Lertkiet Lertchayantee

Data Analyst Portfolio

About Me

With a strong foundation in engineering and experience managing electrical projects in industrial plants, I naturally transitioned into the field of data analytics. My previous role involved reconciling and analyzing large datasets in Excel to inform engineering decisions, which laid the groundwork for my advanced analytical capabilities.

Completing the CareerFoundry Data Analytics program allowed me to deepen my expertise in data cleaning, analysis, and visualization while acquiring technical skills in SQL and Python. Additionally, I honed my ability to create insightful visualizations and design impactful storyboards in Tableau, enabling me to deliver actionable insights and data-driven solutions.

Portfolio Case Study Overview

01

GameCo

Global video game sales analysis

02

Influenza Season

Providing staff analysis for medical facilities

03

Rockbuster Stealth

Strategy analysis for the new online video service

Portfolio Case Study Overview

04

Instacart

Analysis to uncover data on sales patterns

05

Pig E. Bank

Data mining analysis for a global bank

06

House Sales in King County

Insights into House Prices Uncovered

GameCo

Global video game sales analysis

GameCo

Global video game sales analysis

Background

GameCo is a new video game company discovering that sales for the various geographic regions have stayed the same over time.

Methods

- Grouping/sorting/filtering data
- Descriptive analysis
- Data cleaning
- Data visualization

Tools

- MS Excel

Goal

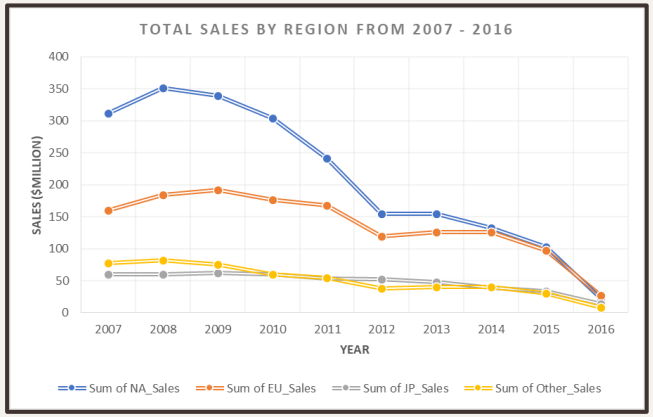
Conduct a comprehensive descriptive analysis of a video game dataset to gain deeper insights into market trends, player preferences, and factors influencing game performance

Data

Data was drawn from the VGChartz website and provided by CareerFoundry

Link

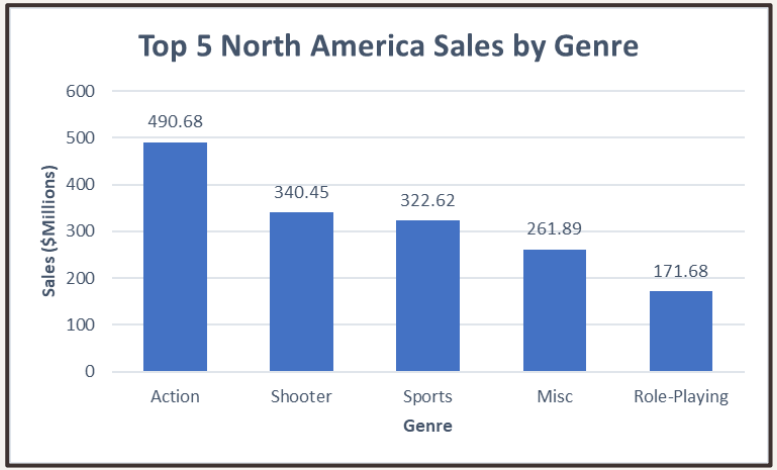
1. Raw Data 2 Additional Files Related to This project



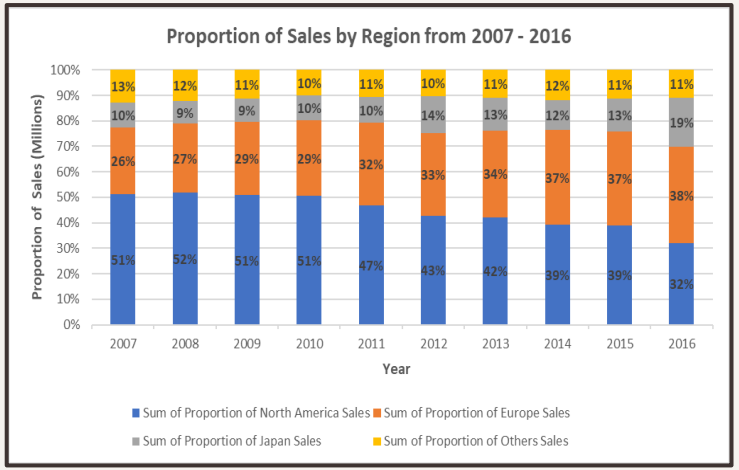
Market data reveals a notable decrease in overall market size, shifts in regional market dominance, and the emergence of growth opportunities in smaller, previously underrepresented regions.

These trends suggest a dynamic and evolving market landscape that requires a more targeted and adaptive approach to marketing strategy.

Result and Insights

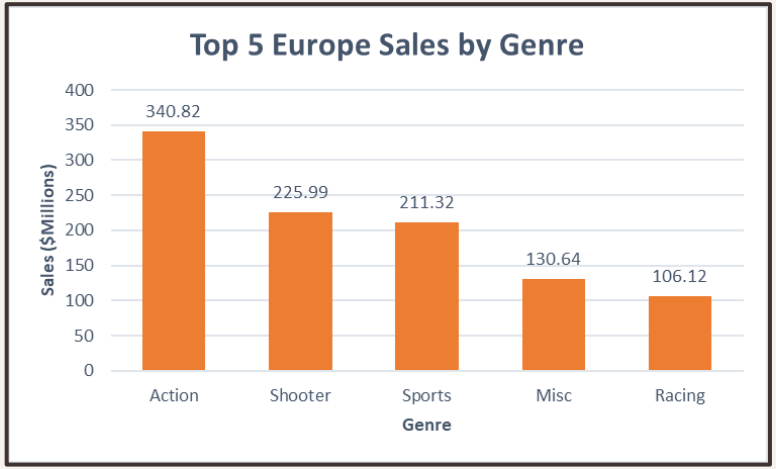


In **North America, action and shooter** games are **the most popular** ones over the years (2007 - 2016).

