IVONNE ASPILCUETA

DATA ANALYST PORTFOLIO

ABOUT ME

Welcome to my Portfolio!

I'm Ivonne Aspilcueta, a data analyst with a background in software development and entertainment business.

I always been fascinated to be curious, find solutions and be creative, skills I gained in my opposing backgrounds that led me to pursue a data analysis career.

My goal is to discover data that cannot be seen by just looking at the data but that requires a series of skills to discover insights and answer business questions.

TABLE OF CONTENTS

GAMECO

A FICTIONAL VIDEO GAME COMPANY REQUESTING GLOBAL SALES ANALYSIS TO HELP WITH DEVELOPING THE MARKETING BUDGET FOR 2017.

GAMECO IS ASSUMING THAT SALES FOR THE VARIOUS GEOGRAPHIC REGIONS HAVE STAYED THE SAME OVER TIME.

PROJECT OVERVIEW

GOALS:

THE PURPOSE OF THE ANALYSIS IS TO LOOK FOR THE SALES BY REGIONS TO PLAN THE MARKETING BUDGET FOR 2017.

KEY QUESTIONS:

GAMECO EXECUTIVES ARE OPEN TO HEARING ANY INSIGHTS YOU CAN PULL FROM THE DATA BUT

ARE SPECIFICALLY INTERESTED IN THESE 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?

SKILLS: Excel

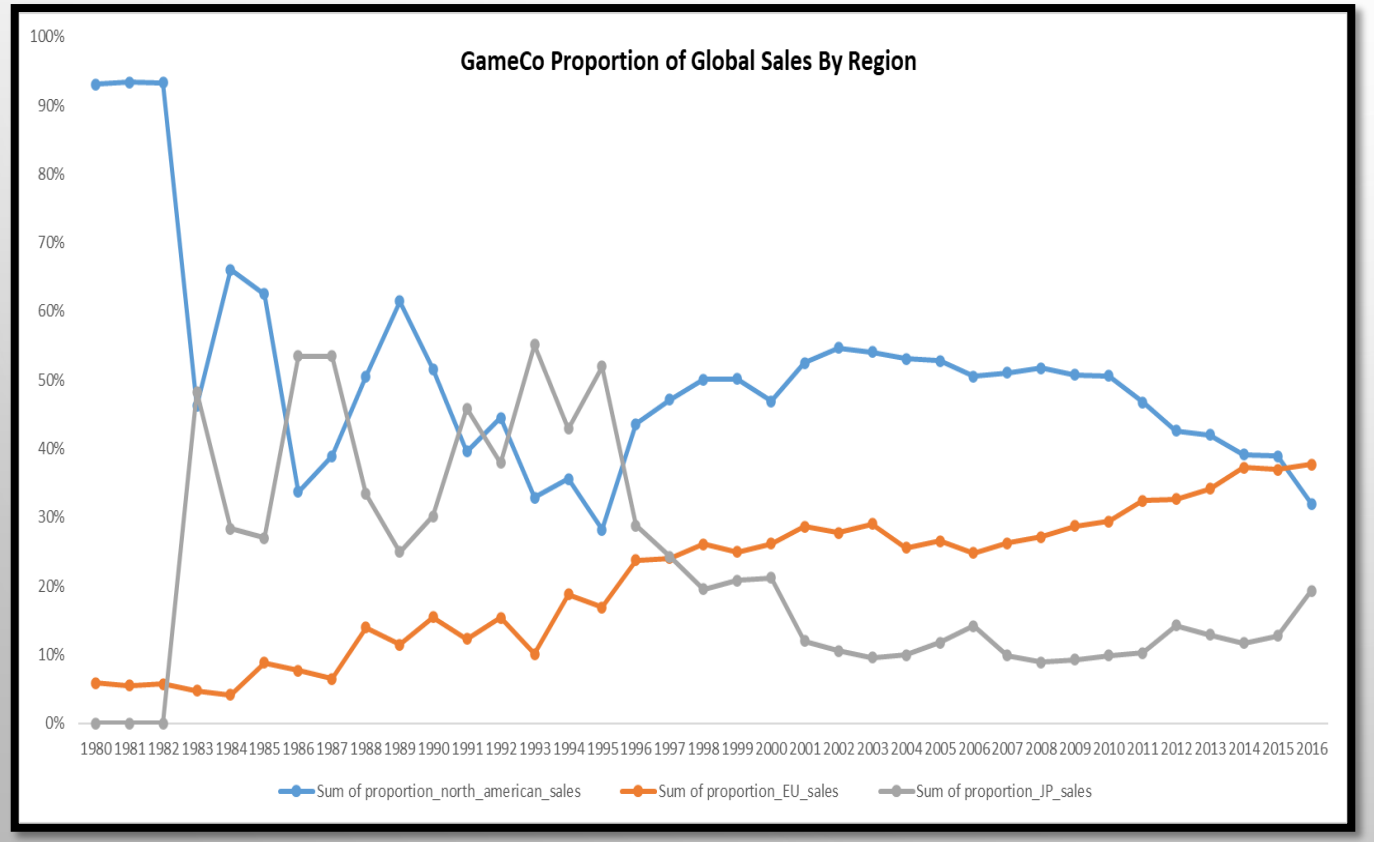
- CLEANING DATA
- GROUPING & SUMMARIZING
- CONDUCTING A DESCRIPTIVE ANALYSIS
- DEVELOPING & VISUALIZING INSIGHTS
- PRESENTING RESULTS

TOOLS:

DATA:

- DATA SOURCED FROM VGCHARTZ
- IT TRACKS TOTAL NUMBER OF UNIT'S GAMES SOLD FROM 1980 TO 2016.
- VGCHARTZ DATA COLLECTION METHODOLOGY
- PROJECT BRIEF
- DOES NOT CONTAIN FINANCIAL FIGURES

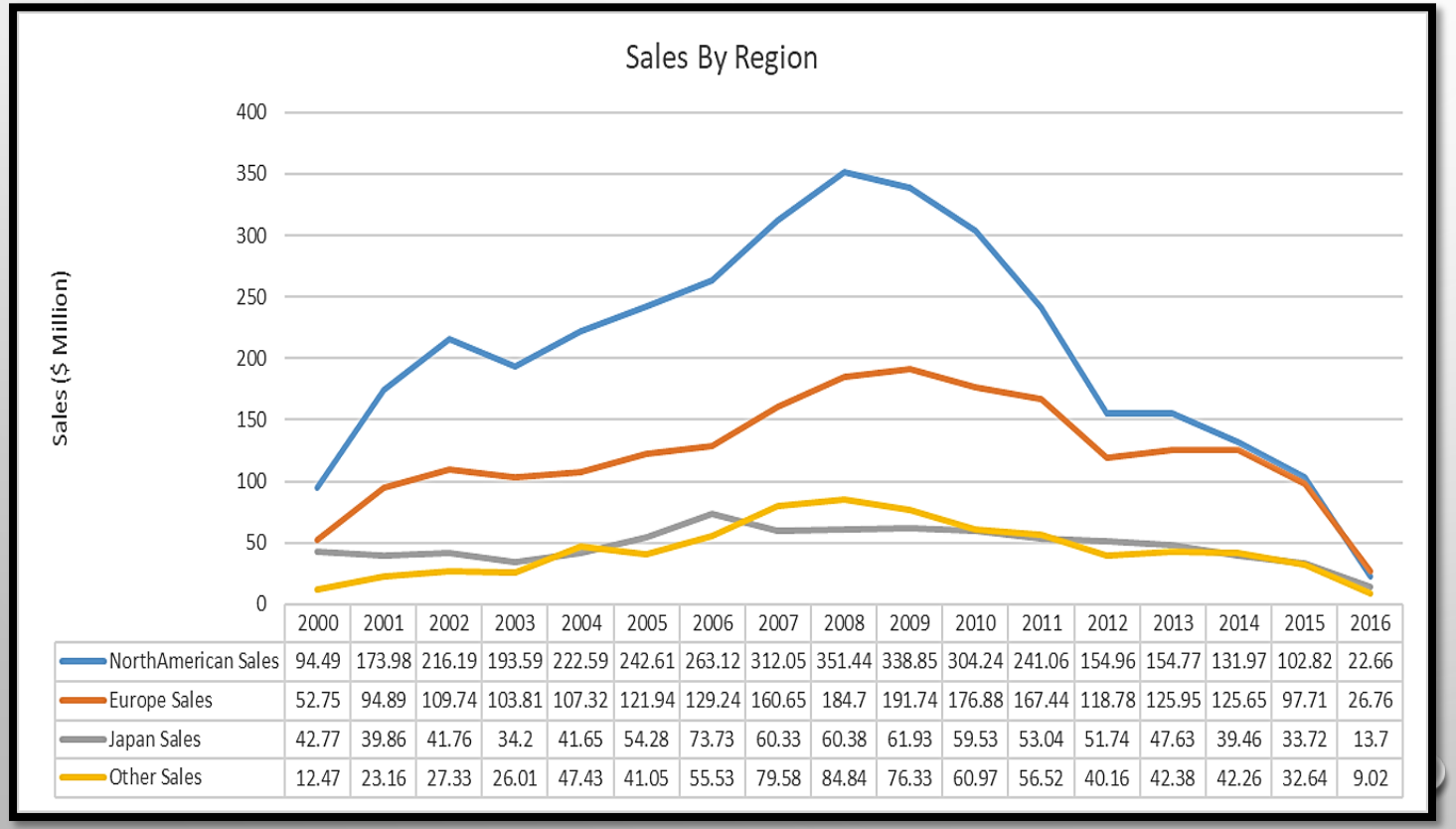
PROPORTION OF GLOBAL SALES BY REGION (1980-2016)



PROPORTION OF GLOBAL SALES HAS CHANGED OVER TIME. GENERALLY NORTH AMERICA IS THE LARGEST MARKET, BUT THIS HAS CHANGED RECENTLY WITH EUROPE NOW MAKING UP THE MOST SALES AT 38% MOST OF THEIR SALES WAS FOR RACING GAMES, PROBABLY BECAUSE BACK 2016 A FEW MAJOR RACING EVENTS HAPPENED LIKE GT WORLD CHALLENGE EUROPE ENDURANCE SCHEDULE SEASON.

REGIONAL SALES HAVE CHANGED OVER TIME

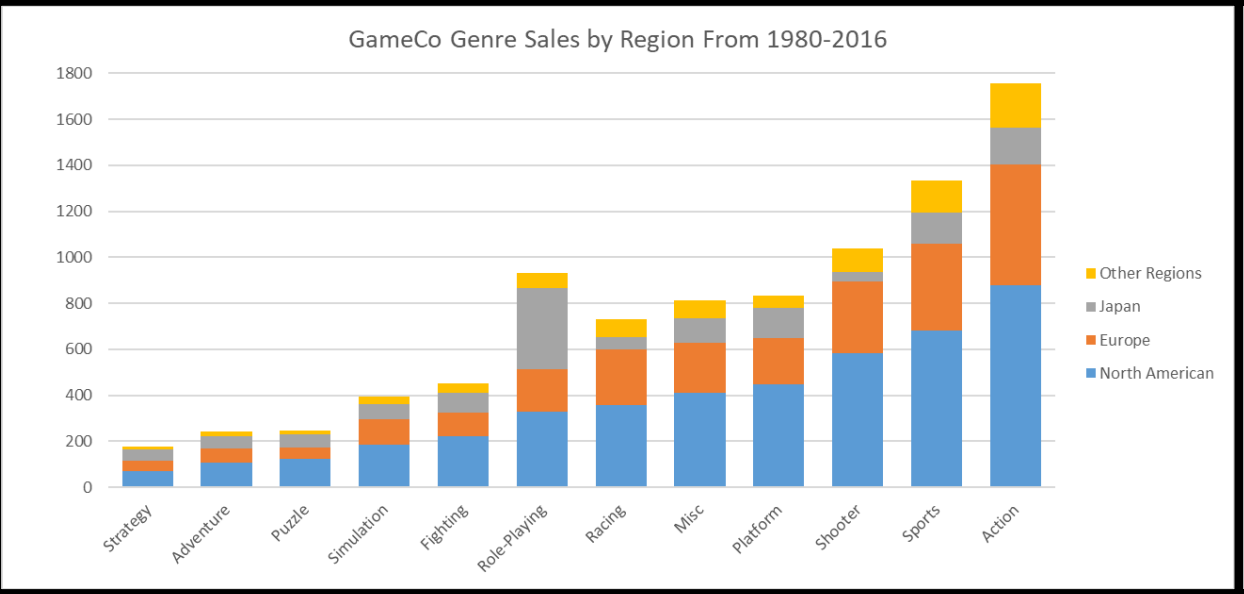
SALES BY REGION (2000-2016)

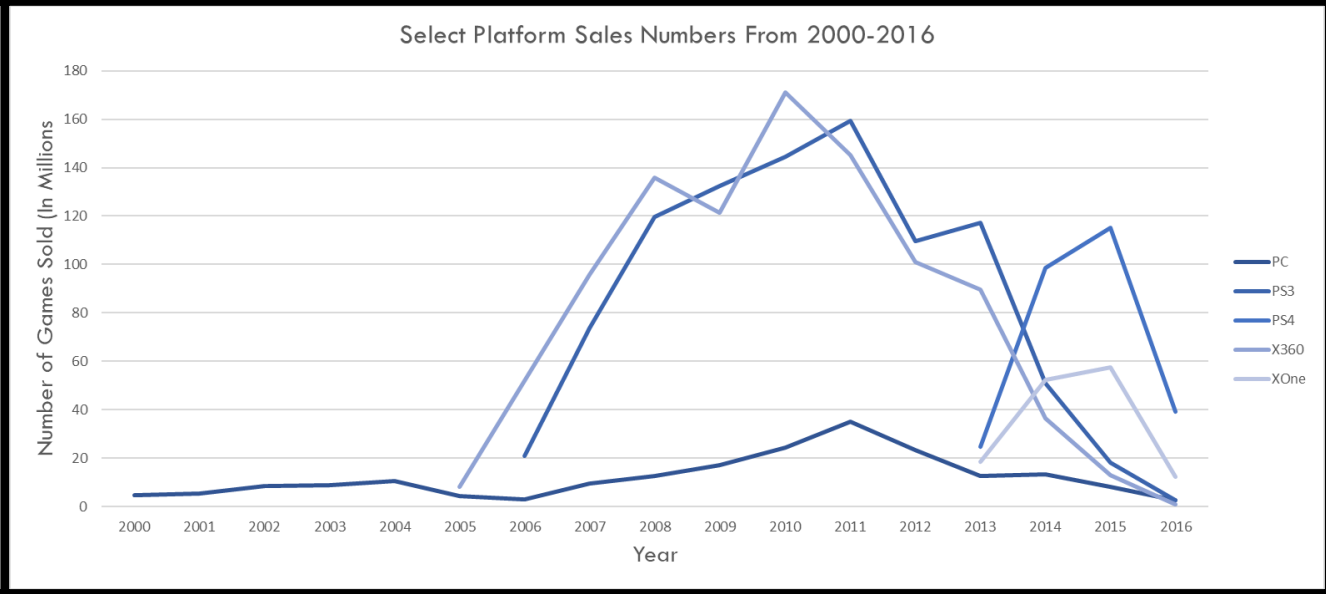


	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
NorthAmerican Sales	94.49	173.98	216.19	193.59	222.59	242.61	263.12	312.05	351.44	338.85	304.24	241.06	154.96	154.77	131.97	102.82	22.66
Europe Sales	52.75	94.89	109.74	103.81	107.32	121.94	129.24	160.65	184.7	191.74	176.88	167.44	118.78	125.95	125.65	97.71	26.76
Japan Sales	42.77	39.86	41.76	34.2	41.65	54.28	73.73	60.33	60.38	61.93	59.53	53.04	51.74	47.63	39.46	33.72	13.7
Other Sales	12.47	23.16	27.33	26.01	47.43	41.05	55.53	79.58	84.84	76.33	60.97	56.52	40.16	42.38	42.26	32.64	9.02

Observing the sales over time, we can see that sales has been declining over the past 5 years, one of the reason can be the advancement of technology on regards to video games, with the innovation of virtual games, now all the videos are played online.

PLATFORMS CHANGE, GENRES REMAIN THE SAME





The most popular genres are Action and Shooter games, but shooter games are still the third highest selling genre in 2016. All three genres are the most popular in Europe and North America. Role playing games with 38% followed by Strategy games with 28% are popular in Japan.

Shooter games with 56% are top of sales in North American and it is also the highest sales comparing with other regions.

Platforms experience high sales for a short period of time before being replaced with a new model. Per example PlayStation and Xbox consoles. The PC is the longest running platform, but only makes up 2.89% of all sales. We can continue to expect this trend with the release of new consoles in the future.

GameCo can also research future consoles that may be more relevant like virtual machines games.

