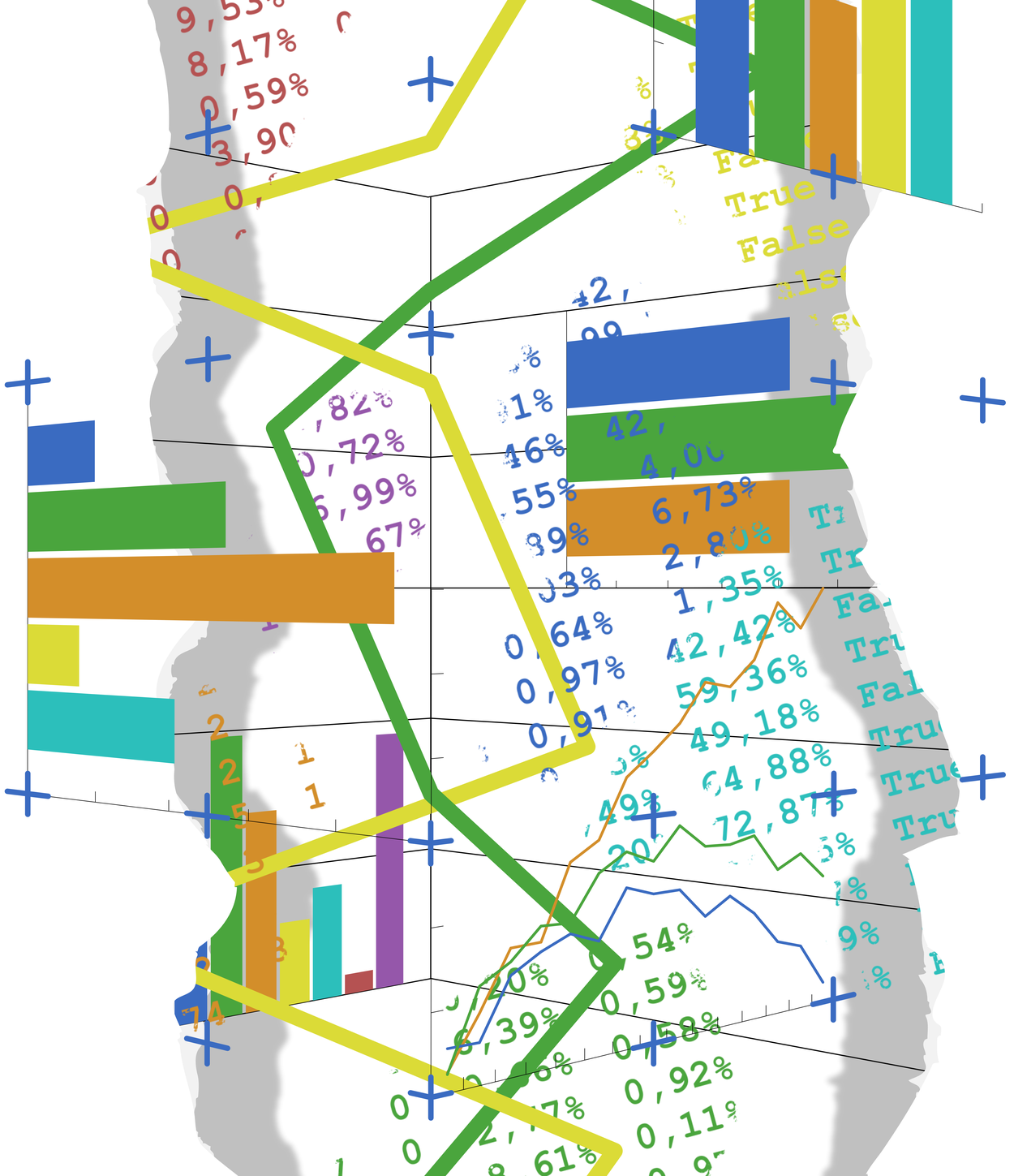
Nils Boden

Data Analytic Portfolio



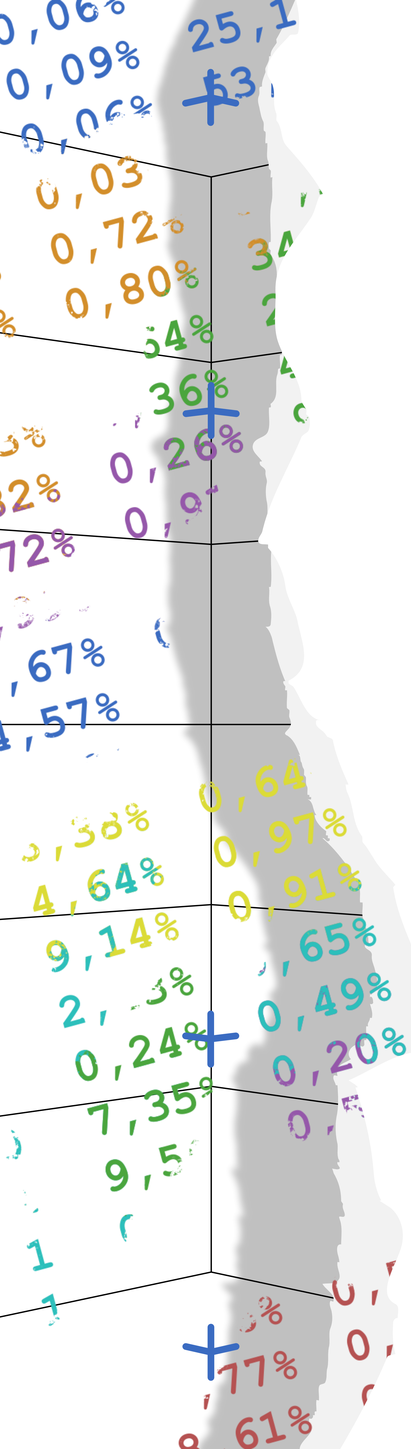
About me...

I'm Nils – a Data Analyst/Scientist driven by curiosity and a thirst for challenges. Coming from a creative background, conception and production of audiovisual media, public relations, and marketing management, I eternalized communicating stories as a key aspect of my past. Now I want to connect my skills with an intellectual challenge in the work of a Data Analyst/Scientist. My eagerness to learn and my high frustration tolerance let me enjoy pressure situations with calmness and ease, making me a reliable and supportive colleague. I'm looking for a fast-paced environment with a team that enables growth and independence on its path to build a better tomorrow.

To never stop coding, I work with codewars (clickable link):

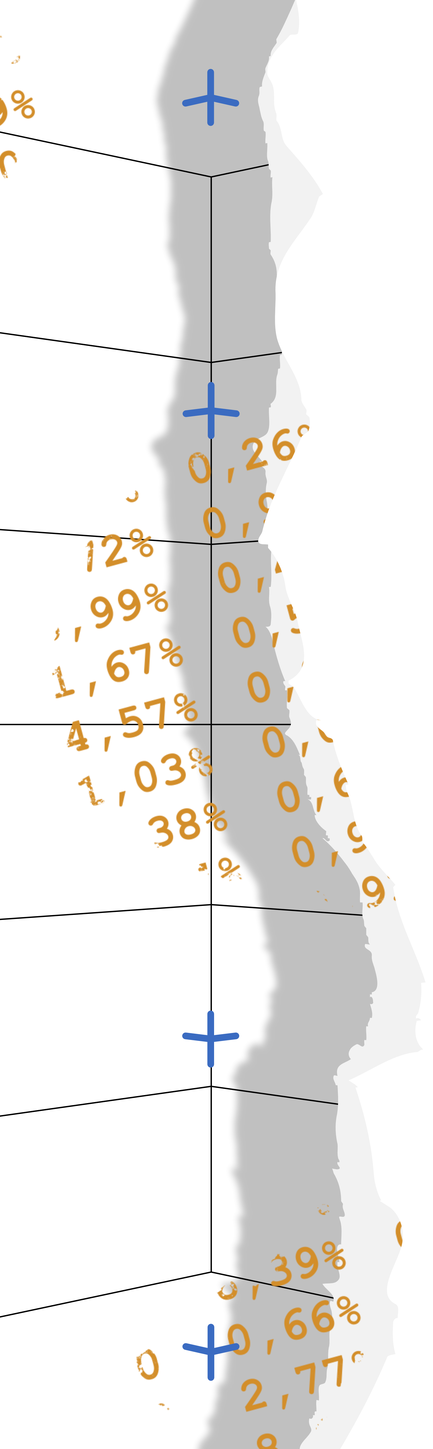
Visit my GitHub or LinkedIn profile to find out more:

