Ngawang Dhundup

DATA ANALYTICS PORTFOLIO

GitHub

+ableau

Hi I’m Ngawang Dhundup

Welcome to my portfolio

Currently, I am looking for a Junior Data Analyst role that gives me the opportunity to learn more about the field, and make impactful inferences to help drive the companies growth.

After years of customer service in many different fields, such as retail, non-profit, and healthcare, I retrained as a data analyst. I always had an interest in customer behavior, and how I could use data to predict the needs and wants of a customer before they even knew it. My psychology major definitely helped urge my curiosity into the human psyche.

View my projects here

Projects

1. Medical cost cluster Analysis

2. InstaCart Basket Analysis

3. Rockbuster Stealth

4. Preparing for Influenza Season

5. GameCo

Medical cost Cluster Analysis

Insurance claims analytical dashboard

OBJECTIVE

To build an interactive dashboard that will visually showcase well-curated results of an advanced exploratory analysis conducted in Python. In this case study we will explore what factors contribute to medical charges in America

DATA

Data was sourced from Kaggle, and scraped from a book " Machine Learning with R" by Brett Lantz. Downloaded here.

SKILLS

- Sourcing open data
- Exploring data analytics
- Geographical Visualizations with Python
- Linear regression machine learning
- Clustering machine learning
- Interactive Dashboard on tableau

TOOLS

Python

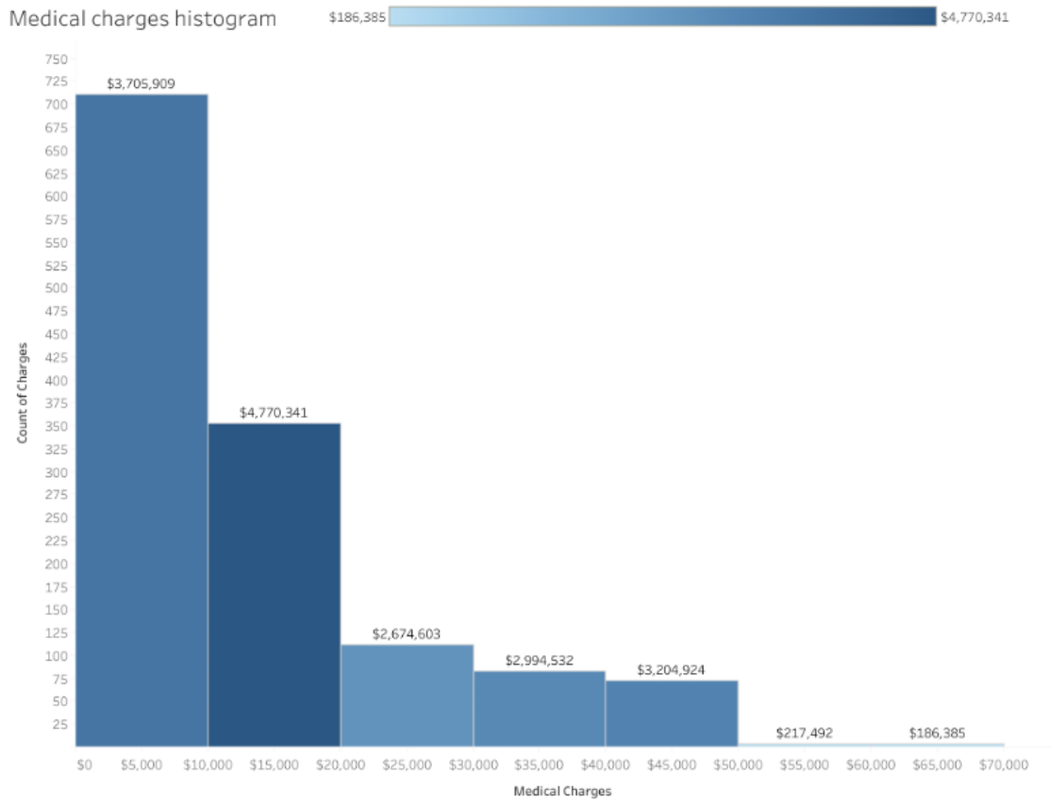
Tableau

Analysis

COMPARING MEDICAL COSTS BETWEEN DIFFERENT DEMOGRAPHICS

	Low Cost	Middle Cost	High Cost
Female	53.63%	37.46%	8.91%
Male	52.74%	32.00%	15.26%

	Low Cost	Middle Cost	High Cost
Non-Smoker	66.89%	32.17%	0.94%
Smoker		44.53%	55.47%

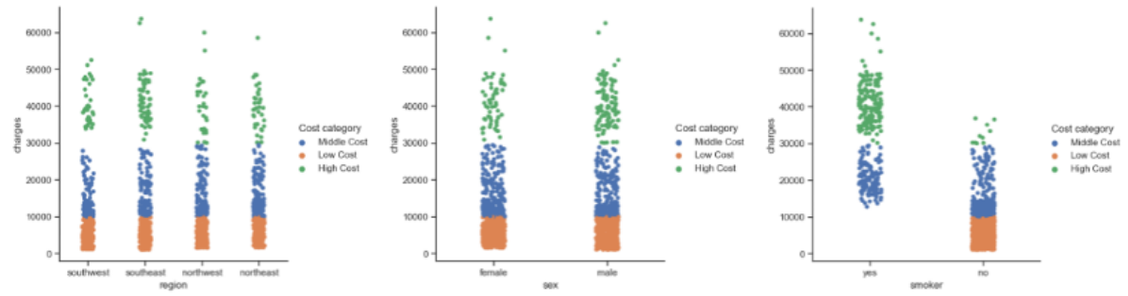


ARE THERE ANY OTHER RELATIONSHIPS TO EXPLORE?

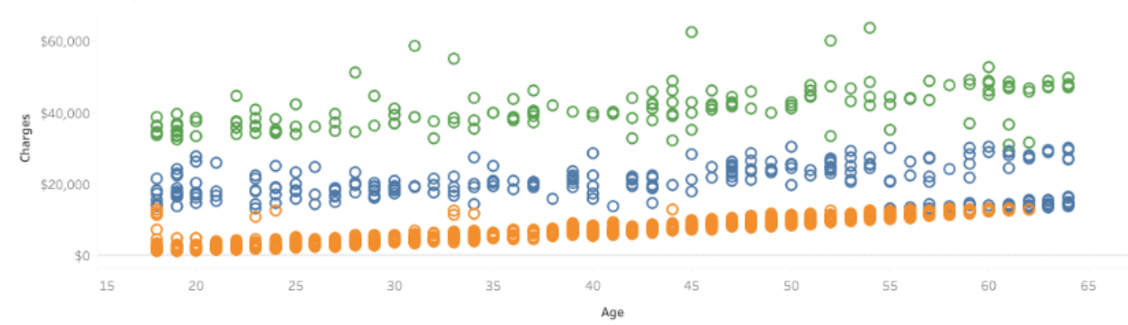
Clustering helps identify and trends and patterns characterised by similarities which arent always obvious. Using the k-means algorithim in Python cost charges was divided into three cost categories (low, middle, high) and compared to three variables (region, sex, smoker)

Region and Sex had no distinguishable differences, however the "smoker" variable had significant clustering patterns.

For **smokers** they had significantly more green and blue clusters, while **non-smokers** had denser orange and blue clusters.



