Data Analyst Portfolio

Chenhao Ma

Self-introduction

Contact Me

GitHub

My name is Chenhao Ma I have previously worked in financial analysis and product management, where I developed strong communication, coordination, and problem-solving skills.

Data analysis is a completely new field for me, and CareerFoundry is an excellent platform that teaches me basic skills and provides real project experience. In this process, starting from clarifying project requirements, to searching for data, cleaning data, analyzing data, generating insights and recommendations, and sharing them with others through presentations, it is challenging but also extremely fulfilling and rewarding. I am also very much looking forward to starting a new career as a data analyst in the future.

Project

GameCo

Analysis of global video game sales

Influenza Season

Preparing medical staffing for upcoming Influenza season

Rockbuster Stealth

Optimizing Business operations through SQL analysis

Instacart

Uncovering information about sales patterns through Python analysis

Pig E. Bank

Predictive analysis of customer retention

KingCounty House Price

Build a housing price prediction model using Python.

01

GameCo Marketing Analysis

Business requirements

Conduct data analysis for GameCo(video game company) to assist them in preparing their market budget for 2017. They want to allocate their marketing budgets across geographies regions to maximize return on investment.

Analysis Process

- Use historical data to see how the share of sales by region changes over time
- For each region, analyze whether other variables such as Platform and Genre can be used to optimize the market budget

Please refer to the link for the completed analysis report.
Final present

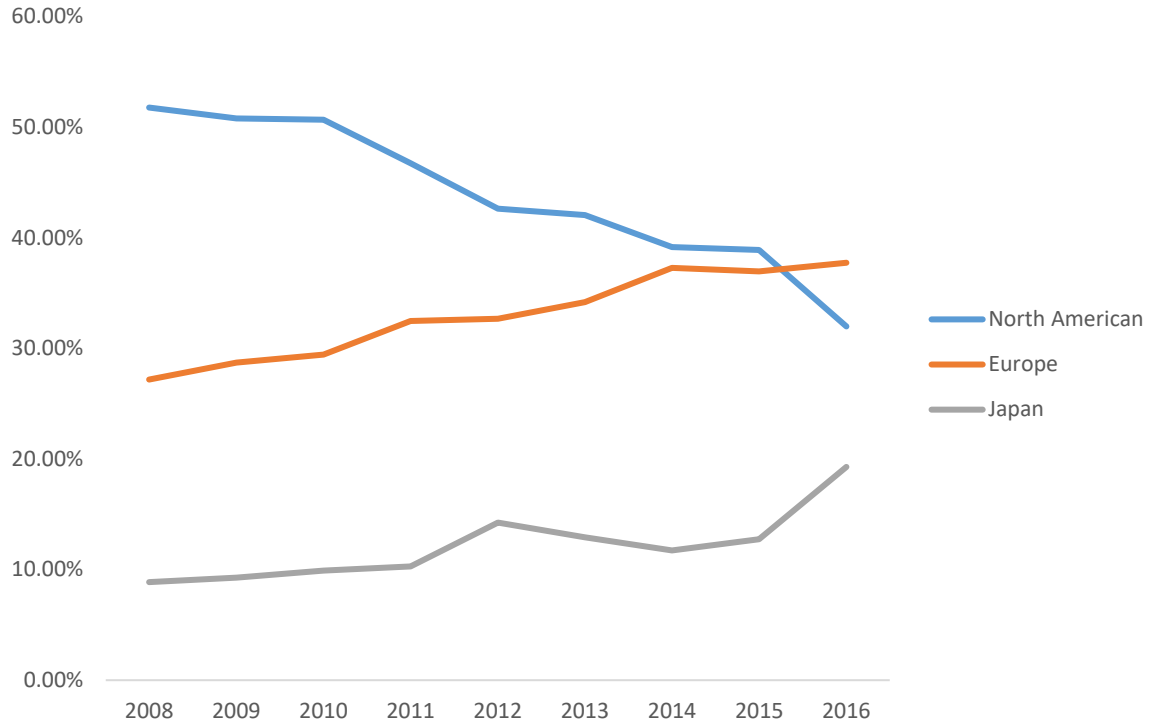
Skill

- ✓ Data Cleaning
- ✓ Data Grouping and Summarizing
- ✓ Descriptive Analysis
- ✓ Pivot Table
- ✓ Visualization Charts in MS Excel/PowerPoint

Tool

- ✓ Excel
- ✓ Power Point

Proportion of sales revenue by region from 2008 to 2016



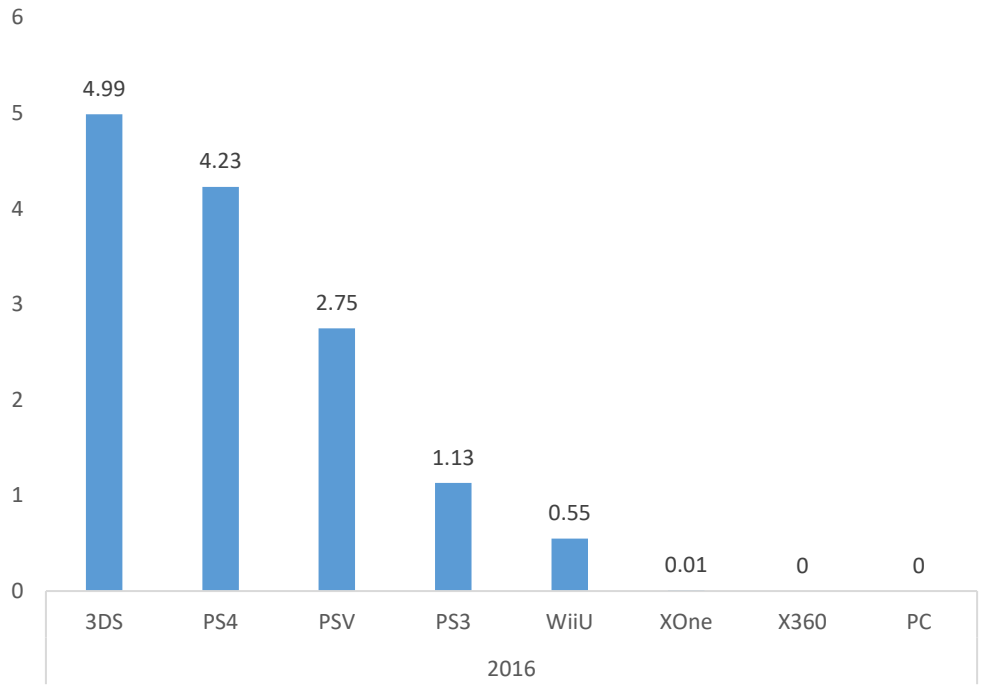
- Use line charts to conduct time series analysis of historical data for each region, understanding the sales trends in different areas to make better predictions for 2017.

2013-2016 Change in market share by regions

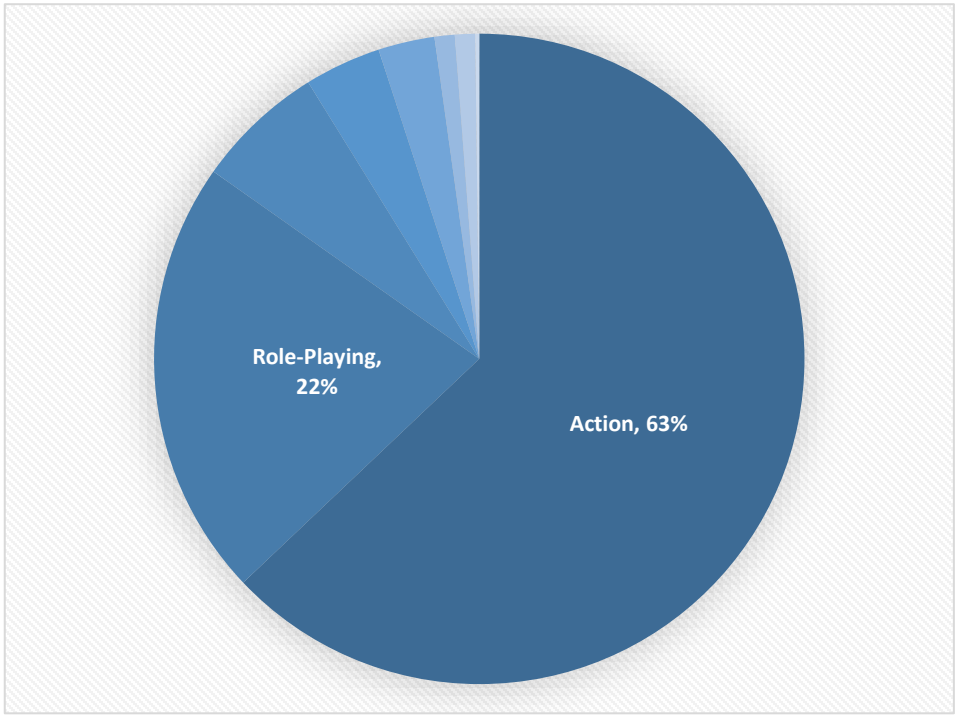
	North American	Europe	Japan
2013-2014	-6.89%	9.09%	-9.39%
2014-2015	-0.68%	-0.87%	8.80%
2015-2016	-17.80%	2.15%	51.27%
AVG	-8.46%	3.45%	16.90%

- Use tables to provide a more detailed listing of specific data.

Japan market sales by Platform in 2016.



The share of 3DS platform sales by Genre in the Japan market 2016



- Identify Platform and Genre as the variables for differentiated marketing in various regions.
- Use bar charts to analyze the top-selling Platform and Genre in each region to determine the marketing focus for each area.

- Use pie charts to analyze the proportion of different types of Genre within the high-selling Platform categories, in order to determine marketing strategies targeted at specific Platform

Using the Japanese market as an example

Insight:

1. The Japanese market has maintained the third largest market share, and in 2016 it got tremendous growth.
2. The Japan market is relatively concentrated in terms of genre, Action, Role-Playing are the top two sales in all Genre, and the sum of the two accounts for 68%.
3. The Japan market is relatively fragmented in terms of platform.
4. Action is favored in the 3DS platform, with a 63% share

Recommendation

1. It is possible to appropriately increase the sales budget for the Japanese market and observe the market trends to determine whether the high growth rate is sustainable.
2. In terms of categories, Action, Role-Playing should enjoy budget priority, In terms of platforms, it should be focused on 3DS, PS4. Bundles of the 3DS platform and the Action can also be considered.

02

Preparing for Influenza season

Business requirements

The United States has an influenza season where more people than usual suffer from the flu. Hospitals and clinics need additional staff to adequately treat extra patients. The medical staffing agency provides this temporary staff. As a data analyst, you will help determine when to send staff and how many to each state.

Analysis Process

- Analyze who is most affected by the Influenza season.
- Analyze where the influenza more severe.
- Analyze when the influenza season is.

Please refer to the link for the completed analysis report.

Tableau Public

Skill

- ✓ Data cleaning, integration & transformation.
- ✓ Statistical hypothesis testing
- ✓ Visual analysis
- ✓ Storytelling in Tableau.

