Vanessa Núñez Peñas

Data Analyst
Project Portfolio

Contact:

About Me

I'm a Data Analyst with experience in machine learning and academic research. I have a strong analytical mind and a passion for storytelling. As an empathetic communicator, I translate complex information into creative and insightful presentations to engage a diverse audience effectively. I have worked for over six years developing a voice assistant that helps millions of customers daily. One of my proudest achievements was to solve biases and constraints in our projects with constructive feedback. Now, I am eager to use my expertise to deep dive analyses to provide clear observations so that people leaders can make more informed decisions.


Projects

- **Airbnb Berlin:** market research for the most well-known company for short-term housing rentals.
- **Instacart:** marketing strategy for an online grocery store.
- **Preparing Influenza Season:** staffing distribution for the flu season in the U.S.
- **Rockbuster Stealth:** answering business questions for an online video rental company.
- **GameCo:** analyzing global video game sales.

Airbnb Berlin

Introduction

Overview

Airbnb is the most well-known company for short-term housing rentals worldwide. Berlin is the third most popular destination in Europe (after London and Paris). In 2019, around 14 million guests visited the capital of Germany, and the number of overnight stays rose to 34.1 million. Exploring the characteristics and the ratings of the listings in the city may help potential customers and hosts determine the best price range.

Process

After cleaning and understanding better the data sourced from publicly available information from Inside Airbnb and scraped by Makeover Monday, I performed all the exploratory and further analysis in Python to derive insights for Airbnb users and hosts in a Tableau Storyboard.

Goal

Developed market research, utilizing supervised and unsupervised machine learning algorithms and geospatial analysis with Python and Tableau.

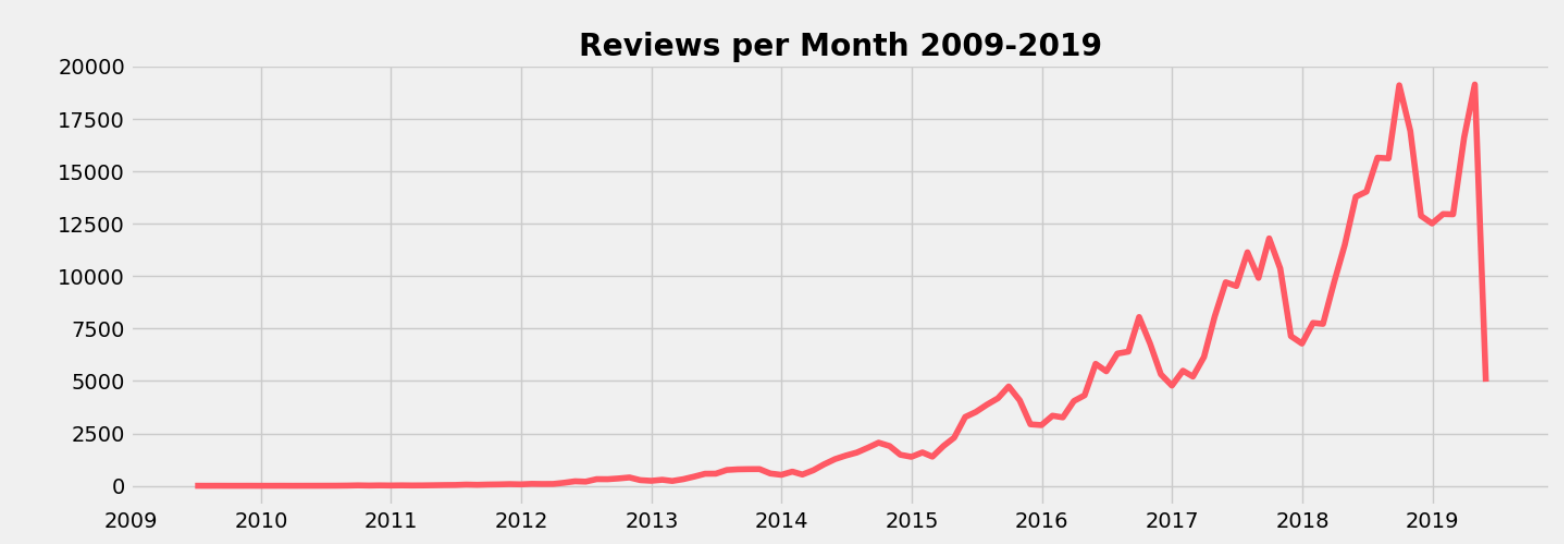
Skills & Tools

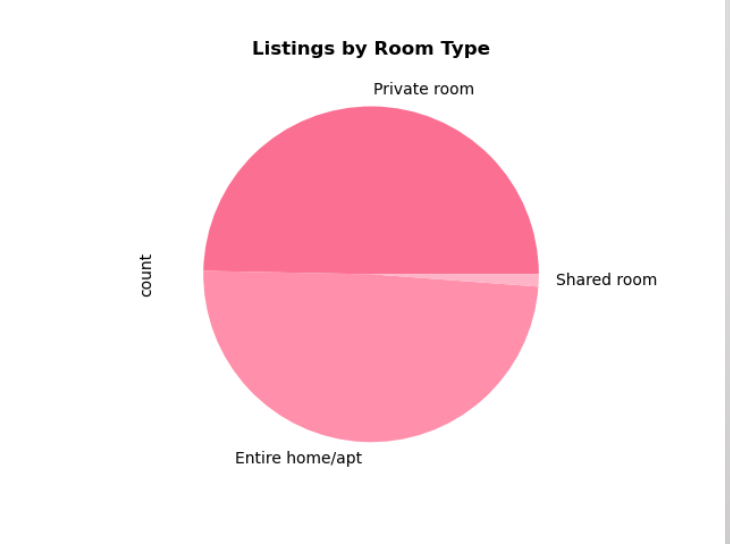
- Python
- Tableau
- Geospatial analysis
- Regressions
- Clustering analysis
- Time series análisis
- Decomposition
- Stationarity

Analysis

The characteristics of this project required a deep understanding of Python and libraries such as Bokeh, Plotly, Dash, Folium, Geopandas, Geoplot, and scikit-learn.

