MANISHA TOMAR

DATA ANALYTICS PORTFOLIO

Projects

1. GameCo.

Global market analysis of video game sales.

2. Preparing for flu Season

Staff deployment planning for influenza season

3. Rockbuster

Launching Rockbuster stealth online movie service

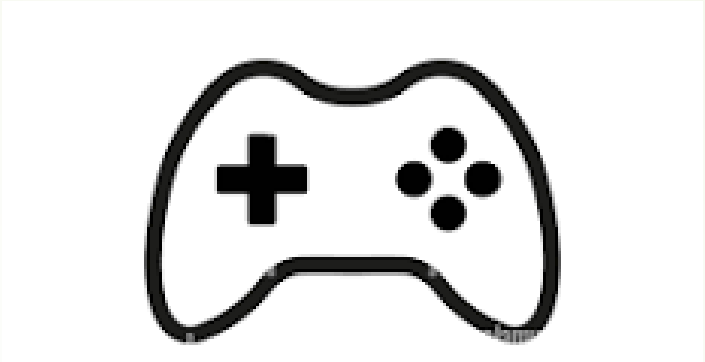
4. Instacart

Market segmentation analysis to uncover sales

5. Pig E. Bank

Analysis customer attrition

6. Conditions Contributing to COVID-19 Deaths



GameCo

ANALYZING GLOBAL VIDEO GAME SALES

GameCo. Market Analysis

ANALYZING GLOBAL VIDEO GAME SALES

About GameCo.

Game Co. is a global gaming leader, offering a diverse range of games for sale and rental across key markets.

Objective

The primary goal of this project was to analyze historical game data to identify key trends and patterns. The insights gained were aimed at predicting market reception for upcoming games, thereby informing strategic decisions for future releases.

Key Steps:

- **Data Cleaning:** Identify and address missing values and duplicates.
- **Exploratory Data Analysis (EDA):** Calculate key statistics, including mean, median, mode, maximum, & minimum.
- **Data Grouping, Filtering, and Summarizing:** Utilize Pivot Tables for efficient data analysis.
- **Data Visualization:** Create a column chart of total global sales by publisher and a line chart depicting average North American sales by year.
- **Type of Analysis:** Determine whether to conduct Descriptive, Diagnostic, Predictive, or Prescriptive analysis.
- **Interpretation:** Analyze results and summarize key findings.

Excel Functions Used: Find & Select, Replace, Replace All, Remove Duplicates.

GOALS

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

DATA LIMITATIONS

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

TOOLS USED

Excel, PowerPoint

GameCo. Market Analysis

ANALYZING GLOBAL VIDEO GAME SALES

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?

Data

- Data source: VGChartz.
- File: Excel - CSV
- Period of data: 1980-2016
- Regions: North America, Europe, Japan, and others
- Information: title, platforms, year, genre, publisher

Skills Applied

- For this project, I used Excel & Tableau for visualizations.
- · Improving data quality
- · Data grouping and summarizing
- · Descriptive analysis, Pivot table
- · Visualization results in Excel and Tableau
- · Presenting results

Challenges

This was my first project, and the main challenge was grasping the business requirements and converting them into meaningful analysis, as I was navigating entirely new territory

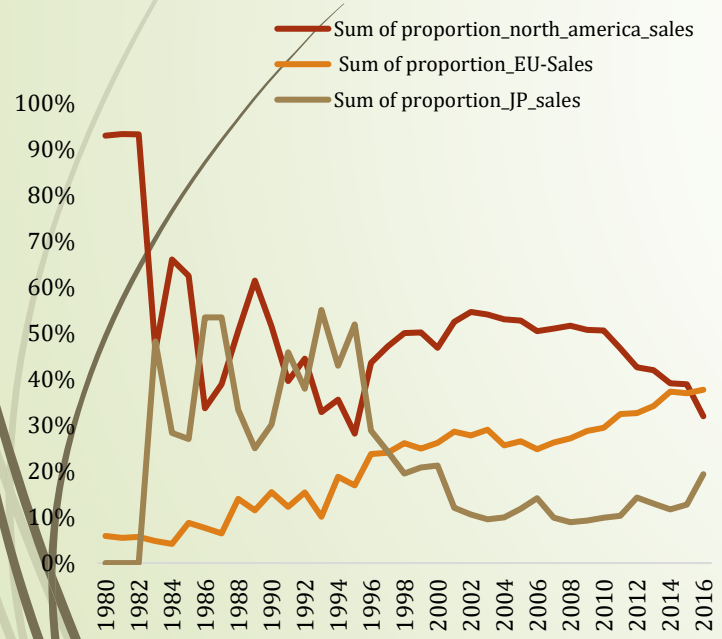
Analysis

6

The final analysis highlights popular game genres and regional sales trends since 1980, offering a detailed breakdown of top genres, platforms, and market share data.

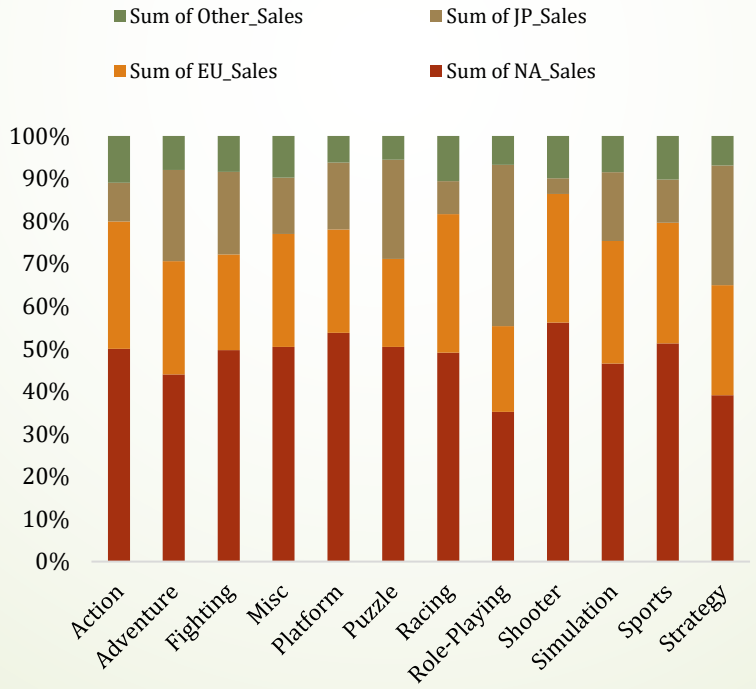
Historical Trends 1980-2016

Sales Over Time in North America, Europe, and Japan



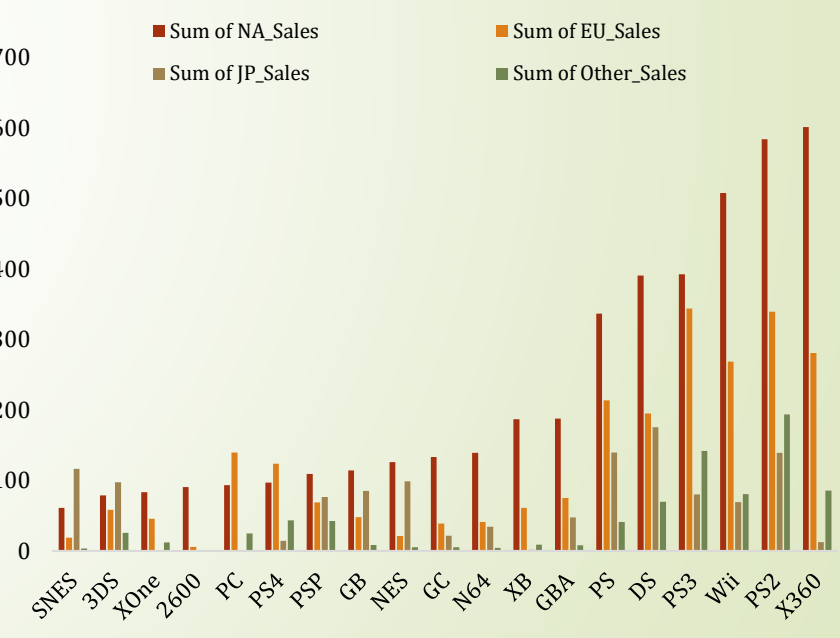
Recent Trends Analysis 2010-2016

Total Sales by Genre with Regional Breakdown Genre



Popular Platforms by Region

The top platforms are PS, DS, PS3, WII, PS2, X360



Insights & Recommendations

KEY LEARNINGS

- Despite declining sales since 2010, Europe surpassed North America in market share after 2015, while Japan experienced a notable sales surge.
- While Action, Shooter, and Sports genres dominate in North America and Europe, Japan has shifted towards Action games, with Role-Playing games remaining the second most popular.
- Platform popularity has shifted significantly; PS2 and X360 have emerged as leaders in the West, replacing former favorites. In Japan, the enduring preference for the 3DS highlights distinct regional differences.
- Post-2015, regional preferences have diverged significantly, highlighting the need for tailored game development and marketing strategies.

RECOMMENDATIONS

Global Marketing Strategy

Adapt marketing strategies to align with regional sales trends and shifting market preferences post-2015.

Develop advertising campaigns that resonate with regional tastes and current trends to enhance engagement and increase market share.

Genre

North America and Europe: Increase focus on Shooters and Sports genres, which continue to perform strongly.

Japan: Leverage the growing interest in Action genres while also prioritizing Role-Playing Games, which remain highly popular.