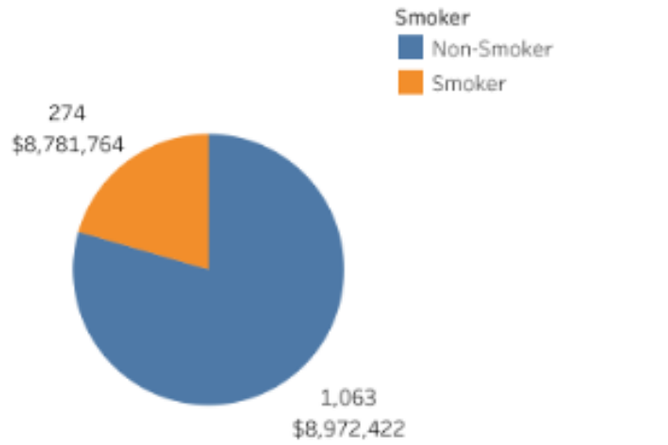
Cluster analysis

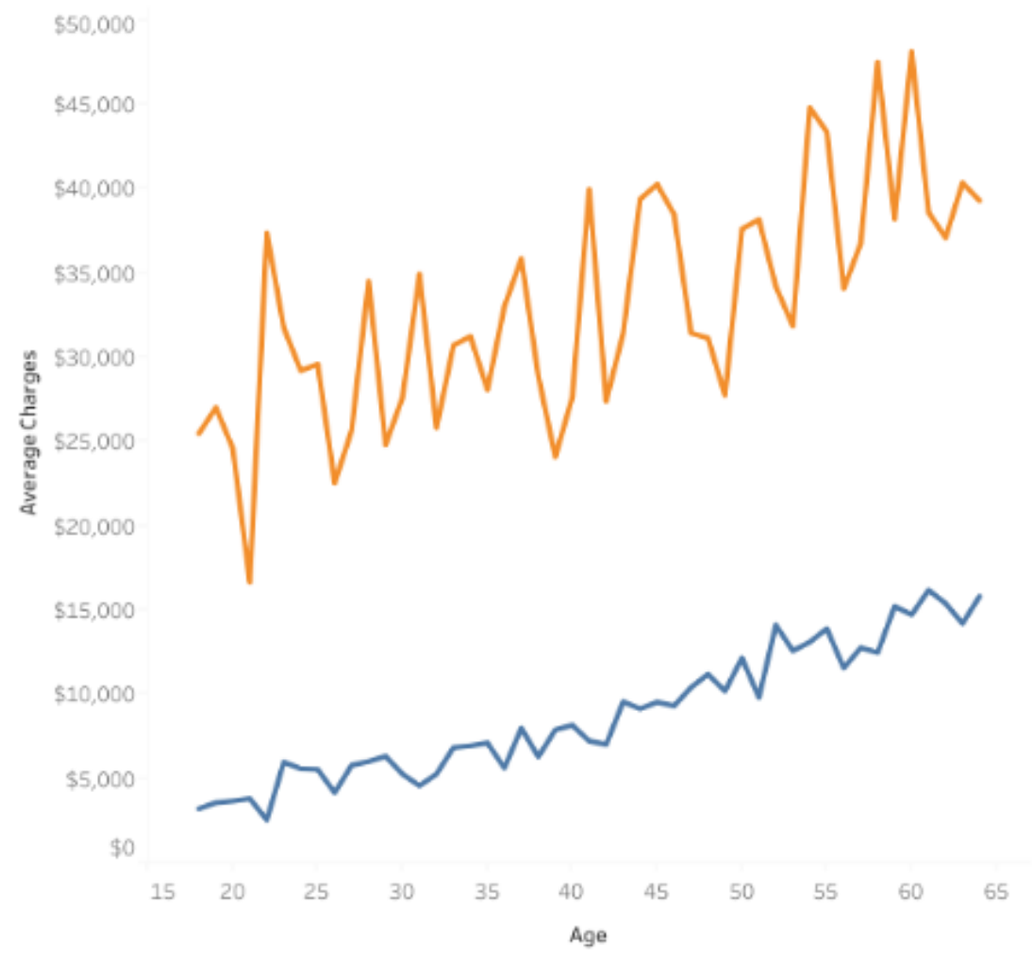


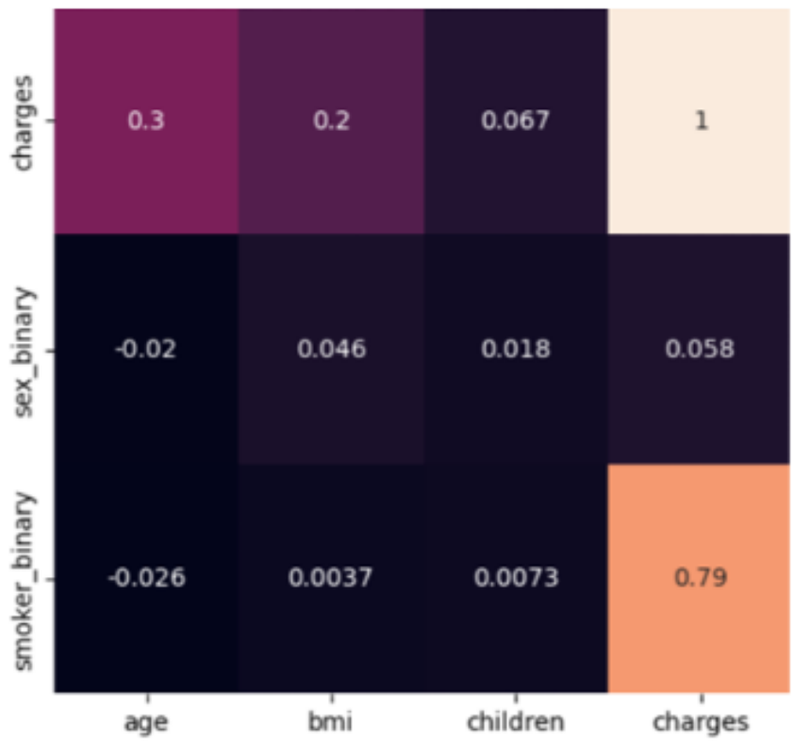
DO YOUR HABITS AFFECT MEDICAL COSTS?

Smokers pay significantly more in medical costs compared to non smokers.

According to the Pie chart, even though non-smokers out weigh the count of smokers by a multiple of 5, Non smokers only pay approximately 200k more in total. In fact on average based on the line graph below someone in their 30s who is a smoker pays 5 times more than their non-smoker counterpart.







Conclusion and next steps

Analysis Findings:

- Smoking is the leading factor for high medical costs
- Age is one of the factors for medical costs. On average the older client has more medical costs than younger clients
- Sex and where you live has no statistical significance to medical costs

Limitations of case study:

- Limited sample size, only 1500 rows
- Difficult to determine clients income in proportion to medical costs
- Limited variables, and no time series available

Next Steps:

- Increase sample size
- Include more variables, such as client income to determine how much the medical charge affects them financially

InstaCart Basket Analysis

Grocery store

OBJECTIVE

Instacart, an online grocery store that operates through an app. Instacart already has very good sales, but they want to uncover more information about their sales patterns. The Instacart stakeholders are most interested in the variety of customers in their database along with their purchasing behaviors.

DATA

Data originates from Instacart's 2017 open - source dataset. Downloaded here.

Data Dictionary here .

SKILLS

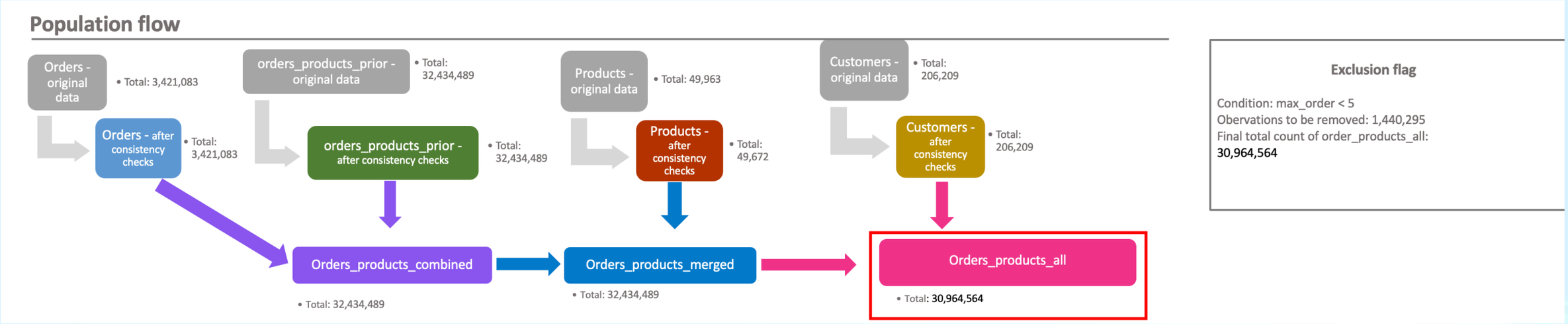
- Data wrangling
- Data merging
- Deriving Variabels
- Grouping data
- Visualizations in Python
- Reporting in Excel

