Saswata Biswas

Data Analytics Portfolio

[Tableau Profile]

[GitHub Profile]

[LinkedIn Profile]

Projects

GameCo Global Video Game Sales

Staff Deployment planning for Influenza Season

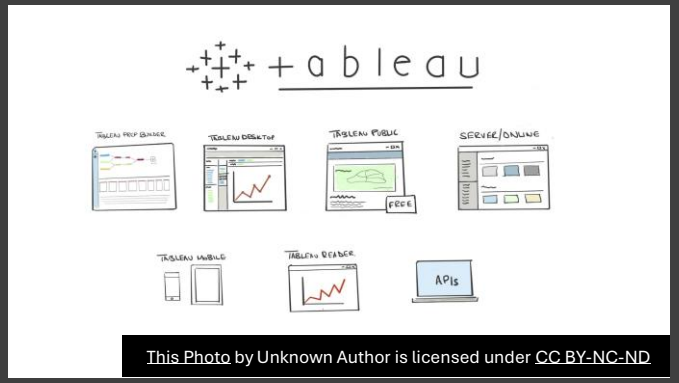
Launching Rockbuster Stealth Online Movie Service

Instacart Market Segmentation to promote Sales

Pig E. Bank Customer Attrition

Future Project

Tools Used



1. Gameco. Global Video Game Sales

Project Overview:
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?

Link for Dataset

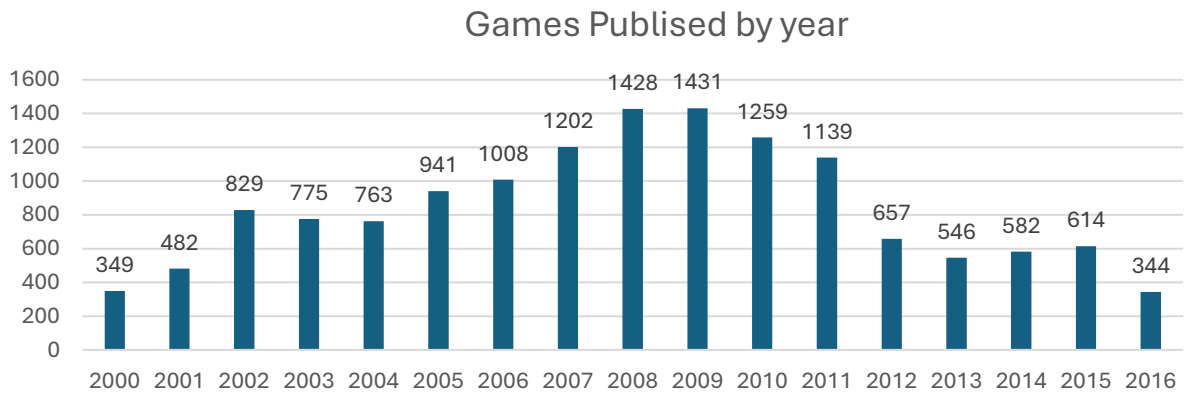
Tools Used:

SKILLS:
- Improving Data quality
- Data grouping and summarizing
- Descriptive analysis
- Pivot table
- Visualization charts in Powerpoint
- Excel/PowerPoint

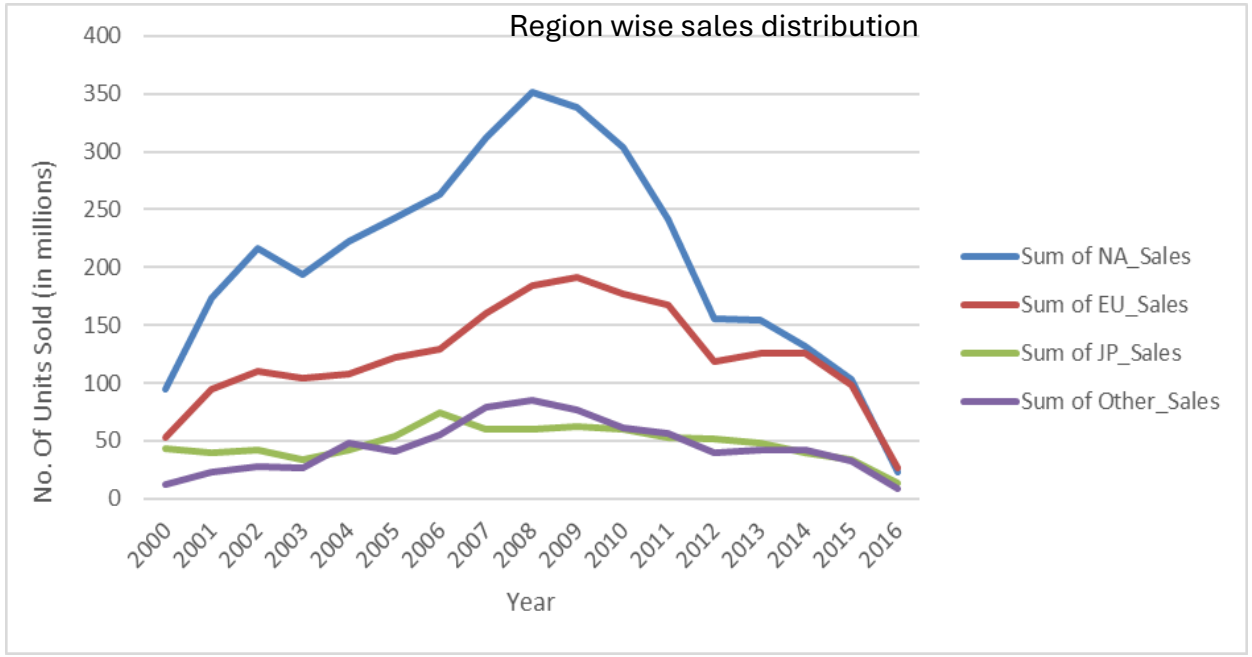
Background: The expectations were Video Games sales have been on the rise over the years and North America led sales globally and certain genres were more popular than the others

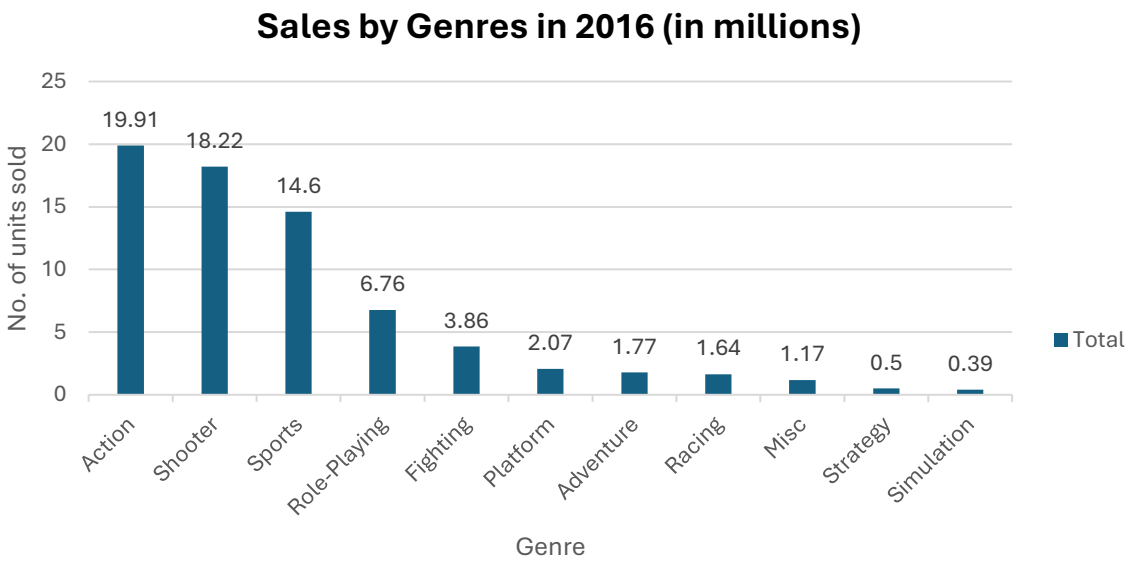
Steps taken:

