Tony Soewignjo

DATA ANALYTIC PORTFOLIO

About me

I have a passion for working with data. My Data Analytic Certification, combined with my education in Computer Science and work experience in Software Industry, provide a strong foundation for a Data Analyst job position. On the top of that, my interpersonal skills from more than 20 years in leading a non-profit organization and working with people from various social background should help to maximize my efficiency in working together with people. I look forward to join a team that would utilize my skills and experiences for the benefit of the organization.

Portfolio

GameCo

Influenza

Rockbuster

Instacart

Boat Sales

Github Repository

- GameCo
- Influenza
- Rockbuster
- Instacart
- Boat Sales

Excel

Tableau

PostgreSQL

Python

Jupiter

Github

Power Point

1. Game Co

Objective

A new video game company, GameCo, wants to use data to inform the development of new games. The goal is 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?
- Have any games decreased or increased in popularity over time?
- How have their sales figures varied between geographic regions over time?

Skill set

- MS Excel
- Grouping data
- Summarizing
- Descriptive analysis
- Visualizing results in Excel

Links

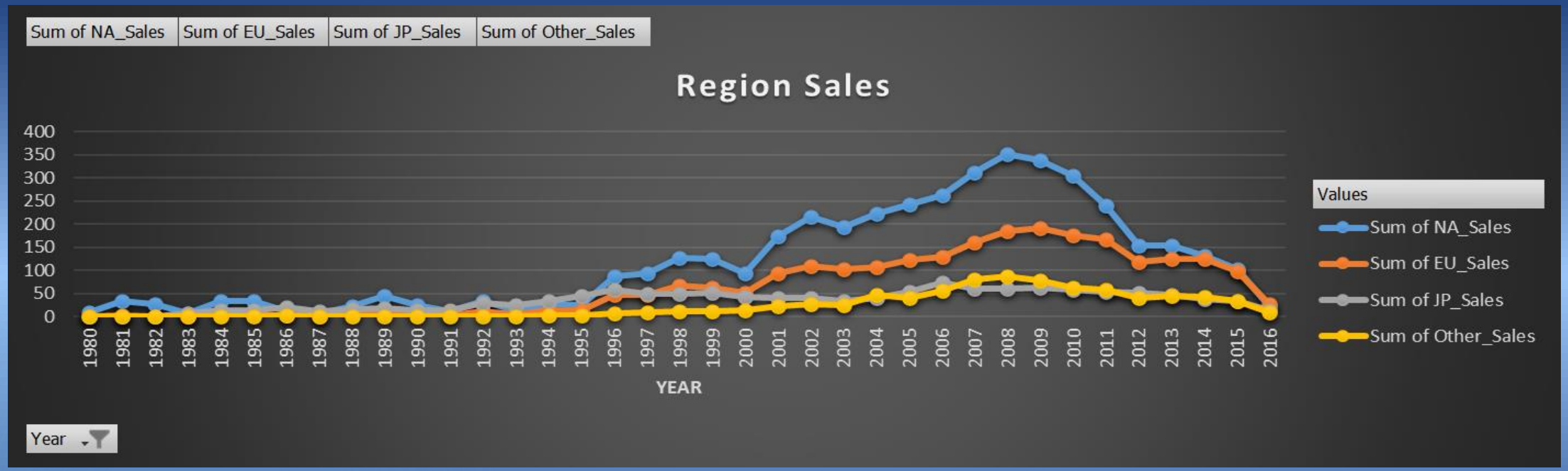
Project Brief

Sales Data

Analysis

Presentation

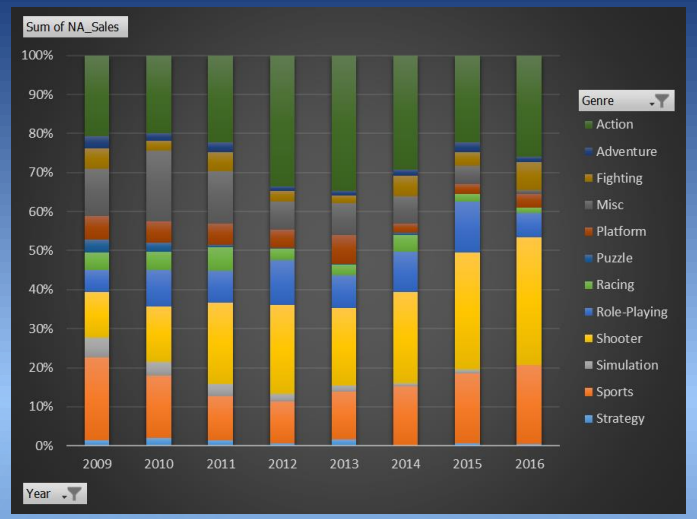
GameCo Regional Sales Data (1980-2016)

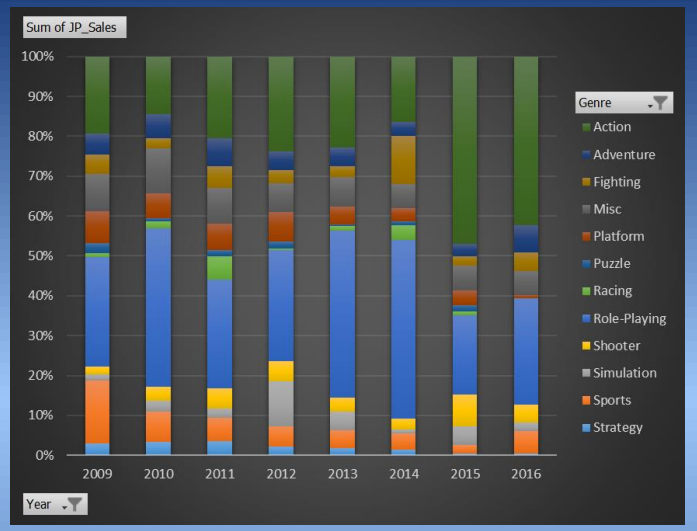


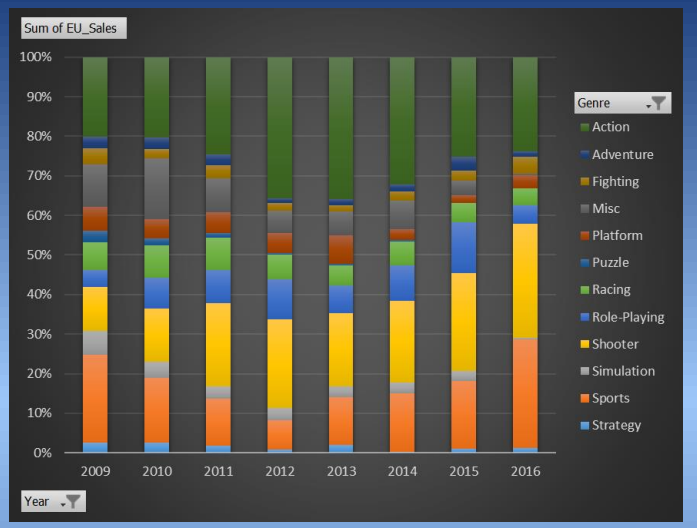
*There is a pattern of increased sales in **North America** and **Europe** since 1995 and reached their peak in 2008. After that sales seemed to decline in all regions until 2016.*

Region	2008	2016
NA	351M	22M
EU	185M	27M
JP	60M	14M
Other	86M	10M

GameCo Genres popularity (2009-2016)







North America:

- **Shooter**
- **Action**
- **Sports**

Japan:

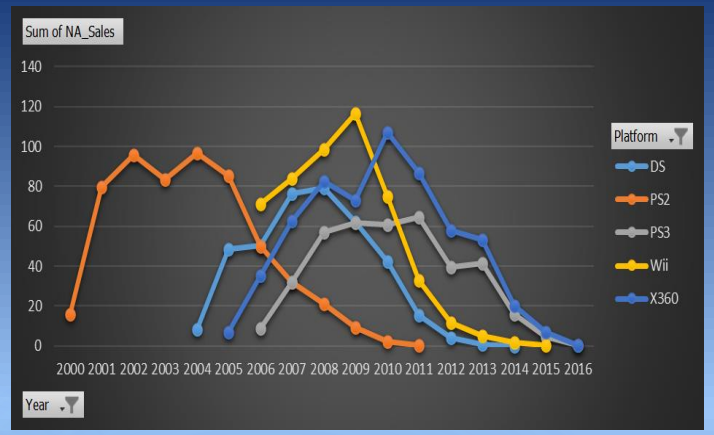
- **Role Playing**
- **Action**
- **Sports**

Europe:

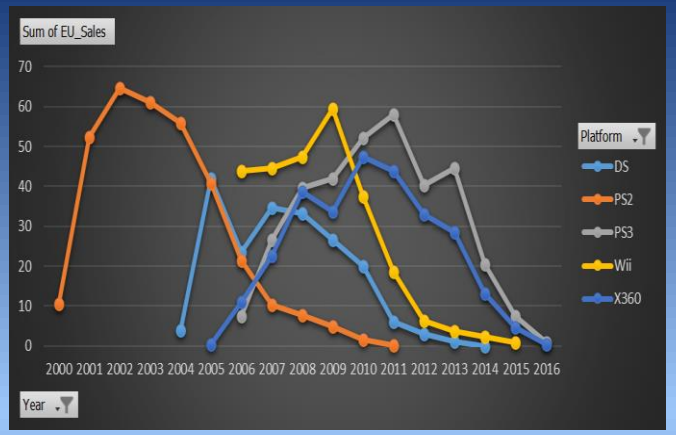
- **Shooter**
- **Action**
- **Sports**

Key point: Genres popularity has been relatively constant in the past 8 years. This consistency might serve as an indicator that **Games Genres are not the main factor in region sales decrease**.

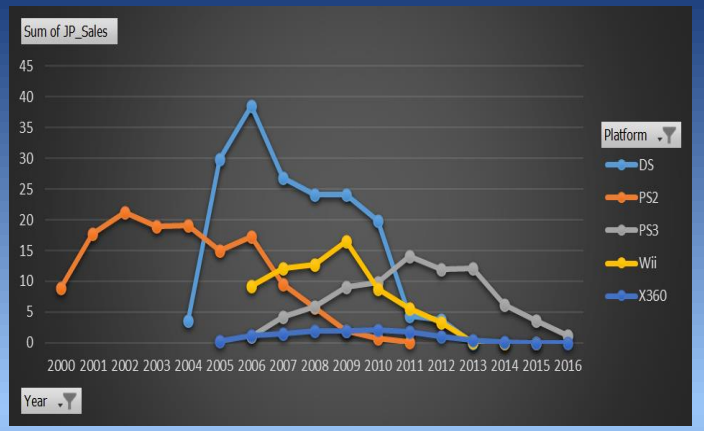
GameCo Top 5 Platform Analysis (2000-2016)



North America



Europe



Japan

Key points:

- Top 5 platform from global sales: **DS, PS2, PS3, Wii, X360**
- Each region shows a similar pattern where each platform will have approximately **8-12 years life cycle**
- This might be a factor in global sales or region sales decline

Recommendations

- GameCo should focus the marketing budget in each region based on **Genre popularity.**
- GameCo should conduct an Analysis of the possibility of **new game platforms** in the market. (ie: mobile gaming)
- GameCo should put extra budget in **North America** region who used to be the major market in game industry.

2. Influenza Season

Objective

The United States has an influenza season where more people than usual suffer from the flu. Some people, particularly those in vulnerable populations, develop serious complications and end up in the hospital. Hospitals and clinics need additional staff to adequately treat these extra patients.

Skill set

- **MS Excel**
- Data cleaning & integration
- Statistical hypothesis
- Visual analysis
- Forecasting
- Storytelling in Tableau