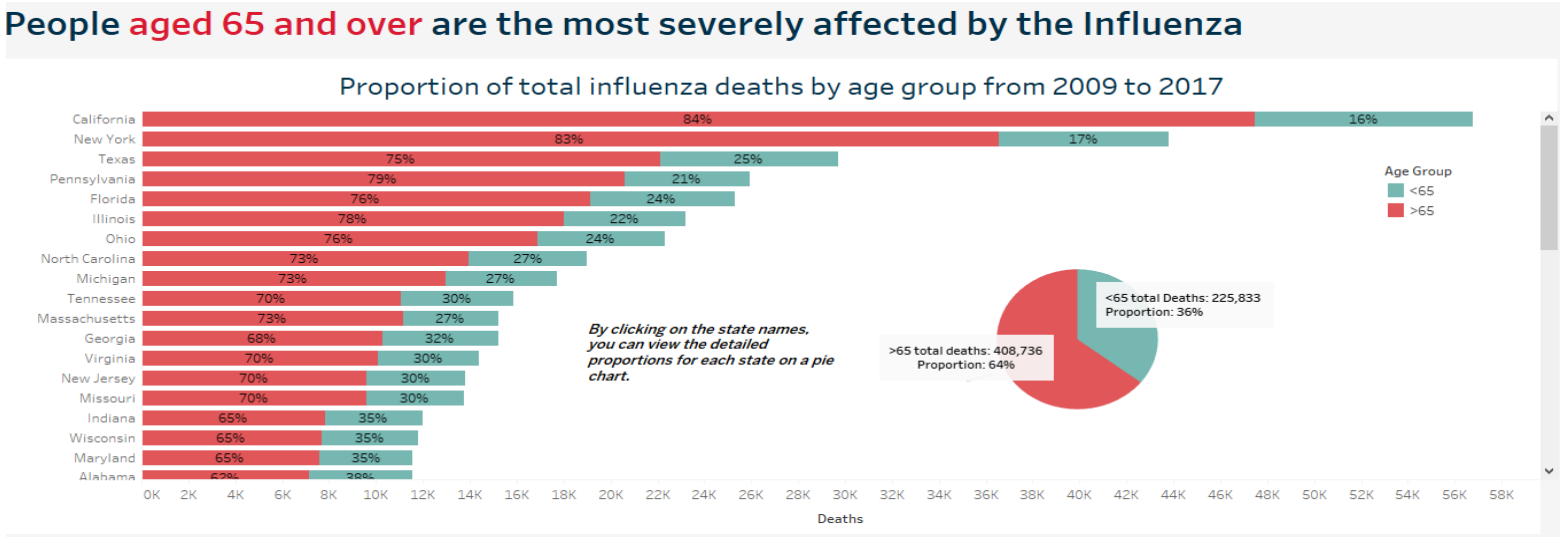
Tool

✓ Tableau ✓ Excel ✓ Power Point

Influenza

Analysis

Research hypothesis	If a patient is over 65 years, then they are more likely to die from influenza.
Independent variable	Age group
Dependent variable	Influenza death rate
Null hypothesis (H_0)	The influenza death rate of patients 65 years or older is less than or equal to the influenza death rate of patients younger than 65 years old.
Alternative hypothesis (H_A)	The influenza death rate of patients 65 years or older is greater than the influenza death rate of patients younger than 65 years old.
Test type	One-tailed
Significance level	0.05



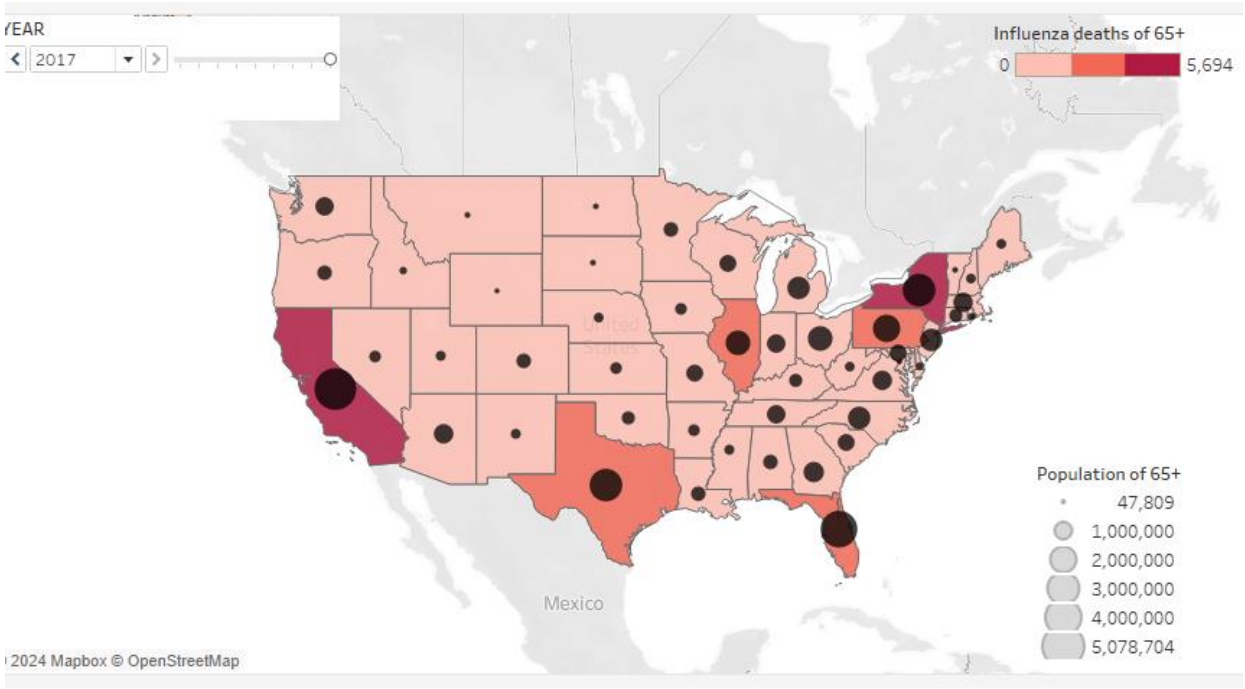
First, use a T-test to verify that the flu mortality rate for the population aged 65 and over is higher than other age groups.

- The result of T-test is that T-Stat is 40.9 and P-value is close to 0 which is less than significance level (0.05).
- The conclusion is we reject the null hypothesis and accept the alternative hypothesis.

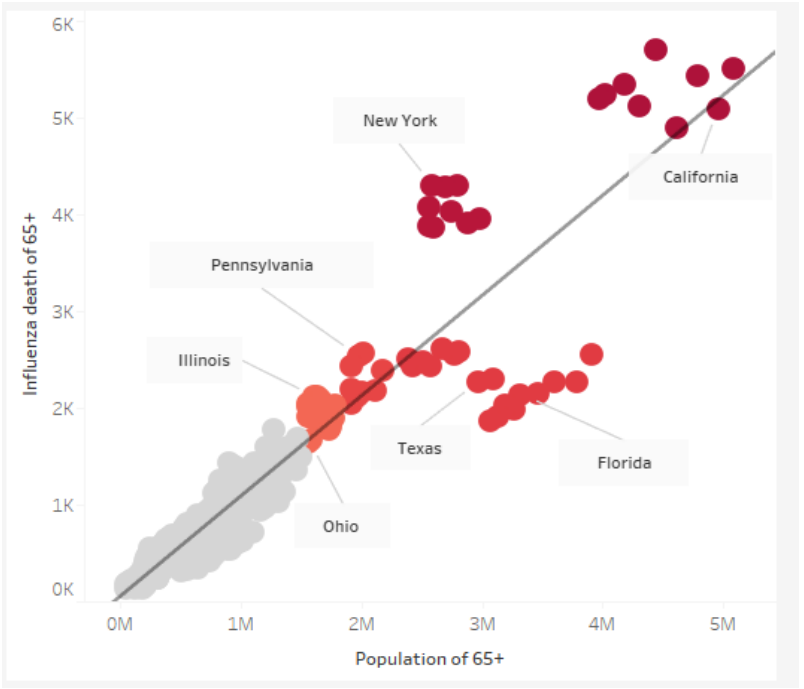
Then, use Tableau to combine bar charts and pie charts to analyze the number of flu deaths in each state and the proportion of deaths among those aged 65 and over. And we can see People aged 65 and over are the most severely affected by the Influenza.

I n f l u e n z a

Analysis



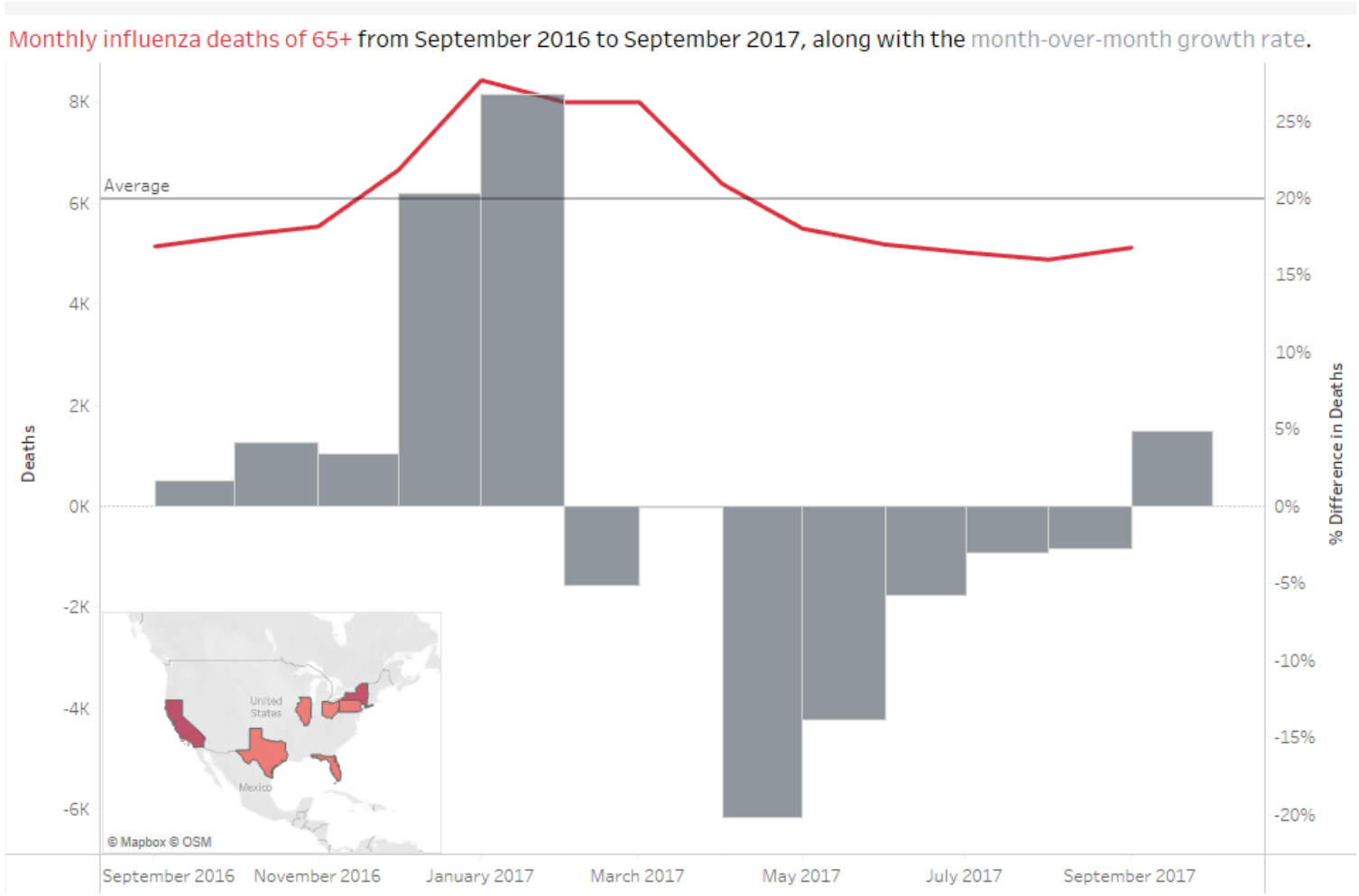
- Use different colors and shapes on the map to represent the total population and the number of flu deaths in each state. And we can see the flu in California, New York, Texas, Pennsylvania, Florida, Illinois and Ohio is more severe



- Use a scatter plot to reveal a strong positive correlation between the number of deaths and the total population.