- 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 column charts of total global sales by genres and games published every year; Line charts of global and regional sales
- Decide on what type of analysis: Descriptive analysis, Diagnostic, Predictive or Prescriptive
- Interpret results and summarize findings


Games Publised by year
1600
1400
1200
1000
800
600
400
200
0
349
482
829
775
763
941
1008
1202
1428
1431
1259
1139
657
546
582
614
344
2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016


Global Sales
Total
No. Of Units Sold (in millions)
800
600
400
200
0
2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016
Year


Region wise sales distribution
No. Of Units Sold (in millions)
400
350
300
250
200
150
100
50
0
2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016
Year
Sum of NA_Sales
Sum of EU_Sales
Sum of JP_Sales
Sum of Other_Sales


Sales by Genres in 2016 (in millions)
No. of units sold
25
20
15
10
5
0
19.91
18.22
14.6
6.76
3.86
2.07
1.77
1.64
1.17
0.5
0.39
Action
Shooter
Sports
Role-Playing
Fighting
Platform
Adventure
Racing
Misc
Strategy
Simulation
Genre
Total

Insights:

- The sales of video games declined since 2008
- The number of games published also declined since 2009
- The genres action, shooter and sports sold almost 53% as compared to the other 8 genres

Recommendations made:

- Publish more games to accomplish higher sales
- Focus more on action, shooter and sports genre games as they are widely popular
- Focus on marketing the games in the North America and Europe regions as they give us more revenue

2. Staff Deployment planning for Influenza Season

Project Overview:
The USA has an increased number of patients during flu season. The medical staffing agency needs to determine when and where to send staff to be better prepared.

Key Questions:

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

Tools Used:

Skills:

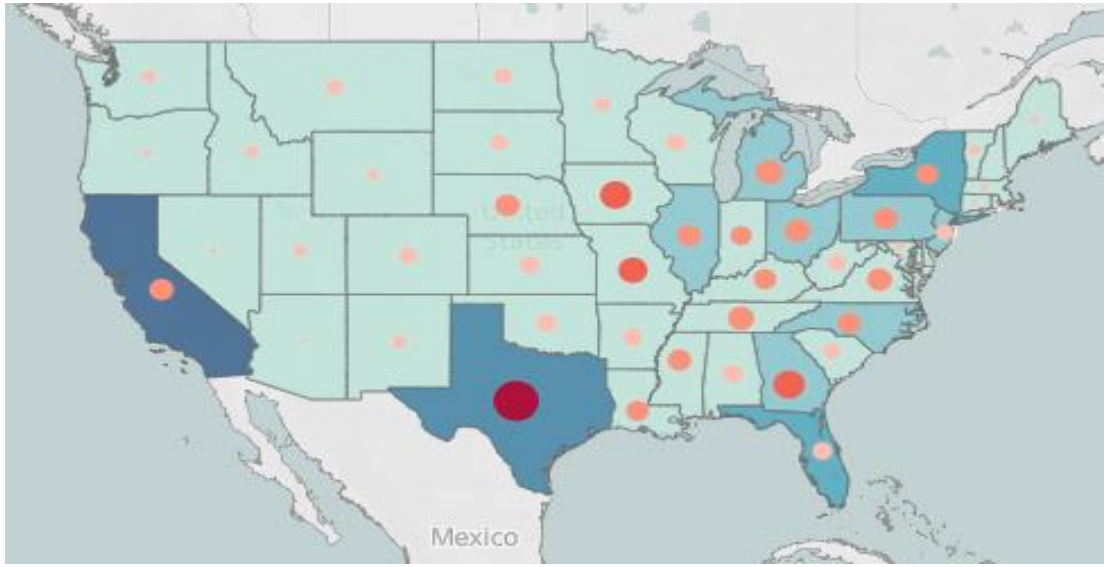
- Translating business requirements
- Data cleaning
- Data integration
- Data transformation
- Statistical hypothesis testing
- Visual analysis
- Forecasting
- Storytelling in Tableau

Analysis Steps:

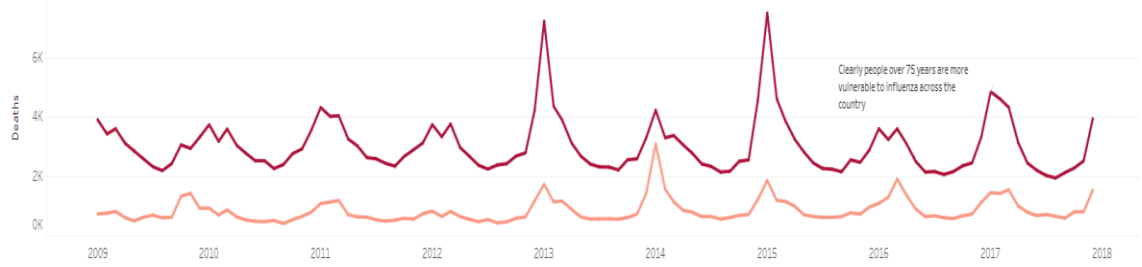
- Finding out who is the most vulnerable population from historical data using Excel
- Finding out where is the most population of vulnerable patients using Excel
- Make visualizations using Tableau to present the data to clients
- Make recommendations for staffing and for next steps

Datasets:

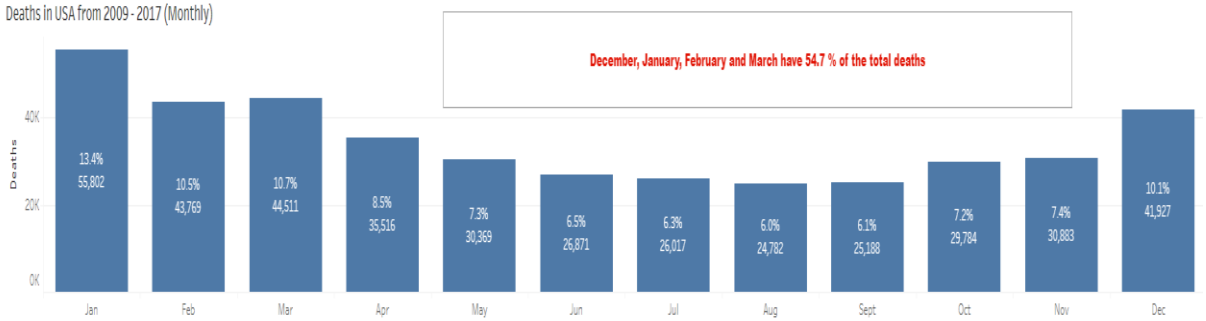
- Influenza Deaths by Geography
- USA population



