SCHAY ESPARZA

PORTFOLIO

DATA ANALYST

PROJECT OVERVIEW

Circulatory System Diseases Mortality

Analysis of deaths due to circulatory system diseases.

GameCo.

Analyzing global video game sales to create a strategy for allocating the upcoming year's marketing budget based on trends over time.

Medical Staffing Agency

Evaluate influenza data to help staffing company plan for the upcoming influenza year.

Rockbuster Stealth LLC

Analyze company data to help launch the new online video rental service.

Instacart Basket

Analysis of customer profiles and purchasing behaviors to develop targeted marketing campaigns.

DELIVERABLE

- *Presentation*

TOOLS

- Excel
- PowerPoint

SKILLS/ PROCEDURES

- Grouping data
- Summarizing data
- Descriptive Analysis
- Visualizing and presenting results in Excel

GAMECO.

A new video game company, GameCo wants to use data to inform the development of new games. Descriptive analysis of video game data is used to better understand how the company's new game might fare in the market.

Italicized text contains link

GAMECO.
PROCEDURES

01 PREPARATION

- Performed data cleaning procedures to ensure accuracy , consistency and clarity of data.
- Creation of new variables derived from given data to offer new insights.

02

Current Understanding

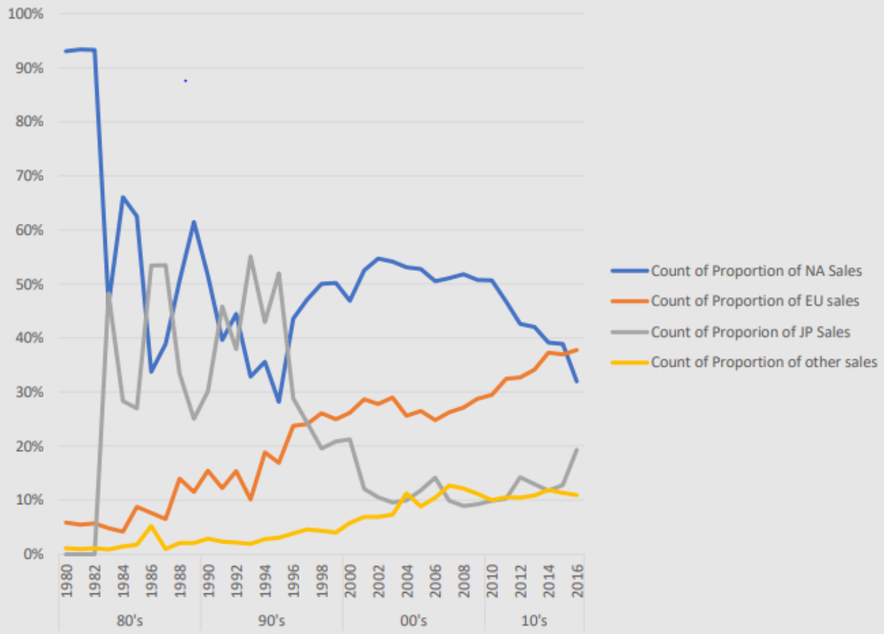
Video game sales have stayed the same over time across geographic region.

ANALYSIS

- Hypothesis Testing (current understanding)
- Performed descriptive analysis of all key variables, including exploring the shape and spread of data via histograms and scatterplots.
- Used pivot tables to group, filter, and sort data to find insights.

GAMECO.
RESULTS

Proportion of Sales per Region over time.



North America (NA)

Fluctuating sales over the years.
Best performing market up until 2016 when proportion of sales was 32% compared to Europe's 38%.

Europe (EU)

Fluctuation per year as well but not as drastic as its North American counterpart and it is showing an upward trend despite the yearly variations.

Japan (JP)

In the 80's to mid 90's comparable to North American Market, with some years even outperforming it.
Faced a steep decline in the late 90's with Europe fully surpassing it and doing worse than Other in some years.
Experienced the most dramatic loss in sales.

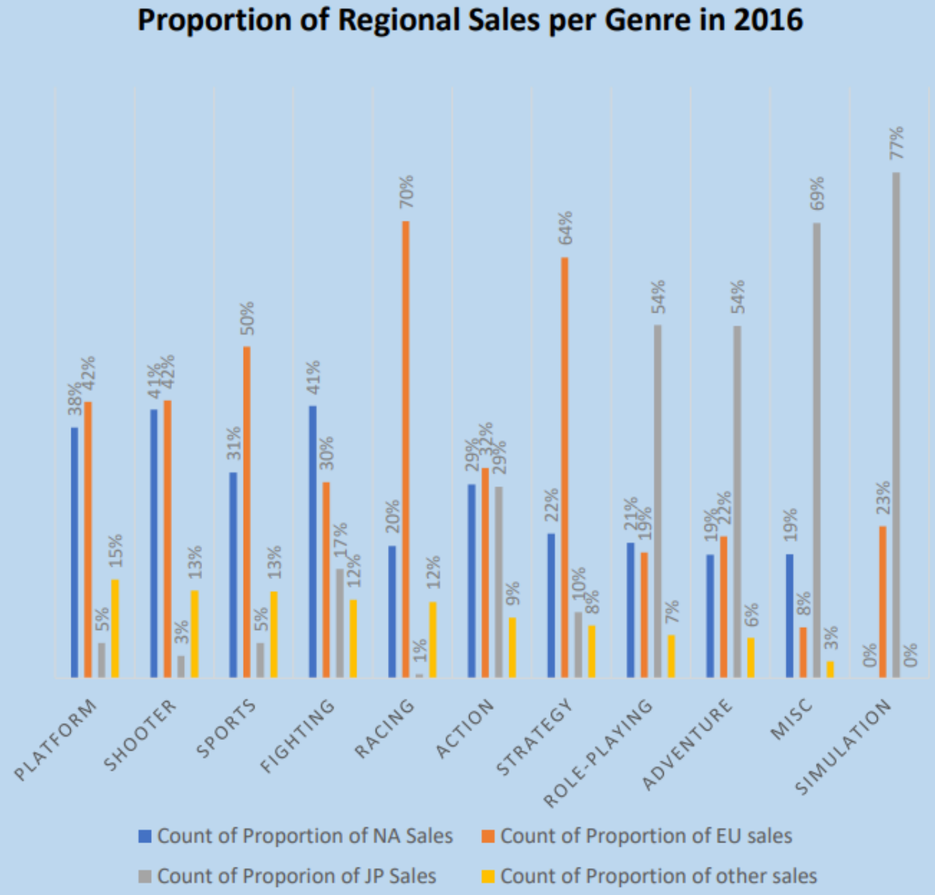
Others

The region that showed the most consistency is Others. It had some peaks and falls but nothing too significant.
Instead, it has a slow rise over the years and plateauing in the 2010's.