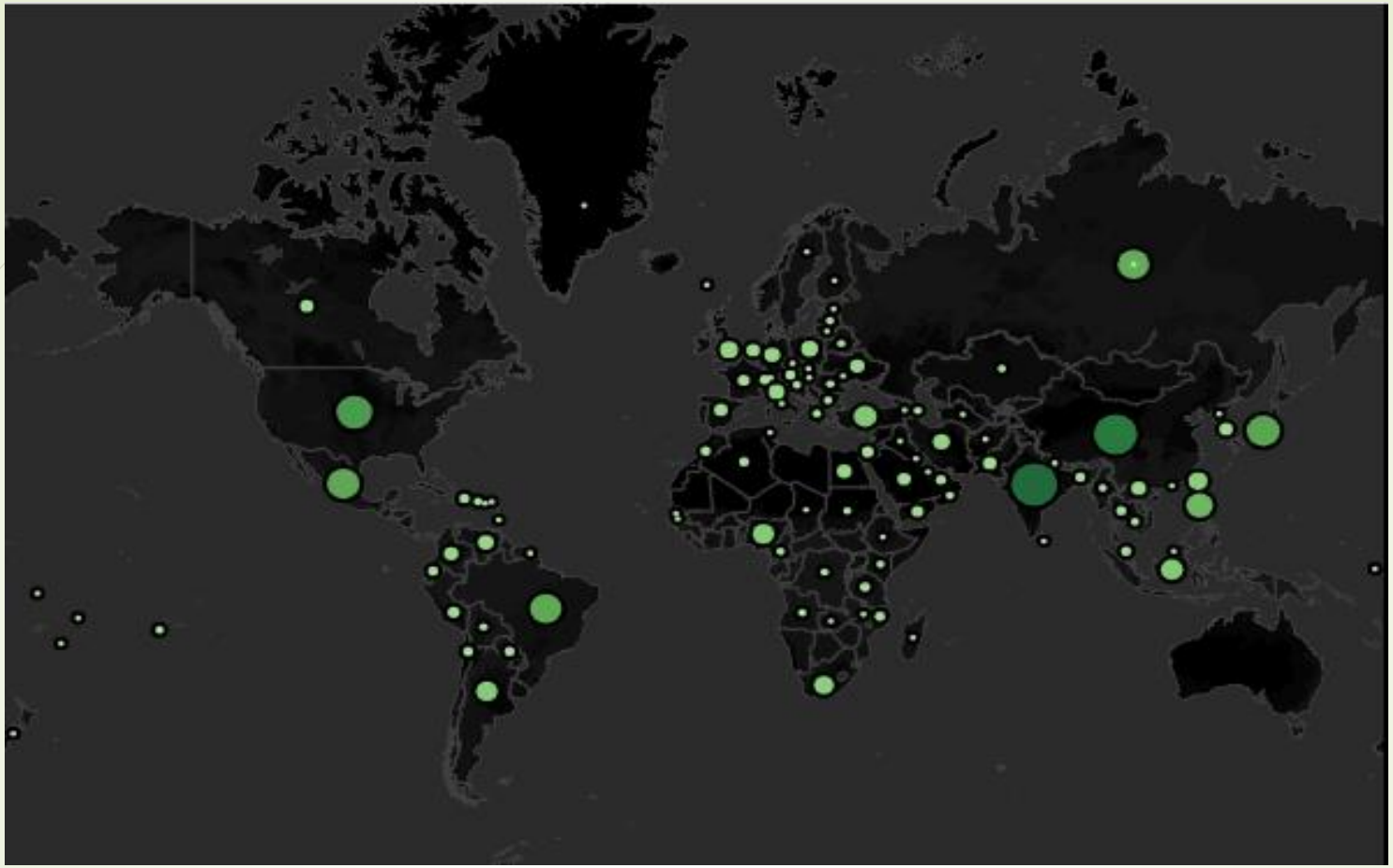
Platform

North America and Europe: Leverage the established console base, particularly the popularity of PS and Xbox One.

Japan: Continue to capitalize on the strong preference for handheld platforms like the 3DS.

General: Stay alert to emerging platforms to swiftly adapt to shifts in platform popularity.

Final Presentation



Flu Season

PREPARING FOR THE FLU SEASON IN THE U.S.

Preparing for Influenza Season

Project Overview: During flu season, U.S. healthcare facilities face increased patient volumes. This project aimed to develop a staffing schedule to address these demands, enhancing preparedness by effectively managing staffing at clinics and hospitals.

Motivation: Influenza season increases patient numbers, especially among vulnerable populations, requiring additional temporary staff from the medical staffing agency.

Scope: Analyze historical influenza data to forecast staffing needs, considering incidence rates, peak seasons, regional variations, and implications for temporary healthcare workers.

Key Steps: Conduct Exploratory Data Analysis (EDA):

- Descriptive statistics: central tendency & distribution.
- Create charts to visualize distribution, relationships, missing values, and outliers.
- Clean data (including imputation) and transform it (e.g., merging spreadsheets).
- Perform hypothesis testing (t-test in Excel).
- Interpret and summarize results.

Excel Functions Used:

- FIND & SELECT, REPLACE WITH, REPLACE ALL, REMOVE DUPLICATES.
- PivotTables and SUMIFS for summarizing and aggregating data.

GOALS

Improve flu season readiness by optimizing staffing needs.

STAKEHOLDERS

- Medical agency frontline staff
- Hospitals and clinics using staffing services
- Influenza patients
- Staffing agency administrators

TOOLS USED

Excel, PowerPoint, Tableau

Preparing for Influenza Season

Key Questions

Support staffing plan with data on medical personnel distribution in the U.S.

Investigate seasonality of influenza across states.

Prioritize states based on vulnerable population size and categorize as low-, medium-, or high-need.

Identify data limitations that hinder analysis.

Data

1. Census Population Dataset Source: US Census Bureau Contents: Population information from the US by country, time, age and gender for 2009-2017.

2. Influenza Deaths Dataset Source: CDC Contents: Information about influenza deaths by age groups in the US by state and time for between 2009-2017.

Skills Applied

- Translating business requirements into analytical questions
- Sourcing relevant datasets
- Data integration and cleaning
- Statistical hypothesis testing, Forecasting ,
- Visual analysis & Storytelling in Tableau
- Presenting results to an audience

Challenges

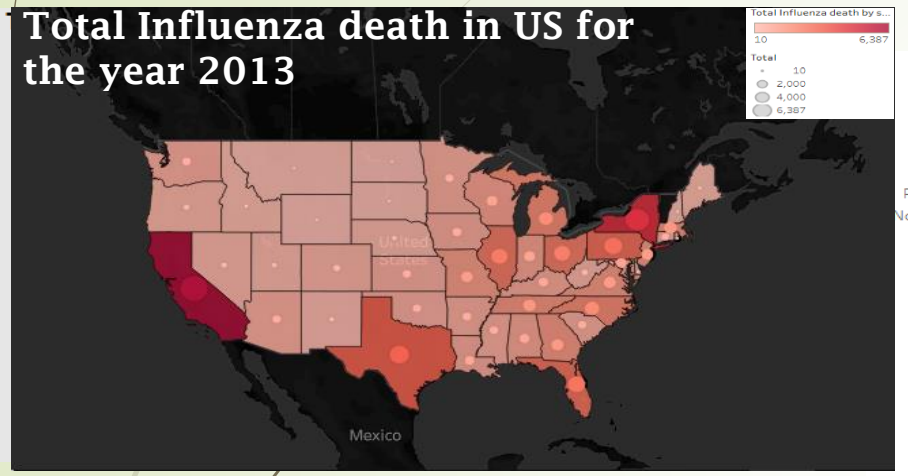
A major challenge in this project was integrating diverse data sources for a cohesive analysis.

Additionally, managing 'suppressed' variables presented difficulties, initially creating uncertainties about their effective use in the analysis.

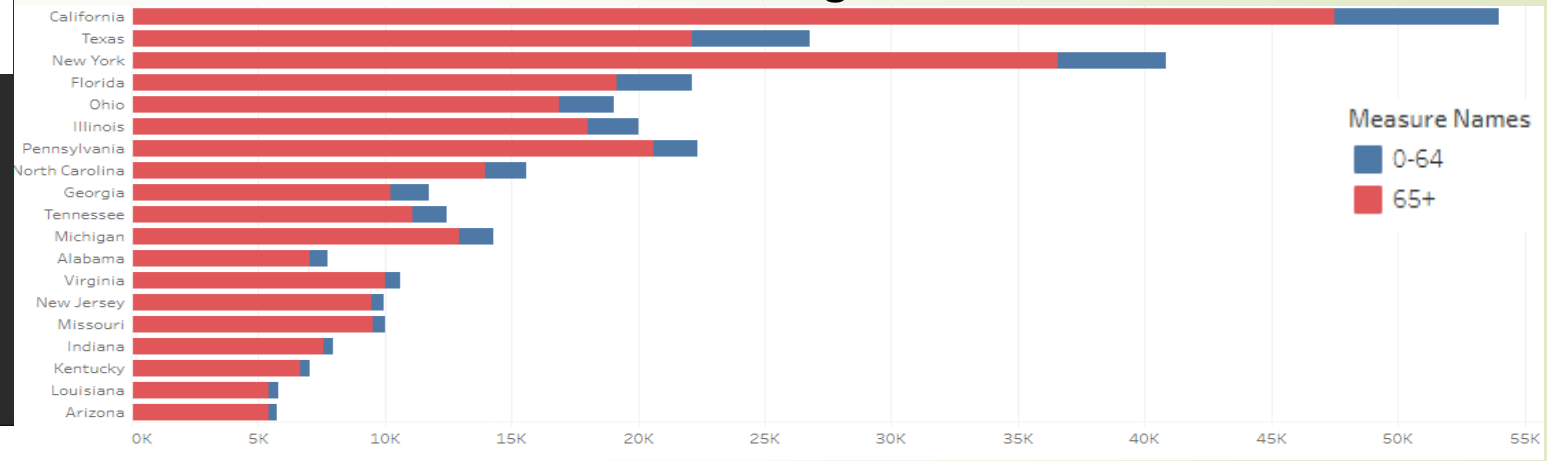
Analysis

In this analysis, staffing needs for a medical staffing agency were forecasted to optimize resource allocation during peak flu periods. A priority map was created, categorizing U.S. states from very high-need to very low-need based on their requirements, and the seasonality of flu outbreaks was analyzed.

