Data Analytics Portfolio

Deepa Sury

Data Analytics Projects

GameCo International Marketing Budget 2017	Analyzing global video game sales from 1980 to 2016 to inform marketing strategies for the upcoming year.
Influenza Prevention Report 2018	Using historical data to prepare for upcoming influenza season.
Rockbuster Stealth LLC Launch Strategy	Gathering insights on customer behaviour and preferences to adapt to the current video streaming market.
Instacart Grocery Basket Analysis	Determining customer profile groups for targeted marketing strategies.
Pig E. Bank – Customer Retention	Data-mining analysis to improve customer retention at global financial institution.
Airbnb Amsterdam 2019	Exploratory analysis of Airbnb listings in Amsterdam for the year 2019.
ClimateWins Weather Prediction	Using machine-learning techniques to predict and classify changing climate conditions in Europe.

GameCo: International Marketing Strategy 2017

Project Goals

GameCo is a video game company interested in using insights from data to inform the development of new games as well as to optimize their marketing campaigns. This project aims to answer key questions such as the geographical distribution of video game sales, the popularity of certain game titles and genres, and which publishers are likely to be the main competitors in certain markets.

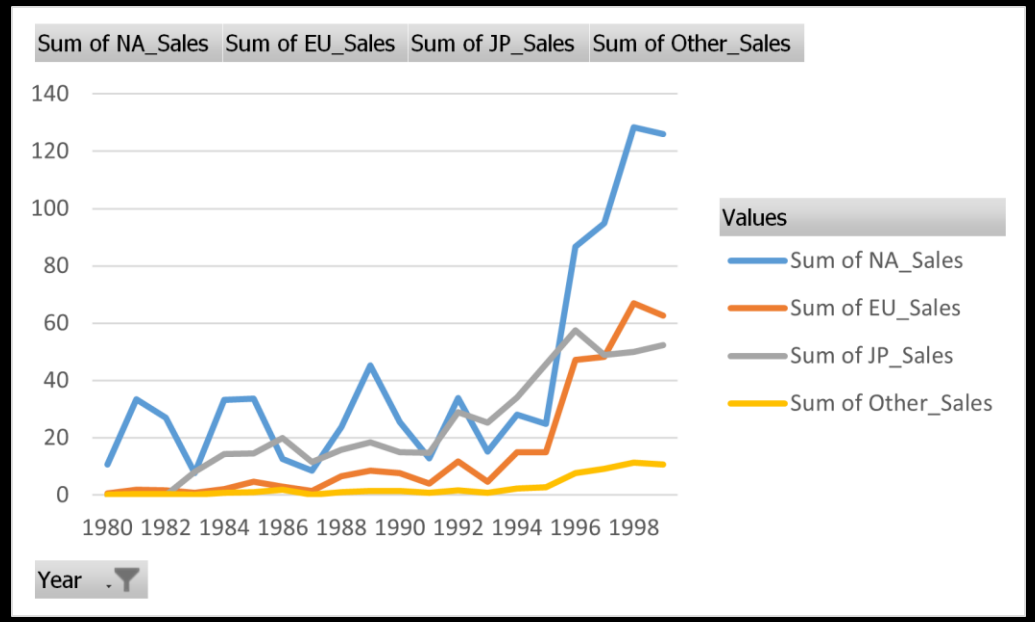
Data Used

The data used for this project was collected from VGChartz and consists of historical sales data of video games across the world from 1980 to 2016. The data-set contains information on games across a variety of genres, platforms and publishers, having sold 10,000 or more copies.

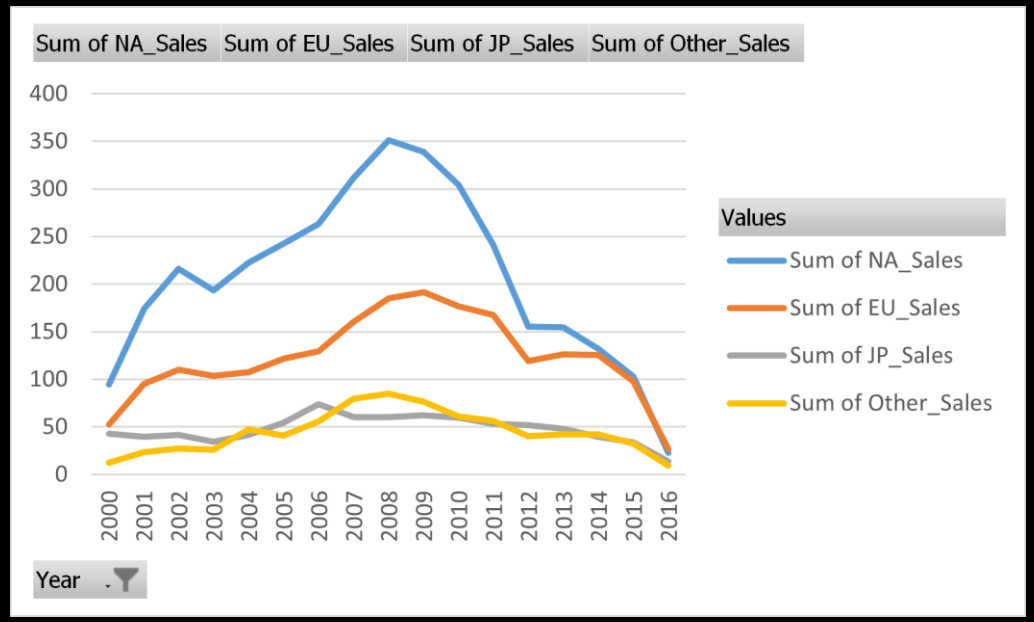
Methods Used

- Basic data sorting, cleaning and filtering techniques in Microsoft Excel
- Pivot Tables
- Descriptive analysis
- Data visualization with Excel

Analysis

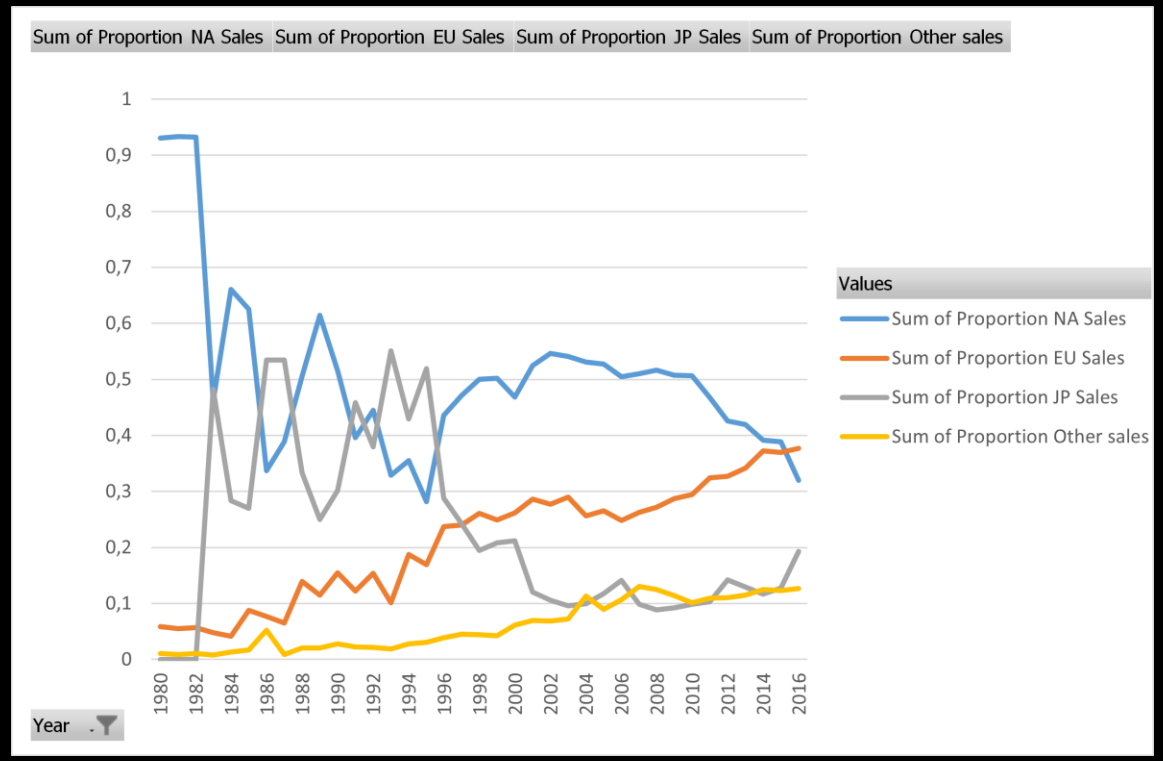


Total annual video game sales by region, 1980-1999.
Sales in all regions except Japan sharply increased in the mid 1990s.
EU sales begin to surpass Japanese sales after 1996.

