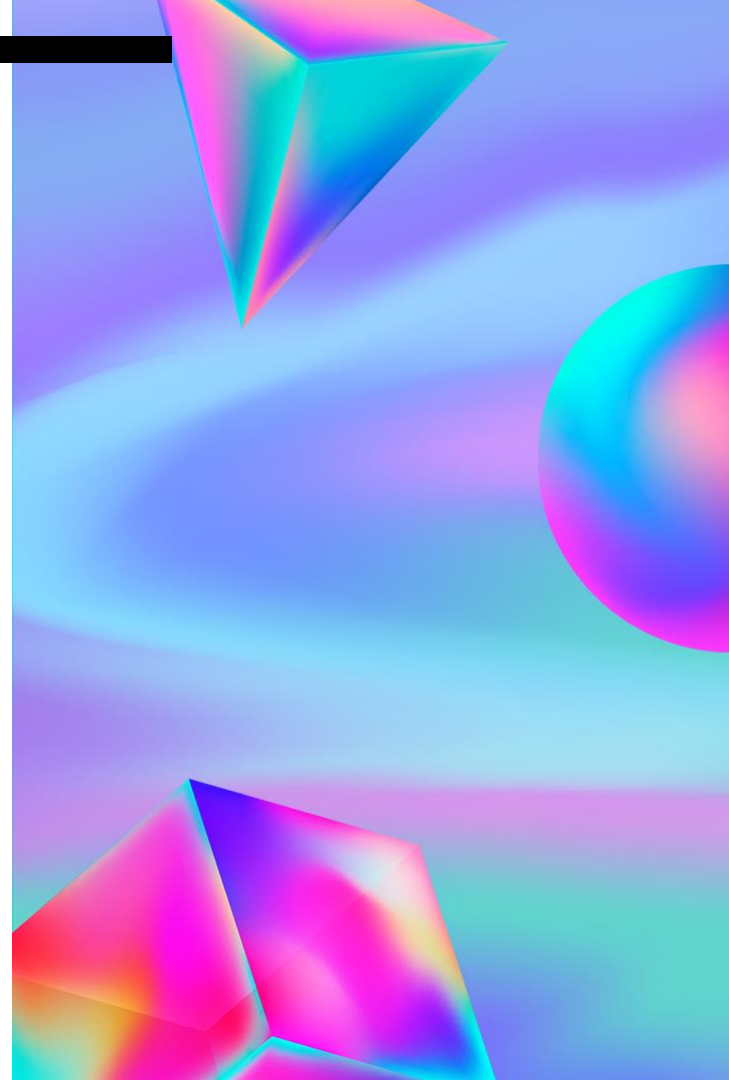


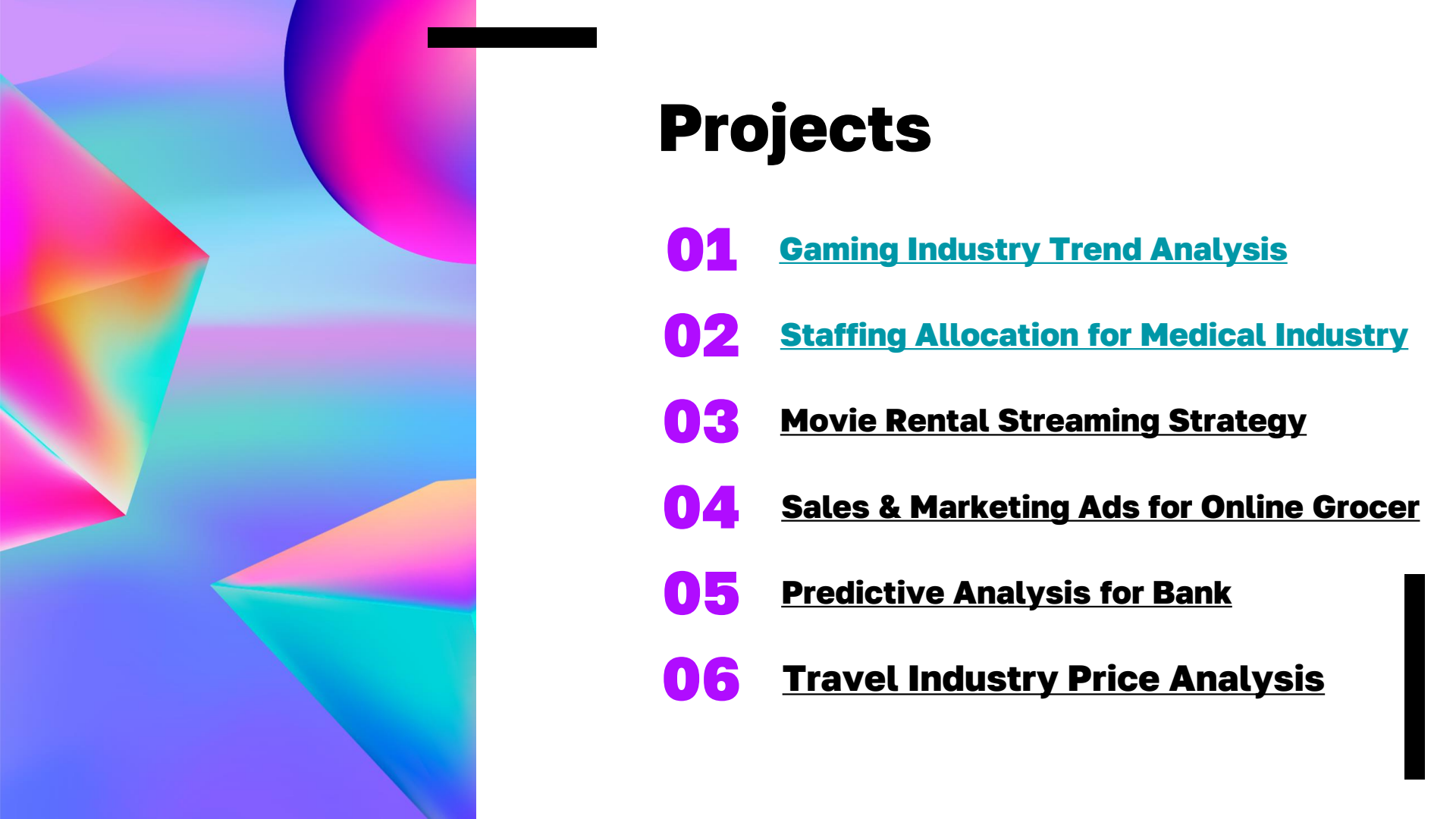
Justina Bui

Data Analysis Portfolio





Projects

- 01** [Gaming Industry Trend Analysis](#)
 - 02** [Staffing Allocation for Medical Industry](#)
 - 03** [Movie Rental Streaming Strategy](#)
 - 04** [Sales & Marketing Ads for Online Grocer](#)
 - 05** [Predictive Analysis for Bank](#)
 - 06** [Travel Industry Price Analysis](#)
- 

01 Gaming Industry Trend Analysis

OVERVIEW: Project Brief by CareerFoundry

Game Co is seeking to leverage video game data to drive its product development and marketing strategy, with a primary focus to boost future sales. The goal is to analyze global historical trends in the gaming console industry and provided marketing and sales recommendations.

DATASET AND LIMITATIONS:

Source: Video Games Sales Dataset by VGChartz

The data set for this project is primarily video games sales across different platforms, genres, and publishing studios. It includes only games that sold at least 10K units (not financial \$) from 1980-2016.

KEY QUESTIONS

1. Have regional sales have remained relatively the same over time?
2. Where are the areas for worth for future investment?
3. Which genres are performing the best?

TOOLS USED AND TECHNIQUES:

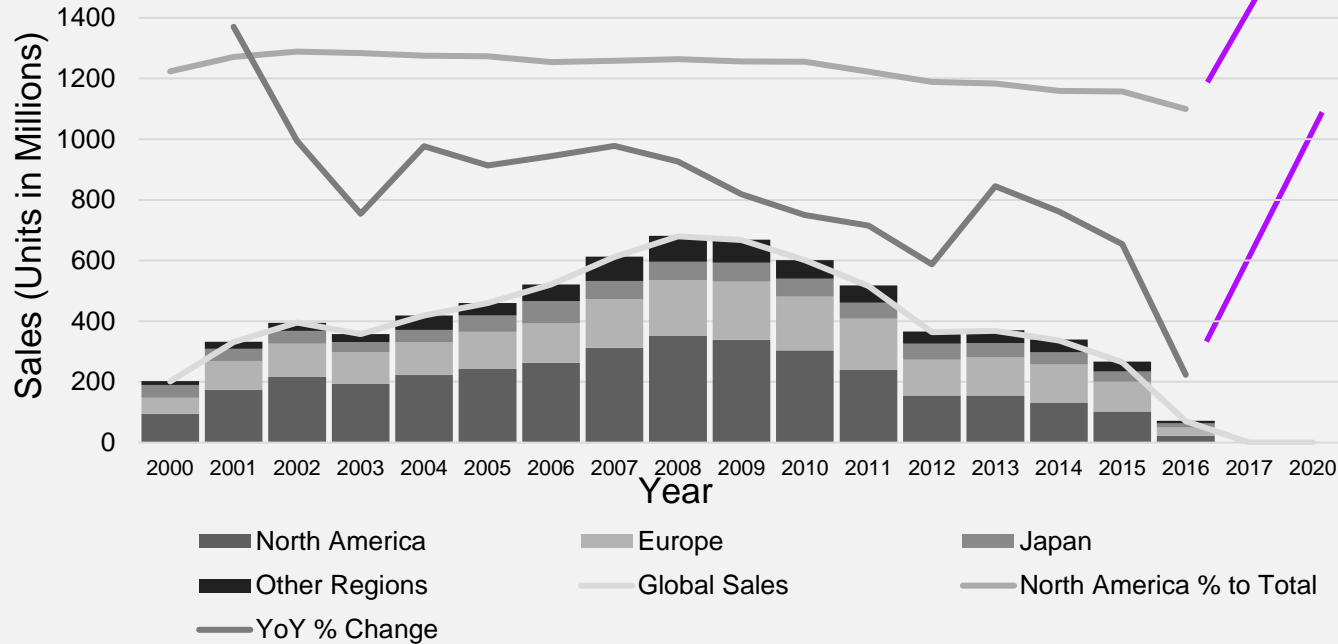
Microsoft Excel, PowerPoint

- Data Integrity, Quality, Consistency Checks
- Descriptive Analysis
- Pivot Tables
- Summarizing and Grouping Data
- Calculated Fields
- Combo Stacked Bar, Line charts



