

Data Analysis Portfolio

by BodyArchives, Inc.
2024

Content

Data Analysis Portfolio

- [About The CEO & Lead Analyst](#)
 - Case Studies
 - [#1 GameCo.](#)
 - [#2 Influenza Season](#)
 - [#3 Rockbuster Stealth LLC](#)
 - [#4 InstaCart Basket](#)
 - [#5 US Real Estate](#)
-



Isom J. Winton

CEO and Lead Analyst

Email: isom@body-archives.com

LinkedIn: <https://www.linkedin.com/in/isomwinton/>

GitHub: <https://github.com/isom17>

At BodyArchives, Inc., under the leadership of CEO and lead data analyst Isom Winton, we bring a wealth of experience from the advertising sector to the realm of data analytics. With a history of fostering key relationships and driving revenue growth for global giants like ByteDance, Spotify, and Dentsu Japan, our team is adept at combining traditional business strategies with cutting-edge data insights.

We're committed to leveraging our extensive background in ad sales and business development, along with our analytical expertise, to enhance decision-making processes and drive success in the fast-paced advertising landscape.

Our objective is to optimize marketing efforts and maximize return on investment through the integration of modern data analysis.

Case Study #1: GameCo

Project Overview

Leverage data analytics to provide strategic insights that will guide the development and market positioning of GameCo's upcoming video game releases. This involves a thorough descriptive analysis of relevant video game datasets to identify market trends, player preferences, and competitive benchmarks.

Key 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?

Data

This data was drawn from [VGChartz](#) website.

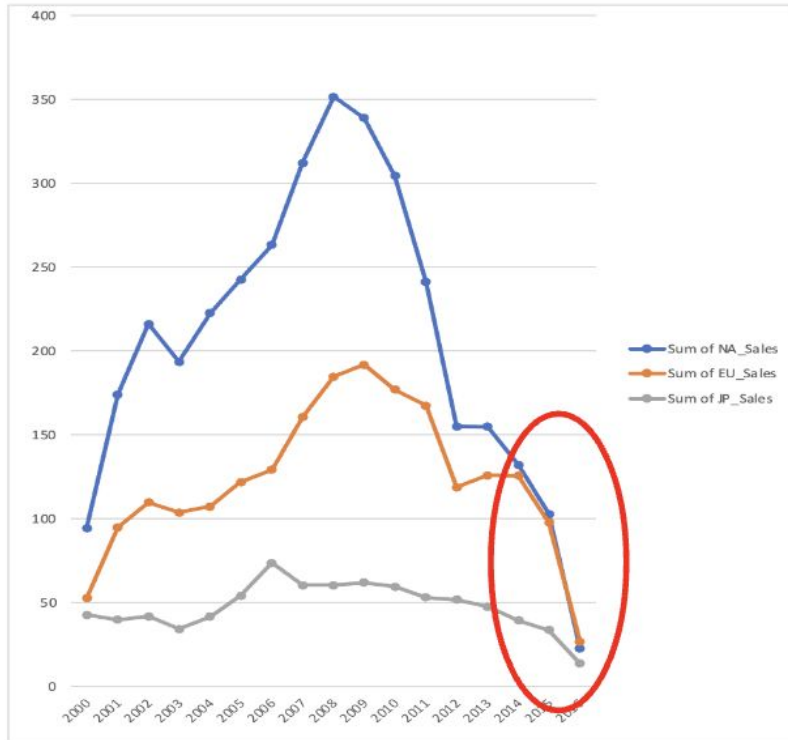
[GameCo Data](#)

Skill Set

- Advanced Excel
- Grouping data
- Summarizing data
- Descriptive analysis
- Visualizing results in Excel
- Presenting results

Total Game Sales by Region (NA, EU, JP)

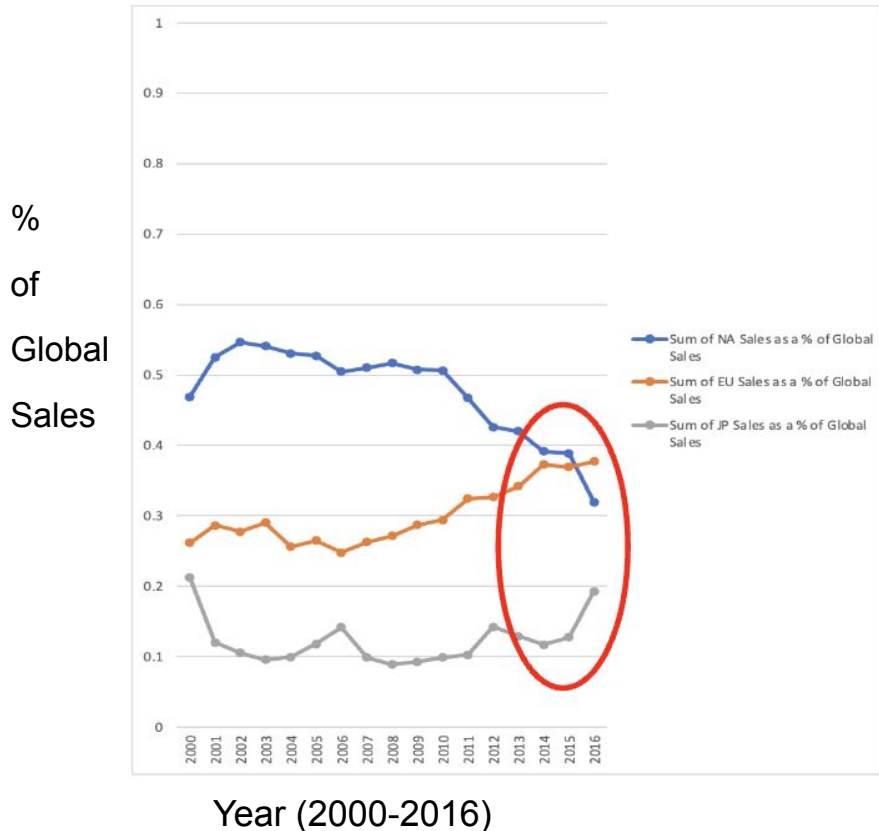
Total Sales
(in
thousand
units)



Year (2000-2016)

The line chart illustrates total sales have been declining in each region over the past few years with NA and EU experiencing the most drastic declines. The data also tells us sales have not been the same over time which challenges our original hypothesis.

Percentage of Global Game Sales of Target Regions (NA, EU, JP)




The line chart shows the percentage of sales contributing to overall global sales in the EU has been steadily increasing over time.

On the other hand, the percentage of sales in the US has been declining since 2010. The amount of sales Japan has contributed has been relatively stagnant since 2000.

Conclusion and Recommendations

Based on the data gathered, we can see our initial assumption of the sales for each region remaining “the same” throughout the recent years has been challenged. Rather, we notice declines in sales for each region based on the line chart in slide #3.

