Data Analytics Portfolio

Case studies

Julia Fortuny Wollny, October 2021

1. GameCo

Overview

The video game company GameCo wants to use data to inform the development of new games.

They have asked for a descriptive analysis of video game data.

They want to answer business questions such as:

- Are certain types of game more popular than others?
- How have sales figures varied between geographic regions over time?

Goal

- Foster a better understanding of how GameCo's new games might fare in the market
- Support marketing team to better allocate budget
- Help financial team keep tab on competitors
- Assist management in understanding swings in the market

Data used

Data set that covers historical sales of video games spanning different platforms, genres and publishing studios.

The data was drawn from the website VGChartz and can be found here.

Process & methodology

Preparation

- Examine data set
- Clean data
- Perform EDA
- Group & summarise data with pivot tables (Excel)
- Obtain first insights
- Form hypothesis
- Wrangle data, incl. deriving new calculated fields

Analysis

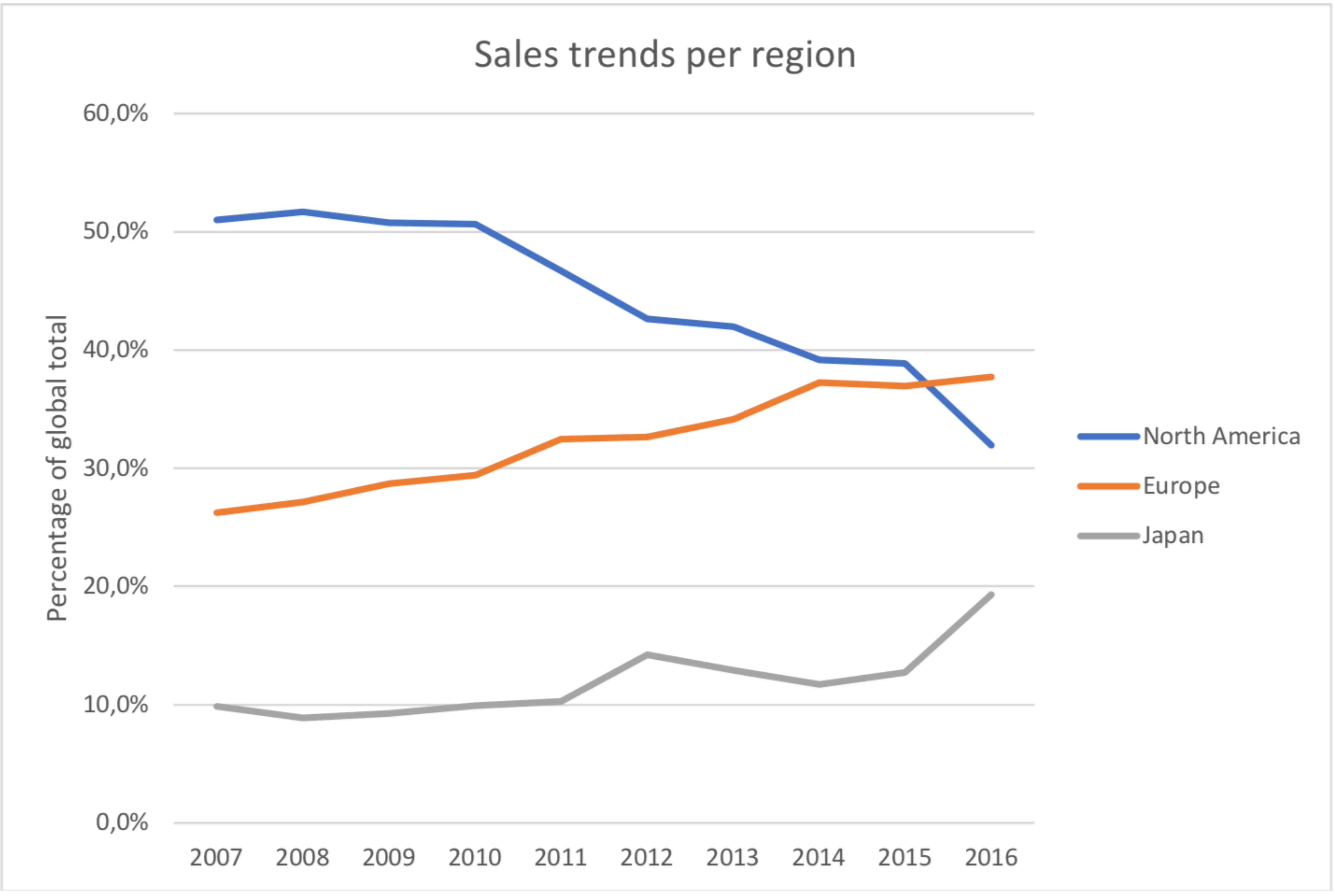
Create a descriptive analysis to answer business questions, including:

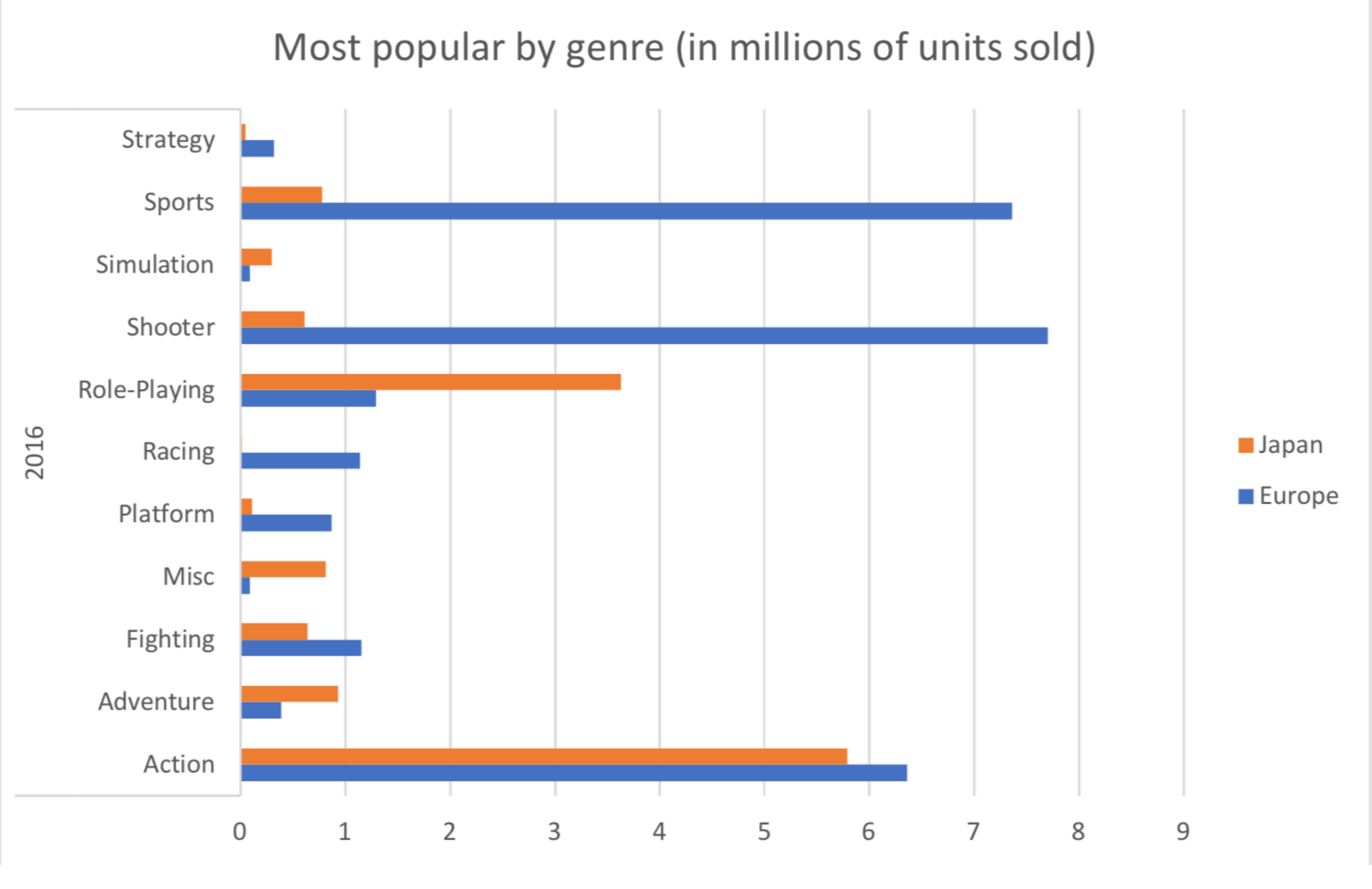
- Line graph to show sales trends per region
- Bar chart to highlight most popular genres by region
- Ranking to find best selling single game in 2016

Presentation

- Summary stats on variables
- Line graph
- Bar chart
- Ranking
- Answers to business questions

The full presentation can be found here.





This slide shows a decreasing sales trend in N. America and rising trends in Europe and Japan between 2007 and 2016.

Here we see that the top 3 popular genres in 2016 in Europe were shooter, sports and action, whereas in Japan they were action, role-playing and sports, in this order.

Skills & tools

Understanding & translating business requirements

Develop & visualise insights

Cleaning & transforming data in Excel

Descriptive analysis

Visualisations in Excel

Filter, group & summarise data in Excel

Pivot tables

Storytelling with data

2. Preparing for influenza season

Overview

The United States has an influenza season where more people than usual suffer from the flu.

Some, particularly vulnerable populations, end up in hospital.

The stakeholders (hospitals, medical frontline staff, patients, clinic & staffing agency administrators) want to proactively plan for influenza season across the country using historical data.

Goal

- Help plan for influenza season
- Examine trends in influenza
- Provide insights to support a staffing plan
- Prioritise states with large vulnerable populations
- Assess data limitations which might influence analysis results

Data used

1. Influenza deaths by geography, time, age and gender. Source: CDC. Download here.
2. Population data by geography. Source: US Census Bureau. Download here.
3. Survey of flu shots in children. Source: CDC. Download here.