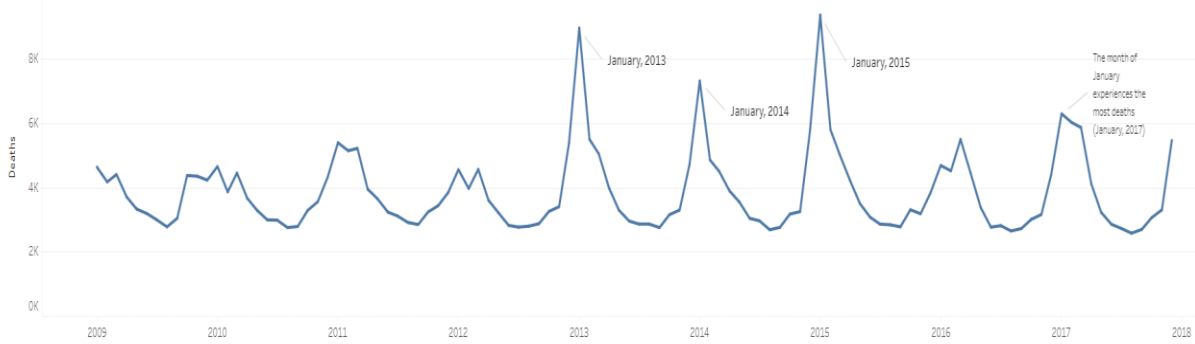
Deaths by age group (2009-2017)



Deaths in USA from 2009 - 2017 (Monthly)

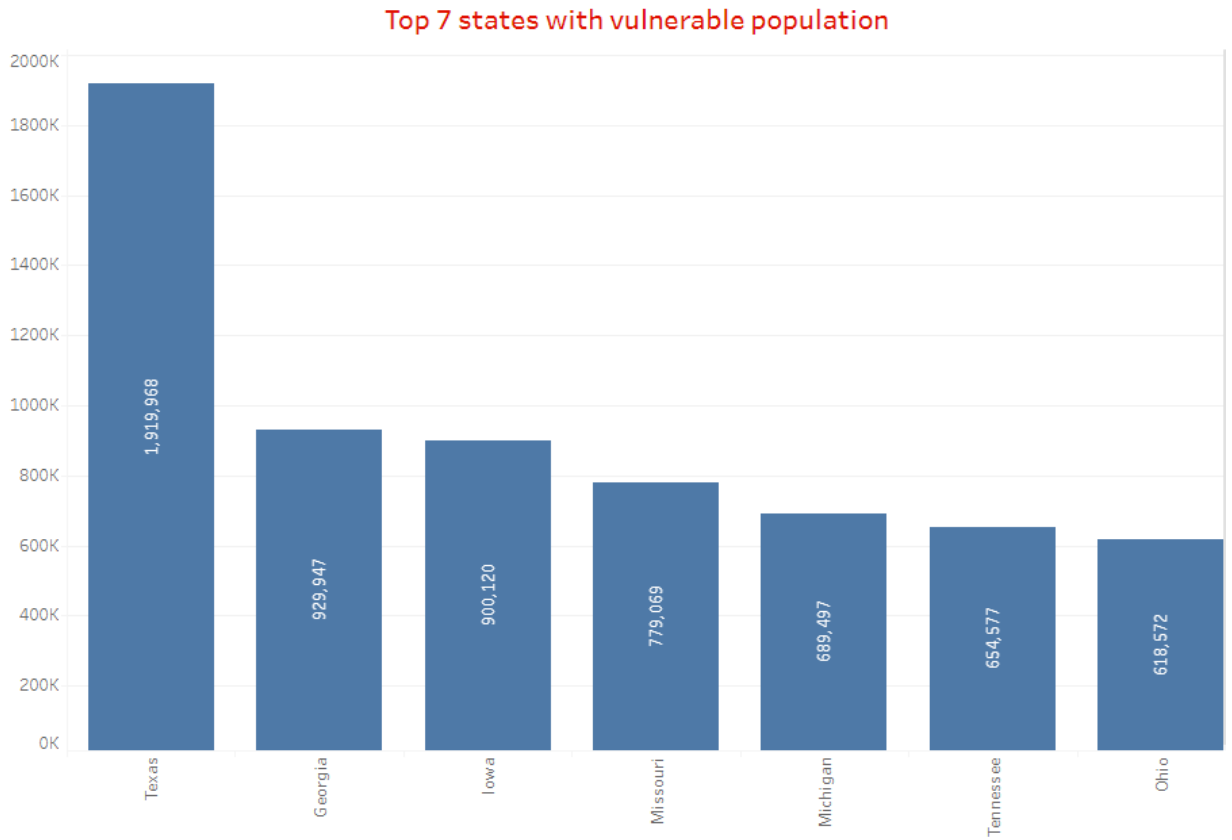


Influenza Deaths in USA by Year (2009-2017)



Insights:

- December, January, February and March account for 54.7% of the total deaths
- The deaths of people over 65 are more than double across all years since 2008 till 2018.
- The top 7 states with the most vulnerable population are Texas, Georgia, Iowa, Missouri, Tennessee and Ohio



Recommendations made:

- Maximize staffing in these 7 states first
- Analyse the effects of influenza on pregnant women, individuals with HIV/AIDs, cancer, heart disease, stroke, diabetes, asthma, and children with neurological disorders.

Links:

- All Visualizations in Tableau
- Video Presentation in Loom

3. Launching Rockbuster Stealth Online Movie Service

Project Overview:
To launch an online video rental service with existing movie licenses.

Key Questions:

- 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?

Tools Used:

Skills:

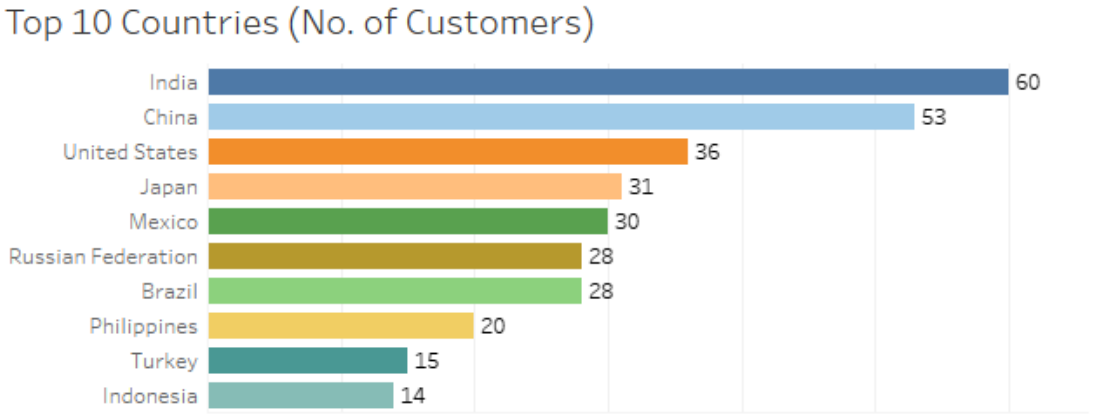
- Using Relational databases
- SQL Database querying
- Data Filtering
- Cleaning and summarizing
- Joining tables
- Subqueries
- Common table expressions (CTE)