TOOLS

Python

Excel

Population Flow



Preparing Data

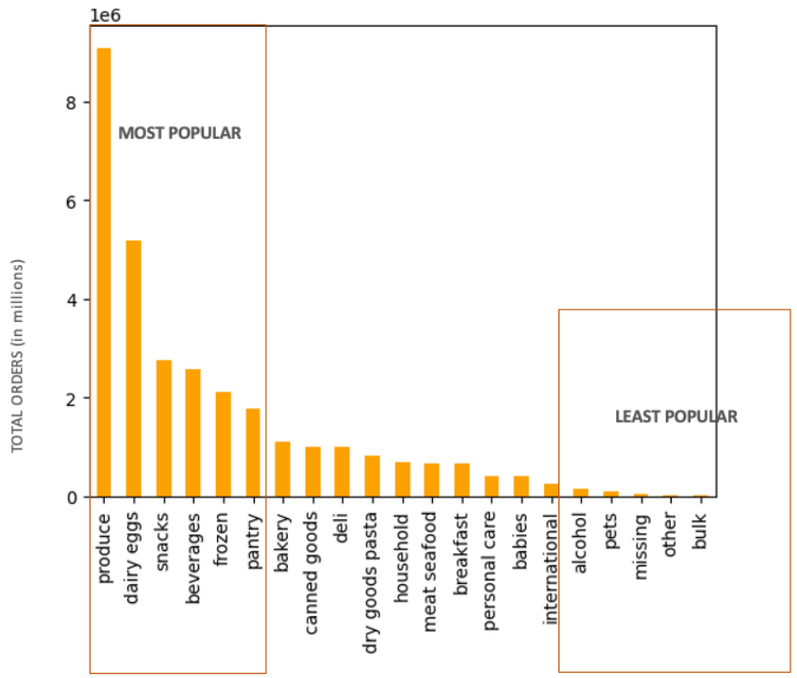
- Derived new columns
- Merged data sets
- Created flags using for loops, If Functions
- Created exclusions exporting certain CSV, and pickle files

Findings

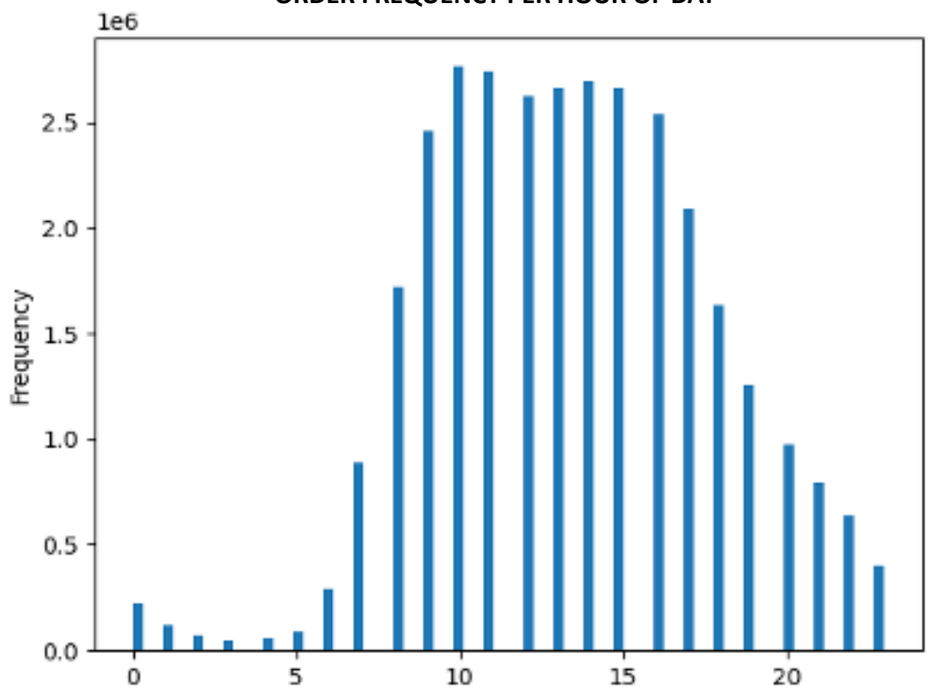
SUM OF ORDERS BY DEPARTMENT

DEPARTMENT	TOTAL ORDER($)
produce	9,079,273
dairy eggs	5,177,182
snacks	2,766,406
beverages	2,571,901
frozen	2,121,731
pantry	1,782,705
bakery	1,120,828
canned goods	1,012,074
deli	1,003,834
dry goods pasta	822,136
household	699,857
meat seafood	674,781
breakfast	670,850
personal care	424,306
babies	410,392
international	255,991
alcohol	144,627
pets	93,060
missing	64,768
other	34,411
bulk	33,451

SUM OF PRODUCTS

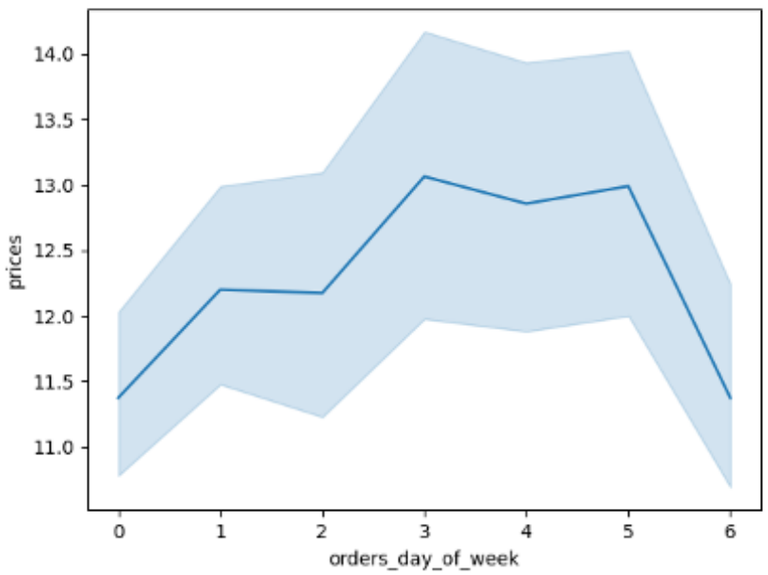


ORDER FREQUENCY PER HOUR OF DAY



- Most orders occur from 9am to 4pm with the peak at 10am
- The busiest days are Saturday, Sunday, and Friday in that order
- Produce and Dairy eggs are the two most popular departments
- International, alcohol, pets and bulk are the least popular

