Case Studies

Eva-Maria Kuck

Projects

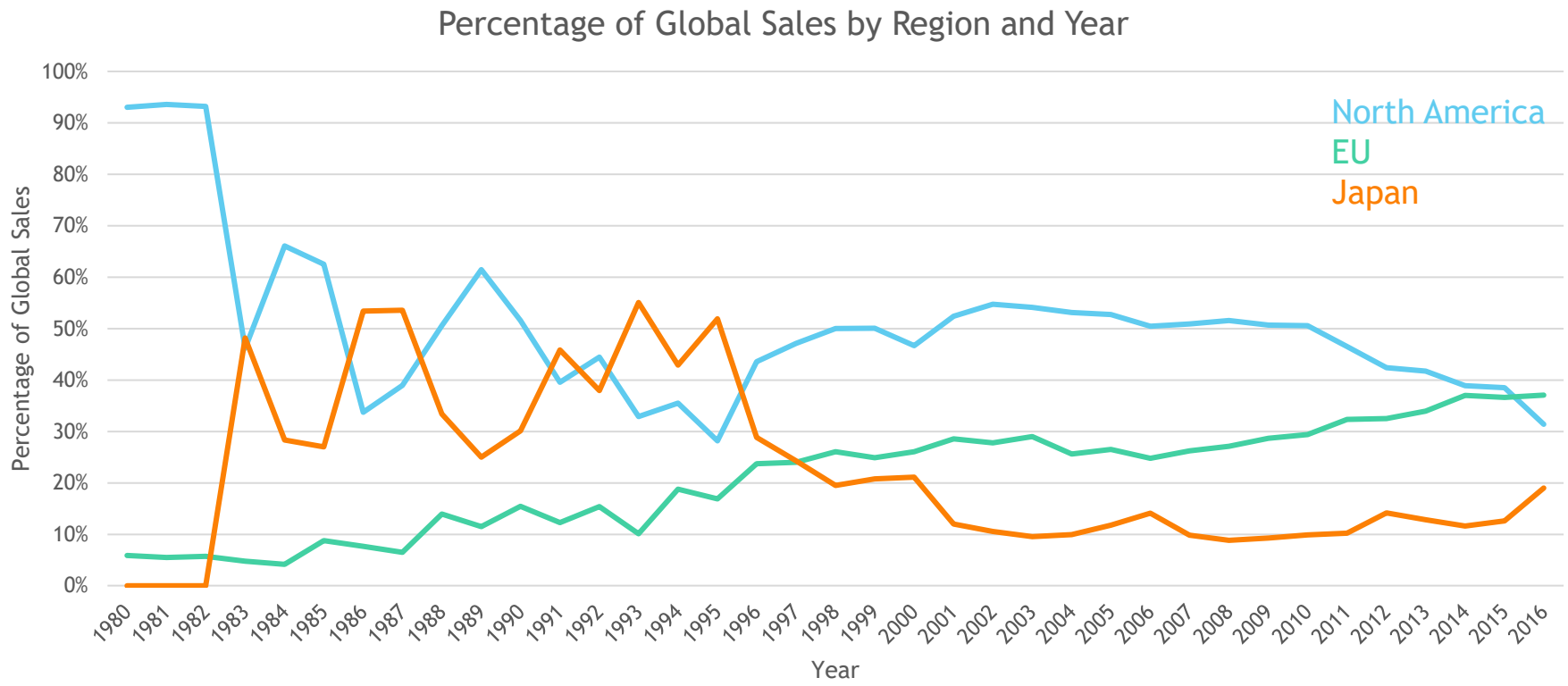
- GameCo
- Medical Staffing Agency
- Rockbuster Stealth LLC
- Instacart
- Achievement 6

GameCo

Overview

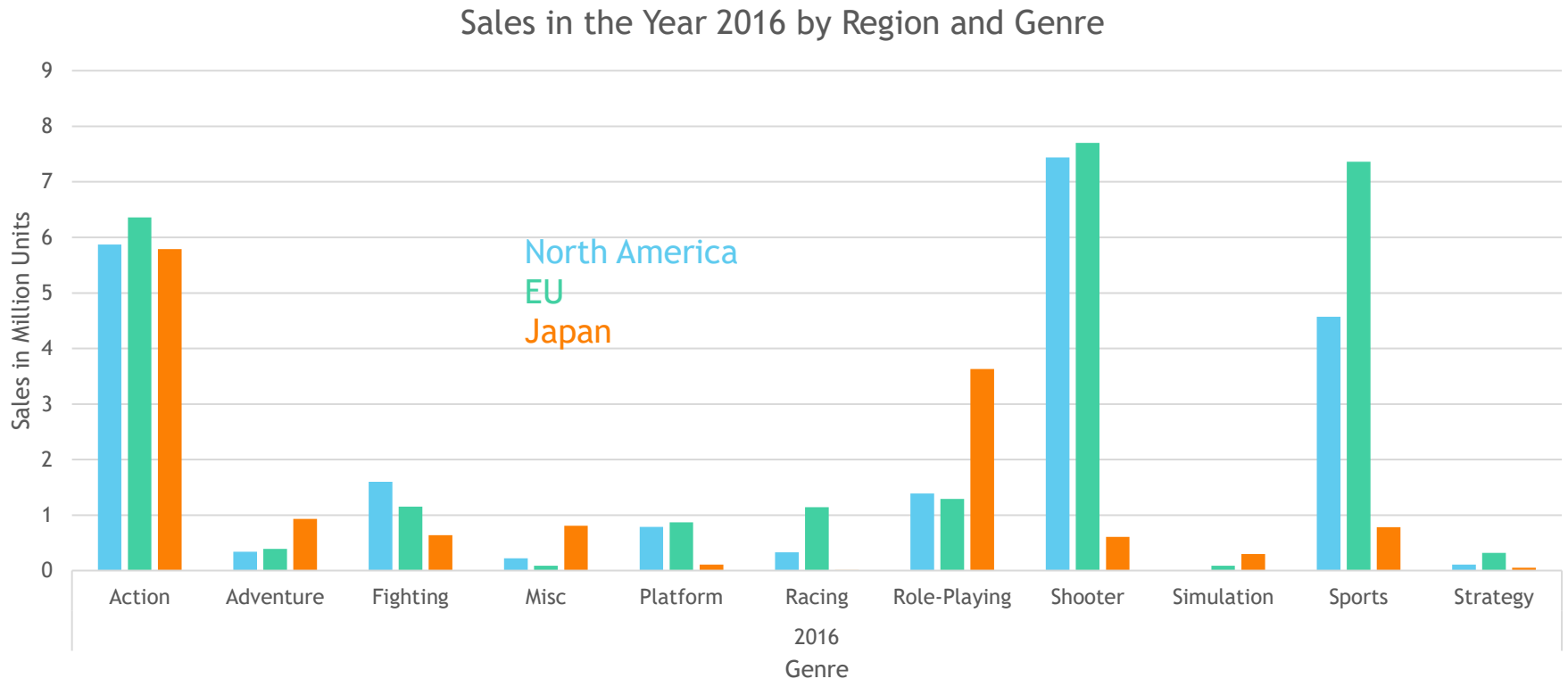
Objective	Data	Tools and Analytical Skills
• Fostering a better understanding of how GameCo's new games might fare in the market • Current understanding: Video games sales have stayed the same for different geographic regions since 1980.	• Includes sales of video games from 1980 – 2016 broken down by different platforms, genres, and publishing studios • Drawn from the website VGChartz	**Excel:** • Cleaning data • Grouping and summarizing data • Descriptive analysis • Visualizing results in Excel **PowerPoint** • Presenting results

Analysis



In the 1980s and 1990s, the Japanese and North American percentages of global sales have varied a lot, and the Japanese proportion of global sales was highest in some years (e.g. 1986 and 1987).
Since around 2010, the percentage of North American sales has declined.
The European proportion of global sales has been continually growing, and 2016 was the first year in which the European proportion of global sales was highest.

