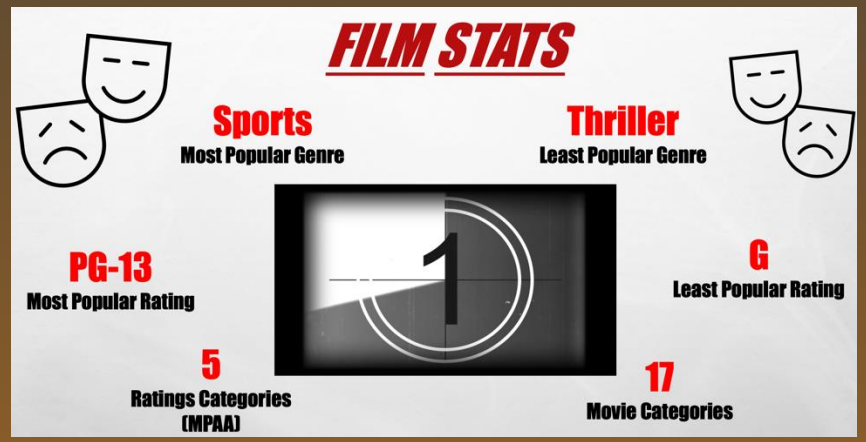
- Excel
- Tableau
- Powerpoint
- SQL
- pgAdmin4
- Anaconda

Analysis

SKILLS:

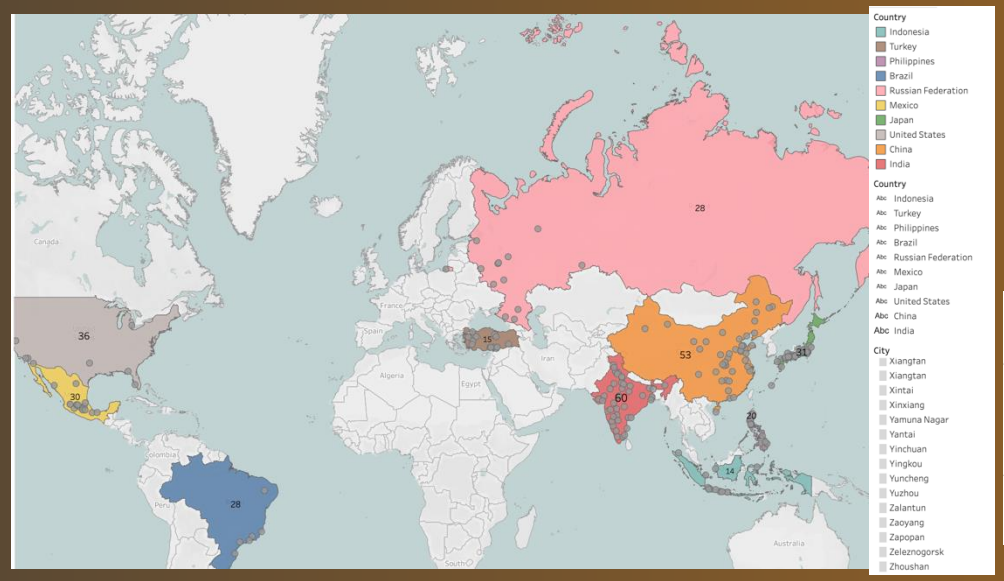
- *SQL*
  - *Database cleaning*
  - *Querying*
  - *Common table expressions (CTE)*
  - *Performing subqueries*
  - *Joining tables*
- *Visualization in Tableau*
- *PowerPoint presentation for results*

