Data Analytics Portfolio

Richa Vijay

January 2022

Introduction

This portfolio contains the case studies from the Data Analytics Program I was enrolled in with CareerFoundry Educational Platfrom in 2021

This Portfolio consists of 4 different projects where I analyzed and created data visualizations , given recommendations for different datasets and business problems for different kind of Industries.

While pursuing these projects I learned popular Data Analytics tools, languages and techniques likes Advanced Excel, SQL, Tableau, Python

All Project Documentation can be found on Github

Table of Contents

Game Co : Analysis of Video Game Popularity and Sales……**Project 1**
Tools : Excel

Preparing for Influenza Season :Analysis of Influenza Season for Staffing Agency……**Project2**
Tools : Excel and Tableau

Rockbuster Stealth Data Analysis : Data Analysis for launching Online Platform……**Project3**
Tools : POSTGRESQL & Tableau Public , Language: SQL

Instacart Grocery Basket Analysis : Analysis of Sales Pattern and Customer Segmentation……**Project4**
Tools : Jupyter ,Language : Python

Project 1 : Game Co

GAME CO

Analysis of Video Game Popularity and Sales

TOOL : EXCEL

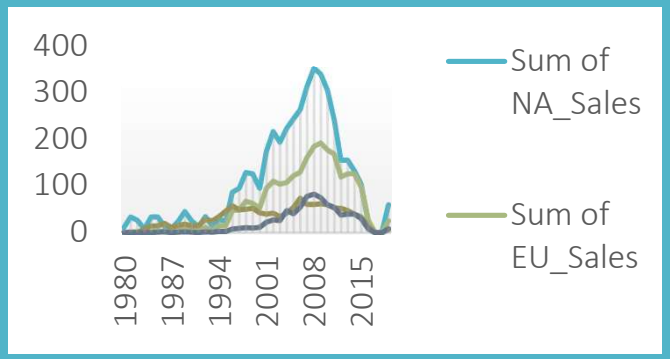
GameCo is a video game company with a strong presence globally with strongest market in North America, Europe and Japan.

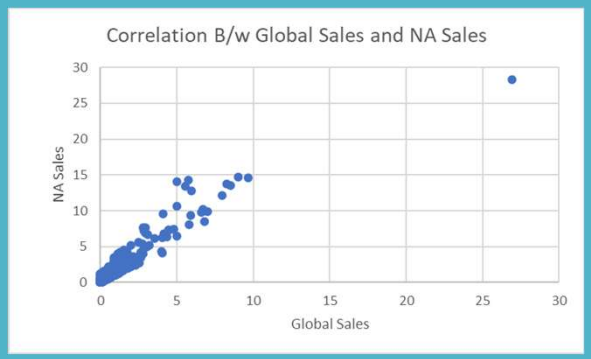
- Project Goal : To perform descriptive analysis of historical video game sales data to foster a better understanding for the development of new video game launches and changes in marketing budget accordingly

- Data Source : VGA Chartz was the main source of data. The dataset used here contains 16599 observations with sales numbers for each game's title from 1980-2020, as well as the game's genre, platform, publisher, and publishing year.

- **Data Exploration** : sorted , filtered , created some visualizations and summarized the data with help of pivot tables and different Excel functions.
- **Data Cleaning and Data Consistency Checks** – Finding and Fixing Incomplete , Duplicate and Missing records using Excel Data cleaning techniques.
- **Data limitations & Data bias Check** : That may have caused due to the data collection methods used and any potential Data bias.
- **Descriptive Statistic Analysis** : Worked with measures of Central tendency , distribution, spread, quartiles, data outliers using various excel functions and visualizations.
- **Data visualizations** : Created histograms, box and whisker plots, and scatterplots

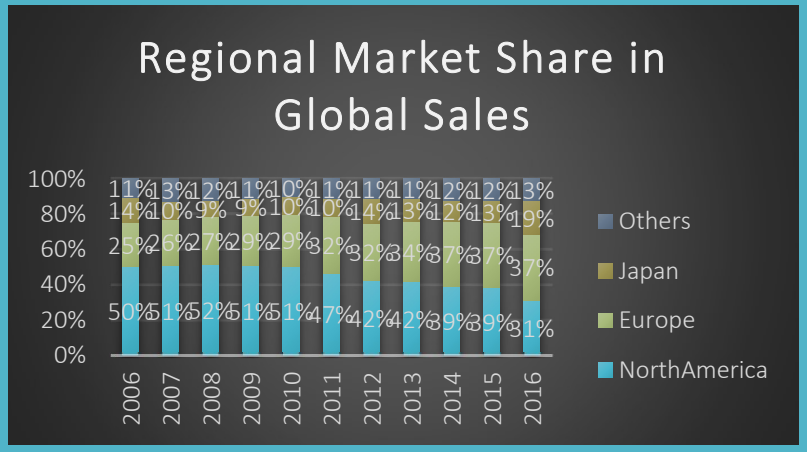
When combined Statistics with Visualizations, I could clearly see how certain factors in data are skewing the results.

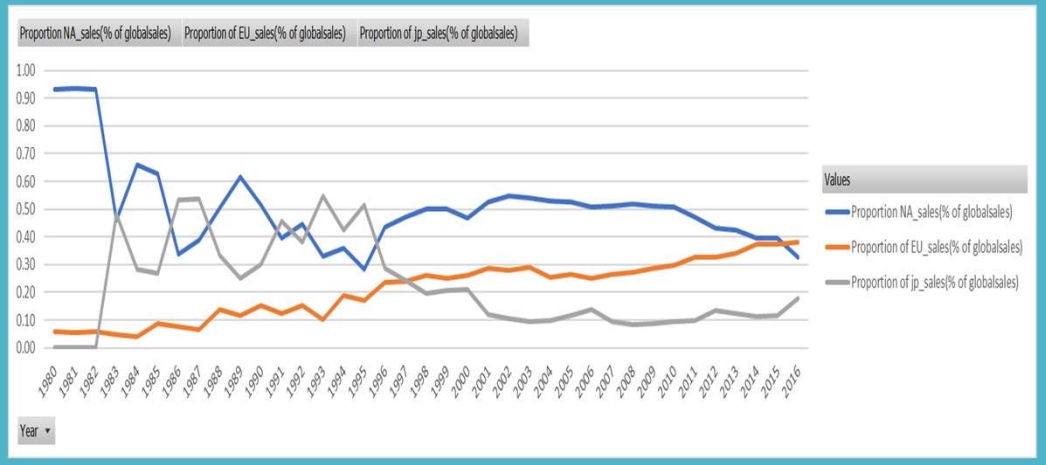




Based on the data analysis , I have drawn my expectations about , how the sales numbers are going to behave overtime, gave my expectations and proved that my expectations are very close to reality.