I n f l u e n z a

Analysis



- Combine line charts and bar charts to analyze the temporal trends of influenza. This approach allows us to clearly identify the number of flu deaths each month and the monthly growth rate of flu deaths. From this analysis, we can determine when flu deaths begin to surge, when they peak, and when they start to decline.
- By using Tableau's interactive features, we can click on high-risk flu states on the map to view the varying flu trends for each state.

Conclusion

1.The proportion of influenza deaths among the population aged over 65 exceeds 90%, making them the most affected group by influenza.

2.Analysis has found that the population of over 65 and the number of deaths in this age group are strongly correlated. States with a large population of over 65 should be where more medical resources are allocated.

3.Analysis has revealed that influenza is seasonal, with the number of deaths typically beginning to rise in October and reaching a peak in January or March, then starting to decline, with the lowest point in August or September.

Recommendation：

1. Geographic dimension: Divide the United States into three priority levels and allocate medical resources based on these levels.

2. Time dimension: Allocate medical resources appropriately based on the peak flu season times in each state.

03

Rockbuster Stealth LLC

Business requirements

Rockbuster Stealth LLC is a movie rental company that used to have stores around the world. The Rockbuster Stealth management team is planning to use its existing movie licenses to launch an online video rental service in order to stay competitive.

Analysis Process

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

Please refer to the link for the completed analysis report:
Final Report

Skill

- ✓ Relational databases
- ✓ SQL
- ✓ Database querying, filtering, cleaning & summarizing
- ✓ Subqueries
- ✓ Common table expression

Tool

✓ SQL ✓ Tableau ✓ Excel ✓ Power Point


Create a data dictionary that clearly displays the contents of each dataset and their connections.

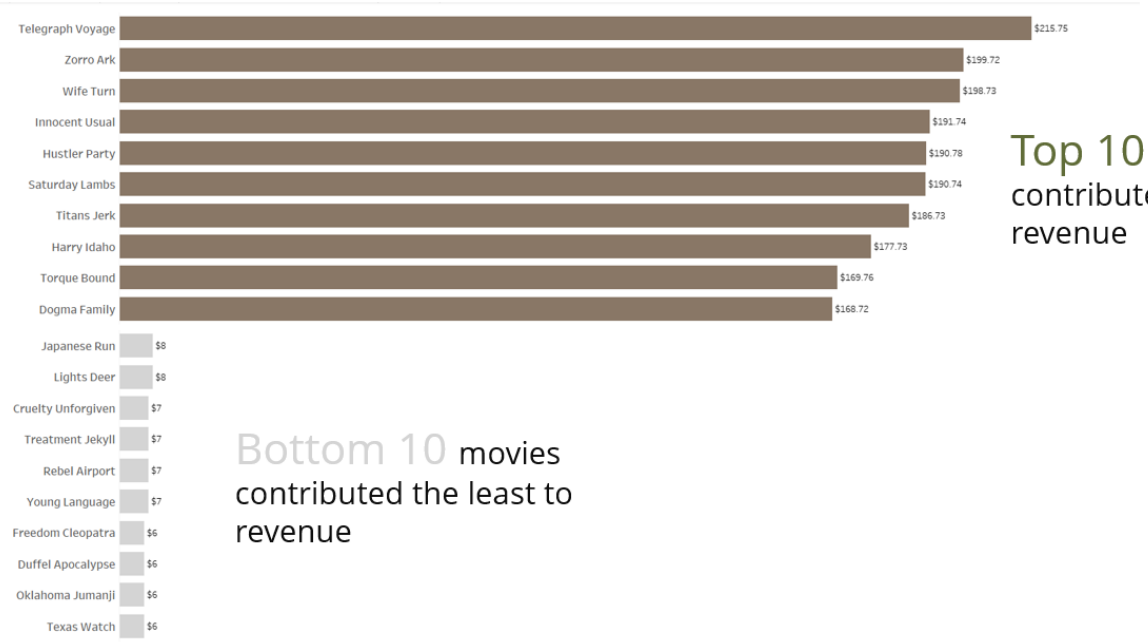
Please refer to the link:
Rockbuster_Dictionary

```
SELECT title, COUNT(*) FROM film GROUP BY title;
SELECT release_year, COUNT(*) FROM film GROUP BY release_year;
SELECT language_id, COUNT(*) FROM film GROUP BY language_id;
SELECT rental_duration, COUNT(*) FROM film GROUP BY rental_duration;
SELECT rental_rate, COUNT(*) FROM film GROUP BY rental_rate;
SELECT length, COUNT(*) FROM film GROUP BY length;
SELECT replacement_cost, COUNT(*) FROM film GROUP BY replacement_cost;
SELECT rating, COUNT(*) FROM film GROUP BY rating;

SELECT store_id, COUNT(*) FROM customer GROUP BY store_id;
SELECT first_name, COUNT(*) FROM customer GROUP BY first_name;
```

```
SELECT R.customer_id,A.address_id,CI.city_id,CO.country_id FROM rental AS R
INNER JOIN customer AS C ON C.customer_id = R.customer_id
INNER JOIN address AS A ON A.address_id = C.address_id
INNER JOIN city AS CI ON CI.city_id = A.city_id
INNER JOIN country CO ON CO.country_id = CI.country_id;

SELECT CO.country, COUNT(DISTINCT R.customer_id) AS number_of_customers, COUNT(R.customer_id) AS rental_frenquency FROM rental AS R
INNER JOIN customer AS C ON C.customer_id = R.customer_id
INNER JOIN address AS A ON A.address_id = C.address_id
INNER JOIN city AS CI ON CI.city_id = A.city_id
INNER JOIN country CO ON CO.country_id = CI.country_id
GROUP BY CO.country
ORDER BY number_of_customers DESC
LIMIT 10;
```

```
SELECT AVG(total_amount_payment) AS average FROM
(   SELECT C.customer_id, C.first_name, C.last_name, CO.country, CI.city, SUM(P.amount) AS total_amount_payment FROM payment AS P
    INNER JOIN customer AS C ON C.customer_id =P.customer_id
    INNER JOIN address AS A ON A.address_id = C.address_id
    INNER JOIN city AS CI ON CI.city_id = A.city_id
    INNER JOIN country CO ON CO.country_id = CI.country_id
    WHERE CI.city IN (
        SELECT CI.city FROM customer AS C
        INNER JOIN address AS A ON A.address_id = C.address_id
        INNER JOIN city AS CI ON CI.city_id = A.city_id
        INNER JOIN country CO ON CO.country_id = CI.country_id
        WHERE CO.country IN (
            SELECT CO.country FROM customer AS C
            INNER JOIN address AS A ON A.address_id = C.address_id
            INNER JOIN city AS CI ON CI.city_id = A.city_id
            INNER JOIN country CO ON CO.country_id = CI.country_id
            GROUP BY CO.country
            ORDER BY COUNT(C.customer_id) DESC
            LIMIT 10)
    GROUP BY CO.country,CI.city
    ORDER BY Count(C.customer_id) DESC
    LIMIT 10)
    GROUP BY C.customer_id,CO.country, CI.city
    ORDER BY total_amount_payment DESC
    LIMIT 5) AS total_amount_paid;
```

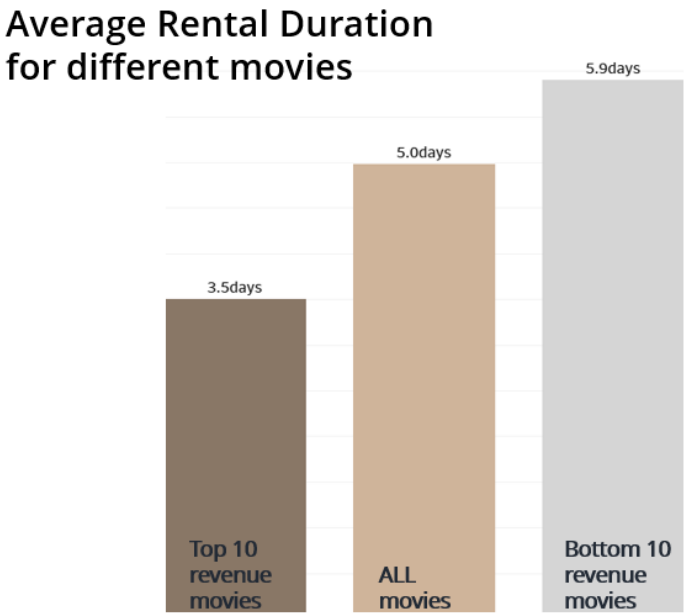
Clean and integrate the data using SQL.
Please refer to the link:
SQL Code



Top 10 movies contributed the most to revenue

Bottom 10 movies contributed the least to revenue

Use SQL to filter and select the top 10 and bottom 10 movies, then display them using Tableau.



Use SQL to calculate the Average Rental Duration by category and display it using a bar chart.

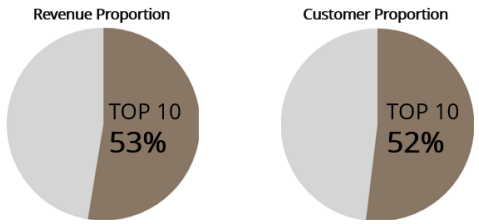
ROCKBUSTER
Analysis

WHERE
Top 10 countries

Top 10 countries with high revenue & customer

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

Top 10 countries proportion of revenue & customer



The revenue and customer proportion of the TOP 10 countries are both **more than 50%** of the total, making it a very important market.