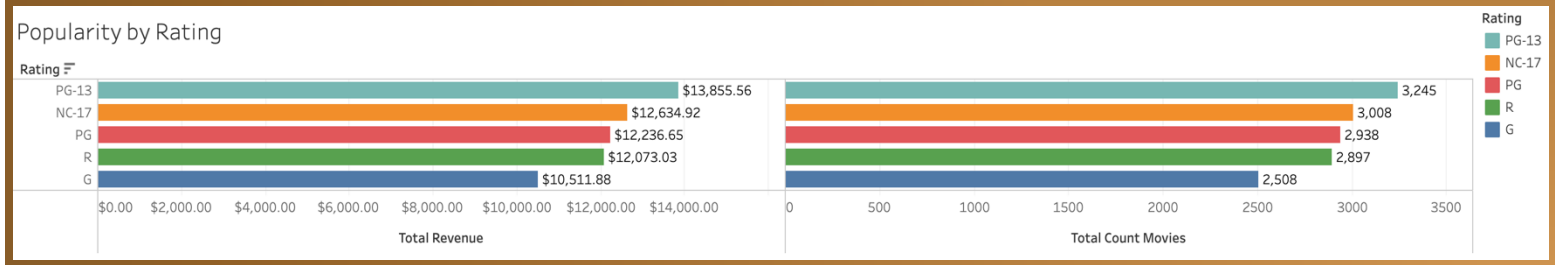




- Initially I conducted basic data quality queries to check for duplicates, missing values, and non-uniform data types and remedied as needed
- I then checked for general statistical information and film statistics as shown above

- I was able to delve information top grossing genres and countries



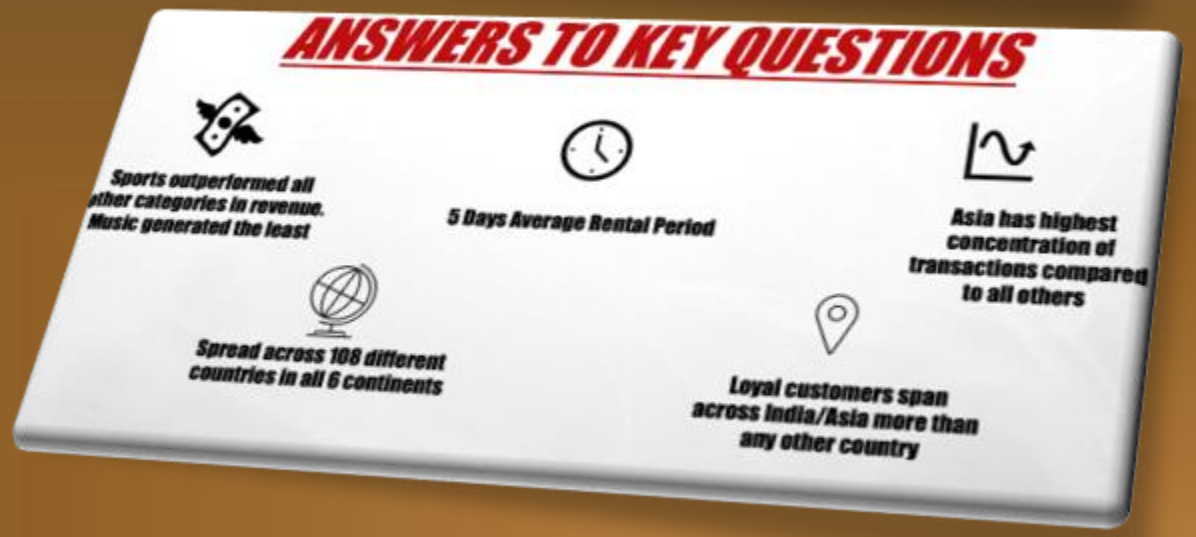


Analysis

- After recording stats related to the films I used multiple joins to acquire information about revenue per genre, countries and customers