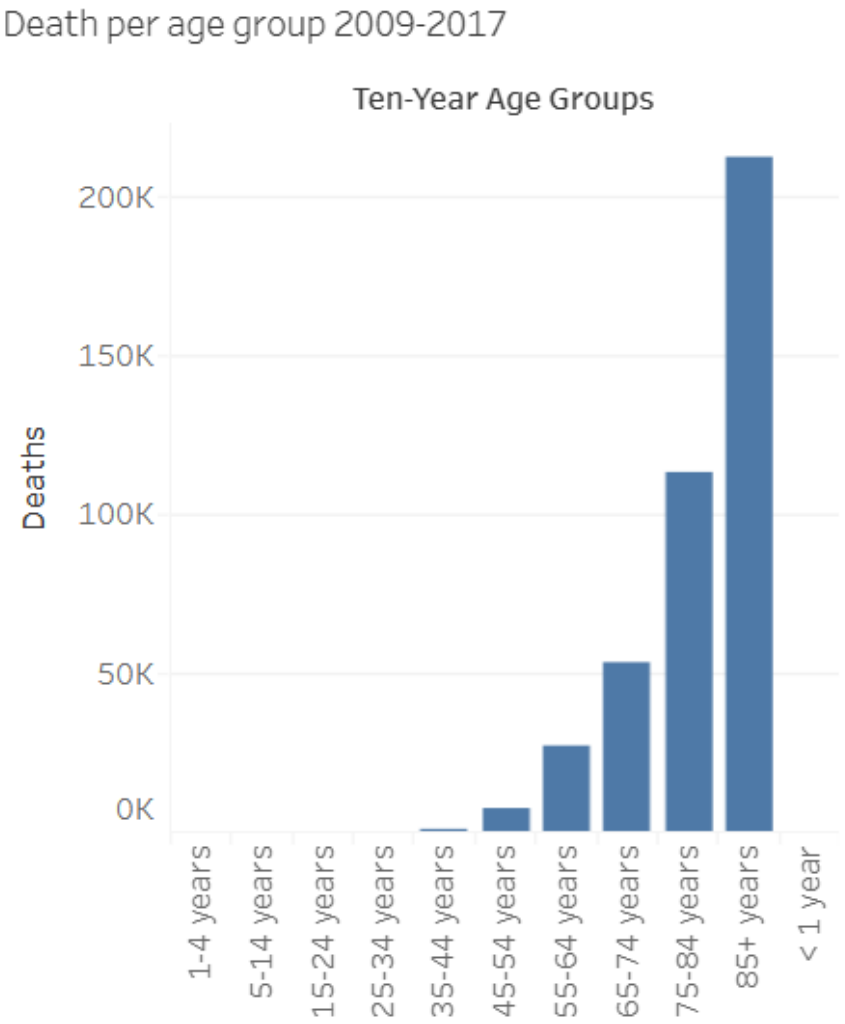
Requirements

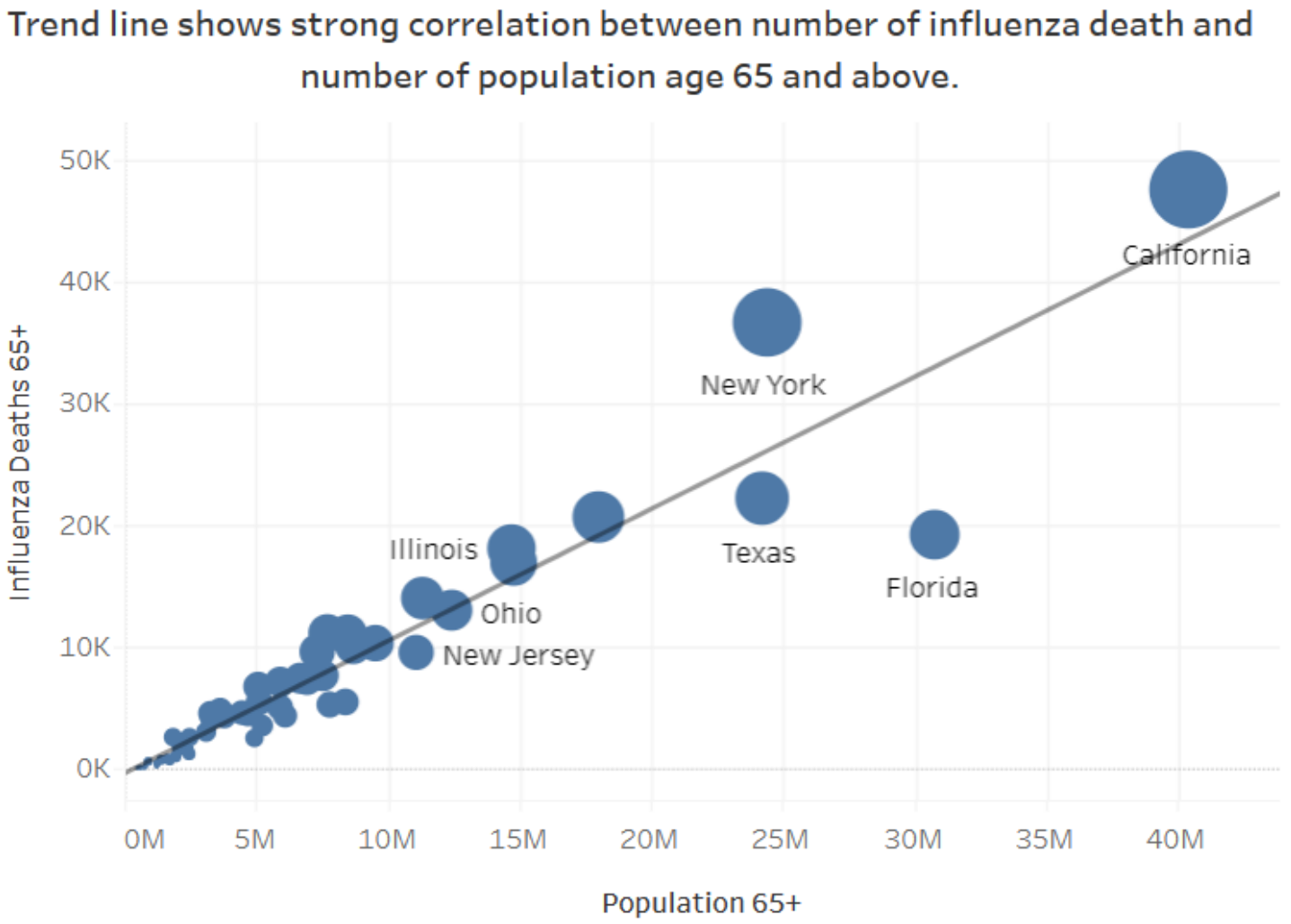
- Provide data that can help inform the timing and spatial distribution of medical personnel throughout the U.S.
- Determine whether influenza occurs seasonally or throughout the entire year.
- Prioritize states with large vulnerable populations.

Links

Project Brief
CDC Data
Census Data
Tableau
Presentation

Do older people have more risk?

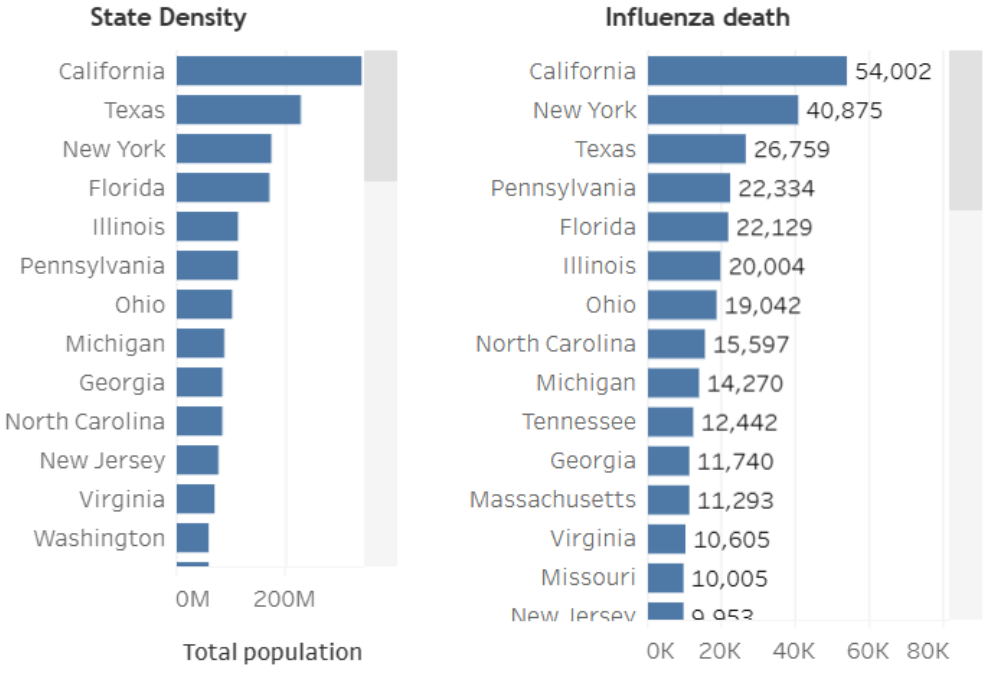




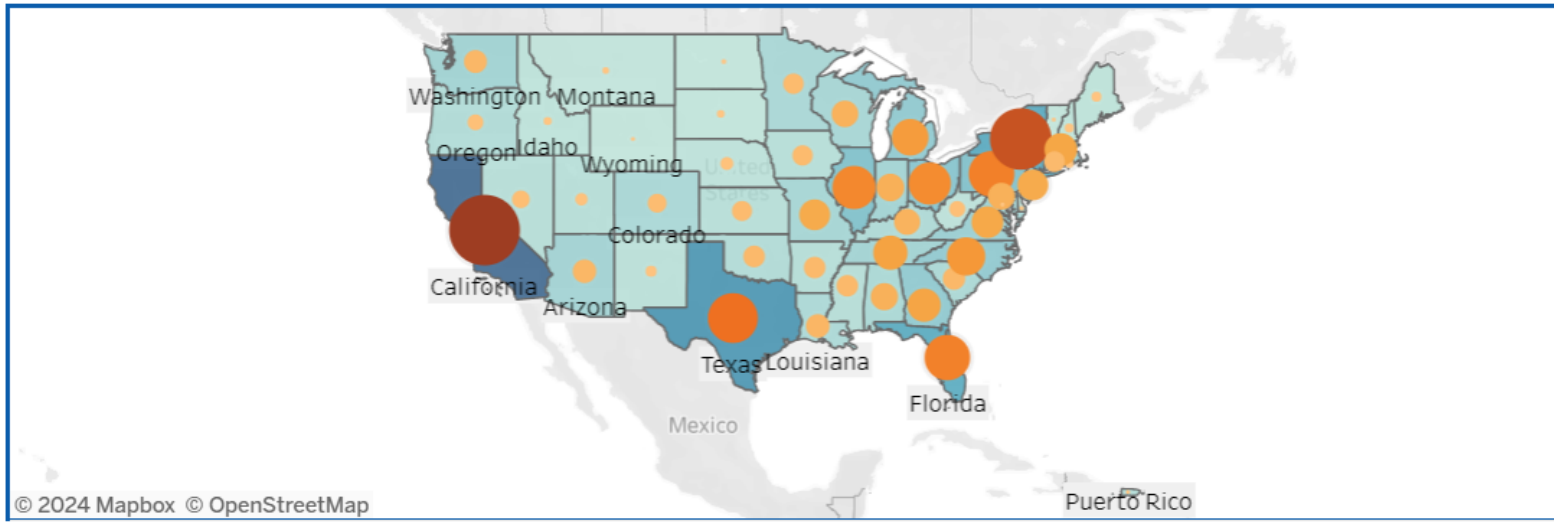
These two graphs prove the hypothesis that there is a strong correlation between the number of vulnerable population and the number of influenza deaths.