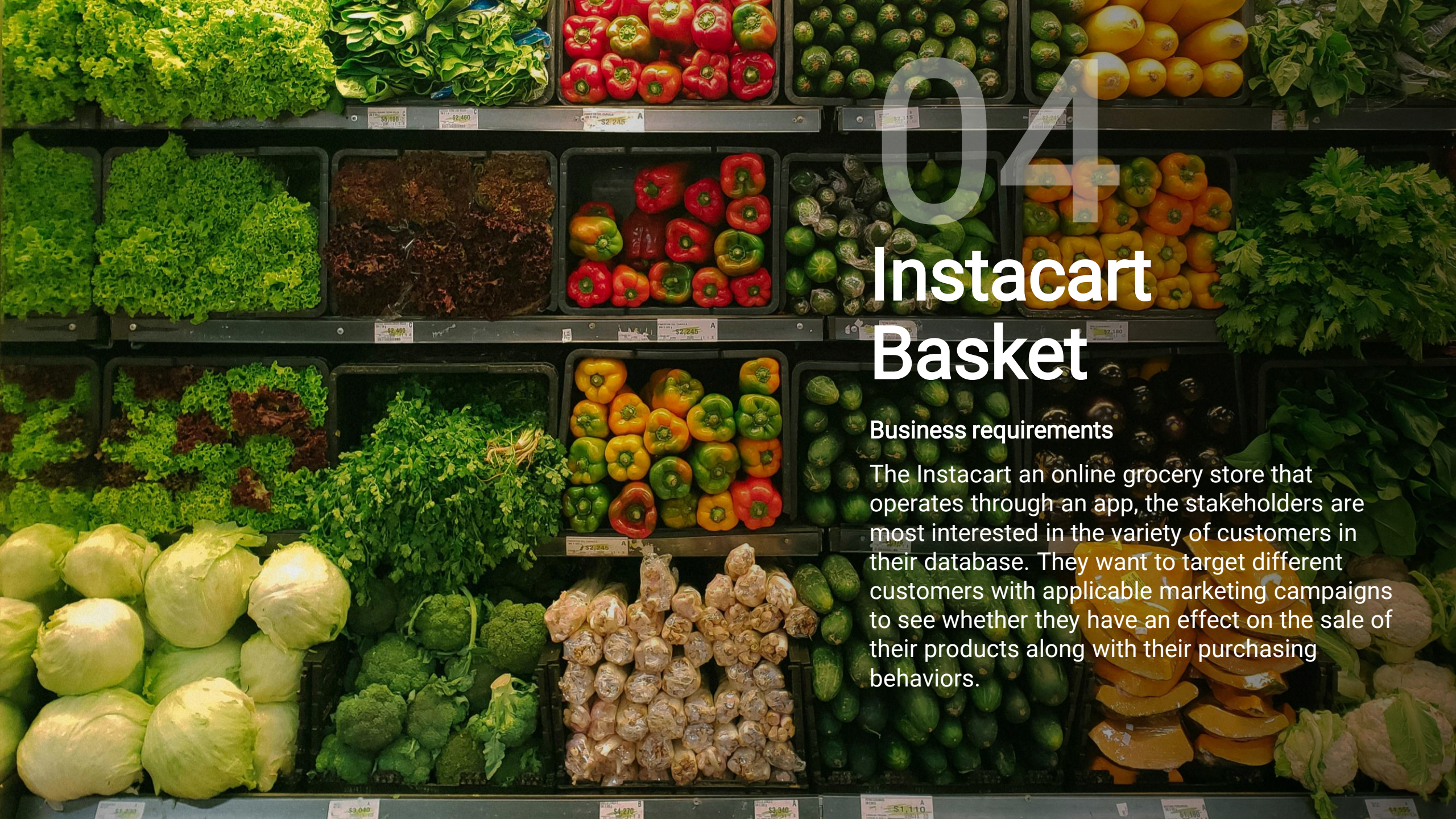
Use SQL to filter and calculate the top 10 countries with high revenue and customer numbers, then use pie charts to display their revenue and customer proportions of total.

WHO
Who are customers with a high lifetime value based?

Top 10 highest revenue-generating customers distribution

First Name	Last Name	City	Country	Revenue
Eleanor	Hunt	Saint-Denis	Runion	211.55
Karl	Seal	Cape Coral	United States	208.58
Marion	Snyder	Santa Brbara doeste	Brazil	194.61
Rhonda	Kennedy	Apeldoorn	Netherlands	191.62
Clara	Shaw	Molodetno	Belarus	189.6
Tommy	Collazo	Qomsheh	Iran	183.63
Ana	Bradley	Memphis	United States	167.67
Curtis	Irby	Richmond Hill	Canada	167.62
Marcia	Dean	Tanza	Philippines	166.61
Mike	Way	Valparai	India	162.67

Use SQL to filter the top 10 highest revenue-generating customers and display their distribution using a map.



04 Instacart Basket

Business requirements

The Instacart an online grocery store that operates through an app, the stakeholders are most interested in the variety of customers in their database. They want to target different customers with applicable marketing campaigns to see whether they have an effect on the sale of their products along with their purchasing behaviors.

Analysis Process

- Clean and integrate the dataset.
- Analyze the busiest days of the week and hours of the day?
- Categorize products based on their prices.
- Analyze the frequency of product orders of different product departments?
- Analyze Do sales figures vary between geographic regions?
- Analyze the different types of customers in their system and how their ordering behaviors differ?

Please refer to the link for the complete analysis report.
Final Report

Skill

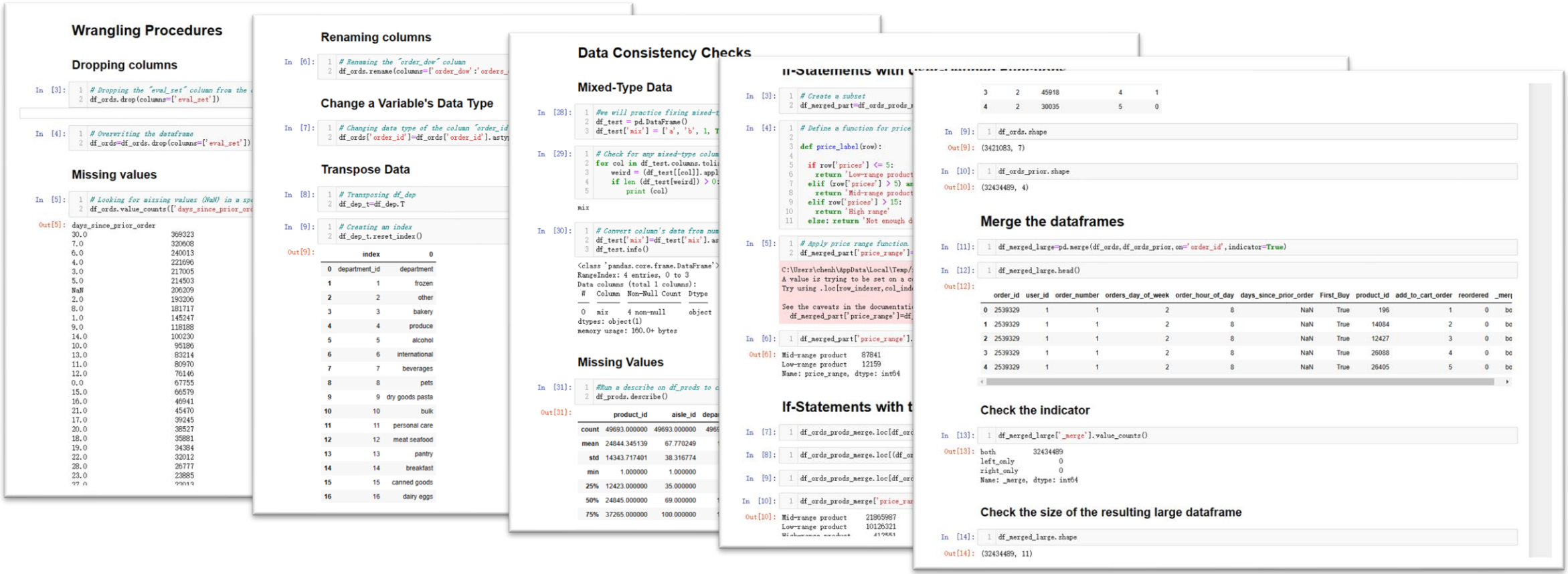
- ✓ Python
- ✓ Data wrangling and merging
- ✓ Deriving variables
- ✓ Grouping datasets
- ✓ Aggregating data
- ✓ Reporting in Excel

Tool

✓ Python ✓ Tableau ✓ Excel ✓ Power Point

Instacart

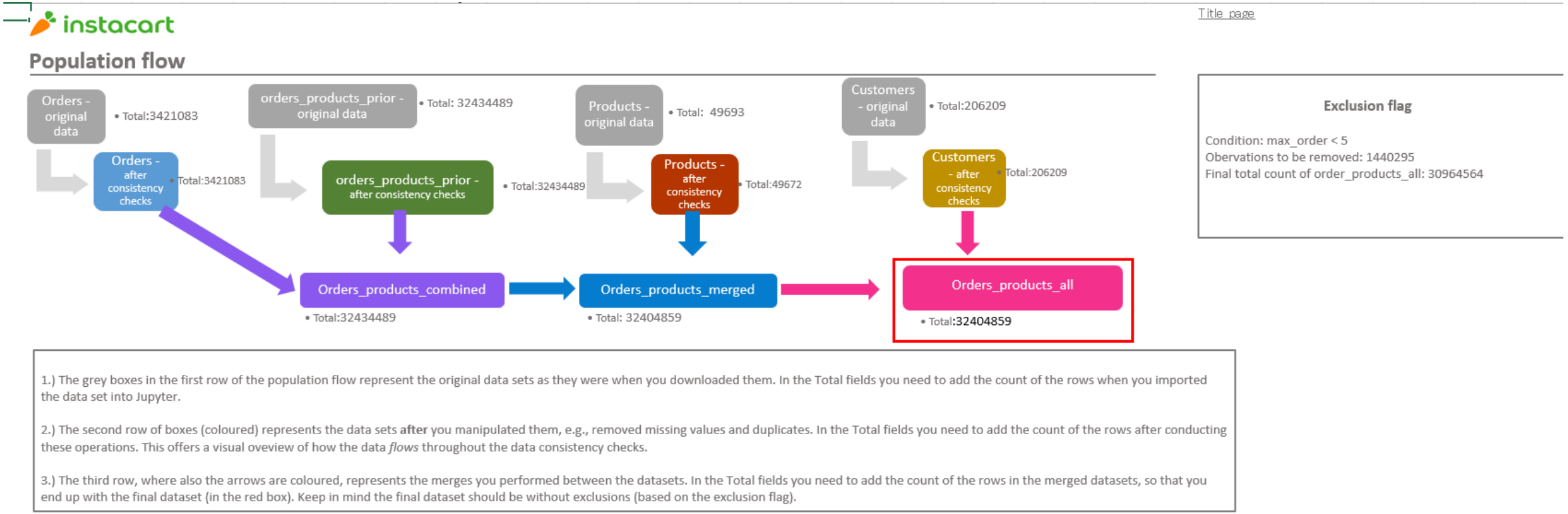
Analysis



Use Python for data cleaning, merging, deriving variables, and grouping datasets.

For detailed Python execution, please refer to the link: Python_code

Create a **Population flow** to make the data cleaning process clearer.



