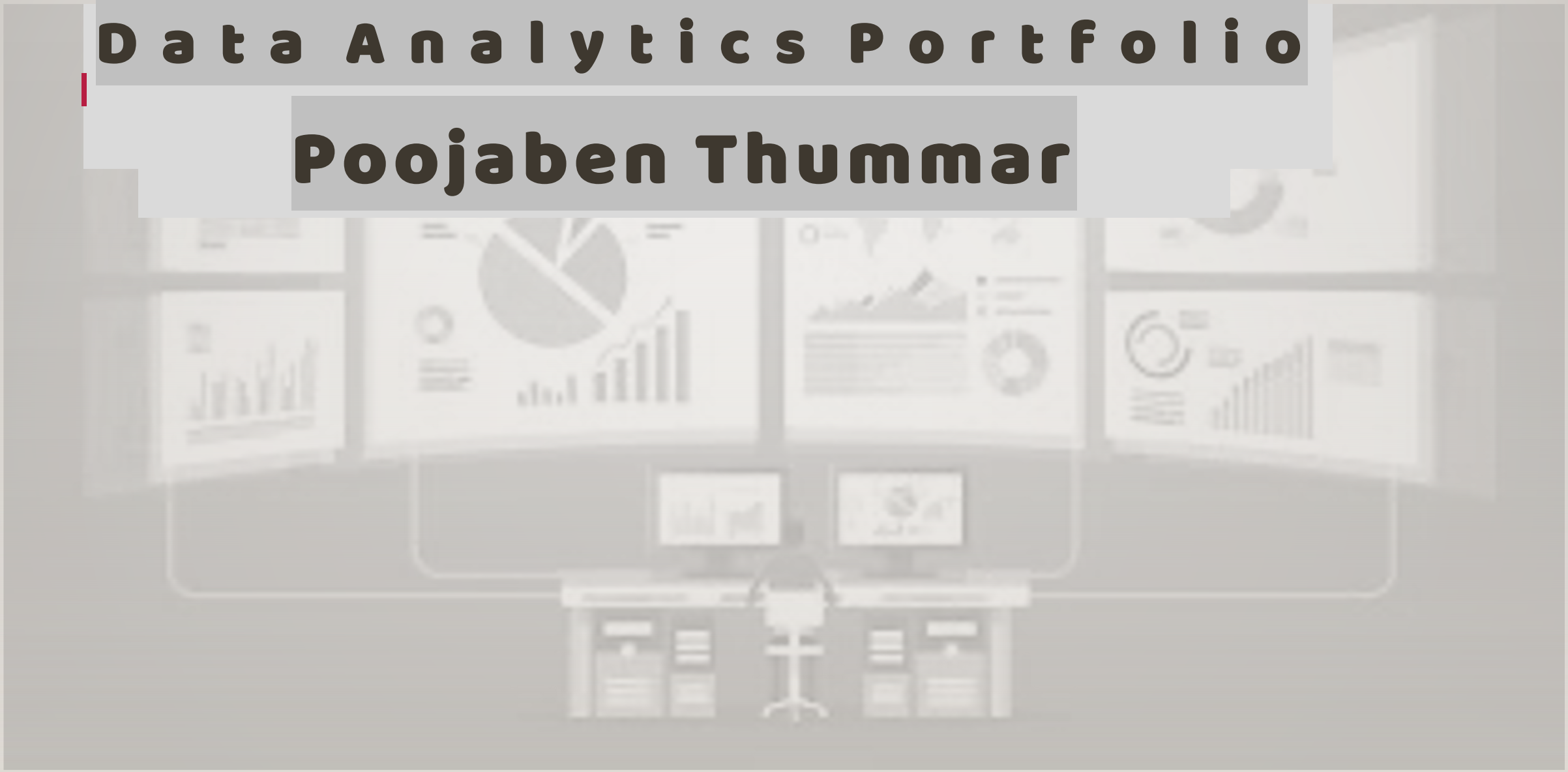
Data Analytics Portfolio

Poojaben Thummar



PROJECTS

1. Global market analysis of video game sales

2. Seasonal Influenza Forecast Staff deployment in preparation for Flu season

3. Rockbuster Stealth LLC Movie rental market analysis for launching a new online video service

4. Instacart Market segmentation analysis to uncover sales

5. Pig E. Bank Customer attrition

6. cancer a disease

1. GAME CO.

PROJECT OVERVIEW:
Game Co. is a global gaming company with market America, Europe, Japan, and other Country, offering its customers a diverse range of games for sale or rent across various categories.

TOOLS USED: Excel, PowerPoint

DATA SETS:
The data set for this project is primarily video games sales across different platforms, genres, publisher and publishing studios.

DATA LIMITATIONS:
Data set represents unit of games sold, not their dollar value.

KEY QUESTIONS
- Are certain types of games more popular than others?
- Have any games decreased or increased in popularity over time?
- How have their sales figures varied between geographic regions over time?

DESCRIPTION:
For this project, the main objective was to uncover significant trends and patterns by analyzing historical data from a game dataset. The aim is to glean insights into the potential market reception for upcoming games.

1. GAME CO.

KEY STEPS:

- Data cleaning: find and address missing values and duplicates;
- Exploratory Data Analysis (EDA): to calculate for mean , meadian ,mode,max,min
- Data grouping, filtering, and summarizing using Pivot Table
- Data Charts: Create a column chart of total global sales by publisher; Line chart of the average North American sales by year
- Decide on what type of analysis: Descriptive analysis, Diagnostic, Predictive or Prescriptive
- Interpret results and summarize findings

Excel functions used: find & select values, replace with, replace all, remove duplicates

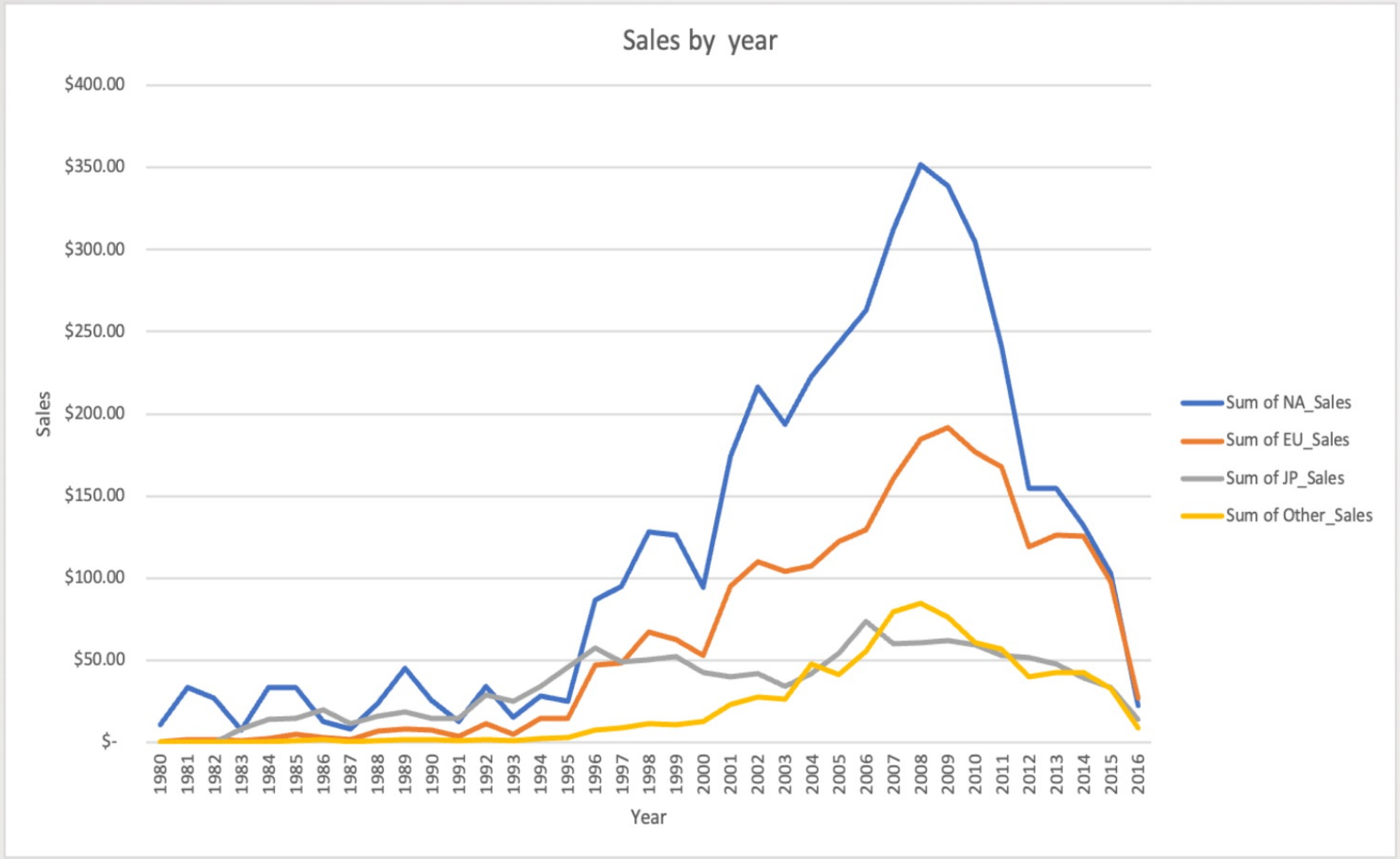
GOALS

Analyze regional and temporal sales trends for informed future decision-Making.

SKILLS:

- Data cleaning
- Data grouping and summarizing
- Descriptive analysis
- Pivot table
- Visualization charts in MS
- Excel/PowerPoint

1. GAME CO.



Data Analysis

1980 – 1995: Global Sales were under $50 million ever single year.

1995: The Rise of North America begins.

1995 – 2008: North American sales rise by over $300 during this time. Europe rises strong as well.

2008 - 2015: The decline in sales from all regions.

2015-2016: North America has always been the consistent leader in sales since 1980, but Europe is set to surpass North America in sales in the next few years.

1. GAME CO.

Early 80's:

North America dominated the market until 1983 when Japan entered the market.

Mid 80's to Mid 90's:

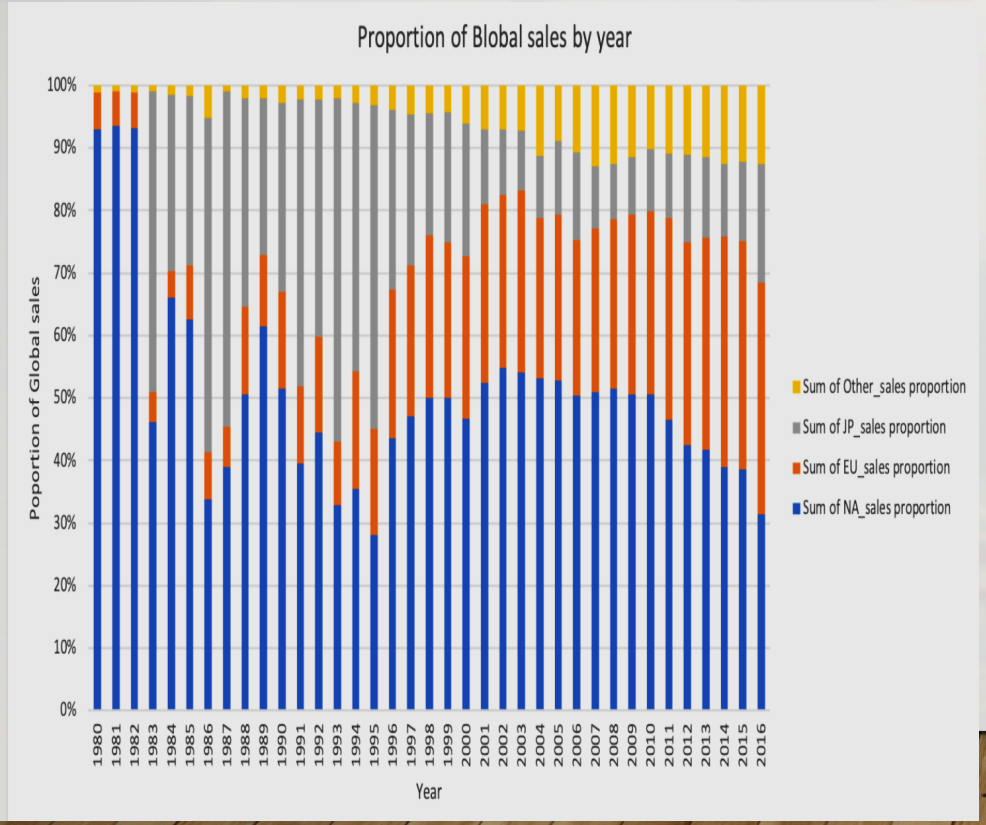
Japan and North America continue to compete for the top spot.

