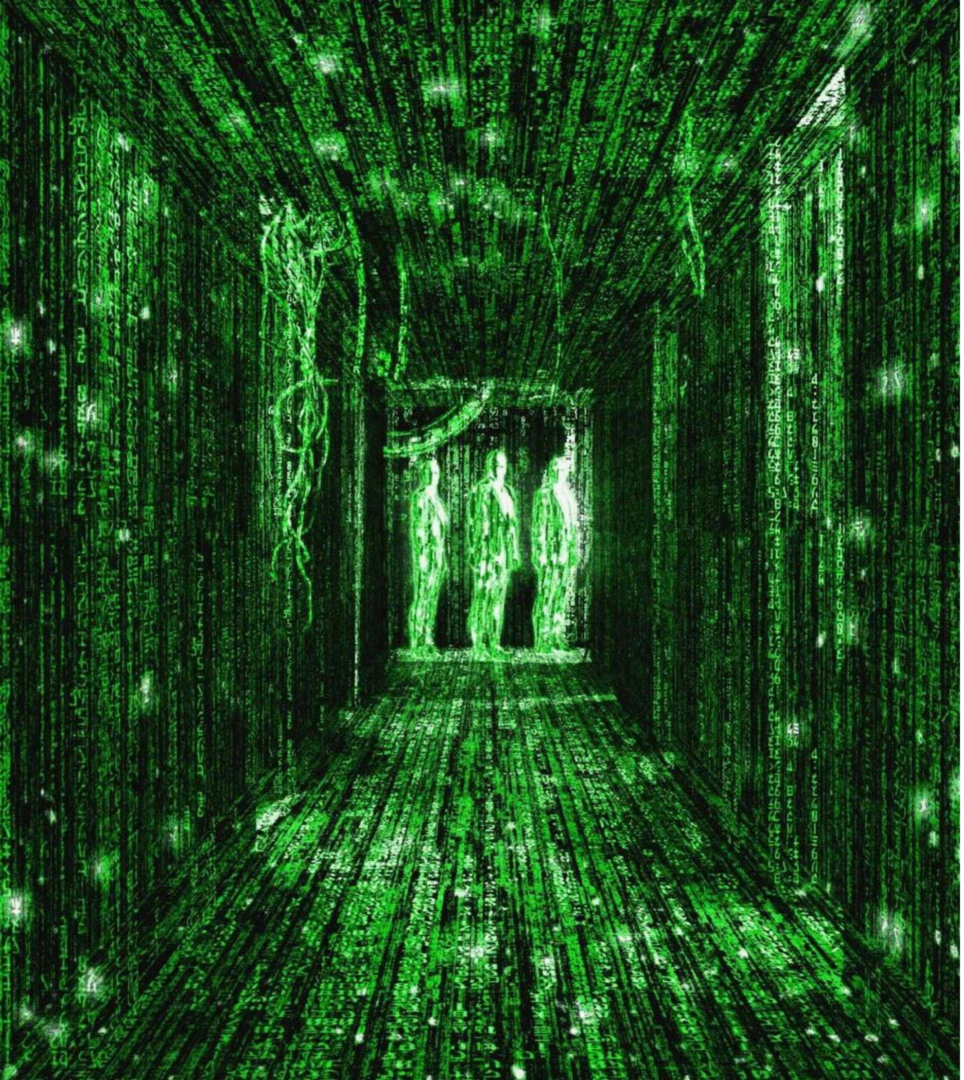


Data Analyst Portfolio

Harrison Genrong Zhong

+ableau

About Me

Hi, I am Harrison Genrong Zhong and I am a data analyst who is passionate about different quantitative domains.

My prior experience in the banking and logistics industries has taught me important skills, such as customer service, problem-solving, time management, sales and data analytics skills.

My goal is to utilize the data skills that I gained and improved from the development of various projects to help businesses strive.

Projects

GameCo

Video game sales data analysis for the planning of marketing budget and the development of new games.

Influenza

Analysis of multiple CDC and US Census Bureau data sets to prepare staffing for the upcoming influenza season.

Instacart

Analysis of the sales data to get a better understanding of the market and the customers in order to improve marketing quality.

Rockbuster Stealth LLC

Analysis of movie rental data to help the company launch the new online platform.

Pig E. Bank

Predictive analysis of customer retention in order to find out the main indicators that customers leave the bank.

New Project

Coming soon.

GameCo

Intro

GameCo is a fictional video game company that wants to use data to plan their sales and marketing budget as well as game development.

Tools

X P

Data Limitations

- Only tracks the total number of units of games sold.
- No data after 2016.

Objective

Perform a descriptive analysis of a video game dataset to foster a better understanding of the market and help GameCo plan their game development.

Main Tasks

- Data Cleaning
- Grouping & Summarizing Data
- Descriptive Analytics
- Data Visualization & Storytelling

Data

- Video Game Sales Data Set

(It tracks the total number of units of games sold in millions.)

- Data Source

(A website that tracks video game shipment information and sales data.)

Main Steps

Data Cleaning

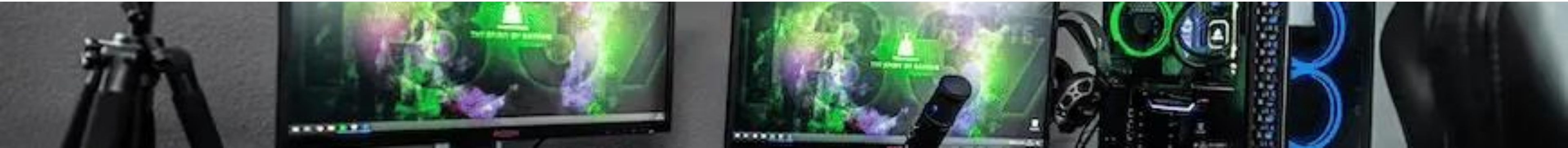
- Removed duplicates, zeros, blanks and empty rows.
- Used mean imputation to fill in the empty cells in partially empty columns.

Data Analysis

- Created pivot tables, slicers, groups and calculated fields for analysis.
- Calculated central tendency, the spread of data and the range of the sales. Also looked for outliers.
- Created bar charts, line charts, box charts and whisker charts for analysis.

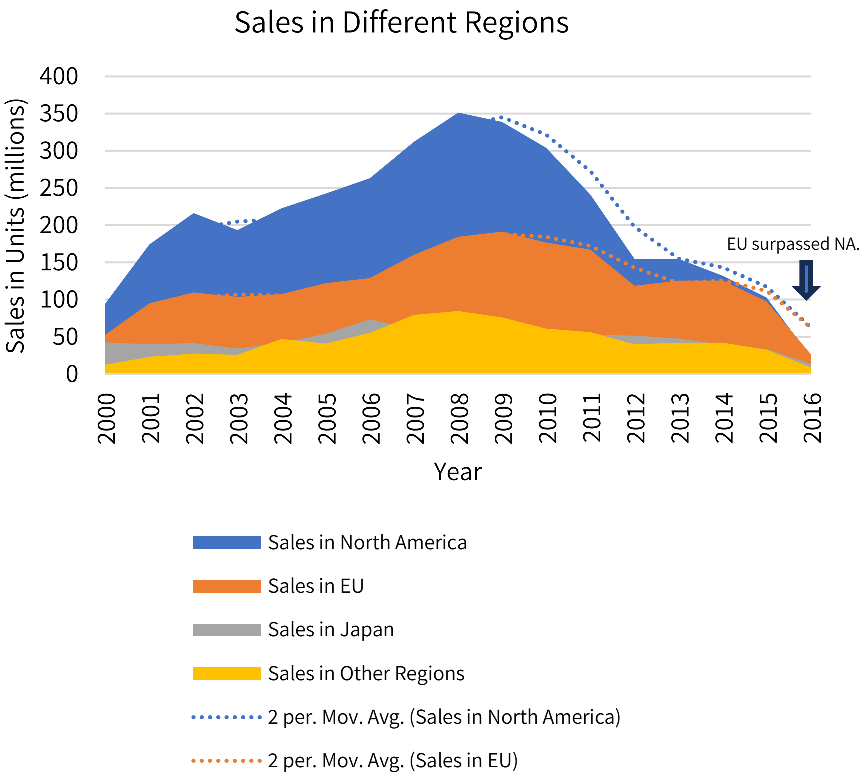
Data Visualization

Created data visualization for my findings.