STATISTIC SUMMARY
GAME CO

Measures of Central Tendency		North American Sales	Europe Sales	Japan Sales	Other Sales	Global Sales
Mean	Average	0.065872093	0.077790698	0.039825581	0.02622093	0.20619186
Median	Refers to the middle value	0	0.01	0.01	0	0.05
Mode	The most frequent value in the data.	0	0	0	0	0.02
RANGE	MAX - MIN	1.3	3.75	1.27	0.69	4.76
Q1	Lower Q1	0	0	0	0	0.02
Q3	Third Q3	0.0375	0.05	0.04	0.03	0.1775
IQR	Measures the spread of the data in a way that's less susceptible to outliers	0.0375	0.05	0.04	0.03	0.1575
MIN OUTLIER	Lower Outlier	-0.05625	-0.075	-0.06	-0.045	-0.21625
MAX OUTLIER	Larger Outlier	0.13630814	0.166686047	0.099738372	0.069331395	0.466787791

CONCLUSION AND RECOMMENDATIONS

- IN CONCLUSION WE CAN SAY ALL GAME SALES ARE DECLINING FOR ALL REGIONS IN THE PAST 3 YEARS. HOWEVER, NORTH AMERICAN IS THE NUMBER ONE ON SALES WITH THE MOST SELLS WHICH FALLS INTO TWO CATEGORIES: FIGHTING GAMES AND SHOOTER GAMES, TO BRING HIGHER SALES NUMBERS FOR THIS REGION, CONDUCT A MARKET RESEARCH TO FIND OUT WHY CONSUMERS STOP BUYING GAMES COMPARING TO PREVIOUS YEARS.
- EUROPE AND JAPAN HAVE THE SAME BEHAVIOR ON REGARDS TO SALES INCREASES, THEY ARE DETERMINED BY BIG EVENTS = CONSUMERS. USE BUDGET TO INCREASE ADVERTISING WHEN BIG EVENTS HAPPENS IN 2017 FOR BOTH REGIONS, MAKE DEALS WITH EVENTS SPONSORS; IF NOT, FOLLOW THE TOP 5 GENRE TO CONDUCT A MARKET RESEARCH TO DEVELOP MORE GAMES AND PLATFORMS FOR THOSE GENRE.
- CONSIDER TO USE PART OF THE BUDGET TO INCREASE THE BUDGET FOR OTHER REGIONS. IN 2016 THE TOP GENRE WAS MISCELLANEOUS THAT MEANS THE CONSUMERS FOR THAT REGION ENJOY A VARIETY OF GAMES THAT DOES NOT NECESSARY FALL INTO SPECIFIC GENRE, PERHAPS THEY GO BY THE MOST POPULAR GAME IN THAT MOMENT. A MARKET RESEARCH IS HIGHLY SUGGESTED HERE.

Final Report: GITHUB Back to Table of Contents

INFLUENZA SEASON

PREPARE FOR THE NEXT INFLUENZA SEASON IN THE UNITED STATES BY HELPING A MEDICAL STAFFING AGENCY DETERMINE WHEN, WHERE, AND HOW MANY STAFF TO SEND TO EACH STATE.

PEOPLE PART OF VULNERABLE POPULATION ARE MORE LIKELY TO DEVELOP COMPLICATIONS AND BECOME HOSPITALIZED FROM THE FLU.

PROJECT OVERVIEW

GOAL:

TO HELP THE MEDICAL STAFFING AGENCY THAT PROVIDES TEMPORARY WORKERS TO CLINICS AND HOSPITALS ON PLANNING FOR INFLUENZA SEASON, AND A TIME WHEN ADDITIONAL STAFF ARE IN HIGH DEMAND.

MOTIVATION:

HOSPITALS AND CLINICS NEED ADDITIONAL STAFF TO ADEQUATELY TREAT PATIENTS, PARTICULARLY THOSE IN VULNERABLE POPULATIONS WHICH ARE DEFINED BY THE CDC (CENTER FOR DISEASE CONTROL AND PREVENTION) AS ADULTS OVER AGE 65, CHILDREN UNDER 5, PREGNANT PEOPLE, INDIVIDUALS WITH HIV/AIDS, CANCER, HEART DISEASE, STROKE, DIABETES, ASTHMA, AND CHILDREN WITH NEUROLOGICAL DISORDERS., WHO DEVELOP SERIOUS COMPLICATIONS AND END UP IN THE HOSPITAL.

OBJECTIVE:

TO DETERMINE THE TIME AND QUANTITY OF TEMPORARY MEDICAL STAFF, TO EACH STATE.

SCOPE:

THE AGENCY COVERS ALL HOSPITALS IN EACH OF THE 50 STATES OF THE UNITED STATES, AND THE PROJECT WILL PLAN FOR THE UPCOMING INFLUENZA SEASON.

SKILLS:

- DESIGNING A DATA RESEARCH PROJECT.
- SOURCING, CLEANING, AND PROFILING DATA.
- DATA INTEGRATION AND TRANSFORMATION.
- STATISTICAL HYPOTHESIS TESTING.
- VISUAL ANALYSIS AND FORECASTING.
- VISUALIZATION IN TABLEAU.

TOOLS:

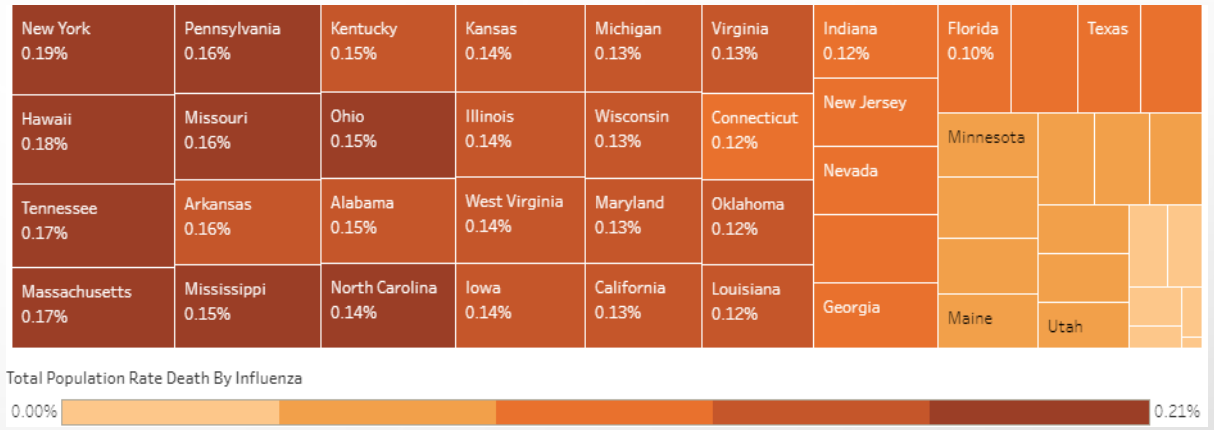
+ a b | e a u

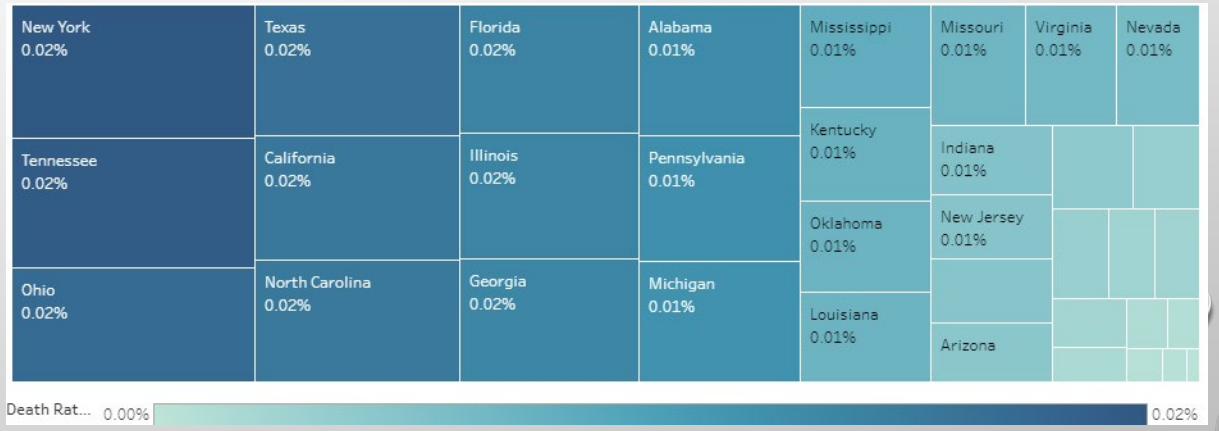
DATA:

- U.S. CENSUS POPULATION DATA BY GEOGRAPHY FROM THE U.S. CENSUS BUREAU AGENCY 2009-2017
- INFLUENZA DEATHS FROM CDC (CENTER FOR DISEASE CONTROL AND PREVENTION) 2009-2017
- INFLUENZA VISITS FROM CDC (FLUVIEW) 2010-2019
- PROJECT BRIEF

POPULATION AGE AND PERCENTAGE OF DEATHS

RELATION BETWEEN TOTAL POPULATION DEATH BY INFLUENZA 65+ YEARS OLD & UNDER 65+ YEARS OLD (2009 - 2017)



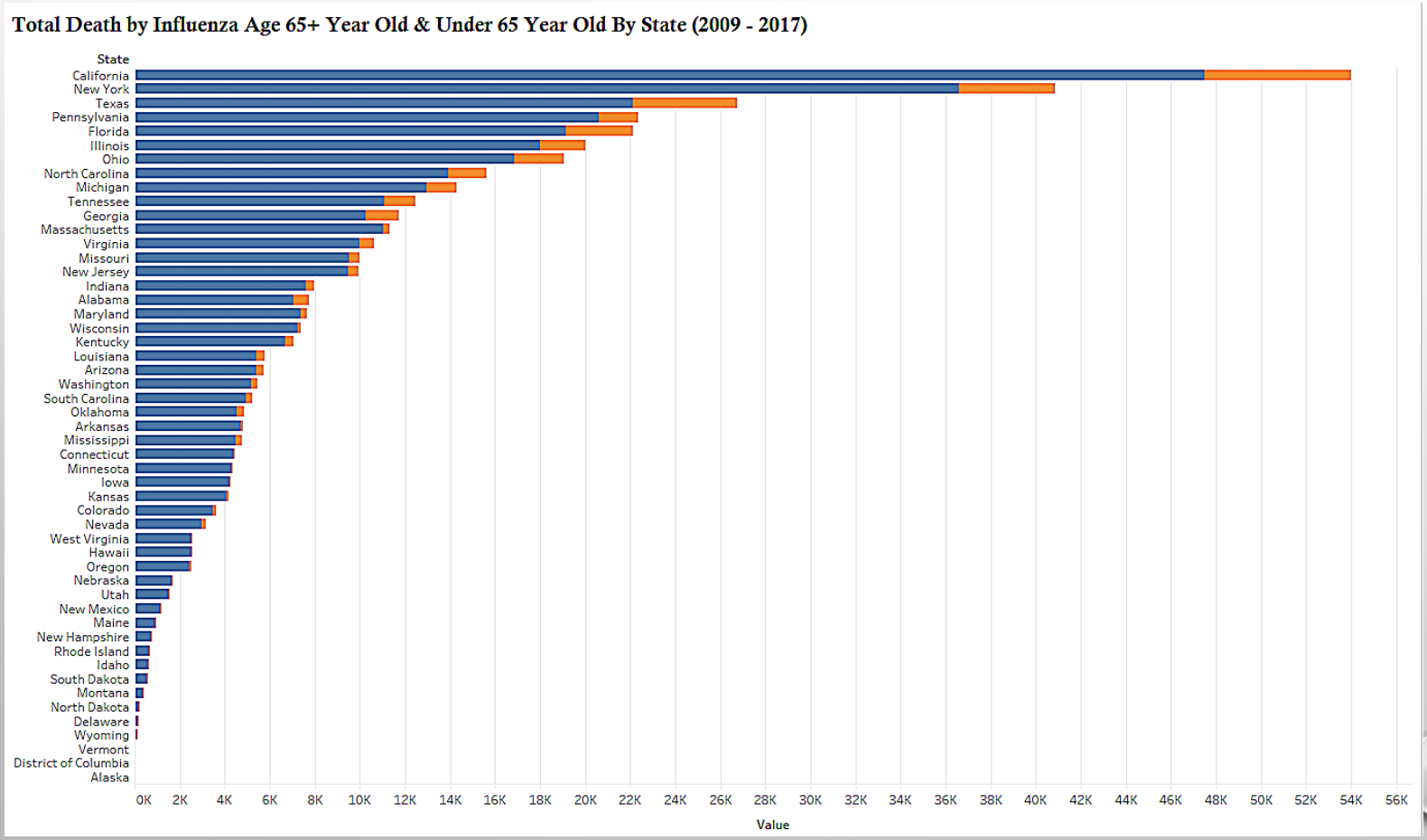


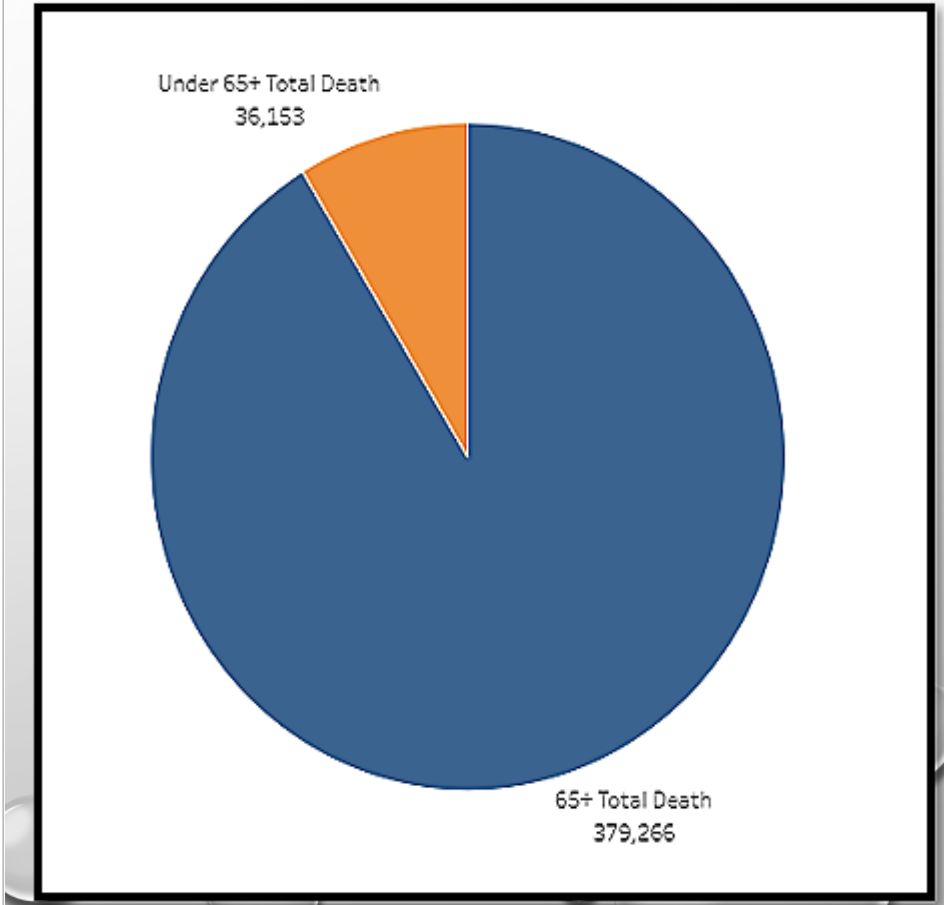
- COMPARISON OF THE POPULATION BY AGE ACROSS THE UNITED STATES. WE CAN SEE THE VULNERABLE POPULATION IS FROM 65+, AS WE CAN SEE THE MOST POPULATED STATES WITH CITIZENS OVER 65+ YEARS OLD HAVE THE HIGHER DEATHS BY INFLUENZA.

VULNERABLE POPULATION ANALYSIS

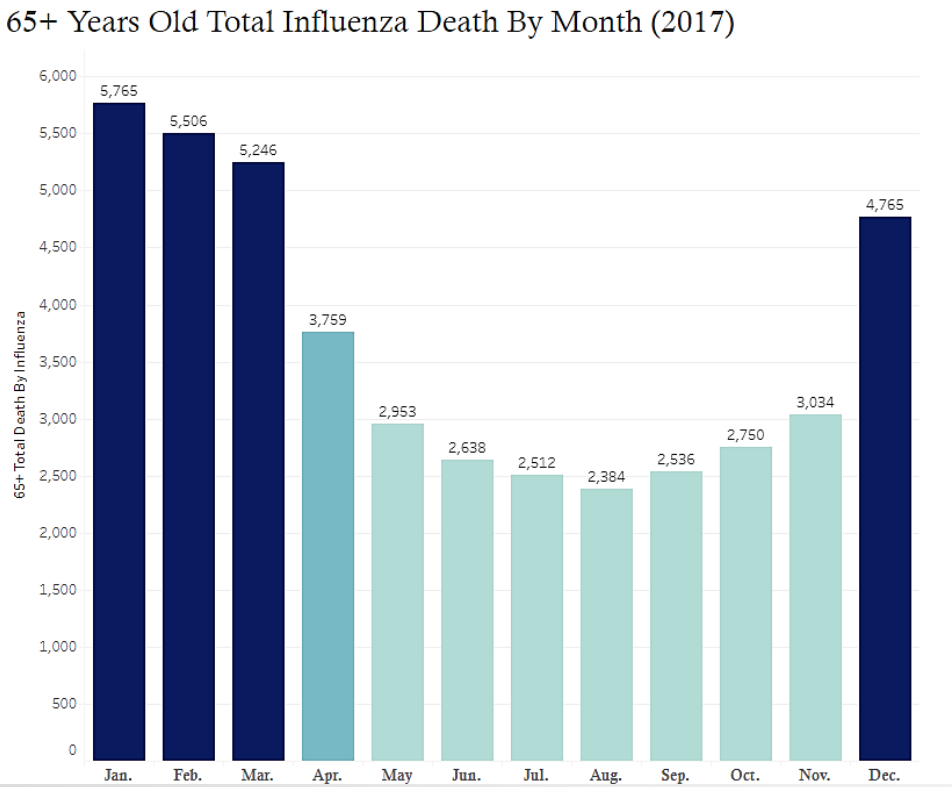
We do not have data for all the vulnerable population (those under age 5 or having certain pre-existing conditions) data was excluded for privacy matters, but we do have data for the population age 65 and over. More people ages 65 and older die from influenza than those under age 65, unfortunately it is for all states from (2009-2017.)

States with the most deaths continue to be those with large populations and large vulnerable populations.

