Sandra Wienecke

Data Analyst Portfolio

Introduction

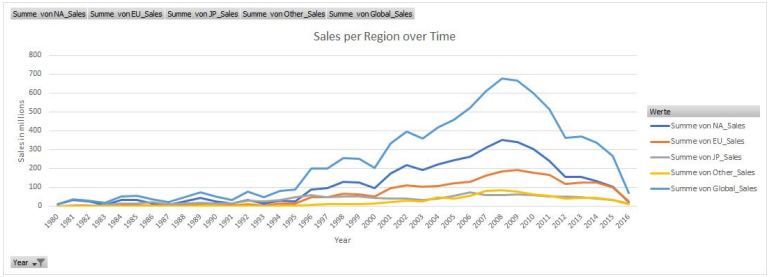
This portfolio includes various projects where I have analyzed and transformed data into understandable and valuable insights.

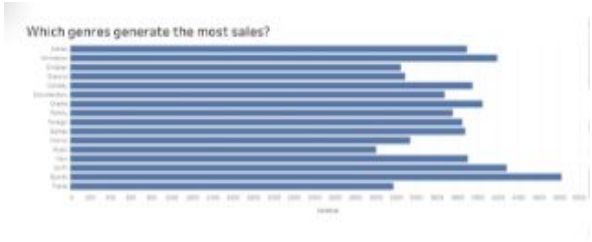
The goal of my analyses was to improve the results so that the stakeholder can make correct decisions and save costs.

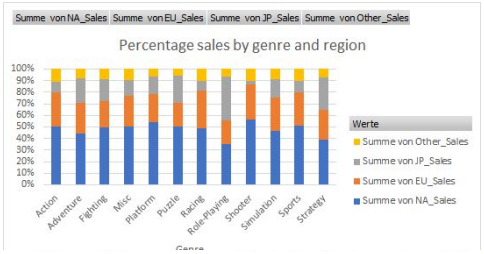
The next two slides provide an overview of the tools and visualizations used in the data analysis.

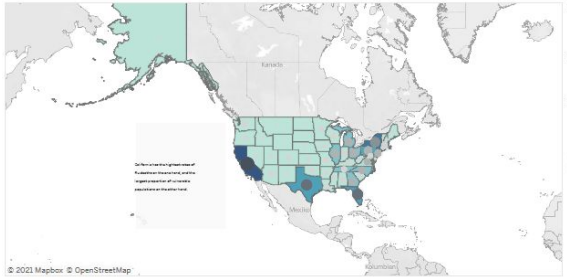
Tools & Libraries

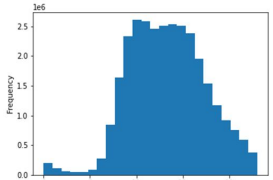
Visualizations

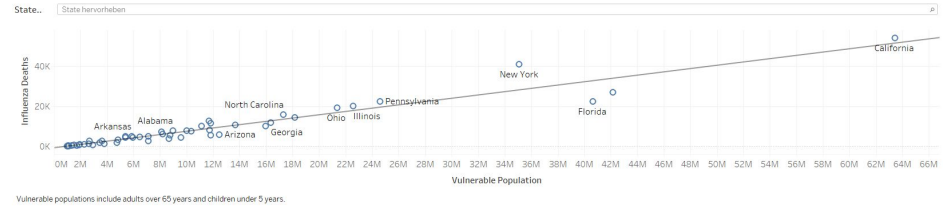


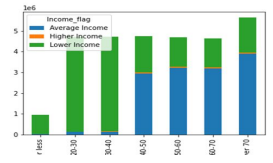


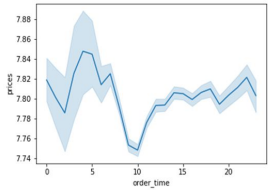












Project Overview

GameCo	Marketing Recommendations Marketing Data & Behavioral Data
Temporary Staff Allocation	For the next flu season (USA) Public Health Data & Demographics
Rockbuster Stealth	Online Service Launch Customer Data & Financial Data
Instacart	Grocery Basket Analysis Customer Data, Behavioral Data & Marketing Data

Case Study 1: GameCo

Marketing Recommendations

Purpose & Context

You're an analyst for a new video game company, GameCo, which wants to use data to inform the development of new games. As such, you've been asked to perform a descriptive analysis of a video game data set to foster a better understanding of how GameCo's new games might fare in the market.

