Alexandra Borschke

Data Analyst Portfolio

Image source: https://iconscout.com/

Projects

- Global solid waste emissions
- Instacart
- Rockbuster Stealth
- Preparing for influenza season in the USA 2018
- GameCo
- Pig E. Bank

! Note that the portfolio contains only parts of the projects' analyses. Links to detailed analyses (and other important materials) can be found in the "Additional sources" at the end of each project's section.

Image Source: https://png.pngtree.com/

Global solid waste emissions 2015-2021

Analysis of waste emissions, their global distribution and relationships with socio-economic factors

Global solid waste emissions 2015-2021

Analysis of waste emissions, their global distribution and relationships with socio-economic factors

Project goal

Determine the leading countries in terms of gas emissions from solid waste generation and analyse relationship between emissions and socio-economic factors, such as emissions & population, as well as relationship between GDP per capita and a country's impact in waste management.

Key objectives

- To conduct a EDA and determine relationships in the data;
- To formulate a hypothesis and conduct regression analysis to test the model;
- To conduct a cluster analysis using k-means to identify groupings in the data;
- To determine leading countries in terms of gas emission generation;
- To conduct a geographical analysis;
- To create visualizations for all analytical steps.

Data sets

- Solid waste emissions
- Population
- GDP per capita
- WMG index

Sources: UN, EPI, Climate TRACE

Limitations:

- data sets on WMG contain data only for 2019-2022;
- 50% of data are missing values;
- Most of the data values are estimates

Tools used

Skills & procedures

- Data wrangling & subsetting
- Data merging
- Data consistency checks
- Deriving variables
- Grouping data
- Aggregating data
- Combining & exporting data
- Regression analysis
- Time series analysis
- Cluster analysis with k-means
- Geo visualizations in Python
- Visualizations in Tableau

Data analysis stages

Data preparation

- ✓ Basic descriptive exploratory tests;
- ✓ Assessing data values and creating new dataframes;
- ✓ Removing duplicates, managing missing values, mix-typed data checks;
- ✓ Merging data sets.

Analysing relationships

- ✓ Conduct a EDA to determine possible correlations between variables and formulate a hypothesis;
- ✓ Conduct a regression analysis to test the model;
- ✓ Conduct a cluster analysis to determine data groupings;
- ✓ Conduct a time series analysis.

Geo-spacial analysis

- ✓ Conducting a geospacial analysis in Python to determine the geographical distribute of data values.

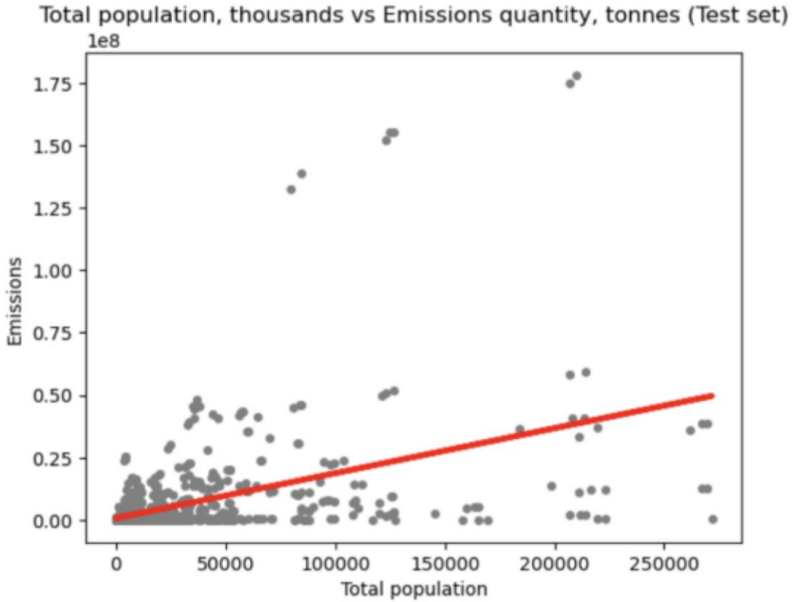
Data visualization

- ✓ Creating histograms, bar charts, line charts, scatterplots, map charts for different variables and relationships between variables in Python and Tableau

Analysis: testing relationships

- **Hypothesis:** The higher the population in a country, the higher the solid waste emissions' generation.

The model proved not to be conclusive (R2 = 0.21), and an addition of more variables (e.g., development level of countries with similarly large populations) to the model is required to find evidence of a linear relationship between population and emissions.



- **Hypothesis:** The higher the GDP per capita of a country, the higher its WMG index.

- A logarithmic regression analysis was conducted to test the hypothesis.
- The relationship between WMG indicator and GDP per capita is non-linear. The vast majority of growth happens before 60 WMG, and afterward, it takes significant growth in GDP to grow the WMG up to 100.