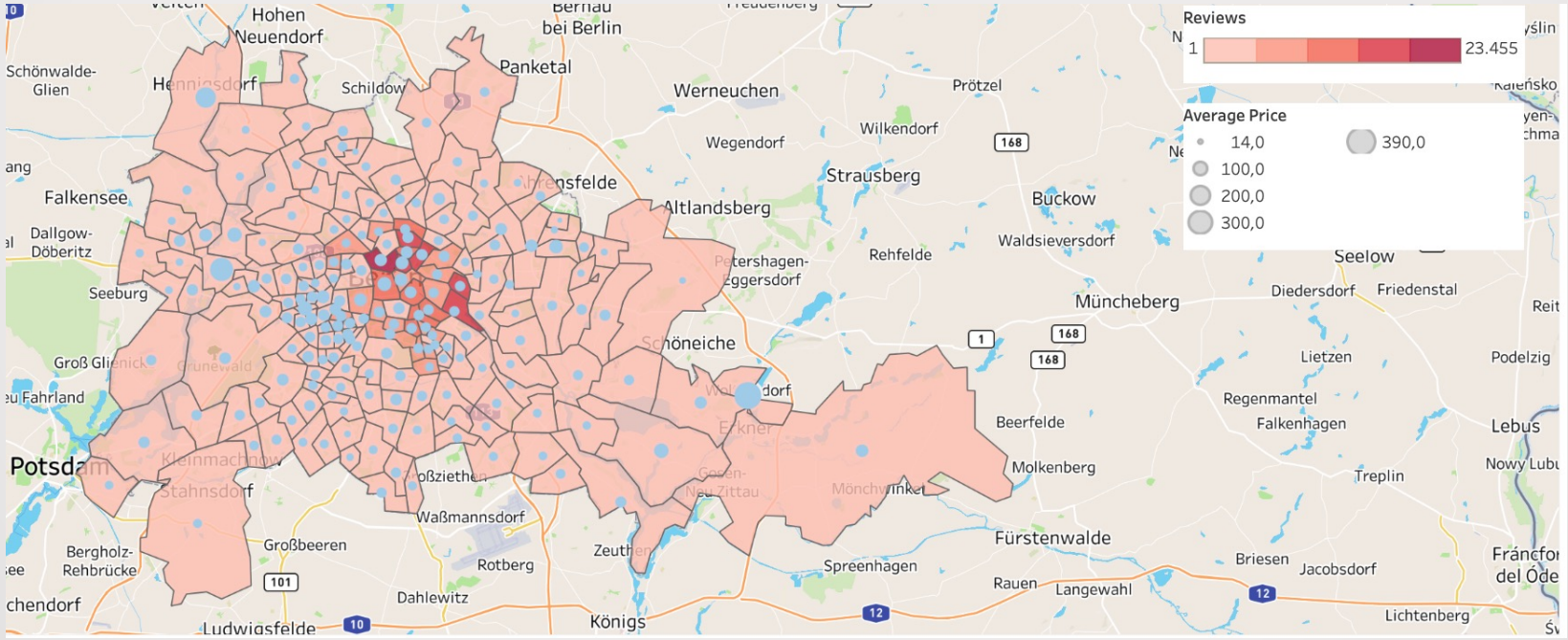
The stage for cleaning and understanding the data was especially important to prepare subsets to use in the next analysis stages. The original data set had 456,961 entries and 47 columns, and it was focused mainly on the reviews. For the purpose of this case study, I grouped the data by listings, reducing the entries to 18,833 entries. The scripts and visualizations created with Python are available for reference in a GitHub repository.





Analysis

The visualizations with Tableau helped me to get a more comprehensive picture, especially with regard to the geospatial analysis.



Results and Restrospective

Based on the analysis performed in Python, I created a Tableau Storyboard to present the more relevant findings. I couldn't find a clear relationship between price and the other available variables in the data, but I created an interactive map to facilitate prospective Airbnb users and host a complete picture of the market in Berlin. Filtering by property type, features, and neighborhood, they can check the average price of the accommodations.

The development of this project was challenging and frustrating at times, but it is also a good example of the importance of the cleaning and the understanding of the data as a starting point of every analysis. I needed to start over twice due to a wrong understanding of the data set and some mistakes while wrangling the data. This, along with the fact that most of the analysis did not give the expected results, have helped me to realize that unexpected results are still results.

Results and Restrospective

Whether a prospective customer looking for a nice place to stay in Berlin or a potential new host seeking guidance on prices, the following should be considered:

- Localization does not influence the price as much as other features, such as the property type.
- Listings close to the city center are the favorite in terms of reviews, something that can be translated into more bookings.

Limitations of the case study:

- We do not know if the prices change over time.
- The data contained a limited number of variables about the characteristics of the listings and the hosts.

Next Steps:

- Analyze further the differences in price or features between listings of hosts of only one listing and hosts of multiple listings.

Instacart

Introduction

Overview

Instacart is a well-known online grocery store in the U.S. and Canada that operates through an app. Instacart already has very good sales, but its marketing and sales teams want to uncover more information about their sales patterns.

Process

I performed an initial data and exploratory analysis of some of their open-source data to derive insights and suggested strategies for better segmentation to ensure Instacart targets the right customer profiles with the appropriate products. The customer data set was created for the purpose of this Project, and consequently, the Personal Identifiable Information (PII) on it is not real.

Goal

E-commerce Buying Patterns: identify the correlation between customer buying behavior and e-commerce marketing strategy using NumPy, pandas, Jupyter, matplotlib, and Excel.

Skills & Tools

- Python
- Data wrangling
- Data merging
- Deriving variables
- Grouping data
- Aggregating data
- Reporting in Excel Population flows

Analysis

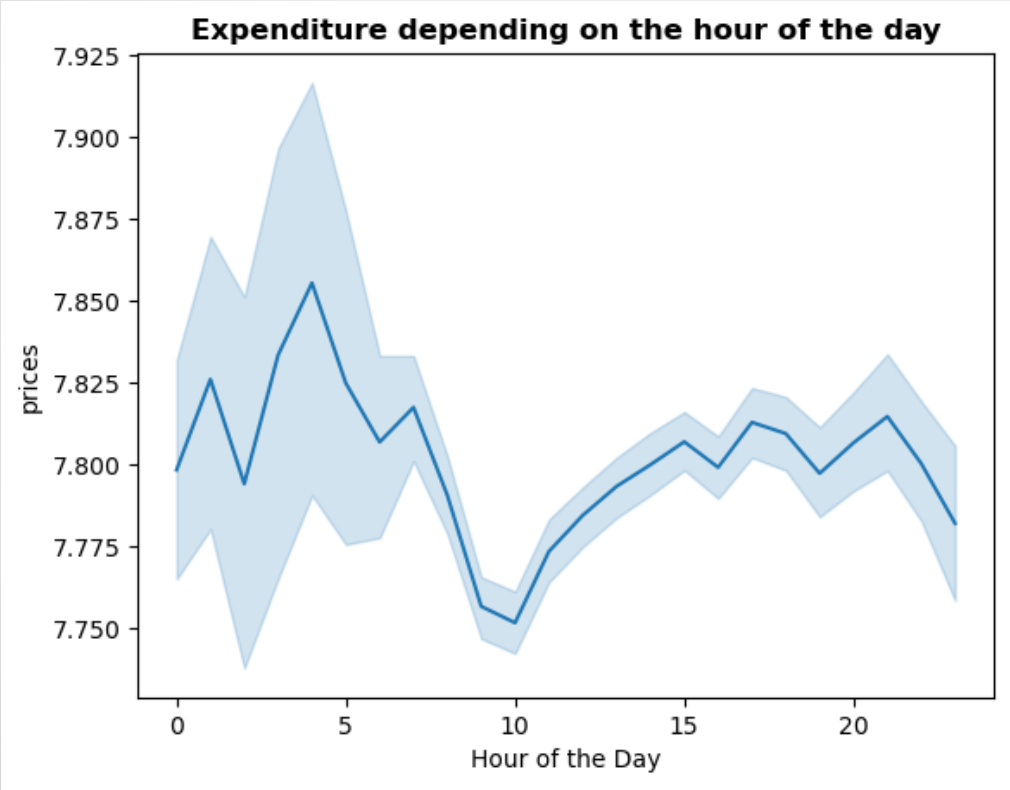
For this project, I conducted all the stages of my analysis with Python in Jupyter Notebook, using libraries such as Pandas, Numpy, Matplotlib, and Seaborn:

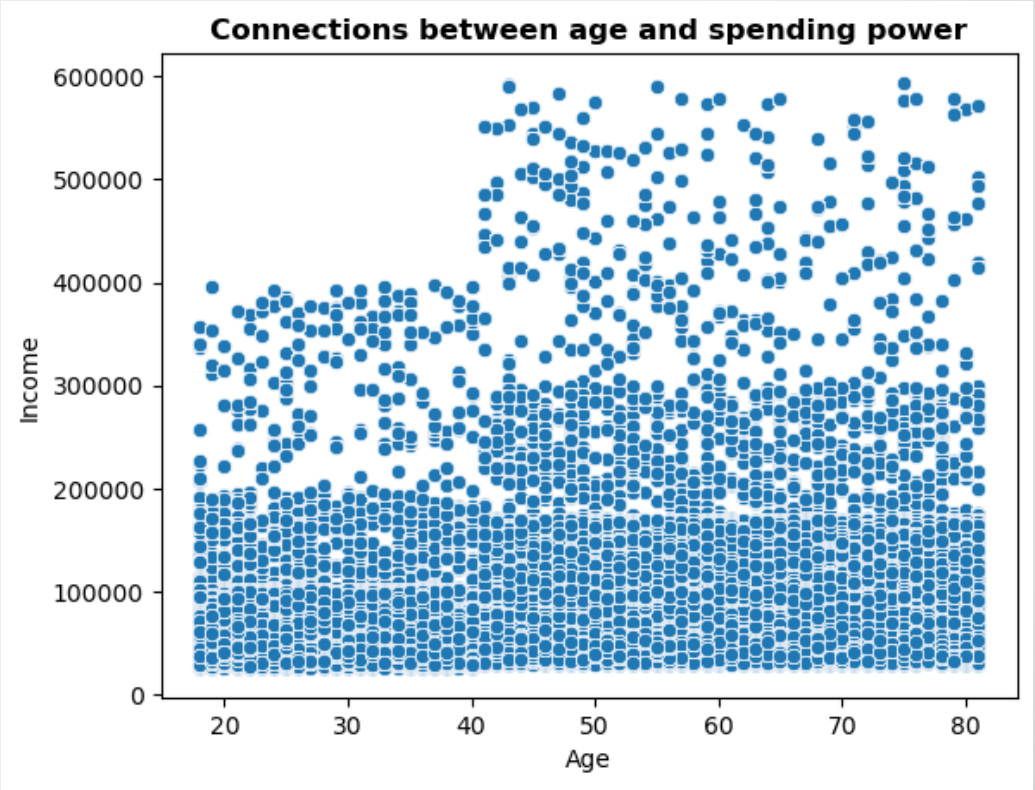
- Data wrangling and subsetting;
- Data consistency checks;
- Combining and exporting data;
- Deriving new variables;
- Creating summary columns of descriptive statistics;
- Data visualizations.

All the scripts and visualizations created with Python are available for reference in a GitHub repository.

Analysis

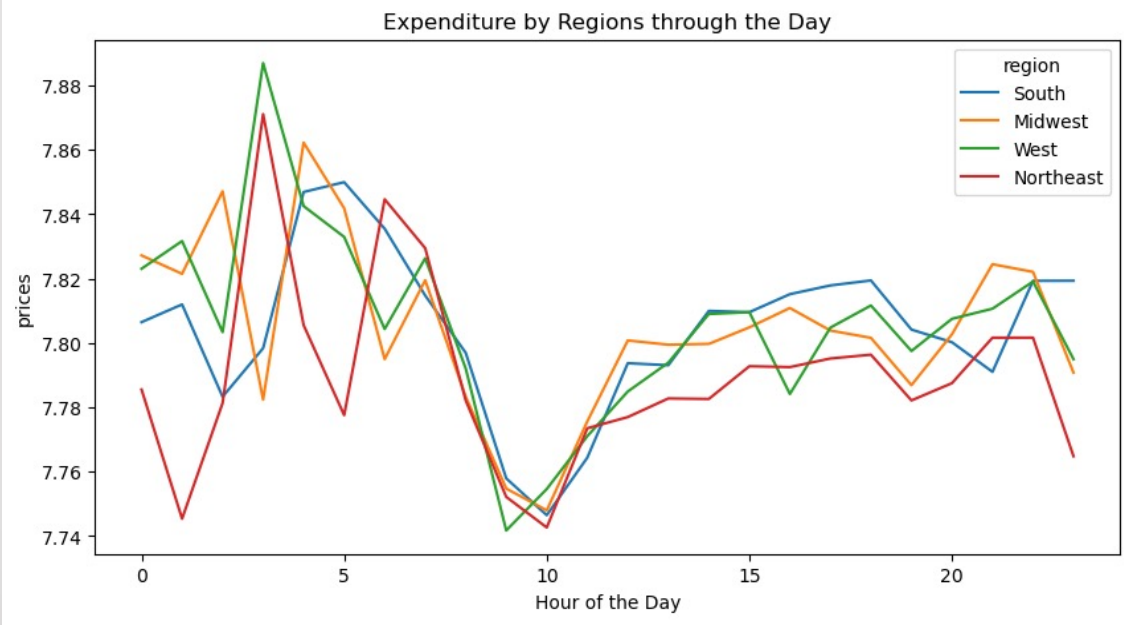
One of the biggest challenges was the data set's size: more than 30 million rows. For some analysis and visualizations, I needed to sample the data to get an accurate representation of the entire population.





Analysis

As the final deliverable for the project, I prepared a report in an Excel document describing the analysis methodology, the results, and the recommendations for Instacart stakeholders about when to schedule more ads to promote what products to whom, considering different customer profiles.



Recommendations

Customer Profile

- The most common Instacart Customer Profile is senior over 65 years, with dependants, mid-income between 80,000-150,000 $, and interested in 'Produce', 'Dairy Eggs' and 'Snacks'.
- The ordering habits are very similar regardless of the customer profile.

Brand Loyalty

- There are no differences in ordering habits based on a customer's loyalty status.
- Reward customers to encourage consumption with special offers or a system of cumulative points redeemable, taking into consideration their loyalty status.

Advertisements

- The time to place ads to attract orders is Tuesday and Wednesday between 19:00 and 7:00h.
- The ads to promote the more expensive products should be placed between 1:00 and 7:00.

Products

- Marketing and sales should focus the ads on products ranged between 5-15 $.
- The most ordered products are from 'Produce', 'Dairy Eggs' and 'Snacks', followed by 'Beverages', 'Frozen' and 'Pantry'.

Retrospective

When I started analyzing the data, I didn't expect to find such a well-distributed loyalty status among the different age groups or no significant differences in ordering habits. The age range of the most common customer profile was also a surprise, especially considering the online nature of the grocery store. This is an excellent example of how easy it is to bias something if we lower our guard. I, personally, have taken it as a great lesson, especially because I thought I was already well aware of data bias due to my experience in machine learning in the last few years and the projects I did in this regard.