01 Gaming Industry Trend Analysis

Game Sales By Region from 2000-2015

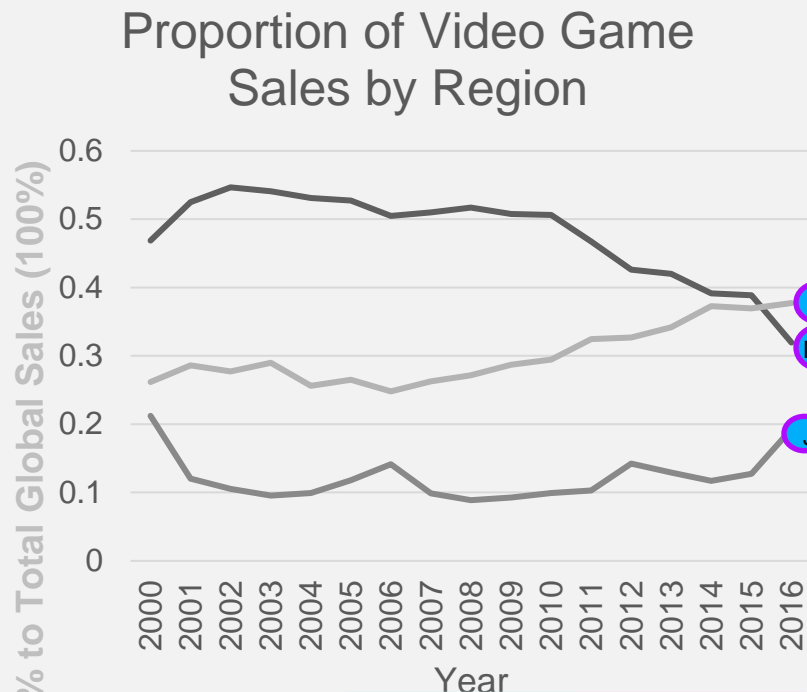


YoY% change is showing dynamic shifts in YoY% change is showing dynamic shifts in the market and showing slight downtrend

North America contributed 45% to total global sales in terms of units sold



01 Gaming Industry Trend Analysis



2.17 bil

3.08 bil

1.92 bil

Units sold over time

Europe sales' contribution to global total has continue to rise since 2010 from 29% contribution to 38% by 2016 (~9% increase).

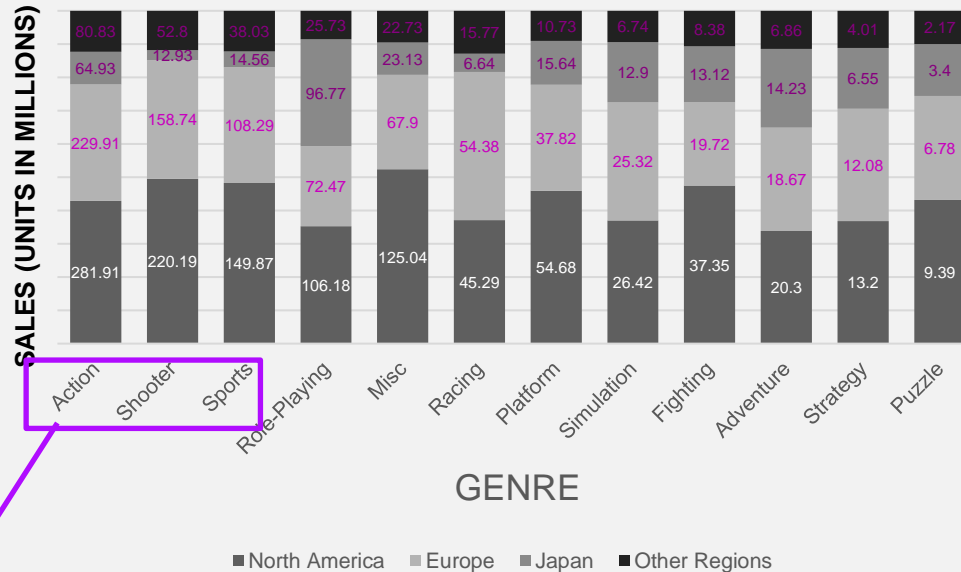
Japan sales' contribution to global total had a similar trend from 2010 at 10% contribution and then to 19% by 2016 (~9% increase).

North America's sales contribution decreased from 51% to 32% (-19% variance). As the gaming industry's total addressable market is shrinking YoY, so is North America's contribution



01 Gaming Industry Trend Analysis

Sales per Region by Genre Breakout For the Last 5 Years (2010-2015)



Action, Shooter, Sports games have been more popular than other genres



01 Gaming Industry Trend Analysis



BUDGET RECOMMENDATIONS FOR 2017

INTEREST

“Action” has been the dominate genre in all North America, Europe, Japan regions. Collaborating with Twitch streamers to co-create a whole new game increases their motivation to share on their platforms vs. streaming most popular Action game: Grand Theft Auto (GTA)

Build “Sport-Role-Play” (SRP) category

Co-create “Action” game with Top 3 Twitch streamers

Hire more European sales representatives

PURCHASE

Make the purchase price slightly below the top performing game name for competitive selling point

\$XX.XX



BUILD AWARENESS

To compete with getting market share from other publishers, GameCo needs to create a new genre that combines “Sport” theme + compelling “Role-Playing” elements to influence Japanese customers who primarily participate in “Role-Playing” genre, to engage in exercise, and to increase brand awareness for family members to join in and play together

DECISION

Increase sales reps in Europe to facilitate region’s growing interest and persuade new customers to GameCo’s “Shooter” games as we leverage increasing population and differentiate in customer service (~462m unit sales in the last decade)



02 Staffing Allocation for Medical Industry

OVERVIEW: Project Brief by CareerFoundry

Analyzed seasonality of when does influenza occur and ramps throughout the year and align existing nursing staff headcount to allocate throughout 50 states of US accordingly.

DATASET AND LIMITATIONS:

Source: Fluview by CDC and US Census Bureau

- Fluview Dataset only till 2019, there is an estimated 80,738 Medical Personnel Available to distribute.
 - Majority of nurses recorded were currently located in the East Coast.
 - Government data of each state's population size of a given year provided with variables such as state, year, age group.
- Time range of data: 2009-2017.

KEY QUESTIONS

1. When is influenza season?
2. Who are most vulnerable?
3. Which states have most vulnerable?



TOOLS USED AND TECHNIQUES:

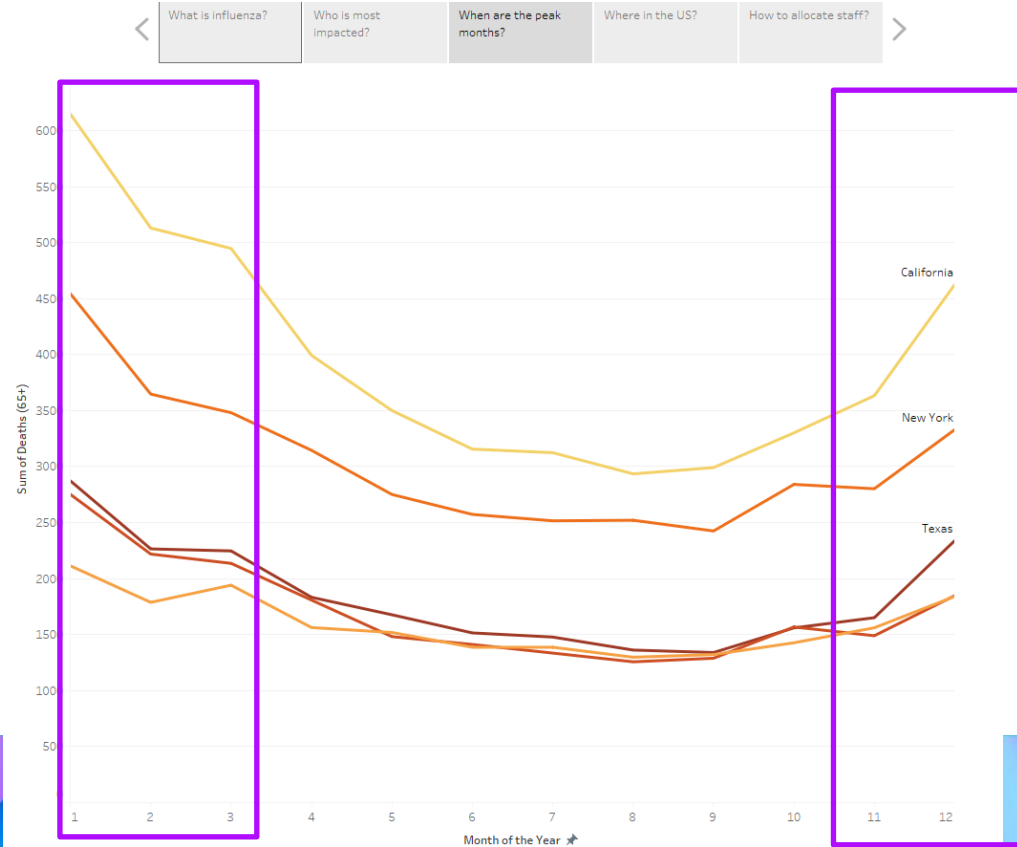
Microsoft Excel, PowerPoint

- Translate business requirements with real world constraints
- Data Integrity, Quality, Consistency Checks
- Statistical Hypothesis Testing
- Descriptive Analysis & Correlation
- Forecasting
- Combo Stacked Bar, Line charts



02 Staffing Allocation for Medical Industry

Tableau Storyboard - How to Allocate Staff



Influenza starts around November and ends in March

California, New York, and Texas have the most recorded deaths



02 Staffing Allocation for Medical Industry

Variables	Deaths 65 years+ & Total Population 65 years+
Proposed Relationship	To test the relationship of # of deaths from senior citizen age (65+ to its state's overall population of that same age group)
Correlation Coefficient	0.94
Strength of Correlation	Strong
Usefulness/ Interpretation	Because the sample data's correlation coefficient = 0.94, it is very close to 1 and therefore showcases if someone is 65 years old or above, the mortality rate (chance of death) is strongly correlated.

Independent variable: Age group Dependent variable: Death rate	<p>Null Hypothesis: The influenza death rate of elderly population (65 years old and up) is less than or equal (\leq) to the influenza death rate of those age 64 and below.</p> <p>Alternative Hypothesis: The influenza death rate of more elderly population (65 years old and up) is greater than ($>$) the influenza death rate of those age 64 and below.</p>	Significance level: $\alpha = 0.05$
---	--	--

T Test: Two-Sample Assuming Unequal Variances

	0-64 years old	65+ years old
Mean	0.000809457%	0.086474735%
Variance	1.00510680496592E-10	1.78562315932869E-07
Observations	450	450
Hypothesized Mean Difference	0	
df	449	
t Stat	-42.9926273815464	
P(T<=t) one-tail	1.42776992102307E-161	
T Critical one-tail	1.64825431724876	
P(T<=t) two-tail	2.85553984204614E-161	
t Critical two-tail	1.96526138811176	

Conducted T-Test to confirm that there is a strong relationship between "Age Group" & "Death Rate" variables

This is one-tailed test because I'm looking for a difference in one direction (influenza death rate of elderly population).