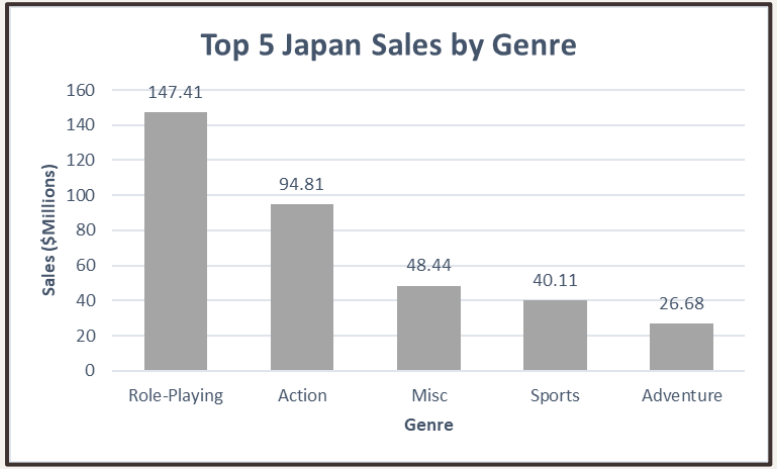


The proportion of sales in **North America has taken majority**, but it **has decreased** over the years

Result and Insights



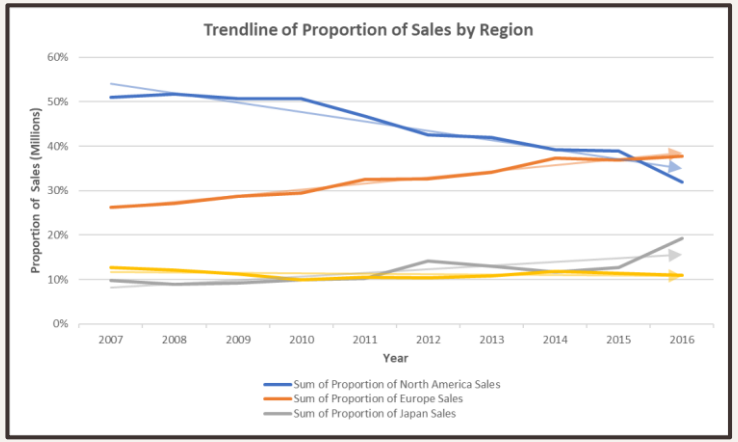
In Europe, action and shooter games are **the most popular** ones over the years (2007 - 2016).



In Japan, Role-Playing game **is the most popular** one over the years (2007 - 2016) due to its culture and technology.

Recommendations

Region	Priority	Budget Allocation	Comment
North America	High	35%	Focus investments on Action and Shooter games, with attention to emerging genres like Fighting.
Japan	High	30%	Prioritize the budget for Role-Playing dominance and support Action.
Europe	Medium	25%	Prioritize Action, Shooter, and Sports games while exploring growth in diverse genres.
Others	Low	10%	Focus on emerging opportunities, especially in Shooter games.



Future Trend

Influenza

Providing staff analysis for medical facilities

Influenza

Providing staff analysis for medical facilities

Background

Hospitals and clinics in the US need additional staff to effectively treat the increased number of patients during the influenza season.

Methods

- Data profiling and integrity
- Data Integration
- Statistical hypothesis testing
- Data visualization with Tableau
- Forecasting

Tools

- MS Excel
- Tableau

Goal

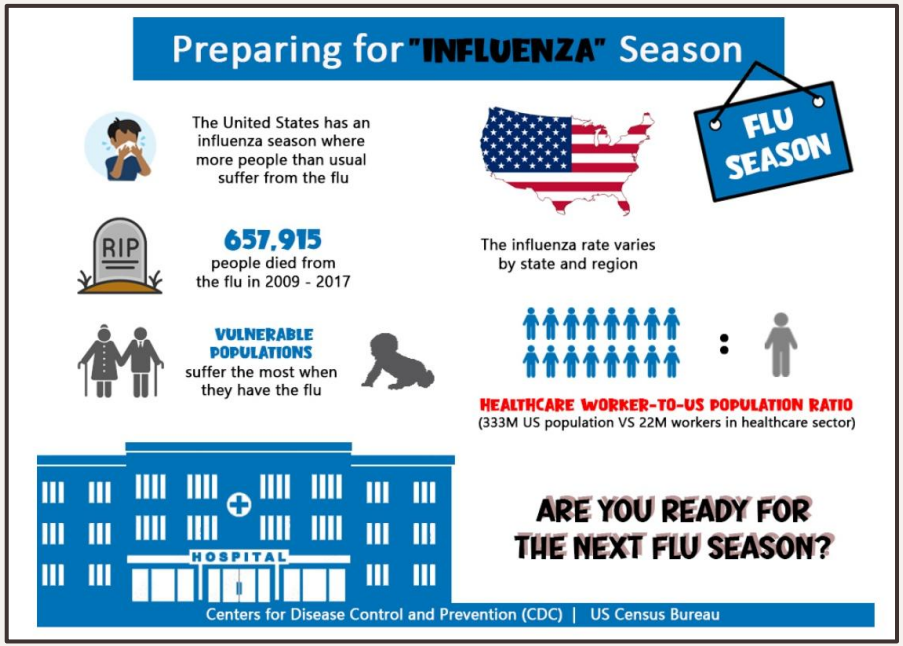
- Develop a plan to forecast additional staff demand for the influenza season.
- Analyze influenza trends to guide staffing needs nationwide.

Data

The dataset was collected from the US Centers for Disease Control and Prevention (CDC) and the US Census Bureau.

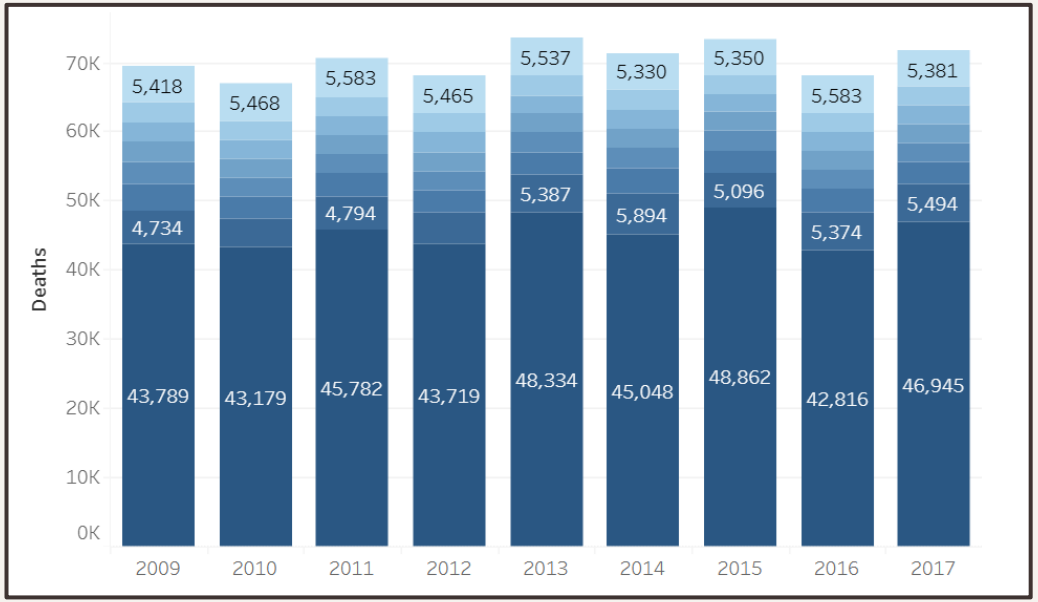
Links

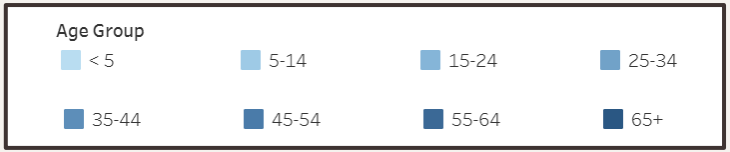
1. Raw Data (CDC)
2. Raw Data (US Census)
3. Tableau Presentation
4. Additional Files Related to This Project



Result and Insights

Influenza Death by Age Group (2009 - 2017)

