Data Analysis Portfolio

Esther Howard

About Me

Hello! I'm Esther, with a decade-long journey in the tech industry where I've thrived as a software and data engineer as well as a web analyst. I am looking to get more experience as a Data Analyst to refine my skills further and grow in my career.

My prior career expertise has honed my abilities in problem-solving, critical thinking, attention to detail, adaptability and proactivity, enabling me to not only analyze data but also to understand it in the context of broader challenges.

Thank you for taking the time to explore my portfolio—I'm thrilled to showcase how my diverse expertise converges to bring value to the field of data analysis!

Esther

Projects

1 ClimateWins

In-depth analysis using machine learning algorithms to predict weather patterns.

2 Gun Violence

Comprehensive analysis using machine learning and time series to find trends and risks.

3 Influenza Season

Influenza data analysis to help national hospital staffing agency.

4 Rockbuster

International movie rental analysis to come with a new online streaming strategy.

5 Instacart

Customer behavior analysis to optimize business strategy.

6 GameCo

Global gaming market analysis to help with business initiatives.

1

ClimateWins

In-depth analysis using machine learning algorithms to predict weather patterns.

Project Overview

- This project aims to Use Machine Learning to help predict the consequences of climate change in Europe

Project Data

- Project Brief
- Dataset
- Python Scripts
- Machine Learning Reports
- Presentation

Techniques Applied

- Data Optimization
- Supervised Machine Learning: KNN, ANN and Decision Tree
- Unsupervised Machine Learning: CNN, RNN and Random Forest
- Presenting Machine Learning results

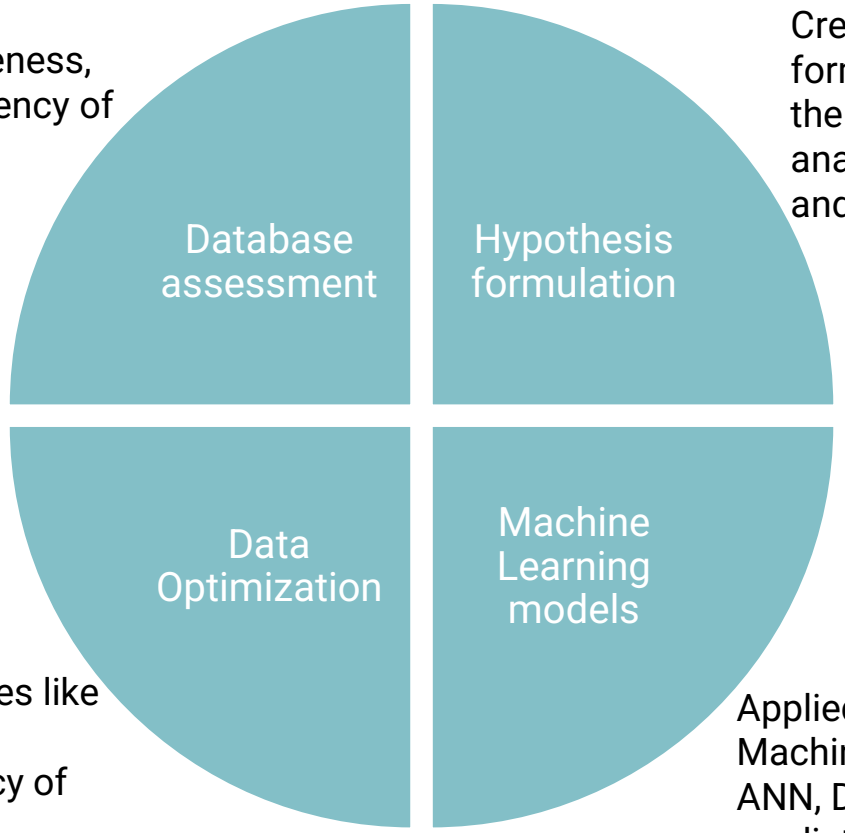
Limitations

- Data available for 18 stations in Europe
- Only 62 years of data available
- Only 11 meteorological patters included in the dataset

Tools

Approach and Methodology

Checked for completeness, accuracy and consistency of the dataset.



Created business questions to help formulate a research hypothesis for the analysis. Also developed an analysis considering the project goals and choose a hypothesis to prove

```
line_marker = dict(color='#101010', width=2)
fig = go.Figure()
fig.add_surface(x=theta1_vals, y=theta0_vals, z=J_vals)
fig.add_scatter3d(x=theta1_history1, y=theta0_history1, z=J_history1, line=line
#The below line adds a graph of just the loss over iterations in a 2D plane
plt.plot(theta0_history1, theta1_history1, 'r+');
fig.update_layout(title='Loss function for different thetas', autosize=True,
                  width=600, height=600, xaxis_title='theta0',
                  yaxis_title='theta1')
fig.show()
```

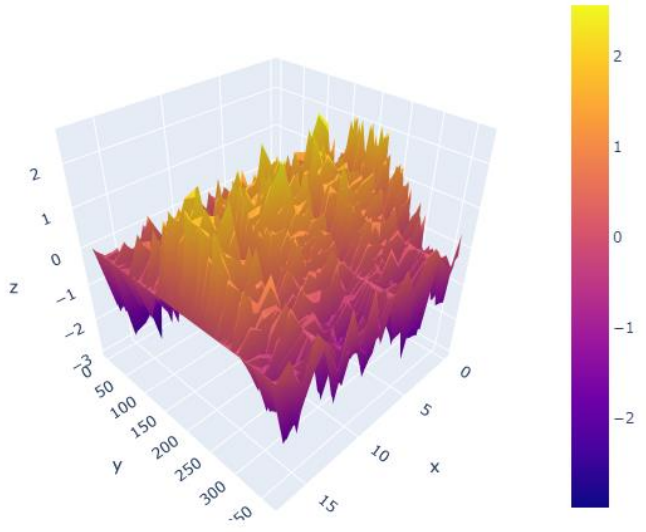
```
# Run Decision Tree classifier
weather_dt = DecisionTreeClassifier(criterion='gini', min_samples_split=2)
weather_dt.fit(X_train, y_train)
figure(figsize=(15,15))
tree.plot_tree(weather_dt)
```

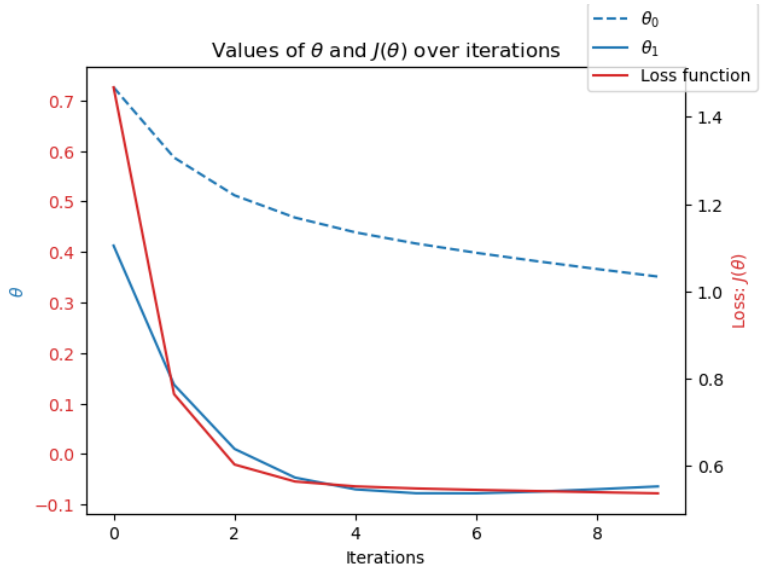
```
# Make predictions on the test set
y_pred_array = classifier.predict(X_test_array)

# Generate the classification report
report = classification_report(y_test_array, y_pred_array)

# Print the classification report
print(report)
```

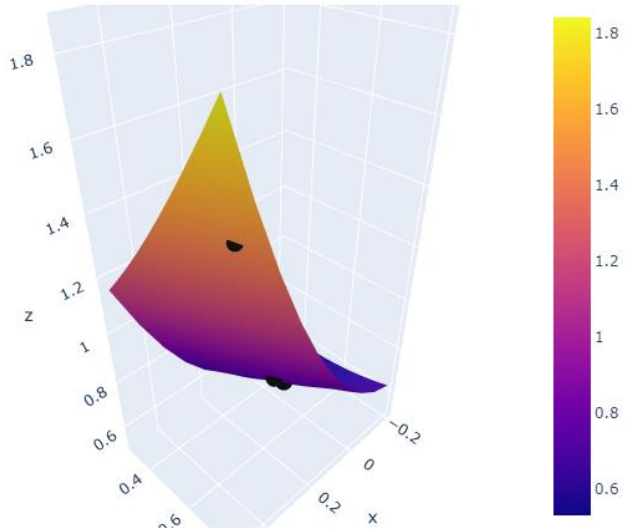
Employed different techniques like Loss Function and Gradient descent to verify the accuracy of ML models

Applied 3 different Supervised Machine Learning techniques (KNN, ANN, Decision Tree) to the dataset to predict weather patterns

Optimization Techniques





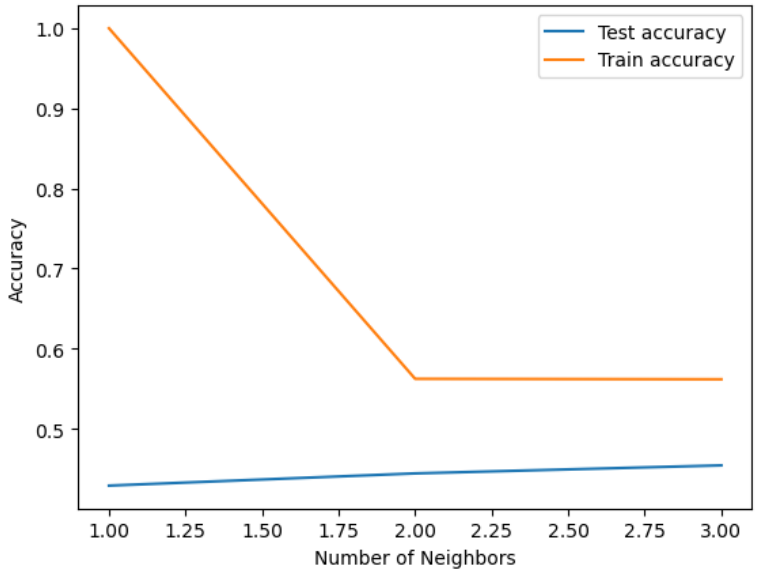


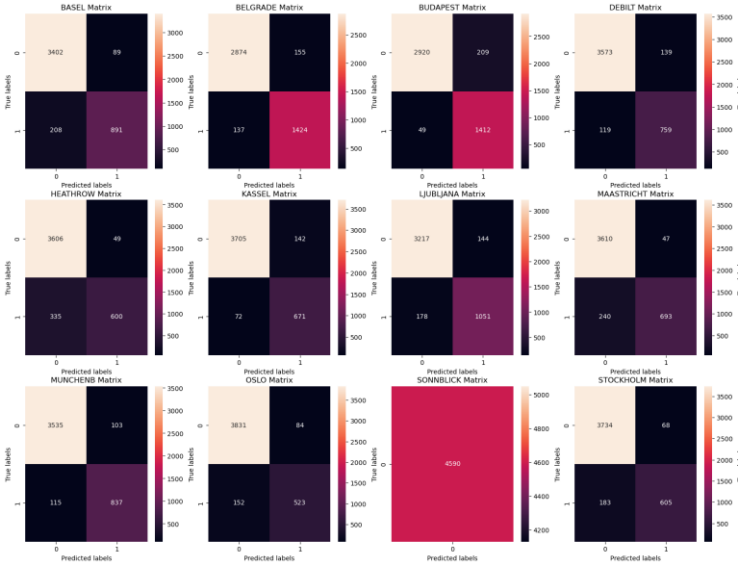
3D visualizations were used to map the weather data for all stations throughout a year

Loss function was employed to assess the deviation between predicted and actual weather data.

Gradient descent was applied to pinpoint the local minimum in the dataset, optimizing the model's performance towards an optimal solution.

Supervised ML models





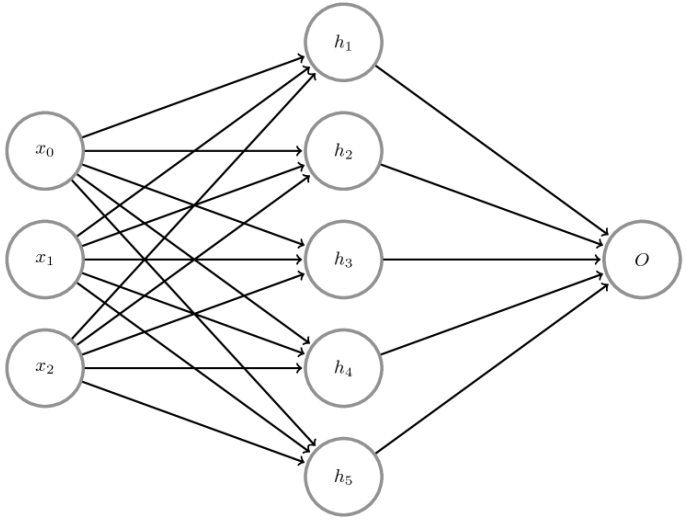
	precision	*Recall*	*f1-score*	*support*
0	0.68	0.69	0.68	1099
1	0.63	0.80	0.71	1561
2	0.65	0.85	0.74	1461
3	0.74	0.73	0.73	878
4	0.72	0.72	0.72	975
5	0.63	0.61	0.62	935
6	0.64	0.55	0.59	743
7	0.62	0.81	0.70	1229
8	0.73	0.74	0.74	933
9	0.74	0.74	0.79	2033
10	0.70	0.61	0.65	952
11	0.64	0.48	0.55	675
12	0.00	0.00	0.00	0
13	0.63	0.46	0.54	788
14	0.00	0.00	0.00	228
micro avg	0.68	0.71	0.69	14490
macro avg	0.58	0.59	0.58	14490
weighted avg	0.67	0.71	0.68	14490
samples avg	0.32	0.33	0.31	14490

KNN was used to classify data points into groups by examining the categories of their closest neighbors.

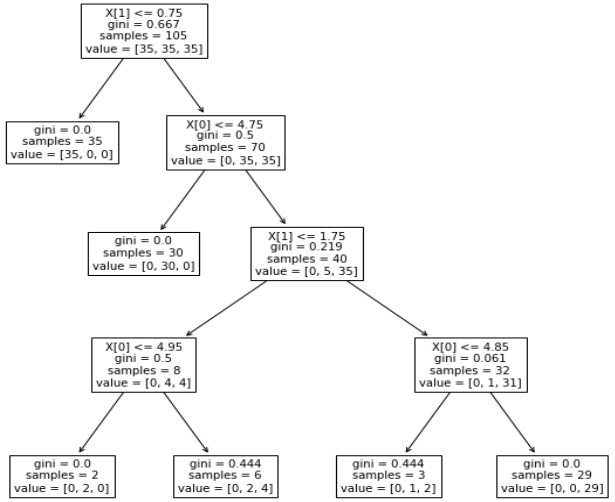
Confusion matrix were used to provide visual insights and showcasing instances where classed were misclassified (confused).

Classification reports were used to evaluate the model's performance across various weather stations in the dataset.

Supervised ML models



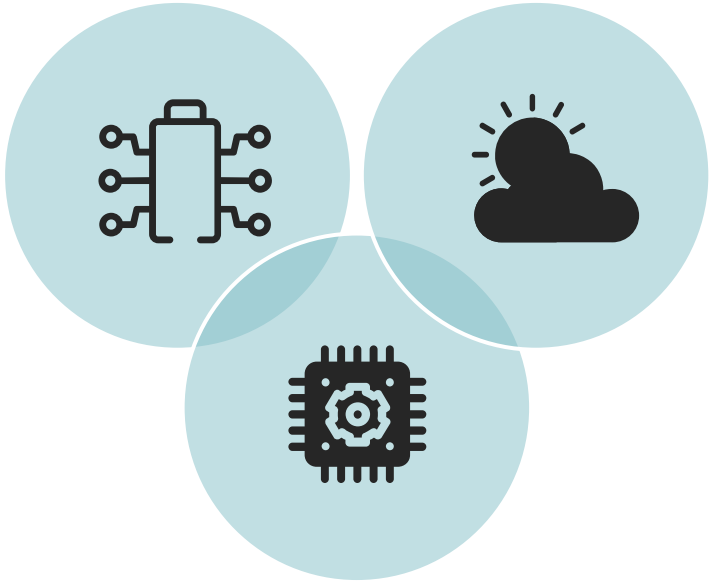
Artificial neural networks (ANN) were used enabling complex computations for predictions and classifications.



Decision trees were employed to progressively refine solutions effectively narrowing down optimal outcomes

Next Steps

Use Artificial Neural Network (ANN) for predicting weather data as it had the highest accuracy rate of the algorithms employed



Further evaluate the weather of all the stations in the dataset over the 60 years' worth of data to determine if Machine Learning can help predict the consequences of climate change.

Use other Machine Learning algorithms like Random Forests and Gradient Boosting Machines to compare results with the already used models.

2

Gun Violence

Comprehensive analysis using machine learning and time series to find trends and risks

Project Overview

- This project focuses on analyzing gun violence incidents data to identify patterns and trends. Through data compilation, exploratory analysis, and predictive modeling, I aim to gain insights into the dynamics of gun violence for academic exploration and understanding.