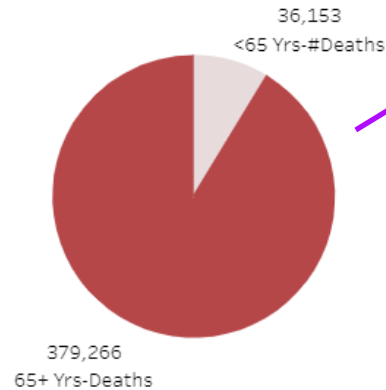
Because the p-value < 0.05 (significance level), this shows that something statistically significant is happening among the influenza death rate of these 2 cohorts divided by age group. There is almost 0% chance the difference in number of death was random. There is almost 100% chance that age does affect the number of deaths. Therefore, given $\alpha = 0.05$, we have greater than 95% confidence level to reject the null hypothesis and accept the alternative: "The influenza death rate of more elderly population (65 years old and up) is greater than ($>$) the influenza death rate of those age 64 and below."



02 Staffing Allocation for Medical Industry

Tableau Storyboard - How to Allocate Staff

Pie Chart of 91% of Influenza Deaths are Aged: 65+ Years Old



Total Sample Size
415,419

9% Aged <65 Influenza Deaths
91% Aged 65+ Influenza Deaths

36,153 379,266

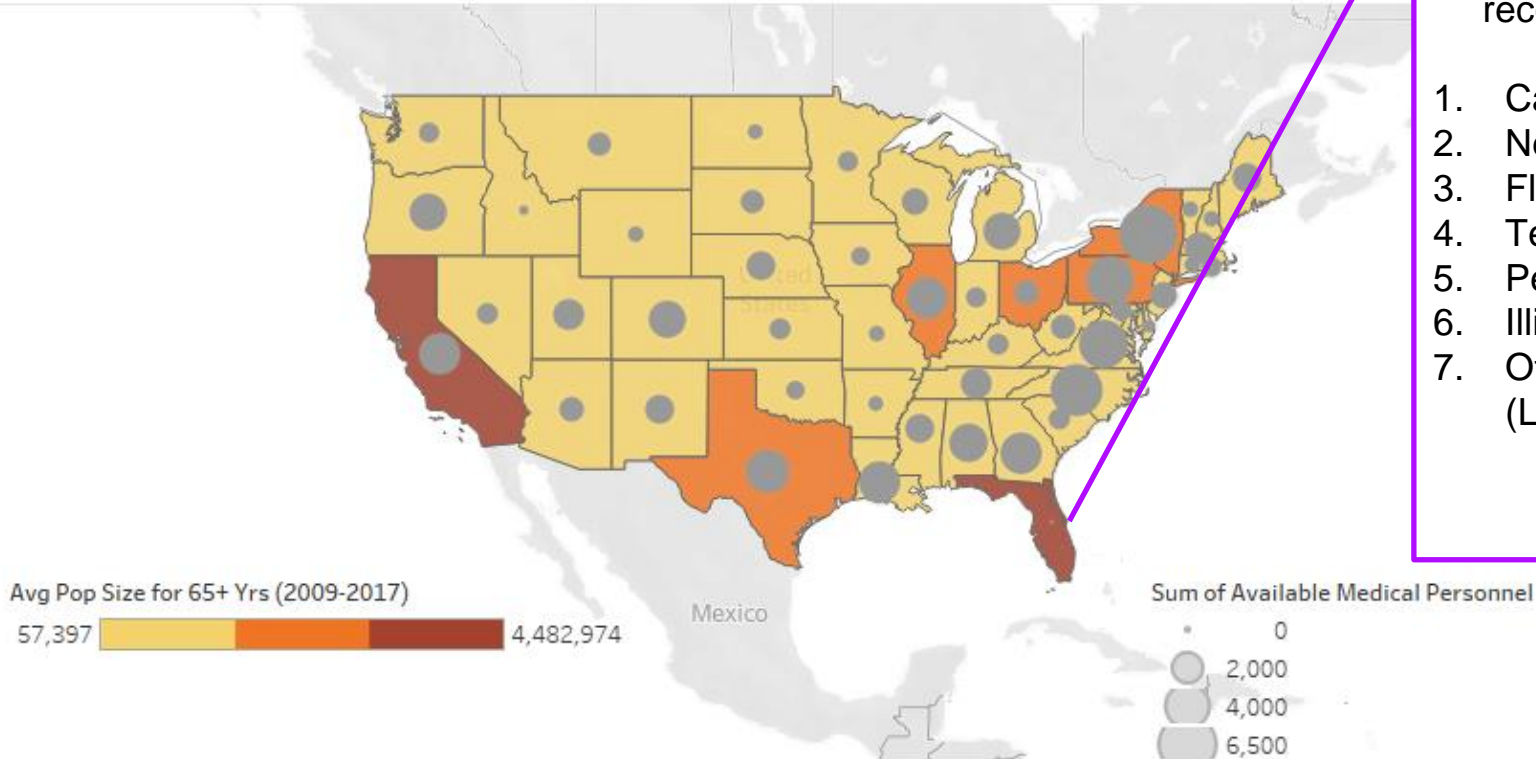
Elderly population (65 yrs+) have higher recorded death than younger population.



02 Staffing Allocation for Medical Industry

Tableau Storyboard - How to Allocate Staff

As of 2019, there is an estimated 80,738 Medical Personnel Available to Distribute
*Most of them are currently located in the East Coast



Where is Influenza Most Prevalent (highest recorded deaths)?:

1. California (High)
2. New York (Med)
3. Florida (Med)
4. Texas (Med)
5. Pennsylvania (Med)
6. Illinois (Med)
7. Other yellow states (Low)



02 Staffing Allocation for Medical Industry

Tableau Storyboard - How to Allocate Staff

How to Allocate Existing Medical

Rank Dense of Current Distribution By State

Provider Count ..	State	
1	New York	6,429
2	North Carolina	5,163
3	Virginia	4,525
4	Pennsylvania	4,299
5	Texas	3,541
6	Georgia	3,274
7	Louisiana	3,243
8	California	3,230
9	Illinois	3,106
10	Alabama	2,807
11	Oregon	2,612
12	Michigan	2,549
13	Colorado	2,521
14	Utah	1,812
15	Massachusetts	1,760
16	Tennessee	1,754
17	Maine	1,689
18	Mississippi	1,611

Recommendations (Starting in November):

HIGH states: Send 70% of available staff from AK or OR--> CA to alleviate peak season of January

MED states: Monitor flu patient activity of top elderly states (FL, TX, NY, PA, IL) and if spikes are happening, migrate 10% of east coast medical personnel (ex: VT, DE, ME, RI) as they are single digit % capacity from 1:1 ratio of staff to patient.

LOW states: Migrate 4% staff from LA, 12% staff from NE medical personnel --> adjacent states AK, MO respectively.



03 Movie Rental Streaming Strategy

OVERVIEW: [Project Brief by CareerFoundry](#)

Rockbuster, a traditional video rental company with a global presence is facing intense competition from major streaming services like Netflix, Hulu, and Disney. In response, Rockbuster aims to launch an online video rental service to remain competitive. Analyzed film title and sales data for global movie rental company as they would like to transition from brick-and-mortar to online services platform.

DATASET AND LIMITATIONS:

Dataset provided by CareerFoundry

- Data covers stores location, film titles languages, staff, inventory, categories, actors, customers, payment history, addresses

KEY QUESTIONS

1. Do sales figures vary between geographic regions?
2. Which countries are Rockbuster customers based in?
3. Which movies contributed the most/least to revenue gain?
4. What was the average rental duration for all videos?

TOOLS USED AND TECHNIQUES:

Microsoft Excel, PowerBI, PowerPoint, DbVisualizer, ProgreSQL

- Relational Databases (ERD)
- Data Dictionary

- SQL

Database Querying

Summarizing and Joining Data

Subqueries & CTEs

- PowerBI

Tree Mapping

Pivot Tables



03 Movie Rental Streaming Strategy

PowerPoint Presentation - Github

WHERE ARE THE ROCKBUSTER CUSTOMERS?

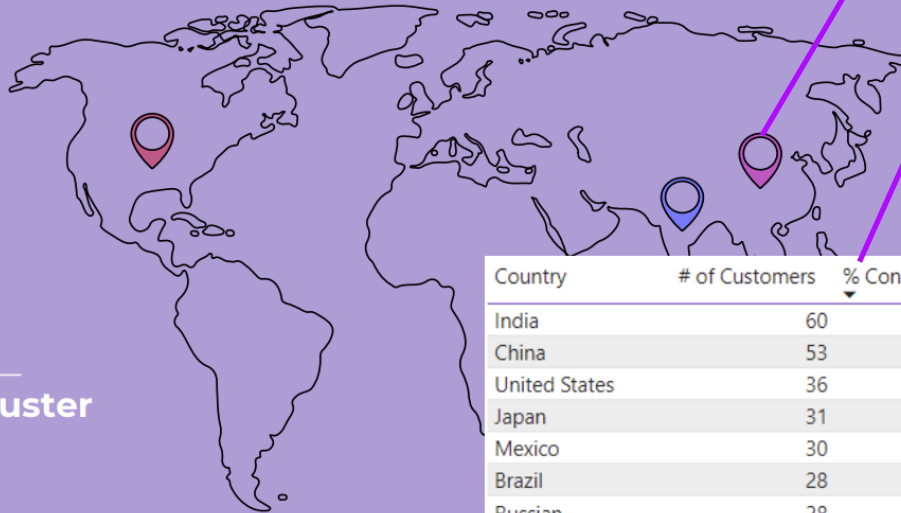
9% INDIA \$6m

9% CHINA \$5m

6% US \$3.7m

~24% of current Rockbuster customer base

~28% are in Japan, Mexico, Brazil, Russia, Philippines, Turkey, Indonesia (of Top 10)



Country	# of Customers	% Contribution
India	60	9.84%
China	53	8.56%
United States	36	6.01%
Japan	31	5.09%
Mexico	30	4.87%
Brazil	28	4.76%
Russian Federation	28	4.51%
Philippines	20	3.62%
Turkey	15	2.44%
Indonesia	14	2.21%