Analysis of Regional Proportional Sales Over Time

Breakdown of regional proportional sales over time. Data is categorized in regions and count of regional sales are analyzed and visualized to study the sales behavior.

GAMECO.
RESULTS



Top 3 Performing Genres per Region

North America:
Fighting, Shooter and Platform

Europe:
Racing, Strategy and Sports

Japan:
Simulation, Misc and Role Play

Other:
Platform, Shooter and Sports

Worst Performing Genre per Region

North America:
Simulation

Europe:
Misc

Japan:
Racing

Other:
Simulation

Analysis of Regional Proportional Sales Per Genre

Analysis of top and worst performing genres per region. This data was used to make marketing and budgeting recommendations .

MEDICAL STAFFING AGENCY

The medical staffing agency is preparing for the next influenza season, The agency provides temporary workers to clinics and hospital on a per diem basis. A staffing plan is utilized to be used to address staffing needs across the country.

DELIVERABLE

- *Tableau Story*
- *Presentation*
- *Interim Report*

TOOLS

- Tableau
- Excel

SKILLS/ PROCEDURES

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

Italicized text contains link

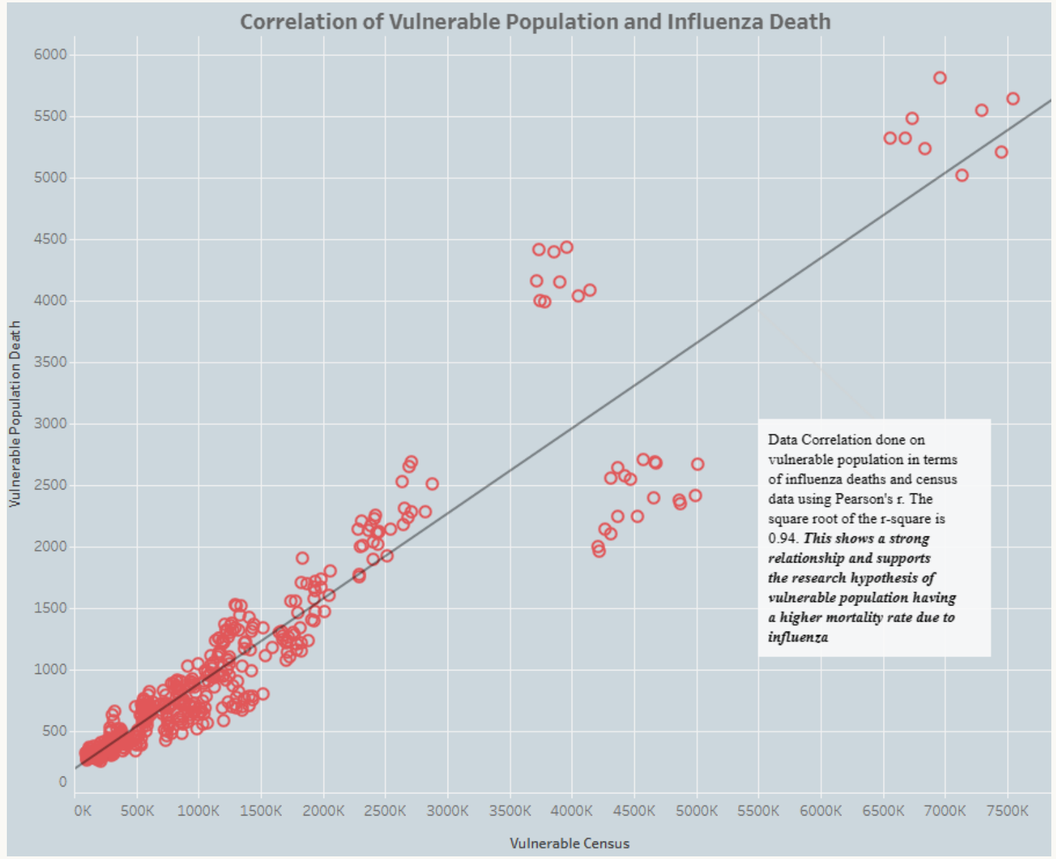
PROCEDURES

01 PREPARATION

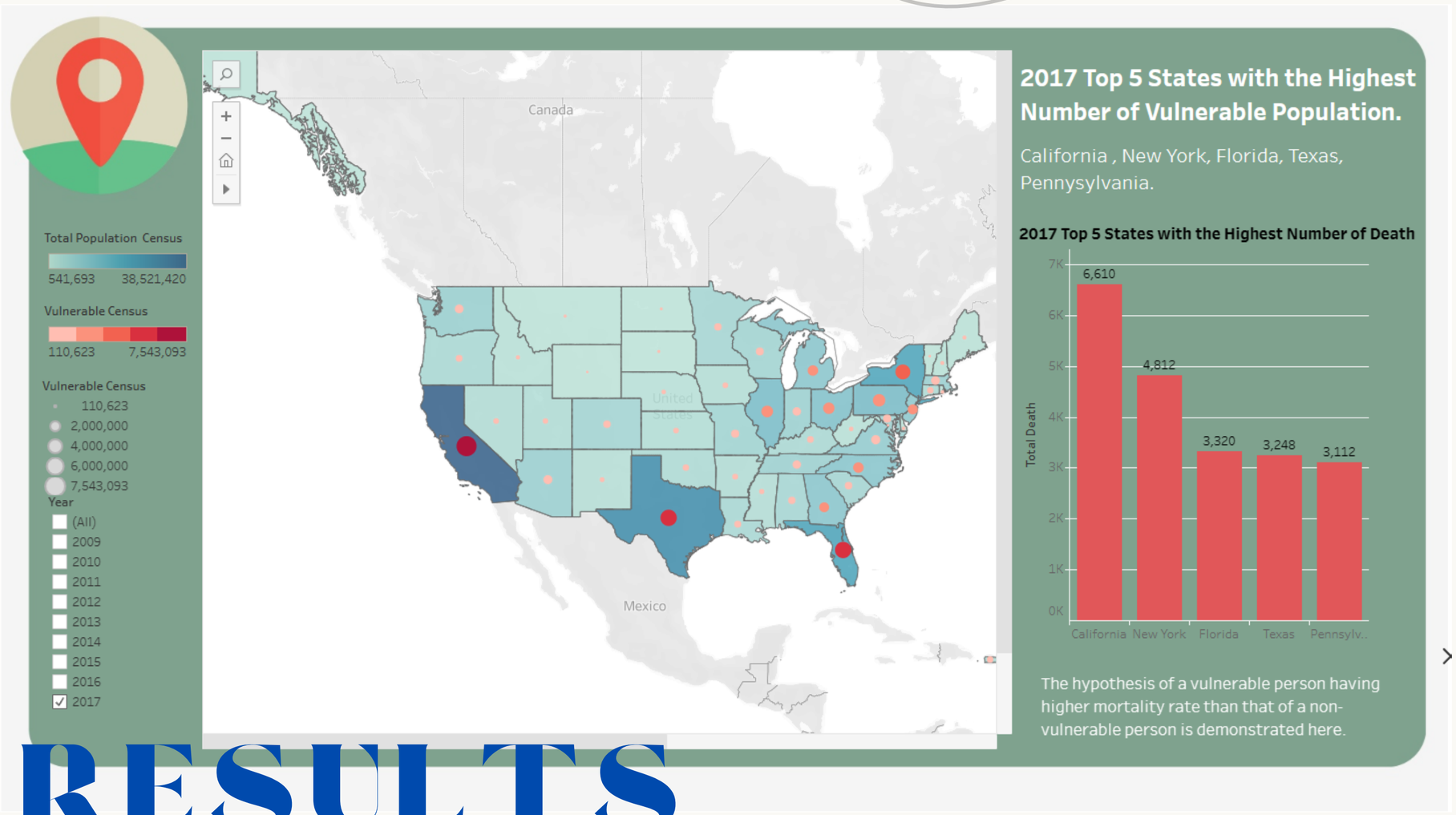
- Translating requirements into a project plan
- Performed data cleaning procedures to ensure accuracy , consistency and clarity of data.
- Creation of new variables derived from given data to offer new insights.
- Data integration from multiple sources

02 ANALYSIS

- Performed statistical hypothesis testing
- Performed visual analysis using Tableau
- Time analysis forecasting



Performing statistical hypothesis testing using Pearson's R. Data is then visualized using a scatterplot.

