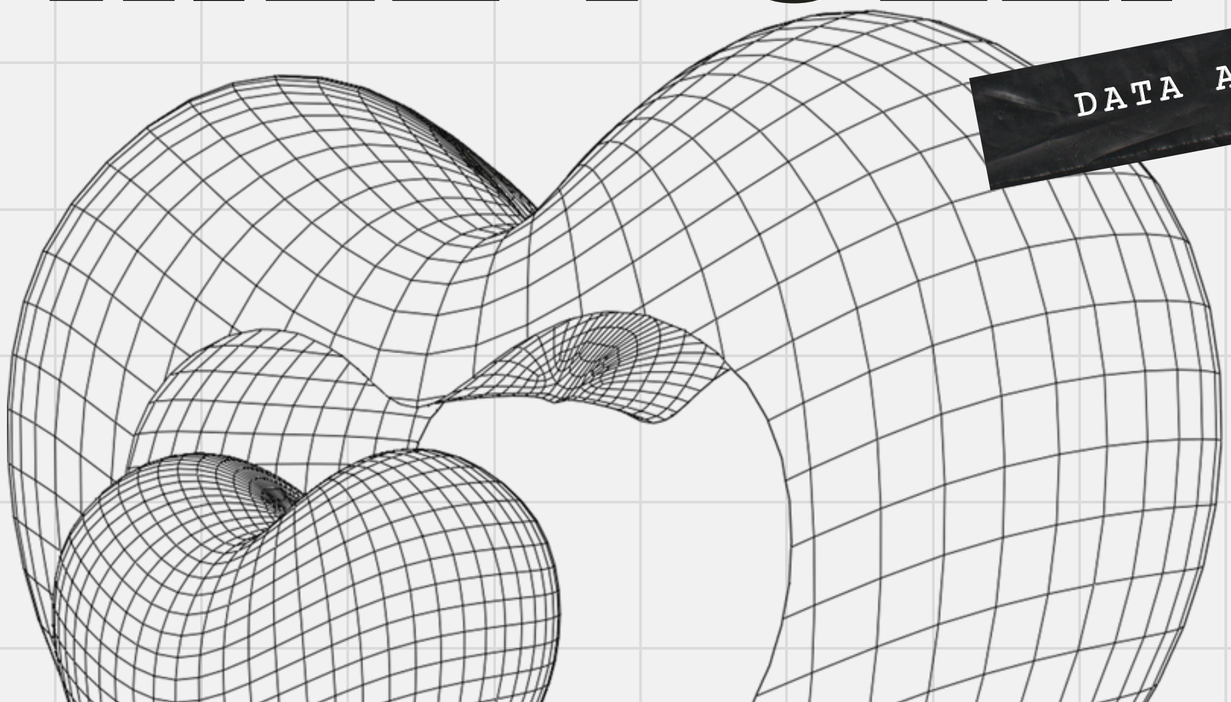
GABRIEL POLLICAR
@POLLICARG42 - GITHUB
POLLICARG42@GMAIL.COM

WINTER 2023

ANALYTICS PORTFOLIO

GABRIEL POLLICAR

DATA ANALYST

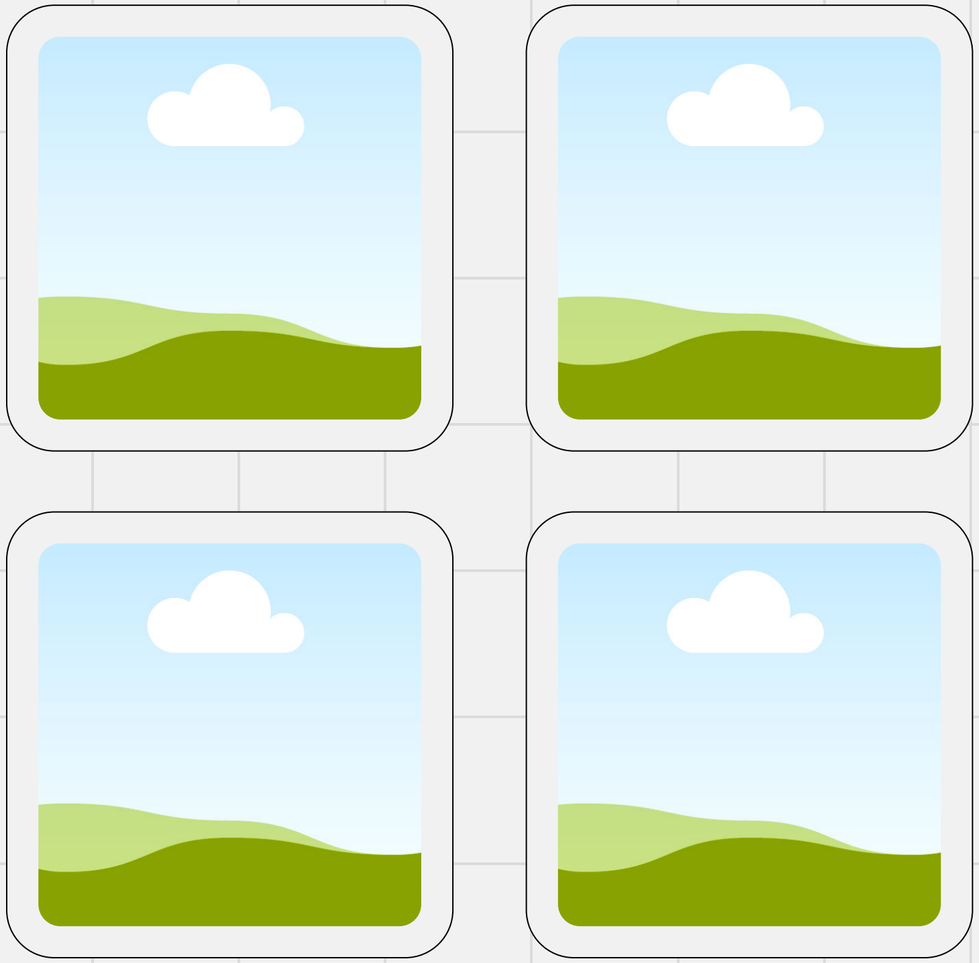


PORTFOLIO

0012 7482901 2744103 0592346 8774510 7255

TABLE OF CONTENTS

@POLLICARG42 - GITHUB



ABOUT ME

Hello, my name is Gabriel Pollicar.

I am a recent graduate of University of Maryland with a Bachelor's of Science Degree in Management Information Systems. I am a goal-oriented individual with a passion for finding solutions for companies using technology and business intelligence.

As a data analyst, I want to utilize my skills in SQL, Python, Excel and Tableau to discover trends and insights that help companies solve problems.

@POLLICARG42 - GITHUB

SKILLSET BREAKDOWN

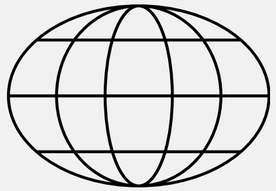
I have proficient skills and experience with these technologies and toolsets:

ANALYTICAL TOOLS	PYTHON	SQL	EXCEL	TABLEAU
JUPYTER	FUNCTIONS AND OPERATIONS	QUERY TOOLS	GRAPHING	CHARTS, GRAPHS
POSTGRES	PANDAS DATAFRAMES	JOINS	PIVOT TABLES	STATISTICAL VISUALIZATIONS
MS SUITE	DATA MINING	SUBQUERIES	DESCRIPTIVE ANALYSIS	SPATIAL AND TEXTUAL ANALYSIS
STATISTICAL ANALYSIS	MATPLOTLIB VISUALIZATIONS	CTE	VISUALIZATIONS	STORYTELLING WITH DASHBOARDS

@POLLICARG42 - GITHUB

PROJECT CASE STUDIES

These are the projects I have completed during Career Foundry's Data Analytics Course.

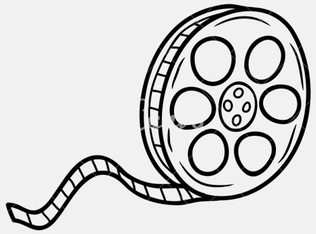


GAMECO SALES

Analyzed Global Video Game Sales using Excel Pivot Tables and Visualization tools.

INFLUENZA STAFFING

Analyzed National healthcare Data to gain insight on key variables driving healthcare staffing shortages across the U.S.



ROCKBUSTER STEALTH

Found business solutions for an online video rental company using SQL tools to query a large relational database.

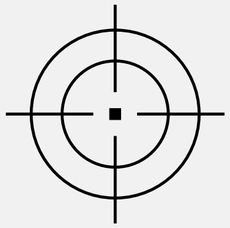
@POLLICARG42 - GITHUB

PROJECT CASE STUDIES

These are the projects I have completed during Career Foundry's Data Analytics Course.

INSTACART

Develop Marketing Strategies based on derived variables from large integrated datasets using Python



GLOBAL BANK MODELING

Created predictive models to identify money laundering by using time series analysis and forecasting on a global bank's customer data.

REAL ESTATE VALUES FORECASTING

Used Python libraries and machine learning tools to model and forecast real estate values using open sourced data.

INFLUENZA FACTORS CASE STUDY

Tools used in this project:

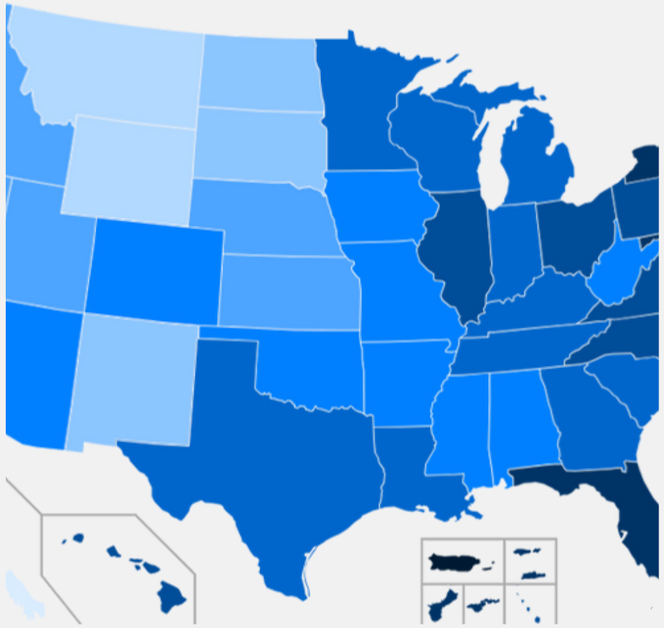
EXCEL

TABLEAU

PROJECT OBJECTIVE

Objective: This project aimed to assess staffing deployment for a healthcare staffing agency using two datasets sourced from the CDC and the US Census.

This project used Excel analysis and Tableau Visualization tools to discover seasonal patterns of influenza virus and develop a strategic plan to allocate medical staff across the United States.



US Census Dataset

I used the US Census Dataset collected by the US Census Bureau. These records measures populations based on residents living in certain geographic regions and include certain characteristics of the population.

CDC Influenza Deaths Dataset

This dataset was gathered by the Center for Disease Control (CDC). It contains information from death certificates that present influenza as the main underlying cause of death.

TOOLS USED IN THIS PROJECT:

EXCEL

TABLEAU

STEP 1

Source, clean, process and analyze excel data.

STEP 2

Develop a Factors Analysis Interim Report showcasing correlation between variables.

STEP 3

Integrate datasets and create statistical visualizations using Tableau tools.

STEP 4

Develop Tableau Storyboard using dashboard and present findings to Stakeholders.

PROCESS OVERVIEW

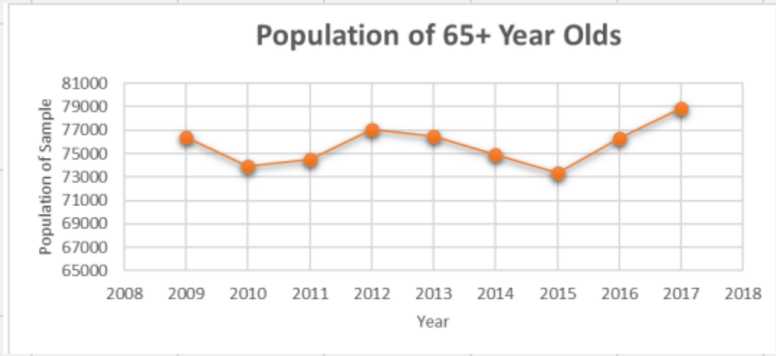
This project began with developing a project plan for the stakeholders of the medical staffing agency. It involved translating business requirements into actionable deliverables in a form of staffing recommendations.

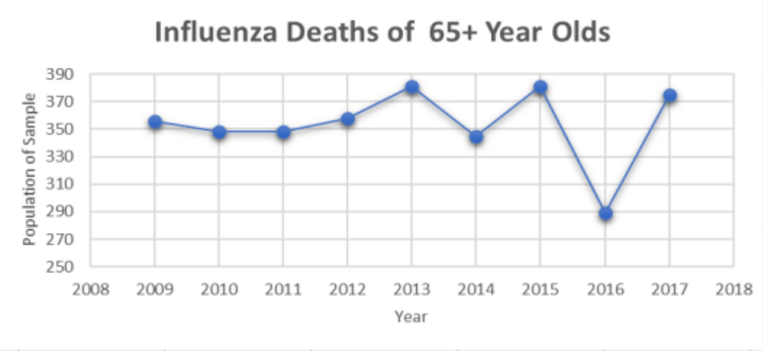
From there, raw data was sourced, cleaned and transformed into usable datasets. Variables were analyzed for correlations in a statistical hypothesis test. Correlations in the data were visualized using Tableau and used to develop a staffing plan. This plan was presented using Storytelling with Tableau Dashboards.

STATISTICAL HYPOTHESIS TESTING

An interim report was completed during halfway of the project. This report presented my collection methods and the data processing that I used to clean and transform the dataset. This document also presented my hypothesis of the specific factors to investigate for influenza. My hypothesis stated that that age was a potential variable for determining influenza rates among the population.

I used Excel pivot tables to integrate the two datasets sourced from the project brief. I then used statistical analysis to prove my hypothesis. I used T-Testing to prove correlation between age and flu rates. Then, by analyzing the P-values and its Significance level alpha variable, I was able to determine that Age groups and rates of influenza deaths were highly correlated with a 95% confidence level.





t-Test: Two-Samples Assuming Unequal Variances	*Age 65 Below*	*Age 65 Above*
Mean	7.93585E-06	0.000847792
Variance	9.97979E-11	1.8946E-07
Observations	459	459
Hypothesized Mean Difference	0	
Degree of Freedom	458	
t Stat	-41.32739773	
P(T<=t) one-tail	6.2532E-157	
t Critical one-tail	1.648187415	
P(T<=t) two-tail	1.2506E-156	
t Critical two-tail	1.965157098	

***Table 2**. t-Test on the 65+ years vs 65+ years old influenza deaths*

TABLEAU DASHBOARDS

Tableau Visualization tools were then used to provide stakeholders with a comprehensive understanding of the factors that caused staffing shortages.

