Kyle Stanford

Data Analyst

Here are some of the **projects** I have worked on:

GameCo
Gaming Company

Influenza Forecasting
Public Health Sector – U.S.

Rockbuster Stealth
Video Rental Company

Instacart Basket Analysis
Online Grocery Shop

Pig E. Bank
Finance Industry

Lung Cancer Survival
Public Health Sector - EU

These are the **tools** I have used:

Excel

PowerPoint

Tableau

SQL

PostgreSQL

GitHub

Python

Juptyer

GameCo

Gaming Company

Tools Used

Context

GameCo is a new video game company with markets across the globe. They want to understand market trends to inform the development and marketing of new games.

Goal

Analyse regional sales trends to support GameCo in making data-driven decisions.

Data (click here for raw data)

Historical sales data of video games that have sold over 10,000 copies between 1980 and 2020 – from VGChartz

Technical Skills

- Data Cleaning
- Pivot Tables
- Data Grouping and Summarising
- Descriptive Analysis
- Visualisations in Excel & PowerPoint
- Storytelling with Data

Approach and Process

1. Data Preparation

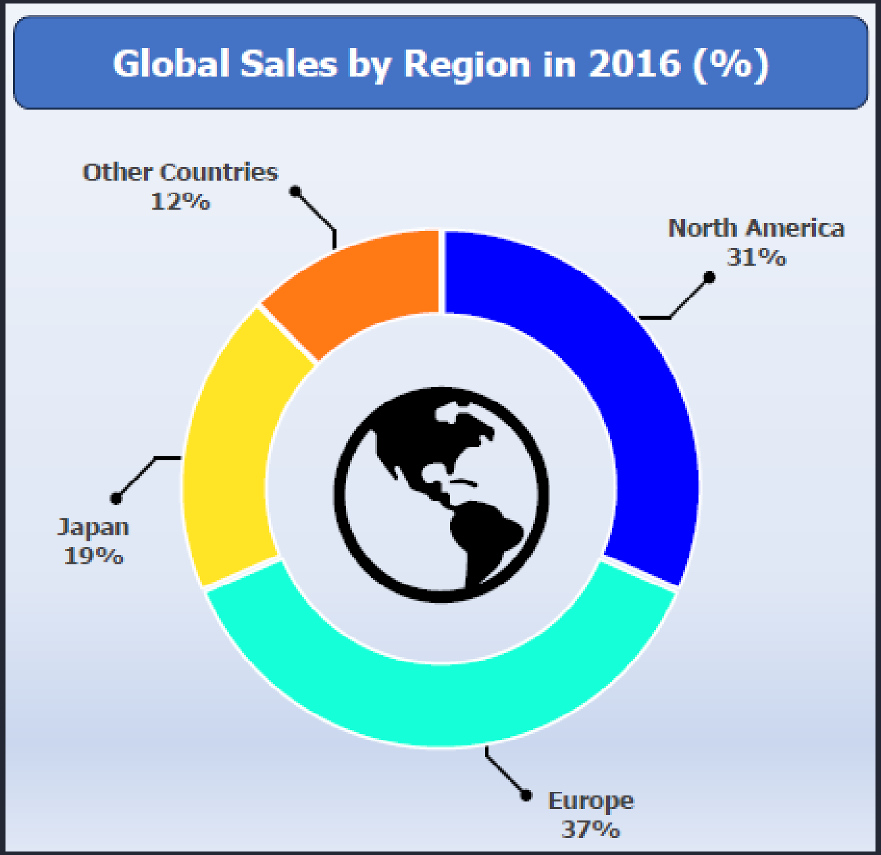
Conducted an exploratory analysis to find and address inconsistencies, duplicates, and missing values.

2. Data Analysis

Utilised pivot tables to group and summarise data by region, genre, and publisher. Filtered data to focus on the last 10 years of consistent data.

3. Visualisation and Presentation

Created doughnut charts, line charts, and 100% stacked column charts to visualise market trends.