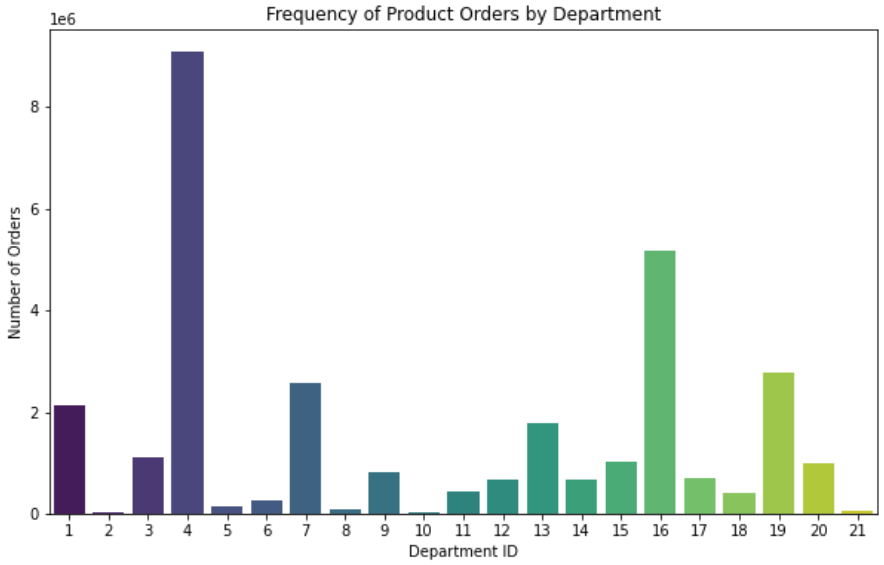
1.) The grey boxes in the first row of the population flow represent the original data sets as they were when you downloaded them. In the Total fields you need to add the count of the rows when you imported the data set into Jupyter.

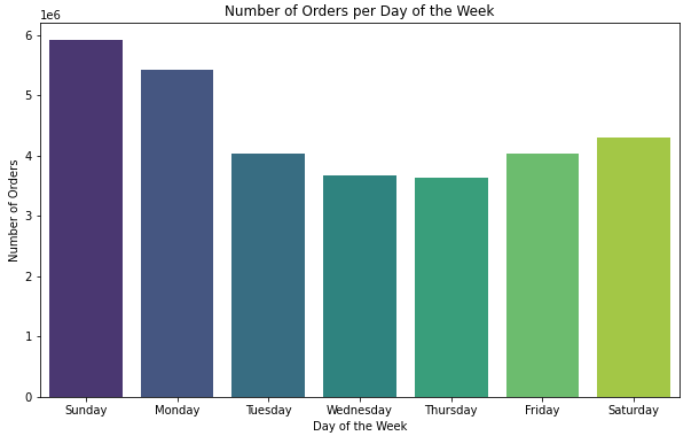
2.) The second row of boxes (coloured) represents the data sets **after** you manipulated them, e.g., removed missing values and duplicates. In the Total fields you need to add the count of the rows after conducting these operations. This offers a visual oveview of how the data *flows* throughout the data consistency checks.

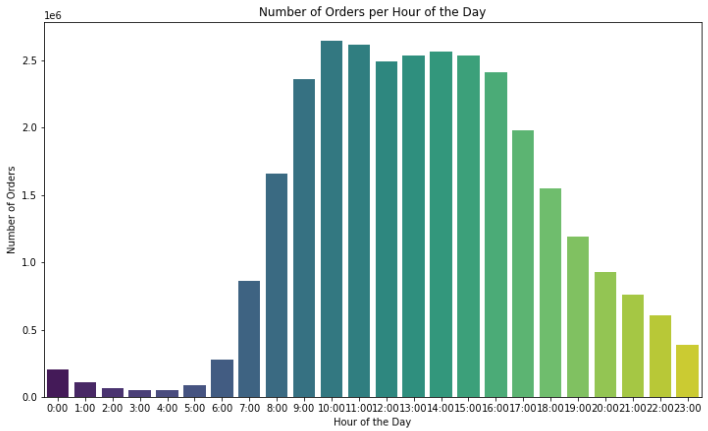
3.) The third row, where also the arrows are coloured, represents the merges you performed between the datasets. In the Total fields you need to add the count of the rows in the merged datasets, so that you end up with the final dataset (in the red box). Keep in mind the final dataset should be without exclusions (based on the exclusion flag).

Categorize products into three classes based on price ranges and use a bar chart to display the order volume for each class.

Additionally, use a bar chart to display the order volume for products from each department.

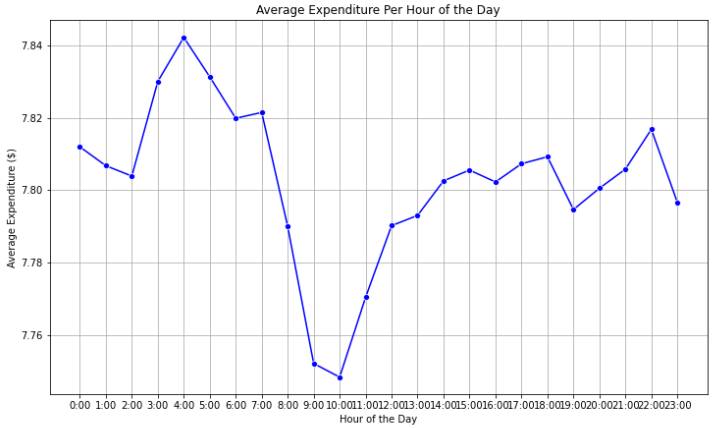


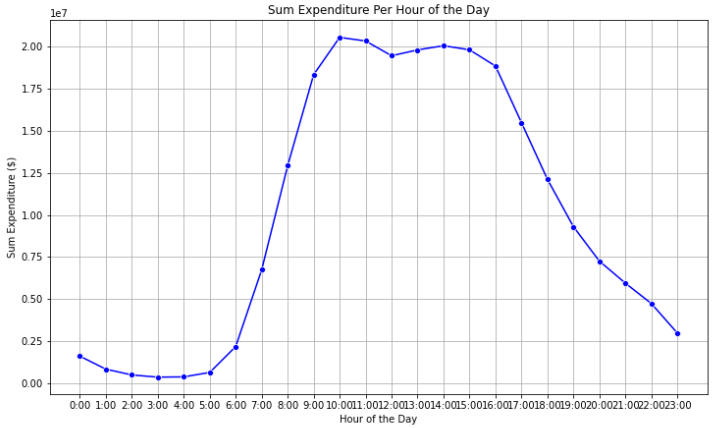




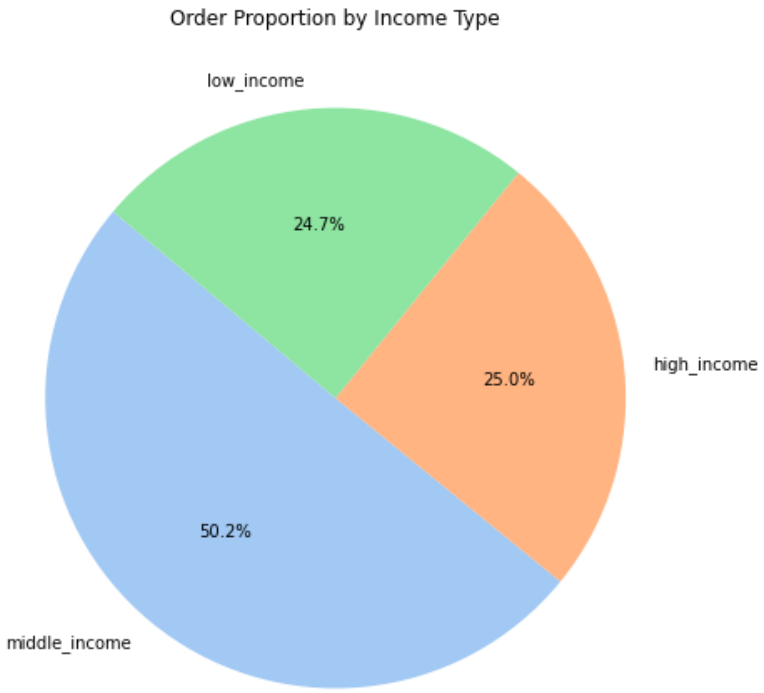
Use bar charts to analyze the order volume for different time periods.

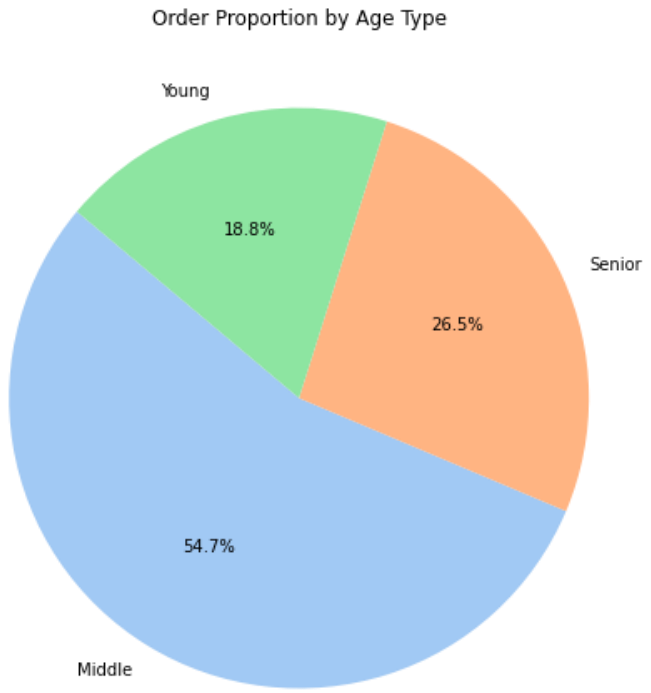
Use line charts to analyze the average and total spending for different time periods.

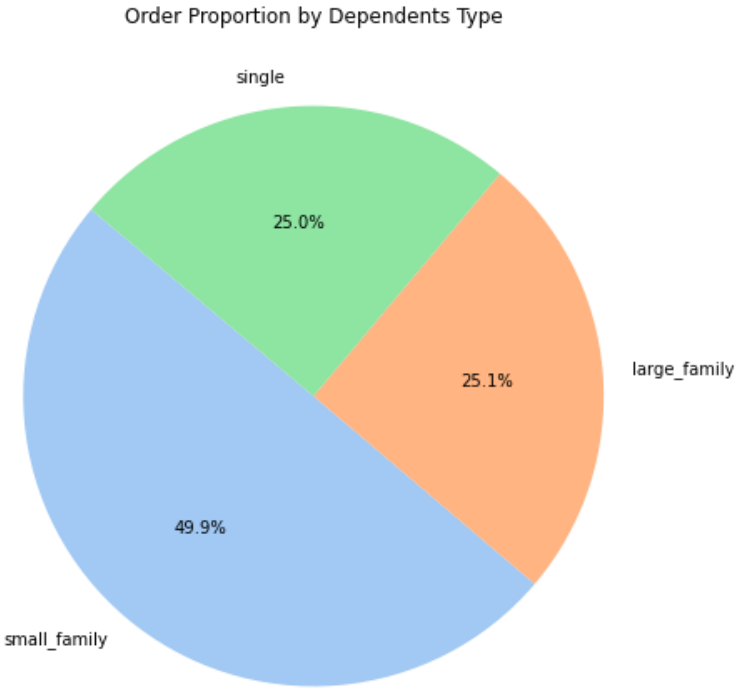




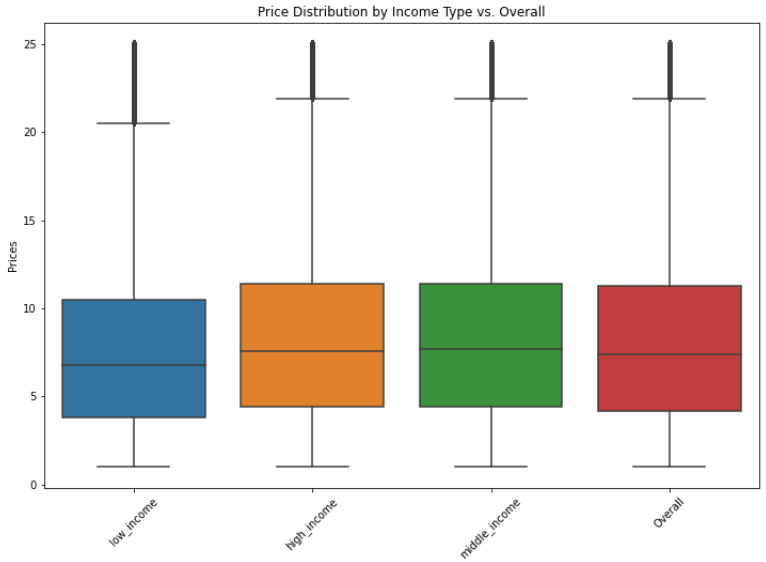
Segment users by region, loyalty, age, income level, and family size.