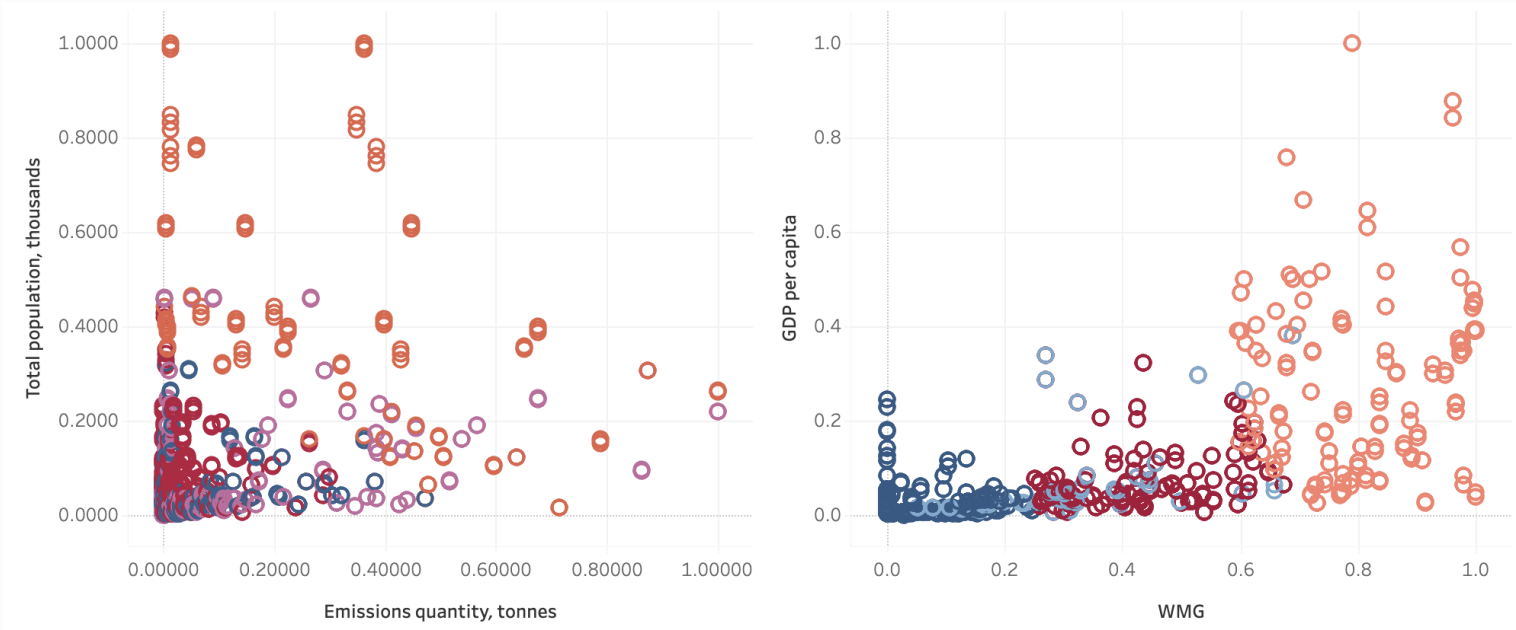
Cluster analysis using k-means

Population - Emissions clusters

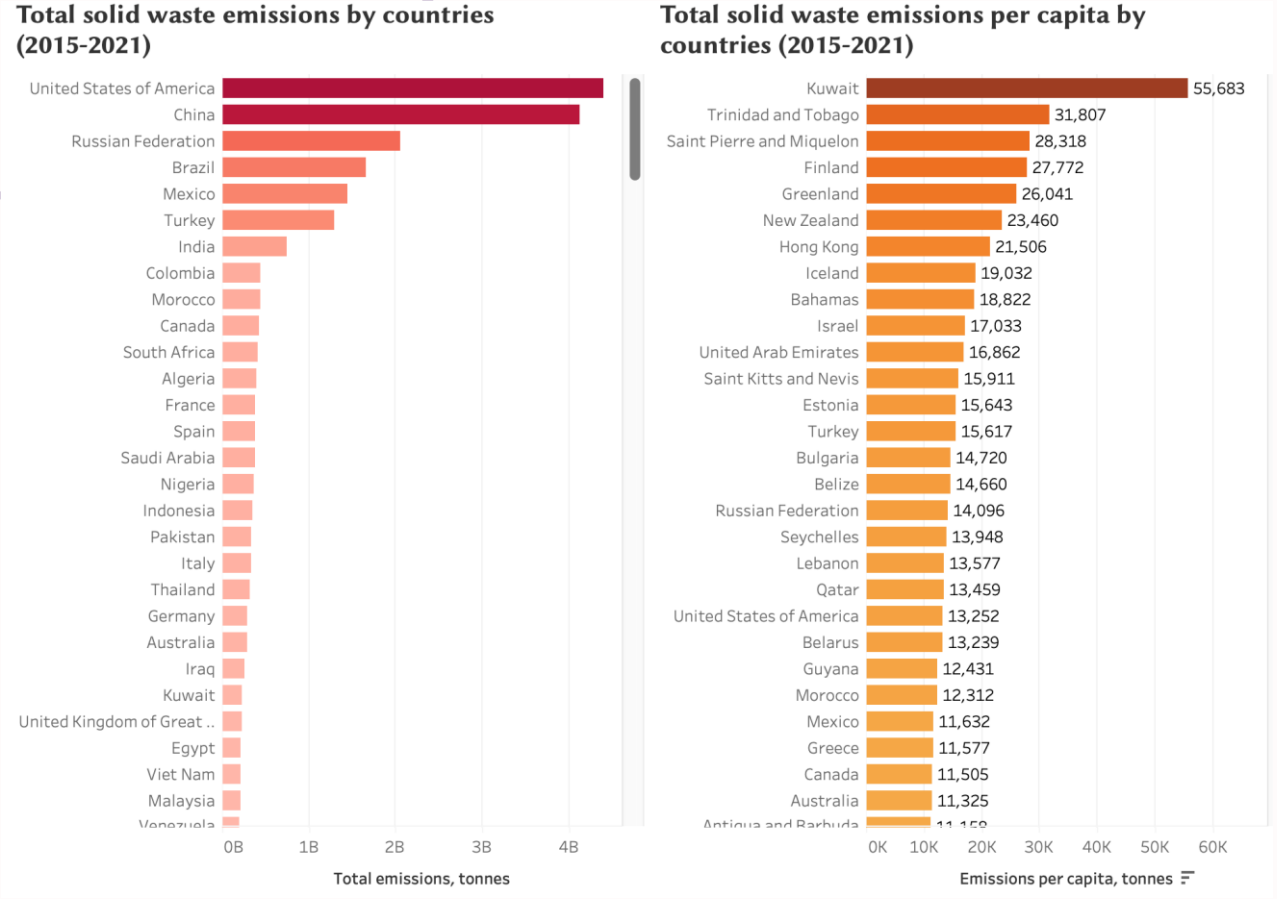
- The highest amounts of emissions are produced by countries with lower populations (parts of orange and light-purple clusters), whereas countries with higher populations tend to produce less emissions in total (orange cluster). There are only a few data points from the orange cluster that indicate both: higher population and relatively high emissions.
- The data points of other tree clusters (dark-red, blue and partially light-purple) are concentrated on the crossroad of lower population and lower emissions.

GDP per capita - WMG index

- There is a cluster (in light-orange) of a few data points that indicates that there are countries that have a higher GDP per capita, which also indicate higher WMG levels.
- However, there are also a lot of data points in another cluster (dark-red) and hald of the light-orange cluster, where higher WMG levels go along with lower GDPs per capita.
- There is a clear evidence though that most of the countries with lower GDP per capita (dark-blue and partially light-blue clusters) also indicate lower WMG levels.



Analysis: geographical distribution of emissions



- There are clear leaders in terms of total solid waste emissions in 2015-2021: USA, China, Russia, Brazil, Turkey, and Mexico.

There is a striking difference between countries in terms of total emissions vs. emissions per capita throughout the years. The countries that produced an insignificant proportion of total emissions remain the leaders in terms of generated emissions per capita.

- In terms of solid waste emissions per capita the leaders are Kuwait, Trinidad & Tobago, Saint Pierre & Miquelon, Finland, Greenland.

To take a closer look at the map chart view my Tableau storyboard:

Summary and next steps

Summary

- There are clear leaders in terms of total solid waste emissions in 2015-2021: USA, China, Russia, Brazil, Turkey, and Mexico.
- In terms of solid waste emissions per capita the leaders are Kuwait, Trinidad & Tobago, Saint Pierre & Miquelon, Finland, Greenland.

- There is no linear relationship between population and emissions produced, however the cluster analysis indicated that there are different groupings that indicate some patterns of relationships between the variables.

- The same applies to the relationship between GDP per capita / WMG index, although this pair indicated a stronger relationship than the one between population and emissions. Again, the cluster analysis produced different groupings that help understand the relationship better.

Further investigation

- Further research and gather more data on socio-economic factors (variables) that might influence solid waste generation.
- Carry out the same project with new variables to indicate relationships and once the relationship is indicated, train a model to predict solid waste generation more accurately.
- Analyse solid waste emissions in separate continents to have a more detailed picture.

Additional sources

Python scripts:

Tableau presentation

Image Source: https://www.instacart.com/

Instacart

Analysis of customer profiles, buyer behavior, and sales for an online grocery delivery service

instacart

is an online grocery store that want to uncover more information about their sales pattern to build a successful marketing campaign

Project goal

Determine customer profiles and buying behaviour, as well as sales patterns in departments to gain insights and give recommendations for targeted marketing campaign to the marketing and sales departments.

Key objectives

Determine:

- busiest days and hours for sales;
- price range groupings;
- the most popular store departments;
- customer ordering profiles based on various features such as brand loyalty, demographic info, order frequency, family status, regions, etc.

Data sets

- Customers
- Orders
- Departments
- Products
- Data dictionary

Sources: Instacart & CareerFoundry

Limitations:
- data sets contain only data as of 2017

Tools used

python, ANACONDA., jupyter, pandas, matplotlib, X, NumPy, seaborn

Skills & procedures

- Data wrangling & subsetting
- Data merging
- Data consistency checks
- Deriving variables
- Grouping data
- Aggregating data
- Combining & exporting data
- Visualizations in Python
- Excel reporting
- Population flows

Image Source: https://www.instacart.com/

Data analysis stages

Data preparation

- ✓ Basic descriptive exploratory tests;
- ✓ Assessing data values and creating new dataframes;
- ✓ Removing duplicates, managing missing values, mix-typed data checks;
- ✓ Merging data sets.