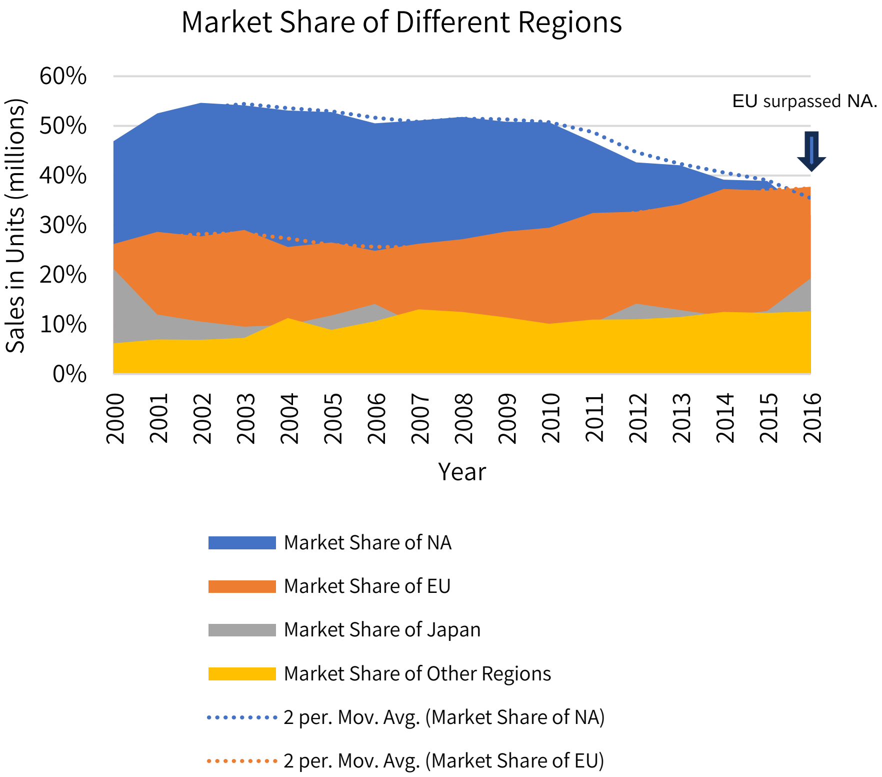
Regional Analysis

Market share and sales in Europe have been growing steadily and surpassed North America in 2016.



Sales in Different Regions

EU surpassed NA.



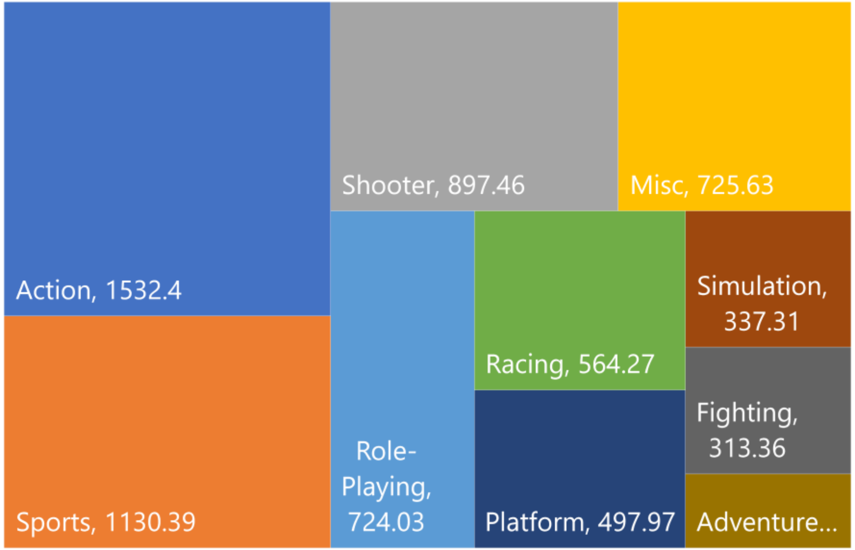
Market Share of Different Regions

EU surpassed NA.

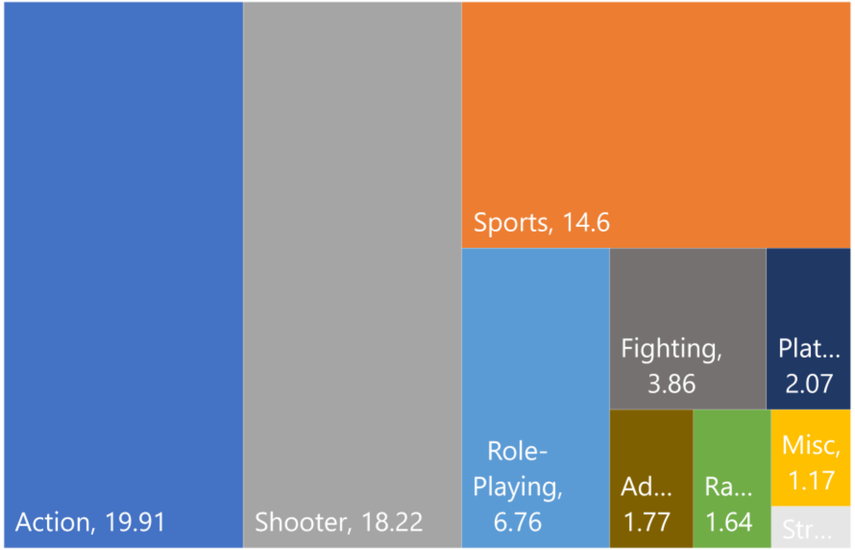
Genre Analysis

The top 3 genres (action, shooter, sports) have been performing well over the years.