Analysis

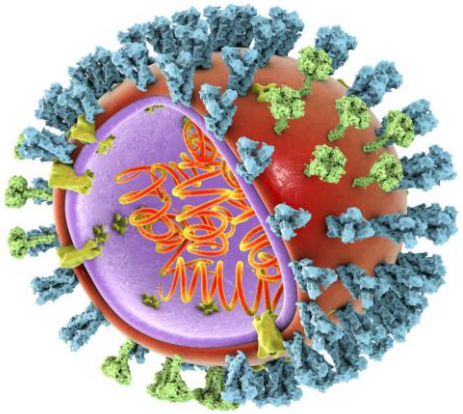


North America and Europe favored Action, Sports and Shooters in 2016.
Japan favored Action and Role-Playing in 2016.

Conclusion and Recommendations

- Europe is a continually growing market for video games
- In Europe, the preferred video game genres in 2016 were action, sports and shooters
- GameCo should increase the percentage of games shipped to the EU, and also increase the marketing in Europe
- Full presentation here

Medical Staffing Agency

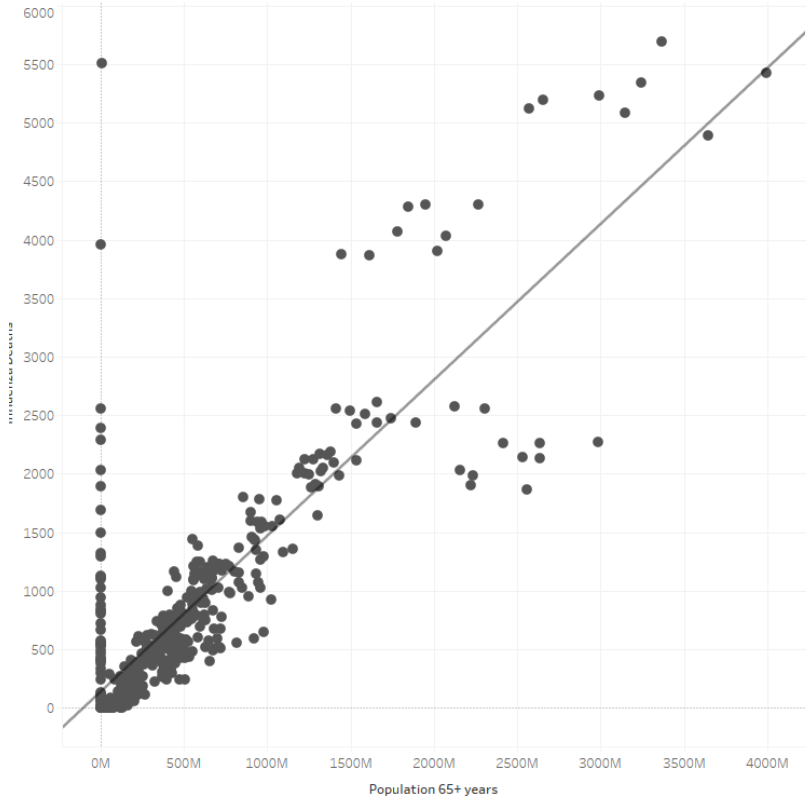


Overview

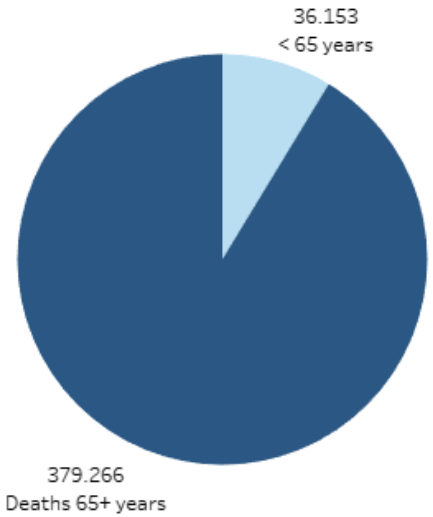
Objective	Data	Tools and Analytical Skills
The United States has an influenza season where more people than usual suffer from the flu. Hospitals need additional staff to adequately treat these extra patients. The medical staffing needs to know: **when** to send staff, and **how many, to each state**.	**US Census Bureau Population Data:** The data includes populations by county, state, gender and age group for each year from 2009 to 2017. **CDC Influenza Deaths Data:** The data contains monthly death counts for influenza-related deaths in the United States from 2009 to 2017 broken down by state and age group.	**Excel:** • Profiling and cleaning data, improving data integrity • Integrating and transforming data • Calculating descriptive statistics • Statistical hypothesis testing **Tableau:** • Composition and comparison charts • Temporal visualizations and forecasting • Statistical visualizations • Spatial analysis

Analysis: Who is Most Vulnerable to Influenza Complications Including Death?

Scatterplot for the relationship between US population aged 65+ and influenza deaths (2009-2017)



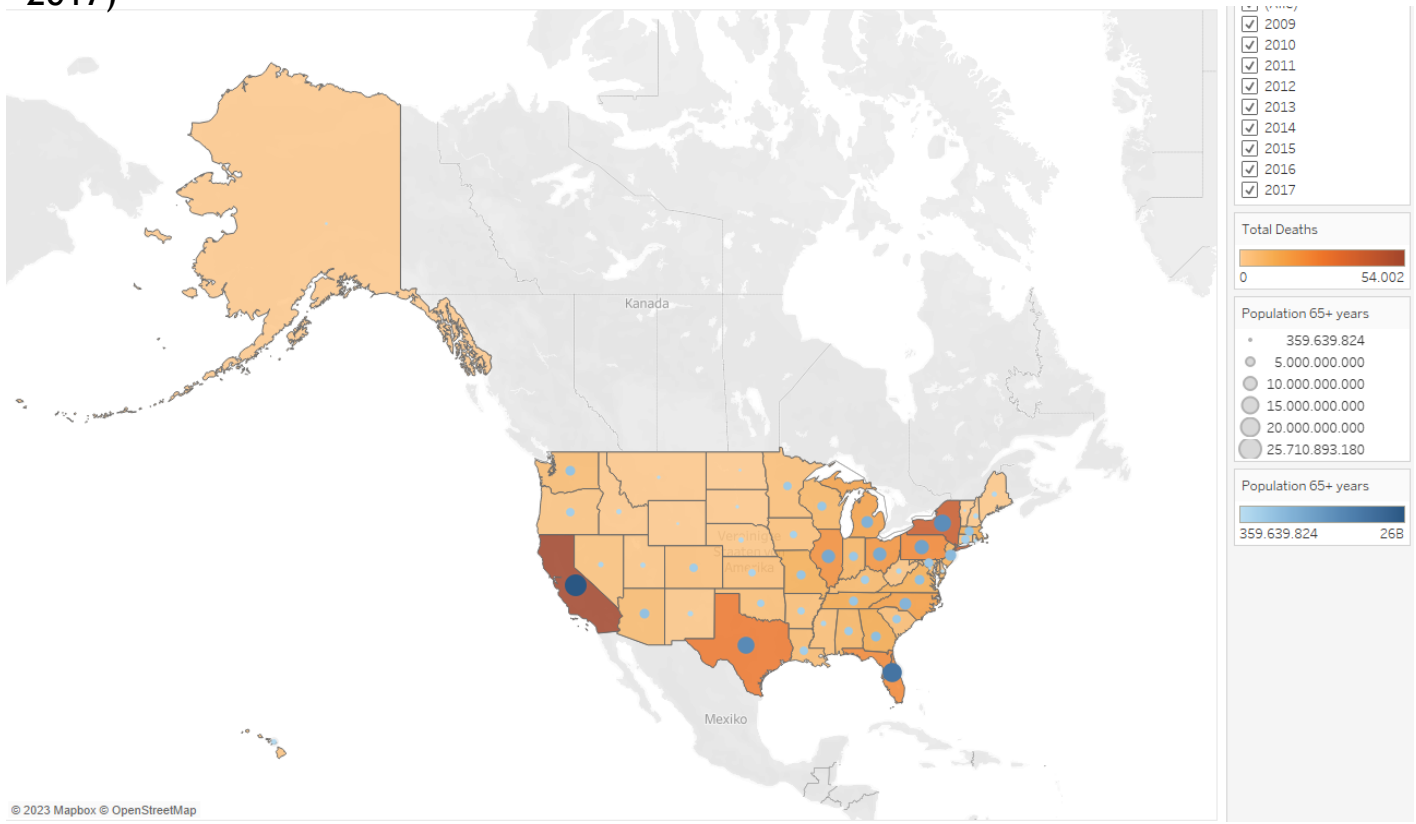
Pie chart for total US influenza deaths by age group (2009-2017)