Moving forward, we recommend allocating the marketing budget based on the percentage each region has contributed to overall sales. The line chart in slide #4 shows the ratio of EU sales contributing to overall sales has steadily increased over the years. Specifically, we’d recommend allocating 40-50% of the marketing budget to the EU market focused on high-revenue generating game genres as well as high-potential game genres.





Case Study #2: Preparing for Influenza Season

Project Overview

Optimize the allocation of temporary medical staff to hospitals across all 50 states during the US influenza season, focusing on meeting increased patient needs, particularly in vulnerable populations. The objective is to determine the timing and number of staff required in each state for the upcoming flu season.

Requirements

- Provide information to support a staffing plan.
- Determine whether influenza occurs seasonally or throughout the entire year.
- Prioritize states with large vulnerable populations.
- Assess data limitations

Data

Influenza deaths by geography, time, age, and gender—Source: [CDC](#)

[Population data set by geography](#)—Source: US Census Bureau

[Final Tableau Dashboard](#)

[Final Presentation Video Link](#)

Skill Set

- Advanced Excel
- Translating business requirements
- Data cleaning
- Data integration
- Data transformation

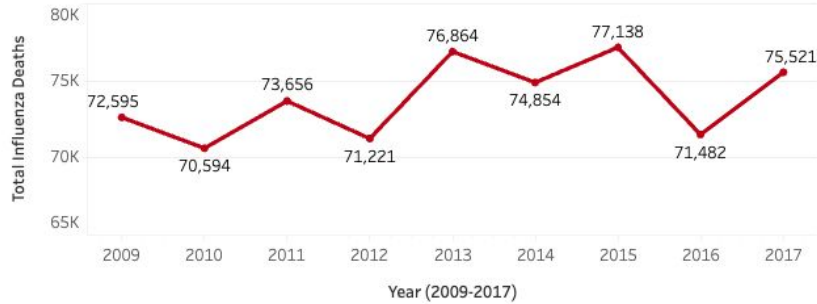
- Statistical hypothesis testing

- Visual analysis
- Forecasting
- Storytelling in Tableau

- Presenting results to an audience

Visualizations

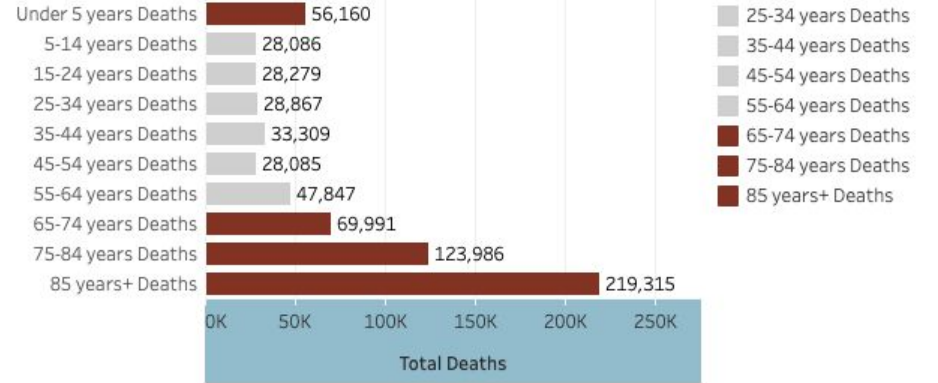
Annual Influenza Deaths in the US



Annual influenza deaths in the United States has remained between 70k to 80k between 2009-2017. Our analysis attempts to provide possible solutions to lower these numbers by focusing on medical facilities' staffing needs in high need areas and high risk populations.

Total Influenza Deaths in the US by Age Group in the US Between 2009-2017

(Vulnerable Deaths = Below 5 or Over 65 Years Old)



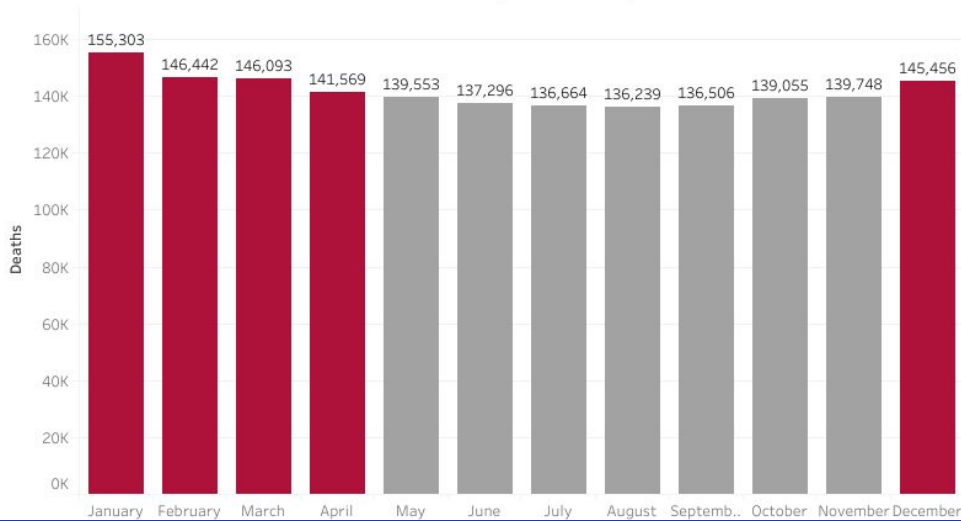
People at greater risk of severe disease or complications when infected are grouped into what is called "**Vulnerable Populations**". These populations include:

- 1) Pregnant women
- 2) **Children under 5 years of age**
- 3) **People ages 65 and older, and**
- 4) Individuals with chronic medical conditions or immunosuppressive conditions.

71% of total influenza deaths were from Vulnerable Populations.

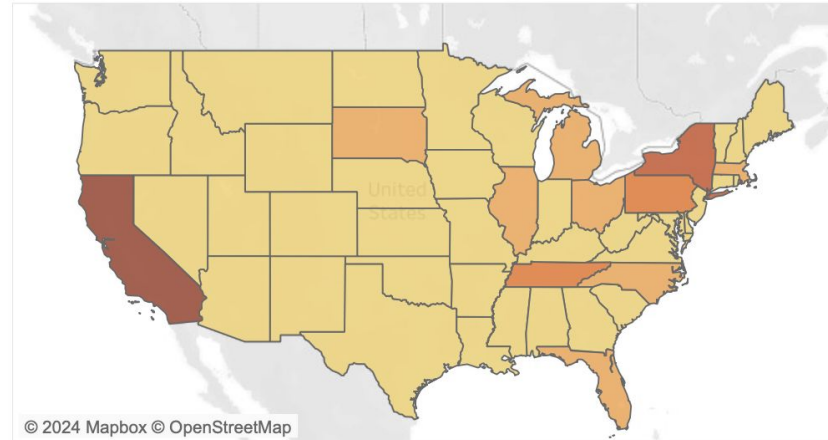
Visualizations

Total Influenza Deaths by Month in the US (2009-2017)



Our analysis demonstrates most of the vulnerable population deaths due to influenza occurred between the months of December to April providing us with a basis to determine seasonality.