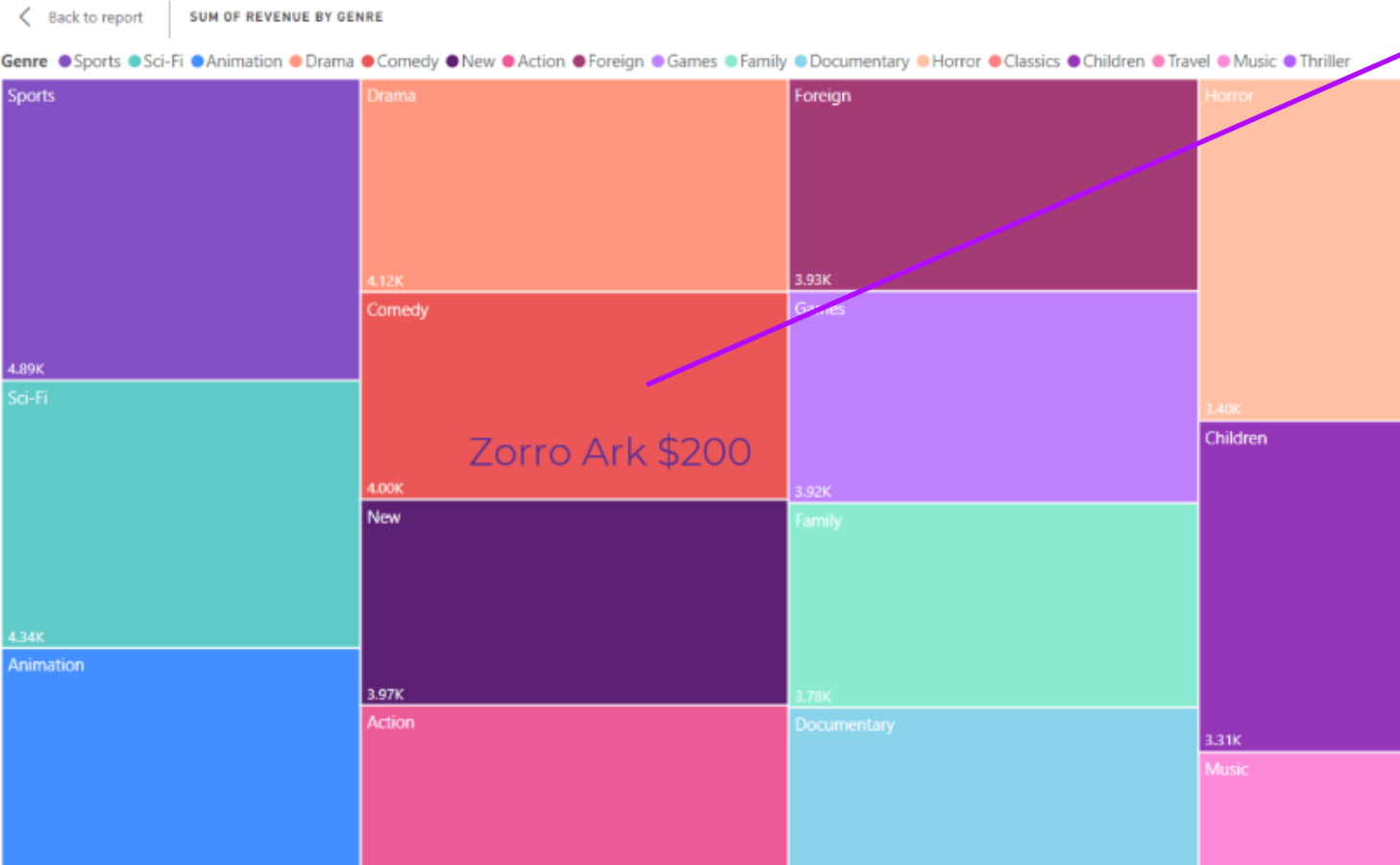
1. India makes up 19% contribution of sales
2. China makes up 17% contribution of sales
3. US is third rank at 12%

Asia (India, China, Japan, Russia, Philippines) roughly contributes to 63% of total sales and 65% of customer base.



03 Movie Rental Streaming Strategy

PowerBI Presentation - Github



1. Sports, Sci-Fi, Animation are top genres
2. “Telegraph”, “Voyage” and “Zorro Ark” films were top performers
3. Opportunity to double down on “comedy” genre



03 Movie Rental Streaming Strategy

PowerPoint Presentation - Github

RECOMMENDATION 1: PILOT PROGRAM

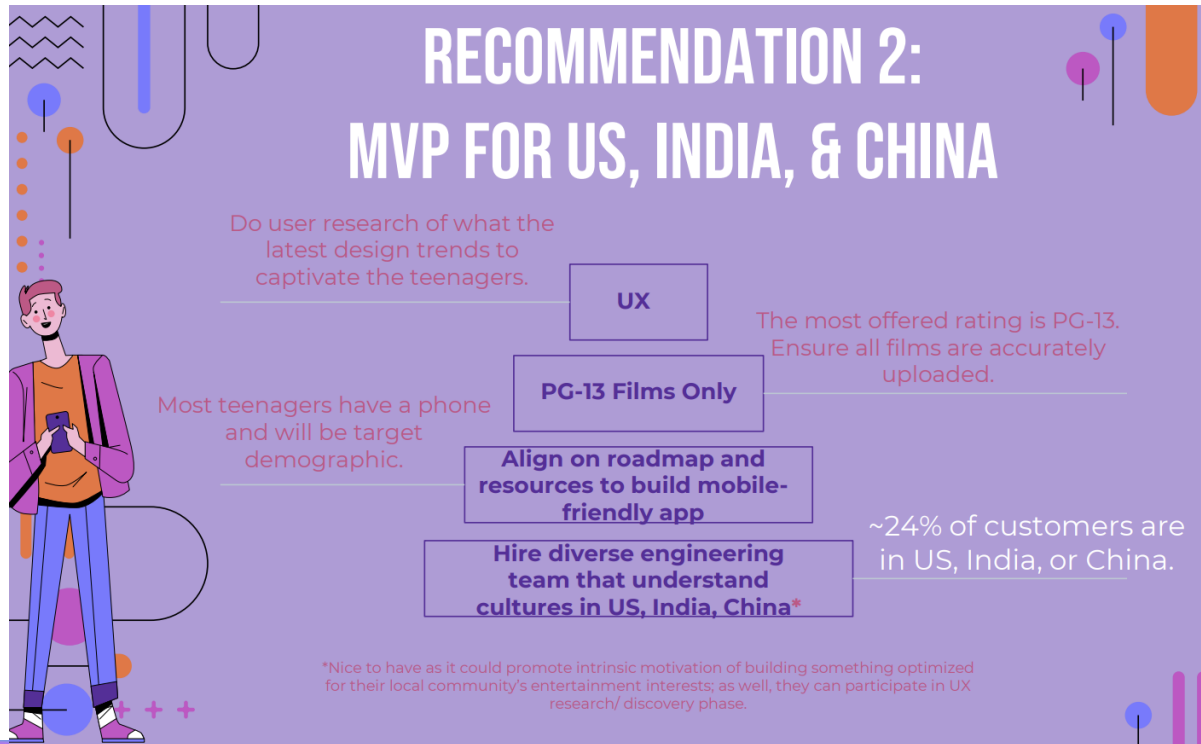
	BASIC	PRO	PREMIUM
Price per month*	\$ 9.99	\$ 19.99	\$ 29.99
Number of simultaneous screens	1	3	6
Unlimited movies	1000	1000	1000
Downloads on tablets and phones	✓	✓	✓

*\$9.99, \$19.99, \$29.99 are replacement costs that current customers are already acquainted with



03 Movie Rental Streaming Strategy

PowerPoint Presentation - Github



03 Movie Rental Streaming Strategy

PowerPoint Presentation - Github

RECOMMENDATION 3: REWARD LOYALTY

CHALLENGE

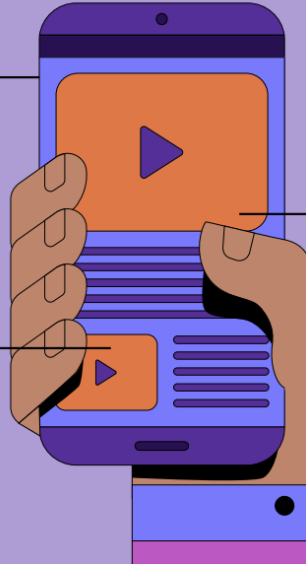
Offer exclusive banner to top spenders when they “refer a friend” and they can earn \$

RESULTS

This will get a flywheel effect to double down on customer growth % in US, China, India.

SOLUTION

Make the button showing the \$ as easy as possible to click



03 Movie Rental Streaming Strategy

PowerPoint Presentation - Github

RECOMMENDATION 4: CONTINUOUS MONITORING



SURVEYS

Release surveys to Top 30 customers what other ratings/ genres do they want Rockbuster to add



REPORTING

Share reports and findings to key decision makers in the business



INVENTORY

Reach out to production companies to get proper licenses of new releases to capitalize on what's on-trend.



04 Sales & Marketing Ads for Online Grocer

OVERVIEW: Project Brief by CareerFoundry

Analyzed product, order, and customer sales data for online food delivery app company (Instacart) as they would like to send targeted ads to continually increase revenue goals.

DATASET AND LIMITATIONS:

Grocery Shopping by Instacart

Data was last updated in 2017

Customer data and the "prices" column in the products data set were fabricated for the purpose of this course

KEY QUESTIONS:

1. What are the busiest days of the week and hours of the day?
2. Are there particular times of the day when people spend the most money?
3. Are there certain types of products that are more popular than others?