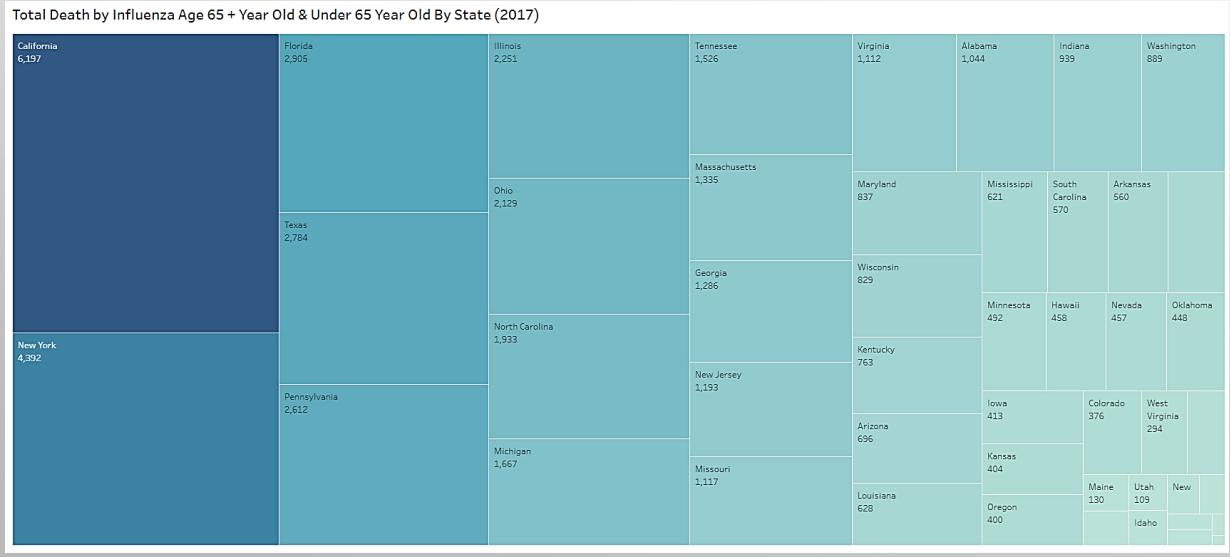




SEASONALITY AND STAFF ALLOCATION

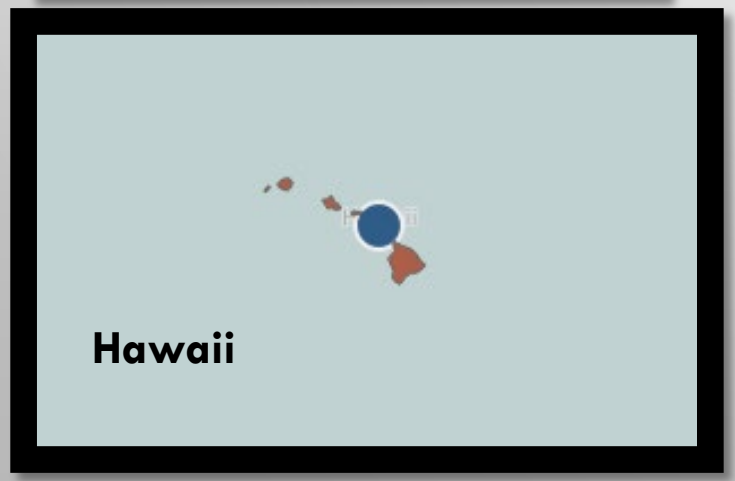


INFLUENZA HIGH SEASON STARTS FROM DECEMBER TO MARCH. THAT IS WHEN INFLUENZA DEATHS ARE AT THEIR HIGHEST. SOME STATES, LIKE FLORIDA AND NEW JERSEY, SHOW AN INCREASE IN THE MONTHS BEFORE THE HIGH SEASON.

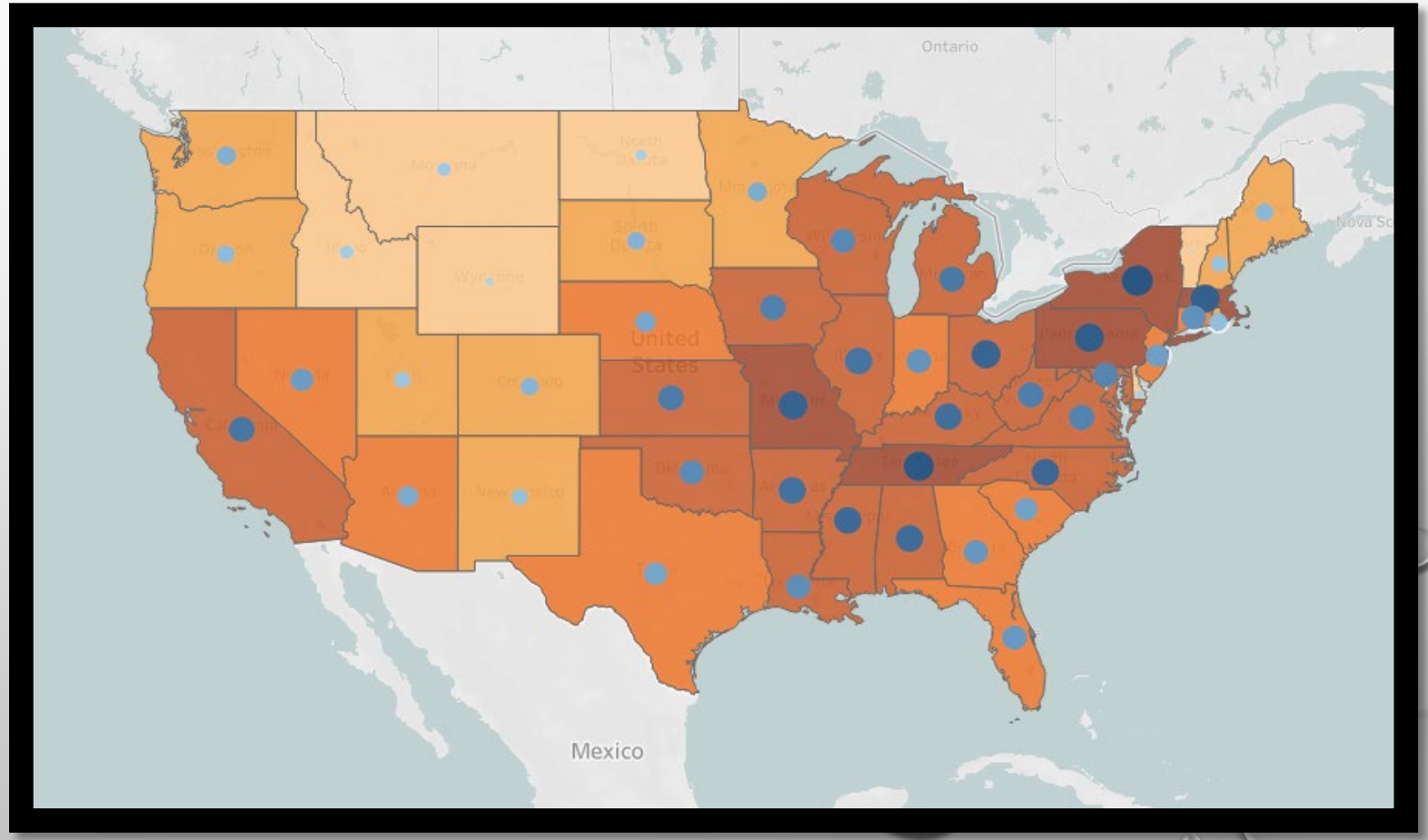


TREE MAP IS BASED ON EACH STATE'S TOTAL VULNERABLE POPULATION BY DEATH IN 2017. WE CAN SEND MORE STAFF TO AREAS WITH MORE PEOPLE AT RISK TO PREVENT DEATHS.

PRIORITIZING STATES



This map breaks down states by priority using the average population of those over 65 and the average number of deaths.
The high priority states are New York, Tennessee, Florida, Hawaii, Missouri, North Carolina.



RECOMMENDATIONS AND NEXT STEPS

1. SEND RESULTS AND DETERMINE NEXT STEPS IN COLLABORATION WITH THE STAKEHOLDERS.

2. FOCUS VACCINATION CAMPAIGNS FOR THE GROUP AGE 65+ BY PRIORITIZING THAT CAMPAIGN IN THE AGE GROUP.

3. INCREASE THE STAFF NUMBER IN THOSE STATES WITH THE HIGHER RATES OF MORTALITY BY INFLUENZA, START SENDING STAFF BY OCTOBER-NOVEMBER.

4. PROVIDE EDUCATIONAL MATERIAL TO INCREASE KNOWLEDGE TO CHANGE THE PERCEPTIONS TOWARDS INFLUENZA TO INCREASE HEALTHY AND PREVENTIVE BEHAVIORAL PRACTICES, ESPECIALLY IN THE MOST VULNERABLE POPULATION.

Final Report: GITHUB Back to Table of Contents

ROCKBUSTER STEALTH LLC

A FICTIONAL MOVIE RENTAL COMPANY WITH PHYSICAL STORES AROUND THE GLOBE, ROCKBUSTER STEALTH LLC IS PLANNING TO LAUNCH AN ONLINE VIDEO RENTAL SERVICE.

PURPOSE OF THE ANALYSIS IS TO HELP WITH THE LAUNCH STRATEGY FOR THE NEW ONLINE VIDEO SERVICE.

PROJECT OVERVIEW

GOALS:

ROCKBUSTER STEALTH LLC PLANS TO LAUNCH AN ONLINE MOVIE RENTAL PLATFORM TO COMPETE WITH OTHER STREAMING SERVICES.

KEY QUESTIONS:

1. WHICH MOVIES CONTRIBUTED THE MOST/LEAST TO REVENUE GAIN?
2. WHICH COUNTRIES ARE ROCKBUSTER CUSTOMERS BASED IN?
3. WHAT WAS THE AVERAGE RENTAL DURATION FOR ALL VIDEOS?
4. DO SALES FIGURES VARY BETWEEN GEOGRAPHIC REGIONS?
5. WHERE ARE CUSTOMERS WITH A HIGH LIFETIME VALUE BASED?
6. WHAT GENRE ARE POPULAR BY REVENUE?

SKILLS:

- DATA EXTRACTION AND CLEANING.
- DATA EXPLORATION.
- DATA PREPROCESSING.
- SQL AND DATABASE MANAGEMENT: QUERYING AND EXTRACTING DATA.
- JOINING TABLES
- PERFORMING SUBQUERIES
- USING CTES (COMMON TABLE EXPRESSIONS).
- CREATING A DATA DICTIONARY

DATA:

- ROCKBUSTER STEALTH LLC DATA
- PROJECT BRIEF

TOOLS:

SUMMARY STATISTICS

Total Revenue
$61,312

Average movie rental days
5

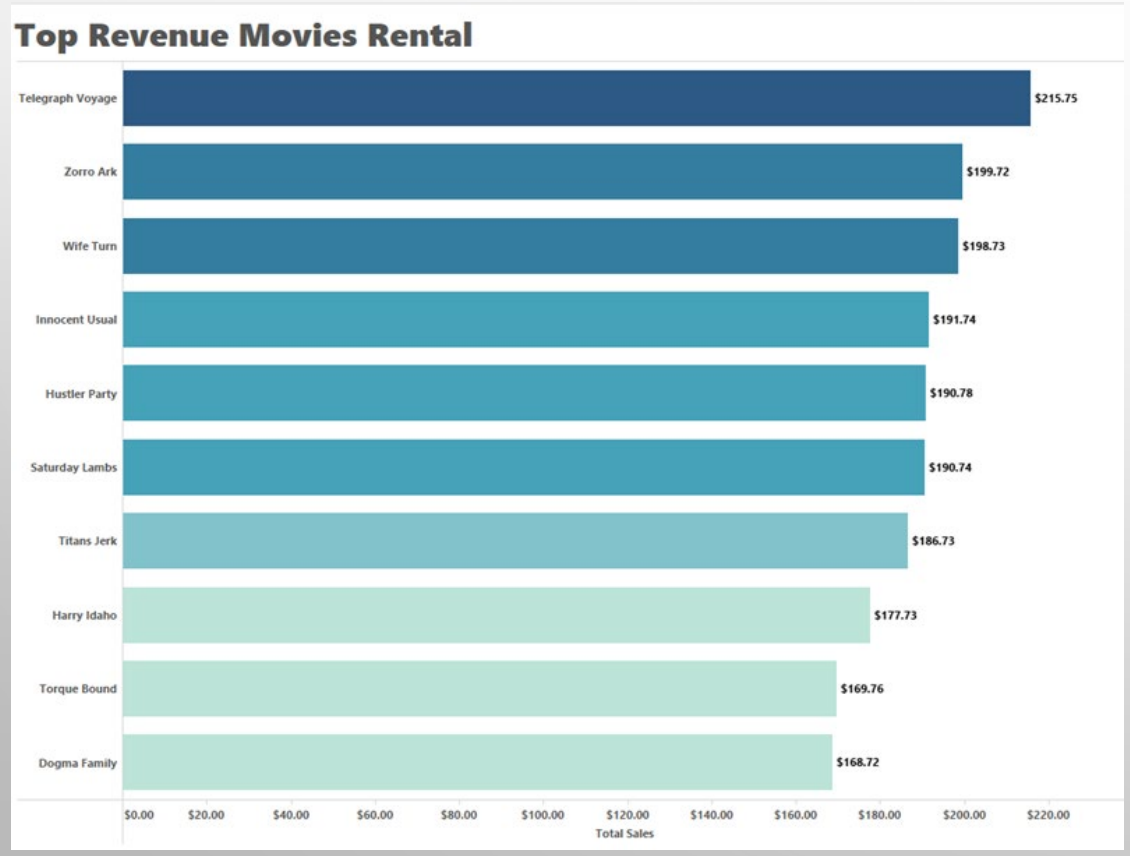
Most common movie rate
PG-13

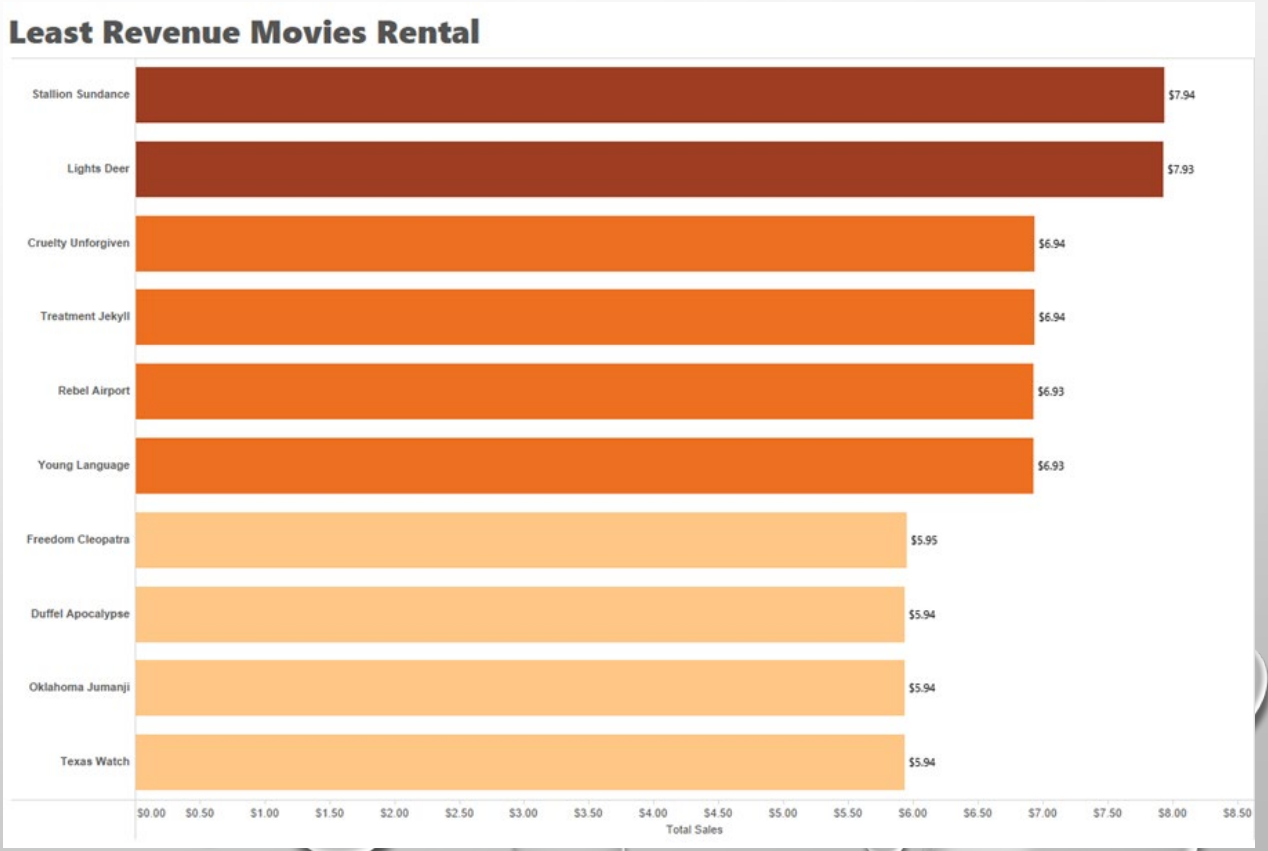
Average replacement cost
$19.98

Total active customers
599

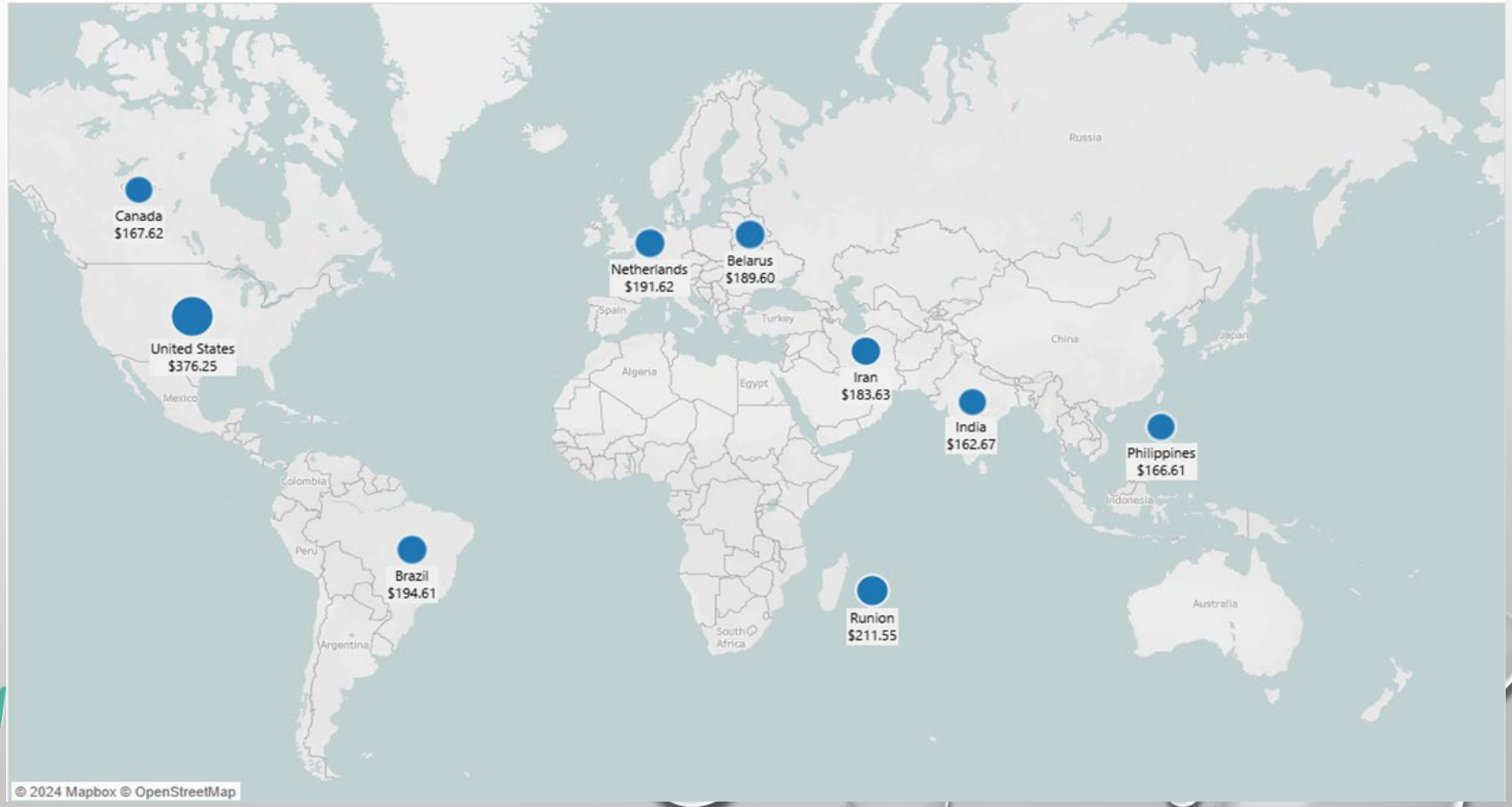
WHAT MOVIES CONTRIBUTED MOST/LEAST TO REVENUE GAIN?

The top 10 revenue generating movies all have the maximum rental rate of $4.99, but their average rental duration is short at 4 days.
The bottom 10 revenue generating movies all have the minimum rental rate of $0.99, and a longer average rental duration of 6 days.

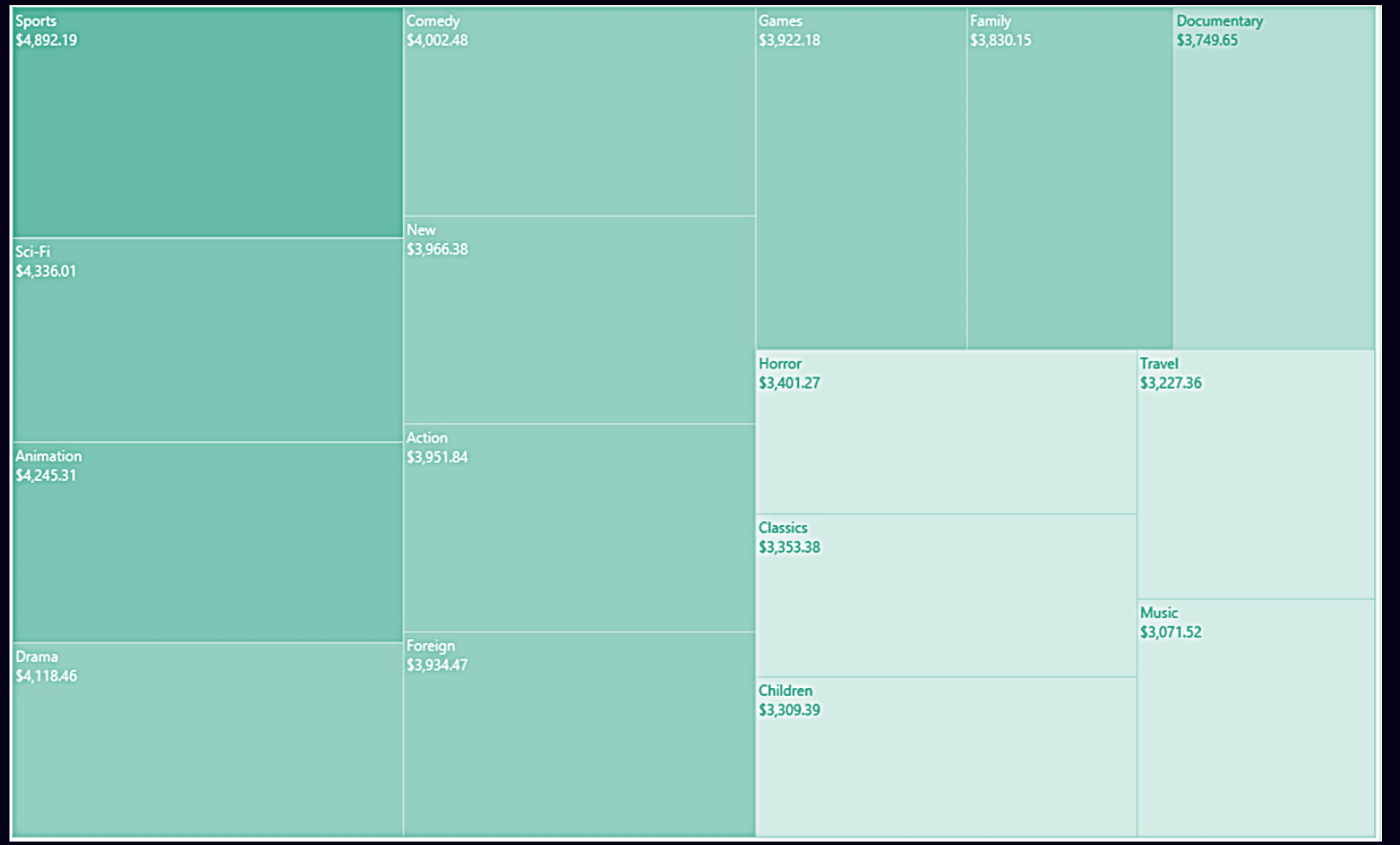




WHERE ARE CUSTOMERS WITH A HIGH LIFETIME VALUE BASED?



WHAT GENRE ARE POPULAR BY REVENUE?



RECOMMENDATIONS

Top-Performing

Focus on the top genres and ratings, like PG_13 and Sports Increase the variety and offer promotions and deals.

Customer Experience

Send surveys to lifetime customers to get their input on the new online video strategy.
Offer annual membership with unlimited streaming to all customers.
Launch promotions and deals to increase customers activity.

Update Inventory

Take the films out of the inventory that produced no revenue and replace them with licensing for new release movies.

Market Expansion

Asia: Since Asia contributes the most revenue in sales, focus investment on marketing and customer engagement in this region could bring substantial returns.
Underperforming Regions: Create strategies to boost revenues in regions like Oceania Central America and The Caribbean, which currently have lower revenue contributions.

Final Report: GITHUB Back to Table of Contents

INSTACART

AN ONLINE GROCERY STORE

THAT OPERATES THROUGH AN APP.

THE ANALYSIS IS TO HELP FINDING MORE INFORMATION ABOUT THEIR SALES PATTERNS, PURCHASES BEHAVIORS AND THE VARIETY OF CUSTOMERS.

PROJECT OVERVIEW

GOALS:

INSTACART, AN ONLINE GROCERY STORE THAT OPERATES THROUGH AN APP. THE INSTACART STAKEHOLDERS ARE CONSIDERING A TARGETED MARKETING STRATEGY, MY ANALYSIS WILL INFORM WHAT THIS STRATEGY MIGHT LOOK LIKE TO ENSURE INSTACART TARGETS THE RIGHT CUSTOMER PROFILES WITH THE APPROPRIATE PRODUCTS.

KEY QUESTIONS:

- WHAT THE BUSIEST DAYS OF THE WEEK AND HOURS OF THE DAY ARE IN ORDER TO SCHEDULE ADS AT TIMES WHEN THERE ARE FEWER ORDERS.
- PARTICULAR TIMES OF THE DAY WHEN PEOPLE SPEND THE MOST MONEY, AS THIS MIGHT INFORM THE TYPE OF PRODUCTS THEY ADVERTISE AT THESE TIMES.
- USE SIMPLER PRICE RANGE GROUPINGS TO HELP DIRECT THEIR EFFORTS.
- ARE THERE CERTAIN TYPES OF PRODUCTS THAT ARE MORE POPULAR THAN OTHERS? THE MARKETING AND SALES TEAMS WANT TO KNOW WHICH DEPARTMENTS HAVE THE HIGHEST FREQUENCY OF PRODUCT ORDERS.
- WHAT'S THE DISTRIBUTION AMONG USERS IN REGARDS TO THEIR BRAND LOYALTY?
- ARE THERE DIFFERENCES IN ORDERING HABITS BASED ON A CUSTOMER'S LOYALTY STATUS?
- IS THERE A CONNECTION BETWEEN AGE AND FAMILY STATUS IN TERMS OF ORDERING HABITS?

THIS IS A LARGE DATASETS, THE LARGE SIZE OF DATASETS, I WILL USE PYTHON TO RETRIEVE THE INFORMATION WE NEED, ANALYZE IT, AND CREATE VISUALS.

SKILLS:

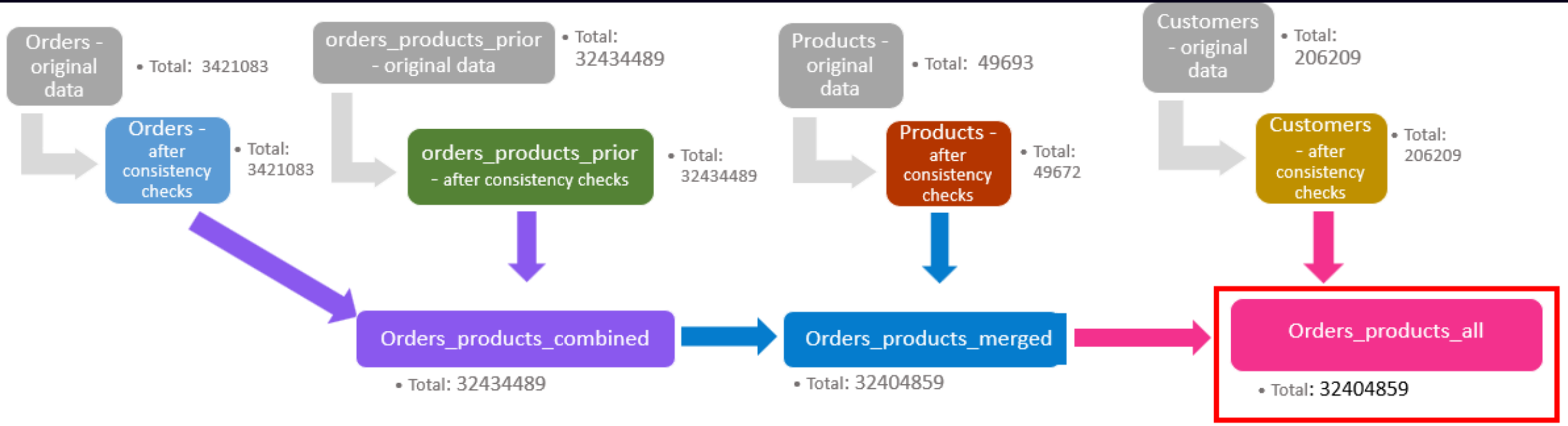
- DATA CLEANING, WRANGLING, SUBSETTING.
- COMBINING & EXPORTING DATA.
- DATA CONSISTENCY CHECKS.
- DERIVING NEW VARIABLES.
- GROUPING AND AGGREGATING DATA WITH PYTHON.
- VISUALIZATION WITH PYTHON.
- CODE ETIQUETTE $ EXCEL REPORTING.

DATA:

- DATA IS THE INSTACART ONLINE GROCERY SHOPPING DATASET 2017, ACCESSED FROM HTTPS://WWW.INSTACART.COM/DATASETS/GROCERY-SHOPPING-2017 VIAKAGGLE ON NOV 15TH, 2023.
- DATA DICTIONARY WAS PROVIDED
- PROJECT BRIEF.

TOOLS:

POPULATION FLOW

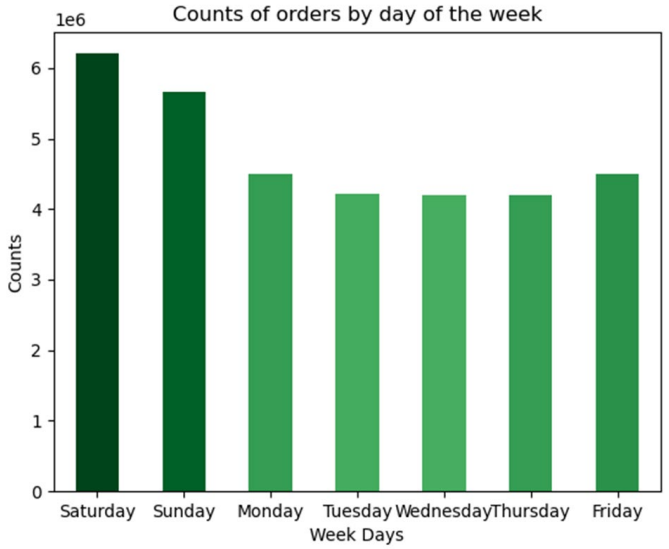


1.) The grey boxes in the first row of the population flow represent the original data sets as they were when I downloaded them.

2.) The second row of boxes (colored) represents the data sets after I manipulated them, e.g., removed missing values and duplicates. This offers a visual overview of how the data flows throughout the data consistency checks.

3.) The third row, where also the arrows are colored, represents the merges I performed between the datasets.

ORDERING AND PRICING FINDINGS

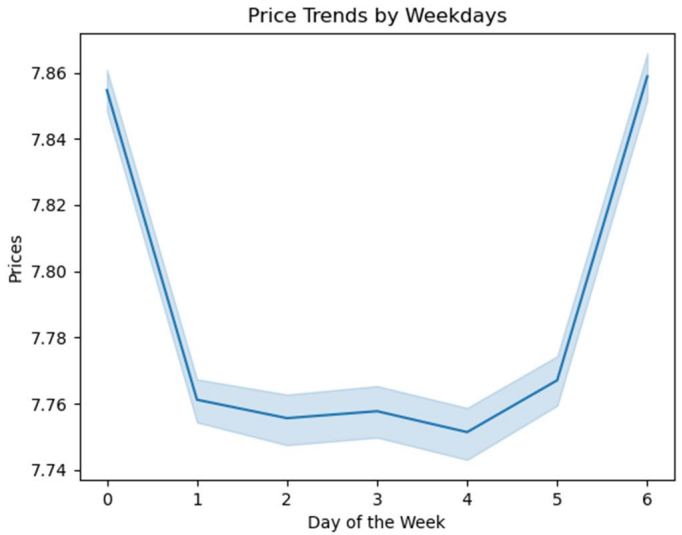


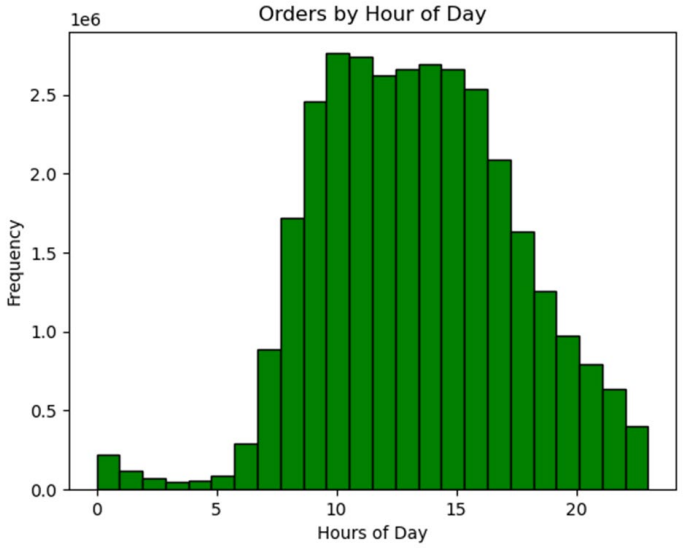
The busiest days of the week are weekends, Saturday and Sunday. Busiest hours of day for ordering are between 9-16 (9:00 am - 4:00 pm).