SQL Query to find most loyal customers and where they live

```
(SELECT
A.customer_id,
A.first_name,
A.last_name,
D.country,
C.city,
SUM(B.amount) AS total_amount_paid
FROM customer A
    JOIN
    payment B ON A.customer_id = B.customer_id
    JOIN
    address E ON A.address_id = E.address_idzzz
    JOIN
    city C ON E.city_id = C.city_id
    JOIN
    country D ON C.country_id = D.country_id
    WHERE C.city IN
                    (SELECT city
                    FROM customer A
                    JOIN address E ON A.address_id = E.address_id JOIN city C ON E.city_id = C.city_id
                    JOIN country D ON C.country_id = D.country_id GROUP BY C.city
                    ORDER BY COUNT(A.customer_id) DESC LIMIT 10)
GROUP BY
A.customer_id, A.first_name, A.last_name, D.country, C.city
ORDER BY total_amount_paid DESC
LIMIT 10)
```



Customer Id	First Name	Last Name	Country	City	Total Amount Paid
148	Eleanor	Hunt	Runion	Saint-Denis	211.55
144	Clara	Shaw	Belarus	Molodetno	189.6
566	Casey	Mena	Turkey	Tokat	130.68
84	Sara	Perry	Mexico	Atlixco	128.7
506	Leslie	Seward	Indonesia	Pontianak	123.72
512	Cecil	Vines	United Kingdom	London	115.74
131	Monica	Hicks	Ukraine	Mukateve	112.73
537	Clinton	Buford	United States	Aurora	98.76
476	Derrick	Bourque	Canada	Gatineau	87.8
521	Roland	South	China	Yingkou	80.77

Results

- ✓ India China and the USA account for 1/4th of the customer base
- ✓ Asia is the region with the largest customer base and sales revenue
- ✓ There are 599 customers in 108 countries spread across all 6 continents
- ✓ Customers rent most of their films at the lowest rental rate of $0.99
- ✓ Genres of movies rented were evenly spread with Sports, Sci-Fi, and Animation at the top of the list & Thriller, Music and Travel films at the bottom
- ✓ NPAA Ratings grossed similar returns with a spread of $3,343.68

Recommendations

- ✓ Before expanding consider partnering with studios to create surplus of content
- ✓ Soft launch product in most popular countries first (India, China, USA)
- ✓ Boost popularity with advertising across all 108 countries Rockbuster is located in
- ✓ Adjust pricing scale through subscriptions or memberships to ensure consistent revenue
- ✓ Offer loyalty programs to top spending customers

Deliverables

- ✓ Data Dictionary explaining the relationships between data tables
- ✓ Excel file containing queries and outputs
- ✓ PowerPoint presentation of results
- ✓ Visualizations created in Tableau
- ✓ Project Repository on GitHub

Instacart Analysis

Intro:

- Instacart is an online grocery store looking to uncover information about their sales
- I performed initial data and exploratory analysis of some of their data to derive insights and suggest strategies for better segmentation based on the provided criteria.

Key Questions:

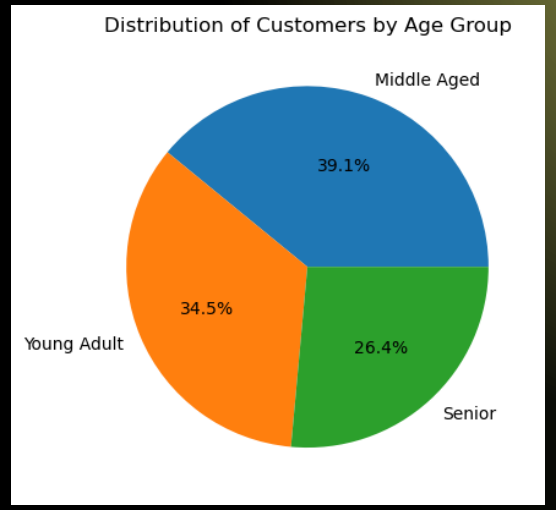
- What are the busiest days of the week and hours of the day?
- Are there times when people tend to spend the most money?
- Are certain products more popular than others?
- Identify different customer profile types and how their habits differ
- Create simpler price range groupings