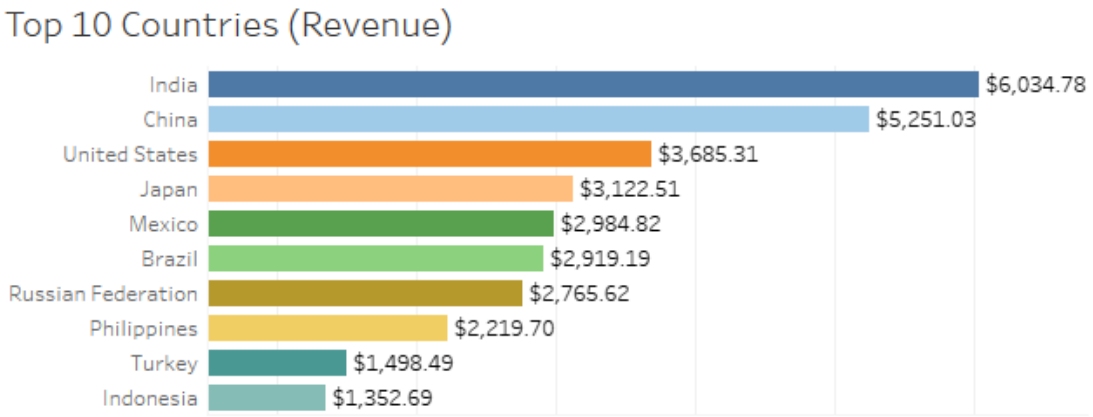
Analysis Steps:

- Uploaded Rockbuster's data into PostgreSQL relational database.
- Create a Data dictionary
- Used SQL to answer key business questions
- Identified top-performing movies and customer locations.
- Create Visualizations of results
- Challenges included complex joins and ensuring data accuracy, which were addressed by thoroughly cleaning and validating the data.

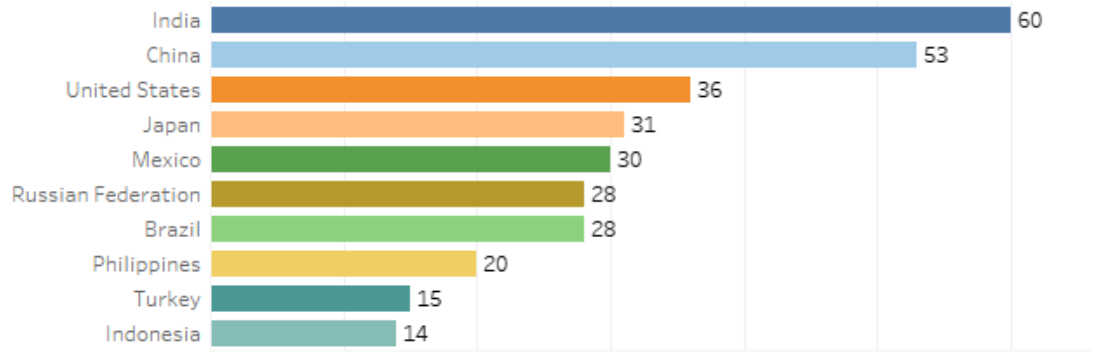
Data Set:

Rockbuster Data

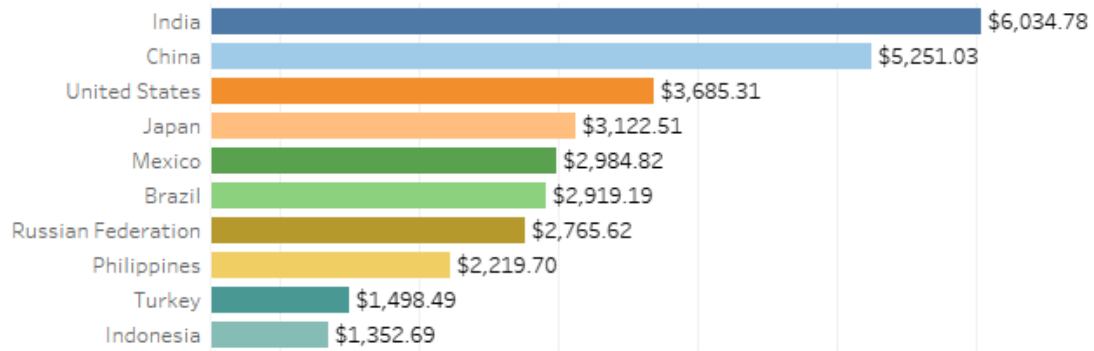




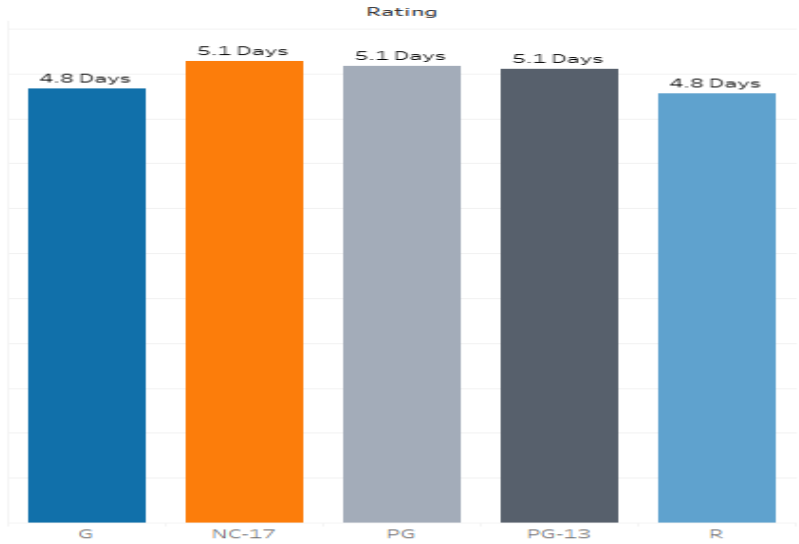
Top 10 Countries (No. of Customers)



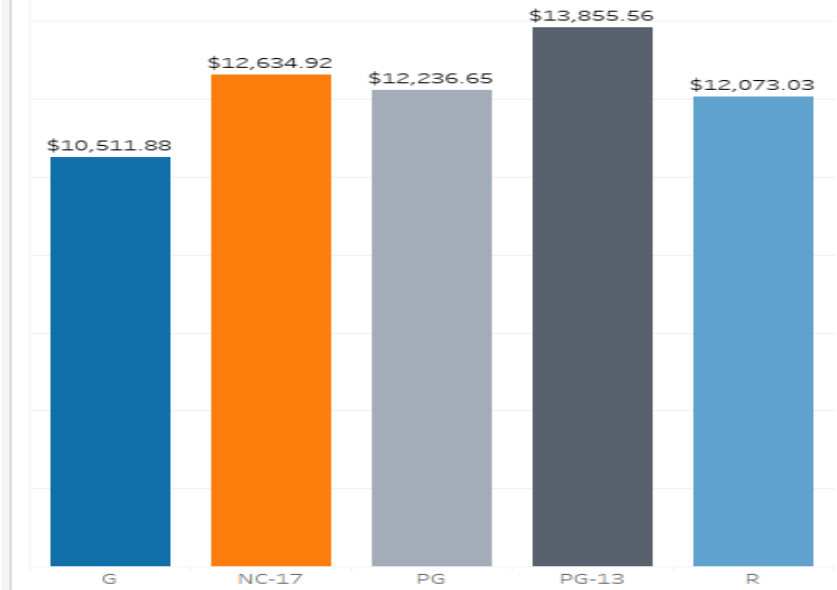
Top 10 Countries (Revenue)