Throughout my journey, I worked with the following programs:



Portfolio Projects

1. GameCo - Video Game Popularity (Excel)
2. MedCo - Preparing for Influenza Season (Excel, Tableau)
3. Rockbuster Stealth Data Analysis Project (PostgreSQL)
4. Instacart grocery Basket Analysis (Python, Jupyter)
5. Metabolites in the Wastewater (Python, Jupyter, Tableau)



GameCo

Challenge

As an analyst for a fictional video game company, my goal was to execute a descriptive analysis of a video game data set. In short, finding insights to better the company‘s understanding of how new games perform in the decreasing market.

Process

Using Excel a first overview about the situation at whole was created. After finding the most popular video game genres in the global market, the situation for these genres was observed. The most promising genre in terms of units sold per count of published title could be identified.

Result

With the new information recommendations were phrased. Due to decreasing overall sales it is important to redefine the marketing budget and to focus on opportunities for future profit. The facts provided by the analysis gave an answer to each topic.

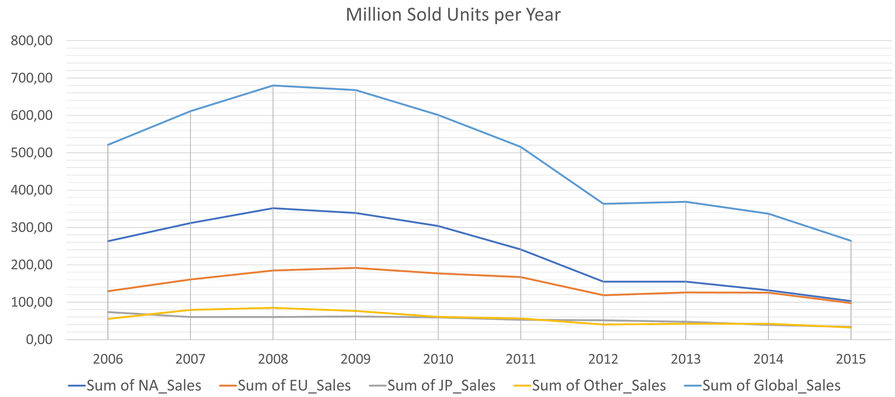
Key aspects of this project:

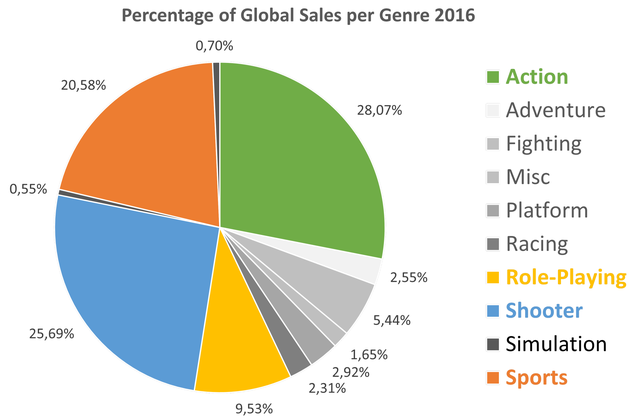
The project was executed with Excel. Conducting a descriptive analysis and give recommendations based on insights were the essential goals of the project.

GameCo | Basic insights

- As the sales throughout the regions decrease, finding opportunities in the market becomes essential

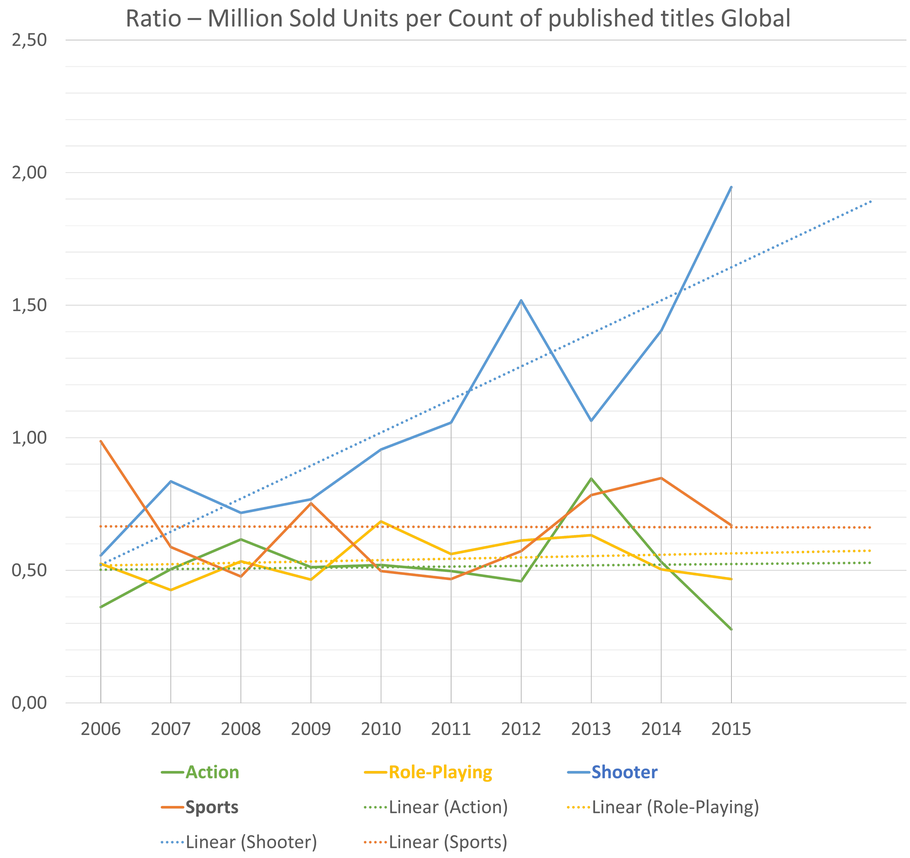
- The most popular video game genres are:
  - Action
  - Role-Playing
  - Shooter
  - Sports