Deriving new variables

- ✓ Create new columns using conditional logic;
- ✓ Creating flags and placing them in new columns;
- ✓ Creating summary columns of descriptive statistics;

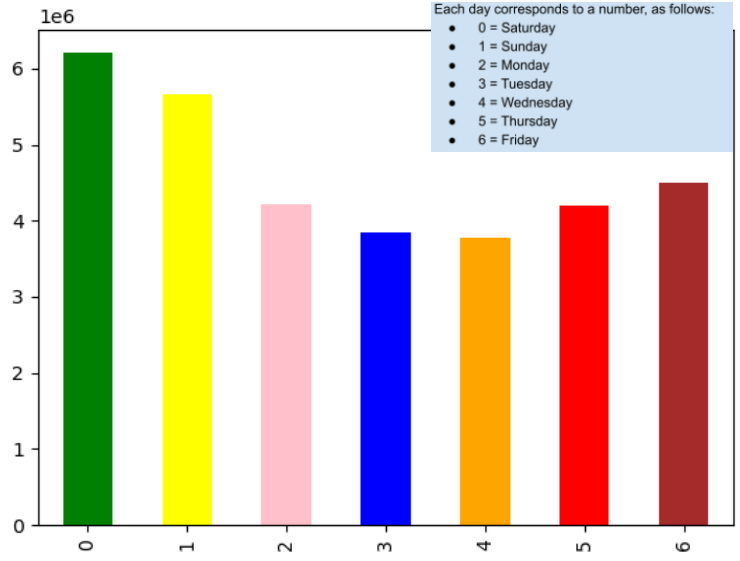
Data visualization

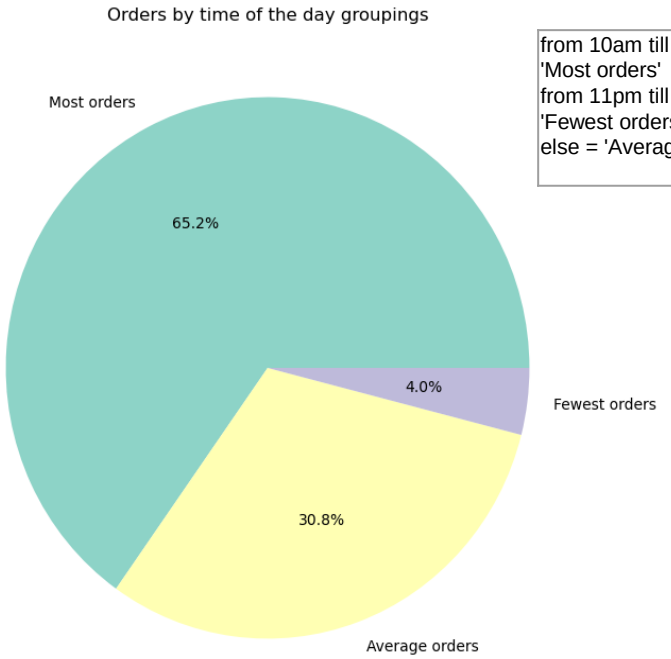
- ✓ Creating histograms, bar charts, line charts, and scatterplots for different variables and relationships between variables

Excel reporting

- ✓ Summarizing analysis findings and describing what connections in the data were found;
- ✓ Creating a report that describes the analysis methodology, the analysis, the results, visualizations, and recommendations.

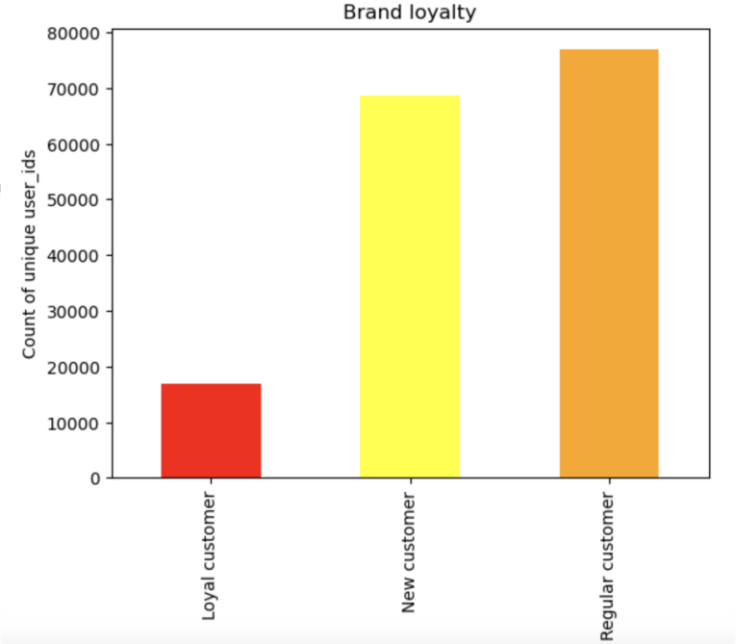
Analysis: busiest days and hours

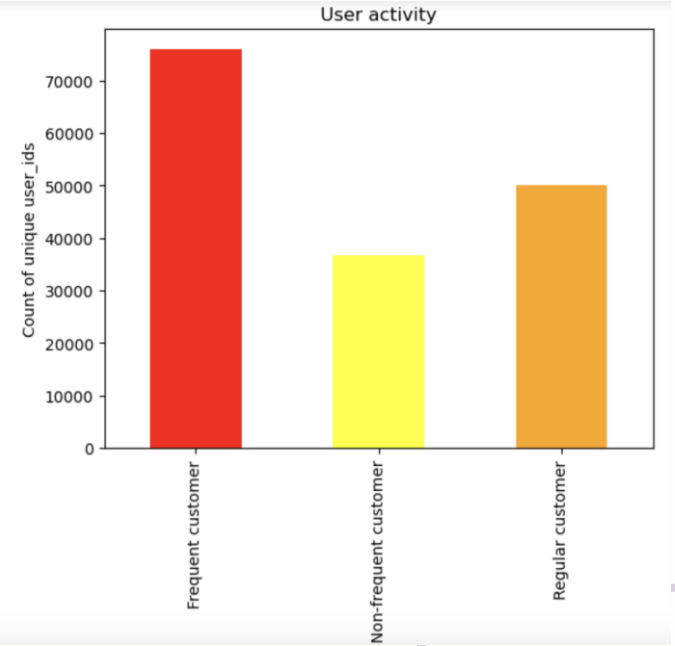




- The busiest days are Saturday and Sunday. The least busy days are Tuesday and Wednesday.
- Most of the orders (65,2%) happened between 10 am and 4 pm, with the peak at 10 am.