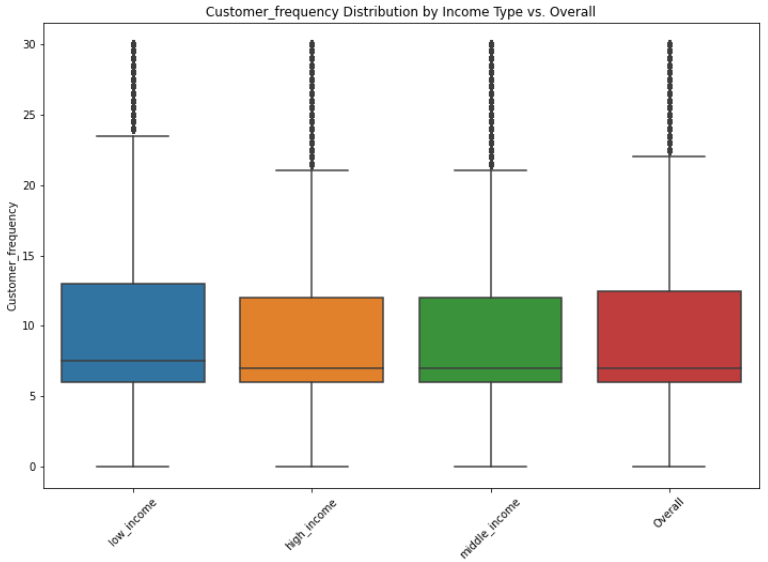


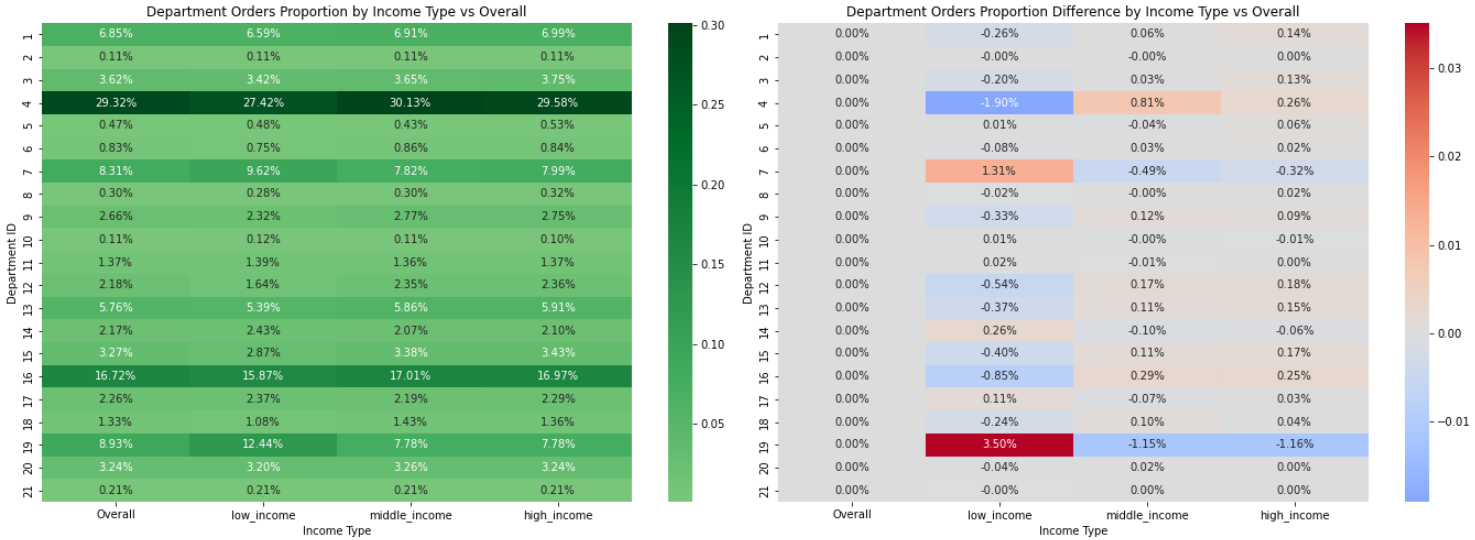


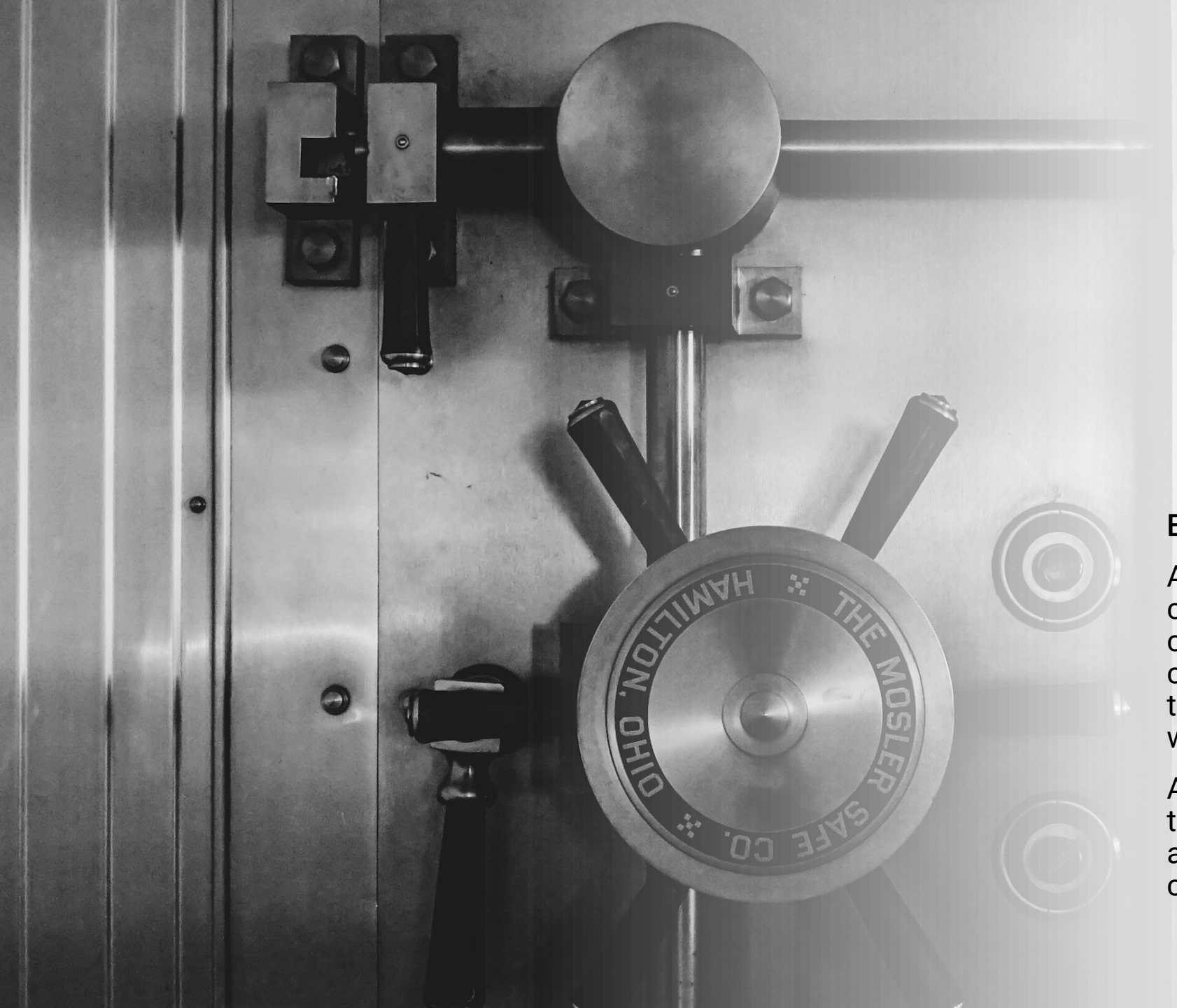


Analyze user consumption behavior based on order quantity proportion, purchase frequency distribution, spending price distribution, and preferences for different product departments.









05

Pig E. Bank

Business requirements

Analytical support will be provided to the company's anti-money-laundering compliance department. This entails conducting various datarelated projects to assess client and transaction risk, while also reporting on relevant metrics.

Additionally, contributions will be made to building and optimizing models aimed at enhancing the efficiency of the bank's compliance program.

Analysis Process

- Clean and integrate the dataset.
- Analyze the retention rates for different categories of users.
- Analyze which variables have the greatest impact on retention rates.
- To build a decision tree

Please refer to the link for the complete analysis report.

Final Report

Skill

- ✓ Proficiency in handling Big Date
- ✓ Understanding of Data Ethics
- ✓ Expertise in Data Mining
- ✓ Capability in Predictive Analysis
- ✓ Competence in Time Series Analysis and Forecasting
- ✓ Proficient use of GitHub

Tool

✓ GitHub ✓ Excel ✓ Power Point

Credit Score Evaluation	IF exited bank		
	0	1	Total
376-533	13.47%	16.67%	14.13%
534-691	50.06%	52.45%	50.55%
692-850	36.47%	30.88%	35.32%
Total	100.00%	100.00%	100.00%

Using pivot table to analyze the retention rates for different categories of users.

Country Evaluation	IF exited bank		
	0	1	Total
France	51.21%	37.75%	48.44%
Germany	23.13%	36.76%	25.93%
Spain	25.67%	25.49%	25.63%
Total	100.00%	100.00%	100.00%

	Current Customers	Exited Customers	Total Customers
	Mean		
Credit Score	652	637	649
Age	38	45	39
Tenure	5	5	5
Balance	$74,830.87	$90,239.22	$78,002.72
Number of Products	2	1	2
Estimated Salary	$98,942.45	$97,155.20	$98,574.54

Analyze the descriptive statistics for retained and departed customers.