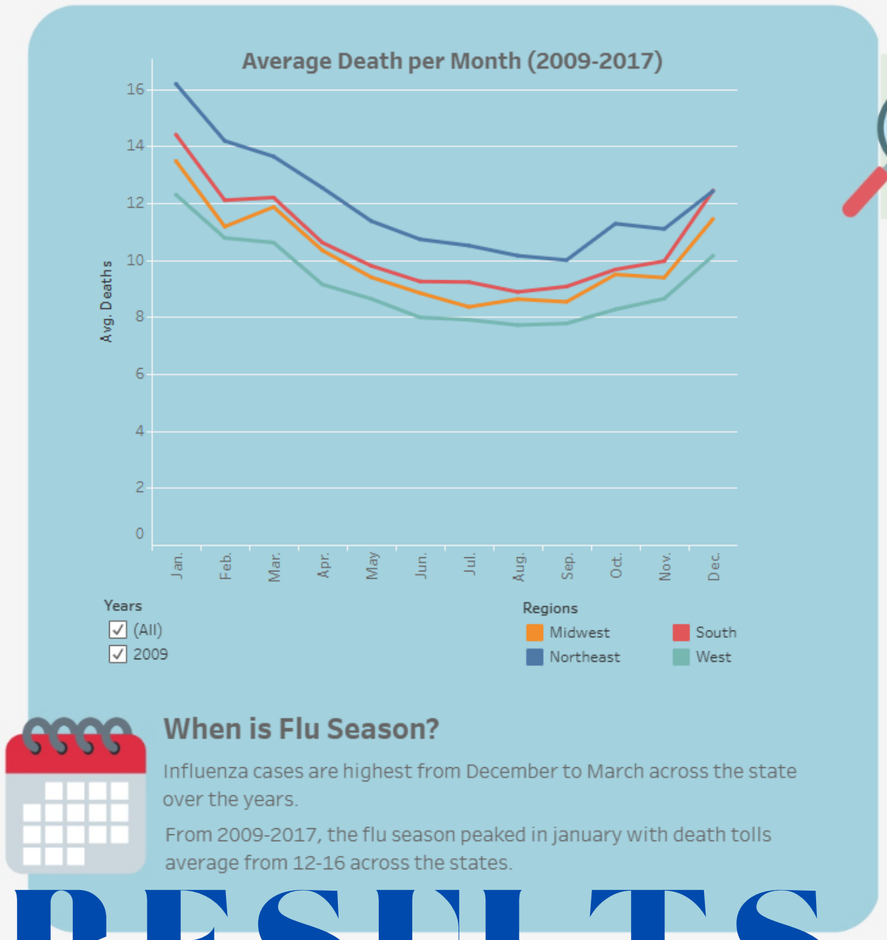


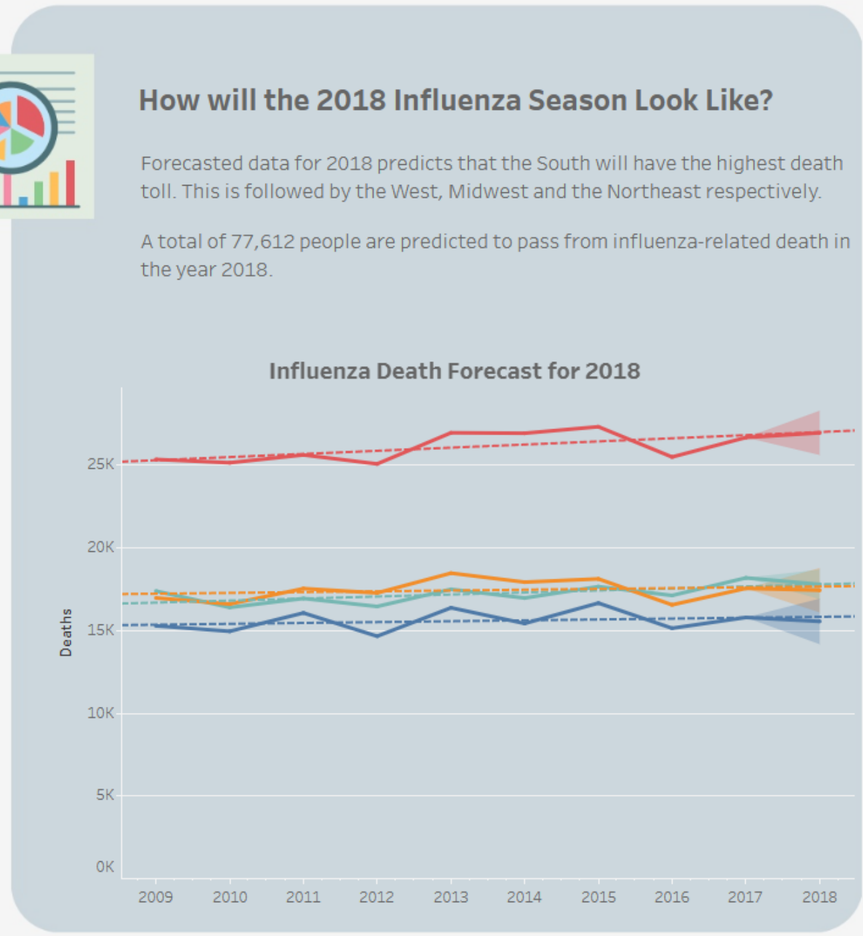
Vulnerable Population Distribution

Map analysis of population distribution broken down in to vulnerable and non-vulnerable population. This is accompanied by bar chart showing the states with the highest death toll.

RESULTS

STAFFING AGENCY





RESULTS
STAFFING AGENCY

Influenza Data in Time

Line chart showing average regional deaths per month. This also shows when flu season occurs.

Influenza season deaths forecasted for 2018 using historical data from 2009-2017.

Italicized text contains link

DELIVERABLE

- *GitHub Repository*
- *Presentation*

TOOLS

- Tableau
- SQL - PostGres

SKILLS/ PROCEDURES

- Database querying
- Filtering
- Cleaning and summarizing
- Joining tables
- Subqueries
- Common table expression

ROCKBUSTER STEALTH LLC

A fictional movie rental company wishes to stay competitive in the current market. The company is looking to use its existing movie licenses to launch an online video rental service in order to stay competitive. The Business Intelligence team is looking into launching a new and effective strategy for the launch of the online video rental service.