Vulnerable Deaths by State (2009-2017)



Allocating resources to the states with the most vulnerable population deaths by grouping them into 2 tiers: High-Need and Medium-Need.

Recommendation/Limitations

Medical staffing companies facilitating the personnel needs across the United States should understand the historical data as is analyzed in this project. Our recommendation is to deliver staff based on the data on "Vulnerable Deaths" in each state between November to April each year. We have broken down the staffing needs of the states into 2 categories: High-need and Medium-need.

High-need:

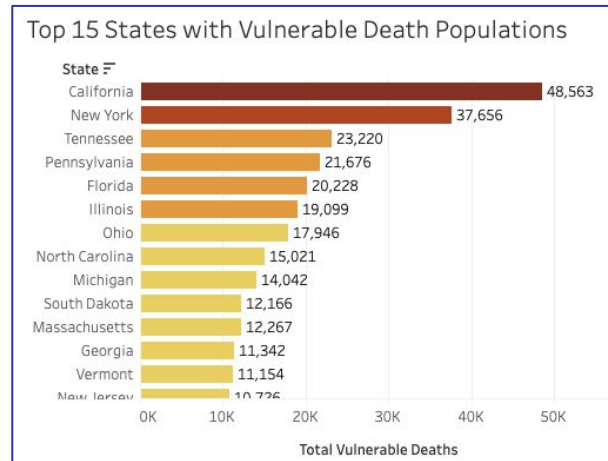
California, New York, Pennsylvania, Tennessee, Illinois, Florida, Ohio

Medium-need:

North Carolina, Michigan, South Dakota, Massachusetts, Georgia, Vermont, New Jersey, Missouri

Data Limitations

Moving forward it would be well advised to understand the vaccine rates of Vulnerable Populations each state to glean insights into preventative care that would further support the reduction of deaths caused by influenza.





#3 Case Study: Rockbuster Stealth LLC

Project Overview

Rockbuster Stealth LLC, a global movie rental company facing competition from streaming platforms, is transitioning to an online video rental service using its existing licenses. I'm tasked with loading the company's data into a relational database management system (RDBMS) and utilizing SQL for data analysis to support various departmental queries and inform the launch strategy. The project's goal is to leverage SQL skills for detailed analysis, contributing to the strategic planning and successful launch of Rockbuster's new online service.

Key Questions

- Which movies contributed the most/least to revenue gain?
- What was the average rental duration for all videos?
- Which countries are Rockbuster customers based in?
- Where are customers with a high lifetime value based?
- Do sales figures vary between geographic regions?

Data

[Rockbuster Data Set](#)

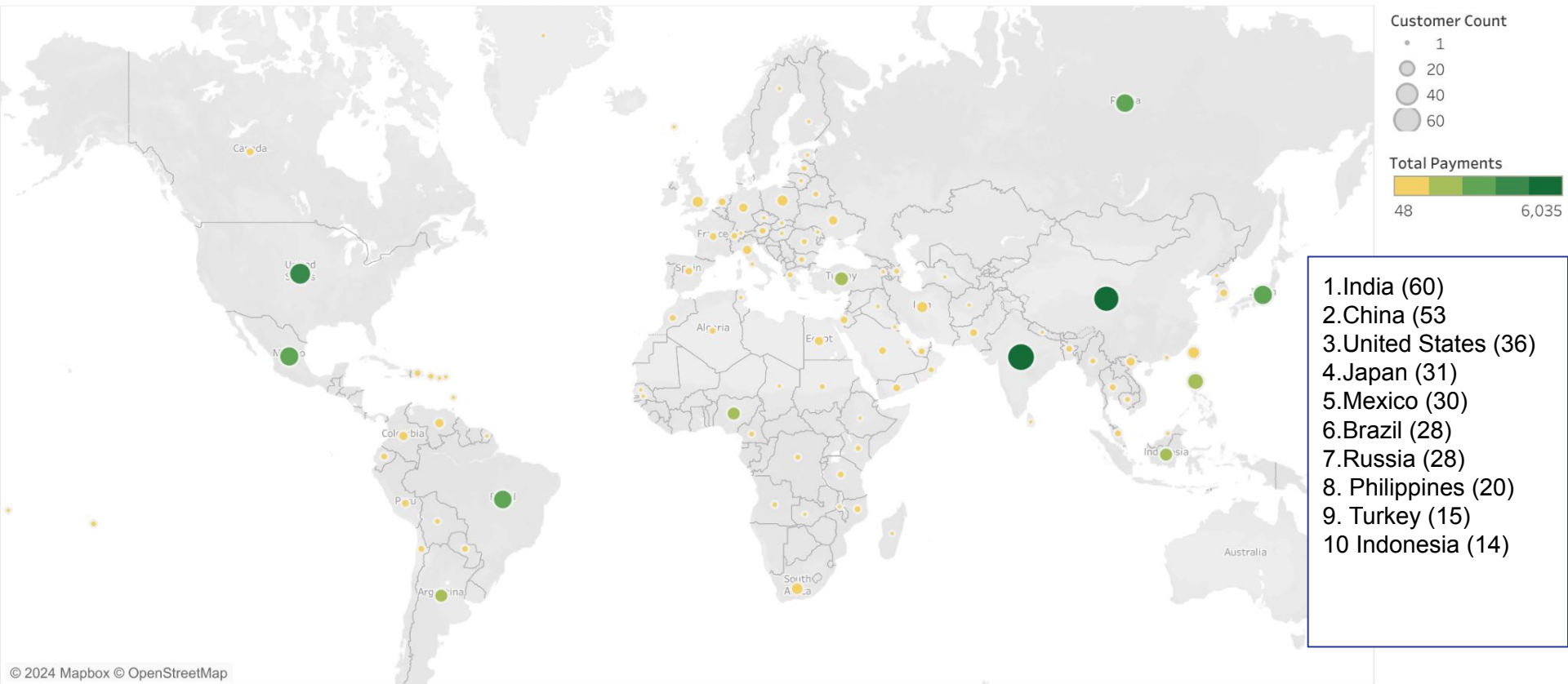
[Final Proposal](#)

[GitHub Repo](#)