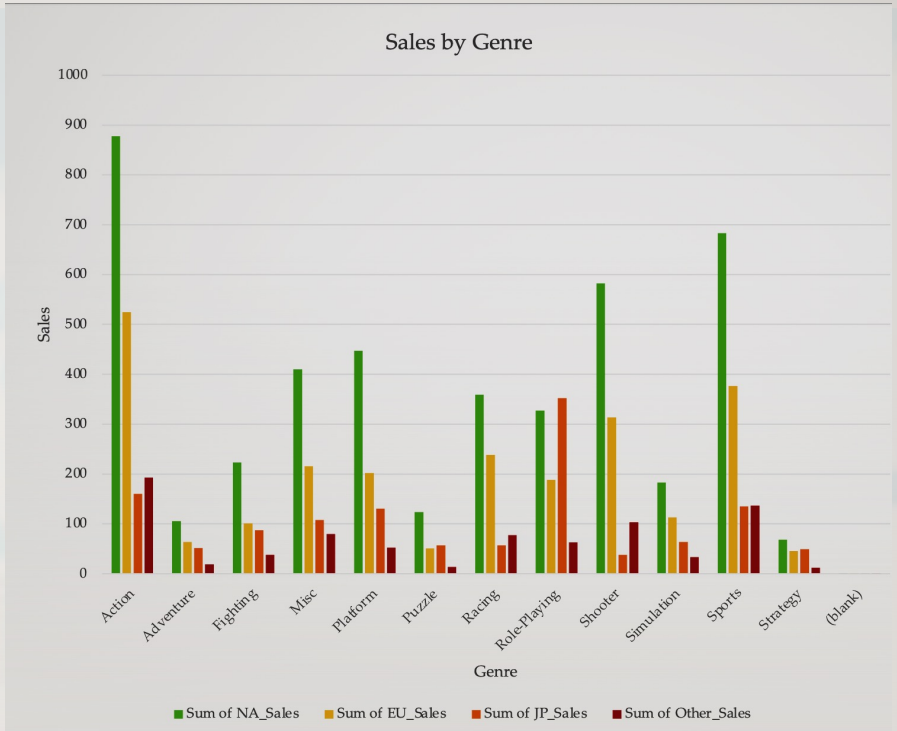
1996:

Japan's great run ends. Europe gains momentum. North America stays steady.

1996 to 2016:

Japan's portion continues to decrease. Europe gains closer and closer to North America. Other countries region continues growth.

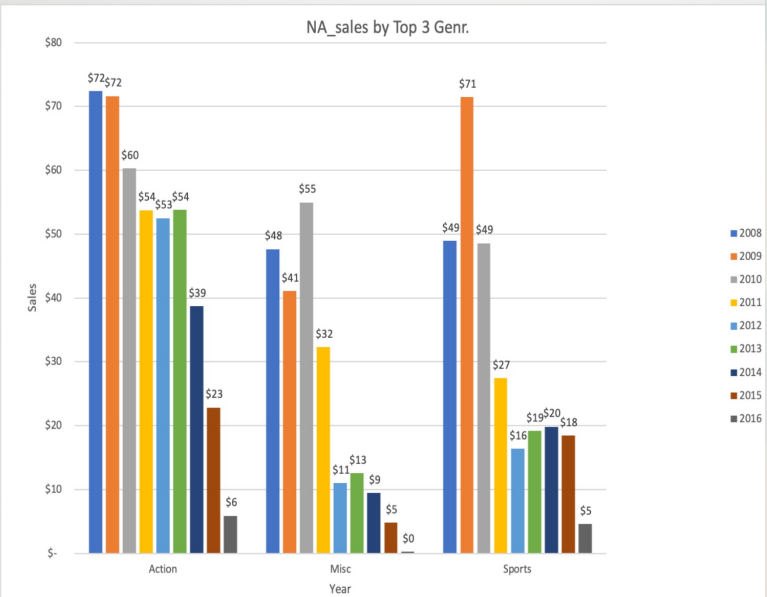




Top 3 genre. By 4 Regions North America, Europe, Japan and Other country(2008-2016).

- Action
- Misc
- Sports

1. GAME CO.



Action Sales continuous drop: Over $72 million in sales during 2008. 2008 is North America's best year of sales. In 2016 Action Sales had dropped down to $6 million.

Misc Sales:

In 2008 sales is $48 million . Year 2010 is highest sales to $48 million.

In 2016 North America have 0 sales.

Sports Sales:

2008: $49 million. Higher $71 million in sales year 2009. 2009 is North America's best year oh sales. 2015 is dropped year with only $5 million lowest sales.

Action Sales:

2008 : $39 million

Highest sales: in 2013 $45 million 2016 : $6 million

Misc Sales:

2008 : $21 million

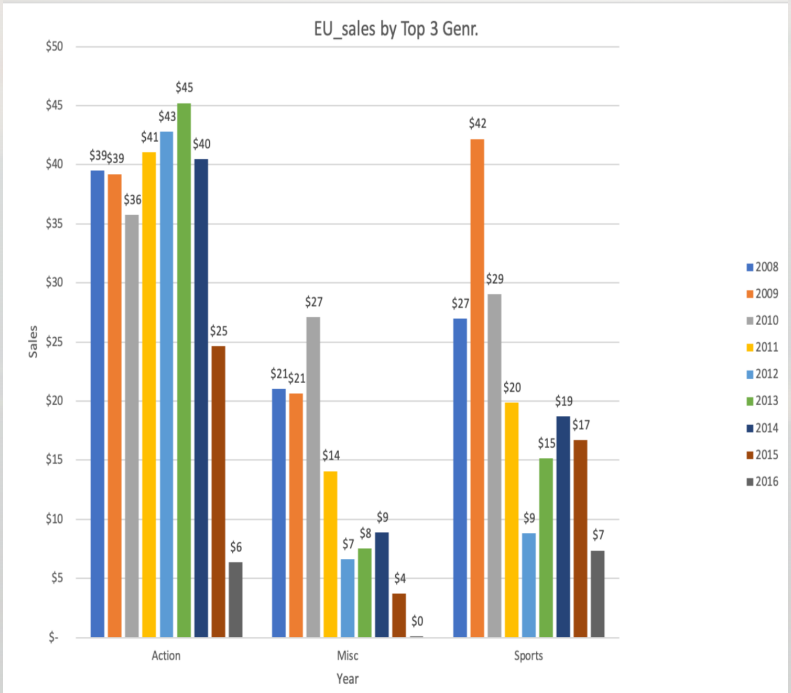
Highest sales : in 2010 $27 million 2016 : $6 million

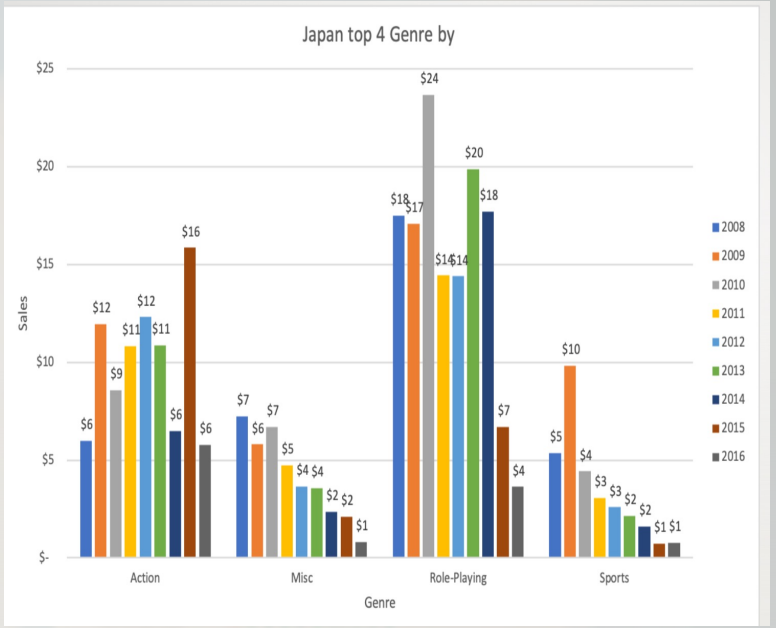
Sports Sales:

2008: $27 million

Highest sales : in 2009 $42 million 2016 : $7 million

We could say that in Europe have most popular genre is Action .





Action Sales:

2008 : $6 million sales

Highest sales: in 2025 $16 millions ales 2016 : $6 million sales

Misc Sales:

2008 : $7 million

Highest sales : in 2010 and 2008 $27 million sales 2016: $1 million sales

Sports Sales:

2008: $5 million sales

Highest sales : in 2009 $10 million sales 2016: $1 million sales

But here is surprising data is that in Japan only Role-playing genre is more popular and famous. Japan have Maximum sales Over the $2o million in year 2010.

1. GAME CO.

RECOMMECATION

North America	• North America can get its momentum back if the focus is shifted towards Action, Sports and Misc. titles. These 3 areas could make or break the N.A. region. • Gameco should maintain their support for North America's video game market.
Europe	• Europe's video game market sales is steadily on the rise, even surpassing North America sales in 2016. • Europe's growth can continue by focusing on it's well-rounded support of multiple genre titles.
Japan	• For Japan's video game market and continue to support the genre Role-playing since it is Japan's most popular genre. Gameco should more focus on genre Role-playing and genre sport as well in Japan.

PROJECT OVERVIEWL

During flu season, U.S. healthcare facilities experience an influx of patients. This project aimed to create a staffing schedule that addresses the critical staffing demands during the flu season.

Goal: Enhance preparedness for the influenza season by effectively managing staffing needs at clinics and hospitals served by the medical staffing agency.

Motivation: The United States has an influenza season where more people than usual suffer from the flu. Some people, particularly those in vulnerable populations, develop serious complications and end up in the hospital. Hospitals and clinics need additional staff to adequately treat these extra patients. The medical staffing agency provides this temporary staff.

Scope: The scope of the analysis encompasses examining historical influenza data, including incidence rates, peak seasons, and regional variations, to forecast staffing needs across the country. It involves identifying patterns, seasonal fluctuations, and high-risk periods for influenza outbreaks. Additionally, the analysis will assess the implications of influenza trends on the demand for temporary healthcare workers, considering factors such as clinic size, patient demographics, and geographic location.

Stakeholder Identification:

- Medical agency frontline staff (nurses, physician assistants, and doctors)
- Hospitals and clinics using the staffing agency services
- Influenza patients
- Staffing agency administrators

TOOLS USED: Excel, PowerPoint, Tableau

KEY QUESTIONS

- When is flu season at its peak?
- Who is most vulnerable?
- Which states have the most residents in vulnerable populations?