Influenza deaths of people over 65 years represent the vast majority of total influenza deaths. On the same lines, there is a strong positive correlation between age over 65 and influenza deaths.

Analysis: In which States is Additional Medical Staff Most Needed?

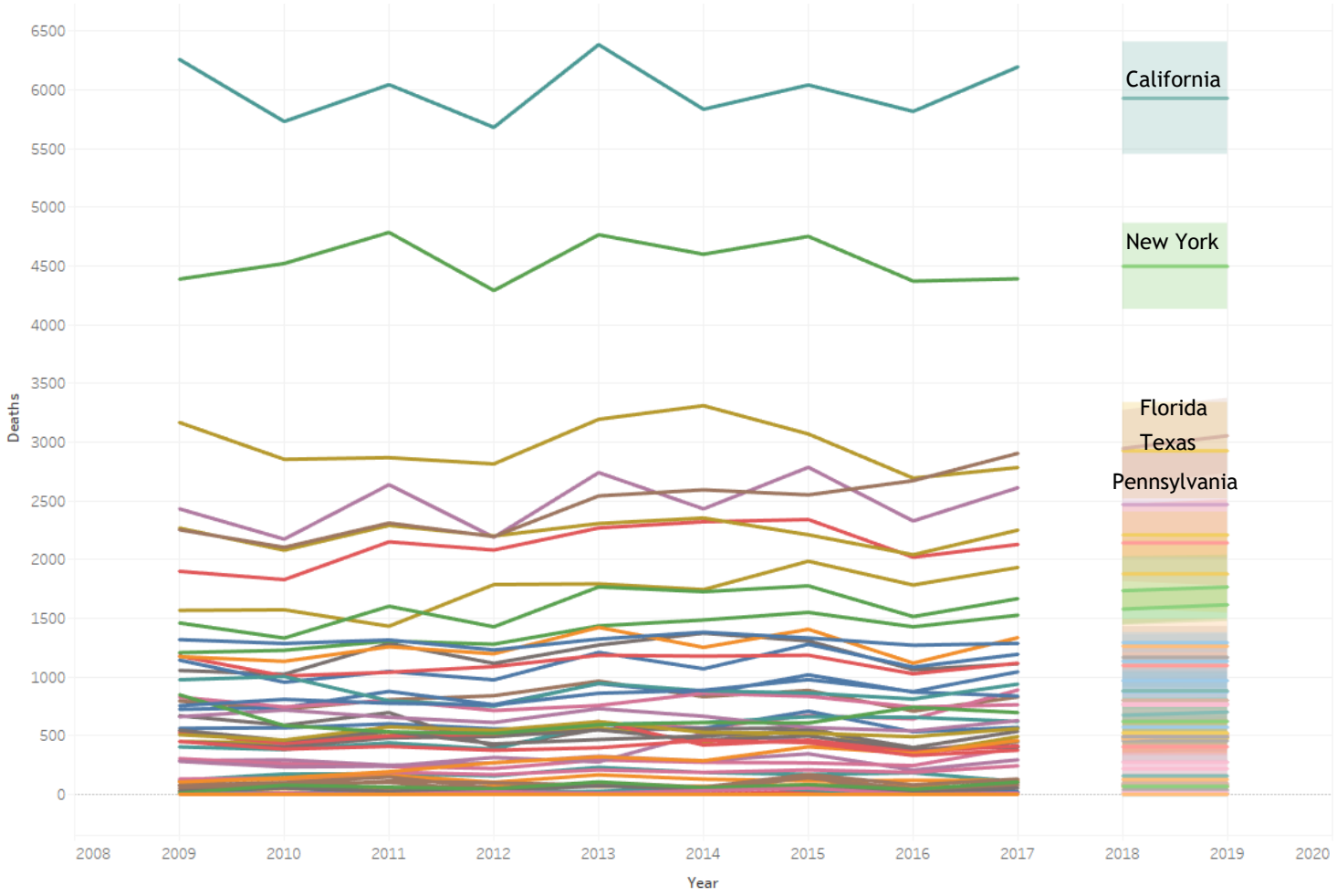
Dual axis map on total influenza deaths and population 65+ years by state (2009-2017)



The top 5 states with the highest number of influenza deaths over time are California, New York, Texas, Pennsylvania and Florida.

Analysis: In which States is Additional Medical Staff Most Needed?

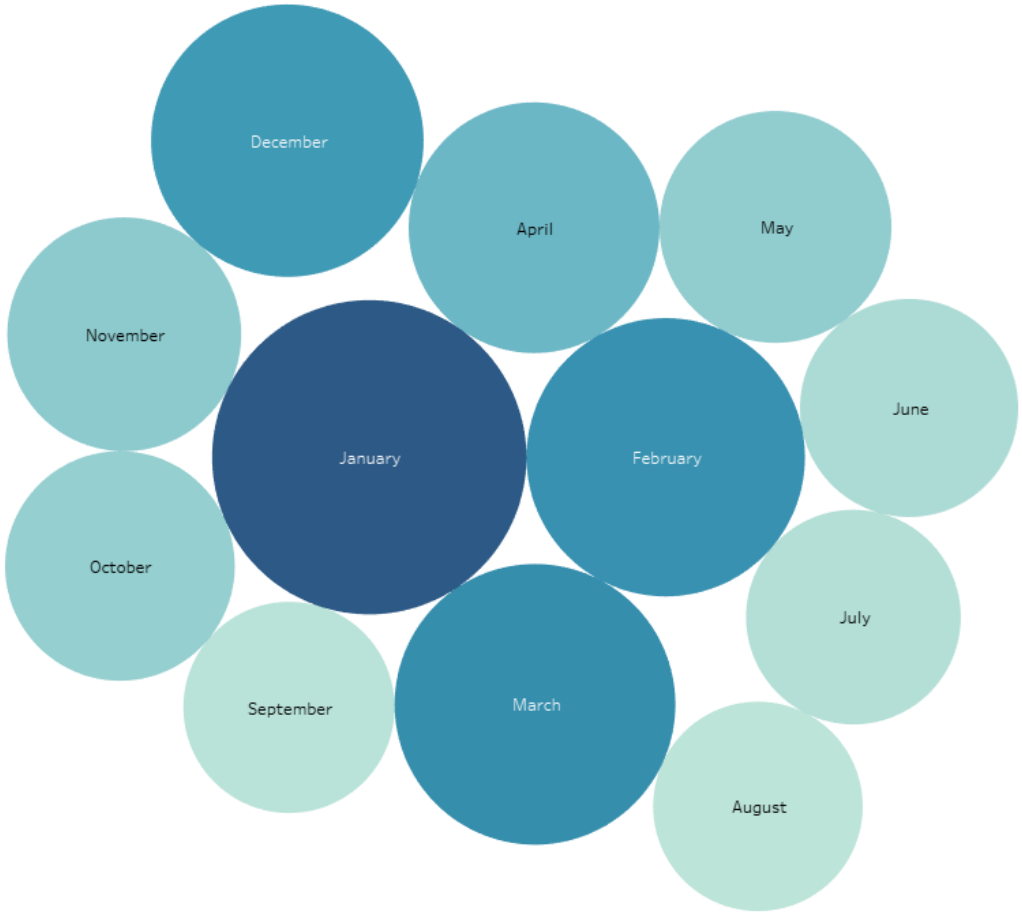
Line chart with 2018 seasonal forecast on influenza deaths by US state (2009-2017)



The top 5 states most affected by influenza deaths in 2018 are California, New York, Florida, Texas and Pennsylvania.