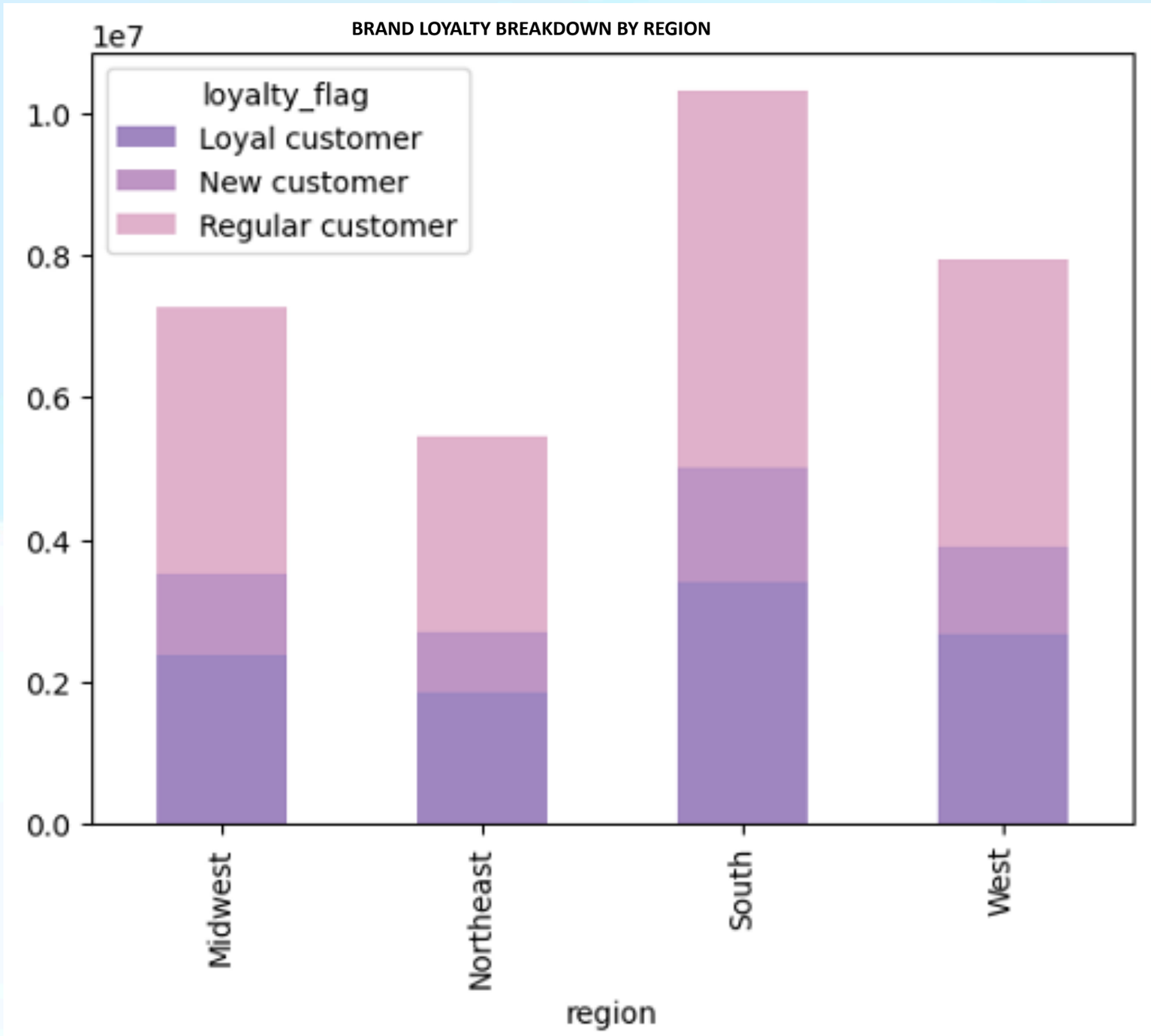
CUSTOMERS' EXPENDITURE PER DAY



LEGEND:

0 - Saturday
1 - Sunday
2 - Monday
3 - Tuesday
4 -Wednesday
5 - Thursday
6 - Friday

Customer Loyalty Findings



- The South region has the majority of customers, and the bulk of Loyal Customers
- The Northeast has the lowest total customer count
- Each region has more regular customers compared to loyal customers which means there is room for growth and acquisition

Insights & Recommendations

My recommendation would be to schedule ads after 8pm based on the fact most orders are placed between 9am to 4 pm with the peak at 10am. Any weekday is a good recommendation because the weekdays are fairly equal in terms of orders per day.

To encourage new customers to turn into regular and loyal customers, we should focus the ads in the South

It is recommended to run more ads for higher priced items such as meat seafood, and alcohol between hours of midnight and 6am

InstaCart can focus on increasing variety of brands and price points for Produce, Dairy eggs, and snacks in order to capitalize on their top performers

Full report can be found here

Rockbuster Stealth

A movie rental company

OBJECTIVE

Business Intelligence analysis on a fictional movie rental company. Rockbuster Stealth is planning to use its existing movie licenses to launch an online video rental service to stay competitive with the other streaming providers. The objective of this analysis is to present insights on Rockbuster's current portfolio. The findings will serve as an input for future company strategy.

DATA

The dataset rests in a SQL database and a data dictionary has been created as a part of this analysis. The database backup file can be downloaded here .

SKILLS

- Database querying
- Filtering, Cleaning, Transformation
- Joining Tables
- Subqueries
- CTEs

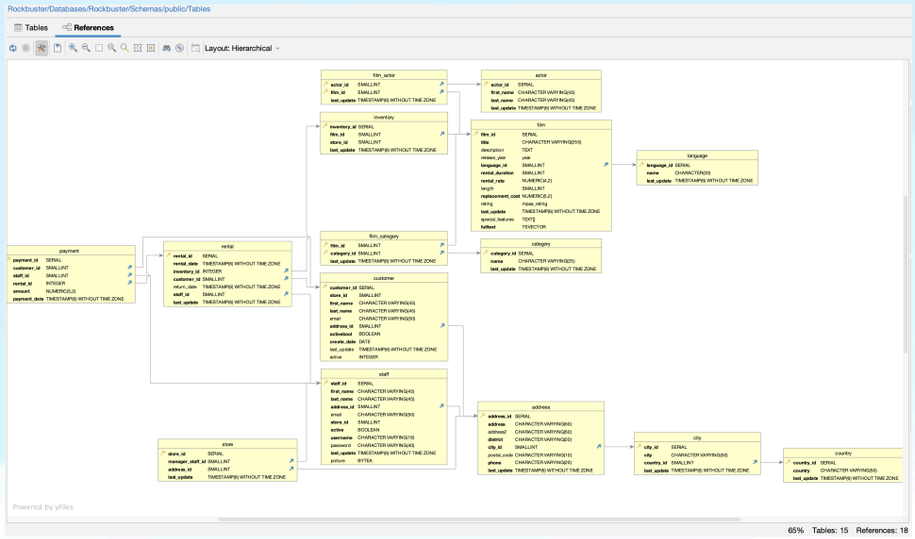
TOOLS

SQL (PostgreSQL)

Tableau

Sample Analysis and Queries

Entity Relationship Diagram created using DB Visualizer



Query searching for duplicates

```
SELECT DISTINCT customer_id,
        store_id,
        first_name,
        last_name,
        email,
        address_id,
        activebool,
        create_date,
        last_update,
        active
FROM customer;

SELECT DISTINCT active
FROM customer
GROUP BY active;
```

Query joining two tables searching for the top 10 Cities within the top 10 Countries

```
SELECT
C.city,
D.country,
Count(A.customer_id) AS customer_count
FROM customer A
INNER JOIN address B ON A.address_id = B.address_id
INNER JOIN city C ON B.city_id = C.city_id
INNER JOIN country D ON C.country_id = D.country_id
WHERE D.country IN ('India', 'China', 'United States', 'Japan', 'Mexico', 'Brazil', 'Russian
Federation', 'Philippines')
GROUP BY C.city, D.country
ORDER BY Count(A.customer_id) DESC
LIMIT 10;
```

Query showcasing CTEs searching for the count of customers within the top 5 Countries

Query | Query History

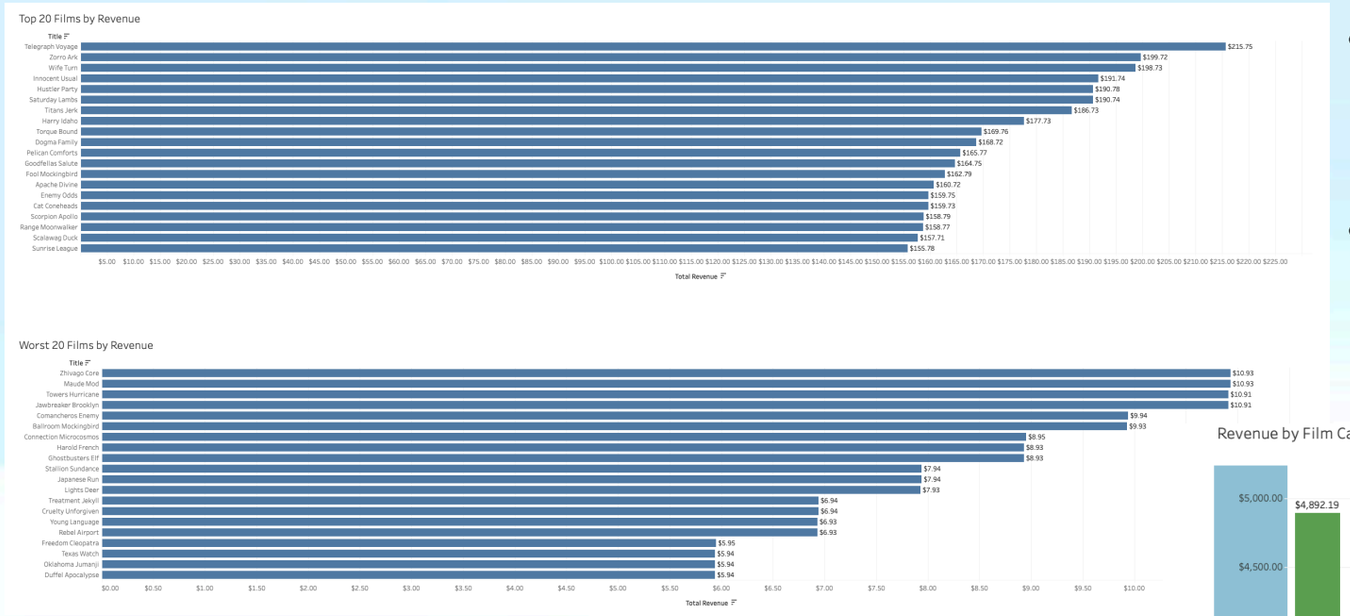
```
--first cte = top 5 customers--
WITH top_5_customers_cte AS
        (SELECT
        A.customer_id,
        A.first_name AS customer_first_name,
        A.last_name AS customer_last_name,
        A.email AS customer_email,
        C.city,
        D.country,
        SUM(E.amount) AS total_paid_by_customer
        FROM customer A
        INNER JOIN address B ON A.address_id = B.address_id
        INNER JOIN city C ON B.city_id = C.city_id
        INNER JOIN country D ON C.country_id = D.country_id
        INNER JOIN payment E ON A.customer_id = E.customer_id
        WHERE C.city IN ('Aurora',
        'Atlixco',
        'Xintai',
        'Adoni',
        'Kurashiki',
        'Dhule (Dhulia)',
        'Pingxiang',
        'Ozamis',
        'Nezahualcyotl',
        'So Leopoldo')
        GROUP BY
        A.customer_id,
        A.first_name,
        A.last_name,
        A.email,
        C.city, D.country
        ORDER BY total_paid_by_customer DESC
        LIMIT 5)
--counting by country--
    SELECT D.country,
    COUNT(DISTINCT A.customer_id) AS all_customer_count,
    COUNT(DISTINCT D.country) AS top_customer_count
    FROM customer A
    INNER JOIN address B ON A.address_id = B.address_id
    INNER JOIN city C ON B.city_id = C.city_id
    INNER JOIN country D ON C.country_id = D.country_id
    LEFT JOIN
    top_5_customers_cte
    ON D.country = top_5_customers_cte.country
    GROUP BY D.country
    ORDER BY all_customer_count DESC
    LIMIT 5
```