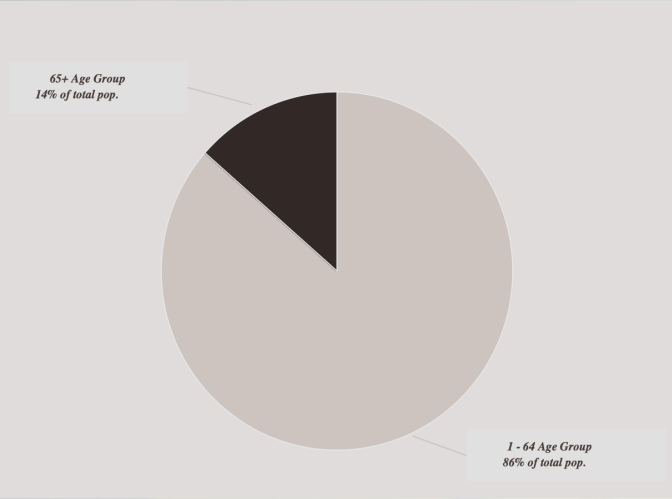
KEY STEPS:

- Conduct Exploratory Data Analysis (EDA), including:
- Descriptive/Summary Statistics: Central tendency & distribution.
- Creating charts to understand distribution, relationships, missing values, and outliers.
- Data cleaning (including imputation) and transformation (merging spreadsheets).
- Hypothesis testing (t-test in Excel).
- nterpret results and summarize findings and summarize findings
- EXCEL FUNCTIONS USED: FIND & SELECT, REPLACE WITH, REPLACE ALL, REMOVE DUPLICATES; used PivotTables and SUMIFS for summarizing and aggregating data

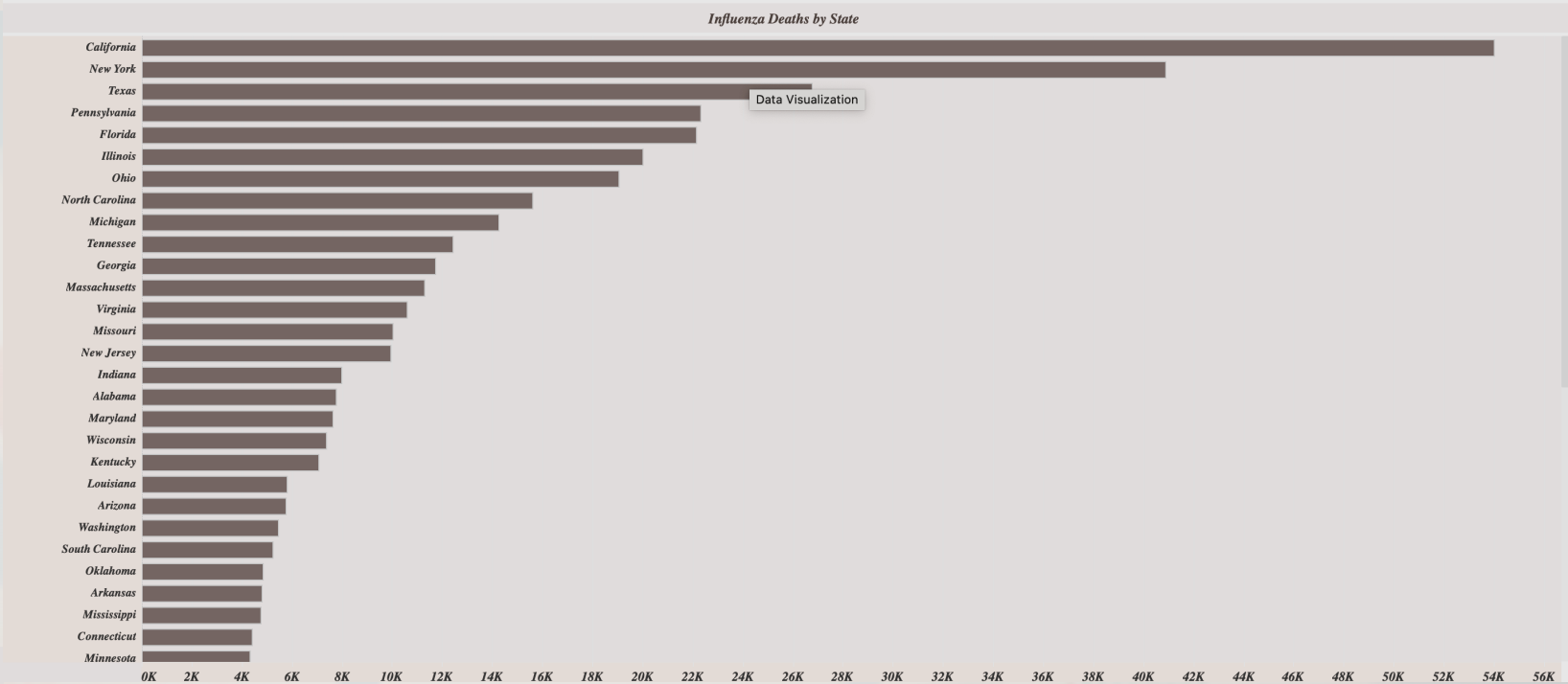
SKILLS

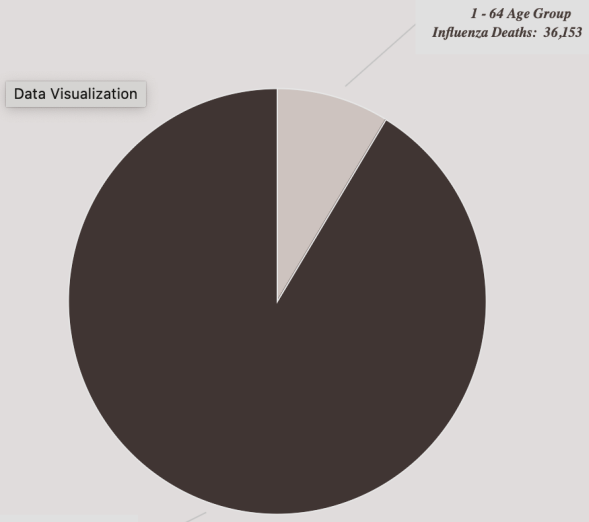
- Data cleaning, integration, and transformation.
- Statistical hypothesis
- Testing
- Visual analysis
- Forecasting
- Storytelling in Tableau
- Presenting results

2. SEASONAL INFLUENZA FORECAST



Most vulnerable population is adults aged 65+



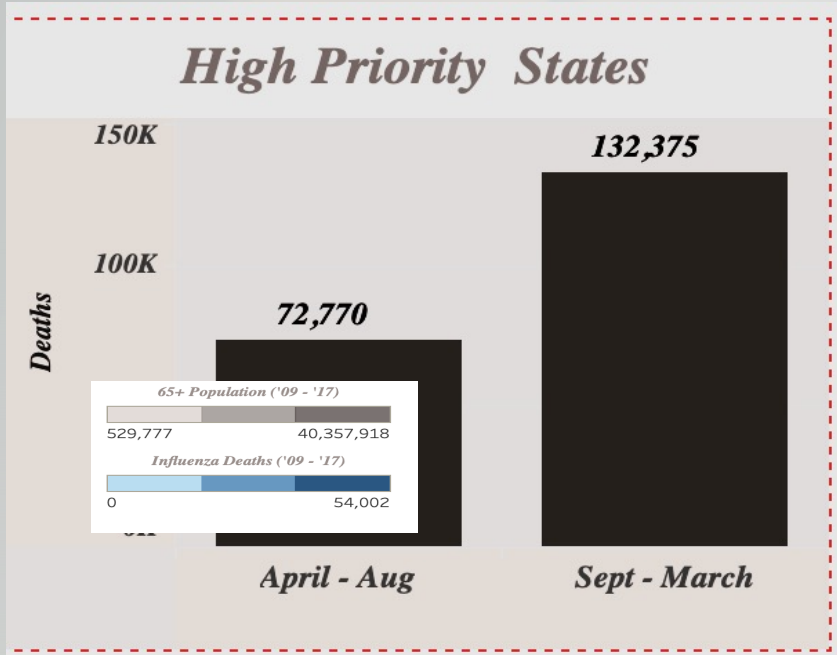


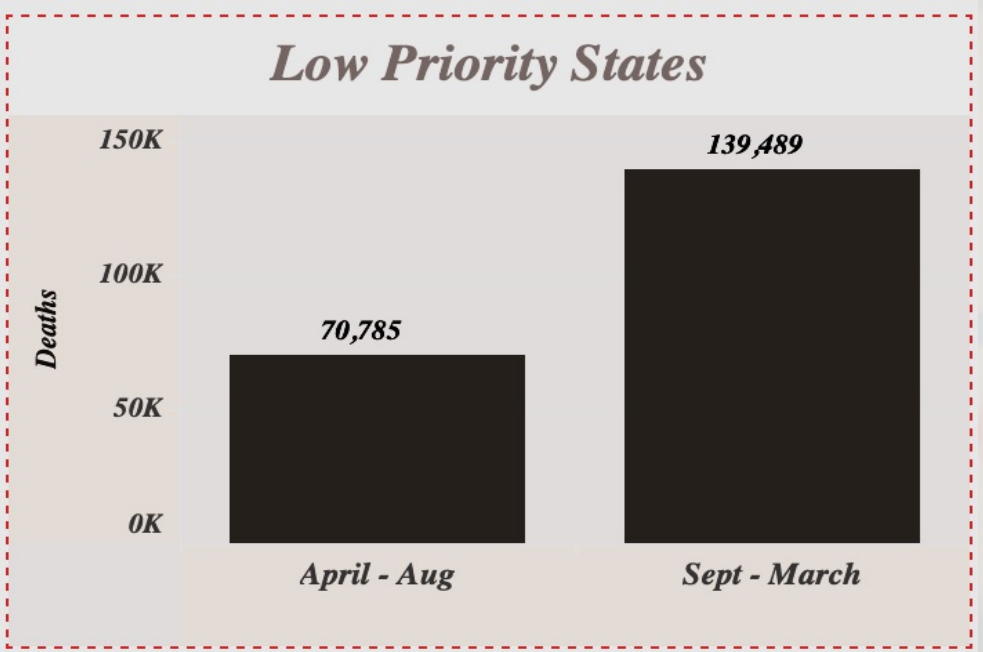
Most vulnerable top 2 states are California, New York, Texas, Pannsylania and Florida.

	Mortality 65+ Years	Population 65+ years
Mean	826	806566
Standard Deviation	1014	880783

	Mortality 65+ Years	Population 65+ Years
Correlation Coefficient	0.94	0.29
Strength of Correlation	Strong relationship	Weak relationship

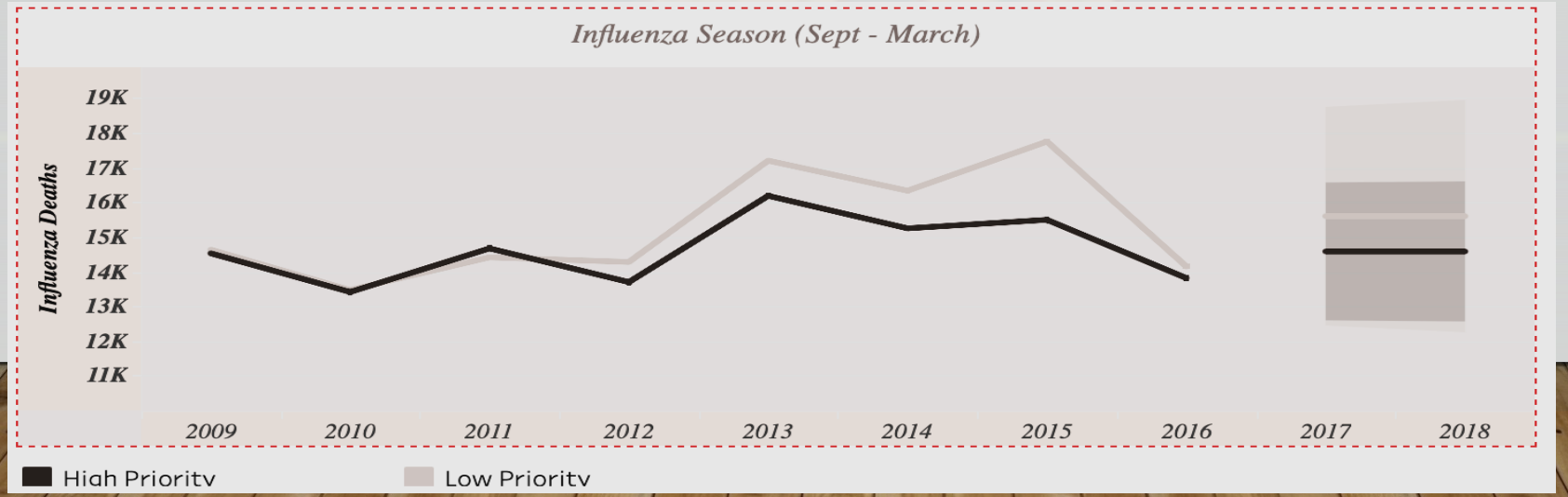
2. SEASONAL INFLUENZA FORECAST





Forcast:

The line charts to the right shows the influenza death forecast for 2018, separated by 'non-influenza season' to 'influenza season'. We can see that during both times of the year the deaths are expected to stay the same. Per this information, we should not see an influx of influenza deaths from the previous year (2017).