Average Rental Duration (Film Rating)



Total Revenue (Film Rating)



Insights:

- 1000 movies available, 42 movies did not generate revenue
- Average rental duration was 5 days
- Total customers were 599 spread across 108 countries
- Revenue generated was directly proportional to number of customers in countries

Recommendations made:

- Remove movies that have not generated any revenue
- Allocate marketing budget in the top 10 countries
- Create customized marketing strategies and loyalty programs for high-value customers, and attract mid-tier customers with specific promotions and incentives
- Advertise the availability of most popular movies
- Get films that are released after 2006 as well

Links:

- All Visualizations in Tableau
- Github Project overview

5. Instacart Market Segmentation to promote Sales

Project Overview:
The USA has an increased number of patients during flu season. The medical staffing agency needs to determine when and where to send staff to be better prepared.

Key Questions:

- Which are the busiest days and hours of the week?
- Are certain products/departments more popular?
- Can products be grouped as per their price tags?

Tools Used:

Excel python jupyter

Skills:

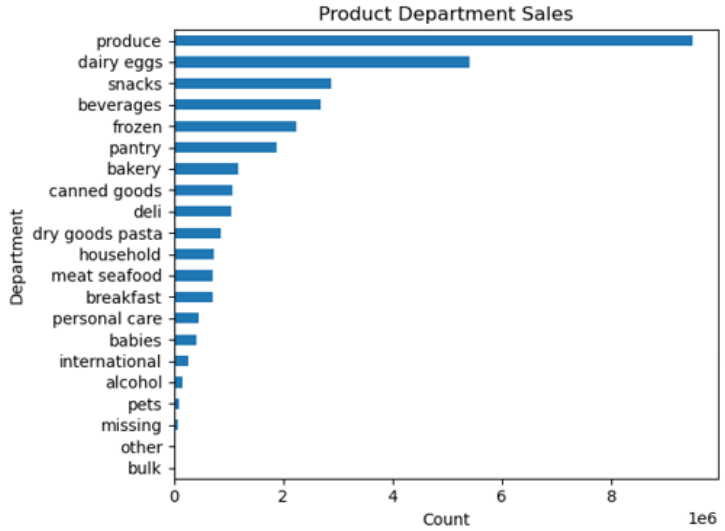
- Translating business requirements
- Data cleaning
- Data integration
- Data transformation
- Statistical hypothesis testing
- Visual analysis
- Forecasting
- Storytelling in Tableau

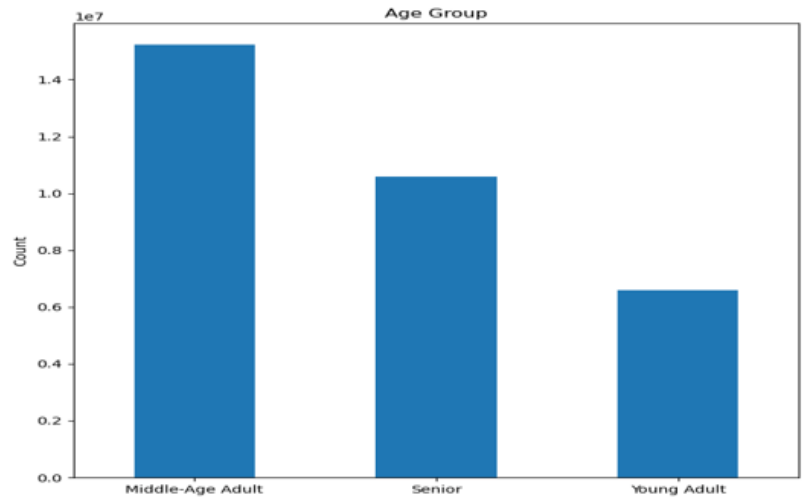
Analysis Steps:

- Installing Anaconda, Jupyter and Python
- Creating project folder and installing python
- Downloading and cleaning data
- Merging datasets for better overview
- Creating new variables
- Analysing dataset and producing recommendations

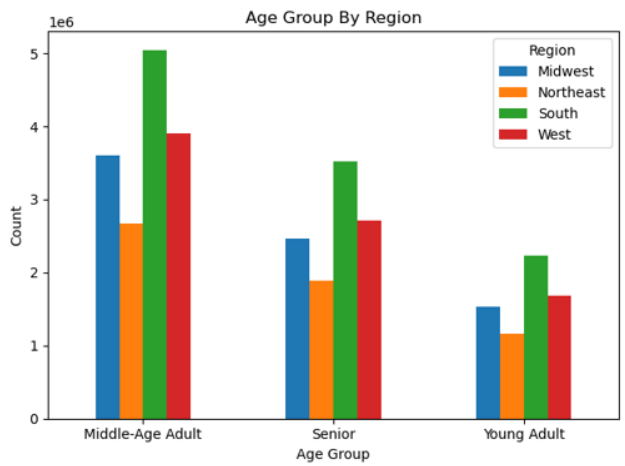
Challenges faced:

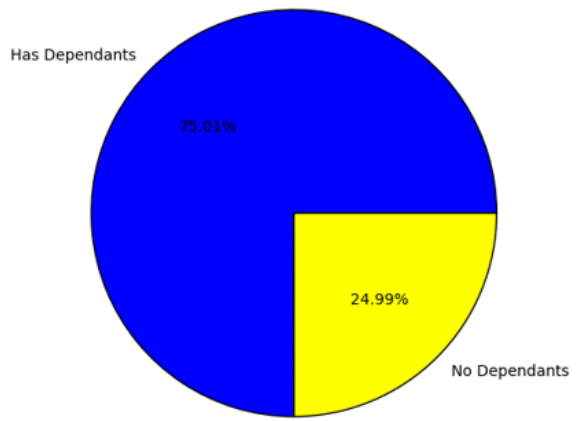
- The libraries require lots of free space in the laptop, cleaned up all non-useful data
- The Jupyter notebook scripts require lots of RAM to run properly on the large data-sets, installed new RAM
- Some visualizations were difficult to produce, used ChatGPT to provide explanation





Age Group vs region





Insights and recommendations:

- The least busy hour is from 11 pm to 6 am and the least busy days are Tuesday and Wednesday. Increase ads during the busiest hour (11 am till 3 pm) and busiest days (Saturday and Sunday) offering some discounts for the least busy hours and days
- Prices see their highest peak at 4 am, but we need to understand that the number of customers are the least during these hours, so ads displayed in these hours will not reach the intended number of customers
- Instacart needs to check the prices that are over 25 dollars as there seems to be some discrepancy there. Most of the prices can be tagged into 3 different tags, Low: 0.1 to 4.99, medium: 5.00 to 15.00 and high: 15.01 to 25.00, anything above this should either be removed or put into Expensive category
- The produce, dairy/eggs and snack departments rule over all other departments across age-groups, family status and regions. They should focus more on these products to increase sales. Also, it would help sales if they could come up with some sort of offers on products which are low on sales to boost them up
- We can clearly see below that more than 80% of our customer are regular or loyal and more than 90% of the customers are frequent.
- There is no correlation based on customer's region. However, based on the population, I would recommend to focus on South region as it has the highest population

Links:

Github Project overview

6. Pig E. Bank Customer Attrition

Project Overview:
Working with the bank's client data to understand the potential reasons for attrition

Key Questions:

- Which customers are more prone to leaving?
- Does age, salary have any relation?