Data:

Video Game Sales data set covers historical sales of video games spanning different platforms, genres, and publishing studios, Source: VGChartz

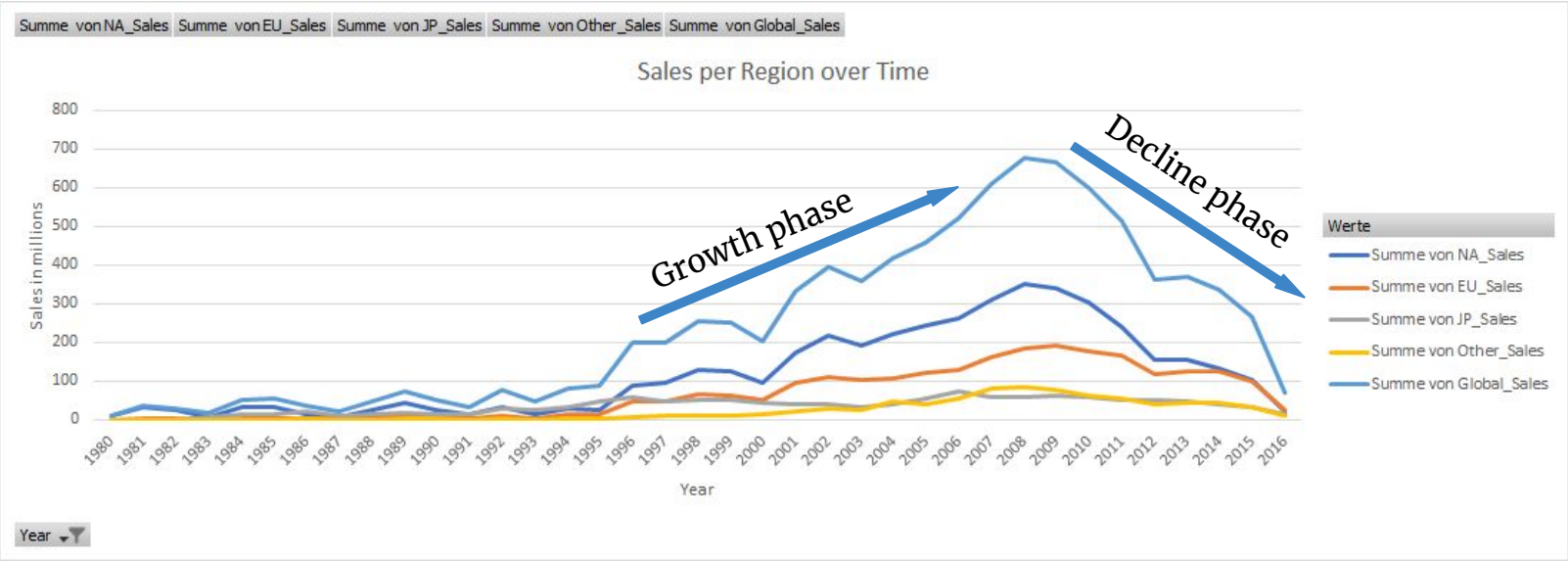
Procedures:

- Data cleaning
- Pivot tables
- Descriptive analysis
- Data visualization & storytelling

Tools: Excel, PowerPoint

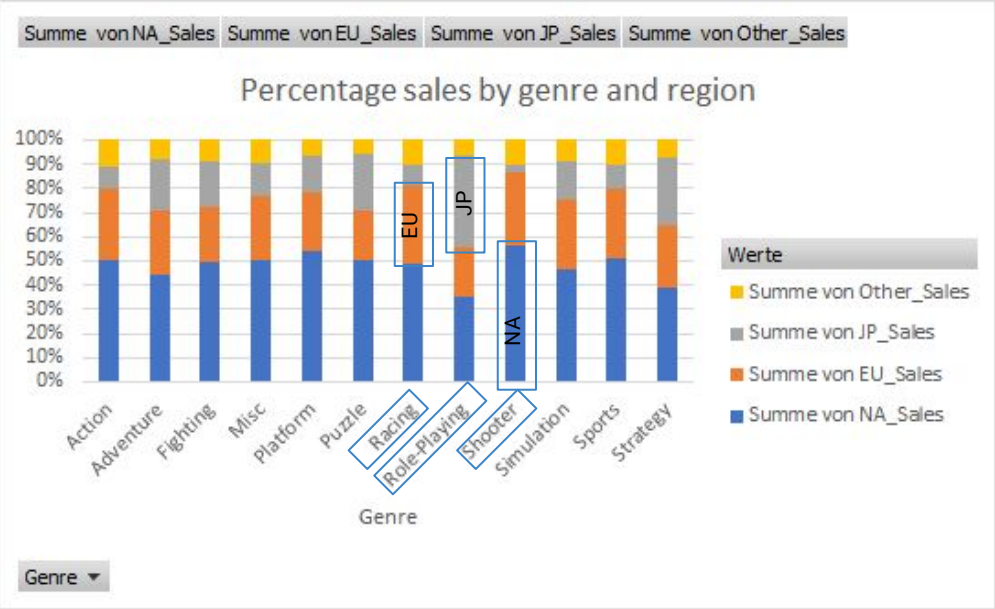
Visualization

Data on GameCo sales per region, from 1980-2016.