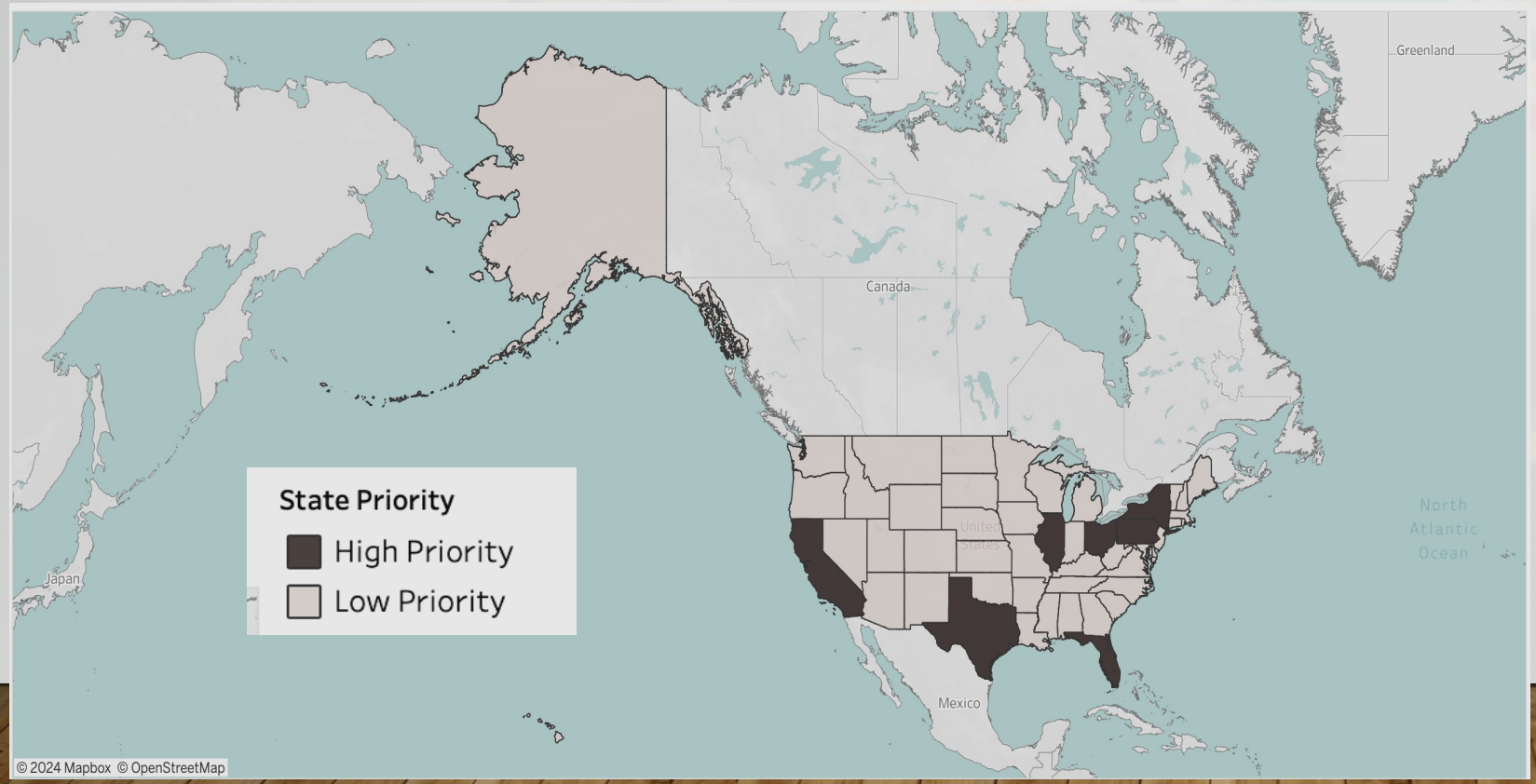
Recommendations

Based on the locations of a higher populas of the vulnerable age group (65 and older) and seven states with higher influenza death counts, it is recommended to staff a larger medical staff in the High Priority states listea below and maintain the normal staffing requirements for the rest of the Unite States. The staffing adjustments will begin in September and end in March

High Priority: California, Florida, New York, Texas, Pennsylvania, Ohio, Illinios

Data Visualization

Low Priority: Remainder of the United States



Staffing

On the left shows the increase in influenza deaths from 'non-influenza season' to 'influenza season'. For both state levels, the influenza death count nearly doubles. We can assume the patient visit would be on a similar trajectory

Per the World Health Organization, there should be 2.5 physicians or nurses per 1,000 people.

This should give an estimate of how many medical staff needed for the influx of patients.

3. Rockbuster Stealth LLC

Project Overview :
Strategic Analysis and Data-Driven Recommendations for Rockbuster Stealth LLC
Background:
Rockbuster Stealth LLC, a global movie rental company, has traditionally operated throughphysical stores. Faced with rising competition from digital streaming platforms like Netflix and Amazon Prime, Rockbuster is exploring a shift to an online video rental model.
Objective:
This presentation aims to leverage data analytics to answer critical business questions that will inform Rockbuster's strategy for 2020 as they transition to an online service model. These insights will guide decision-making processes and strategic planning.

TOOLS USED: Excel, DBVisualizer, SQL/PostgreSQL

KEY QUESTIONS
- Do sales figures vary between geographic regions?
- Which countries are Rockbuster customers based in?
- Which movies contributed the most/least to revenue?
- What was the average rental duration for all videos?

DATA SETS
Population (U.S. Census) & Influenza Deaths (U.S. CDC)

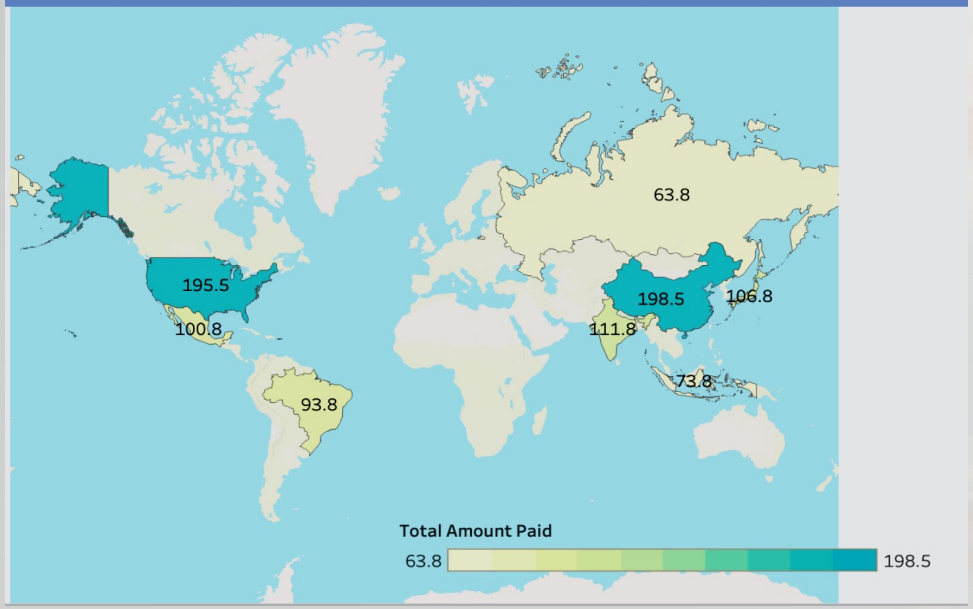
SKILLS
- Relational databases
- SQL
- Database querying, filtering, cleaning,and summarizing subqueries
- Common Table expressions (CTEs)

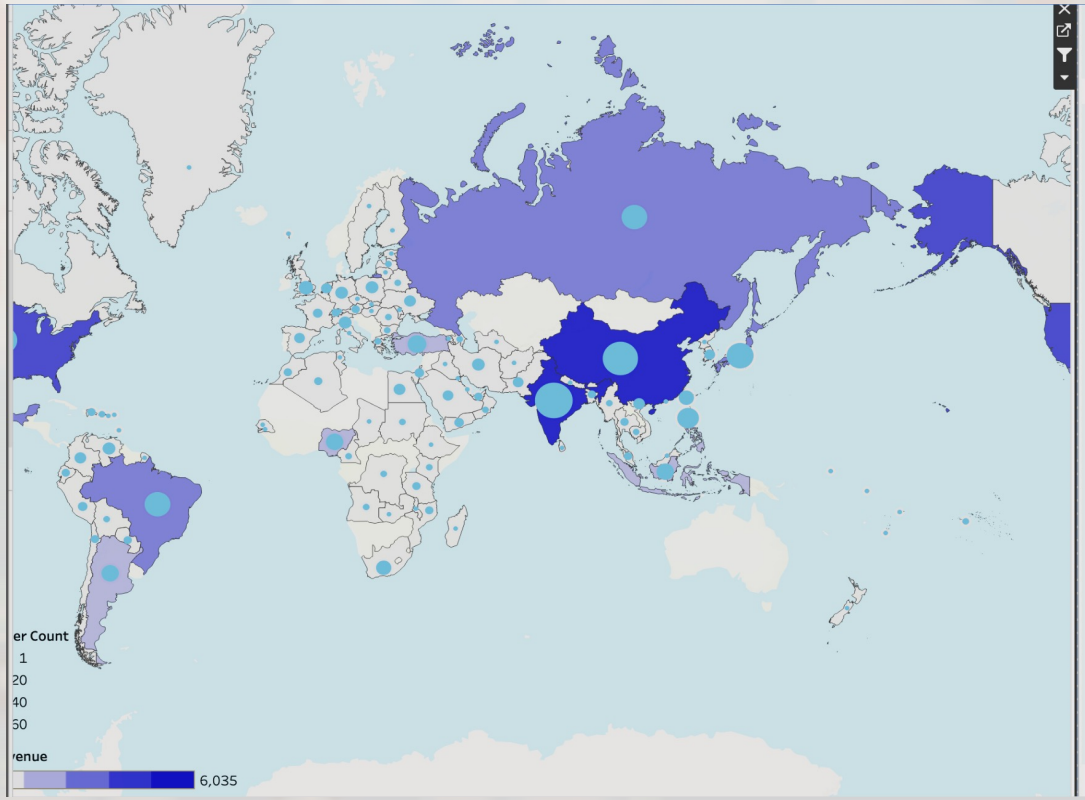
KEY STEPS
- Data Extraction: Used SELECT, JOIN, and WHERE clauses to extract relevant data from multiple tables.
- Data dictionary: Created an Enterprise Relationship Diagram (ERD) using DBVisualizer.
- Extract Enterprise Relationship Diagram (ERD)
- Extracted ERD from Rockbuster's database using DBVisualizer
- Data Aggregation: Employed GROUP BY and aggregate functions like SUM, AVG, and COUNT to summarize data.
- Data Filtering: Utilized HAVING clauses to filter aggregated results.
- Data Visualization: Created bar charts, line charts, scatter plots, among others.