Skill Set

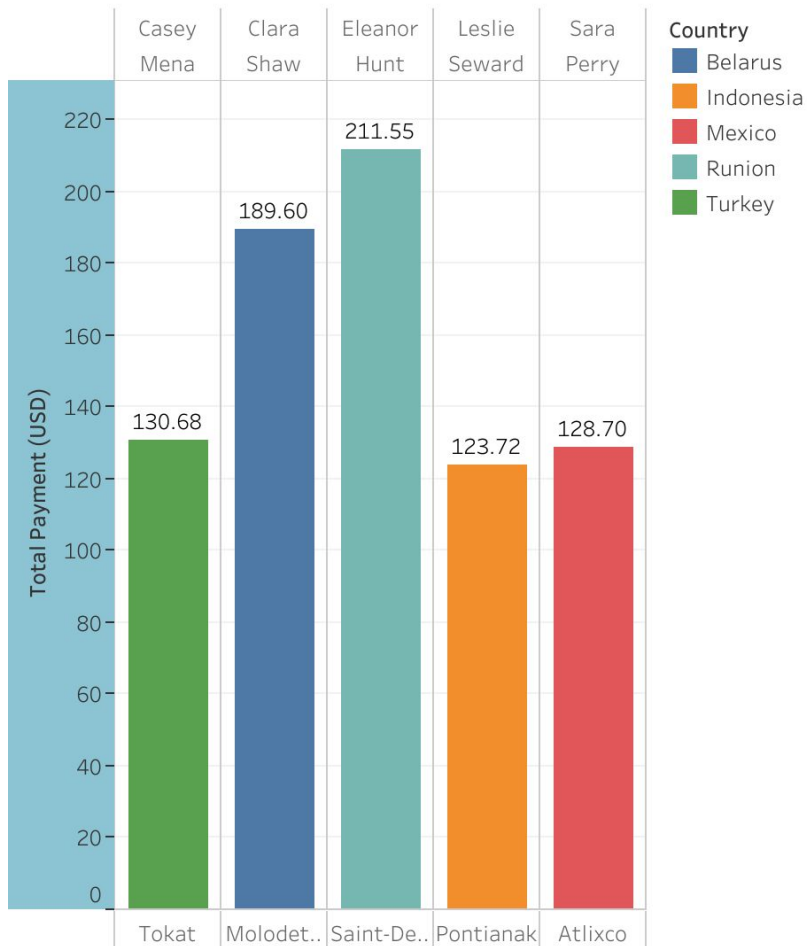
- Relational database Management
- PostgreSQL
- Database querying
- Filtering, cleaning and summarizing
- Joining tables
- Subqueries
- Common table expressions
- Tableau

Geographical Distribution of Customers and Revenue Generated



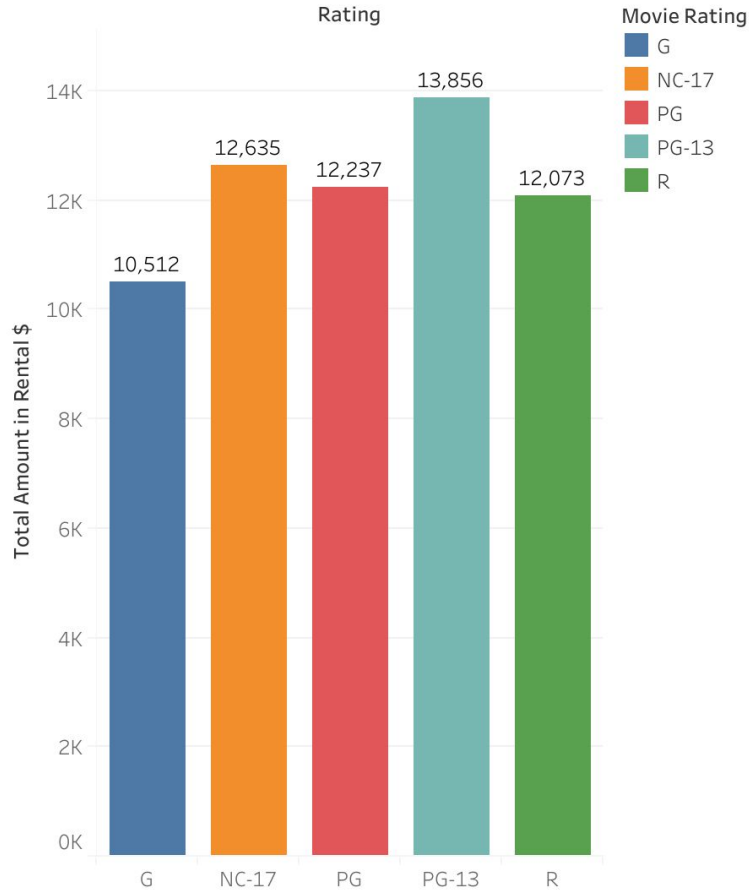
The top 10 markets in terms of revenue are indicated by the color of circle on the map while the countries with the most customers are will have a larger circle circumference. We can in infer the countries with the largest number of customers, large circle size, will also generate the most revenue, dark green color. India, China and the United State are the top revenue generating and have the most customers.

Top 5 Customers and Their Location



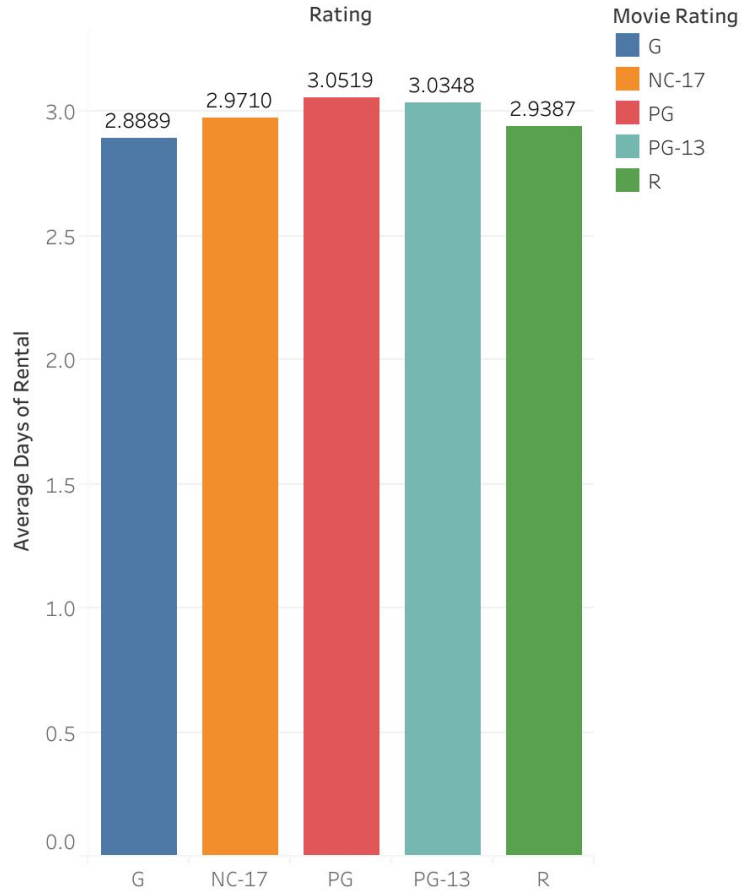
The bar chart to the right illustrates where the top revenue generating, highest LTV customers reside. Both Mexico and Indonesia are in our list of top customer count and average amount of revenue generation.