Process & methodology

Preparation

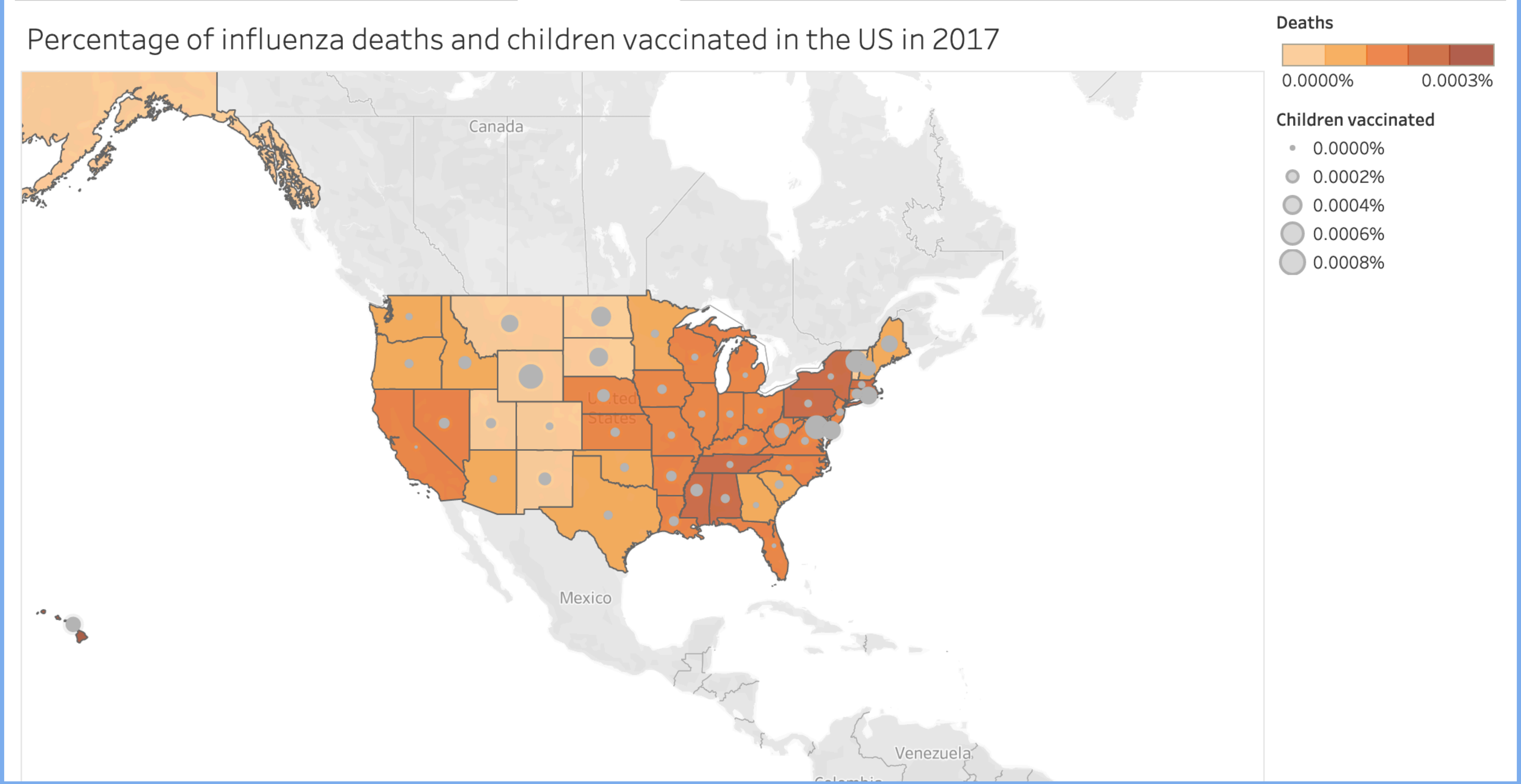
- Distil business requirements and requests into questions
- Design a data research project
- Source & curate data
- Data profiling & integrity checks
- Measure data quality
- Transform & integrate data

Analysis

- Conduct statistical analysis
- Formulate statistical hypothesis
- Test hypothesis & interpret results
- Translate results into visualisations

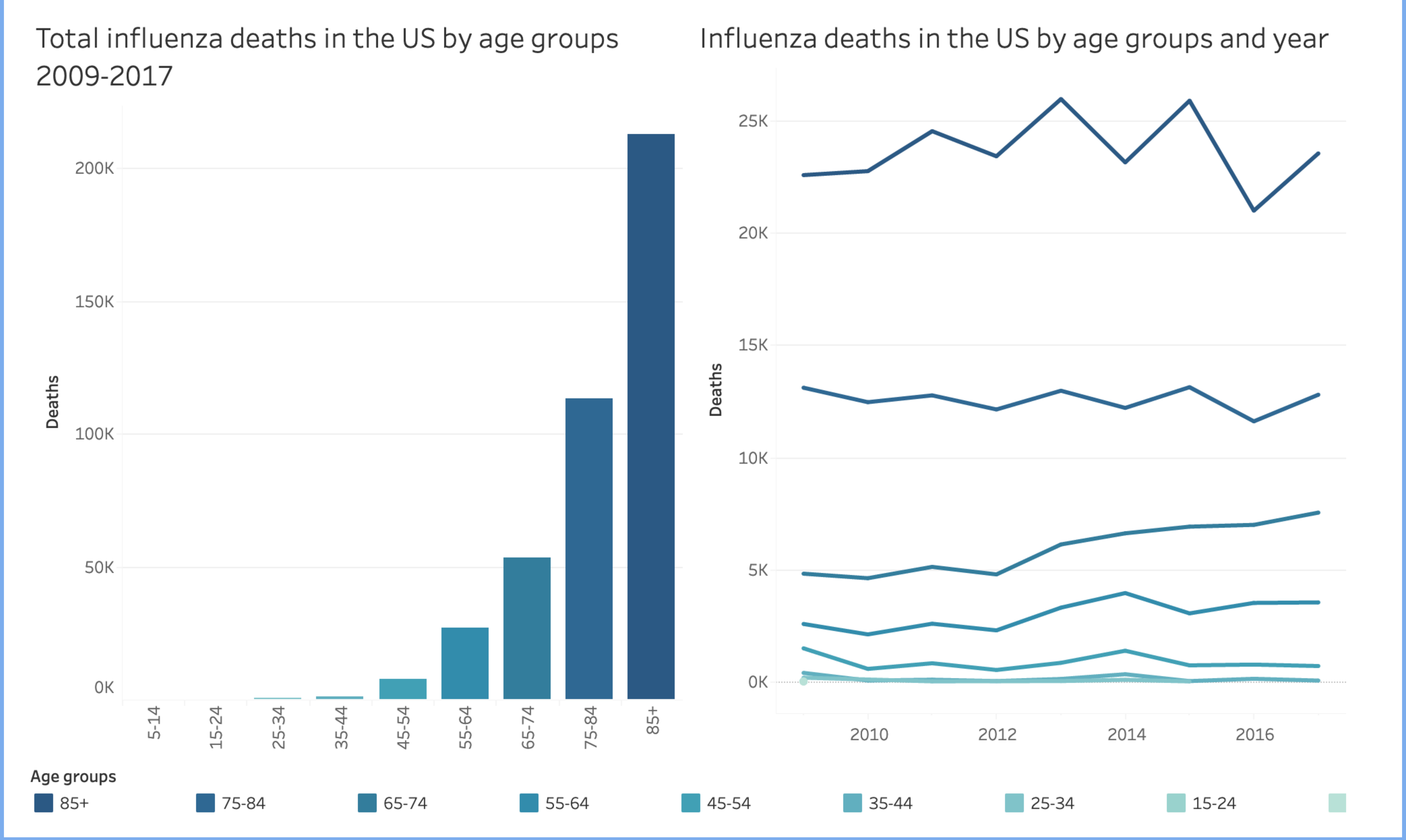
Presentation

- Compelling Tableau presentation including spatial & temporal visualisations, conclusions, recommendations & next steps. See it here.
- Video presentation considering the audience (stakeholders). Link here.