Where to send workers?

US states population density and influenza death from 2009 to 2017



*** Dark BLUE state: represent area with higher population.
*** Dark RED circle: represent area with higher influenza death.



States with higher population and higher influenza death should receive more medical workers.

The above graphs show that the size of influenza death is not always determined by the number of population density, but it is also determined by other factor such as weather condition. States on the northern part tend to have higher proportion of influenza deaths.

Conclusion

Factors that impact influenza season:

1. Population density and number of influenza death should the main factor to be considered when sending medical workers. Put extra attention to states with high density and colder weather such as New York.

2. Put attention also to states with high percentage of vulnerable population such as Florida. Although the data shows Florida with a lower percentage of influenza death, it could be expected that there are more influenza patients in that area.

3. Analysis shows that typical influenza season last from December to March.

Next step:

1. Research the impact of weather into influenza season.
2. Perform Flu Shot Analysis.
3. Set up meetings with medical staff administrators.

3. Rockbuster

Objective

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, the management team is planning to launch an online video rental service in order to stay competitive.

Skills

- **SQL Queries**
- Relational databases
- Joining tables, subqueries
- Common table expressions

Key questions

- Which movies contributed the most/least to revenue gain?
- Where are customers with a high lifetime value based?
- Do sales vary between geographic regions?

Links

Project Brief
DB Schema
Data Dictionary
Presentation

DATA OVERVIEW

Total movie **1000**	Total actor **200**	Total genre **16**
Total customer **599**	Total country **109**	Total inventory **4581**
Total staff **2**	Total store **2**	Total revenue **$61,312**

REGIONAL ANALYSIS

country	total_customer	total_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	2219.7
Turkey	15	1498.49
Indonesia	14	1352.69

Top five countries based on revenue:

- **India**
- **China**
- **United States**
- **Japan**
- **Mexico**

Note: there is a **strong correlation** between total revenue and total customer.

CUSTOMER ANALYSIS

Note:
The top 10 global customers are distributed almost evenly in 9 countries. This shows that **top customer's locations** are spread out randomly and not confined to any regions.

Location of TOP TEN Global Customers

first_name	last_name	city	country	total_payment
Eleanor	Hunt	Saint-Denis	Runion	211.55
Karl	Seal	Cape Coral	United States	208.58
Marion	Snyder	Santa Brbara dOeste	Brazil	194.61
Rhonda	Kennedy	Apeldoorn	Netherlands	191.62
Clara	Shaw	Molodetno	Belarus	189.6
Tommy	Collazo	Qomsheh	Iran	183.63
Ana	Bradley	Memphis	United States	167.67
Curtis	Irby	Richmond Hill	Canada	167.62
Marcia	Dean	Tanza	Philippines	166.61
Mike	Way	Valparai	India	162.67

RECOMMENDATION

- Build a **unique marketing strategy** in each region, based on genre popularity in that region.
- Focus on adding new movies with **PG-13 and NC-17 ratings**.
- Focus on renting movies in the top five categories: **Sports**, **Sci-Fi**, **Animation** **Drama**, **Comedy.**
- Provide movies with languages available in the top five countries: **India**, **China**, **United States**, **Japan**, **Mexico**.

- Next Step: Further research with movie data beyond 2006.

4. Instacart Basket Analysis

Objective

Instacart wants to uncover more information about their sales patterns. The goal is to perform an initial data and exploratory analysis of some of their data in order to derive insights and suggest strategies for better segmentation based on the provided criteria.

Skills

- **Python**
- Jupiter Notebook
- Data wrangling & merging
- Aggregating data
- Reporting in Excel
- Population flows

Key questions

- Busiest days of the week and hours of the day
- Distribution among users in regard to their brand loyalty
- Differences in ordering habits of different customer profiles