Project Data

- Project Brief
- Dataset
- Python Scripts
- Tableau presentation

Techniques Applied

- Exploring relationships
- Geographical analysis
- Supervised Machine Learning: Regression
- Unsupervised Machine Learning: Clustering
- Time series analysis
- Creating data dashboards

Limitations

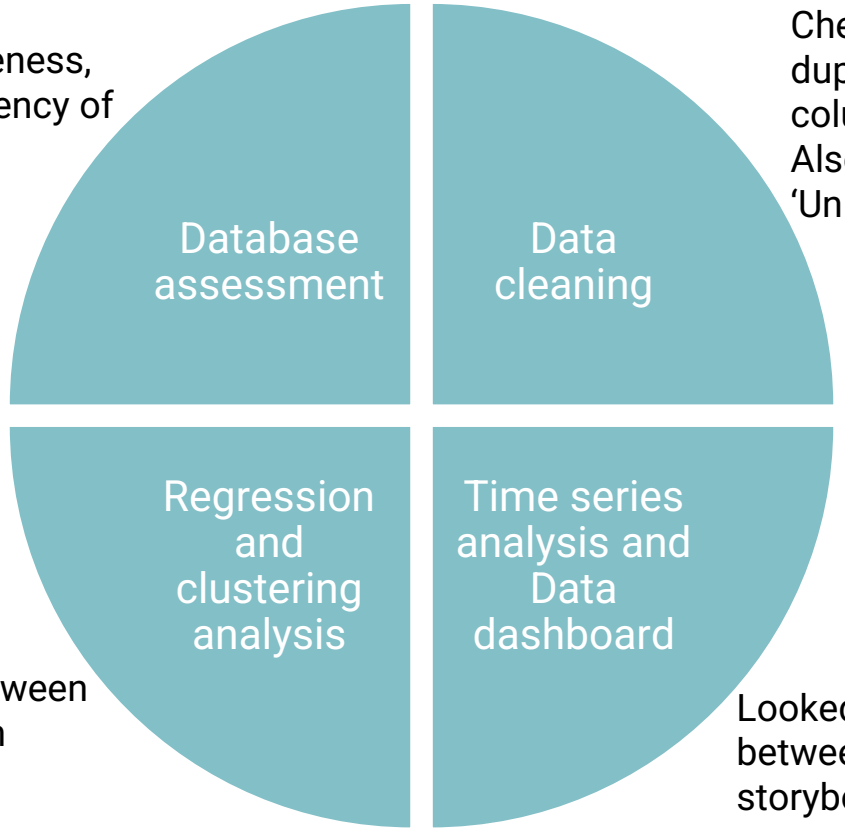
- Data available only between 01/2013 and 03/2018
- Missing demographics data
- No differentiation between perpetrator and victim

Tools

Approach and Methodology

Checked for completeness, accuracy and consistency of the dataset.

Checked for missing values and duplicates besides standardizing columns names.
Also imputed missing values for 'Unknown'.




```
rmse = mean_squared_error(y_train, y_predicted_train)
r2 = r2_score(y_train, y_predicted_train)
```

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

```
# create linear regression object
regression = LinearRegression()

# fit regression to training set
regression.fit(x_train, y_train)
```

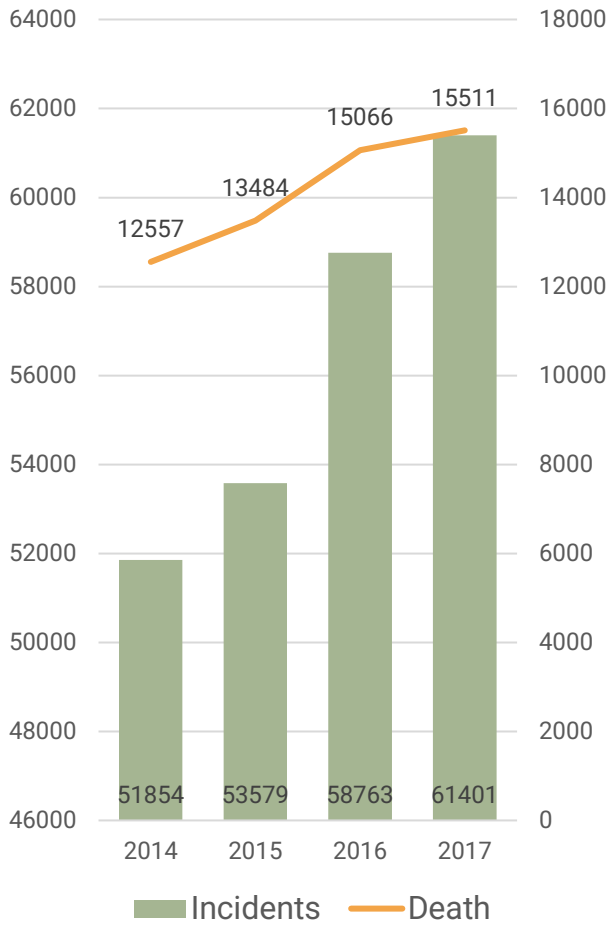
```
# Create the k-means object.

kmeans = KMeans(n_clusters = 3)
```

```
# Fit the k-means object to the data.

kmeans.fit(df_clustering)
```

Explored relationships between variables using correlation heatmaps, pair plots and categorical plots.

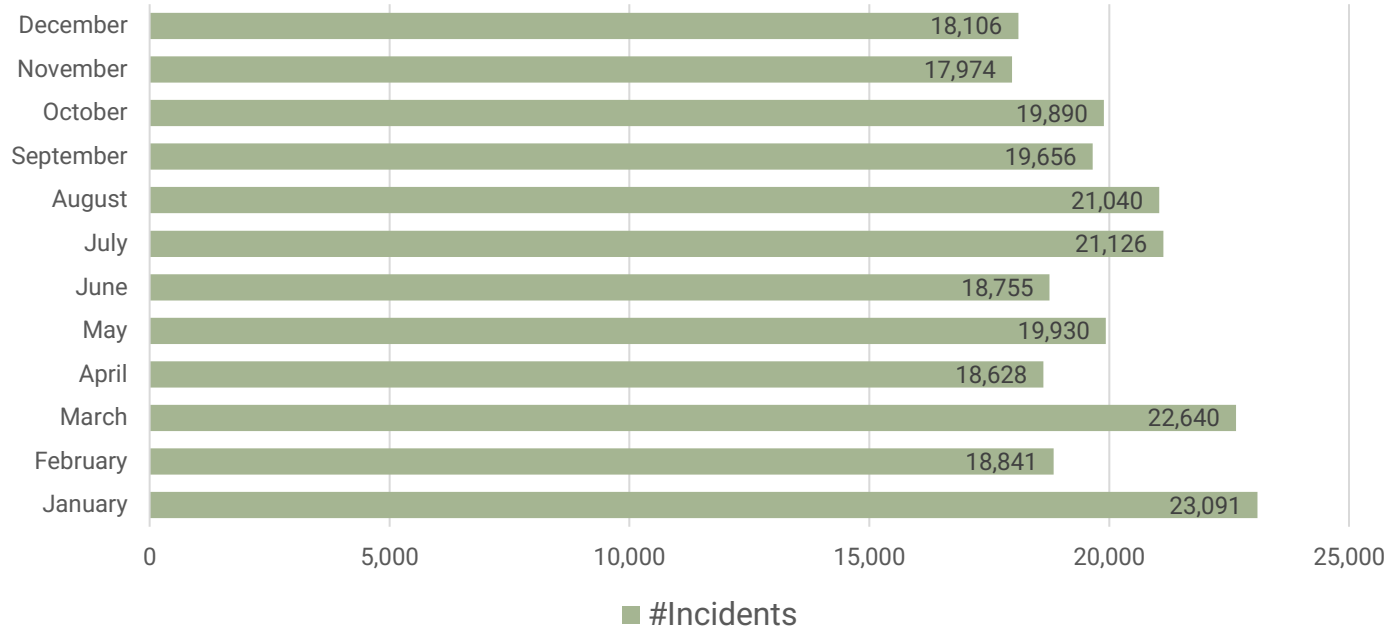
Looked further into relationships between years. Formed a Tableau storyboard to present findings.

Time Analysis

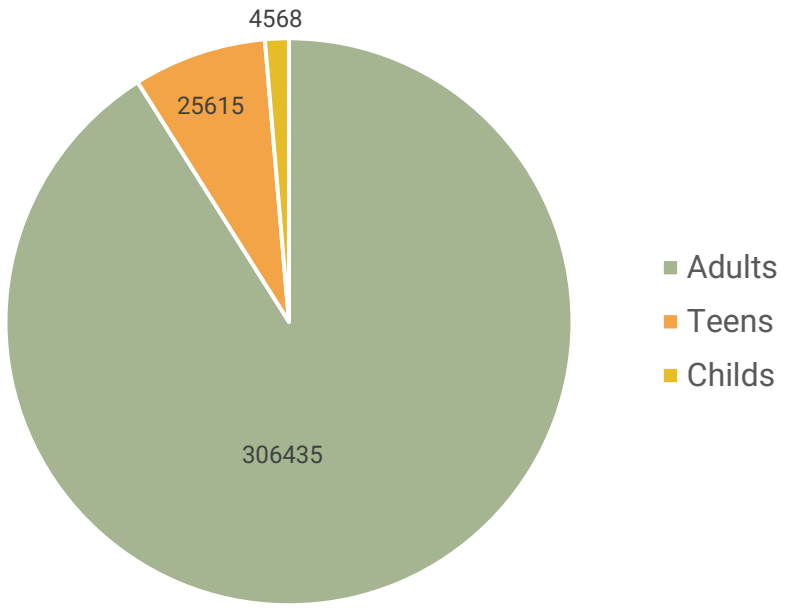


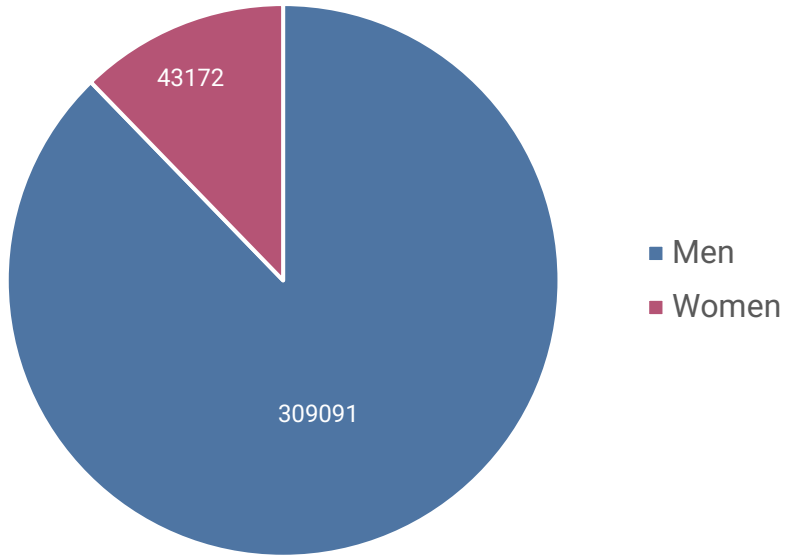
Number of incidents and number of deaths is on the rise