Doughnut chart showing distribution of video game sales in 2016

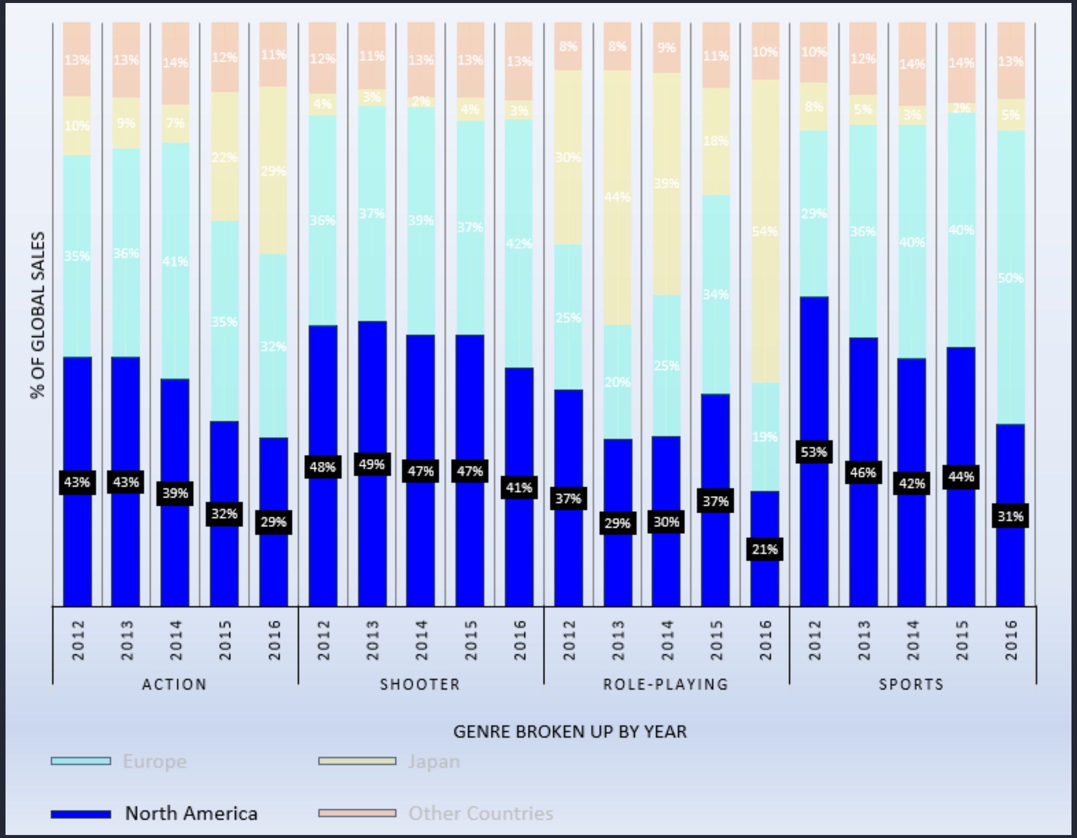
Challenges and Solutions

Complex Analysis

Used pivot tables and calculated fields to answer business questions requiring complex data groupings.

