Stephanie Ugwuanya

Data Analytics Portfolio

Introduction

“With extensive experience in the world of finance I know that cashflow and profit are extremely important to any business, however, I have always been interested to know more about the components that drive this. Data is the future! I have thoroughly enjoyed embarking on these projects and bringing the numbers to life.”

- Stephanie Ugwuanya

Projects

01. Games Co

X P

02. Influenza Season

X P

tableau

03. Rockbuster Stealth

PostgreSQL P

tableau

04. InstaCart

CO jupyter

P

05. Pig E Bank

CO X

01. Games Co

Project Objective

GamesCo is a games company that wants to develop a new game.
They require indepth analysis on their customers preferences across regions to gain a better understanding on how their new game might fare in the market.

Data Sets

The dataset covered historical sales of video games (for games that sold more than 10,000 copies) spanning different platforms, genres, and publishing studios

It tracks the total number of units of games sold between 1980 to 2016.

Key Question/ Hypothesis

Hypothesis: 'Sales for regions have stayed the same over time'

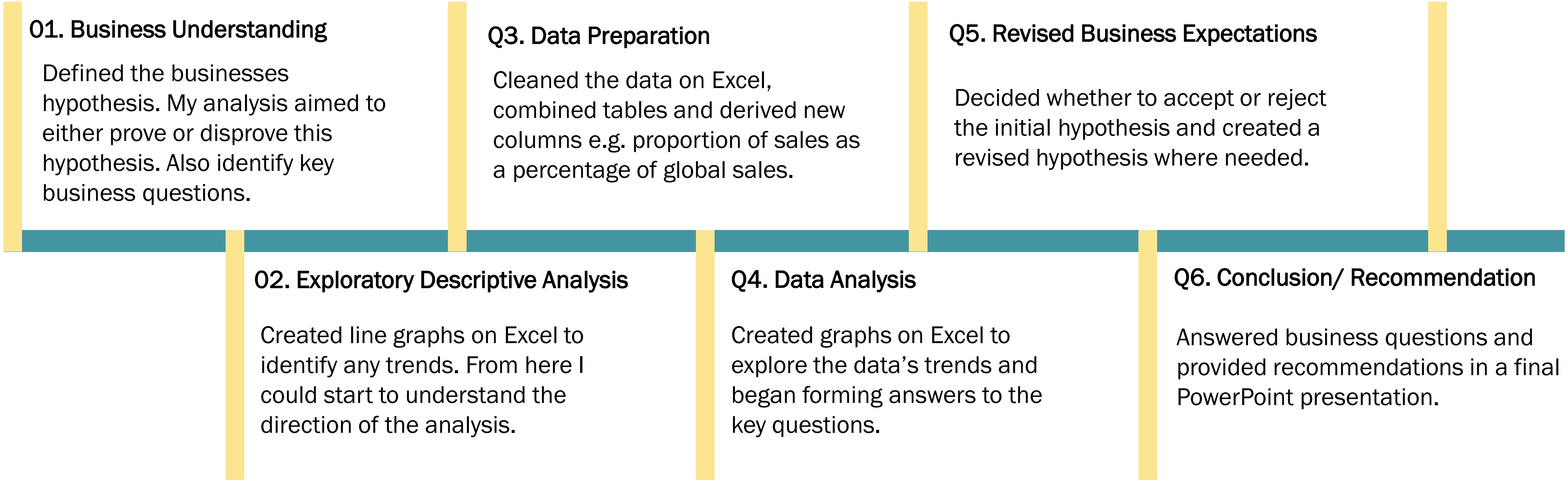
- Are certain types of games more popular than others?
- What other publishers will likely be the main competitors in certain markets?
- Have any games decreased or increased in popularity over time?
- How have their sales figures varied between geographic regions over time?

Limitations

The dataset provided the number of games sold and not the financial figures e.g. profits generated. This would have been an interesting metric to explore and could have helped further answer the key questions.

Tools Used

Key Steps



01. Business Understanding

Defined the businesses hypothesis. My analysis aimed to either prove or disprove this hypothesis. Also identify key business questions.

02. Exploratory Descriptive Analysis

Created line graphs on Excel to identify any trends. From here I could start to understand the direction of the analysis.

Q3. Data Preparation

Cleaned the data on Excel, combined tables and derived new columns e.g. proportion of sales as a percentage of global sales.

Q4. Data Analysis

Created graphs on Excel to explore the data's trends and began forming answers to the key questions.

Q5. Revised Business Expectations

Decided whether to accept or reject the initial hypothesis and created a revised hypothesis where needed.

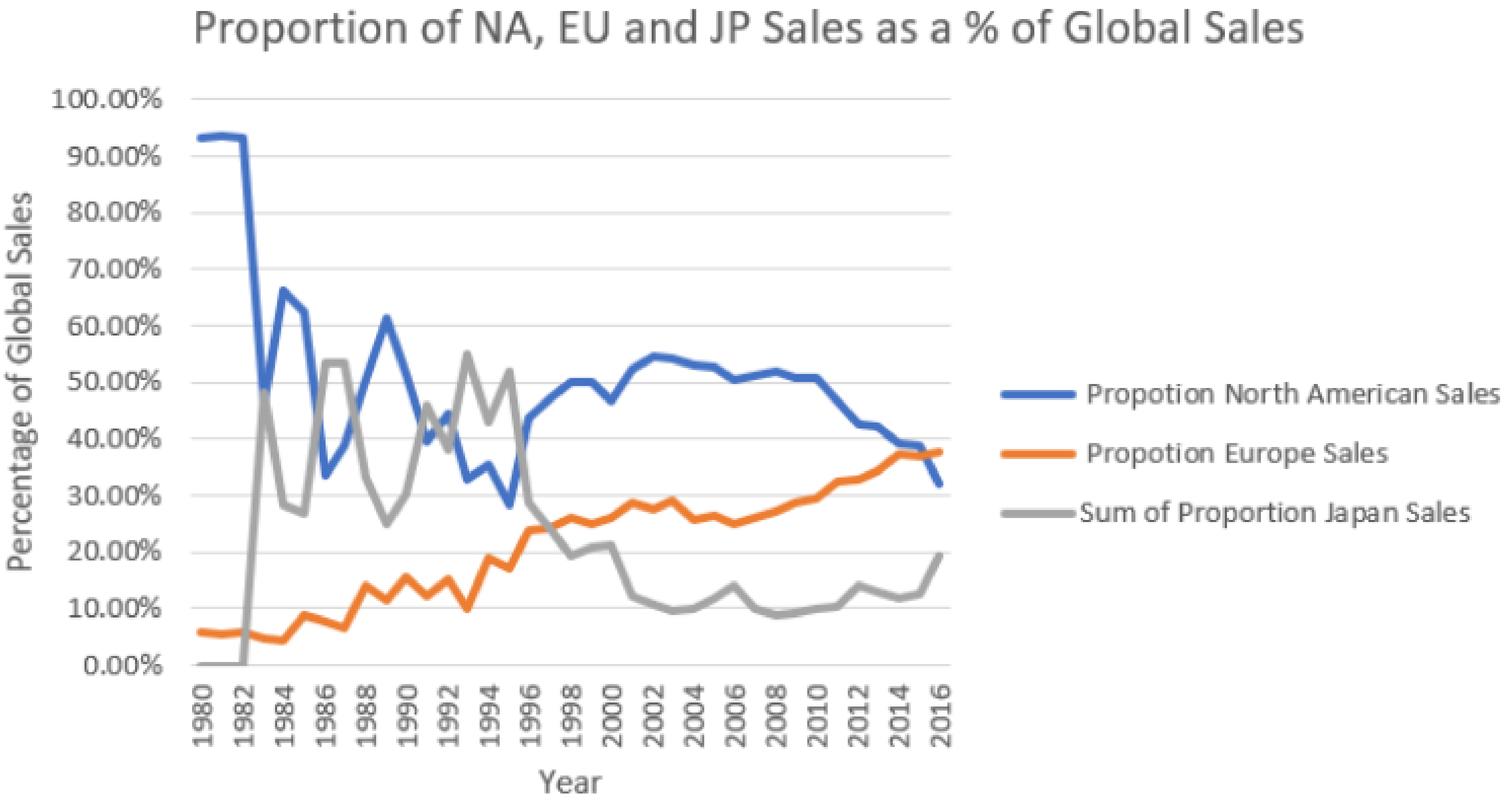
Q6. Conclusion/ Recommendation

Answered business questions and provided recommendations in a final PowerPoint presentation.

01. GamesCo Analysis

This graph shows that between 1998 and 2008 North American Sales were consistently the largest contributor to Global Sales at around 50%.