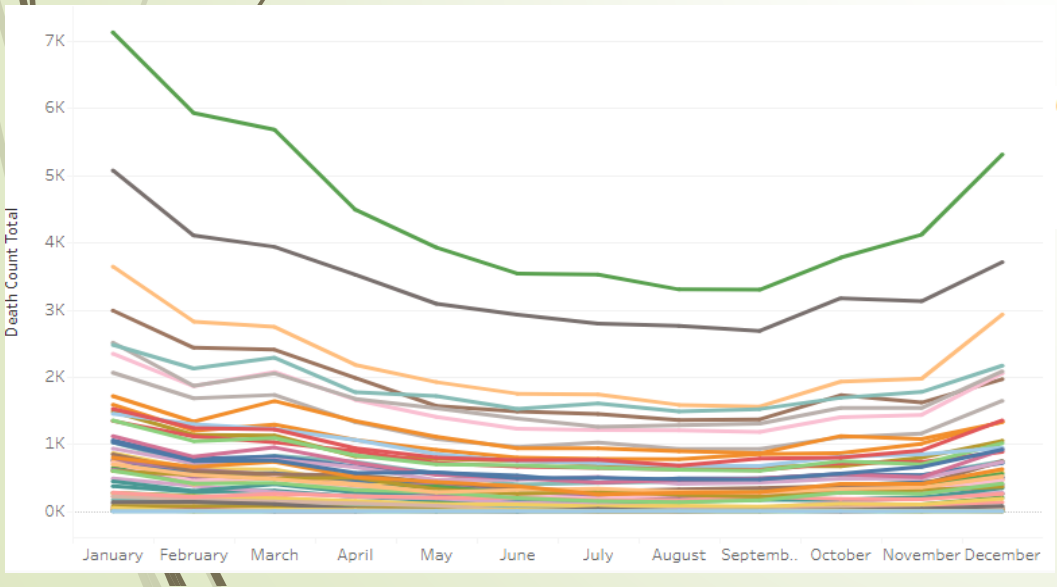
Total Influenza death in US for the year 2013



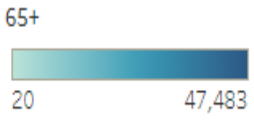
Influenza Death & Age

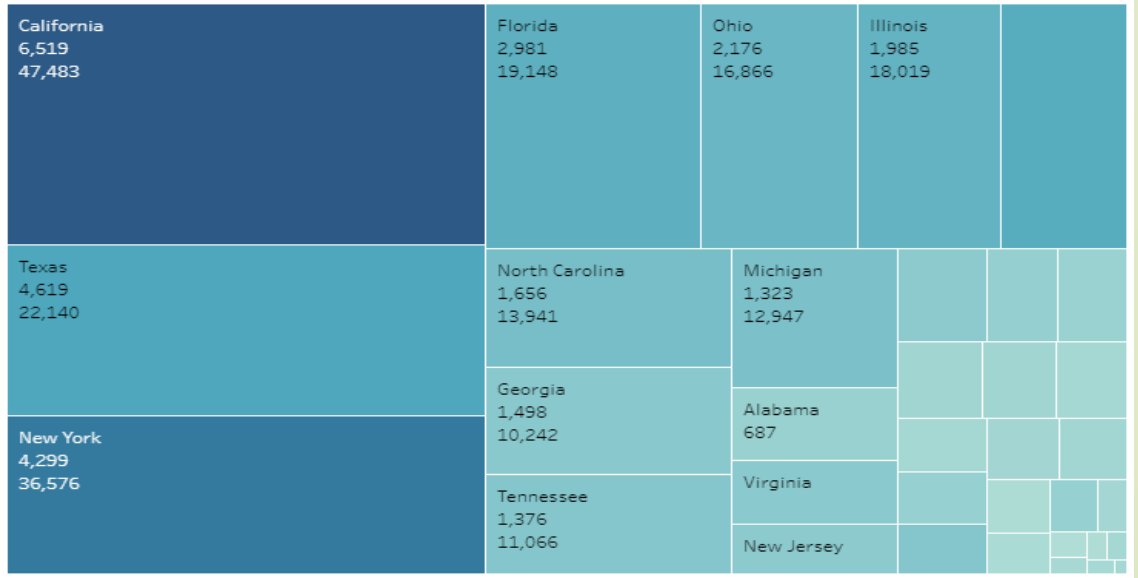


Seasonality



Priority Map





Insights & Recommendations

KEY LEARNINGS

- States with a larger population of individuals aged 65 and older experience a higher number of influenza-related deaths.
- Influenza deaths usually increase in December, peak in January, and stay high through February and March, consistently across states and regions.
- The virus's impact on vulnerable age groups underscores the need for targeted interventions and preventative measures.

RECOMMENDATIONS

Priority

- States with significant populations aged 65 and older should boost their medical staffing during the influenza season. The most critical states are California, New York, Texas, and Pennsylvania.

Seasonality

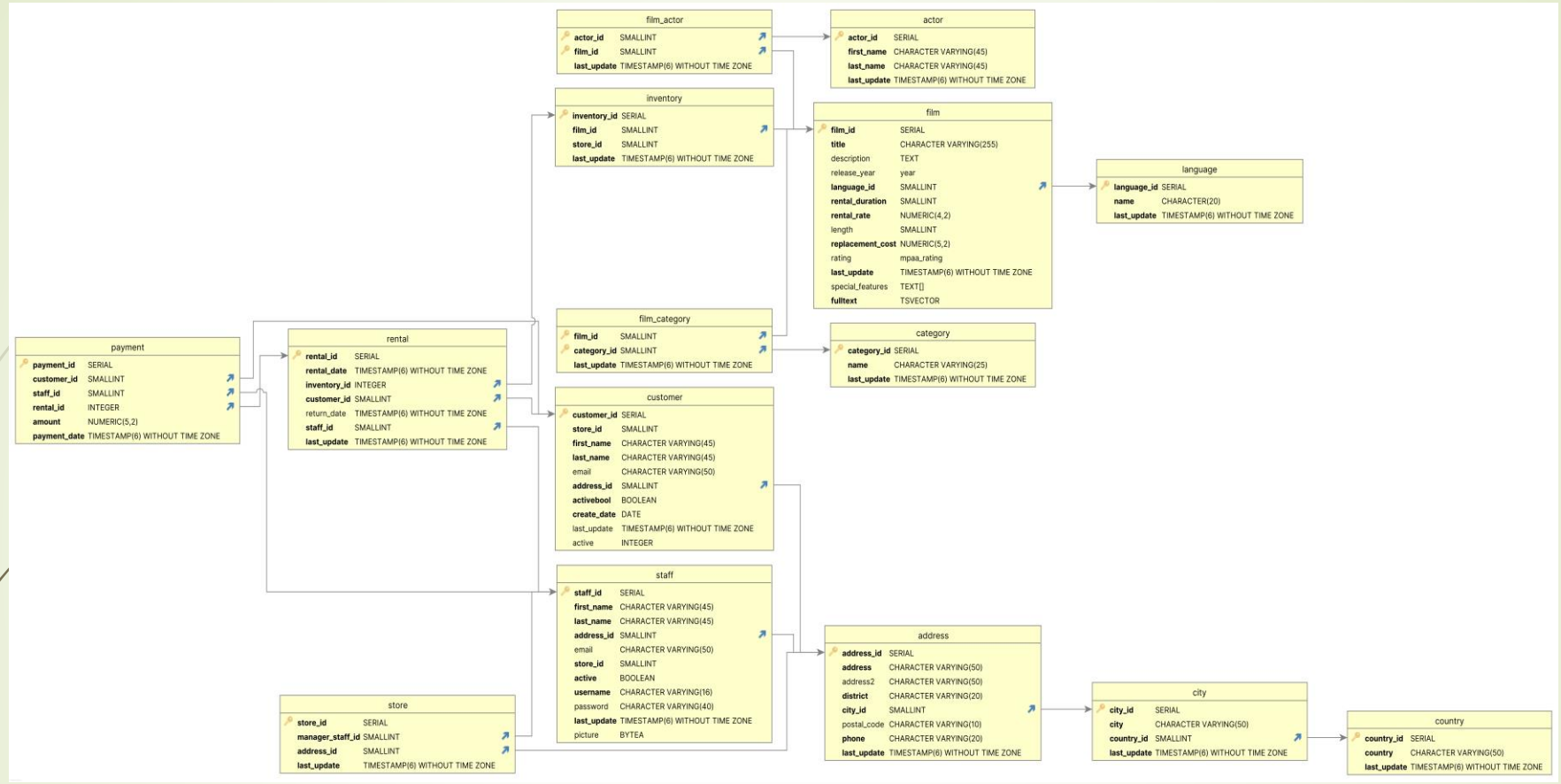
- To prepare for the influenza season, which runs from December to March, medical staff should be allocated to each state according to its priority level.

Further Analysis

- States with smaller populations or fewer individuals over 65 but higher influenza death rates should be examined more closely to identify other contributing factors.
- Future analysis should investigate the impact of influenza on additional vulnerable groups, such as pregnant women, individuals with HIV/AIDS, and those with other conditions.

LINKS & DELIVERABLES

Tableau Story Board Interim Report



Rockbuster Stealth LLC

Transition to Online Video Rental Service

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 relied on physical stores. With rising competition from digital platforms like Netflix and Amazon Prime, Rockbuster is exploring a shift to online video rentals.