In the final presentation for the management board of GameCo, I showcased my findings and put them in the context of the goal of the analysis. The understanding of GameCo was, that sales for the various geographic regions have stayed the same over time. By conducting the analysis, I challenged their expectations and recommended changes in the marketing budget for the next year.





GAME CO

Analysis of Video Game Popularity and Sales

Recommendations :

- Game Sales across the different regions have been fluctuating overtime. After the peak of 2008 , there is a decline of almost 89.4% in Global sales and all the regions recorded their lowest sales in year 2016

- We need budget revision for all the regions as well as the further investigation to find out the root cause for decline in sales of video games.

- Marketing budget revisions should be made after consideration of regional sales trends ,genre popularity, consumer preference , Mode of playing (Online V/s Console based V/s Mobile based)

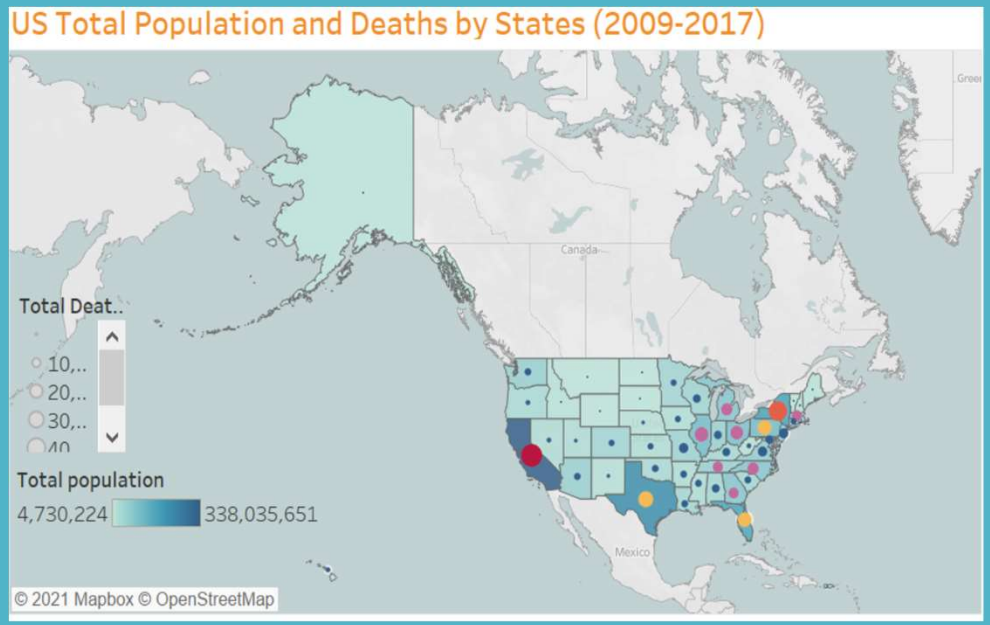
Please find the copy of complete presentation here

Project 2
Preparing for Influenza Season

Preparing for Influenza Season

Analysis of Influenza Season for Staffing Agency

Tools : Excel and Tableau



Project Motivation : The United States has an influenza season where more people than usual suffer from the flu, particularly vulnerable population, and hospital needs additional staff to treat these patients. Main objective is to determine when to send staff and how many to each of the 50 States.

I had a chance to expand on Excel analytical functions and work on visualizations in Tableau Public while working on this project.

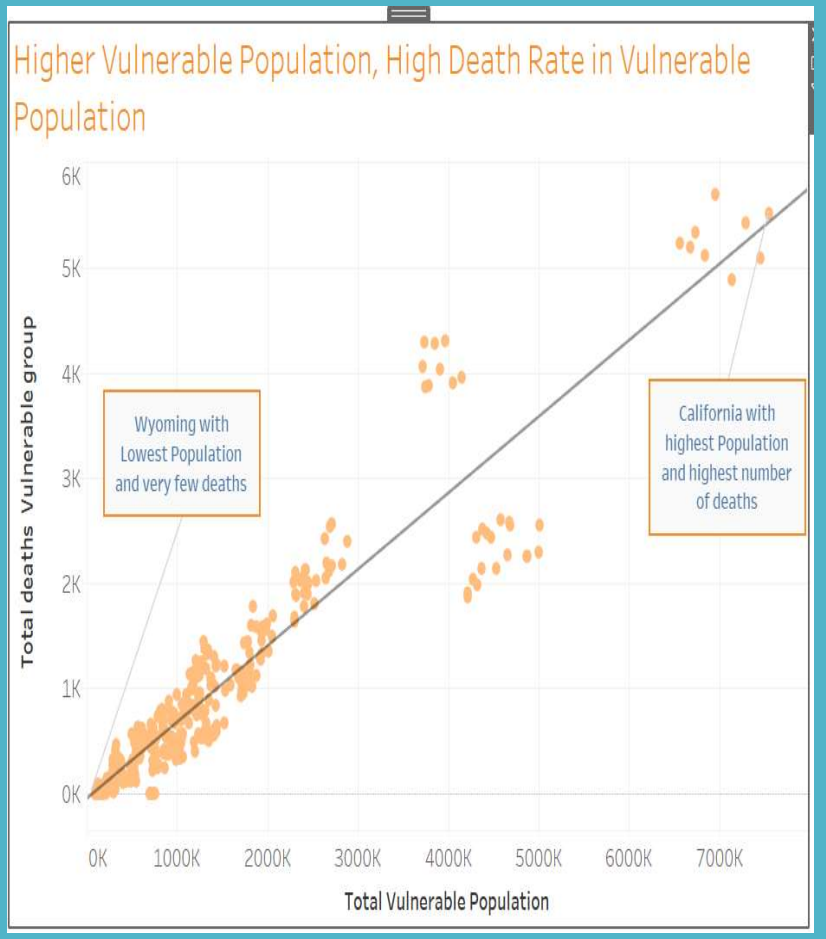
Data Source : The data about demography(Population), Influenza deaths, Influenza tests, flu shots and doctor's visits were collected in 5 separate sets. Sources of the data are US Census Bureau and CDC.

In this Project beginning ,I started **translating business requirements** into clarifying and funnel questions from data perspective that guide me through analysis , followed by project plan creation.