January and March are the months with more incidents



Demographics

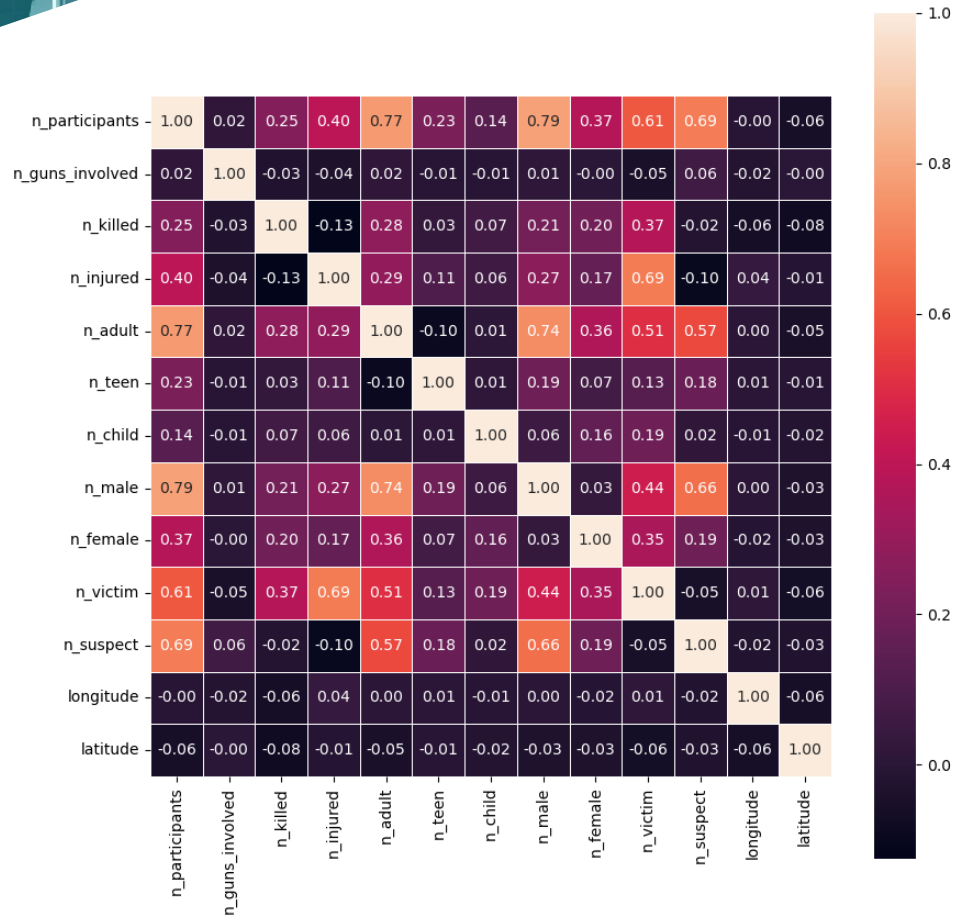


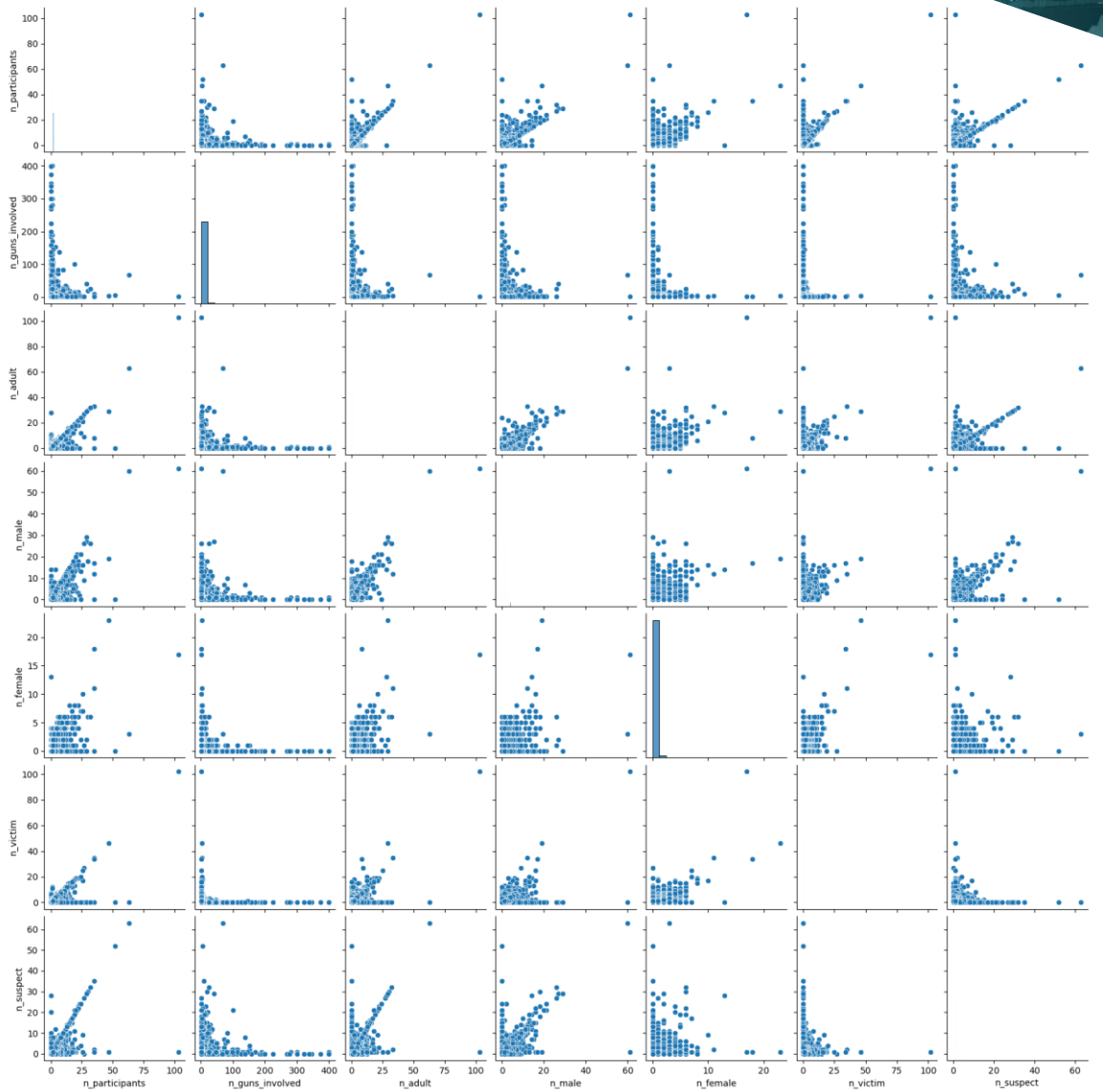


The age group with more incidents are Adults

The gender with more incidents are Men

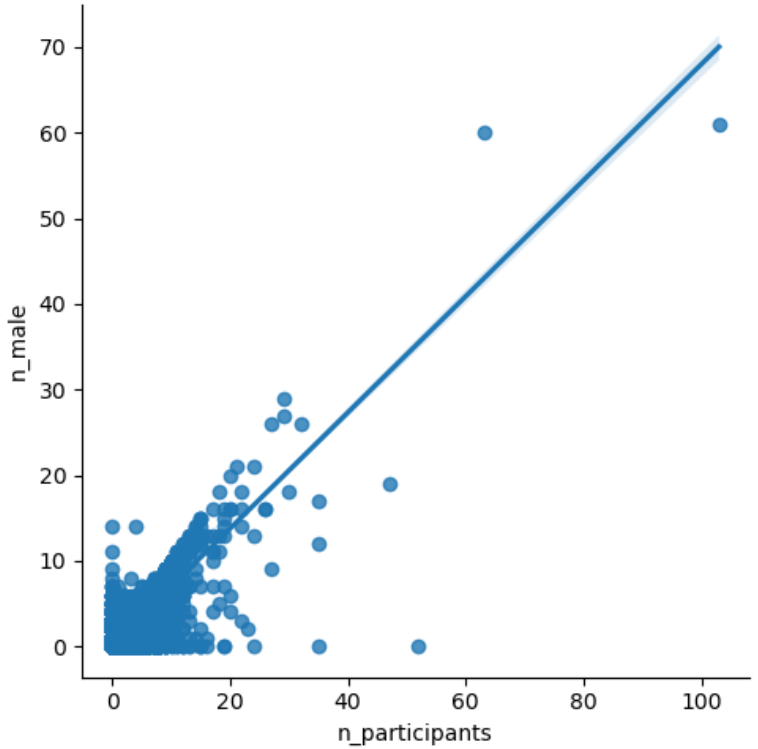
Regression and Clustering Analysis



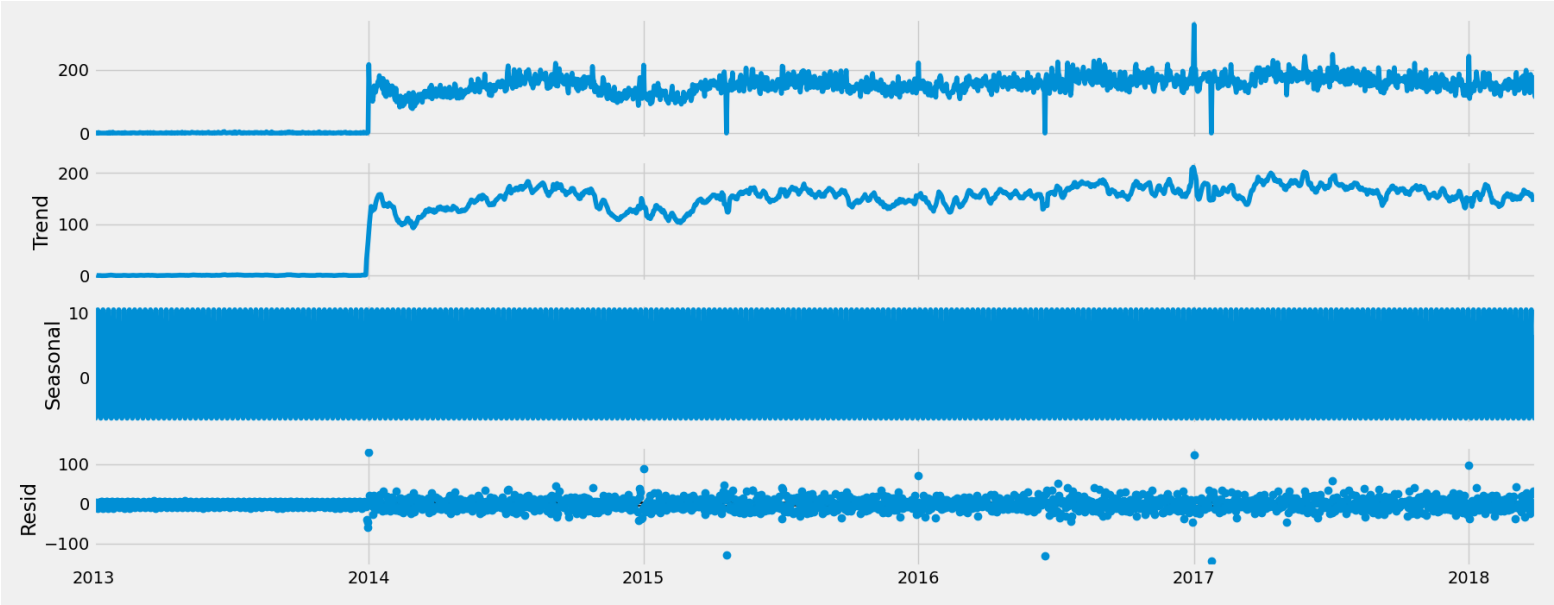


I used a correlation map and a pair plot to analyze the relationships between quantitative variables like number of participants and number of males.
The analysis revealed a strong positive correlation between these 2 variables.

Time series Analysis

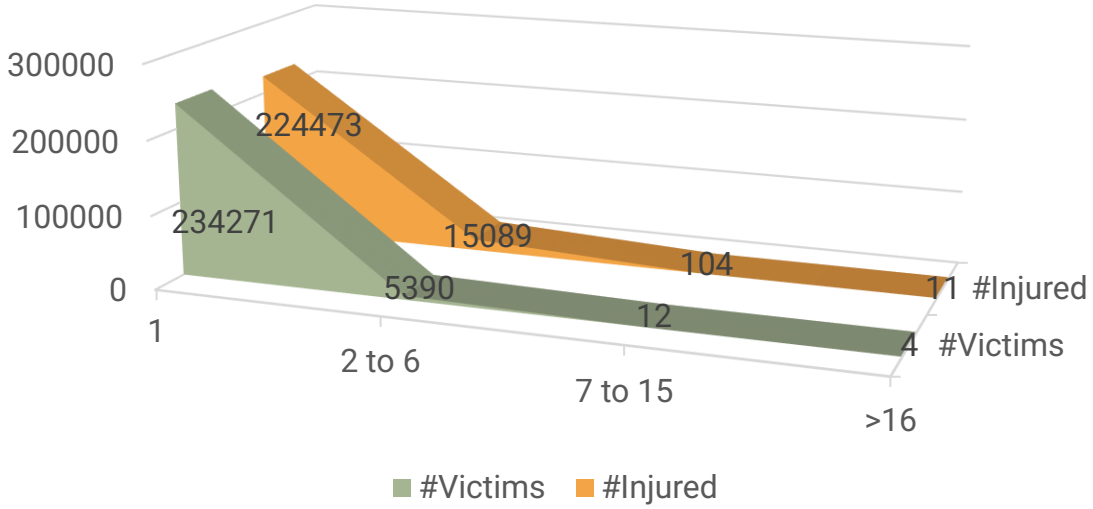


The linear regression analysis identifies a strong correlation between number of participants and number of males as the heatmap and pair plot showed.



The decomposition of the time series analysis allows for the assessment of individual components such as seasonality to identify possible trends. It is evident that gun incidents fluctuate over time.

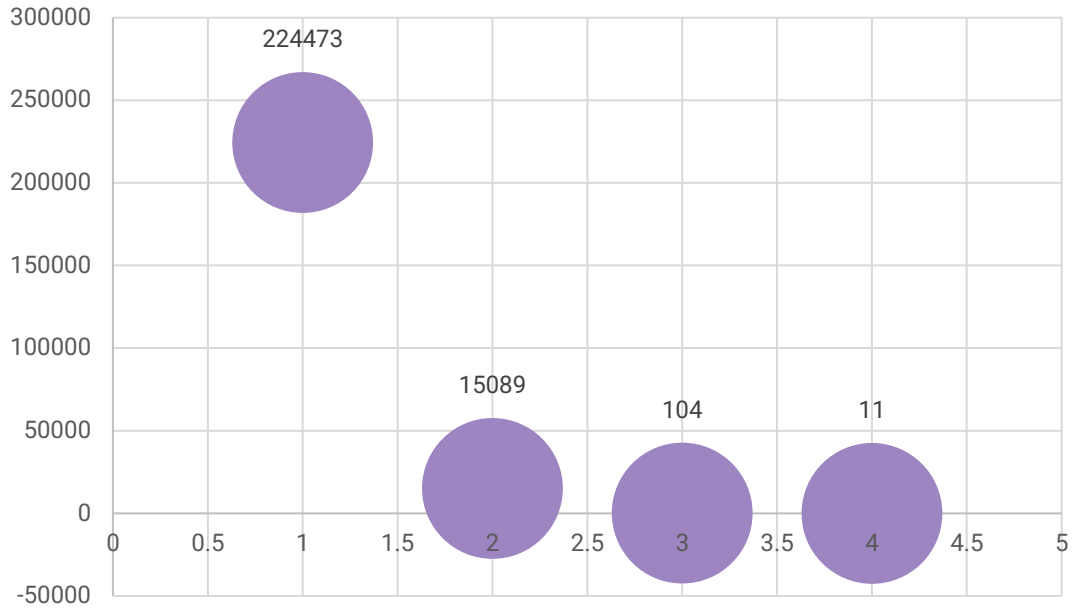
Incident Analysis



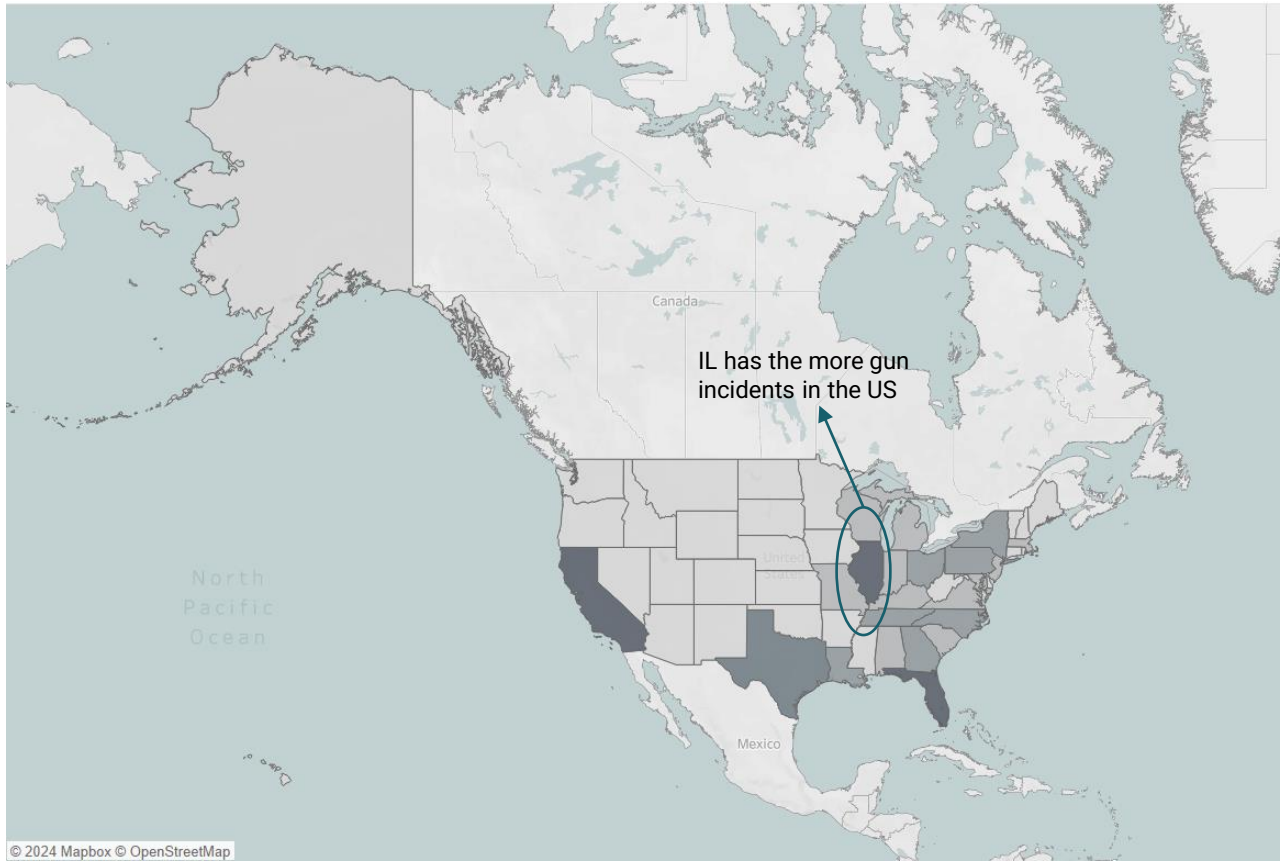
Most incidents result in 1 death

Most incidents result in 1 injured

Most incidents are associated with 1 gun



Regional Trends

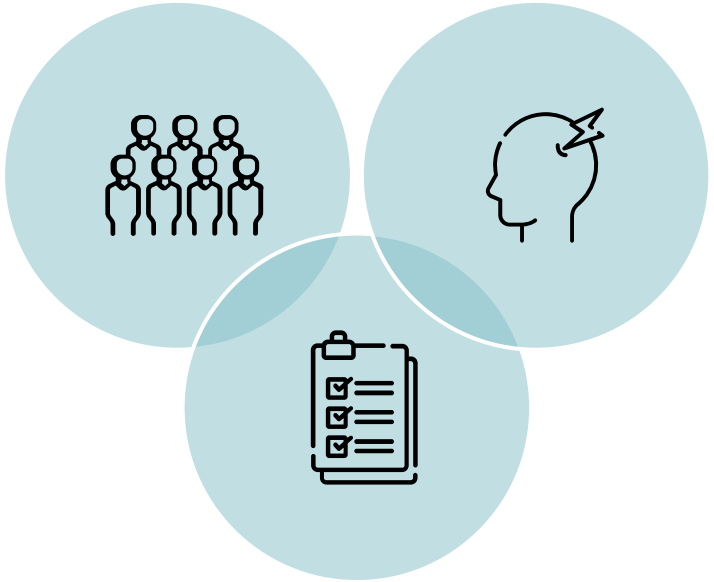


Top cities in top 3 states with more incidents

Illinois	California	Texas
Chicago	Oakland	Houston
Peoria	Los Angeles	San Antonio
Rockford	Fresno	Dallas
Chicago (Englewood)	Bakersfield	Corpus Christi
Springfield	Stockton	Austin
Champaign	Sacramento	Fort Worth
Joliet	San Francisco	Amarillo
Aurora	San Diego	Lubbock
Kankakee	Long Beach	Killeen
Chicago (Roseland)	Salinas	El Paso

Next Steps

Community engagement: foster dialogue and engagement within communities affected by gun incidents to understand local concerns and priorities.



Research and evaluation: research to better understand the effectiveness of different practices and programs for preventing gun incidents.

Comprehensive data collection: enhance data collection to gather more comprehensive and accurate information like demographics of victims and perpetrators and circumstances surrounding each incident for further analysis.

3

Influenza Season

Influenza data analysis to help a national hospital staffing agency.

Project Overview

- Analyze historical influenza trends in the US to assist medical staffing agency
- Learn how to interpret business requirements to guide a data analysis
- Explore the different types of data visualizations and best practices to keep them accessible

Project Data

- Project Brief
- Datasets
- Interim Report
- Final Report
- Tableau Presentation

Techniques Applied

- Sourcing the right data
- Data profiling
- Data quality measures
- Data transformation
- Data integration
- Statistical analysis
- Consolidating analytical techniques
- Data visualization
- Spatial analysis and visualization
- Storytelling

Limitations

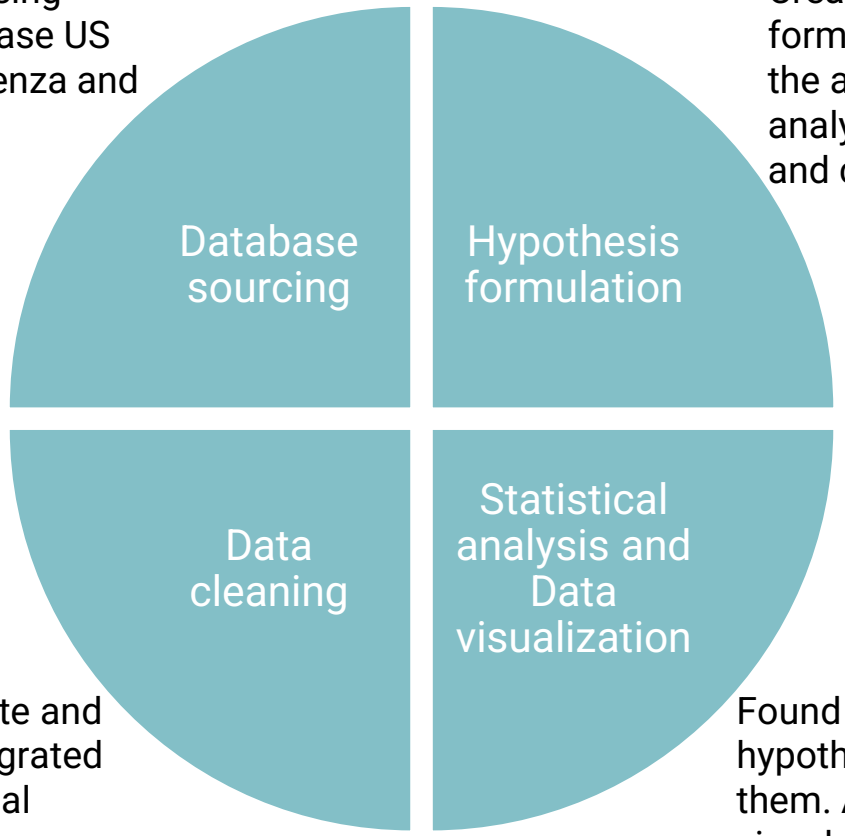
- Influenza mortality data suppressed for confidentiality
- Records missing from many states
- Datasets only cover years 2009-2017
- Vulnerable population data is partially missing, this forced the project to focus only on vulnerable people of 65+ years

Tools

Approach and Methodology

Sourced pertinent datasets using trustworthy sources. In this case US government website for influenza and vaccines.