Amount of Revenue Generated Per Movie Rating



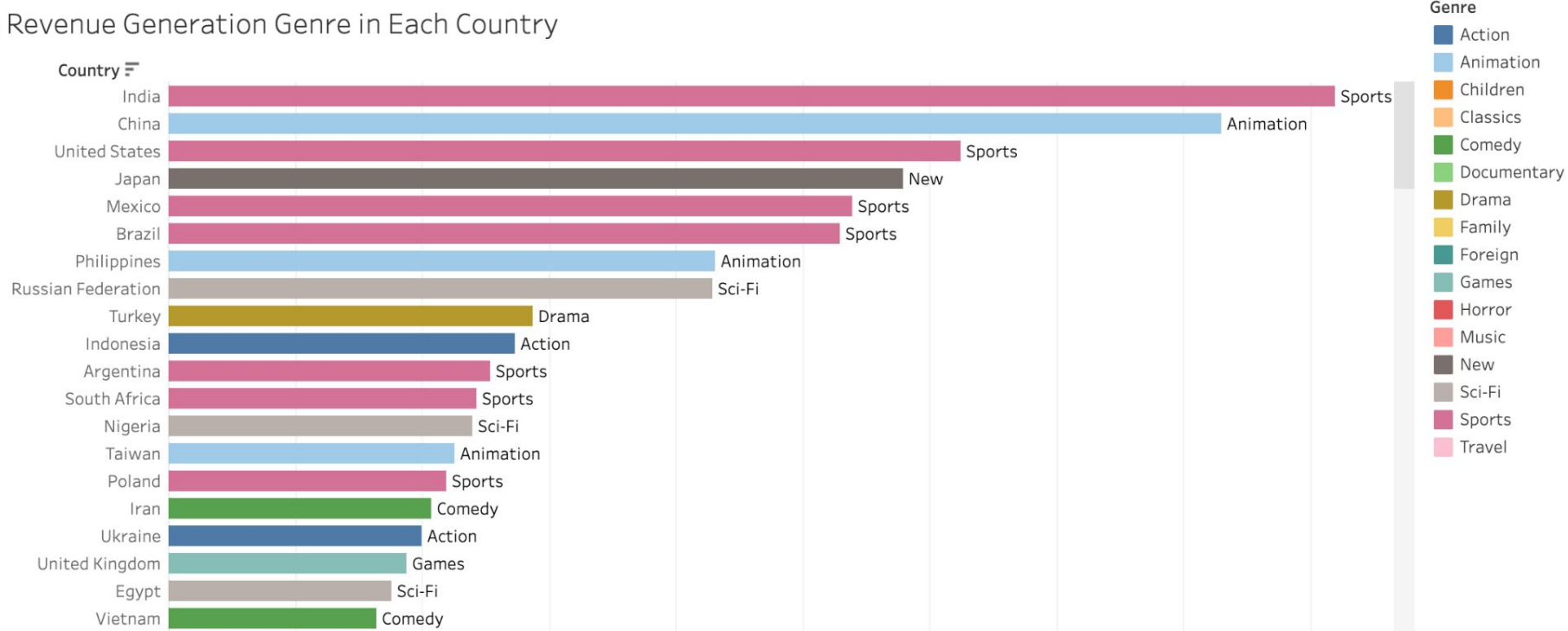
PG-13 is the greatest revenue generating genre amongst the 5 rating categories with \$13,856.

Average Rental Rate (Days) Per Movie Rating



PG and PG-13 rated movies have the highest average rental rate amongst the rating categories.

Top Revenue Generation Genre in Each Country



The bar chart above shows the highest revenue generating movie genre for each country. We notice the top genre for the top revenue generating markets is Sports.

Recommendations Moving Forward

Target Markets

- Focus on the top revenue generating markets such as India, China and the United States with an emphasis on on the top revenue generating movie genre and language in those countries.

Movie Genre & Rating

- Sports, Sci-Fi and Animation along with PG and PG-13 movies are the top revenue generating genres and ratings respectively so it would be best to focus on promoting genres with these ratings in mid-tier markets.

Licensing, Pricing & Loyalty

- Decide on which movies and movie genres are not contributing to enough to revenue and discontinue those licenses.
- Increase the prices on popular movies, genres and reward customers with high lifetime values.

#4 Case Study: InstaCart Basket

Project Overview

Conduct a detailed exploratory analysis to understand sales patterns, aiming to enhance customer segmentation and personalize marketing strategies. The objective is to analyze customer behaviors and purchasing trends to inform targeted marketing campaigns, ensuring the right customer profiles are approached with suitable products. Key areas of focus include identifying the busiest shopping days and hours, understanding spending habits across different times, categorizing products by price ranges & departments, and examining brand loyalty.

Key Questions

1. What are the busiest shopping days and hours for Instacart, and how can this information optimize ad scheduling?
2. During which times of day do customers tend to spend the most, and how can this influence targeted advertising strategies?
3. How can Instacart categorize its diverse product range into simpler price brackets to streamline marketing efforts?
4. Which product departments see the highest order frequencies, indicating their popularity among customers?
5. How do customer demographics, including loyalty status, region, age, and family status, impact their purchasing behaviors and preferences?

Data

[InstaCart Population](#)
[Flow Excel](#)

[InstaCart Project Folder](#)

[GitHub Repo](#)

Skill Set

Python Programming

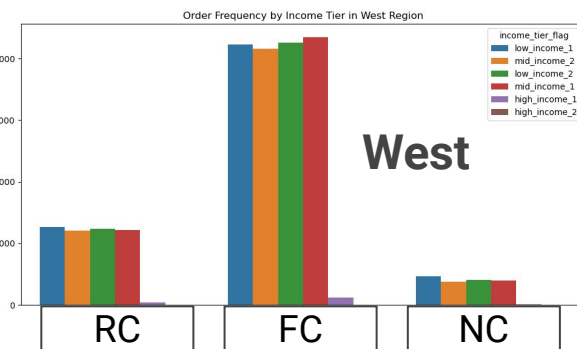
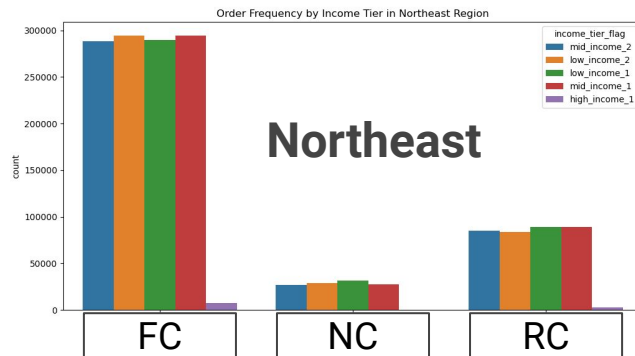
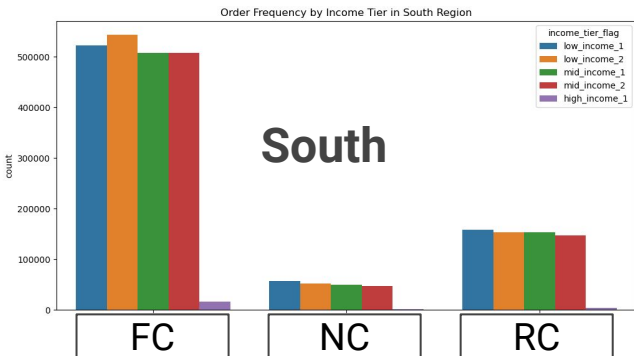
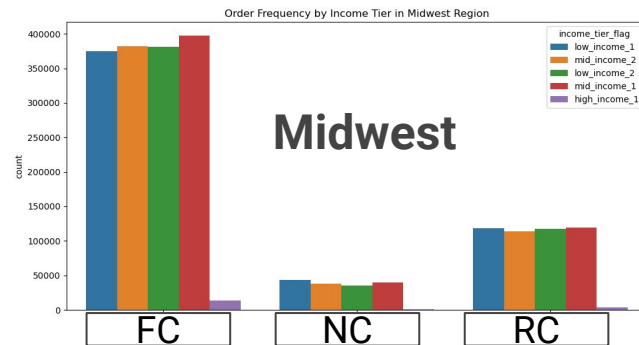
Data wrangling, Data merging, Deriving variables, Grouping data, Aggregating data, Creating graphs, Correlation analysis