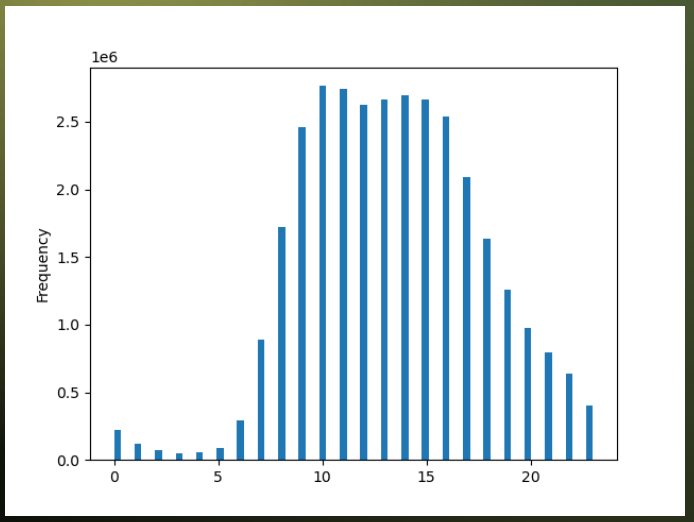
Dataset & Tools:

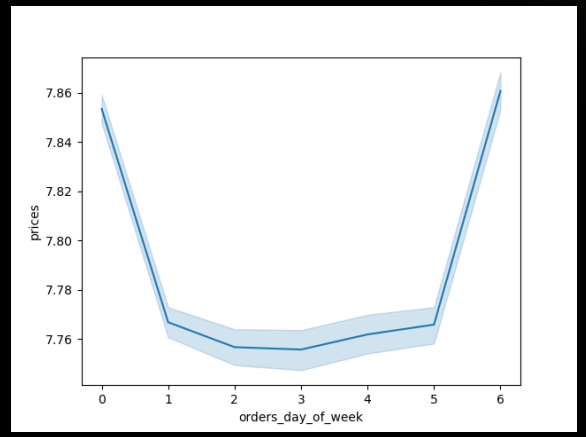
- Multiple open-source Instacart datasets
  - Instacart is a real company however Career Foundry has fabricated data for educational purposes
- Analysis was done by Jupyter notebooks and Anaconda libraries manager
- Excel for general analysis, Tableau for visualizations and PowerPoint for presentation
- Python was the primary analysis tool with accompanying libraries
  - OS
  - Pandas
  - NumPy
  - Matplotlib
  - Scipy
  - Seaborn

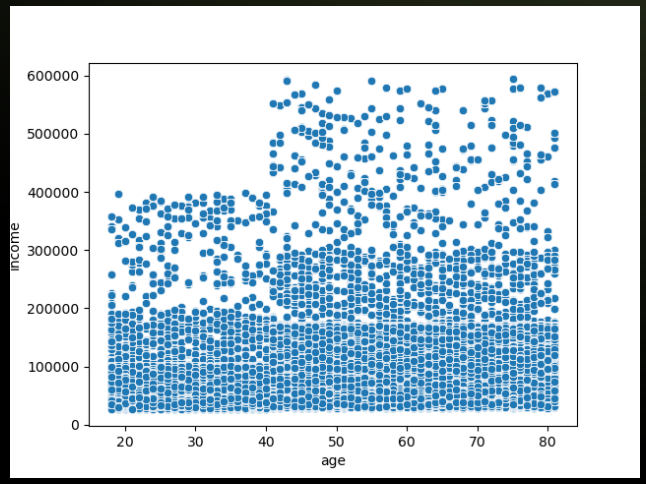
Analysis

- Cleaned, concatenated, and merged data-frames
- Performed basic data wrangling
  - Consistency Checks
  - Deleted duplicates
  - Dropping and renaming columns
  - Changing data types
  - Addressing missing values
- Derived new columns
  - i.e ‘age_groups’ and ‘spending_flag’
- Created 7 customer profiles based on age, income, and number of dependents
  - i.e ‘Married young parent’ or ‘single senior adult’ etc…
- Created visualizations for deep insight on data
  - i.e. time of day orders were placed, most popular order type, and age/income ratio