TOOLS USED AND TECHNIQUES:

Microsoft Excel, Jupyter Notebook, Python (libraries: Pandas, Numpy, Seaborn, Matplotlib)

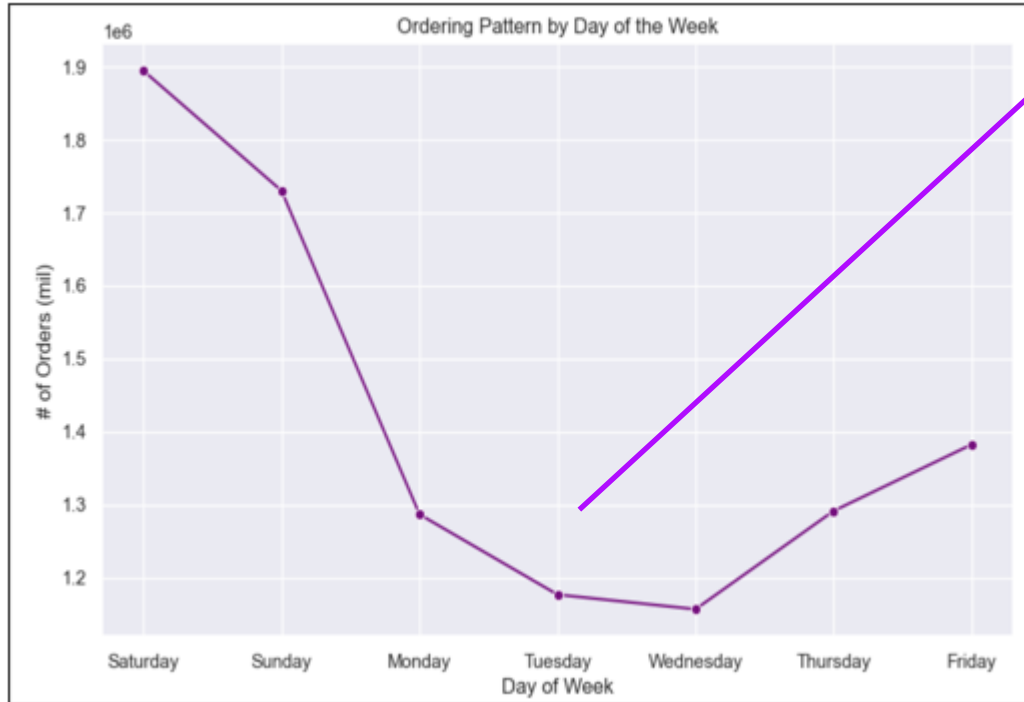
- Data wrangling, subsetting, and consistency checks
- Combining data
- Deriving new variables
- Grouping data
- Aggregating variables
- Data visualization – stacked bar chart, histogram, line
- Excel reporting



04 Sales & Marketing Ads for Online Grocer

Excel Reporting - Github

Fig 1.1 - Busiest Days of the Week



Most busy: Saturday, Sunday (weekend)
Average busy: Friday, Thursday, Monday
Least busy: Tuesday, Wednesday (middle of the week)

Most orders occur on the weekend (Saturday and Sunday). *Schedule ads in the middle of the week to boost more sales (Tuesday and Wednesday)*

Most orders happen between 10am- 3pm (order count above 80,000). Most stores are closed and people are sleeping between midnight- 6am (order count below 10,000).

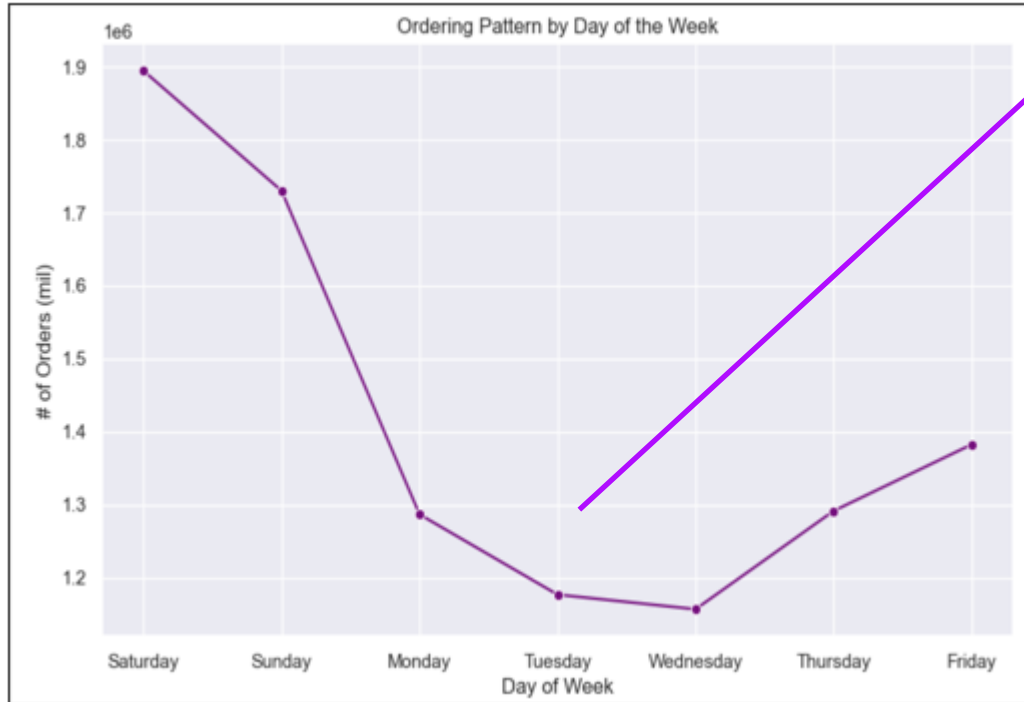
Schedule ads for goods in popular depts: produce, dairy egg, and beverages between 7am-9am to reward early birds ordering in the most popular categories



04 Sales & Marketing Ads for Online Grocer

Excel Reporting - Github

Fig 1.1 - Busiest Days of the Week



Most busy: Saturday, Sunday (weekend)
Average busy: Friday, Thursday, Monday
Least busy: Tuesday, Wednesday (middle of the week)

Most orders occur on the weekend (Saturday and Sunday). *Schedule ads in the middle of the week to boost more sales (Tuesday and Wednesday)*

Most orders happen between 10am- 3pm (order count above 80,000). Most stores are closed and people are sleeping between midnight- 6am (order count below 10,000).

Schedule ads for goods in popular depts: produce, dairy egg, and beverages between 7am-9am to reward early birds ordering in the most popular categories



04 Sales & Marketing Ads for Online Grocer

Excel Reporting - Github

Fig 2.1 - Popular Hours of Ordering

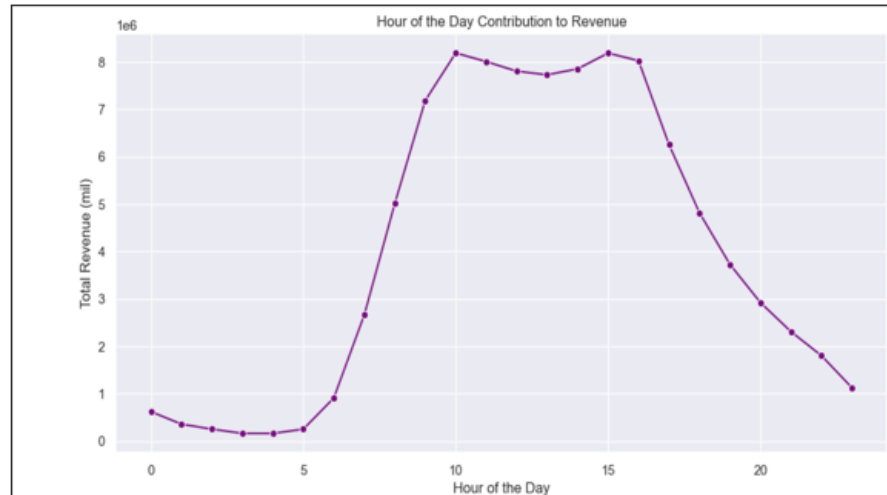
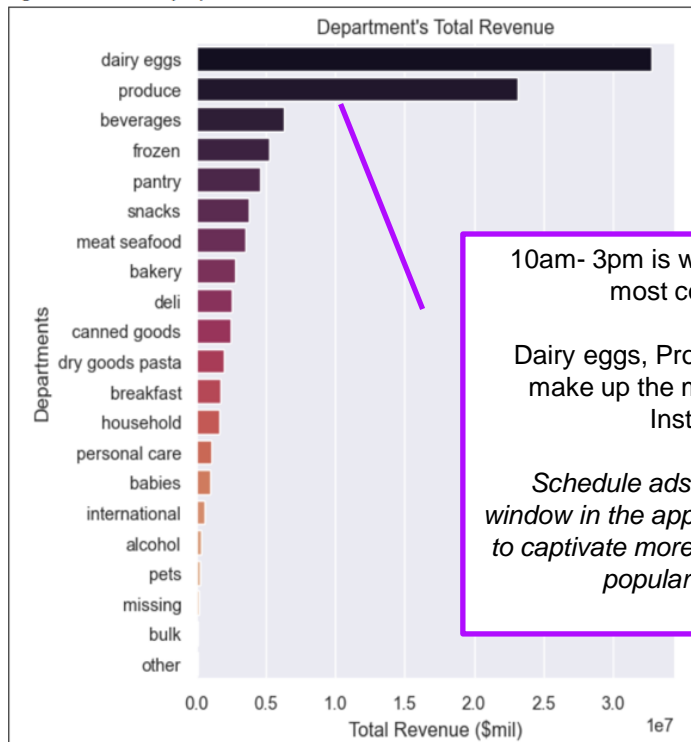


Fig 4.1 - Total Revenue by Department



10am- 3pm is window of revenue most computed.

Dairy eggs, Produce, Beverages make up the most revenue for Instacart.

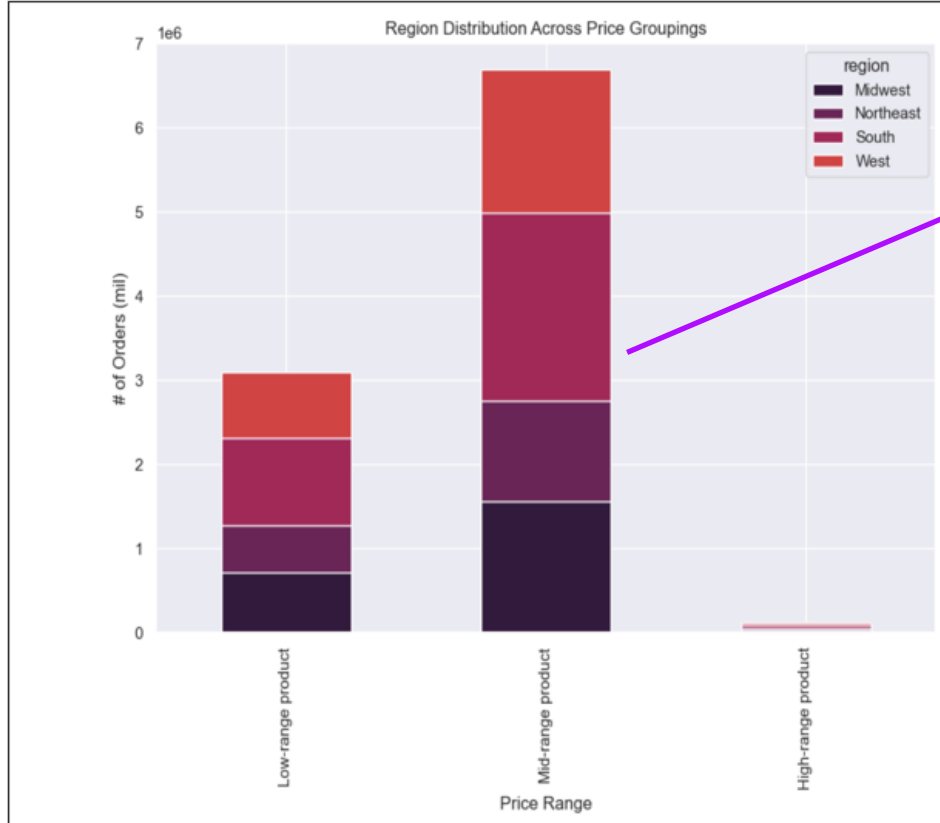
Schedule ads in this specified window in the app using egg products to captivate more spending for these popular products.