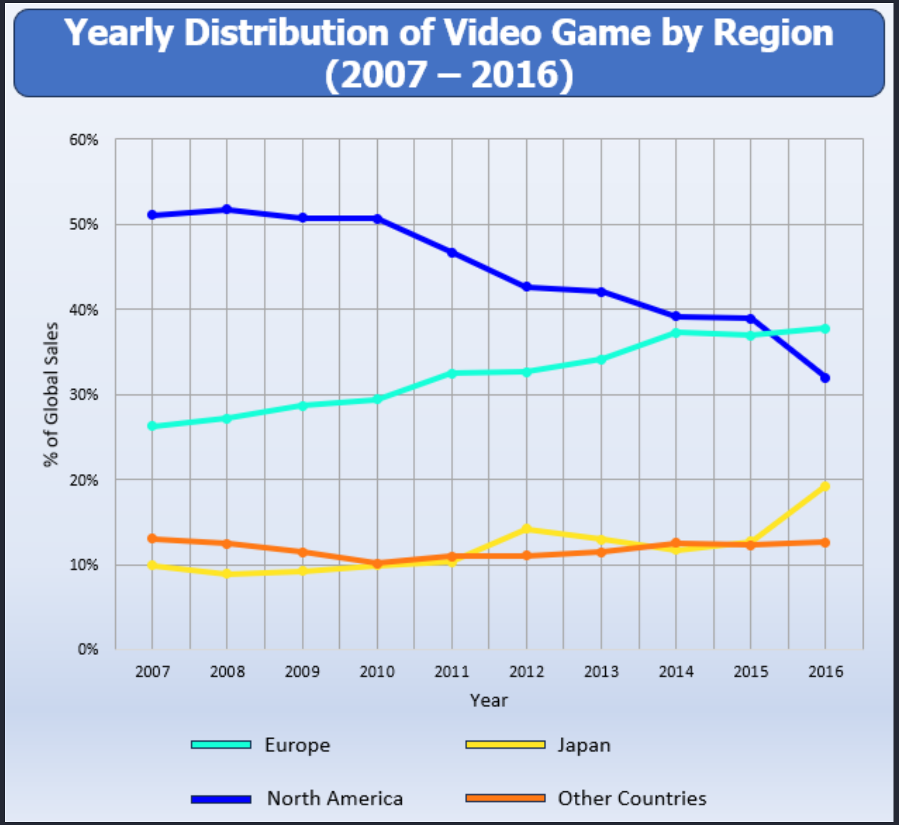
Translating Data to Insights

Utilised advanced formatting options in MS PPT to construct clear data visualisations and communicate findings to stakeholders.



100% Stacked Column Chart showing global sales distribution by genre from 2012 to 2016. Formatted to focus on downward trend in North American market across all genres.

Results and Deliverables



This line chart shows general shifts in market dynamics between North America, Europe, Japan, and other countries over the last decade's worth of full data

Main Findings

- **Europe's** market share overall raised steadily but was most notable in the "shooter" and "sport" genres.
- Increased sales of "role-playing" games contributed to a successful year in 2016 in **Japan**.
- Sales in **North America** have continued to decline. This was consistent in all genres.

Deliverables

1. **Excel Report:** containing cleaned data, pivot tables, and charts.
2. **Final Presentation:** summarising key findings and recommendations to GameCo.
3. **Project Reflections Document:** outlining analytical processes which lead to each insight.

Influenza Forecasting

Public Health Sector – U.S.

Tools Used

Context

A medical staffing agency, responsible for providing temporary workers to hospitals on an as-needed basis, requires assistance in planning for influenza season.

Goal

Analyse influenza trends focusing on vulnerable populations, especially those over 65-years-old, to proactively plan for staffing needs across the country.

Data

1. Influenza Deaths by Geography (Source – CDC)
2. U.S. Census Data (Source – US Census Bureau)

Technical Skills

- Data cleaning, integration, and transformation
- Statistical hypothesis testing
- Visual analysis
- Forecasting in Tableau
- Storyboards in Tableau
- Presenting results

Approach and Process

Data Analysis

1. **Hypothesis Testing** – determined if the mortality rate for older individuals (65+) was higher than other ages.
2. **Visual analysis** – used Tableau Public to identify influenza seasonality and distributions of vulnerable populations across the United States.

Data Preparation

1. Cleansed both datasets of inconsistencies, duplicate values, and addressed missing values.
2. Integrated datasets by using **VLOOKUP.**
3. Grouped age ranges and normalised influenza mortality rates for data analysis.