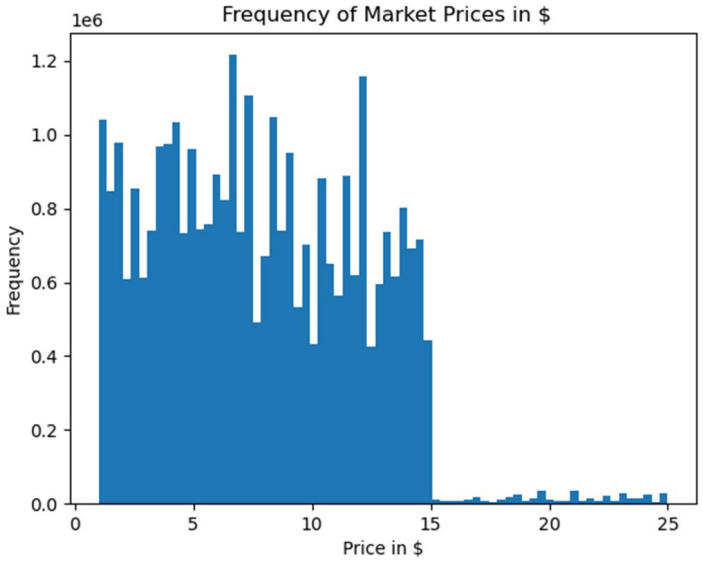
Ads should be run on weekdays before 9AM or after 4PM.

The peaks on days zero and six mean that most money is spent on Friday and Saturday.
This might be due to people stocking up on things before the weekend.

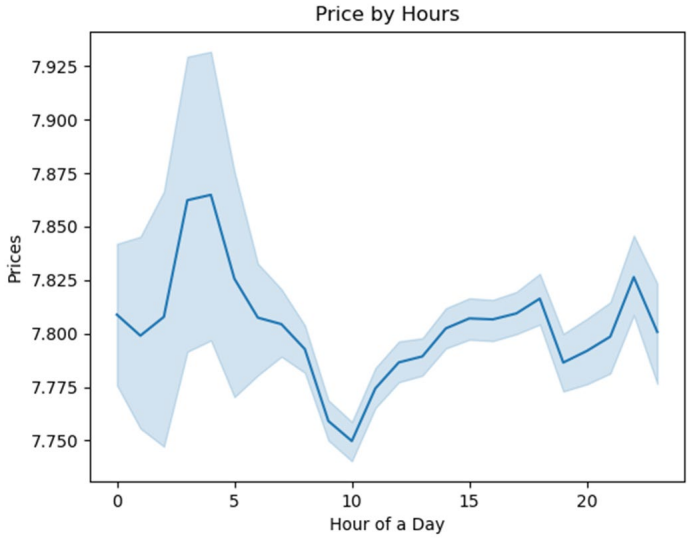
Price volume distribution appear to be steady and do not change much during the day. It fluctuates between 7.750 and 7.850.







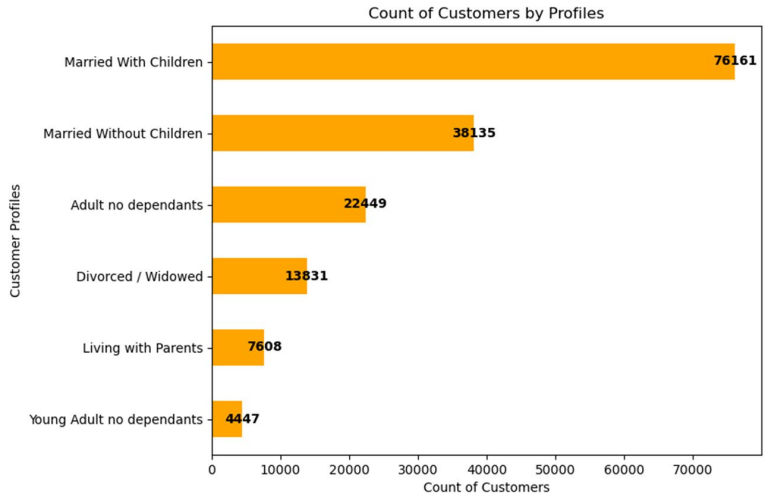
Most products are between $1 and $15, while a few are higher priced at $15 to $25.



CUSTOMER PROFILE

In terms of Loyalty, Regular customers orders more products in terms of order counts than loyal customers or new customer.

Married with children are most of the customers who shop on Instacart.



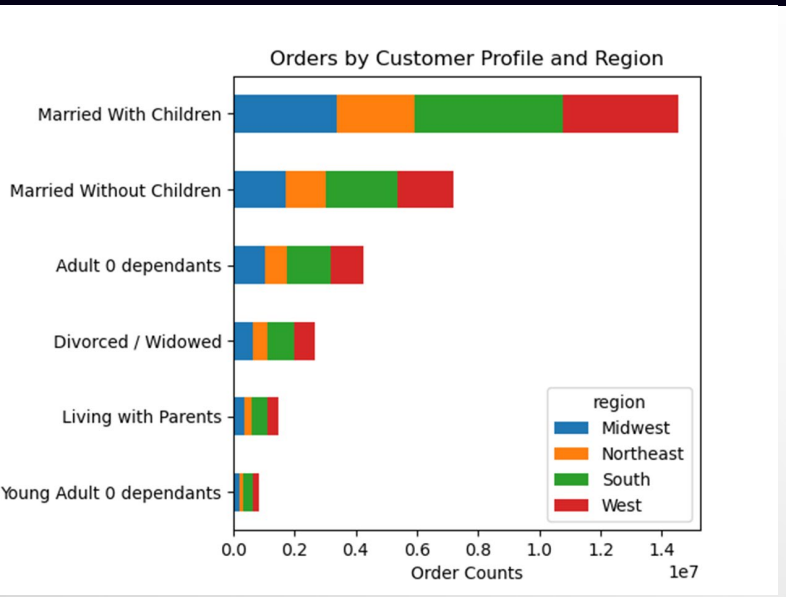
Based on the scatter plot, an income increases after the age of 40+

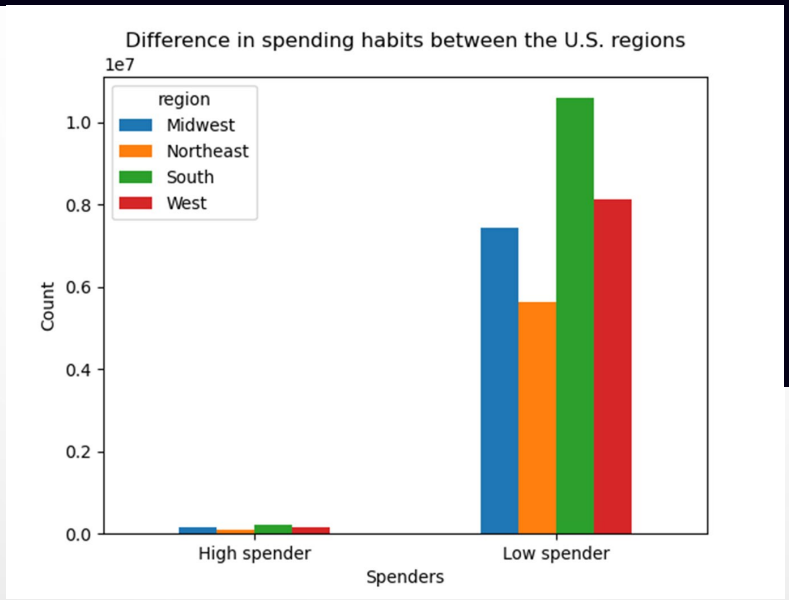
Married with children also bring the highest revenue to Instacart.

Revenue By Customer Profiles

customer_profile	Revenue	Count	Revenue/Count
Adult 0 dependants	33230281.9	4268305	7.79
Divorced / Widowed	20632363.5	2644831	7.8
Living with Parents	11552531.1	1480979	7.8
Married With Children	113156281.6	14530234	7.79
Married Without Children	56205930.2	7209966	7.8
Young Adult 0 dependants	6416847.3	825372	7.77

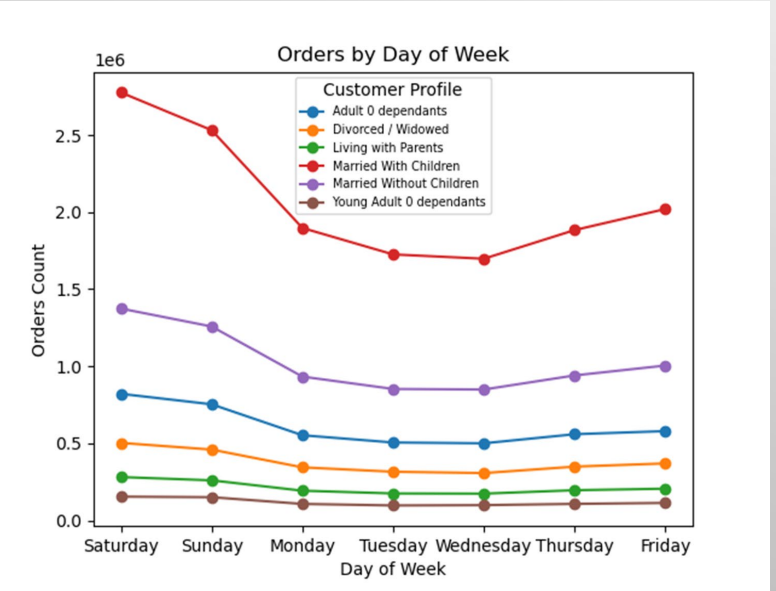
PROFILING BY REGIONS & SPENDERS

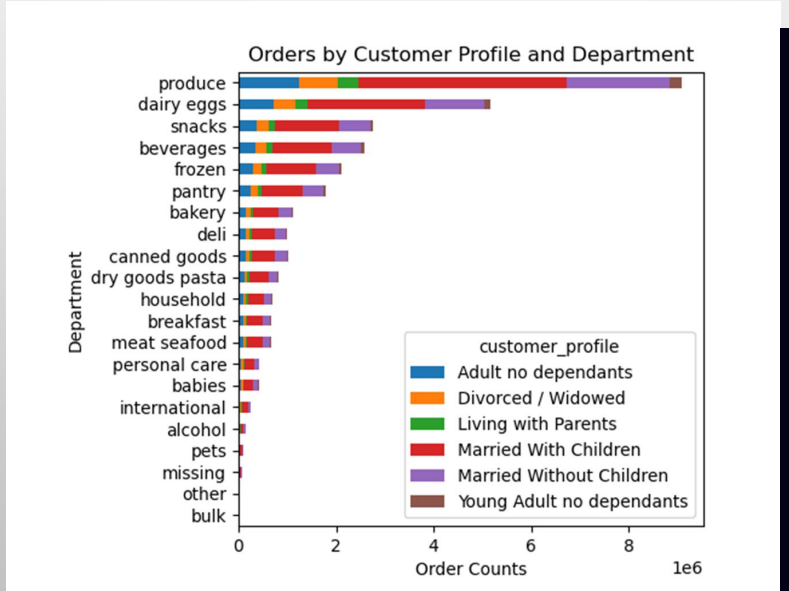




Customers married with children order more in any region comparing to the other profiles.

Regarding spending habits seems that customers do not spend much in pricey products, most likely they will spend on the basics they will need.



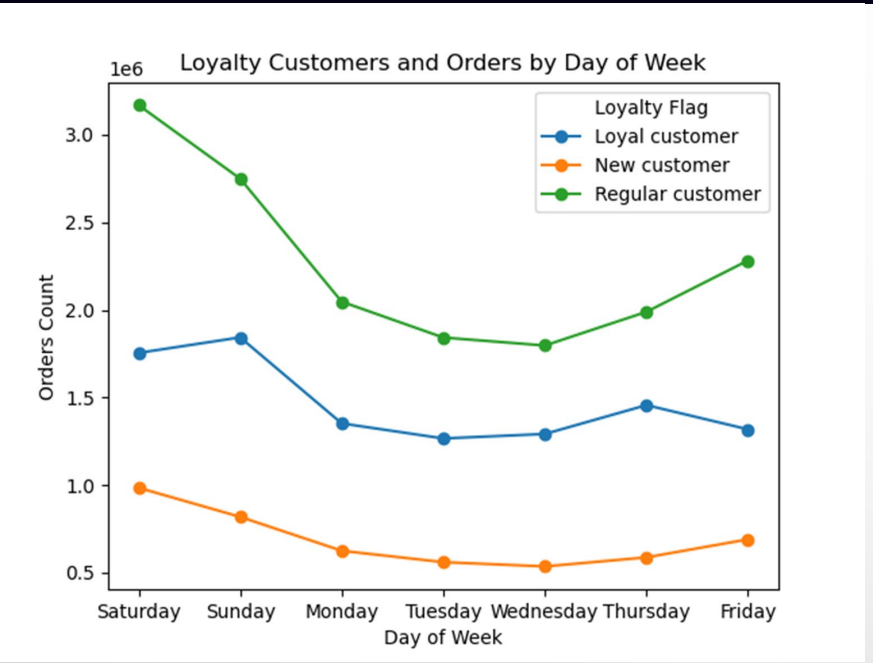


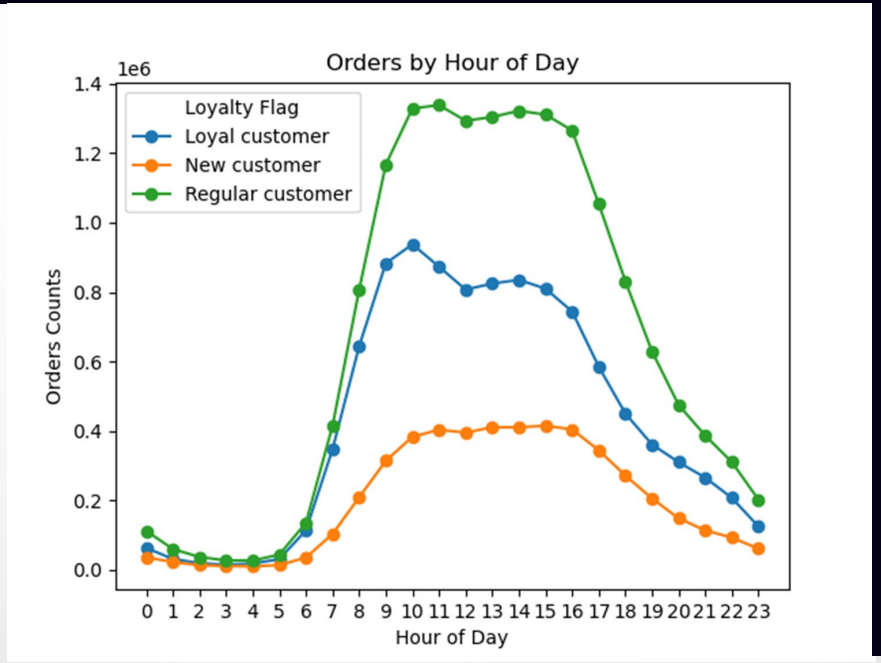
Customers married with children are the ones that have most of the orders placed in Instacart.

The peak days are Thursday, Friday and Saturday, then we see a declination of orders on Sunday, one of the reason could be they save that day for family gatherings in comparison with the other customers profiles which they pretty much maintain the same flow.

Produce and dairy eggs are the top orders from all customers, but the ones with most orders are the customers married with children.

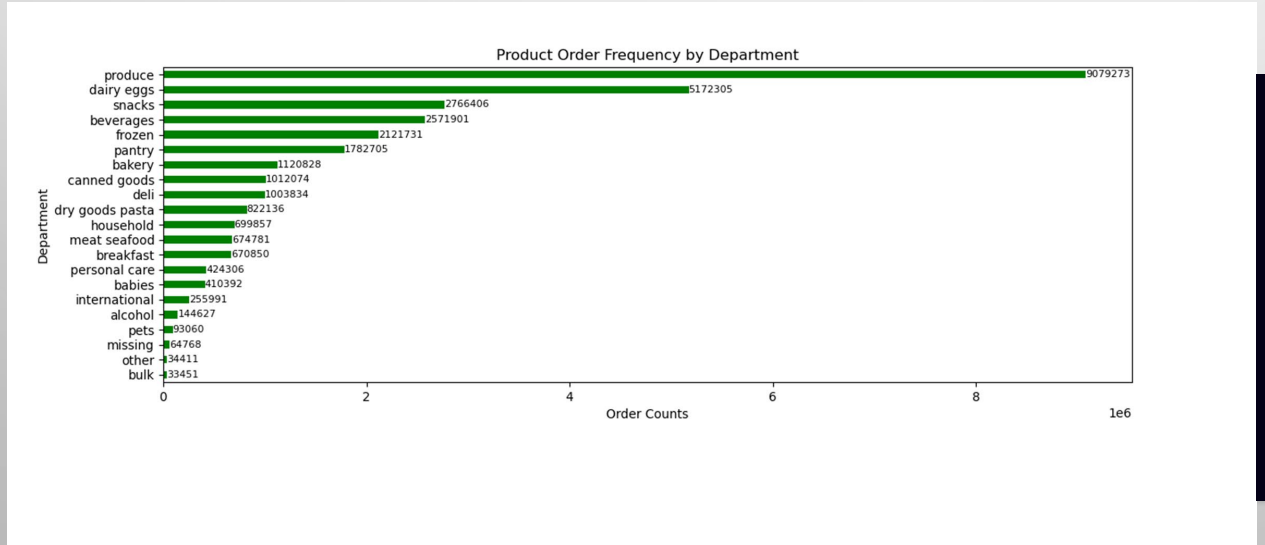
ADDITIONAL ANALYSIS TO ANSWER KEY QUESTIONS





Regular customers are the ones who orders more, we know now they are not big spenders, but they are the ones who order more frequent, especially during the weekend.

And between 9:00 am to 4pm are the most sales and where you can find the customers to be the majority regulars.



The top 5 departments frequently visit are: Produce, dairy eggs, snacks, beverages and frozen.