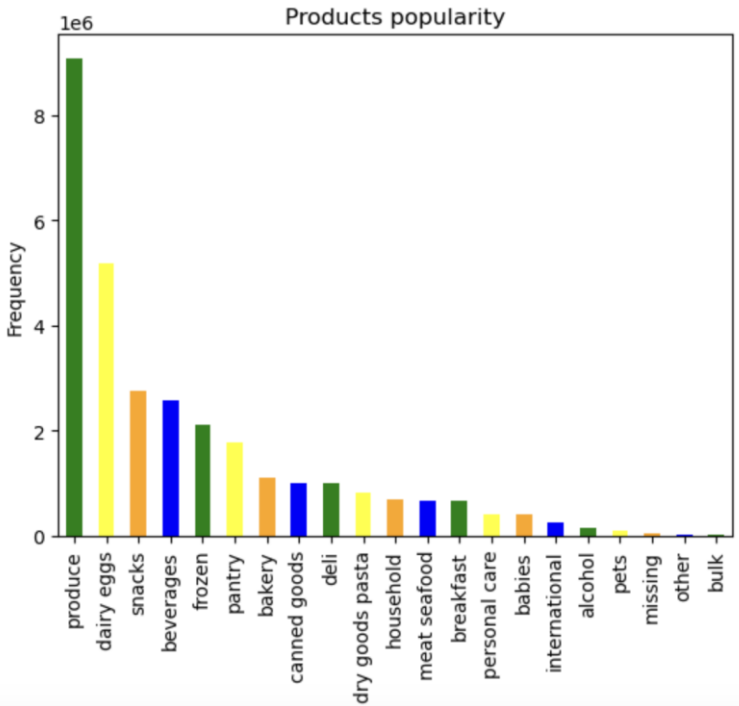
Analysis: customer profiles

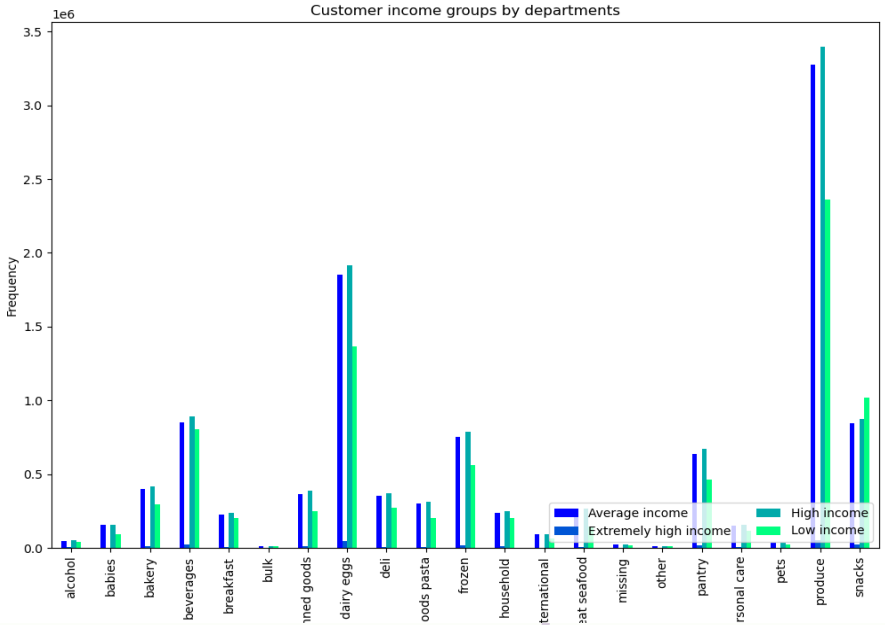




- Most of the Instacart customers are represented by regular (brand loyalty) and frequent customers (user activity), that is those customers who have placed between 10 and 40 orders and those who wait 10-20 days till placing a new order.
- New customers (those who have placed less than 10 orders so far) also constitute a large proportion of customers, who represent an untapped segment and should be especially targeted by the marketing campaign in order to turn them into regular customers first.

Analysis: department popularity





- Produce is the most popular department, with more than 8mln. items sold, followed by dairy eggs (about 5mln items), snacks and beverages (ca.2-2,5 mln. items), and frozen and pantry products.
- Customers with high income consume more products in each department in general than other income groups. However, in terms of snacks, low income group demonstrates the highest consumption.

Some recommendations for Instacart

Marketing campaign: timing

Introduce marketing campaign during Tuesday and Wednesday and also daily between average orders hours and fewest orders hours to increase demand. Create discounts for expensive items during Fridays and Saturdays.

Marketing campaign: customer focus

- Attract more customers with low income by offering discounts and introducing a section in the app with only cheap items.
- Introduce discounts and create special offers for new/non-frequent customers to make them stay on the platform and become regular/frequent or loyal customers.

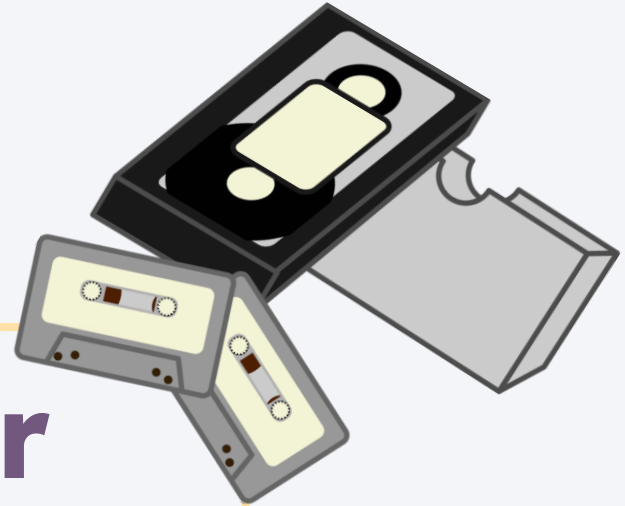
Marketing campaign: department focus

- Offer promotions in the departments with a lower demand
- Introduce more high-range products and advertise them to gain more high spenders.

Additional sources

Python scripts:

Detailed final report:



Rockbuster Stealth

Helping the BI department with the launch of a new online video rental service

Rockbuster Stealth

is a fictitious movie rental company that wants to launch an online video rental service to stay competitive on the market

Project goal

To help the BI department with the launch strategy for the new online video service: conduct a data analysis and give recommendations based on its results.

Key objectives

- To conduct an EDA and describe patterns in the data, such as the average rental rate, revenue, number of active customers, etc.

To determine:

- Top 10 countries and cities with the highest number of customers in 2006.
- Top 10 countries with the highest revenue and the 5 most loyal customers in 2006.
- The most popular film genres in 2006.

Data sets

- Rockbuster Sales Data Set

Source: CareerFoundry

Limitations:
- data set contains data only for 2006

Tools used

SQL

PostgreSQL

tableau

DbVisualizer

Skills & procedures

- Understanding relational databases
- Analysing Entity Relationship Diagram (ERD)
- Database querying in SQL
- Filtering, cleaning and summarizing data in SQL
- Joining tables
- Coding subqueries
- Coding common table expressions (CTE)
- Creating a data dictionary
- Visualizations of results in Tableau

Data analysis stages

Database analysis

- ✓ Setting up a database environment using PostgreSQL;
- ✓ Analysing keys and indexes of the database;
- ✓ Extracting an ERD and starting creating a data dictionary.

Cleaning & Summarizing data in SQL

- ✓ Conducting basic CRUD operations and SQL commands;
- ✓ Organizing, sorting, and filtering data;
- ✓ Identifying and cleaning dirty data;
- ✓ Creating a profile of summary statistics (EDA).