- **Data Integrity, Accuracy and Consistency Checks :** Excel tools used : sorting, filtering, Pivots, Visualization and various statistical functions.

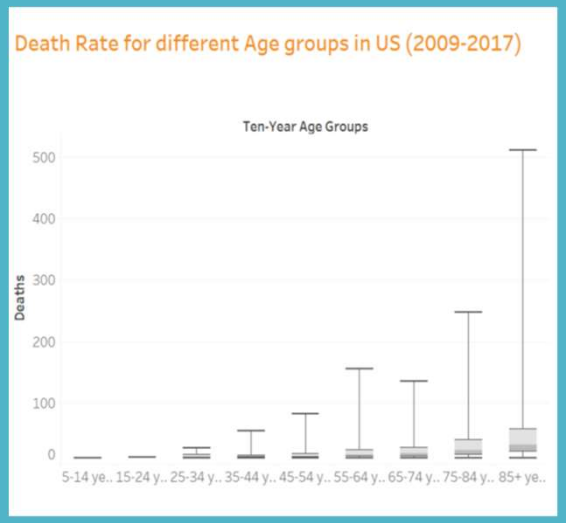
- **Data Integration and Mapping** : Since data was spread between two sets, used V- Look ups and Pivots to map and integrate data

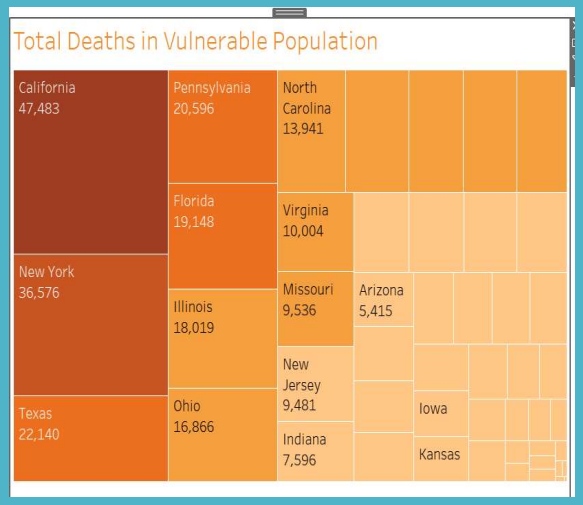
- **Inferential statistical analysis :** Analysis where I learned how to turn how to turn my research hypothesis into a testable statistical hypothesis.

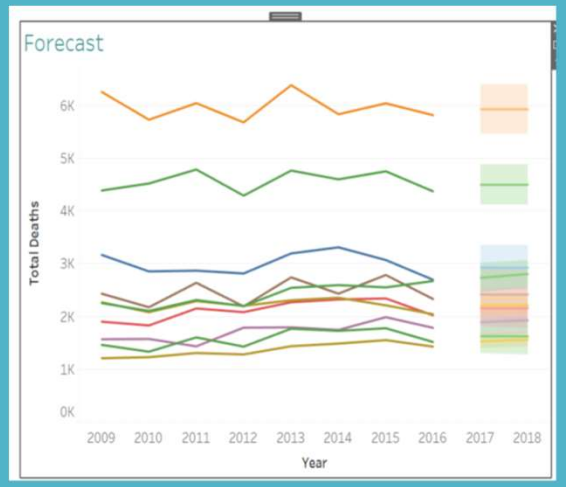


- **Tableau Visualizations :** The second part of the influenza project concentrated on to draw insights from the same Influenza data using data visualization techniques of Tableau. I created composition & comparison charts(Pie, Bar, Column, Tree Map), temporal, spatial, and textual charts.

- **Forecasting :** Since the scope of the project required looking for insights to prepare for the next season, I also generated a forecast and created some Statistical visualization like histogram , Box Plots to look at the distribution of variable and correlation between different variables.







PREPARING FOR INFLUENZA SEASON - Analysis

Reccomendations...

- From the Correlation between Population and number of deaths we can conclude that States with higher Population have higher number of deaths and they need the allocation of additional staff
- Vulnerable Population (65+ years) needs special focus as they are most prone to develop Chronic Flu Symptoms , States with higher Vulnerable Population must be the top priority for additonal Staffing during the peak Influenza Season.
- Seasonality of Flu and Forecasting for different States should be taken under consideration for additional Staff planning.
- Conducting short surveys with the Permanent & Temporary staff as wll as with the Patients, post execution of the Staffing Plan . This will help in analyzing the effectiveness and areas of improvement based on the feedback received.

In Tableau's dashboard, I produced report with my findings along with recommendations and video-recorded the insights for the stakeholders

Project 3
Rockbuster Stealth Data Analysis

ROCKBUSTER STEALTH DATA ANALYSIS

Data Analysis for launching Online Platform

Tools : POSTGRESQL & TABLEAU PUBLIC

LANGUAGE : SQL

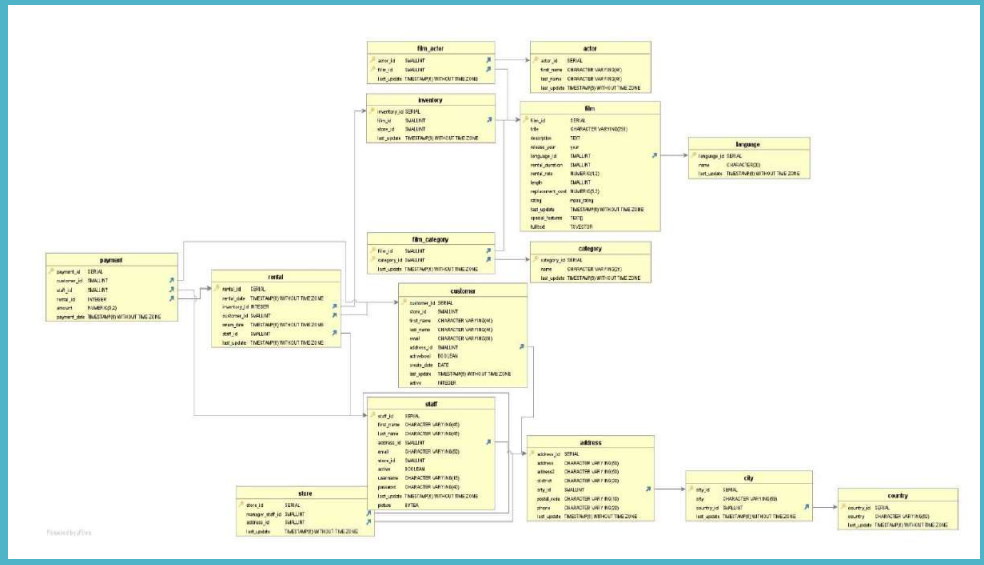
Rockbuster Stealth LLC is a movie rental company that used to have stores around the world. Facing stiff competition from streaming services such as Netflix and Amazon Prime, RockBuster's Management is planning to use its existing movie licenses to launch an online video rental service in order to stay competitive.