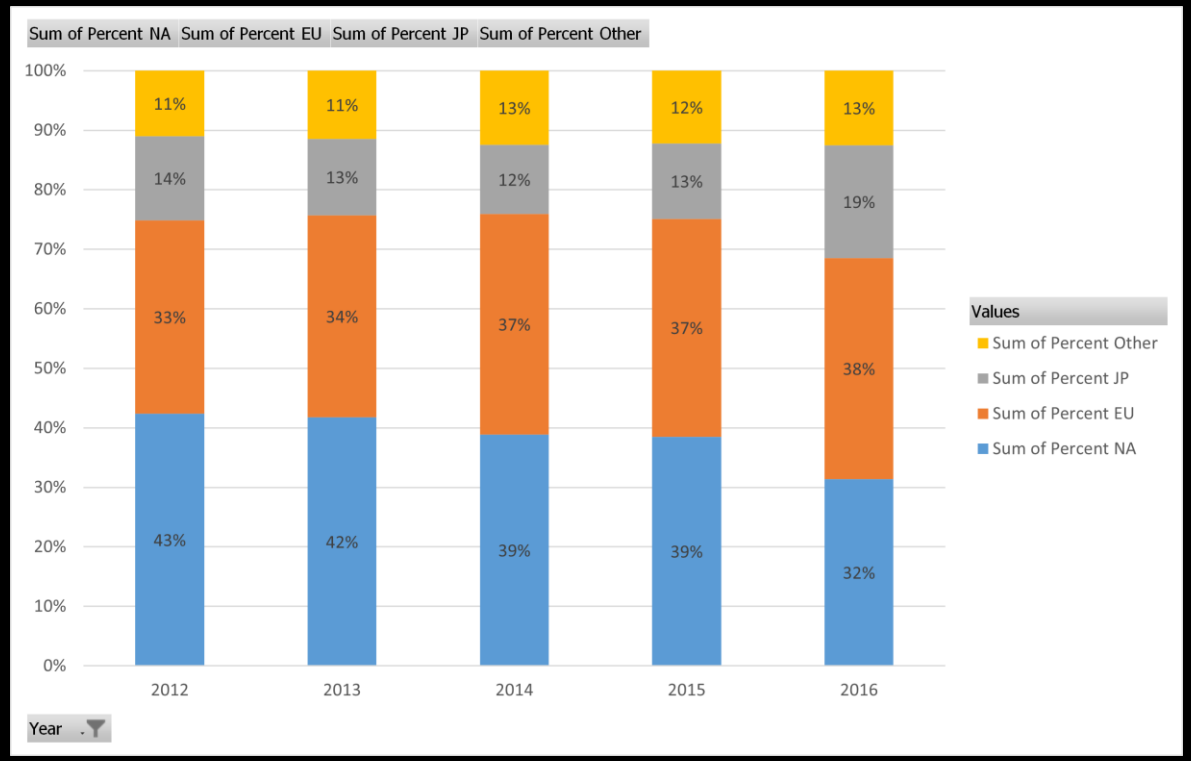


Total annual video game sales by region, 2000-2016.
Sales in all regions except Japan continue to increase until 2008.
Japanese sales remain relatively stagnant, at one point showing the lowest numbers out of all regions.
From 2015-2016, sales across all regions appear to be decreasing sharply.

Analysis



Regional proportion of global video game sales, 1980-2016.
Between 1980-1982, North America dominated global sales.
The proportion of Japanese sales peaked in 1982, 1986-1988 and at various points in the 1990s.
EU and other regions have shown a steady overall increase in their share of global video game sales.



Regional distribution of global video game sales, 2012-2016.
Japanese proportion of global sales increased 6% in 2016.
Proportion of EU sales have steadily increased, proportion of North American sales have steadily decreased.
Proportion of sales from other regions remains relatively constant.

Insights

The largest share of sales comes from Europe (38%) and not North America (32%)

The biggest changes to the regional makeup of global video game sales can be seen in the mid-2000s.

As in previous years, European sales, as well as sales from other parts of the world, continue to steadily increase.

Japanese sales, after several years of decline and stagnation, have begun to increase once again.

Sales of video games are challenged across regions by the rise of digitally-downloaded software.

Recommendations

- Allocate more of the budget towards overseas marketing.
- Focus especially on regions outside of North America, Europe and Japan (listed under the “Other” category in the sales data), as many countries known to be emerging markets are in these regions.
- Develop strategies to revive customer interest in physical video game media.
- Focus these efforts on the North American market at first, as it continues to make up a significant enough share of global sales, despite declining numbers in recent years.

Influenza Prevention Measures for 2018

Project Goals

- The objective of this project is to help a medical staffing agency allocate temporary workers to hospitals ahead of the upcoming influenza season.
- Using insights from historical influenza data, this project helps identify areas that are likely to need additional staffing with factors such as high rates of mortality and large numbers of inhabitants aged 75 and older.

Data Used

- Influenza deaths by geography, 2009-2017 (CDC).
- US Census data by geography, time, age and gender (US Census Bureau).

Methods Used

Data profiling, checking data for integrity and applying additional quality measures.

Transforming and integrating data-sets from different sources (using tools such as Microsoft Excel vlookup and other functions).

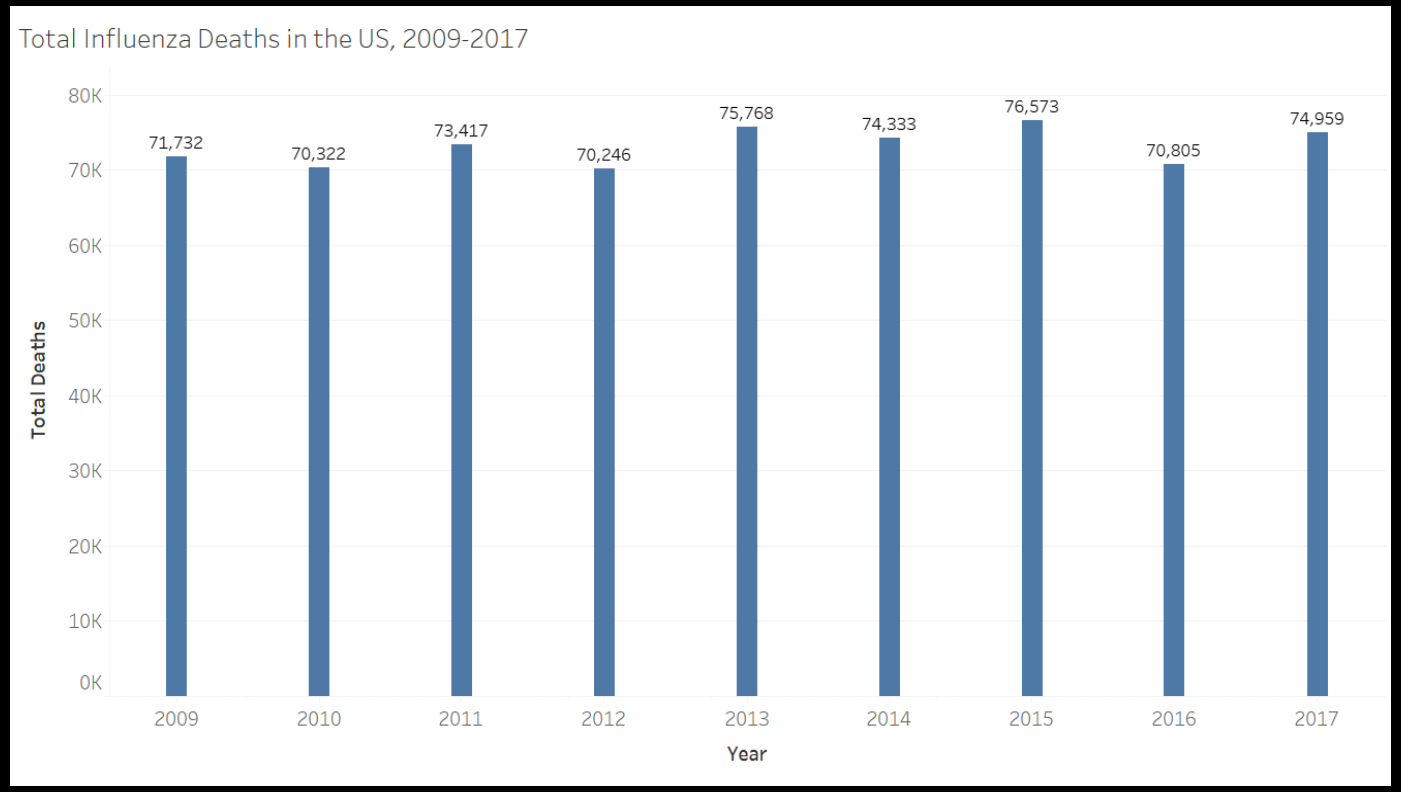
Conducting statistical analyses (calculating variance and standard deviation, testing for correlation).

Formulating and testing statistical hypotheses.

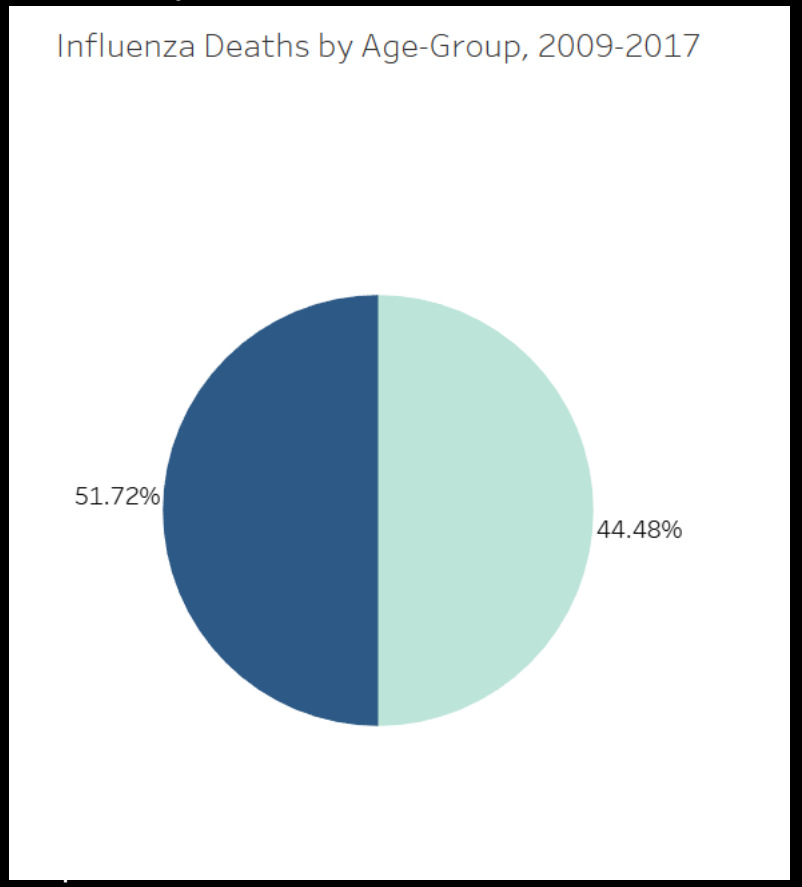
Data visualization and storytelling using Tableau.