Objective:
This project aims to leverage data analytics to address key business questions and shape Rockbuster's 2020 strategy for transitioning to an online service model. The insights gathered will be crucial in guiding decision-making and strategic planning.

Key Steps:
- **Data Extraction:** Utilized SELECT, JOIN, and WHERE clauses to retrieve relevant data from multiple tables.
- **Data Dictionary:** Developed an Enterprise Relationship Diagram (ERD) using DbVisualizer.
- **ERD Extraction:** Extracted the ERD from Rockbuster's database using DbVisualizer.
- **Data Aggregation:** Applied GROUP BY and aggregate functions like SUM, AVG, and COUNT to summarize data.
- **Data Filtering:** Used HAVING clauses to filter aggregated results.
- **Data Visualization:** Created bar charts, line charts, scatter plots, and other visualizations

TOOLS USED

Excel
PowerPoint
Tableau
DbVisualizer
SQL/PostgreSQL

Rockbuster Stealth LLC

Key Questions

1. Which movies contributed the most/least to revenue gain?
2. What was the average rental duration for all videos?
3. Which countries are Rockbuster customers based in?
4. Where are customers with a high lifetime value based?
5. Do sales figures vary between geographic regions?

Data

Data Source: Postgres Tutorial Data Set

Skills Applied

- Relational databases , SQL (Link to Queries)
- Creating a data dictionary, Database querying
- Data filtering, Data cleaning and summarizing Joining tables Subqueries
- Common table expressions
- Presentation

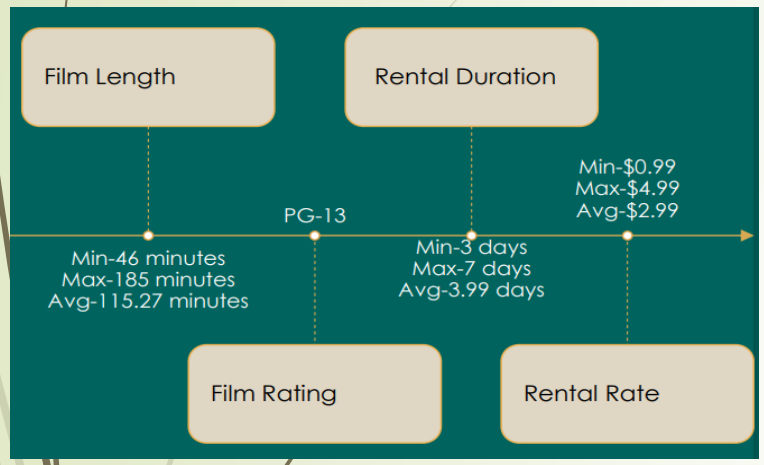
Challenges

Challenges included managing complex queries and maintaining data accuracy.

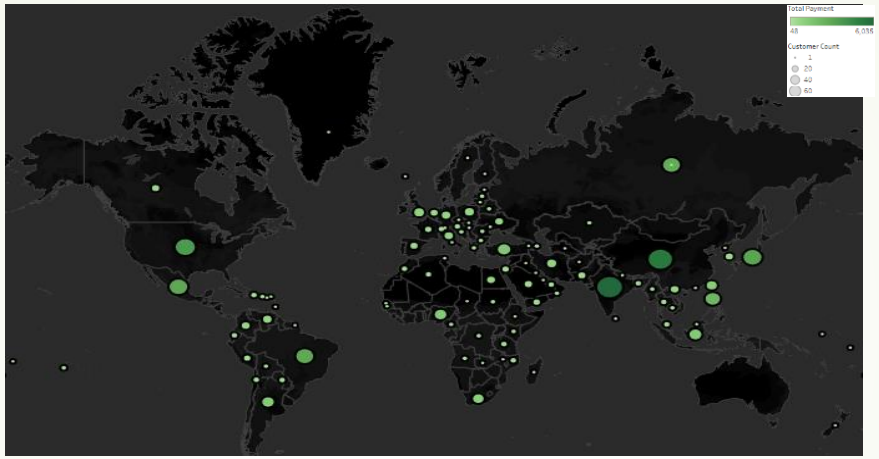
Analysis

Rockbuster's data was efficiently managed within a PostgreSQL database, with data integrity enhanced through meticulous cleaning and validation. SQL was utilized to uncover key insights into top-performing movies and strategic customer locations, which were then visualized using Tableau. The comprehensive analysis has been documented in a detailed report, including SQL code in Excel, and is supported by an ERD and data dictionary.

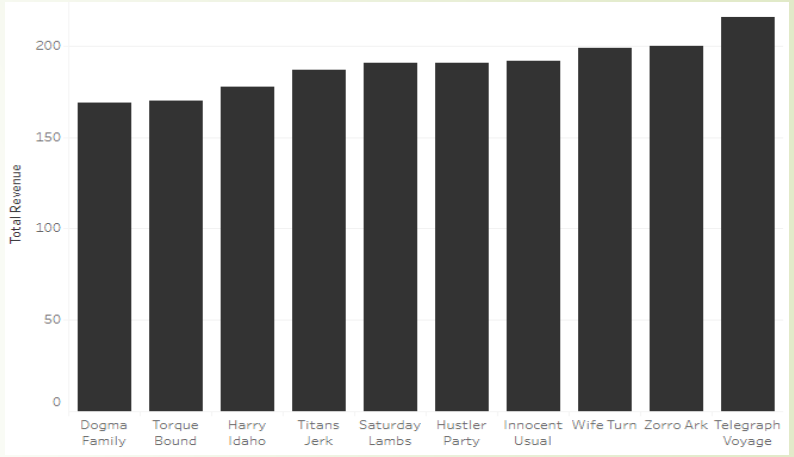
Descriptive Analysis



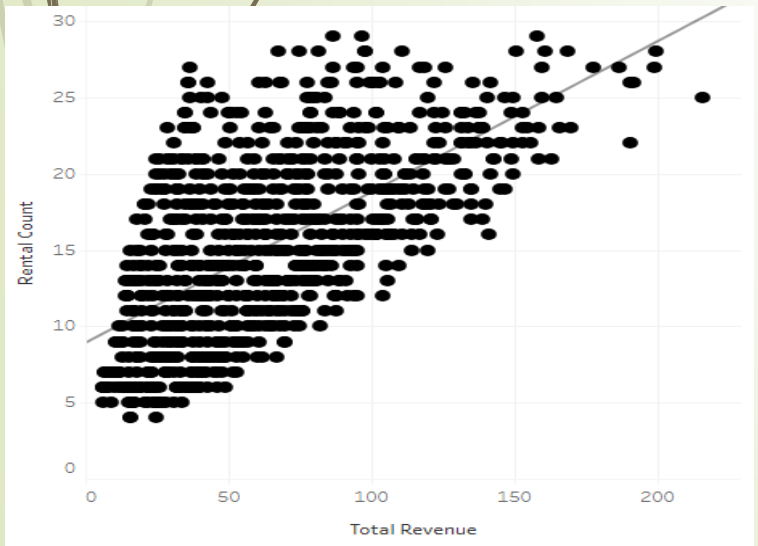
Rockbuster Customers Worldwide (Link)



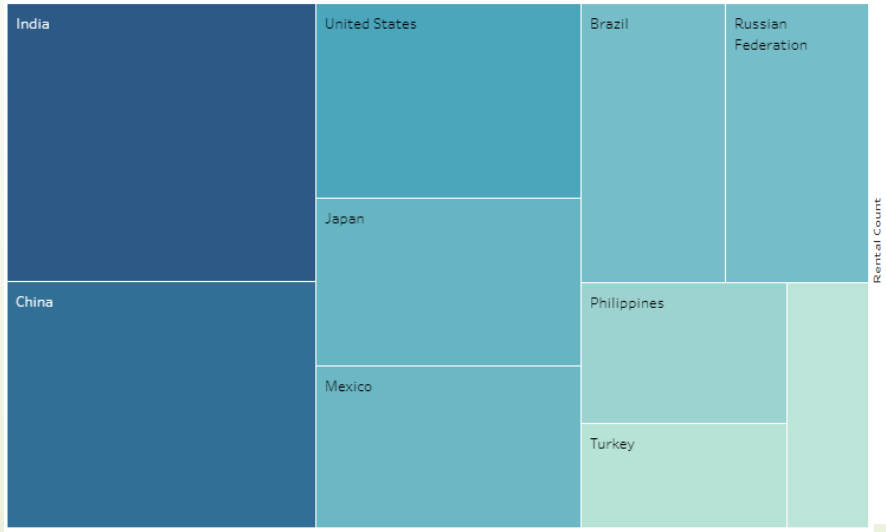
Top 10 Movies Generated Highest Revenue (Link)



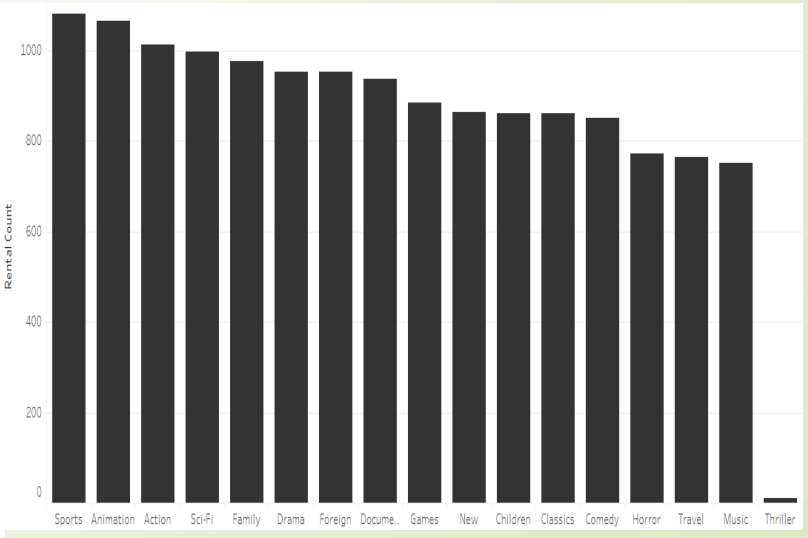
Rental Frequency and Revenue (Link)



Top 10 Customers (Link)



Highest Genre Performance (Link)



Insights & Recommendations

KEY LEARNINGS

1. Telegraph Voyage contributed the most revenue gain while Texas Watch contributed the least.

2. The average rental duration across all videos is 7 days, indicating consistent consumer engagement.

3. Sports, sci-fi, and animation genres are the most popular. However, different regions have different preferences.

4. Top customers are coming from all around the globe.

5. India, China, and the United States lead in both customer numbers and revenue share, driving significant market impact.