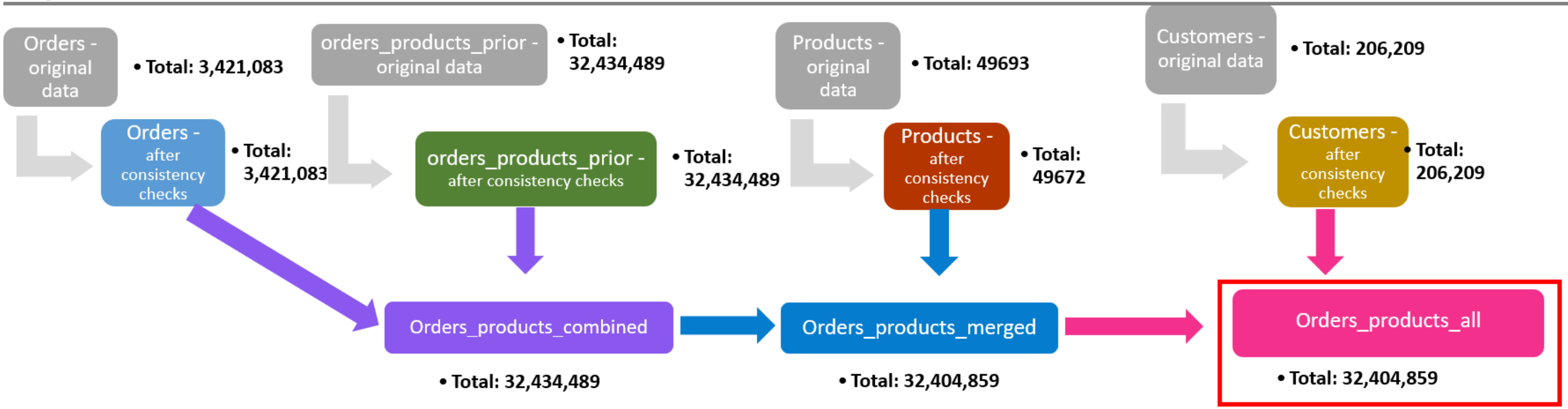
Links

Project Brief
Instacart Data
Customer Data
Python Scripts
Final Report

Data integrity check

instacart

Population flow



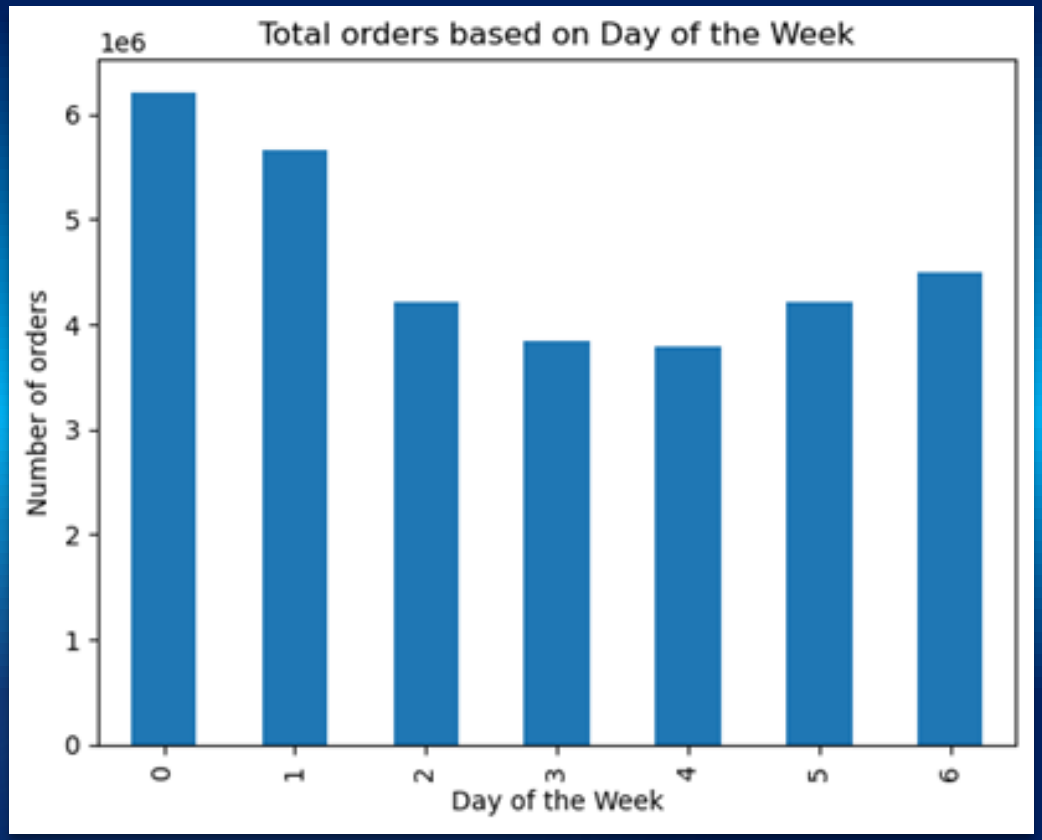
1. The grey boxes in the first row represent the original data sets as they were downloaded.
2. The second row of boxes (colored) represents the data sets after they were manipulated.
3. The third row represents the merges that were performed between the datasets.

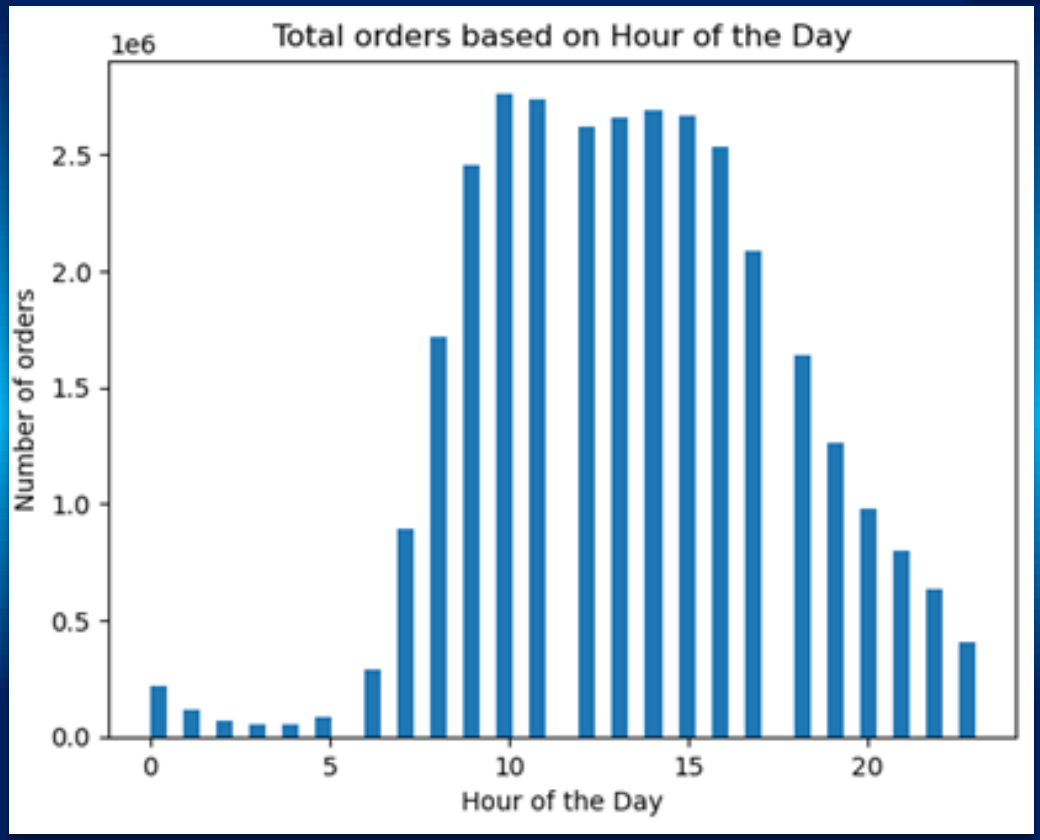
Column derivations and aggregations

Dataset	New column	Column/s it was derived from	Conditions
ord_prod_combined	price_label	prices	Low-range: <=5 Mid-range: > 5, <= 15 High-range: > 15
ord_prod_combined	busiest_day	orders_day_of_week	Saturday: Most busy Wednesday: Least busy Others: Regular busy
ord_prod_combined	two_busiest_days	orders_day_of_week	Sat & Sun: Most busy Tue & Wed: Least busy Others: Regular busy
ord_prod_combined	busiest_period_of_day	order_hour_of_day	most_orders = [10,11,14,15,13,12,16,9] least_orders = [3,4,2,5,1,0,6]
ord_prod_combined	max_order	user_id & order_number	max order number from each user_id
ord_prod_combined	loyalty_flag	max_order	Loyal: > 40 max order Regular: > 10, <= 40 New: <= 10
ord_prod_combined	avg_spend	user_id & prices	average prices per user_id
ord_prod_combined	spend_flag	avg_spend	High spender: >=10 Low spender: <10
ord_prod_combined	median_order_day	user_id & days_since_prior_order	median of days_since_prior_order per user_id
ord_prod_combined	freq_customer	median_order_day	Freq customer: <=10 Reg customer: >10, <=20 Non freq customer: > 20
ord_prod_combined	region	state	grouping of various states in 4 regions
ord_prod_combined	activity_flag	max_order	High activity: >= 5 Low activity: <5
ord_prod_combined	income_flag	income	Low income: <= 60k Mid income: > 60k, <= 150k High income: > 150k
ord_prod_combined	sum_per_order	prices & order_id	total spending per order_id

New columns were added to the final dataset to create new categories of data, for the purpose of better analysis.