Top 10 Genres 2000 - 2016

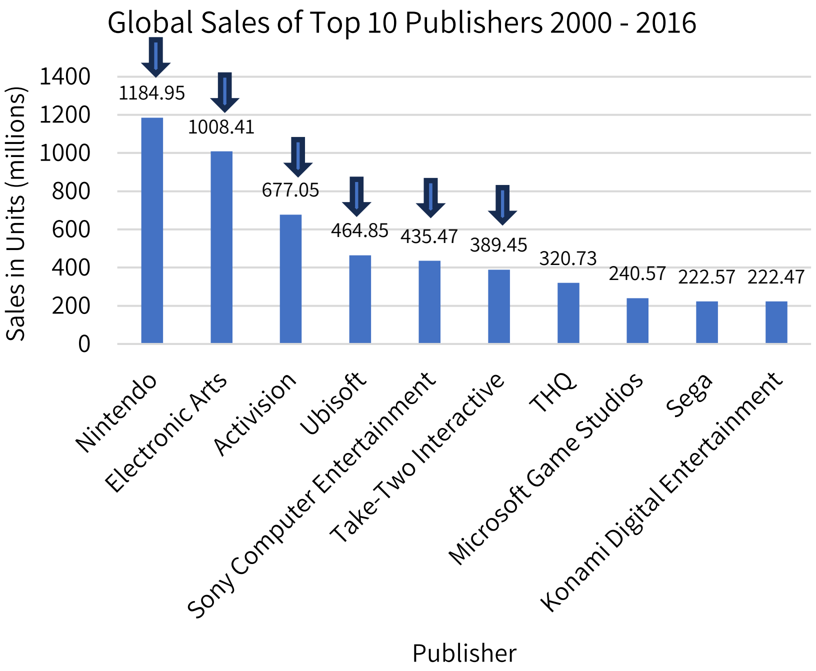


Top 10 Genres in 2016

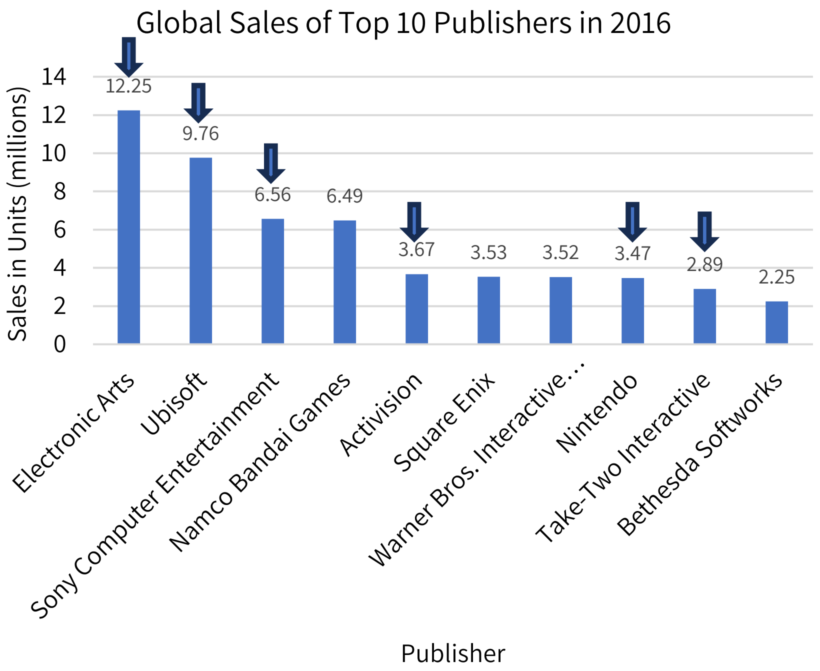


Publisher Analysis

Nintendo, EA, Activision, Ubisoft, Sony and Take-Two have been performing well over the years.



Global Sales of Top 10 Publishers 2000 - 2016



Global Sales of Top 10 Publishers in 2016

Recommendations

01 Sales and Marketing Budget

- Allocate sales and marketing budgets according to the market share percentages in 2016.
- North America 32%
- Europe 38%
- Japan 19%
- Other Regions 13%

02 Game Development

Take top 3 genres (action, shooter, sports) into consideration when developing new games.

03 Market Research

Conduct research on Nintendo, EA, Activision, Sony and Take-Two to find out how they have been consistently performing well over the years.

Influenza

Intro

A medical staffing agency needs to plan for the upcoming influenza season.

Tools & Deliverables

Interim Report
Tableau Dashboard

Data Limitations

- Death counts of 9 or fewer were not included.
- Population numbers are estimates.
- Flu shot data was collected through phone surveys, which might affect reliability.

Objective

Analyze datasets from CDC and the US Census Bureau and help the staffing agency plan for the influenza season on an as-needed basis.

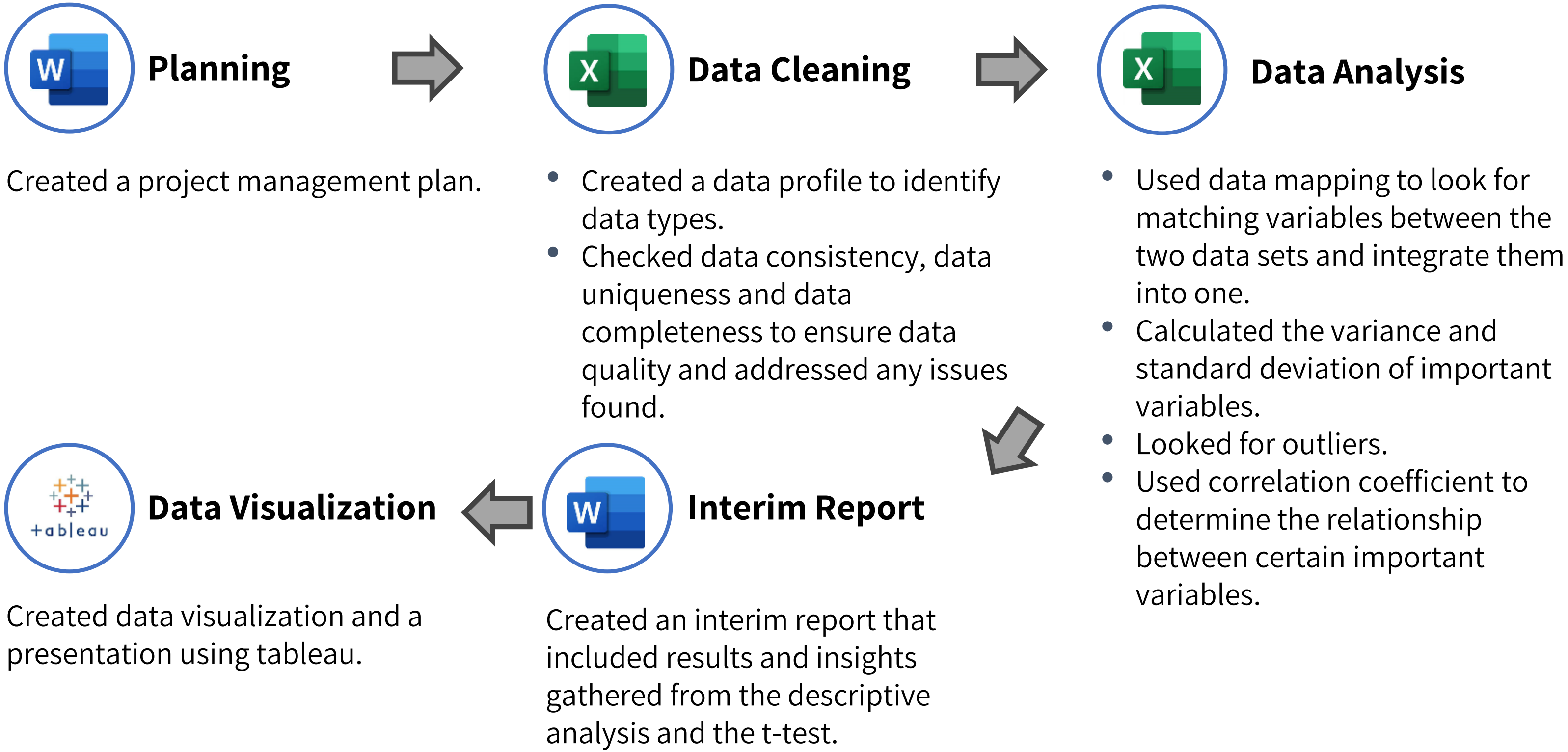
Main Tasks

- Data Sourcing
- Data Profiling
- Data Quality Measures
- Data Transformation
- Statistical Analysis
- Data Analysis & Visualization with Tableau

Data

- Influenza Deaths by Geography (CDC)
- Population Data by Geography, Time, Age, and Gender (US Census Bureau)
- Influenza Visit Data Set (CDC)
- Survey of Flu Shot Rates in Children (CDC)

Main Steps



Planning

Created a project management plan.

Data Cleaning

- Created a data profile to identify data types.
- Checked data consistency, data uniqueness and data completeness to ensure data quality and addressed any issues found.

Data Analysis

- Used data mapping to look for matching variables between the two data sets and integrate them into one.
- Calculated the variance and standard deviation of important variables.
- Looked for outliers.
- Used correlation coefficient to determine the relationship between certain important variables.

Interim Report

Created an interim report that included results and insights gathered from the descriptive analysis and the t-test.

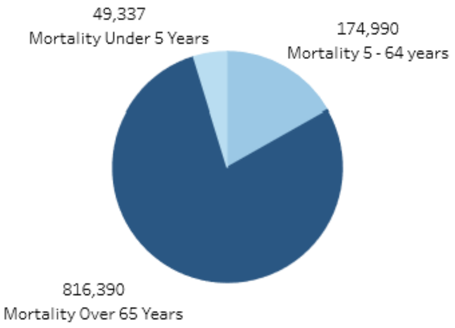
Data Visualization

Created data visualization and a presentation using tableau.

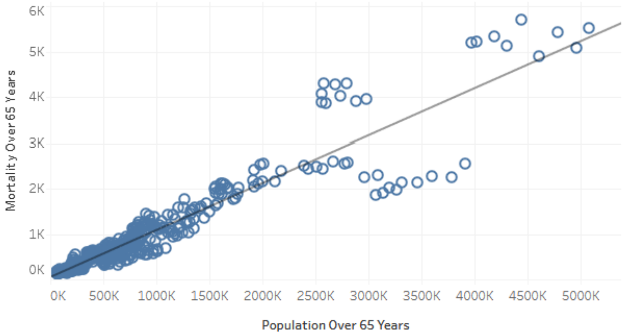
Statistical Analysis

- The population over 65 years old and the mortalities over 65 years old are positively correlated.
- The mortalities over 65 years old are more than the rest of the age groups combined.