Moving forward, I would like to do new projects focused on customer segmentation and profiling. Personalizing experiences according to perceptions, interests, attitudes, or behavior is a compelling strategy where data analytics have much to offer. But at the same time, it is a topic that may present important challenges from an ethical standpoint. Data privacy and security must always encompass responsible data analysis, and this is a topic where should be hand by hand creating a positive impact on both business and society as a whole.

Influenza Season

Introduction

Overview

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

Process

Work through the planning and preparation phases, sourcing the right data in CDC and US Census Bureau, profiling and checking data integrity issues, integrating data into one cohesive data set, conducting statistical analyses, and drawing insights using data visualization techniques in Tableau Public.

Goal

Strategically assigned key medical staff ahead of the flu season using Excel, VLOOKUP, and Tableau for data visualization.

Skills & Tools

- Tableau
- Excel
- Translating business requirements
- Data cleaning
- Data integration
- Data transformation
- Statistical hypothesis testing
- Visual analysis
- Forecasting

Analysis

My starting point was to design the data research project and to formulate the research hypothesis.

Then, I created a data profile for each data set and implemented data quality measures to ensure optimal data integration into one cohesive data set.

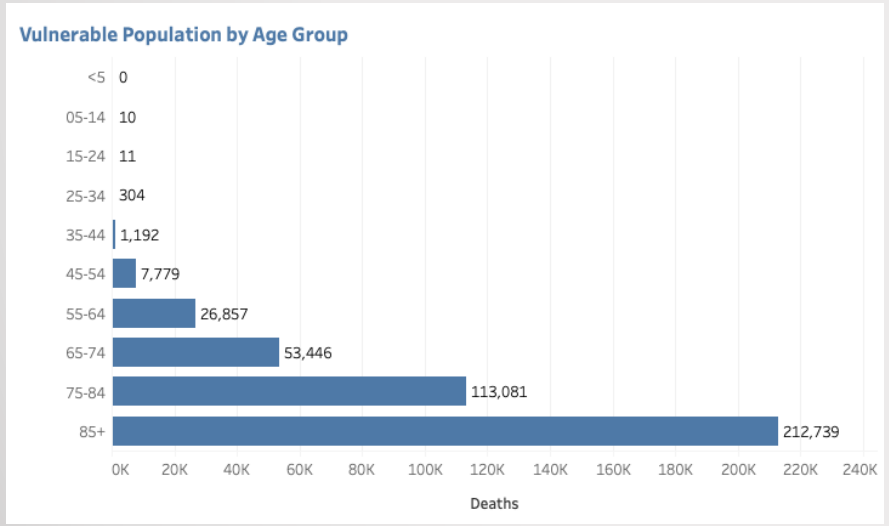
Descriptive analysis helped me understand the data variables and their relationships.

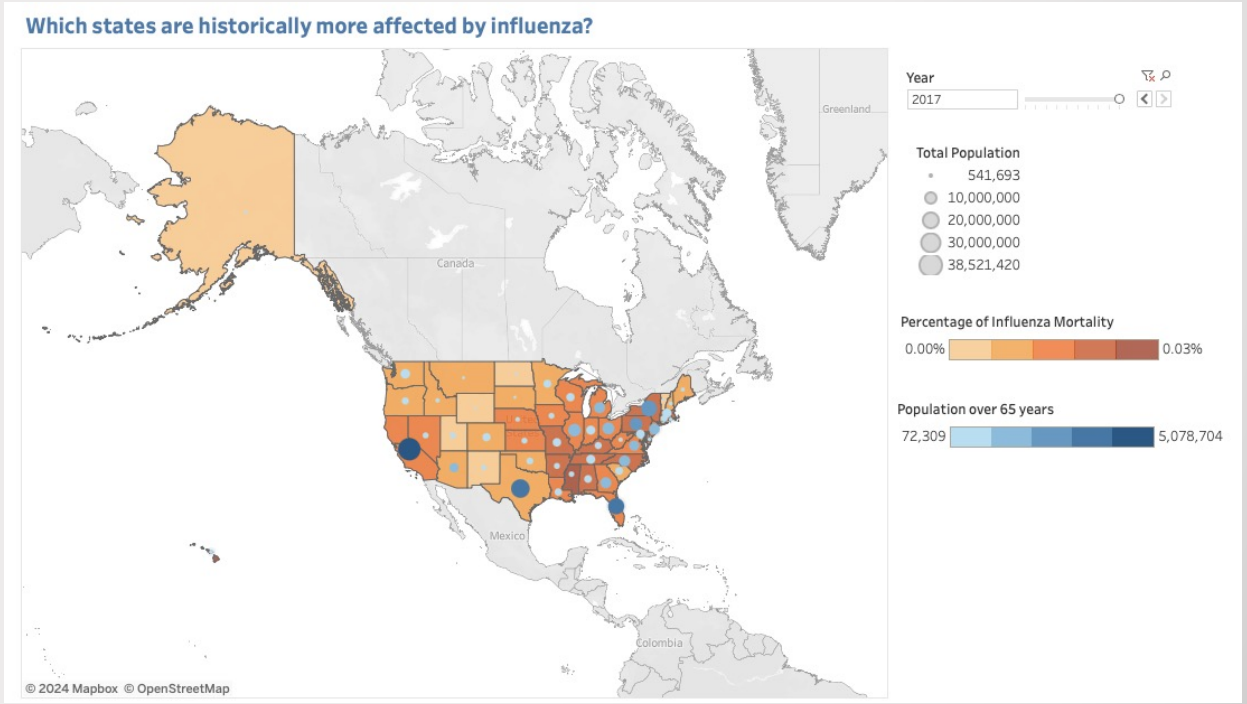
1. Influenza complications and mortality increase notably if flu patients are over 65 years old.
2. If patients are contagious by influenza in determined states with specific demographic conditions, the spread of the virus is higher than in others.
3. If people at high risk are vaccinated, influenza complications are less, and the medical staffing needs are not overcrowded.

All this process and the Excel workbooks with the corresponding analysis and findings are consolidated in an Interim Report available for consultation in a GitHub Repository.

Analysis

I did different visualizations with Tableau Public to understand the data and identify trends. Composition, temporal, statistical, and geospatial charts helped me to gain new perspectives and prepare the final presentation for the stakeholders. You can find the final storyboard in my Tableau Public profile.