Data Output | Messages | Notifications

	country character varying (50)	all_customer_count bigint	top_customer_count bigint
1	India	60	1
2	China	53	1
3	United States	36	1
4	Japan	31	1
5	Mexico	30	1

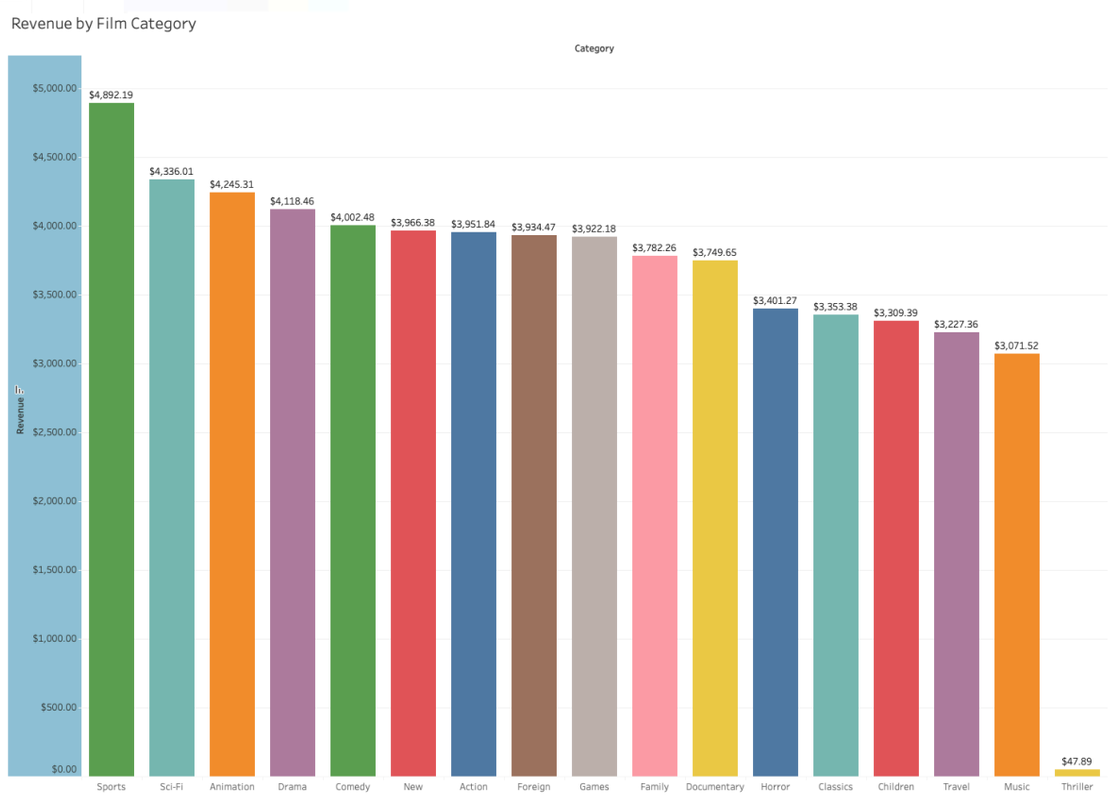
Findings

Revenues by Film Titles and Categories



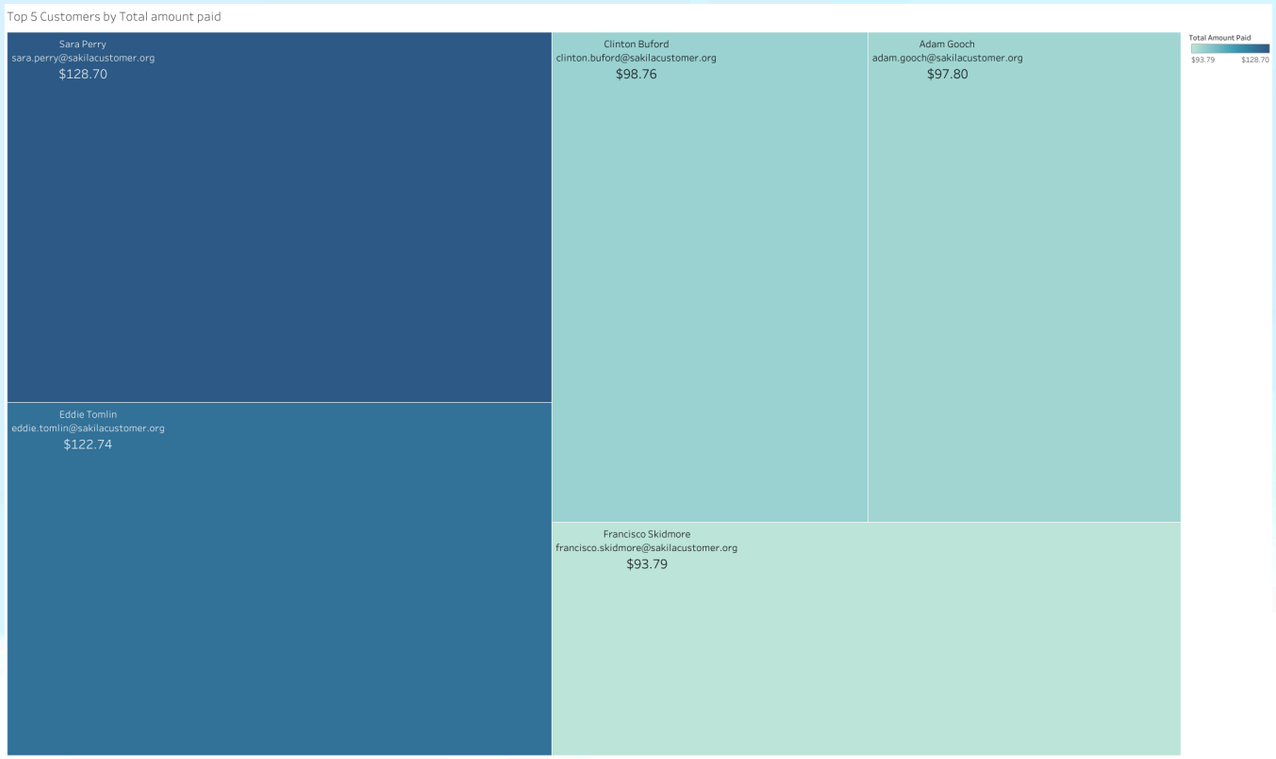
- The best film by revenue was the “Telegraph Voyage” sitting at $215.75
- The worst performing film was the “Duffel Apocalypse” sitting at $5.94

- Sports category performs the best with total revenue at $4,892.19
- The worst performing category is “Thriller” with total revenue at $47.89