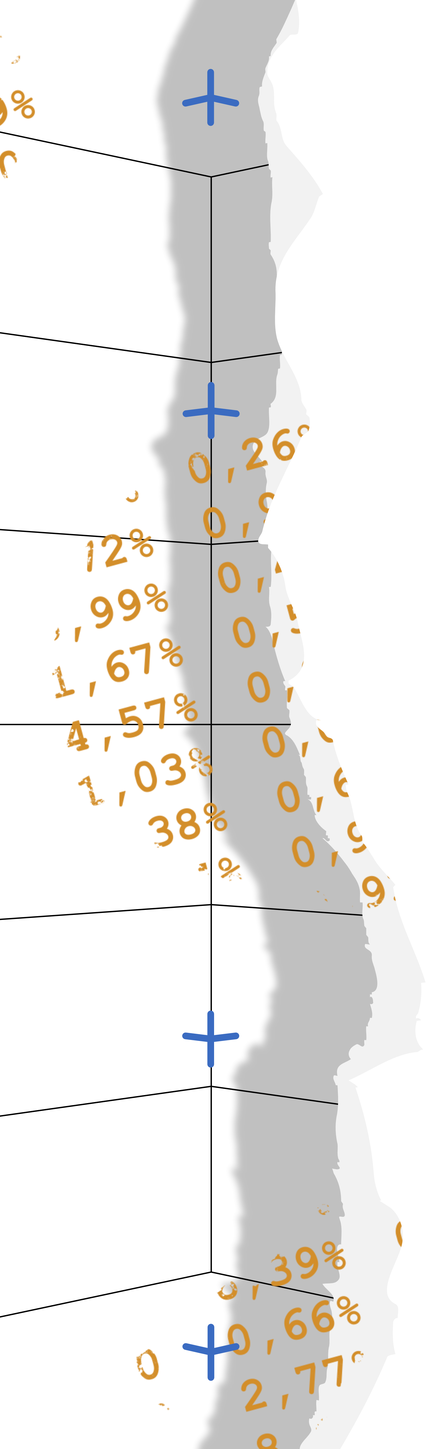




GameCo | Finding future potential

- Focusing on the most popular genres:
  - Shooter sells more units per published title than any other genre
  - In 2015, 34 Shooter games were published and 66.15 million units were sold
  - Trend is increasing steadily
- Despite the general trend of decreasing sales, the Shooter genre offers future potential to be used





GameCo | Recommendations

Decreasing Sales:

- Compared with past years, the trend towards decreasing sales should be investigated further
- Are people switching to different systems (Mobile Gaming, Streaming, Cloud-Gaming Services,…) or is there an overall decrease in popularity of Video Games?

Marketing Budget:

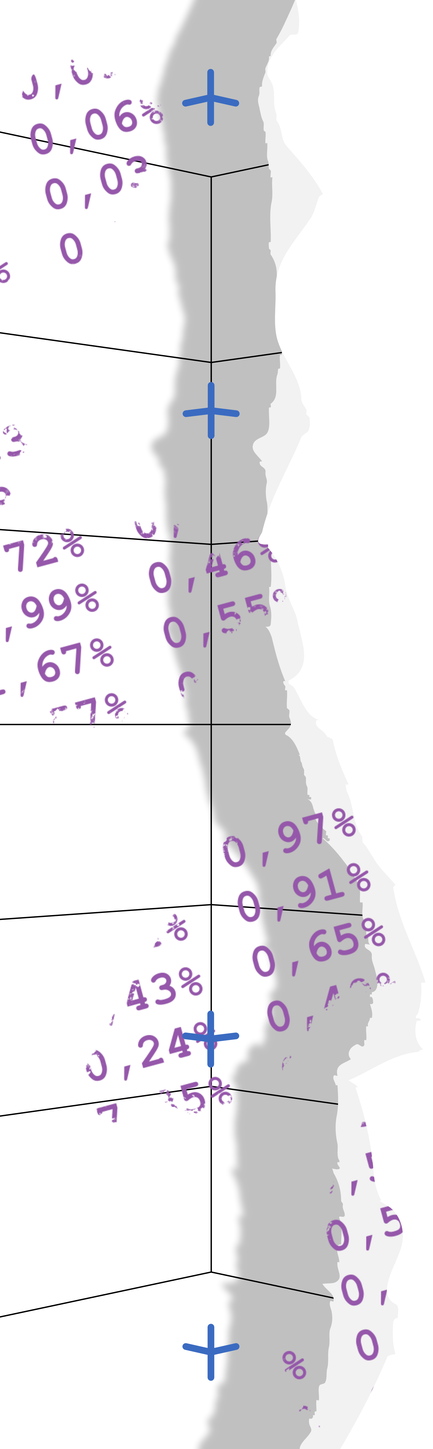
- EU and NA have still the highest market share
- EU will likely surpass NA in the coming years, increasing marketing budget for EU is advised
- JP and Other are decreasing slowly but steadily

Focusing on Shooter Games:

- The newfound popularity for Shooter Games gives an opportunity to gain from the trend of increasing sales per game

The data set was created using information from: https://www.vgchartz.com/

For further information visit the project presentation on my GitHub:



MedCo | Preparing for Influenza

Challenge

Helping a medical staffing agency which provides temporary workers to hospitals and clinics on an on-need basis, my goal was to perform an analysis on past clinical data helping developing a staffing plan for the influenza season in the US.

Process

Multiple datasets from the CDC (Centre for Disease Control and Prevention) and the US Census Bureau were cleaned, utilized, and merged. Using the demographic information and data about previous influenza deaths, a high risk population was statistically significant identified. A staffing plan was created, based on these two KPIs

Result

With the created staffing plan, all states could be sorted into three priority groups. Each group needs a different amount of staffing support, based on past data. Furthermore, the calendar weeks in which the support is needed were identified.

Key aspects of this project:

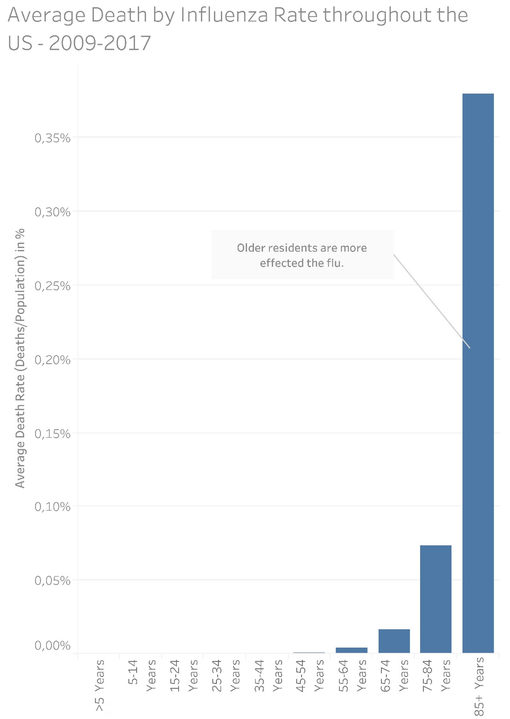
The important parts here, were to merge multiple datasetswith Excel and to export the datasets into Tableau to visualize the key insights of the project.

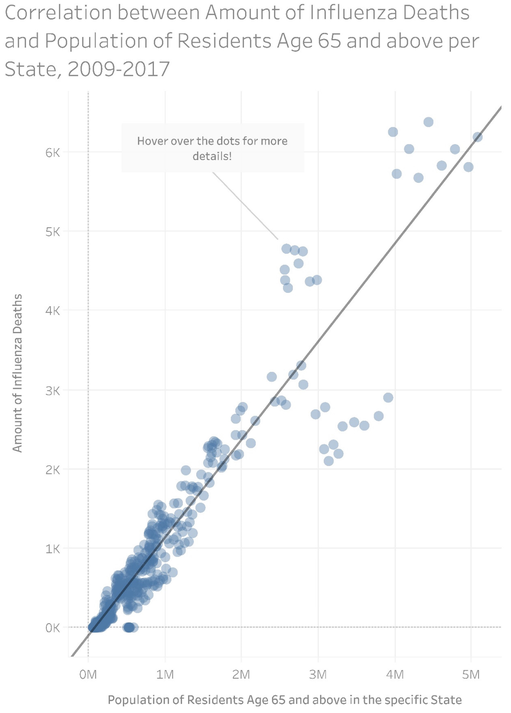
MedCo | High risiko population

- On the left chart we can see, that the most deaths occur in the population above 64 years