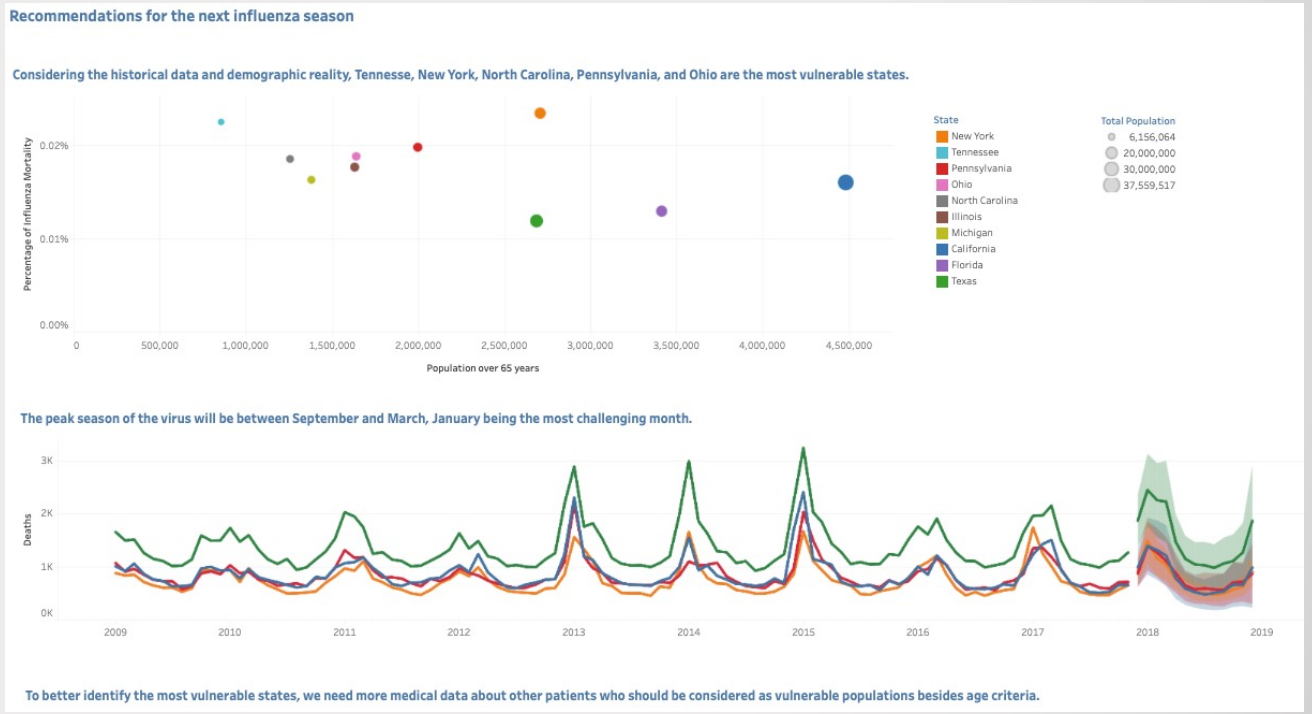


Recommendations

With the available data and its insights, I could identify where and when hospitals and clinics would need more medical staff:

- I recommend sending more staff and resources to Tennessee, New York, North Carolina, Pennsylvania, and Ohio as the most vulnerable states.

- Preparing for the end of autumn will be crucial, increasing the staff and resources for December, January, February, and March.



Retrospective

- These conclusions come from analyzing the available data, mainly from the U.S. census and influenza mortality. I see two essential data limitations:
  - There is insufficient medical data about other patients who should be considered vulnerable populations besides age criteria or vaccination rates among the total population.
  - It would be interesting to have more information on influenza deaths by county to provide better recommendations about where to send more staff and resources. Those would be more efficient, especially in the bigger states with more considerable demographic differences.

- The big challenge of this project was integrating the different available data sets into one, mainly because the values were not uniform. That's why the stage of preparing the data before combining everything into one cohesive data set was so important.

- To monitor the impact of the recommended staffing changes, getting feedback monthly from medical staff and patients would be interesting. In parallel form, having numbers regarding the occupancy rates during the season and the information of when, where, and how many overwhelming hospitals have suffered from understaffing weekly could help to monitor the recommendations and make changes accordingly for the next influenza season.

Rockbuster Stealth

Introduction

Overview

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

Process

Load all of Rockbuster's data into a relational database management system (RDBMS) to use SQL to analyze the data and answer any ad-hoc business questions other departments may have, from inventory to customer insights. Compile the analysis results into an easily digestible format, which will be presented to the Management Board.

Goal

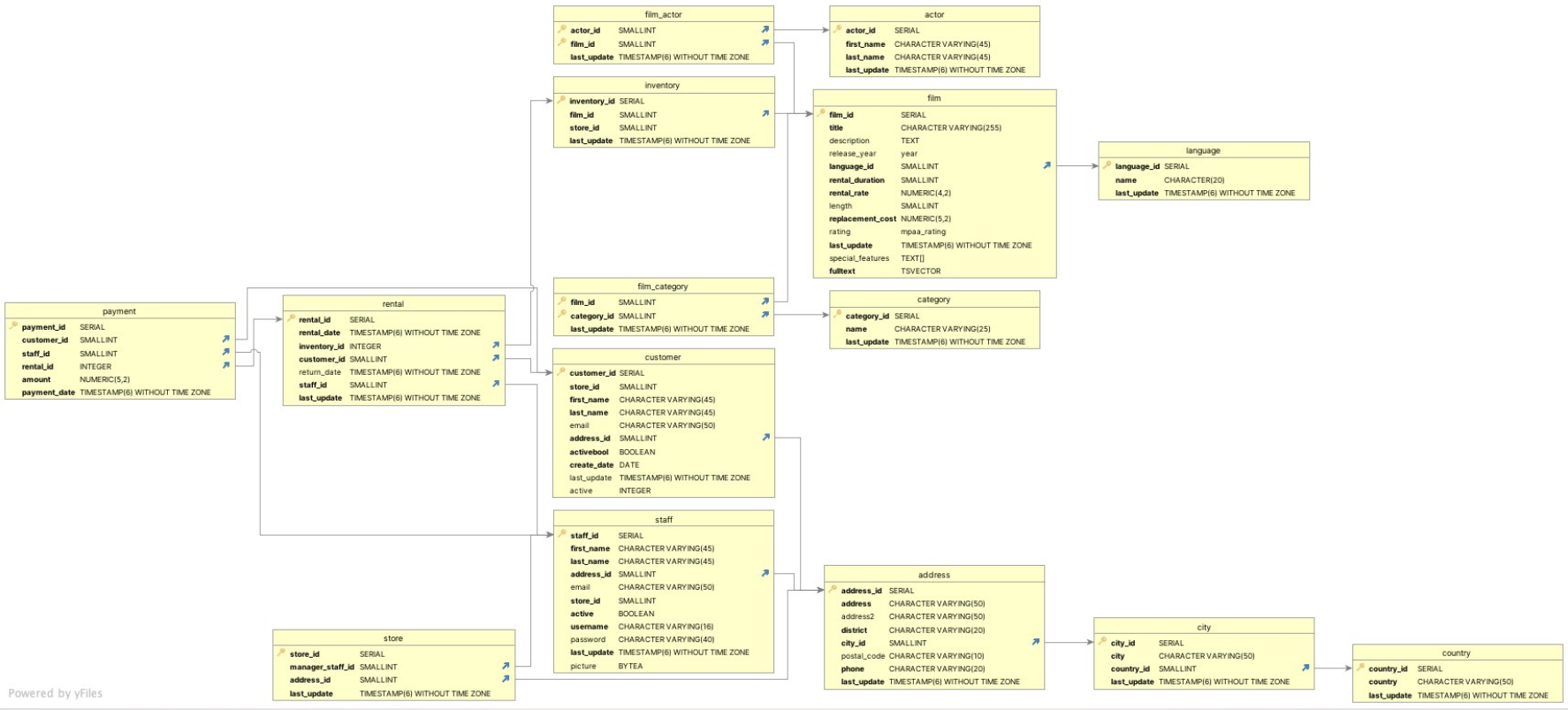
Market Research: conducted analysis to identify opportunities for a movie rental shop's transition to an online service using Tableau, Excel, and SQL queries.