The Data suggest that each region didn't stay the same over time. Each region had phases of growth and decline in sales. This contrast with the current understand of the business.

Visualization



Genre preferences vary across each sales regions. This requires a targeted marketing:

- In NA Shooter games are the most popular genre
- In the EU Racing had the most sales
- In Japan (and the Others) role-playing games are popular

Conclusion & Recommendation

Marketing budget should be distributed by considering regional trends and preferred game genres:

- EU has been gaining market share for the past years with focus on genre action and sports. So with further marketing budget the growth can be continued to expand.
- NA is the second largest but declining market. The major genres are action and sport. With more marketing focus, NA can possibly reactivate.
- JP is the most shrinking market. Further investigation is required to understand the underlying factors for the declines in the last years. Nevertheless, marketing campaigns e.g. for the most popular genre role playing can strengthen brand awareness.

Game Co - Deliverables

- Final Project Presentation
- Project Reflections

Case Study 2: Temporary Staff Allocation

Preparing for the next flu season in the US

Purpose & Context

The United States has an influenza season where more people than usual suffer from the flu. Some people, particularly those in vulnerable populations, develop serious complications and end up in the hospital. Hospitals and clinics need additional staff to adequately treat these extra patients. The medical staffing agency provides this temporary staff. Determine when to send staff, and how many, to each state.

Data:

- Influenza deaths by geography, time, age, and gender date set, Source: CDC
- Population data by geography from the US Census Bureau

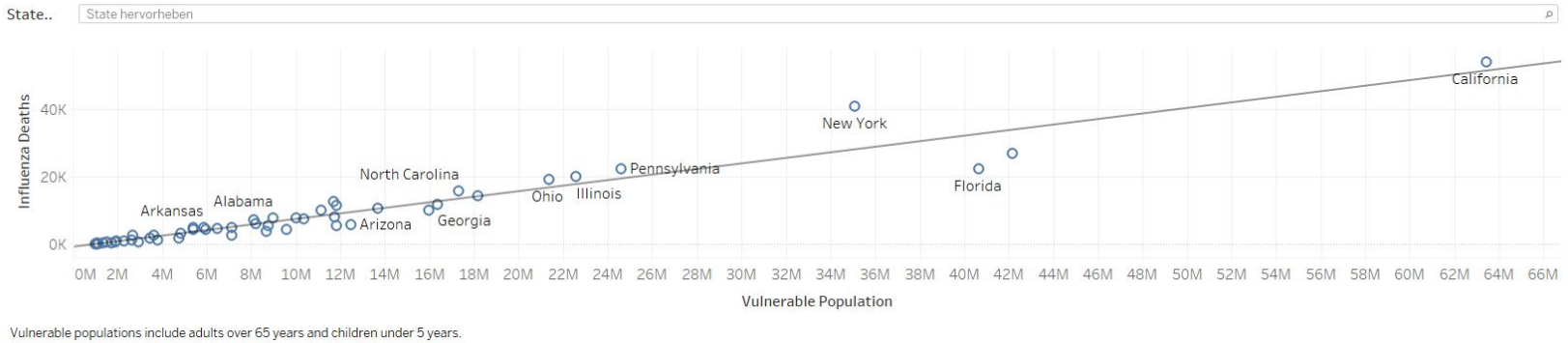
Procedures:

- Data cleaning
- Data interpretation & transformation
- Statistical hypothesis testing
- Visual analysis & storytelling with Tableau