Analysis

Full visual presentation here: https://public.tableau.com/app/profile/d.s5411/viz/Exercise2_9_17259903842140/Story1



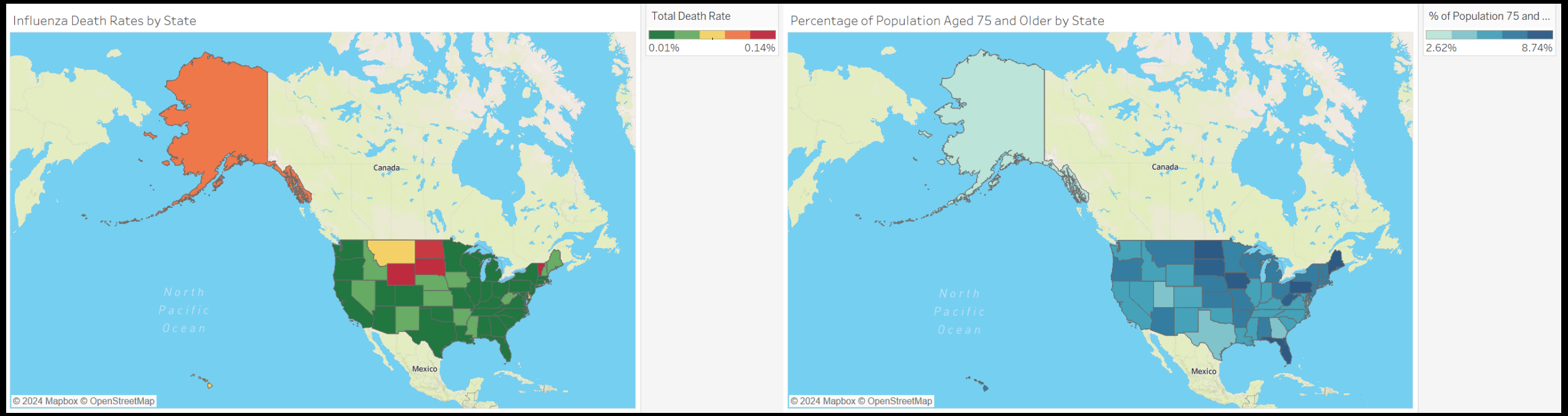
Total influenza-related deaths per year in the US, 2009-2017.



persons aged 75 and older (darker portion).

Analysis

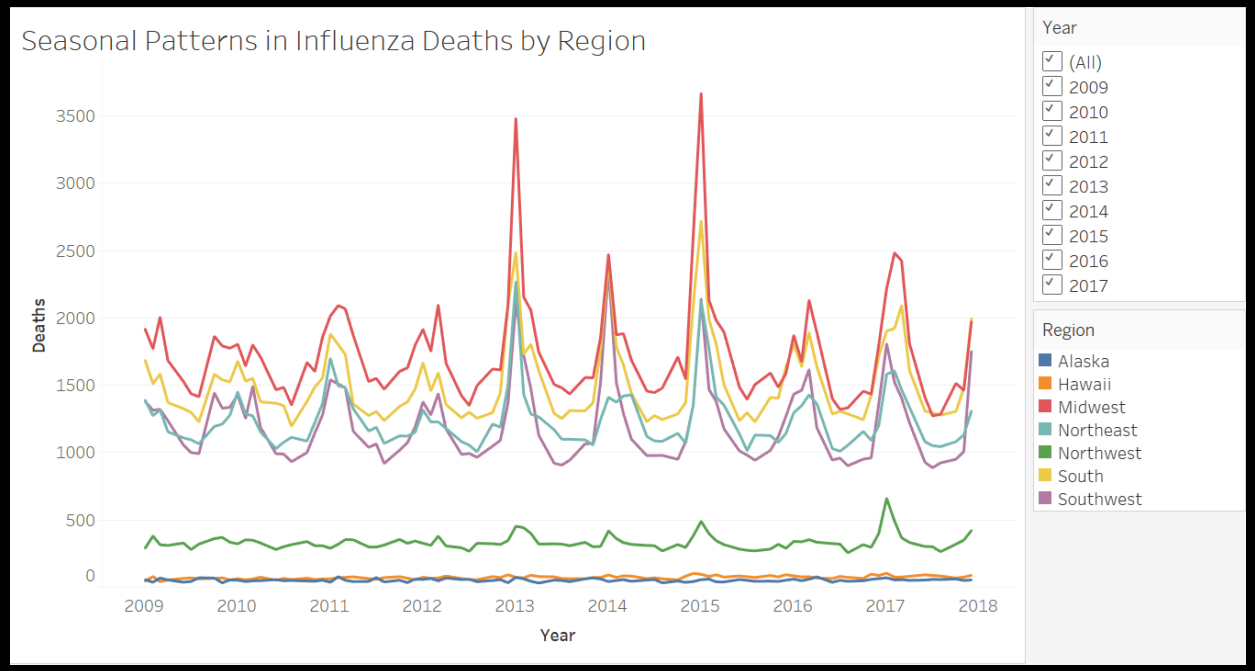
Full visual presentation here: https://public.tableau.com/app/profile/d.s5411/viz/Exercise2_9_17259903842140/Story1



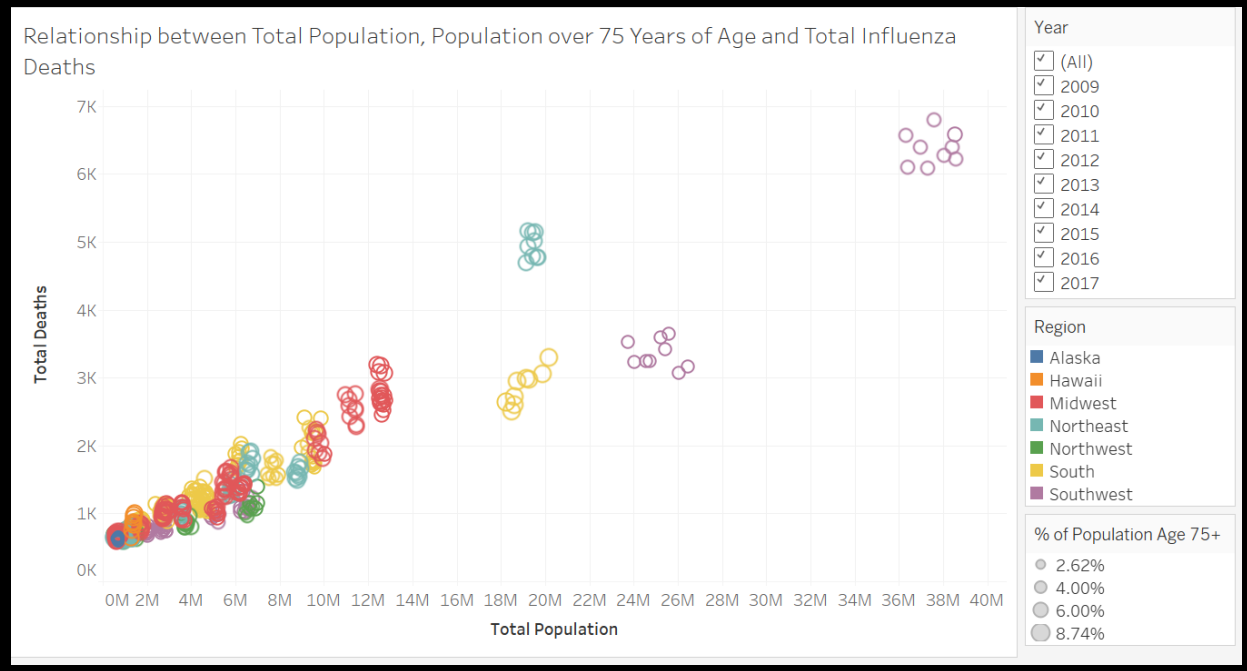
Maps depicting the distribution of death rates and percentage of the population aged 75 and older across states.
These maps aim to visualize regions which are more likely to be affected by seasonal influenza and where more medical staff should be allocated during influenza season.

Analysis

Full visual presentation here: https://public.tableau.com/app/profile/d.s5411/viz/Exercise2_9_17259903842140/Story1



Monthly influenza deaths by region, 2009-2017.
This chart shows the seasonal pattern in the numbers of influenza deaths over the years.



Comparing the total number of influenza deaths and the percentage of the population aged 75 and over between 2009 and 2017.

Insights

Individuals aged 75 and older accounted for just over half of all influenza deaths between 2009 and 2017.

States with the largest populations also had the highest number of deaths.

States with the highest death rates (number of deaths divided by the population) also showed the largest percentage of inhabitants aged 75 and older.

Large populations and a high percentage of elderly inhabitants can both significantly increase a region's vulnerability to seasonal influenza.

Recommendations

- More medical staff should be allocated to regions with higher death rates (Alaska, Vermont and certain states in the Midwest).
- Highly-populated states like New York and California showed the highest death counts. Because influenza is a highly-transmissible disease, these states should have more medical staff on hand to deal with the higher number of severe infections.
- Promotion of preventative hygiene measures, such as routine vaccination, hand-washing and masking, aimed at the elderly and anyone in close contact with them.

Rockbuster Stealth LLC Launch Strategy

Objective

- Assist Rockbuster LLC with the launch strategy of their online streaming service.
- Use data to determine which titles contributed to revenue gain, how long movies are rented on average, sales figures and customer numbers across geographic regions.

Data Used

Rockbuster Stealth LLC data on customers, inventory, sales and payments contained in a RDBMS (relational database management system) and analyzed using SQL.