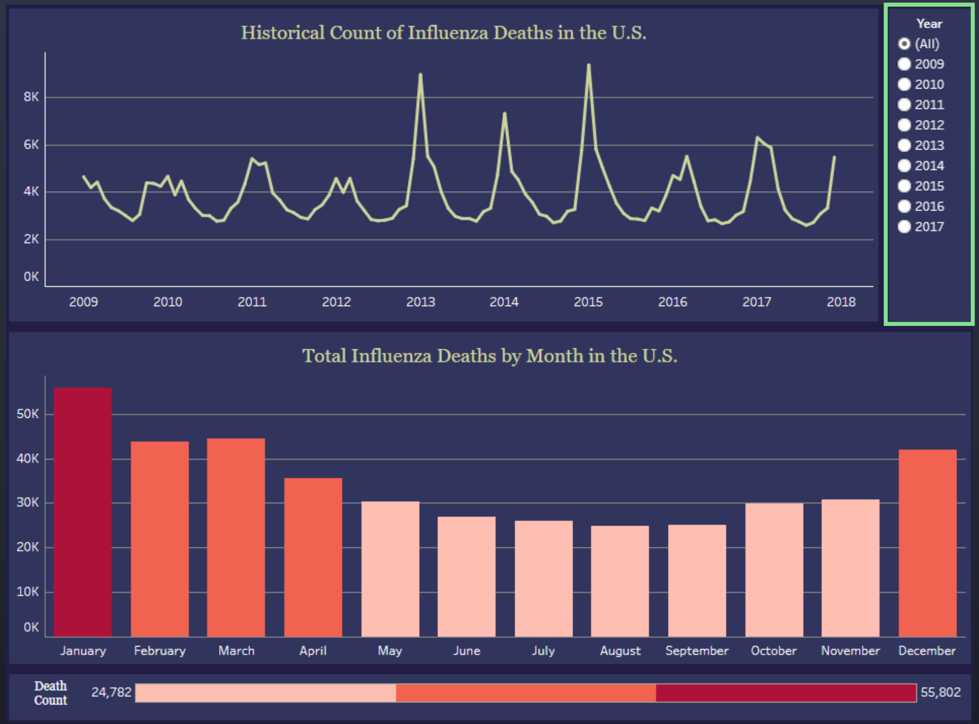
Visualisations and Presentation

1. Created various charts including **combination maps** in Tableau Public to show findings.
2. Presented Tableau Storyboard along with recommendations to stakeholders.

Challenges and Solutions

Interactive Data Presentation

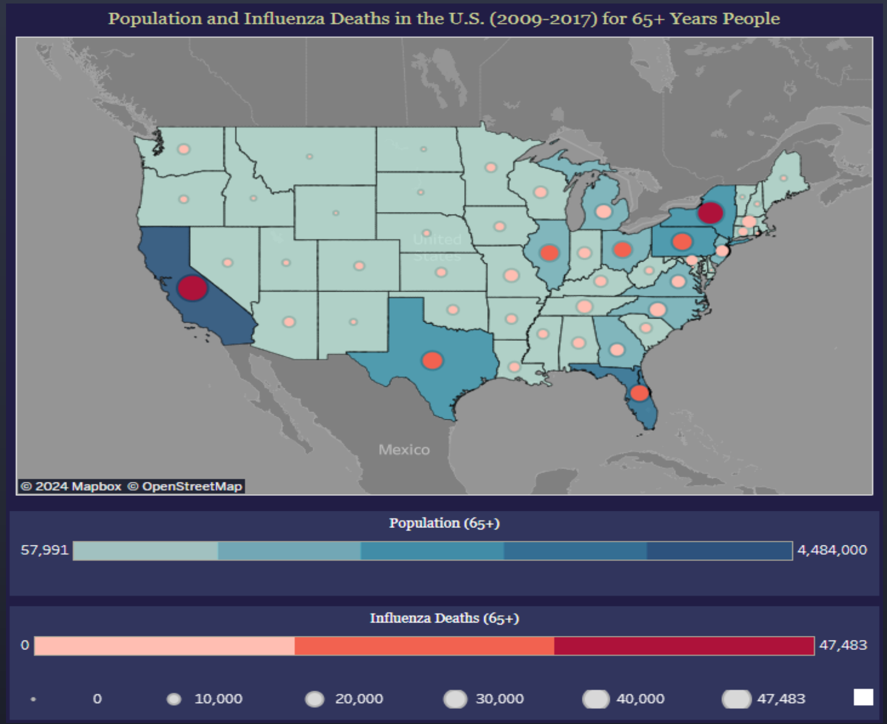
To highlight shifts and consistencies of influenza seasonality over many years, interactive filters were added to the Tableau Storyboard.