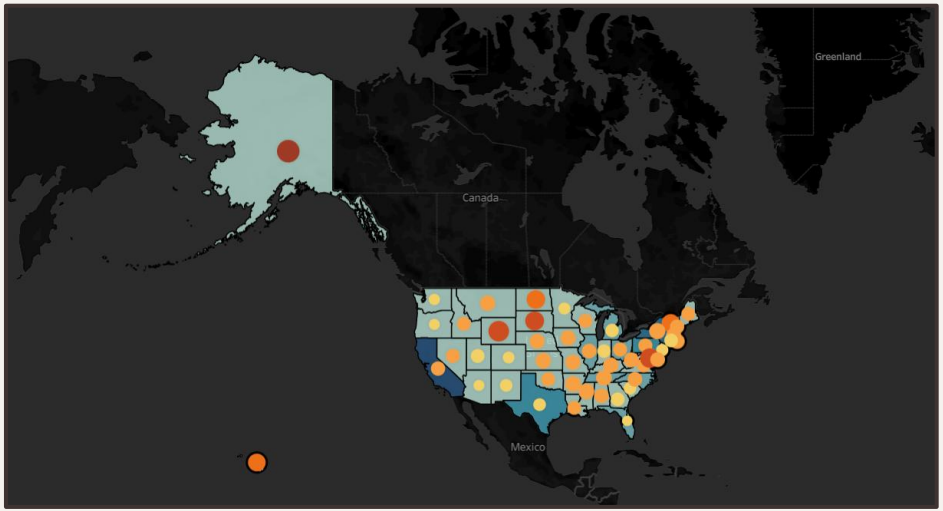


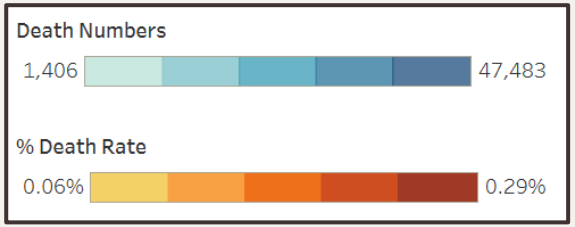


Age groups under 5 and over 65 have significantly **higher total death rates** compared to other age groups. Together, these two age groups account for over 70% of the total influenza-related deaths.

Result and Insights

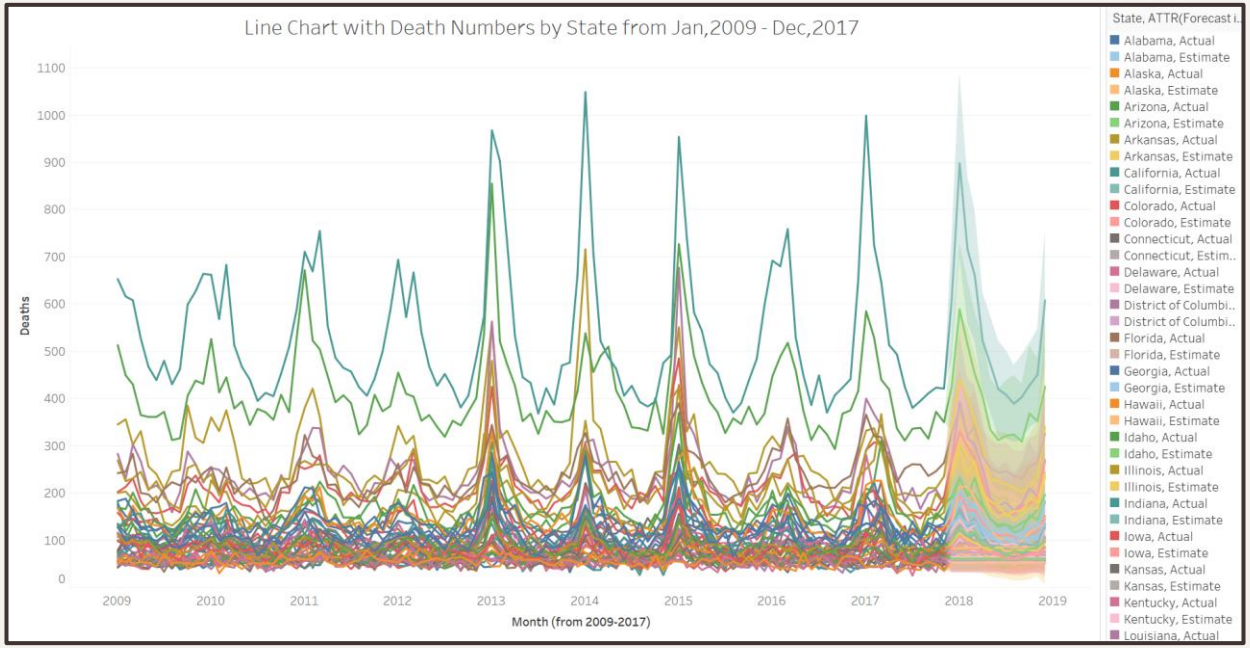
Influenza Death Numbers VS Death Rates of 65+ by State (2009 - 2017)





1. The states with the most deaths are California, New York and Texas.
2. The states with the highest death rates are Alaska, Wyoming and Hawaii.

Result and Insights



Based on the analyzed data**, the highest death numbers** occur in the first and fourth quarters, which correspond to **the fall and winter seasons**.

Recommendations

Staffing Based on Risk:

Clinics should allocate additional staff and resources to individuals aged 55 and older, particularly during the first and fourth quarters, when the risk of influenza-related deaths is highest.

Geographic Focus:

Health systems in the ten most affected states (as identified in our analysis) should receive extra resources and staff, especially during flu season, to help reduce mortality rates.

Continuous Monitoring:

The eight states predicted to experience rising influenza-related deaths post-2017 should be continuously monitored to ensure their preparedness, as influenza death rates tend to peak during the fall and winter seasons.

Vaccination and Education:

Efforts to strengthen vaccination campaigns for high-risk groups should be prioritized, with a focus on educating the public—particularly in states most heavily affected by influenza.

Data-Driven Planning:

Predictive models should be used to plan staffing and resource allocation in advance of peak influenza season, ensuring that clinics are adequately prepared to handle the increased demand.

Rockbuster Stealth

Strategy analysis for the new online video service

Rockbuster Stealth

Strategy analysis for the new online video service

Background

Rockbuster Stealth is facing stiff competition from streaming services. The management team is planning to launch an online video rental service to stay competitive.