Percentage of influenza deaths and children vaccinated in the US in 2017

This choropleth map shows the percentage of flu deaths in 2017 by state, as well as the rate of vaccinated children.



These charts show the total influenza deaths in the US by age groups and by year.

Skills & Tools

Understanding & translating business requirements

Sourcing & curating data

Designing a data research project

Data profiling, integrity & quality checks

Data transformation & integration

Statistical analysis

Statistical hypothesis testing

Composition & comparison charts

Temporal visualisations & forecasting

Spatial analysis

Presenting findings to stakeholders

Storytelling

Tableau

Excel

3. Rockbuster Stealth Data Analysis Project

Overview

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.

Goal

- Find the top paying customers worldwide in order to target them for a marketing campaign
- Answering business questions such as:
  - Which countries are Rockbuster customers based in?
  - Do sales figures vary between regions?
- Compiling results into digestible format
- Present results to management board

Data used

Data set with information on Rockbuster's film inventory, customers, payments and more.

It can be downloaded here.

Process & methodology

Preparation

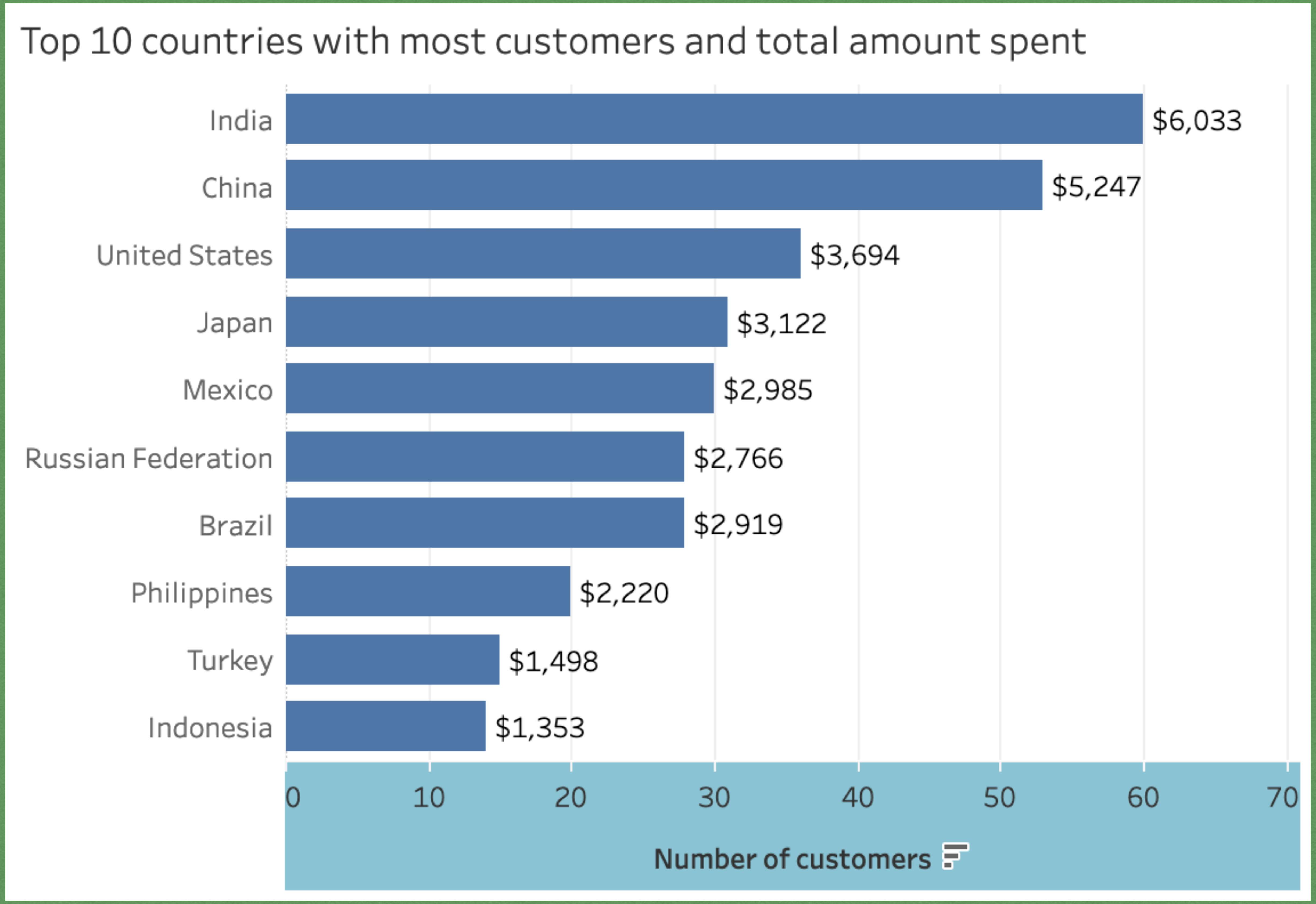
- Set up SQL database environment
- Extract entity relationship diagram (ERD). Find it here
- Create data profile & summary statistics
- Use SQL commands to clean data