- The scatterplot on the right investigates the impact of population of residents above 64 and the amount of influenza deaths per state

- With an R-Value of around 0,95 - meaning that there's a strong correlation
  - Therefore, If the population of seniors (residents above 64) in a state increases, the amount of influenza deaths increases as well





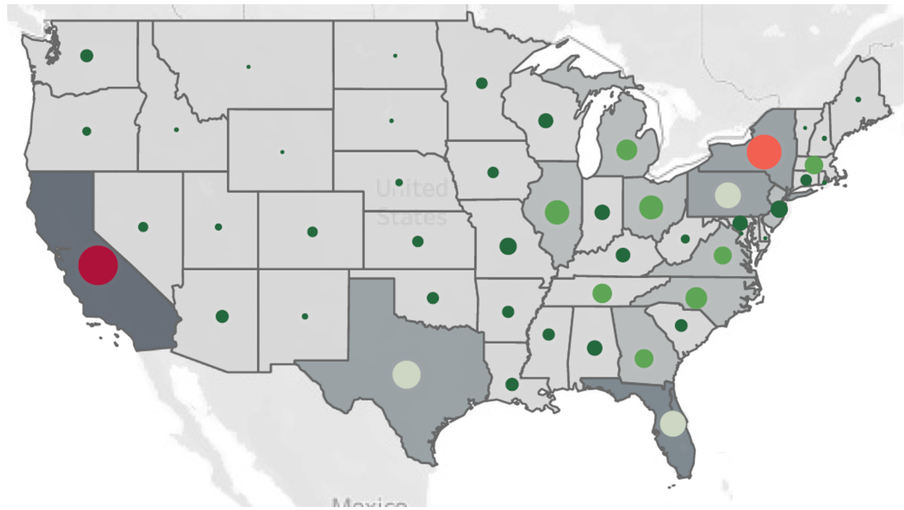
MedCo | The risk in the states

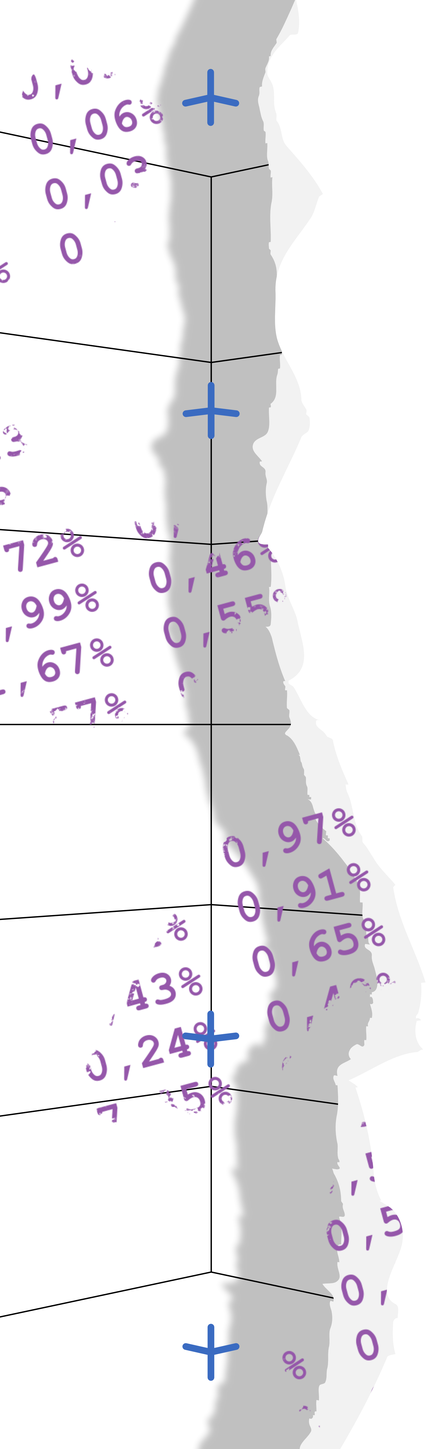
- The map on the right shows the average number of influenza deaths (green to red) and the average population of residents above 64 (shades of grey) per state

- Using these two variables, we can assess that states with a high number of influenza deaths and a high population of seniors pose a risk for a severe influenza season

- For an easier communication the states have been summarized into 3 categories
  - Decisive for influenza deaths is the population of seniors, summarizing the states with the highest population of elders and most deaths into the high priority group:
    - California, Florida, New York, Pennsylvania and Texas
  - The mid priority group consists of below average deaths and population of elders:
    - Georgia, Illinois, Massachusetts, Michigan, New Jersey, North Carolina, Ohio, Tennessee and Virginia
  - The low priority group consists of the least deaths and the least population of elders:
    - The rest

- **Theoretically a risk assessment for every state could be created, but due to the limitations of the project data this was excluded.**





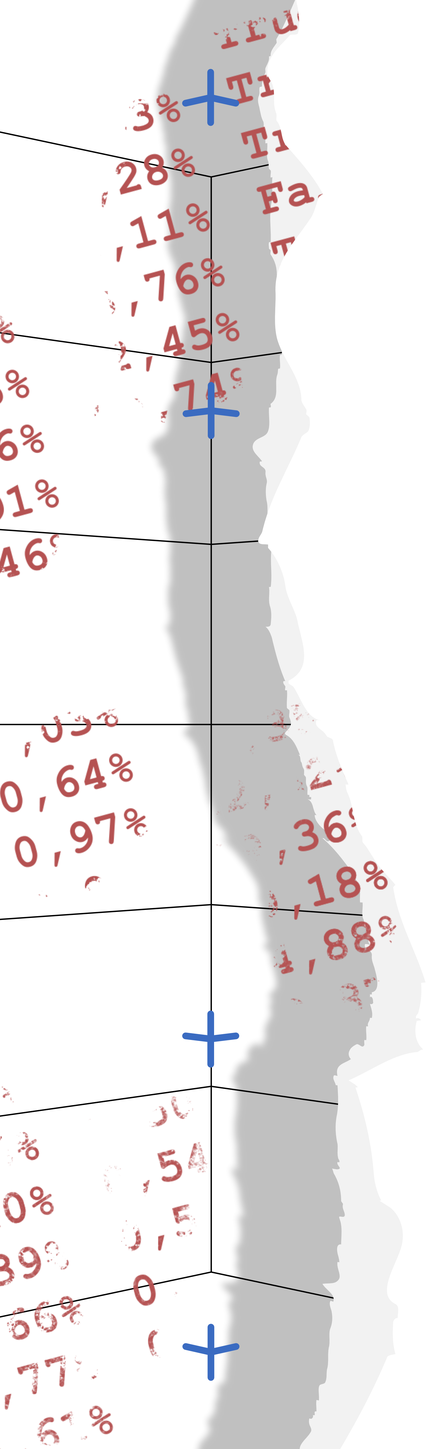
MedCo | Recommendations

- As mentioned before, high priority states should receive the best staffing support, followed by mid and low priority states
  - However, the high priority group should be split into a ranking, since the demands differ slightly looking at past data, in conclusion the rating is as follows: California, New York, Florida, Pennsylvania and Texas
  - Regarding the past data it is likely, that California will have the most deaths followed by New York

- The staff is needed around the calendar weeks 1-14 and 46-53 as described in the Tableau presentation

The Data was provided from the Center for Disease Control and Prevention (1, 2, 3) and the US Census Bureau.