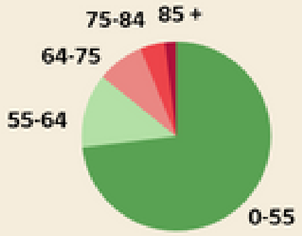
This dashboard was generated to show Age as a determining factor of Hospitalization rates for influenza.

Is Age a factor that affects Hospitalization Rates for Influenza?

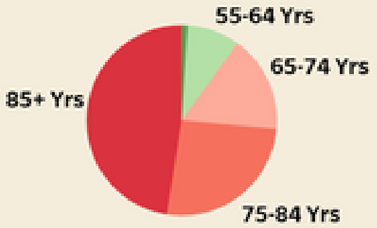
The Relationship Between Influenza Reported Deaths of Young/Adult (5-64) and Senior (64-85+) age groups:

Senior Citizens make up 8.5% of the population of the United States but encompass 91% of Influenza Hospitalizations.

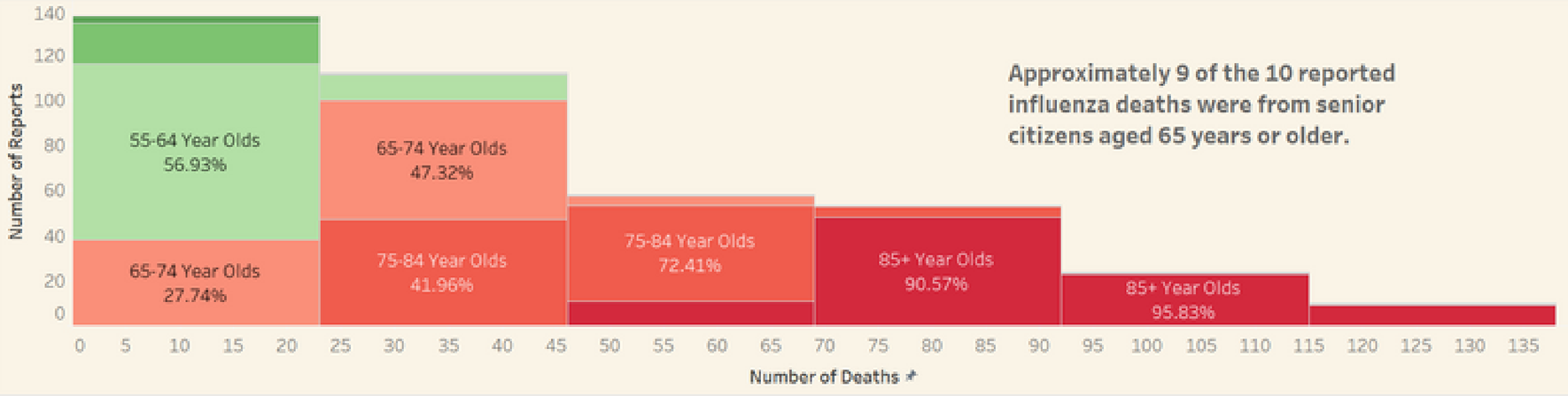
Of those hospitalizations, 85+ Year Olds make up more than half of those deaths despite being the smallest size age group.



Population by Age



Influenza Deaths by Age

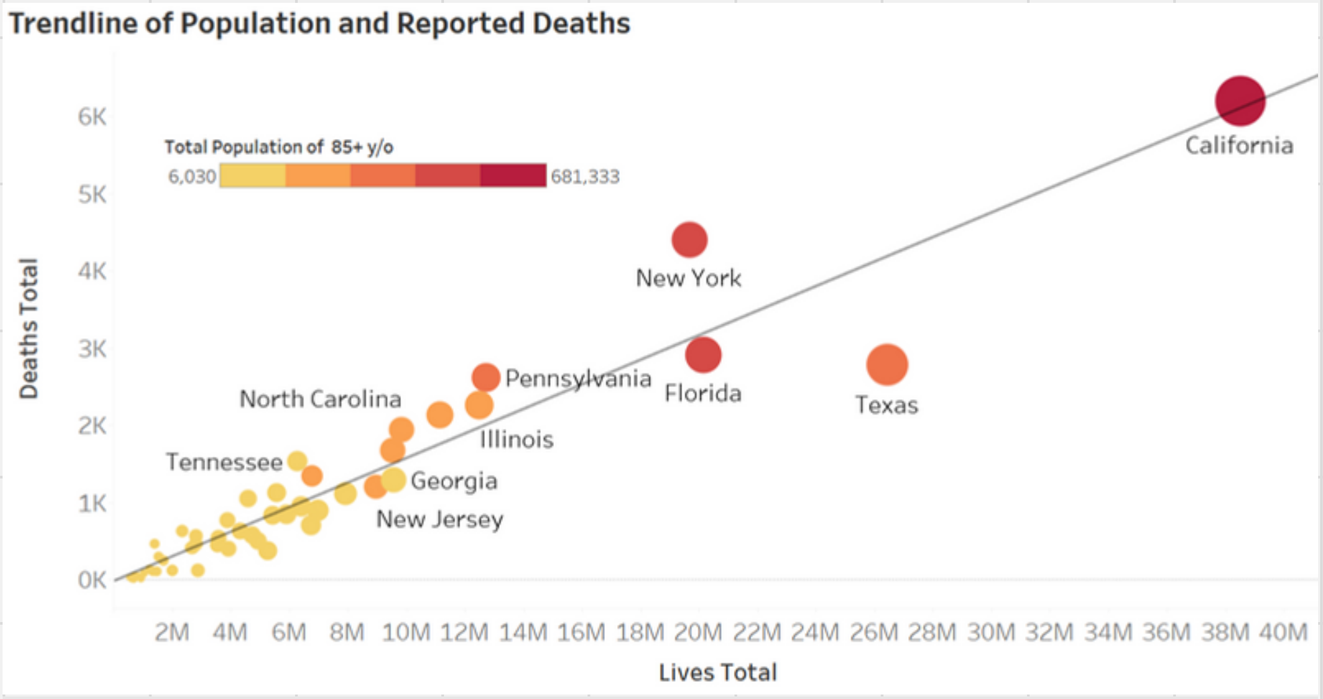


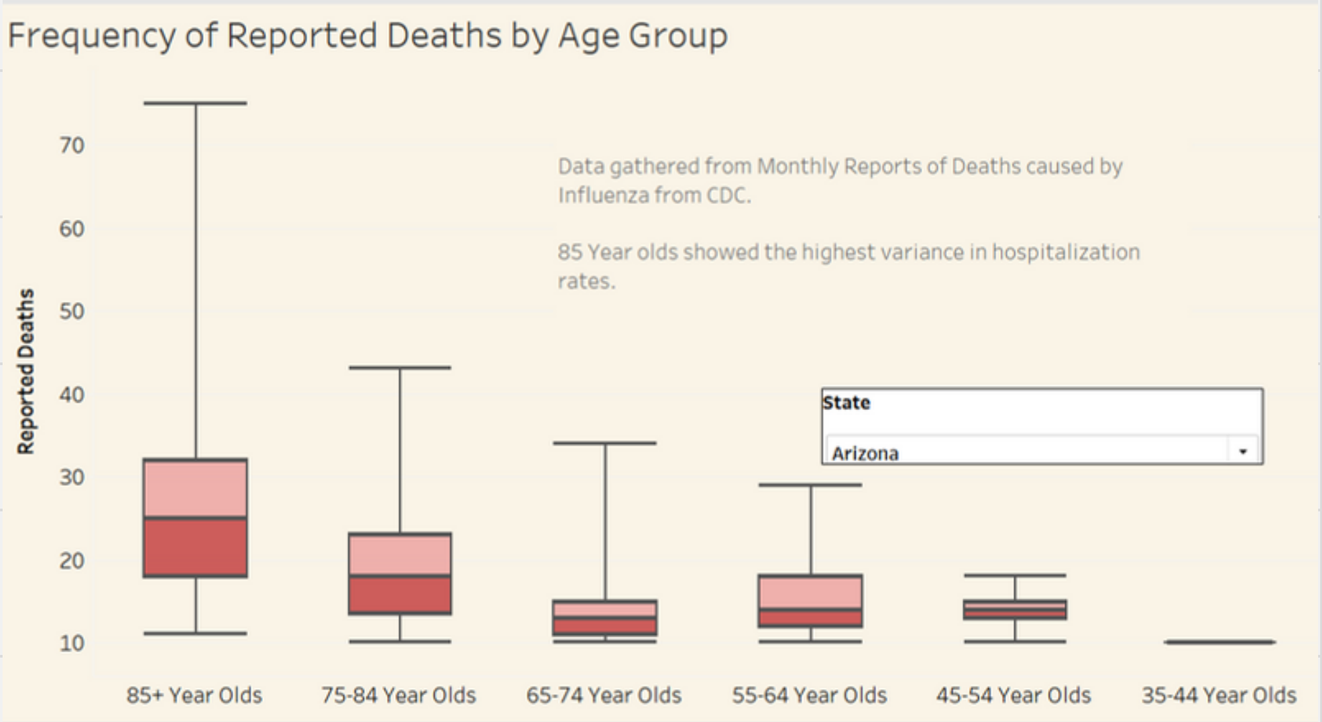
Approximately 9 of the 10 reported influenza deaths were from senior citizens aged 65 years or older.

INSIGHTS FROM VISUALIZATIONS IN TABLEAU

By using statistical tools and Tableau, I was able to show my analysis of the multiple factors that caused staffing shortages during influenza. For instance, the trendline above shows rates of hospitalizations per state according to population size of elderly individuals. By using color grades across a trendline, I was able to determine the states that were at risk.

This histogram below also shows variances in reported deaths by age group. This was used to help identify age groups that were potential causes of uncertainty in hospitalizations.

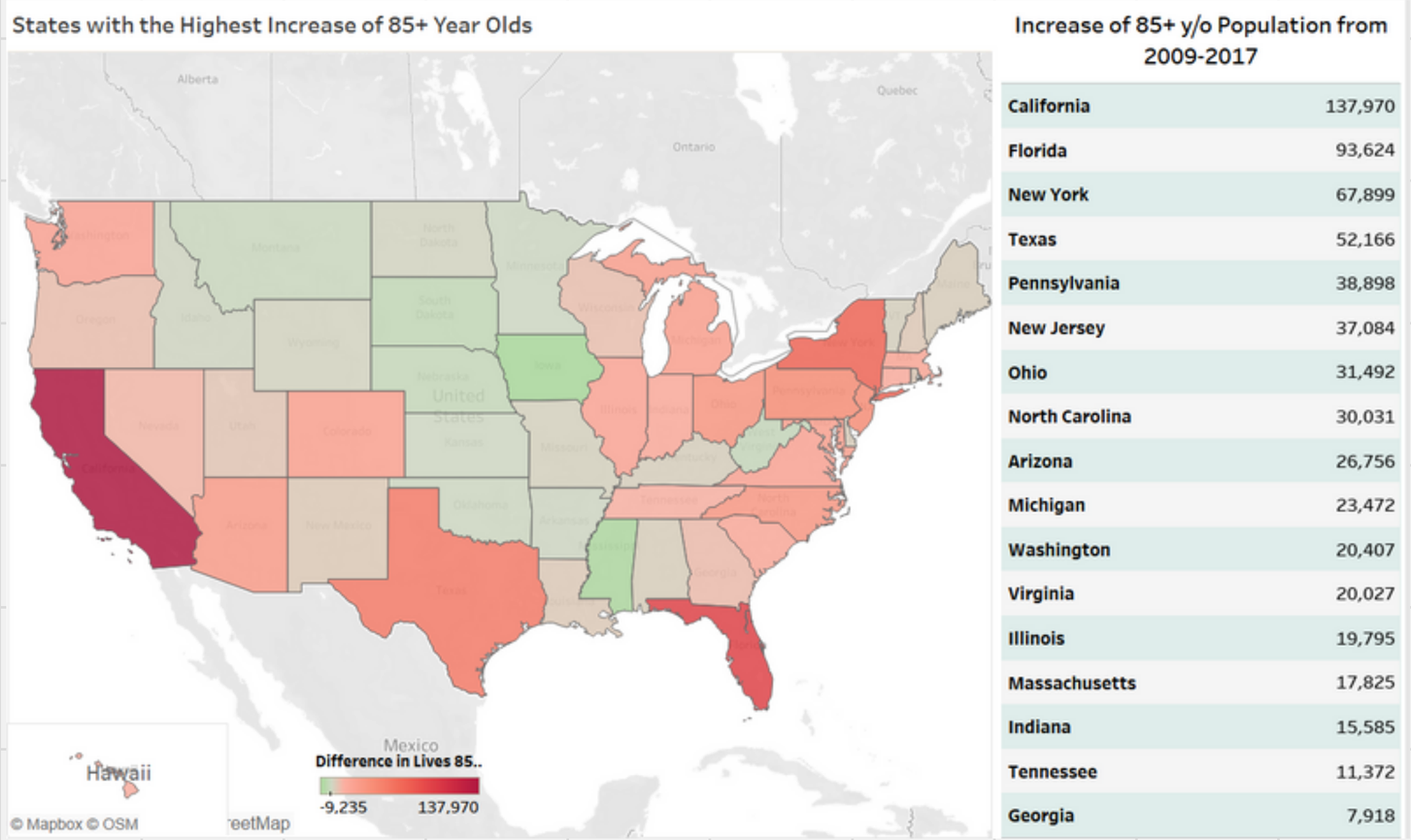




INSIGHTS FROM VISUALIZATIONS IN TABLEAU

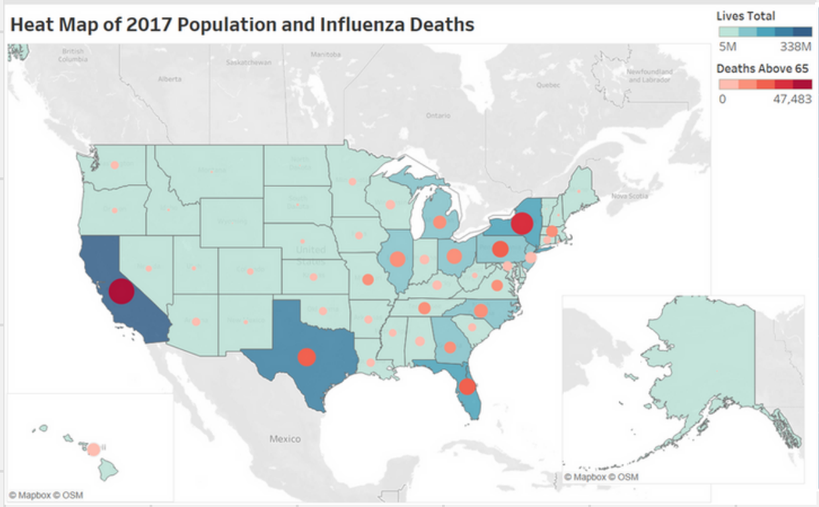
By using Spatial Analysis in Tableau, I was able to showcase different variables derived from my analysis. One variable was the rate of increasing elderly populations for each state. This gave stakeholders insight into which state was the most at risk for spikes of influenza hospitalizations.