04 Sales & Marketing Ads for Online Grocer

Excel Reporting - Github

Fig 3.1 - Price range preferences among US customers



Mid-range is the most bought price range in all the 4 US regions.

Schedule ads about buying mid-range products to all regions during less busy days of Tuesdays- Wednesdays to boost sales.



04 Sales & Marketing Ads for Online Grocer

Excel Reporting - Github

Fig 5.2- Loyalty vs. Regular vs. New by US Region

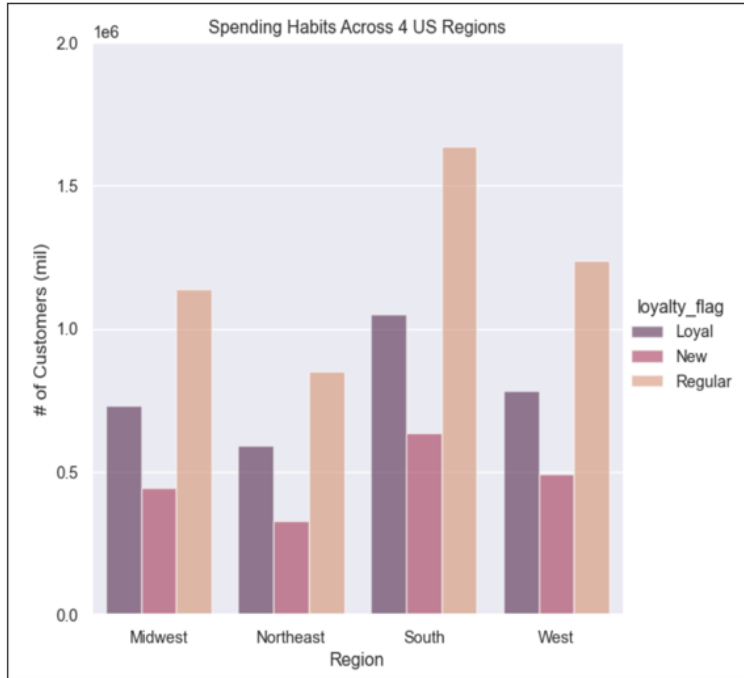
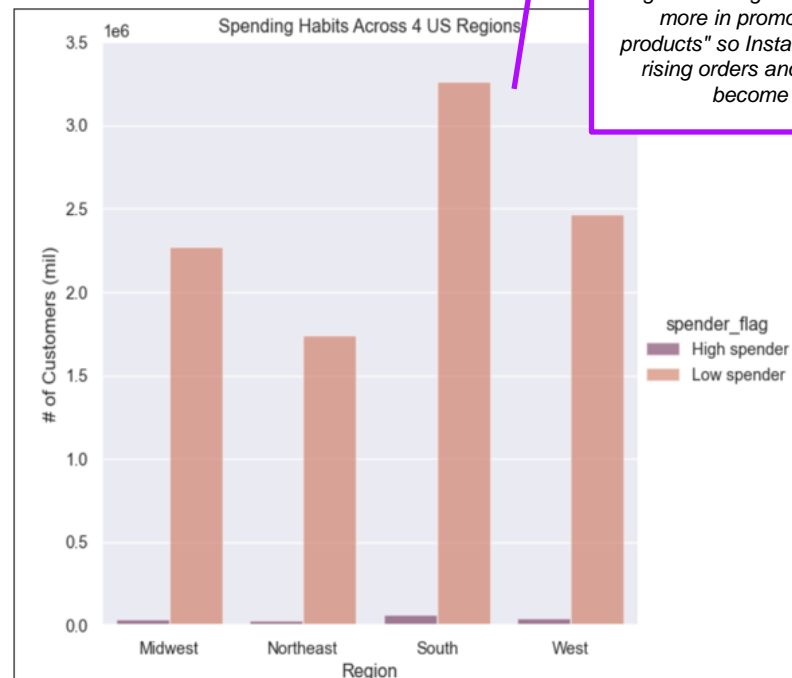


Fig 5.3- Low vs. High spender distribution



Overall, the ratio between Loyal vs. Regular vs. New are similar across all 4 Regions where 10-40 orders per customer is dominant trend. The most loyal users who order 40+ orders are predominantly residing in the South.

Target the "Regular customers in the South" more in promotions of "Low spend products" so Instacart can see an uptick in rising orders and gamify how they can become "Loyal status".



05 Predictive Analysis for Bank

OVERVIEW: Project Brief by CareerFoundry

Analyzed clients' sensitive information data for global bank to assess what specific attributes will lead to client leaving and closing their account.

DATASET AND LIMITATIONS:

Client Database provided by CareerFoundry

Customer demographics are limited to gender, age, and country, account balance, estimated salary, membership status

No transaction data is available for further analysis

KEY QUESTIONS:

1. What are the leading factors that contribute to client loss?

TOOLS USED AND TECHNIQUES:

Microsoft Excel

- Data Integrity, Quality, Consistency Checks
- Descriptive Analysis
- Pivot Tables
- Grouping Data
- Decision Tree



05 Predictive Analysis for Bank

The bank take care of all age group ranges 18-82, but overall average age of loyal customer is ~39 years old.

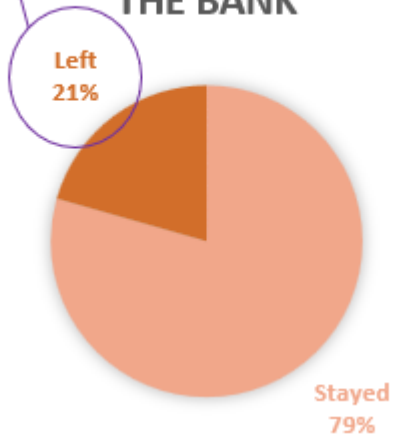
Looking briefly of the descriptive statistics of "Existing Customers", it looks like the average credit score of someone leaving is lower than someone staying (637 vs. 649) and their average balance ~\$12k higher than a loyal customer.

	Current Customers				Exited Customers				Total Customers		
	Min	Max	Mean		Min	Max	Mean		Min	Max	Mean
Credit Score	411	850	652		376	850	637		376	850	649
Country											
Gender											
Age	18	82	38		22	69	45		18	82	39
Tenure	0	10	5		0	10	5		0	10	5
Balance	\$0.00	\$197,041.80	\$74,830.87		\$0.00	\$213,146.20	\$90,239.22		\$0.00	\$213,146.20	\$78,002.72
#Products	1	3	2		1	4	1		1	0	0
Has Credit Card?											
Is Active Member?											
Estimated Salary	\$ 371.05	\$ 199,661.50	\$ 98,942.45		\$ 417.41	\$ 199,725.39	\$ 97,155.20		\$ 371.05	\$ 199,725.39	\$ 98,574.54

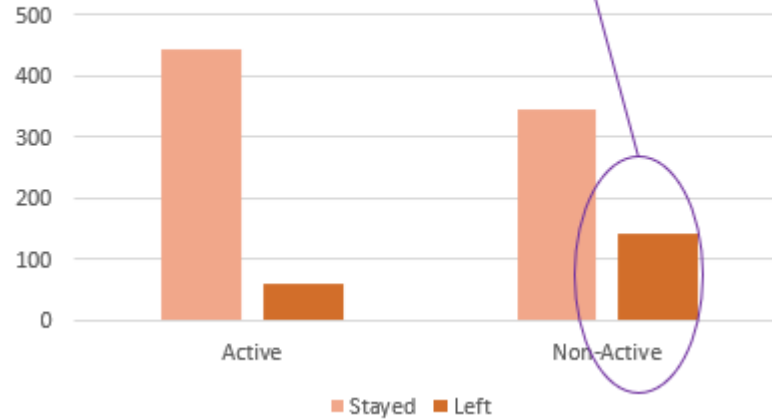


05 Predictive Analysis for Bank

21% OF CUSTOMERS LEFT
THE BANK

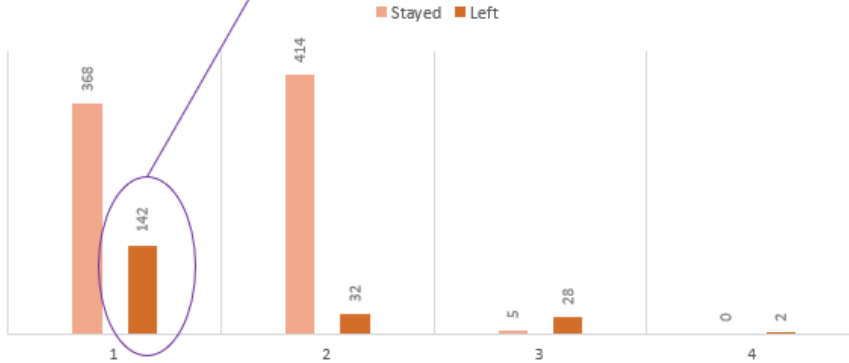


30% OF NON-ACTIVE MEMBERS
DECIDED TO LEAVE BANK

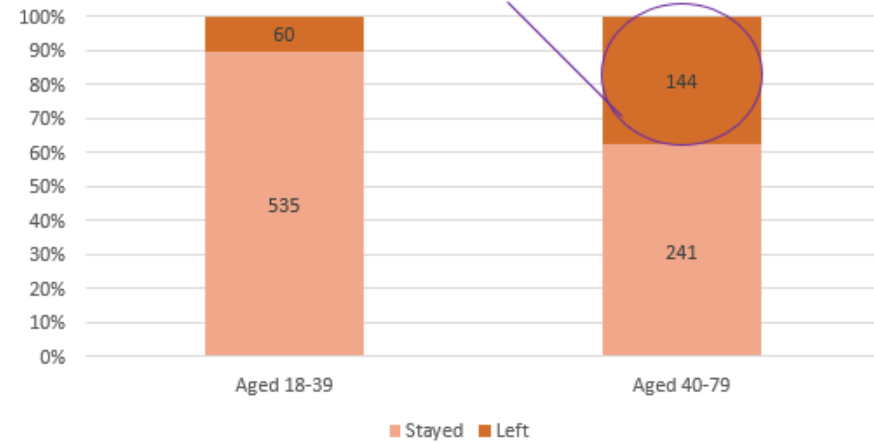


05 Predictive Analysis for Bank

28% OF CUSTOMERS WHO LEFT ONLY HAD 1 PRODUCT USAGE



37% CUSTOMERS AGED 40+ DECIDED TO LEAVE BANK



05 Predictive Analysis for Bank

60% OF CUSTOMERS WHO LEFT WERE FEMALE



Overall, through this customer retention analysis, the top leading factors attributing to someone leaving the bank are:

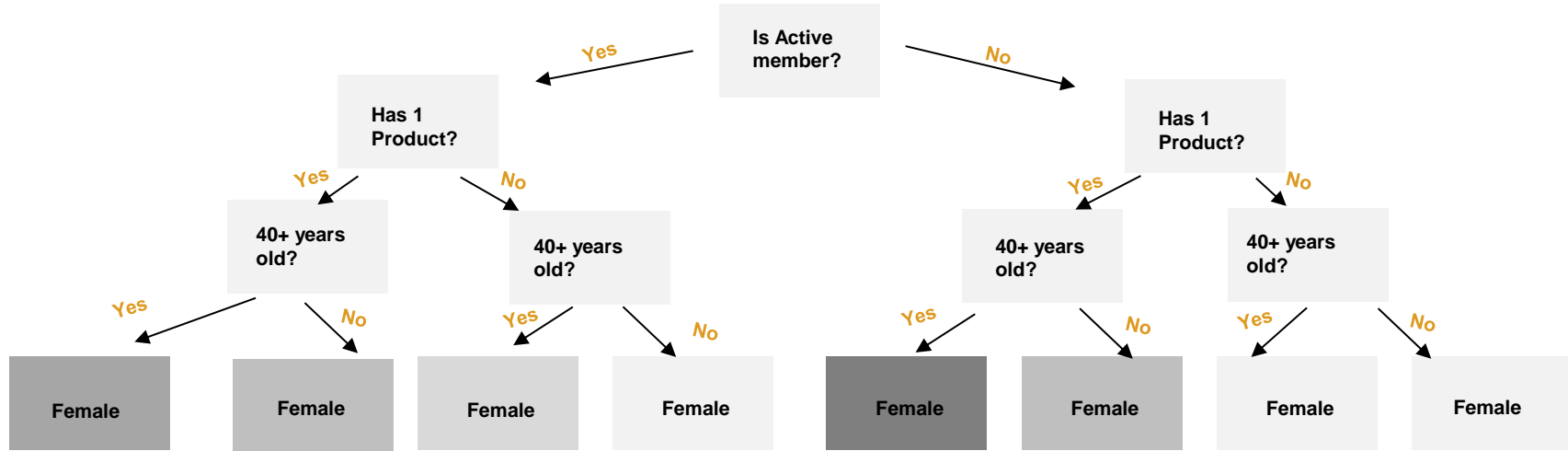
1. Female
2. Aged 40+
3. Used only 1 product from Bank
4. Non-Active

Building out a decision tree with these factors can show the likelihood of a new customer with these attributes could more likely leave the bank and would advise marketing team to target this cohort of promos for retention purposes.



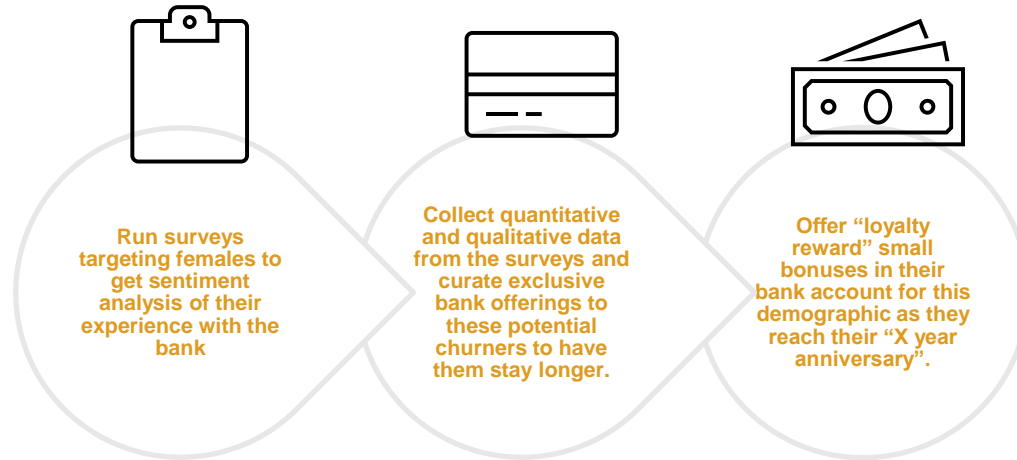
05 Predictive Analysis for Bank

“Will Customer Leave Bank?” Decision Tree



05 Predictive Analysis for Bank

Recommendations for bank's marketing team:



06 Travel Industry Price Analysis

OVERVIEW: Project Brief by CareerFoundry

Airbnb offers short-term rentals all over the world. This project is to analyze key factors influencing listing prices of 20,000+ units in Amsterdam, Netherlands.

DATASET AND LIMITATIONS:

Source: InsideAirbnb & Github

- Data is open-source from Kaggle dated 2018
- CSV format: property listings with coordinates and property specific variables
- Json format: Geodata for Amsterdam

KEY QUESTIONS

1. What is the average price of listings by neighborhood?
2. What are the influencing factors on price?

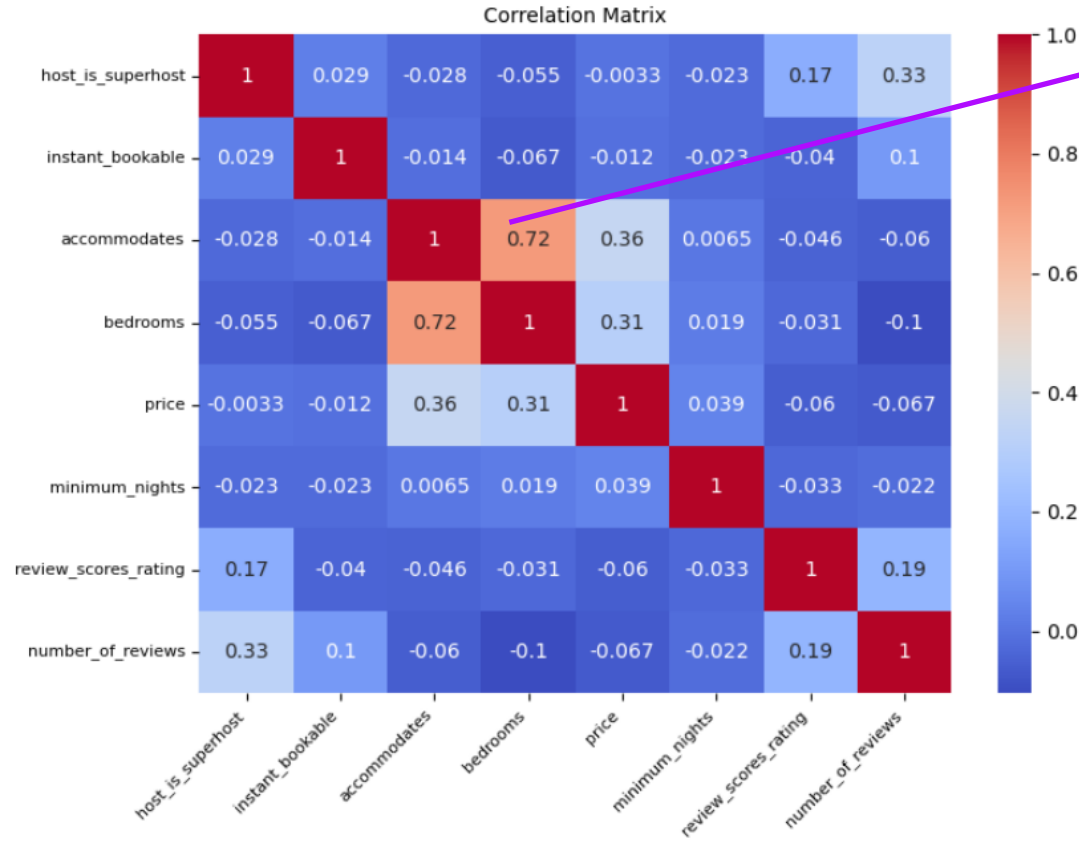
TOOLS USED AND TECHNIQUES:

Python, PowerBI

- Data Sourcing
- Integrity, Quality, Consistency Checks
- Descriptive Analysis
- Correlation Matrix
- Linear Regression
- Clustering Analysis



06 Travel Industry Pricing Analysis



Exploring correlation: There is a high correlation of 0.72 number of beds increase, the listing accommodates increase to more people, which is practical and makes sense.

There is a moderately positive correlation of 0.36 of number of accommodates, price increase. A correlation of 0.33 suggests that listings with more reviews are moderately more likely to have a superhost.

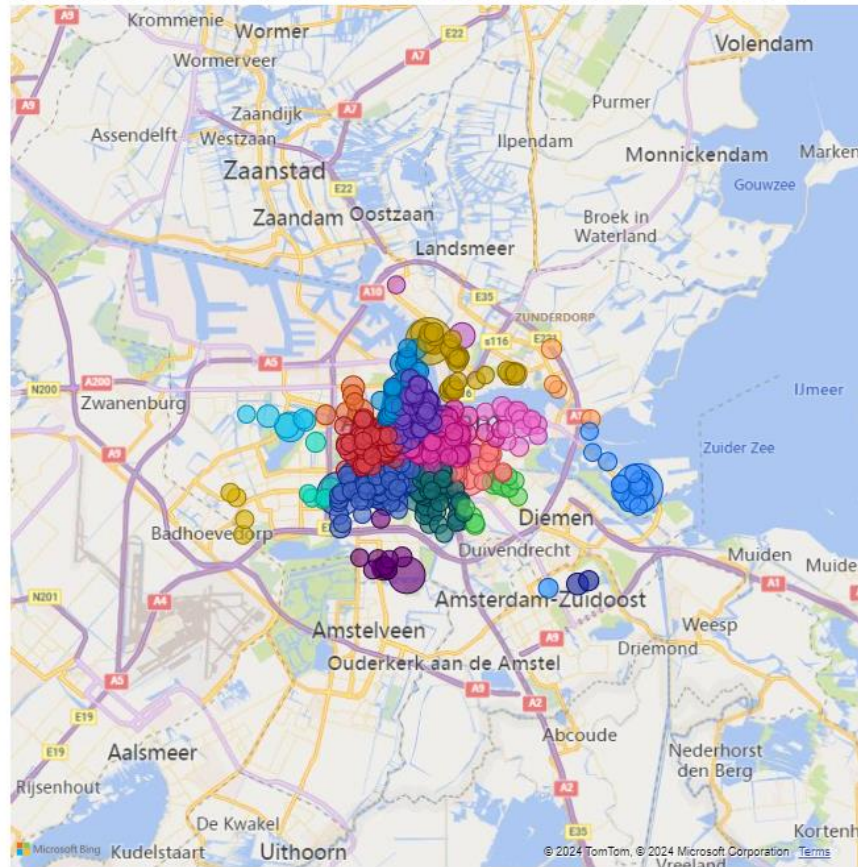
There is a weak correlation of 0.1 if the listing has "instant_bookable" capability, it does not affect "number_of_reviews". There is virtually no linear relationship between "host_is_superhost" and "price" as it is nearly 0 (0.0033).