Skills & Tools

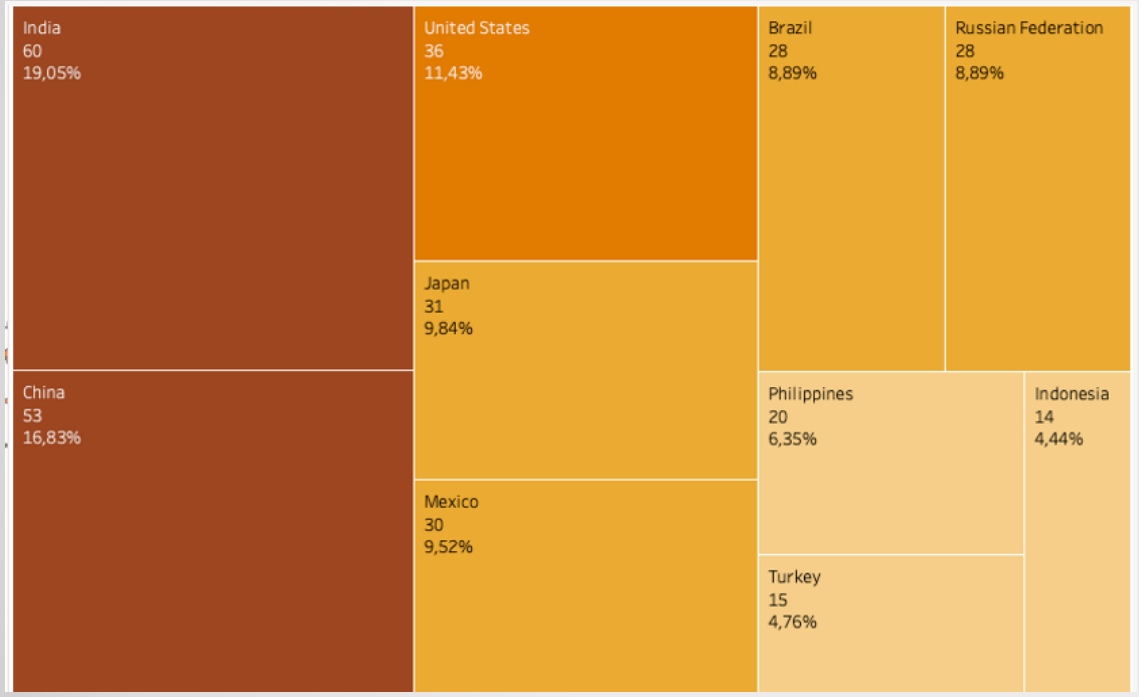
- Relational databases
- SQL
- Database querying
- Filtering
- Cleaning and summarizing
- Joining tables
- Subqueries
- Common table expressions

Analysis

My starting point was to extract an entity relationship diagram of the data and create a first draft of a data dictionary to understand the overall database structure better.



Analysis



Rockbuster's Customers Around the World

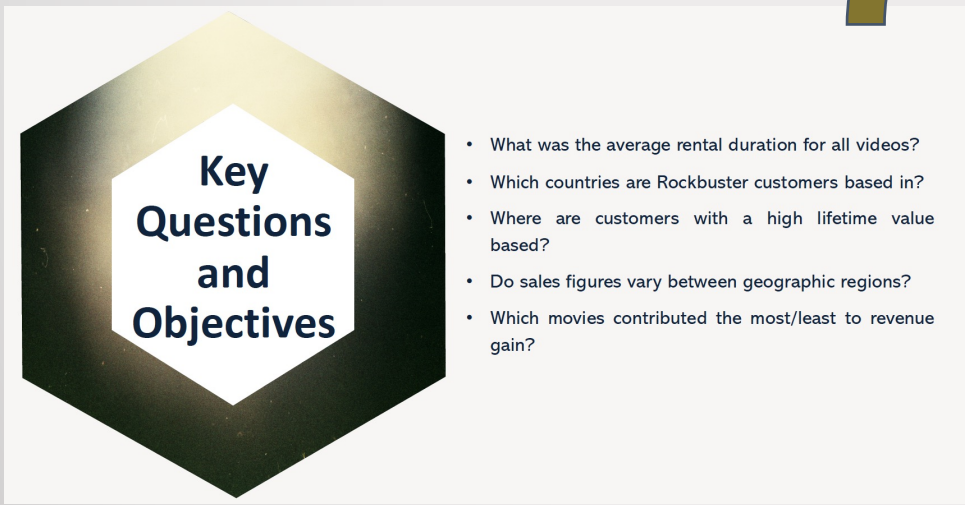
To find the answers to the business questions, I used SQL commands in PostgreSQL to organize and sort the data, clean the data, create a data profile of summary statistics, join tables, and perform subqueries and common table expressions.

To communicate my findings effectively, I transformed my SQL results into visualizations with Tableau Public.

You can find the SQL queries in my GitHub repository, and the visualizations to support my final presentation to the Management Board in my Tableau Public profile.