-Reporting in Excel

-Population flows

Visualizations

RC= Regular Customer
FC = Frequent Customer
NC - Non-frequent Customer



Over 70% of Instacart customers are 'Frequent customer' as categorized by the median number of days before their next order is less than or equal to 10 days. 'Regular customer' are categorized by having made their last order between 10 and 20 days while the 'Non-frequent customer' is characterized by taking over 20 days until their next order. By region, we can see the South has the largest number of 'Frequent customer'. Our graphs also show there is not a major difference between the annual income between 'Frequent customer' which shows the platform appeals to all family income types.

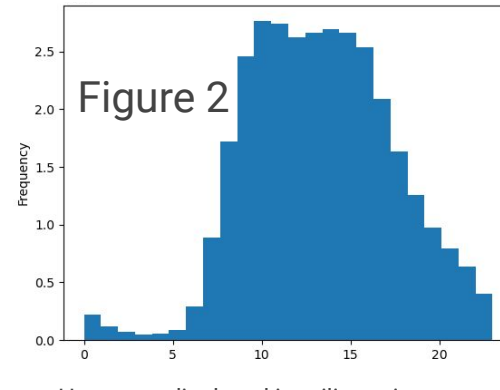
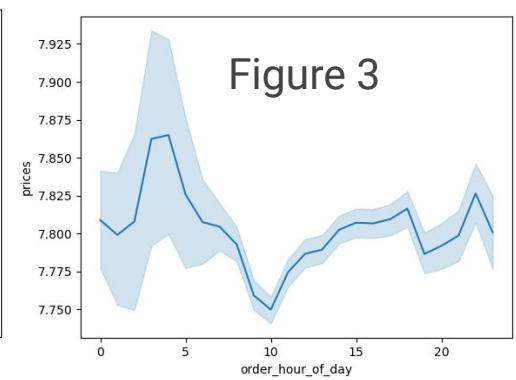
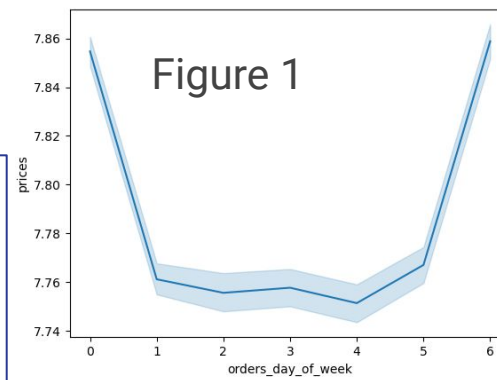
Visualizations

Figure 1- This graph shows the frequency of the amount of orders each day of the week receives. 0 = Saturday, 1= Sunday, 2= Monday...6= Friday. We can see Saturday and Sunday receive the most orders while Wednesday and Tuesday receive the least amount of orders.

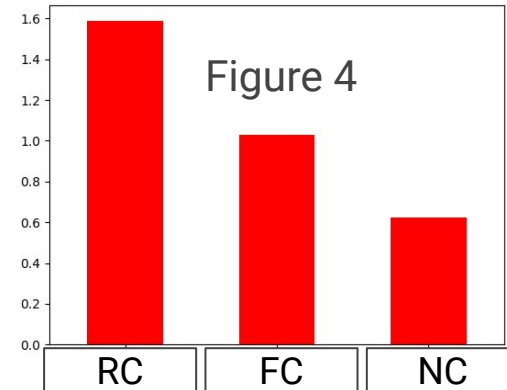
Figure 2 - This is a histogram of the most popular order hours of the day. This graphs shows us the majority of orders happen between 9am to 3pm.

Figure 3- Although the most popular times to order are between 10am to 3pm, the evening and early mornings receive the highest order dollar amounts.

Figure 4- This bar chart shows that Regular Customers have placed the most orders overall.



Hours are displayed in military time on the x-axis



From left to right (Bar 1 = Regular Customer, Bar 2= Loyalty Customer and Bar 3 = Non-frequent Customer)

Visualizations

Age Group Ratio

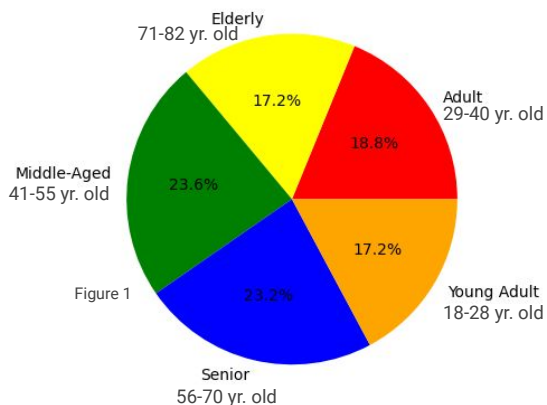


Figure 1

Revenue by Top 20 Customer Profiles

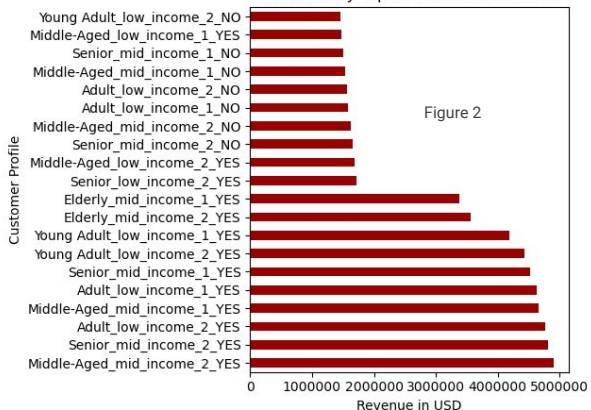


Figure 2

Income Tiers:

Low Income 1: < \$67,321
 Low Income 2: \$67,321 - \$96,775
 Mid Income 1: \$96,776 - \$128,105
 Mid Income 2: \$128,106 - \$220,681
 High Income 1: \$220,682 - \$593,900
 High Income 2: ≥ \$593,901

Dependants in the Household:

Yes = 1 or more dependants
 No = Zero dependants

Popular Departments by Region

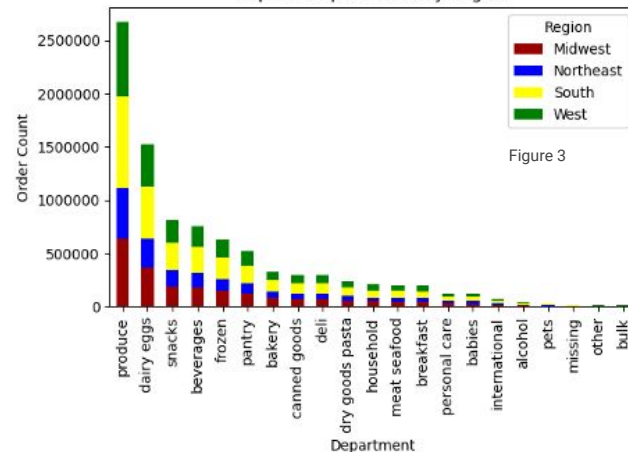


Figure 3

The graphs above provide more insight into the demographics and ordering preferences of InstaCart's customer base. The age demographics are fairly evenly distributed however, we see that more mid-income customers are contributing more to the overall revenue of the business as displayed in Figure 2. Figure 3 illustrates the most popular departments customers order from across regions and the produce department is by far the most sought after.

Recommendations for InstaCart

Ad Timing

Schedule advertising campaigns during less busy order times to capture attention when potential customers are more likely to be browsing. Specifically, consider late nights and early mornings before 5am, as well as on Wednesdays and Tuesdays which are the least busy days.

Popular Departments

Tailor ads to highlight products from the most popular departments, namely produce, dairy eggs, snacks, beverages, and frozen, as these have the highest frequency of orders.

Loyal Customers

Since loyal customers constitute 67.5% of the customer base and have a lower average spend than new customers, create specialized offers for them during off-peak hours to increase their average spend. This could include special deals or highlighting unique products from the popular departments.

Recommendations for InstaCart

Upselling New Customers

Implement bulk discounts, package deals, and other incentives for new customers, who currently have the highest average spend and order frequency. The goal is to encourage them to increase their cart size and frequency of orders, converting them into regular or loyal customers.

Regional Targeting

Given that the South has the largest number of 'Frequent customers', regional ad campaigns could be more heavily targeted there. However, since the income level does not significantly vary between customer types, it suggests that the platform has wide appeal, and campaigns should not be overly segmented by income.

Demographic Focus

Tailor ad campaigns to Young Adults, Adults, and Middle-Aged groups with low to mid incomes, as these demographics are the most active and revenue-generating on the platform. Ensure the ads resonate with these age groups, possibly by highlighting convenience, value, and variety.

Summary

InstaCart Analysis

Instacart should consider a two-pronged approach: (1) Create value-driven campaigns for new customers to boost their lifetime value, and (2) Develop tailored offers for loyal customers to increase their spending during less busy times, with a focus on the most popular product categories. Additionally, the company should leverage regional and demographic insights to optimize the targeting and content of their ad campaigns.



Case Study #5: US Real Estate

Project Overview (Capstone Project)

This case study focuses on real estate properties in the United States that were for sale in 2021. Our case study attempts to uncover what variable(s) have a significant impact on the price of a property such as bedroom count, bathroom count or the size of the home itself.

Requirements

- Be open source
- Come from an authoritative source
- Include non-anonymized column names
- Be no more than 3 years old
- Contain at least 2 continuous variables
- Contain at least 2 categorical variables
- Contain at least 1,500 rows
- Include a geographical component with at least 2 different values (e.g., countries, continents, U.S. states, cities, latitude and longitude values)

Data

-The dataset we employed was sourced on [Kaggle](#) and collected from <https://www.realtor.com> which is a real estate listing website operated by News Corp subsidiary called Move, Inc. based in Santa Clara, California.

-[Tableau Final Dashboard](#)

-[GitHub Repo](#)

Skill Set

- Python Programming (scikit-learn, pandas, matplotlib, seaborn)
- Exploratory analysis through visualizations (scatterplots, correlation heatmaps, pair plots, and categorical plots)
- Geospatial analysis using a shapefile
- Regression analysis
- Cluster analysis
- Time-series analysis
- Tableau (Analysis narrative and final results)