PROCEDURES

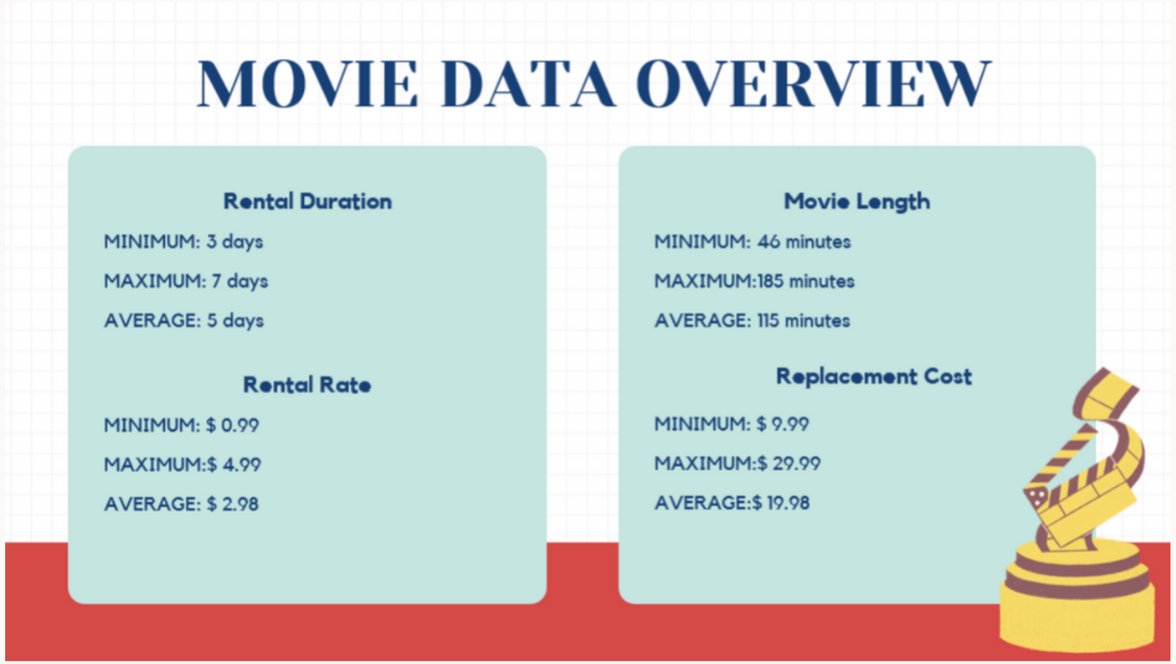
01 PREPARATION

- Imported data into a relational database system
- Generated Entity-Relationship Diagram (ERD) to visualize relationships between entities
- Development of data dictionary which shows connections between the relational database tables.

02 ANALYSIS

- Use of SQL to query structured database.
- Utilization of filtering, grouping, ordering, aggregating functions on SQL to query data.
- Use of subqueries, joins and Common Table Expressions (CTEs) on SQL.

.



Statistical data derived from SQL functions. Visualized in a summary format to show an overview of the data.

RESULTS

Average Rental Duration per Genre

The average rental duration for the different genres as represented by a bubble chart. Visualized data is derived from SQL query of data using joins and group by function.


```
1  SELECT C.name,AVG(A.rental_duration)
2  FROM film A
3  INNER JOIN film_category B on A.film_id=B.film_id
4  INNER JOIN category C on B.category_id=C.category_id
5  GROUP BY name
```

What was the average rental duration for all videos?