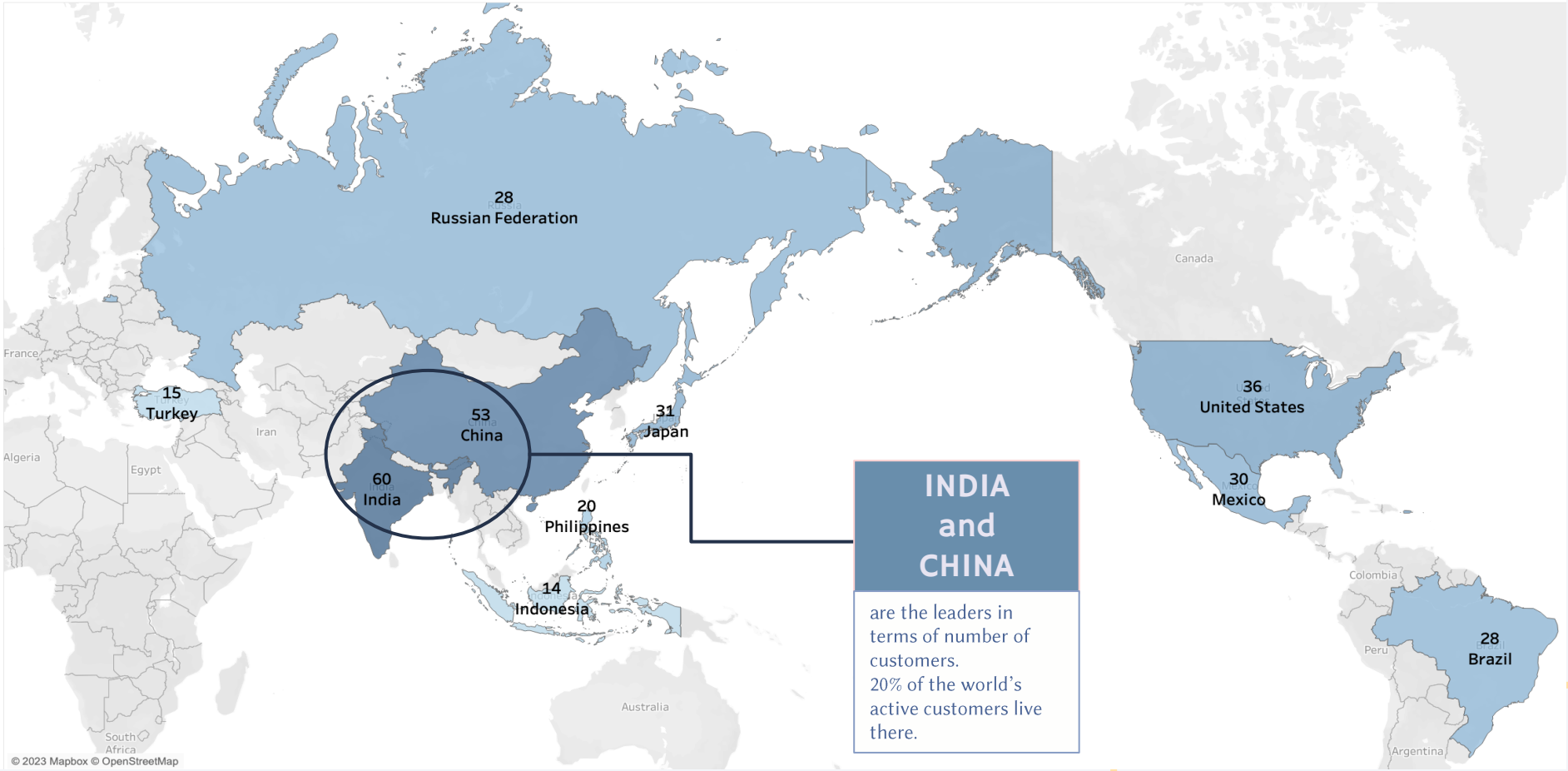
Answering business questions

- ✓ Using SQL to join tables;
- ✓ Writing subqueries and applying common table expressions;
- ✓ Answering complex business questions (key objectives) by using CTEs, subquieries, and joins.

Presenting results

- ✓ Creating visualizations and a storyboard in Tableau;
- ✓ Creating a final presentation with analysis results and recommendations for Rockbuster Stealth in a PPT.

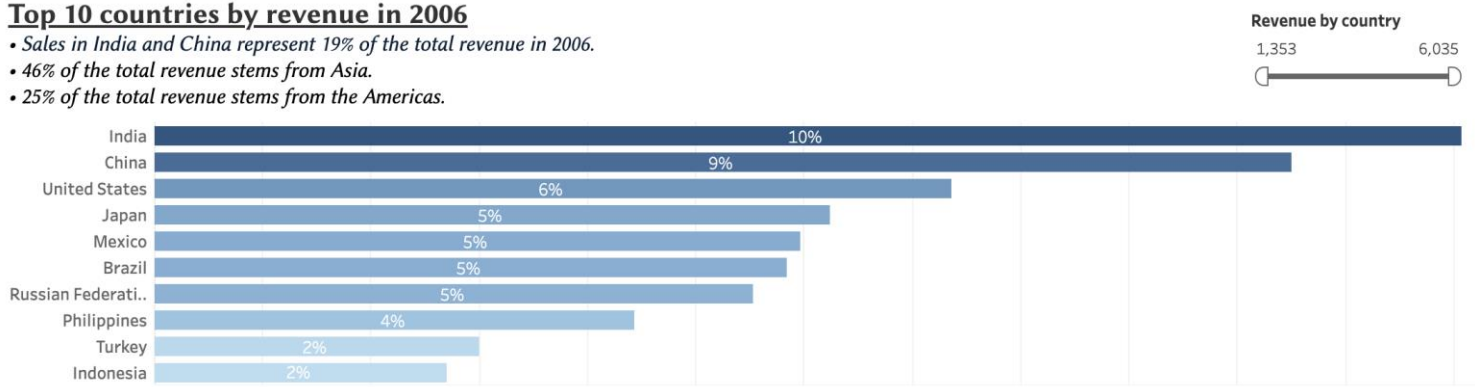
Analysis: top 10 countries by customer number



Analysis: revenue 2006

Top 10 countries by revenue in 2006

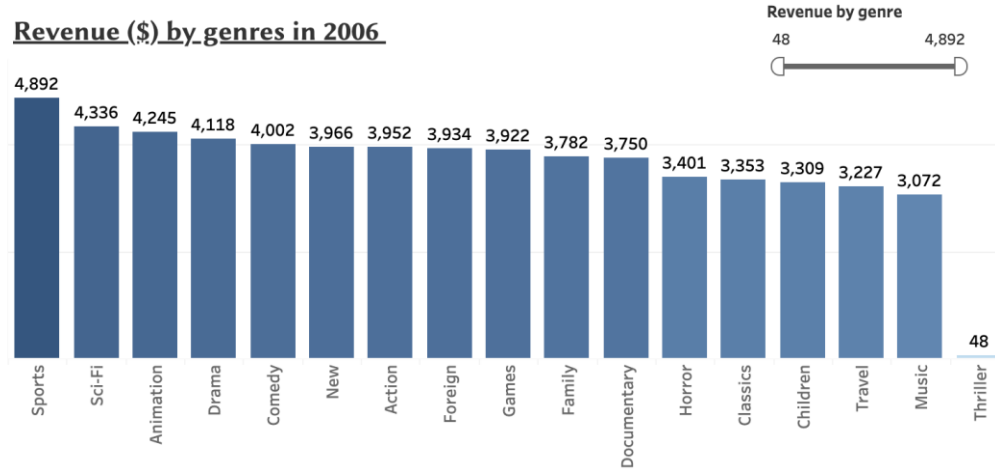
- *Sales in India and China represent 19% of the total revenue in 2006.*
- *46% of the total revenue stems from Asia.*
- *25% of the total revenue stems from the Americas.*



The most loyal customers of 2006 by revenue

First na..	Last na..	Country	City	
Arlene	Harvey	India	Ambatt..	111.76
Kyle	Spurlock	China	Shanwei	109.71
Marlene	Welch	Japan	Iwaki	106.77
Glen	Talbert	Mexico	Acua	100.77
Clinton	Buford	United States	Aurora	98.76

Revenue ($) by genres in 2006



Recommendations for launching an online rental service

Films

Focus on marketing films of the following popular genres: *Sports, Sci-Fi, and Animation, Comedy, Drama and New.*

Market expansion

- Actively advertise in Europe, Africa, and Oceania to attract new customers.
- Keep being active in Asia and the Americas.

Loyalty program

Introduce a loyalty program – thank your most loyal customers with discounts and special offers. Especially focus on customers from Europe, Africa, and Oceania, where the customer activity is still low. Remain active in Asia and the Americas.

Reduce prices

- Reduce the price for a rental by $0,30 - 0,50 to gain more customers.
- Introduce a subscription service that is more beneficial for customers in terms of costs.