Visualizations

Figure 1

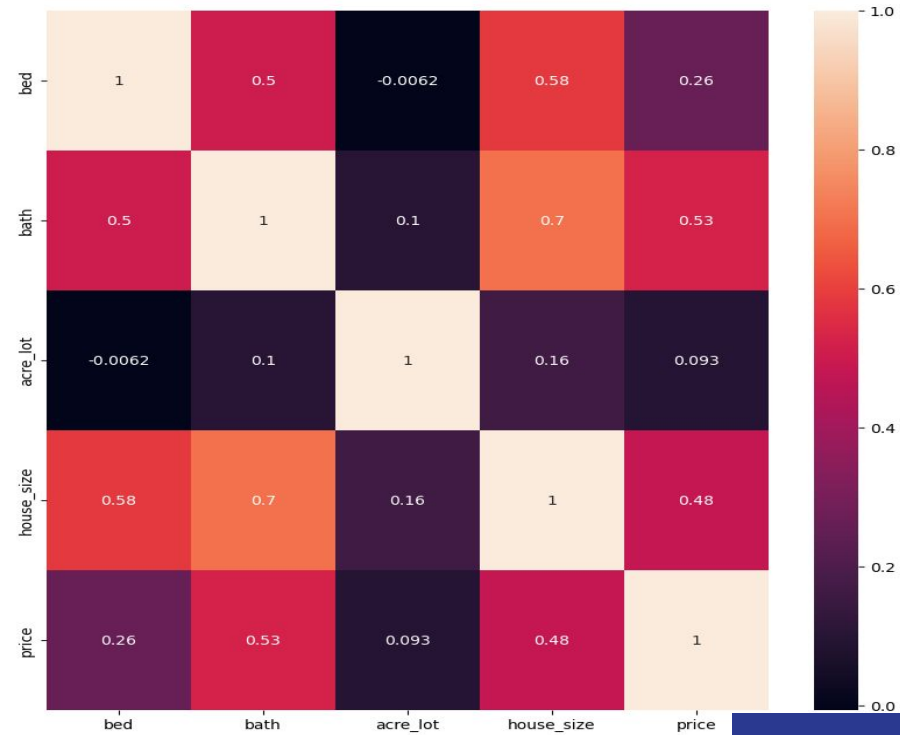


Figure 2

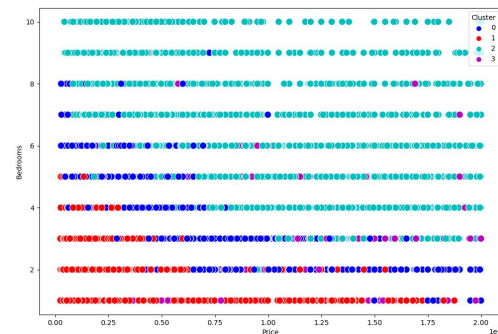
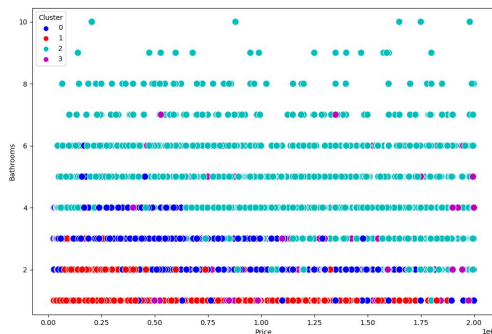
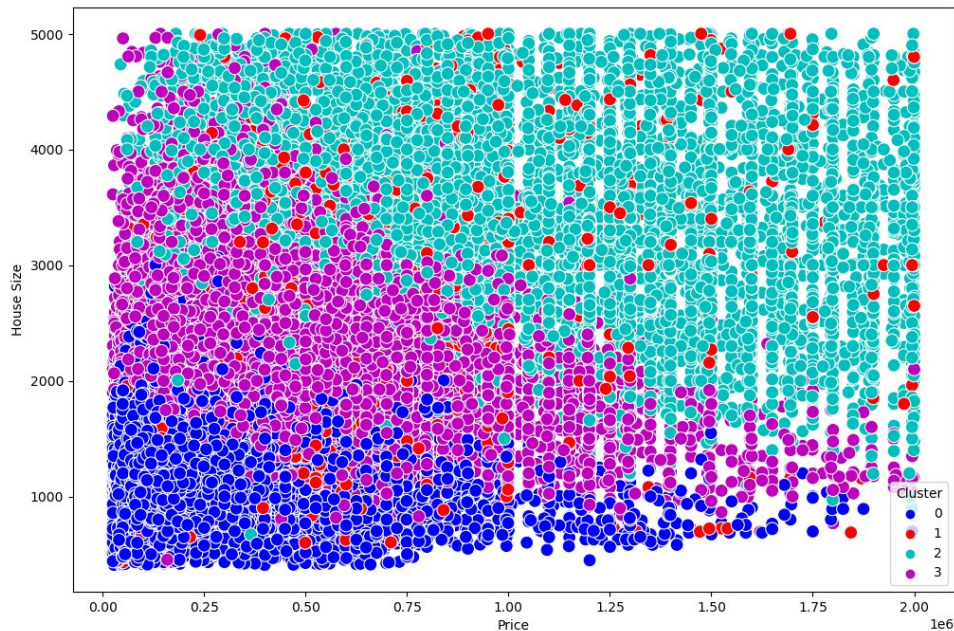
Performed exploratory analysis by creating a pair plot (fig.1) of the variables in the dataset to determine positive or negative trends. We can see observe positive correlations between "Price" and key variables of 'bed', 'bath' and 'house_size' based on these scatter plots but needed to calculate the significance statistically to better understand the strength of the correlations. Our correlation matrix (fig.2) proves only moderate relationships between the price of properties and other key variables which suggests there are many data points that fall outside of our hypothesis and need to explore other methods to explain our data.

Visualizations

Our correlation analysis was not sufficient to explain the relationship between property price and other key variables so we initiated cluster analysis.

Our cluster analysis created 4 groups for each of the key variables' comparisons against price. The analysis further confirms the diversity of property in our dataset.

The data suggests the properties in this dataset vary significantly across different geographical locations, particularly between urban and rural areas.



Recommendation/Limitations

US Property in the Northeast

-The data analysis confirms there are only moderately strong correlations between the key variables and market price suggesting varying property types and features across different geographical locations, particularly between urban and rural areas.

-In this scenario, the best use of the data is for a real estate agent to offer this [Dashboard](#) to their potential clients looking for properties in specific Northeast states, with a specific budget and other features.

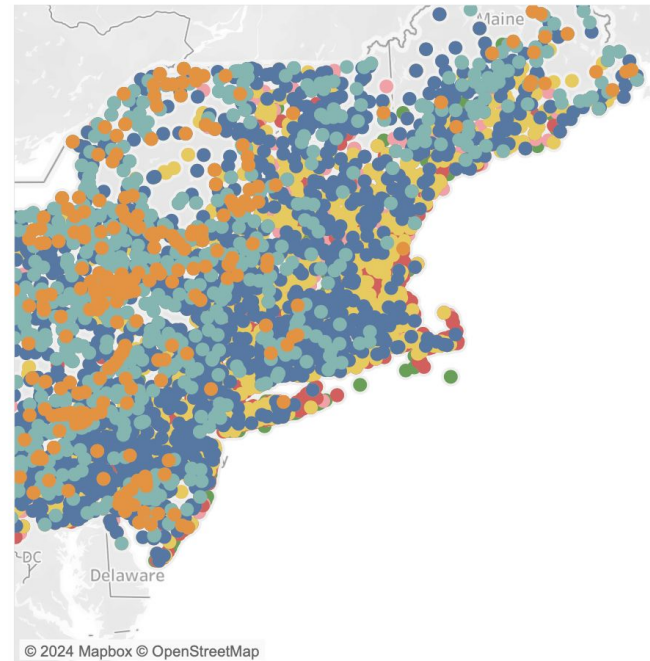
Limitations of the Case Study

-We only had data focused on the Northeast of the United States.
-The data only had a limited number of variables. Data on how long a property had been on the market and the year a property was built would have provided deeper insights.

Next Steps:

-Gather more information that includes samples from all of the states.
-Received updated information on housing prices.

For Sale US Real Estate Properties- Prices and Features as of 2021 w/ Northeast Region Focus



Price Category Key

- \$10,000 - \$49,999
- \$50,000 - \$99,999
- \$100,000 - \$249,999
- \$250,000 - \$399,999
- \$400,000 - \$649,999
- \$650,000 - \$799,999
- \$800,000+

Price Category

Multiple values

State (Select)

Multiple values

of Beds (Select)

To Null

of Baths (Select)

Non-Null Values Only

House Sq.Ft. (Select)

400 to 5,000

Thank You

Email: isom@body-archives.com

LinkedIn: <https://www.linkedin.com/in/isomwinton/>

GitHub: <https://github.com/isom17>