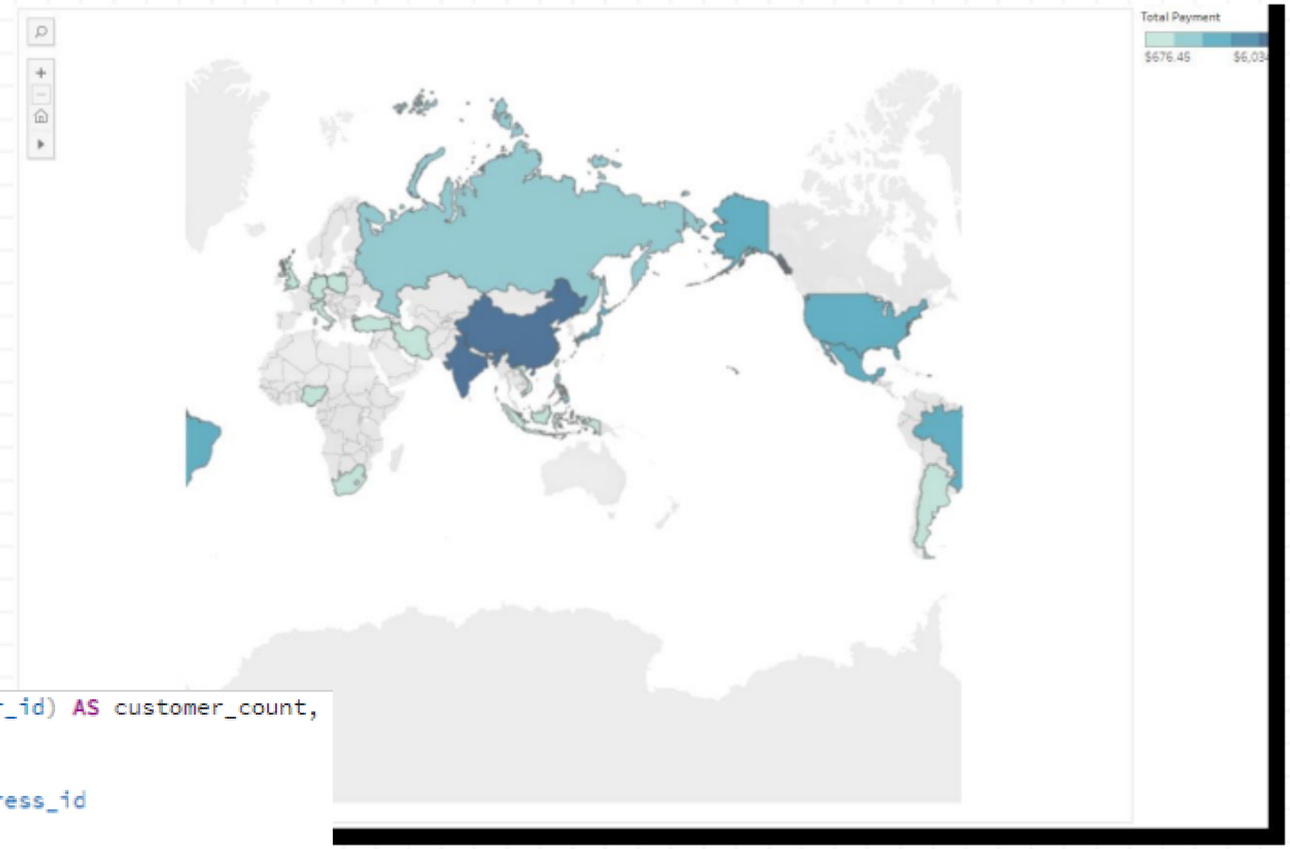
- The longest average rental duration is 6 days for thriller.
- Most movies are rented out for an average of 5 days.

RESULTS

Sales figures across different geographical regions

Sales figure of different regions as represented in a heat map. Data is derived from SQL queries using joins, group by, order by and limit functions.

Do sales figures vary between geographic regions?



TOP 20 COUNTRIES AND THEIR REVENUE

TOP 5 COUNTRIES WITH HIGHEST REVENUES.

India	$6034.78
China	$5251.03
United States	$3685.31
Japan	$3122.51
Mexico	$2984.82

Sales figure does vary between geographic region.

Currently dominated by the Asian Market with 3 out of the top 5 slots being Asian countries.

```
SELECT D.country, COUNT(DISTINCT A.customer_id) AS customer_count,
SUM(E.amount) AS total_payment
FROM customer A
INNER JOIN address B on A.address_id=B.address_id
INNER JOIN city C on B.city_id=C.city_id
INNER JOIN country D on C.country_id=D.country_id
INNER JOIN payment E on A.customer_id=E.customer_id
GROUP BY D.country
ORDER BY Total_Payment DESC
LIMIT 20
```

INSTACART

An online grocery store wants to uncover more information about their sales patterns. Use of exploratory analysis of their data to derive insights and suggest strategies. The company is interested in learning about the customer profiles and behaviors to target better marketing campaigns.

DELIVERABLE

- *GitHub Repository*
- *Report*

TOOLS

- Python - Pandas, NumPy, Seaborn, Matplotlib

SKILLS/ PROCEDURES

- Dara Wrangling and merging
- Deriving variables
- Grouping data
- Aggregating data
- Reporting in Excel
- Population Flows

Italicized text contains link

PROCEDURES

01 PREPARATION

- Cleaned data for accuracy, consistency and clarity.
- Data wrangling (data type conversion, transposing, etc.)
- Merging and sub setting data frames.
- Deriving of new variables.