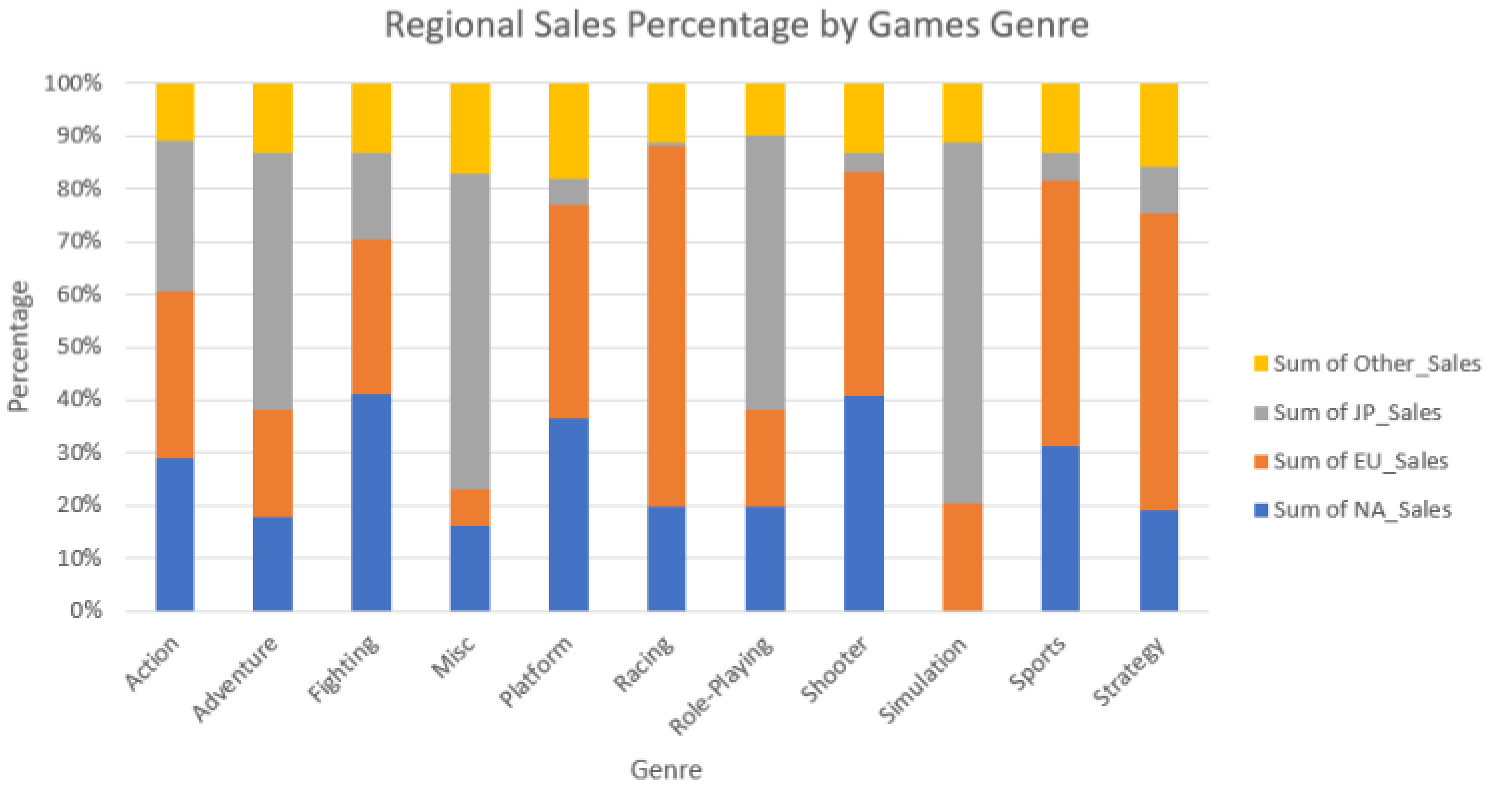
European Sales contributions have been consistently rising throughout the period surpassing North American Sales in 2016.



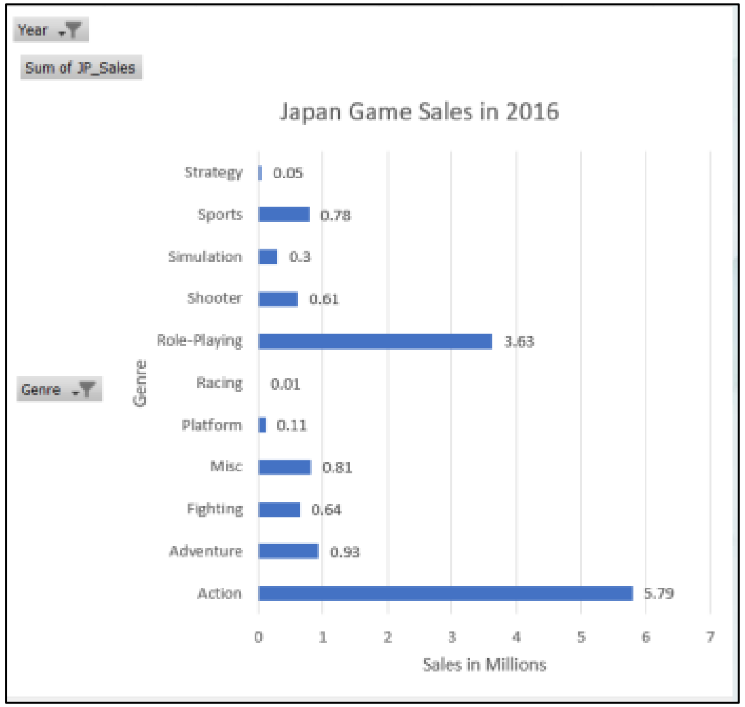
01. Game Co Analysis

We can see that for most genres, North America and Europe generated the most sales in terms of percentage of total sales per genre.

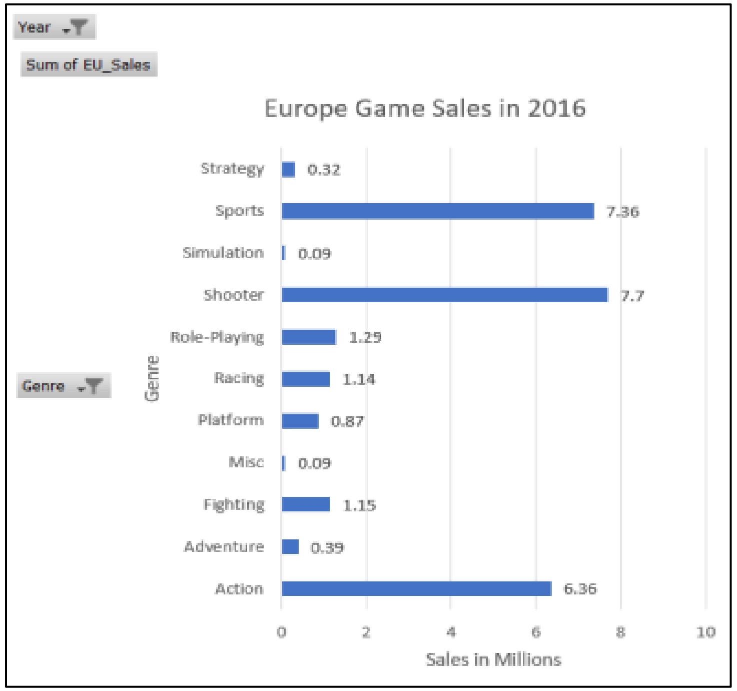
However, the roleplaying game genres is a huge market in Japan.



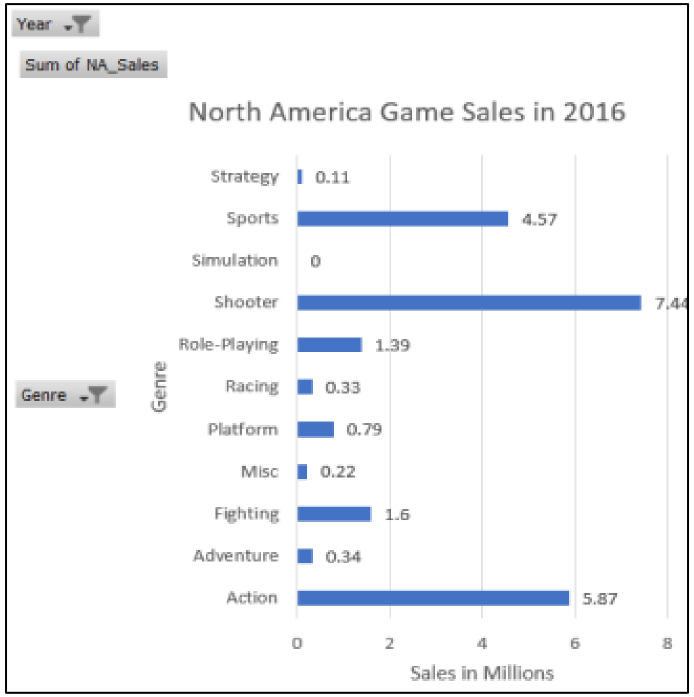
01. Game Co Analysis - Regional Sum of Sales in 2016



Action games are most popular in Japan at 5.79 million sales , however Role-Playing games are most popular in Japan compared to the other regions at 3.63 million.