Analysis: In which Months of the Year is Additional Medical Staff Most Needed?

Influenza deaths by month of year (over all US states, 2009-2017)



Most influenza deaths occur in January, followed by February, March and December. The typical influenza season is in wintertime.

Conclusions and Recommandations

Conclusions:

- Individuals of 65 years and older are a vulnerable group to the influenza virus.
- The US states with the highest number of influenza deaths over the last 9 years are California, New York, Texas, Pennsylvania and Florida.
- Influenza mortality is highest in the months of December until March.

Recommendations:

- The medical staffing agency should send extra staff to the states of California, New York, Florida, Texas and Pennsylvania (in descending order), especially from December to March, to help with high influenza cases and avoid influenza mortalities.

Interim report here

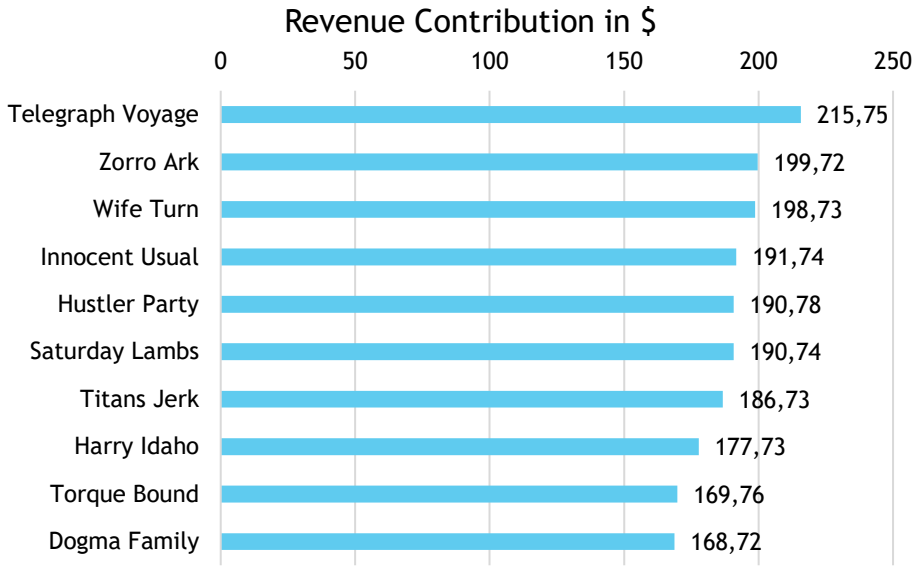
Tableau visualizations here

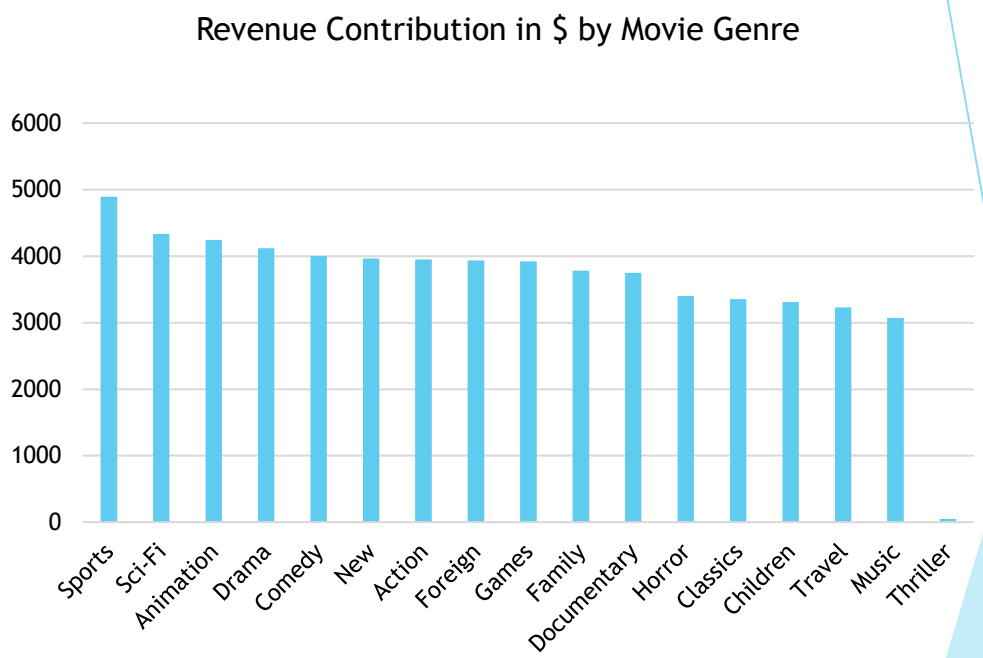
Rockbuster Stealth LLC

Overview

Objective	Data	Tools and Analytical Skills
Helping Rockbuster Stealth's Business Intelligence (BI) department with the launch strategy for the new online video service in order to stay competitive	Data set created for the purpose of this project, that contains information about Rockbuster's film inventory, customers, and payments, among other things.	**SQL:** • Relational databases • Database querying • Filtering • Cleaning and summarizing • Joining tables • Subqueries • Common table expressions **Tableau and PowerPoint:** • Presenting results

Analysis: Which movies and movie genres contributed the most to revenue gain?