instacart

Column derivations and aggregations

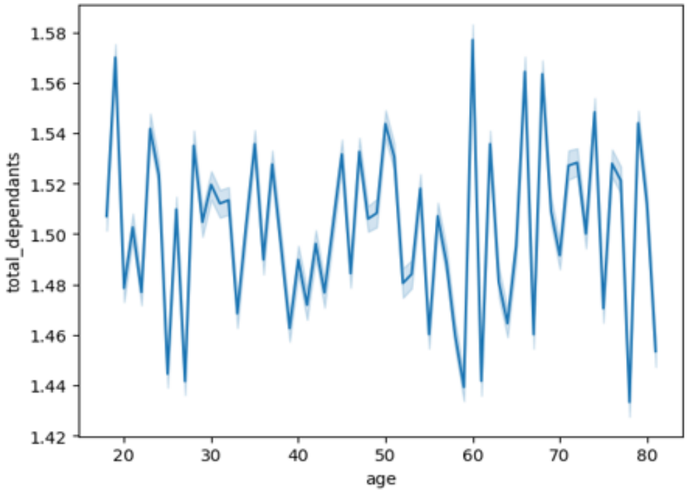
Dataset	New column	Column/s it was derived from	Conditions
orders_products_merged.pkl	price_range	prices	'price <= 5 : 'Low range product', price > 5 and <= 15 : 'Mid range product', price > 15: 'High range product'
orders_products_merged.pkl	busiest_day	orders_day_of_week	orders_day_of_the_week 0 = 'Busiest day', 4 = 'Least busy', else: 'Regularly busy
orders_products_merged.pkl	busiest_days	orders_day_of_week	orders_day_of_the_week 0 or 1 = 'Busiest days', 4 or 3 = 'Slowest days ', else: 'Regularly busy'
orders_products_merged.pkl	busiest_period_of_da	order_hour_of_day	if hour in [10, 11, 14, 15, 13, 12, 16, 9] then 'Most orders', elif hour in [23, 6, 0, 1, 5, 2, 4, 3] then 'Fewest Orders' , everything else is 'Avergae orders'
orders_products_merged_upd	max_order	order_number	displays total number of orders each customer has placed
orders_products_merged_upd	loyalty_flag	max_orders	max_orders > 40 = 'Loyal customer', 40 >= max_order > 10 : 'Regular customer', else: 'New customer'
orders_products_merged_upd	spending_flag	avg_ord_prices	mean_spending < 10 : 'Low spender' igh-spender, else Low-spender
orders_products_merged_upd	customer_frequency	*median frequency of days since p*	Dataframe grouped by 'user_id' and transformed by median of 'days_since_prior_order'
orders_products_merged_upd	avg_price	prices	the mean price of items purchased by each user
customer_merge.pkl	region	state	the states grouped in regions
customer_merge.pkl	low_order_flag	max_orders	order frquennoy based on max order, if below 5, low order and if above 5, high order customer.
customer_merge.pkl	age_group	age	grouped in different age groups, 18-40 years old, 41-65 years old and 65+ years old
customer_merge.pkl	income_group	income	grouped in 65k and below , 65k-125k and 125k and above based on their incomes
customer_merge.pkl	household_status	total_dependants	depedning on number of dependents, grouped in households

02 ANALYSIS

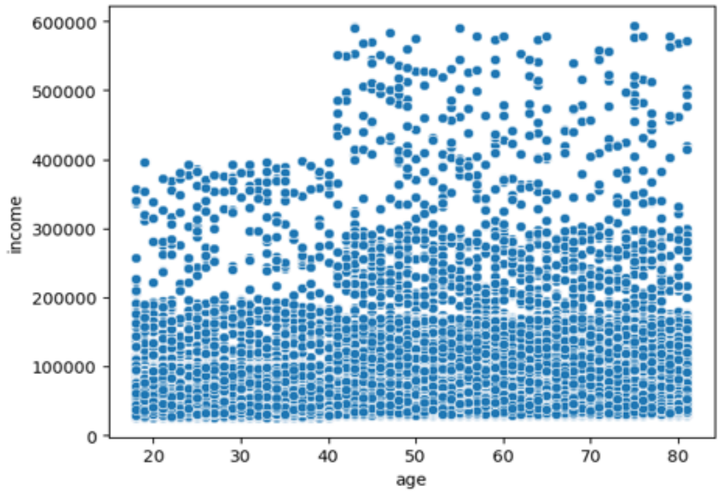
- Filter, sort, group and aggregate data ti answer key questions.
- Performed exploratory analysis on each profile/ flag to uncover additional insights.
- Compilation of findings in a final report.

INSTACART

4. Exploratory analysis of customer demographics to inform the targeted marketing campaigns.

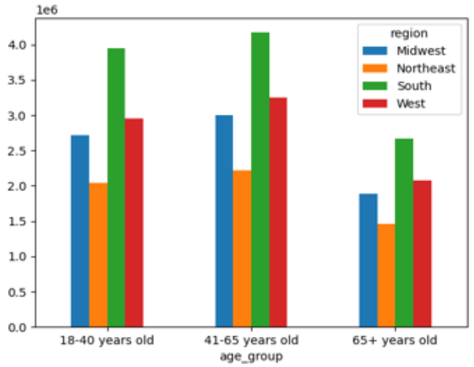


Not much difference between the different age groups and total number of dependants

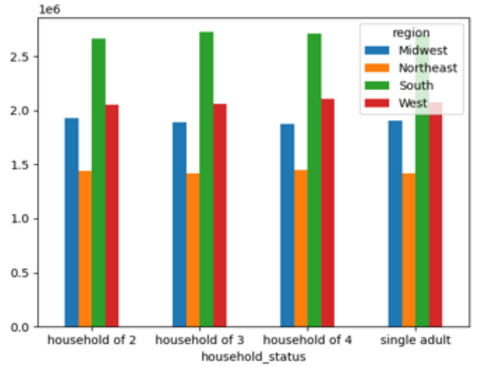


Although most people earn roughly around 200.000 and below, the likelihood of earning more than that increases in age. this makes sense as people generally start to earn more as they gain more experience as they age.

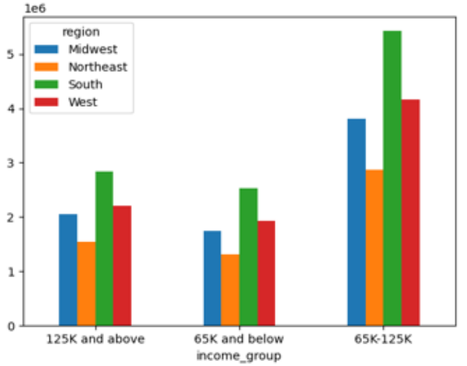
6. Regional Customer Profiles



The biggest customer age group is the 41-65 years old. The distribution of regions per age group is consistently the same with the South being the highest followed by the West, Midwest and then the Northeast.