Created business questions to help formulate a research hypothesis for the analysis. Also developed an analysis considering the project goals and choose a hypothesis to prove



Hypothesis:

"If patient is senior patient (65+ years) then the patient is at higher risk of contracting influenza and/or dying of influenza"

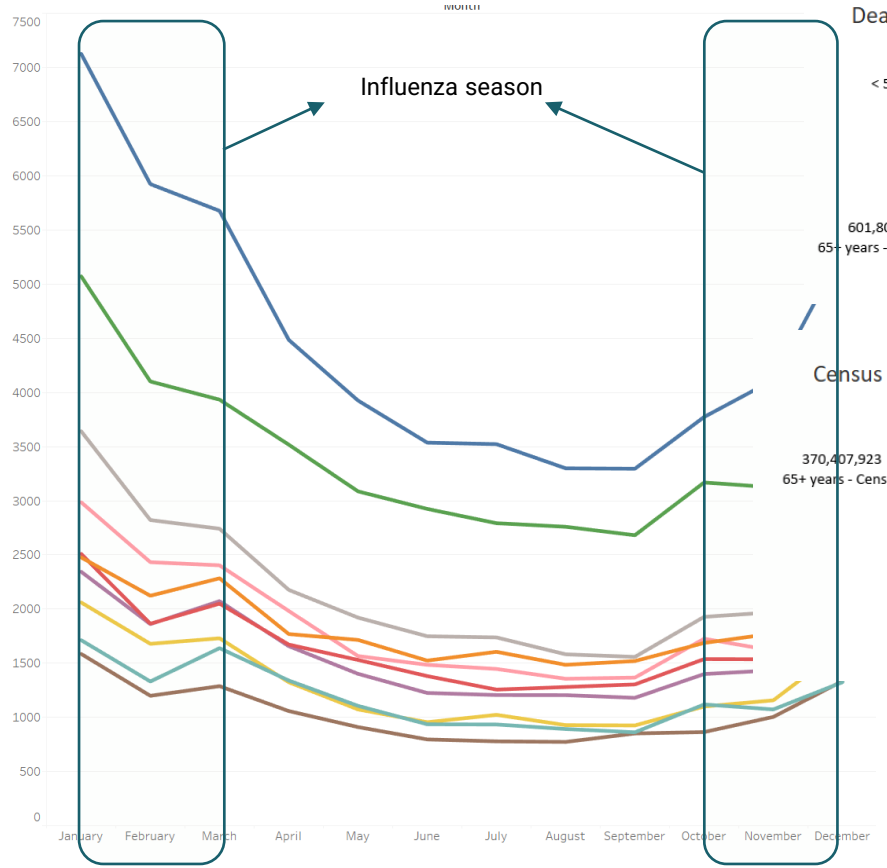
t-Test: Two-Sample Assuming Unequal Variances

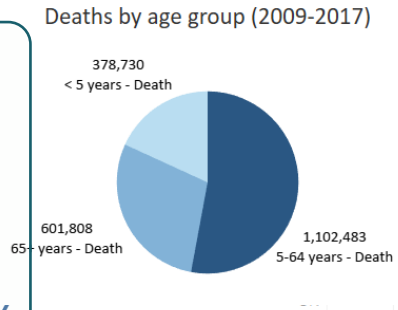
	0-64 years	*65+ years*
Mean	3227.04	1311.13
Variance	9964.77	617945.65
Observations	459	459
Hypothesized Mean Difference	0	
df	473	
t Stat	51.80	
P(T<=t) one-tail	2.23E-197	
t Critical one-tail	1.65	
P(T<=t) two-tail	4.45E-197	
t Critical two-tail	1.96	

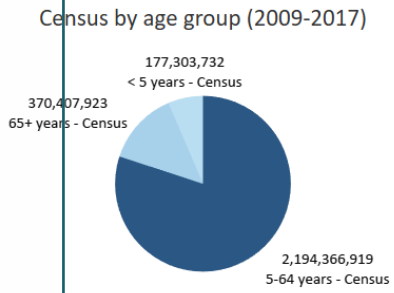
Transformed datasets by state and year for the analysis and integrated the 3 datasets used into a final one.

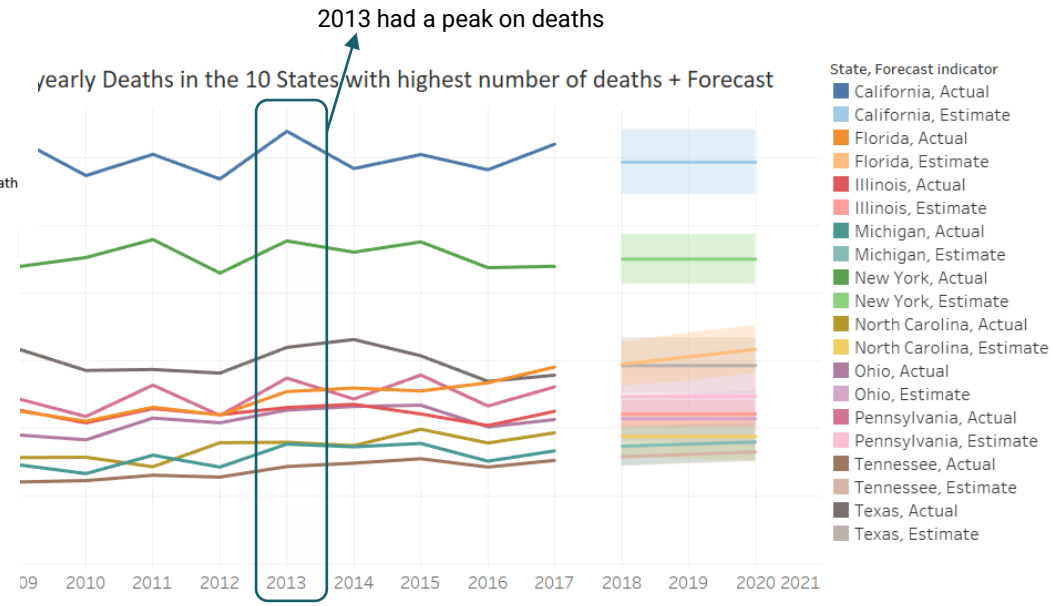
Found correlations between hypothesis and conduct t-test to prove them. Also designed and sequenced visuals using storytelling principles.

Influenza Season

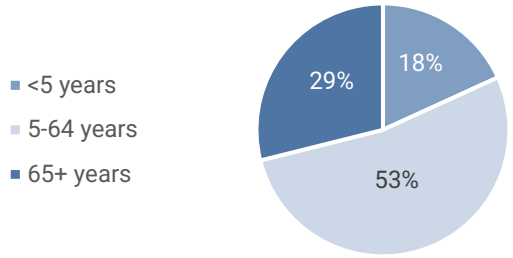




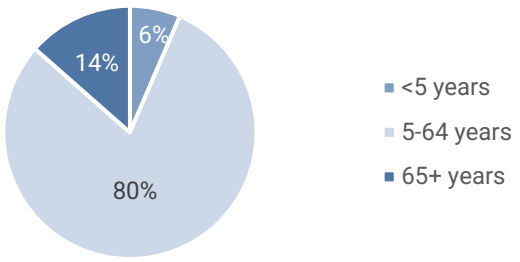




Deaths by age group (2008-2017)



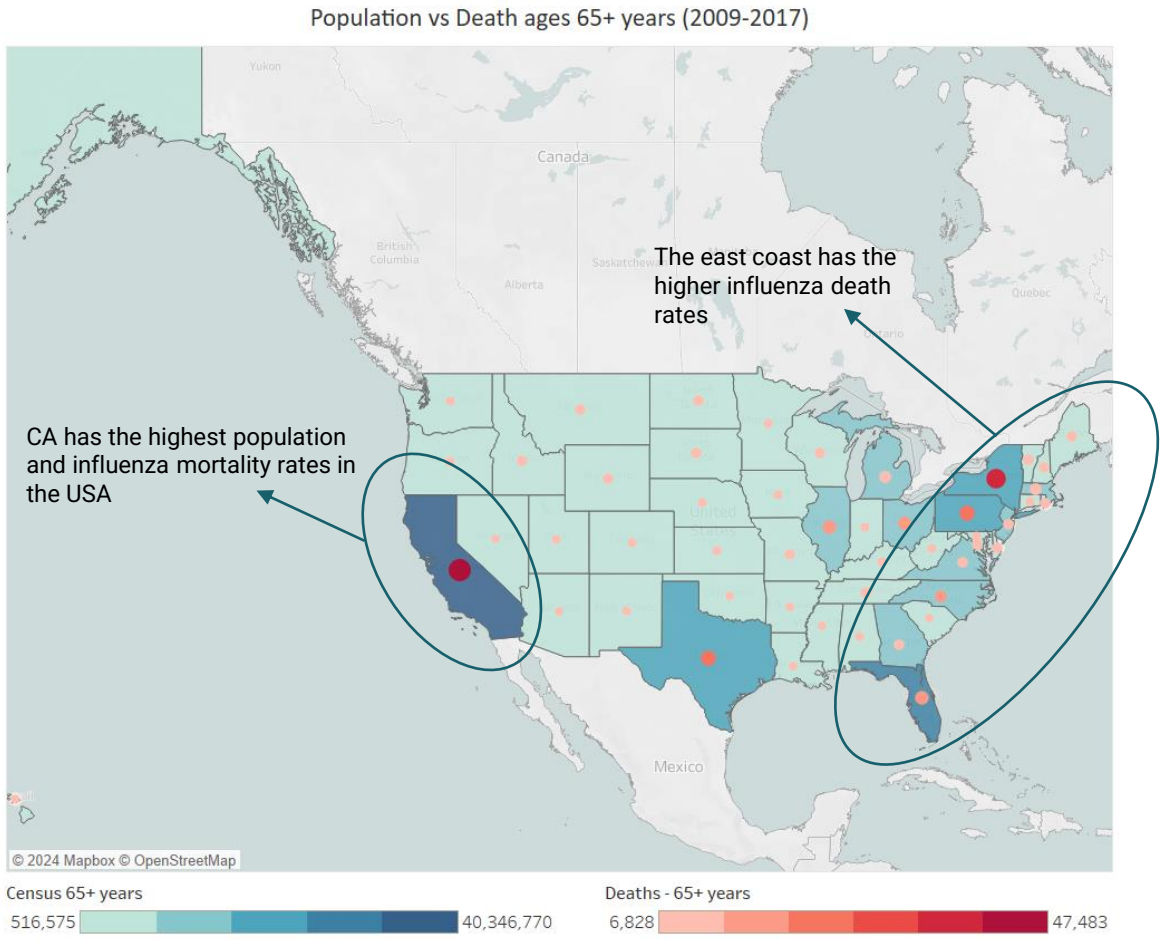
Census by age group (2008-2017)

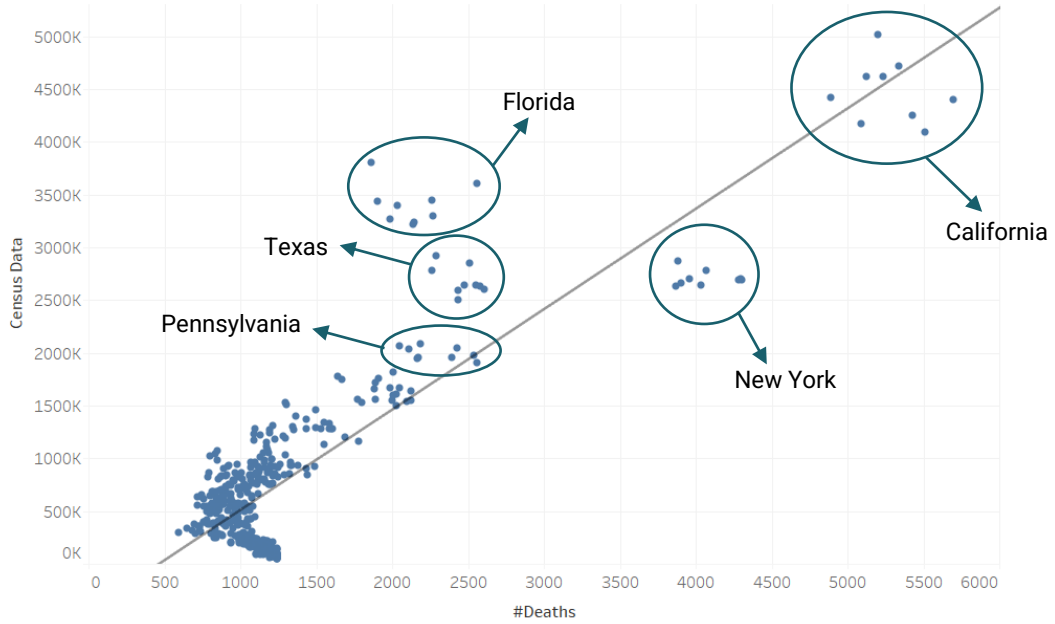


Influenza season starts around November and ends in March, overlapping with the cold months of winter.
The forecast for following years has some states increasing number of deaths.

47% of total deaths are in only 20% of the total population

Spatial Analysis and Mortality

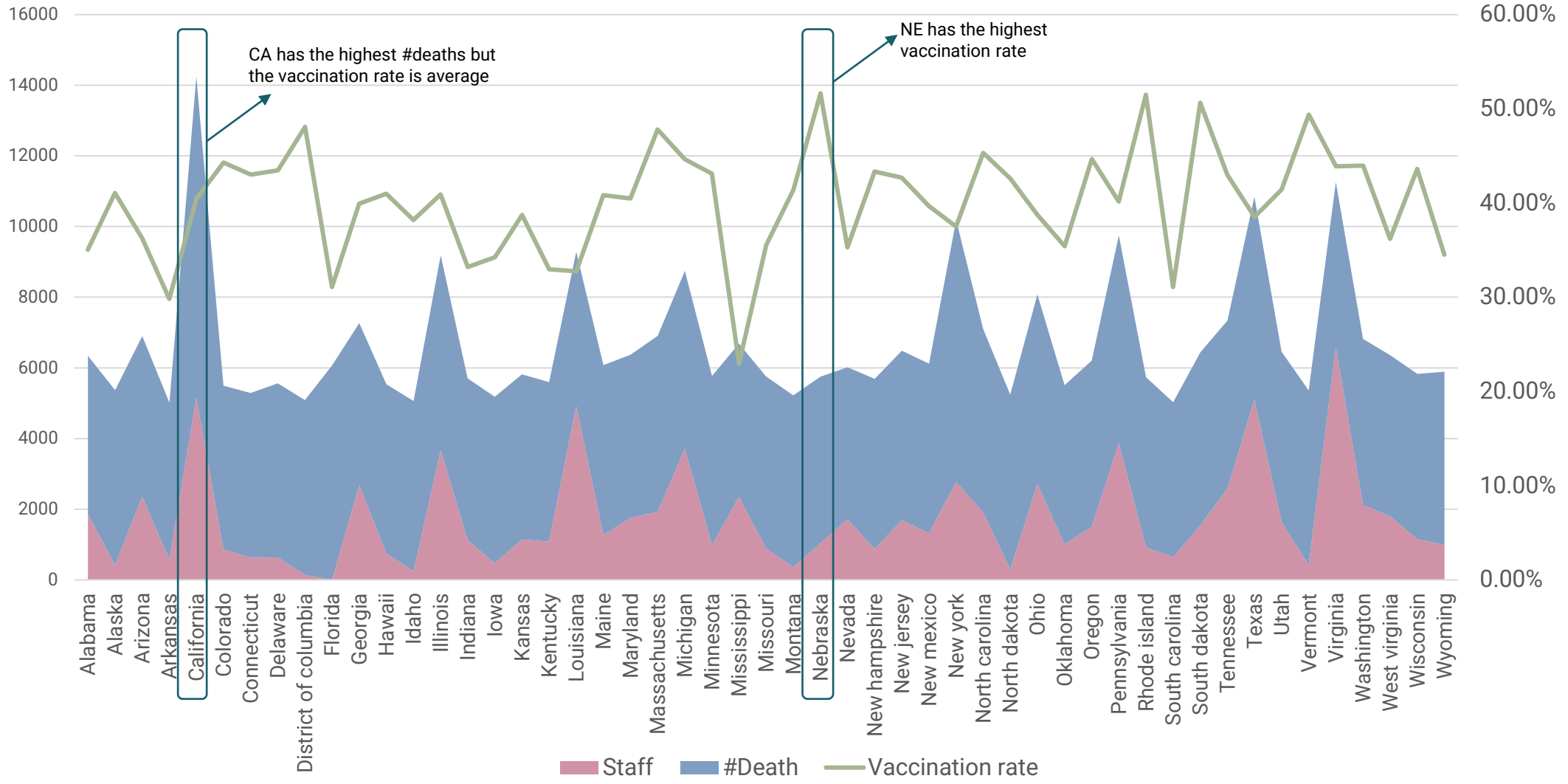




States with higher # of 65+ years population
California, Florida, New York, Texas, Pennsylvania, Ohio, Illinois, Michigan, North Carolina and New Jersey