Mortality Pie Chart 2009 - 2017



The Ralationship between Mortality and Population over 65 Yeas Old 2009 - 2017



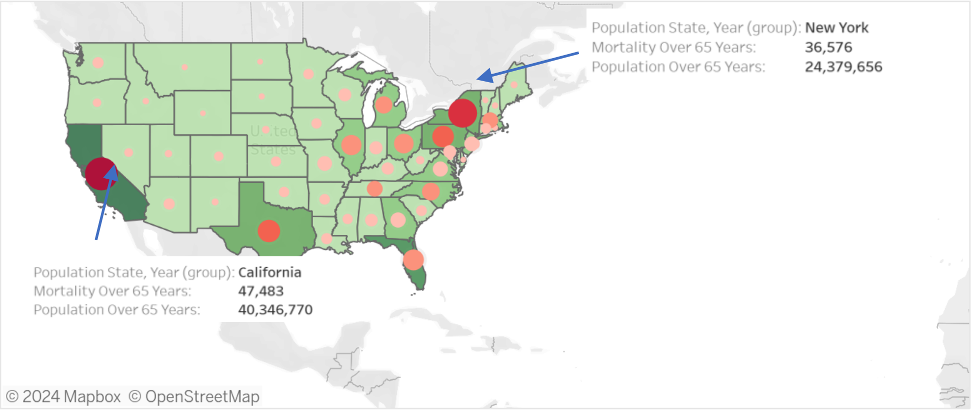
Data Spread (Over 65 Years Old)		
Variable	Death over 65 Years Old	Population over 65 Years Old
Data Set Name	Influenza Mortality	Census Population
Sample or Population?	Population (based on death certificates)	Sample (based on surveys)
Variance	966931.14	799546871054.48
Standard Deviation	983.33	894173.85
Mean	896.38	814676.89
Outlier Lower Bound	-2053.60	-1867844.66
Outlier Higher Bound	3846.36	3497198.43
Counts of Outliers	18	12
Counts of All Values	450	450
Outlier Percentage	0.04	0.03
Correlation		
Correlation Coefficient	0.94	
Strength of Correlation	Strong relationship.	
Interpretation	Since the correlation coefficient is 0.94, they have a strong relationship. It shows that population over 65 years old and deaths over 65 years old are directly corelated.	

The closer the absolute value of the coefficient is to 1, the stronger the relationship.

Mortality Analysis

California and New York have the highest mortalities over 65 years old.

Mortality and Population over 65 by State 2009 - 2017

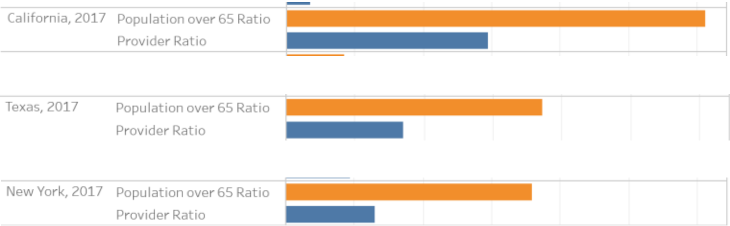


Population State, Year (group): New York
Mortality Over 65 Years: 36,576
Population Over 65 Years: 24,379,656

Population State, Year (group): California
Mortality Over 65 Years: 47,483
Population Over 65 Years: 40,346,770

- Many states had excessive staff.
- Many states didn’t have enough staff.
- California, New York and Texas didn’t even have half of the required staff.

(Population over 65 Ratio = Population over 65 each State / Total Population over 65, Provider Ratio = Number of Sentinel Providers each State / Total Sentinel Providers)

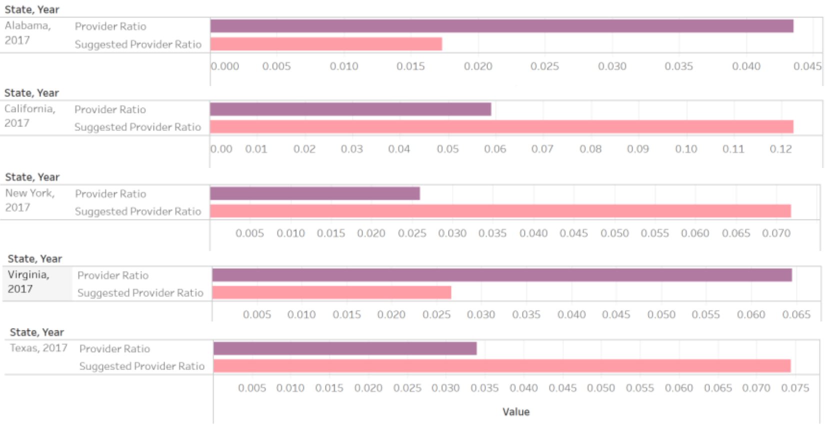


Recommendations

01 Staffing Ratio

- Calculate staffing ratio based on population over 65 (most susceptible population).
- Suggested Staffing Ratio = Population over 65 each State / Total Population over 65
- Alabama, California, New York, Virginia and Texas will have the most significant changes.

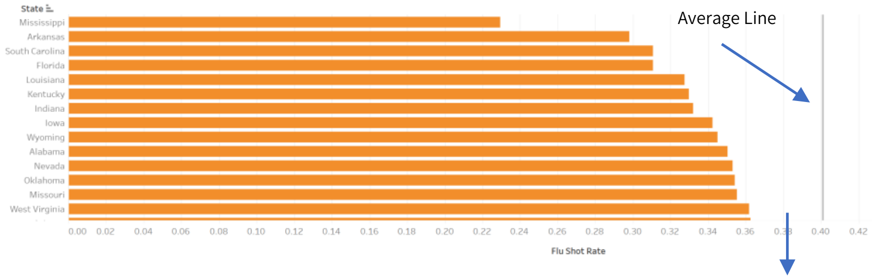
Comparison between Provider Ratios 2017 and Suggested Ratios



02 Flu Shots

States that are below the average line should push for higher flu shot coverage.

Children under 2 Year Old Flu Shot Rates 2017 from Low to High



Average Line

A lot of the states are below the average line in terms of flu shot rates.

Instacart

Intro

Instacart is an online grocery store that operates through an app.

Tools & Deliverables

ANACONDA jupyter pandas NumPy Matplotlib seaborn

Project Github Page
Presentation

Data

- Customer Data Set (CareerFoundry)
- Data Dictionary
- The Instacart Online Grocery Shopping Dataset 2017 (Kaggle)

Objective

Perform an initial data and exploratory analysis of some of their data in order to derive insights and suggest strategies for better segmentation.

Main Tasks

- Data Wrangling & Subsetting
- Data Consistency Checks
- Combing & Exporting Data
- Deriving New Variables
- Grouping Data & Aggregating Variables
- Data Visualization with Python
- Excel Reporting

Data Limitations

Only contains data from 2017.

Main Steps

Data Analysis

- Created a project folder.
- Ran data consistency checks (missing values and duplicates).
- Performed data wrangling, which included dropping, renaming, and changing data types of certain columns.
- Derived new columns from original columns and merged datasets.

Data Visualization

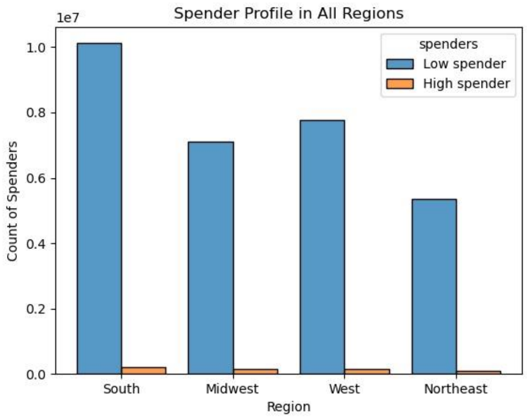
Created visualization and gained useful insights.