Methods

- PostgreSQL
- Data cleaning in SQL
- Joining Tables
- Subqueries & CTEs

Tools

- MS Excel
- Tableau
- SQL

Goal

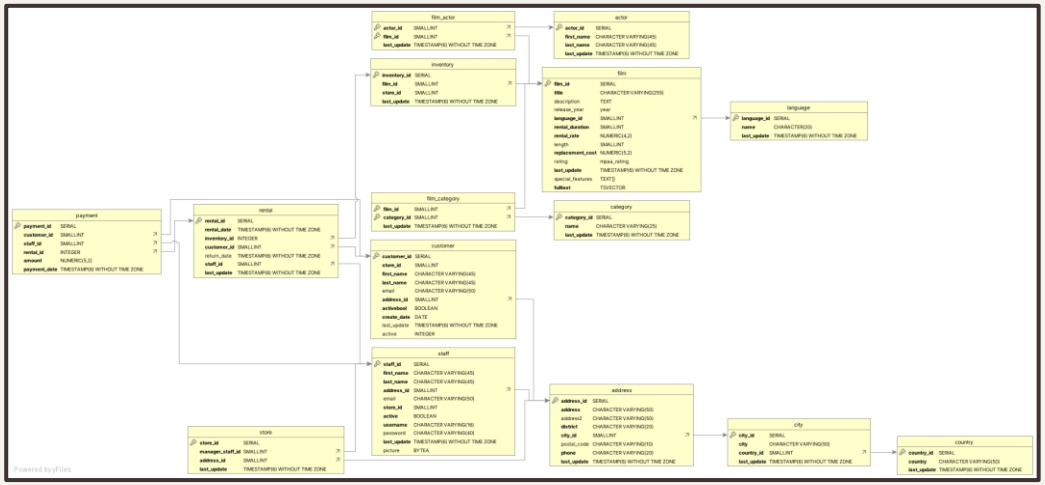
Provide data-driven insights to assist the management team of Rockbuster Stealth LLC in strategically launching their new online video rental service.

Data

The dataset includes information about Rockbuster's film inventory, customers, payments, and other relevant details, as illustrated in the data diagram below.

Links

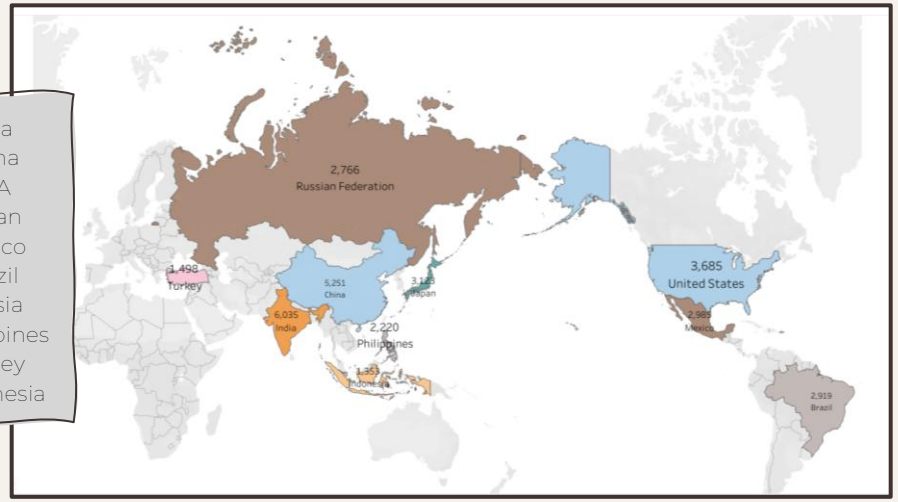
1. Tableau Visualizations 2. Additional Filed Related to This Project



Result and Insights

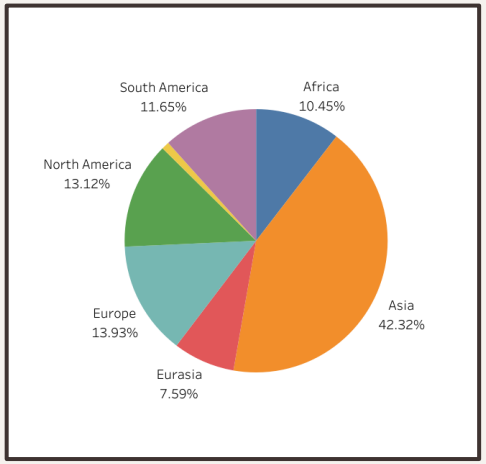
Analyzing Our Customer Distribution: A Global Overview

1. India
2. China
3. USA
4. Japan
5. Mexico
6. Brazil
7. Russia
8. Philippines
9. Turkey
10. Indonesia



Rockbuster operates in **109 countries worldwide**, with **just 10 countries accounting for more than 50% of the total** customer base.

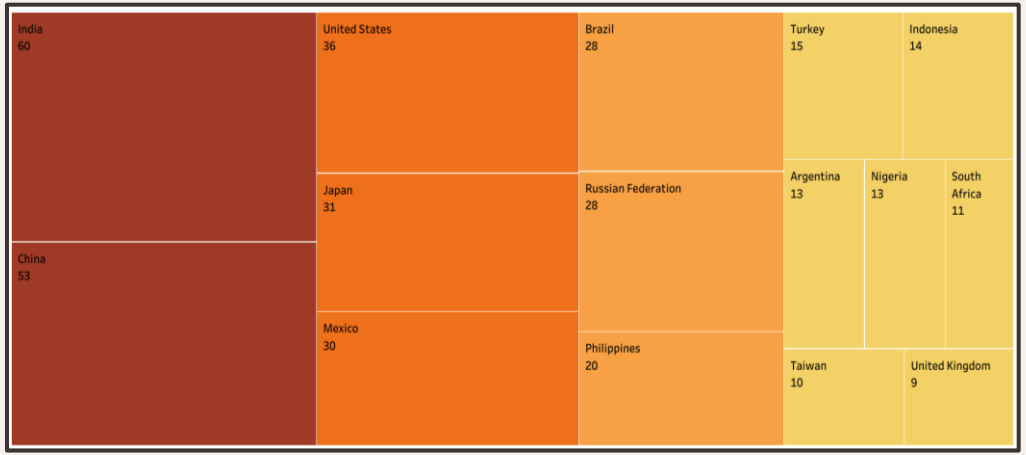
Customer Distribution Analysis: Regional Overview



Asia has the **highest sales**, accounting for over 40%, while **Oceania** has the **lowest**, with less than 8%.

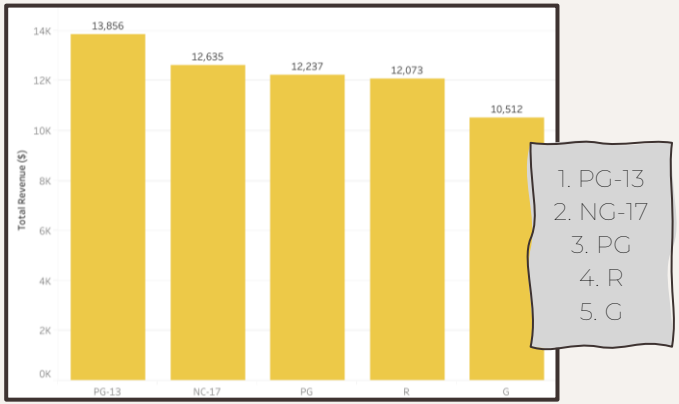
Result and Insights

Analyzing Our Customer Distribution:
Overview of Countries Where Customers Are Based



India, China, the USA, Japan, and Mexico are the **top five countries** where Rockbuster **customers are based**, each with a customer range of 30 to 60.

Customer Distribution Analysis: Top-Performing Genres by Revenue



PG-13 rating has **the highest total revenue**. It has 9.66% in total revenue higher than the second one (NC-17).