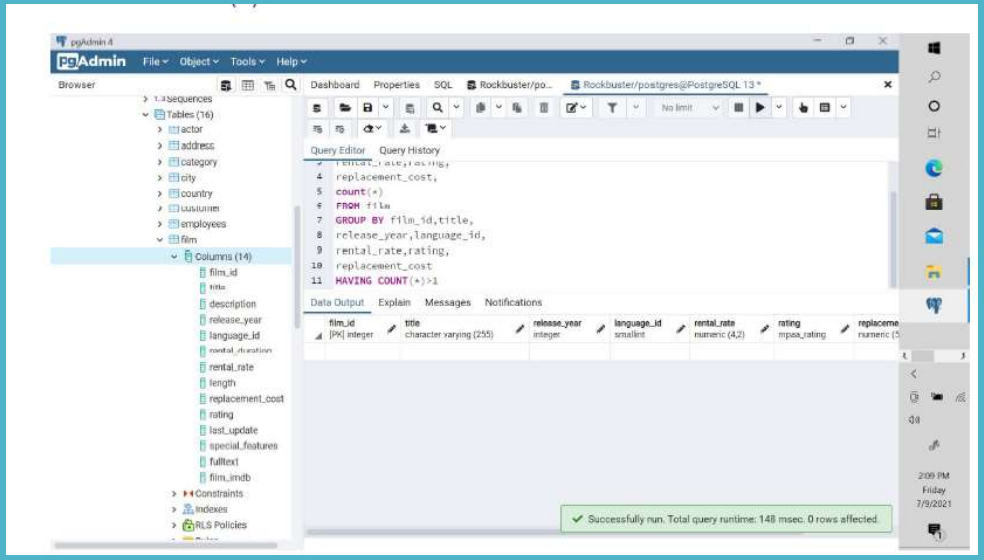
I worked on the launch strategy for the new online video service and this project helped me in learning SQL commands incrementally

Project Goal : To acquire a good understanding of the various data points. This has been achieved by exploring the inventory and revenues of the stores by using the relational database management system pgAdmin4 within the PostgreSQL. The findings have been communicated and visualized in Tableau and PowerPoint.

Data Source : The dataset is around 3 MB and contains several files. No Source for this data was provided.

In the beginning of the project , it was important to know the relationship that exists in the Rockbuster's database and overview of all the tables and to execute this, I extracted The snowflake-type entity relationship diagram with DbVisualizer and created Data dictionary.





Thereafter started preparing the data using CRUD commands - ordering, grouping , filtering followed by data cleaning and data consistency checks (looking for missing values, duplicate records).After that wrote few queries for calculating descriptive statistics of the data.

Finally , by writing queries, **Subqueries combined with joints** I was able to find answers for more advanced questions.

Using **Subqueries**, I was able to write queries that are more dynamic, and data driven, and they give the flexibility to the statements even when the input data is changing frequently.

Later those advanced questions were answered through **CTEs** to learn the advantages of using CTEs over subqueries

```
select
AVG(Total_amount_paid.Total_amount_paid) as average
from
(SELECT A.customer_id,
B.first_name,
B.last_name,
E.country,
D.city,
SUM(A.amount) AS Total_amount_paid
FROM payment A
INNER JOIN customer B on A.customer_id=B.customer_id
INNER JOIN address C on B.address_id=C.address_id
INNER JOIN city D on C.city_id=D.city_id
INNER JOIN country E on D.country_id=E.country_id
WHERE D.CITY IN('Aurora','Tokat','Tarsus','Atlixco','Emeishan','Pontianak','Shimoga','Apar
GROUP BY A.customer_id,first_name,last_name,country,city
ORDER BY Total_amount_paid DESC
LIMIT 5 ) as Total_amount_paid;
```

Data Output | Explain | Messages | Notifications

	average numeric
1	120.3220000000000000

Query Editor | Query History

```
WITH Total_amount_paid (customer_id,first_name,last_name,country,city,amount
AS
(SELECT A.customer_id,
B.first_name,
B.last_name,
E.country,
D.city,
SUM(A.amount) AS amount_paid
FROM payment A
INNER JOIN customer B on A.customer_id=B.customer_id
INNER JOIN address C on B.address_id=C.address_id
INNER JOIN city D on C.city_id=D.city_id
INNER JOIN country E on D.country_id=E.country_id
WHERE D.CITY IN('Aurora','Tokat','Tarsus','Atlixco','Emeishan','Pontianak','
GROUP BY A.customer_id,first_name,last_name,country,city
ORDER BY amount_paid DESC
LIMIT 5)
SELECT AVG(amount_cte) from Total_amount_paid;
```

While working for this project , I understood the advantages and disadvantages of using EXCEL functions and SQL commands to reach the same results in data analysis process and also learned when it is appropriate to use EXCEL or SQL.

My findings for this project were presented through Excel for the technical stakeholders and then used Tableau Public to create Visualizations to compile those findings into the presentation.

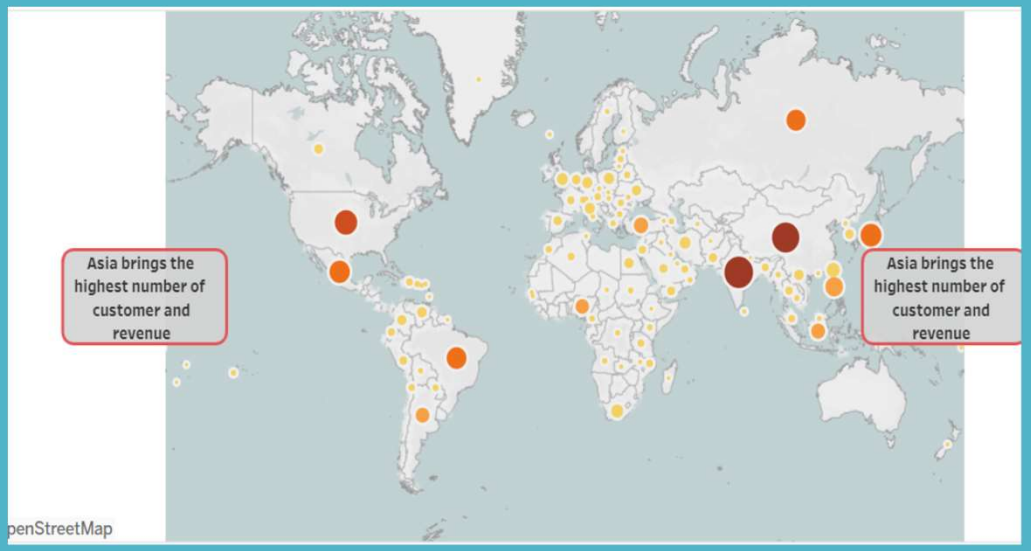
Descriptive Statistics from Customer Table