For further information visit the project presentation on my GitHub:



Rockbuster | Streaming & SQL

Challenge

Hired by fictional Rockbuster Stealth‘s business intelligence department, my goal was to perform an analysis to further improve the online video rental store‘s understanding about its movie inventory and its customer base.

Process

Starting with the cleaning of data and a basic descriptive analysis, using SQL, the movie categories were ranked by popularity and earnings. Furthermore, the countries with the most profits were ranked and further analyzed to gain the insights needed for the management board.

Result

With the insights of popular categories and individual spending behaviour of customers in high earning regions, the Rockbuster Stealth management board can specialize its marketing efforts for more effective results.

Key aspects of this project:
This project was conducted with PostgreSQL, the goal was to write basic and advanced queries, exporting datasets into Tableau and to visualize the newfound insights.

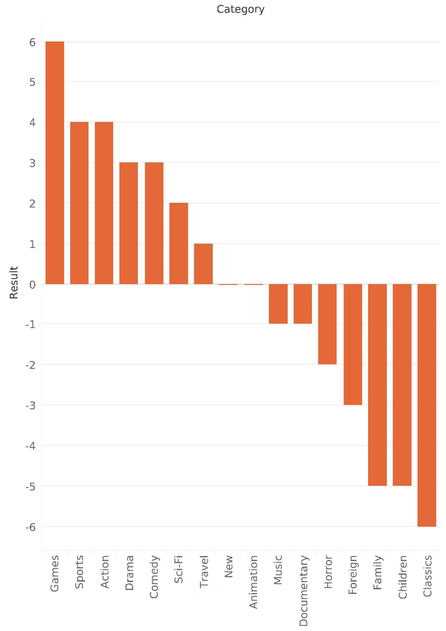
Rockbuster | Category ranking

- To extract the most popular genres (a.k.a. categories), I set up a ranking on the top/bot 100/250 movies grouped by category based on total earnings

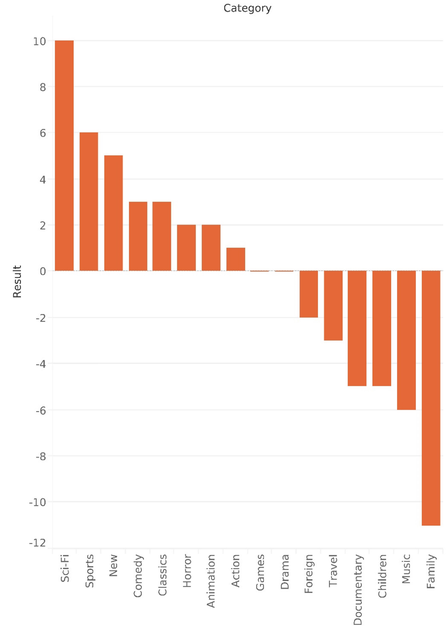
- I subtracted the top lists rank with the bot lists rank of each category to see which categories are popular based on their counts in the top/bot 100/250 list

- Popular genres are:
  - Sci-Fi, Sports, Comedy and Action are popular genres
- Unpopular genres are:
  - Family, Children, Foreign, Music and Documentary are unpopular genres

Subtracting the rank of the Bot 100 Categories by Earnings off the Top 100 Categories by Earnings



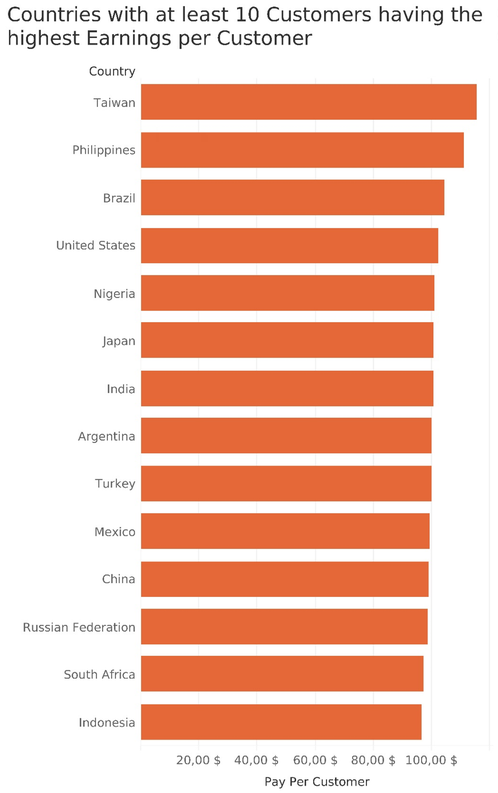
Subtracting the rank of the Bot 250 Categories by Earnings off the Top 250 Categories by Earnings

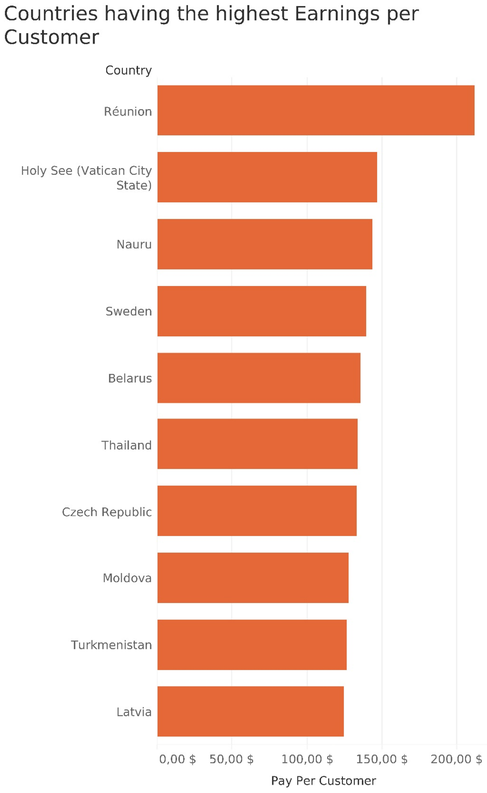


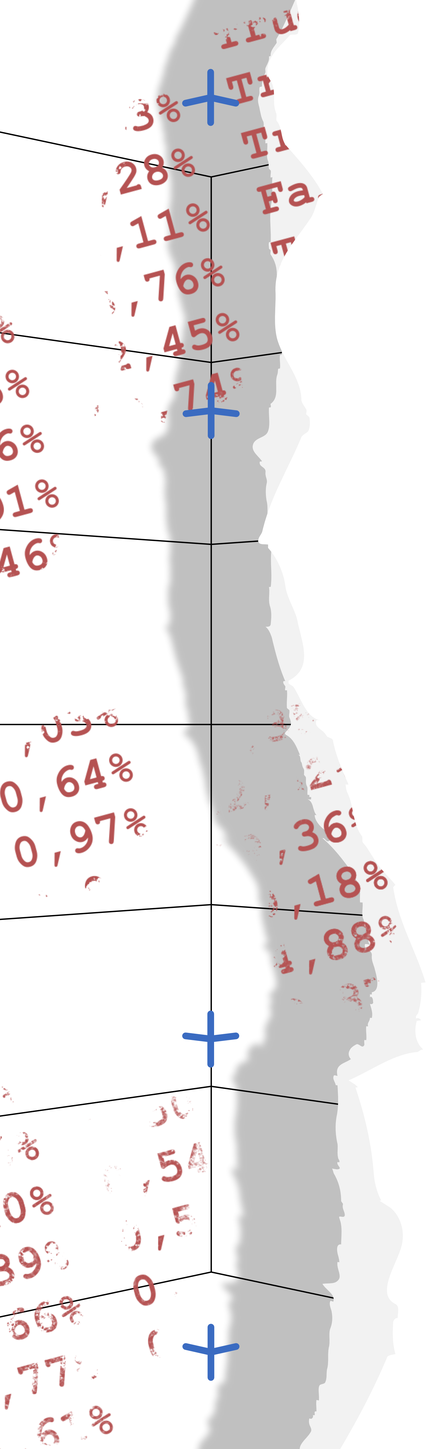
Rockbuster | Pay per customer

- The top 5 countries with highest earnings per customer while having at least 10 customers are:
  - Taiwan, Philippines, Brazil, United States and Nigeria
  - The maximum pay per customer value is inhabited by Taiwan with ~ 120 $ / customer