Tools Used:

Excel

GitHub

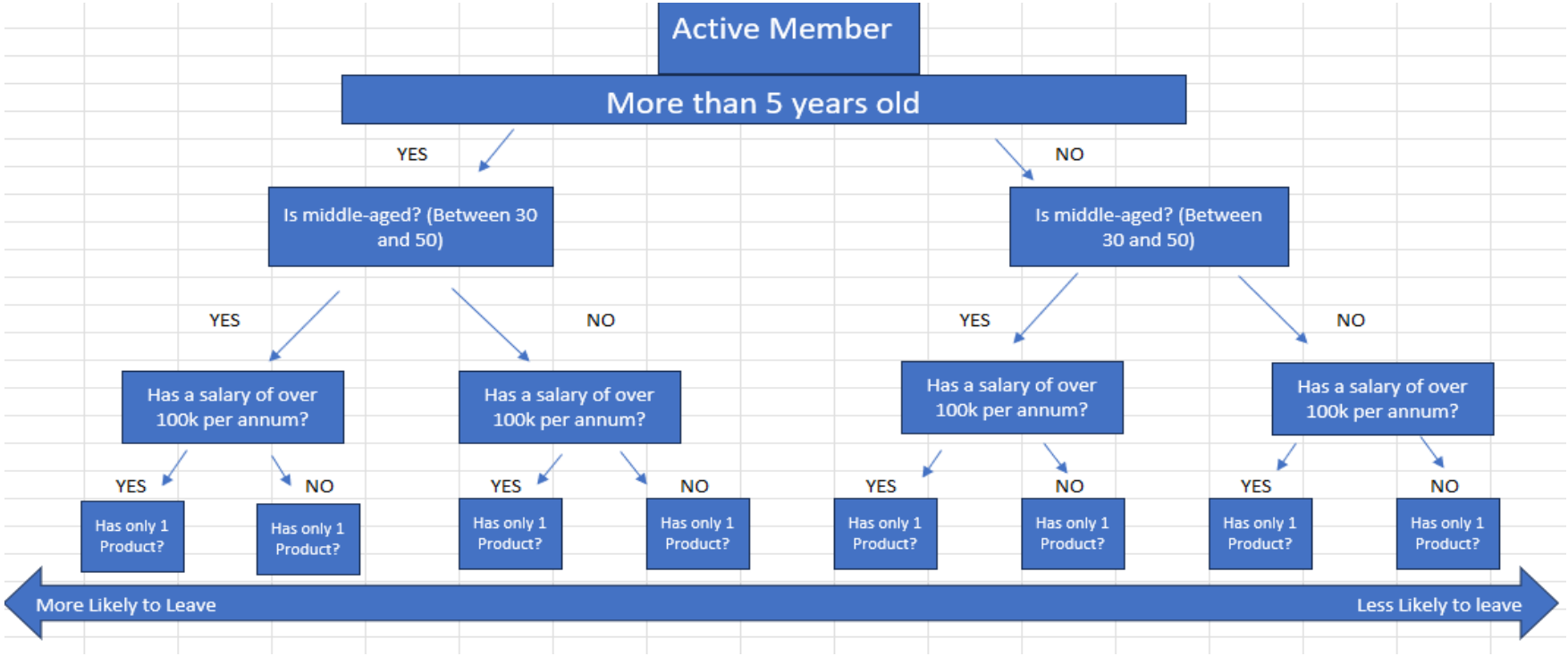
Skills:

- Big data
- Data ethics
- Data mining
- Predictive analysis
- Time series analysis & forecasting
- Decision tree
- Using GitHub

Analysis Steps:

- Download and clean data set
- Separate client data into former and existing clients
- Find patterns in the former client dataset
- Make a decision tree

Statistics	Former	Current	All
Average Age	45	37	39
Max Age	69	82	82
Min Age	2	2	2
Average Salary	97155.20	98942.45	98574.54
Max Salary	199725.39	199661.50	199725.39
Min Salary	417.41	371.05	371.05
Average Credit Score	637	652	649
Max Credit score	850	850	850
Min Credit Score	376	411	376
Average Balance	90239.22	74830.87	78002.72
Max Balance	213146.20	197041.80	213146.20
Minimum Balance	0.00	0.00	0.00



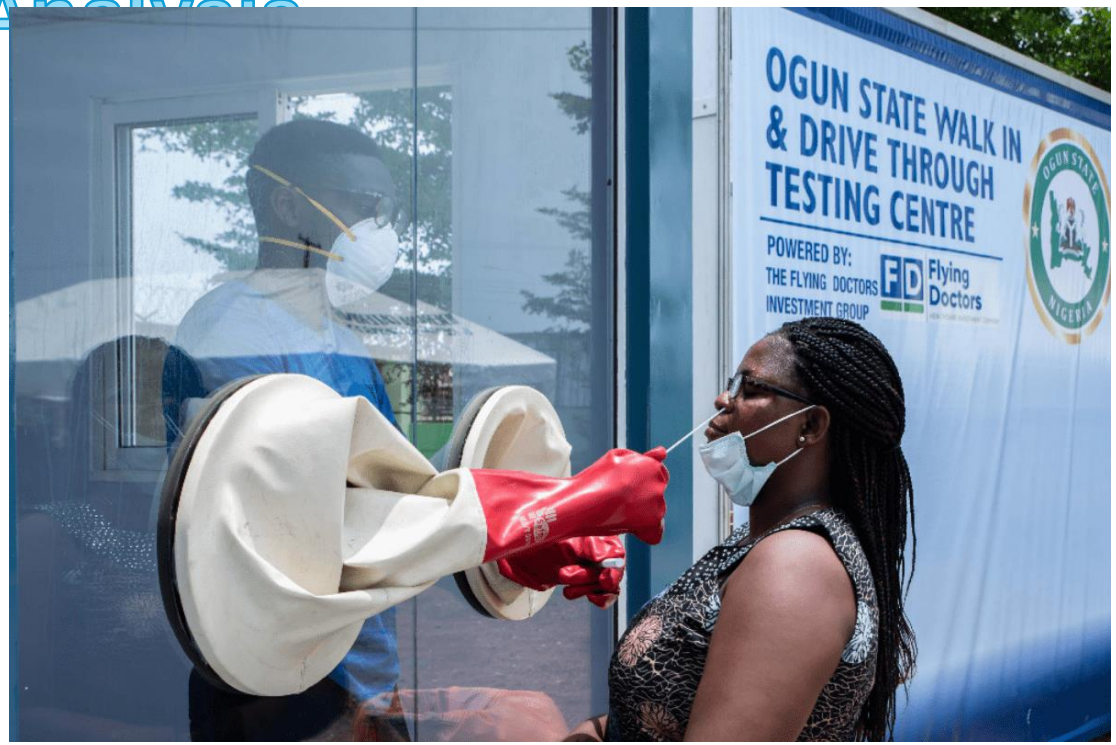
Insights: Factors leading the client loss:

- 78% of clients were 5 years or above
- 65% of clients were middle aged (Between 30 and 50 years old)
- 60% of clients had a high salary range (above $10,000,0)
- 62% of the clients had a medium credit score (between 500 and 700)
- 70% of clients had only one product.