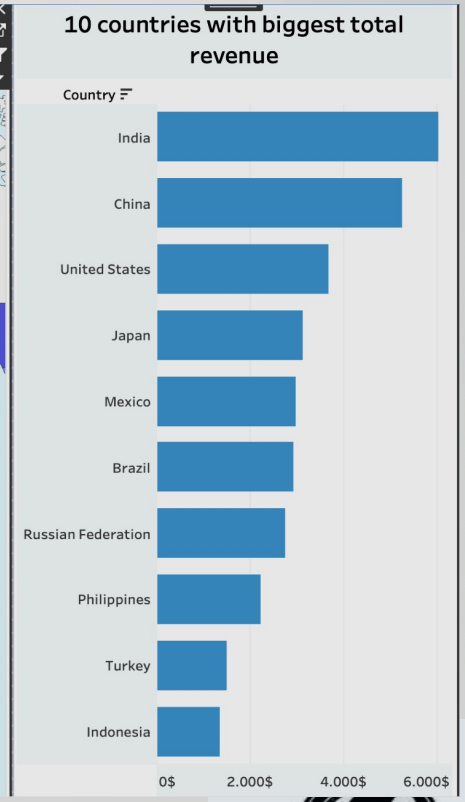
3. Rockbuster Stealth LLC

What is the average rental duration for all films?

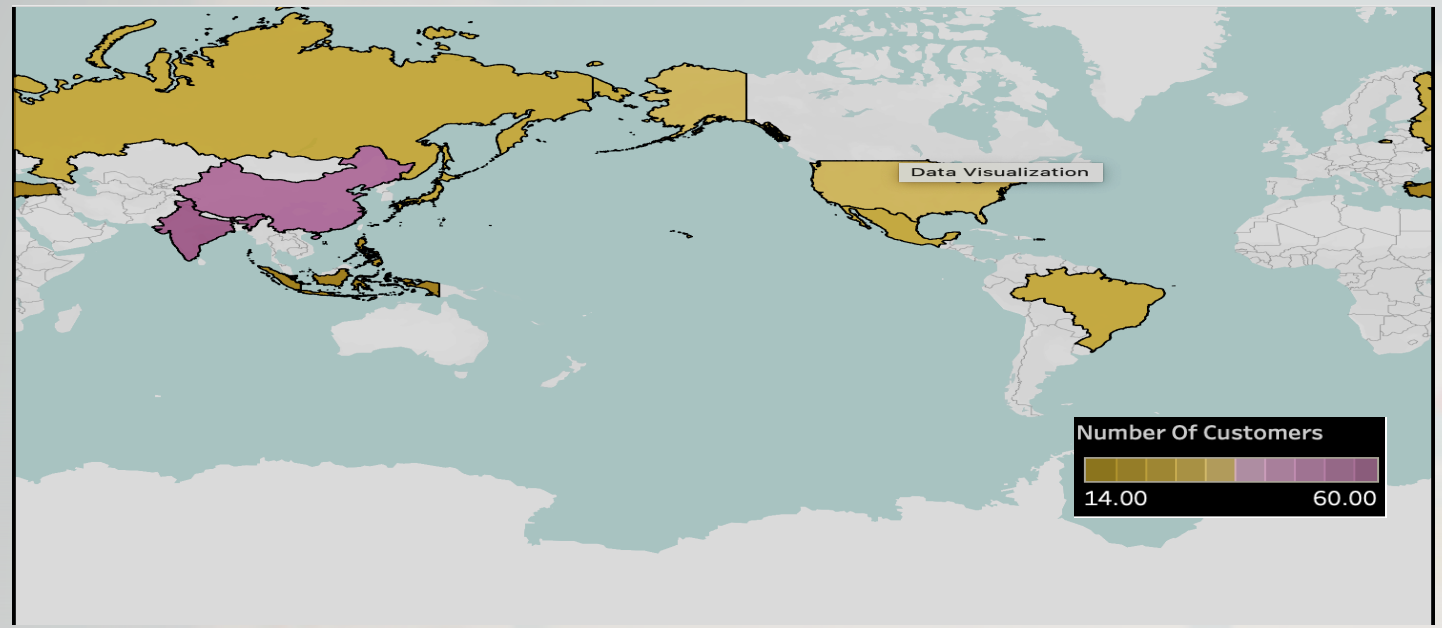
Where are customers with a high lifetime value based?





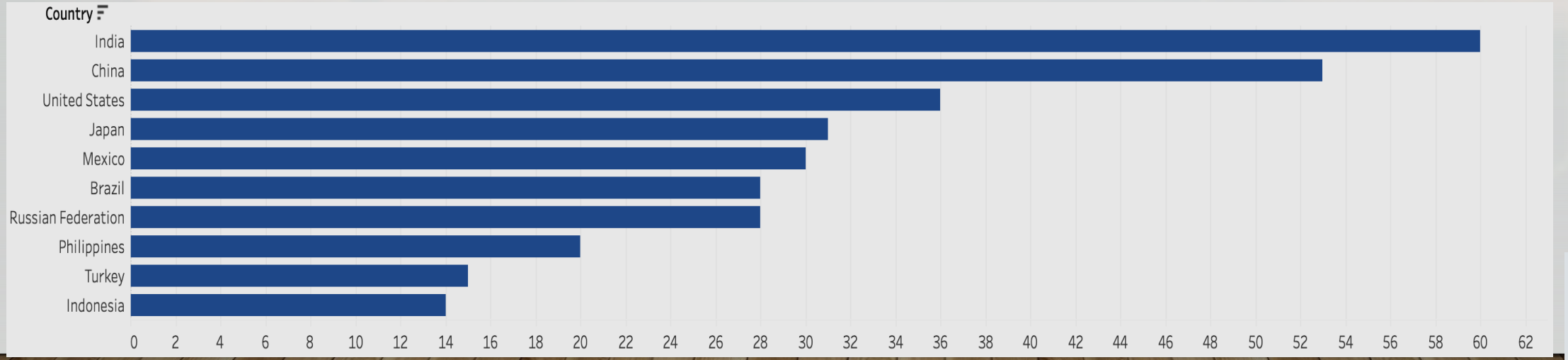


3. Rockbuster Stealth LLC



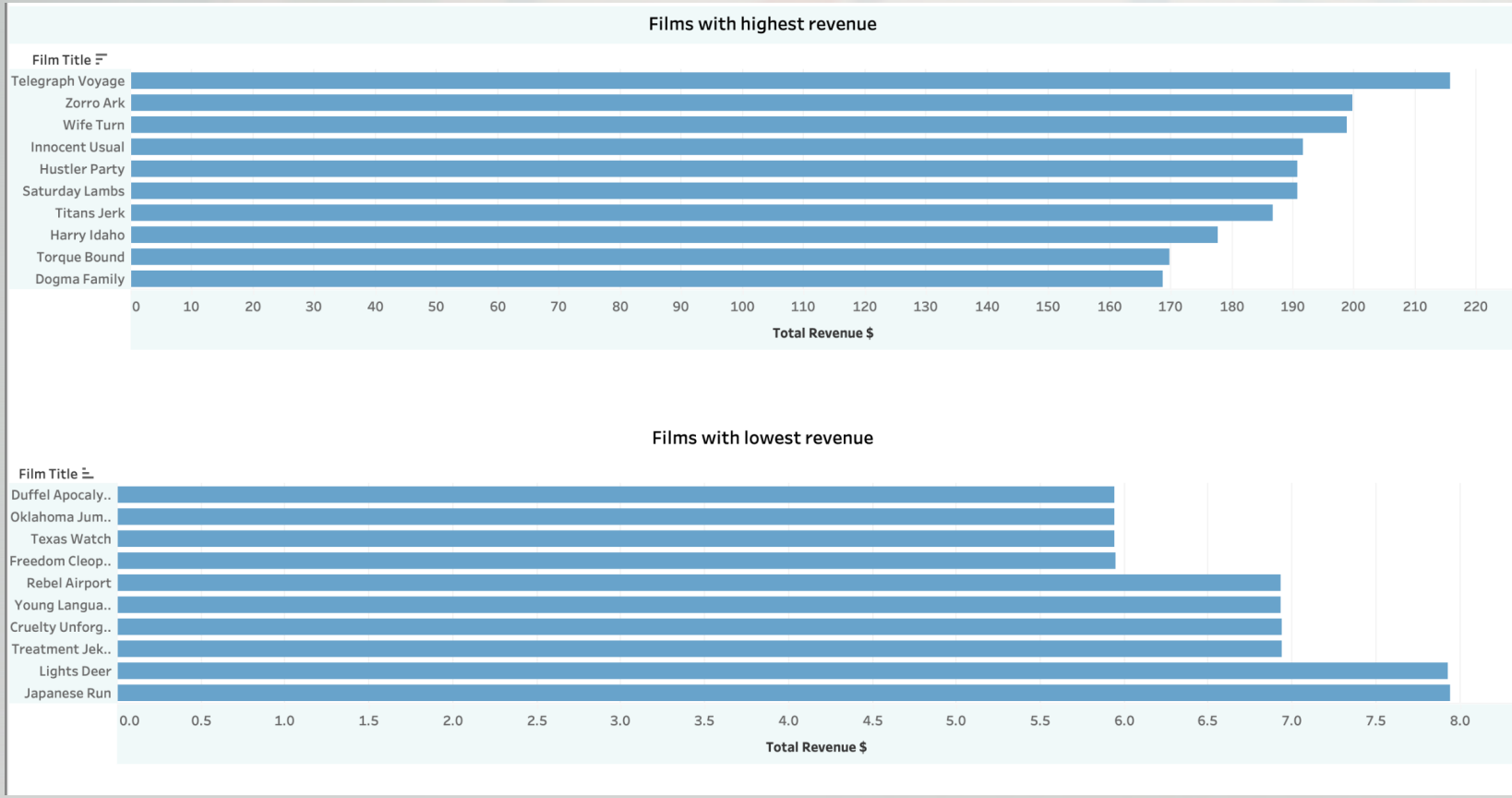
India and China are the countries with most customers and the highest revenue.

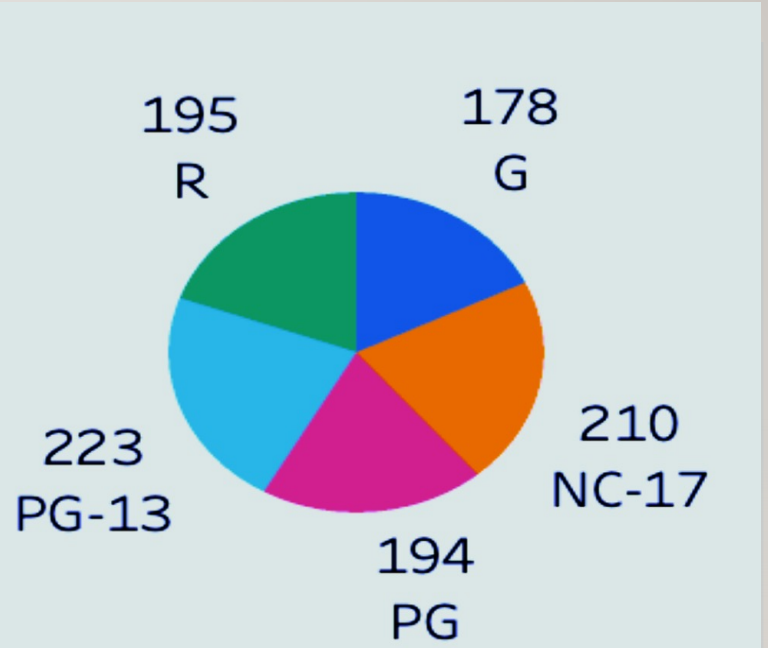
- **Top 10 countries:**
- India, China, US, Japan, Mexico, Brazil, Russia, Philippines, Turkey and Indonesia



3. Rockbuster Stealth LLC

Which movies contributed the most/least to our revenue?





What is most common Movie Rating?

RECOMMENDATIONS

- Concentrate on marketing efforts in countries with the highest customer base on largest popularity. This involves analysing market data to identify regions with the largest customer concentrations and tailoring marketing campaigns to effectively reach and engage these key demographics.

- Create customized marketing strategies and loyalty programs for high-value customers For instance ,Festival offer and attract mid-tier customers with specific promotions and incentives.

- The most purchased movie genres are SPORTS, SCI-FI, and ANIMATION. Therefore, concentrating on producing more films in from these categories would boost revenue.

- Feature new movies, as well as movies from last 10 years.