SQL Code Examples

Joining Tables

Query | Query History

```sql
SELECT
    D.country,
    C.city,
    COUNT(customer_id) AS customer_numbers_by_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
WHERE D.country IN ('India', 'China', 'United States', 'Japan', 'Mexico' ,
          'Brazil', 'Russian Federation', 'Philippines', 'Turkey', 'Indonesia')
GROUP BY D.country, C.city
ORDER BY customer_numbers_by_country DESC
LIMIT 10
```

Aggregate Functions

Query | Query History

```sql
SELECT rating,
       COUNT(film_ID) AS count_of_movies,
       AVG(rental_rate) AS average_movie_rental_rate,
       MAX(rental_duration) AS maximum_rental_duration,
       MIN(rental_duration) AS minimum_rental_duration
FROM film
WHERE rating in ('PG','G')
GROUP BY rating
```

Common Table Expression

Query | Query History

```sql
-- 1st CTE
WITH top_10_countries AS (
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 ),

-- 2nd CTE
top_10_cities AS (
SELECT D.country, 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 D.country IN (SELECT * FROM top_10_countries)
GROUP BY D.country, C.city
ORDER BY COUNT(customer_id) DESC
LIMIT 10),

-- 3rd CTE
top_5_customers AS (
SELECT E.customer_id, A.first_name, A.last_name, C.city, D.country,
   SUM(E.amount) AS Total_Amount_Paid
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
INNER JOIN payment E ON A.customer_id = E.customer_id
WHERE (D.country,C.city) IN (SELECT * FROM top_10_cities)
GROUP BY E.customer_id, A.first_name, A.last_name, C.city, D.country
ORDER BY Total_Amount_Paid DESC
LIMIT 5)

-- The main statement
SELECT
ROUND(AVG(total_amount_paid),2) AS average
FROM top_5_customers
```

Recommendations

Geographic Market Expansion

- Asia's total revenue accounts for over 40%. Consider investing in strategies to maintain customer engagement in this region, such as implementing a loyalty program.
- Develop strategies to boost revenues in Oceania and South Africa.

Targeted Marketing by Rating

- PG-13 and NC-17 rated movies generate the highest revenues. Rockbuster should focus marketing efforts on promoting movies with these ratings.
- Investigate the reasons behind the low revenue from G-rated movies, despite their broad age appeal. Additionally, create promotional campaigns to increase sales for these movies.

Expand Inventory of High-Revenue Movies

- Analyze the top 5 least revenue-generating movies to determine if they should remain in stock. Would it be more beneficial to stock more high-demand, high-revenue movies instead of keeping low-demand movies on hand?

Optimize Long Rental Duration

- The data shows that movies with shorter rental durations generate more revenue. Explore solutions to increase revenue from long-duration rentals while maintaining customer satisfaction, such as adjusting prices or increasing the availability of these movies in the inventory.

Instacart

Analysis to uncover data on sales patterns

Instacart

Analysis to uncover data on sales patterns

Background

Instacart stakeholders are interested in customer variety and purchasing behaviors to target different segments with marketing campaigns and assess their impact on product sales.

Methods

- Data wrangling and subsetting
- Merging data frames
- Grouping and aggregating data
- Data visualization with Python

Tools

- MS Excel
- Python

Goal

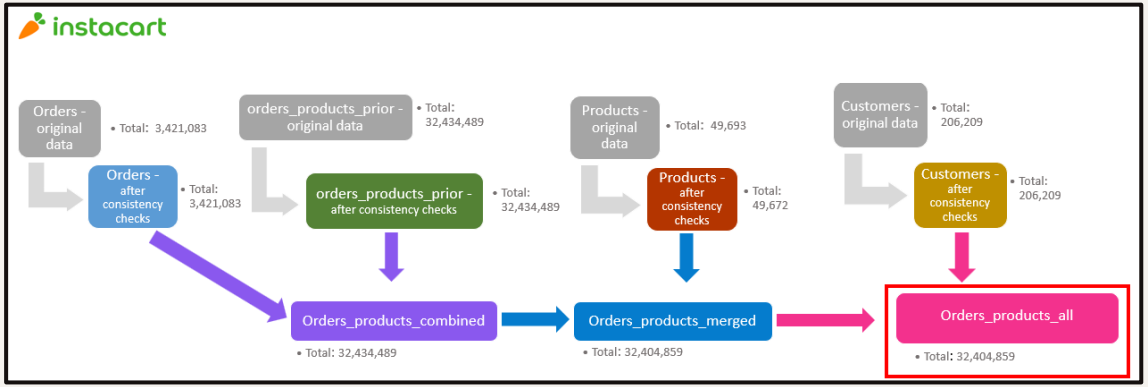
Performing initial data exploration and analysis to uncover sales patterns, aiming to generate insights and recommend strategies for better customer segmentation.

Data

The dataset was collected from an open-source platform provided by Instacart.

Links

1. Raw Data 2. Python Scripts, Visualizations and Additional Files



Instacart's expense dataset includes crucial information related to customers, payments, and inventory. The visualization above highlights the population and the necessary dataset integrations for conducting this analysis.

Analytic Process

1. Data Wrangling: Explored and cleaned the datasets for analysis.

2. Data Merging: Merged multiple datasets into a unified dataset using the merge function.

3. Deriving Variables: Created new variables through if-statements and for-loops.

4. Grouping and Aggregating Data: Grouped the data using the groupby function and performed aggregations with agg, transform, and loc functions.

5. Visualization: Developed bar, line, and bubble charts using Python for data visualization.

6. Report Writing: Compiled an Excel report summarizing results and key insights.

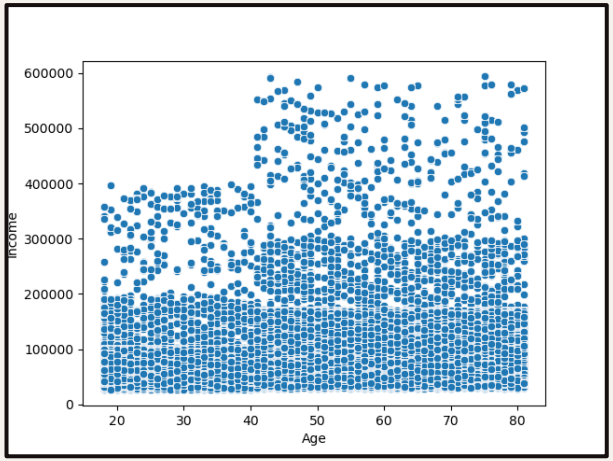
instacart

Consistency checks

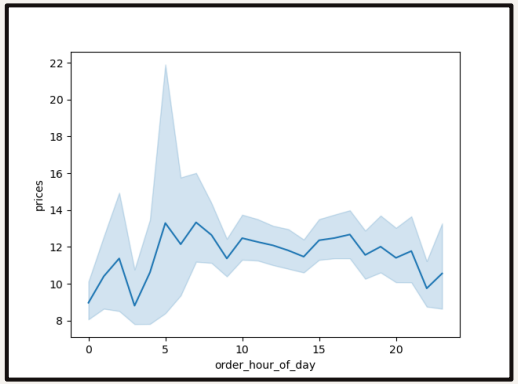
Dataset	Missing values	Missing values treatment	Duplicates
orders	206,209 missing values in days_since_prior_order column.	Flag them as a first orders in a different column.	No duplicates.
products	16 missing values in product_name column.	1. Save rows with missing values in df_nan variable. 2. Filter out those 16 missing values and save into df_prods_clean variable.	Found 5 duplicated rows and save them into df_dups variable.
orders_products_prior	No missing values.	No missing values.	No duplicates.
customers	11,259 missing values in First Name column.	Leave them as they are because first names can be unique and cannot be substituted with any other values.	No duplicates.