The top 5 department with minimum visits are: International, Alcohol, pets, others and bulk.

RECOMMENDATIONS

1. If you are targeting ads while people are already spending, ads should be targeted for Fri-Sun from 9am-4pm. The departments should focus on Produce (especially organic), Dairy & Eggs, Snacks, and Beverages.

2. Different groups of customers can be targeted thanks to our flags we created. These include:
a. Loyalty (how many orders placed)
b. Spending (how much they spend on average)
c. Frequency (how often they order)
d. Region (where do they live)

3. Regular Customers order more frequently, follow by Loyal Customer and the last New Customers. A savings program to reward members for ordering more frequently would increase consistency of sales. A survey for the new customers to get to know their shopping behavior, and Survey in general to find out the shopping habits of all customers.

4. In All 4 regions most of the orders are made by regular customers. Ads to target the loyal and new customer in all regions to increase sales and incentivized more orders placed by these two groups.

5. In all 4 regions low spenders dominate the orders trends. With a saving program and special offers for those products that are not popular may incentive all customers to spend more.

Final Report: GITHUB Back to Table of Contents

PIG E. BANK

A FICTIONAL GLOBAL BANK LOOKING TO INCREASE CLIENT RETENTION.

CUSTOMER DATA IS ANALYZED TO IDENTIFY FACTORS THAT CONTRIBUTE TO CLIENT EXITED. THESE FACTORS ARE THEN MODELED IN A DECISION TREE.

PROJECT OVERVIEW

GOALS:

THE GOAL IS TO ASSIST THE SALES TEAM OF PIG E. BANK INCREASE CUSTOMER RETENTION, BY IDENTIFYING THE LEADING INDICATORS THAT A CUSTOMER WILL LEAVE THE BANK.

TO DO THIS I WILL USE DATA MINING TECHNIQUES TO ASSESS THE QUALITY OF OUR DATA, CLEAN IT, AND THEN GENERATE SOME DESCRIPTIVE STATISTICS. FROM THERE I LOOKED AT DEMOGRAPHICS FIRST (AGE, GENDER, COUNTRY) AND THEN USAGE (ACTIVITY, NUMBER OF PRODUCTS, TENURE, BALANCES, SALARY, CREDIT CARD STATUS, CREDIT SCORE) TO SEE WHAT MIGHT BE TOP FACTORS IN CLIENTS EXITED.

THE TOP EXIT FACTORS IS PRESENTED IN A DECISION TREE MODEL AS PART OF A FINAL REPORT.

SKILLS:

- BIG DATA AND DATA ETHICS
- DATA MINING
- DATA QUALITY ASSESSMENT AND CLEANING
- DESCRIPTIVE STATISTICS
- PREDICTIVE ANALYSIS
- TIME SERIES ANALYSIS AND FORECASTING

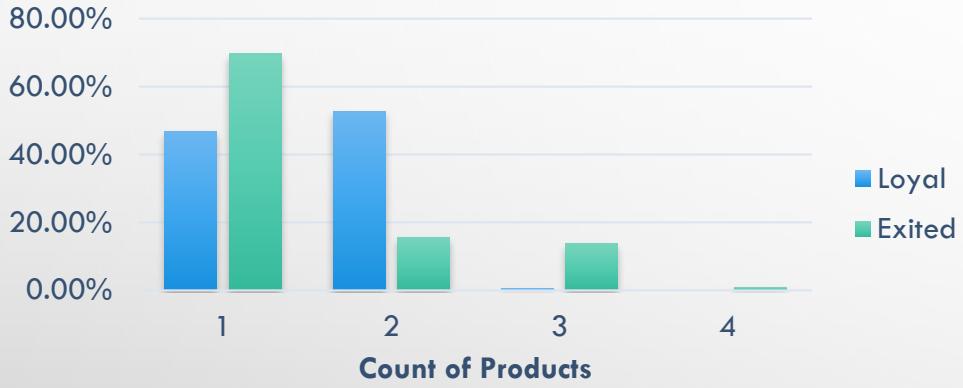
DATA:

- CUSTOMER DATA PROVIDED BY CAREER FOUNDRY.

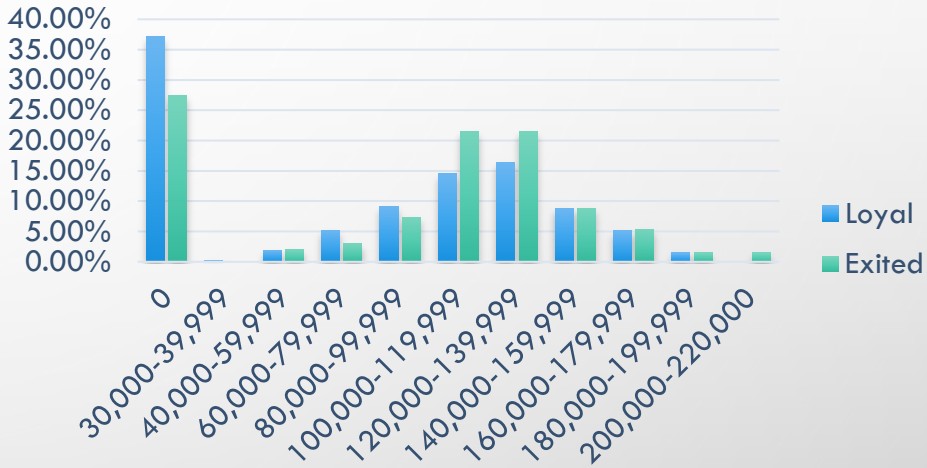
TOOLS:

LOYALTY BY ACTIVITY, BALANCE, PRODUCT

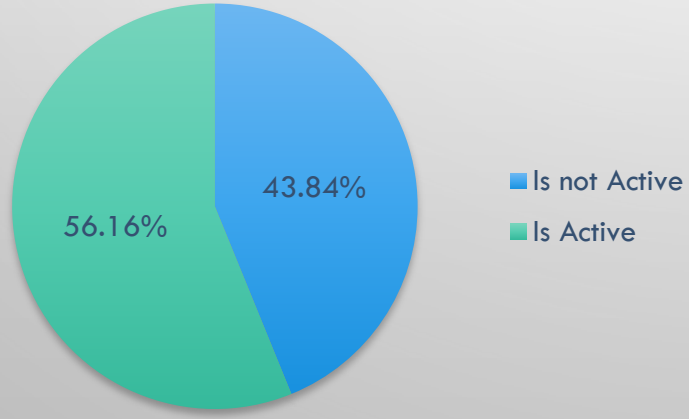
Loyalty By Products



Loyalty By Balance



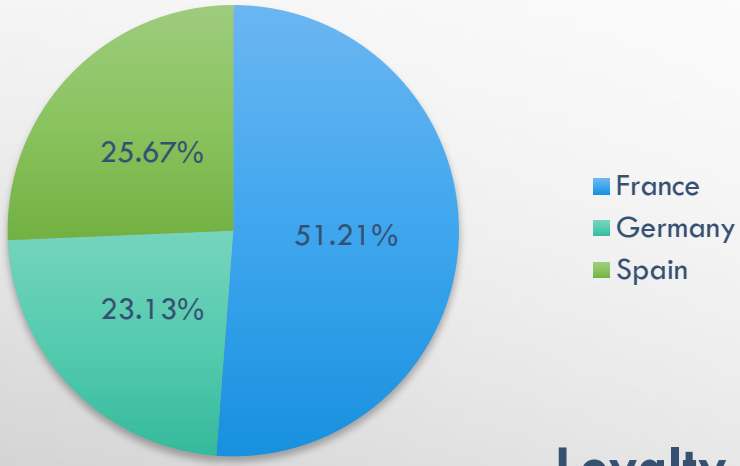
Loyalty By Activity



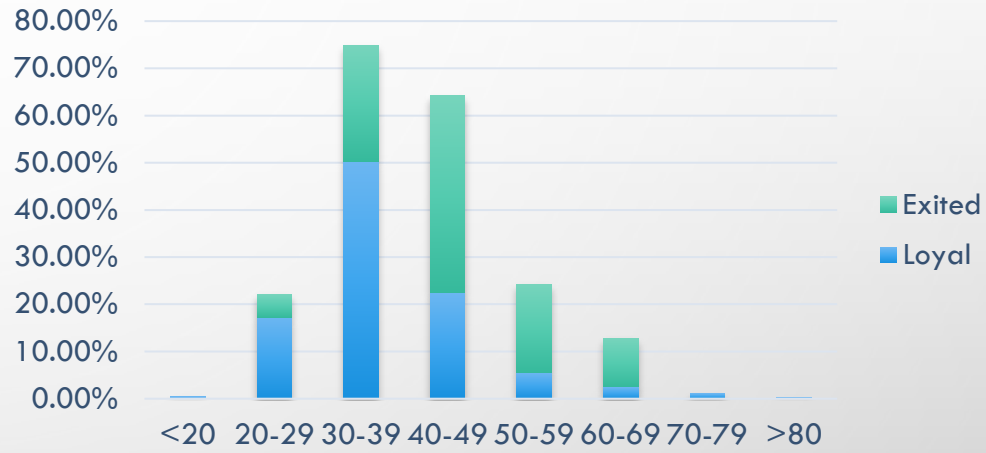
Leading Factors	
IsActive	70% of exited clients were not active, while 56% of loyal clients are active
Number Of Products	70% of exited clients had only 1 product, while 53% of loyal clients have 2 and more products
Balance	22% of exited clients leave with balances between 100,00 – 140,000

LOYALTY BY COUNTRY, AGE, GENDER

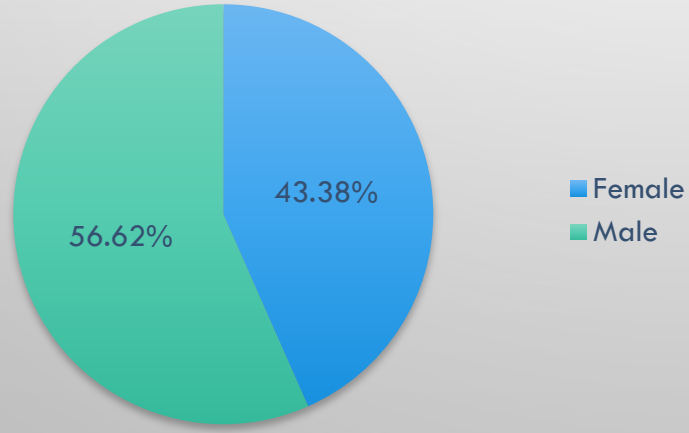
Loyalty By Country



Loyalty By Age



Loyalty By Gender



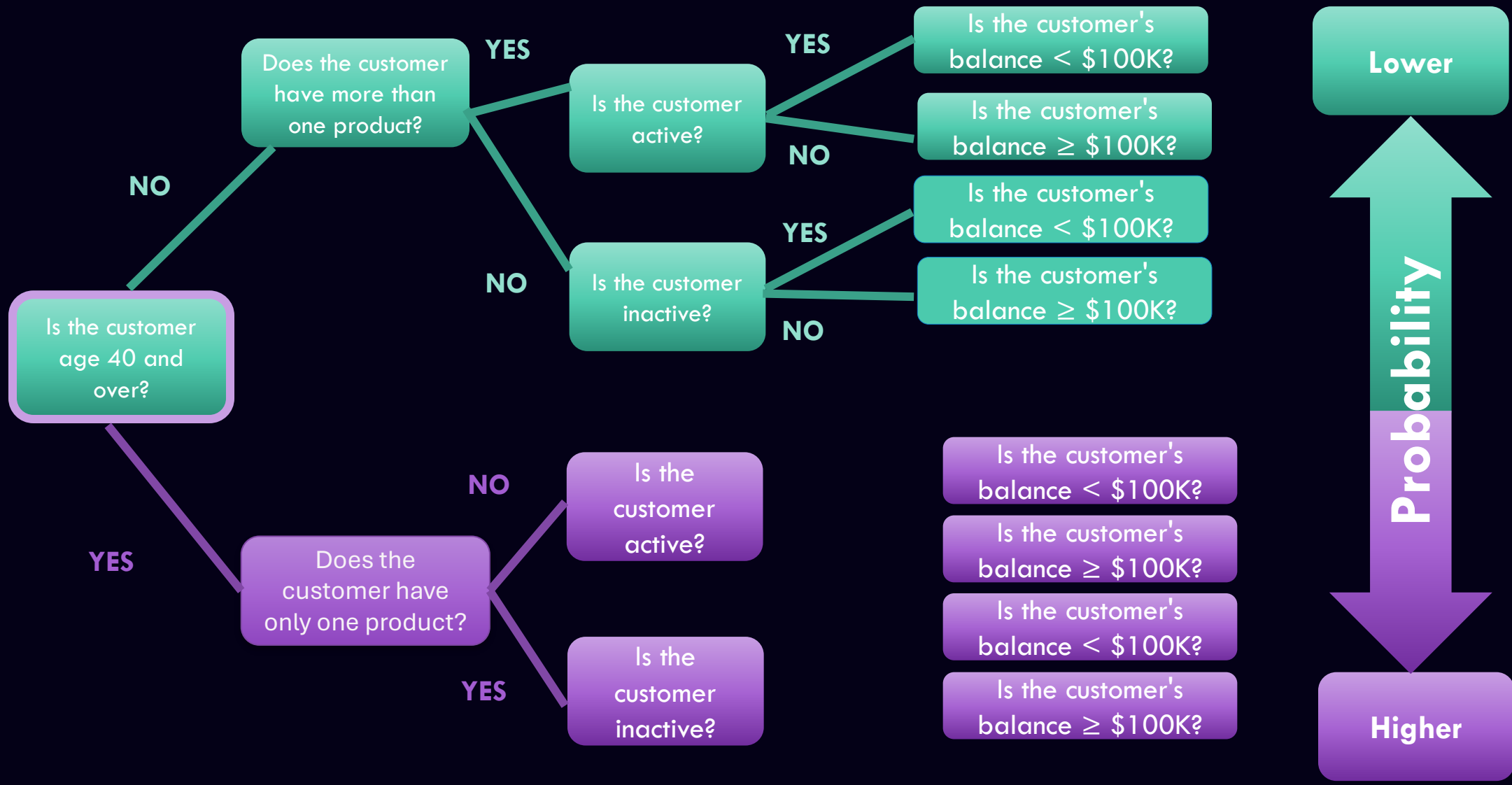
Leading Factors	
Country	37% of clients from Germany and 38% from Spain have exited the bank.
Age	86% of exited clients are 30 or above years old, while 67% of loyal clients are under 40 years old
Gender	59% of exited clients were female, while 57% of loyal clients are males

STATISTICS

LOYAL CLIENTS						
	Credit Score	Age	Tenure	Balance	Number Of Products	Estimated Salary
MIN	411	18	0	$ -	1	$ 371.05
MAX	850	82	10	$ 197,041.80	3	$ 199,661.50
MEAN	**652**	**38**	**5**	**$ 74,830.87**	**2**	**$ 98,942.83**

EXITED CLIENTS						
	Credit Score	Age	Tenure	Balance	NumOfProducts	Estimated Salary
MIN	376	22	0	$ -	1	$ 417.41
MAX	850	69	10	$ 213,146.20	4	$ 199,725.39
MEAN	**637**	**45**	**5**	**$ 90,239.22**	**1**	**$ 97,155.20**

DECISION TREE: WILL A CUSTOMER LEAVE THE BANK?



RECOMMEDATIONS

•Pig E. Bank should consider connecting with inactive customers before these customers make a permanent decision to permanently close their accounts . Providing suitable offers and rewards especially to female customers with an effective approach.

•Priority efforts should be given to the German customers, which experiences the highest percentage of customer closing accounts at Pig E. Bank.

•Conducting a survey for all their customers to understand why they are closing their accounts. Create an especial survey to understand what are the financing needs of female customers who's tend to leave the bank more frequently than their male customers.

Final Report Back to Table of Contents

CHOCOLATE BAR RATING

CHOCOLATE IS ONE OF THE MOST POPULAR CANDIES IN THE WORLD.

EACH YEAR, THE UNITED STATES CONSUMERS EAT MORE THAN 2.8 BILLIONS POUNDS.

HOWEVER, NOT ALL CHOCOLATE BARS ARE CREATED EQUAL!

THIS DATASET CONTAINS EXPERT RATINGS OF OVER 2,700 INDIVIDUAL CHOCOLATE BARS, ALONG WITH INFORMATION ON THEIR BEAN ORIGIN, PERCENTAGE OF COCOA, THE VARIETY OF CHOCOLATE BEAN USED AND WHERE THE BEANS WERE GROWN.

PROJECT OVERVIEW

GOALS:

THE GOAL IS TO LEARN ABOUT CHOCOLATES. IDENTIFY THE COUNTRIES THAT PRODUCE THE BEST COCA BEANS, KNOWING MORE ABOUT THE RELATIONSHIP BETWEEN COCA SOLIDS PERCENTAGE AND RATINGS AND IF THE AMOUNT AND SELECTION OF THE INGREDIENTS MATTER TO THE QUALITY OF THE TASTE.

KEY QUESTIONS:

1. WHAT IS THE AVERAGE RATING BY COUNTRY OF ORIGIN?
2. WHERE ARE THE BEST COCOA BEANS GROWN?
3. WHICH COUNTRIES PRODUCE THE HIGHEST-RATED BARS?
4. WHAT ARE THE TOP COMPANIES THAT PRODUCE MOST CHOCOLATE BARS?
5. IS THERE A CORRELATION BETWEEN COCA PERCENTAGE AND CHOCOLATE RATINGS?
6. IS THE CACAO BEAN'S ORIGIN AN INDICATOR OF QUALITY?
7. WHICH ARE THE MOST COMMON INGREDIENTS USED?
8. WHAT PERCENTAGE OF COCOA HAS THE HIGHEST RATING?