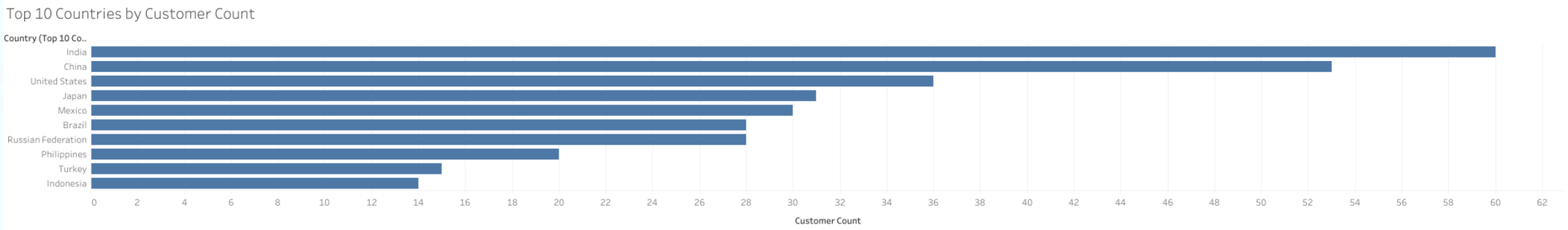


Customer overview

Who are the most loyal customers



- Top 5 customers to reward as a promo for retention, and to boost customer loyalty (customer information has been anonymized)
- Top 10 countries to focus marketing on to increase market share
- India has the largest customer count



Insights & Recommendations

Reduce expenses by removing low performing films such as “Duffle Apocalypse”, “Oklahoma Jumanji” and others found in the worst 20 films chart

Use promos and discounts in larger markets such as India and China to obtain more new and loyal customers

Allocate more budget on films that perform well such as the top 20 films, and categories such as sports and sci-fi

Full report can be found here

Preparing for Influenza Season

Medical agency

OBJECTIVE

To help a medical staffing agency that provides temporary workers to clinics and hospitals on an as-needed basis. The analysis will help plan for influenza season, a time when additional staff are in high demand. The final results will examine trends in influenza and how they can be used to proactively plan for staffing needs across the country.

DATA

Influenza deaths by geography, time, age, and gender from CDC. Dataset can be downloaded here.

Population data by geography from US Census Bureau. Dataset can be downloaded here.

SKILLS

- Data Cleaning, Transformation
- Data integration
- Data Visualization
- Statistical Hypothesis Testing

TOOLS

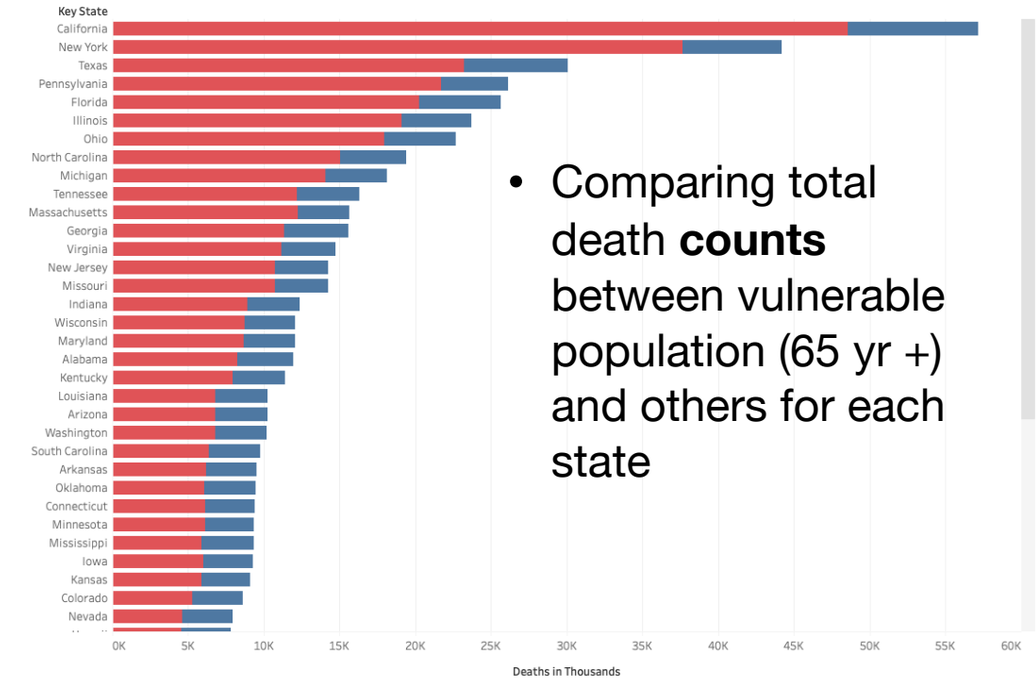
Excel

Tableau

Findings

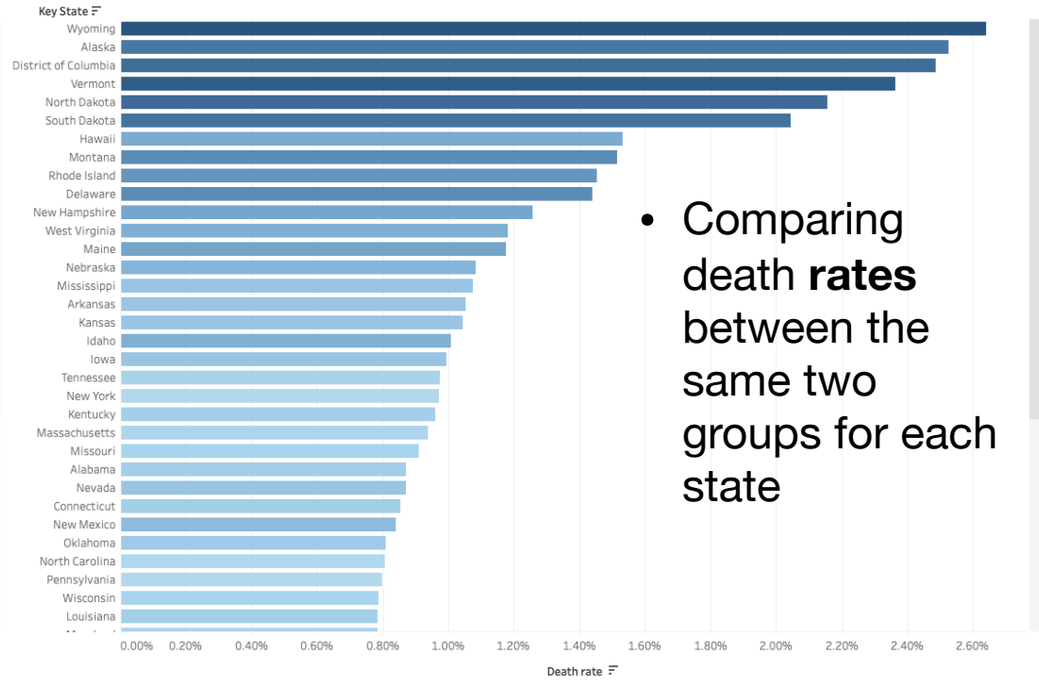
Relationships between two groups during the influenza season

Deaths for Vulnerable and Not Vulnerable populations (2009-2017)



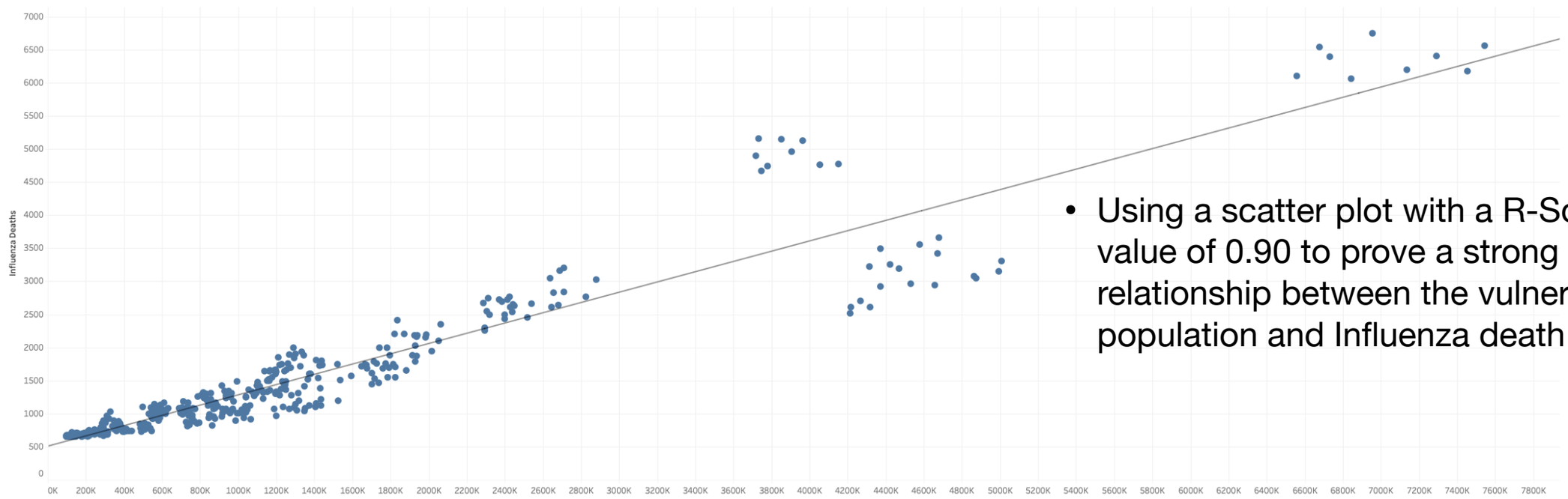
- Comparing total death **counts** between vulnerable population (65 yr +) and others for each state