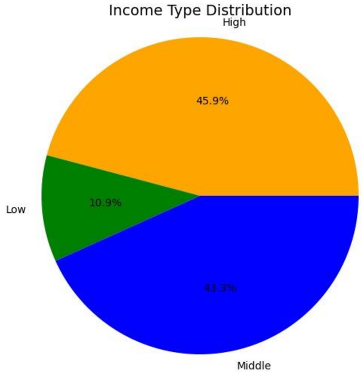
Presentation

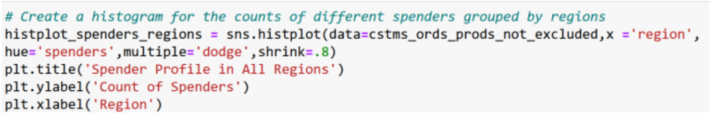
Created a presentation to present my findings.

Customer Analysis

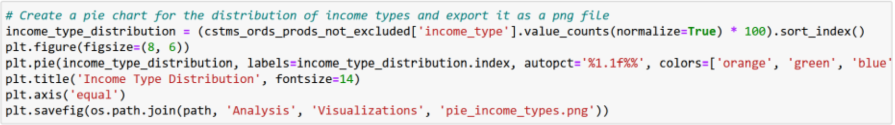
- Low spenders (<$10) outperformed high spenders (>=$10) even though there is a much larger amount of high-income earners than low-income earners.
- The Northeast significantly underperformed in terms of sales.







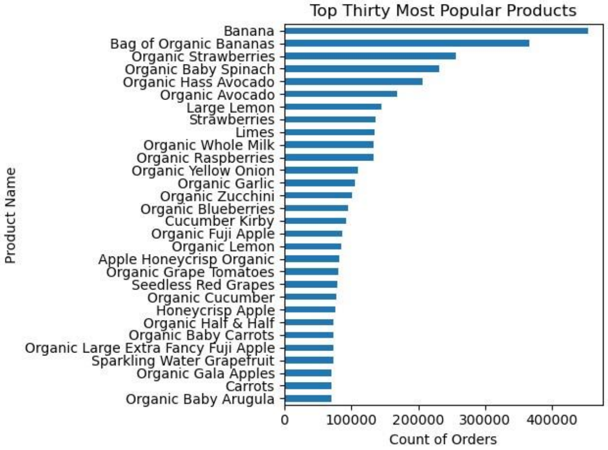
```
# Create a histogram for the counts of different spenders grouped by regions
histplot_spenders_regions = sns.histplot(data=cstms_ords_prods_not_excluded,x ='region',
hue='spenders',multiple='dodge',shrink=.8)
plt.title('Spender Profile in All Regions')
plt.ylabel('Count of Spenders')
plt.xlabel('Region')
```

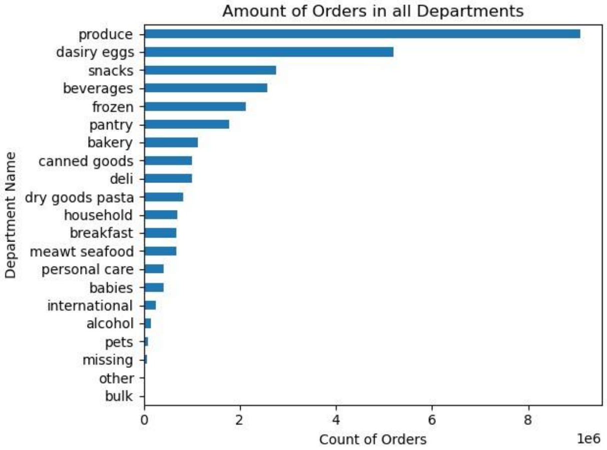


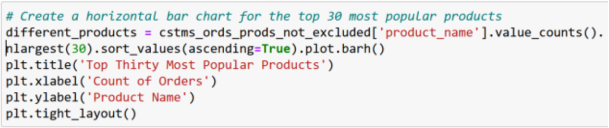
```
# Create a pie chart for the distribution of income types and export it as a png file
income_type_distribution = (cstms_ords_prods_not_excluded['income_type'].value_counts(normalize=True) * 100).sort_index()
plt.figure(figsize=(8, 6))
plt.pie(income_type_distribution, labels=income_type_distribution.index, autopct='%1.1f%%', colors=['orange', 'green', 'blue'
plt.title('Income Type Distribution', fontsize=14)
plt.axis('equal')
plt.savefig(os.path.join(path, 'Analysis', 'Visualizations', 'pie_income_types.png'))
```

Department Analysis

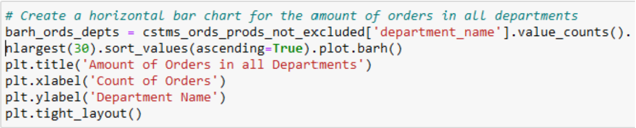
- Produce is the best-selling department, and it outperforms other departments.
- The best-selling products include a lot of organic products.







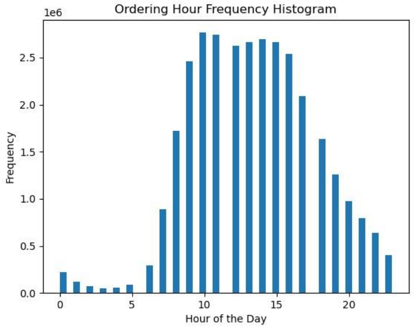
```
# Create a horizontal bar chart for the top 30 most popular products
different_products = cstms_ords_prods_not_excluded['product_name'].value_counts().
nlargest(30).sort_values(ascending=True).plot.barh()
plt.title('Top Thirty Most Popular Products')
plt.xlabel('Count of Orders')
plt.ylabel('Product Name')
plt.tight_layout()
```

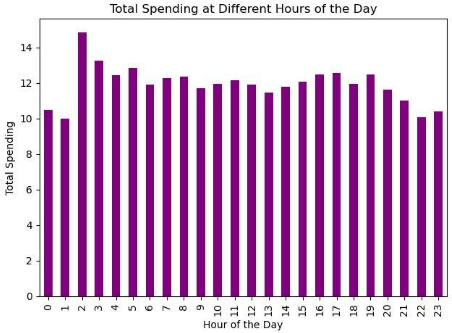


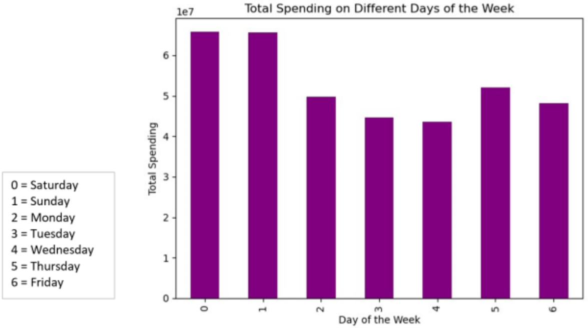
```
# Create a horizontal bar chart for the amount of orders in all departments
barh_ords_depts = cstms_ords_prods_not_excluded['department_name'].value_counts().
nlargest(30).sort_values(ascending=True).plot.barh()
plt.title('Amount of Orders in all Departments')
plt.xlabel('Count of Orders')
plt.ylabel('Department Name')
plt.tight_layout()
```

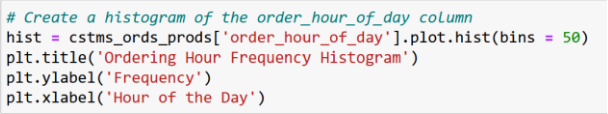
Frequency Analysis

- The slowest period is 12am – 6am but the highest spending hours are 2am and 3am.
- Tuesday and Wednesday are the lowest spending days of the week.