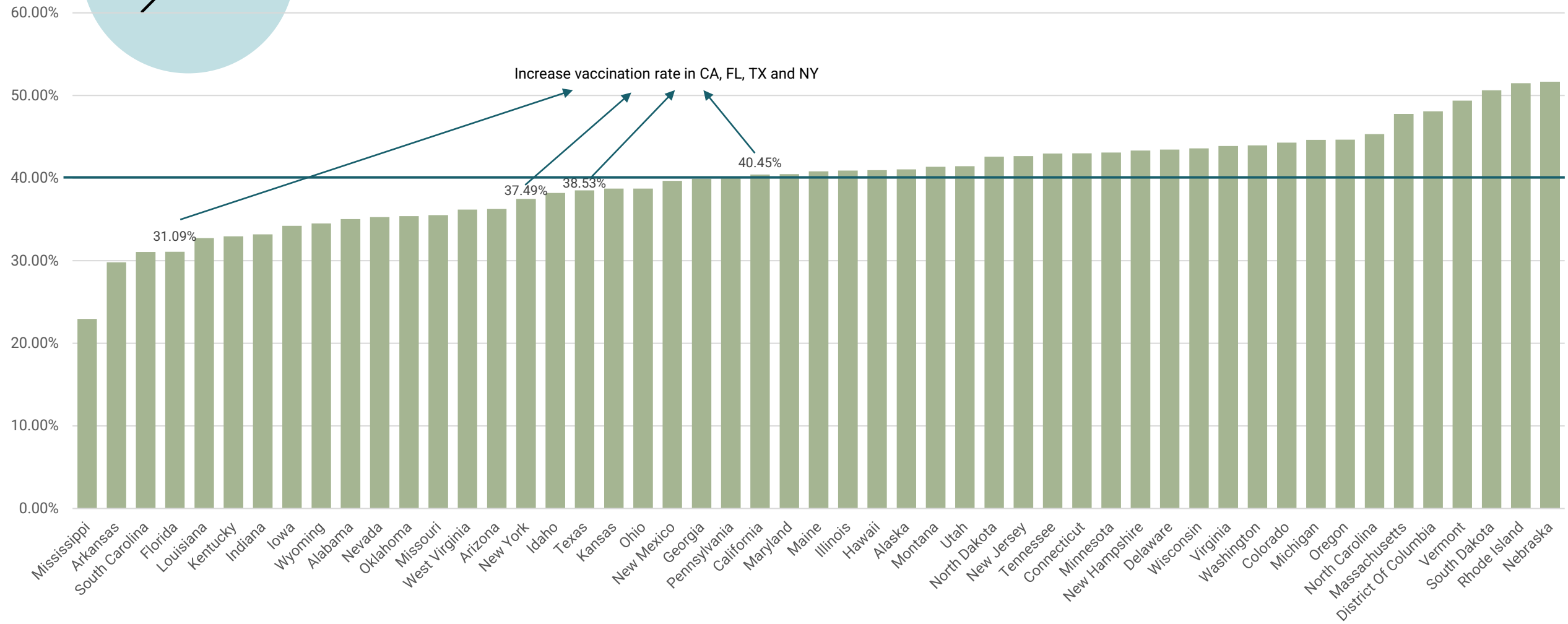
States with higher #deaths on 65+ years population
California, New York, Texas, Pennsylvania, Florida, Illinois, Ohio, North Carolina, Michigan, Massachusetts

Vaccination rates and Staffing



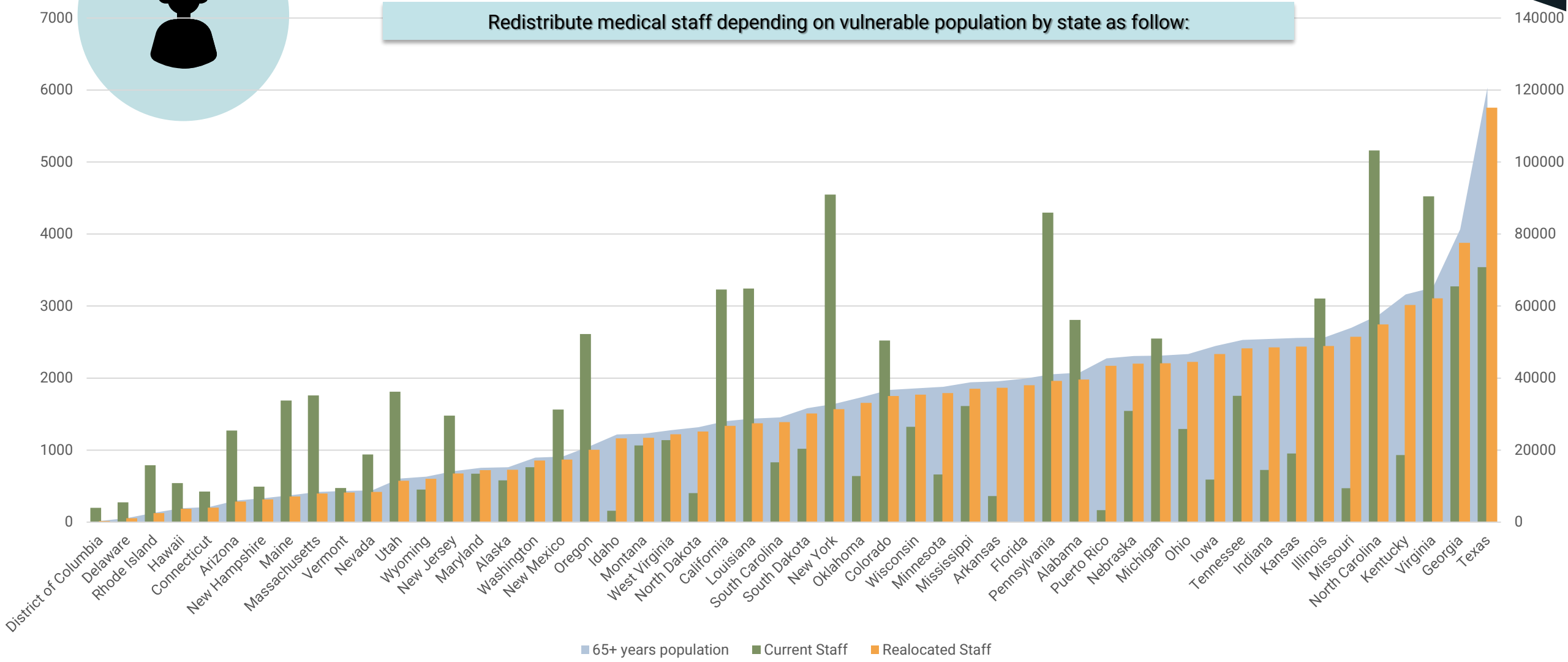
Recommendations - Vaccines

Avg vaccination rate in the country: 40%
Increase vaccination rate to 55% in states with high number of deaths



Recommendations - Staffing

Redistribute medical staff depending on vulnerable population by state as follow:



4

Rockbuster

International movie rental analysis to produce a new online streaming strategy

Project Overview

- Give insights to management team ahead of their online video rental launch
- Develop database querying skills while mastering SQL

Project Data

- Project Brief
- Dataset
- Data Dictionary
- Queries
- Presentation

Techniques Applied

- Database querying in SQL
- Filtering data
- Summarizing data
- Cleaning data
- Joining tables
- Performing subqueries
- Presenting SQL results

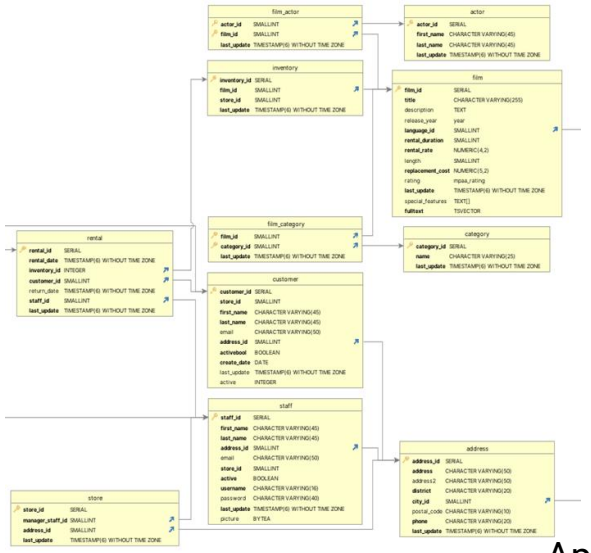
Limitations

- Movies available only in English
- Transactions missing payment history

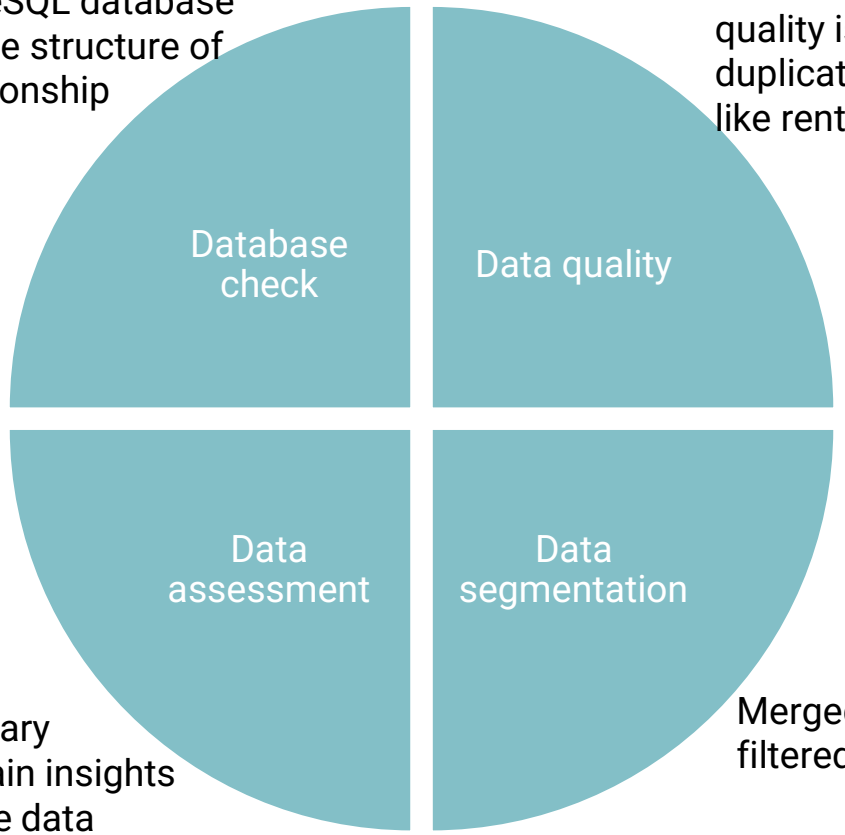
Tools

Approach and Methodology

Checked PostreSQL database and analyzed the structure of the Entity Relationship Diagram



Applied summary statistics to gain insights of the available data



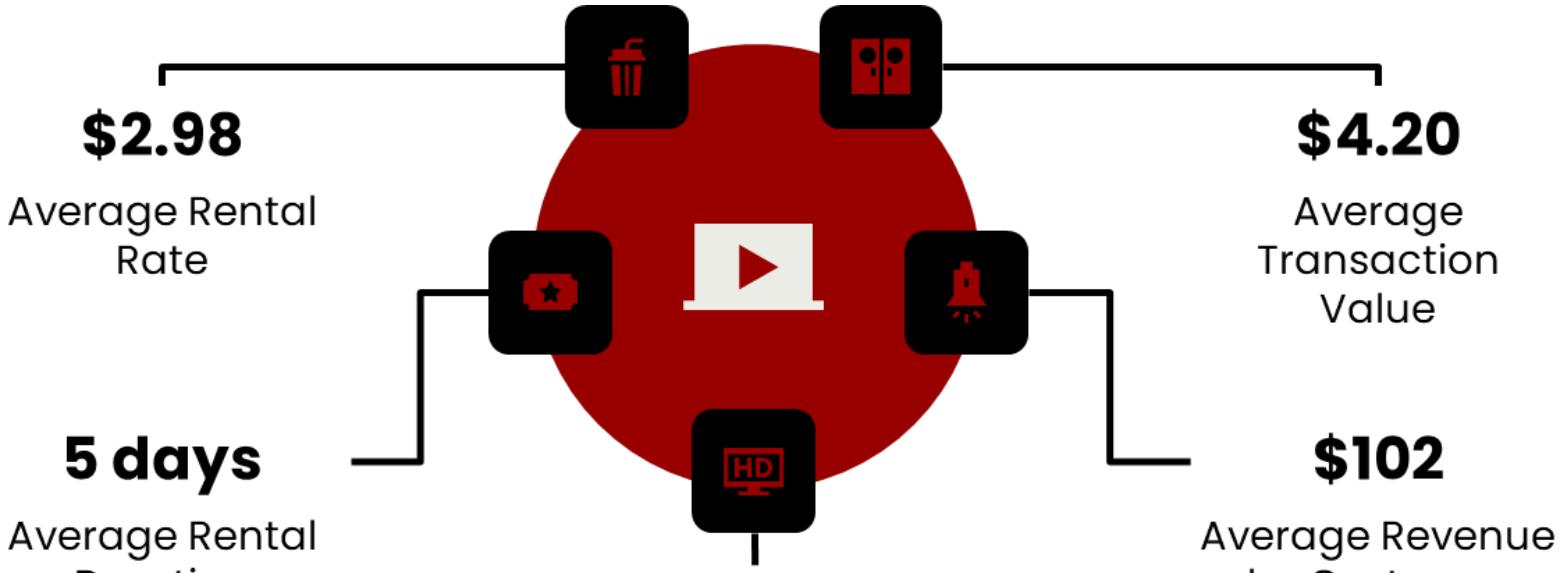
Evaluated and resolved data quality issues such as duplicates and missing values, like rental data missing in 2005

```
SELECT * FROM customer
WHERE first_name IS NULL or first_name = '';
SELECT * FROM customer
WHERE last_name IS NULL or last_name = '';
SELECT * FROM film
WHERE title IS NULL or title = '';
SELECT * FROM payment
WHERE amount IS NULL;
```

```
SELECT payment_id, customer_id, count(*) FROM payment
GROUP BY payment_id, customer_id
HAVING count(*) > 1
```

Merged subsets and filtered them for analysis

Business Overview



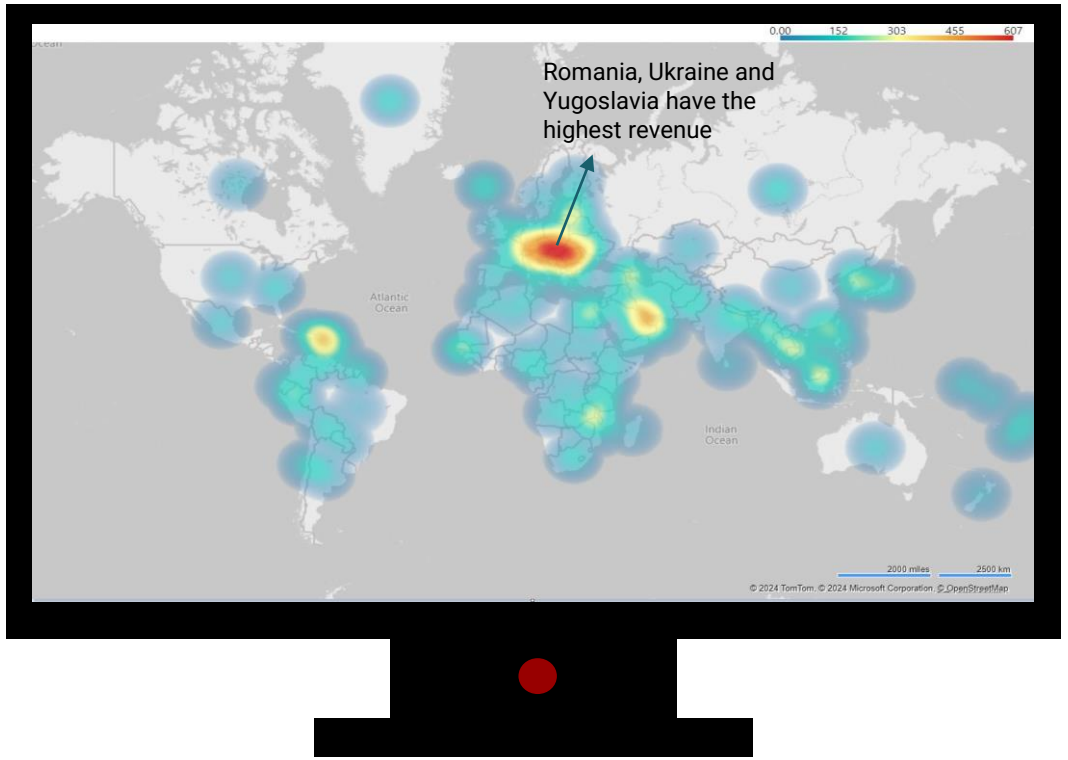
1000 films

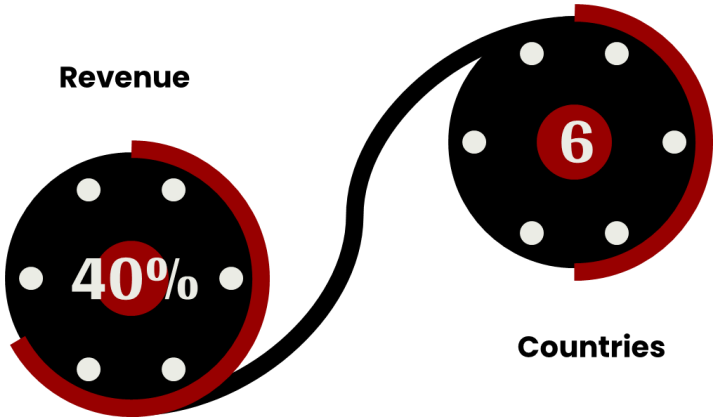
17 genres

5 ratings

1 language

Revenue by countries

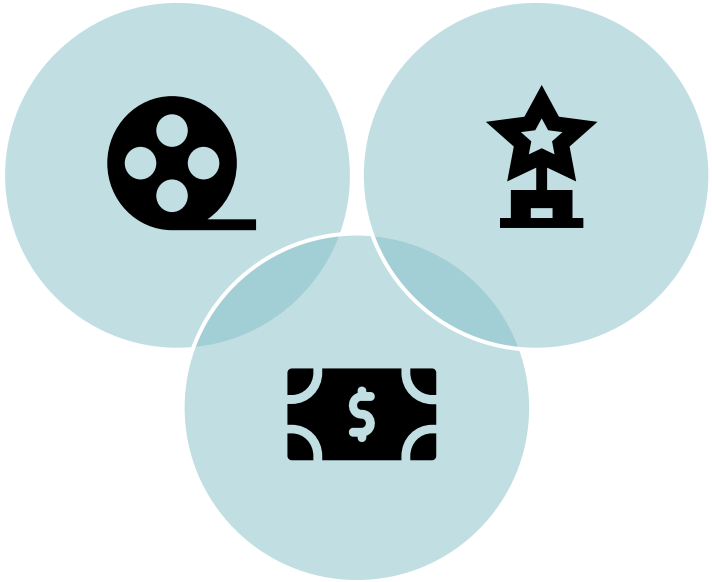




The countries with highest revenue generated are Romania, Ukraine, Yugoslavia, UAE, Egypt and Hong Kong, as seen in the map.
These 6 countries account for the 40% of total revenue.