- Feature movies of various lengths and consider the data for average length of rented film i.e. 115 minutes

- Expanding the availability of movies in multiple language, based on the preferences of different countries can significantly contribute to revenue growth. Therefore, it is crucial to prioritize the development of language-specific versions for movies.

4.Instacart Merket:

PROJECT OVERVIEW

Instacart, a leading online grocery storestrategic, enables customers to order groceries through an app. By analyzing historical data, the company aimed to boost sales by adopting a more approach to targeting customers and improving segmentation.

TOOLS USED: Jupyter, Python, Anaconda, Python Libraries

KEY QUESTIONS

- What are the busiest days of the week and hours of the day?
- know whether there are particular times of the day when people spend the most money, as this might inform the type of products they advertise at these times.
- simpler price range groupings to help direct their efforts.

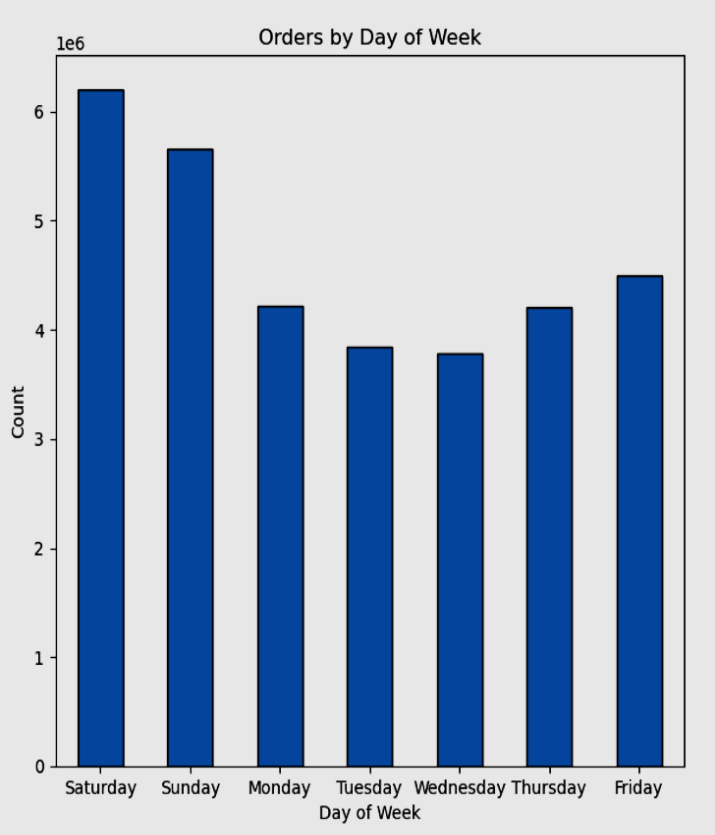
KEY STEPS

- Write Python Scripts: Used Jupyter Notebook to write Python scripts and document the analysis.
- Import: Imported datasets and necessary Python libraries.
- Data cleaning and data wrangling: Utilized Pandas and NumPy for data cleaning, transformation, and preparation.
- Agrregastion Functions
- Export:eported dataset to pickle file and CSV file.

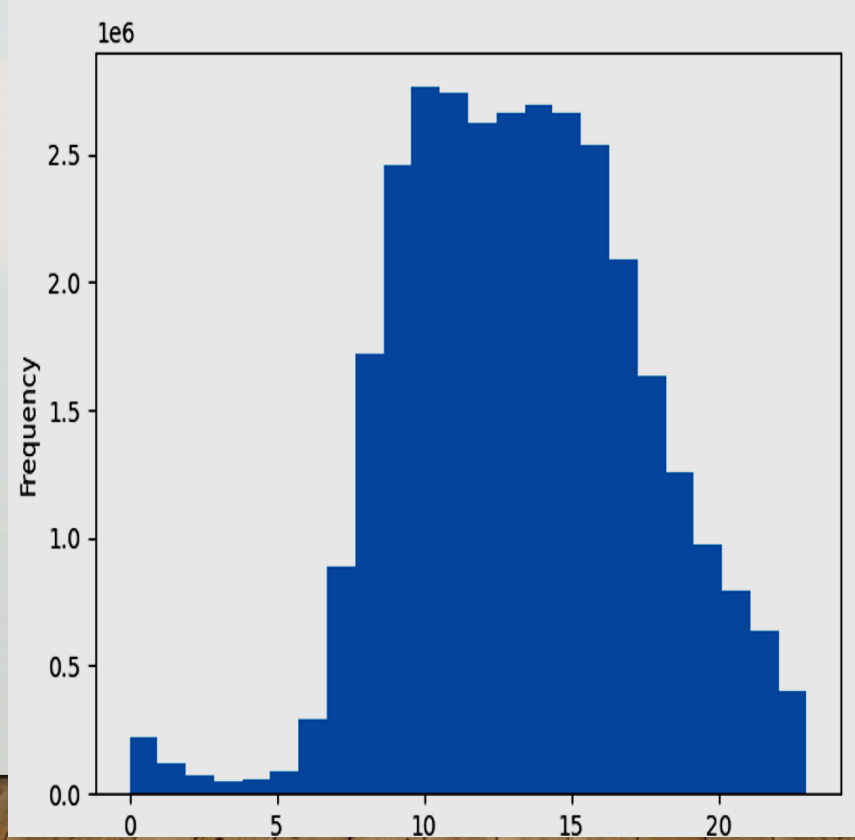
Dictionary: performed data quality and consistency checks, including frequency counts.
Exploratory Data Analysis (EDA): Derived new variables, data manipulation, grouping, and aggregating.
Visualization: Created graphs and visualizations using Matplotlib, Seaborn, and Scipy.
Export data frames: Exported cleaned and processed data frames for reporting.

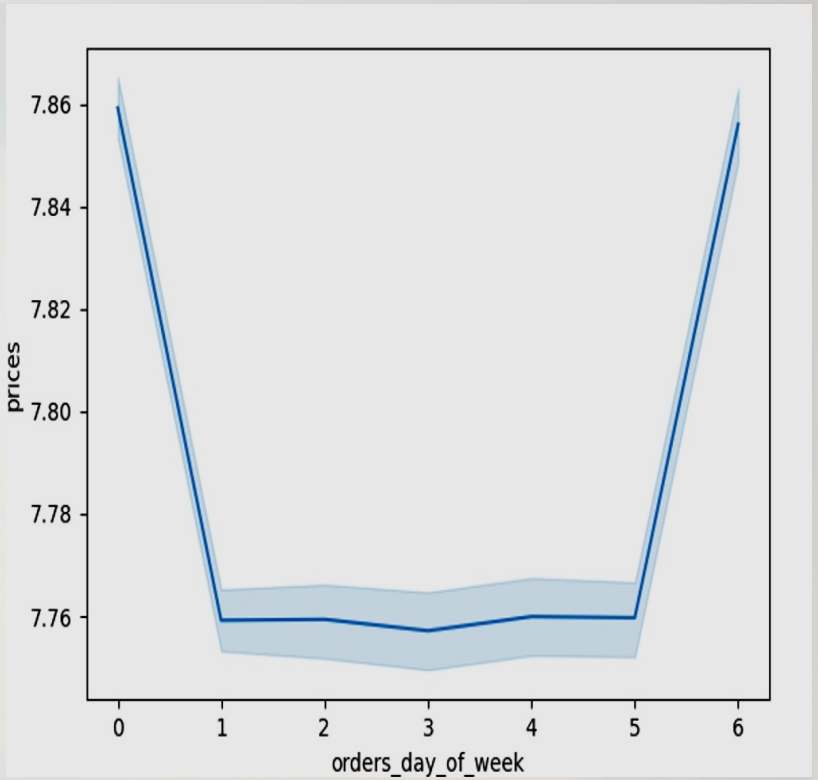
4.Instacart Merket:

Saturday and Sunday are busiest day in the week



The hours of 10am to 3pm with the fewest amount of order in the early mornings from 12am to 5am. After 5am, orders begin to see an increase in activity to their peak times of 10am to 3pm and then begin to decline after 4pm.

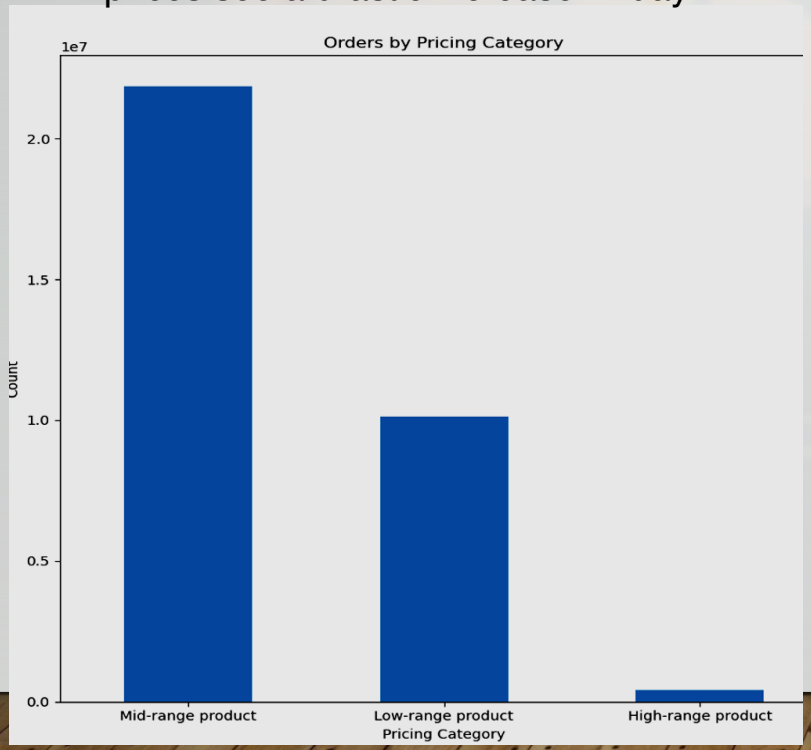




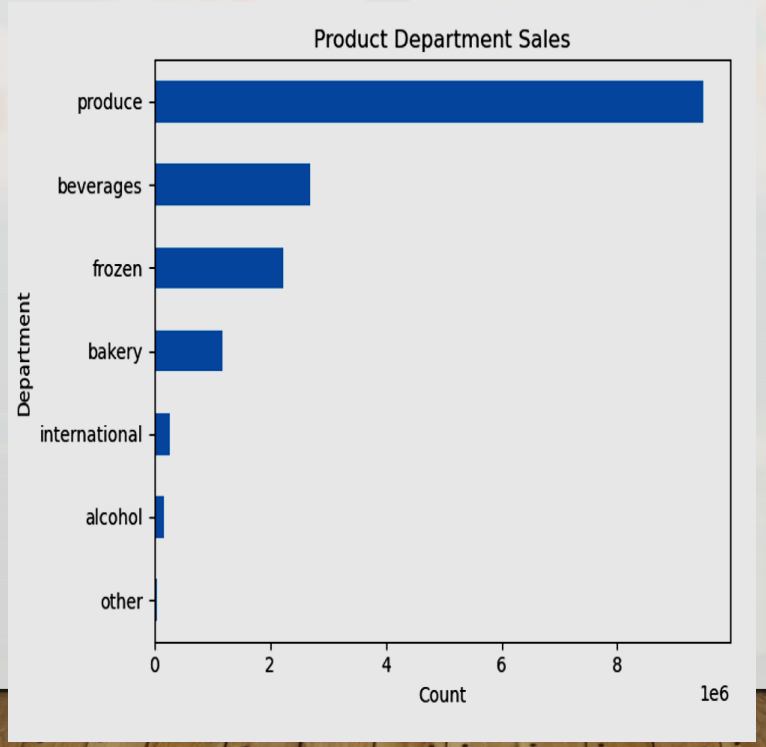
highest peaks on Friday and Saturday with a sharp decline on Sunday that continues through Monday. After Monday, we see a gradual increase from Tuesday where prices stay steady between Tuesday and Wednesday and begins to see a gradual increase for Thursday until prices see a drastic increase Friday.