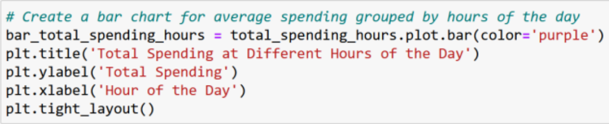




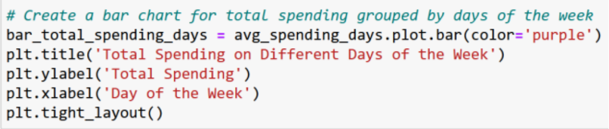




```
# Create a histogram of the order_hour_of_day column
hist = cstms_ords_prods['order_hour_of_day'].plot.hist(bins = 50)
plt.title('Ordering Hour Frequency Histogram')
plt.ylabel('Frequency')
plt.xlabel('Hour of the Day')
```



```
# Create a bar chart for average spending grouped by hours of the day
bar_total_spending_hours = total_spending_hours.plot.bar(color='purple')
plt.title('Total Spending at Different Hours of the Day')
plt.ylabel('Total Spending')
plt.xlabel('Hour of the Day')
plt.tight_layout()
```



```
# Create a bar chart for total spending grouped by days of the week
bar_total_spending_days = avg_spending_days.plot.bar(color='purple')
plt.title('Total Spending on Different Days of the Week')
plt.ylabel('Total Spending')
plt.xlabel('Day of the Week')
plt.tight_layout()
```

Recommendations

01 Marketing

Increase marketing in the Northeast region to attract new customers and increase sales (currently the bottom region in all aspects).

02 Promotions

- Promotions from 10pm to 6am because those are the slowest hours.
- 2am and 3am are the highest spending hours. Promotions for produce, dairy & eggs and beverages (the top 3 best selling departments) during those hours can encourage customers to spend even more.
- Tuesday and Wednesday generate the least sales. Promotions for those days could boost sales.

03 Other

- Increase more organic options since they are the best-selling category.
- Send surveys to high income earners to gather their preferences.

Rockbuster Stealth LLC

Intro

Rockbuster Stealth LLC is a fictional movie rental company that is planning to use its existing movie licenses to launch an online video rental service.

Objective

Analyze data, gain insights and answer questions in order to help the company with the launch of the new online platform.

Tools & Deliverabels

Rockbuster ERD
Rockbuster Data Outputs
Rockbuster Data Dictionary
Rockbuster Presentation

Main Tasks

- Data Querying, Filtering, Summarizing and Cleaning in SQL.
- Joining Tables
- Subqueries
- Performing Common Table Expressions
- Data Dictionary
- Presentation

Data

Rockbuster Dataset
(It contains film inventory, customers, and payments, among other things.)

Data Limitations

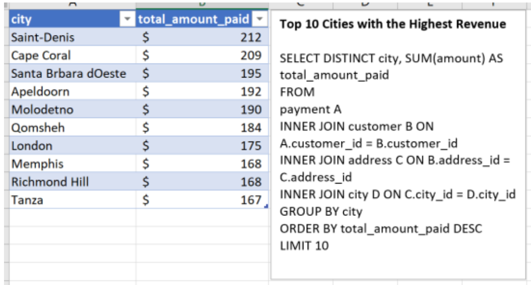
It only contains 3 months of data ranging from February 2017 to May 2017.

Main Steps

ERD

Created an entity relationship diagram.

SQL Example



city	total_amount_paid
Saint-Denis	$ 212
Cape Coral	$ 209
Santa Brbara dOeste	$ 195
Apeldoorn	$ 192
Molodetno	$ 190
Qomsheh	$ 184
London	$ 175
Memphis	$ 168
Richmond Hill	$ 168
Tanza	$ 167

Top 10 Cities with the Highest Revenue

```
SELECT DISTINCT city, SUM(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
GROUP BY city
ORDER BY total_amount_paid DESC
LIMIT 10
```

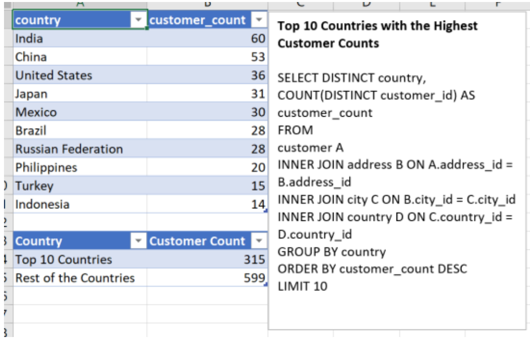
Data cleaning and Analysis

- Refined queries, ordered data and grouped data.
- Filtered data to answer initial questions.
- Cleaned data (duplicates and missing values)
- Summarized data.
- Joined tables
- Ran subqueries and common table expressions to answer key questions.

Data Dictionary and Presentation

- Created a data dictionary using Word.
- Created a presentation using PowerPoint.

SQL Example



country	customer_count
India	60
China	53
United States	36
Japan	31
Mexico	30
Brazil	28
Russian Federation	28
Philippines	20
Turkey	15
Indonesia	14

Country	Customer Count
Top 10 Countries	315
Rest of the Countries	599

Top 10 Countries with the Highest Customer Counts

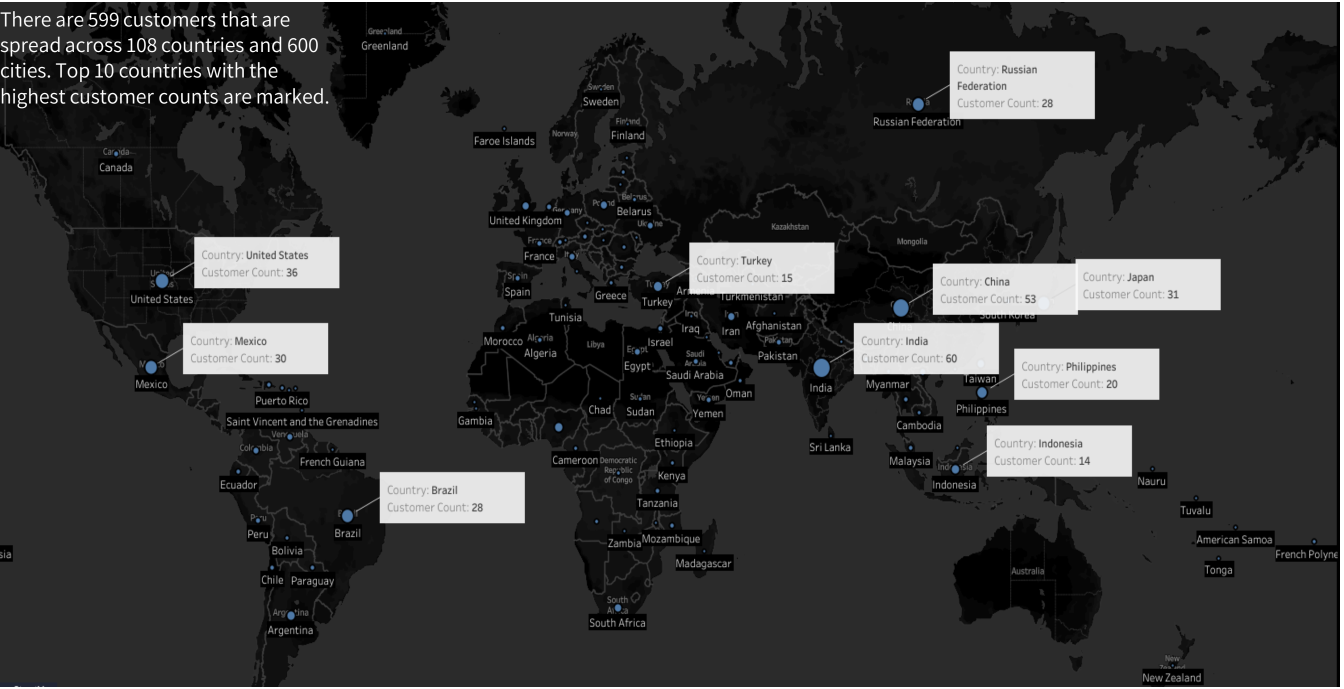
```
SELECT DISTINCT country,
COUNT(DISTINCT customer_id) AS
customer_count
FROM
customer A
INNER JOIN address B ON A.address_id =
B.address_id
INNER JOIN city C ON B.city_id = C.city_id
INNER JOIN country D ON C.country_id =
D.country_id
GROUP BY country
ORDER BY customer_count DESC
LIMIT 10
```

Business Overview

- Customer Count: 599
- Count of Film Titles: 1000
- Average, Longest and Shortest Rental Duration: 5 Days, 7 Days, 3 Days
- Average, Highest and Lowest Rental Rate: $3, $5, $1
- Average Rental Revenue per Customer: $102
- Total Revenue: $61312
- Amount of Language Selections: 6 (English, Italian, Japanese, Mandarin, French, German)

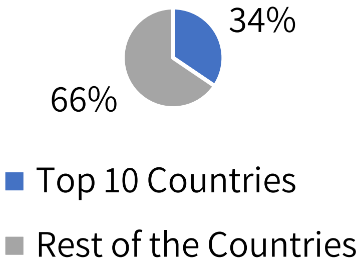
Top 10 Country Analysis

Top 10 countries with highest customer counts take up 34% of the total population and total revenue.