Tools: Excel, Tableau

Visualization

Focus of analysis: Vulnerable population & influenza deaths



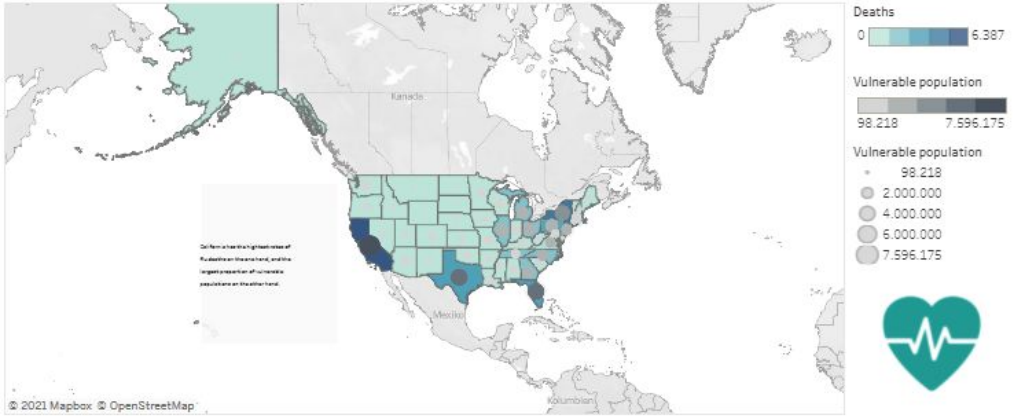
There is a strong correlation between number of influenza deaths and vulnerable population.

The greater the proportion of vulnerable population, the greater the number of people who die from influenza.

Conclusion & Recommendation

Where to send additional medical staff for the next influenza season?

The analysis shows that California, Texas, and Florida are the states with the most influenza deaths and the largest proportion of vulnerable populations.



Additional medical staff should be sent to these three states.

Temporary Staff Allocation - Deliverables

- Interim Report
- Presentation for Business Manager in Tableau
- Record of the presentation

Case Study 3: Rockbuster Stealth

Online Service Launch

Purpose & Context

Rockbuster Stealth LLC is a movie rental company that used to have stores around the world. Facing stiff competition from streaming services such as Netflix and Amazon Prime, the Rockbuster Stealth management team is planning to use its existing movie licenses to launch an online video rental service in order to stay competitive.

Data:

Rockbuster data set is a relational database which contains information about Rockbuster's film inventory, customers, and payments

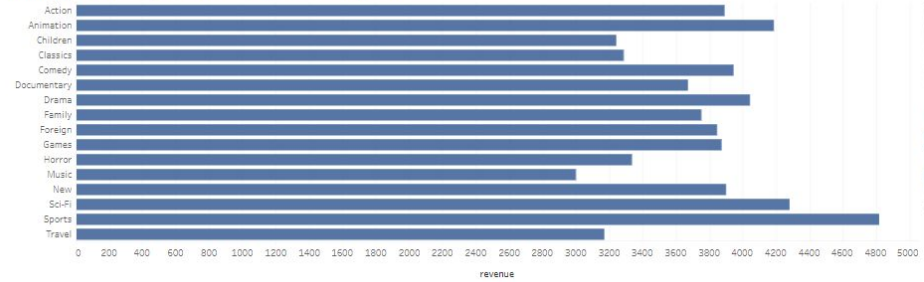
Procedures:

- Summarizing and cleaning data
- Filtering data
- Joining tables
- Subqueries and CTEs

Tools: SQL, DB Visualizer, Tableau

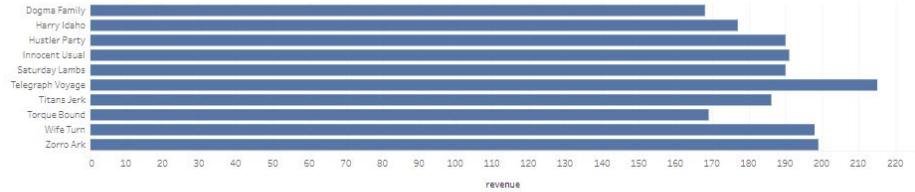
Visualization

Which genres generate the most sales?



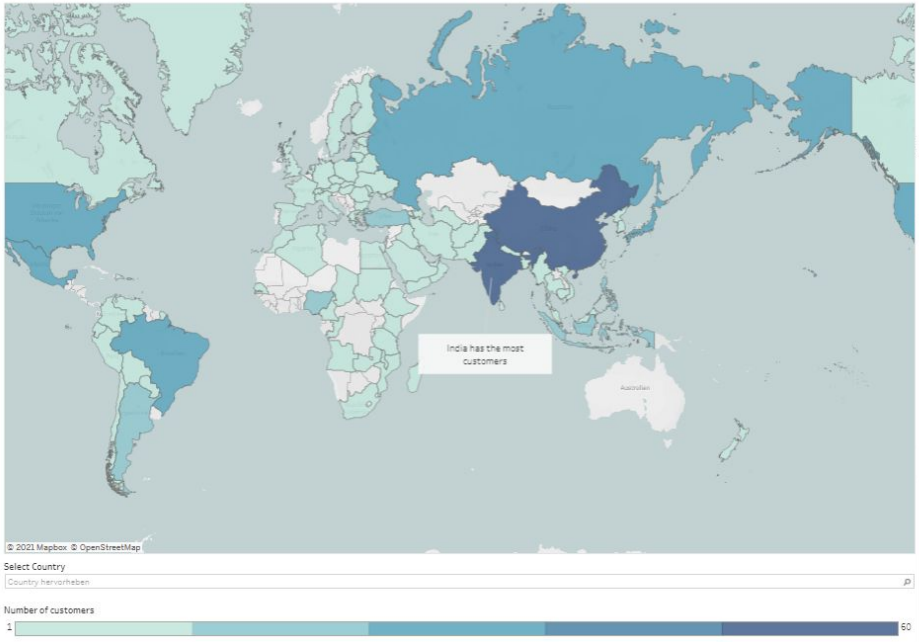
At Rockbuster, the customer can choose between 1000 films and 20 different film genres.