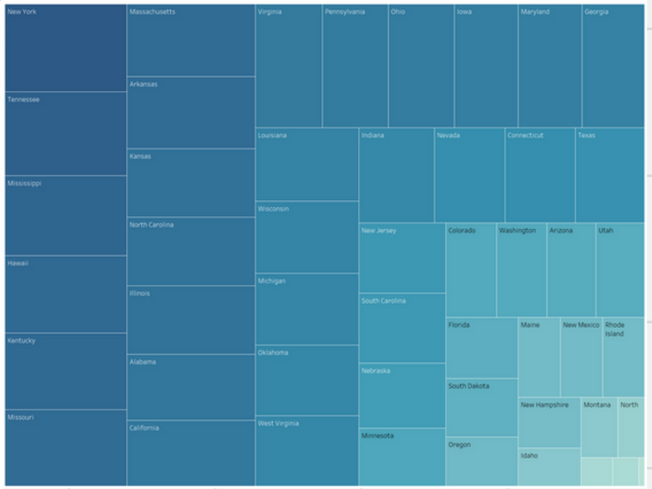
The visualizations shown here were made from Tableau. These spatial maps helped display the concentrations of increasing influenza rates across the U.S. These could help stakeholders identify regions where medical staffing is most needed.



Increase of 85+ y/o Population from 2009-2017

California	137,970
Florida	93,624
New York	67,899
Texas	52,166
Pennsylvania	38,898
New Jersey	37,084
Ohio	31,492
North Carolina	30,031
Arizona	26,756
Michigan	23,472
Washington	20,407
Virginia	20,027
Illinois	19,795
Massachusetts	17,825
Indiana	15,585
Tennessee	11,372
Georgia	7,918





FINAL PROJECT DELIVERABLES

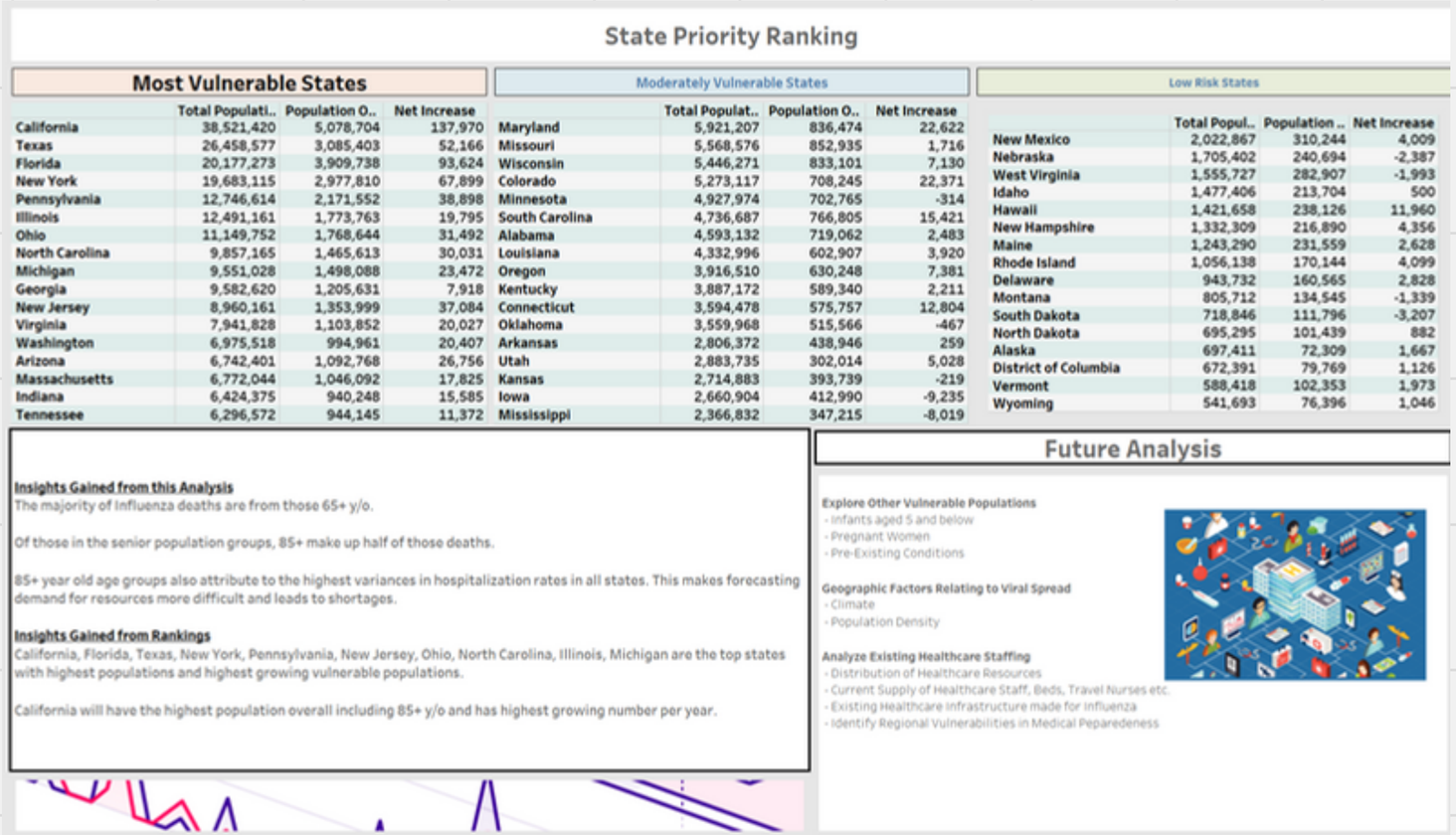
This project provided stakeholders with an Interim Report detailing the data processing and statistical analysis done on the causes and variables of influenza staff shortages. A tableau storyboard was then created to showcase the results and recommendations of the analysis. In addition, I also provided a presentation video of myself presenting my findings and results.

This is the final recommendations page of the Tableau storyboard presentation. It provides the state rankings required by the project along with the main insights and recommendations for future analysis.

INTERIM REPORT: LINK

STORYBOARD VIDEO PRESENTATION:LINK

TABLEAU STORYBOARD: LINK



State Priority Ranking

Most Vulnerable States

	Total Populati..	Population O..	Net Increase
California	38,521,420	5,078,704	137,970
Texas	26,458,577	3,085,403	52,166
Florida	20,177,273	3,909,738	93,624
New York	19,683,115	2,977,810	67,899
Pennsylvania	12,746,614	2,171,552	38,898
Illinois	12,491,161	1,773,763	19,795
Ohio	11,149,752	1,768,644	31,492
North Carolina	9,857,165	1,465,613	30,031
Michigan	9,551,028	1,498,088	23,472
Georgia	9,582,620	1,205,631	7,918
New Jersey	8,960,161	1,353,999	37,084
Virginia	7,941,828	1,103,852	20,027
Washington	6,975,518	994,961	20,407
Arizona	6,742,401	1,092,768	26,756
Massachusetts	6,772,044	1,046,092	17,825
Indiana	6,424,375	940,248	15,585
Tennessee	6,296,572	944,145	11,372

Moderately Vulnerable States

	Total Populat..	Population O..	Net Increase
Maryland	5,921,207	836,474	22,622
Missouri	5,568,576	852,935	1,716
Wisconsin	5,446,271	833,101	7,130
Colorado	5,273,117	708,245	22,371
Minnesota	4,927,974	702,765	-314
South Carolina	4,736,687	766,805	15,421
Alabama	4,593,132	719,062	2,483
Louisiana	4,332,996	602,907	3,920
Oregon	3,916,510	630,248	7,381
Kentucky	3,887,172	589,340	2,211
Connecticut	3,594,478	575,757	12,804
Oklahoma	3,559,968	515,566	-467
Arkansas	2,806,372	438,946	259
Utah	2,883,735	302,014	5,028
Kansas	2,714,883	393,739	-219
Iowa	2,660,904	412,990	-9,235
Mississippi	2,366,832	347,215	-8,019

Low Risk States

	Total Popul..	Population ..	Net Increase
New Mexico	2,022,867	310,244	4,009
Nebraska	1,705,402	240,694	-2,387
West Virginia	1,555,727	282,907	-1,993
Idaho	1,477,406	213,704	500
Hawaii	1,421,658	238,126	11,960
New Hampshire	1,332,309	216,890	4,356
Maine	1,243,290	231,559	2,628
Rhode Island	1,056,138	170,144	4,099
Delaware	943,732	160,565	2,828
Montana	805,712	134,545	-1,339
South Dakota	718,846	111,796	-3,207
North Dakota	695,295	101,439	882
Alaska	697,411	72,309	1,667
District of Columbia	672,391	79,769	1,126
Vermont	588,418	102,353	1,973
Wyoming	541,693	76,396	1,046

Insights Gained from this Analysis

The majority of Influenza deaths are from those 65+ y/o.

Of those in the senior population groups, 85+ make up half of those deaths.

85+ year old age groups also attribute to the highest variances in hospitalization rates in all states. This makes forecasting demand for resources more difficult and leads to shortages.

Insights Gained from Rankings

California, Florida, Texas, New York, Pennsylvania, New Jersey, Ohio, North Carolina, Illinois, Michigan are the top states with highest populations and highest growing vulnerable populations.

California will have the highest population overall including 85+ y/o and has highest growing number per year.

Future Analysis

Explore Other Vulnerable Populations
- Infants aged 5 and below
- Pregnant Women
- Pre-Existing Conditions

Geographic Factors Relating to Viral Spread
- Climate
- Population Density

Analyze Existing Healthcare Staffing
- Distribution of Healthcare Resources
- Current Supply of Healthcare Staff, Beds, Travel Nurses etc.
- Existing Healthcare Infrastructure made for Influenza
- Identify Regional Vulnerabilities in Medical Peparedeness

ROCKBUSTER PROJECT CASE STUDY

Tools used in this project:

EXCEL

POSTGRES

SQL

TABLEAU

PROJECT OBJECTIVES

The Rockbuster project consisted of a large relational database with sales and shipment data from Rockbuster Stealth Company. This company wanted to expand their business towards streaming, and required key questions to be answered by a data analyst regarding their sources of revenue.

Mainly, this project aims to answer this question using SQL to query information from tables organized in a relational database, and then presenting the findings in a presentation with Tableau visuals.

PROJECT BRIEF: LINK

QUESTIONS ASKED IN THE ANALYSIS:

- WHAT ARE THE AVERAGE, MINIMUM AND MAXIMUM RENTAL DURATIONS?
- AT WHAT RATE ARE MOST MOVIES CHARGING PER RENTAL?
- WHAT ARE THE AVERAGE MOVIE LENGTHS?
- WHAT ARE THE TOP 5 COUNTRIES AND CITIES THAT HOLD THE MOST CUSTOMERS?
- WHO ARE OUR TOP 5 MOST PAYING LOYAL CUSTOMERS?
- WHAT CAN WE FIND OUT ABOUT THEM?

TOOLS USED IN THIS PROJECT:

EXCEL

POSTGRES

SQL

TABLEAU

STEP 1

Find the Top Revenue Countries

STEP 2

Within Those Countries, Find the Highest Revenue Cities.

STEP 3

Find the Top Paying Customers in those High Revenue Cities.

STEP 4

Analyze their Purchases and Develop Insights for Sales and Marketing.

ANALYSIS PROCEDURE

Rentals, Rates and Films

- Understand and summarize trends in rental rates and film attributes
- Generate insights on average rental rates, duration of rentals and average fees.

Customer Locations

- Analyze global consumer trends
- Visualize distribution of customers across the globe
- Target Top 10 countries with most customers and revenue

USING SQL QUERY

For this project, I used Postgres and SQL to analyze the company's relational database.