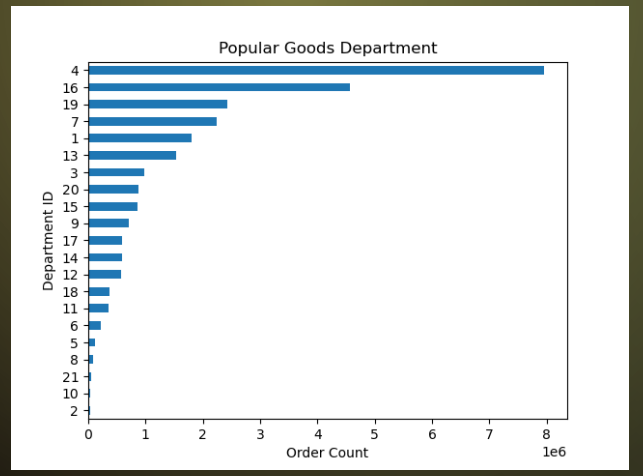




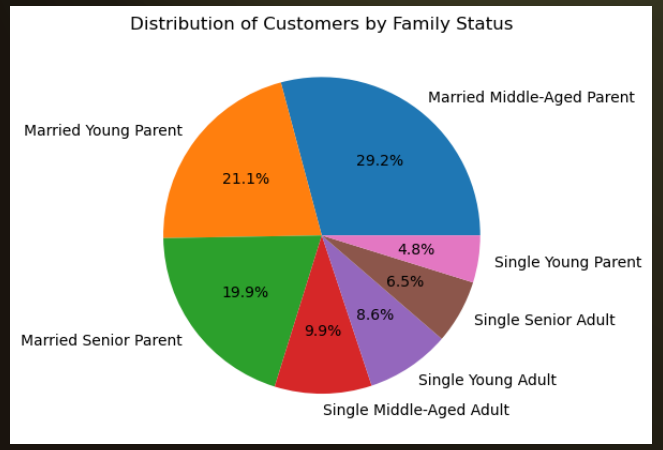


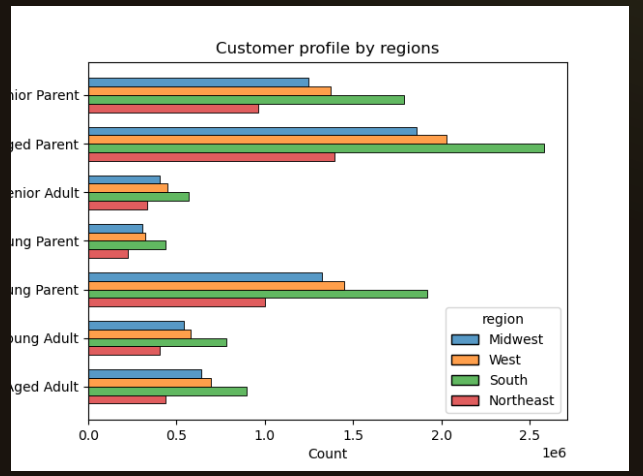
Results

- High income groups spend the most in Instacart with middle income following closely behind
- Customers with children/ dependents make up 75% of customer base
- Married couples utilize Instacart more than single persons
- Spending across regions is proportional
- Produce is the most popular department
- Identified key aspects of data for marketing purposes
  - Popular times and days
  - Most lucrative demographics and regions



department_id	department
1	frozen
2	other
3	bakery
4	produce
5	alcohol
6	international
7	beverages
8	pets
9	dry goods pasta
10	bulk
11	personal care
12	meat seafood
13	pantry
14	breakfast
15	canned goods
16	dairy eggs
17	household
18	babies
19	snacks
20	deli
21	missing