Methods Used

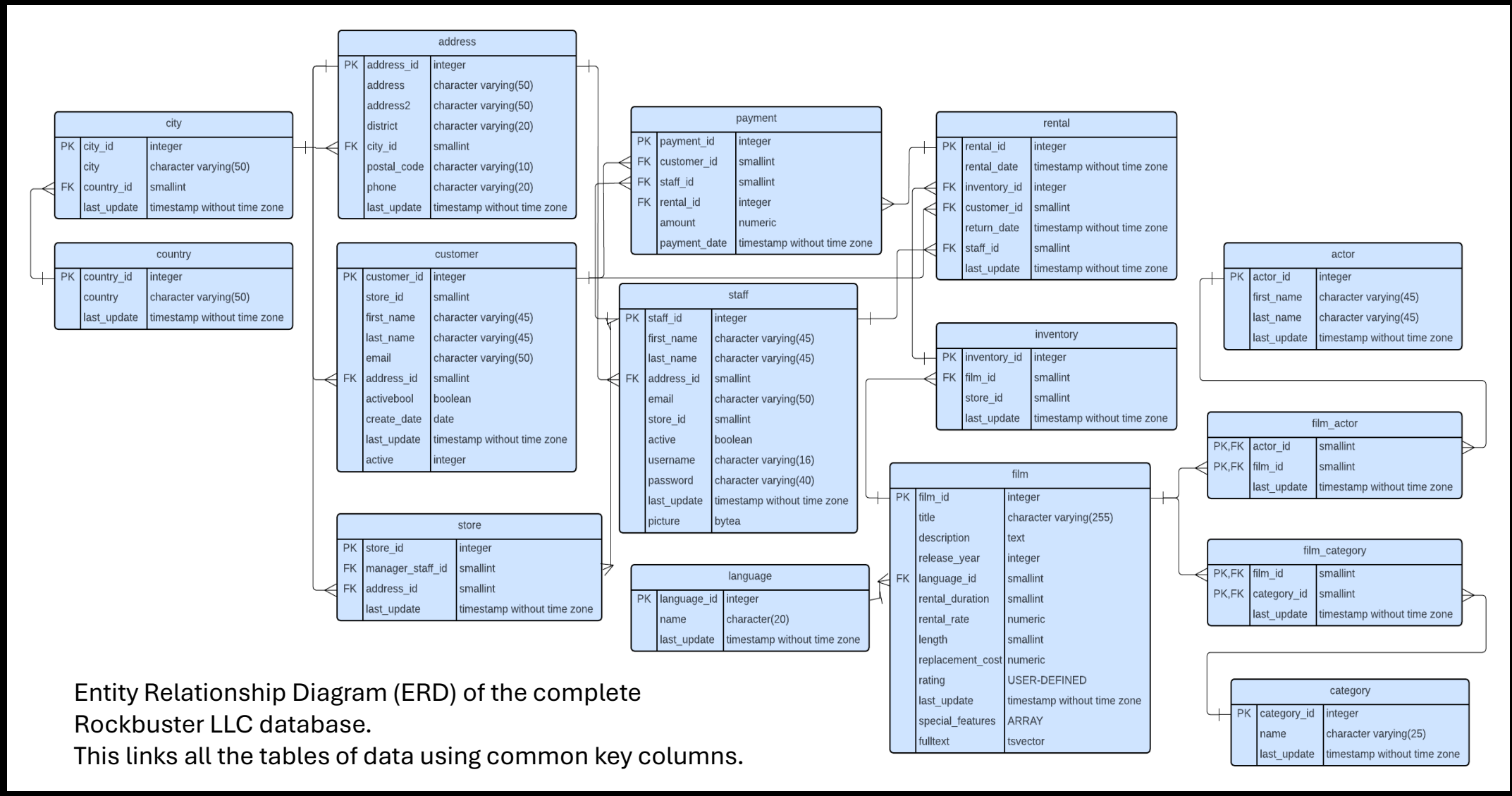
Extracted an entity relationship diagram (ERD).

Database querying, filtering, summarizing and cleaning data; joining tables, performing subqueries and using common table expressions (CTEs) with SQL using PostgreSQL.

Visualizing insights using Tableau.

SQL queries and data dictionary can be viewed on GitHub.

Entity Relationship Diagram (ERD)



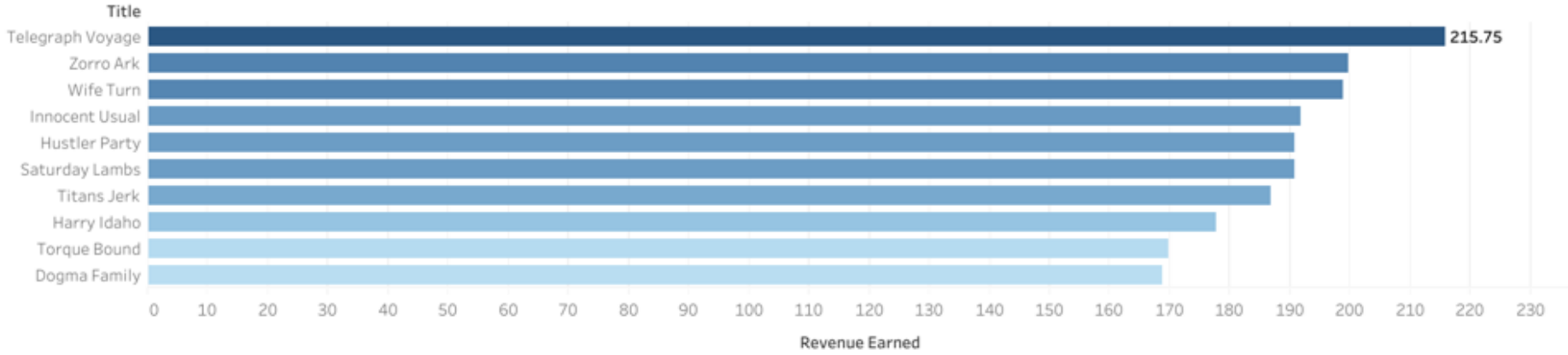
Entity Relationship Diagram (ERD) of the complete Rockbuster LLC database.
This links all the tables of data using common key columns.

Analysis

Highest Earning Film Genres

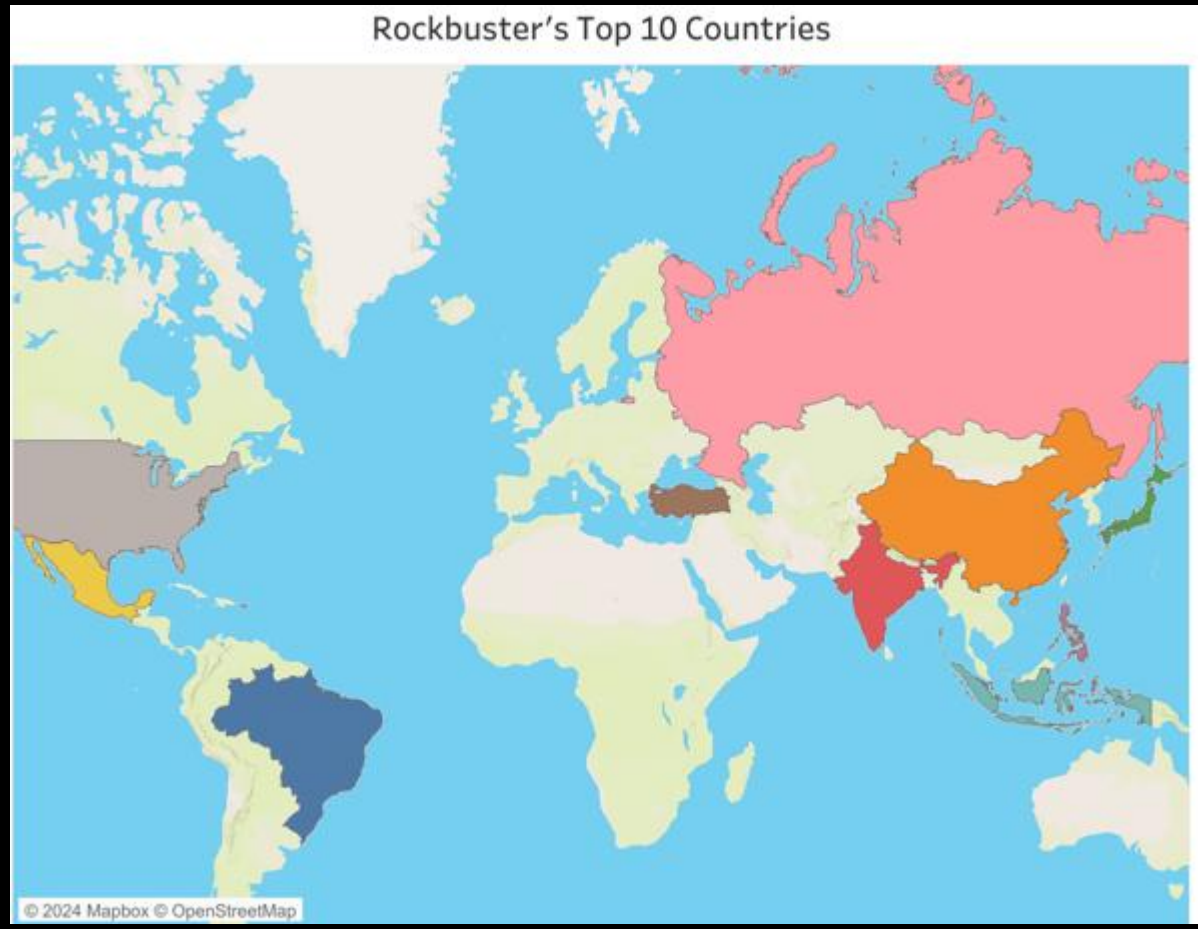
Sports films earned the highest amount of revenue at Rockbuster,
followed by **sci-fi, animation** and **drama.**

Highest Earning Titles



Rockbuster's highest-earning title is the music film **Telegraph Voyage,**
followed by the comedy **Zorro Ark** and the documentary **Wife Turn.**

Analysis



The top 10 countries with the most Rockbuster customers are: India, China, United States, Japan, Mexico, Brazil, Russian Federation, Philippines, Turkey and Indonesia.