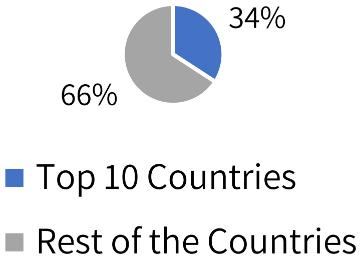


There are 599 customers that are spread across 108 countries and 600 cities. Top 10 countries with the highest customer counts are marked.

Customer Count

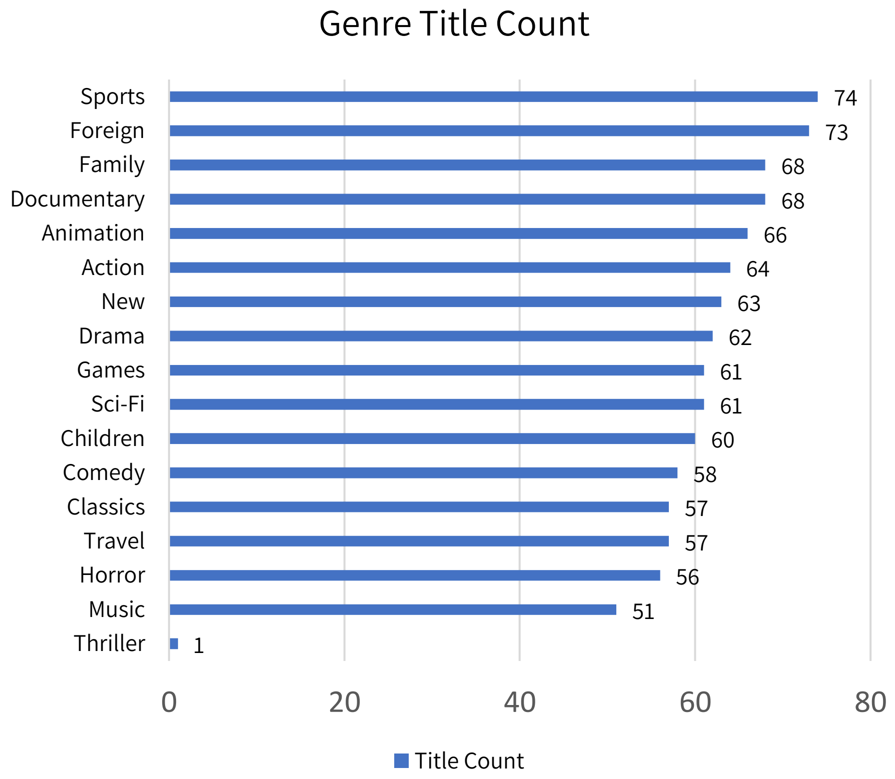


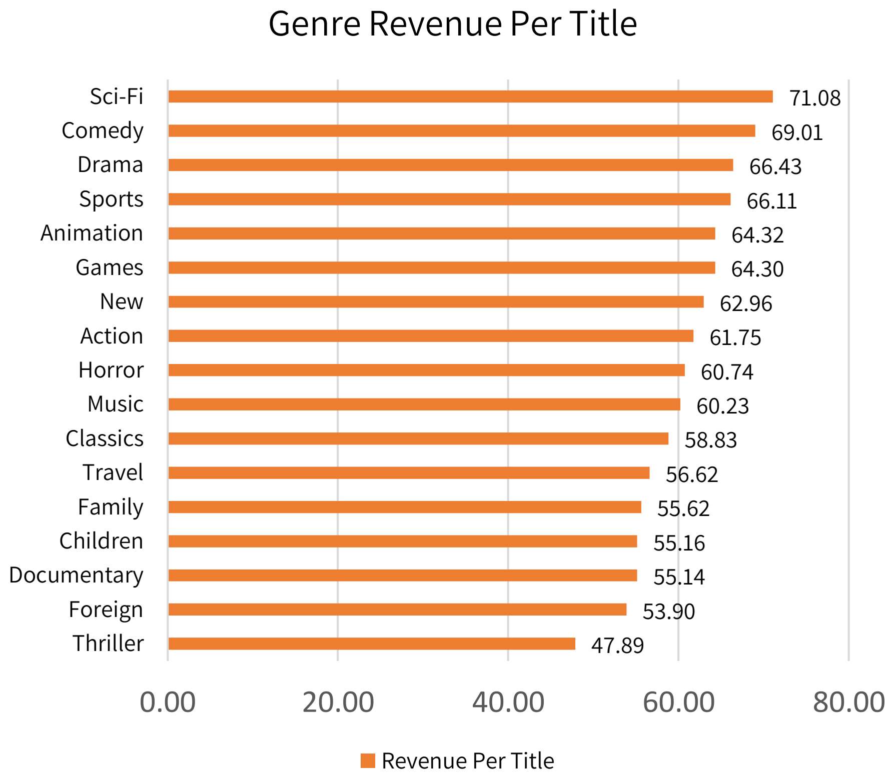
Revenue



Genre Analysis

- Drama, sci-fi, horror and music have potential because they did really well despite lower title counts.
- Foreign, family and documentary did poorly despite high title counts.
- Thriller doesn't have a big enough sample size.





Recommendations

01 Languages

Add more languages that can cover the top 10 countries.

02 Movie Collection

- Add more titles to the thriller genre to increase the sample size and test out the popularity.
- Add more titles to the genres that have potential in order to build a better movie collection.
- Send out surveys to gather customers' preferences.

03 Marketing

Social media or a referral program that can increase the current low customer count.

Pig E. Bank

Intro

Pig E. Bank is a fictional bank that wants to analyze their data and improve customer retention.

Tools

Pig E. Bank Data Analysis

Data

Client Data Set (CareerFoundry)

Objective

Analyze the data and identify the leading indicators that a customer will leave the bank.

Main Tasks

- Data Mining
- Predictive Analysis
- Time Series Analysis & Forecasting

Data Limitations

Customer demographics are limited. Doesn’t have transactional data, banking products and the occupation that each customer has.

Main Steps

Data Cleaning

Checked for missing values, errors and inconsistencies and cleaned them.

Data Analysis

Grouped the data and used pivot tables to identify the top 3 to 4 factors that lead to clients leaving.

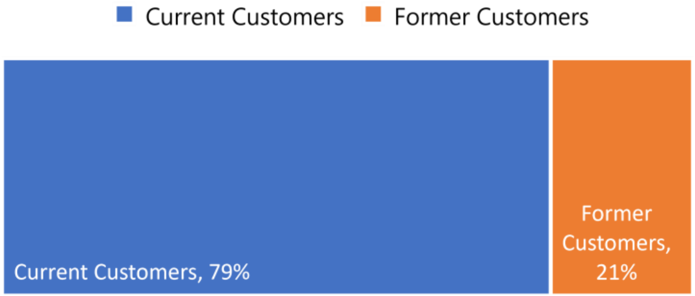
Decision Tree

Created a decision tree based on the findings.

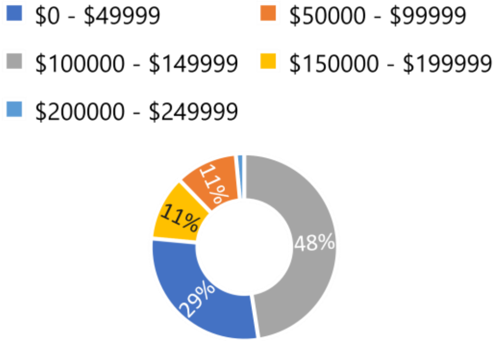
Former Customer Analysis

- 48% of former customers have a $100000 - $149999 bank balance.
- A majority of the former customers live in Germany and France.
- Most of the former customers are in the 40 – 59 age range.