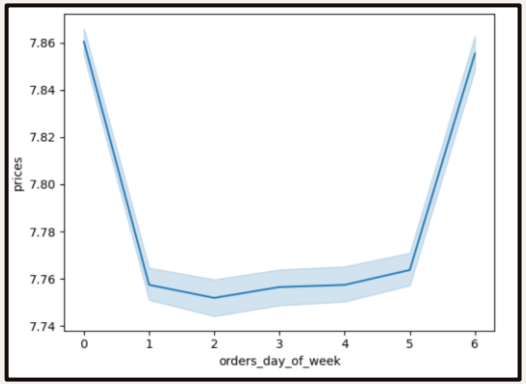
Summary of consistency checks, which is part of the process **performed in Python**. This step ensures that the data follows expected rules and patterns.

Result and Insights



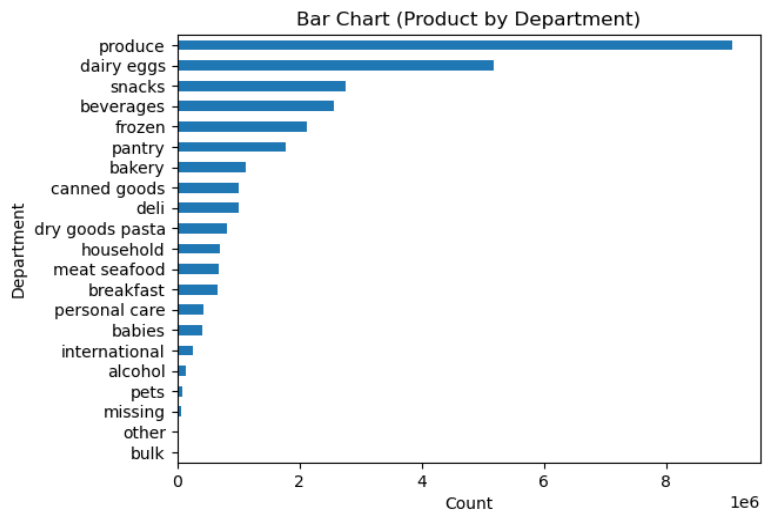
The scatterplot indicates that individuals aged 40 and above are more likely to earn higher incomes, particularly between $40,000 and $60,000.

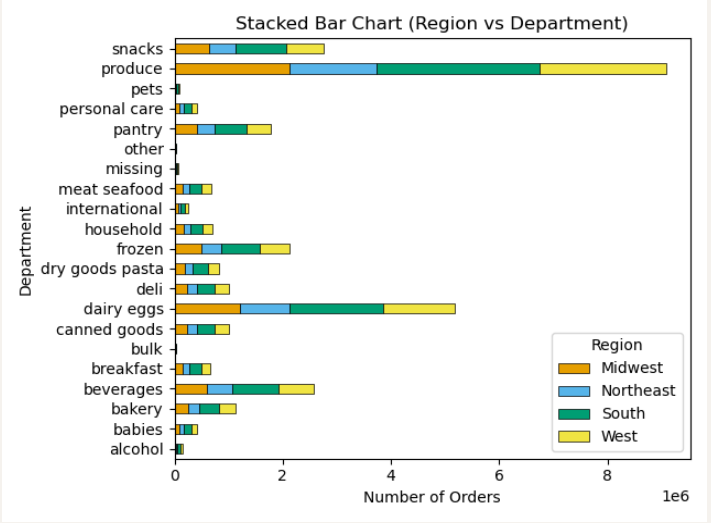




The charts show that prices are higher between 5 and 8 hours. Additionally, the error bands suggest that prices are statistically more likely to fall within the higher range at hour 5 than at other times.

Result and Insights





The charts show that product from produce department contributes the highest frequency on this dataset followed by dairy eggs and snacks. The demographic in terms of department product shows that the Southern region has the most potential on purchasing items across all regions

Recommendations

1. Launch promotions to **upsell on weekdays**, targeting slow hours (**7 p.m. to 5 a.m.**).
2. Loyal customers are most likely to return. **A loyalty program could encourage more shopping.**
3. The **new customer group contributes the least**. We should focus on promoting to them.
4. **Married people have greater potential**. Instacart should upsell family-related products.
5. The **Southern region leads in contributions**, so Instacart should prioritize it. For the **Northeastern region**, which has the **least contributions**, promotions could help boost sales.

Pig E. Bank

Data mining analysis for a global bank

Pig E. Bank

Data mining analysis for a global bank

Background

Pig E. Bank is a global financial institution with an anti-money-laundering compliance department. They develop models to lag suspicious transactions indicative of money laundering and fraud.

Methods

- Exploratory data analysis
- Data mining process
- Data modeling
- Time series analysis
- Predictive analysis

Tools

- MS Excel
- MS Pivot Table

Goal

Conduct a comprehensive analysis of bank client data to identify the key factors influencing clients' decisions to leave the bank.

Data

The data was collected from the sales team at Pig E. Bank. The dataset was created as a case study by CareerFoundry.

Links

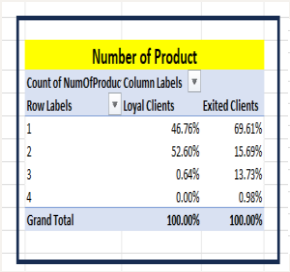
1. Raw Data 2. Additional Files Related to This Project

Loyal clients

	Credit Score	Age	Tenure	Balance	NumOfProducts	Estimated Salary
Min	411	18	0	$ -	1	$ 371.05
Max	850	82	10	$197,041.80	3	$ 199,661.50
Mean	652	38	5	$ 74,830.87	2	$ 98,942.45

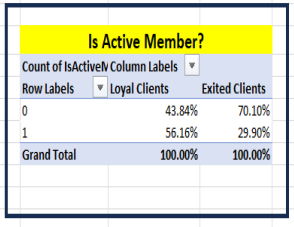
Exited clients

	Credit Score	Age	Tenure	Balance	NumOfProducts	Estimated Salary
Min	376	22	0	$ -	1	$ 417.41
Max	850	69	10	$213,146.20	4	$ 199,725.39
Mean	639	45	5	$ 90,101.69	1	$ 96,676.39



Number of Product

Count of NumOfProduc	Column Labels	
Row Labels	Loyal Clients	Exited Clients
1	46.76%	69.61%
2	52.60%	15.69%
3	0.64%	13.73%
4	0.00%	0.98%
Grand Total	100.00%	100.00%



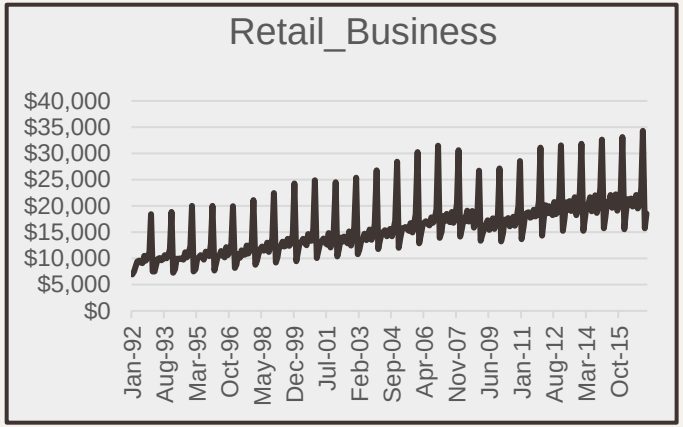
Is Active Member?