The top 10 cities within the top 10 countries with the most Rockbuster customers are: Aurora (United States), Acua (Mexico), Citrus Heights (United States), Iwaki (Japan), Ambattur (India), Shanwei (China), So Leopoldo (Brazil), Teboksary (Russian Federation), Tianjin (China) and Cianjur (Indonesia).

Insights

Rockbuster generated a total of 61312.04 in rental revenue.

Rockbuster's customers can be found in 109 countries around the world.

India is home to the largest number of Rockbuster customers (60).

Customers have an inventory of 1000 films to choose from.

Sports is the best-performing film genre, followed by sci-fi, animation and drama.

Films are rented from Rockbuster for an average of 4.99 days, at an average rate of 2.98 per day.

Recommendations

- Expand the film inventory to include films in languages other than English, as this is a crucial step towards establishing and maintaining an international customer base.
- Research the customer preferences within the top 10 countries and develop localization strategies based on these findings.
- Develop and implement customer loyalty programs to encourage more frequent rentals and to reward existing top customers.

Instacart Grocery Basket Analysis

Objective

- Derive more information about Instacart sales patterns through the analysis of customer, order and product data.
- Inform decisions regarding customer profiling and market segmentation.

Data Used

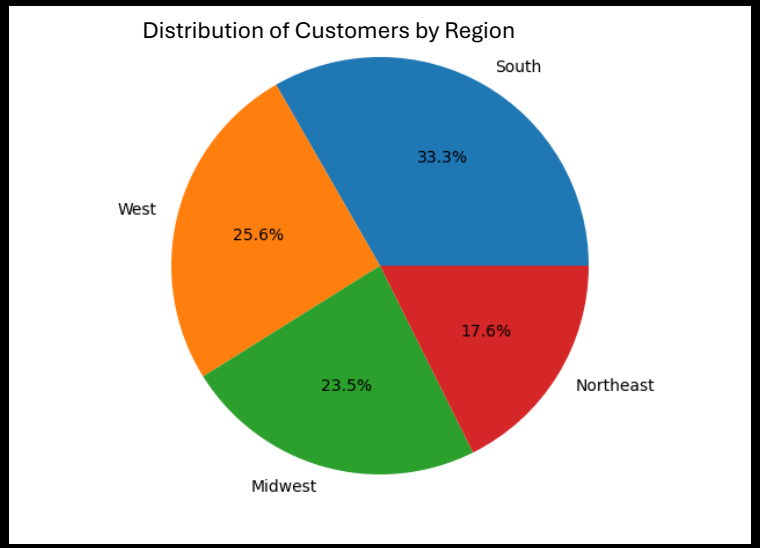
"The Instacart Online Grocery Shopping Dataset 2017", Accessed from www.instacart.com/datasets/grocery-shopping-2017 via Kaggle on 25 November 2024.
Note that, while Instacart is a real company, the data used for this project is fictional.

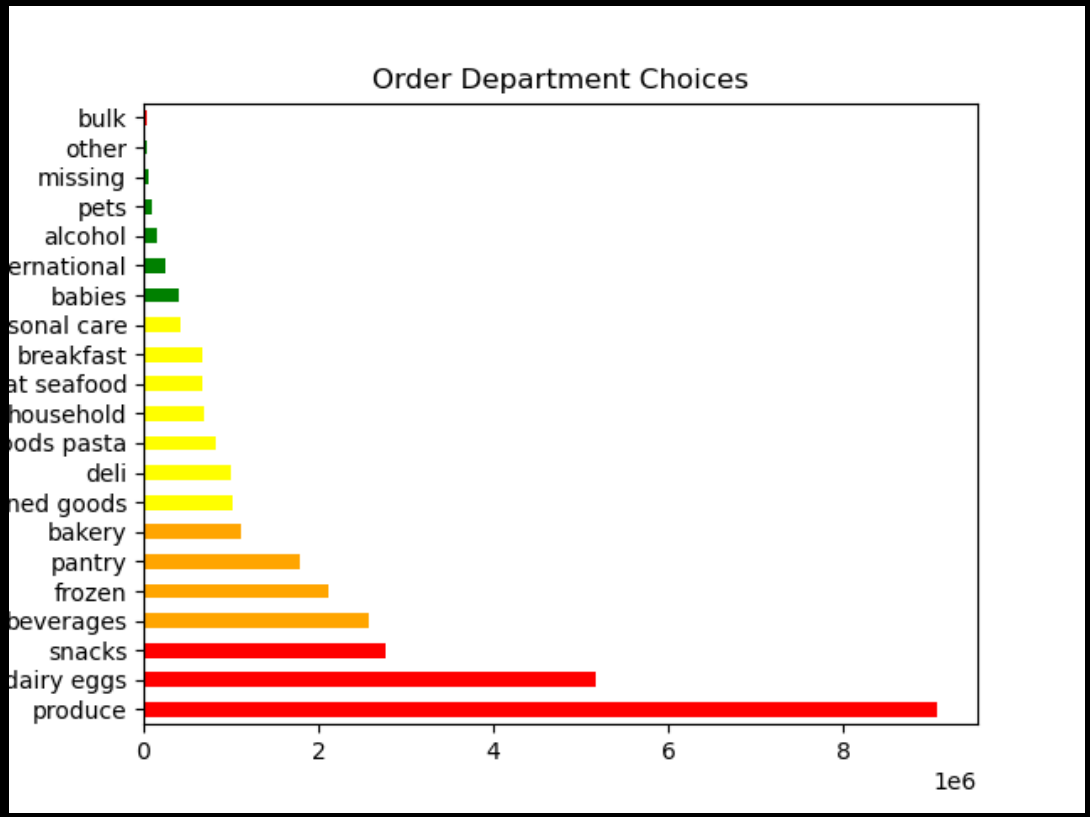
Methods Used

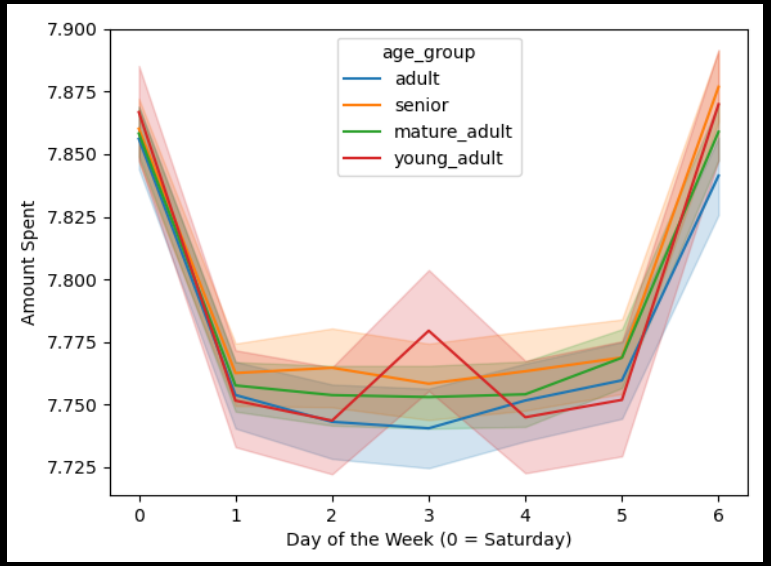
- Data wrangling, subsetting, merging, grouping, aggregating; checking for consistency and deriving new variables, using Python with pandas, NumPy and os libraries.
- Data visualization using Python with matplotlib, scipy and seaborn libraries.

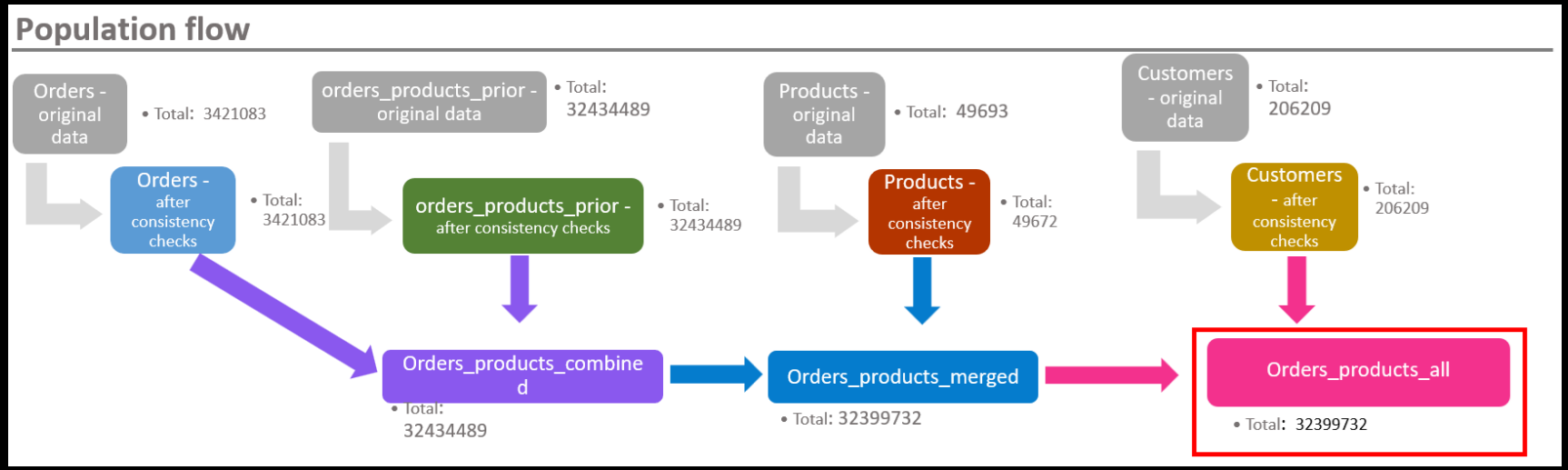
Analysis

Some examples of visuals created for this project.
Full project, including scripts and visuals, can be viewed on GitHub









Insights

Instacart receives the highest number of orders between 10:00 AM and 4:00 PM, and on weekends.

This can be observed across all customer profiles.

Customers spend the most money in the evening and at night.