	Rental Duration	
Minimum	Average	Maximum
3	4.985	7

modal_value_rating
PG-13

	Rental Rate	
Minimum	Average	Maximum
0.99	2.98	4.99

	Length	
Minimum	Average	Maximum
46	115.272	185



ROCKBUSTER STEALTH DATA ANALYSIS

Recommendations

1. Launching Online movie rental platform needs lots of addition of movies in different genres , speciallly the ones which are very popular like thriller and there are hardly any thriller movies in the existing collection to offer to customers.

2. Asia is the major contributor to the existing revenue and customer base, so the new strategy should have the special focus targeting the customer in this region considering the potential of the area. Same applies to the North American and Europe region. Planning is required to penetrate the untapped markets like Australia. Australian market has lot of potential being tech savy and this can considerably increase the revenue and customer base .

3. Customer Loyalty program should be introduced with attractive features to retain the existing customers and for new client addition as well.

Please find the full report here

Project 4
Instacart Grocery Basket Analysis

INSTACART GROCERY BASKET ANALYSIS

Analysis of Sales Pattern and Customer Segmentation

TOOL : JUPYTER

LANGUAGE : PYTHON

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.

Project Goal : To derived insights and suggest strategies for better customer segmentation based on the provided criteria.

Skills : Data wrangling & Subsetting , Data consistency checks, Combining & Exporting Data ,Deriving new variables, Grouping Data and Aggregation, & Data Visualization.

Data Source : open-source data sets from Instacart from year 2017
The consumer data and the prices of the products were both fabricated for learning purposes. Some of the datasets contain over 32M observations.

In order to draw insights, as expected by Instacart stakeholders, transformation procedures like deriving new columns using if-statements, loc and for-loops functions, as well as grouping and aggregating methods were applied.

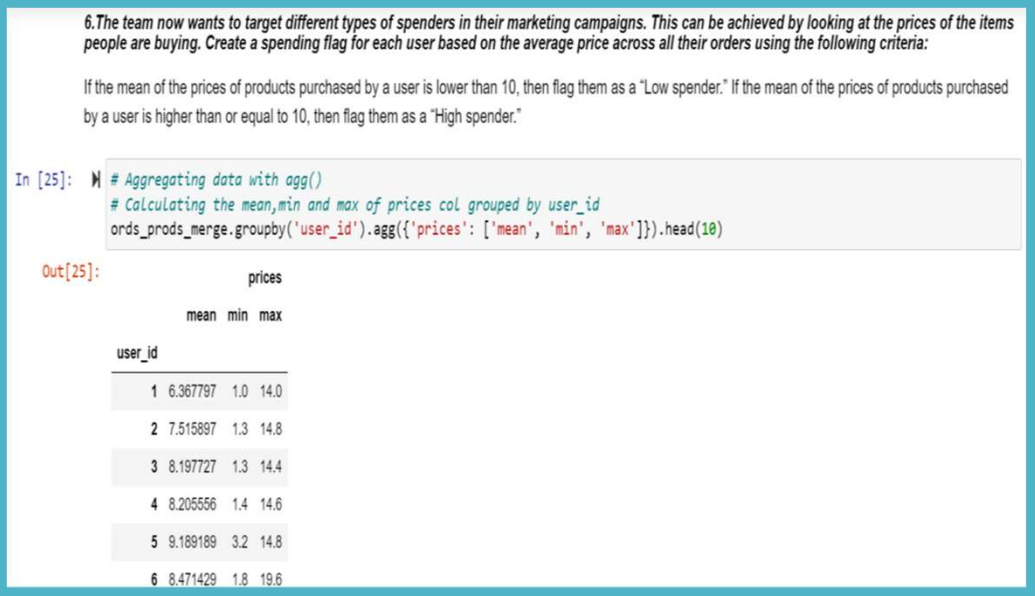
Ques 3 The Instacart officers are interested in comparing customer behavior in different geographic areas. Create a regional segmentation of the data.

In [25]:
```
# Creating filter using else/if for each region

region = []

for state in allcomb_df1['state']:
    if (state == 'Maine') or (state == 'New Hampshire') or (state == 'Vermont') or (state == 'Massachusetts') or (state == 'R
        region.append('Northeast')
    elif (state == 'Idaho') or (state == 'Montana') or (state == 'Wyoming') or (state == 'Nevada') or (state == 'Utah') or (s
        region.append('West')
    elif (state == 'Wisconsin') or (state == 'Michigan') or (state == 'Illinois') or (state == 'Indiana') or (state == 'Ohio'
        region.append('Midwest')
    elif (state == 'Delaware') or (state == 'Maryland') or (state == 'District of Columbia') or (state == 'Virginia') or (sta
        region.append('South')
    else:
        region.append('Stateless')
```

Fig 1 : Regional Segmentation of Customer Data using else/if



6.The team now wants to target different types of spenders in their marketing campaigns. This can be achieved by looking at the prices of the items people are buying. Create a spending flag for each user based on the average price across all their orders using the following criteria:

If the mean of the prices of products purchased by a user is lower than 10, then flag them as a "Low spender." If the mean of the prices of products purchased by a user is higher than or equal to 10, then flag them as a "High spender."