Top 10 films by revenue



Visualization

countries



Rockbuster counts customers in 109 countries.

Top 3 Countries

1. India
2. China
3. United States

Top 5 customers

Customer Id	First Name	Last Name	Country	City	Total Amount Paid
148	[redacted]		Runion	Saint-Denis	211.55
526	[redacted]		United States	Cape Coral	208.58
178	[redacted]		Brazil	Santa Brbara dO...	194.61
137	[redacted]		Netherlands	Apeldoorn	191.62
144	[redacted]		Belarus	Molodetno	189.60

Conclusion & Recommendation

- Focus on and expand the genres that generate the most revenue (Sports, Sci-Fi, Animation)
- Continue to stock and advertise the best films by turnover (Telegraph Voyage, Zorro Ark, Wife Turn)
- Treat best customers in a special way so that they continue to be Rockbuster's best customers, e.g. special gifts, customer programs, etc.
- Focus on countries with the most customers such as India, China and United States

Rockbuster Stealth - Deliverables

- GitHub repository
- Tableau Visualization
- Rockbuster Data Dictionary

Case Study 4: Instacart

Grocery Basket Analysis

Purpose & Context

Instacart is an existing online grocery store that operates through an app. Stakeholders are interested in the variety of customers in their database and their purchasing behaviors. They want to target different customers with applicable marketing campaigns.

Data:

- Open-sourced data sets from Instacart
- Customers data set includes information about Instacart's orders, customers and products
- Data Dictionary
- The Instacart Online Grocery Shopping Dataset 2017

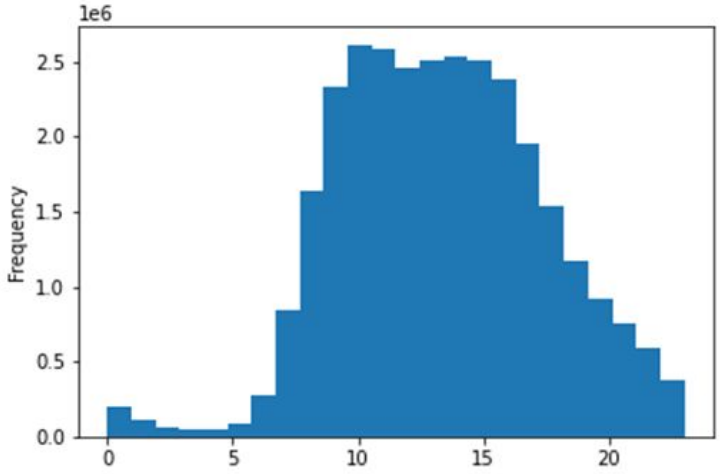
Procedures:

- Data wrangling & Combining data
- Deriving new variables
- Grouping data and aggregating variables

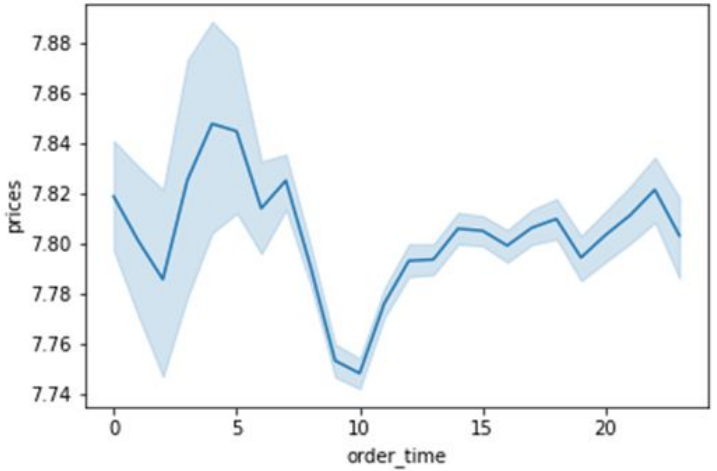
Tools:

- Python and Jupyter

Visualization

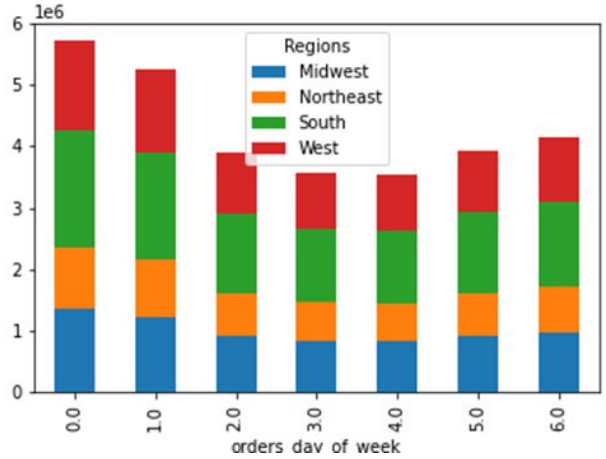


The hours with the most orders are from about 10 to 16. So there are placed in the middle of the day.

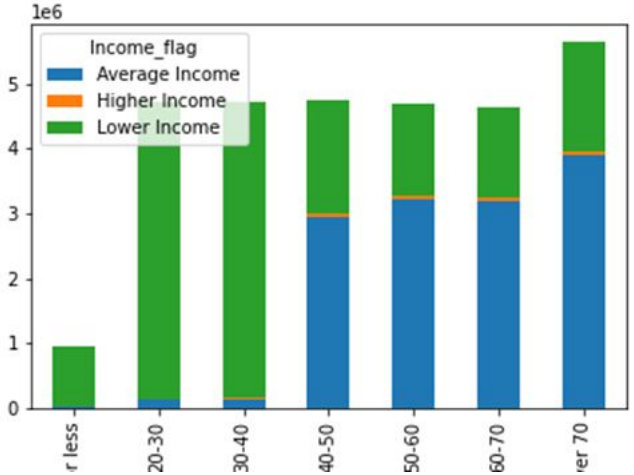


In the graph, we see the order time in relation to the price. The expensive products tend to be bought at the beginning and end of the day.

Visualization



The bar chart shows the orders day of the week for the different regions. You can see that the regions are all buying in roughly similar proportions on that day.



From the age group 40-50 years, the share of average income increases significantly. In all age groups there are only very few with Higher Income.

Conclusion & Recommendation

- Off-peak times are an important time to replenish our products in the warehouses and to plan orders with the manufacturers.
- It is interesting to understand exactly which products are bought at the beginning of the day. This way, they could better position themselves with the products to increase sales.
- All regions show similar shopping behavior. One should therefore make similar advertising campaigns in the regions. And in addition, the warehouses should also be filled similarly.
- We know that our younger customers have a lower income and the older the customers get, the share of the average income increases. This helps us to better tailor the advertising to our different client groups.

Instacart - Deliverables

- GitHub repository
- Final Report

Case Study 5: Immobilienscout24

Apartment rental offers in Germany

Purpose & Context

Rental offers data set was scraped from a private person from Immobilienscout24. The data set contains information about apartment rental offers in Germany, including their characteristics, e.g. location, costs, building fabric & energy certificate, equipment.

Data:

- Open-sourced from Kaggle
- Rental offers data set contains offers from the dates 2018-09-22, 2019-05-10 and 2019-10-08.

Tools:

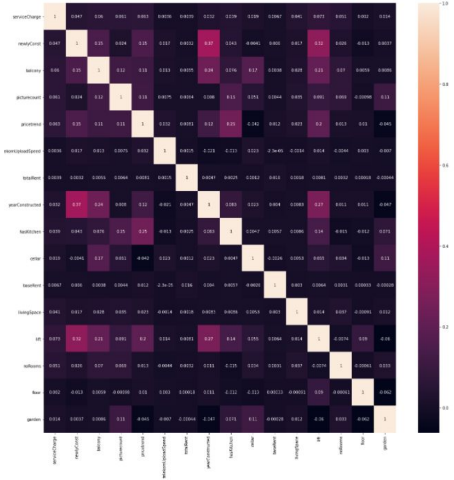
- Python and Jupyter
- Tableau

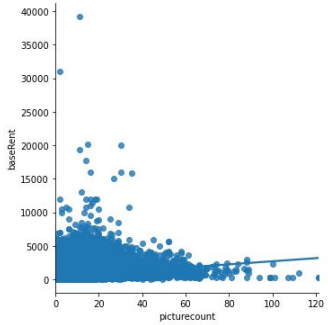
Procedures:

- Data cleaning and wrangling
- Exploratory Analysis
- Machine Learning techniques: Regression Analysis and Clustering
- Geospatial Analysis
- Analyzing Time Series Data

Visualization

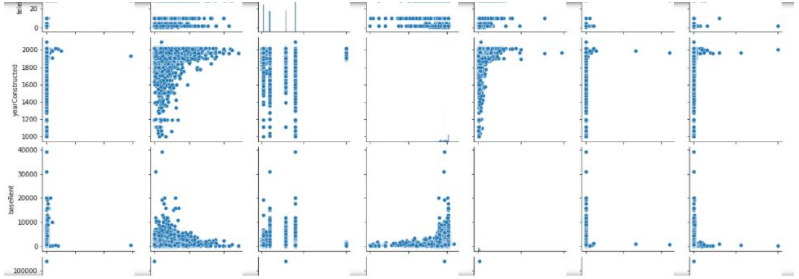
BaseRent - What factors have the most impact on the base rent of a flat,
For Example - noRooms: 0,16 The coefficient indicates a weak positive correlation. This could be interpreted as “the more rooms a flat has, the higher the base rent”, and vice versa - “the less rooms in a flat, the lower the price.”





The trend line is a bit steep, so there is a very small upward trend. There are many points not close to the line, and in addition we see one outlier for picturecount (over 100) and that the most flat offers have less than 20 pictures.

In most plots, you see the points in a vertical or horizontal line. I want to investigate the variables yearconstructed, picturecount and baseRent further because they show an upward or downward trend in the plot.

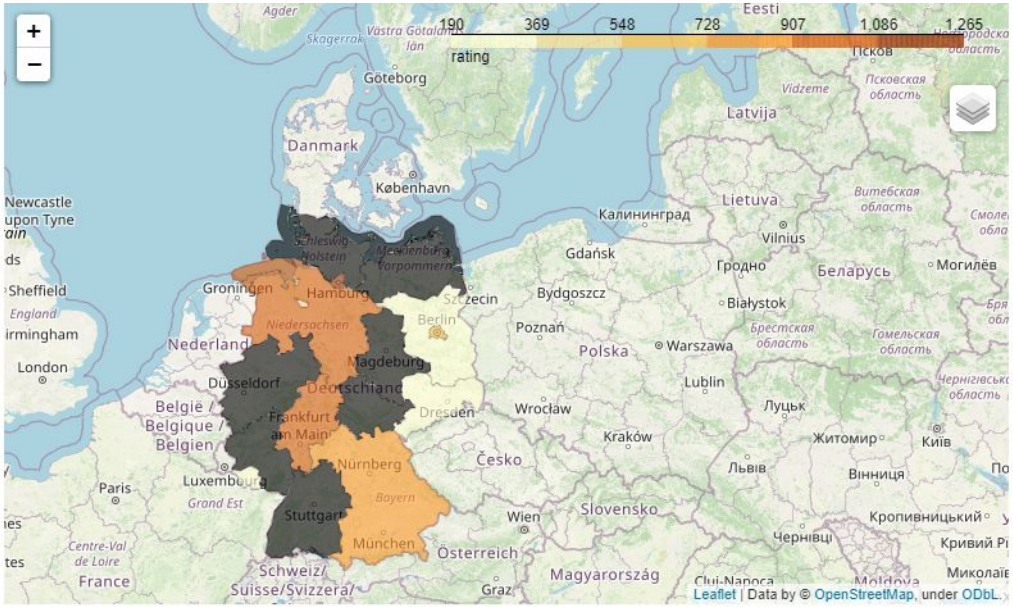


Visualization

Where can people live most cheaply/ expensive?

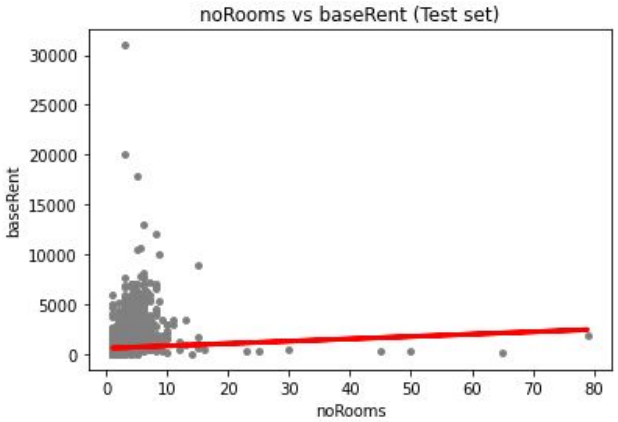
For the analysis, I used .geojson file from XY for Germany.