As seen previously, the household status were split into even quarters. This chart shows that even split and it looks like regionally, the split is also consistent.



Most of the customers fall between the 65k-125k income group followed by those in the 125k and above group and the the 65k and below group.
Regional distribution of income is consistent with the south being the most, followed by west,midwest and then northeast.

RESULTS

Italicized text contains link

DELIVERABLE

- *Github Repository*
- *Tableau Dashboard*
- *Tableau Storyboard*

TOOLS

- Tableau
- Python

SKILLS/ PROCEDURES

- Sourcing Data
- Exploratory Analysis
- Geospatial Analysis
- Time Series Analysis
- Linear Regression Analysis
- Clustering Analysis
- Advance Dashboard Design

CIRCULATORY SYSTEM DISEASES – MORTALITY

Heart disease is the leading cause of death in the United States of America. Analysis done to show how death toll throughout the years have been and who are at more risk of death due to this disease

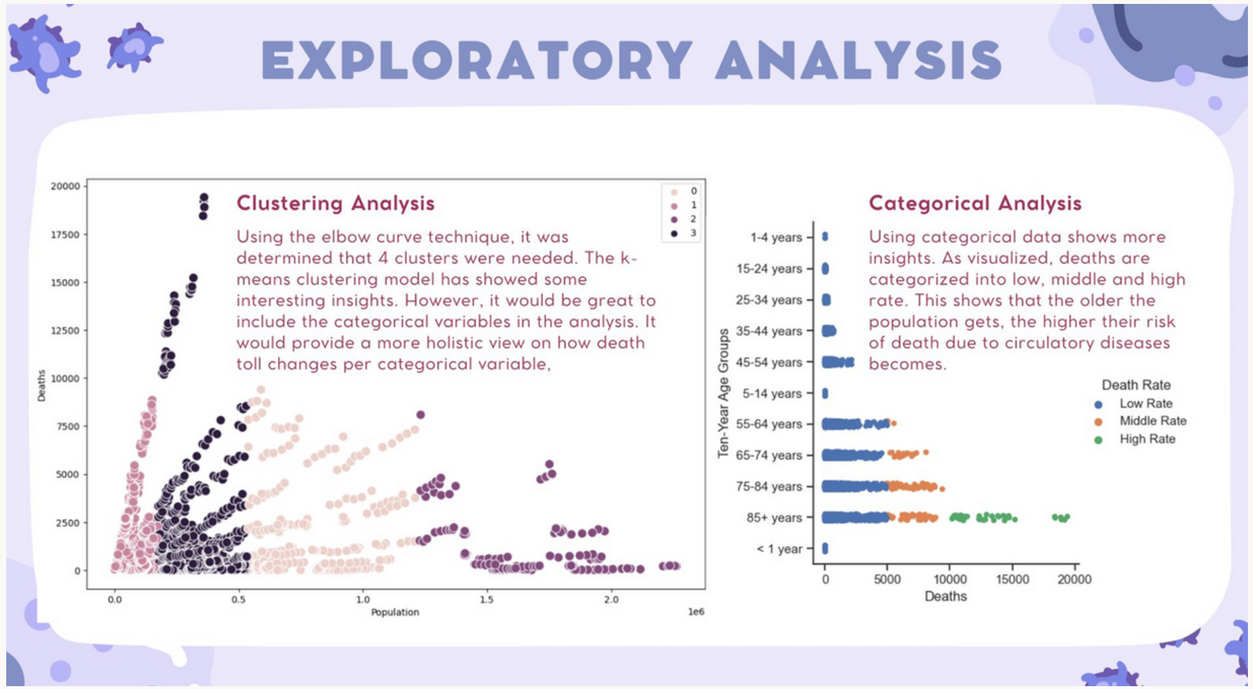
PROCEDURES

01 PREPARATION

- Sourcing data that meets data requirements for the project.
- Cleaned data for accuracy, consistency and clarity.
- Data wrangling

02 ANALYSIS

- Performed exploratory analysis for relationships (Correlation heat map and pair plots)
- Performed Geospatial Analysis
- Performed Time Series Analysis
- Performed Linear Regression and K-Clustering