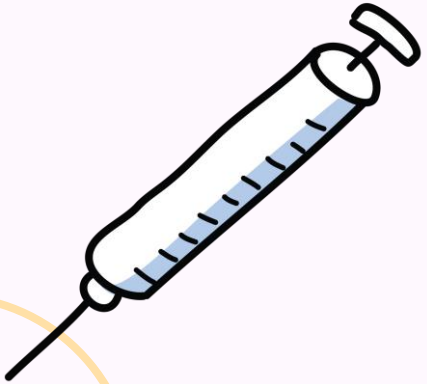
Additional sources

SQL code:

PostgreSQL

Tableau storyboard:

Data dictionary:



Preparing for the 2018 influenza season in the USA

Analyzing trends in influenza and determining where to send additional medical staff

Preparing for influenza season in the USA

Project goal

Determine when and where to provide extra medical staff for US hospitals for the next influenza season.

Key objectives

Determine:

- whether influenza occurs seasonally or throughout the entire year;
- forecast influenza season for 2018
- characteristics of vulnerable population and their locations (US states);
- which states need more medical staff.

Data sets

Influenza deaths by geography, time, age, and gender
Source: CDC

Population data by geography
Source: US Census

Limitations:
- 80% of influenza death data is suppressed
- manual data collection might lead to errors
- time lag

Tools used

Skills & procedures

- Data quality check, profiling & cleaning
- Data integration & transformation
- Data grouping & summarizing
- Statistical hypothesis testing (t-test)
- Forecasting
- Visual analysis
- Visualizing results & storytelling in Tableau

Data analysis stages

Designing a research project

- ✓ Translating business requirements into a research project via formulating research questions and hypotheses
- ✓ Sourcing the relevant data for the project
- ✓ Preparing a project management plan

Data preparation

- ✓ Checking for integrity & quality issues
- ✓ Cleaning data, removing duplicates and handling missing values
- ✓ Transforming data and integrating two data sets
- ✓ Deriving new variables

Statistical hypothesis testing

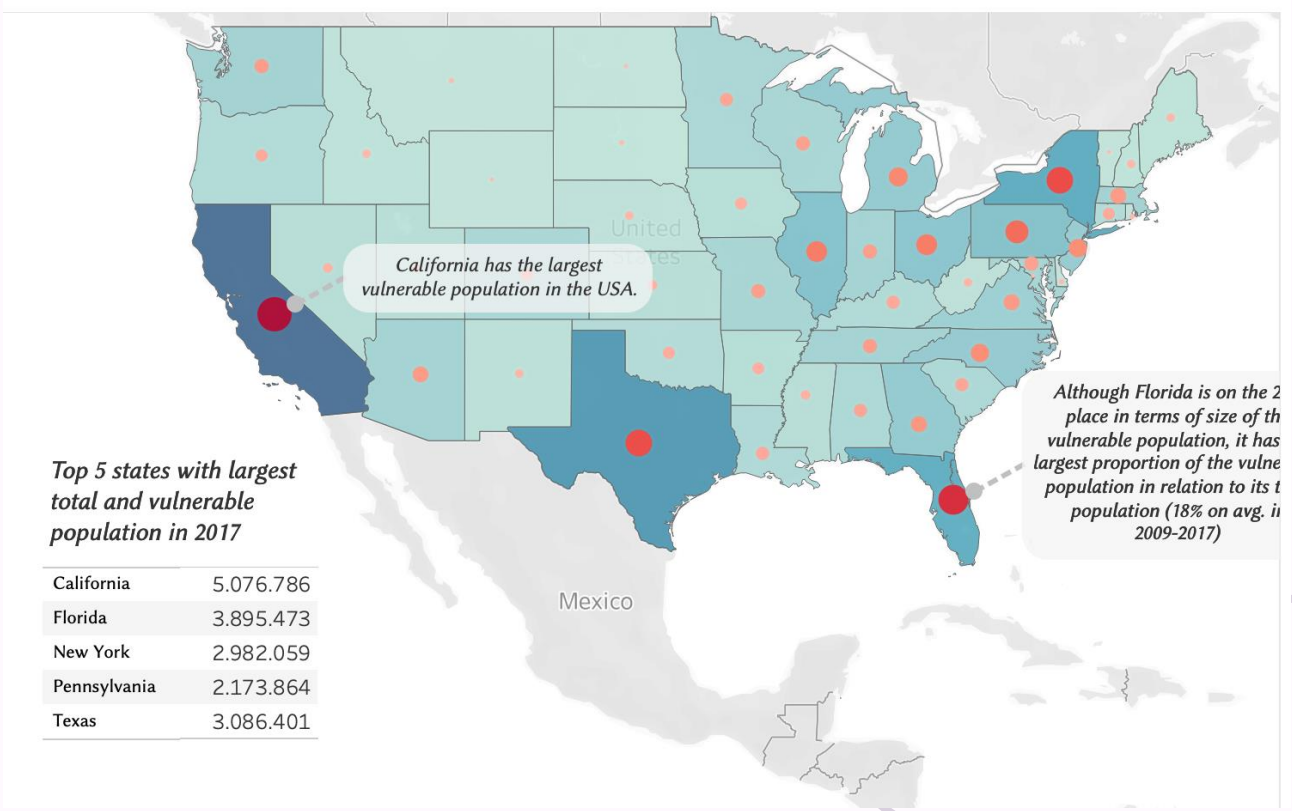
- ✓ Calculating variance and data spread (variance, st.dev., outliers) for two variables in question
- ✓ Calculating correlation between two variables
- ✓ Conducting a two-sample t-test to test the hypothesis
- ✓ Interpreting the results

Data visualization & Storytelling in Tableau

- ✓ Creating composition & comparison charts
- ✓ Temporal visualizations and forecasting
- ✓ Statistical visualizations: box plots, histograms, bubble charts, & scatter plots
- ✓ Spatial analysis
- ✓ Creating a narrative and recording a video presenting results

Analysis: influenza vulnerable group

- The statistical testing indicated the vulnerable group is represented by the US population older than 65.
- The spatial analysis visualized on the right showed that California, Florida, New York, Pennsylvania & Texas are the states with the highest number of vulnerable population.



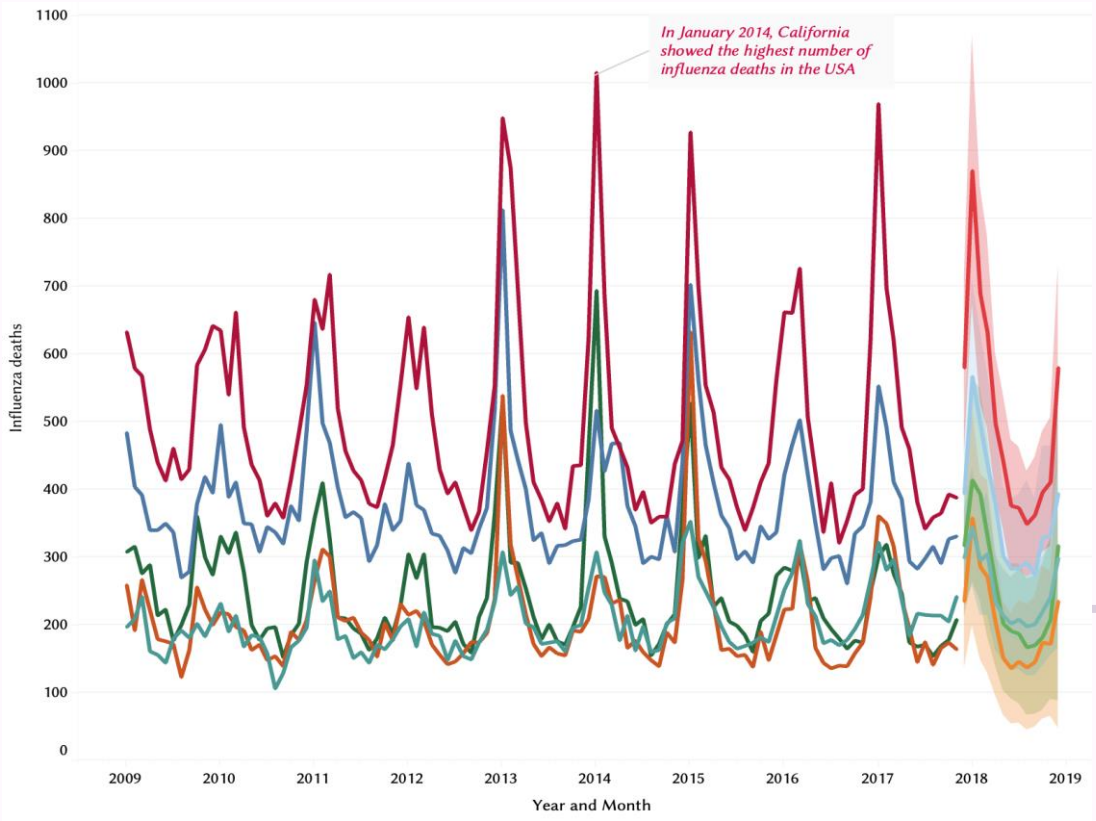
Top 5 states with largest total and vulnerable population in 2017