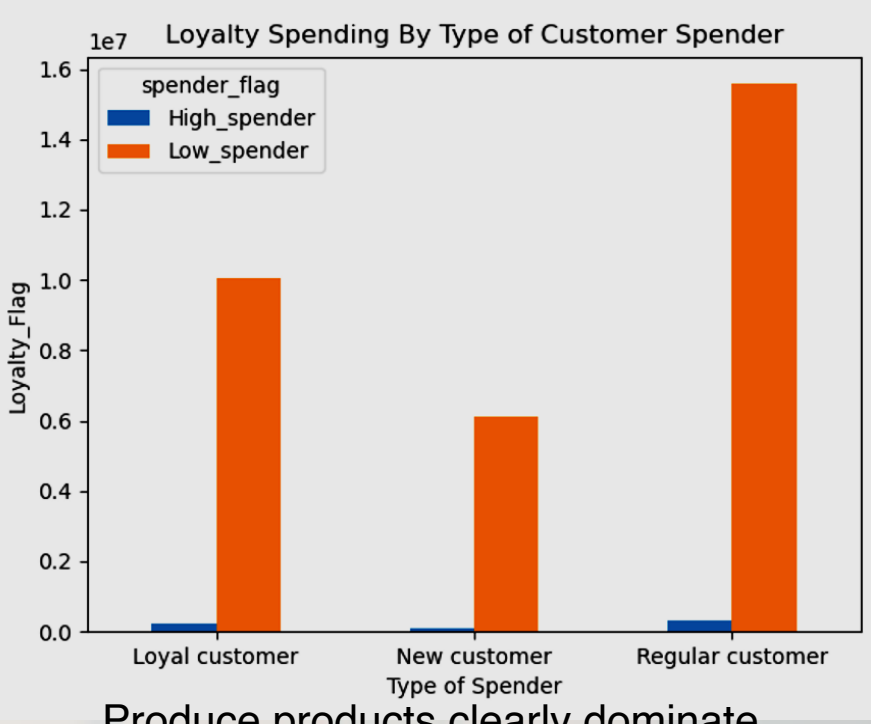
4.Instacart Merket:

highest peaks on Friday and Saturday with a sharp decline on Sunday that continues through Monday. After Monday, we see a gradual increase from Tuesday where prices stay steady between Tuesday and Wednesday and begins to see a gradual increase for Thursday until prices see a drastic increase Friday.



Produce products clearly dominate product sales over any other department.

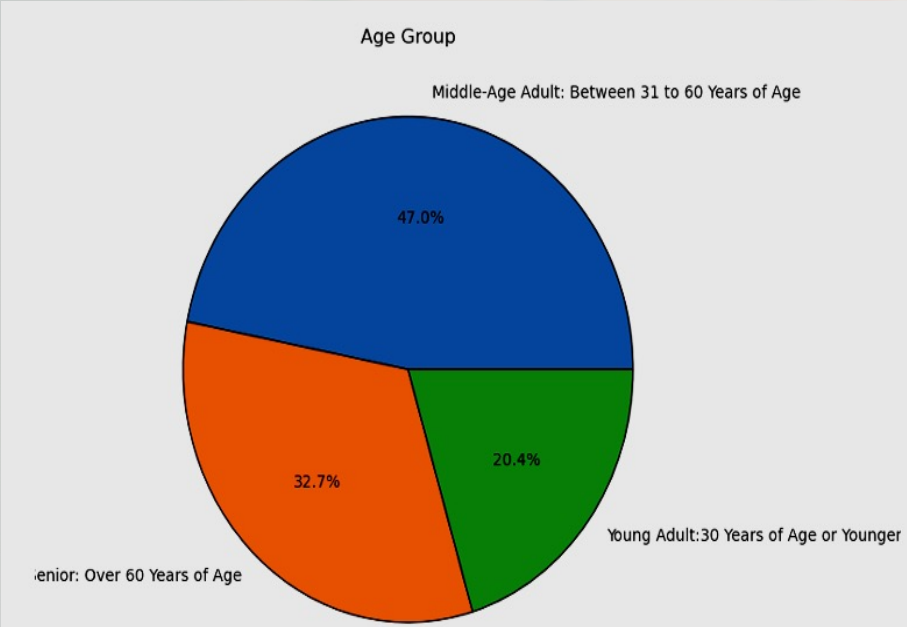


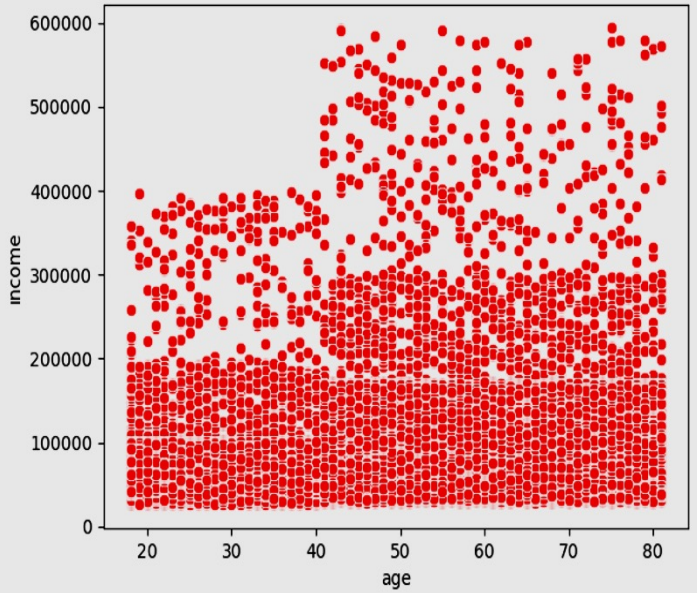


Produce products clearly dominate product sales over any other department.

4.Instacart Merket:

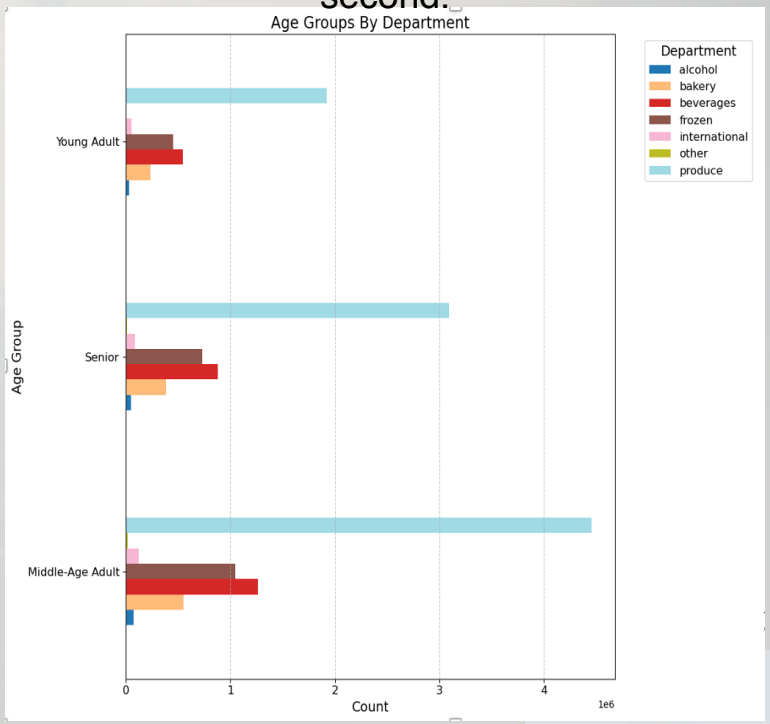
Produce products clearly dominate product sales over any other department.





Produce products clearly dominate product sales over any other department, after the age of 40, incomes begin to see an increase for certain customers compared to younger ages.

The middle age group holds primary sells in all product departments with produce being the main selling department. This middle-age group also holds the highest number of loyalty flag customers with regular being first and loyal customers being second.



RECOMMENDATIONS:

- some customers have very large income, design premium product lines and luxury items to see if Instacart can attract more high-income customers.
- As the scatterplot is mostly dense around the lower and middle incomes, consider developing a budget friendly product line for essential items that might appeal to both low and middle income shoppers trying to save on money.
- To encourage more young adult shoppers, develop trendy product lines and budget friendly options to attract more young adult shoppers.
- As customer spending is similar to each age group and income group, it is recommended to get better feedback on customer department preferences.
- spending is similar amongst age groups and income groups, tailor marketing strategies to the top 10 performing departments mentioned earlier to ensure customer attention, until better insight to customer thoughts can be obtained.

4.Pig E Bank:

PROJECT OVERVIEW

Instacart, a leading online grocery store, enables customers to order groceries through an app.By analyzing historical data, the company aimed to boost sales by adopting a more strategic approach to targeting customers and improving segmentation.

TOOLS USED: Jupyter, Python, Anaconda, Python Libraries

KEY QUESTIONS:

- What are the busiest days of the week and hours of the day?
- Are there particular times of the day when people spend the most money?
- Are there certain types of products that are more popular than others?

KEY STEPS:

- Write Python Scripts: Used Jupyter Notebook to write Python scripts and document the analysis.
- Import: Imported datasets and necessary Python libraries.
- Data cleaning and data wrangling: Utilized Pandas and NumPy for data cleaning, transformation, and preparation.
- dictionary; performed data quality and consistency checks, including frequency counts.
- Exploratory Data Analysis (EDA): Derived new variables, data manipulation, grouping, and aggregating.
- Visualization: Created graphs and visualizations using Matplotlib, Seaborn, and Scipy.
- Export: Exported datasets in xlsx File.

SKILLS:

- Python
- Data wrangling and merging
- Deriving variables
- Grouping dataset
- Reporting in Excel
- Population flows

4.Pig E Bank:

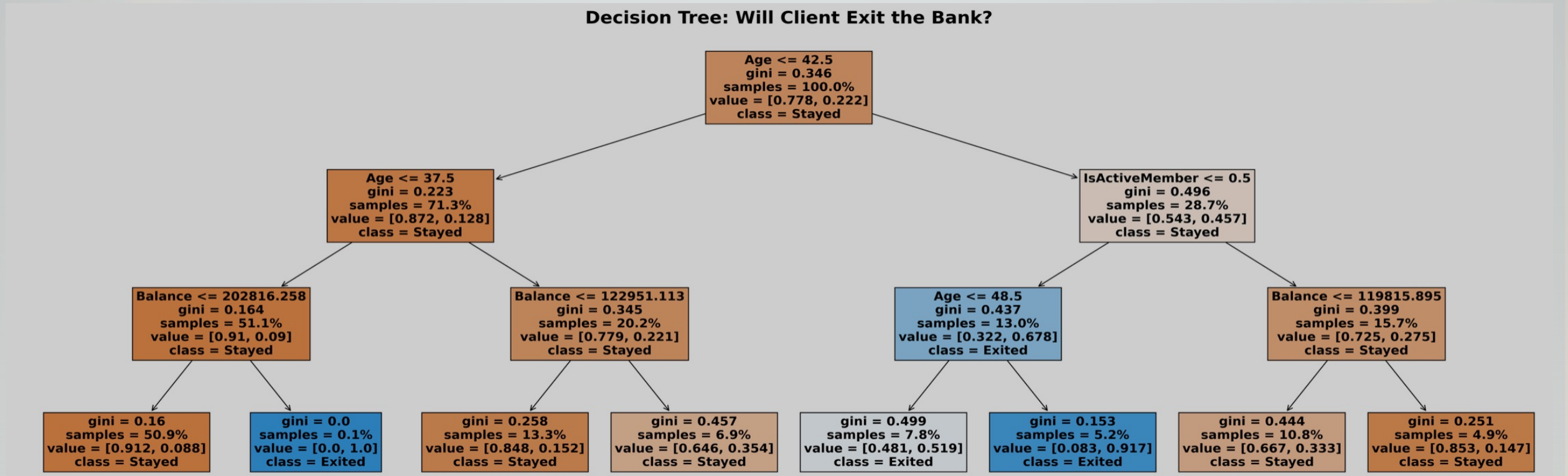
Calculate the proportion of each category for Country and Gender for clients who left

```
Country
France     0.377451
Germany    0.367647
Spain      0.254902
Name: proportion, dtype: float64
Gender
Female    0.593137
Male      0.406863
Name: proportion, dtype: float64
```

Calculate the proportion of each category for Country and Gender for clients who left

```
Country
France     0.512071
Spain      0.256671
Germany    0.231258
Name: proportion, dtype: float64
Gender
Male      0.566709
Female    0.433291
Name: proportion, dtype: float64
```

4.Pig E Bank:



The refined decision tree model has an accuracy of 83.6% for predicting if a client will leave the bank based on the information available.