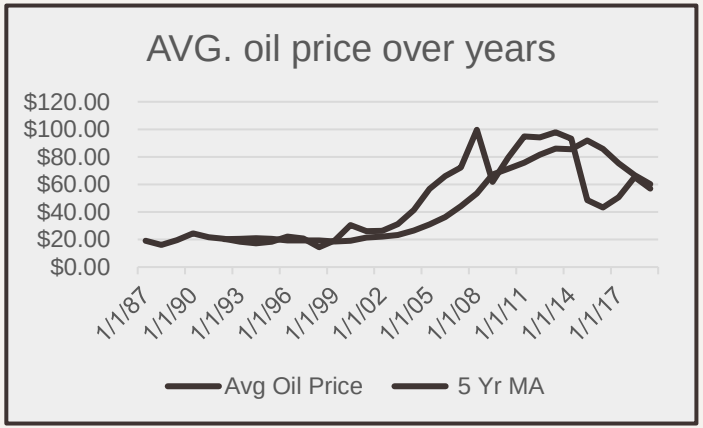
Count of IsActiveM	Column Labels	
Row Labels	Loyal Clients	Exited Clients
0	43.84%	70.10%
1	56.16%	29.90%
Grand Total	100.00%	100.00%

The descriptive analysis method (left) and Pivot Tables tool (right) were used in this project to analyze the data.

Result and Insights

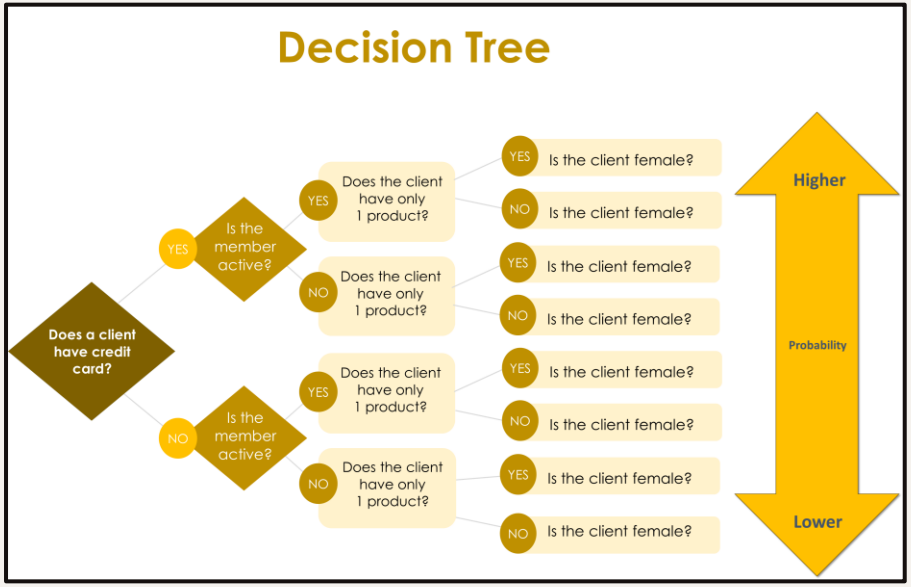
This **time series chart (right)** depicts the monthly revenue trend for the retail business from 1992 to 2016. A clear **seasonal pattern** is evident, with revenue consistently **spiking** during the holiday season in **November and December** each year, likely due to increased consumer spending during that period. Over time, there is a noticeable upward trend in the baseline revenue, indicating steady business growth year over year.





The chart (left) shows **non-stationary** characteristics with fluctuations and changing mean and variance over time. From **1987 to 1998, the data was stable** but began rising in 1999, followed by a sharp drop in 2008. Overall, there was an **upward trend with fluctuations from 1987 to 2008**. Between **2008 and 2019**, prices swung **unpredictably**, complicating future forecasting.

Recommendations



A decision tree was used to estimate the probability of customers leaving the bank.

The most impactful factors that cause customers to leave the bank

Has Credit Card? → Is Active Member? → Number of Product → Gender

1. Build a loyal base that uses more products and stays engaged.
2. Identify reasons for inactivity and potential product improvements.
3. Examine what competitors offer to understand their advantage over Pig E. Bank.

House Sales in King County

Insights into House Prices Uncovered

House Sales in King County

Insights into House Prices Uncovered

Background

As a data analyst, my role involves analyzing King County house sales data to develop predictive models for sale prices and determine the key factors affecting them.

Methods

- Relationships exploring
- Geographical visualizations
- Machine learning regression
- Machine learning clustering

Tools

- MS Excel
- Python
- Tableau

Goal

Performing exploratory data analysis to identify factors affecting King County house prices, followed by building a machine learning model for price prediction.

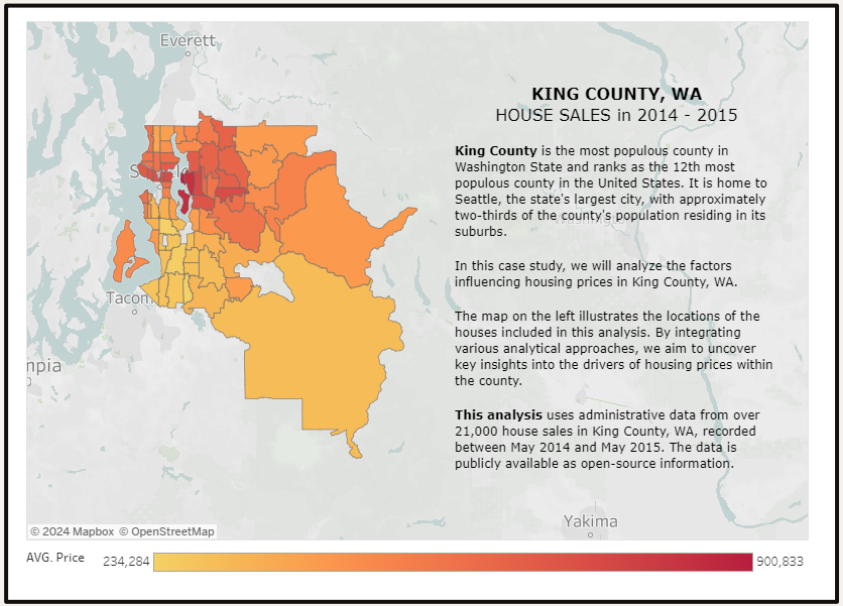
Data

This dataset, sourced from Kaggle.com, contains house sale prices for King County, including Seattle. It covers homes sold between May 2014 and May 2015.