Shooter games and Action games are most popular in Europe.



Shooter games are most popular in North America at 7.44 million sales.

01. Game Co Analysis - Recommendations

Revised Hypothesis:

- *Sales for regions differ greatly over time, with some regions contributing a much larger contribution to Global Sales than others.*
- *The popularity of certain genres varies depending on the region.*

Recommendations

- A larger proportion of the marketing budget should go towards the Europe market, as we have seen that the Europe locations have been contributing an increasing percentage to global sales.
- We should try and stimulate the North American market. In recent years it has been declining and it has in the past been a huge contributor to overall global sales.
- The Japanese market is falling and has been contributing a lower percentage to global sales compared to the other locations for some years, however they have a huge Role-Playing game market that could be utilised to stimulate more sales.

02. Influenza Season

Project 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. My task was to determine the temporary staffing levels needed across the USA during the influenza season.

Data Sets

- **Influenza deaths by geography** Source: CDC
- **Population data by geography, time, age, and gender** Source: US Census Bureau
- **Counts of influenza laboratory test results by state (survey)** Source: CDC (Fluview)
- **Survey of flu shot rates in children** Source: CDC

Key Question/ Hypothesis

Hypothesis:
'States with a higher proportion of a vulnerable population (Over the age of 65) have the highest death rates.'

When do we need to send out temporary staff to hospitals?
How many temporary staff members do we need out to hospitals?
Which states need the temporary staff at their hospitals?

Limitations

The staffing agency has a limited number of nurses, physician assistants, and doctors on staff.

There's no money to hire additional medical personnel.

Tools Used

Key Steps

01 Business Understanding

In this project I decided upon the hypothesis to give my analysis direction.

'States with a higher proportion of a vulnerable population have the highest death rates'.

A project management plan was also created.

02. Data Understanding

Data sources were assessed for their legitimacy, and research was done on how the data was collected.

03. Data Preparation

Data cleaning was conducted on Excel and tables were combined to create a final dataset.

Additionally new columns were derived to help with the analysis,

04. Exploratory Descriptive Analysis

EDA was conducted on Tableau to identify any initial trends.

05. Data Analysis

Graphs were created on Tableau to answer the business questions.

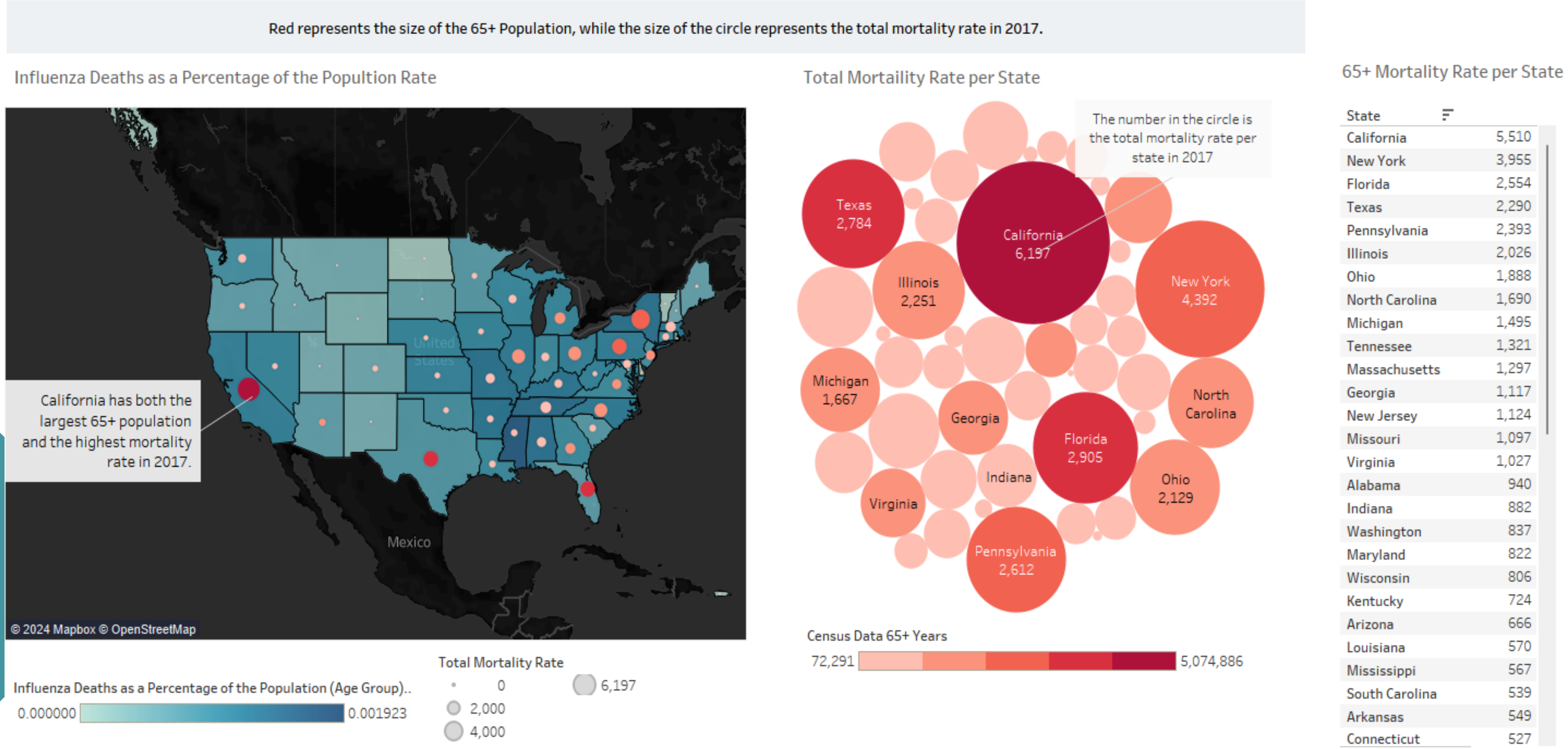
06. Revised Business Expectations

From the analysis conducted I decided whether to accept or reject the hypothesis and created a revised hypothesis if needed.

Q7. Conclusion/ Recommendation

A story board was created on Tableau to display results and conclusions along with a verbal presentation created on Vimeo to explain my findings.

Influenza Mortality by State 2017 Data



Red represents the size of the 65+ Population, while the size of the circle represents the total mortality rate in 2017.

Influenza Deaths as a Percentage of the Popultion Rate

California has both the largest 65+ population and the highest mortality rate in 2017.

United States
Mexico
© 2024 Mapbox © OpenStreetMap
Influenza Deaths as a Percentage of the Population (Age Group)..
0.000000
0.001923
Total Mortality Rate
0
2,000
4,000
6,197

Total Mortaility Rate per State

The number in the circle is the total mortality rate per state in 2017

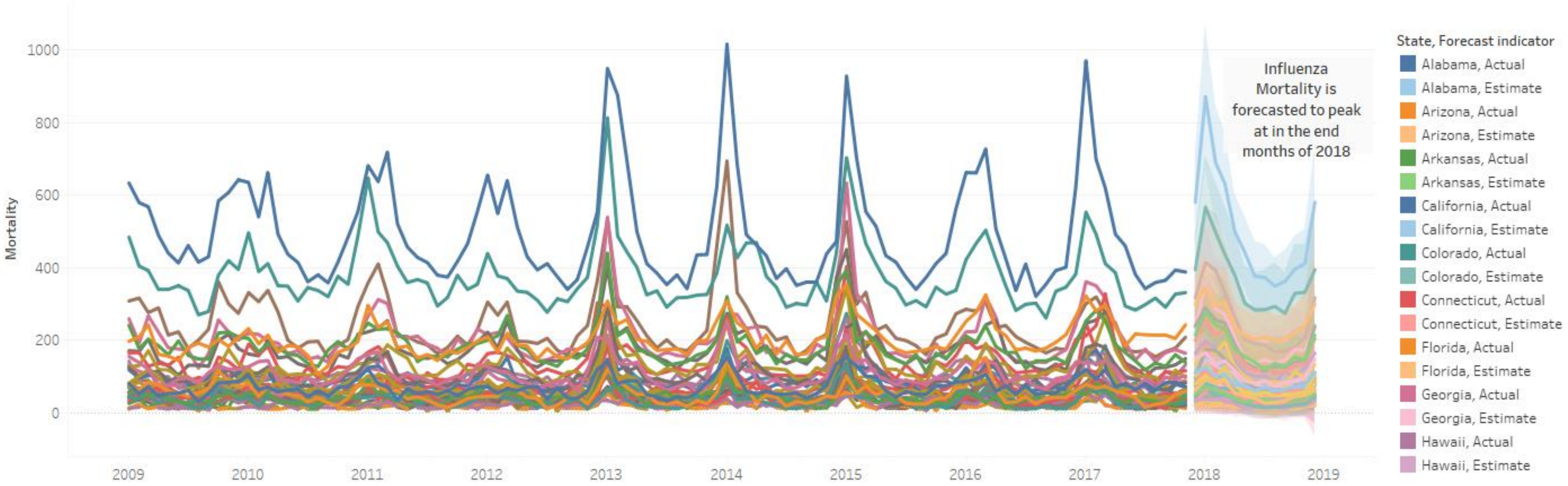
65+ Mortality Rate per State

State	
California	5,510
New York	3,955
Florida	2,554
Texas	2,290
Pennsylvania	2,393
Illinois	2,026
Ohio	1,888
North Carolina	1,690
Michigan	1,495
Tennessee	1,321
Massachusetts	1,297
Georgia	1,117
New Jersey	1,124
Missouri	1,097
Virginia	1,027
Alabama	940
Indiana	882
Washington	837
Maryland	822
Wisconsin	806
Kentucky	724
Arizona	666
Louisiana	570
Mississippi	567
South Carolina	539
Arkansas	549
Connecticut	527

The top 5 states with the highest influenza mortality rate in 2017 were: California, New York, Florida, Texas and Pennsylvania . We can see on the map and bubble chart that these states also had the largest 65+ populations compared to other states.

