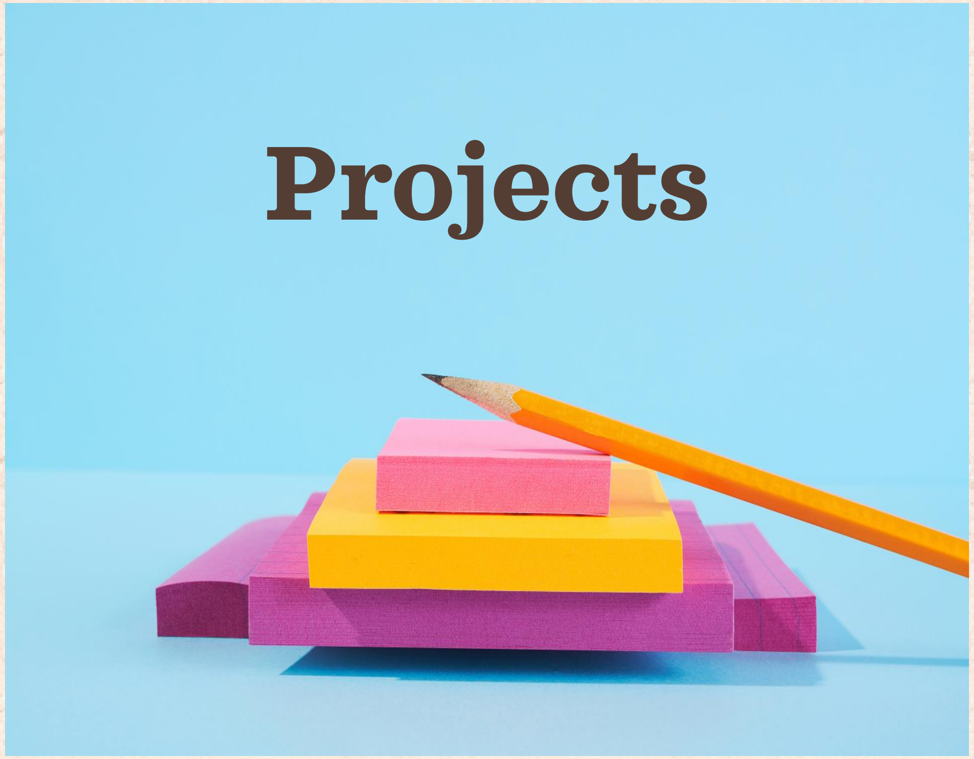
Data Analytics Portfolio

Shruthi Abraham



Projects

GAMECO
Analyzing Global Video Game Sales

INFLUENZA SEASON
Preparation

ROCKBUSTER STEALTH
Business Analysis for Online Video Rental Company

INSTACART BASKET
Marketing Strategy for Online Grocery Store

PIG E BANK
Ethical and Predictive Analysis for Global Bank

US HOUSING PRICES
Influencing Factors Analysis

1. GAMECO

Key Questions:

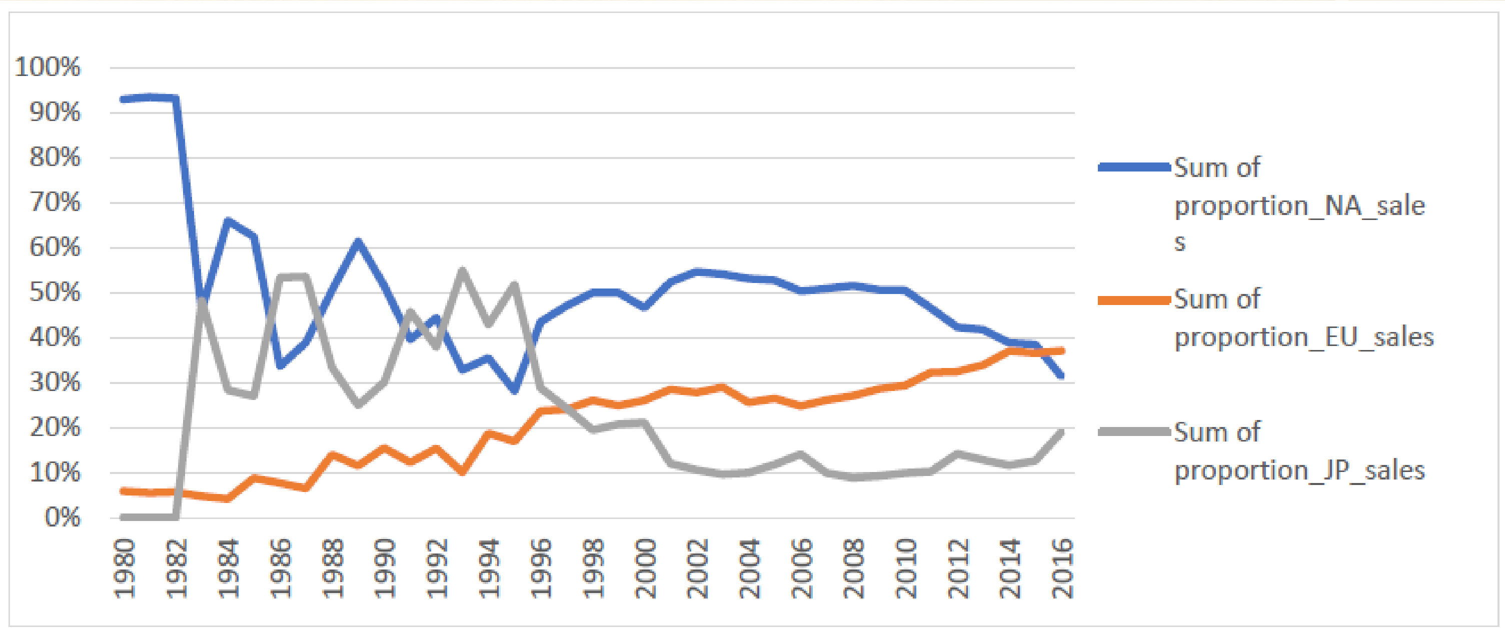
- Are Certain Types Of Games More Popular Than Others?
- What Other Publishers Will Likely Be The Main Competitors In Certain Markets?
- Have Any Games Decreased Or Increased In Popularity Over Time?
- How Have Their Sales Figures Varied Between Geographic Regions Over Time?

Tools and Skills:

Excel
Grouping And Summarizing Data
Descriptive Analysis
Data Visualization

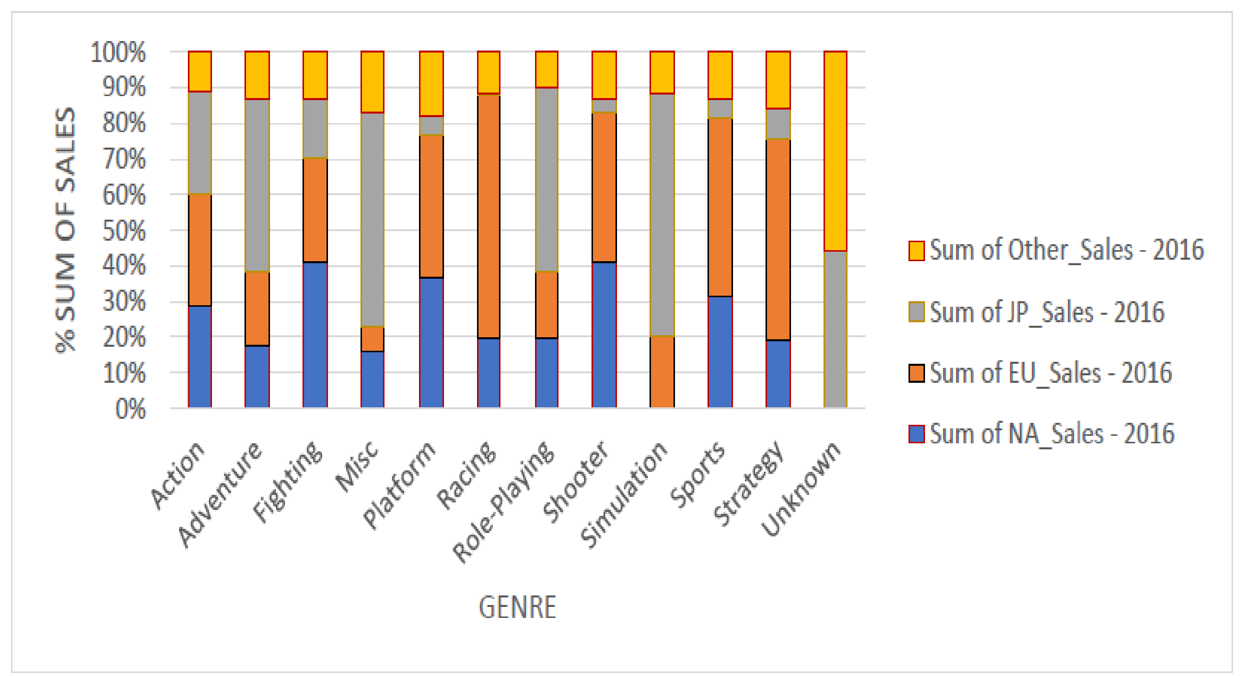
Analysis

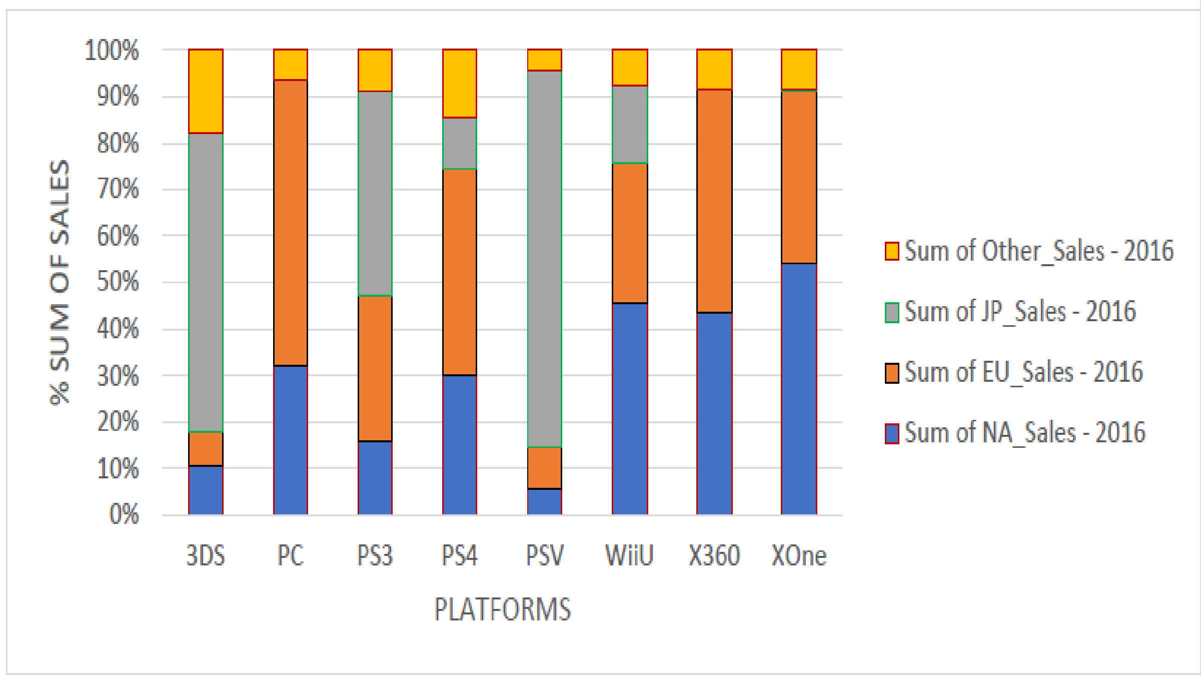
- Data was checked for anomalies and then cleaned.
- Grouping and summarizations are mainly done on understanding the sales for different regions
- Next summarization was done on the proportion of sales of each region towards the global sales.



- North America has dominated sales during most of the time
- From 2013, there is decline in sales for North America and Europe
- Japan also shows sales declination, but not as much as North America and Europe

Genre And Platforms Distributions (2016)





Recommendation:

Redistribution of the budget can be done by focusing more on:

- North America : Shooter and Fighting games with platforms XONE and WiiU
- Europe: Racing and Strategy with platforms PC and X360
- Japan : Simulation and Misc with platforms PSV and 3DS
- Action games can also be considered since it is popular in all regions

Projects

2. INFLUENZA SEASON

Key Questions:

- Provide information to support a staffing plan during flu season for States of US
- Determine whether influenza occurs seasonally or throughout the entire year. If seasonal, does it start and end at the same time (month) in every state?
- Prioritize states with large vulnerable populations. Consider categorizing each state as low-, medium-, or high-need based on its vulnerable population count

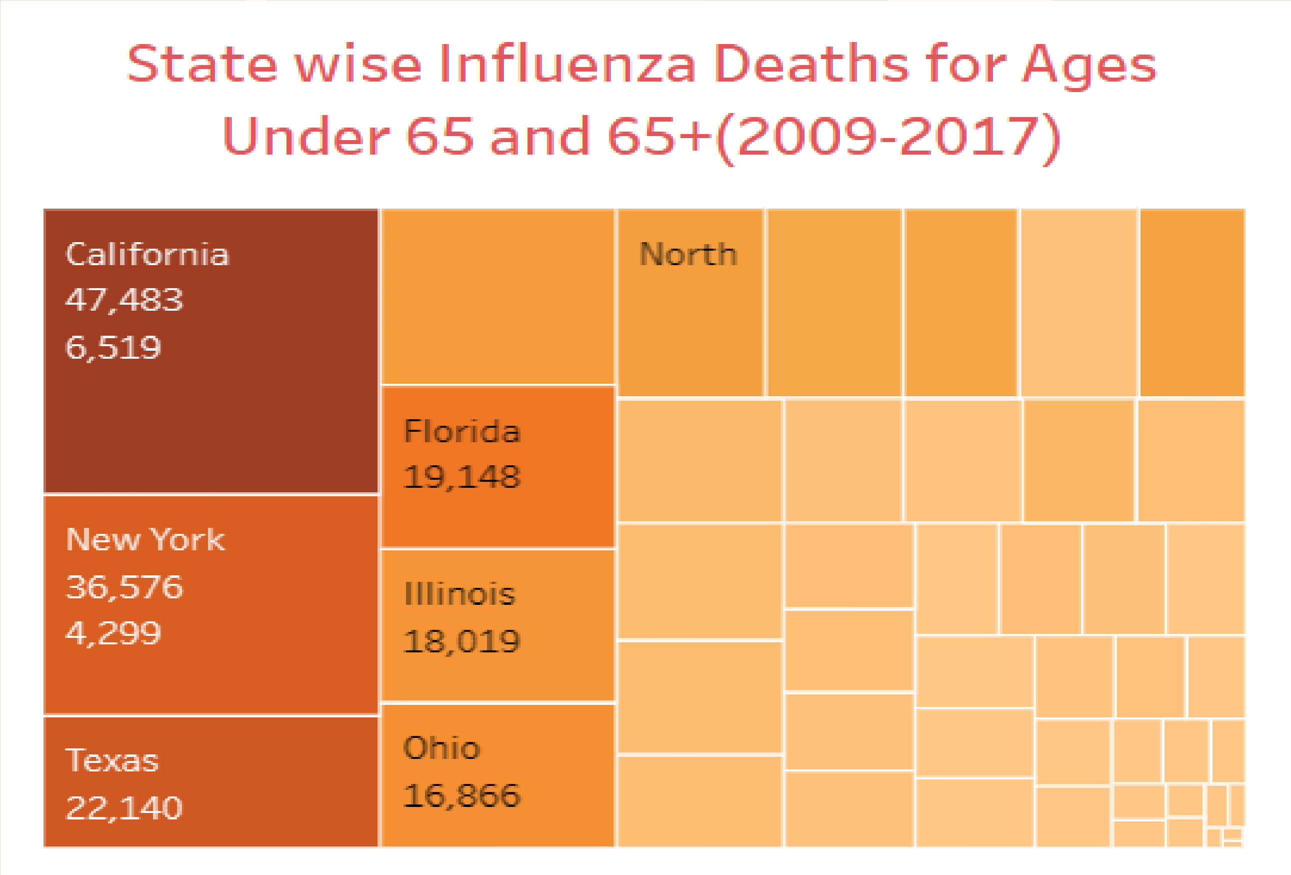
Tools and Skills:

Tableau
Data Cleaning And Integration
Data Transformation
Statistical Hypothesis Testing
Visual Analysis
Forecasting
Storytelling In Tableau

Analysis

Hypothesis:

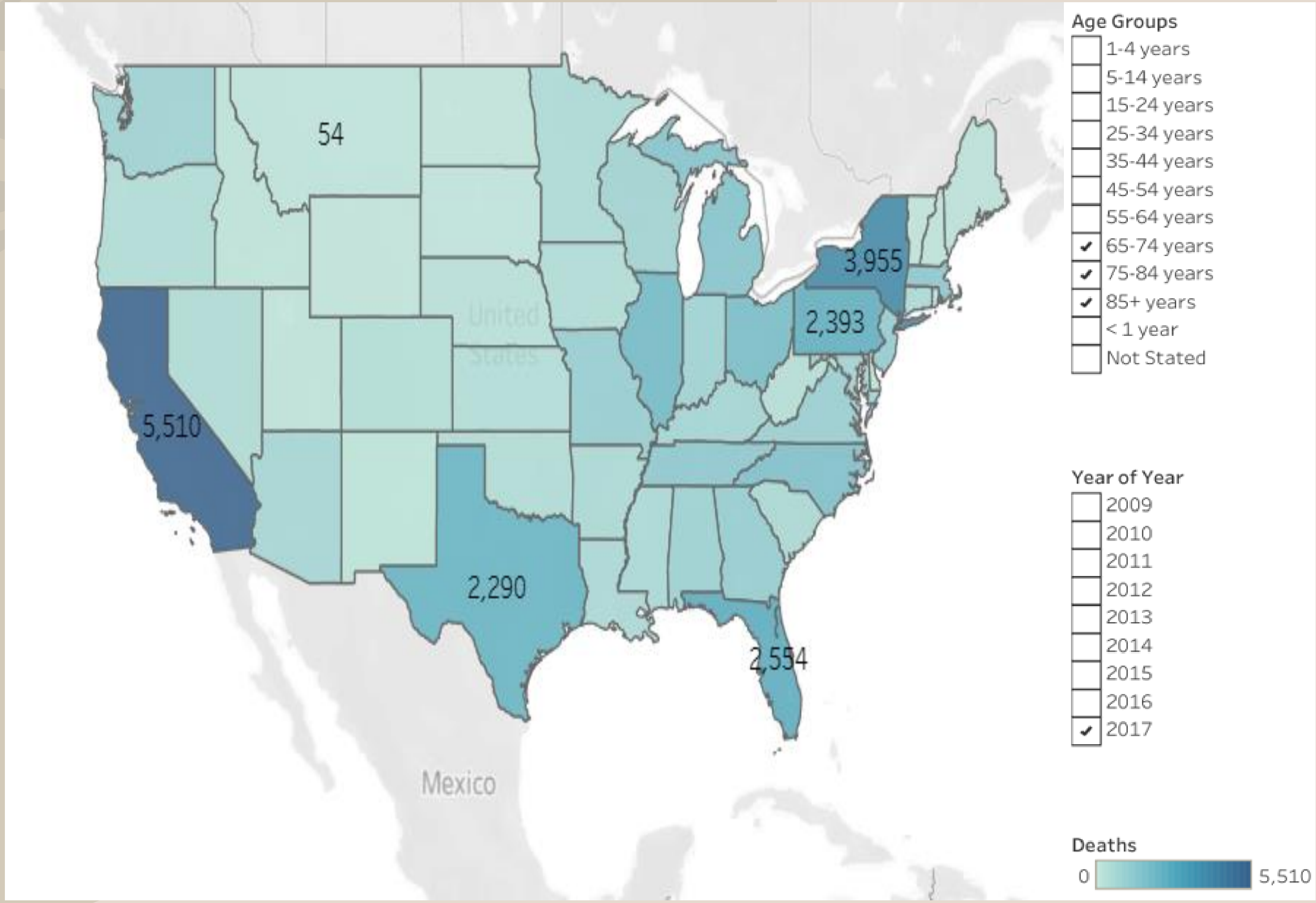
- Older populations are more vulnerable to complications related to the flu.
- If a state has a larger percentage of the population over 65 years old, then the number of influenza related deaths will be higher



State wise Influenza Deaths for Ages Under 65 and 65+(2009-2017)

- The distribution of Deaths for various States are assessed
- The high risk States are California, New York, Texas, Florida and Illinois

Spatial Distributions for Influenza based on Deaths



Age Groups

- □ 1-4 years
- □ 5-14 years
- □ 15-24 years
- □ 25-34 years
- □ 35-44 years
- □ 45-54 years
- □ 55-64 years
- ☑ 65-74 years
- ☑ 75-84 years
- ☑ 85+ years
- □ < 1 year
- □ Not Stated

Year of Year

- □ 2009
- □ 2010
- □ 2011
- □ 2012
- □ 2013
- □ 2014
- □ 2015
- □ 2016
- ☑ 2017

Deaths

0 5,510

Recommendation

- The analysis is done based on death distribution and Major risk areas are identified
- More staffs are to high risk areas - California, New York, Texas, Florida and Illinois
- Medium risk areas are Ohio, Tennessee, North Carolina, Michigan and Missouri.
- Peak season is November to January

Projects

3. ROCKBUSTER STEALTH

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 and Skills :

Data Cleaning And Summarizing
Relational Databases
SQL
SQL JOINs
Subqueries
Common Table Expressions