California	5.076.786
Florida	3.895.473
New York	2.982.059
Pennsylvania	2.173.864
Texas	3.086.401

View full Tableau storyboard here:

Analysis: influenza season forecast for 2018

- Winter and early spring are the peak seasons of influenza activity in the USA: 45 % of total deaths from influenza occurred between December and March (2009-2017).
- January is the month with the highest numbers of influenza deaths.
- As in the previous years (2009-2017), winter and early spring will be influenza seasons in the USA in all states.



View full Tableau storyboard here:

Recommendations for the upcoming influenza season 2018

When to send additional medical staff?

December, January, February, and March 2018

Where is additional medical staff needed?

First consider highly populated states that have higher numbers of vulnerable populations, such as California, Florida, Texas, New York, and Pennsylvania.
These states will also need more influenza vaccines supplied.

Preparations outside influenza season (April – September)

- Staff training and evaluation; Staff distribution in hospitals
- Influenza vaccination marketing
- Plan influenza vaccine supply

Additional sources

Project management plan and interim report:

Detailed data analysis:

Tableau storyboard:

Video presentation:

GameCo

Analyzing global video game market trends

GameCo

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

Project goal

To foster, with a help of a descriptive analysis, a better understanding of how GameCo's new games might fare in the market.

Key objectives

To determine:

- global and regional trends in video game sales
- regional markets with current rapid growth.
- popular video game genres and platforms.

Data set

Historical sales of video games spanning different platforms, genres, and publishing studios.

Source: VGChartz

Limitations:
Sales figures can differ from the official statistics of manufacturers due to a different method of sales calculation. See the source for more information.

Tools used

Techniques

- Data sorting, filtering, and cleaning
- Grouping & summarizing data
- Deriving new variables
- Descriptive analysis
- Visualizing results in Excel
- Presenting results

Data analysis stages

Data cleaning

- ✓ Removing duplicates
- ✓ Removing erroneous data values
- ✓ Imputing missing values
- ✓ Correcting inconsistencies in formating

Data grouping and summarizing

- ✓ Using pivot table to gain general data insights
- ✓ Grouping data
- ✓ Applying filters to look at specific segments
- ✓ Deriving new variables

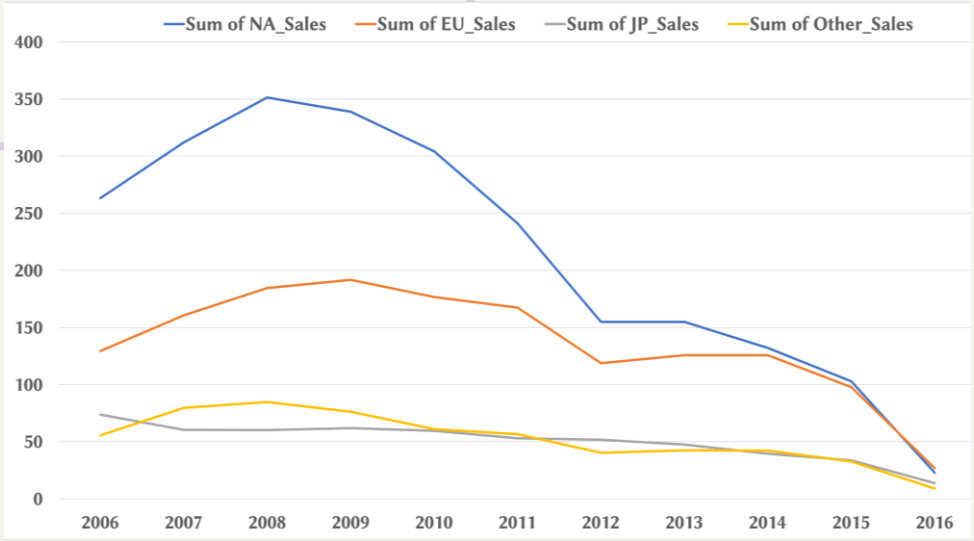
Descriptive analysis

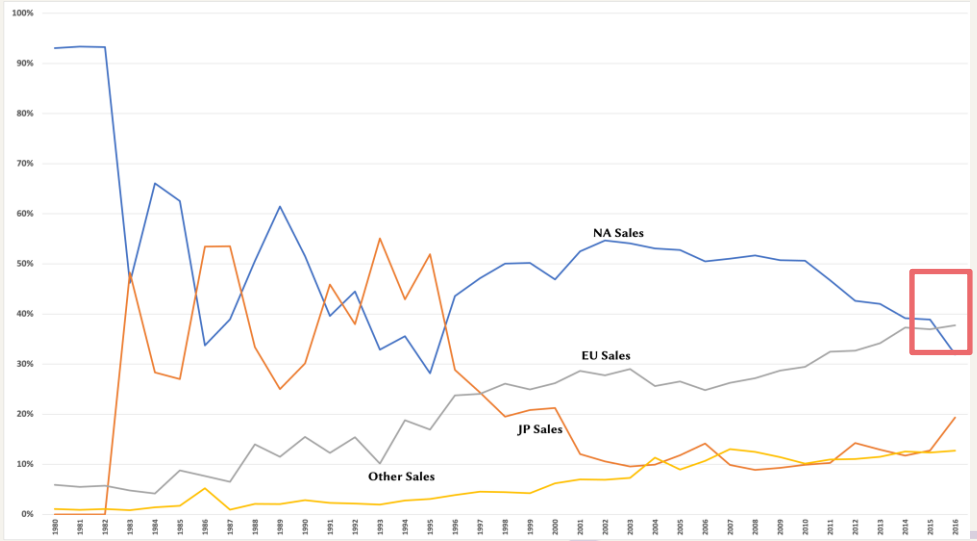
- ✓ Analyzing basic statistics features: mean, median, mode
- ✓ Analyzing data distibution and skeweness
- ✓ Identifying outliers

Developing insights & visualizations

- ✓ Conducting analysis of market trends according to key objectives
- ✓ Creating visualizations
- ✓ Consolidation of project deliverables (reflections, presentation)