Recommendations

- Focus ads on weekends to capitalize when traffic is highest
- Promote higher/medium price range products during the late night early mornings targeting the more sporadic spenders.
- Offer mid week discounts to boost purchases during slow periods

Project deliverables

- Jupyter Notebooks
- Stakeholder reports via Excel
- GitHub repository

New York City Analysis

Intro:

- This study focuses in on Airbnb listings across New York City in regards, but not limited to, pricing and popularity of listings, availability regarding season, and review patterns of customers to !help gain the best insight and guide customers find the vacation of their dreams

Key Questions

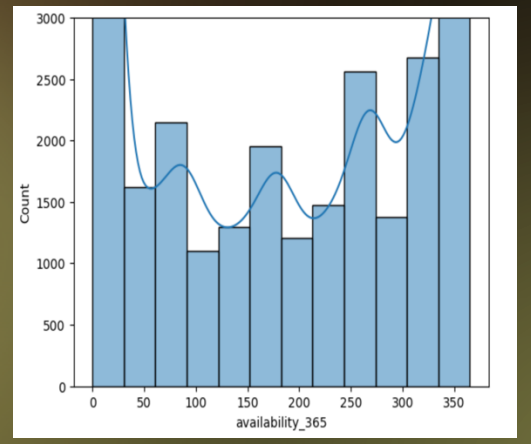
1. What are the most popular bookings based off the room type?
2. Which city/ neighborhood are the most popular listings booked?
3. What price ranges are the properties that are booked the most?
4. How does availability & minimum night stays effect customer rental decisions?

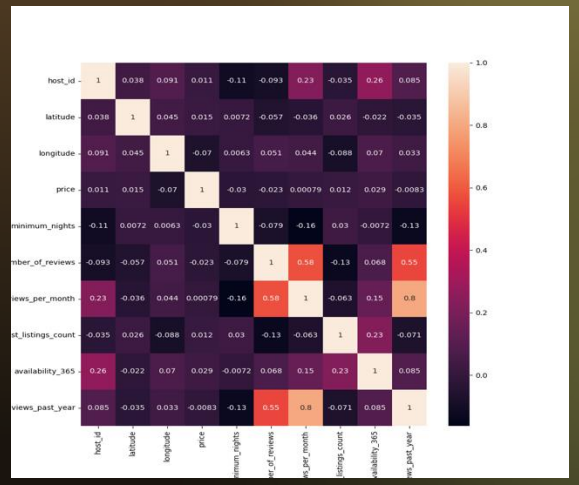
Data set and Tools:

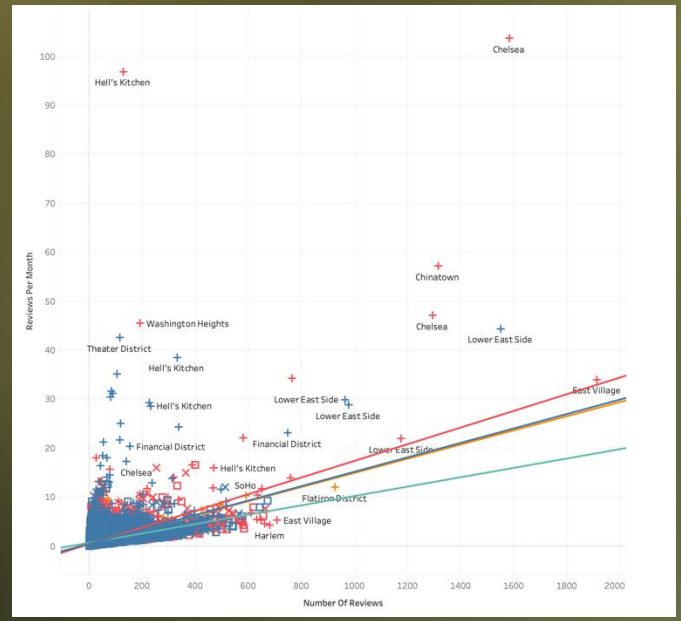
- Airbnb is a real-world rental site where individuals or companies list properties to rent for certain periods of time
- **https://insideairbnb.com/get-the-data/**
- Analysis was done in Jupyter notebooks and Python libraries manager with accompanying libraries
  - OS
  - Pandas
  - NumPy
  - Matplotlib
  - Scipy
  - Seaborn
  - Json & Geojson
  - Folium
  - Statsmodel.api
  - Quandl
- Visualizations created in Tableau