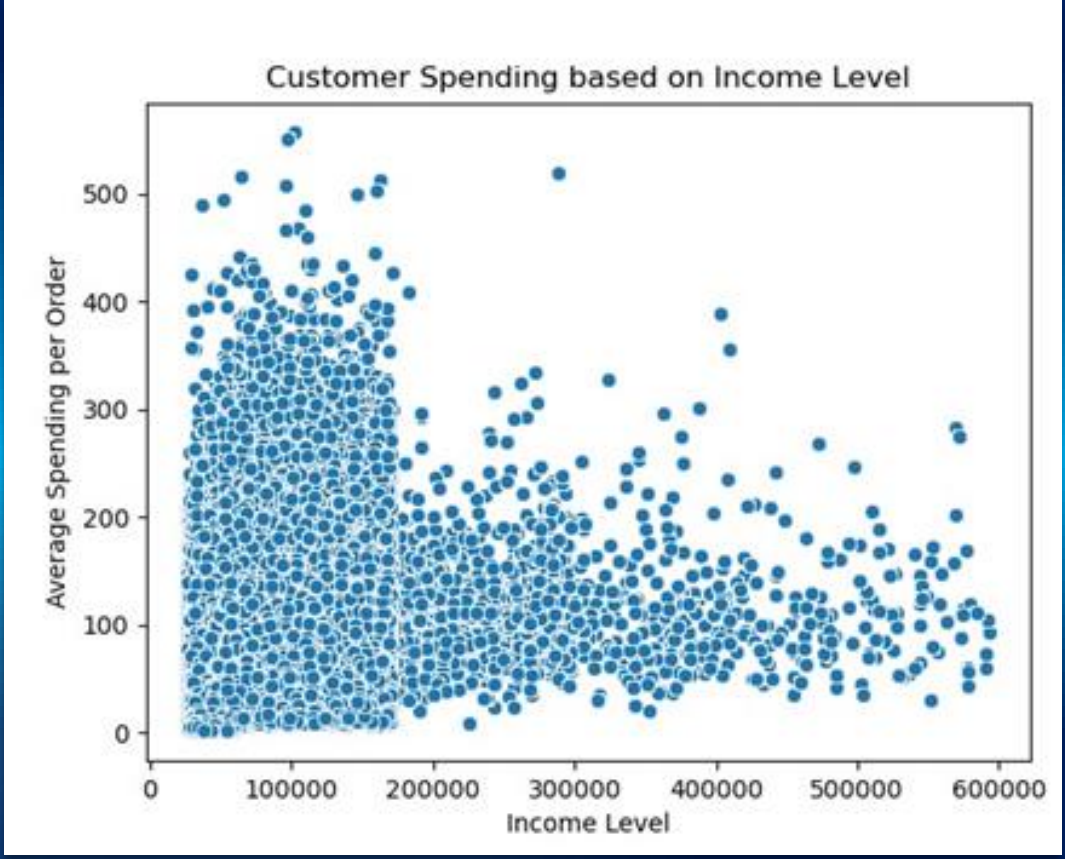
Customer Order Analysis

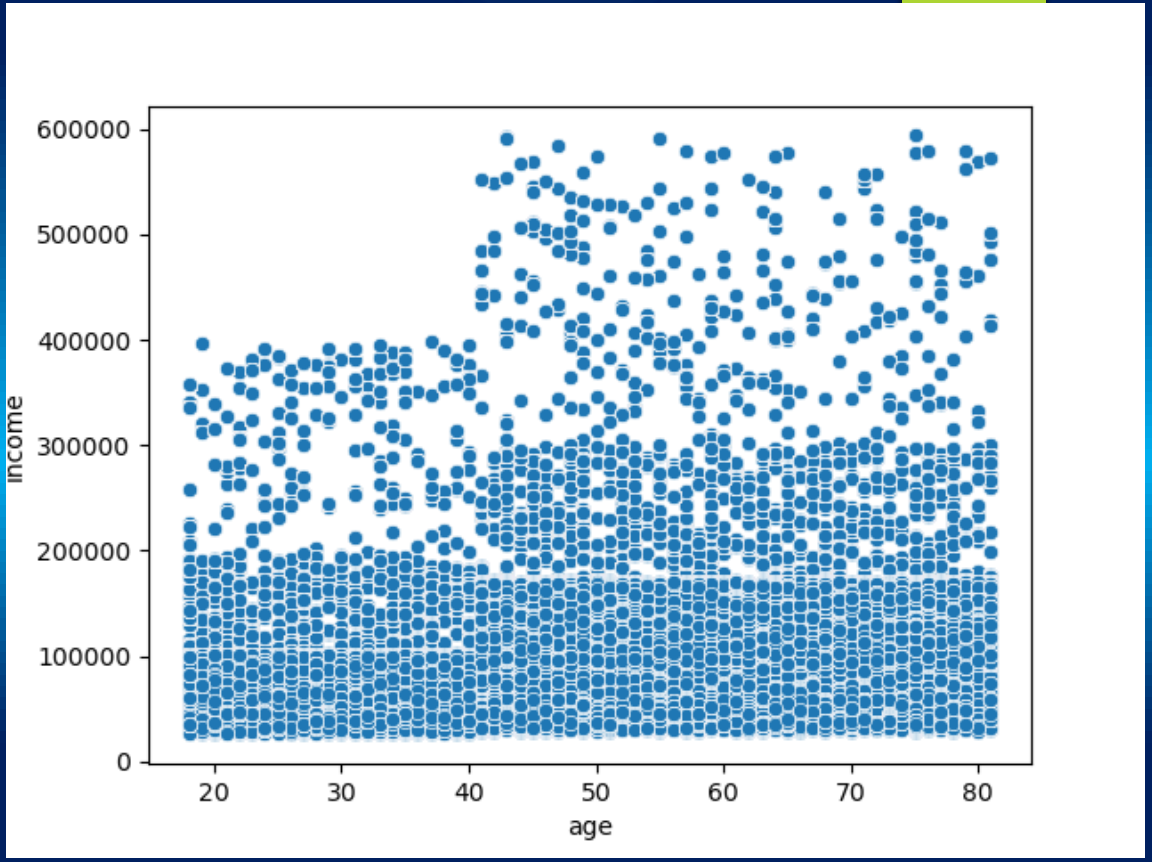




Left bar chart shows that Weekends are the busiest order of the week, and chart on the right tells the busiest order hours are from 9AM to 6PM. This information will allow Instacart to strategize right timing for ad campaign, and prepare better for inventory management.

Customer Profile Analysis





Scatter plot on the left shows that most orders come from customers with income of less than 200k. Plot on the right describes that there are significant jump of income for some customers over age of 40. This data will allow Instacart to target older customers with higher income, who may not spend as much as the younger ones.

Recommendations

- Sales team may want to run more ads on the weekdays between 6PM to 6AM where orders are the slowest.
- Most sales are from Mid-range price products. Marketing may want to advertise more on Low and High-range products.
- "Marketing should focus to advertise to customers over 40 because they have significantly higher income. But also maintain relationship with customers with income under $200k because they are the largest demographic.
- Sales team might try to find more products that will be suitable for people with higher income."

5. Boat Sales Analysis

Objective

You are working as a data analyst for a yacht and boat sales website. The marketing team is preparing a weekly newsletter for boat owners. The newsletter is designed to help sellers to get more views of their boat, as well as stay on top of market trends.

Skills

- **Python**
- **Linear Regression** with sklearn library
- **Clustering** with Kmeans library
- **Time Series** with statmodels library
- Tableau Dashboard