We can also see from the blue shading on the map that these states are among the states with highest influenza deaths as a percentage of the population in 2017.

Mortality Rate Predictions 2018



Influenza Mortality is forecasted to peak at in the end months of 2018

This forecast predicted mortality rates in 2018. We can see from the forecast that mortality rates were predicted to follow the same pattern as previous years. The mortality rate was set to rise at the end of the year (Oct-Dec) – winter months.

02. Influenza Season - Recommendations

Recommendations

Firstly, there should be a focus on supplying additional staff to states with historically high death rates - California, New York, Florida, Texas, Pennsylvania, Illinois, Ohio, North Carolina, Michigan and Tennessee.

Secondly, there should also be a focus on supplying additional staff to states with a high 65 + populations. As we found from our analysis these states often have the highest death rates. The only state with a high 65+ population that was not included in the highest death rates in 2017 list was New Jersey.

We recommend this state is also included in the priority list as we know the link between a high 65+ population and a high mortality rate.

Next Steps

Influenza Vaccination rollouts to the 65+ population should be prioritised before the winter months (Oct-Dec) to effectively reduce possible influenza mortality in this age range.

Further investigation should be conducted on New Jersey in terms of factors that have contributed to a high 65+ population but a lower mortality rate than other states in its category.

Hypothesis Accepted

States with a higher proportion of a vulnerable population have the highest death rates.

03. Rockbuster Stealth

Project 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 Rockbuster Stealth management team is planning to use its existing movie licenses to launch an online video rental service in order to stay competitive. My analysis was to help decide upon the best strategy to tackle the online movie market.

Data Sets

The dataset was provided by Rockbuster Steath and contains data on film and customer information e.g.:

film inventory, customers, and payments.

Key Question/ Hypothesis

- 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

PostgreSQL

tableau

Key Steps

01 Business Understanding

This project was focused on answering the business questions. These business questions shaped the direction of my analysis.

02. Data Understanding

The Entity Relationship Diagram (ERD) was reviewed on PostGreSQL and a Data Dictionary was created.

03. Data Preparation

Data cleaning was conducting using SQL e.g. searching for any null values or non-uniform column names.

04. Exploratory Descriptive Analysis

EDA was conducted using SQL looking at the mean median, mode, max and min of each column to see any initial trends.

05. Data Analysis

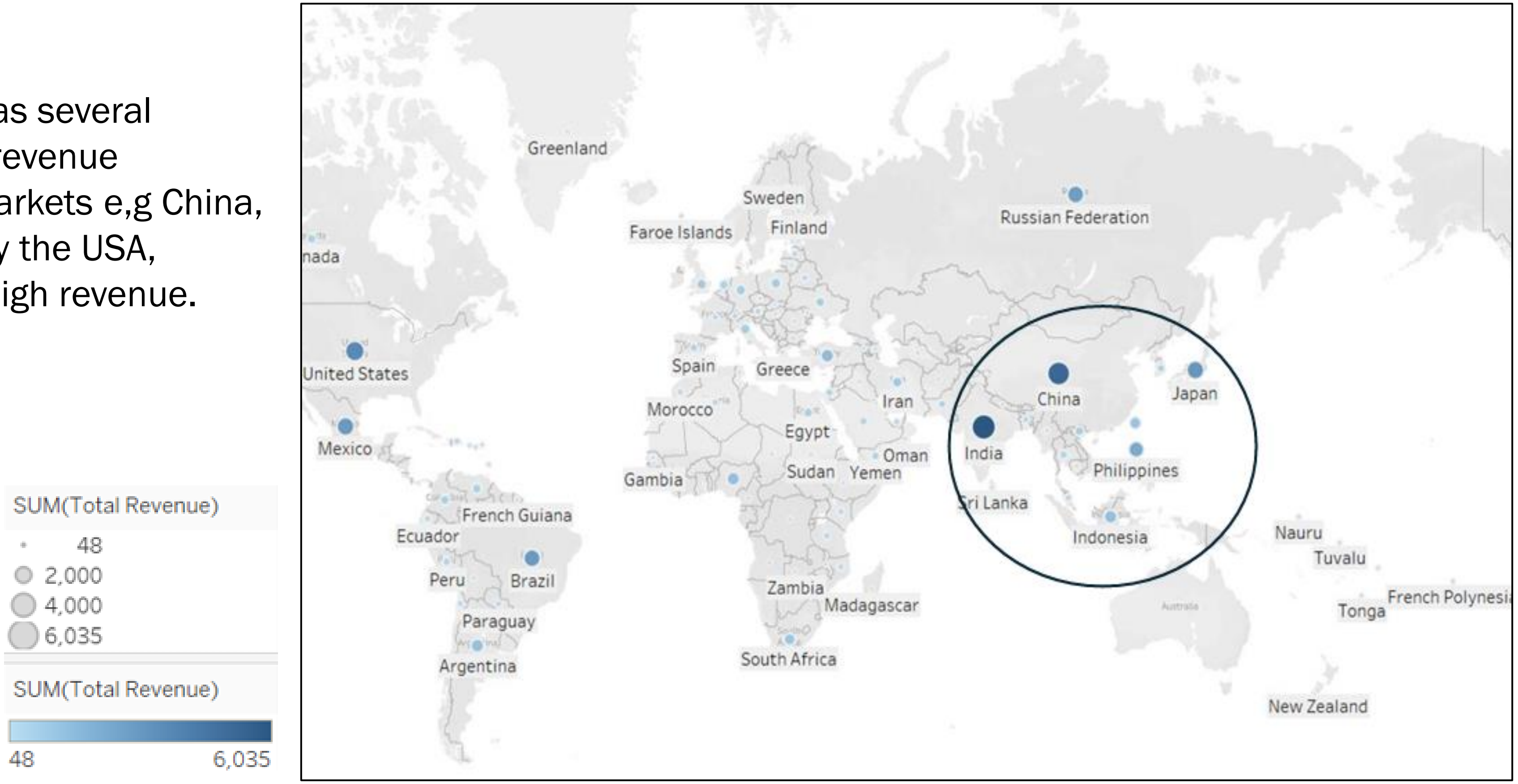
Analysis was conducted using a combination on SQL code and Tableau Visualisations.

Q6. Conclusion/ Recommendation

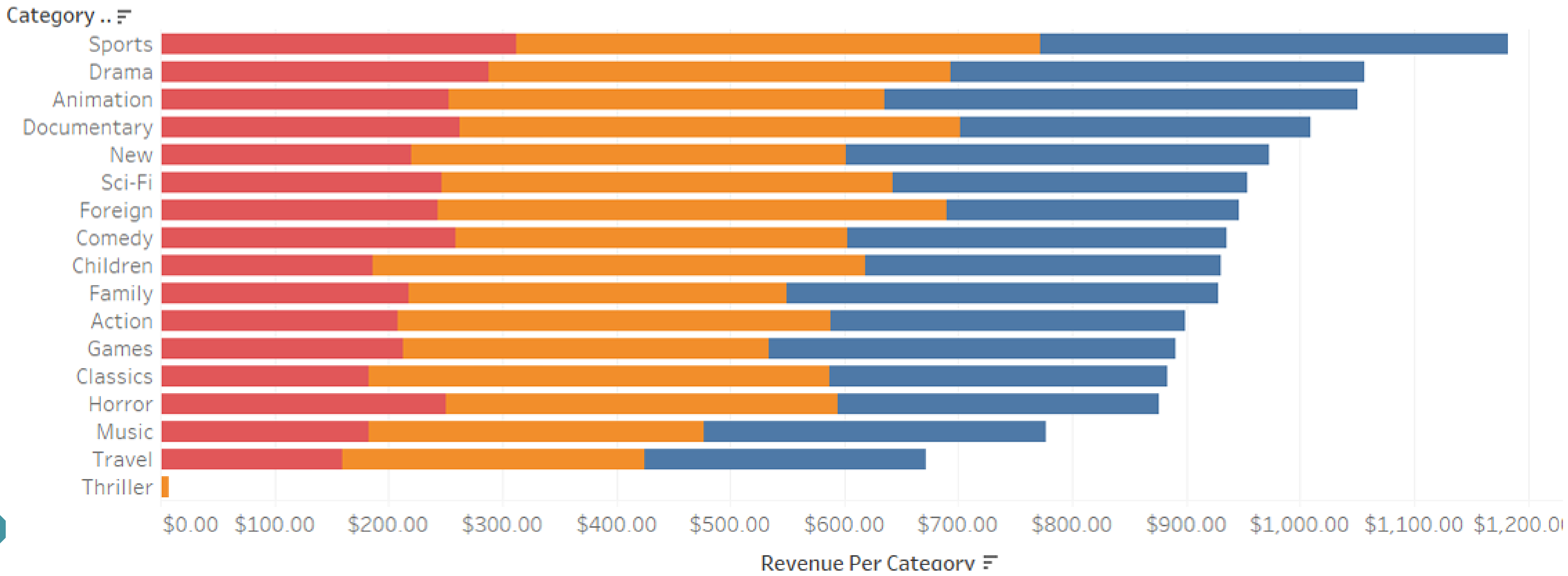
A PowerPoint presentation was created to display the results and recommendations. All business questions were answered.