US Vulnerable population Death Rate (2009-2017)



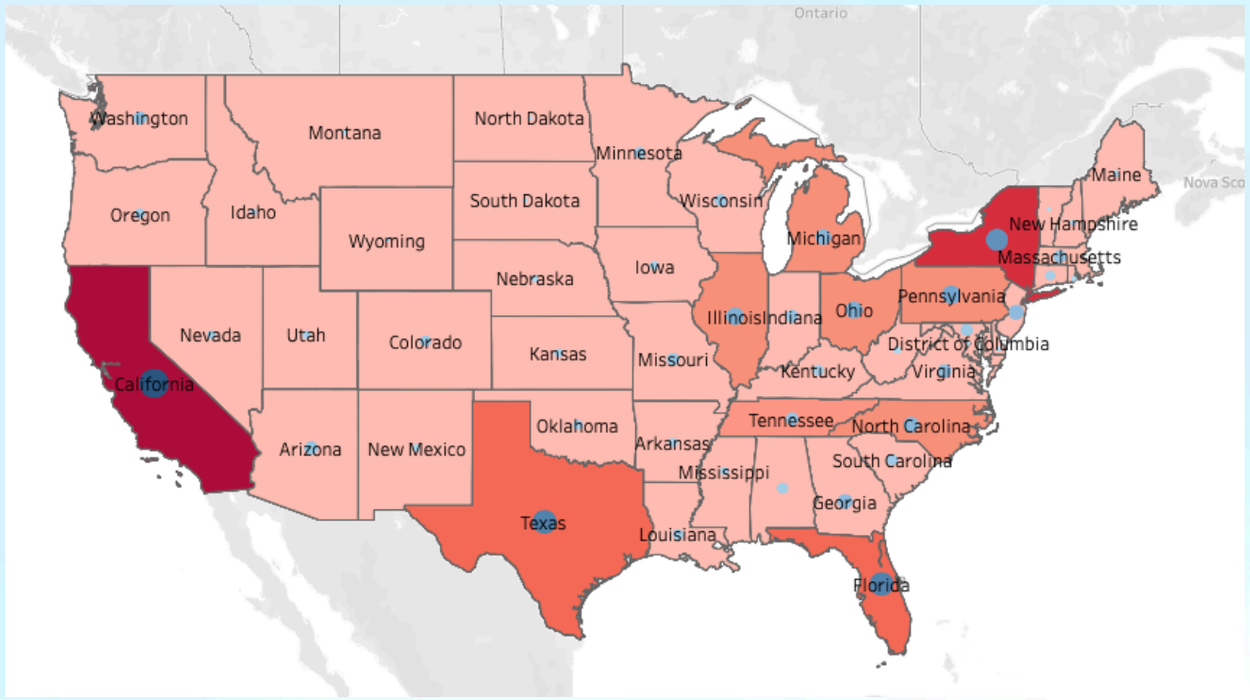
- Comparing death **rates** between the same two groups for each state

The Relationship between Vulnerable population and Total Influenza Deaths



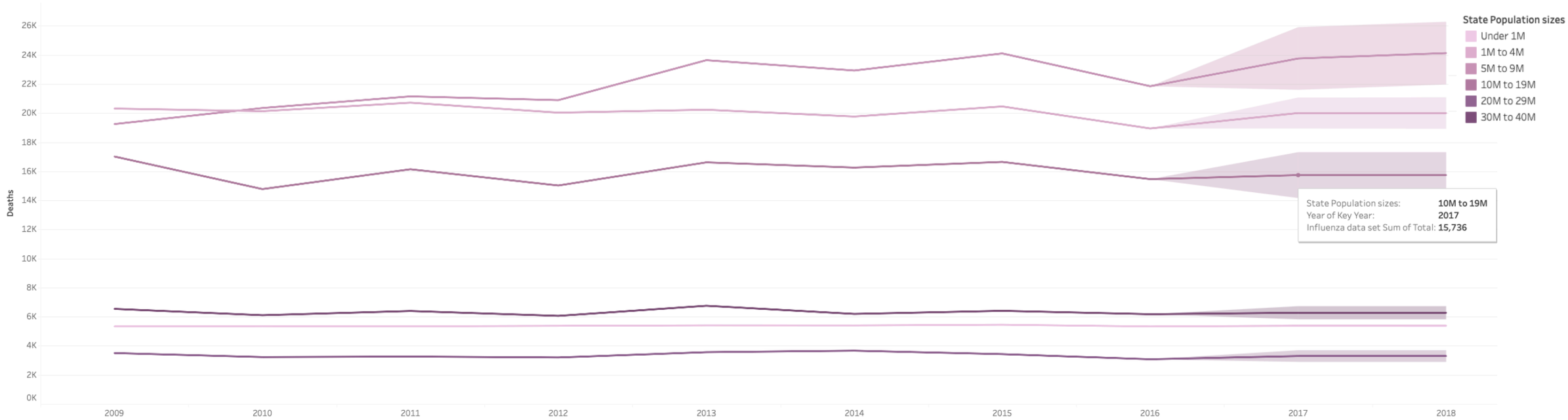
- Using a scatter plot with a R-Squared value of 0.90 to prove a strong relationship between the vulnerable population and Influenza death

Deaths across the Nation



- A forecast of influenza deaths grouped by state population for the next year using the past ten years as historical data
- Proving deaths will continually increase unless something is done

Forecast Influenza Deaths in US by State population (without seasonality)



Insights & Recommendations

States with a bigger population and with higher share of vulnerable populations are more likely to be most affected

While BIG States like CA,TX,NY, FL, and PA have the highest death counts, smaller states like WY, AK, DC, VT, ND Alaska, actually have a higher death rate, so both factors need to be taken into account when planning to prevent future tragedies

Areas with larger vulnerable population need more funding, since there is a direct correlation between the vulnerable population and death counts

Consider early intervention and prevention programs (vaccinations and hygiene awareness) for states with relatively low population of > 65 years old but with high death rates.

Full report can be found here

GameCo

A video game company

OBJECTIVE

Business Intelligence on video game market as input for GameCo's new marketing strategy and budget plan