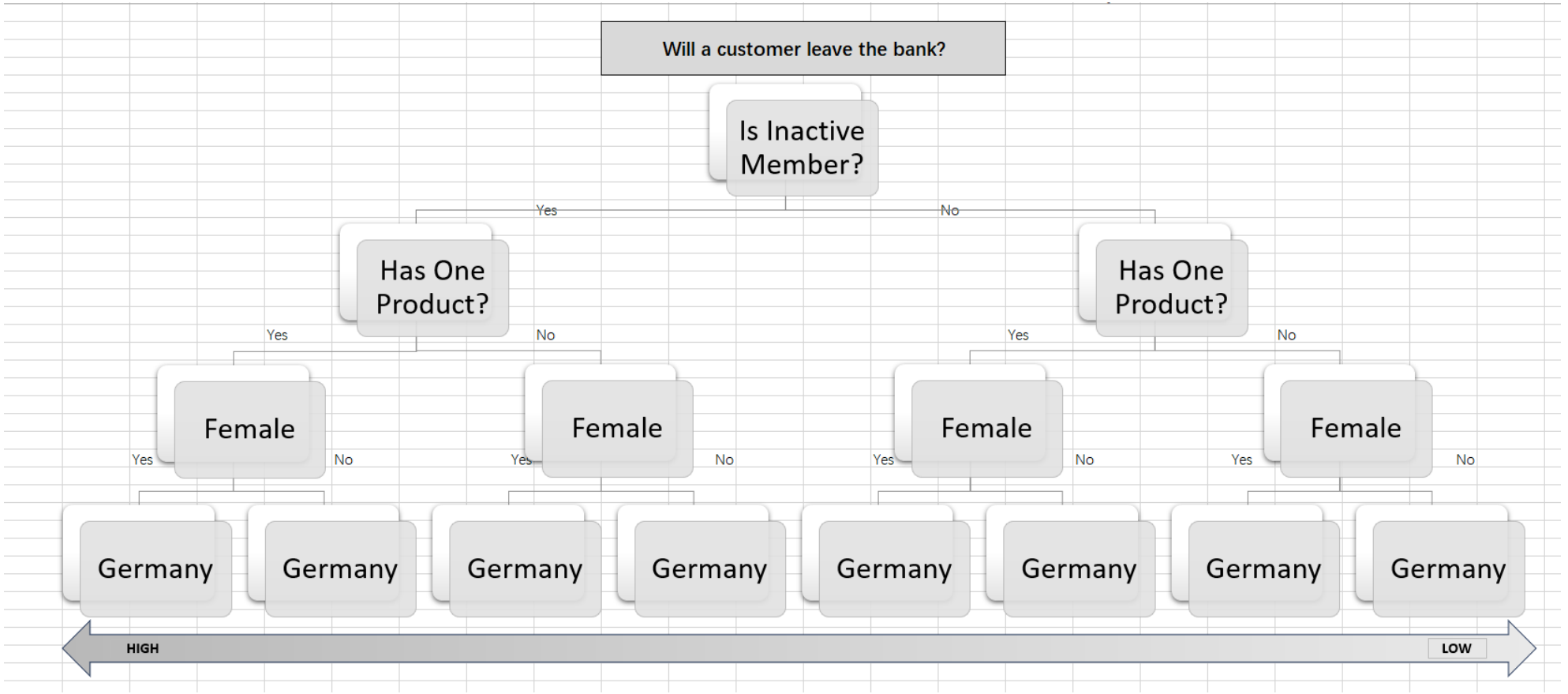
Analysis

Item		Current Customers	Exited Customers	Total Customers	Difference between Exited customers and the total
Credit Score Evaluation	376-533	13.47%	16.67%	14.13%	2.54%
	534-691	50.06%	52.45%	50.55%	1.90%
	692-850	36.47%	30.88%	35.32%	-4.44%
Country Evaluation	France	51.21%	37.75%	48.44%	-10.69%
	Germany	23.13%	36.76%	25.93%	10.83%
	Spain	25.67%	25.49%	25.63%	-0.14%
Gender Evaluation	Female	43.33%	59.31%	46.62%	12.69%
	Male	56.67%	40.69%	53.38%	-12.69%
Tenure Evaluation	0-4	42.95%	49.02%	44.20%	4.82%
	5-10	57.05%	50.98%	55.80%	-4.82%
Balance Evaluation	0-60000	39.39%	29.41%	37.34%	-7.92%
	60000-120000	28.84%	31.86%	29.47%	2.40%
	120000-180000	30.24%	35.78%	31.38%	4.40%
	180000-240000	1.52%	2.94%	1.82%	1.12%
Num of Products Evaluation	1	46.76%	69.61%	51.46%	18.14%
	2	52.60%	15.69%	45.01%	-29.32%
	3	0.64%	13.73%	3.33%	10.40%
	4	0.00%	0.98%	0.20%	0.78%
Has Credit card Evaluation	0	29.35%	29.41%	29.36%	0.05%
	1	70.65%	70.59%	70.64%	-0.05%
Active Member Evaluation	0	43.84%	70.10%	49.24%	20.85%
	1	56.16%	29.90%	50.76%	-20.85%
Salary Evaluation	371.05-50371.05	24.65%	28.43%	25.43%	3.00%
	50371.05-100371.05	26.94%	21.57%	25.83%	-4.26%
	100371.05-150371.05	26.05%	26.47%	26.14%	0.34%
	150371.05-200371.05	22.36%	23.53%	22.60%	0.93%

Use the difference between Exited Customer and Total in each items for comparison. The four factors with the largest differences will be the top 4 factors that leading factors that contribute to client loss.

They are:
1. If customer is active member.
2.If number of products = 1
3.If customer is female.
4.If customer is from Germany

Build a decision tree



06

King County House Price

Business requirements

We have the 2014-2015 housing sales data for King County. Upon observation, the price fluctuations are significant, with the difference between the highest and lowest prices reaching $7,625,000.

Based on this, we conducted the analysis. We hope it will help homebuyers better understand the price differences in various areas of King County and the main factors affecting housing prices. This should enable homebuyers to have a better assessment of the prices of the houses they are interested in purchasing.

Analysis Process

- Clean and integrate the dataset.
- Geographic Factors Analysis: Analyzing the differences in housing prices across different zip code areas.
- Exploratory Data Analysis, analyze the correlation between each variable and the price.
- Select and build a model to reasonably predict housing prices.

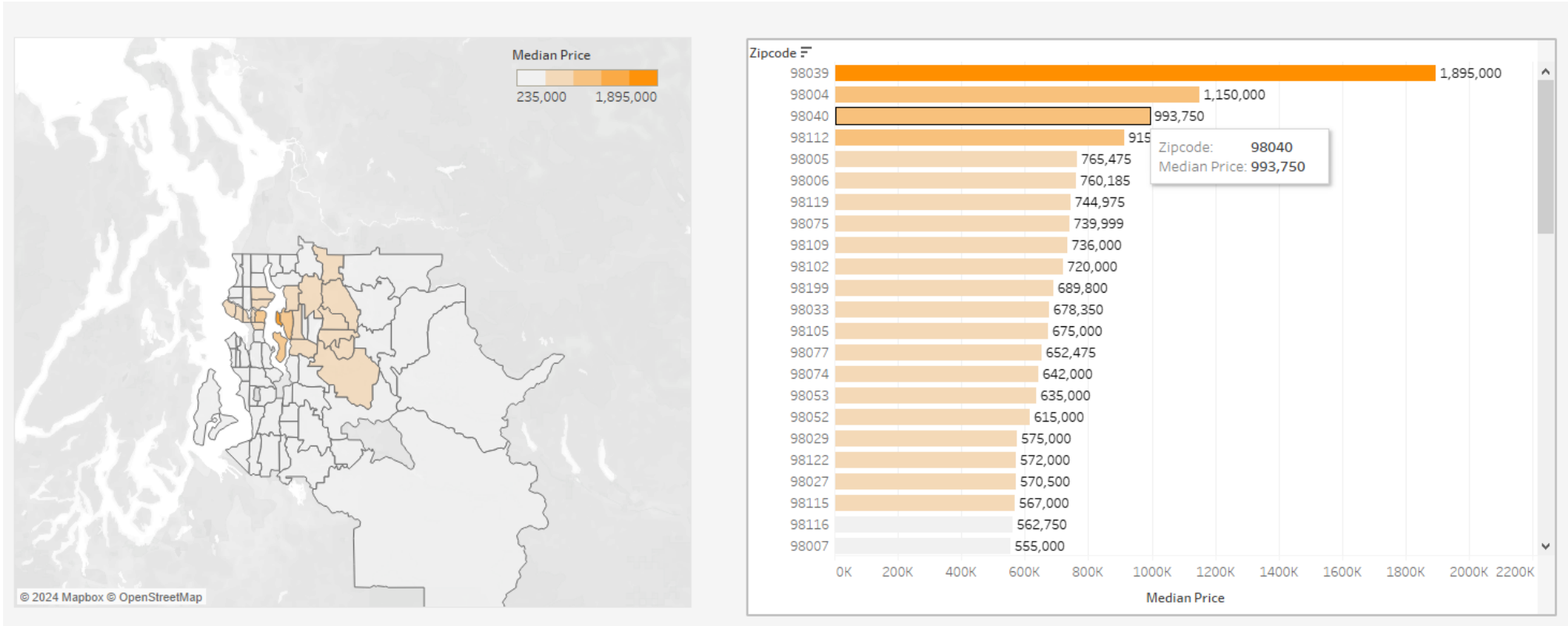
Please refer to the detailed report at the following link:
Tableau Dashboard

Skill

- ✓ Data wrangling and merging
- ✓ Geographic Visualization in Python
- ✓ Machine Learning-Regression
- ✓ Machine-Learning-Clustering
- ✓ Sourcing & Analyzing Time Series Data
- ✓ Select and build a model