Total Revenue Per Country

We can see that the Asian market has several countries that have generated high revenue compared to the rest of the world markets e,g China, Philippines, India and Japan. Notably the USA, Mexico and Brazil also contributed high revenue.



Revenue by Genre China, India and USA

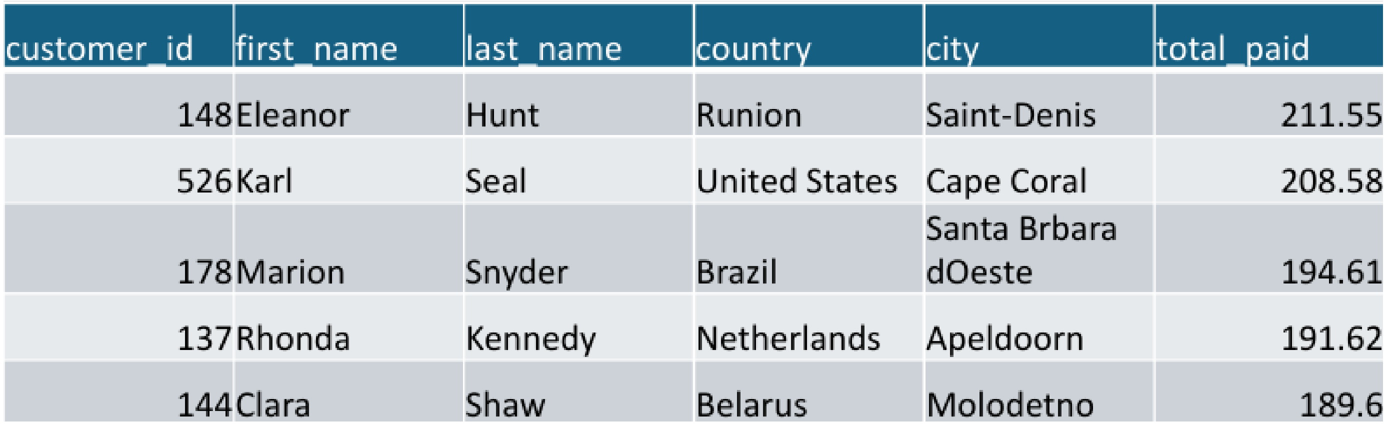


China, India and USA were identified as the top 3 countries in terms of revenue.

- Thriller movies contribute significantly less to the revenue.
- Popularity among other genres are well balanced across China, India and USA.
- Sports movies are the most popular movies among the top 3 countries when looking at the total revenue generated.

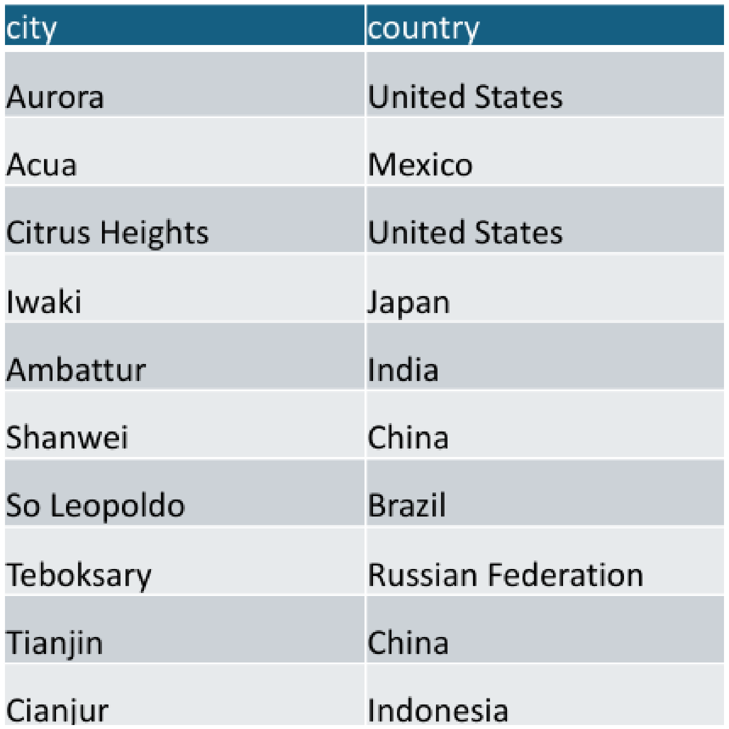
Customers with a high lifetime value

Top 5 customers from the top 10 cities who’ve paid the highest



customer_id	first_name	last_name	country	city	total_paid
148	Eleanor	Hunt	Runion	Saint-Denis	211.55
526	Karl	Seal	United States	Cape Coral	208.58
178	Marion	Snyder	Brazil	Santa Brbara dOeste	194.61
137	Rhonda	Kennedy	Netherlands	Apeldoorn	191.62
144	Clara	Shaw	Belarus	Molodetno	189.6

Top 10 cities that fall within the top 10 countries



city	country
Aurora	United States
Acua	Mexico
Citrus Heights	United States
Iwaki	Japan
Ambattur	India
Shanwei	China
So Leopoldo	Brazil
Teboksary	Russian Federation
Tianjin	China
Cianjur	Indonesia

These tables were generated using SQL code and look at the customers who have contributed greatly to the overall revenue of the company.

We have also identified the top 10 countries that have contributed the most revenue along with the specific city.

03. Influenza Season - Recommendations

Focus on continued growth in the Asian market. China and India specifically generate high revenue. Growth in these markets can be achieved by ensure there is access to a wide range of movies with the relevant subtitles in both countries.

We should also focus on the American markets (North and South America) Specifically USA, Mexico and Brazil as these countries have generated high revenue. This could be done through increased advertising/ online offers.

We currently have five customers that have contributed greatly to the revenue, they should be rewarded with discounts/vouchers to encourage their continued usage.

The most popular genres in terms of revenue are sports, drama and animation, we should ensure that we have a wide range of movies in these genres to further revenue gains.

The rental duration for thriller movies are higher than any other genre. We will need to investigate this further as thriller movies are not as popular as other genres.

04. InstaCart Bucket

Project Objective

Instacart is 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. My analysis was centred around revealing patters in customer behaviour by focusing on segmenting the data into customer profiles to derive insight.

.

Data Sets

Data provided by Career Foundary:
Customers Dataset

Data provided by Instacart:
Data Dictionary

.

Tools Used

Key Question/ Hypothesis

Sales Team Questions

- The busiest days of the week and hours of the day are (i.e., the days and times with the most orders)
- Times of the day when people spend the most money.
- Instacart has a lot of products with different price tags. Marketing and sales want to use simpler price range groupings to help direct their efforts.

Marketing Team Questions

- What's the distribution among users in regards to their brand loyalty (i.e., how often do they return to Instacart)?
- Are there differences in ordering habits based on a customer's loyalty status?
- Are there differences in ordering habits based on a customer's region?
- Is there a connection between age and family status in terms of ordering habits?
- What different classifications does the demographic information suggest? Age? Income? Certain types of goods? Family status?
- What differences can you find in ordering habits of different customer profiles?

Key Steps

01 Business Understanding

This project was focused of answering the business questions. These business questions shaped the direction on my analysis.

02. Data Understanding

The Data Dictionary was reviewed along with the data types of columns on Python.

03. Data Preparation

Extensive data cleaning was conducted using Python. New columns were derived, flags were created to form custome profiles and datasets merged. All changes were recorded on the final excel sheet.

04. Exploratory Descriptive Analysis

EDA was conducted on all columns using Python.