RECOMMENDATIONS

Region/Country Specific

- Increase marketing efforts and tailor promotional campaigns specifically for high-impact markets: the broader Asian region, and countries like India, China, and the USA.

Genre Specific

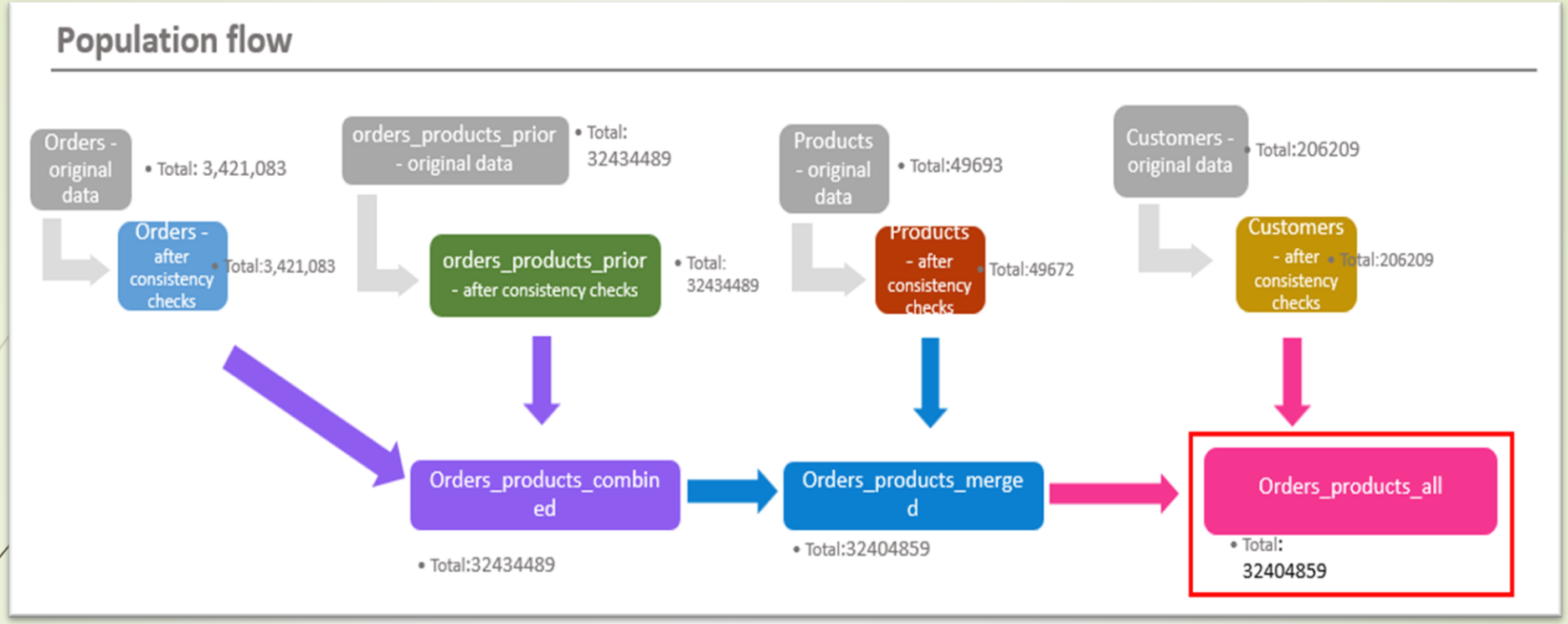
- Capitalize on regional differences in genre preferences by promoting Animation and Sports heavily in Asia, and Sci-Fi in the USA and Europe.

Customer Specific

- Develop a global loyalty program tailored to reward top customers. Implement flexible rental options to cater to the average 5-day rental duration.

LINKS & DELIVERABLES

Final Presentation GitHub Repository Data Dictionary Tableau Storyboard



Instacart

Marketing Strategy for an Online Grocery Store

Instacart

Project Overview:
Instacart, a top online grocery platform, enables customers to order groceries through an app. The company sought to increase sales by leveraging historical data to implement a more strategic approach to customer targeting and segmentation.

Objective:
This project aimed to conduct an in-depth analysis of Instacart's sales data to refine their marketing strategy, with a focus on precise customer targeting and effective segmentation to drive sales growth.

Key Steps:
•**Write Python Scripts:** Developed Python scripts and documented the analysis using Jupyter Notebook.
•**Import:** Loaded datasets and essential Python libraries.
•**Data Cleaning and Wrangling:** Employed Pandas and NumPy for data cleaning, transformation, and preparation.
•**Aggregation Functions:** Applied aggregation functions for data analysis.
•**Export:** Exported datasets to pickle and CSV files.
•**Data Quality and Consistency:** Conducted data quality checks, including frequency counts.
•**Exploratory Data Analysis (EDA):** Performed data manipulation, grouping, aggregation, and derived new variables.
•**Visualization:** Created graphs and visualizations using Matplotlib, Seaborn, and SciPy.
•**Export Data Frames:** Exported cleaned and processed data frames for reporting.

TOOLS USED

Jupyter
Python
Anaconda
Python Libraries

Instacart

Key Questions

1. What are the busiest days of the week and hours of the day?

2. At what times of the day do people tend to spend the most money?

3. How can simple price range groupings be used to optimize marketing and sales efforts?

4. Which types of products are most popular?

5. What are the characteristics and spending habits of different customer profiles

Skills Applied

Python Programming:
Utilized in Jupyter Notebook for all coding and analysis tasks.
Data wrangling, subsetting, filtering, and summarizing with Pandas.
Data merging and consistency checks to ensure accuracy.
Deriving new variables and complex data transformations.
Grouping and aggregating data for detailed insights.
Data Visualization: Advanced visualizations created using Matplotlib and Seaborn.
Presentation of analytical results through clear and effective charts.

Data

Data Sets: Customers: Analyzed for purchasing patterns and loyalty.
Orders: Studied to determine busy times and spending habits.
Products: Categorized to understand popularity and sales impact.
Departments: Analyzed for sales volume per department.
Data Citations: "The Instacart Online Grocery Shopping Dataset 2017", Accessed via Kaggle. "Customers Data Set", Provided by CareerFoundry.

Challenges

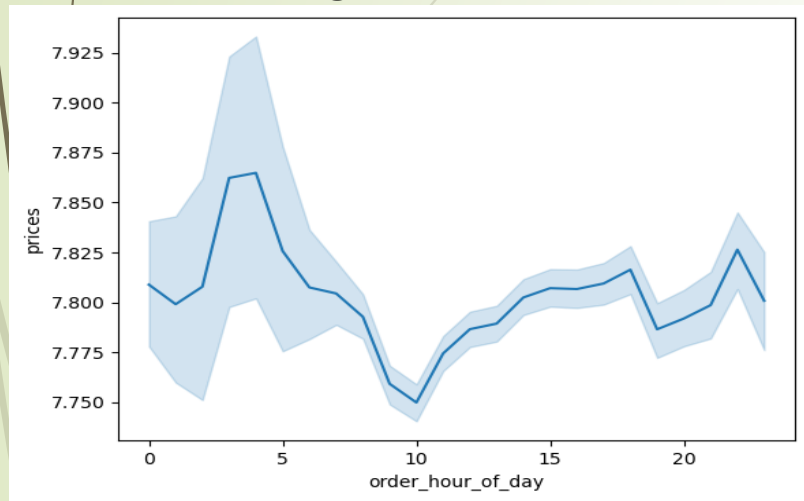
Managing the dataset's 30 million records presented significant challenges in cleaning, standardizing, and formatting for analysis.

Memory shortages were addressed through efficient coding practices, ensuring effective data transformation.

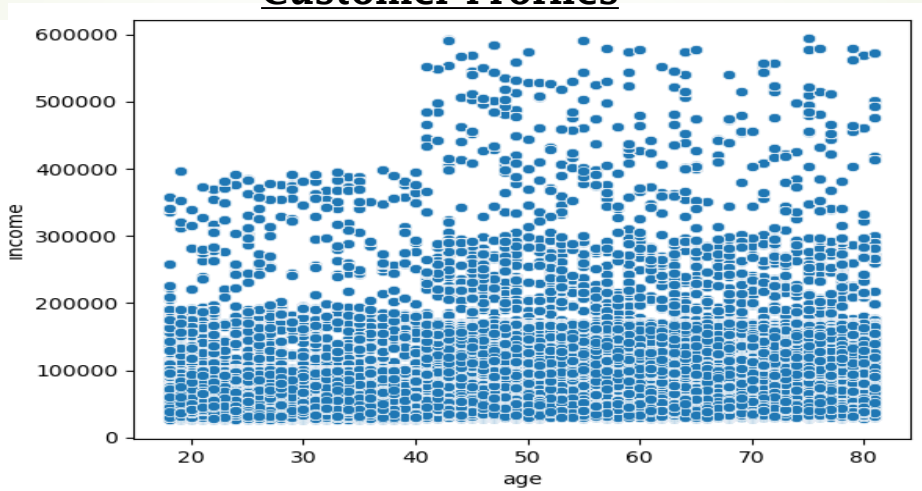
Analysis

The process began with the cleaning and merging of multiple datasets to ensure accuracy, utilizing Pandas for data wrangling, aggregation, and feature derivation. Key trends, such as peak shopping times and popular product categories, were identified, and customer purchasing behaviors were explored. Outliers were removed to refine the dataset. The comprehensive analysis provided actionable insights for crafting targeted marketing campaigns.