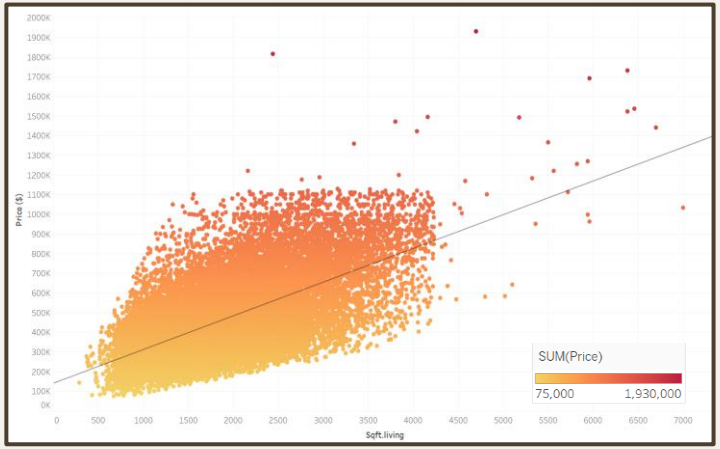
Links

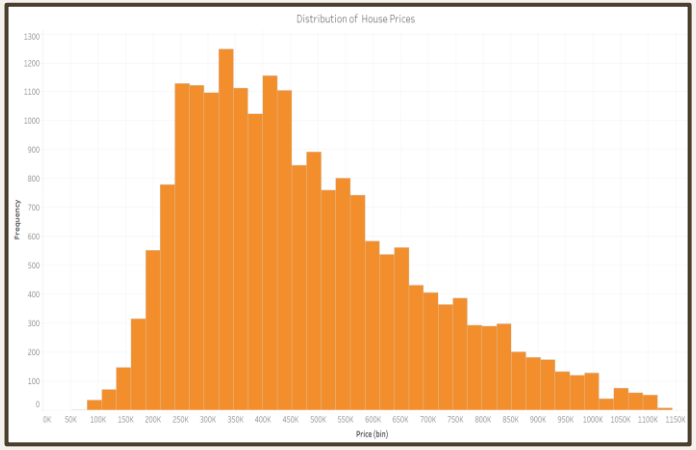
1. Raw Data
2. Tableau
3. Python Scripts and Additional Files Related To This Project

A choropleth map was created using Python to aid in the analysis for this project.



Analytic Process: Exploratory Analysis





The distribution of house prices, as shown in the chart on the left, reveals a wide range, with the majority of prices falling between **$200,000 and $450,000**. This suggests that certain factors may be driving the variability in house prices.

To explore this further, we begin by examining potential **linear relationships** between variables. The analysis indicates **no significant relationship** between the year a house was built and its price. However, the scatterplot above reveals a noticeable trend between **living space (sqft_living)** and house prices. Specifically, house prices (dependent variable) tend to increase as the size of the living space (independent variable) increases.

This upward trend leads us to the following hypothesis: **As the living space (sqft_living) increases, house prices increase.**

Analytic Process: Linear Regression


```
# Create objects that contain the model summary statistics.

# This is the mean squared error
rmse = mean_squared_error(y1_test, y1_predicted)

# This is the R2 score.
r2 = r2_score(y1_test, y1_predicted) # This is the R2 score.
```

```
print('Slope:' ,regression.coef_)
print('Mean squared error: ', rmse)
print('R2 score: ', r2)
```

```
Slope: [[170.49090601]]
Mean squared error:  26279471646.580242
R2 score:  0.3837514452734949
```

Supervised machine learning regression was applied in this analysis. As shown in the chart on the left, the red regression line from the test set aligns with the hypothesis: ***"Increasing the number of square feet (sqft_living) leads to higher house prices."*** It is evident that most house prices increase as living space expands.

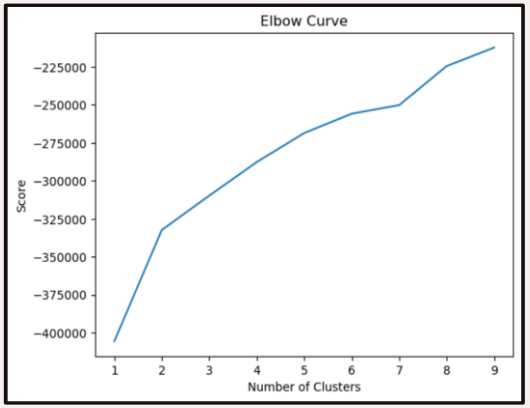
However, some data points indicate exceptions where larger living spaces do not correlate with higher prices. To assess the model's accuracy, we must use metrics beyond visual inspection. As shown in the lower-left code, metrics like **MSE (Mean Squared Error) and R^2 score** are useful for evaluation.

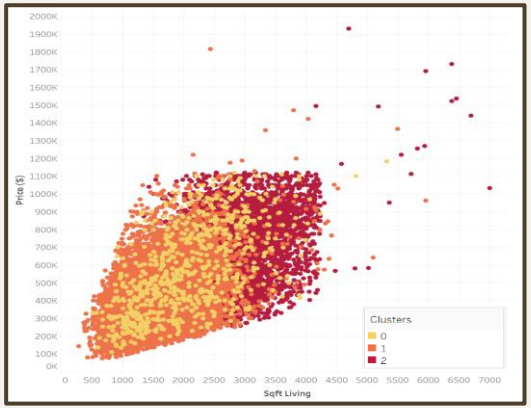
Key observations:

1. The positive slope supports the hypothesis that larger living spaces generally lead to higher house prices.
2. An R^2 score of 0.38 suggests a weak to moderate relationship, with living space accounting for only 38% of the variation in prices.

The scattered data points around the regression line indicate that the relationship is not entirely linear. Since linear regression does not fully capture the data, exploring alternative modeling approaches is recommended.

Analytic Process: Cluster Analysis





Elbow technique and k-means method were used for this **unsupervised machine learning analysis**.
The ideal number of clusters is where the "elbow" appears where the score's improvement slows. In this chart, the elbow is at 3 clusters, testing with clustering with 3 clusters is appropriate as a starting point.

K-means predicts the 'clusters' column (from the standardized dataset) to the original dataset. This step links each cluster to the original dataset for better representation. It will be useful in the final analysis to calculate the real average and median values of each variable, providing meaningful statistical insights.

Cluster 2 (Dark red)

This cluster stands out with the highest statistics across almost all variables. It represents luxurious houses with expensive prices, located in affluent neighborhoods and offering full amenities.

Cluster 1 (Orange)

This cluster clearly has the lowest statistical values across most variables. It indicates affordable houses with the lowest prices, older properties, and fewer amenities, likely located in less wealthy neighborhoods.

Cluster 0 (Yellow)

This cluster represents mid-range houses, with most variables falling at moderate levels. However, the analysis suggests that this cluster skews toward the lower and more affordable house category, despite being classified as mid-range.

Summary and Recommendations

1. Linear Regression:

Applying linear regression in machine learning helps identify significant relationships between variables in the dataset. From this method, we observed that increasing the number of square feet (sqft_living) generally leads to higher house prices. However, the regression output indicates that the relationship is not entirely linear. The data points scattered around the regression line suggest that other factors also influence house prices. Exploring additional techniques such as clustering analysis can provide deeper insights.

2. Clustering Analysis:

Clustering analysis allows us to uncover insights that may be overlooked with linear regression. This method treats all variables equally without introducing bias. By applying this machine learning algorithm, we identified three distinct groups of houses: luxurious, mid-range, and affordable.

3. Key Insights:

- Houses with the highest price per square foot are predominantly located in densely populated areas, such as Seattle and Bellevue.
- Square footage (sqft_living) has the strongest influence on house prices, while other factors, such as the number of bedrooms/bathrooms, property condition, and year built, have a relatively minor impact.

4. Next Steps:

- Update the Dataset: Collect recent home sales data to analyze price trends over time and validate current market insights.
- Incorporate Additional Features: Examine the impact of additional factors, such as proximity to amenities, school districts, and neighborhood crime rates, on house prices.

THANK YOU

Connect with me

Email

Tableau

GitHub