Recommendations

- Add movies in more languages
- Explore adding titles adapted to different countries:
  - Documentaries in India
  - Animation in China
  - Sports in the USA



- Pricing strategy depending on location.
- 2 price tiers:
  - Monthly subscription ($5.99 to $12.99/month)
  - Pay per view ($2.99 to $4.99/movie)

- Start marketing campaign to announce changes to current customers.
- Soft launch of streaming service.

5

Instacart

Customer behavior analysis to optimize business strategy

Project Overview

- Conduct an exploratory analysis on customer's behavior and sales patterns to find insights and suggest new strategies for marketing and sales teams.
- Develop Python skills to conduct advanced analysis

Project Data

- Project Brief
- Datasets
- Final Report
- Scripts

Techniques Applied

- Data wrangling and subsetting
- Data consistency checks
- Combining data
- Deriving new variables
- Grouping data
- Aggregating variables
- Data visualization
- Excel reporting

Limitations

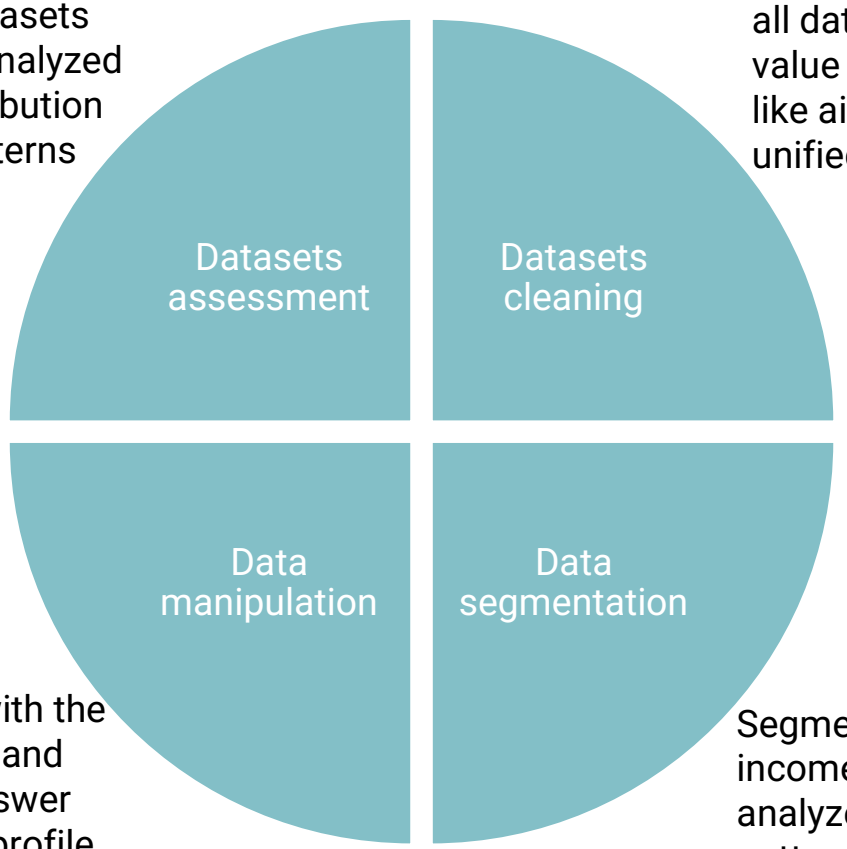
- Data available only since 2017
- Data demographics are limited to income, age and marital state.
- Missing dates in orders.

Tools

Approach and Methodology

Checked for completeness, accuracy and consistency for the 5 datasets needed in this project. Also analyzed summary statistics and distribution graphs to detect unusual patterns

Implemented data cleaning methods across all datasets such as impute any incorrect value and dropping unnecessary columns like aisle_id. Also merged all datasets into a unified one for easier data manipulation.




```
df_final_age_income_vs_dependents = pd.crosstab(df_final['age_income_profile'],
df_final_age_income_vs_dependents
```

number_of_dependants	0	1	2	3
age_income_profile				
high-income adult	3385	3414	3294	3395
high-income senior	2373	2319	2380	2390
high-income young-adult	72	64	84	81
low-income adult	3676	3676	3617	3685
low-income senior	1798	1838	1793	1801
low-income young-adult	3096	2997	3111	3017
mid-income adult	17086	17227	16978	17204
mid-income senior	9487	9517	9578	9455
mid-income young-adult	10628	10477	10647	10566

```
df_final_with_departments = df_final.merge(df_departments, on='department_id', how='left', indicator=True)
df_final_with_departments
```

	product_id	product_name	department_id	prices	order_id	user_id	order_number	order_day_of_week	order_hour_of_day
0	1	Chocolate Sandwich Cookies	19	5.8	3139998	26711	28	6	11
1	1	Chocolate Sandwich Cookies	19	5.8	1977647	33890	30	6	17
2	1	Chocolate Sandwich Cookies	19	5.8	389851	65803	2	0	21
3	1	Chocolate Sandwich Cookies	19	5.8	652770	125935	1	3	13
4	1	Chocolate Sandwich Cookies	19	5.8	1813452	130797	3	4	17

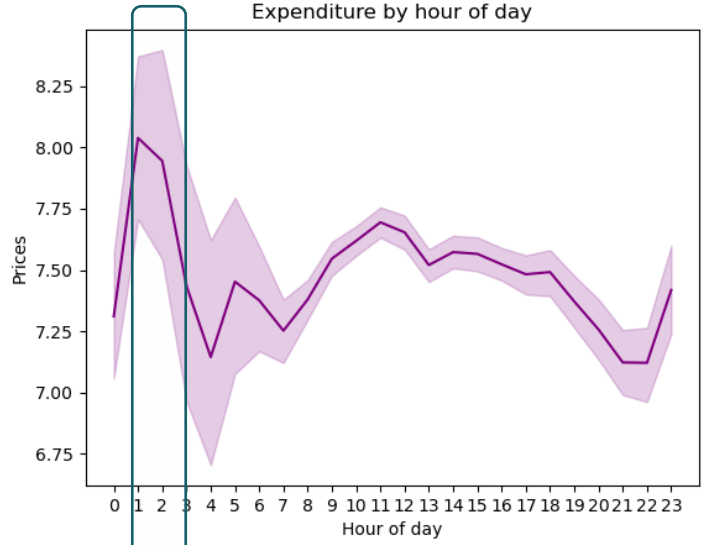
Added new columns to help with the analysis such as income_flag and grouped data to be able to answer business questions, like age_profile.

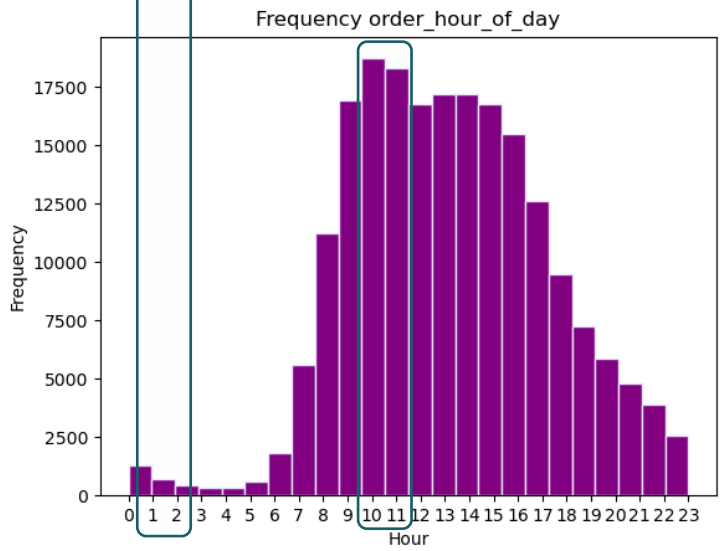
Segmented the datasets into regions, income levels and loyalty and analyzed those segments to find patterns

```
df_final.loc[(df_final['max_order'] < 5), 'activity_type'] = "Low activity"
df_final.loc[(df_final['max_order'] >= 5), 'activity_type'] = "High activity"
```

```
df_final['activity_type'].value_counts(dropna=False)
```

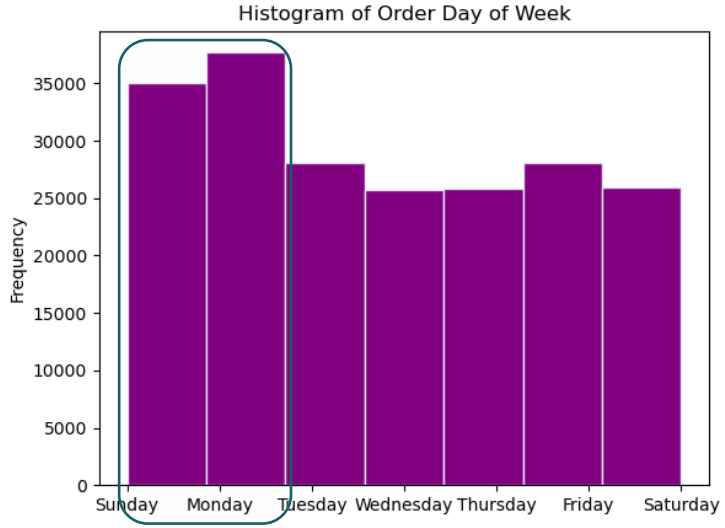
Shopping trends





1am to 3am are the times with a smaller number of orders. It also happens to be the time of the day when customers spend the most money.

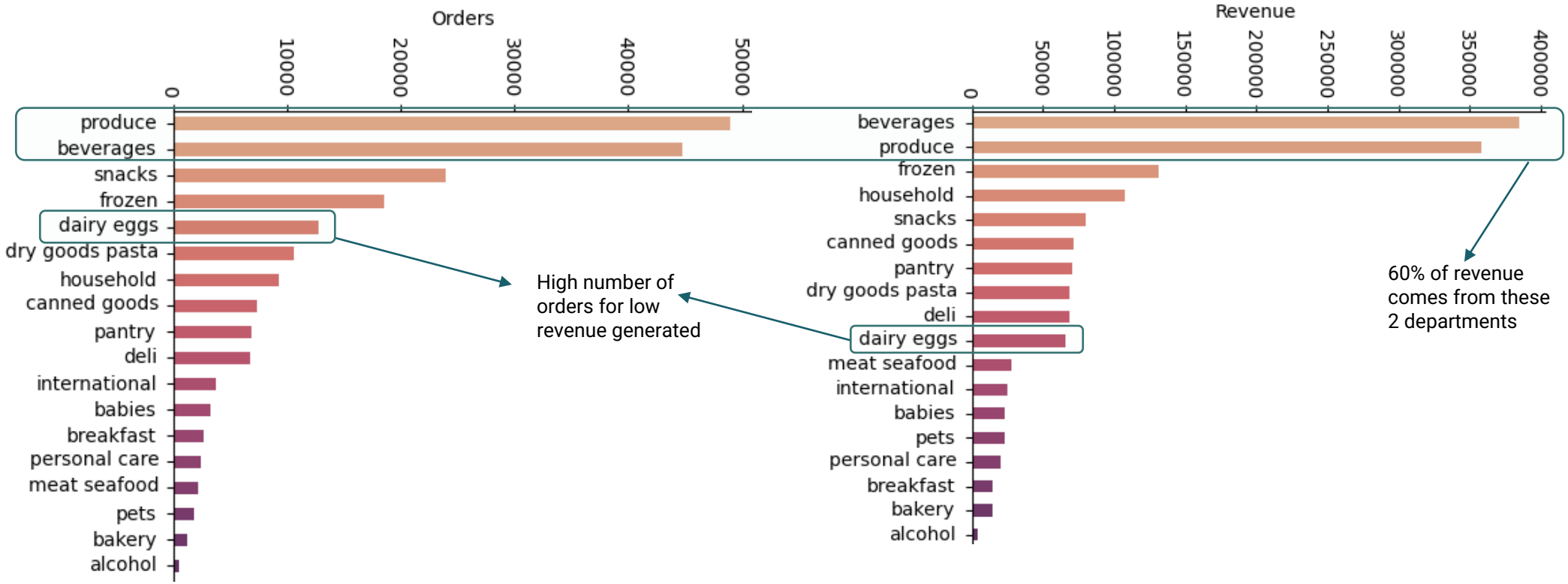
10 am and 11am are the hours of the days with more orders.



Monday and Sunday are the days of the week with more orders

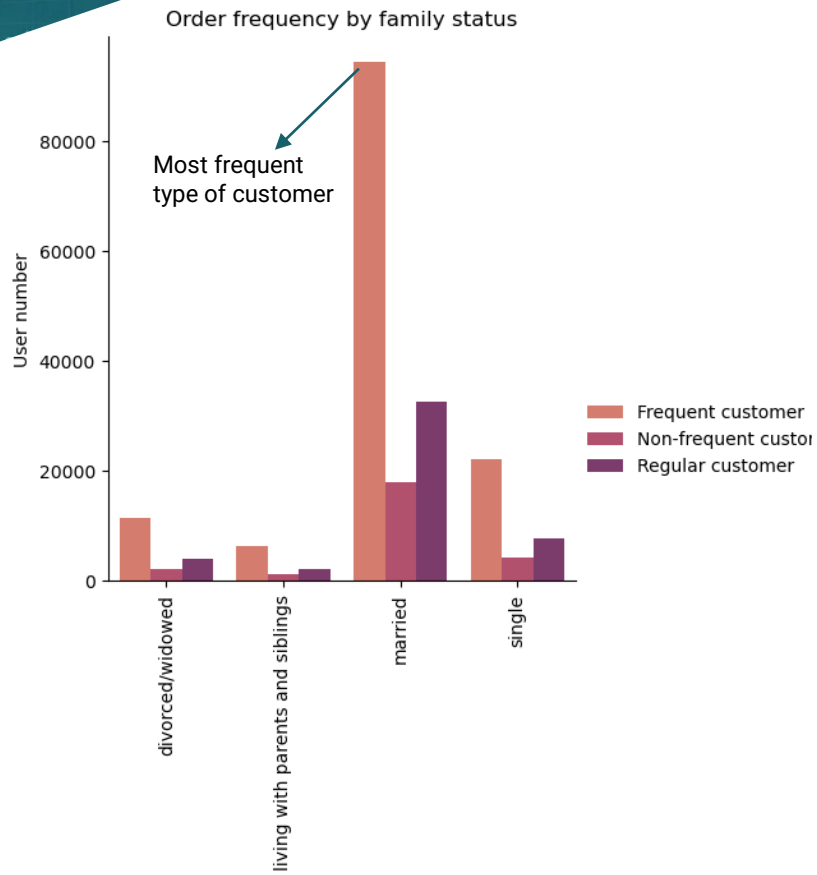
Wednesdays and Thursdays are quieter days

Departments

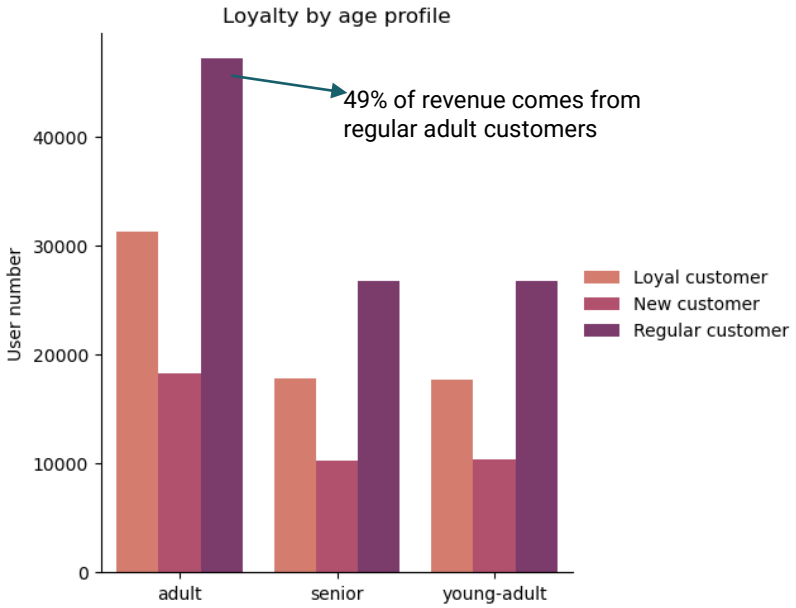


The top 2 departments in both number of orders and revenue are Produce and Beverages

Customer Segmentation



Married folks are the customers with more orders, generating 70% of revenue.
Individuals living with parents and siblings order the least in the platform, generating only 6% of revenue

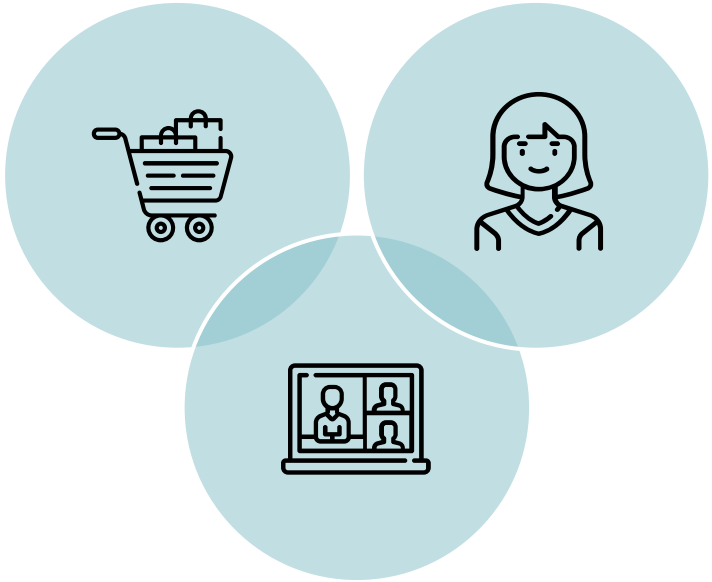


Adults are the type of customers that use the platform the most. They are also the most loyal ones.
Seniors and young adults have about the same numbers for loyalty and customer type.