Perishable grocery items, such as produce and dairy/eggs, are the most ordered items among all customer profiles.

The South is the region with the largest number of customers.

Married, mature adults between age 40 and 64 make up the largest group of customers.

The majority of customers earn at least a middle-level income ($50,000 or more per year).

Recommendations

- Increase promotions and offer discounts on Thursdays and Fridays, as this is when new customers are most active.
- Increase promotion and marketing efforts towards younger customers (between age 18 and 39).

Pig E. Bank – Customer Retention

Objective

- Determine the characteristics of customers that exit Pig E. Bank and compare them with those of remaining customers.
- Identify factors that make a customer more likely to exit the bank and develop retention strategies based on these findings.

Data Used

Fictional data provided by CareerFoundry.

Methods Used

Data ethics and security principles

Identifying and dealing with bias

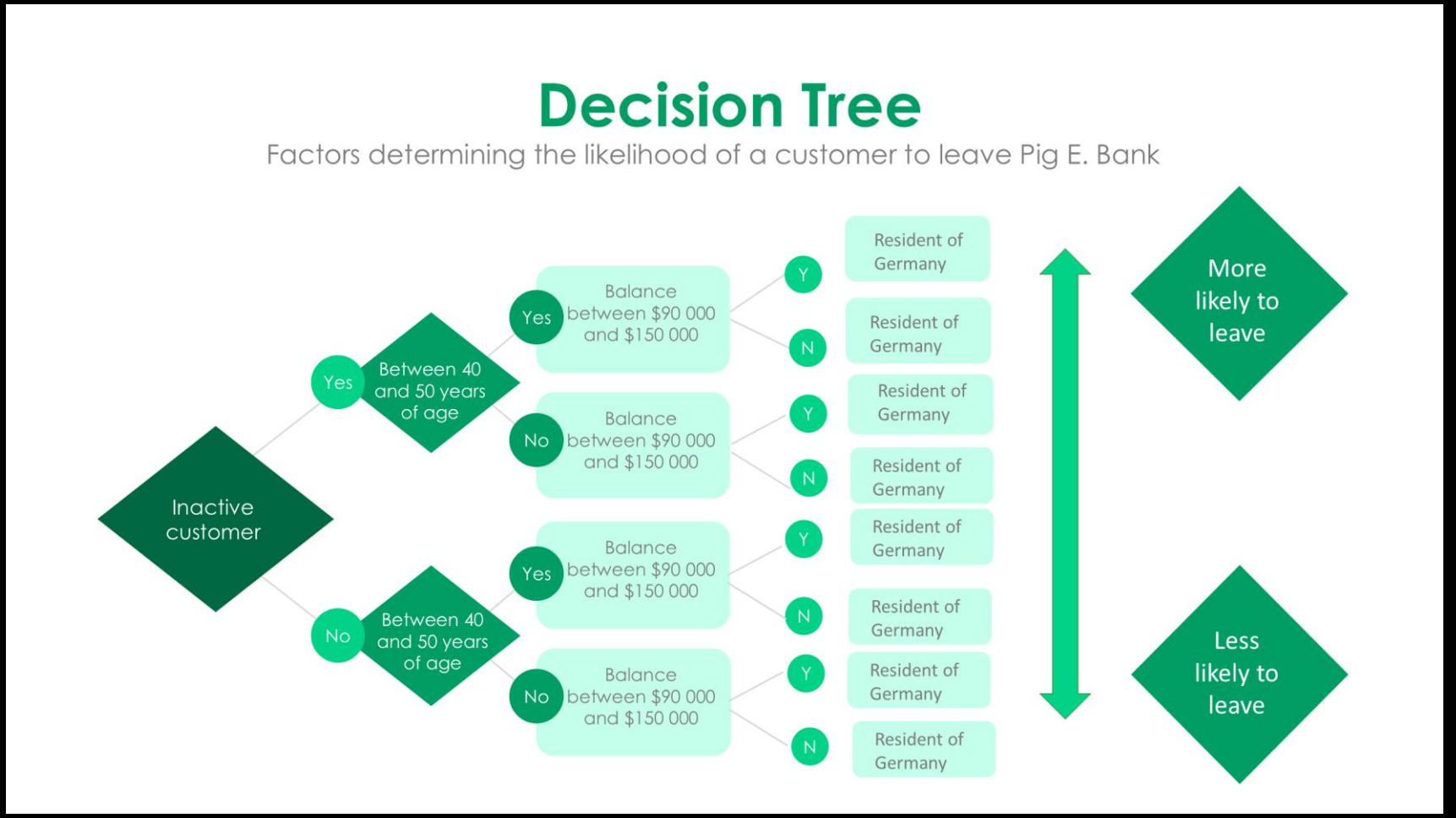
Data cleaning and descriptive statistics for data-mining

Decision-tree models

Linear and logistic regression

Time-series analysis and forecasting

Analysis



	All Customers			Exited Customers			Remaining Customers		
Variable	**Minimum**	**Maximum**	**Mean**	**Minimum**	**Maximum**	**Mean**	**Minimum**	**Maximum**	**Mean**
Credit Score	376	850	649	376	850	637	411	850	652
Age	18	82	39	22	69	45	18	82	38
Tenure	0	10	5	0	10	5	0	10	5
Balance	$0,00	$213 146,20	$78 002,72	$0,00	$213 146,20	$90 239,22	$0,00	$197 041,80	$74 830,87
NumOfProducts	1	4	2	1	4	1	1	3	2
EstimatedSalary	$0,00	$199 725,39	$98 375,60	$417,41	$199 725,39	$97 155,20	$0,00	$199 661,50	$98 691,95

Insights

The majority of customers who exited the bank (70%) were listed as inactive.

Germany has a higher proportion of customers leaving the bank compared to total customers and customers remaining.

Customers who exited have an overall higher average age (45)

More women (59.3%) than men exited the bank

The biggest age group among customers who exited the bank were between 42 and 51

Customers exiting the bank have a higher average balance compared to those remaining and to total customers.

Recommendations

- Do further research into the customer groups most likely to leave (inactive customers, female, age 42-51, resident of Germany) to determine their reasons for exiting the bank. This could take the form of, for example, a customer satisfaction survey, or analyzing existing customer feedback if possible.
- Develop targeted strategies to retain customers belonging to any of these groups based on the findings of this research.

Airbnb Amsterdam 2019

Objective

- Identify different characteristics of neighbourhoods and boroughs in Amsterdam through the analysis of Airbnb rental listings for the year 2019.
- Determine which criteria are most important to Airbnb guests when choosing their accommodation in Amsterdam.

Data Used

"Airbnb Amsterdam". Retrieved from Kaggle on 21 December, 2024. https://www.kaggle.com/datasets/erikbruin/airbnb-amsterdam/

"Boroughs of Amsterdam", Wikipedia: https://en.wikipedia.org/wiki/Boroughs_of_Amsterdam

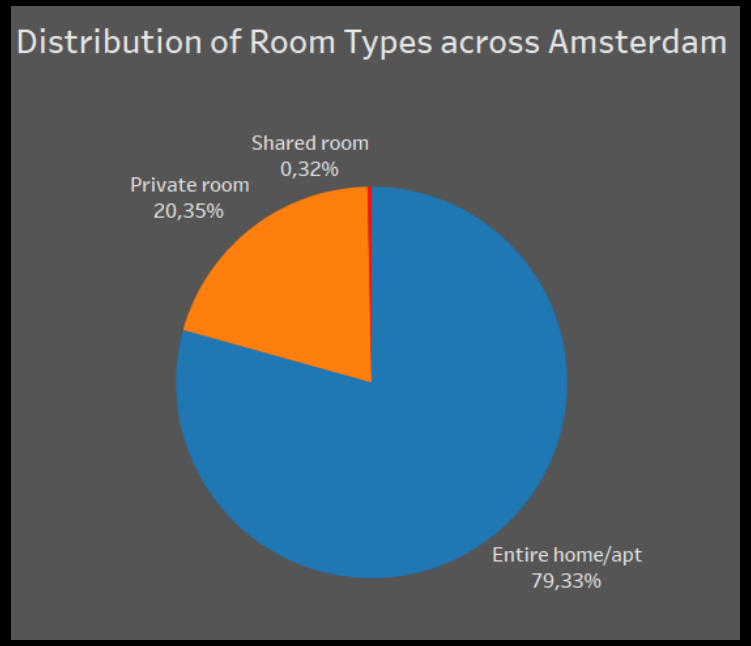
Methods Used

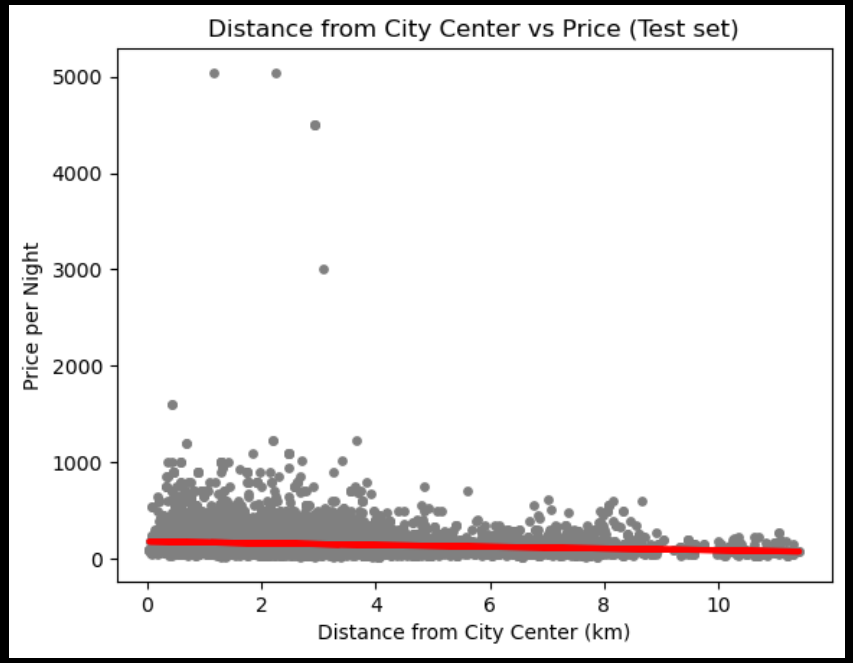
- Data-cleaning, wrangling, subsetting, aggregating and merging with Python, using pandas and NumPy libraries.
- Application of structured (linear regression) and unstructured (k-means clustering) machine-learning algorithms, using SciKitLearn library.
- Time-series analysis: decomposing time-series data, checking for and adjusting stationarity, using statsmodels library.
- Geospatial visualization of GEOJSON data, using folium library.
- Visualizing data relationships using matplotlib and seaborn libraries.
- Creating interactive dashboards on Tableau.