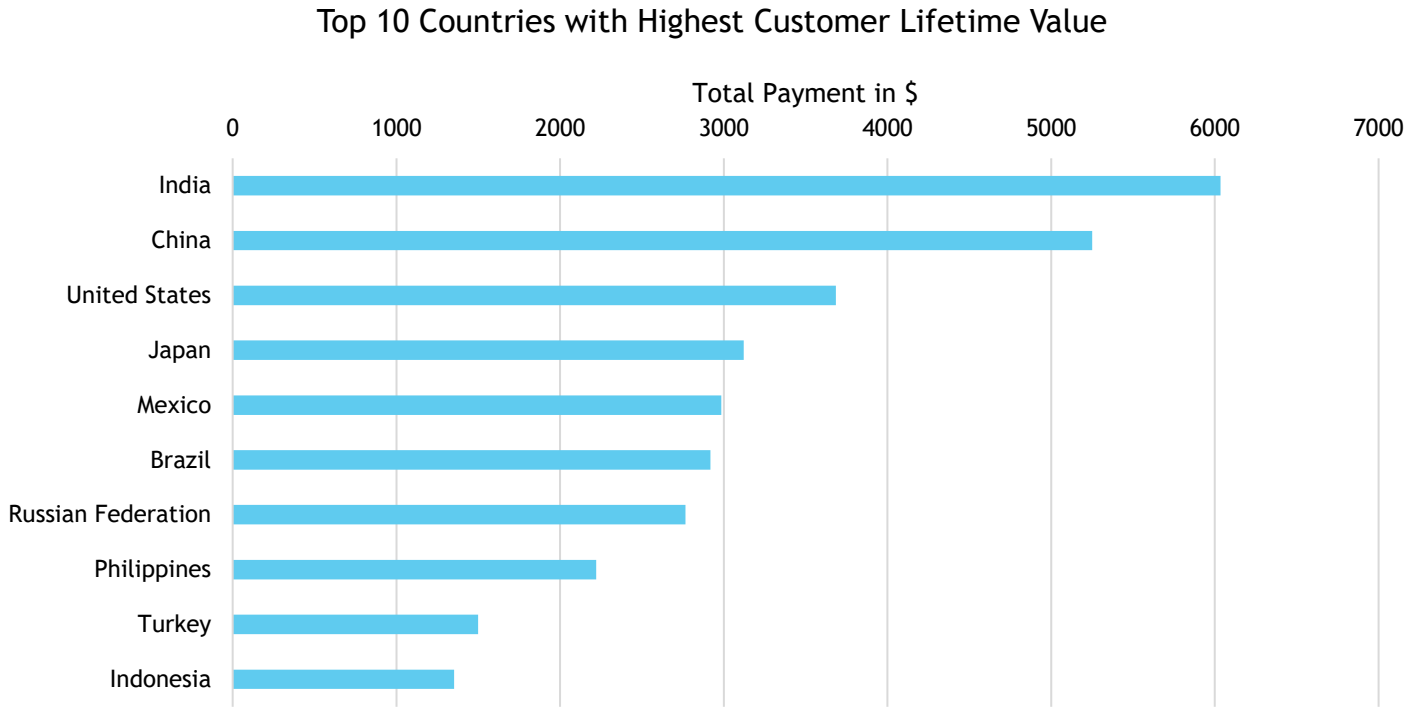




Telegraph Voyage countributed the most to revenue with $215,75.

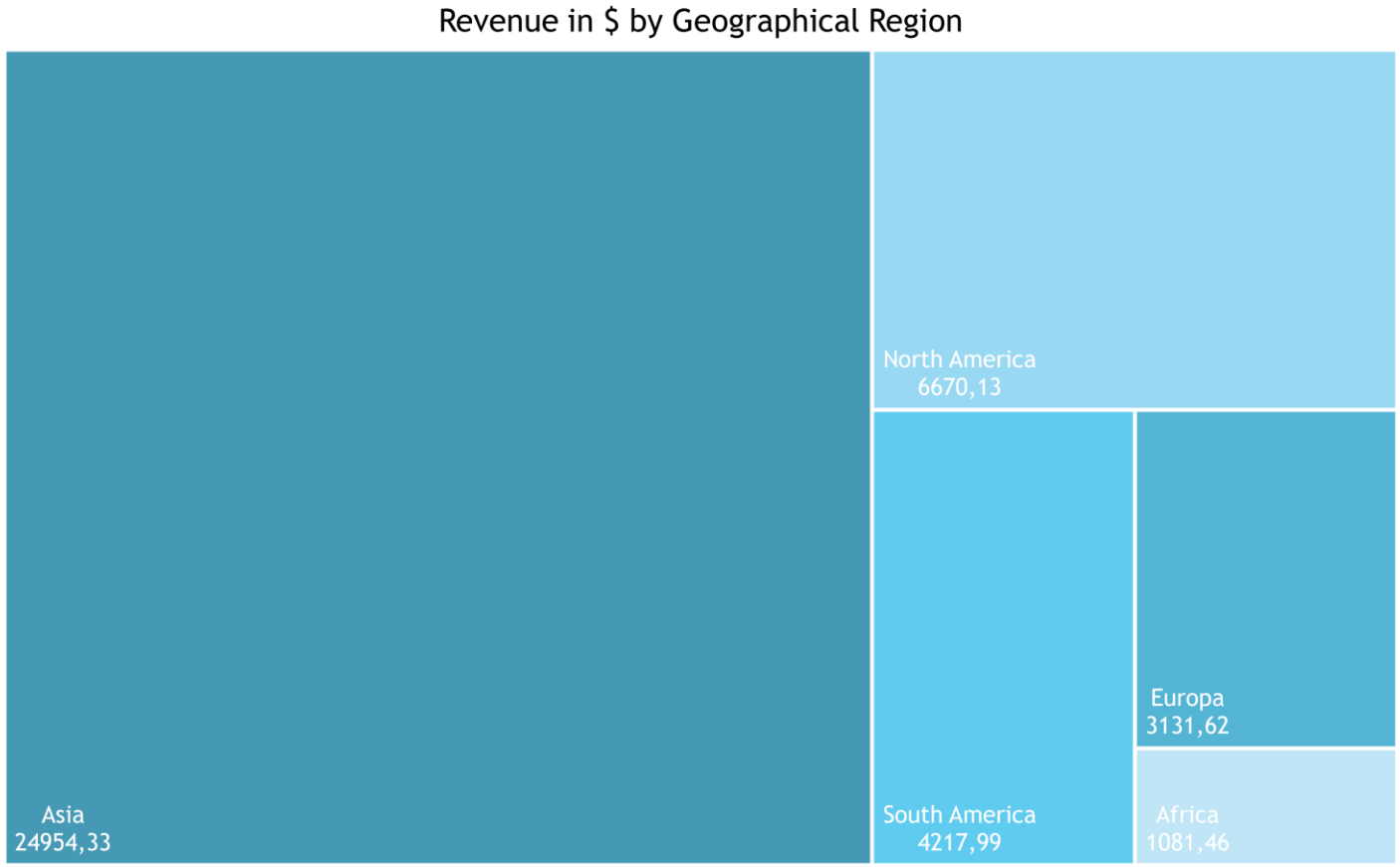
Top 5 Movie Genres:
1. Sports
2. Sci-Fi
3. Animation
4. Drama
5. Comedy

Analysis: Where are customers with a high lifetime value based?



Customers with the highest lifetime value are based in India, China, United States, Japan and Mexico (Top 5).

Analysis: Do sales figures vary between geographic regions?



Asia has by far the most revenue followed by North America and South America.

Conlusions and Recommendations

Conclusions:

- The top 5 most popular genres: Sports, Sci-Fi, Animation, Drama, Comedy
- The top 3 most profitable regions: Asia, North America, South America
- The top 5 countries with highest customer counts and highest customer lifetime value: India, China, United States, Japan, Mexico

Recommendations:

- Similar movies as the movies with highest revenue contribution should be added to the inventory
- Focusing on marketing in Asia, the most profitable region, followed by the other named regions
- Focusing on marketing in high revenue genres, such as Sports, Sci-Fi, Animation, Drama, Comedy
- Rewarding loyal customers by offering them discounts on rentals or special subscription plans

Full presentation here, Tableau visualizations here, GitHub repository here

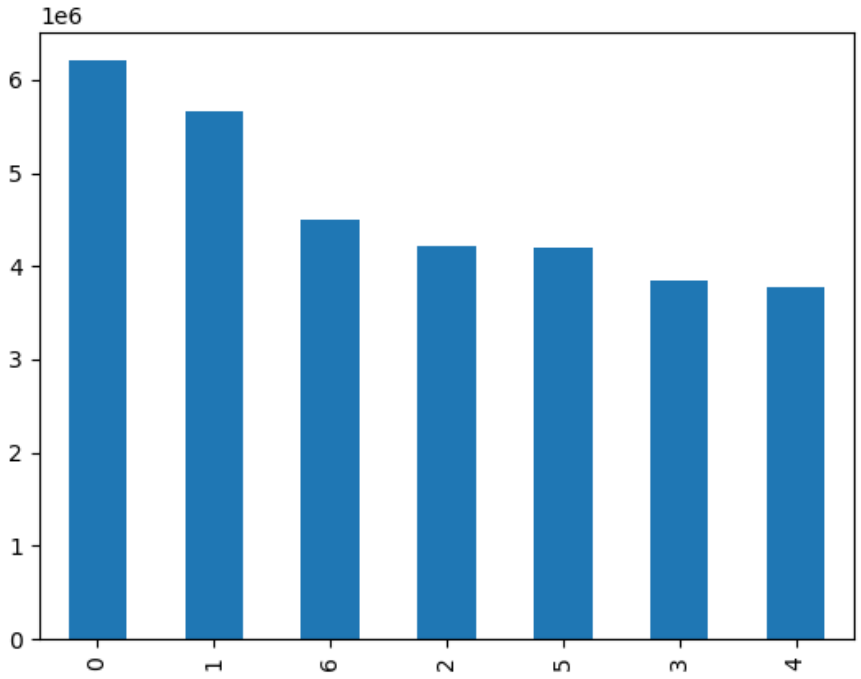
Instacart

Overview

Objective	Data	Tools and Analytical Skills
Analyzing grocery basket data in order to derive insights, uncover more information about sales patterns and suggest strategies for better segmentation	Open-source data sets from Instacart, complemented by a customer data set which was created for the purpose of this project	**Python:** • Data wrangling and subsetting • Data consistency checks • Data merging • Deriving variables • Grouping aata • Aggregating data • Data visualization **Excel:** • Presenting results • Population flows