Recommendations made:

- Reward programs for clients over 5 years
- Target middle aged customers with a high salary range to have more than one product

6. Covid-19 Analysis



Project Overview:
Choosing and analysis of an open-source data and analysing it for insights. I chose Covid data as I have vast experience in health insurance

Key Questions:

- Who are the vulnerable population?
- Does time have any relation to the outbreak?

Tools Used:

python™ jupyter GitHub

+ableau

Excel

Skills:

- Sourcing open data
- Data ethics
- Data mining
- Predictive analysis
- Time series analysis & forecasting
- Using GitHub
- Using Python
- Using Tableau

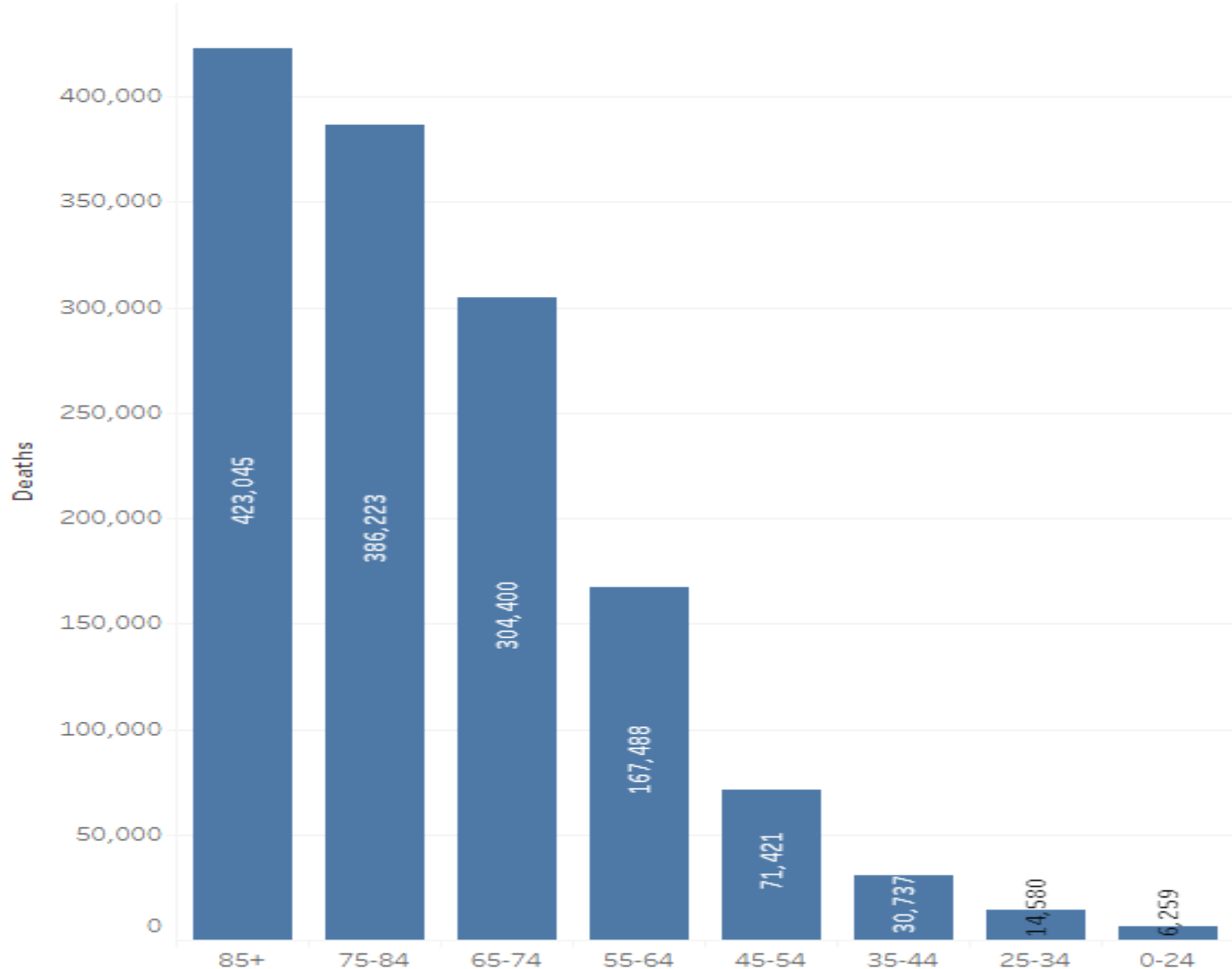
Analysis Steps:

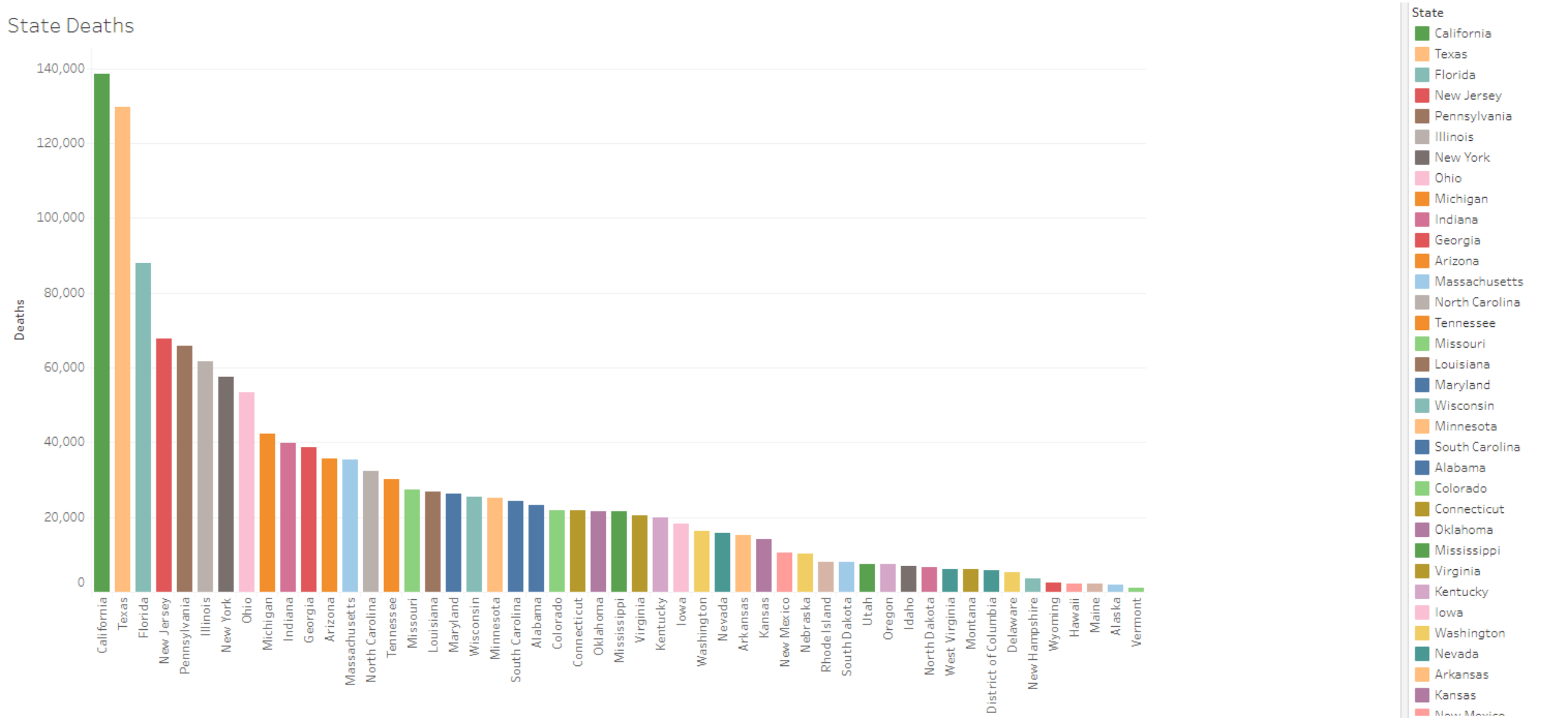
- Data choosing and cleaning
- Exploring relationships
- Geographically representing data
- Regression analysis
- Clustering analysis
- Time series data
- Creating dashboards