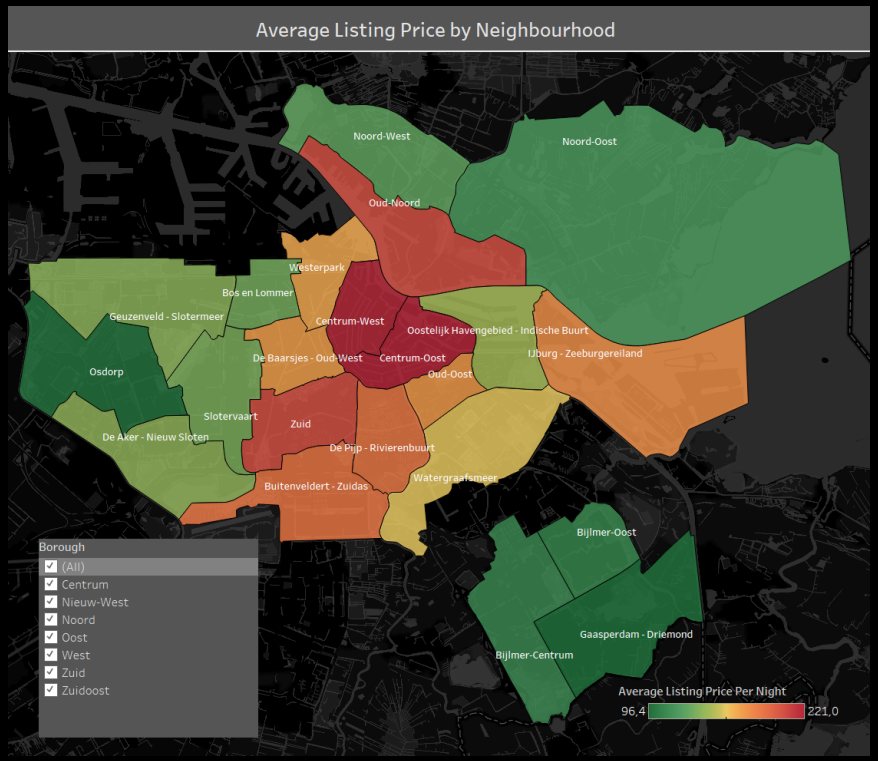
Analysis

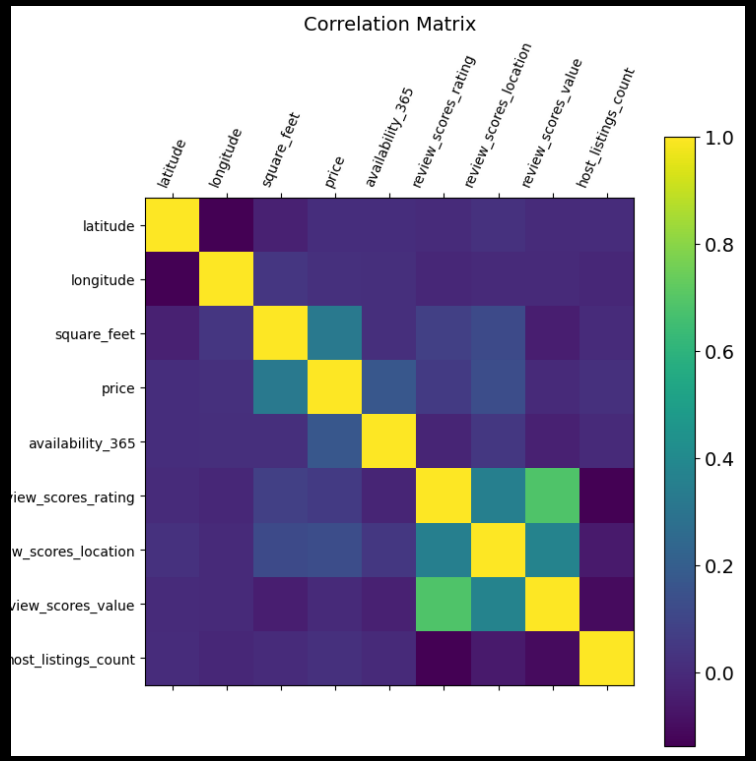
Some examples of visualizations created for this project.

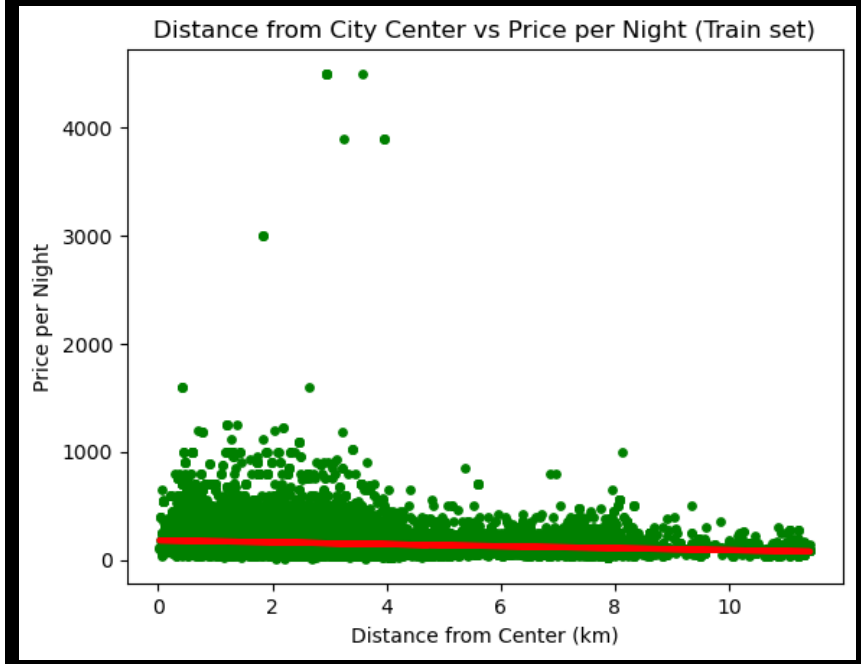
- See GitHub for complete project scripts and visualizations.
- See Tableau Public for interactive data dashboards.











Insights

- The most popular neighbourhoods in Amsterdam for short-term guests are located in the boroughs of Amsterdam-West and Amsterdam-Zuid.
- The center of the city has higher rates of vacancy than surrounding areas likely due to its higher-than-average prices.
- Listings are more likely to be available across the city during the months of January and February.
- The vast majority of Airbnb listings advertise entire homes and apartments. Shared rooms make up less than 1% of total Amsterdam listings.
- Listings offered by hosts with multiple listings are more likely to receive a low review score from guests.

Recommendations

- Visitors can plan to visit Amsterdam during the winter months.
- Visitors should be aware of the higher prices if they choose to stay in the city center.
- Visitors should consider staying in less-popular and less-central neighbourhoods if they are interested in saving money.
- Hosts listing multiple accommodations for rent should put more effort into maintaining and caring for their properties in order to provide the best experience they can for their guests.
- Residents of certain neighbourhoods with particularly high demand can consider becoming an Airbnb host if they wish to.

ClimateWins: Predicting Weather Conditions and Climate Change

Objective

- Identify abnormal weather patterns in Europe and their increase in frequency over recent years.
- Generate predictions for climate conditions in Europe over the next 25 to 50 years.
- Determine which machine-learning techniques (or combinations thereof) are best suited for these purposes.

Data Used

Climate Data extracted from the European Climate Assessment and Data Set Project: https://www.ecad.eu/

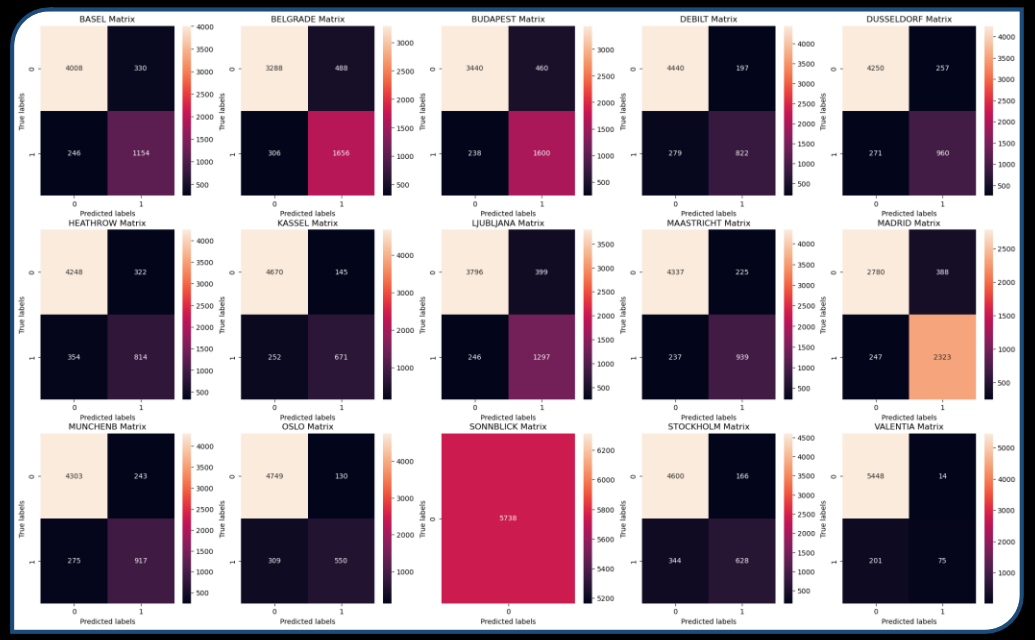
"Pleasant Weather Data", answers to a fictional survey regarding weather conditions, created for the purpose of this project.