05. Data Analysis

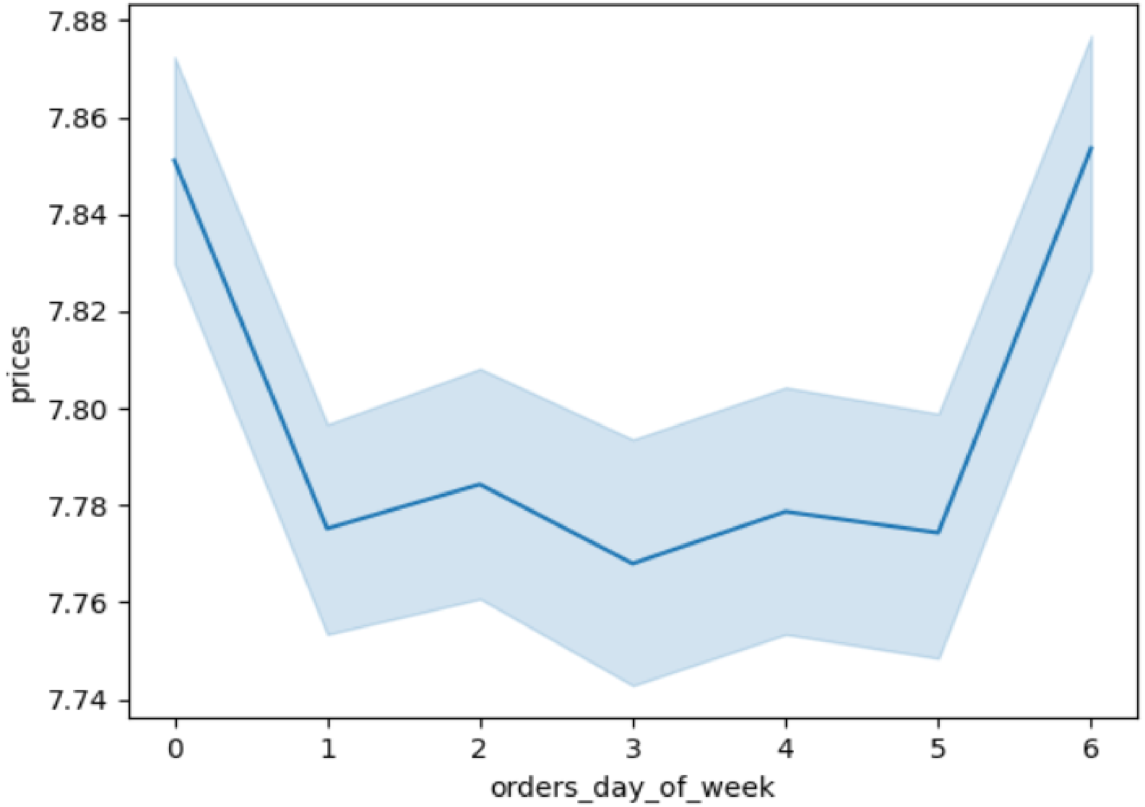
Analysis was conducted using Python. Graphs were also created on python and a final report was created to answer the business questions on Excel.

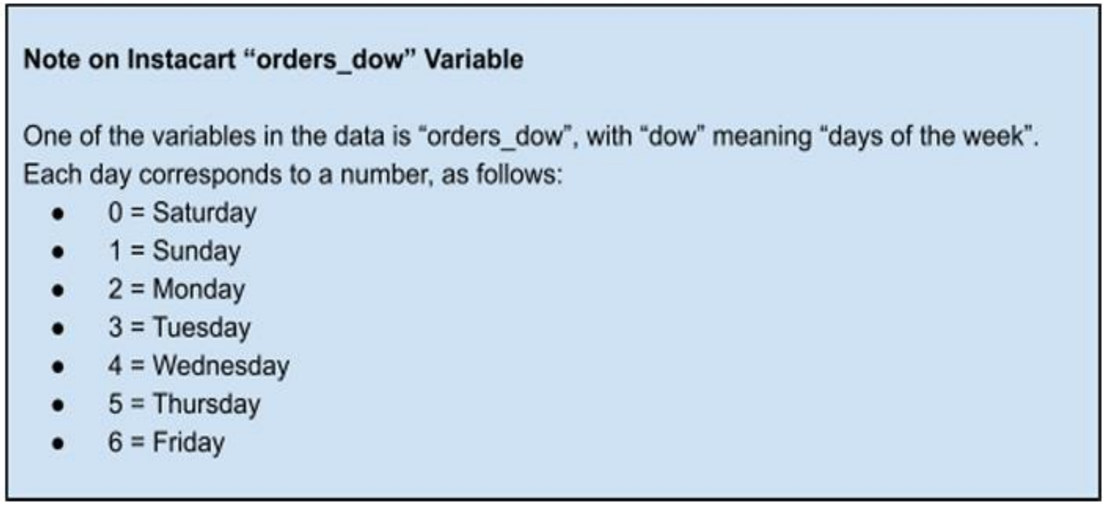
Q6. Conclusion/ Recommendation

All changes to the datasets, recommendation and conclusions were recorded on the final report on Excel.

04. Aggregate Orders by Day of the week

Aggregate Orders by Order Day of the Week





Note on Instacart “orders_dow” Variable

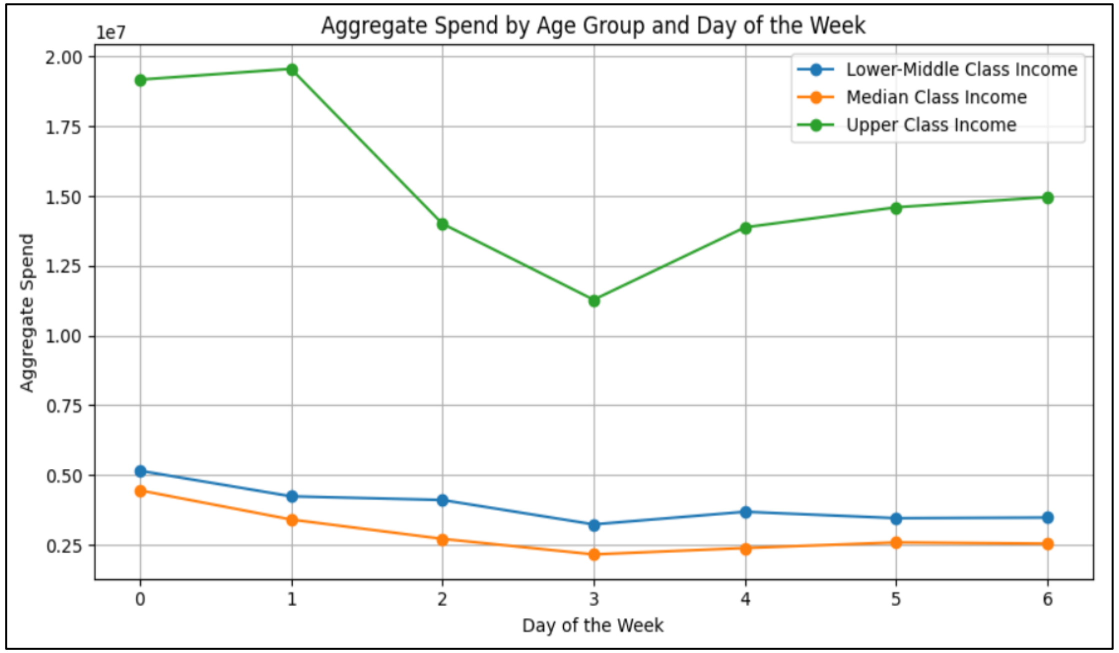
One of the variables in the data is “orders_dow”, with “dow” meaning “days of the week”. Each day corresponds to a number, as follows:

- 0 = Saturday
- 1 = Sunday
- 2 = Monday
- 3 = Tuesday
- 4 = Wednesday
- 5 = Thursday
- 6 = Friday

This line graph shows the overall order totals across the week. We can see that generally there are more orders on Friday and Saturday (0 an 1) ,