This example shows some of the SQL queries I used to find certain information from the database. Because of the nature of relational databases, I used multiple nested subqueries, join statements and CTEs to get the necessary data to answer questions from Rockbuster's stakeholders. The result of this query shows the top 5 customers based on total spent with Rockbuster.

```
SELECT A.customer_id,
                A.first_name,
                A.last_name,
                d.country,
                c.city,
                SUM(amount)
FROM customer A
INNER JOIN payment E ON A.customer_id = E.customer_id
INNER JOIN address B ON A.address_id = B.address_id
INNER JOIN city C ON B.city_id = C.city_id
INNER JOIN country D ON C.country_id = D.country_id
WHERE c.city IN(SELECT c.city
        FROM customer A
        INNER JOIN address B ON A.address_id = B.address_id
        INNER JOIN city C ON B.city_id = C.city_id
        INNER JOIN country D ON C.country_id = D.country_id
        WHERE country IN (SELECT D.country
                FROM customer A
                INNER JOIN address B ON A.address_id = B.address_id
                INNER JOIN city C ON B.city_id = C.city_id
                INNER JOIN country D ON C.country_id = D.country_id
                GROUP By d.country
                ORDER BY COUNT(customer_id) DESC
                LIMIT 10)
        GROUP By d.country, c.city
        ORDER BY COUNT(A.customer_id) DESC
        LIMIT 10)
GROUP By A.customer_id, A.first_name, A.last_name, d.country, c.city
ORDER BY SUM(amount) DESC
LIMIT 5;
```

Example SQL Code Used in Analysis

Customer_ID	First_Name	Last_Name	Country	City	Total Spent
225	Arlene	Harvey	India	Ambattur	$ 111.76
424	Kyle	Spurlock	China	Shanwei	$ 109.71
240	Marlene	Welch	Japan	Iwaki	$ 106.77
486	Glen	Talbert	Mexico	Acua	$ 100.77
537	Clinton	Buford	United States	Aurora	$ 98.76

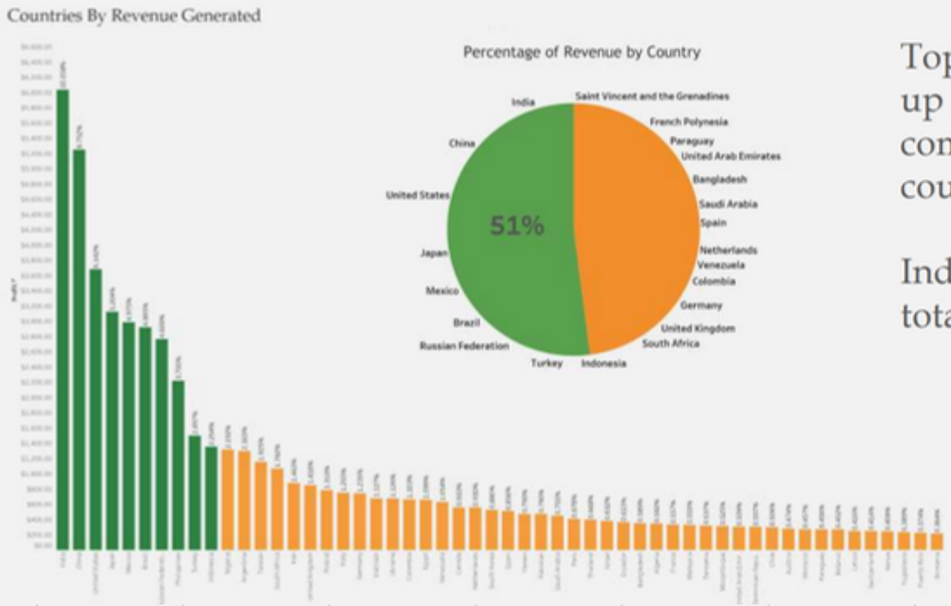
Result of SQL Code Output

INSIGHTS FROM SQL QUERY

By using SQL query, I was able to generate listings of top revenue countries and cities, and highest paying customers. Furthermore, I was also able to generate new datasets from merging and joining existing data. This was used to gain new insight on revenues generated.

The example below shows countries by generated revenue and I was able to discover that the top 10 countries generated more than half of all of Rockbuster's revenue compared to the 108 total countries.

Country	Customers	Total Revenue
India	60	$6034.78
China	53	$5251.03
United States	36	$3685.31
Japan	31	$3122.51
Mexico	30	$2984.82
Brazil	28	$2919.19
Russian Federation	28	$2765.62
Philippines	20	$2219.7
Turkey	15	$1498.49
Indonesia	14	$1352.69



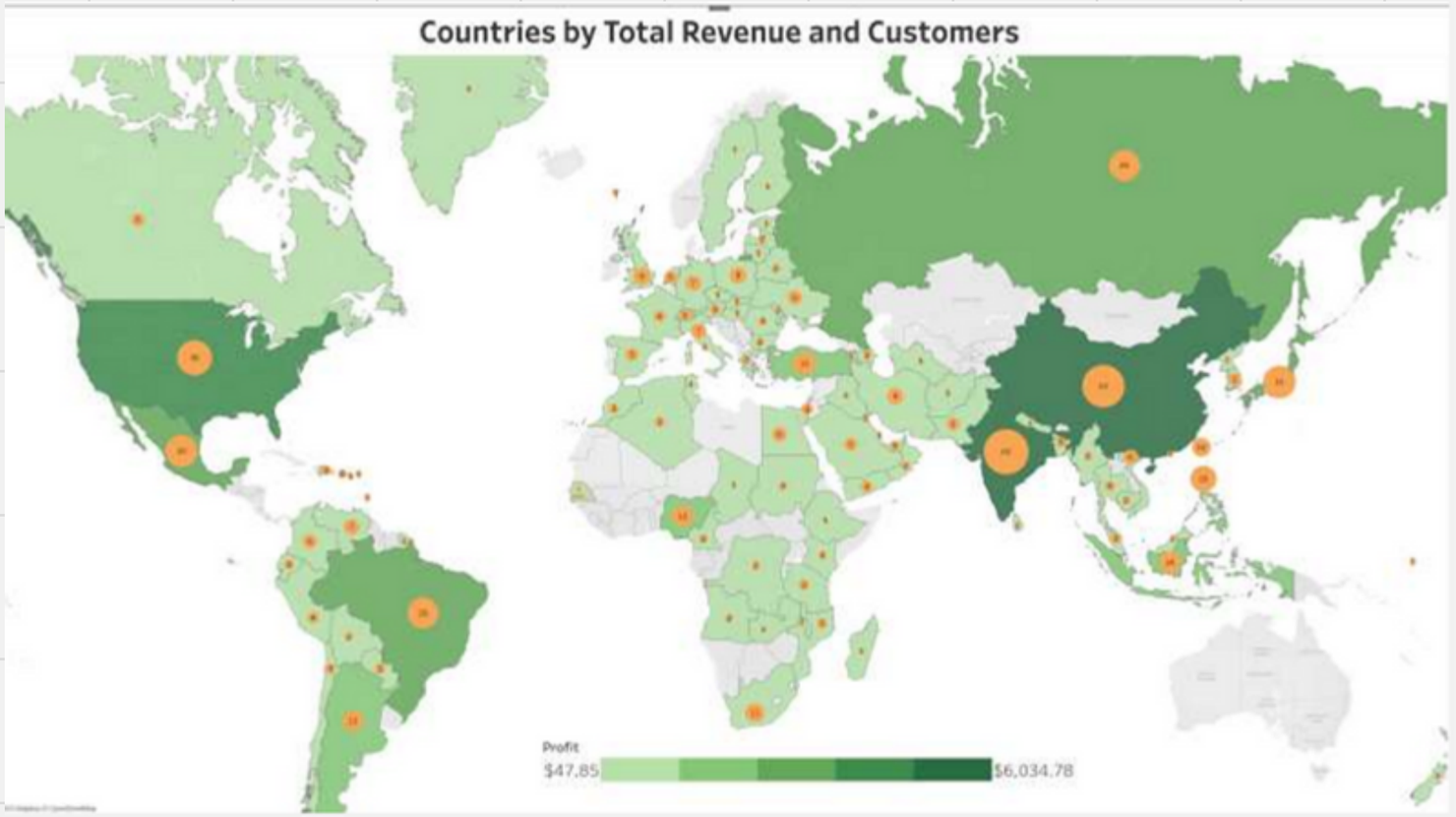
Top 10 Countries make up 51% of total revenue combined of 108 countries.

India makes up 10% of all total revenue gained.

MORE INSIGHTS AND DISCOVERIES

Another deliverable for this project was to provide charts from new datasets that were produced from SQL Queries. These new datasets were created by merging multiple tables across Rockbuster's relational database.

I also used Tableau to generate spatial visualizations to showcase these new datasets. This example demonstrates the distributions of revenue to the stakeholders. In the example, I used map visuals to show concentrations of Rockbuster's highest revenue countries.



ustomer ID	First Name	Country	City	Total Payments
5	Arlene Harvey	India	Ambattur	$111
4	Kyle Spurlock	China	Shanwei	$109
0	Marlene Welch	Japan	Iwaki	$106
6	Glen Talbert	Mexico	Acua	$100
7	Clinton Buford	United States	Aurora	$98

Customer ID	Name	Most Watched Genre	Most Watched Rating
225	Arlene Harvey	Drama	NC-17
424	Kyle Spurlock	New	PG-13
240	Marlene Welch	Drama	PG-13
486	Glen Talbert	Children	R
537	Clinton Buford	Family	PG-13

FINAL PROJECT DELIVERABLES

The deliverables for this project includes a PowerPoint presentation of the visualizations, results and insights that were gained from the analysis. This final slide from the presentation shows my recommendations to stakeholders, as well as proposals for future analysis.

In conclusion, I was able to answer questions from stakeholders by utilizing my skills in SQL Query and Tableau to extract information from RockBuster's large relational database. I was able to find key insights to RockBuster's customer profiles and regions.

POWERPOINT PRESENTATION: LINK

Recommendations

- Procure movies that are most interesting to high paying customers: PG-13 movies and Family, Children, and Drama genres
- Focus marketing on consumers in Top 10 Countries:

India, China, United States, Japan, Mexico,

Brazil, Russia, Philippines, Turkey, Indonesia

Further Analysis

- Analyze most popular films (Genre and Rating) for all consumers in Top 10 Countries (Especially India, China and United States)
- Determine factors contributing to revenue per city: population, age, latitude etc.
- Focus on high paying customers and generate more insights from other factors (rental rates, rental duration, store locations etc.)

INSTACART PROJECT

Completed 2023

Tools used in this project:

EXCEL

PYTHON

JUPYTER NOTEBOOK

PROJECT OBJECTIVES

The Instacart project consisted of multiple large datasets with sales, ordering and customer account data from Instacart, a grocery delivery company. This company wanted to uncover more information about their sales patterns, and required key questions to be answered by a data analyst.

Mainly, this project's aim was to answer these question using SQL to query information from tables organized in a relational database, and then presenting the findings in a presentation with Tableau visuals.

PROJECT BRIEF: LINK

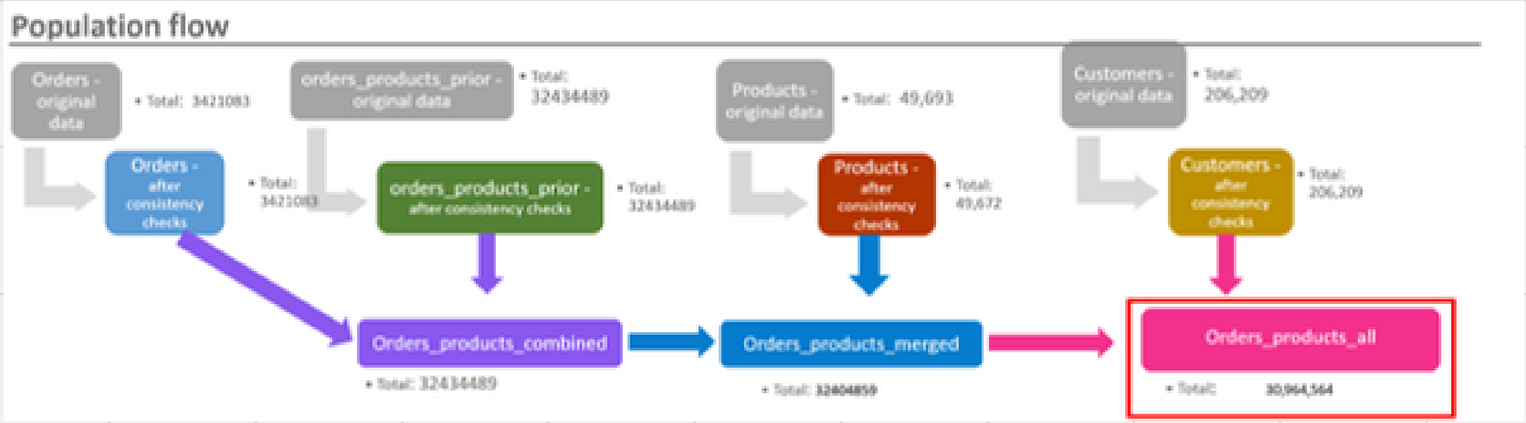
7 KEY QUESTIONS ASKED BY MARKETING TEAM:

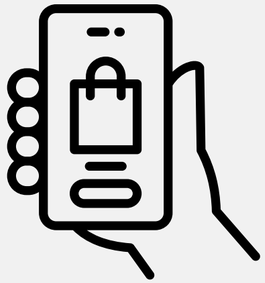
- THE SALES TEAM NEEDS TO KNOW THE **BUSIEST DAYS OF THE WEEK AND HOURS OF THE DAY** ARE TO SCHEDULE ADS AT TIMES WHEN THERE ARE FEWER ORDERS.
- ARE THERE PARTICULAR **TIMES OF THE DAY WHEN PEOPLE SPEND THE MOST MONEY**? AS THIS MIGHT INFORM THE TYPE OF PRODUCTS WE ADVERTISE AT THESE TIMES.
- INSTACART HAS A LOT OF **PRODUCTS WITH DIFFERENT PRICE TAGS**. SALES WANTS TO USE **SIMPLER PRICE RANGE GROUPINGS** TO HELP DIRECT THEIR EFFORTS
- ARE THERE CERTAIN TYPES OF **PRODUCTS THAT ARE MORE POPULAR THAN OTHERS**? THE MARKETING TEAM WANTS TO KNOW WHICH DEPARTMENTS HAVE THE **HIGHEST FREQUENCY OF PRODUCT ORDERS**.
- ARE THERE **DIFFERENCES IN ORDERING HABITS** BASED ON A CUSTOMER'S **LOYALTY STATUS**?
- ARE THERE DIFFERENCES IN **ORDERING HABITS** BASED ON A **CUSTOMER'S REGION**?
- WHAT DIFFERENCES CAN YOU FIND IN THE **ORDERING HABITS** OF DIFFERENT **CUSTOMER PROFILES**?

DATA CLEANING AND TRANSFORMATION

This project used large datasets containing more than 30 million entries of product orders. These were cleaned using Pandas Dataframe functions. Here I used Pandas data wrangling tools to drop missing values, correct erroneous inputs or exclude data requested by stakeholders. I documented my data cleaning process using a Population Flow chart to communicate my data changes.

The next step was to merge all the datasets together into one large table. This would allow me to generate graphs and visualizations used to support my findings.



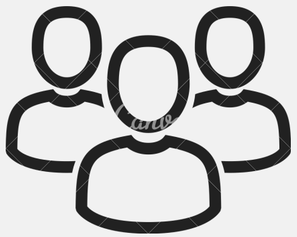


Merged Orders Dataset

This was the largest dataset which contained 30 million rows of customer orders, This dataset was used to merge all other datasets into. The final table provided ordering data with specific product information and customer information to be used for analysis.

Products Dataset

This dataset provided product information on available to purchase in the Instacart App. This table included product id numbers, names, and prices of products.