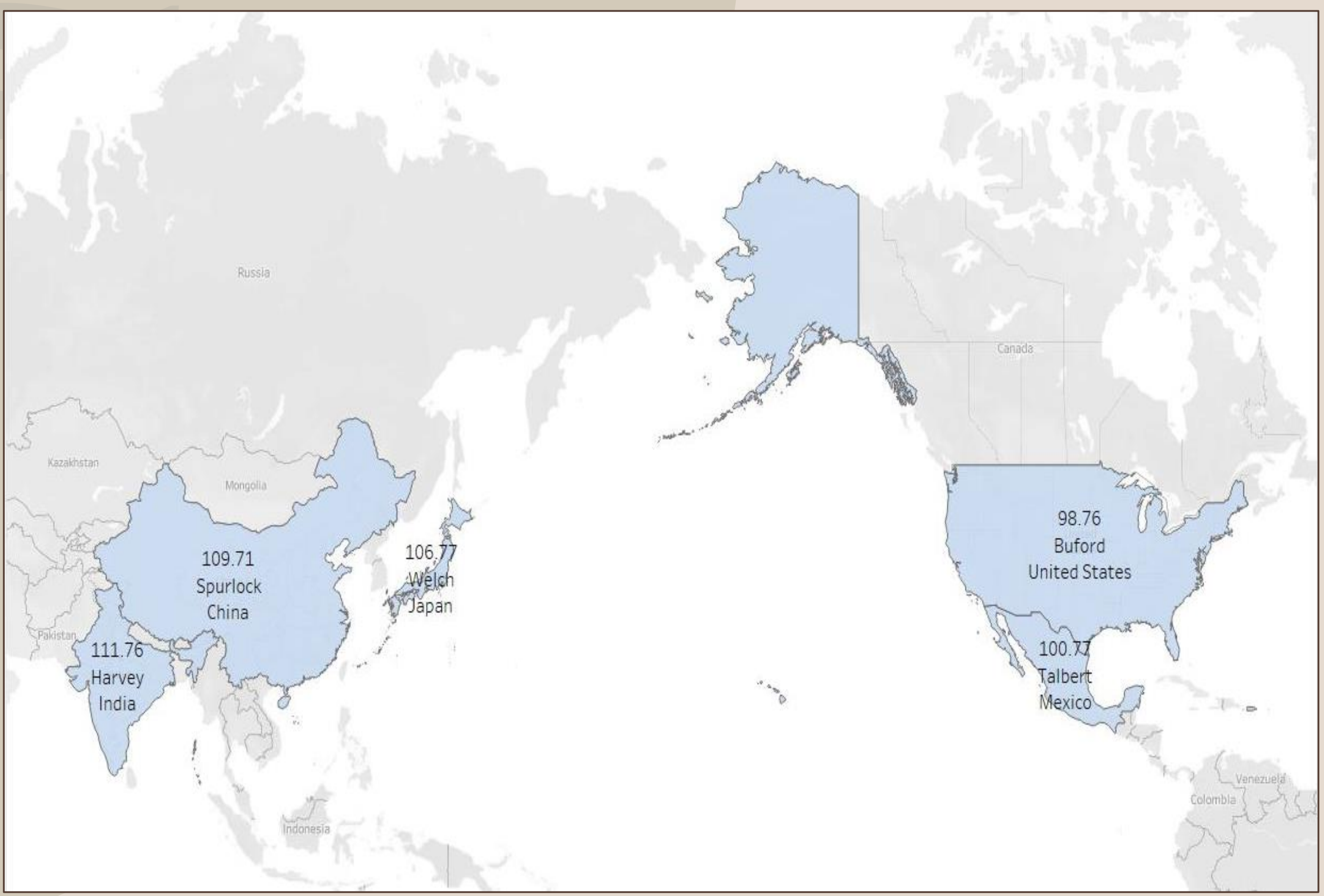
Analysis

- ➢ Database querying done to find the top 10 countries where Rockbuster customers are in.
- ➢ Top 5 Cities are found by using Subquery
- ➢ Top 5 customers are found using SQL JOIN for Loyalty program
- ➢ Implemented Common Table Expression (CTE) in queries

COUNTRY	COUNT	REVENUE
India	60	6034,78
China	53	5251,03
United States	36	3685,31
Japan	31	3122,51
Mexico	30	2984,82
Brazil	28	2919,19
Russian Federation	28	2765,62
Philippines	20	2219,70
Turkey	15	1498,49
Indonesia	14	1352,69

Top 10 Countries based on customers and Revenue

Spatial Distribution for Top 5 Customers and Countries



Recommendation

- Concentrate business more on top Countries like India, China, US, Japan and Mexico based on the number of customers and more revenue
- Include genres that are most popular and those comes in the Top5 list
- Knowing the Top5 customers, taking feedback from them regarding their interests can be helpful to determine the likes of people in similar regions

Projects

4. INSTACART BASKET

Key Questions:

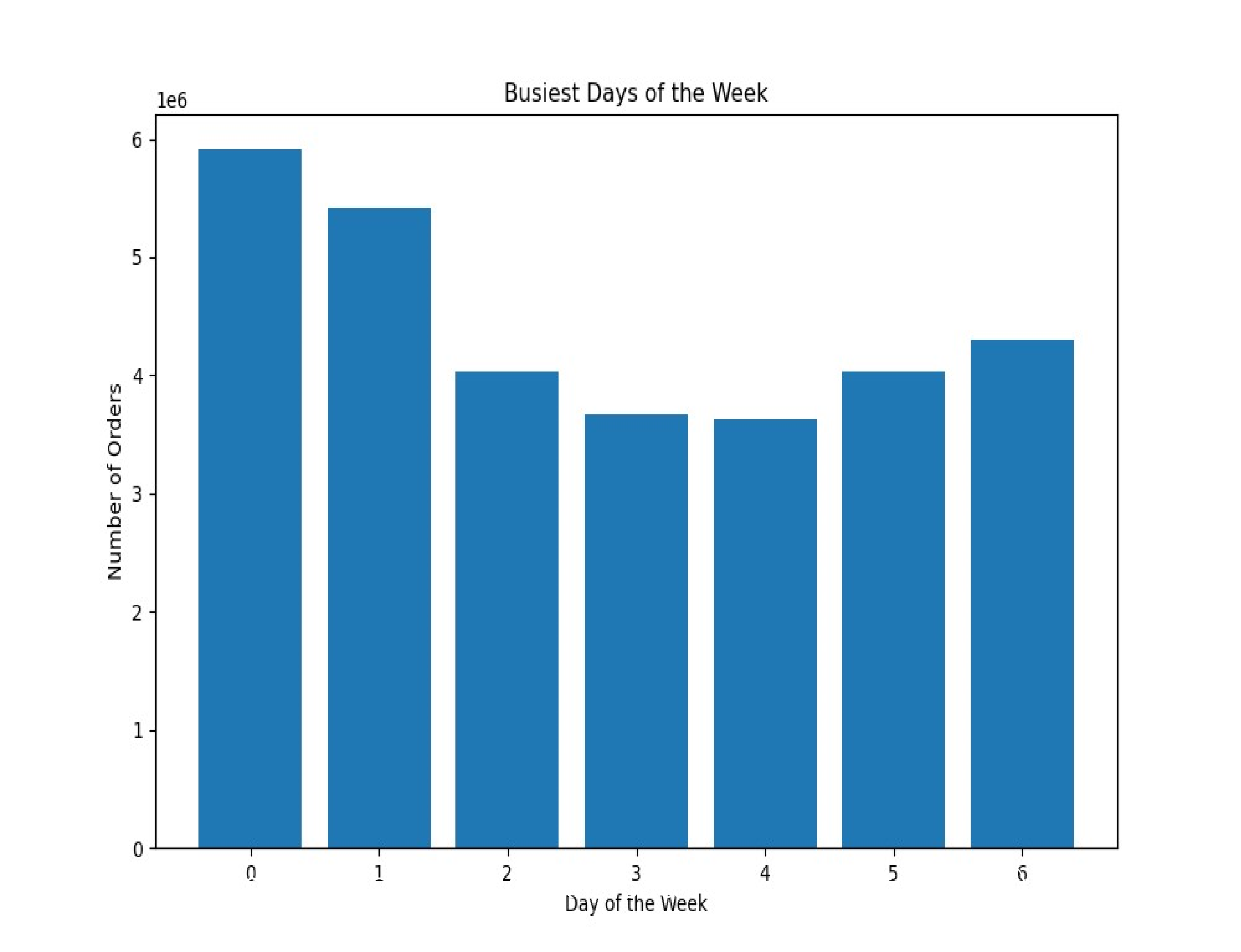
- What are the busiest days and hours of the week?
- Are there particular times of the day when people spend the most money?
- Are there certain types of products that are more popular than others?
- What different types of customers can be identified and how do their ordering behaviors differ?

Tools and Skills:

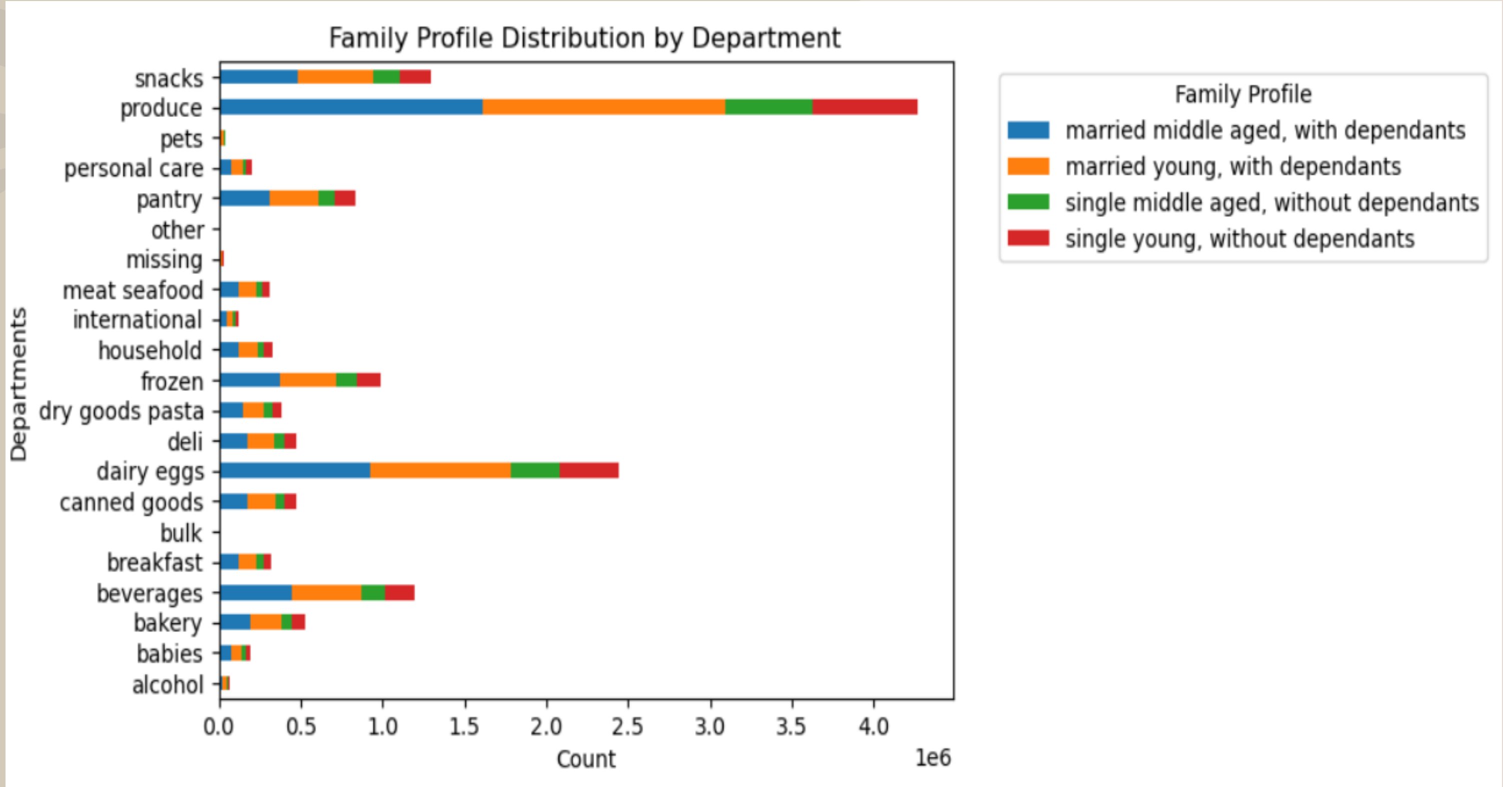
Python
Data Wrangling
Data Merging
Deriving New Variables
Data Grouping and Aggregation
Reporting In Excel
Population Flows

Analysis

- Population Flow: Get the number of records for each dataset, before and after each merge or update
- Check for Data consistency
- Perform Data Wrangling
- Derive new columns from Existing Columns to reach to some new conclusions
- Merge different datasets to get meaningful new dataset
- Create visualizations for the data



Department wise Family Profile Distribution



- The most consumed good is Produce, followed by dairy and eggs
- Married, middle aged with dependants are major in customers

Projects

5. PIG E BANK

Key Question:

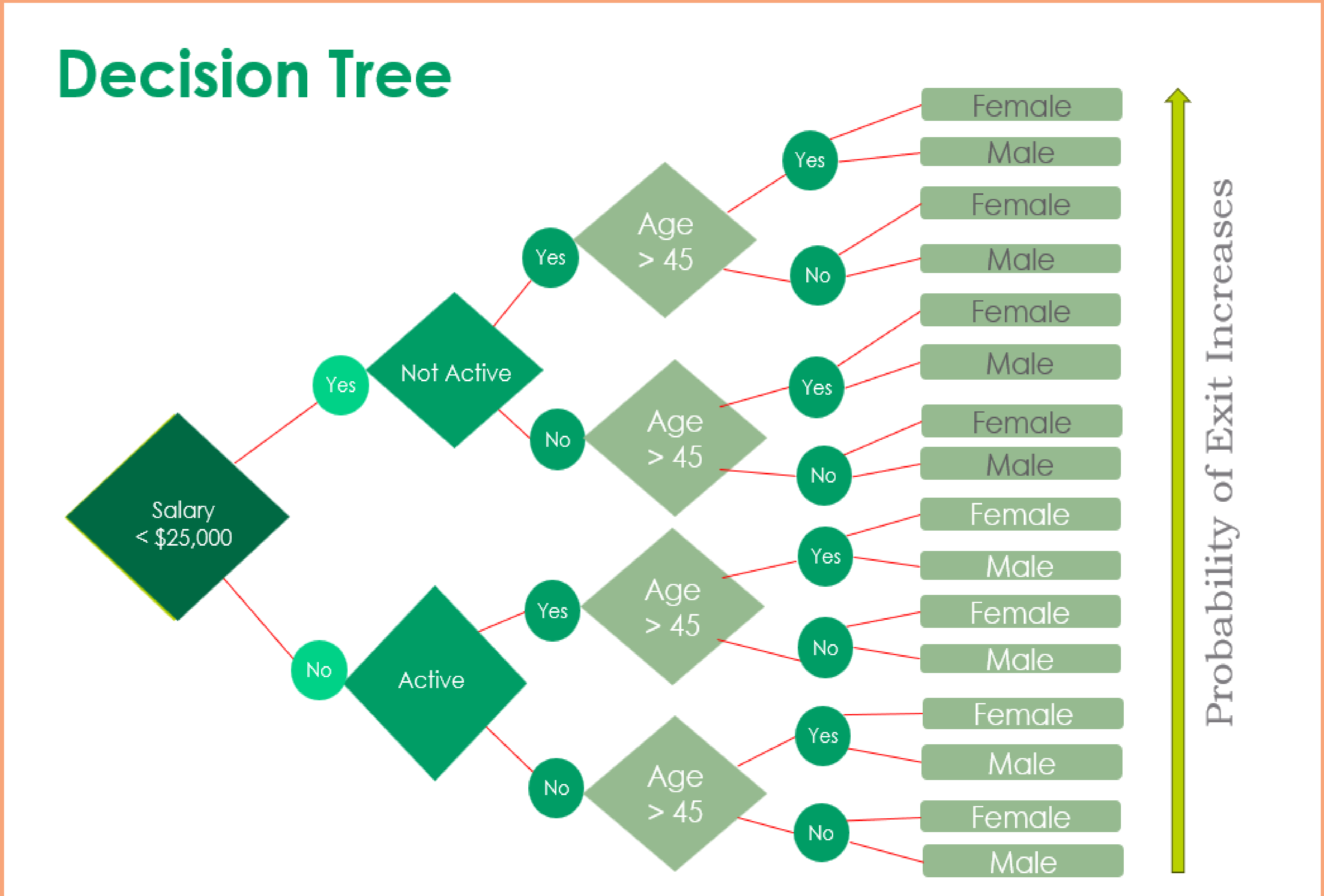
- What are the possible reasons for clients leaving the bank?