Analysis

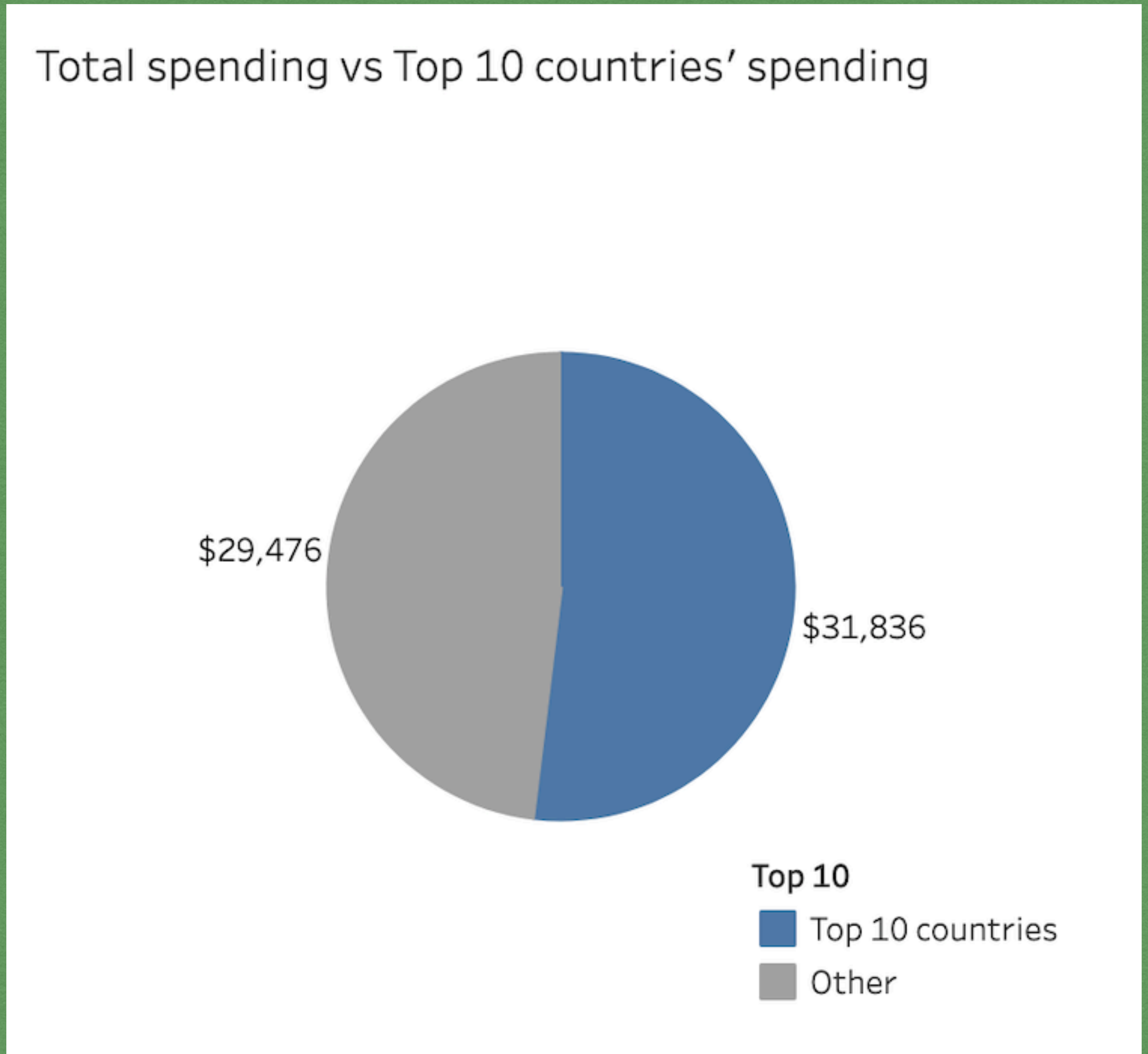
- Extract necessary data to answer business questions
- Order, group, sort & filter data in PostgreSQL
- Write subqueries, CTEs
- Perform table joins
- Answer business questions

Presentation

- Translate results into visualisations
- Formulate recommendations & next steps
- Create a presentation in Tableau. See it here
- Build a data dictionary. Download here



India, China and the United States are the countries with most customers and the highest spending.



This pie chart shows how the spending of the top 10 countries compares to the total spending

Skills & Tools

Write common SQL commands

Perform basic CRUD operations

Order, limit, group data

Filter data using WHERE and HAVING

Clean data SQL

Create a data profile & summary statistics

Perform joins

Write subqueries & common table expressions

Present results to technical colleagues in Excel

Create data dictionary

Produce a compelling presentation

PostgreSQL

PgAdmin

DbVisualizer

Excel

Tableau

Read my code on Github!

4. Instacart Grocery Basket Analysis

Overview

Instacart is an online grocery store that operates through an app.

They already have very good sales but they want to uncover more information about sales patterns.

Goal

- Help the marketing team better segment Instacart's customer base and improve sales
- Answer business questions such as:
  - Which are the busiest times of the day?
  - Are certain types of products more popular than others?
  - What are the ordering habits of different customer profiles?

Data used

Open source data provided by Instacart, including 30+ million rows of information such as products sold, price, time of the day and many more.

It can be found here ("The Instacart Online Grocery Shopping Dataset 2017" accessed from https://www.instacart.com/datasets/grovery-shopping-2017 on July 4th 2021).