Analysis

- ***Methods:***
  - Cleaned data sets
  - Performed basic data wrangling
    - Consistency Checks
    - Deleted duplicates
    - Dropping and renaming columns
    - Changing data types
    - Addressing missing values
  - Organized variables into structure suitable for Tableau
  - Explored relationships using Python:
    - Correlation Matrices
    - Scatterplots
    - Cat Plots
    - Histograms
    - Maps
    - Linear Regression
    - Clustering
  - Created visual insights using Tableau Based off:
    - Reviews
    - Room Type
    - Price
    - Neighborhoods
    - Availability
    - Seasonality
    - Hosts
    - Minimum Nights
    - Licensing

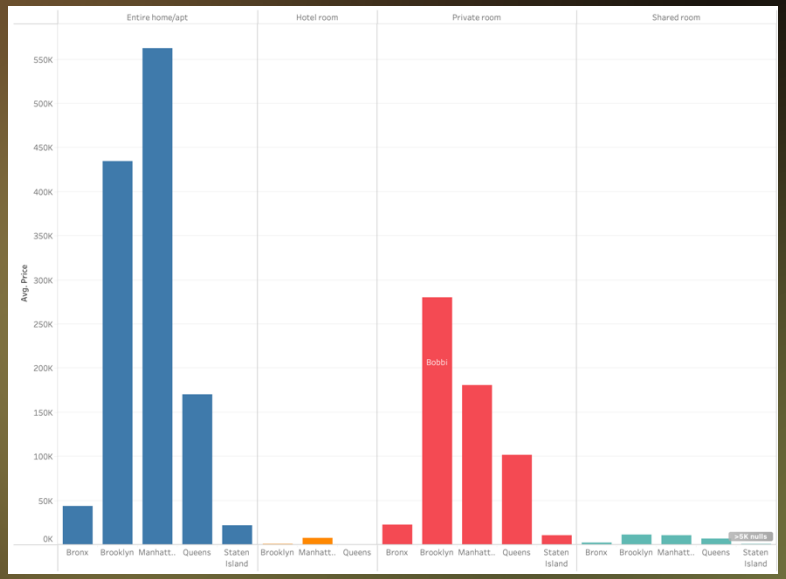


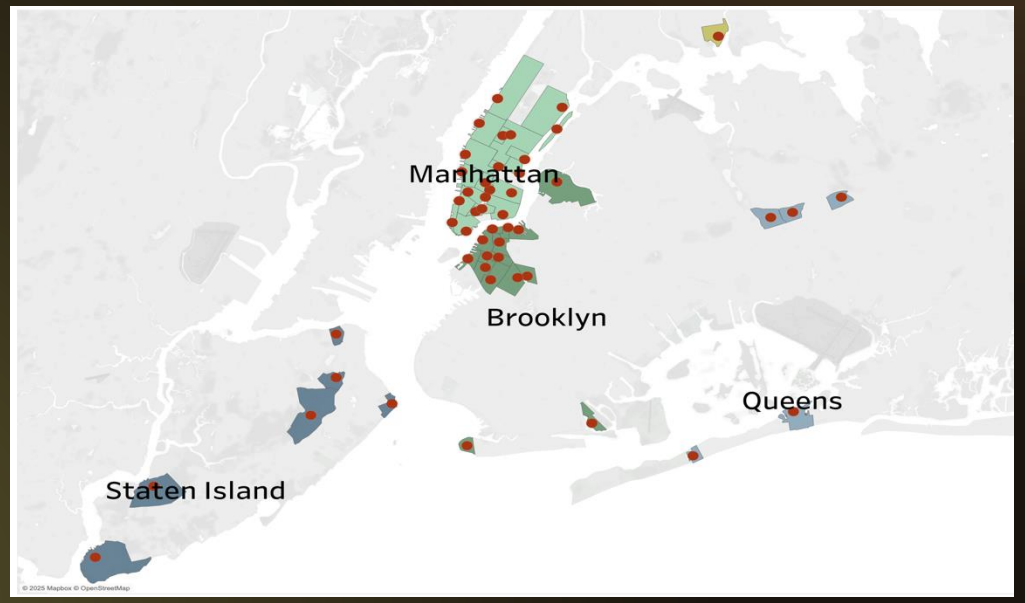


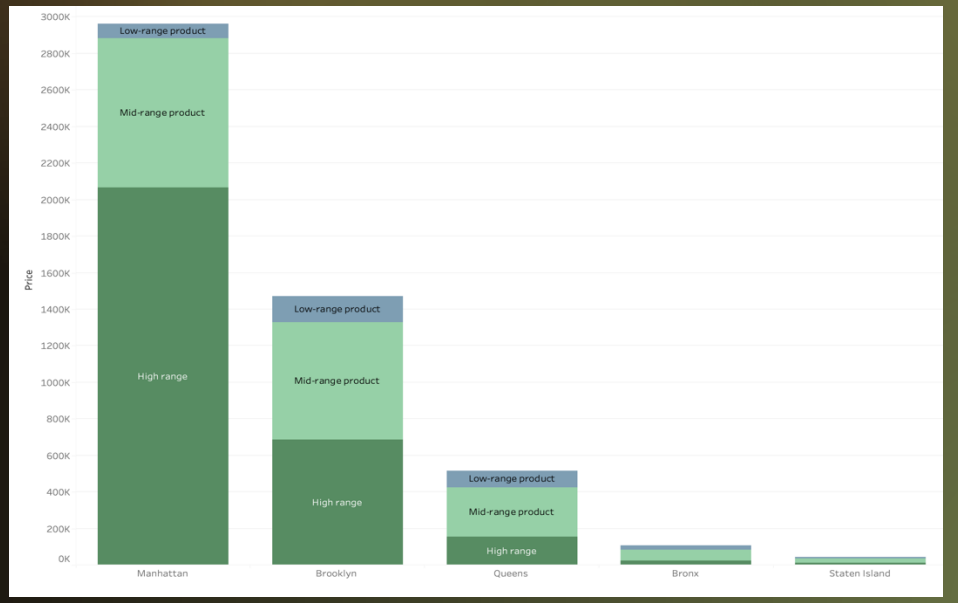


Results

- ✓ Manhattan and Brooklyn are the most expensive neighborhoods and Staten Island brings up the rear as cheapest
- ✓ Manhattan has the most listings of all areas that compromise New York City
- ✓ Renters are more likely to visit NYC in the warmer months
- ✓ The host with the most listings is A company called Blueground
  - ✓ Their listings have a minimum of 31, 60, or 90 days
  - ✓ New York requires licenses for short-term rentals effectively protecting housing for residents
- ✓ The most popular type of stay is Full Bedroom or apartments and private rooms indicating privacy as a necessity
- ✓ High price range listings are the most prominent type of listing available







Recommendations

- As a customer, search for the areas in NYC that have the most positive reviews and safest neighborhoods
- Start looking in the Manhattan area because it has the most listings of all price ranges
- The best times to visit NYC are in the warmer months but there is more availability during the last quarter (Oct-Dec)
- Hotel rooms are best for short term stays, for longer stays contact Blueground

Project Deliverables

- Jupyter Notebooks
- Tableau Workbooks/Website
- Github Repository

Thanks for your time!

Evan Carr

Contact me via

LinkedIn

Gmail

GitHub

Discord