TO DO THIS I WILL CLEAN AND CONDITION THE DATA TO FIT INTO A SUPERVISED LEARNING ALGORITHM; USING PYTHON AND EDA USING VISUALIZATIONS.

SKILLS:

- SOURCE OPEN DATA
- DATA CLEANING, WRANGLING AND SUBSETTING.
- CREATING GEOGRAPHICAL VISUALIZATIONS.
- SUPERVISED MACHINE LEARNING WITH LINEAR REGRESSION.
- UNSUPERVISED MACHINE LEARNING WITH K-MEANS CLUSTERING.
- SOURCING AND ANALYZING TIME-SERIES DATA.
- CREATING A DASHBOARD.

DATA:

- CHOCOLATE BAR 2020 DATASET FROM FLAVORS OF CACAO, A DATASET OUTLINING OVER 2,700 TYPES OF PLAIN DARK CHOCOLATE BARS, THEIR RATINGS, INGREDIENTS, AND TASTES. THE DATASET USED HERE HAVE BEEN ACQUIRED FROM RACHAEL TATMAN'S CHOCOLATE BAR RATINGS DATASET ON KAGGLE.

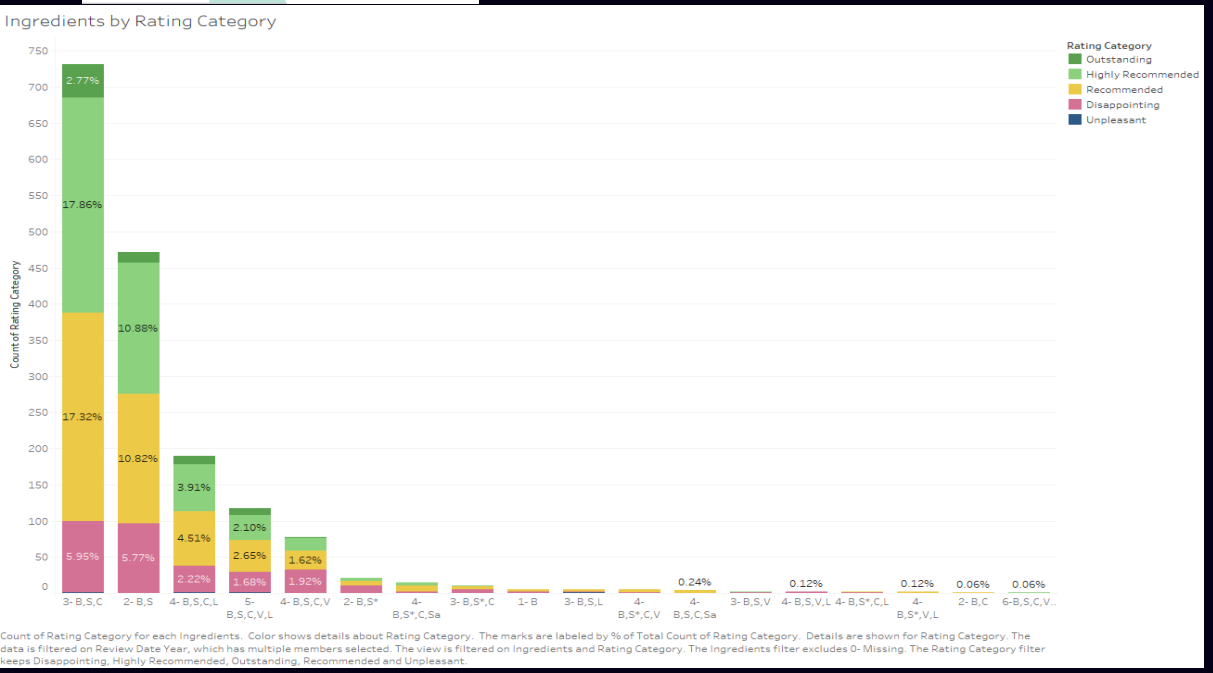
TOOLS:

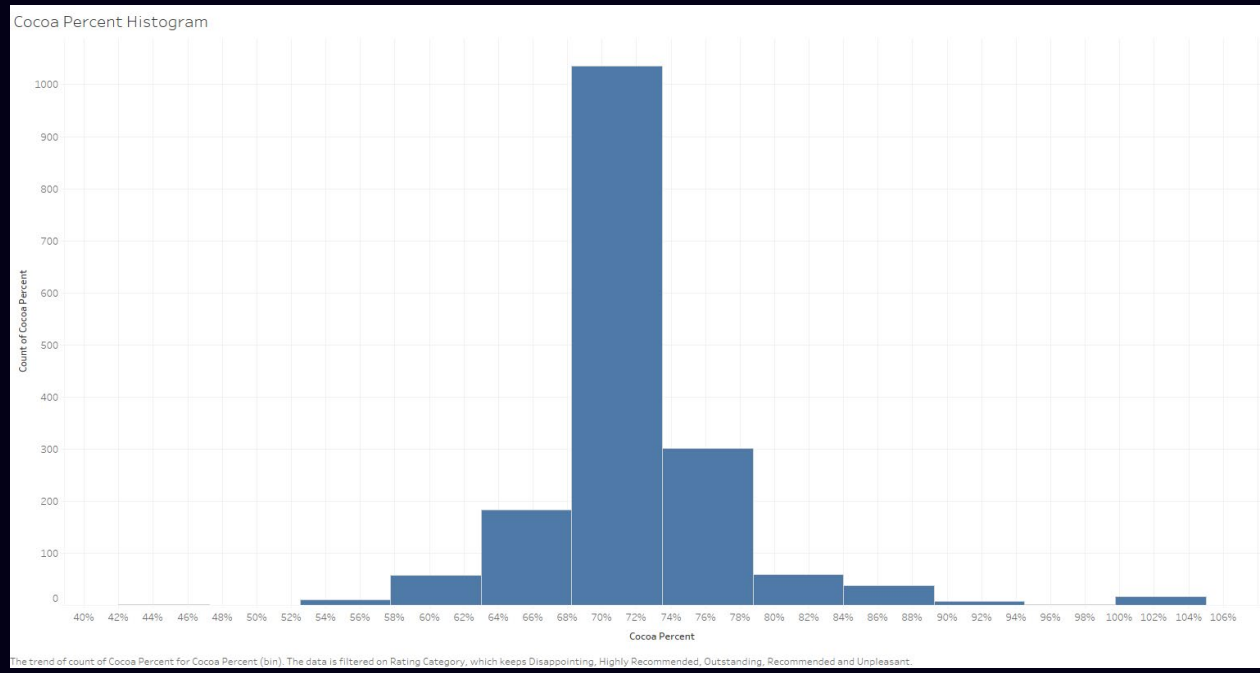
EXPLORATORY ANALYSIS: WHAT IS INSIDE OF A CHOCOLATE BAR?

6-B,S,C,V,L,Sa	3.50
3- B,S,C	3.31
2- B,S	3.23
4- B,S,C,L	3.22
5- B,S,C,V,L	3.14
4- B,S*,C,Sa	3.13
4- B,S,C,Sa	3.13
4- B,S*,V,L	3.13
2- B,C	3.00
3- B,S,V	3.00
4- B,S,C,V	3.00
4- B,S*,C,V	3.00
3- B,S*,C	2.98
2- B,S*	2.90
4- B,S*,C,L	2.88
1- B	2.85
3- B,S,L	2.70
4- B,S,V,L	2.63

Avg. Rating
2.63 3.50

The main ingredient in a chocolate bar is the cacao bean, which is categorized based on the percent of cocoa present in the bar. in the histogram, we can see that the most common percentage of cocoa is between 65 and 70%. The most common and highly rated combination of ingredients are bean, sugar and cocoa butter, vanilla, lecithin and salt; followed by cacao bean, sugar and cocoa butter. The ratio between rating categories is similar for the top 3 combinations with just some minor changes.



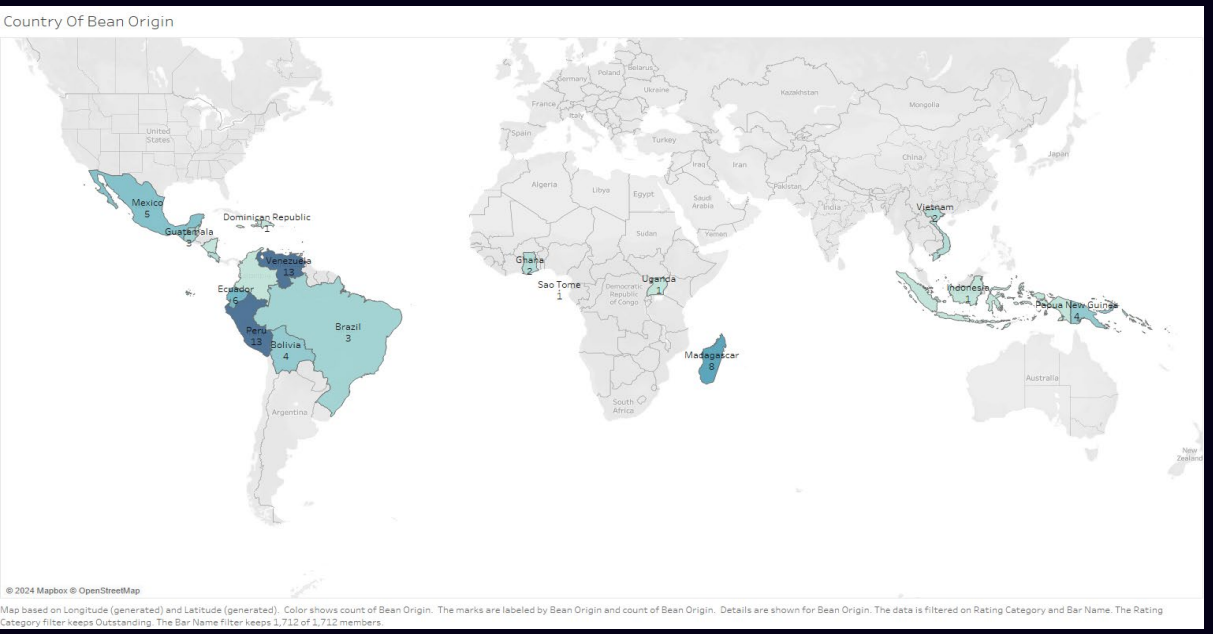


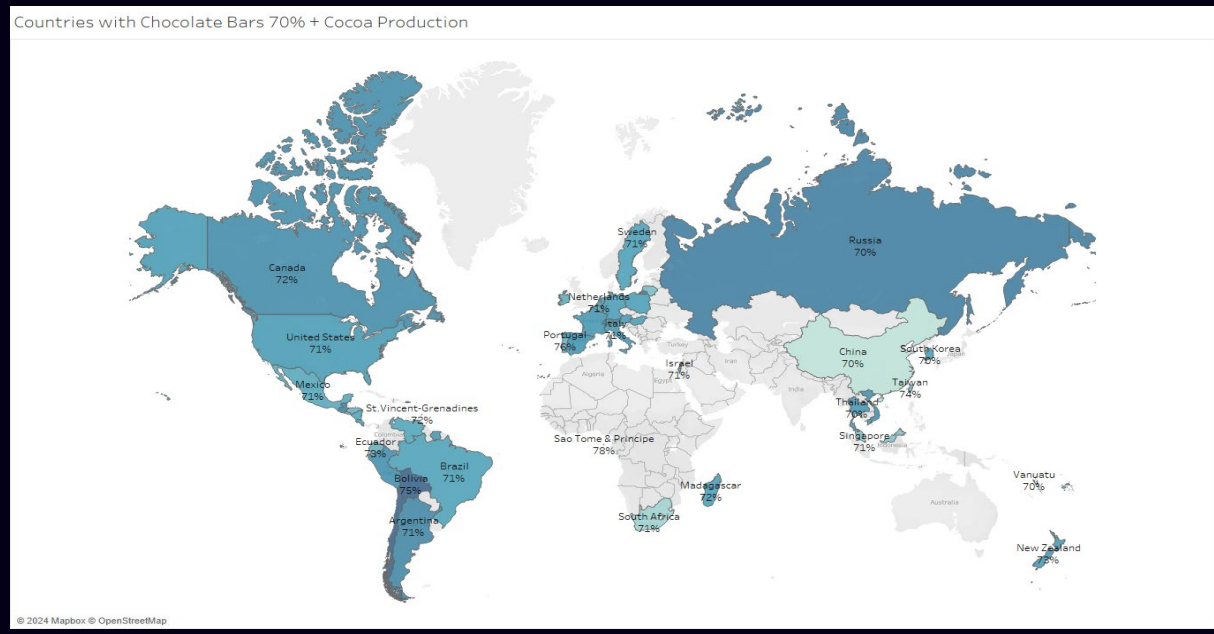
Final Report: GITHUB

Back to Table of Contents

EXPLORATORY ANALYSIS: WHERE THE COCOA IS FROM?

The cacao bean can be considered the most important part in determining the quality and flavor of a chocolate bar. By looking at the different locations that cultivate cacao beans, we can see that most of them come from South America, which include Venezuela, Peru and Ecuador. The two countries with the highest average bar rating are Tobago and Sao Tome & Principe with an avg rating of 3.5 with 1 and 2 reviews, respectively . Solomon Islands is the next highest rated location with a rating of 3.45 and 10 reviews. Venezuela and Peru are the countries with highest rated 'Outstanding bars with 13. We can see all the countries that produce chocolate bars with 70%+ cocoa percent.



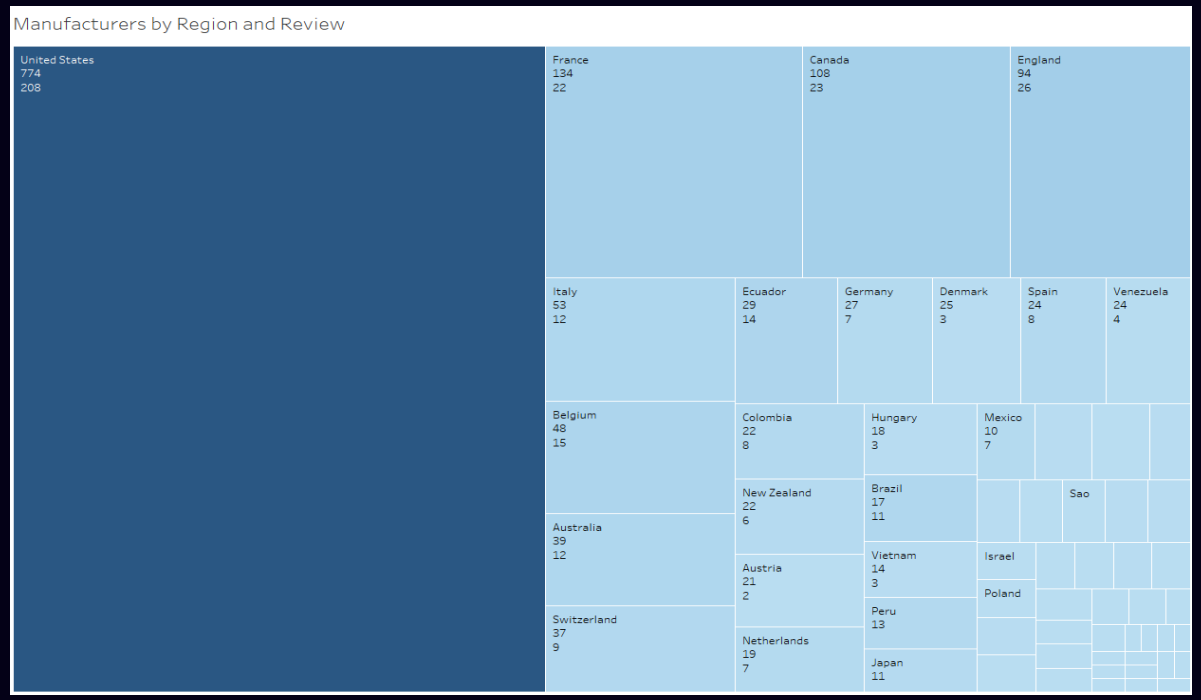


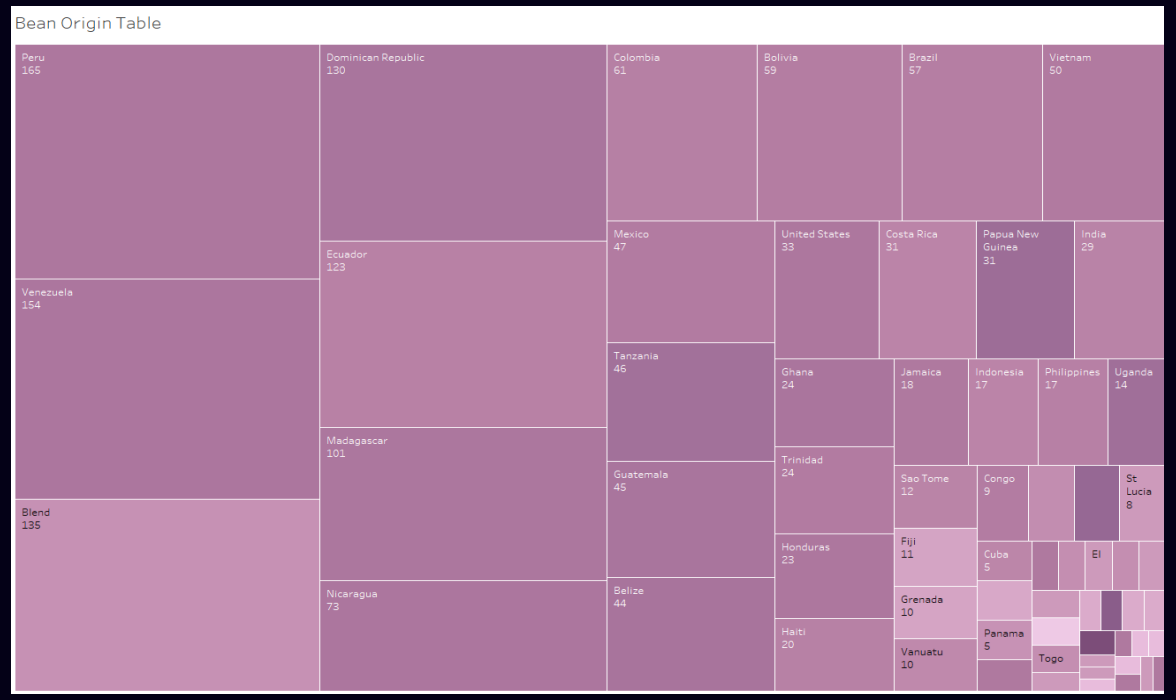
Final Report: GITHUB

Back to Table of Contents

EXPLORATORY ANALYSIS: WHO MAKES THE BEST BARS

The ratings indicate that United States is the manufacturer country favorite between the ratings category as highly recommended, outstanding and recommended.
For the Bean origin country under the rating category, the highly recommended and outstanding goes to Peru as for recommended goes to Venezuela.