Customer Info Dataset

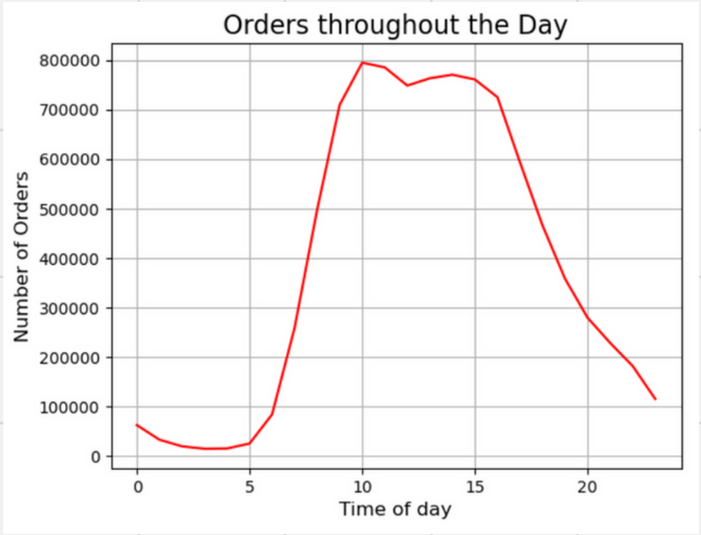
This dataset included customer information such as account number, names, addresses, and account age. Information such as names and full addresses were removed to maintain PII privacy.

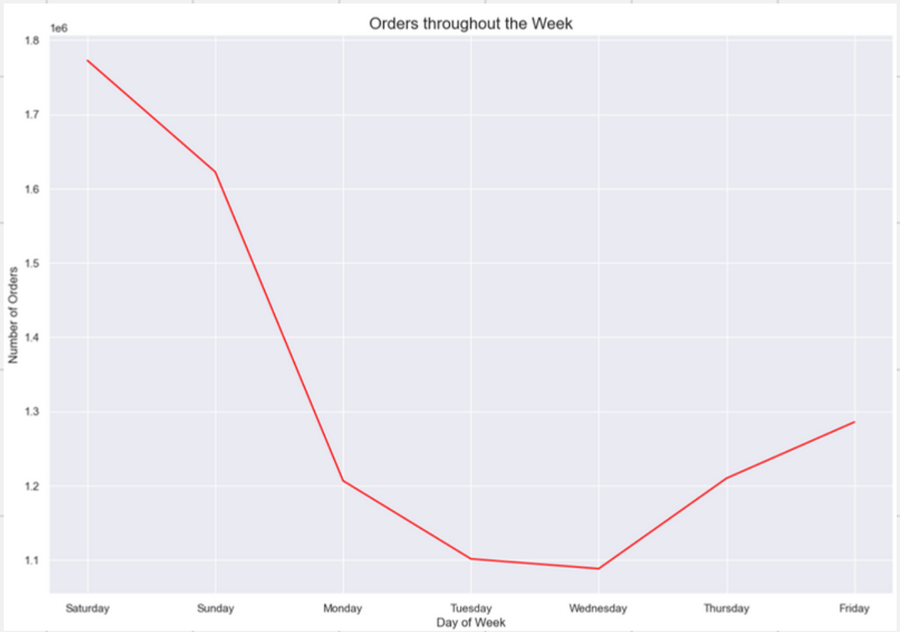
Datasets used for the Project

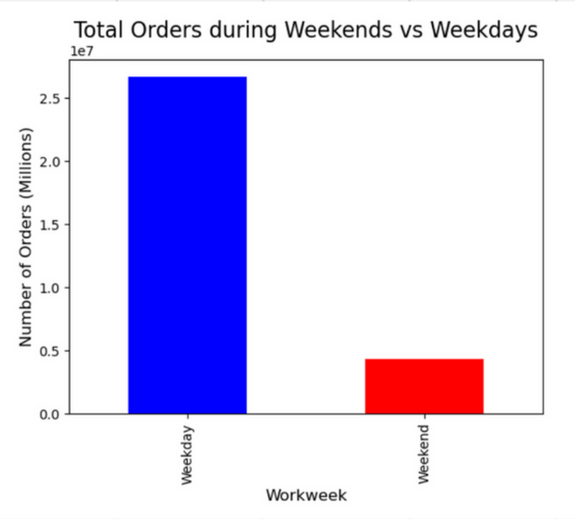
INSIGHTS ON ORDERING HABITS

These visualizations were made using pandas dataframes and matplotlib visualization functions. They were developed by comparing time measurements in the integrated dataset according to order volume. Variables that were measured include purchasing volume throughout different hours of the day and specific days of the week.

These graphs were used to develop findings of customer ordering habits. For example, order volume according to each day of the week shows that Wednesdays were the days that brought the least revenue. Insights like these were used to answer questions for the sales and marketing departments.





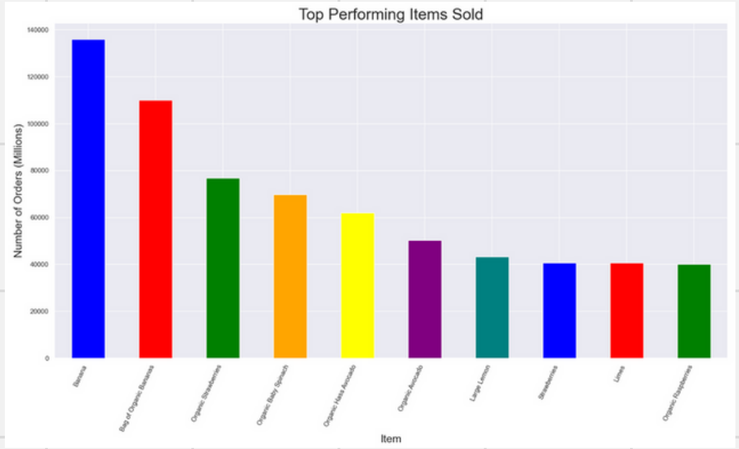


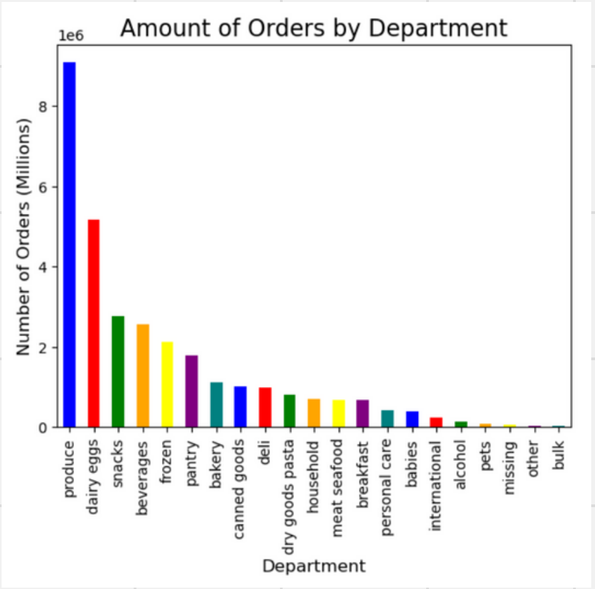
INSIGHTS ON REGIONS AND PRICE GROUPS

In this project, I also analyzed different variables together to discover new trends and patterns in customer purchasing behavior. I used Pandas functions, such as crosstabs, to find commonalities of top revenue items, regions, and income groups.

For example, a crosstab of customers flagged by their income showed that only 2% of customers are high spenders which was consistent across all regions.

Insights like these could help the sales and marketing team develop better marketing strategies with data driven decisions.





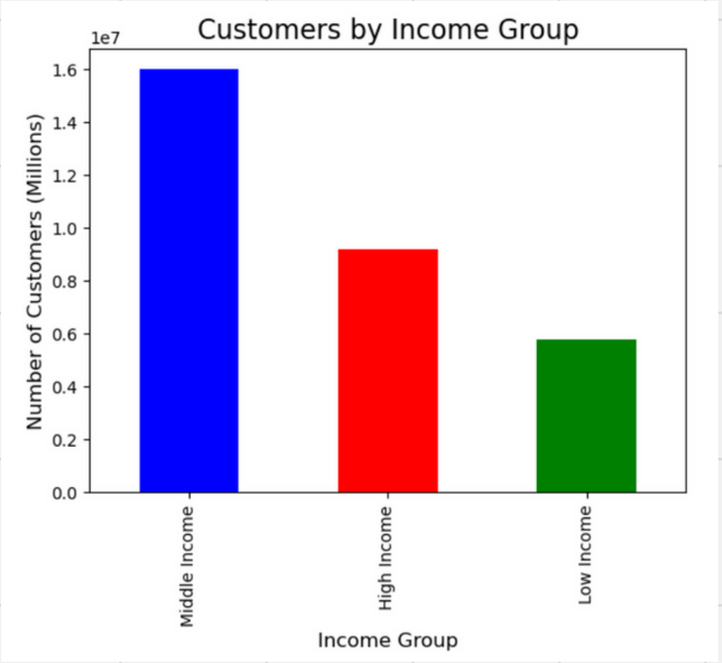
Spending_Flag	High_spender	Low_spender
Region		
Midwest	155975	7441350
Northeast	108225	5614511
South	209691	10582194
West	160354	8132559

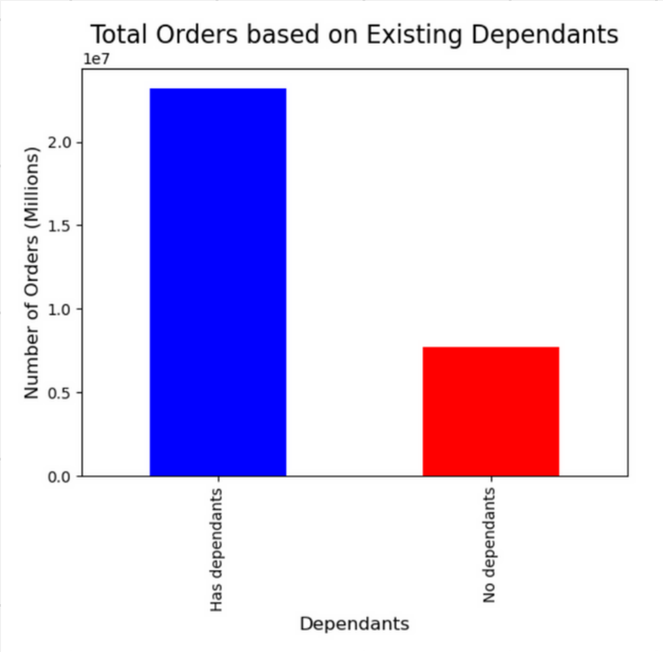
FINDINGS FROM CUSTOMER PROFILING

In this project, I also worked on profiling customers by specific variables to gain an understanding of different purchasing behaviors. Customer information such as income, number of dependents, and Instacart loyalty status were examined.

These customer profiles were analyzed with other attributes to discover trends based on different customers. For example, middle class customers with average incomes brought in the most revenue and were the majority of Instacart Customers.

Customers not subscribed to Instacart's loyalty program also brought in the most revenue compared to other groups.