Tools and Skills:

Understanding Big Data
Sources Of Bias In Data
Decision Tree
Time Series Analysis And Forecasting
Regression
Classification

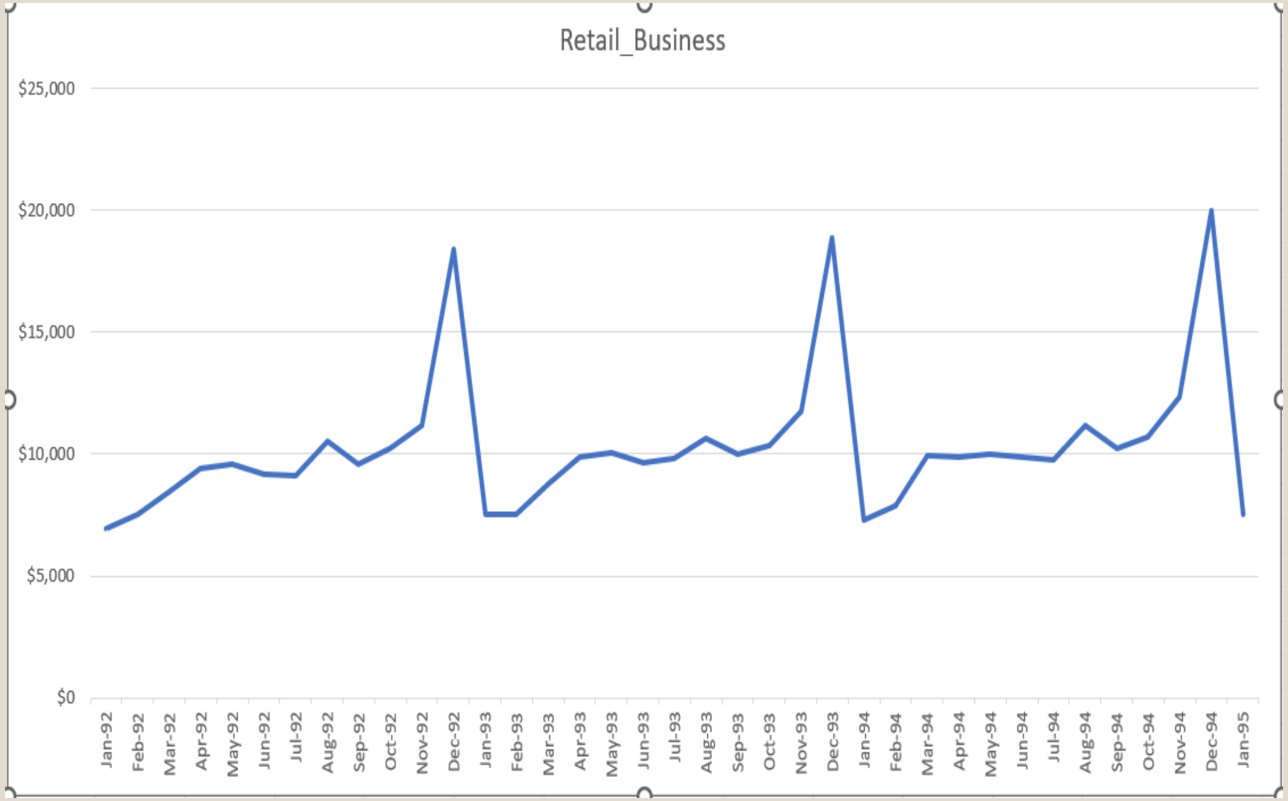
Analysis

- ➢ Analyzed the causes for clients leaving the bank
- ➢ Salary: When Salary is < $25000, there are more chance of clients leaving
- ➢ Inactive customers exit more
- ➢ Age >45 has more tendency to exit
- ➢ Females are more to exit



Decision Tree

Understanding Time Series Retail Business



- ➢ Represents Seasonality which has peaks during some months Nov to Dec (holiday season)and then a drop after that, during January.
- ➢ It repeats for the years followed.

Recommendations:

- ➢ Stock up during the peak seasons so that they didn’t run out of stock during the peak times.

- ➢ Decrease the stock after that so that there won’t be any unsold items flooding the shelves.

Projects

6. US HOUSE PRICING

Key Questions:

- What are the main factors affecting House pricing?
- Do the zip code density and zip code population play a vital role in determining house price?

Tools and Skills:

Sourcing Open Data
Exploring Relationships
Geographical Visualizations using Python
Supervised Machine Learning - Regression
Unsupervised Machine Learning-Clustering

Exploratory Analysis

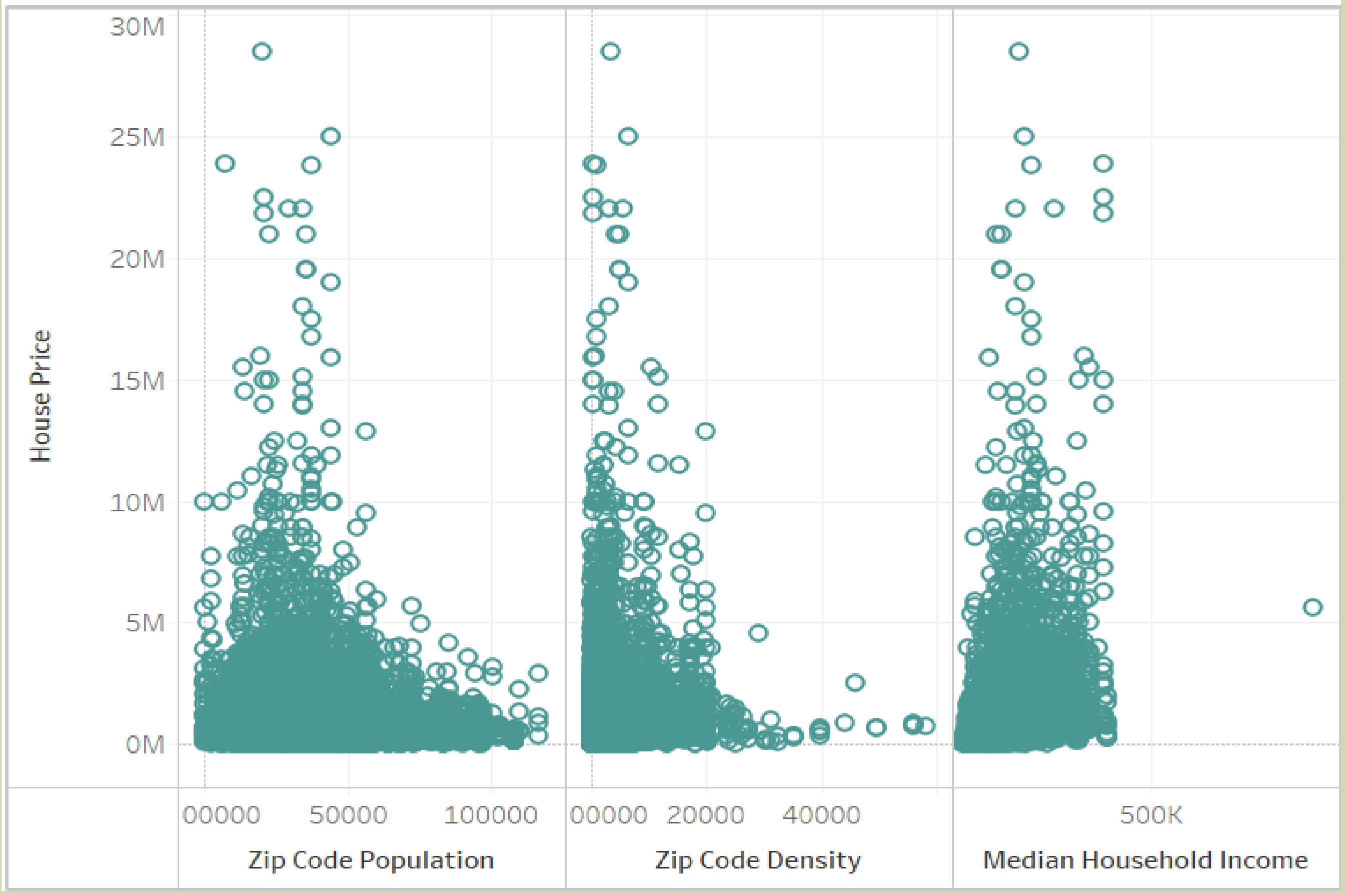
- ➢ As an initial step, the linear relationship between price and other variables like Zip Code Population, Zip Code Density and Median Household Income were analyzed.

- ➢ No linear relationship identified between Median Household Income and Price of House.

- ➢ The linear like relationship between Zip Code Population and House Price were further explored.

Hypothesis:

- ➢ When the Zip Code is low or medium populated, there is a tendency of higher house price than that of highly populated Zip Code.