Tool

✓ Python ✓ Tableau ✓ Excel ✓ Power Point

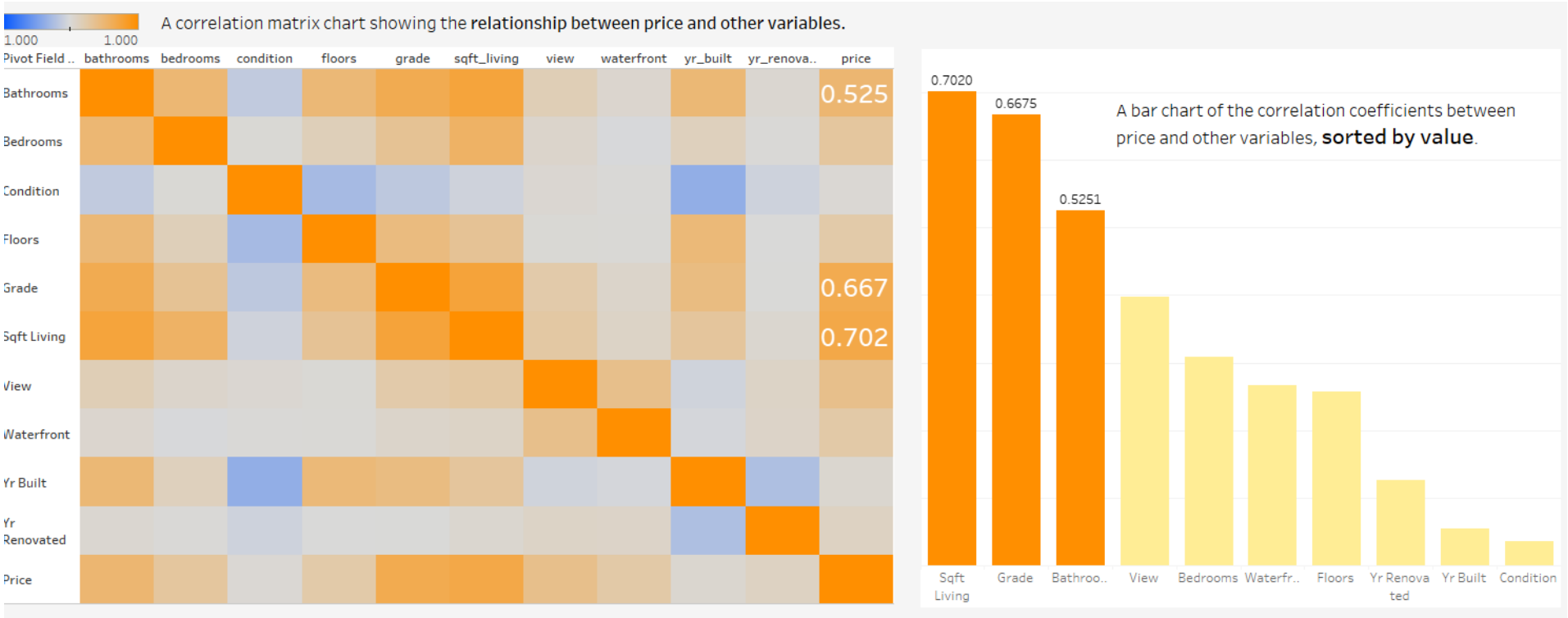


Using Tableau's map feature, display the median housing prices across different zip code areas, and combine this with a bar chart to show specific information. This allows us to clearly see which areas have high housing prices and which areas have low housing prices.

Note: This process can also be completed using Python. Please refer to the link. [Code]

K i n g C o u n t y

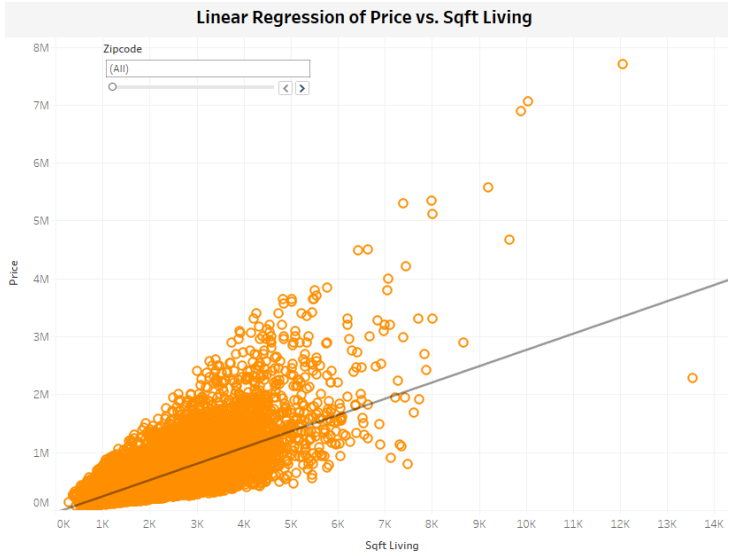
Analysis



Combining Python and Tableau, use Python to organize the data and Tableau to visualize it. Create a correlation matrix of various variables and housing prices, and then select the three variables with the highest correlation coefficients for further analysis.

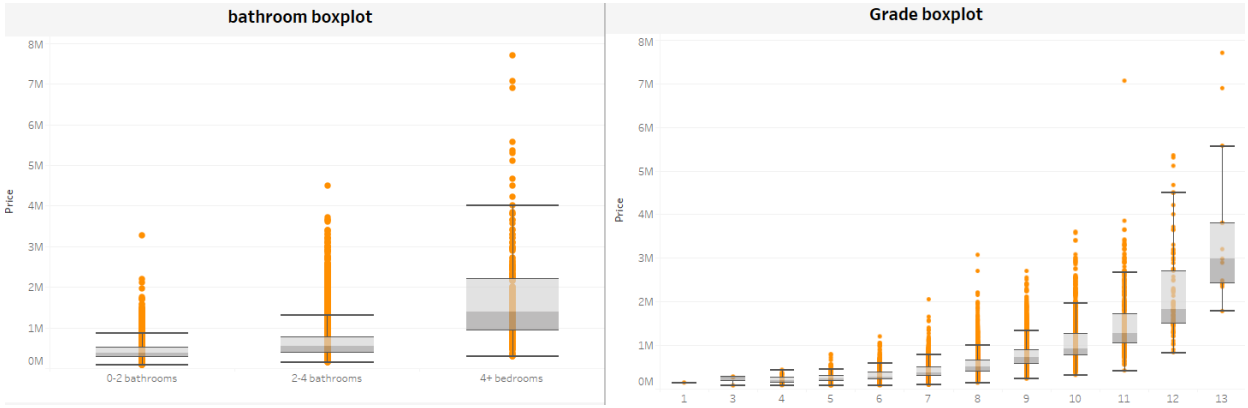
K i n g C o u n t y

Analysis



For the continuous variable SQFT, we create a scatter plot and analyze its relationship with price through linear regression. Additionally, we include the zip code as a filter to view the situation in specific zip code areas.

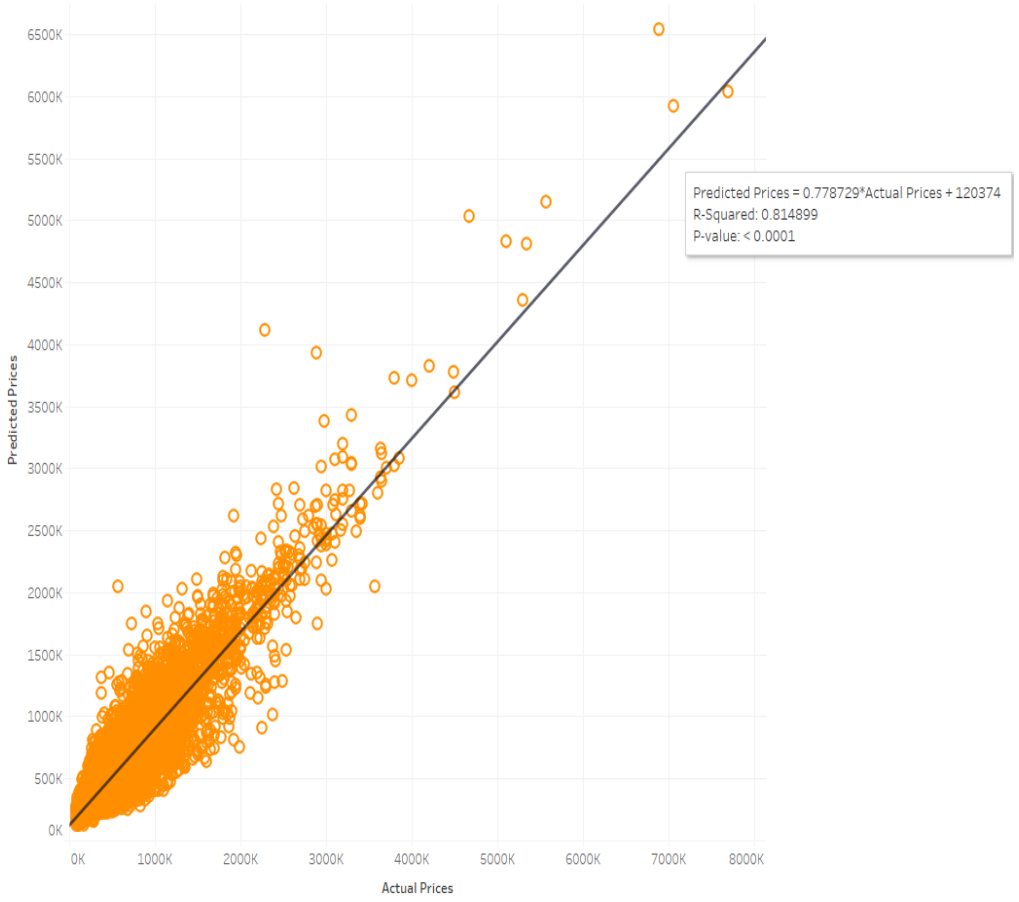
For the discrete variables, such as the number of bathrooms and grade, we use boxplots to analyze the variation in housing price distribution across different categories.



K i n g C o u n t y

Analysis

```python
X = df[['grade', 'bathrooms', 'sqft_living']]
y = df['price']

#Divide the dataset into the training set and the test set
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Normalised data
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

# Define the model
models = {
    'Linear Regression': LinearRegression(),
    'Polynomial Regression': PolynomialFeatures(degree=2),
    'Decision Tree': DecisionTreeRegressor(random_state=42),
    'Random Forest': RandomForestRegressor(random_state=42),
    'Gradient Boosting': GradientBoostingRegressor(random_state=42),
    'Support Vector Regressor': SVR(),
    'K-Nearest Neighbors': KNeighborsRegressor(),
    'Ridge Regression': Ridge(),
    'Lasso Regression': Lasso(),
    'Elastic Net Regression': ElasticNet()
}

# Evaluate model performance
results = {}
for name, model in models.items():
    if name == 'Polynomial Regression':
        poly = PolynomialFeatures(degree=2)
        X_train_poly = poly.fit_transform(X_train_scaled)
        X_test_poly = poly.transform(X_test_scaled)
        lin_reg = LinearRegression()
        lin_reg.fit(X_train_poly, y_train)
        y_train_pred = lin_reg.predict(X_train_poly)
        y_test_pred = lin_reg.predict(X_test_poly)
        train_r2 = r2_score(y_train, y_train_pred)
        test_r2 = r2_score(y_test, y_test_pred)
    else:
        model.fit(X_train_scaled, y_train)
        y_train_pred = model.predict(X_train_scaled)
        y_test_pred = model.predict(X_test_scaled)
        train_r2 = r2_score(y_train, y_train_pred)
        test_r2 = r2_score(y_test, y_test_pred)

    results[name] = {'train_r2': train_r2, 'test_r2': test_r2}

for model, metrics in results.items():
    print(f"{model}:")
    print(f"Train R²: {metrics['train_r2']}")
    print(f"Test R²: {metrics['test_r2']}\n")
```



Use Python to test and select the model.
Finally, we have determined that the 'Random Forest' is the most suitable model.

Visualize the model results using Tableau.

Based on the analysis, we draw the following conclusions:

1. There are significant differences in housing prices across different zip code areas in King County. The most expensive areas are 98039, 98004, and 98040. around the lake, while the cheapest areas are 98168, 98002, and 98032. in the southwest corner.

2. Housing prices have a strong positive correlation with living square footage (living-SQFT). In specific zip code areas, this correlation becomes even stronger.

3.Housing prices are positively correlated with the number of bathrooms, and as the number of bathrooms increases, the price variability also becomes larger.

4.Housing prices are positively correlated with grade, and as the grade increases, the price variability also becomes larger.

5.Based on the above factors, we have developed a housing price prediction model that can serve as a reference for homebuyers.


Thank You !