In [25]:
```
# Aggregating data with agg()
# Calculating the mean,min and max of prices col grouped by user_id
ords_prods_merge.groupby('user_id').agg({'prices': ['mean', 'min', 'max']}).head(10)
```

Out[25]:

	prices		
	mean	min	max
user_id			
1	6.367797	1.0	14.0
2	7.515897	1.3	14.8
3	8.197727	1.3	14.4
4	8.205556	1.4	14.6
5	9.189189	3.2	14.8
6	8.471429	1.8	19.6

Fig 2 : Aggregating Price data to target different kind of spenders in their marketing campaigns

Finally, for **Data visualizations** of my analysis, I used bar charts, horizontally and stacked bar charts, pie charts, line charts, histograms, and scatterplots and created them in **Jupyter**.

```
#Creating object "bar" to save bar chart
bar=ords_prods_merge['order_dow'].value_counts().sor
```

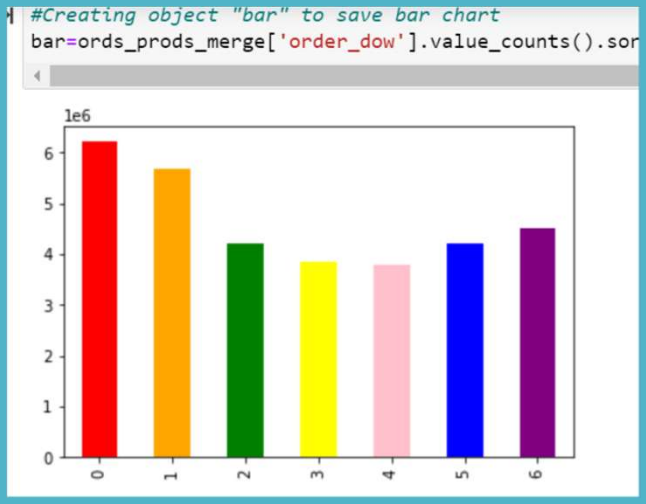


Fig 1 : Created order frequency bar chart for column day of week

```
scatter = sns.scatterplot(data = df, x = 'age', y = 'income')
```

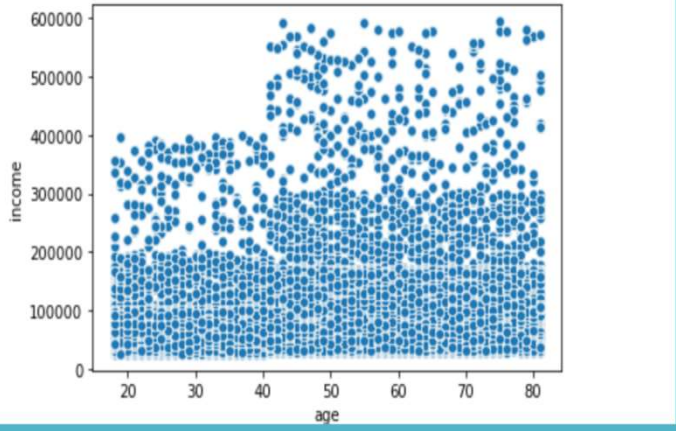


Fig 2 : Created Scatter Plot to show relationship between Income & Age

```
label1 = 'Regular Customer', 'Loyal Customer', 'New Customer'
size1 = loyalty_counts

fig1, loyal_pie = plt.subplots()
loyal_pie.pie(size1, labels=label1, autopct='%1.1f%%')
loyal_pie.axis('equal')
loyal_pie.set_title('Customer Mix by Loyalty')

plt.show()
```

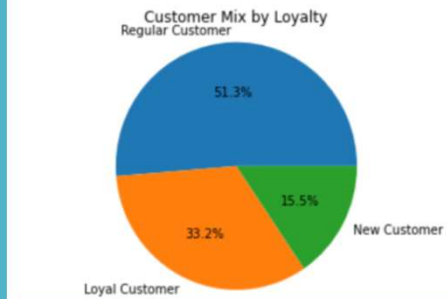


Fig 3 : Created Pie Chart to show the customer mix by Loyalty

Finally, the entire analysis process was documented in form of Excel Reporting that contains population flow, describes all wrangling and deriving operations, shows visualizations of results along with recommendations for the new marketing strategies.

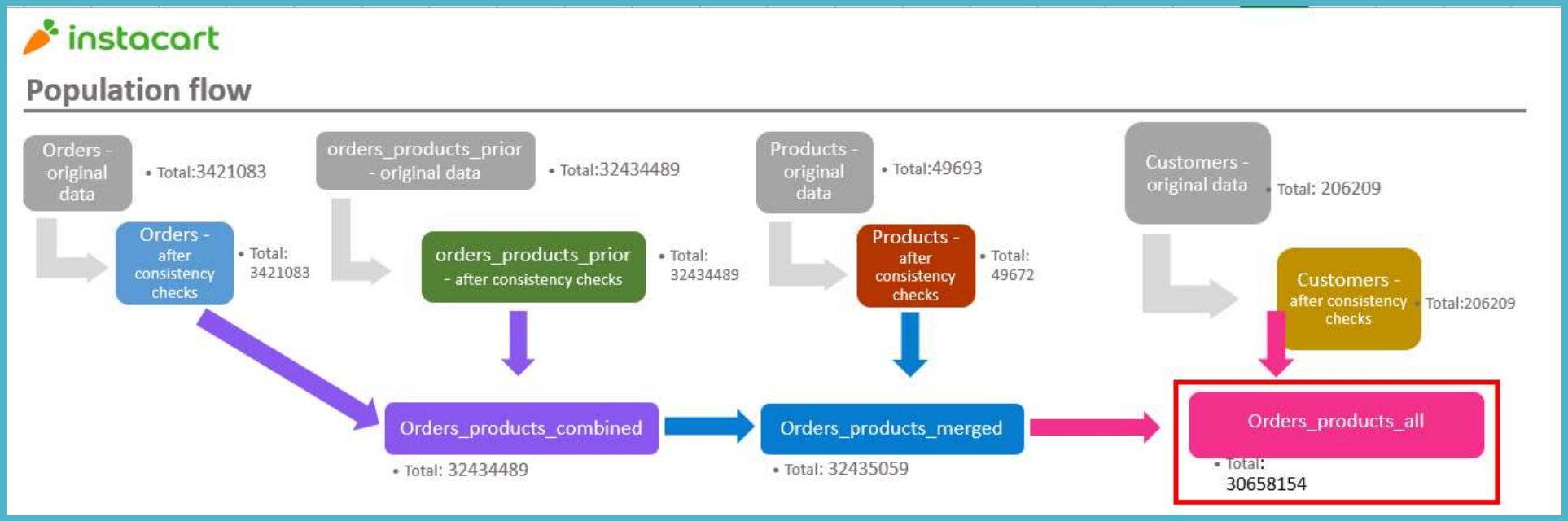


Fig : Population flow for the Instacart Project

Instacart Grocery Basket Analysis

Recommendations

- Saturday and Sunday are the busiest days, and 10 a.m to 3 p.m are the busiest hours. So the slowest days and slowest hours are perfect for targeting advertisements and promotions (upcoming) so that families can plan ahead for weekend deals. Slowest days can be celebrated for special weekday deals.