- However, some countries have much higher earnings per customer, such as:
  - Réunion, Vatican State, Nauru, Sweden and Belarus, Thailand, Czech Republic, Moldova, Turkmenistan and Latvia
  - All exceed the ~ 120 $ / customer from Taiwan
  - In these countries are few customers giving high earnings to Rockbuster







Rockbuster | Recommendations

- Focusing on more popular genres, when licensing new movies, like Sci-Fi, Sports, Comedy and Action can impact the sales positively

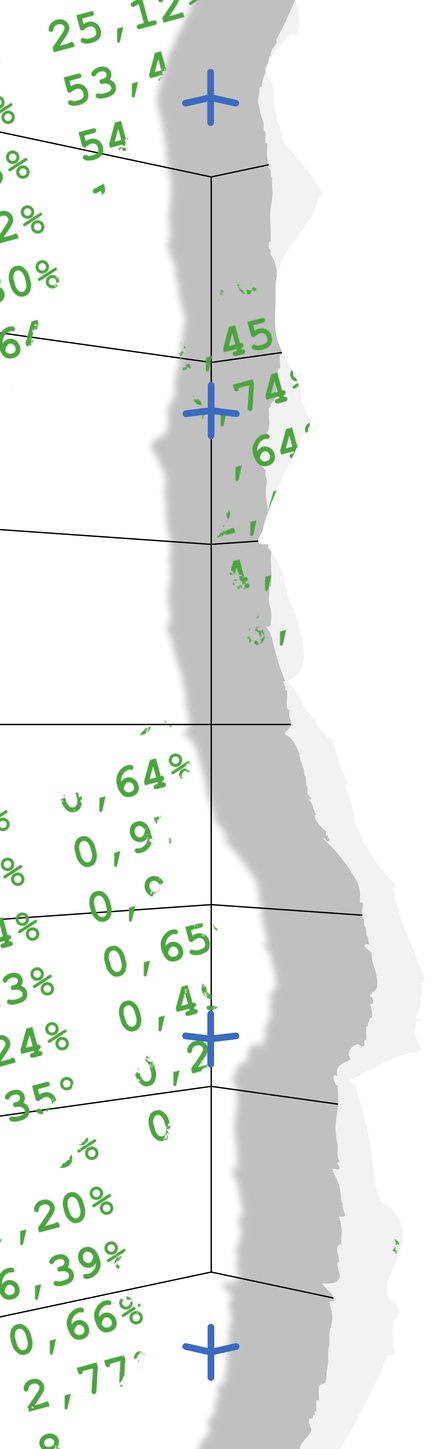
- Putting less focus on genres, when licensing new movies, like Family, Children, Foreign, Music and Documentary can impact the earnings by movie

- Taiwan, Philippines, Brazil, United States and Nigeria offer a general high amount of earnings per customer and provide potential for greater future earnings
  - Increasing the marketing budget is advised

- Despite having few customers some countries have exceptionally high earnings per customer and give an opportunity to increase the client base to gain better earnings

The data stems from PostgreSQL Tutorial.

For further information visit the project folder on my GitHub.



Instacart Grocery Analysis

Challenge

Working for Instacart, an online grocery store, my task was to gain new insights about their sales patterns and their customer order behaviour. Considering a targeted marketing strategy, the Instacart stakeholders needed information about customer groups in question.

Process

After merging and cleaning multiple datasets, I created a variable system that sorts every customer into 512 possible profiles using 6 variables with 3 times 4, and 3 times 2 different ordinal numbers. Looking at the sum of orders of each profile, I identified high ordering profiles, used for comparison with the rest of the customers to identify key demographic indicators.

Result

Using the new demographic indicators, I can provide valuable information to the stakeholders that give opportunity to increase the earnings off of these customers and the efficiency of the marketing towards these customers.

Key aspects of this project:

The whole project was conducted with Python, merging and cleaning multiple different datasets and gaining insights using pandas, numpy, matplotlib, and seaborn were key parts of the process.

Why 512 profiles?

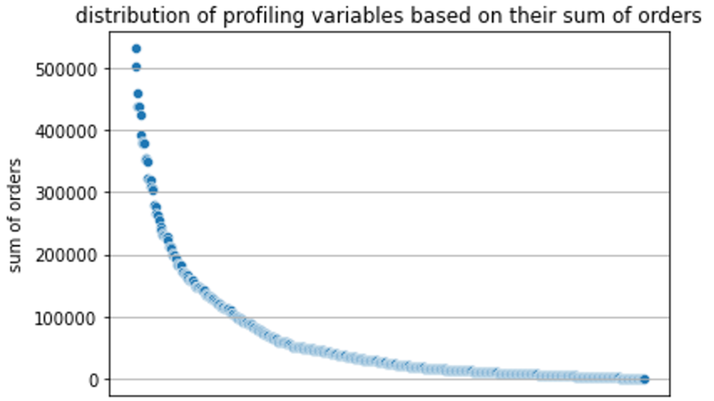
- After merging all datasets together, 6 columns were created that flagged the data with 3 times 4 and 3 times 2 different nominal variables. With this, the data could be splitted/sorted into either 3 times 4 different or 3 times 2 different groups. However, this would only allow to draw insights on either the splitted groups together or each. Therefore, a new column was created which used the previous variables to create 1 out of the 512 possible customer profile variables.

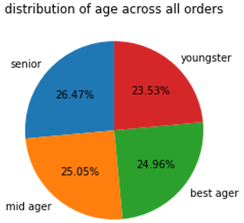
  Having sorted the data into the profiles, it was possible to draw conclusions not based on the 6 different columns at once, but on all the profiles together.

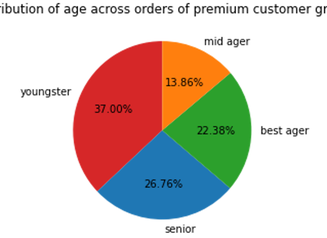
- In short, instead of splitting the data into either 2 or 4 parts, I splitted it into 512 possible parts.