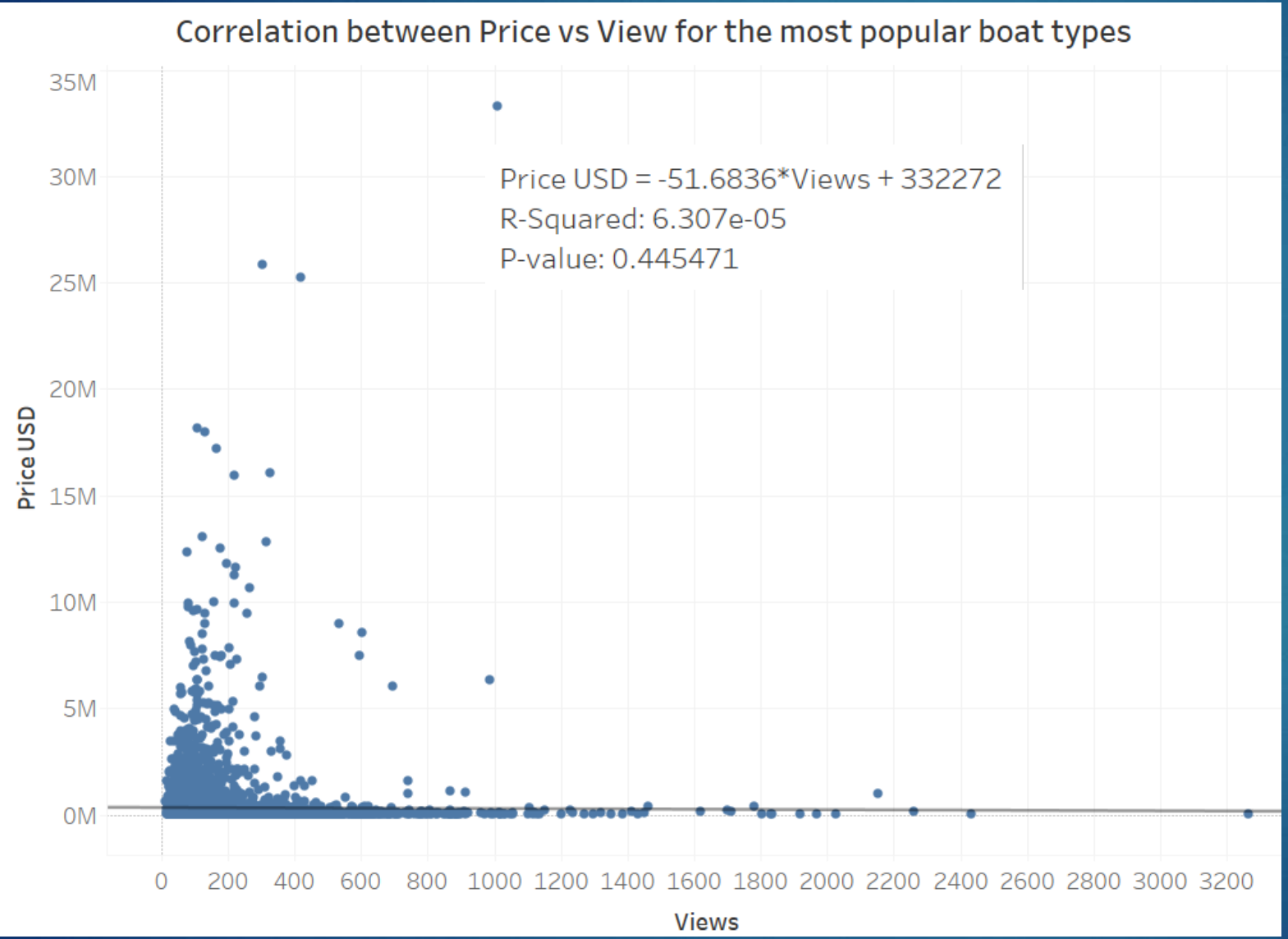
Key questions

- Which regions get most of user views?
- Is there a particular relationship between boat price and views?
- Are there common features among the most viewed boats?

Links

Project Brief
Dataset
Python Scripts
Tableau Dashboard

Linear Regression Analysis



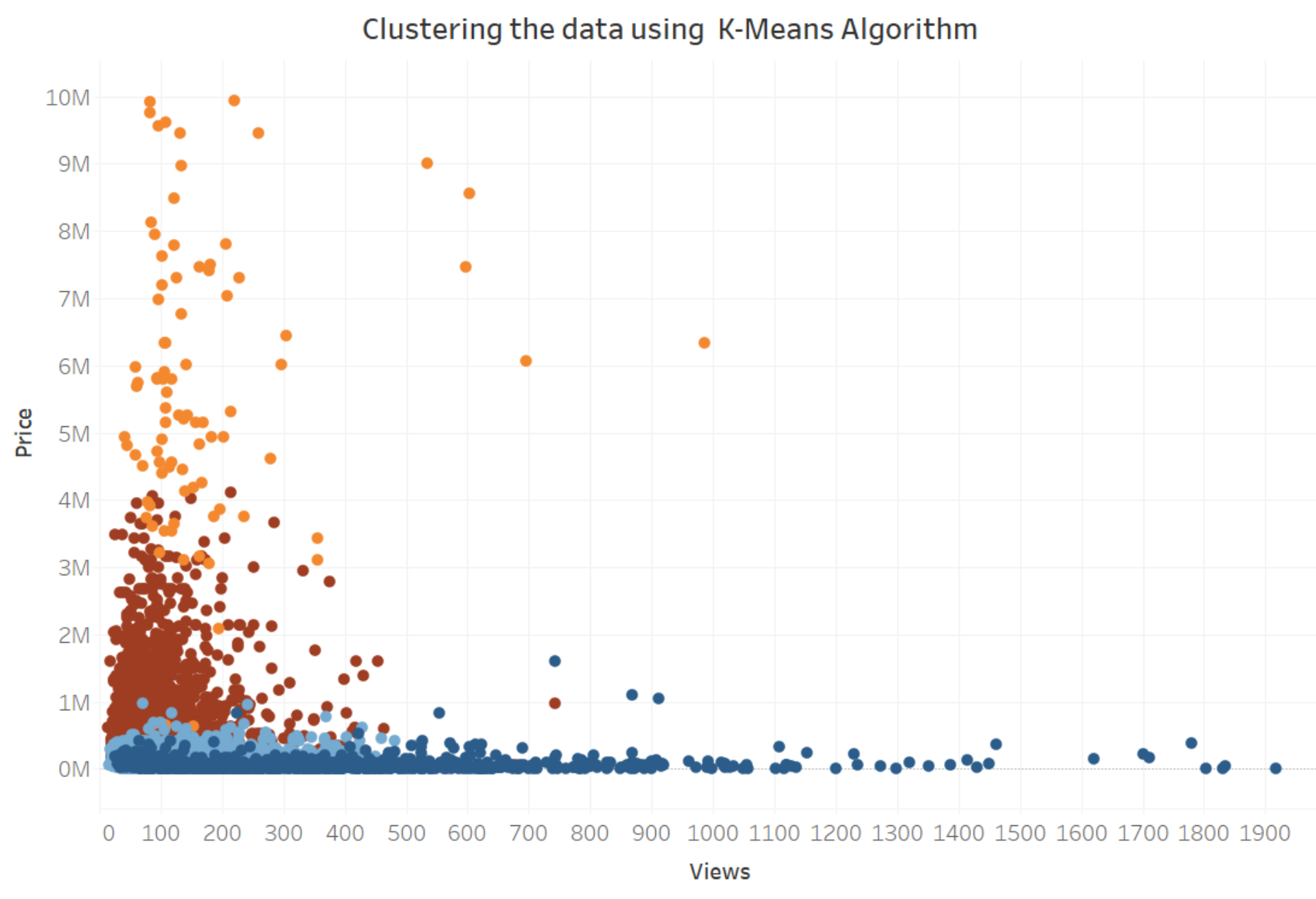
Observation

In this analysis the **p-value is 0.445471** which is much larger than 0.05, therefore the **Null Hypothesis cannot be rejected**. The null hypothesis is often stated as the assumption that there is no change, no difference between two groups, or no relationship between two variables.

Conclusion

This linear regression analysis cannot prove any correlation between Boat Price and Boat View. Next we will perform Clustering Analysis to see any relationship in smaller sets of data.