The cheapest flats are in Saxony and Brandenburg. The most expensive flats are in Baden-Württemberg, Rhineland-Palatinate, Thuringia, Saxony-Anhalt, Mecklenburg-Western Pomerania, North Rhine-Westphalia and Schleswig-Holstein.

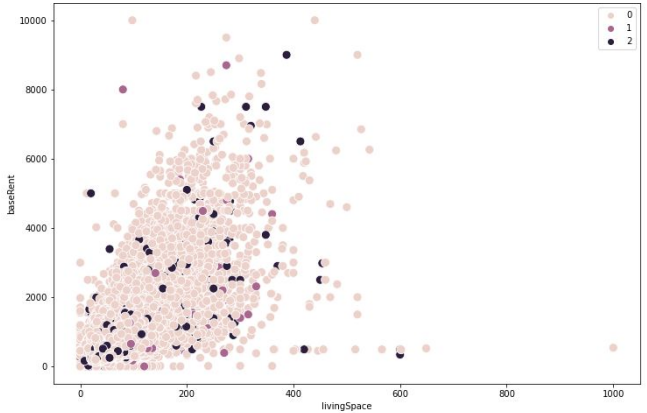


Visualization

We can see that the linear model represented by the red regression line does not cover part of the data points. In the area where the number of rooms is small, there are many data points that indicate a higher price, which contradicts my hypothesis.

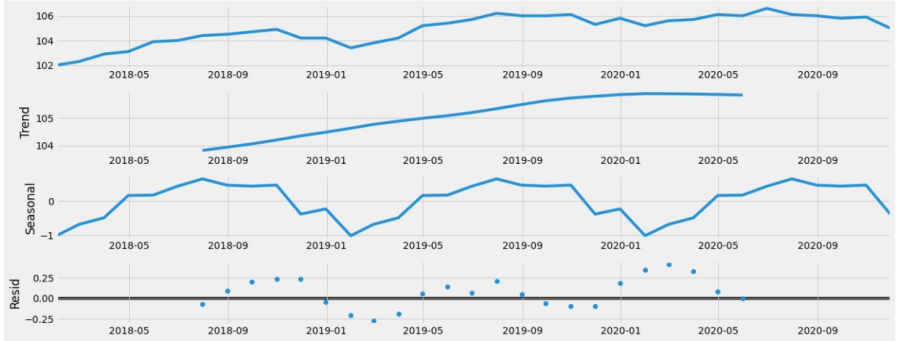


The Cluster analysis shows that the cluster with the most points is the 0/pink cluster. This main cluster is mainly located between 0 and 400 living space and 0 and 7000 baseRent. The other two clusters, 1/purple and 2/dark purple, can only be seen very sporadically in the graph.



Visualization

Here you can see the plot of the separate components for the time series analyze. The first diagram, the level, shows the data itself (including all its components).
The second graph shows the underlying rising trend. The trend increases gradually and linearly after 2018-09. It seems to be a dominant trend.
With seasonality, one sees the seasonal fluctuations represented by the three bulges/curves that repeat similarly at regular intervals. Since the curve is not flat, there is seasonality. You can see that there is not much unexplained noise.




```
Dickey-Fuller Stationarity test:
Test Statistic                  -2.305469
p-value                          0.170210
Number of Lags Used              2.000000
Number of Observations Used     32.000000
Critical Value (1%)             -3.653520
Critical Value (5%)             -2.957219
Critical Value (10%)            -2.617588
dtype: float64
```

<u>Dickey-Fuller Stationarity Test:</u>

The Critical Value 5% has a value of -2.957219. This is a bit smaller than the test statistic, which is -2.305469, which means that I can't reject the null hypothesis. So the data is non-stationary.

Conclusion & Recommendation

Recommendations

The cluster analysis has shown that the data produces at least one category "more housing through more rooms and/or through more space". This category is also related to more costs (higher service charge and higher baseRent).

Unfortunately, other categories cannot be identified through this analysis, as the values in the other clusters are very close to each other.

Next Steps

- Gather more data from other platforms
- Analyze the impact of other/additional variables on the base rent in Germany

Immoscout24 - Apartment rental offers in Germany

- GitHub repository
- Final Report

Find me online

GitHub

Tableau

LinkedIn

Xing