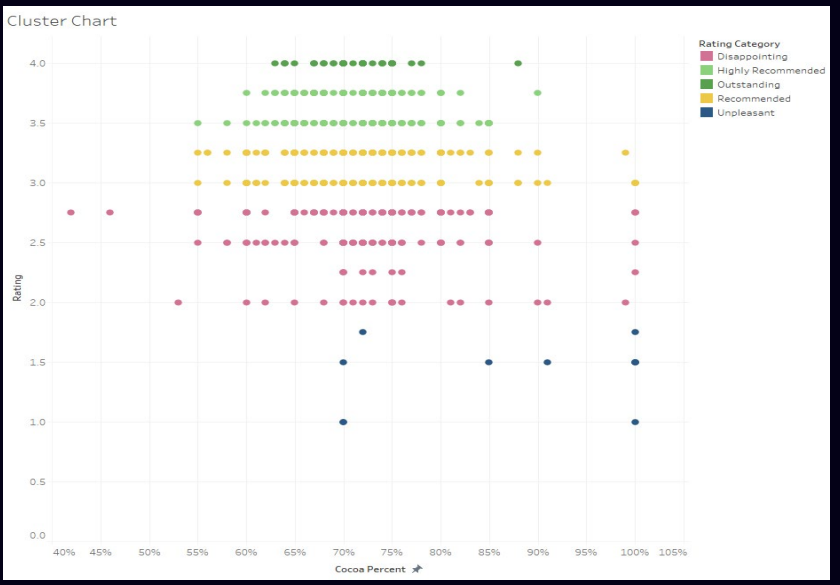




Final Report: GITHUB

Back to Table of Contents

LINEAR REGRESSION: IS MORE COCOA PERCENT BETTER?



Hypothesis:

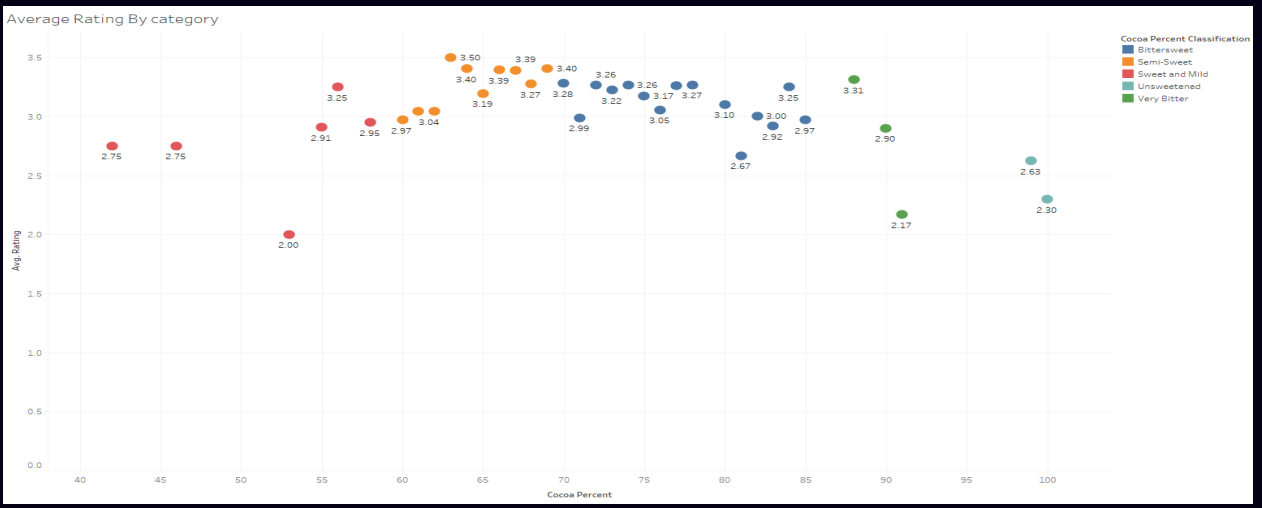
-If the cocoa percentage is high, then the better the rates they received?

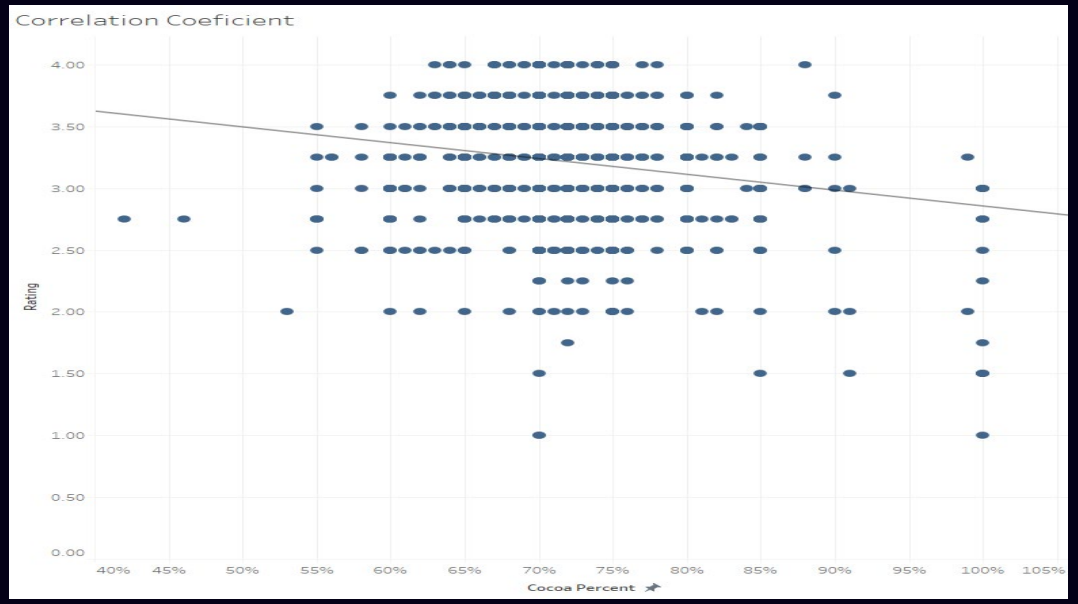
Results:

The linear regression represent a weak negative correlation between chocolate bar rating and cocoa percent. This means that there is no significant relationship between the cocoa percent and rating, we can reject the null hypothesis.

Cluster Analysis:

Chocolate bars are usually categorized by the amount of cocoa percent in a bar. Looking at the average rating of bars by cacao percent, we can see that semi-sweet and bittersweet chocolate bars are the most represented and have the least varied avg ratings. The highest avg rating of cocoa percent is 63% with 3.50, followed by 69% with 3.40. The highest rated 'Outstanding' chocolate bars contain between 63% and 88% cocoa percent.

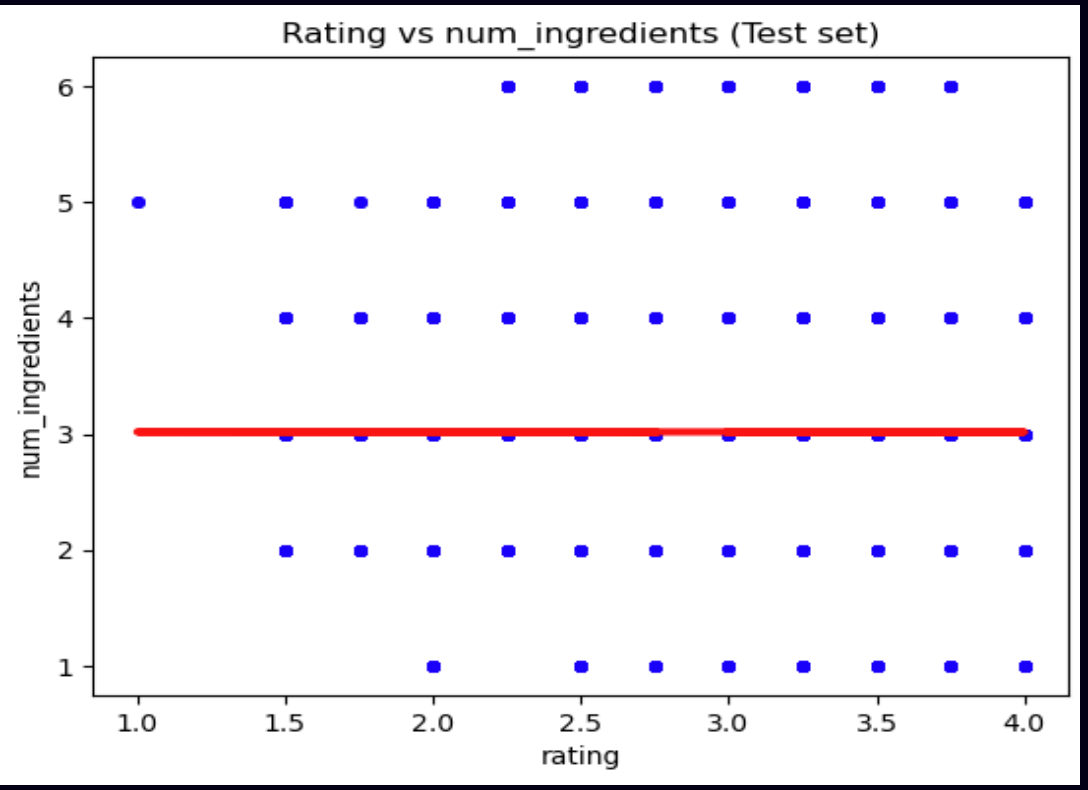




Final Report: GITHUB

Back to Table of Contents

LINEAR REGRESSION: IS MORE INGREDIENTS BETTER?



Hypothesis:

-If the number of ingredients affect the ratings?

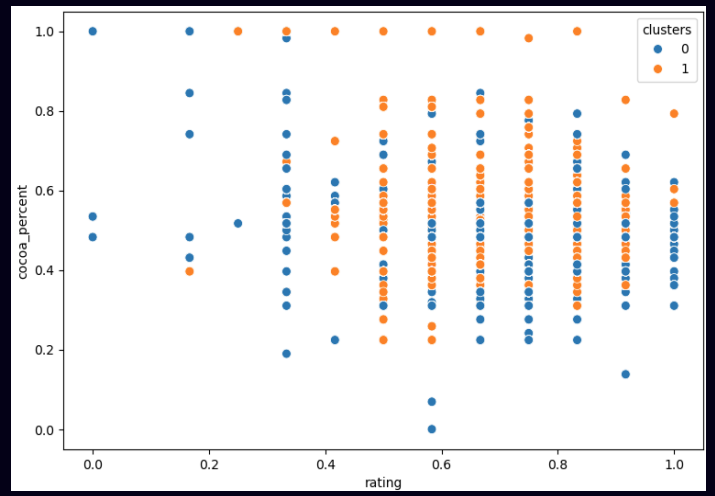
Results:

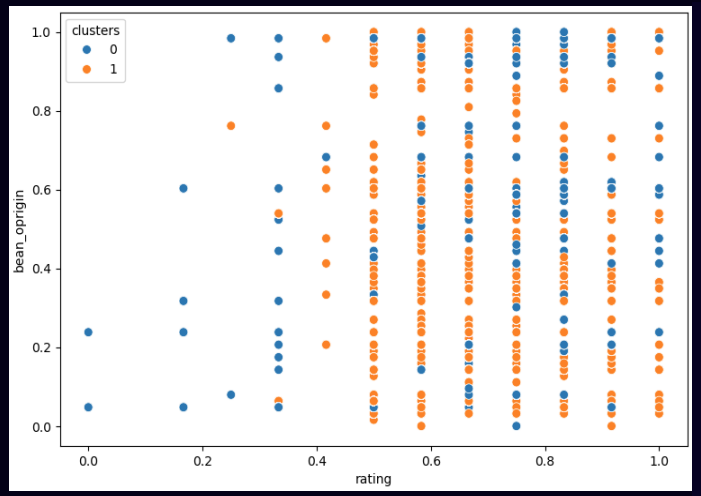
The linear regression relationship isn't purely linear This means that there is no significant relationship between the number of ingredients and rating, we can reject the null hypothesis.

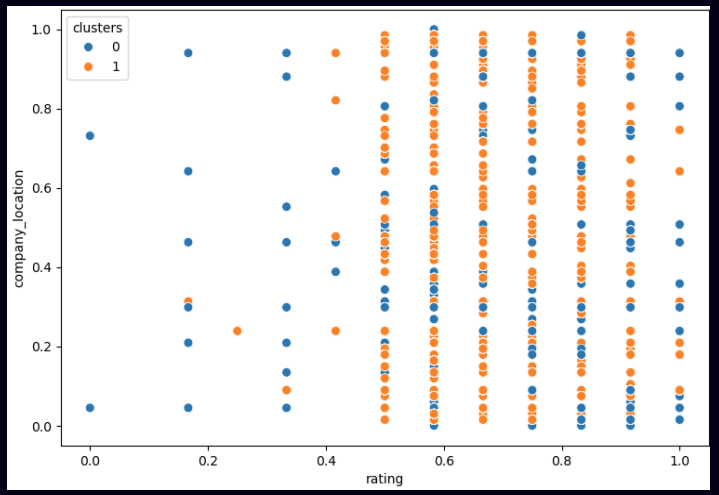
Final Report: GITHUB

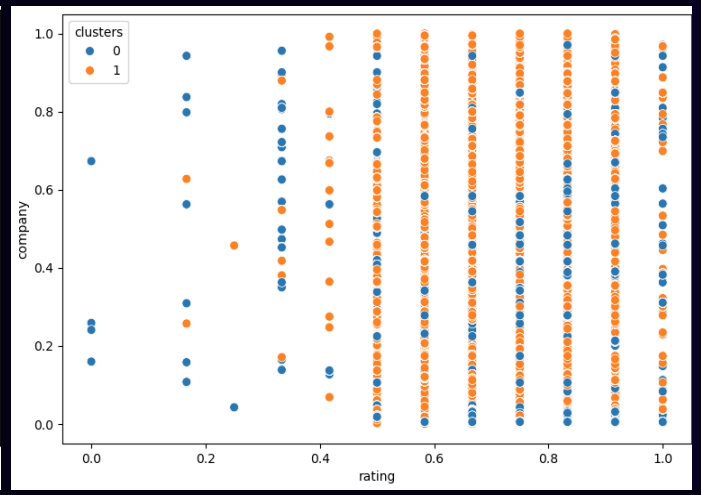
Back to Table of Contents

UNSUPERVISED MACHINE LEARNING: GAP STATISTIC METHOD (K-MEANS CLUSTERING)









Cocoa Percent and Ratings:
Clusters tend to concentrate to the middle right side, this aligns with the conclusion of my hypothesis, rating does not affect cocoa_percent

Bean Origin and Ratings:
Clusters tend to concentrate to the middle right side, I see more concentration of bean_origin than actually ratings,
This aligns with the conclusion the rating does not affect bean_origin.

Company Location and Ratings:
Clusters tend to concentrate to the middle right side. I can see a slight relation with the company location and ratings.

Company and Ratings:
This last cluster is very interesting because there is more concentration between ratings and company to have a little more relation comparing with the other clusters.
Which make me conclude that ratings could create some type of Bias in this case.

Final Report: GITHUB

Back to Table of Contents

RECOMMEDATIONS

Conclusion

- The most common chocolate bars are with cocoa percent between 65 and 70%.
- The most common and important ingredients in the highest performing chocolate bars are the cacao bean, cacao butter, vanilla, lecithin and salt.
- The most common and highest rated cacao beans origin come from Peru and Venezuela.
- The percentage of cocoa in a chocolate bar is not directly related to its rating. The best rated chocolate bars have between 63% and 88% cocoa percent.
- Semi-sweet (63%-88%) chocolate bars contain two of the highest average rated percentages (53% and 69%) with few variability between the average ratings.
- The U.S.A has the most chocolate manufacturer companies, as well as the most 'Outstanding' ratings.
- Soma (Canada) has the most number of reviews, as well as the most 'Outstanding' ratings score. Arete (U.S.A.) has the 3rd most number of reviews and the 3rd most 'Outstanding' ratings.

Recommendations

It is recommended to open the ratings survey to the regular consumer instead of a limited population group in this case expert opinions., with that on my mind we can determine if the cocoa percent has anything to do with the flavor, quality coming from regular consumer perspective, so, we can project revenues and variety.

Limitations

The analysis of this project is limited by the nature of the dataset, as the data is gathered by assigning a rating based by expert opinions which can lead to Biases. Because of this there is an absence of quantitive and continuous variables that could be used to perform a more thorough analysis. Another limitation is the size of the dataset, which limits the amount of insight that can be gathered from measuring averages and performing time analysis decomposition.

Data Source:

Chocolate bar ratings 2022 (kaggle.com)

Data originates from the Flavours of Cacao website

Final Report: GITHUB

Back to Table of Contents

THANK YOU!

Back to top

Contact