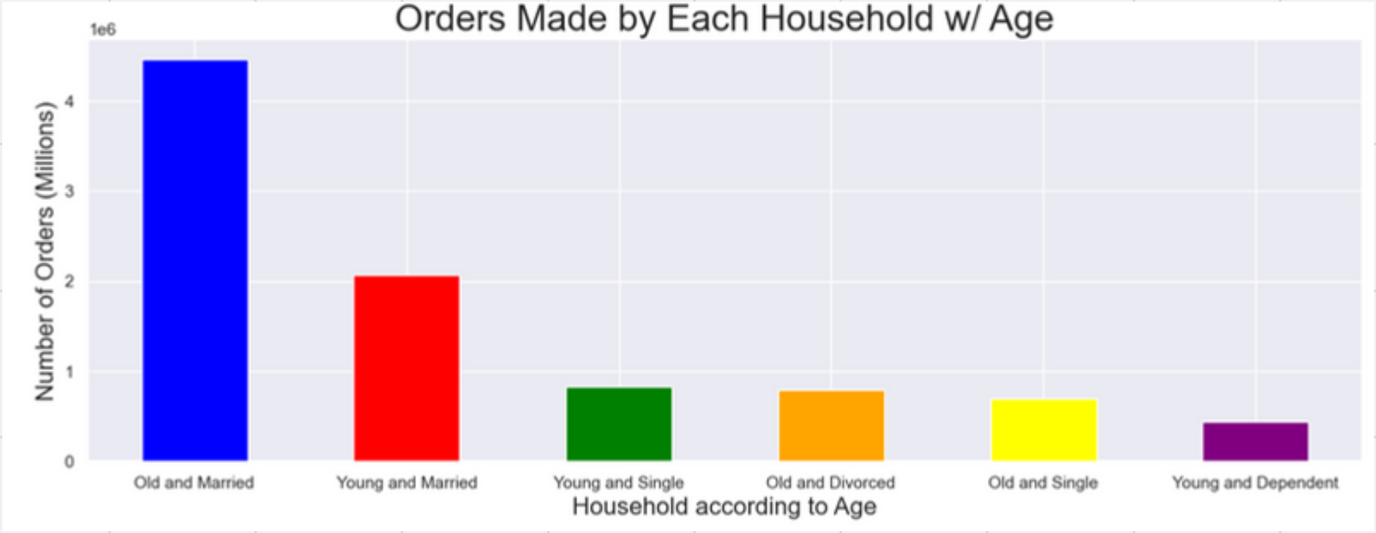


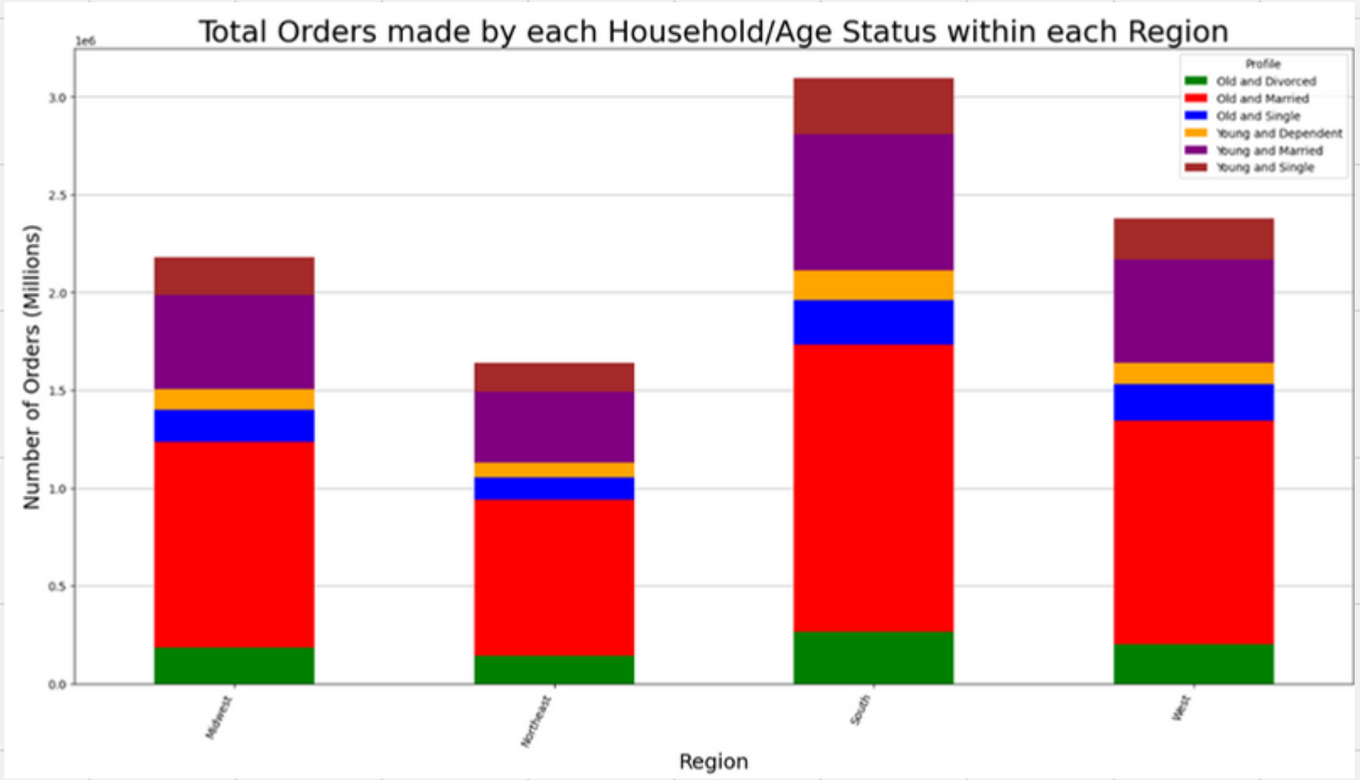


FINDINGS FROM HOUSEHOLD PROFILING

These visualizations were made by analyzing household as a factor for different variables. Customers were grouped into household according to their family status. Then each family status was split into age groups. This was used to find insight on different purchasing behaviors of older vs younger groups.

For example, married households were found to have the highest volume of purchases, but it was also found that adults older than 40 who were married bring far more revenue than younger married couples. This was shown in the graph above.





EXCEL PRESENTATION:LINK

RECOMMENDATIONS AND CONCLUSIONS

The deliverables for this project involved creating visualizations and a comprehensive Excel presentation report that would be sent to the sales and marketing team to communicate my findings. The excel report described my analysis methodology, results, and recommendations for Instacart Shareholders.

This shows a part of the final page of recommendations for Instacart Shareholders based on the Key questions that were stated in the project objectives. Overall, this project showed my ability to gather business applicable insight from big datasets using Python programming and visualization.

RECOMMENDATIONS TO MARKETING TEAM

KEY QUESTION 1

- PEAK ORDER TIMES: 10AM-3PM; LOWEST ORDERS AROUND 4AM.
- SATURDAYS BUSIEST, TUESDAYS AND WEDNESDAYS SLOWEST.
- MARKETING FOCUS: 10AM-3PM, WEEKENDS; BOOST SALES ON TUESDAYS AND WEDNESDAYS.

KEY QUESTION 2

- PEAK REVENUE TIMES: 10AM AND 3PM; CORRELATE WITH PEAK ORDER TIMES.
- IMPLEMENT MARKETING DURING 10AM-3PM FOR ACTIVE INSTACART USERS.
- AVOID MARKETING/SALES DURING LOW USAGE HOURS (MIDNIGHT-6AM).

KEY QUESTION 3

- MAJORITY OF ORDERS FOR LOWER-PRICED ITEMS; $1-$5 AND $10-$15 RANGES POPULAR.
- LOWEST PURCHASES FOR ITEMS OVER $15; FOCUS ON $1-$15 PRICE RANGE.
- ADJUST MARKETING STRATEGY TO HIGHLIGHT LOW TO MODERATE-PRICED ITEMS.

KEY QUESTION 4

- BANANAS MOST POPULAR, FOLLOWED BY ORGANIC BANANAS; TOP 10 ITEMS ARE MAINLY PRODUCE
- MARKET PRODUCE ITEMS, ESPECIALLY ORGANIC VERSIONS LIKE BANANAS, STRAWBERRIES, AND LEMONS. ALIGN MARKETING STRATEGY WITH TRENDS OF TOP 10 SOLD PRODUCTS FOR BETTER OUTREACH.

KEY QUESTION 5

- REGULAR CUSTOMERS MAKE THE MOST ORDERS, ALIGNING WITH THEIR MAJORITY IN THE DATASET.
- EACH LOYALTY GROUP FOLLOWS THE OVERALL ORDER TREND, WITH REGULAR CUSTOMERS LEADING.
- INSTACART CAN BOOST REVENUE BY OFFERING LOYALTY PROGRAMS, ESPECIALLY TARGETING REGULAR CUSTOMERS.

GAMECO DATA ANALYSIS CASE STUDY

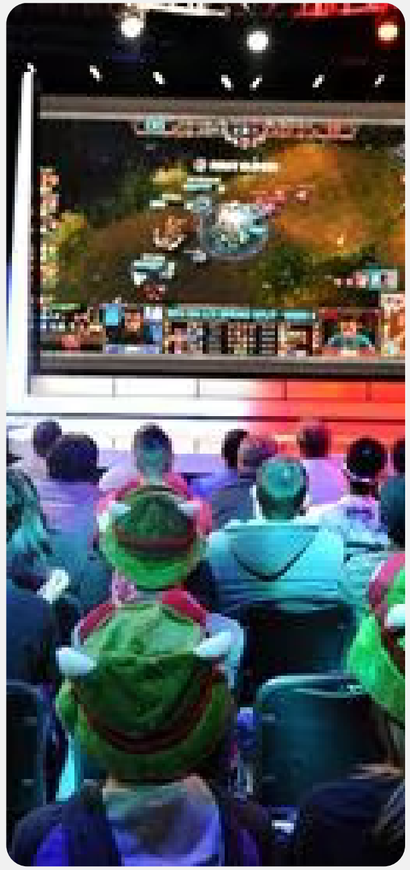
Completed 2023

Tools used in this project:

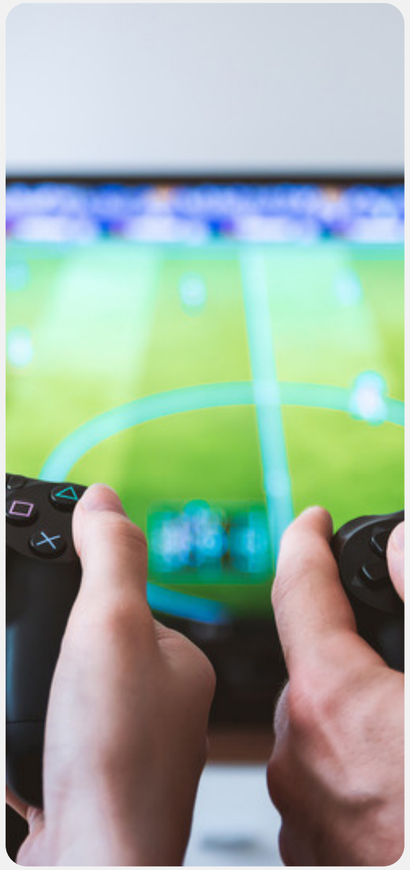
EXCEL

PROJECT OBJECTIVES

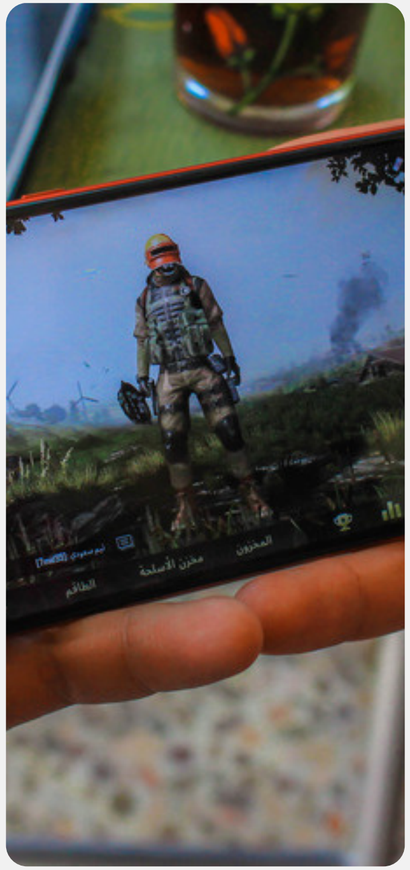
For this project, I focused on finding insights and developing visualizations that could help executives at GameCo develop better solutions to increase their sales growth.
Specifically, I wanted to find niche markets that had previously shown potential to become large top selling divisions at GameCo. The main areas I wanted to focus on was the region, genre and top games sold.



Find the Best Regional Growth Opportunities within European Markets



Find the Top Performing Genres in the Global Market

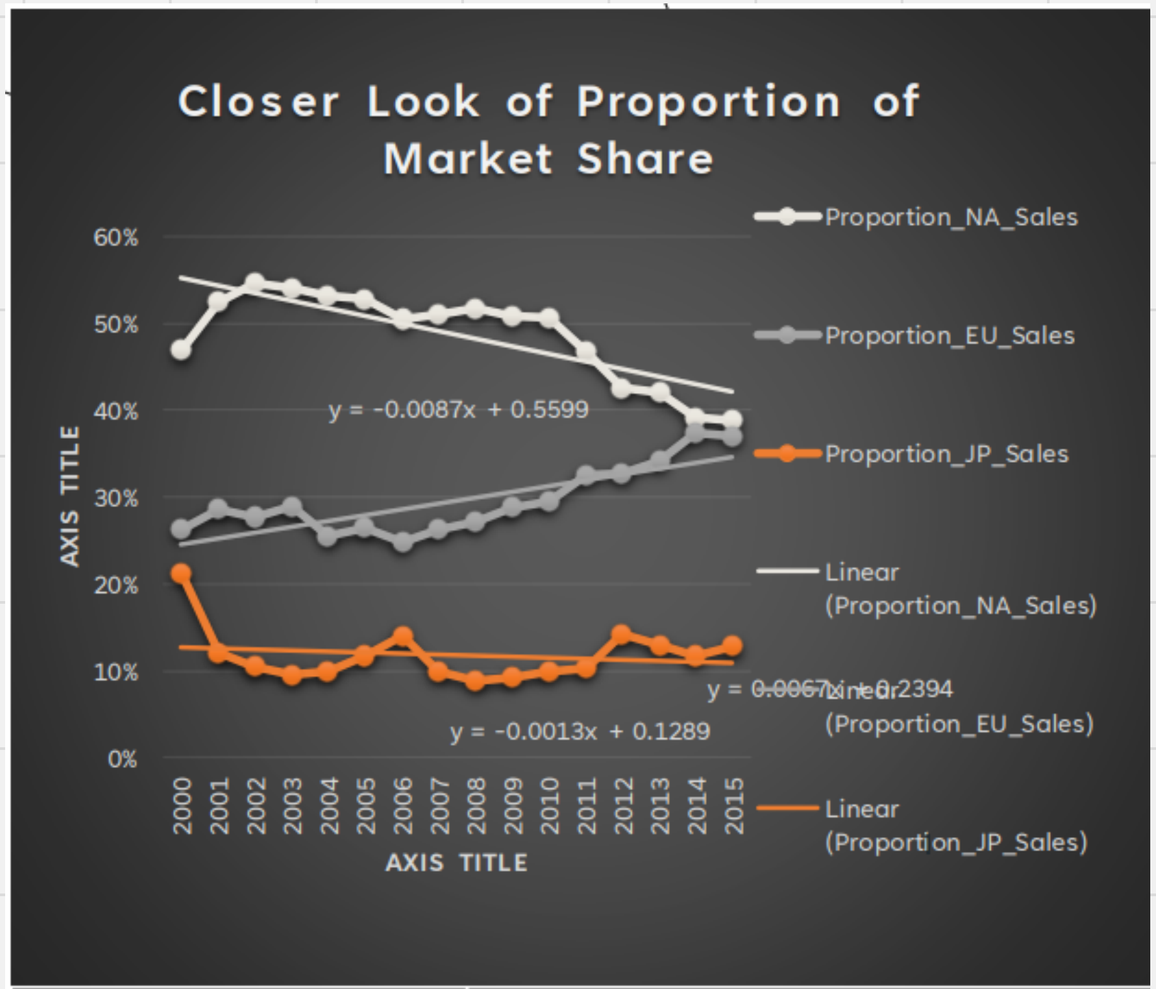


Find the Top Performing Games in the Top Genres in European Markets

DATA INSIGHTS AND DISCOVERIES

For this project I used Excel Pivot Tables to derive new variables from the company's sales data. This led to insights on proportional growths of market share for each of the company's regional sales.

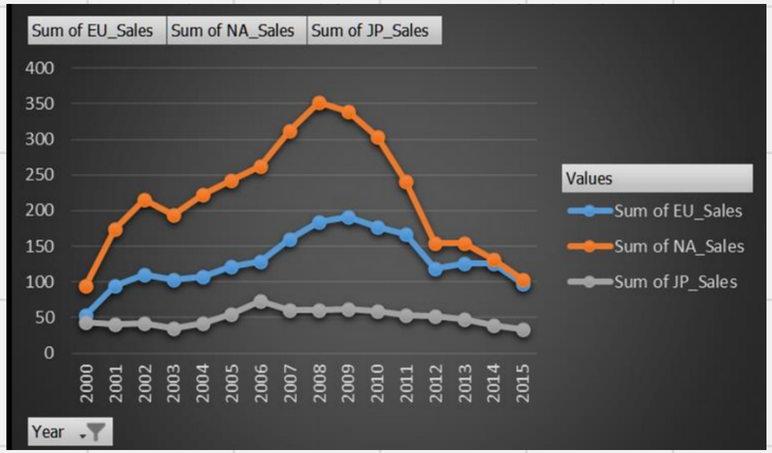
This analysis provided insight on pinpointing regions with continuous growth rates. This can bring value to stakeholders who are trying to understand more about potential investment opportunities in their company.