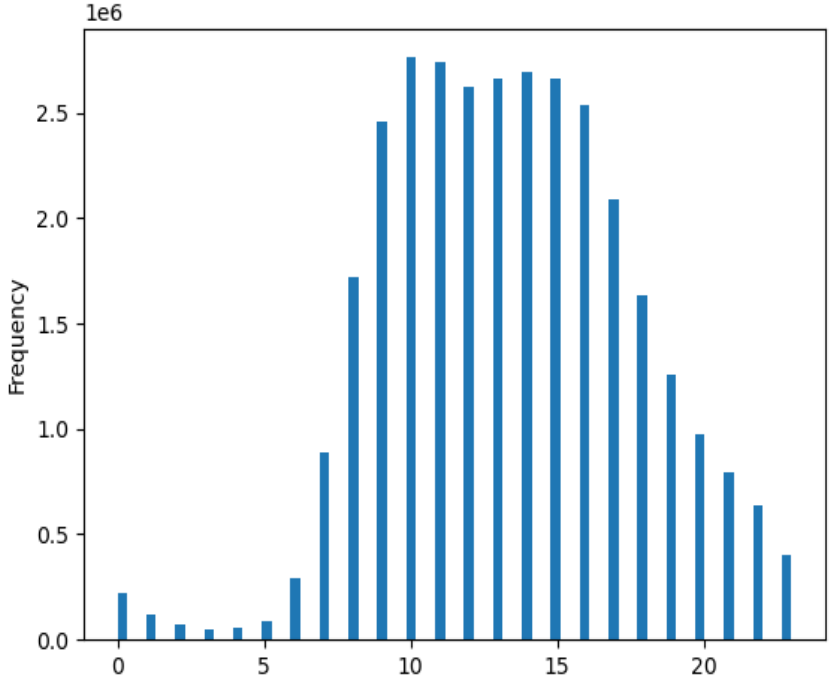
Analysis

Order Frequency by Day of the Week



Saturday (0), Sunday (1) and Friday (6) are the busiest days, and Wednesday (4) and Tuesday (3) are the days with least orders. Generally speaking, the weekend is the time of week with most orders.

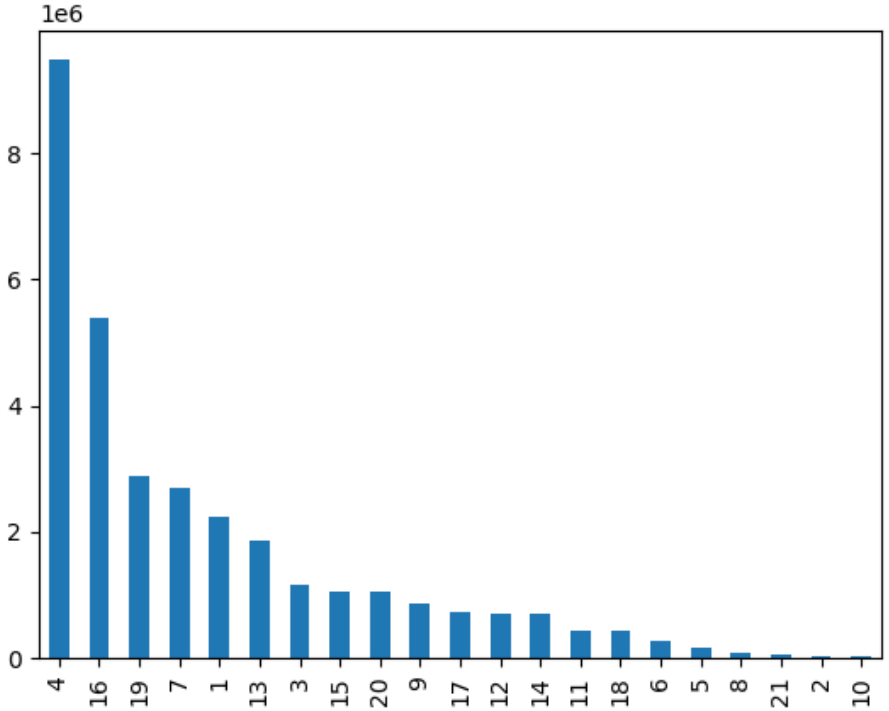
Order Frequency by Hour of the Day



Orders are at their highest from 9 am until 5 pm and then drop steadily to reach a low between midnight and 6 pm.

Analysis

Amount of Products Bought by Department Number

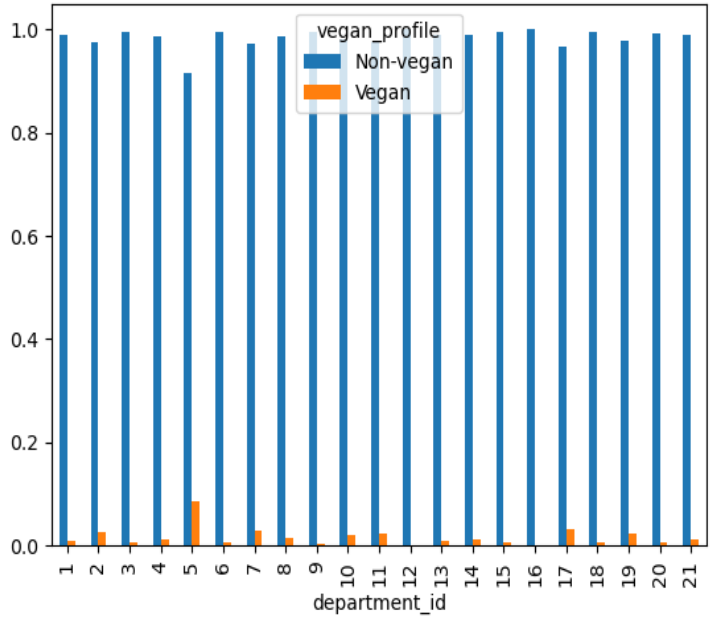


One can see that products from the departments "produce" (4) and "eggs/dairy" (16) are the most popular, followed by the products from the departments "snacks" (19), "beverage" (7), "frozen" (1), and "pantry" (13).