Process & methodology

Preparation

- Wrangle & subset data with Python
- Clean & check data
- Clearly document each step in Jupyter Notebook maintaining coding etiquette

Analysis

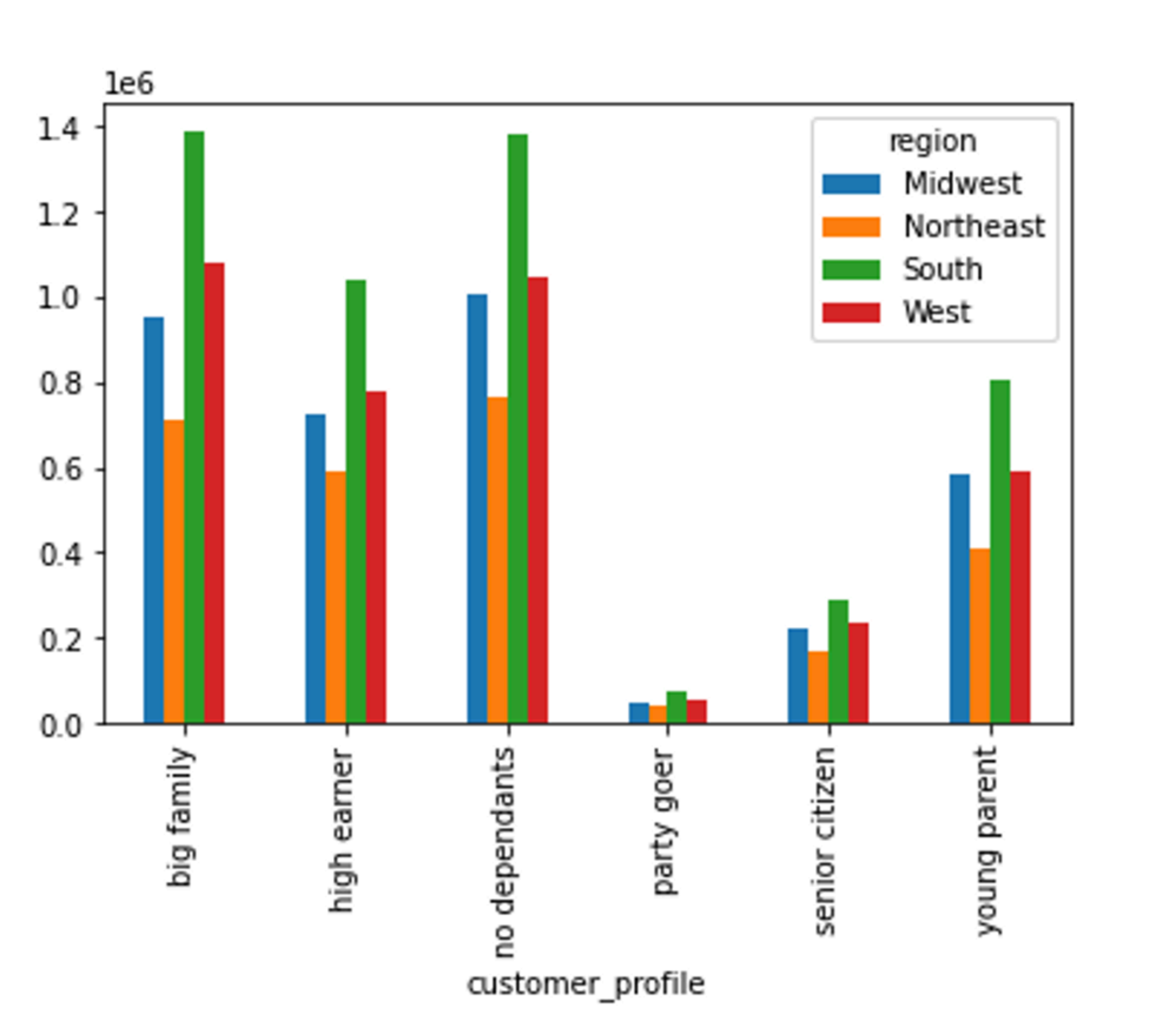
- Group, aggregate data
- Derive new variables
- Create flags
- Produce statistical visualisations to interpret results
- Population flow
- Answer business questions