Region	Annual Growth Rate
North America	- 0.67%
Europe	+ 0.93%
Japan	- 0.58%

CONCLUSION AND RECOMMENDATIONS

The project deliverables were summarized and presented using Microsoft PowerPoint. The slides included Excel visualizations and a summary of the data analysis procedure used to support the insights and recommendations provided.



LINK TO PRESENTATION:LINK

REAL ESTATE VALUES PROJECT

Completed 2024

Tools used in this project:

EXCEL

PYTHON

JUPYTER NOTEBOOK

PROJECT OBJECTIVES

This analysis looked at real estate values and analyzed the factors that influenced them. The project consisted of analyzing open sourced data from multiple open sources such as the Quandl library. Here I used Python to consolidate, edit, and analyze data and generate insights. I used statistical tools, visualization libraries, and machine learning to analyze the variables that influence real estate prices from real world data. The deliverable for this project was a Tableau Presentation showcasing the analytics results.

Study's Limitations:

For this project, the data was limited to properties available from Realtor.com and Zillow.com. They were also limited to properties only in certain Northeastern states therefore, they can pose geographical and source bias.

PROJECT BRIEF: LINK

MOTIVATION:

MANY REAL ESTATE COMPANIES AND INVESTORS ARE ALWAYS LOOKING FOR FAIR PRICED OR UNDERVALUED REAL ESTATE INVESTMENTS. FROM OUR ANALYSIS, WE CAN CREATE MODELSBASED ON VARIABLES THAT DRIVE HOUSING PRICES, THEN USE THESE MODELS TO DETERMINE FUTURE LISTINGS AS UNDERVALUED OR OVERVALUED.

OBJECTIVE:

- ISOLATE THE KEY VARIABLES THAT CORRELATES THE MOST TO HOUSING COSTS. THEN USE THESE VARIABLES TO GENERATE CLUSTER MODELS OF SPECIFIC HOUSING PROFILES.
- USE THE PROFILES TO DETERMINE FUTURE PROPERTY VALUE.

SCOPE:

THIS ANALYSIS COVERS REAL ESTATE PROPERTY VALUES FOR MIDDLE CLASS SINGLE FAMILY HOMES IN THE NORTHEASTERN UNITED STATES. SPECIFICALLY, STATES SUCH AS NEW JERSEY, MASSACHUSETTS AND NEW YORK.

DATA SETS USED

PROPERTY LISTINGS FROM REALTOR.COM

14 MILLION ENTRIES OF PROPERTIES IN THE NORTHEAST REGION.
CONTAINS PROPERTY PRICE LISTINGS, NUMBER OF BEDS/BATHROOMS, FLOOR SPACE,LOT SIZE 2009-2023, SOURCED FROM KAGGLE.

MEDIAN PROPERTY VALUES OVER TIME FROM ZILLOW.COM

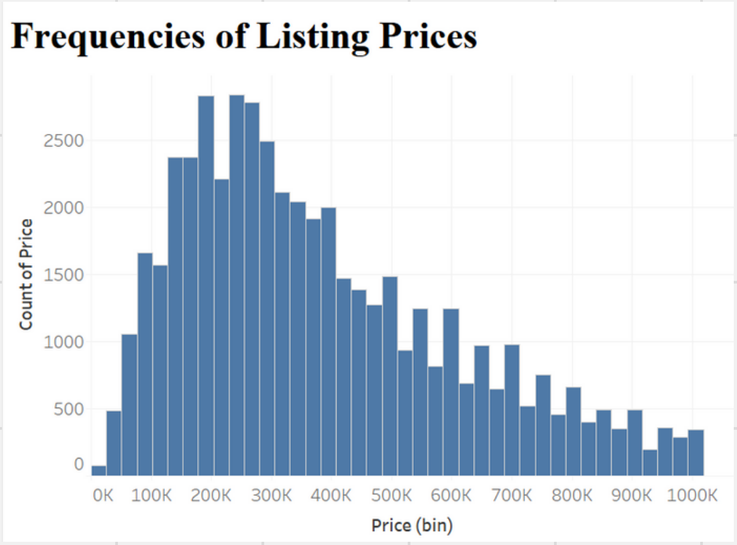
MEDIAN VALUE OF PROPERTIES SOLD IN THE PAST FOR ALL UNITED STATES REGIONS - 2008-2023
SOURCED FROM QUANDL LIBRARY - NASDAQ DATA LINK

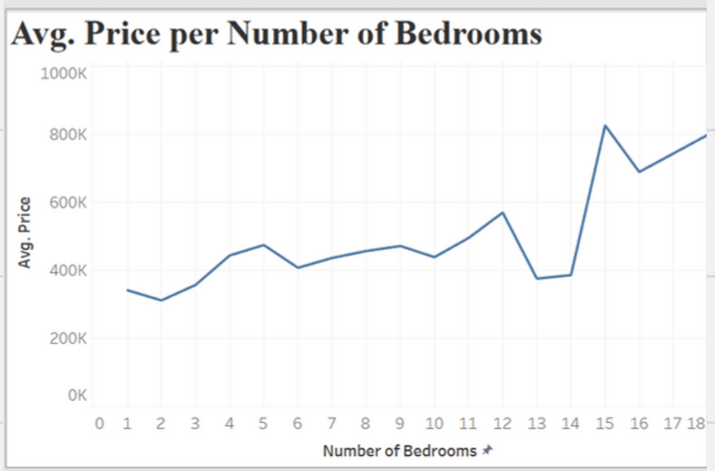
SUMMARY STATISTICS OF HOUSING PROPERTIES

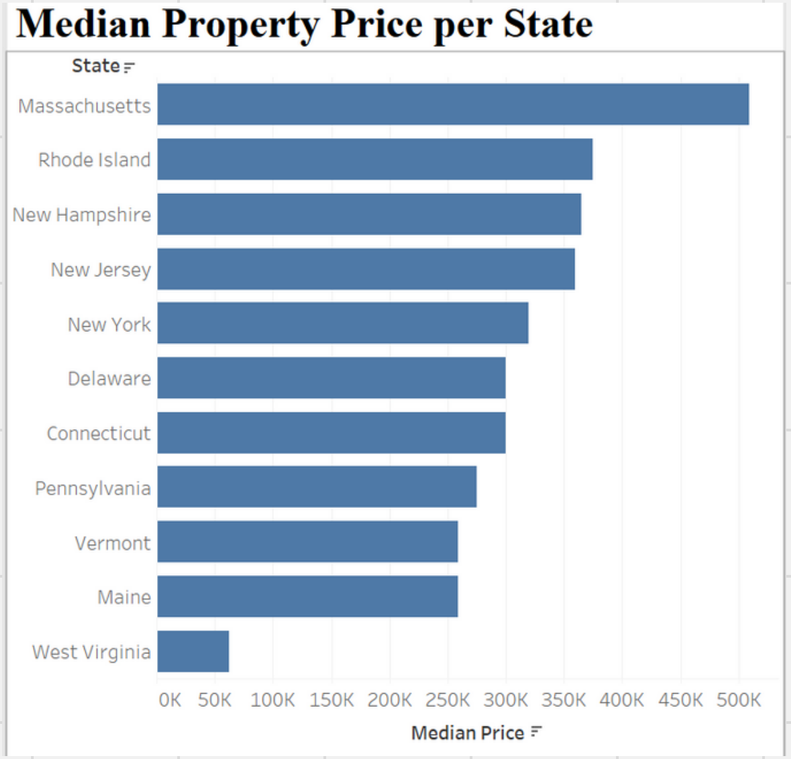
For this project, I began the analysis by using python to gain summary statistics and basic exploratory data to gain an understanding of the variables. The variables available from the datasets included the price of the property and characteristics of the house such as floor space and lot area.

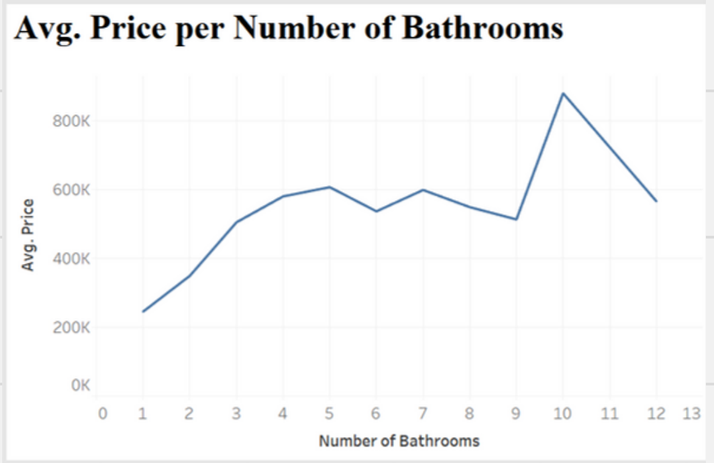
Findings and Results from the analysis:

- Massachusetts has the highest average prices
- Lot Size does not seem to have a correlation with property price
- Number of Bathrooms and Bedrooms correlate with higher prices
- The median property price is around $250K for properties under $1 million.







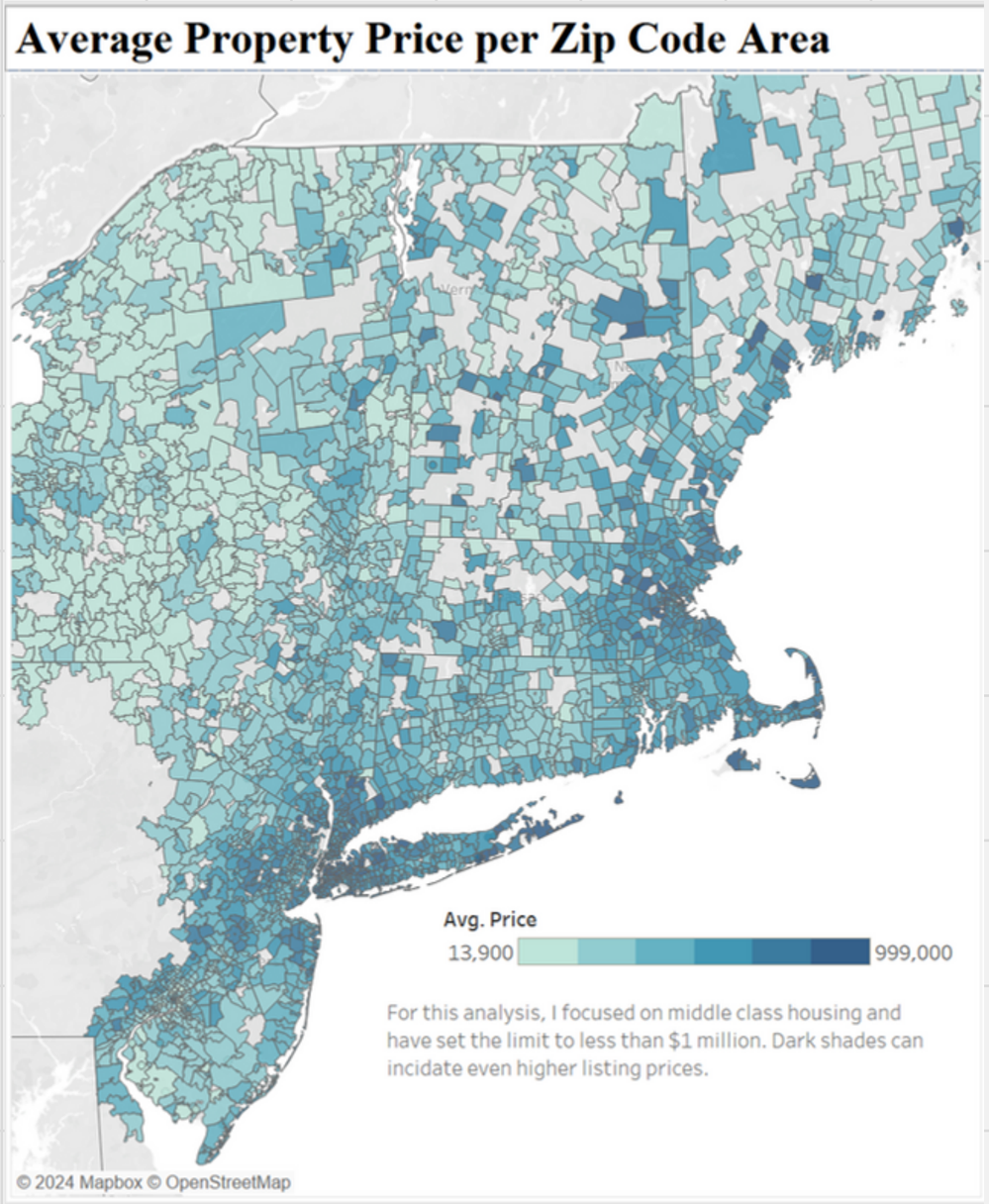


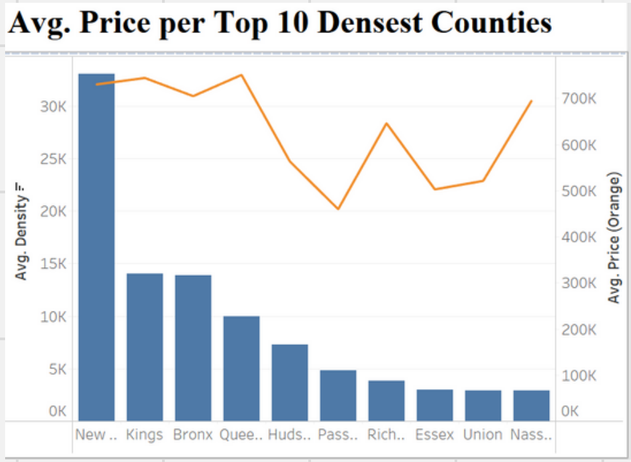
GEOSPATIAL ANALYSIS

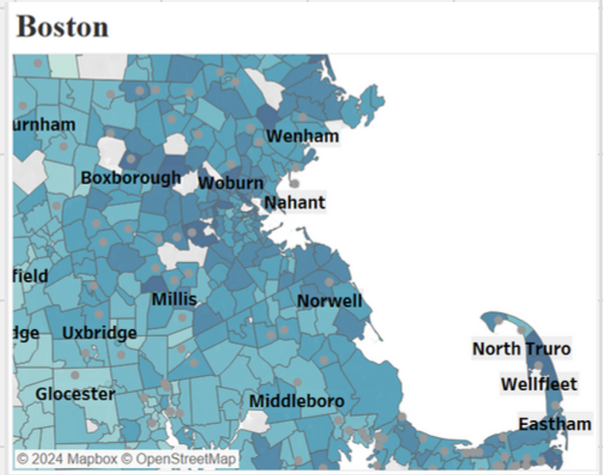
This project also includes geographic analysis of real estate values. These images were created using python visualization tools and were derived from merging datasets with Census data sourced from open resources. Here, I was able to see certain relationships between geography and average housing values.