Clustering Analysis

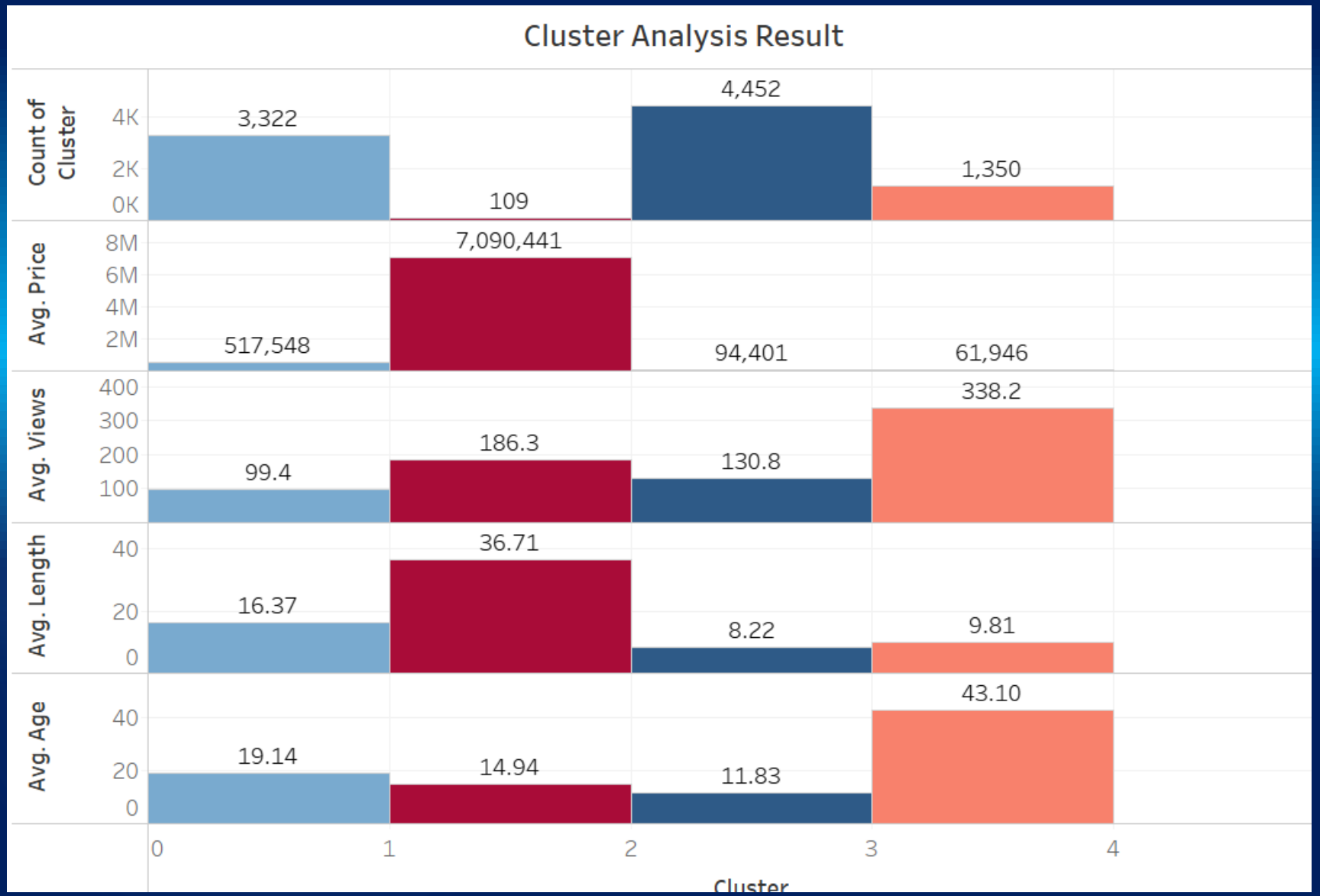


Observation

Using K-Means the dataset is divided into 4 clusters.

1. The most expensive cluster (**Orange**) are new, large and luxury boats such as Mega Yatch, Flybridge, and Catamaran.
2. The medium price cluster (**Dark Red & Light Blue**) are ranging from small to medium size boats, and they are not very popular.
3. The least expensive cluster (**Dark Blue**) are small and older boats. These are the most popular boats.

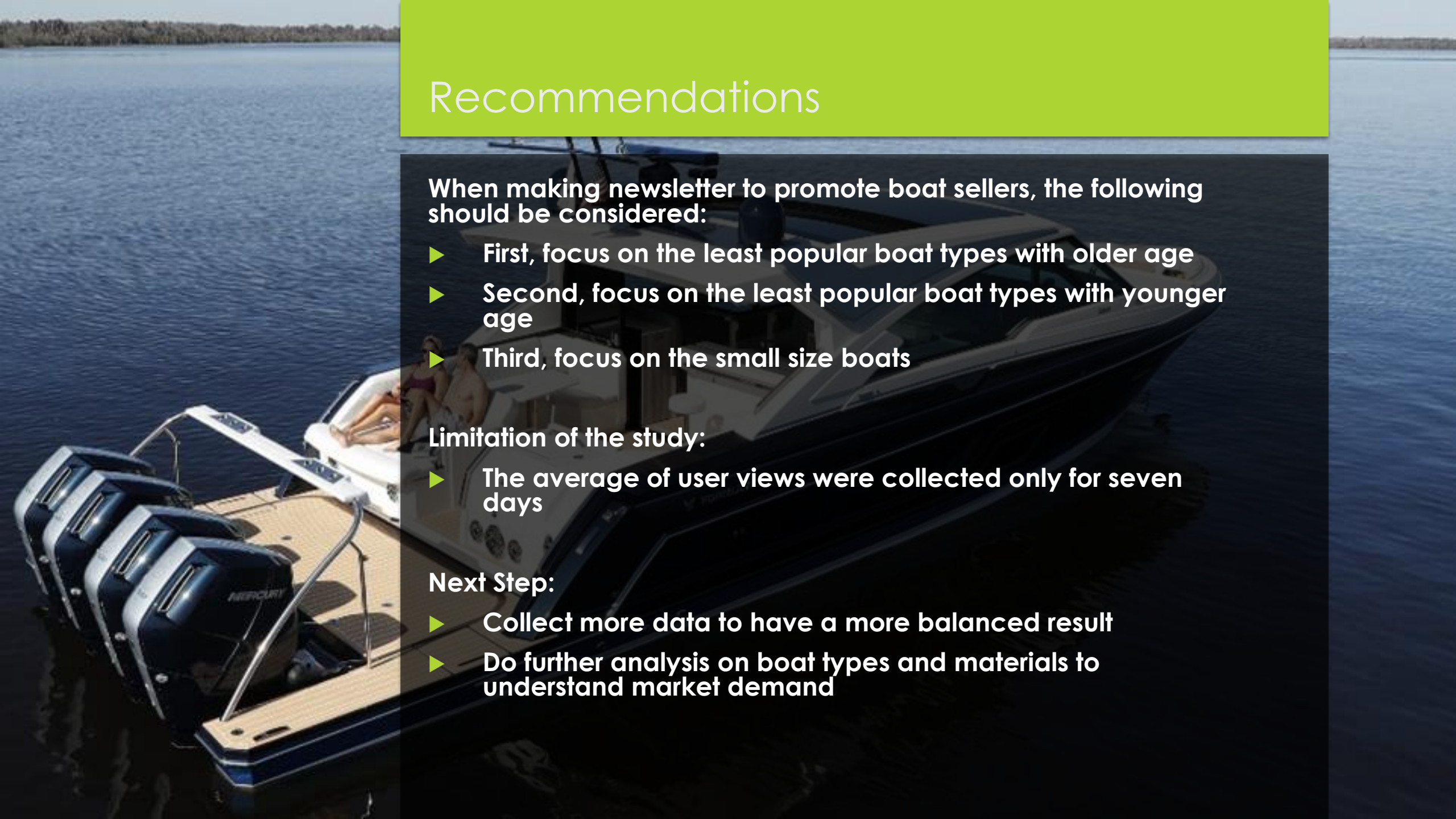
Clustering Analysis Result



Observations

1. The most popular boats (**Orange**) are small boats with lowest price. People in this category are seriously looking to buy small boats probably for leisure purpose.
2. The second most popular boats (**Red**) are very expensive and large boats. People who look these kind of boats are probably just browsing around to see the latest edition of luxury boats.
3. The least popular boats (**Dark and Light Blue**) are the largest portion of the boat market with a combination of small and medium size boats.

Recommendations



When making newsletter to promote boat sellers, the following should be considered:

- First, focus on the least popular boat types with older age
- Second, focus on the least popular boat types with younger age
- Third, focus on the small size boats

Limitation of the study:

- The average of user views were collected only for seven days

Next Step:

- Collect more data to have a more balanced result
- Do further analysis on boat types and materials to understand market demand

Thank you!

Contact

tsoewignjo@yahoo.com
917-881-8360
LinkedIn
Ig: @tonysoewignjo