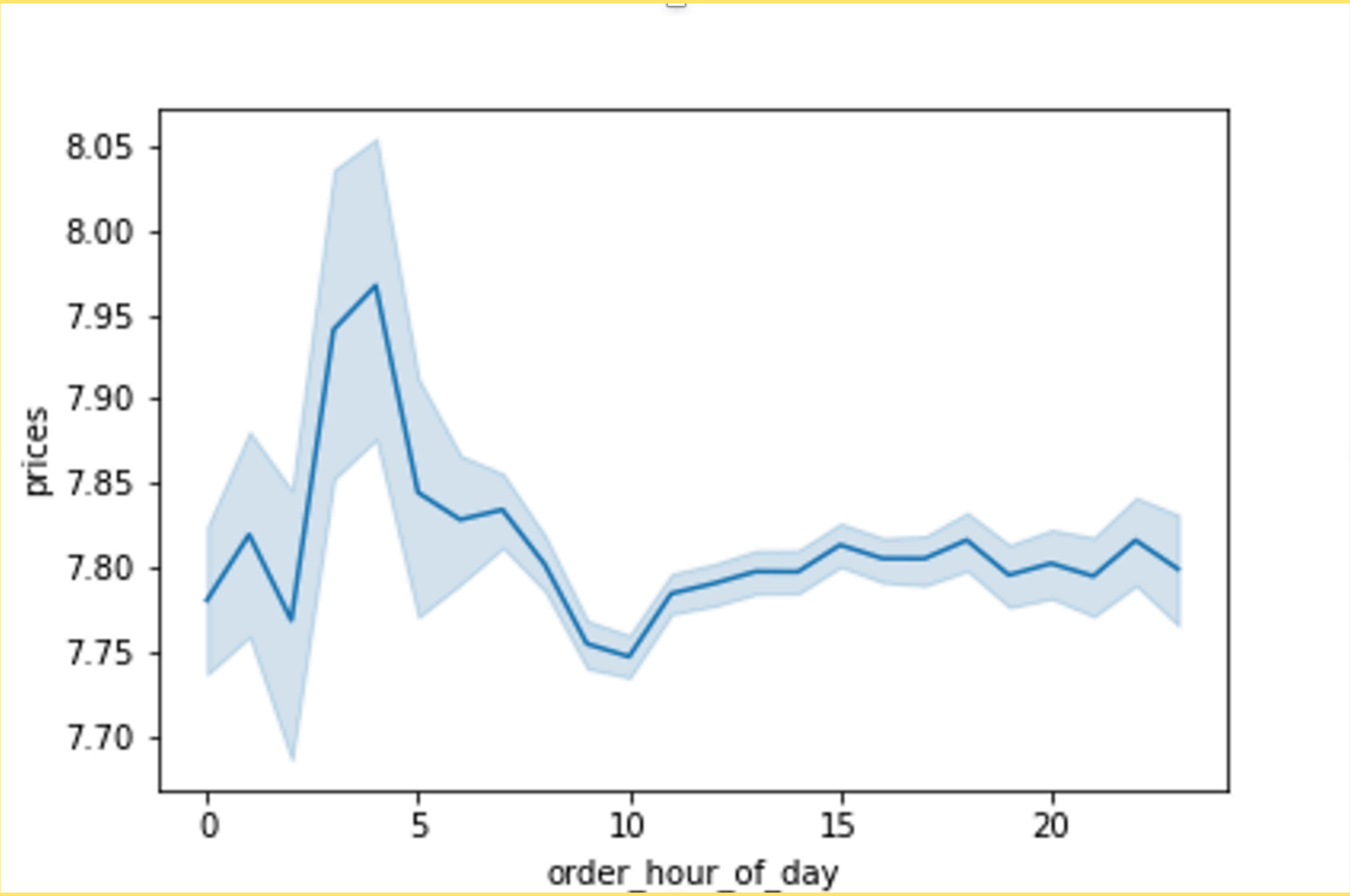
Presentation

Visualise answers & results with Python, including:

- Bar charts
- Line charts
- Histogram

Final report including population flow can be found here.





This chart shows the customer profile distribution by region. All in all most customers fall in the category "no dependants".

This line chart shows that 5 am is the time of the day at which the most expensive items are bought.

Skills & Tools

- Wrangling & subsetting data with Python
- Consistency checks
- Combining & exporting data
- Deriving new variables
- Grouping & aggregating variables
- Data visualisation in Python
- Reporting in Excel
- Population flows
- Jupyter Notebook
- Anaconda libraries manager
- Python libraries Pandas & NumPy
- Matplotlib, Scipy & Seaborn
- Excel

Read my code on Github!

4. Berlin Airbnb Case Study

Overview

Berlin has a chronic shortage of available and affordable long-term rental apartments.

Airbnb has been blamed for facilitating the commercial exploitation of apartments, which could otherwise be used as homes for residents.

In this case study, I explore the impact of commercial hosts on the Berlin rental market.

Goal

- Help a legal company make a case for local tenants' rights to a safe and affordable home
- Answer questions such as:
  - Which are the most popular neighbourhoods?
  - How can we identify commercial hosts?
  - What impact do they have on the local Berlin rental market?

Data used

Open source data from Insideairbnb, including 19,000+ listings from Airbnb in Berlin scraped in July 2021. Data includes price, availability, neighbourhoods and reviews.

It can be found here ("Inside Airbnb") and is licensed under Creative Commons CC0 1.0 Universal (CC0 1.0) "Public Domain Dedication".

I also used geodata from Funkeaktiv, that can be found here and has License: CC-BY.

Process & methodology

Preparation

- Source data
- Wrangle data & check consistency with Python
- Conduct visual exploratory analysis
- Derive new variables

Analysis

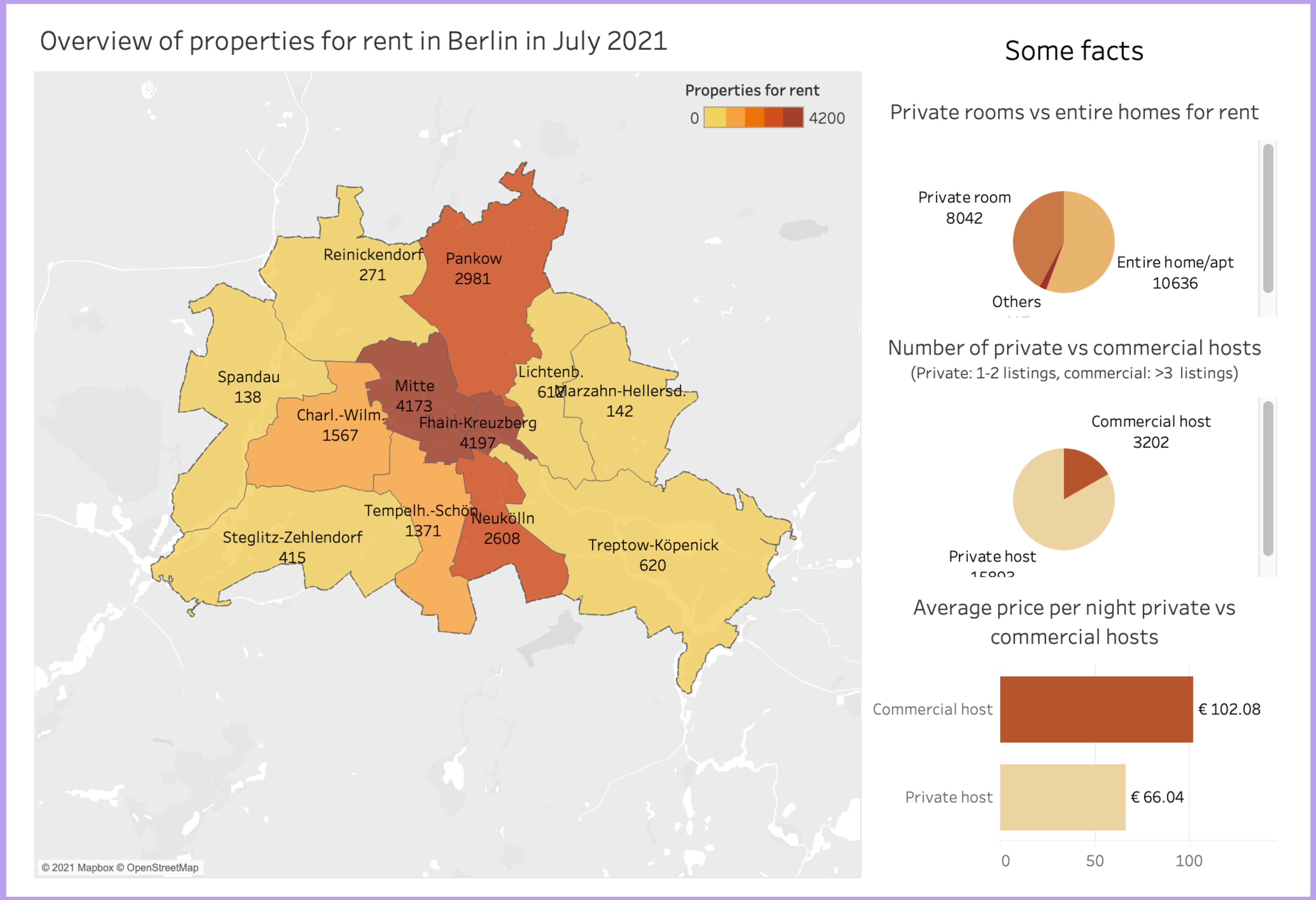
- Supervised machine learning: linear regression
- Unsupervised machine learning: cluster analysis
- Spatial analysis
- Time series analysis
- Statistical visualisations in Python