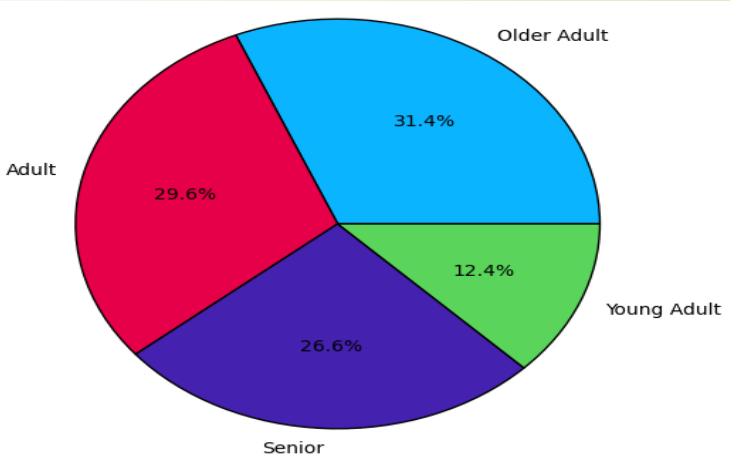
Ordering Patterns

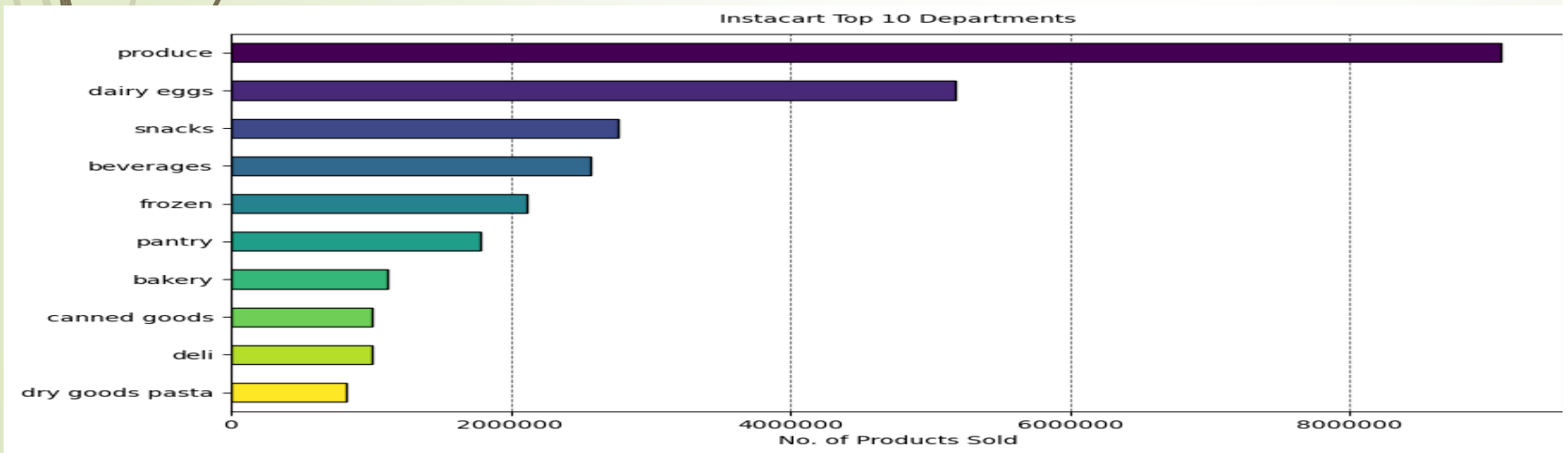


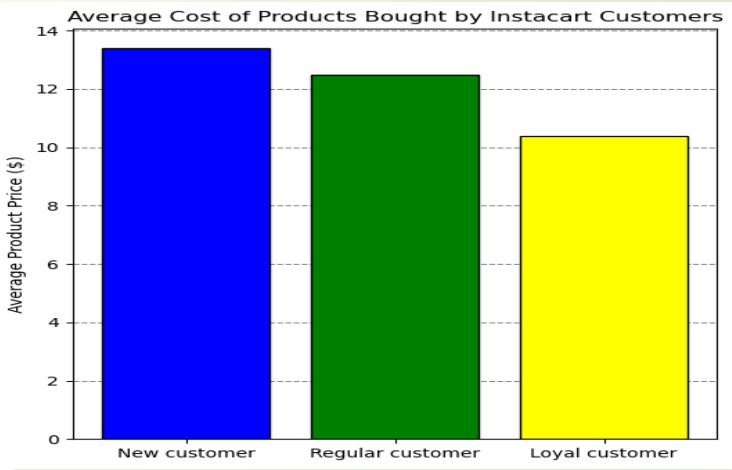
Customer Profiles



Sales Distribution







Insights & Recommendations

KEY LEARNINGS

1. Busiest ordering days are Saturday and Sunday. Busiest hours are between 9AM-4PM.

2. People tend to order more expensive items before 6-7 AM

3. The majority of the products people order are low-range (<5 USD) & mid-range (5-10 USD).

4. Most popular products are fresh food products (produce, meat and sea food, dairy and eggs, deli, bakery).

5. The most significant variation in shopping behavior among customer groups relates to income, particularly in the average price of orders.

RECOMMENDATIONS

Marketing and Sales

Target ads during lower order volumes in the early evening on weekdays to capitalize on increased social media usage. Focus on promoting high-end products early in the morning and late at night, using time-limited offers to create a sense of urgency.

Customer Profiles

Utilize loyalty data to tailor campaigns, particularly promoting premium products to high-income customers like Gen X and Baby Boomers. Implement referral programs that reward loyal customers for bringing in new shoppers, focusing on bulk purchase promotions for families.

Further Analysis

Explore variations in the busiest shopping hours and days, and analyze purchasing trends for different product types during specific times. Examine underperforming high-range products to understand customer purchasing behaviors and preferences better.

LINKS & DELIVERABLES

GitHub Repository

Final Report

Global Bank

ANTI-MONEY LAUNDERING PROJECT

Global Bank

Pig E. Bank is a global bank dedicated to providing exceptional financial services.

Objective The aim of this project was to perform an in-depth analysis of customer satisfaction data at Pig E. Bank. This analysis targeted the identification of factors leading to customer attrition with the ultimate goal of developing robust strategies to improve customer retention.

Key Steps:
•**Write Python Scripts:** Developed Python scripts and documented the analysis using Jupyter Notebook.
•**Import:** Loaded datasets and necessary Python libraries.
•**Data Cleaning and Wrangling:** Used Pandas and NumPy for data cleaning, transformation, and preparation.
•**Data Quality and Consistency:** Conducted data quality checks, including frequency counts.
•**Exploratory Data Analysis (EDA):** Performed data manipulation, grouping, aggregation, and derived new variables.
•**Visualization:** Created graphs and visualizations using Matplotlib, Seaborn, and SciPy.
•**Export:** Exported datasets to XLSX files.

TOOLS USED

Jupyter
Python
Anaconda
Python Libraries

Global Bank

Key Questions

What are the key risk-factors in identifying customers who are most likely to churn?

Data

Data source: Career Foundry

Skills Applied

- Big data
- Data ethics
- Data mining
- Predictive analysis
- Time series analysis and forecasting
- Using GitHub

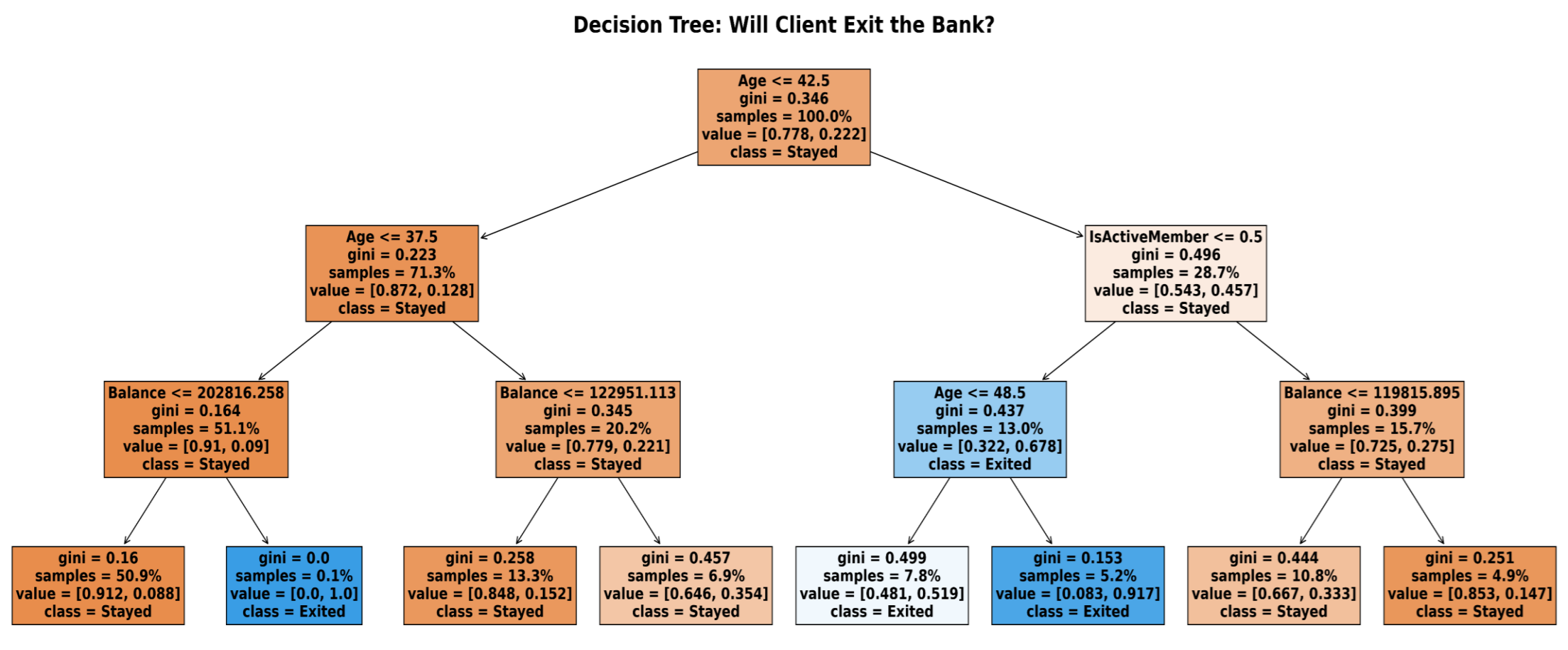
Challenges

Challenges included managing complex queries and maintaining data accuracy.