04. Aggregate Orders by Day of the week

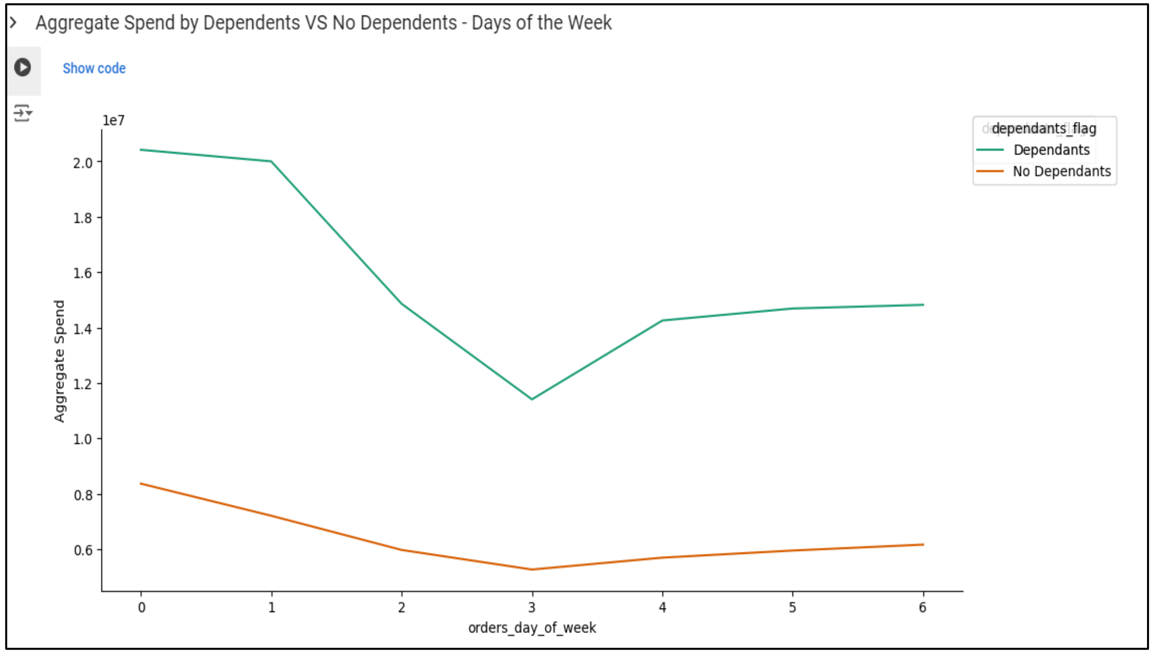
Aggregate Spend Order Days of the Week - Income Group



Among the Upper Class Income earners we can see the most fluctuation in aggregate spend. The most popular day to make orders is Sunday with Saturday following. Another popular day is Friday.

The least popular day among the Upper Class earners to make an orders is Tuesday.

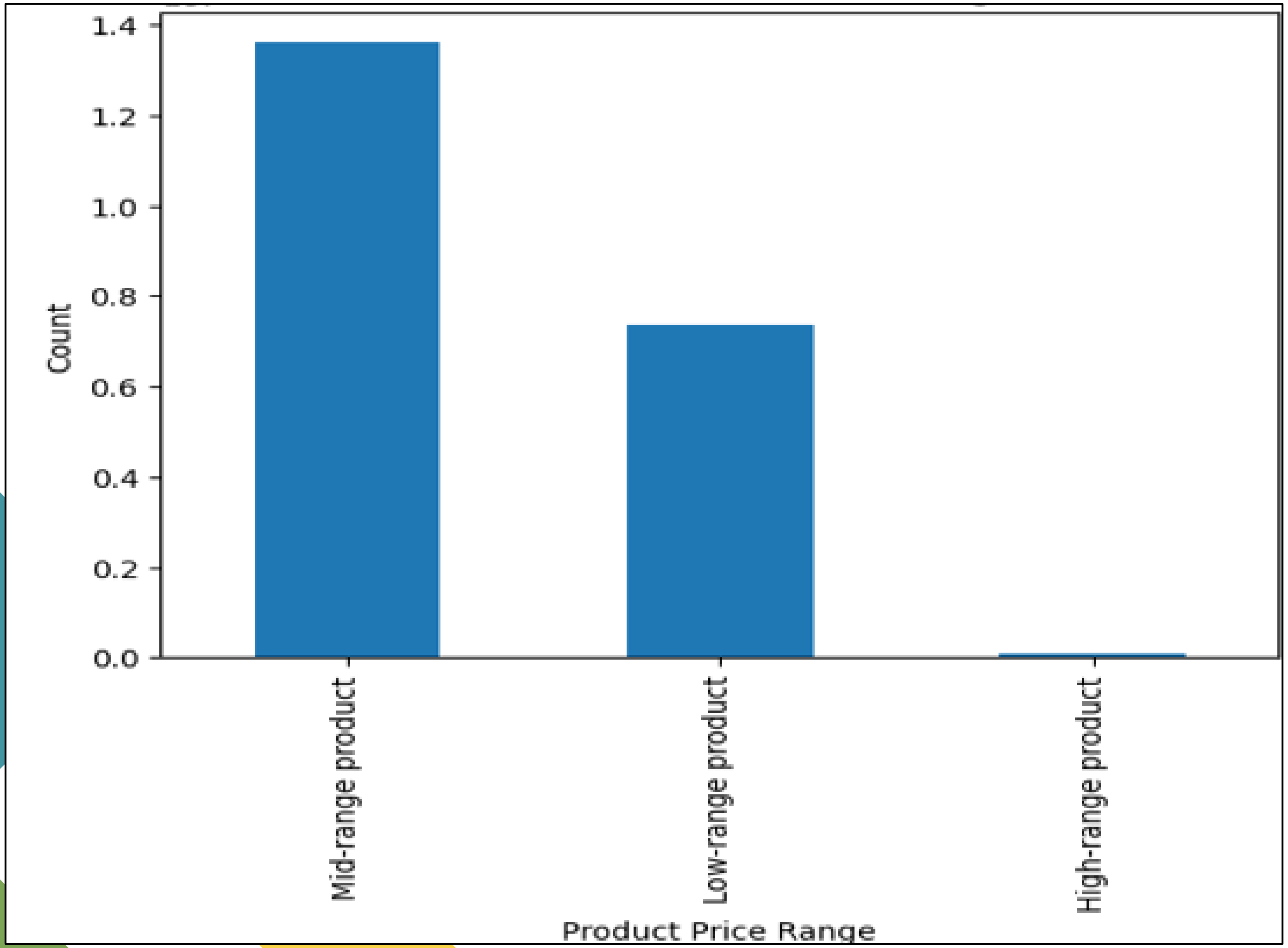
Aggregate Spend Order Days of the Week - Dependents vs No Dependents



We can see a similar pattern in terms of popular order days among customer with dependents/no dependents.

Friday and Saturday are the most popular days to make orders while Tuesday is not a popular order day among both groups.

04. Aggregate Orders by Day of the week



This graph shows the overall count of products in each price range. We can see at InstaCart there a high number of products in the mid-product price range. Conversely, there is a very low count of high-range products.

04. InstaCart Bucket– Recommendations

Question 1 - The sales team needs to know what the busiest days of the week and hours of the day are (i.e., the days and times with the most orders) in order to schedule ads at times when there are fewer orders.

- Ads can be scheduled on Tuesday as this is when the lowest aggregate spend is generated by the largest customer base - the Upper Class income group and those with Dependents . We should also refrain from scheduling ads on Friday Saturday or Sunday as a higher aggregate spend is generated across both customer groups.

Question 2 - They also want to know whether there are particular times of the day when people spend the most money, as this might inform the type of products they advertise at these times.

- The busiest times of the day are between 9-5 which are standard work hours in the USA. We will look more closely at the products that are popular among customer groups below.

Question 3 - Are there certain types of products that are more popular than others?
Notable department popularity among customer profiles:

- Young Aduls and Adults - dairy eggs
- Older Adults - pantry and frozen
- Senior - beverages and snacks
- Dependents - snack and canned food
- No dependents - frozen goods

We should aim resources to encourage further spending in these department across customer profiles by targeted ads and promotions.

Question 4 - Instacart has a lot of products with different price tags. Marketing and sales want to use simpler price range groupings to help direct their efforts.

Across both customer profiles the mid-range products were most popular. Therefore, resources should be used to continue to support this popularity through targeted adverts and offers. It would be beneficial to target resources to the high-range products as boasting popularity among these products could have the biggest effect on profits.

05. Pig E Bank

Project Objective

Pig E is a Global bank interested in the factors that contribute to a customer staying or leaving the bank. This project focused on the data – how data is properly managed which delved into the importance of Data Ethics and the effects on bias that we should minimise as an Analyst. A key feature was the creation of a decision tree to predict the probability of a customer leaving the bank.

Data Sets

Data Set Provided by CareerFoundary.

Tools Used

Key Steps

01 Business Understanding

Understood the main aim of the analysis – to identify the reasons people stay/leave the bank.

02. Data Understanding

Frequency tables were created to give an overview of the data on Excel.

03. Data Preparation

Data clean was conducted using Python e.g. removing duplicated lines and addressing missing values.

04. Exploratory Descriptive Analysis

EDA was conducted using Excel on the numerical column of the dataset.

05. Data Analysis

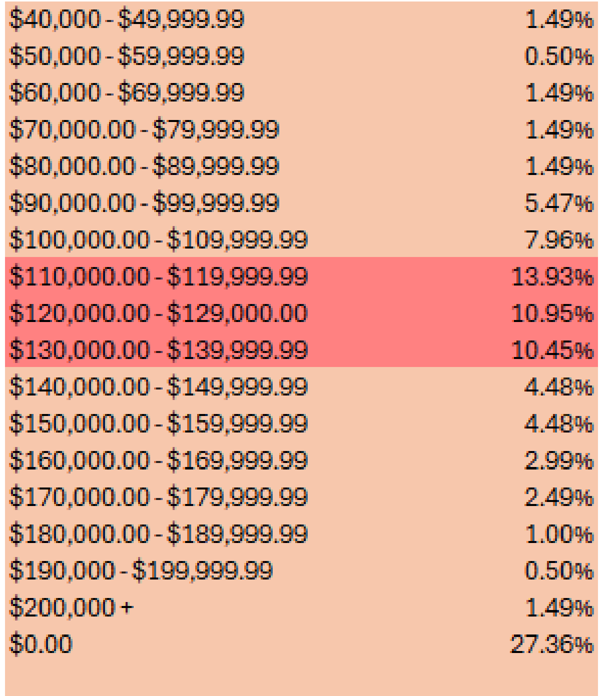
Data analysis was conducted on Excel and a Decision Tree was created.