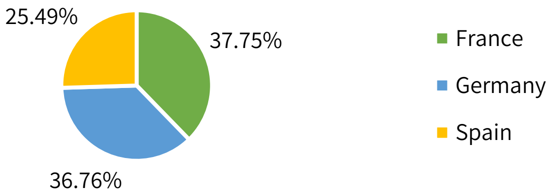
Customer Retention



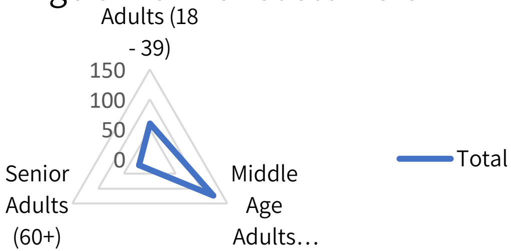
Former Customer Account Balances



Former Customer Countries

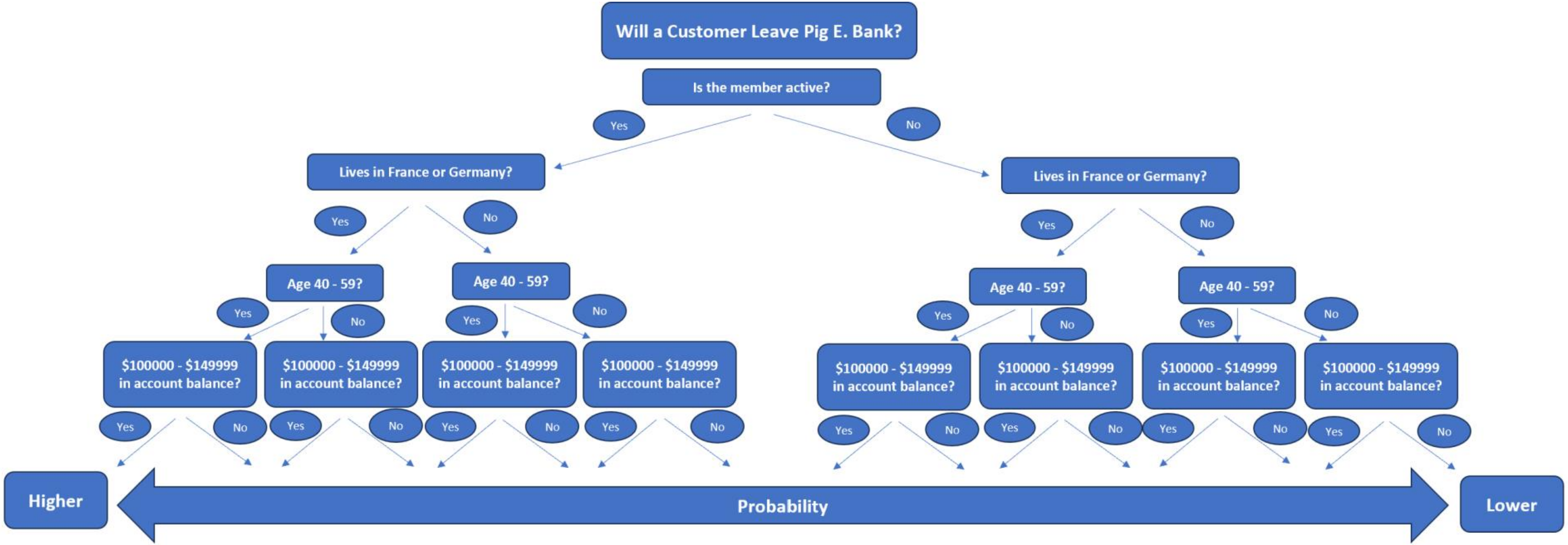


Age of Former Customers



Decision Tree Analysis

There are four main indicators. They go from top to bottom according to how important they are. The most important indicator is at the top. The least significant one is at the bottom.



Will a Customer Leave Pig E. Bank?
Is the member active?
Yes
No
Lives in France or Germany?
Yes
No
Age 40 - 59?
Yes
No
$100000 - $149999 in account balance?
Yes
No
Lives in France or Germany?
Yes
No
Age 40 - 59?
Yes
No
$100000 - $149999 in account balance?
Yes
No
Higher
Probability
Lower

Recommendations

01 Customer Engagement

Increase the quality of customer engagement and product suggestions by sending surveys to customers to gather their preferences.

02 Regions

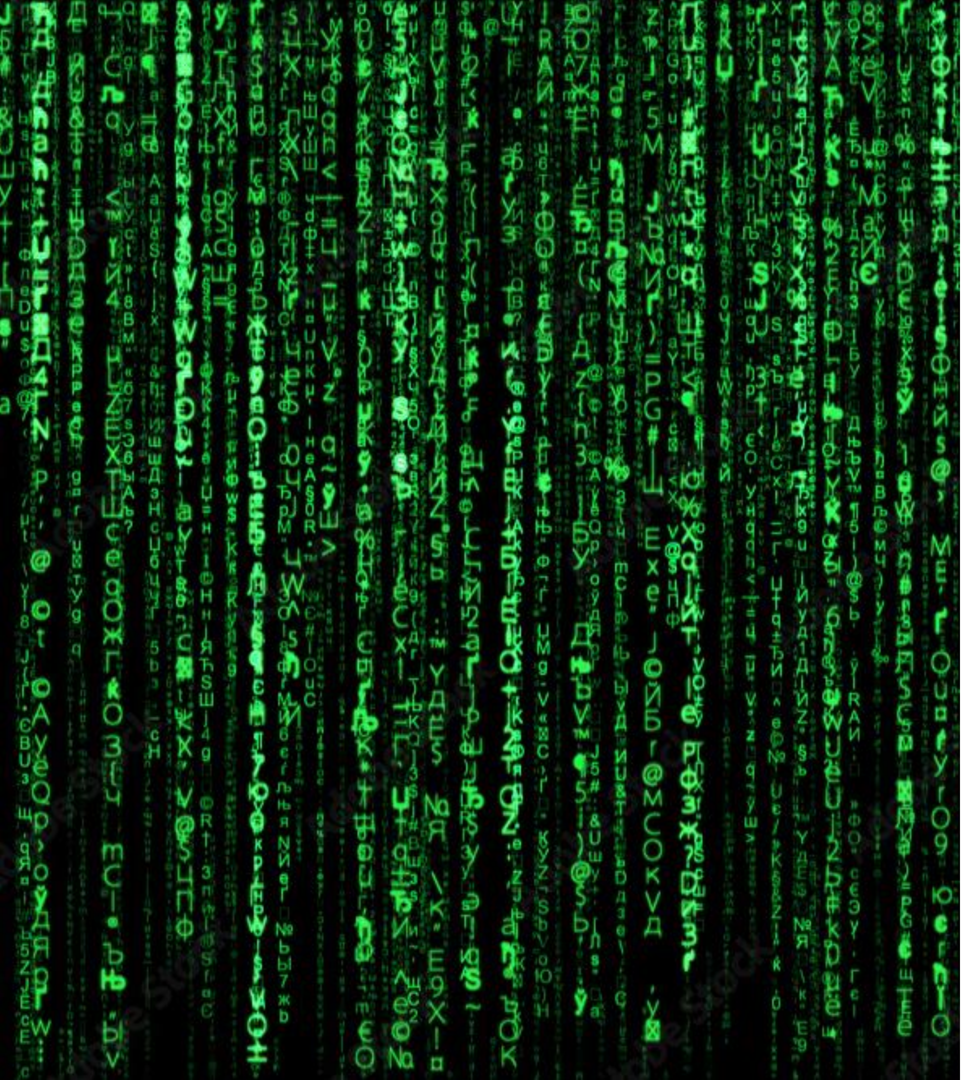
Conduct market research on France and Germany to find out why they a low customer retention rate.

03 40 – 59 Age Group

Send surveys to customers in the 40 – 59 age group in order to conduct research.

04 $100000 - $149999 Account Balance

Send surveys to customers with a $100000 - $149999 account balance to gather more data and conduct research.



Thank you!

Harrison Genrong Zhong