Analysis

The characteristics of structured and unstructured data were initially analyzed. Data was cleaned and mined using Excel, Python, and GitHub, with predictive models and time-series forecasts being developed. Data bias and ethical issues were addressed to ensure the integrity and security of the data.

A decision tree was created, highlighting key factors leading to customer churn, such as Age, Account Balance, Credit Score, and Salary.



Insights & Recommendations

KEY LEARNINGS

Being an inactive member seems to be a major contributing factor in leaving the bank.

A higher proportion of the people who left the bank are above age 45.

A larger proportion of the former customers have higher account balance (between 100k-140k and also balances more than 150k).

Majority of the former customers held only one product.

RECOMMENDATIONS

Activity

Increase customer engagement through loyalty programs, personalized offers, and regular communication.

Age

Develop tailored financial products and services that cater specifically to the needs of people above 45.

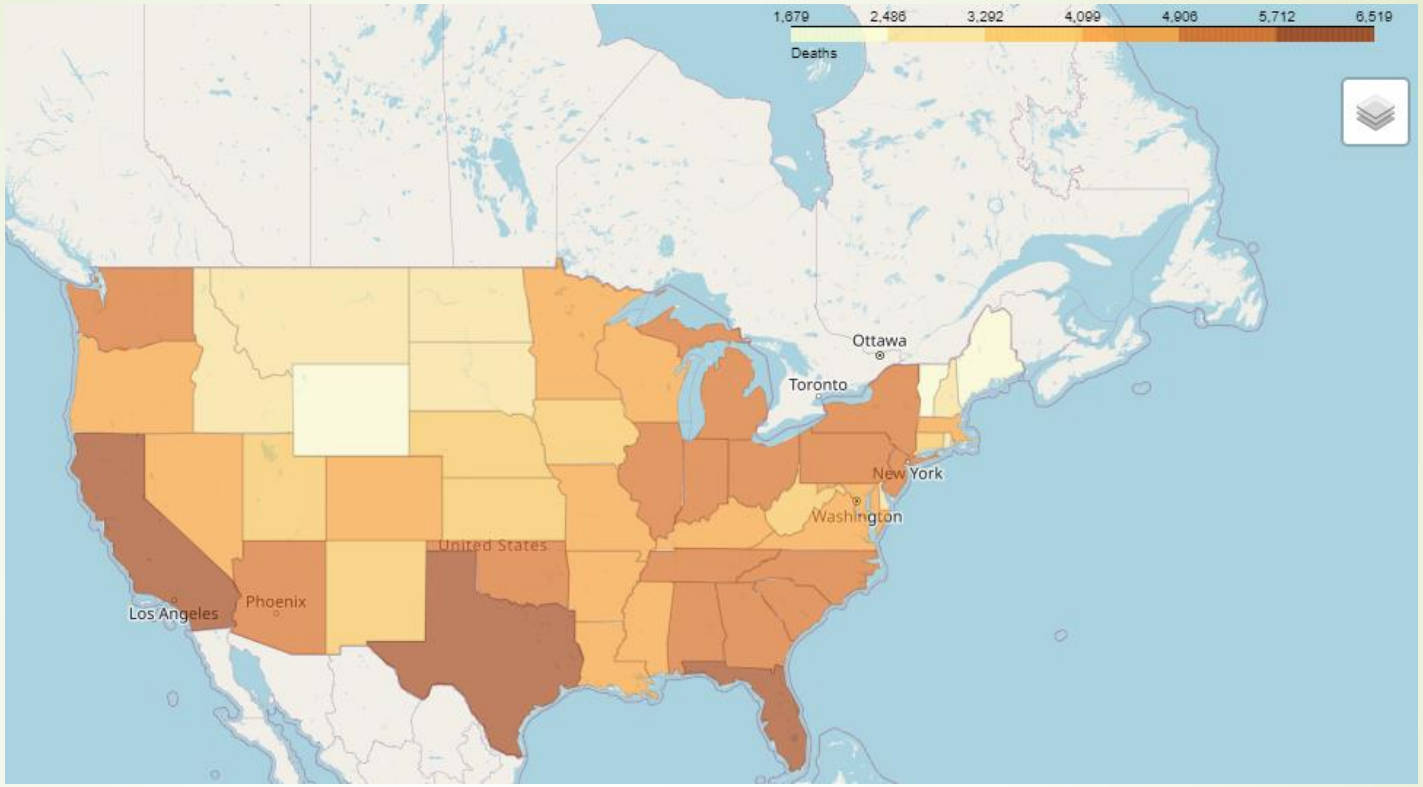
Account Balance

Enhance personalized financial advisory services tailored for the financial needs of people with higher balances.

Products

Encourage product diversification among customers.

plot_data.html



Conditions Contributing to COVID-19 Deaths

Conditions Contributing to COVID-19 Deaths, by State and Age, Provisional 2020-2023

The goal of this project is to analyze the factors contributing to COVID-19 deaths in the U.S., utilizing the CDC's dataset titled "Conditions Contributing to COVID-19 Deaths, by State and Age, Provisional 2020-2023." The project aims to examine demographic, geographic, and temporal trends in COVID-19 mortality, with a focus on how various conditions influenced death counts over time and across different states. This analysis will provide valuable insights into the pandemic's impact on different populations and the role specific conditions played in mortality rates.

Key Steps:

- **Write Python Scripts:** Developed Python scripts and documented the analysis using Jupyter Notebook.
- **Import:** Loaded datasets and necessary Python libraries.
- **Data Cleaning and Wrangling:** Used Pandas and NumPy for data cleaning, transformation, and preparation.
- **Data Quality and Consistency:** Conducted data quality checks, including frequency counts.
- **Exploratory Data Analysis (EDA):** Performed data manipulation.
- **Visualization:** Created graphs and visualizations using Matplotlib, Seaborn, and SciPy.
- **Export:** Exported datasets to XLSX files.

TOOLS USED

Jupyter
Python
Anaconda
Python Libraries

Conditions Contributing to COVID-19 Deaths, by State and Age, Provisional 2020-2023

Key Questions

- Does a higher number of mentions of specific conditions on death certificates lead to a significantly higher COVID-19 death count?

Data

Data source: Career Foundry

Skills Applied

- Exploratory data analysis
- Geographical visualizations with Python
- Regression analysis on Python
- Clustering on Python
- Sourcing open data
- Sourcing time series data
- Data dashboards

Challenges

Challenges included managing complex queries and maintaining data accuracy.

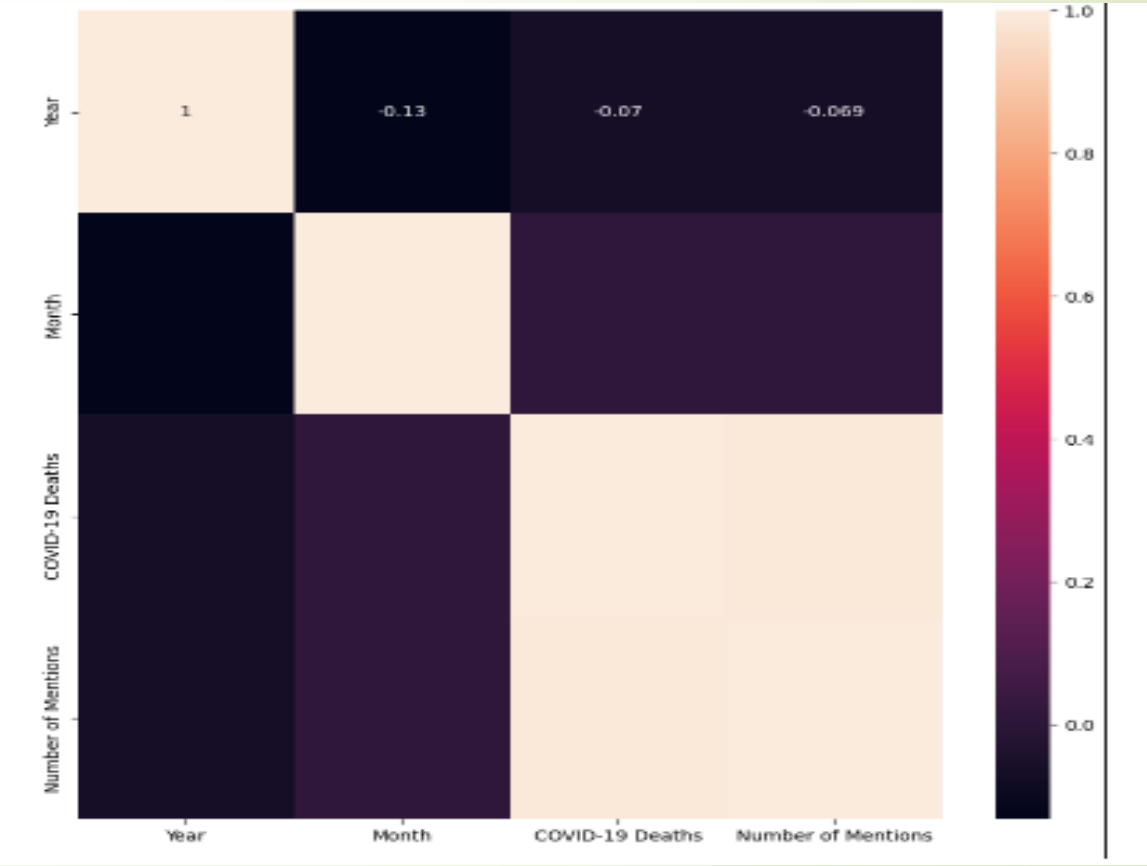
Analysis

I analyzed the COVID-19 data of U.S. from 2020 to 2023.