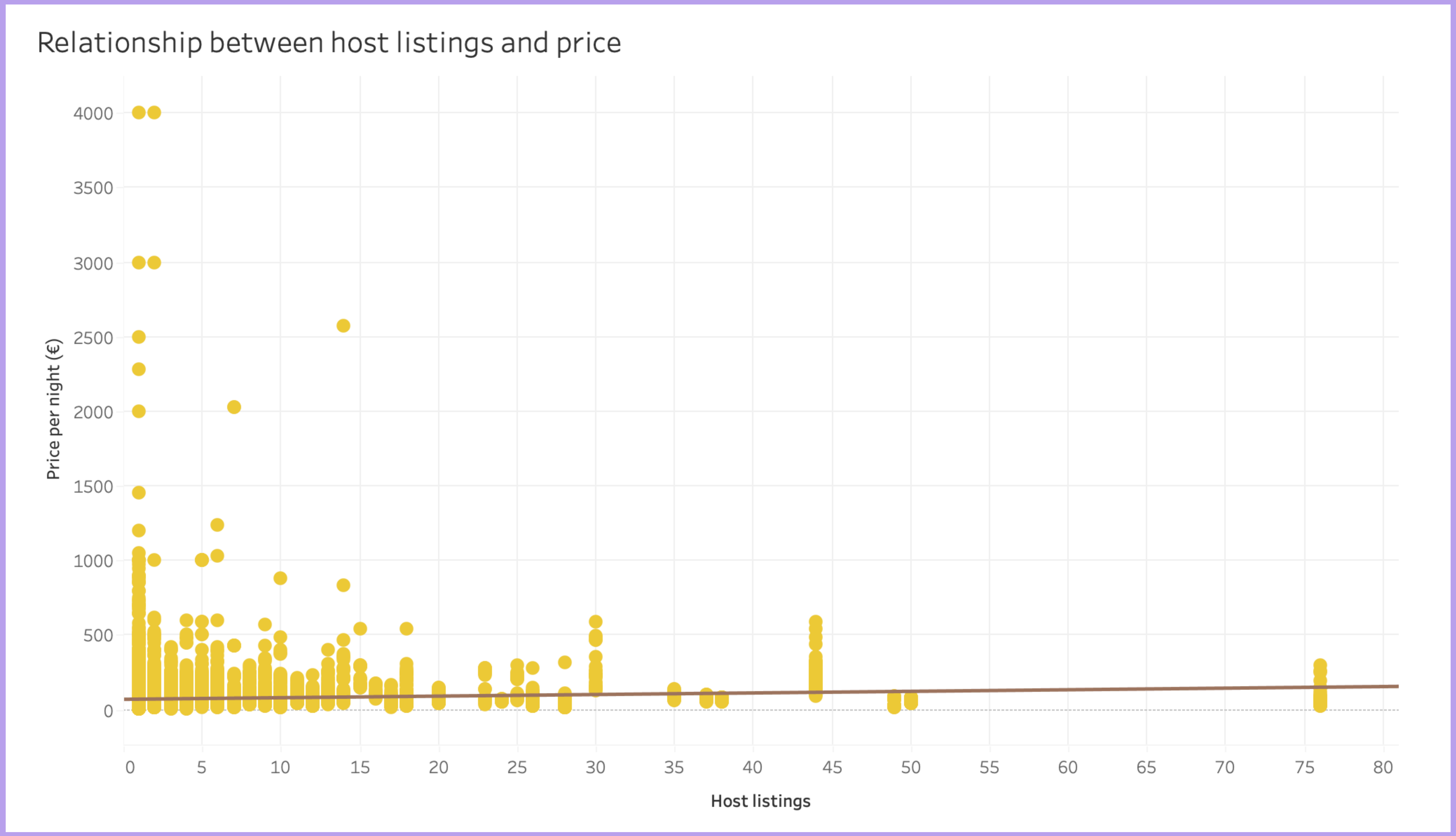
Presentation

Visualise analytical journey & key results in Tableau, including:

- Advanced dashboards
- Choropleth & symbol maps
- Pie & bar charts
- Scatterplots

The final presentation can found here.





On this dashboard we can see that the prices of commercial hosts are significantly higher than those of private hosts.

Here I perform a linear regression to test this relationship between the number of host listings and price.

Skills & Tools

- Scikit, Folium, Pylab libraries
- Advanced dashboard design
- Linear regression
- Cluster analysis
- Time series analysis
- Visual EDA with Python
- Geographical visualisations with Python
- Supervised machine learning
- Unsupervised machine learning

Read my code on Github!

Thank you!

fortunyjulia@gmail.com
https://www.linkedin.com/in/juliafortuny/
https://public.tableau.com/app/profile/julia.fortuny
https://github.com/juliafor/

Julia Fortuny Wollny, October 2021