DATA

Video Game Sales from VGChartz. Click here for downloadable data

SKILLS

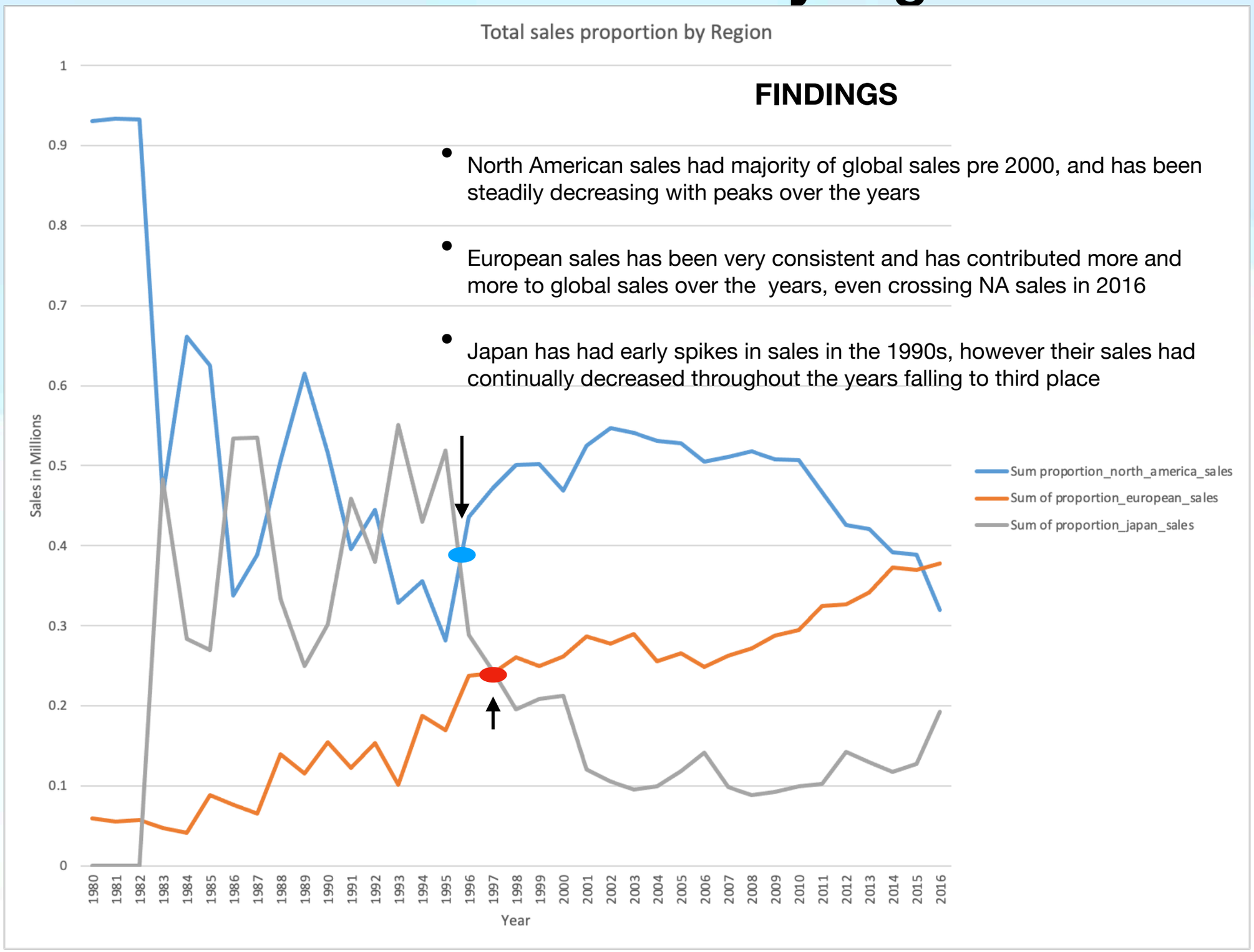
- Data Cleaning
- Data grouping
- Data Visualization
- Descriptive Analysis

TOOLS

Excel

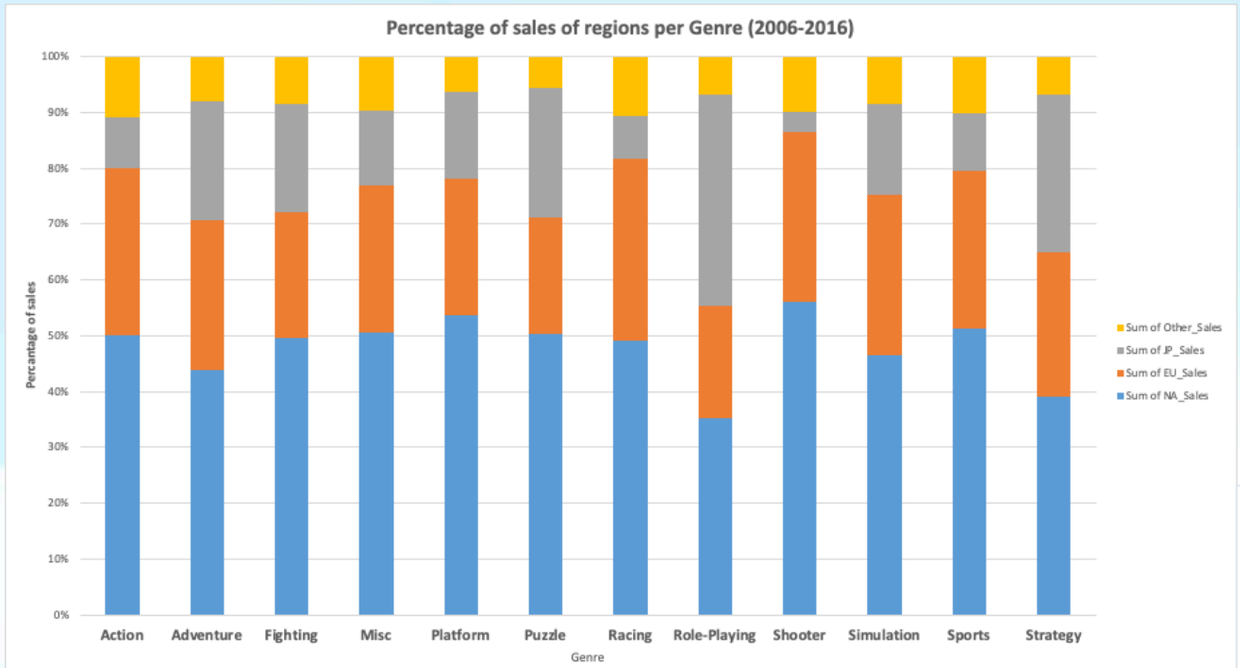
Video game sales (1980 - 2016)

Historical trend of total sales by region



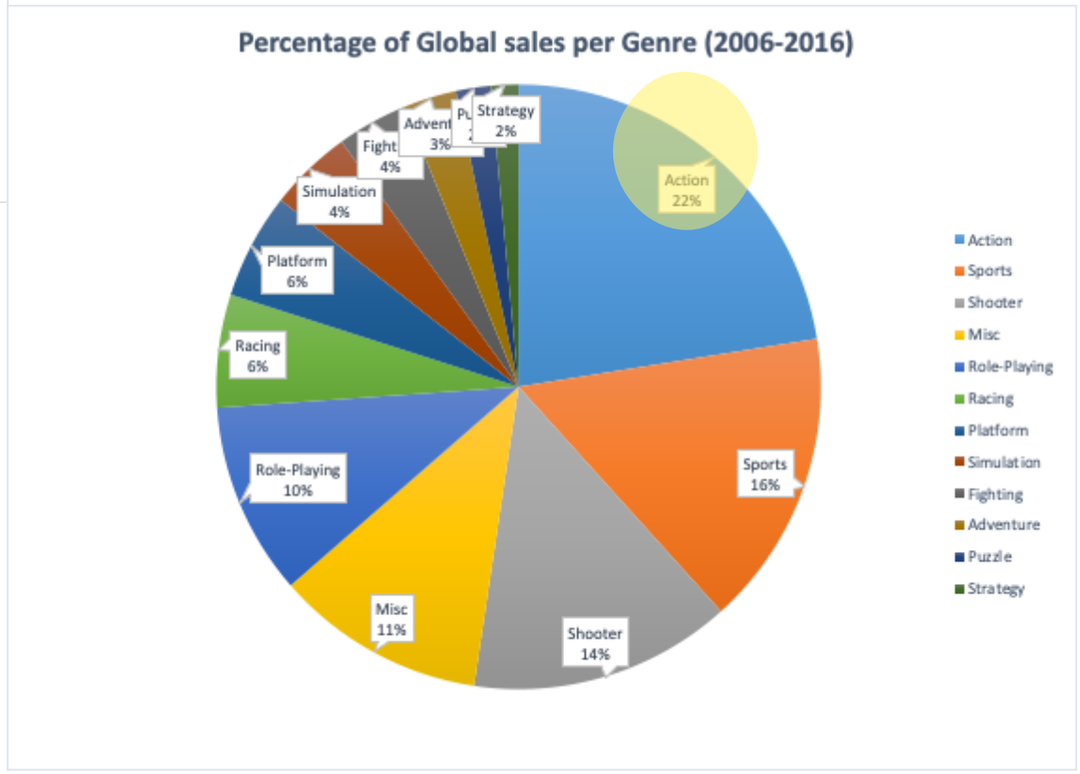
Most Popular Genres

Percentage of sales by genre



- Conducted data cleaning by filtering out N/A values, typos, and duplicates
- Created a pivot table and conducted further descriptive analysis, looking for outliers
- Using a pivot table I created a line graph looking for the total sales per region for the years 1980 to 2016

- NA sales have over 50% of market share in Action, Platform, Shooter, and Sports genre
- EU sales have the most sales in Action, Sports, and Misc.
- JP sales have a noticeable market share in Role-Playing, and Action games



Insights & Recommendations

I would recommend GameCo continue investing into the Europe region in order to capture more of the accessible market share

Specific Genres perform significantly better in certain regions compared to others

Role Playing games perform the best in the Japanese markets, While Sports and Shooter genres make up the majority of sales for the North American market

Full report can be found here

THANK YOU

NGAWANG DHUNDUP

GitHub

+ableau