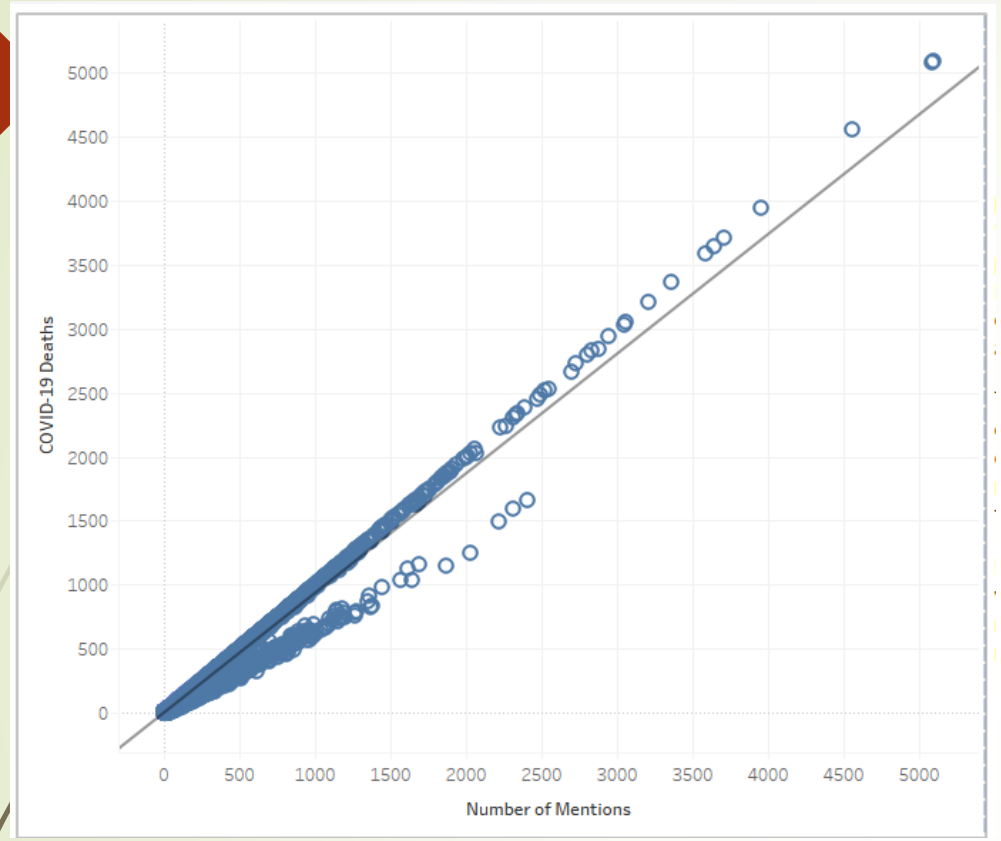
The following results are delivered:
· A detailed repository on GitHub with all Python Code
· A thorough data storyboard on Tableau including all necessary visualizations, demonstrating the objective, hypothesis, test results, and conclusion of the analysis.

1. Exploratory analysis of variables

The analysis revealed a correlation value of 0.989 between "COVID-19 Deaths" and "Number of Mentions," indicating a strong positive correlation between the two variables. This suggests that as the number of COVID-19 deaths increases, the number of mentions of contributing conditions tends to rise proportionally. However, it is essential to recognize that correlation does not imply causation. Although the two variables are closely related, this statistical relationship does not necessarily mean that one causes the other; rather, they tend to move together.

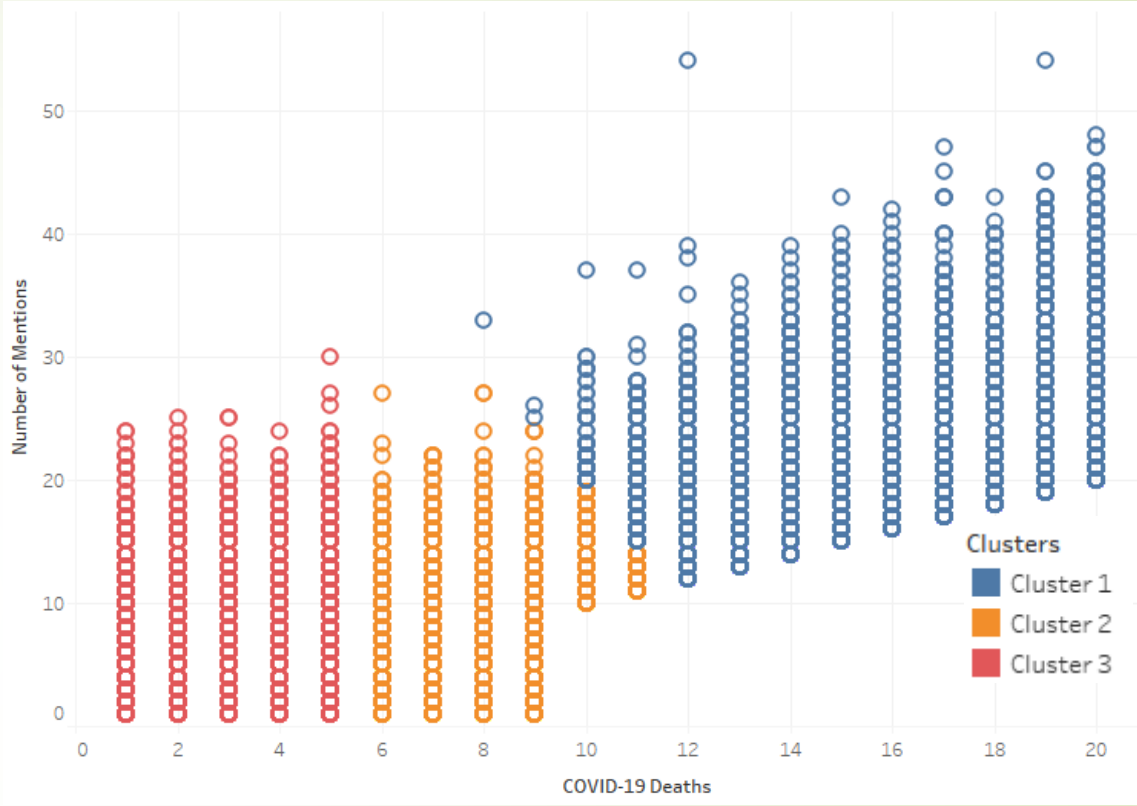


2. Regression Analysis



I conducted a regression analysis to examine the relationship between "COVID-19 Deaths" (independent variable) and "Number of Mentions" (dependent variable). All the data points are close to the regression line exhibiting the Strong Relationship. This suggests that as the number of COVID-19 deaths increases, the number of mentions tends to increase as well.

3. Cluster Analysis



Cluster 1 (Low Deaths, High Mentions) could represent countries that handled the pandemic well but received media attention for their efforts.
Cluster 2 (High Deaths, High Mentions) might highlight the countries hit hardest by the pandemic, receiving corresponding media attention.
Cluster 3 (Low Deaths, Low Mentions) might represent regions with minimal impact and visibility.

Insights & Recommendations

Summary
The close alignment of data points with the regression line in your analysis of "COVID-19 Deaths" and "Number of Mentions" indicates a strong positive relationship. This suggests that as COVID-19 deaths increase, the number of mentions increases proportionally, and the model is effective for predicting and understanding this relationship.

Recommendations: Strengthen global communication to highlight effective pandemic strategies, especially in underrepresented regions, and ensure equitable resource distribution.
Future Research: Investigate outliers and explore factors like healthcare infrastructure or policy effectiveness that influenced pandemic outcomes beyond media attention.

From a **sampling point of view**, bias could occur if the dataset over-represents certain regions or time periods, leading to an inaccurate reflection of the global situation by emphasizing areas with more available or easily accessible data while underreporting others.

Limitations and Ethical Considerations
The dataset is provisional, so conclusions may require updates. It has reporting delays, inconsistent state standards, data suppression, and potential for double counting of deaths with multiple conditions. Ethical considerations include privacy risks, sensitivity of death-related data, and the possibility of demographic underrepresentation or geographic biases.

Conclusions

Journey

My journey in data analytics was showcased, with complex data being transformed into impactful stories.

Skills

My skills in data cleaning, analysis, and visualization are highlighted by each project.

Future

Advanced predictive models are integrated, and dashboards are refined for improved decision-making

Experience

A unique perspective in data storytelling is brought, driving strategic decisions and delivering actionable insights.

Thank you