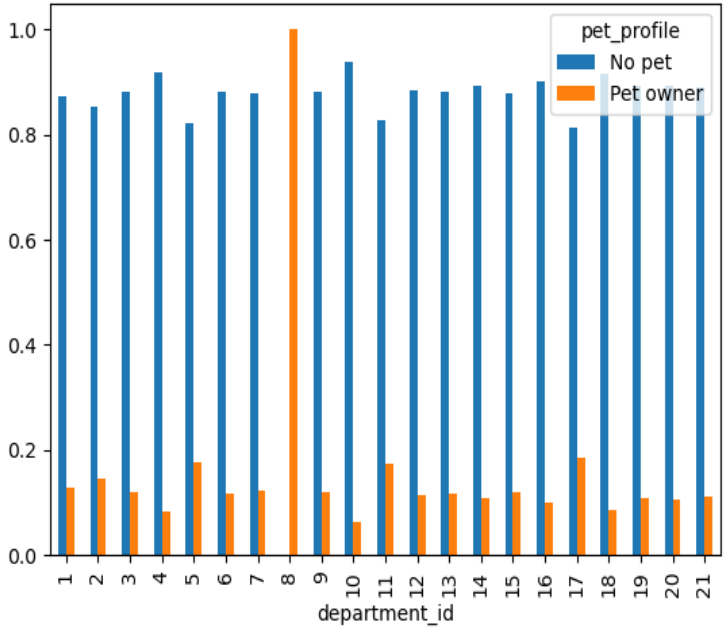
The departments are numbered as followed:
1: department: frozen
2: department: other
3: department: bakery
4: department: produce
5: department: alcohol
6: department: international
7: department: beverages
8: department: pets
9: department: dry goods/ pasta
10: department: bulk
11: department: personal care
12: department: meat/ seafood
13: department: pantry
14: department: breakfast
15: department: canned goods
16: department: dairy/ eggs
17: department: household
18: department: babies
19: department: snacks
20: department: deli
21: department: missing

Analysis

Amount of Products Bought by Department Number and Veganism/Non-Veganism



Amount of Products Bought by Department Number and Pet Ownership



Vegans spend relatively less on "frozen" (1), "bakery" (3), "dry goods/pasta" (9), "pantry" (13), "canned goods" (15), "deli" (20), and obviously, they don't buy meat (12) and dairy (16). Vegans spend relatively more money on "produce" (4), "alcohol" (5), "beverages" (7), "personal care" (11), "household" (17) and "snacks" (19).

Pet owners spend relatively more on "frozen" (1), "beverages" (7), of course "pets" (8), "personal care" (11), "pantry" (13), and "household" (17). They spend relatively less on "produce" (4), and "dairy/eggs" (16).

Conclusions and Recommendations

Conclusions:

- The weekend and between 9 am and 5 pm are the times with most orders.
- Products from the departments "produce" (4) and "eggs/dairy" (16) are the most popular.
- Vegans differ from non-vegans and pet owners differ from non-pet owners in respect to their ordering habits.

Recommendations:

- To increase revenue, Instacart should place ads on Tuesday and Wednesday (when orders are low) between 9am and 5pm to reach as many customers as possible.
- Instacart could either advertise the already pretty well sold products or put more advertizing effort into the not so well sold items from like "bakery", "dry goods/pasta" or "household". Probably more exploration of the customers' preferences would make sense.
- It might be reasonable for Instacart to profile and adress their customers based on their diet (vegan or non-vegan) and their pet ownership.

Final report here

GitHub repository here

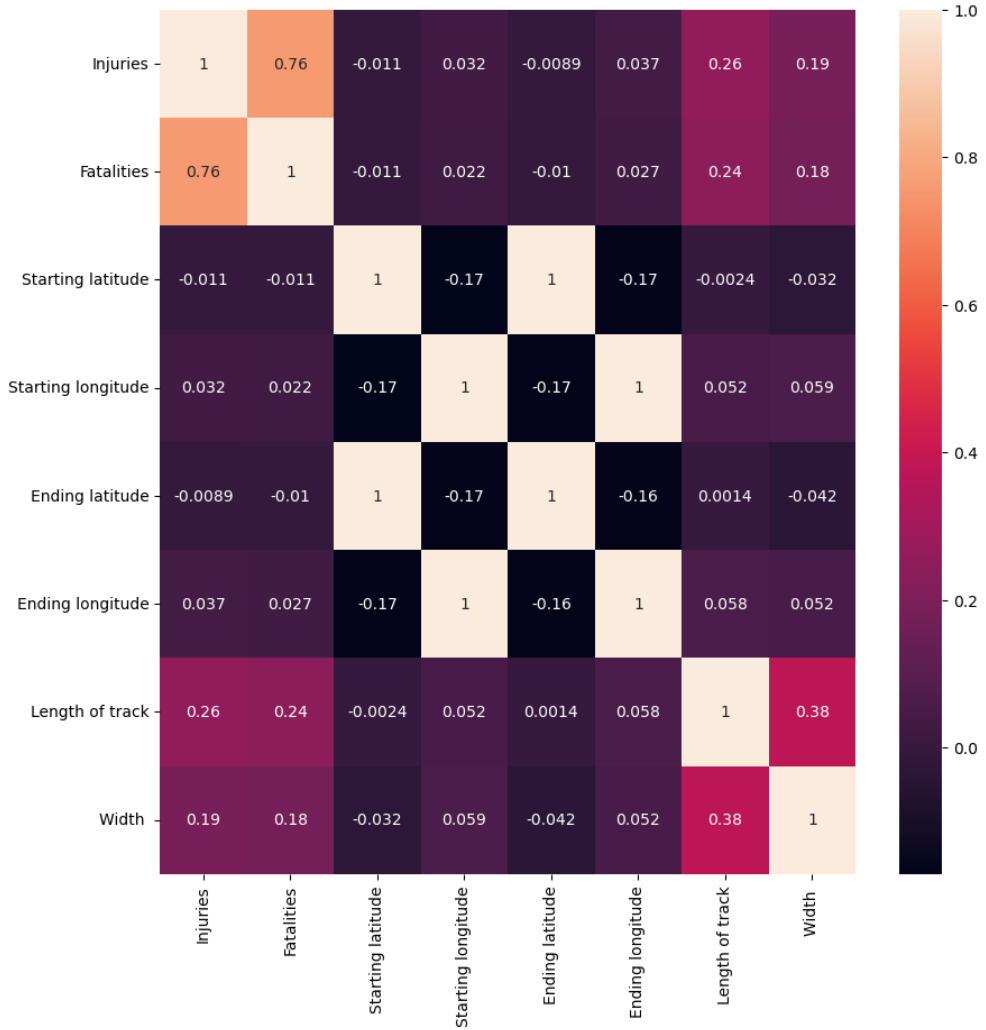
US Tornados

Overview

Objective	Data	Tools and Analytical Skills
Analyzing US tornado data (1950-2021) in order to derive insights about influential factors on the number of injuries caused by tornados. Finding clusters of tornados that share similar features.	Open-source data sets from Kaggle, derived from a dataset produced by the Storm Prediction Center of the National Oceanic and Atmospheric Administration (NOAA), a US government agency.	**Python:** • Exploring relationships • Geographical visualizations • Linear regression • Clustering • Preparing time series data **Tableau:** • Presenting results • Storytelling