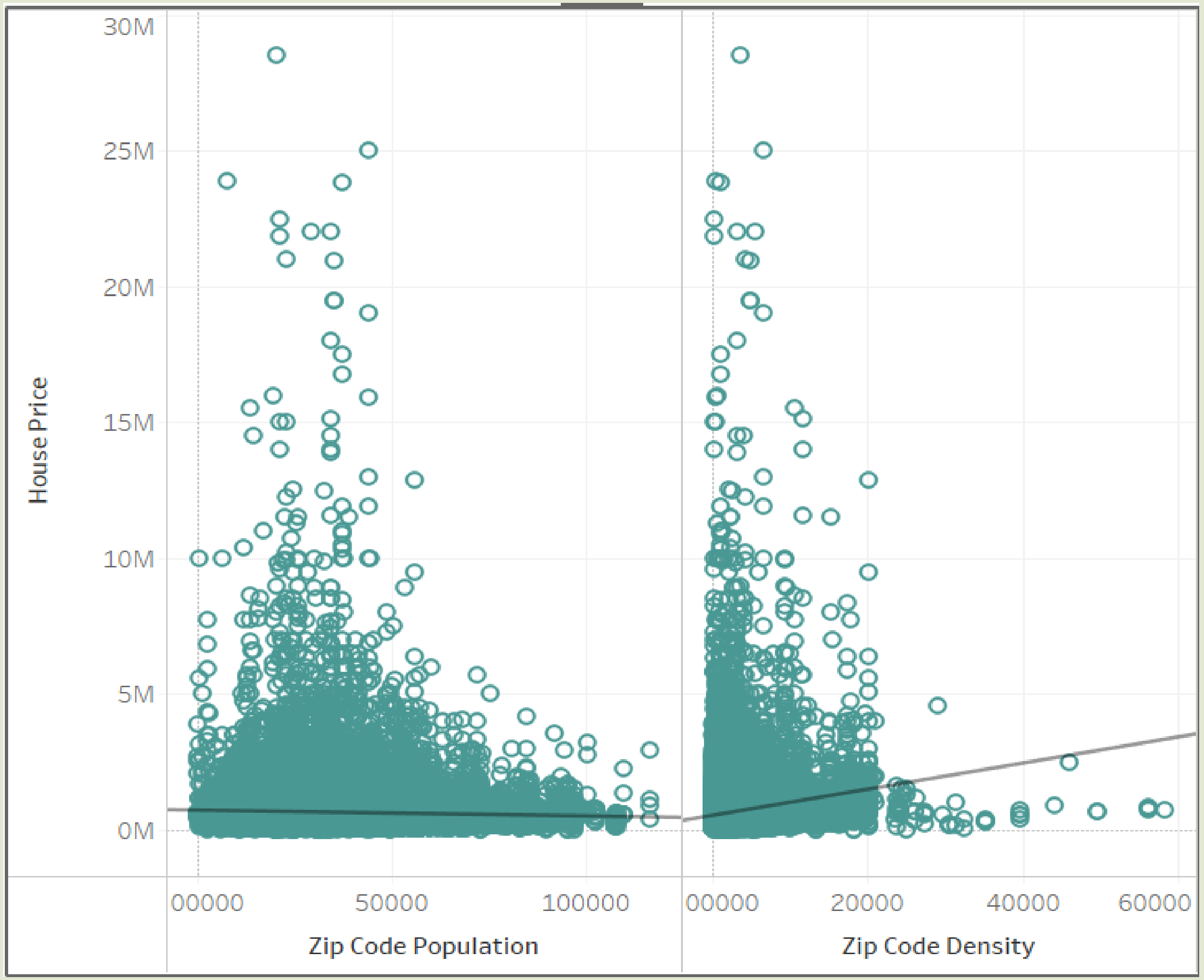
Linear Regression

- The hypothesis is tested using Linear Regression model on the variables Zip Code Population and House Price.

- The fit of the model was confirmed by taking model summary statistics between test output and the predicted output.

- The Root Mean Squared Error (RMSE) was too high, and the variance r-squared value was too low.

- This indicated that the data is not linear, or the model is of poor fit.