06 Travel Industry Pricing Analysis

Average of price by neighborhood, latitude and longitude

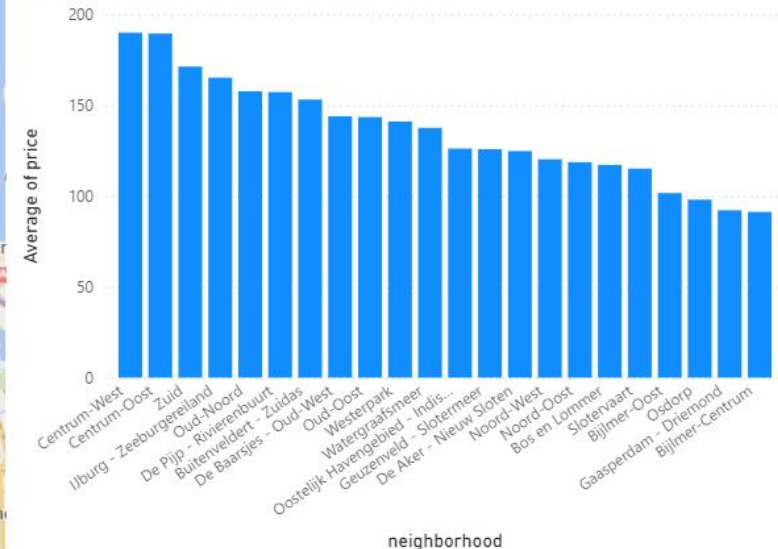
neighborhood ● Bijlmer-... ● Bijlmer-... ● Bos en L... ● Buitenv... ● Centru... ● Centru... ● De Aker - ...



There are so many Airbnb listings in Amsterdam, Netherlands. The average price to stay close to epicenter is ~\$190/night. If you want to stay farther such as Bilmer-Centrum, you can book for ~\$90/night.

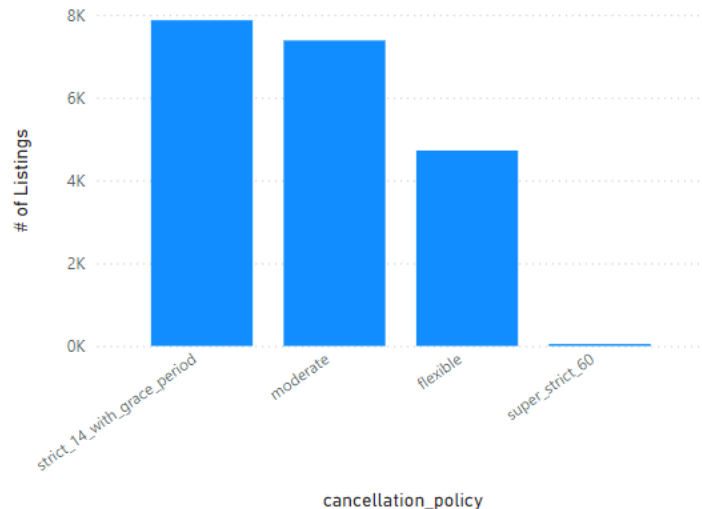
Let's determine if there is a relationship of these attributes to pricing: cancellation, property type, number of reviews and price in the next slides.

Average of price by neighbourhood_cleansed

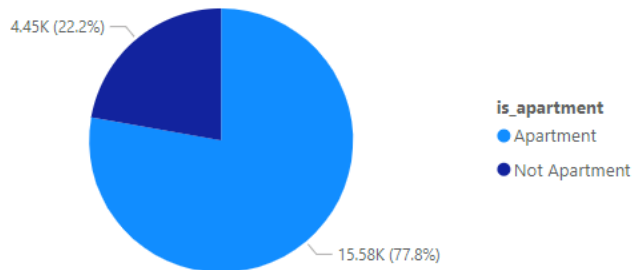


06 Travel Industry Pricing Analysis

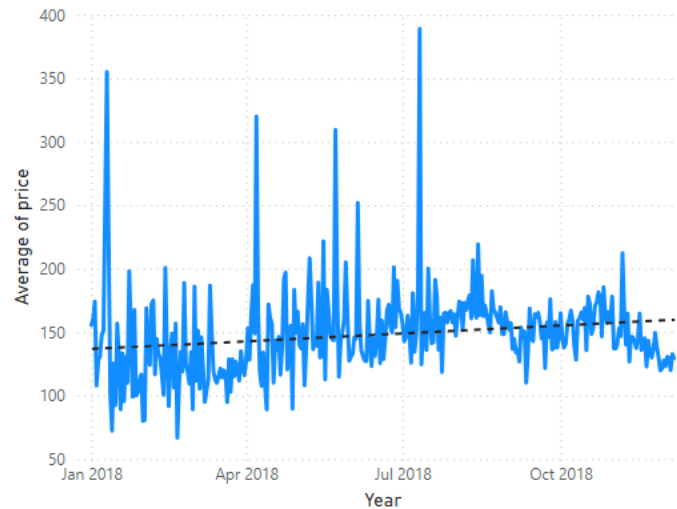
of Listings by cancellation_policy



Count of id by is_apartment



Average of price by Year, Quarter, Month and Day



Over 70% of listings are either "strict" or "moderate" and a majority of the offerings are "apartments". Average prices of booked places have increased throughout 2018 year.



06 Travel Industry Pricing Analysis

number_of_reviews and price



"Number_of_reviews" x "Price" Scatterplot

Hypothesis: If the Airbnb listing has more reviews, then it is more expensive.

I thought places with Airbnb listings that had many more reviews than its competitors would cost more, but that is not the case. Most places with many reviews are priced pretty low; there are only a few expensive ones that stand out. There is no clear pattern connecting the "number of reviews" to the "price".

This is useful insight because listing's number of reviews will not automatically tell me it will be more expensive. I would have to explore the data further to understand the quality vs. quantity of reviews.



06 Travel Industry Pricing Analysis

number_of_reviews and price



"Number_of_reviews" x "Price" Scatterplot

Hypothesis: If the Airbnb listing has more reviews, then it is more expensive.

I thought places with Airbnb listings that had many more reviews than its competitors would cost more, but that is not the case. Most places with many reviews are priced pretty low; there are only a few expensive ones that stand out. There is no clear pattern connecting the "number of reviews" to the "price".

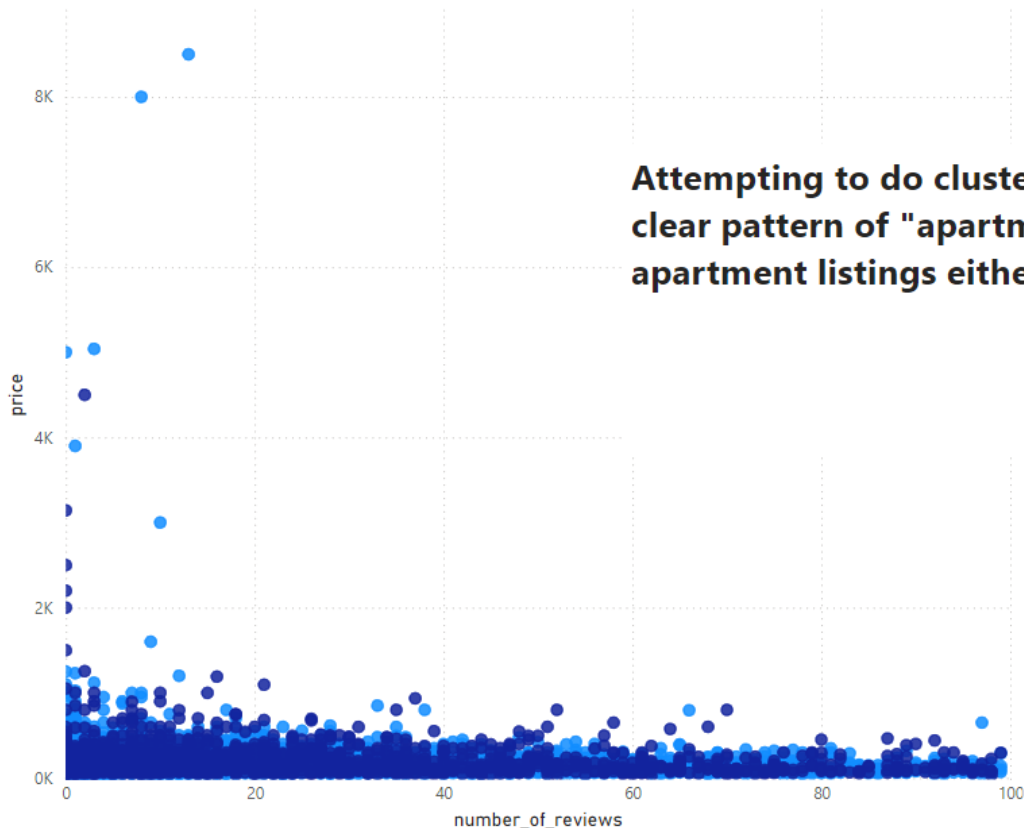
This is useful insight because listing's number of reviews will not automatically tell me it will be more expensive. I would have to explore the data further to understand the quality vs. quantity of reviews.



06 Travel Industry Pricing Analysis

is_apartment, number_of_reviews and price

is_apartment ● Apartment ● Not Apartment



06 Travel Industry Pricing Analysis

Data Limitations & Recommendations:



- The data contained a limited amount of variables upon which to conduct the analysis
- This publicly available data only extracted up to listings as of 2018 (~6 years outdated)



1. Model Enhancement

To better capture the complexities of listing pricing dynamics in Amsterdam, further refinement of the model is recommended. This could mean segmenting the data more create additional variables that may have been overlooked in the initial analysis.



2. Temporal Analysis

Dissecting other seasonal trends in Amsterdam listings beyond 2018 will make the model more accurate of what is the latest average price hosts are charging tourists.



3. Enrich data

Dissecting the object form data of "amenities" and quantify how many "amenities" does the listing offer could add more insight to the model if other sentimental qualities affect the pricing such as "Host greets you"



Resources

CREDITS: Slidesgo.com, Flaticons, Freepik, Google Fonts

Illustrations

- [Gradient style instagram posts template](#)
- [3d geometric shapes background](#)
- [Floating geometric shapes brochure collection](#)
- [3d geometric models landing page](#)

Font packs

- [Golos Text Extra Bold](#)
- [Didact Gothic Regular](#)