Analysis

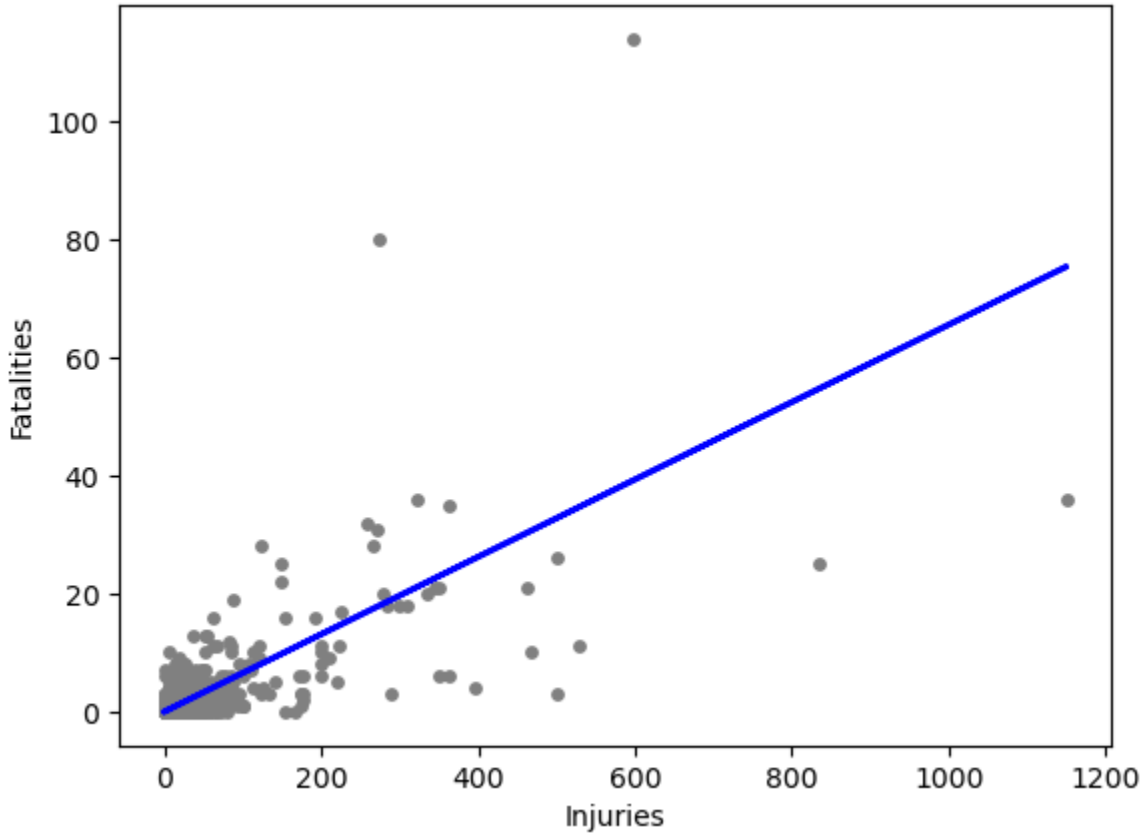
Correlation Heatmap for all Variables in the Dataset



We can see that interestingly the length of track and width of a tornado are only to a small amount positively correlated with injuries and fatalities caused by a tornado. However, injuries and fatalities are highly and positively correlated with each other (0.76).

Analysis

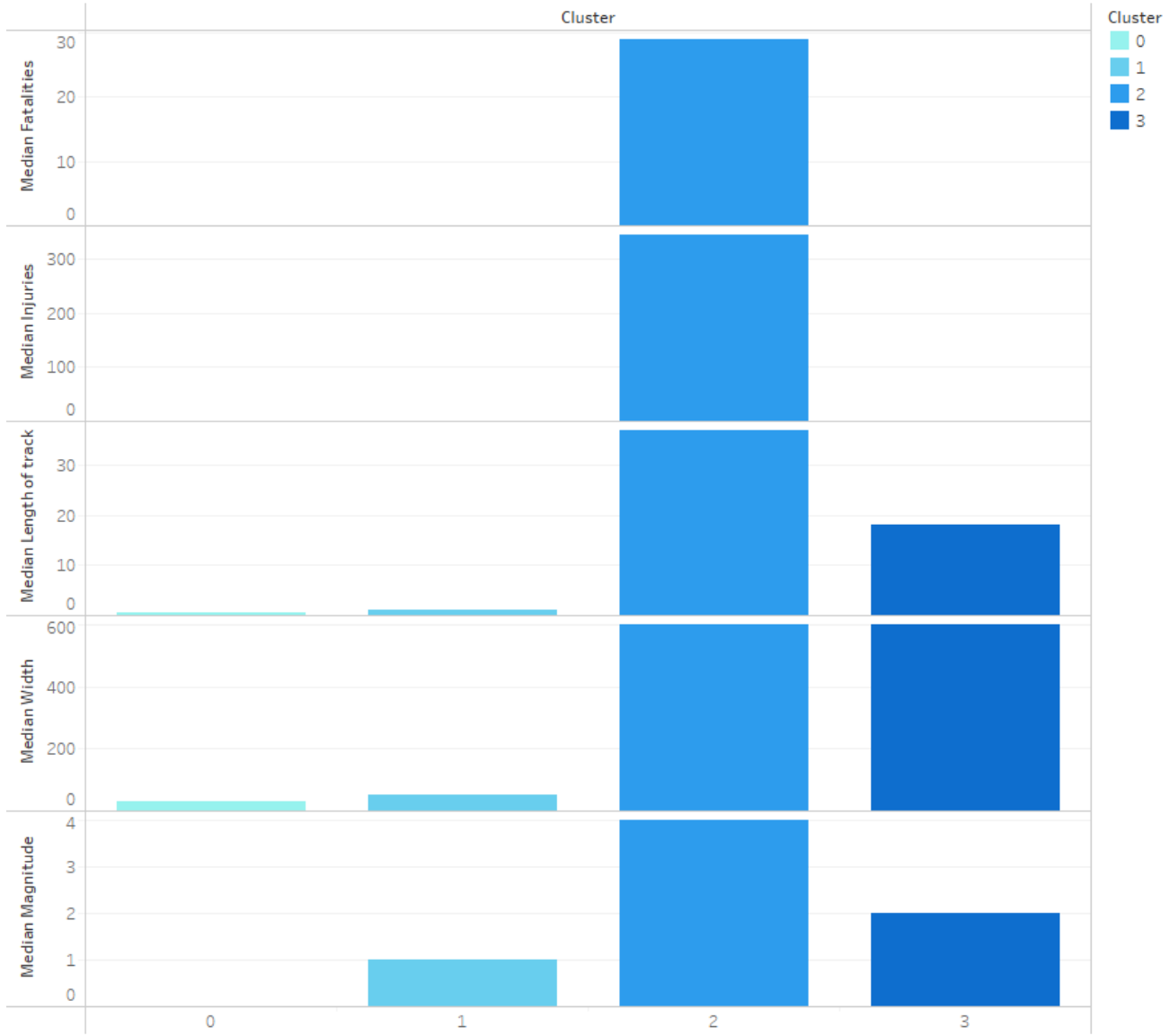
Linear Regression of Injuries on Fatalities caused by a Tornado (US, 1950-2021)



A linear regression was performed to test the hypothesis: *The higher the number of fatalities, the higher the number of injuries.*

The results showed that fatalities explained only **58.03%** of the variance of injuries. A linear regression cannot entirely explain the relationship between fatalities and injuries, as some data points are far away from the regression line. Therefore, a clustering approach was tried as well.

Analysis



A cluster analysis was conducted which revealed the following differences between the tornado clusters:
The most devastating tornado cluster (cluster 2) is the one with by far the highest median fatalities and injuries. It also has the highest median magnitude, median length of track and (together with cluster 3) median width.

Cluster 3 is medium devastating with a median magnitude of 2, a high width, and the second highest length of track. However, the median injuries and fatalities are zero.

Clusters 0 and 1 are the least destructive tornado clusters with cluster 1 being slightly bigger (length and width) and slightly more destructive than cluster 0 in terms of magnitude. Again, the median injuries and fatalities of both clusters are zero.

Conclusions and Next Steps

Conclusion:

- Neither length of track, nor width, nor magnitude have a substantial relationship with injuries caused by a tornado.
- However, the number of fatalities caused by a tornado can explain the number of injuries by 58%.
- Additional cluster analysis revealed i.a., that tornados from the most destructive cluster are usually bigger, cause more damage to buildings and by far the most injuries and deaths. They occur mostly in the midwest of the US.

Next steps:

- Performing a multiple regression with multiple variables which can better explain the amount of injuries caused by a tornado and therefore are of higher practical value
- Performing time-series forecasting

Tableau storyboard here

GitHub repository here