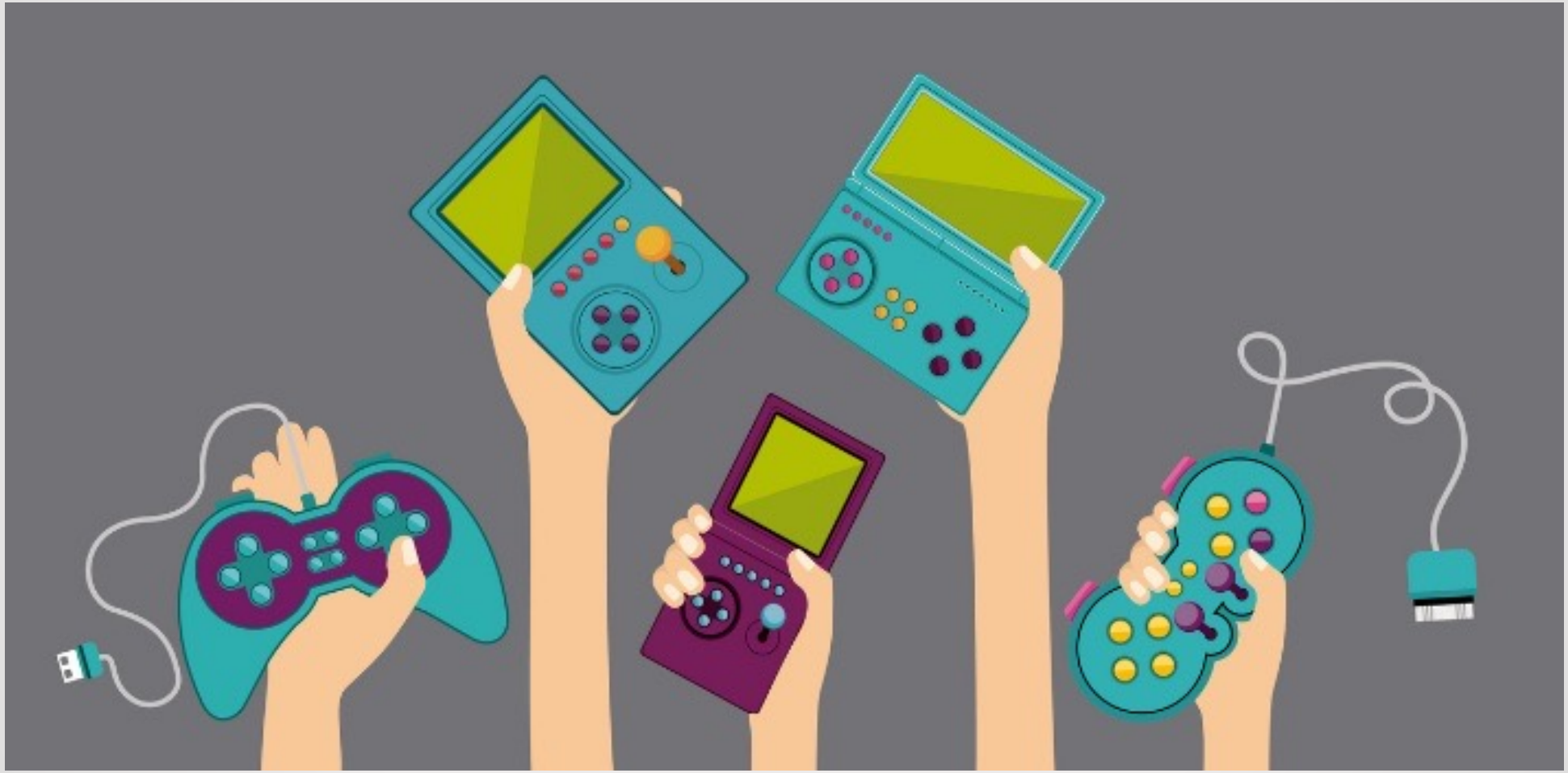
Recommendations

With the key questions and the objectives in mind, I collected the findings and the visualizations in an easy-to-consume report to present the results to Rockbuster management. The final presentation is available in the GitHub repository.



Rental	Market	Customers	Movies	Launch
The online rental video service should offer a duration of 5 days.	Focus on Asia and North America as main markets, especially India, China and the United States.	Loyalty program where customers get different benefits depending on their rank: Better rental rate for customers with higher lifetime value.	Sports, Sci-Fi and Animation as the strongest genres of the platform.	Trial Online Rental Video Service in India, China and the United States.

GameCo



Introduction

Overview

GameCo is a new video game company that wants to use data to inform the development of new games.

Process

End-to-end analysis using a data set covering historical video game sales (1980-2016).

Goal

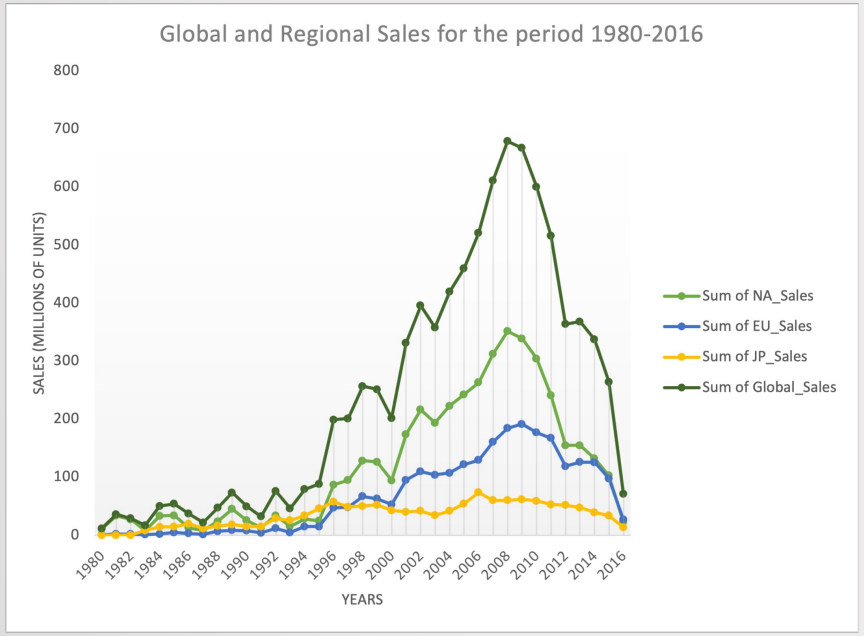
Foster a better understanding of how GameCo's new games might fare in the market.

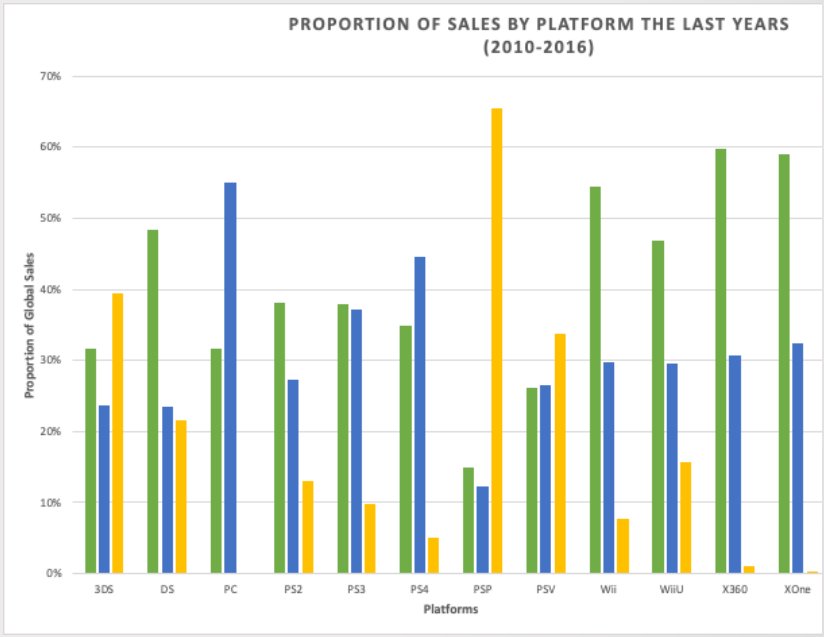
Skills & Tools

- Excel
- Grouping data
- Summarizing data
- Descriptive analysis
- Visualizing results in Excel
- Presenting results

Analysis

After cleaning and grouping the data, I summarized it and conducted a first descriptive analysis to gain insights into global and regional sales.