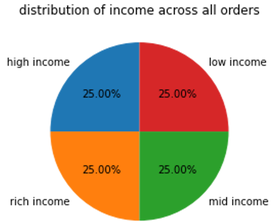
512 profiles

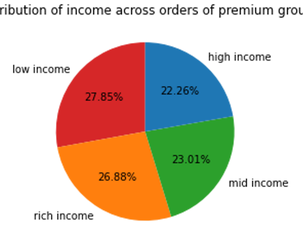
- Using the demographic information from the customers, I created 512 possible customer profiles
- Out of these 512 profiles only 509 existed, furthermore 26 profiles are responsible for around 1/3 of all orders (= premium profiles)
- The chart on the top right demonstrates the distribution of profiles based on their sum of orders
- The 26 premium profiles were used to compare their behaviour with the behaviour of all customers
- The premium profiles are mostly Youngsters and Seniors with low and rich income, having 1-3 dependants in the household, are married, have no babies, and no pets









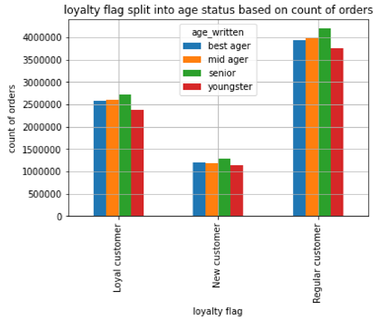


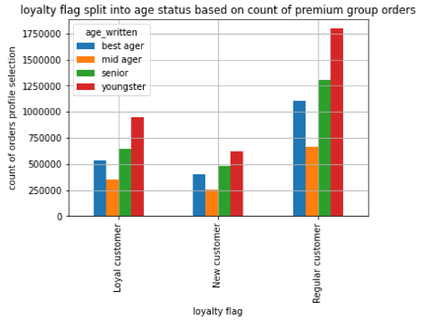
Comparing the profile groups

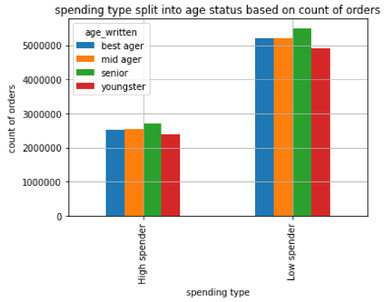
- Comparing the behaviours of the customers (left) with the premium group customers (right) shows the difference of the profile groups

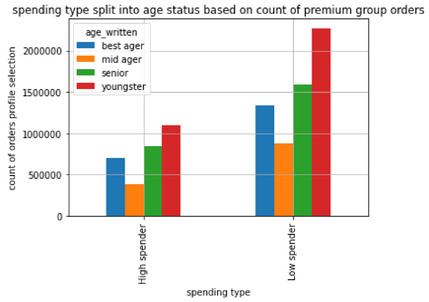
- The order behaviour of all customers suggests, that the youngsters have always the least amount of loyal, regular, and new customers, the least amount of high spenders, and low spenders, and the least amount of non-frequent, regular, and frequent customers

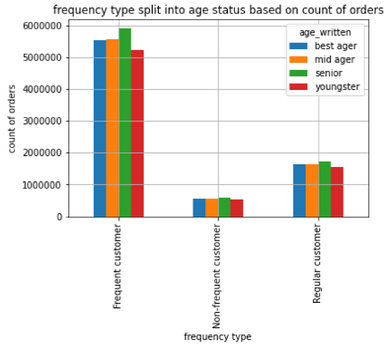
- However, the order behaviour of the premium customers shows, that the youngsters dominate in every group across every loyalty, spending, and frequency profile.

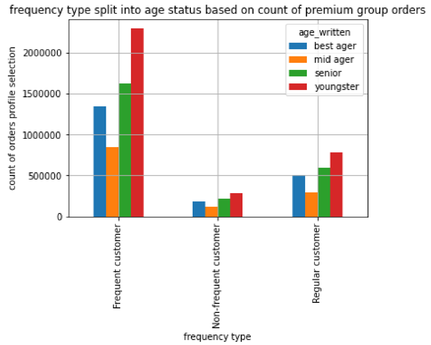


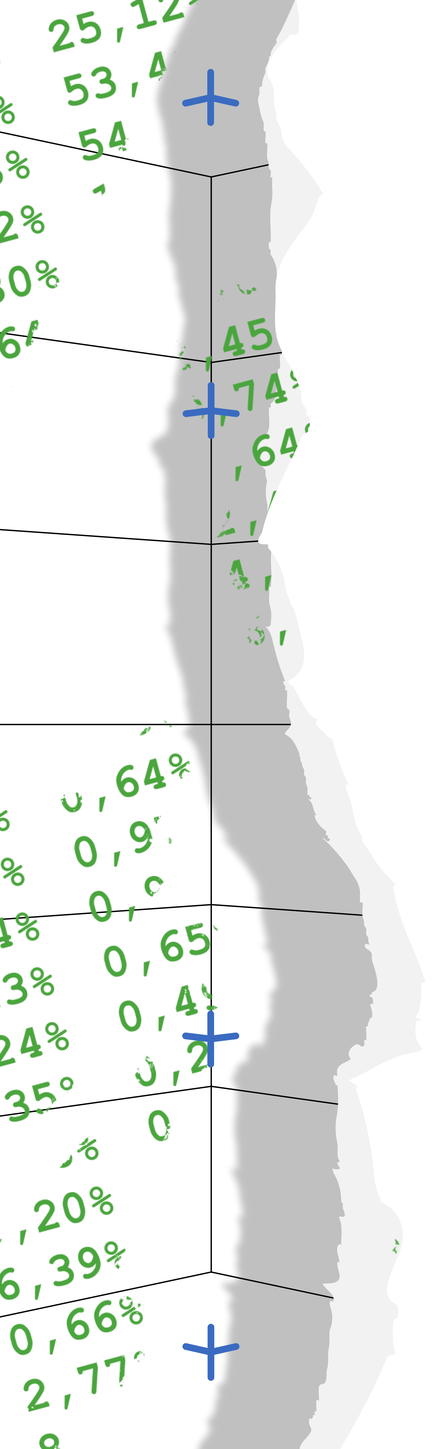












Instacart | Recommendations

- The customers ordering the most of Instacart are usually, low-income youngsters being married, having 1-3 dependants in their household, having no babies, and no pets

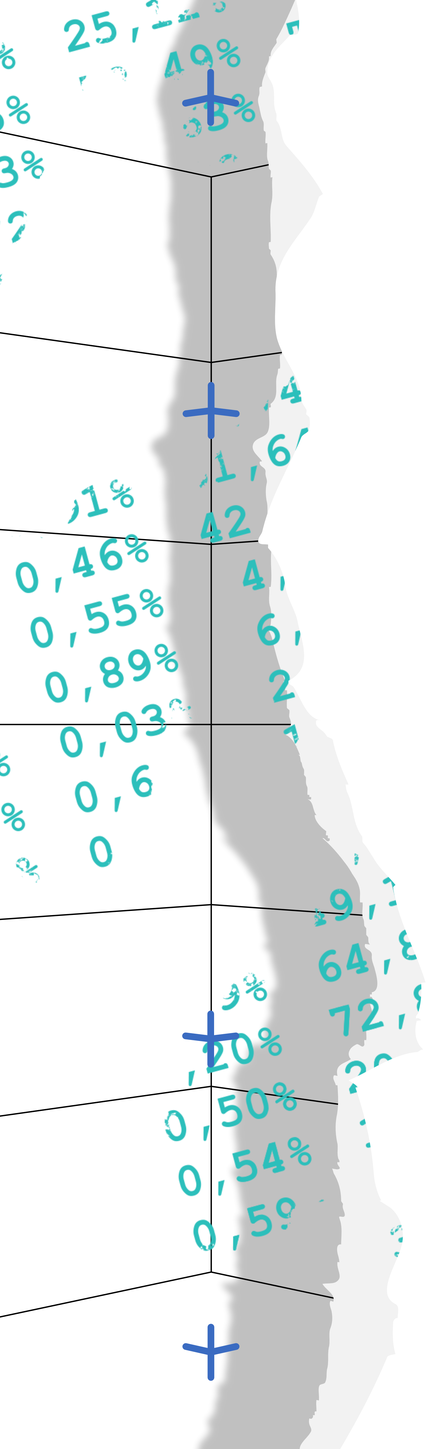
- The youngsters also dominate in every customer profile across the premium customer selection

- Increasing the earnings off youngsters and increasing the client base in the youngster age group using the demographical information of the premium group can positively impact Instacart‘s performance.