- Focusing on struggling departments is going to increase the loyalty customer base. Deals and special days can be planned for those struggling departments after discussing with vendors and that should definitely target the specific customer set and give a sales boost.

- 33.2 % customers are loyal customers and 51.3% as our regular customers. To erase this thin line between Regular and Loyal customers , We need to reward our regular customers with some reward points or may be a mail coupon for their next purchase, to convert them to loyal customers and increase our loyalty base

Please find the full excel reporting , project scripts with data link here

Project 4
Covid -19 Pandemic Analysis

COVID 19 PANDEMIC ANALYSIS

Advanced analytics & Dashboard design

TOOL : JUPYTER & TABLEAU PUBLIC

LANGUAGE : PYTHON

Covid 19 is an ongoing global pandemic and has become one of the deadliest pandemic in history, here we will uncover how the different countries across the globe are dealing with this.

Project Goal :To get the insights about Covid 19 Pandemic throughout the World in terms of mortality and vaccination drives of different countries using different analytical approaches

Skills : Data cleaning and Wrangling, Exploratory analysis, Machine Learning Techniques – regression and clustering ,Geospatial analysis, Analysing time series data.

Data Source. : open-sourced from ourworldindata.org. This data describes the country wise details for the important variables like cases, deaths, vaccination and many more from beginning of the Pandemic till 10th Dec 2021. The original data set has 138727 Rows and 28 Columns

For drawing insights, I first started with exploring the relationships between different variables in data, then I worked on the strongest relationships. While analyzing Number of People vaccinated & Mortality rate, I formed the hypothesis – " Countries with high vaccination rates have low mortality "

Later I tested that hypothesis using regression analysis in python and found out that the relationship between the variables was not entirely linear and there were many outliers

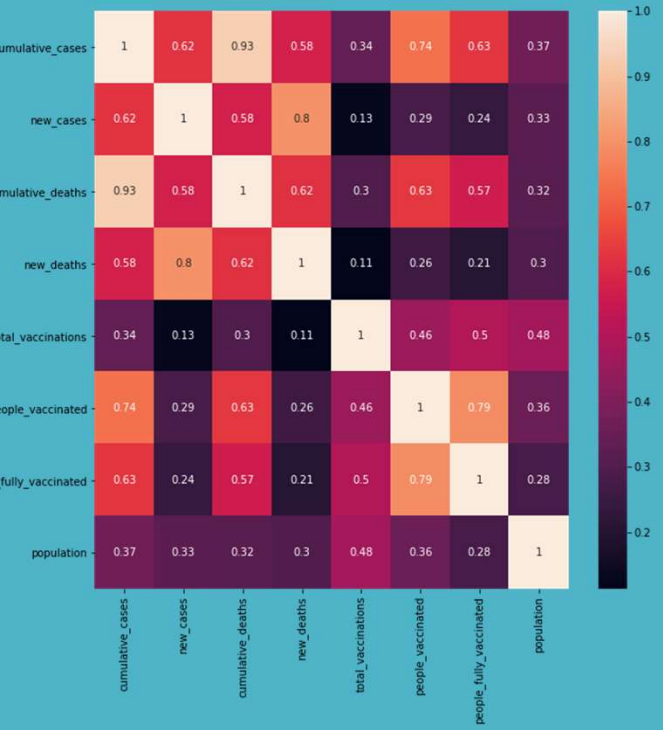


Fig 1 : Correlation Matrix

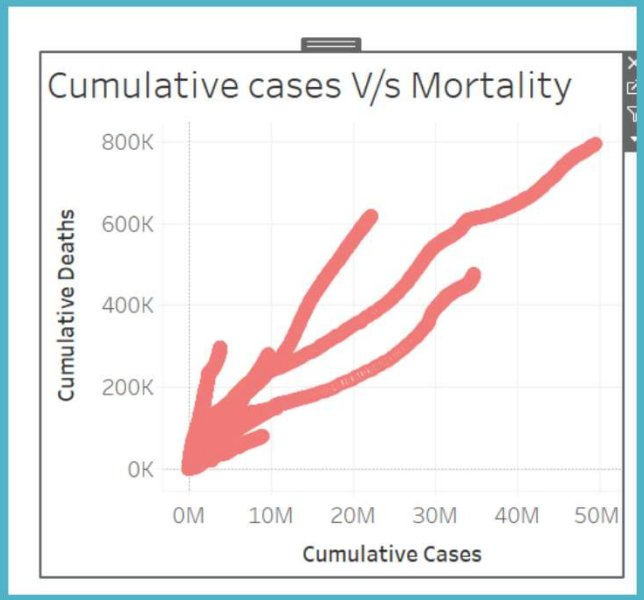


Fig 2 : Exploring relationship between cumulative cases & Mortality using scatter plot

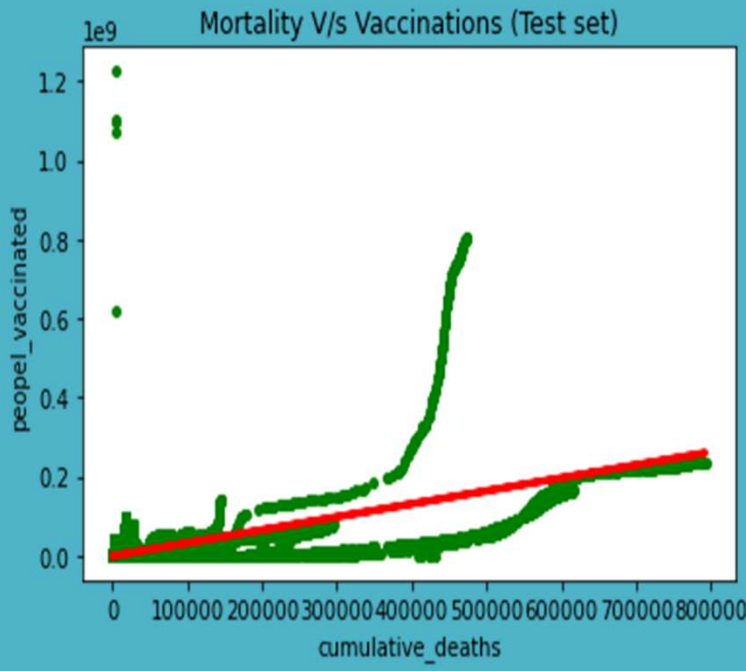


Fig 3 : Linear Regression Analysis

As Linear regression wasn't enough to fully explain the data, so I used another non linear approach to analyze the data – Cluster Analysis where I was able to group the data into three different clusters and by further analyzing the differences in clusters , I was able to identify the where are these clusters located and how differently all the countries are coping with the Pandemic