Q6. Conclusion/ Recommendation

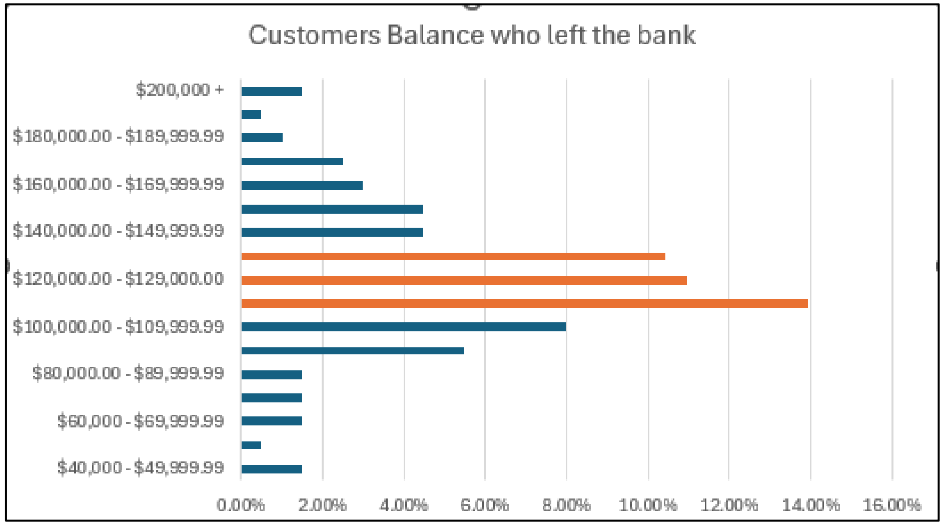
A final Excel document was created outlining the data cleaning process, the results of the analysis and the decision tree created.

05. Pig E Bank Analysis

Customer who left – Bank Balance



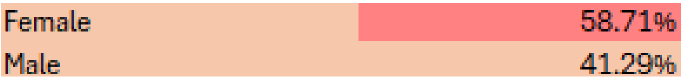
$40,000 - $49,999.99	1.49%
$50,000 - $59,999.99	0.50%
$60,000 - $69,999.99	1.49%
$70,000.00 - $79,999.99	1.49%
$80,000.00 - $89,999.99	1.49%
$90,000.00 - $99,999.99	5.47%
$100,000.00 - $109,999.99	7.96%
$110,000.00 - $119,999.99	13.93%
$120,000.00 - $129,000.00	10.95%
$130,000.00 - $139,999.99	10.45%
$140,000.00 - $149,999.99	4.48%
$150,000.00 - $159,999.99	4.48%
$160,000.00 - $169,999.99	2.99%
$170,000.00 - $179,999.99	2.49%
$180,000.00 - $189,999.99	1.00%
$190,000 - $199,999.99	0.50%
$200,000 +	1.49%
$0.00	27.36%



This shows that customers with a bank balance in this range in red on the bar chart are at a higher risk of leaving the bank at 32.32%.

The next highest group of customers leaving the bank are those with no money in the bank at 27.36%.

Customer who left – Gender

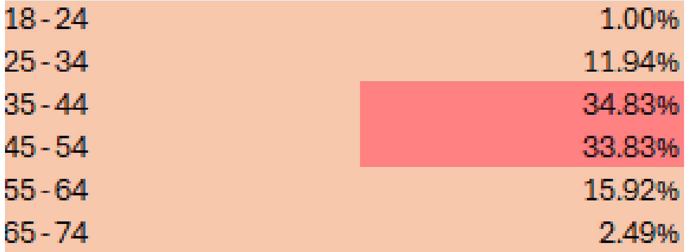


Female	58.71%
Male	41.29%

Female customers are more likely to leave the bank with 58.71% leaving the bank.

Female customer make up a smaller percentage of total customers but a higher percentage of customers who left the bank.

Customer who left – Age

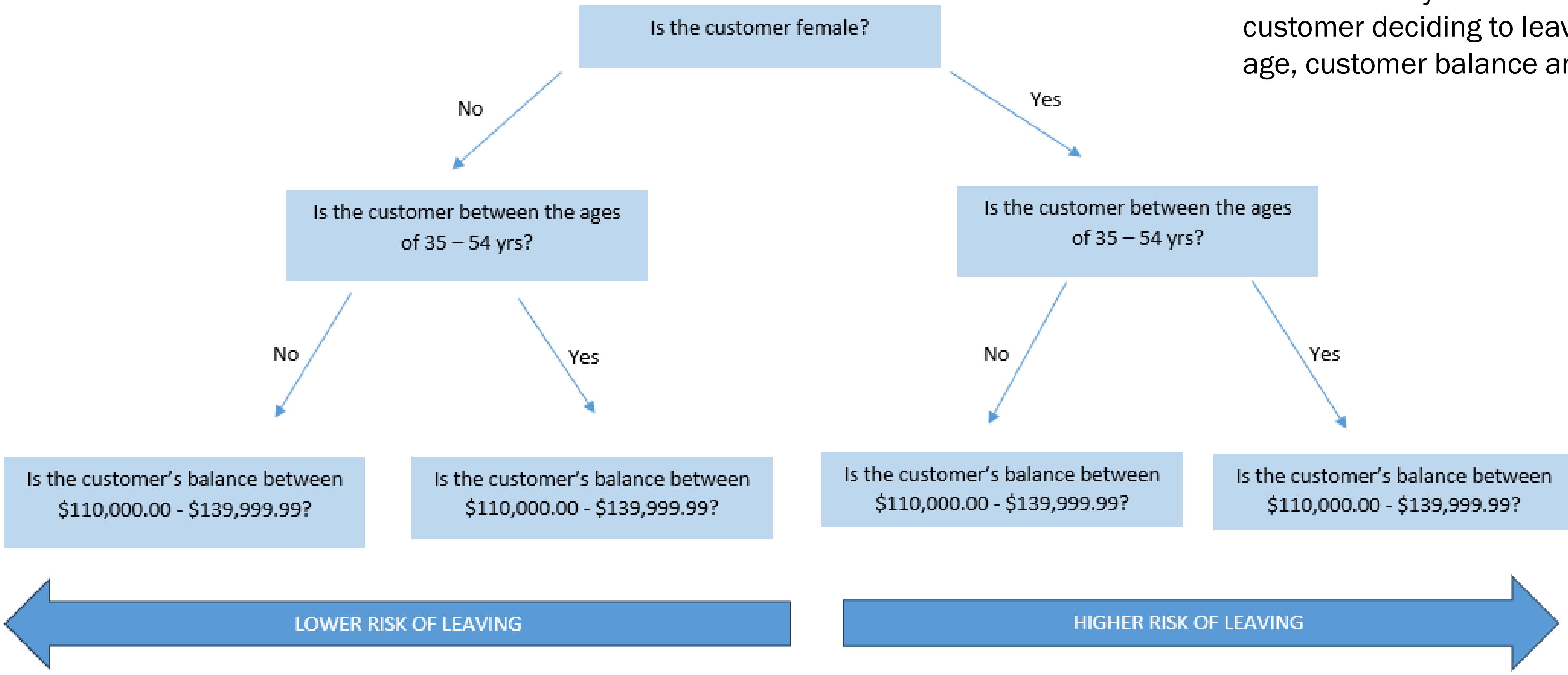


18 - 24	1.00%
25 - 34	11.94%
35 - 44	34.83%
45 - 54	33.83%
55 - 64	15.92%
65 - 74	2.49%

The highest concentration of customers that left the bank were in the 35- 54 age range at 68%.

05. Decision Tree

Risk of Customer Leaving the Bank



Is the customer female?

No

Yes

Is the customer between the ages of 35 – 54 yrs?

Is the customer between the ages of 35 – 54 yrs?

No

Yes

No

Yes

Is the customer's balance between $110,000.00 - $139,999.99?

Is the customer's balance between $110,000.00 - $139,999.99?

Is the customer's balance between $110,000.00 - $139,999.99?

Is the customer's balance between $110,000.00 - $139,999.99?

LOWER RISK OF LEAVING

HIGHER RISK OF LEAVING

The contents of the Decision Tree was generated from the Excel analysis on the factors that contribute to a customer deciding to leave the bank. We can see that age, customer balance and gender are key factors.

Thank you

Email : stephanie-u@hotmail.co.uk

LinkedIn

GitHub