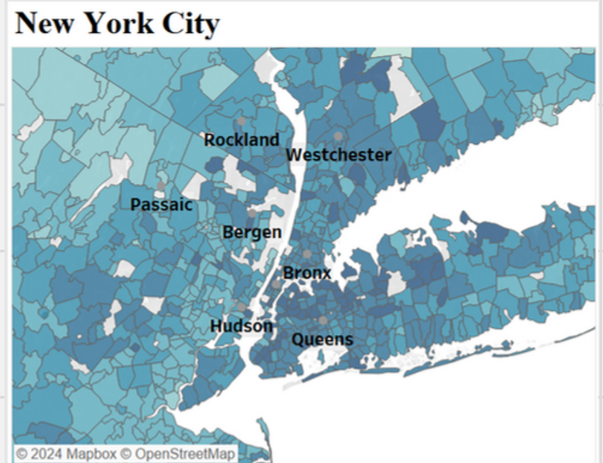
Observations Made:

- The highest prices seem to consolidate around high density urban areas. For our data, mainly in New York City and Boston.
- The bottom graph shows the top 5 counties with the highest population density and their average listing prices below $1million. Almost all are above $500k.
- This shows that population density is also a strong variable for predicting high real estate prices.
- Location is also a major factor, it seems properties near the coastline are valued higher.









FACTORS ANALYSIS

Hypothesis: A property's attributes have a major impact on its price relative that can be used to create predictive models. These can include innate factors such as number of bathrooms, total floor space, or lot size. This also includes external factors such as population density of the neighborhood.

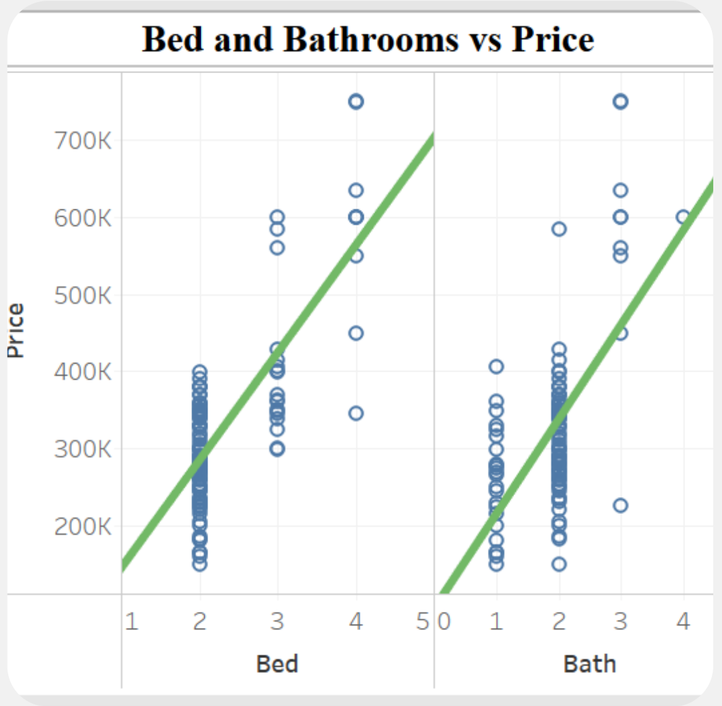
Objective: This analysis aimed to find more linear dependences between variables from our housing data that correlates well with property values.

Results

- Found no correlation between lot size and housing price.

- There is a moderate relationship between the number of beds/bathrooms and the total floor space of the house. This is increased when focusing on small areas such as measuring prices listed within a zip-code.

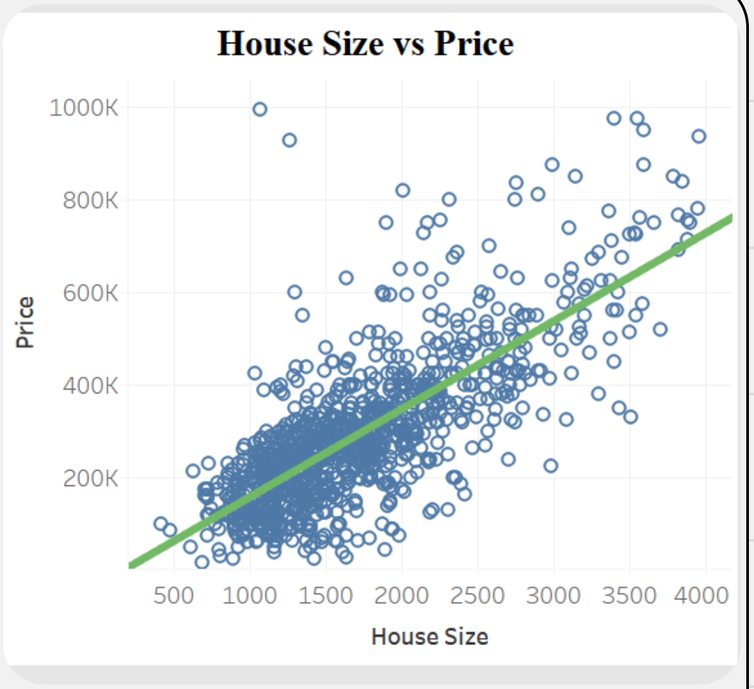
- Further study of House Size shows very low correlation when analyzing all available state data. Specifying into a singular state increases correlation.



Amenities vs Price

Bedrooms
R2 Value = .30 - Moderate Correlation
P-Value = .001 - Statistically Significant
Mean Squared Error - 4.0e10

Bathrooms
R2 Value = .40 - Moderate Correlation
P-Value = .001 - Statistically Significant
Mean Absolute Error - 4.1e10



Total Floor Space vs Price

House Size
R2 Value = .57 - High Correlation
P-Value = .001 - Statistically Significant
Mean Squared Error - 2.29e10

These values also show that there are more correlating factors that must be included to allow us to accurately value a property.

USING MACHINE LEARNING

For this project, I used unsupervised machine learning to develop clusters around datapoints of all the property variables. This identified 4 major clusters of housing profiles. These properties had many variables in common including price range.

Cluster 0 - Light Pink
- Relatively smaller homes, lower square footage
- Lower level price range
- Mostly lower sized lot areas
- Moderate amount of bedrooms and bathrooms

Could be Small Single Family Homes

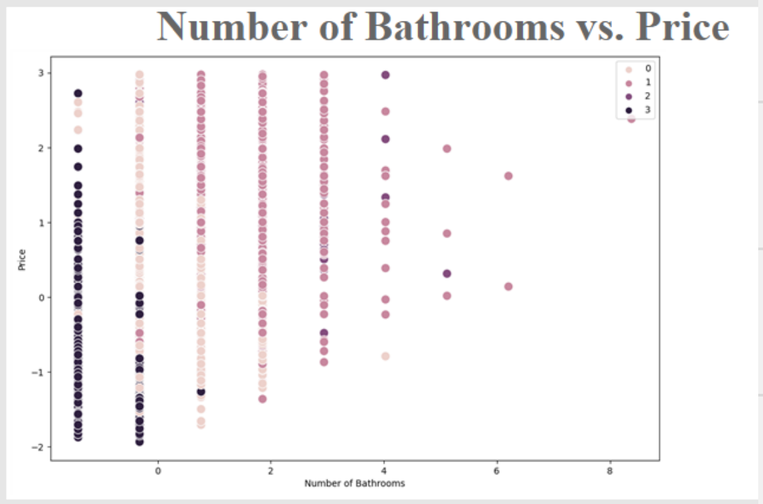
Cluster 1 Pink
- Highest Value homes
- Relatively larger homes
- Wide range of bathrooms and bedrooms.
- Wide range of lot sizes

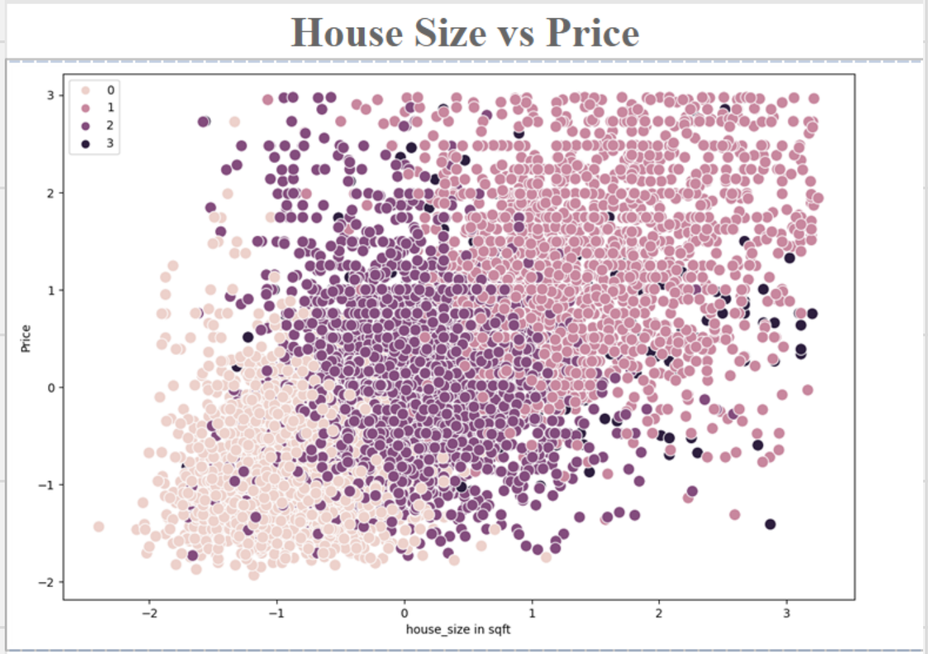
Could be Large family homes, multifamily properties

Cluster 2 - Dark Purple
- Medium priced homes,
- Medium house size, average total square foot
- Low acreages means lower lot sizes

Could be: Townhomes, individual apartments, suburban homes

Cluster 3 - Black
- Lower price range
- Wide range of House Sizes
- Lower lot sizes.
- Low number of bedrooms and bathrooms
- Could be rural properties such as Ranch Homes





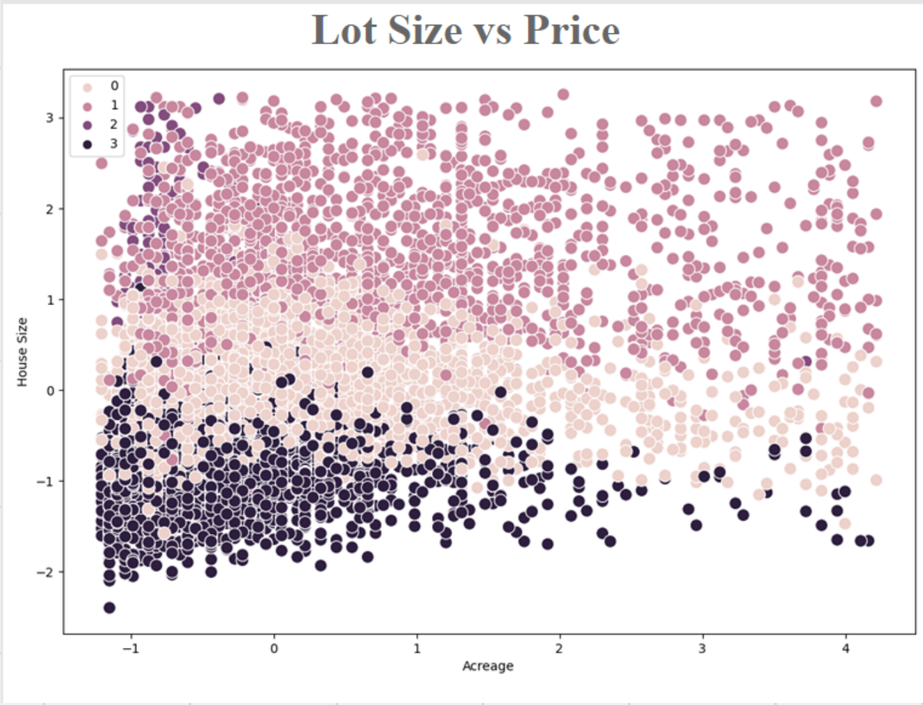


TABLEAU PRESENTATION: LINK
GITHUB LINK: @POLLICARG42

RECOMMENDATIONS AND CONCLUSIONS

The deliverables for this project involved creating visualizations and presenting them in a Tableau Storyboard. This deliverable would be presented to an analytics team working in real estate where I would show the methods and analysis that were used to develop my findings and generate my models.

These slide excerpts show my conclusions, recommendations and next steps of the report. I found some variables that correlated highly with real estate prices, and some that did not. I then identified many more variables that needed to be included and areas where we might source them.

Next Steps

Gather and Analyze More Housing Data with more Variables
- This analysis used only 1 source of data for its housing properties. Multiple real estate listing services could offer more data and more variables to analyze.

Gather External Data, Census, Real Estate Market Data
Many other outside factors affect the housing listing price. Gathering external data to incorporate the environment will improve the models.

Identify and Improve Housing Profiles
ML clusters found profiles for certain groups of properties. Investigating further to identify why these clusters were chosen, and more commonalities of these clusters can help improve their valuations.

Conclusions and Findings

Internal Variables Discovered
- **Number of Bedrooms and Bathrooms** correlate highly with property value
- **Total Floor Space** has a high correlation to listing prices

External Variables Discovered
- **Population Density** has a strong relationship to price
- Proximity to **urban centers, cities or coastlines** has an effect on value

Clusters Discovered
- Certain types of homes with similar internal variables seem to concentrate around similar price ranges
- **4 Clusters found** with their own common characteristics, these could be profiled into **home types**.

Recommendations

Identify More Internal Variables
Large MSE errors found in analysis points to other variables that may affect housing prices as well. These can be more internal variables of the property. These can be characteristics such as garage size, roof quality, recent remodels, safety, infrastructure. These characteristics can affect the value of home to buyers and will improve modeling future property values.

Identify More External Variables:
- Other external factors that are involved in housing prices must be analyzed. These can include factors attractive to home buyers, such as distance to school, amenities, and local public services.
- Considering factors with each area zone such as tax rates, zoning laws, local amenities can make more efficient models.

THANK YOU