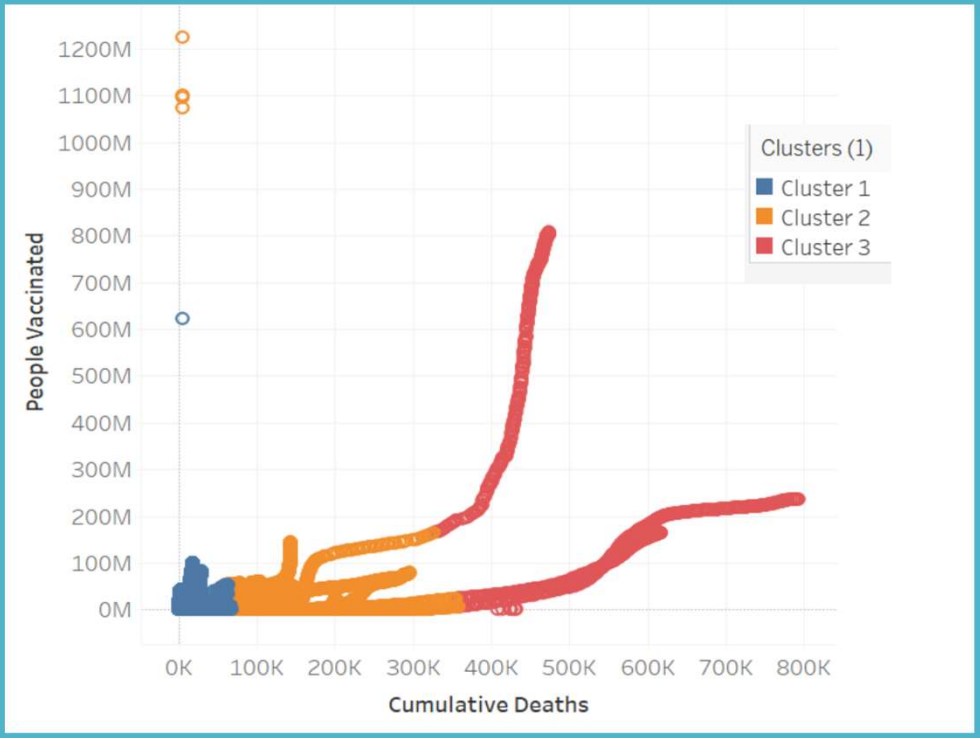
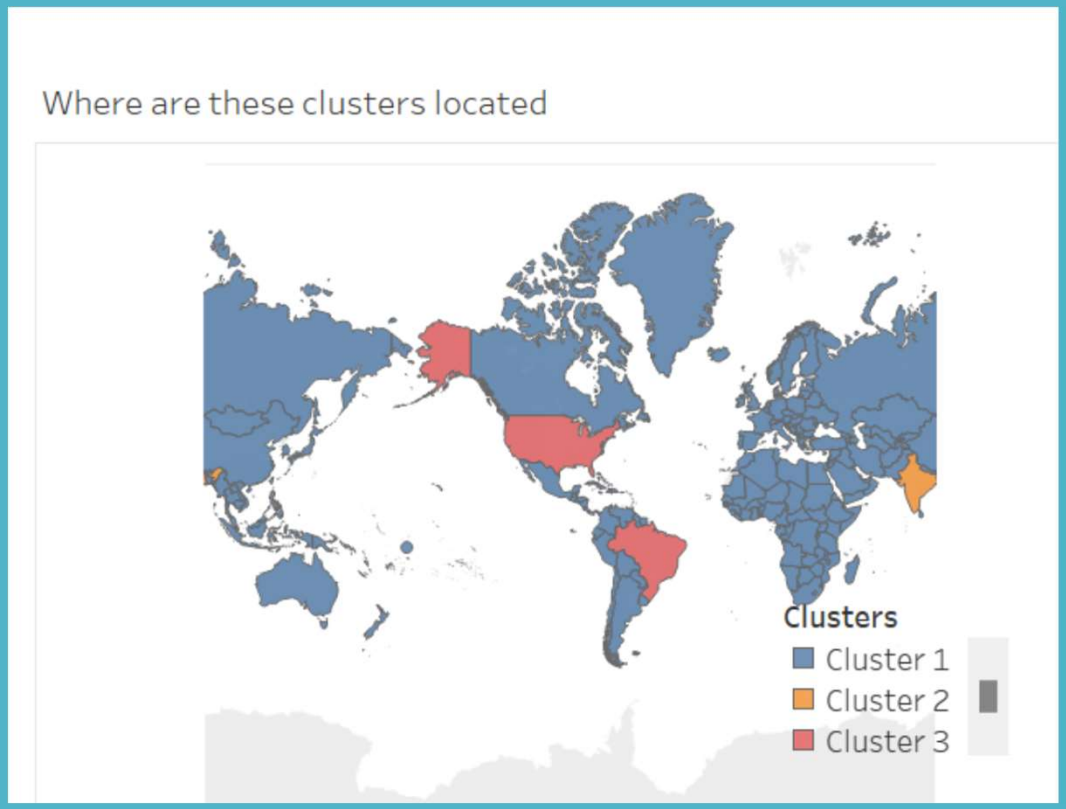


Fig 1 : Cluster Analysis

Fig 2 : Differences between Clusters



Click to see Tableau Story

Click to see project on Github

Key findings :

with most cases have most fatality, and the data points we have so far for these variables shows that vaccination is not making much difference in controlling mortality rate.

I was able to identify three clusters – countries with high fatality and majority population Vaccinated, countries with high fatality but under vaccinated, countries with moderate fatality and under vaccinated

Limitations :

There weren't enough datapoints to yield a highly significant result.
Data Collection bias

Next Steps :

-Gather more datapoints on these variables and run the cluster analysis again.
-Analyze the impact of additional variables

Thank You

FIND ME ONLINE….