The data used is “The Instacart Online Grocery Shopping Dataset 2017”, Accessed from https://www.instacart.com/datasets/grocery-shopping-2017 on 08.06.2022 [DD-MM-YYYY].

For further information click here for to the project on my GitHub

To get directly to the final report click here



Metabolites in the Wastewater

Challenge

In this open data project, the goal was to analyze the drug consumption situation in Europe during the pandemic, using datasets of metabolites in the wastewater. Furthermore, the correlation of specific drug consumption relationships was observed.

Process

After cleaning, merging, and a descriptive analysis, heatmaps were created to identify possible relationships. The drug consumption combinations with the highest correlation were further analyzed and tested. Afterwards machine learning algorithms were used to conduct a forecast.

Result

With the newfound insights, the impact of the COVID-19 pandemic on the drug consumption could be observed. Looking at the correlations between the drug consumption combinations, the fight against drug abuse has to be viewed as a whole project, and not as an individual conflict for each drug.

Key aspects of this project:

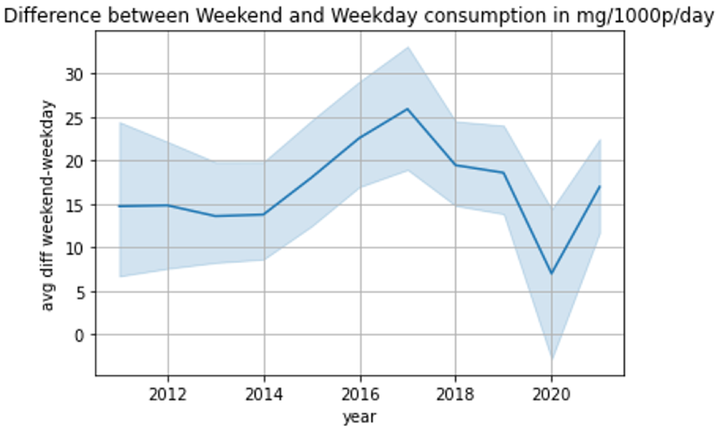
The open data project was conducted with public available data and Python. Furthermore, linear regression, kmean clustering and ARMA/ ARIMA forecasting were advanced techniques, using machine learning for prediction, executed in this project.

The impact of COVID-19

- To test the impact of the pandemic in the consumption behaviour, a chart was created showing the difference between weekend and weekday consumption in milligramme per 1000 persons per day.

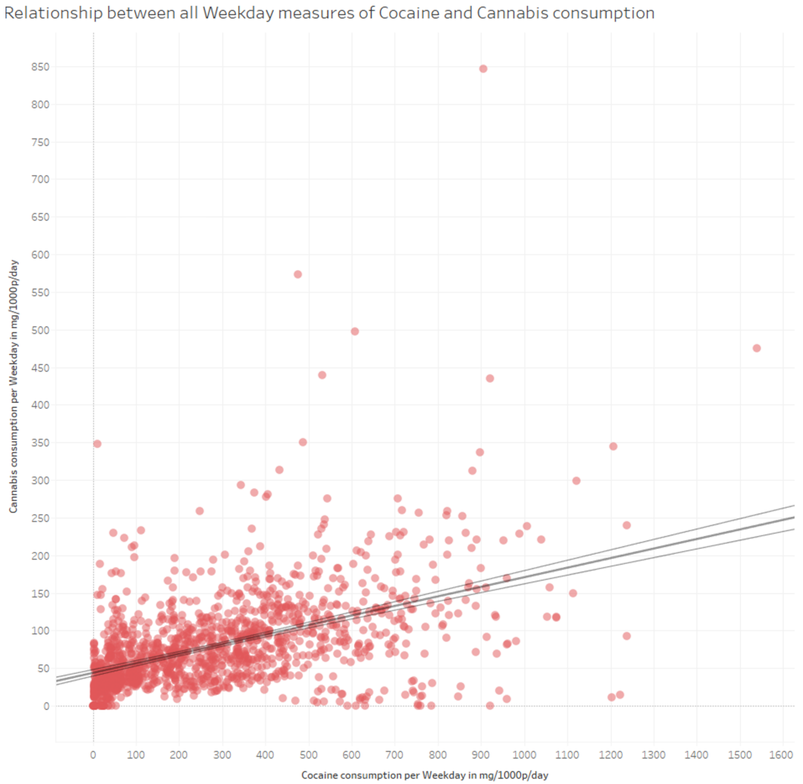
- The formula for the difference is:
  - Weekend – weekday = value

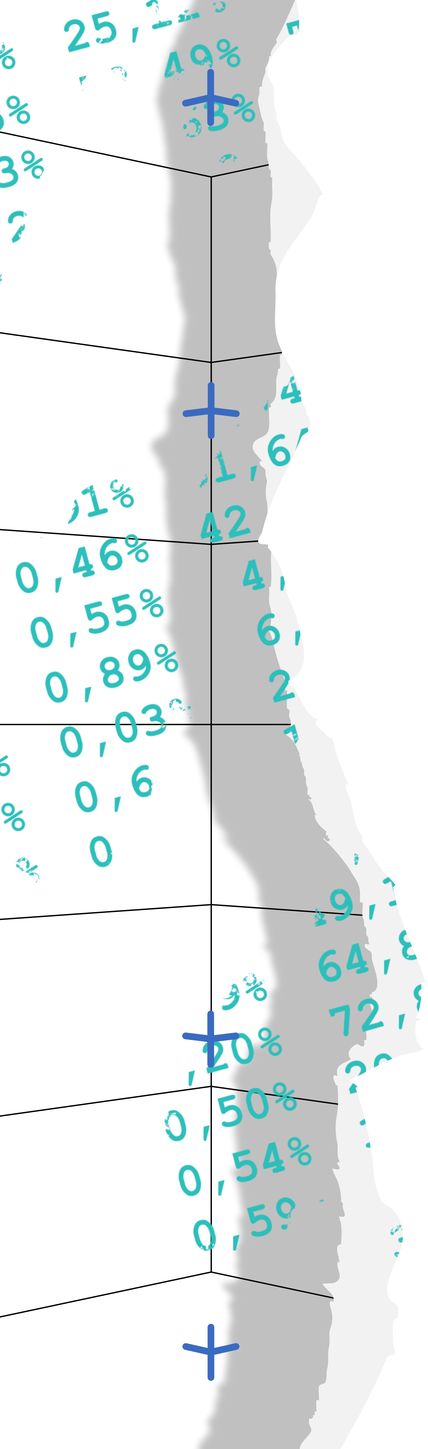
- As we can see, the average difference shows a nadir in the year 2020, the pandemic year with the lockdowns in various states.



Drug consumption combinations

- To test the consumption behaviours of specific drug combinations, all measurements were taken and tested.
- Here the combination Cocaine and Cannabis are an example for the two other combinations (Cocaine x MDMA and MDMA x Cannabis) – the results are similar.
- The chart shows us R ≈ 0,495982, meaning the combination has a moderate correlation.
- This result was likewise with the two other drug combinations.





Metabolites | Recommendations

- In 2020 most european states reintroduced the lockdown as a method to control the spread of COVID. This impacted the drug consumption as the average weekday consumptions converged with the average weekend consumption. As the weekday consumption is usually lower than the weekend consumption, more drugs were consumed during the weekdays.

- In the end, although a certain correlation is provable, it is also an evidence that other factors are relevant as well. Therefore, we can‘t look at the consumption levels together but at the individual levels and research for further insights that explain the levels of each metabolite/drug.

The data used was published by the European Monitoring Centre for Drugs and Drug Addiction and the SCORE network.

For the Tableau presentation regarding this project click here.

For further information click here to the project folder on my GitHub.