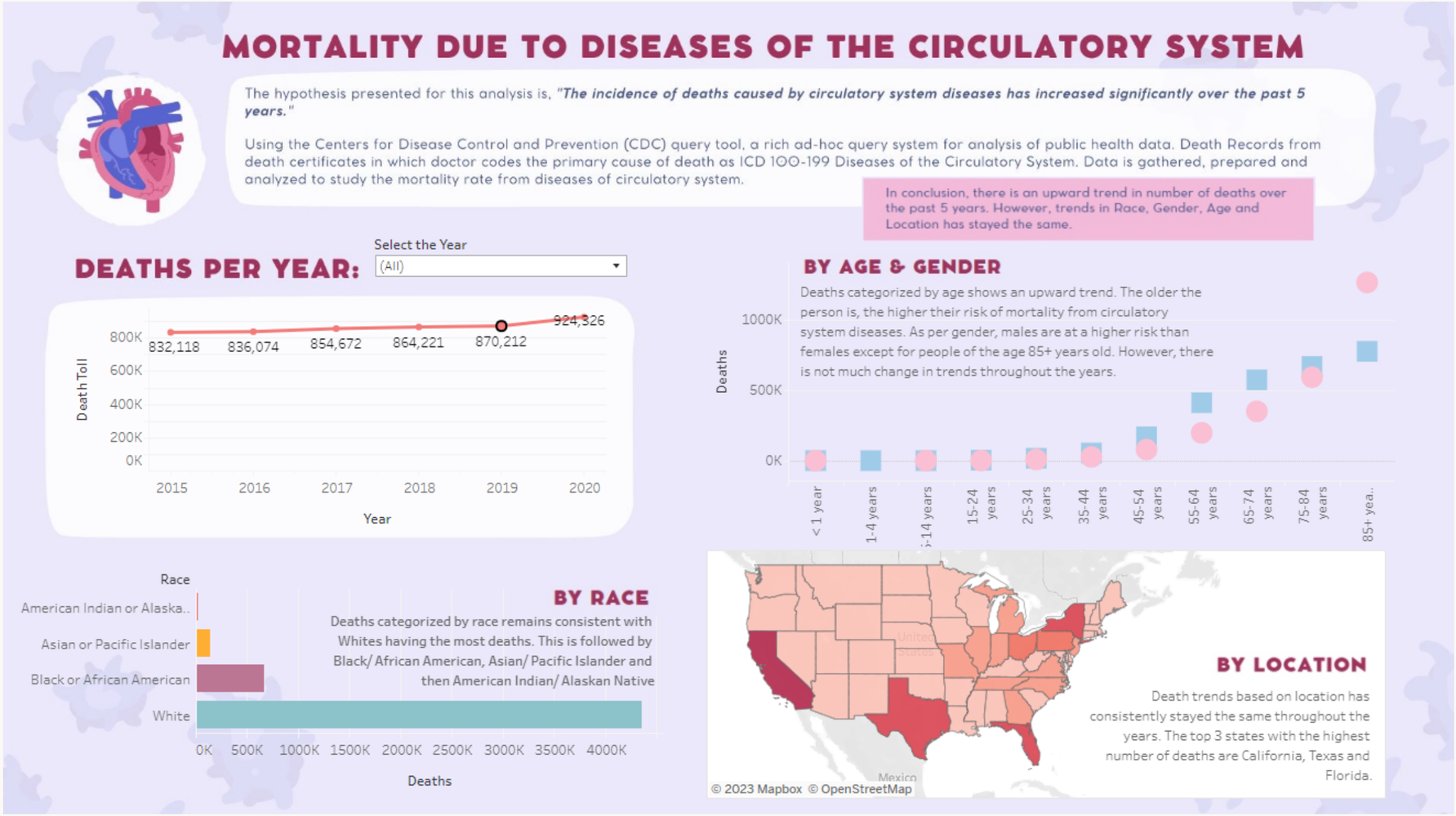


Exploratory analysis using clustering and categorical analysis of the data.

RESULTS

Interactive Dashboard

A Tableau dashboard that visualized death in different perspectives filtered throughout the year. Depending on the year selected, data for that time will be shown.

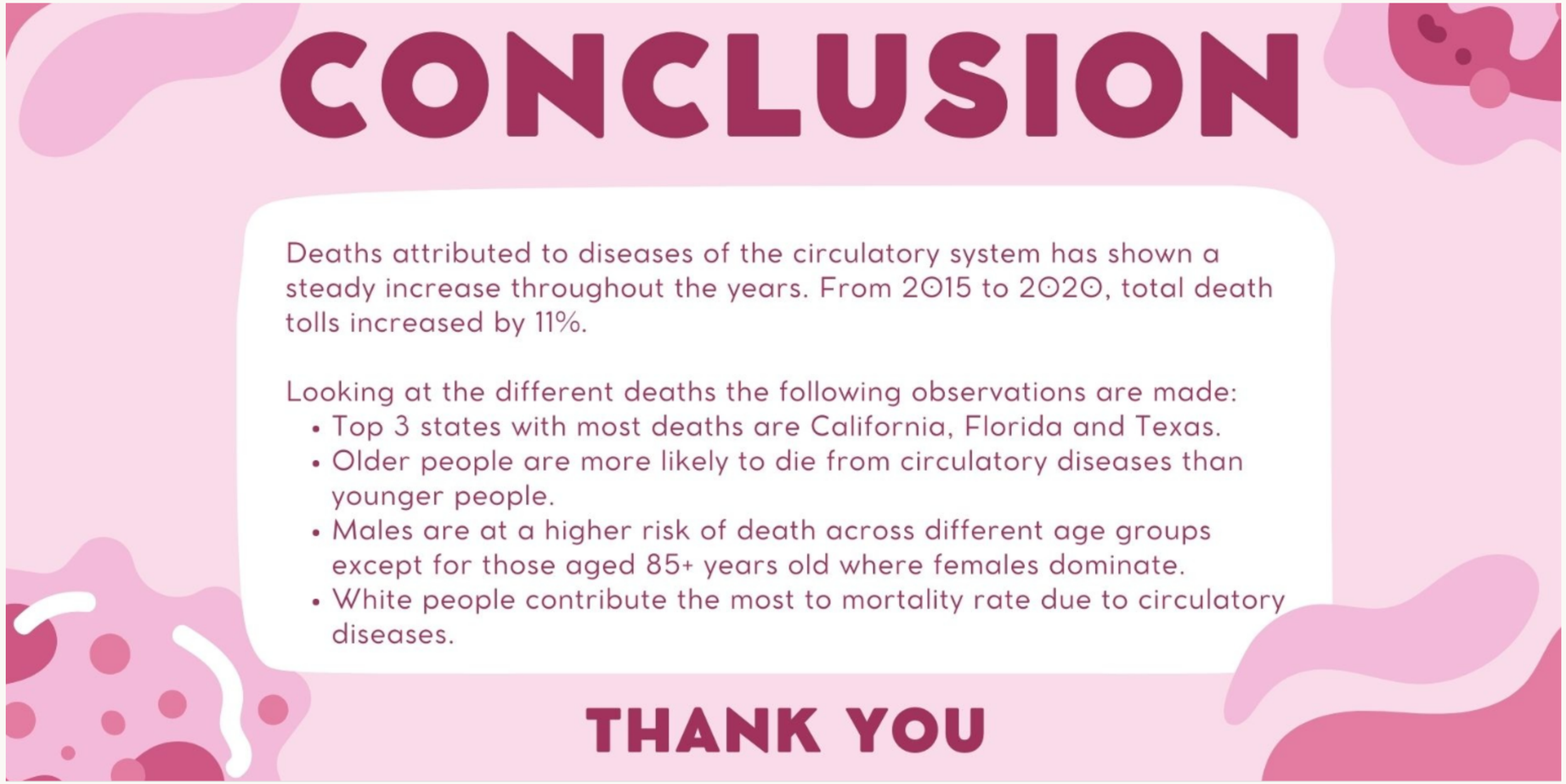


CIRCULATORY SYSTEM DISEASES – MORTALITY

RESULTS

Tableau Storyboard

A Tableau storyboard showing the conclusion and findings of the analysis.



CIRCULATORY SYSTEM DISEASES – MORTALITY

thank you

COMING SOON

MORE PROJECTS COMING SOON

CONNECT WITH ME

RESUME

LINKEDIN

GITHUB