Analysis: market trends

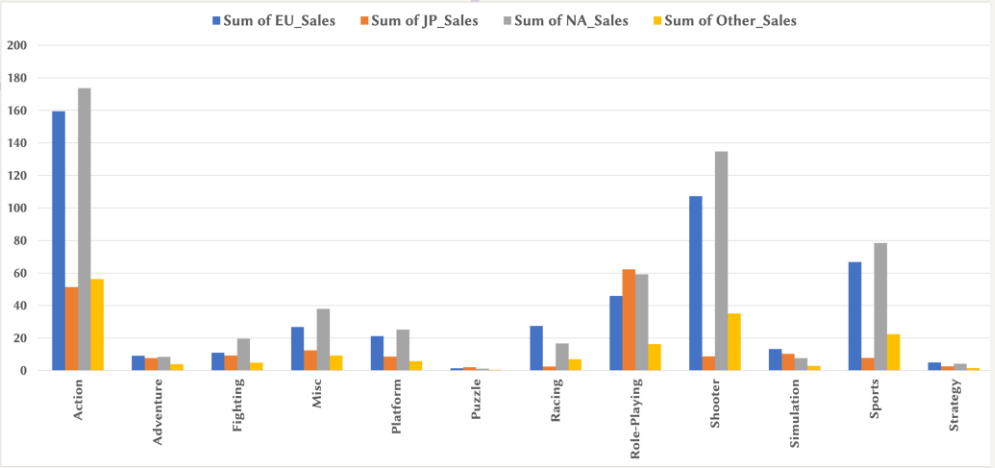


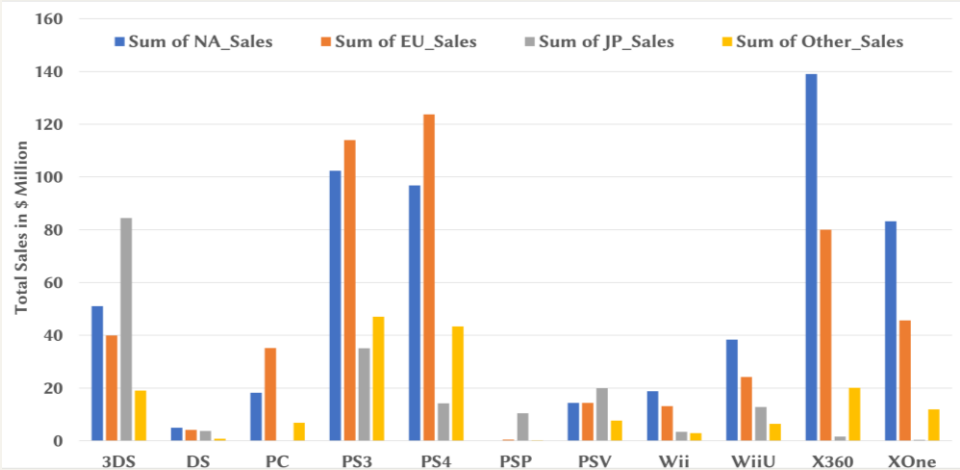


- There has been an overall decline of the videogames' sales in all regions as of 2008.
- NA shows the strongest drop of sales between 2008 and 2016. The sales in EU and JP show a more gradual decline.

- With a continous growth since 2006, EU takes over the leading position in the market in 2016. JP alo shows growths in 2016.
- NA loses its market leader position in 2016.

Analysis: popular genres and platforms





- The most successful genres for EU: Action, Shooter, Sports
- For NA: Action, Shooter, Sports, Platform, Misc, Fighting
- For JP: Role-Playing, Puzzle, Action

- The most profitable platforms for EU: X360, PS3, PS4, XOne
- For NA: X360, PS3, PS3, XOne
- For JP: 3DS, PS3, PSV

Recommendations to GameCo for 2017

European market

The major focus should be on the development of the EU market since it is currently gaining its leader position.
Focus on the development and marketing of games of the following genres: Action, Shooter, and Sports. Produce and market more games for Sony and Microsoft platforms.

North American market

Focus on investing into development and marketing of the most popular games in NA (Action, Shooter, Sports, Platform, Misc, Fighting) on Sony and Microsoft platforms to boost sales and in order to regain a strong position of NA in the global market.

Japanese market

Since JP is showing growth in 2016, focus on investing more into the development of the market by producing and marketing more Role-Playing, Puzzle, and Action games for Nintendo and Sony platforms.

Additional sources

Detailed data analysis:

Final presentation:

Project reflections:

Pig E. Bank

Identifying clients who can potentially leave the bank

Pig E. Bank

is a fictitious bank that wants to integrate a data mining algorithm to predict which clients might leave the bank

Project goal

Using a data mining mechanism, build a decision tree to predict which clients might leave the bank.

Key objectives

- Calculate basic statistics to understand the data
- Identify factors for leaving the bank by using pivot tables
- Analyse statistical information for 2 groups of clients (those who left vs. those who stayed)
- Build a decision tree

Data set

Client Data Set

Source: CareerFoundry

Tools used

Techniques

- Data sorting, filtering, and cleaning
- Grouping & summarizing data
- Descriptive analysis
- Data ethics
- Data mining
- Building a decision tree as a data mining algorithm

Data analysis stages

Data cleaning

- ✓ Removing duplicates
- ✓ Removing erroneous data values
- ✓ Handling missing and sensitive PII data
- ✓ Correcting inconsistencies in formating

Performing a statistical analysis

- ✓ Calculating basic statistics to understand the data
- ✓ Dividing clients into 2 groups and analyse their statistics using pivot tables

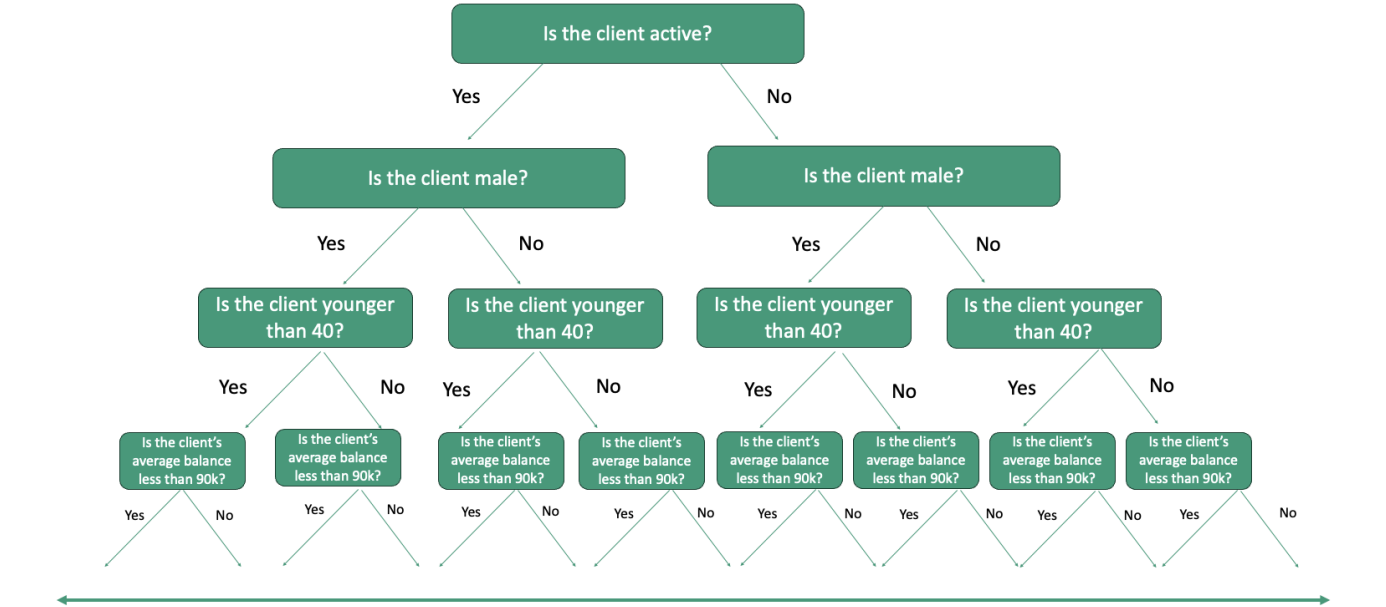
Data mining algorithm

- ✓ Determine the leading factors that contribute to client loss
- ✓ Analyse which factors have which impact and build a decision tree

Documenting the analysis

- ✓ Documenting the analysis in Excel

Decision tree



Additional sources

Detailed data analysis:

Documentation:

Do you have any questions?

alexandra.borschke@gmail.com