Customers with income under 180k spend more in purchases than higher incomes.

Recommendations

- Ads to be run 1-3am since they are the slowest but also the times people spend more money.
- Increase ads on least popular products like pantry on weekends between 9am-3pm to leverage fast-moving goods like eggs and beverages.



- Leverage loyalty with a rewards program to incentivize more purchases.
- Create a line of luxury products to attract high-income customers.

- Bulk items are the lowest sellers, explore partnership with stores that sell bulk items.
- Further analysis in young-adults' interests since they are the lowest customer base but have high potential for sales.

6

Pig E. Bank

Customer retention analysis for a finance company

Project Overview

- Conduct an exploratory analysis on customer's behavior and sales patterns to find insights and suggest new strategies for marketing and sales teams.
- Develop Python skills to conduct advanced analysis

Project Data

- Project Brief
- Final Report

Techniques Applied

- Data ethics
- Predictive analysis
- Data mining
- Time series analysis
- Grouping Forecasting

Limitations

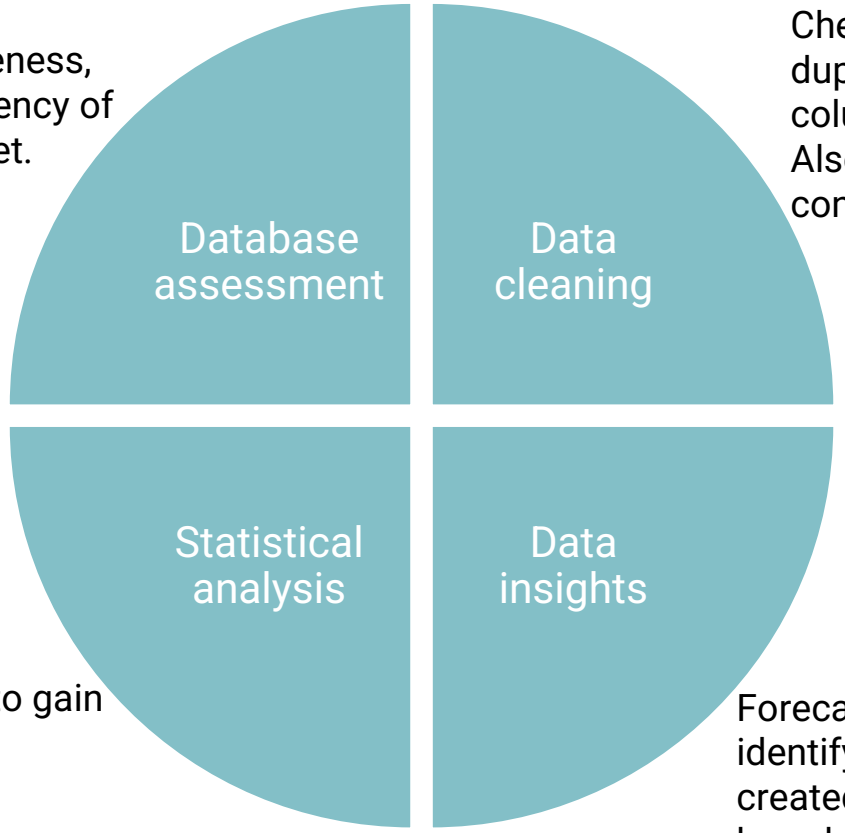
- Customer demographics are limited to country, gender, age and salary.
- Customer bank information is limited to account balance, membership status and number of products.
- No transaction data is available for analysis.

Tools

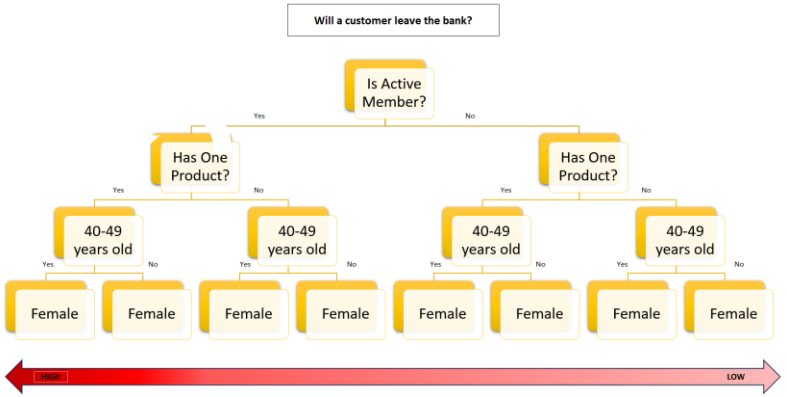
Approach and Methodology

Checked for completeness, accuracy and consistency of the Pig E. Bank dataset.

Checked for missing values and duplicates besides standardizing columns like gender and country. Also updated names of columns for consistency



Top factors for churn
Is Active Member
# of Products
High income
Gender



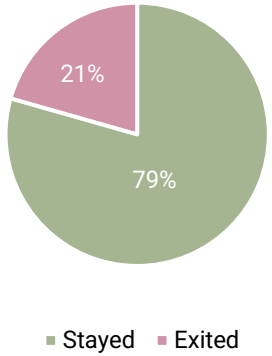
Applied summary statistics to gain insights on available data.

Forecasted customer exiting the bank, identifying key predictors. Also created a decision tree after factors have been found.