Challenges faced:

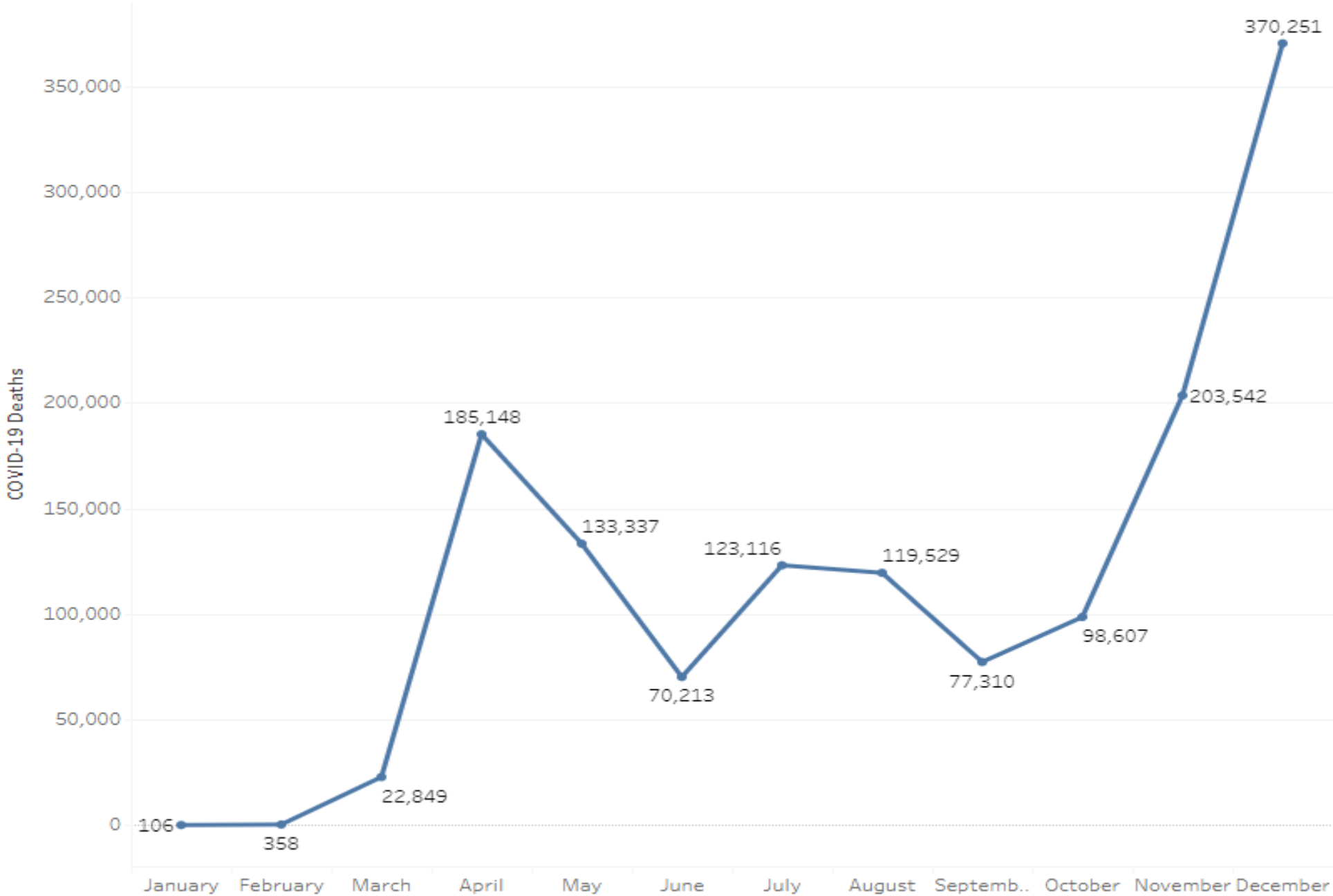
- Choosing data as there were multiple requirements
- Time-constraint as I had to finish learning and delivering within 5 days. Worked on an average for 16 hours everyday.
- The geographic visualization faced multiple errors, used chatGPT to provide explanation

Age Group Deaths




State Deaths
Deaths
140,000
120,000
100,000
80,000
60,000
40,000
20,000
0
California
Texas
Florida
New Jersey
Pennsylvania
Illinois
New York
Ohio
Michigan
Indiana
Georgia
Arizona
Massachusetts
North Carolina
Tennessee
Missouri
Louisiana
Maryland
Wisconsin
Minnesota
South Carolina
Alabama
Colorado
Connecticut
Oklahoma
Mississippi
Virginia
Kentucky
Iowa
Washington
Nevada
Arkansas
Kansas
New Mexico
Nebraska
Rhode Island
South Dakota
Utah
Oregon
Idaho
North Dakota
West Virginia
Montana
District of Columbia
Delaware
New Hampshire
Wyoming
Hawaii
Maine
Alaska
Vermont
State
California
Texas
Florida
New Jersey
Pennsylvania
Illinois
New York
Ohio
Michigan
Indiana
Georgia
Arizona
Massachusetts
North Carolina
Tennessee
Missouri
Louisiana
Maryland
Wisconsin
Minnesota
South Carolina
Alabama
Colorado
Connecticut
Oklahoma
Mississippi
Virginia
Kentucky
Iowa
Washington
Nevada
Arkansas
Kansas

Monthly Deaths



Insights:

1. Influenza and pneumonia and cardiovascular diseases are most common conditions to COVID-19 deaths across all age groups, especially older age groups affected. Some conditions like ischemic heart disease, diabetes, and renal failure are also high COVID-19 deaths in older age groups.
2. Older population consistently show higher death counts for multiple conditions.
3. The winter months are likely to cause a spike in the infection
4. California, Texas, Florida, New Jersey, Pennsylvania are the top 5 affected states mainly because they are also the top 5 most populated states.

Recommendations:

1. Allocate more resources to areas which have more population
2. General population should be more careful during winter months
3. People with pre-existing conditions as mentioned above should be extra careful

- Github Project Overview
- Tableau Project Overview

Thank you