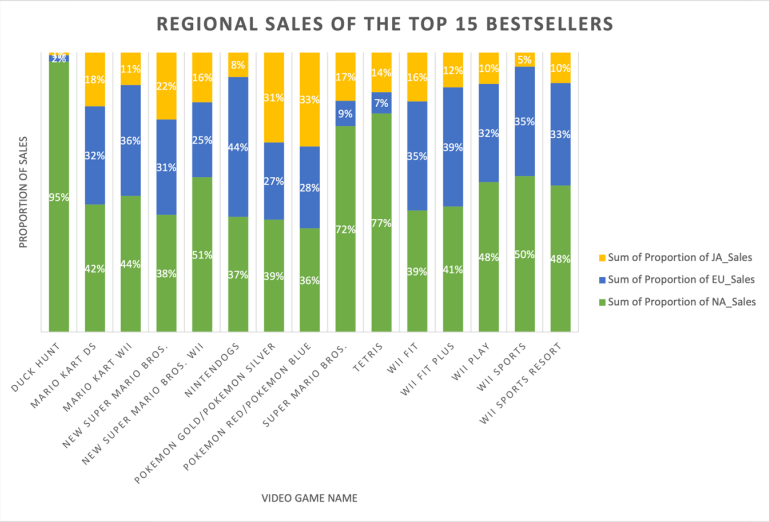


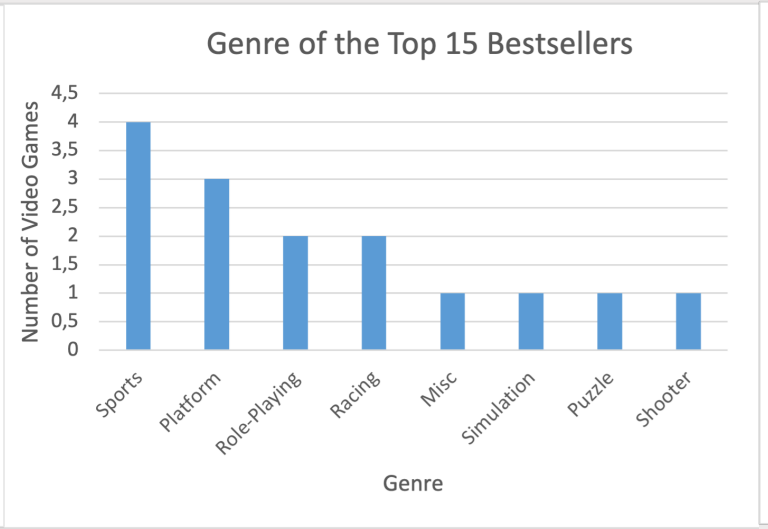


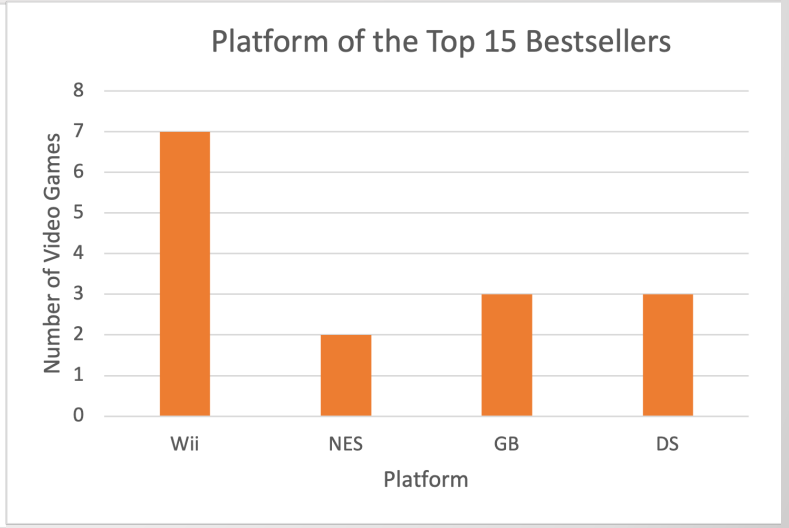
Analysis

Considering the industry's critical technical changes, I focused on the last two decades. However, I also decided to explore what have historically been the bestsellers of video games.

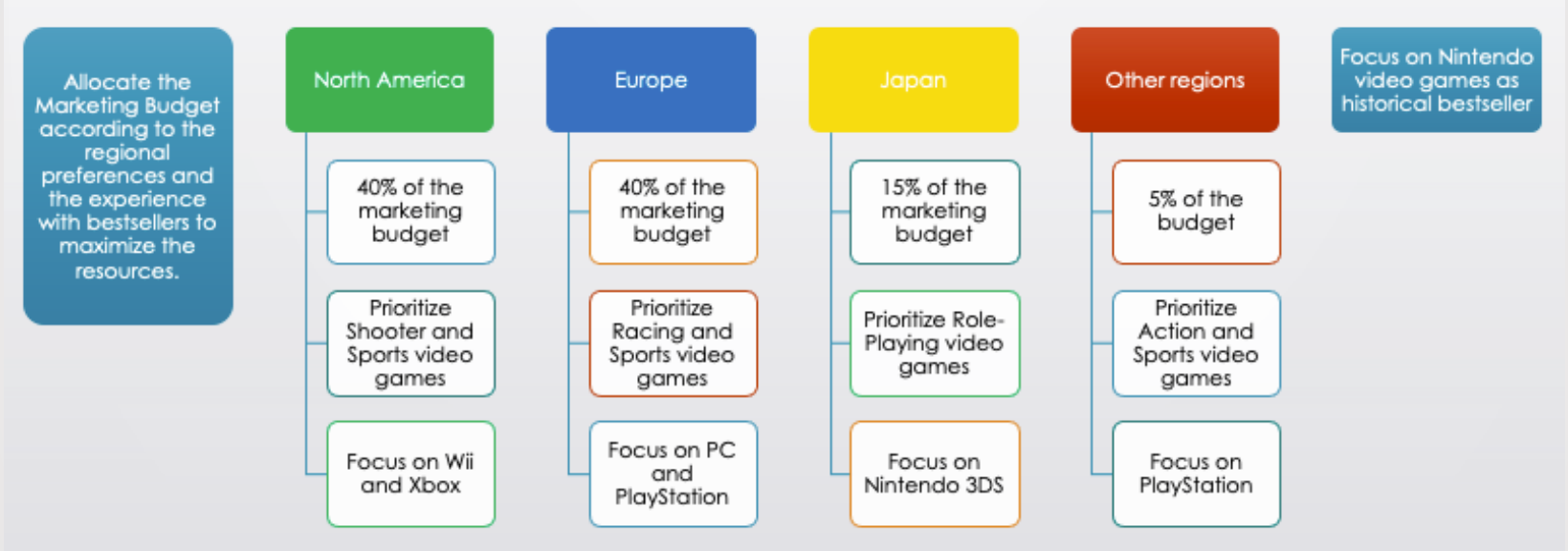
The Excel workbooks and the visualizations are available for consultation in a [GitHub Repository](#).







Recommendations



Conducting an Exploratory Data Analysis helped me to see how, contrary to the expectations, the sales of the three regions have not stayed the same over time, and the popularity of genres and platforms varies by region. All the insights and recommendations for GameCo's executive board to plan better the marketing budget for next year among the regions to maximize return on investment are collected in a Final Report.

Retrospective

I decided to run this analysis in Excel because the data set size was manageable, only 16,601 rows, and because it is a vital tool that a good analyst should be an expert in, too. The analysis and visualizations could be more impactful using Python or Tableau. Still, it is essential to have a solid understanding of this spreadsheet application and a working knowledge of its various formulas, functions, and data visualization capabilities.

This Project was not challenging from the technical perspective but from the personal one: I am not a person who is into videogames, so I needed to do some research to understand the data better, be able to run a better analysis and prepare an impactful final report with the best recommendations possible.

The fields in which an analyst can work are genuinely unlimited. Wherever there's information, there exists a chance to analyze that information. I, like everybody, certainly have more interest in some fields than others, but as a professional data analyst, I am prepared to work in different areas.

Thank you!

Do you have any questions or comments?

vanessa.nupe@gmail.com

Vanessa Núñez Peñas