Next steps:

- Pig E Bank should take measures to incentivize customers to use their account more, such as reward programs.
- Pig E Bank should research and address the issues that lead 45-54 year old customers to churn.
- PE Bank should motivate customers to acquire more than 1 product. This may make the customers more involved with the bank and less likely to churn.
- PE Bank should focus on making its service more female-friendly.

6. Cancer a Disease

PROJECT OVERVIEW:

This dataset contains data about lung cancer Mortality. This database is comprehensive collection of patient information, specifically focused on individuals diagnosed with cancer. It is designed to facilitate the analysis of various factors that may influence cancer prognosis and treatment outcomes. The database includes a range of demographic, medical, and treatment-related variables, capturing essential details about each patient's condition and history.

CONTEXT:

Lung cancer is the leading cause of cancer death in men and second in women. Predictive models can help determine patient chance of survival.

TOOLS USED: Jupyter, Python, Anaconda, Python Libraries

KEY QUESTIONS:

- Which health markers are most strongly associated with lung cancer survival?
- Is there an association between treatment received and lung cancer survival?
- Which age groups are most affected by lung cancer?

GOAL:

Analysis health indicators, demographic data, and treatment-related variables of lung cancer patients to determine which factors increase survival rates.

TECHNICAL SKILLS

- Sourcing Open Data
- Correlation Heatmaps and Scatterplots
- Geospatial Analysis with JSON files
- Linear Regression Analysis in Python
- Cluster Analysis (k-means)
- Tableau Dashboard Creation

6. Cancer a Disease

	Age	Asthma	BMI	Cholesterol Level	Cirrhosis	Country GDP (per capita)	Country Life Expectancy	Country Population	Days to Start Treatment	Family History	Has Other Cancer	Hypertension	Survived	Treatment Duration (days)
Age	1.000	0.000	-0.001	-0.001	-0.001	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.001	-0.001
Asthma	0.000	1.000	0.000	0.000	0.053	0.001	0.000	0.000	0.001	-0.001	0.040	0.108	0.000	-0.006
BMI	-0.001	0.000	1.000	0.747	-0.001	0.000	0.001	0.001	0.000	0.000	0.000	0.000	0.000	-0.007
Cholesterol Level	-0.001	0.000	0.747	1.000	-0.001	0.001	0.001	0.001	0.000	0.000	0.000	-0.001	0.001	-0.009
Cirrhosis	-0.001	0.053	-0.001	-0.001	1.000	0.000	-0.001	0.000	-0.001	0.001	0.023	0.097	0.000	-0.004
Country GDP (per capita)	0.000	0.001	0.000	0.001	0.000	1.000	0.593	-0.018	-0.001	0.000	0.001	0.001	0.001	-0.011
Country Life Expectancy	0.001	0.000	0.001	0.001	-0.001	0.593	1.000	0.263	0.000	0.000	0.000	0.000	-0.001	0.001
Country Population	0.000	0.000	0.001	0.001	0.000	-0.018	0.263	1.000	0.000	0.000	0.000	0.000	-0.001	0.000
Days to Start Treatment	0.000	0.001	0.000	0.000	-0.001	-0.001	0.000	0.000	1.000	0.000	0.000	0.000	-0.001	0.124
Family History	0.000	-0.001	0.000	0.000	0.001	0.000	0.000	0.000	0.000	1.000	-0.001	0.000	0.001	-0.001
Has Other Cancer	0.000	0.040	0.000	0.000	0.023	0.001	0.000	0.000	0.000	-0.001	1.000	0.072	-0.002	-0.002
Hypertension	0.000	0.108	0.000	-0.001	0.097	0.001	0.000	0.000	0.000	0.000	0.072	1.000	0.001	-0.011
Survived	0.001	0.000	0.000	0.001	0.000	0.001	-0.001	-0.001	-0.001	0.001	-0.002	0.001	1.000	-0.001
Treatment Duration (days)	-0.001	-0.006	-0.007	-0.009	-0.004	-0.011	0.001	0.000	0.124	-0.001	-0.002	-0.011	-0.001	1.000

Pearson's Correlation Coefficient

-0.018 1.000

patient survival was poorly correlated with all other variable. After liaising with a mentor, it was decided that:

- Proceeding analyses could focus on relationships between treatment related variables.
- Data would be wrangled in later steps to obtain and compare survival rates.

6. Cancer a Disease

Linear Regression Analysis:

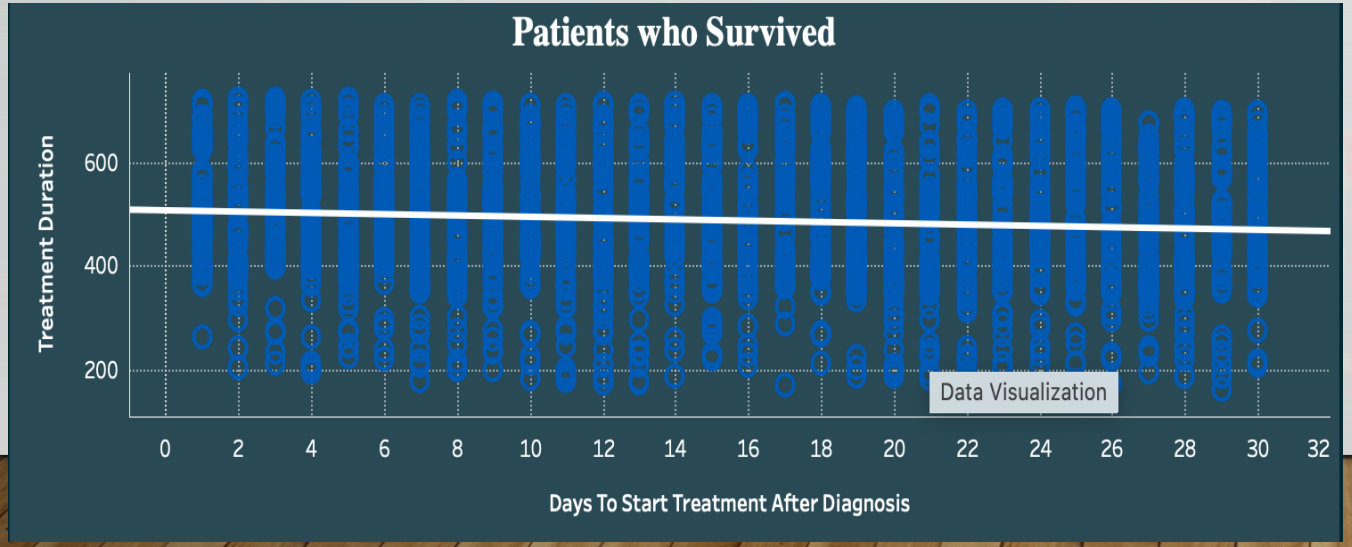
• More than 99% of the data's variance could not be explained by the model.

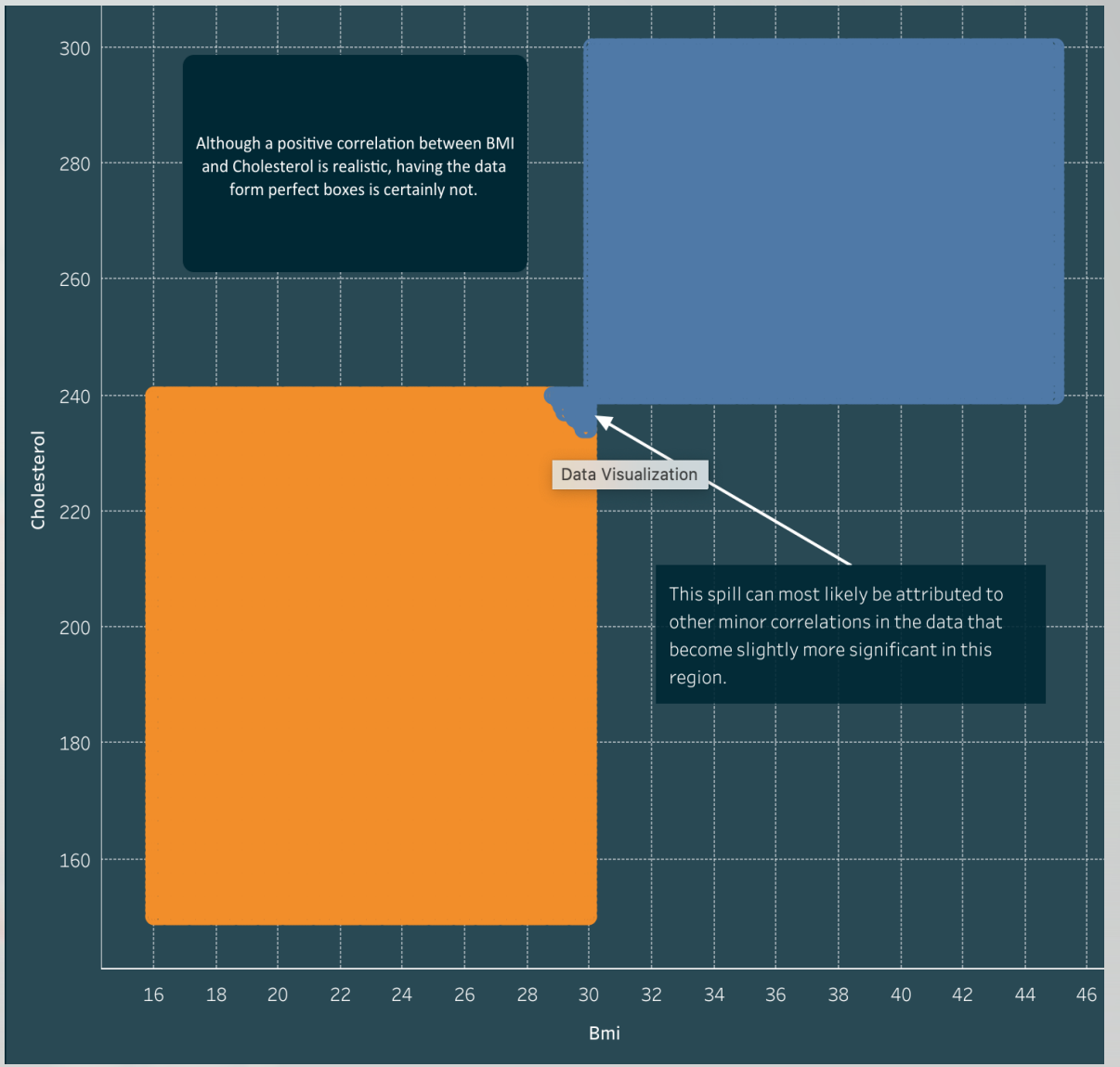
• Data was subdivided based on categorical variables aiming to reduce the variance and yield new insights.

Cluster Analysis:

• Clusters determined by algorithm were heavily determined by only two variables. Analysis on scatterplots yielded insignificant results.

• Additional clusters were added to the algorithm but still yielded insignificant results.





Next Steps:

•Alternate, non-artifical, data should be sourced and analysed (although finding data of a similar structure may be difficult to find due to data privacy laws). The analyses conducted here could be re-run on alternate data with modifications.

•If alternate data is not available, variable selection may be reviewed and previously overlooked variables may be explored.

•Given the abundance of categorical variables, conducting further subgroup analyses on more specific patient groups may yield significant results

•Collaborate with domain experts such as epidemiologists and public health professionals to refine the analysis and identify potential gaps and alternative approaches.


Thank you!