Methods Used

- Data-cleaning, wrangling, subsetting, aggregating and merging with Python, using pandas and NumPy libraries.
- Optimization techniques for data and hyperparameter values, such as gradient descent and Bayesian optimizer.
- Supervised machine-learning algorithms such as k-nearest neighbours, decision trees and ANNs (artificial neural networks).
- Unsupervised machine-learning techniques such as random forest classifier using SciKitLearn.
- Deep-learning models using Keras, such as convolution neural networks (CNNs) and generative-adversarial networks (GANs).

Analysis

- Some examples of visualizations and analysis performed for this project.
- Full set of Python scripts, visualizations and reports can be found on GitHub.



Pred	BASEL	BELGRADE	BUDAPEST	DEBILT	HEATHROW	KASSEL	LJUBLJANA	\
True								
BASEL	3	1216	309	132	4	3	1431	
BELGRADE	0	783	31	0	0	0	276	
BUDAPEST	0	141	9	0	0	0	64	
DEBILT	0	55	3	0	0	0	23	
DUSSELDORF	0	12	1	0	0	0	16	
HEATHROW	0	26	2	1	0	0	49	
KASSEL	0	8	1	0	0	0	2	
LJUBLJANA	0	28	2	0	0	0	31	
MAASTRICHT	0	2	0	0	0	0	6	
MADRID	0	126	13	13	0	0	253	
MUNCHENB	0	5	1	0	0	0	2	
OSLO	0	3	0	0	0	0	1	
STOCKHOLM	0	3	0	0	0	0	1	
VALENTIA	0	0	0	0	0	0	0	

Pred	MAASTRICHT	MADRID	MUNCHENB	OSLO	SONNBLICK	STOCKHOLM	VALENTIA
True							
BASEL	59	294	4	205	10	1	11
BELGRADE	0	2	0	0	0	0	0
BUDAPEST	0	0	0	0	0	0	0
DEBILT	0	1	0	0	0	0	0
DUSSELDORF	0	0	0	0	0	0	0
HEATHROW	1	3	0	0	0	0	0
KASSEL	0	0	0	0	0	0	0
LJUBLJANA	0	0	0	0	0	0	0
MAASTRICHT	0	1	0	0	0	0	0
MADRID	5	48	0	0	0	0	0
MUNCHENB	0	0	0	0	0	0	0
OSLO	0	1	0	0	0	0	0
STOCKHOLM	0	0	0	0	0	0	0
VALENTIA	1	0	0	0	0	0	0

Weather Station	Accurate Prediction 0	Accurate Prediction 1	False Positive	False Negative	Total Accuracy
Basel	4008	1154	246	330	90%
Belgrade	3288	1656	**488**	306	86%
Budapest	3440	1600	460	238	88%
Debilt	4440	822	197	279	92%
Dusseldorf	4250	960	257	271	91%
Heathrow	4248	814	322	**354**	88%
Kassel	4670	671	145	252	93%
Ljubljana	3796	1297	399	246	89%
Maastricht	4337	939	225	237	92%
Madrid	2780	**2323**	388	247	89%
Munchen	4303	917	243	275	91%
Oslo	4749	550	130	309	92%
Sonnblick	**5738**	0	0	0	**100%**
Stockholm	4600	628	166	344	91%
Valentia	**5448**	75	14	201	**96%**
				Average	**91%**

Pred	BASEL	BELGRADE	BUDAPEST	DEBILT	DUSSELDORF	HEATHROW	KASSEL	\
True								
BASEL	3561	66	6	10	7	7	0	
BELGRADE	120	944	7	4	1	2	0	
BUDAPEST	19	23	146	12	1	5	0	
DEBILT	9	5	7	56	2	1	0	
DUSSELDORF	4	2	1	2	5	13	0	
HEATHROW	8	2	1	1	5	55	0	
KASSEL	5	2	1	0	3	0	0	
LJUBLJANA	12	2	4	0	0	2	1	
MAASTRICHT	2	0	0	0	0	1	0	
MADRID	38	11	8	0	0	5	2	
MUNCHENB	5	1	0	0	0	0	0	
OSLO	3	0	0	0	0	0	0	
STOCKHOLM	2	0	0	0	0	0	0	
VALENTIA	1	0	0	0	0	0	0	

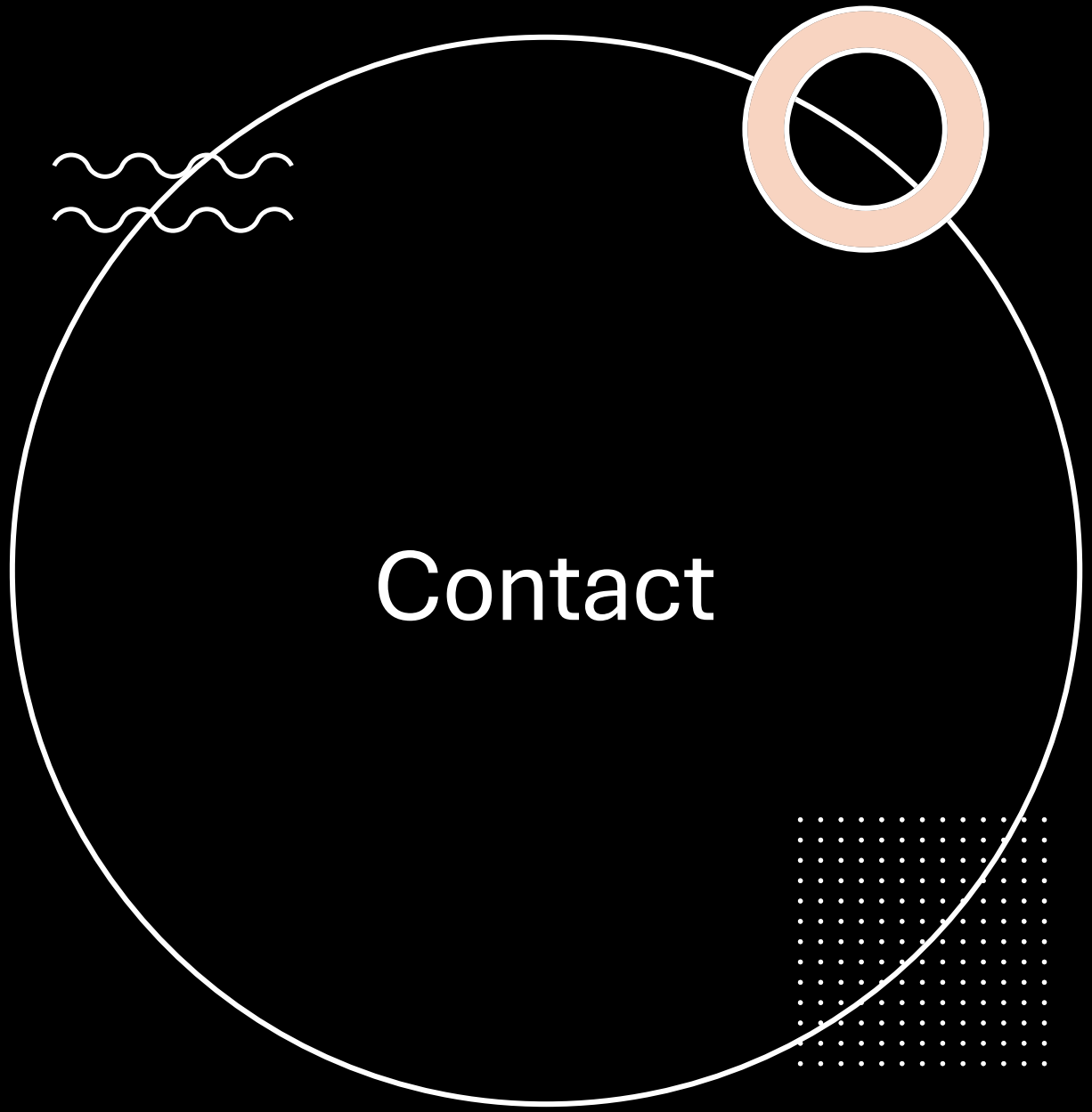
Pred	LJUBLJANA	MADRID	MUNCHENB	OSLO
True				
BASEL	1	22	1	1
BELGRADE	3	11	0	0
BUDAPEST	1	6	1	0
DEBILT	1	1	0	0
DUSSELDORF	0	2	0	0
HEATHROW	1	8	0	1
KASSEL	0	0	0	0
LJUBLJANA	26	13	0	1
MAASTRICHT	0	6	0	0
MADRID	5	389	0	0
MUNCHENB	0	0	2	0
OSLO	0	2	0	0
STOCKHOLM	0	0	1	1
VALENTIA	0	0	0	0

Insights

- Classification algorithms such as k-nearest neighbours are the most suitable for identifying and predicting changing weather conditions within a region.
- Unsupervised classification models, such as random forests, work best when applied to a smaller, less-varied set of data, e.g. climate data from a single city or region.
- Optimized data and hyperparameter values ensure that a machine-learning model performs as intended and generates accurate results.
- Visual data can be accurately interpreted using a convolutional neural network (CNN) and may be the most effective way to identify and predict changing weather conditions in Europe.

Recommendations

- Use a generative adversarial network (GAN) consisting of at least one CNN to analyze weather satellite and radar images and create accurate visual predictions of future weather conditions in Europe.
- Use a random forest model to identify weather patterns within a smaller region and, using a second set of data such as survey answers or healthcare data, determine the impact of these weather conditions on the local population.



Contact

E-mail:
suryd@protonmail.com

GitHub:
https://github.com/sryds