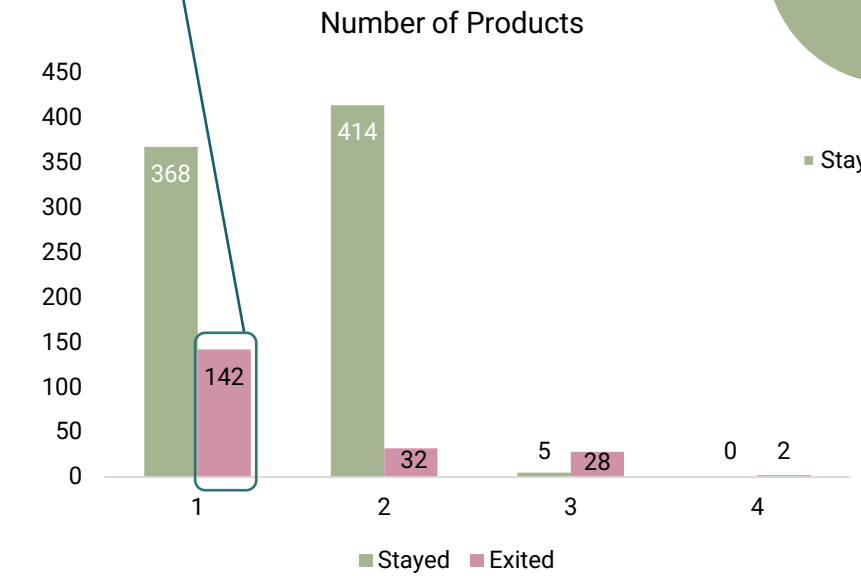
Report

Customer Retention Analysis

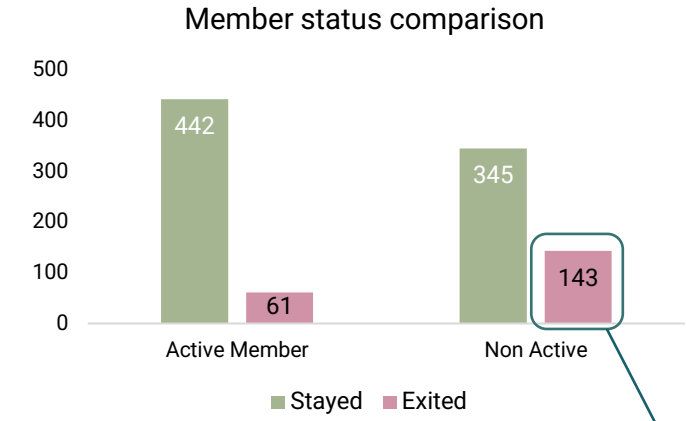
21% of customers left the bank



28% of customers with 1 product have churned

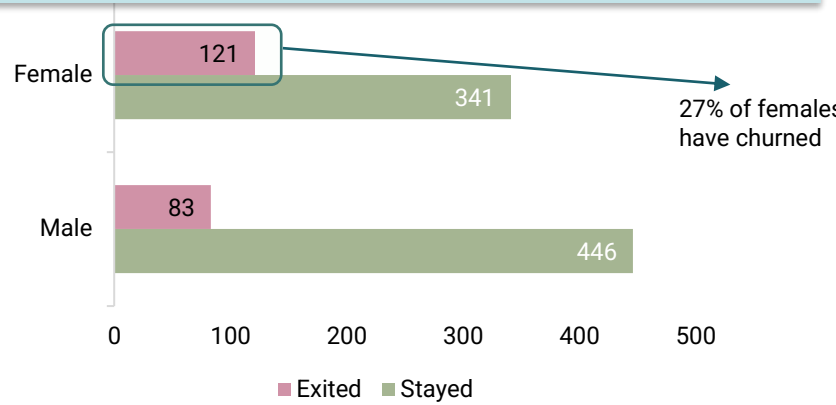


70% of churned customers had only 1 product



30% of non active members have churned

60% of churned customers are female

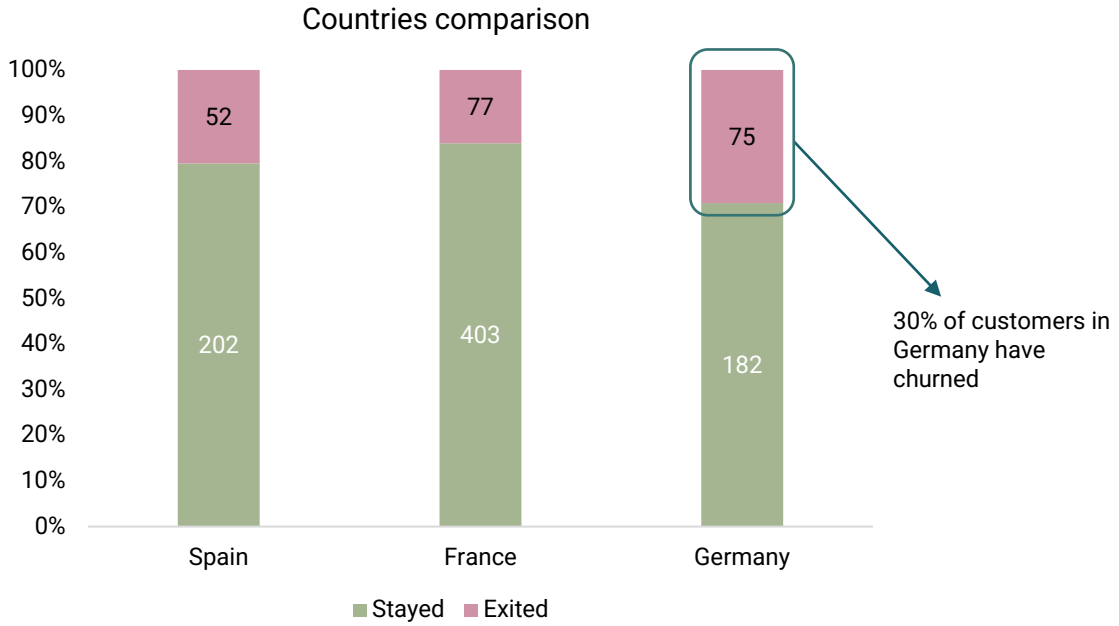


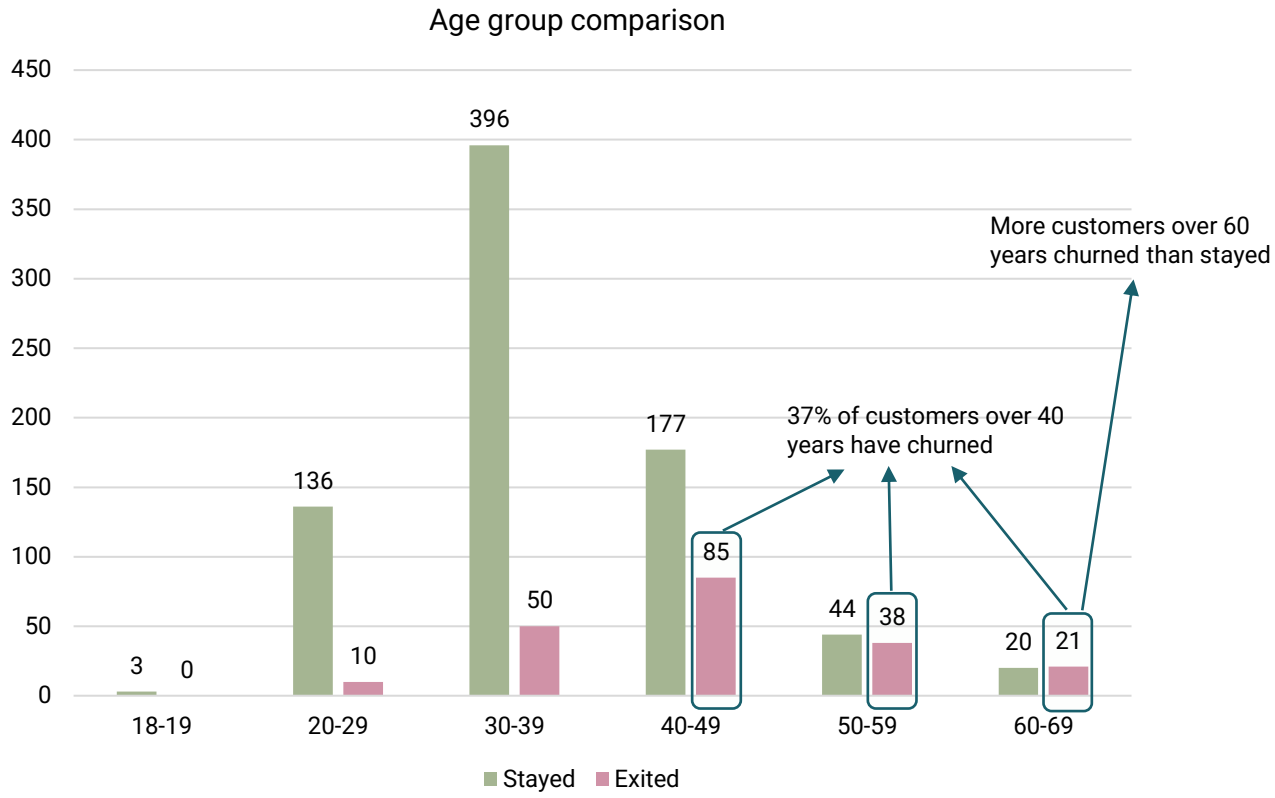
27% of females have churned

70% of churned customers had only 1 product

Customer Retention Analysis

Germany has lost the most customers

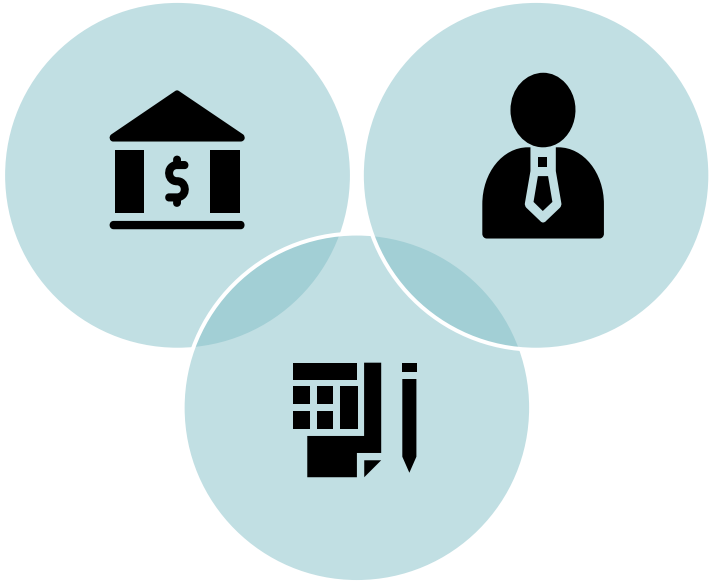




Customers over 40 years old are churning

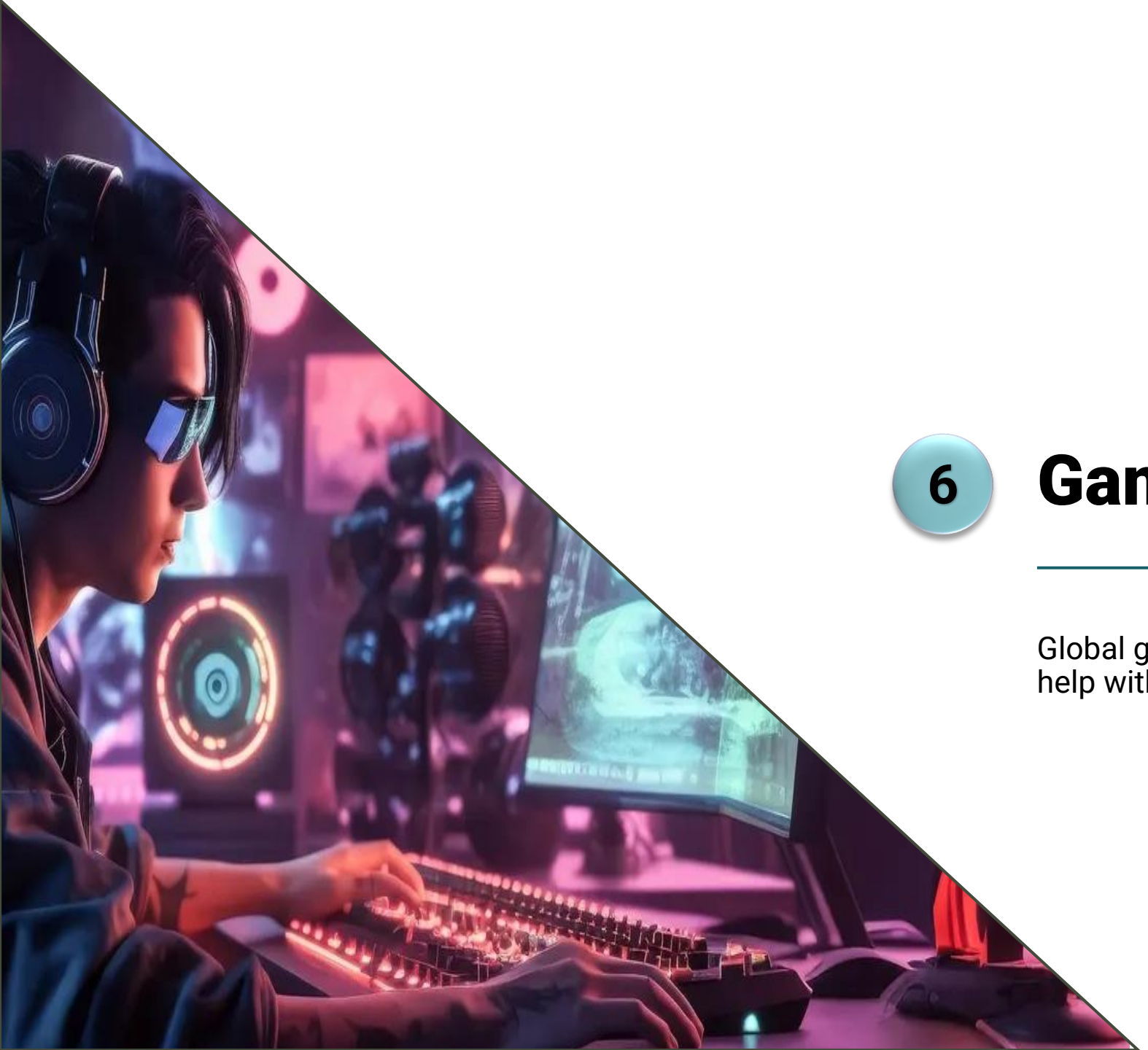
Recommendations

- Increase interaction so members stay active with the bank.
- Re-engage with inactive customers.



- Increase interaction with females and customers over 40 years old.
- Run surveys to understand their needs.

- Run targeted campaigns for customers with only one product.
- Showcase additional bank offerings with exclusive incentives.



6

GameCo

Global gaming market analysis to help with business initiatives

Project Overview

- Perform a descriptive analysis to find insights on the current gaming market
- Inform product development and sales about new strategies

Project Data

- Project Brief
- Dataset
- Presentation
- Report

Techniques Applied

- Understanding datasets
- Cleaning data
- Grouping and summarizing data
- Analyze data and develop insights
- Conduct descriptive analysis
- Visualize data insights
- Storytelling with data

Limitations

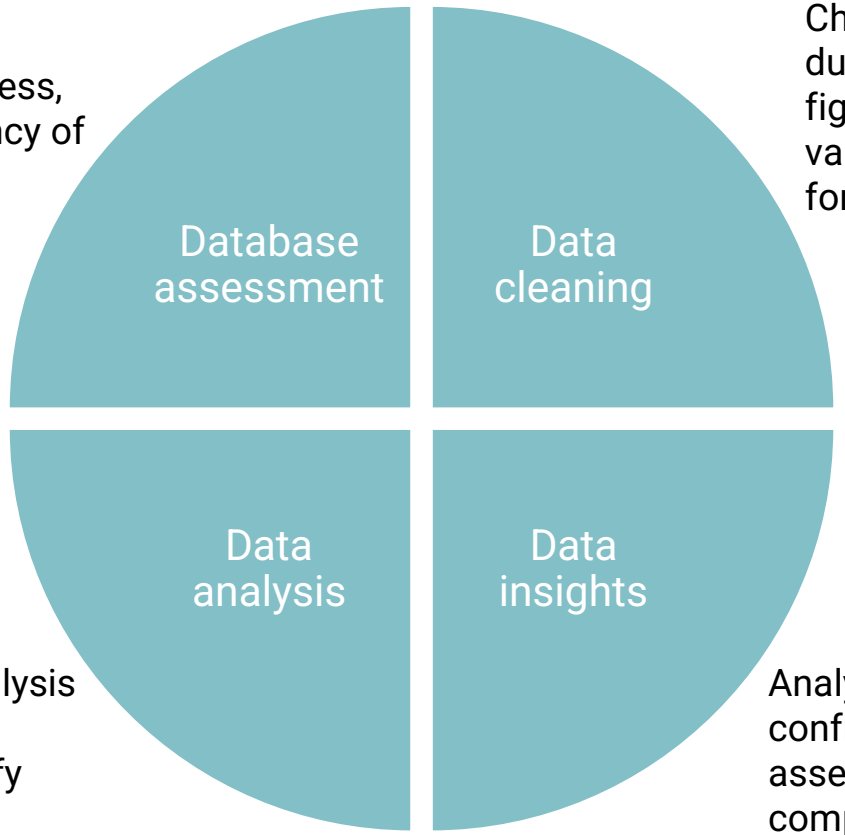
- Data available only for physical units sold.
- No data after year 2016
- No streaming data.

Tools

Approach and Methodology

Checked for completeness, accuracy and consistency of the data.

	Mean	Median	Mode			
NA_Sales	0.264907	0.08	0		Q1 NA_Sales	0
EU_Sales	0.146785	0.02	0		Q2 NA_Sales	0.08
JP_Sales	0.077821	0	0		Q3 NA_Sales	0.24
Other_Sal	0.04956	0.01	0		IQR lower side	-0.36
Global_Sa	0.537895	0.17	0.02		IQR higher side	0.6



Checked for missing values and duplicates. For example, annual sales figures across regions has missing values. Also updated column names for consistency

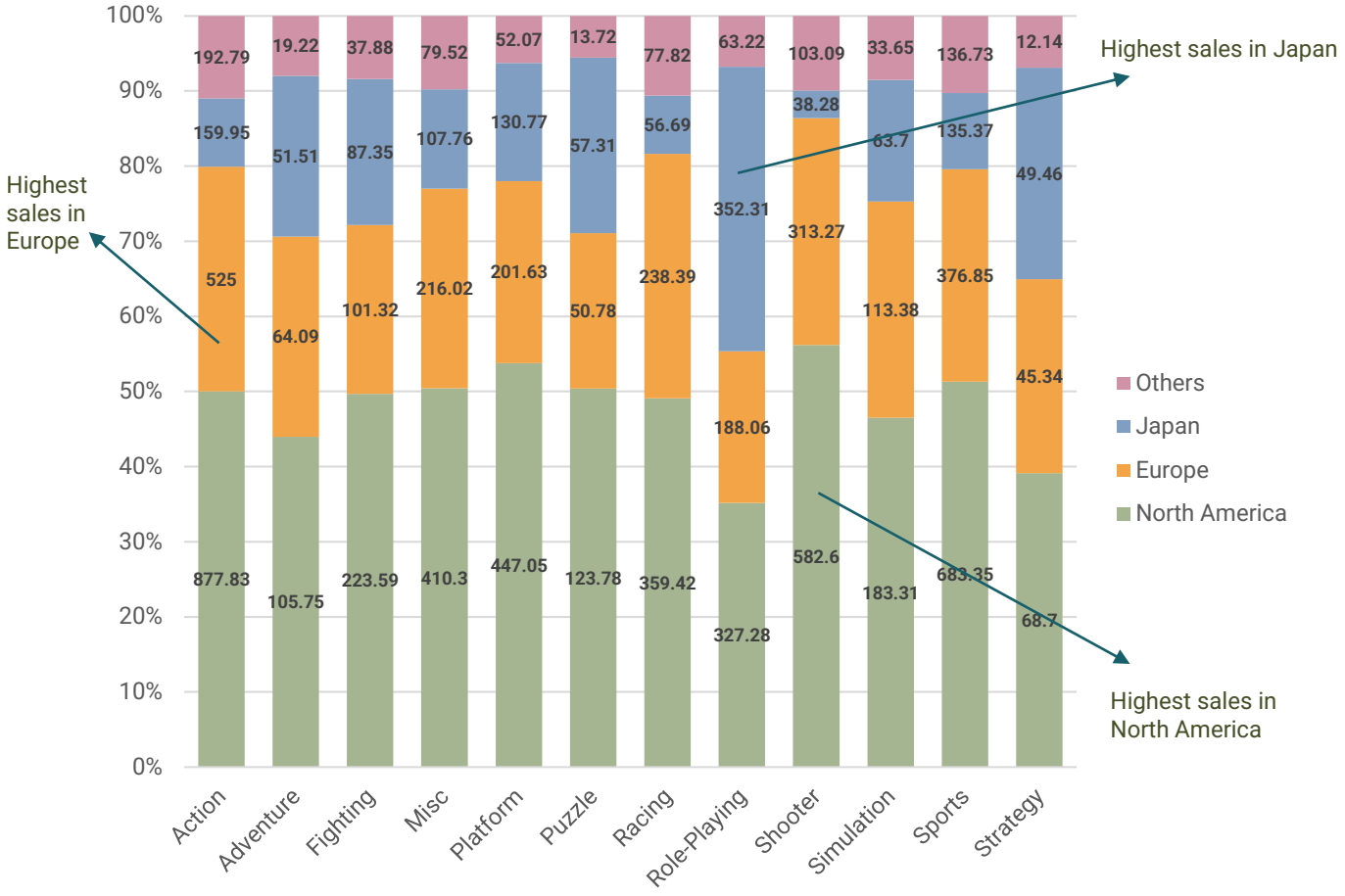
Performed a descriptive analysis like central tendency and distribution graphs to identify outliers and sales range.

Analyzed the cleaned dataset to confirm regional sales trends and assessed leading genres, games and competitors.

Regional Sales Trends

- Global Sales are in decline since 2008.
- North American sales have been trending down.
- Meanwhile sales in Europe and Japan have gone up.
- Europe takes over North America for highest sales in 2016

Regional Genre and Game Trends



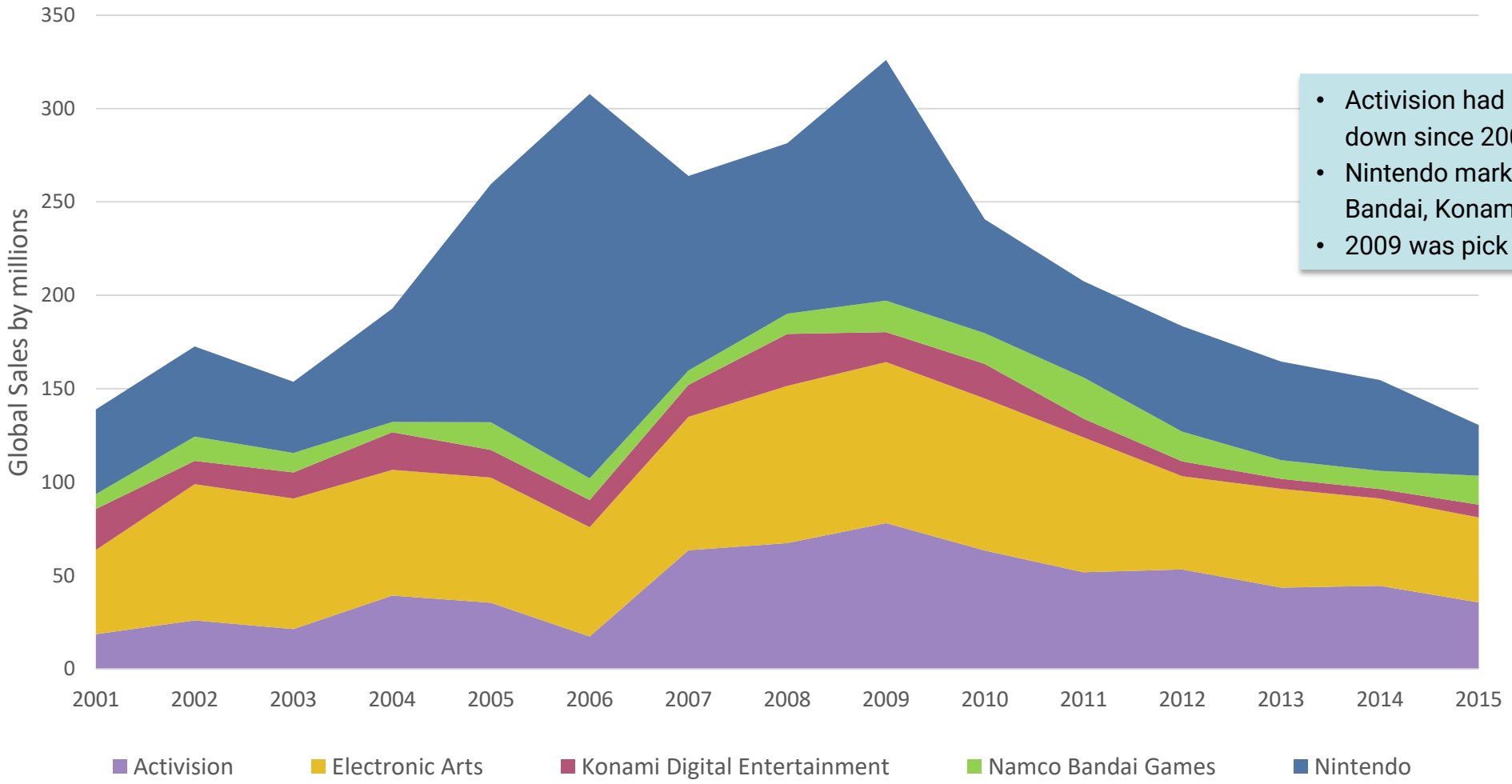
- North America and Europe highest sales: Action, Sports and Shooter games.
- Japan highest sales: Role-playing, Action and Sports.

- North America and Europe game with highest sales is Wii Sports
- Japan's game with highest sales is Pokémon Red/Blue

Sales in millions by game

Game	North America	Europe	Japan
Wii Sports	42	30	4
Grand Theft Auto V	23	24	2
Super Mario Bros	32	5	7
Mario Kart Wii	16	13	4
Pokémon Red/Blue	11	9	10
Duck Hunt	27	1	1
Pokémon Gold/Silver	9	6	7

Regional Publishers Preferences



- Activision had steady growth but has been trending down since 2009.
- Nintendo market share was taking over by Namco Bandai, Konami Digital and Electronic Arts in 2007
- 2009 was pick sales for the top 5 publishers

Recommendations

- Increase stock for Nintendo games, specially increase stock for Mario Bros and Pokémon games

- Reallocation of marketing budget as follow
  - Europe 35%
  - Japan 25%
  - North America 40%
- Start analysis to understand digital trends

- Focus budget on Action, Sports and Shooter genres for Europe and North America
- Focus budget for role-playing and action for Japan

Thanks!

Esther Howard