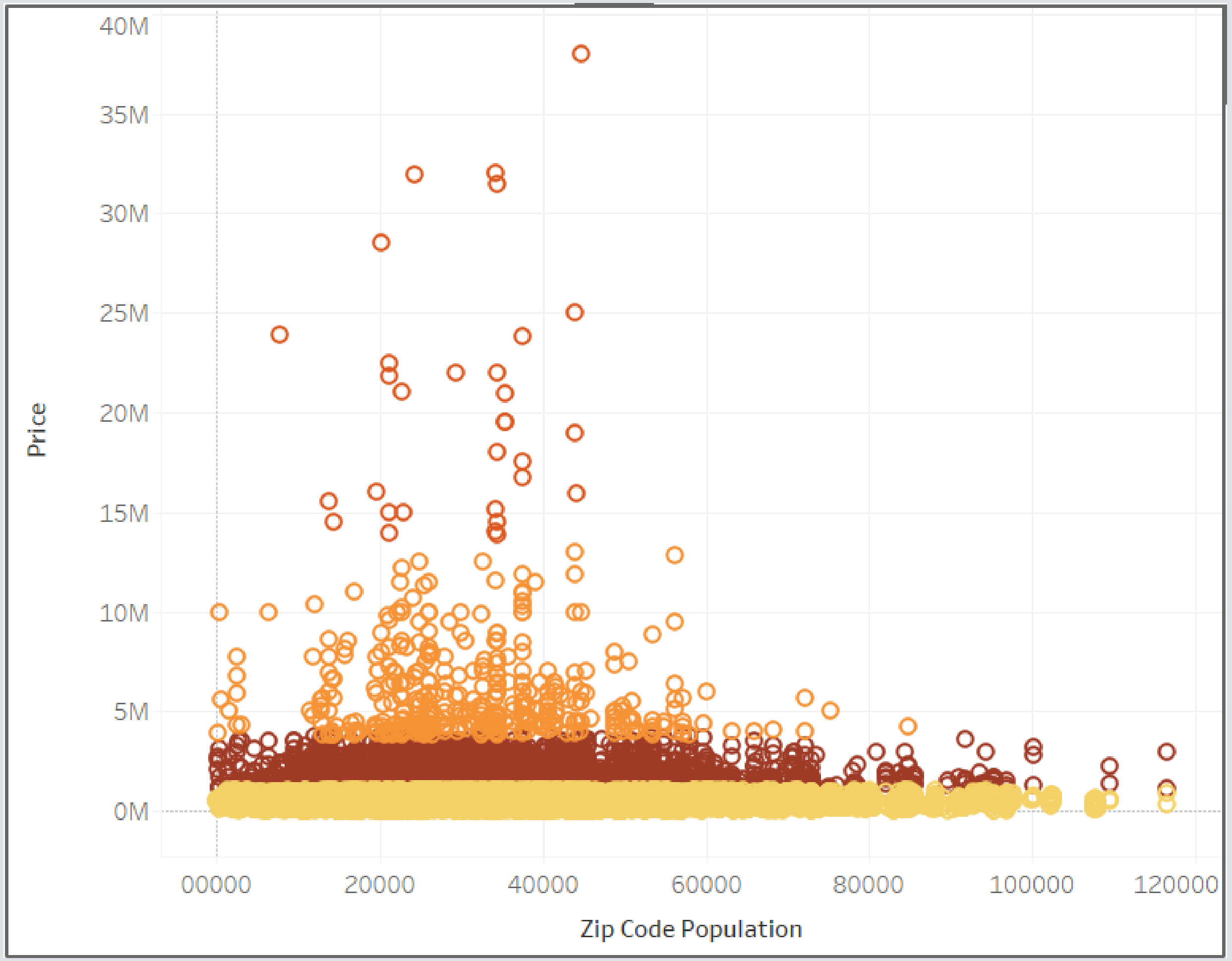


Cluster Analysis

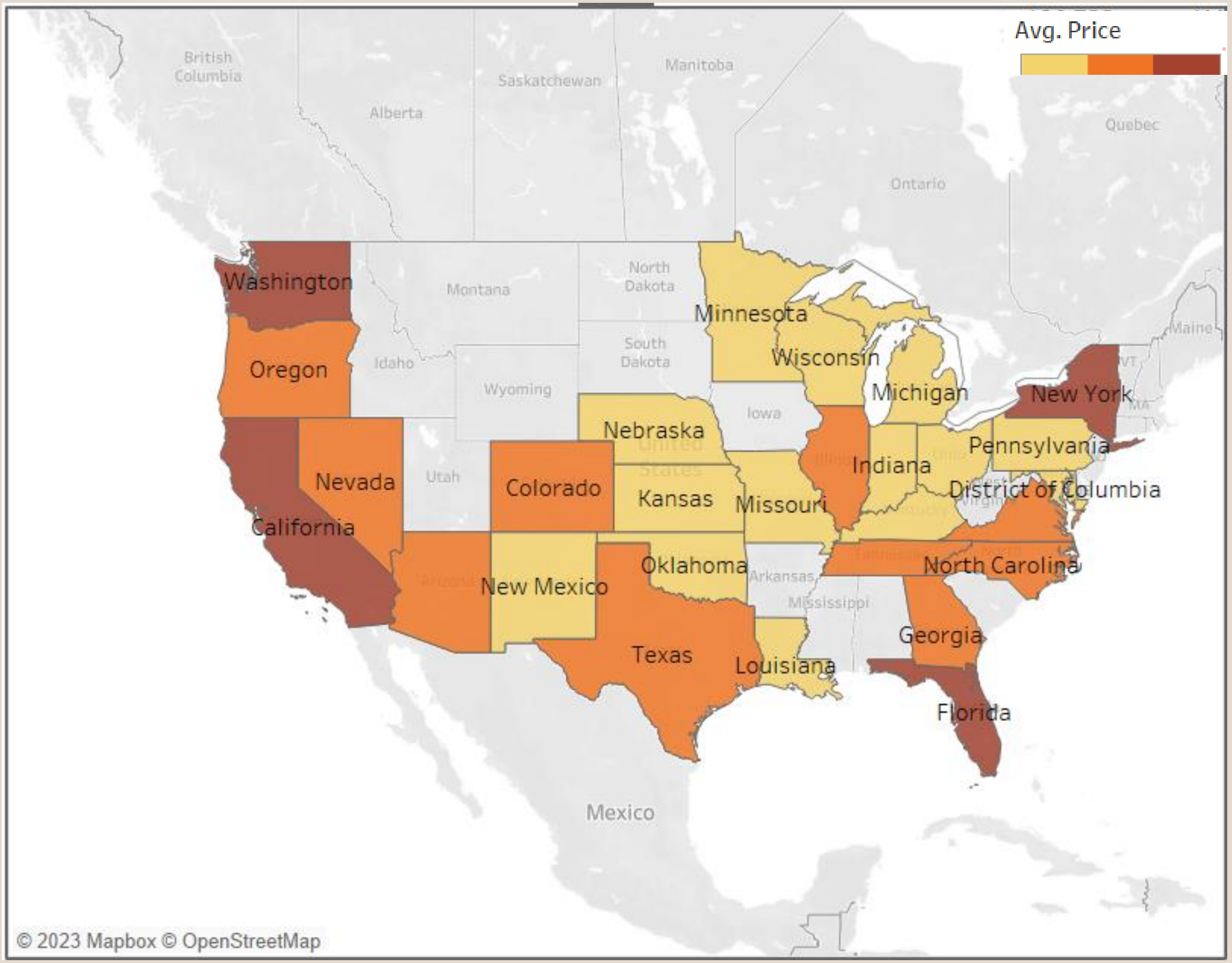
- ➢ Data was analyzed by converting them to clusters using the k-means model
- ➢ 4 clusters based on the elbow curve (0,1,2,3)
- ➢ A pattern or trend for data points in a particular cluster:

FINDINGS:

- ➢ Clusters 0: Lowest price and have almost all the Zip Code Populations.
- ➢ Cluster 1: Medium price and Zip Code Population upto 80000
- ➢ Cluster 2: Higher price and Zip Code Population upto 45000
- ➢ Clusters 3: Lowest price and have almost all the Zip Code Populations.



State wise House Price and Zip Code Population



© 2023 Mapbox © OpenStreetMap

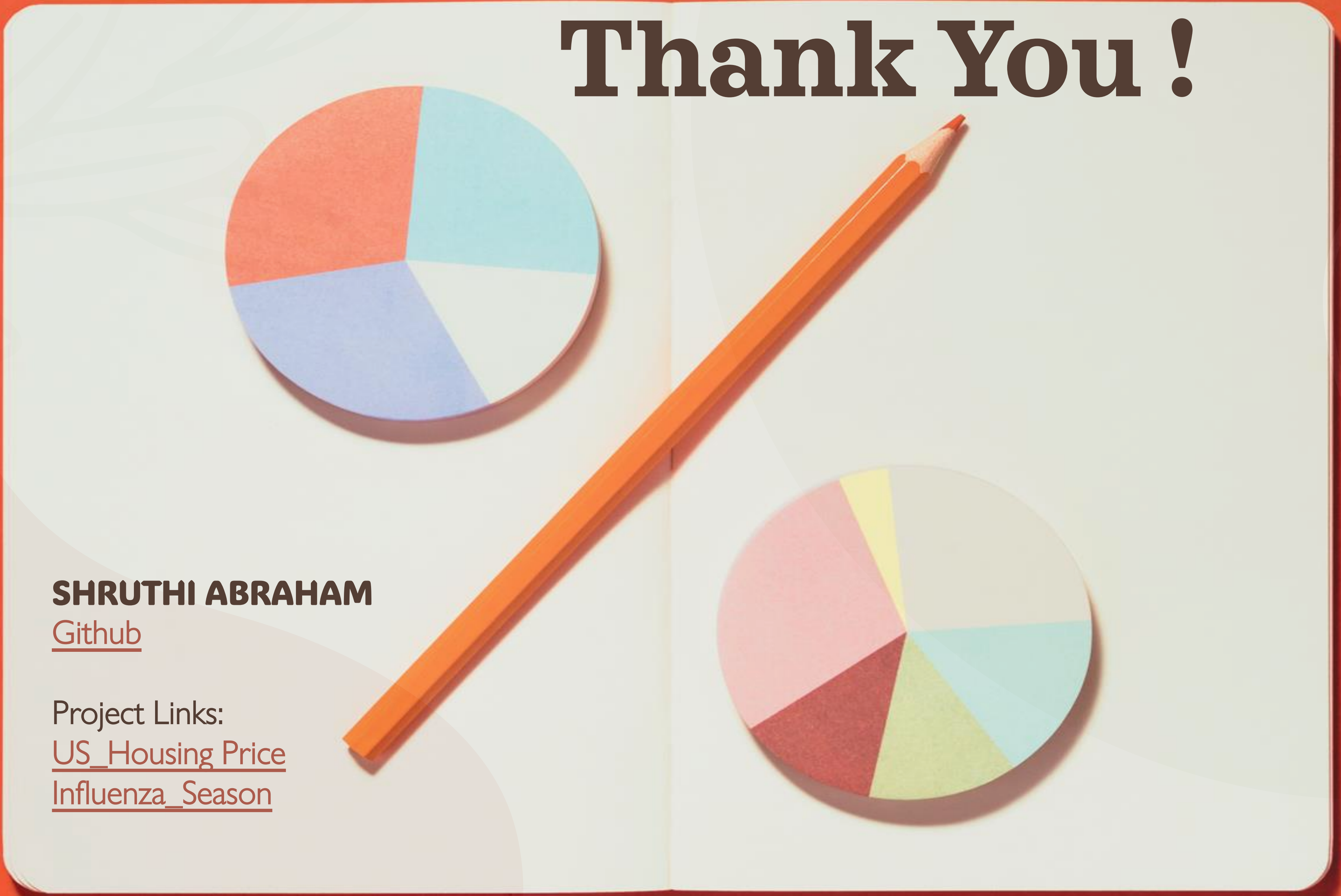
Highest Average Price: California, Washington, Florida and New York
Medium average Price: Oregon, Nevada, Arizona, Colorado, Texas, Indiana, North Carolina and Georgia.

Recommendations:

- Selection of a State to buy a house can be made by considering the average price distribution.
- Within a state, the Zip code Population has an impact on the prices because lower Zip Code Population tend to have high prices.

Projects

Thank You !



SHRUTHI ABRAHAM
Github

Project Links:
US_Housing Price
Influenza_Season