Graphs illustrating influenza seasonality with Year filter

Multivariable Spatial Analysis

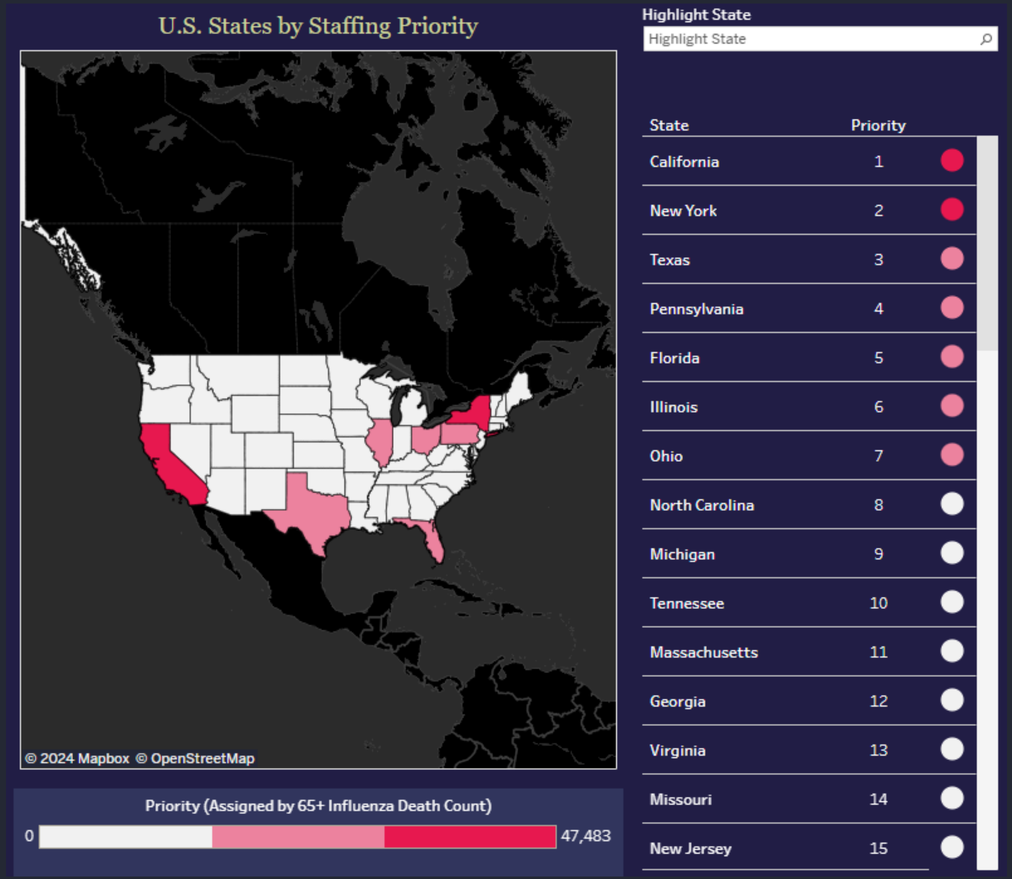
To compare each states' total vulnerable population to their vulnerable population mortality count, combination maps in Tableau were utilised.



Population and Influenza Deaths in the U.S. (2009-2017) for 65+ Years People
Mexico
© 2024 Mapbox © OpenStreetMap
Population (65+)
57,991
4,484,000
Influenza Deaths (65+)
0
47,483
0
10,000
20,000
30,000
40,000
47,483

Combination map showing distribution of older citizen population (blue) and older citizen influenza deaths (red)

Results and Deliverables



U.S. States by Staffing Priority

Highlight State

Highlight State

State	Priority	
California	1	●
New York	2	●
Texas	3	●
Pennsylvania	4	●
Florida	5	●
Illinois	6	●
Ohio	7	●
North Carolina	8	○
Michigan	9	○
Tennessee	10	○
Massachusetts	11	○
Georgia	12	○
Virginia	13	○
Missouri	14	○
New Jersey	15	○

© 2024 Mapbox © OpenStreetMap

Priority (Assigned by 65+ Influenza Death Count)

Choropleth map showing states by staffing priority, accompanied with a scrollable list and priority rankings for clarity. A highlight field has also been added for ease of use.

Results

- Categorised states by urgency of staffing needs.
- Identified peak staffing needs will occur from December to March (influenza season).
- Recommendation for further analysis regarding hospital and clinic staff-patient ratios.

Deliverables

1. **Interim Report:** project progress and findings, including data limitations, descriptive and statistical analysis, and next steps.
2. **Tableau Storyboard:** interactive storyboard containing visualisations and recommendations.

   Link to Tableau storyboard
3. **Video Presentation:** Screencast walkthrough of storyboard with explanations of project analysis and insights.

   Link to video presentation

Rockbuster Stealth

Video Rental Company

Context

Rockbuster Stealth LLC is a movie rental business with stores around the world. It's trying to compete with popular streaming services by launching an online video rental service.

Goal

Assist Rockbuster in launching their online video rental service by using SQL to analyse Rockbuster's data and answer ad-hoc business questions.

Tools Used

Data (click here for zip folder)

Rockbuster Relational Database containing 15 connected tables.

Technical Skills

- Database querying:
  - ➢ Filtering
  - ➢ Joins
  - ➢ Common Table Expressions
  - ➢ Summarising
  - ➢ Subqueries
- Database cleaning
- Data profiling and creating a Data Dictionary
- Data visualisation in Tableau Public

Approach and Process

Data Understanding

Data dictionary and **entity relationship diagram** created to best understand Rockbuster's database.

Data Preparation

Consistency checks made to ensure no duplicates, no missing values, and uniformity across multiple tables.

Data Analysis

Various **SQL queries** made to answer simple and complex business questions, e.g., *"Which movies contributed most/least to revenue gain"?*

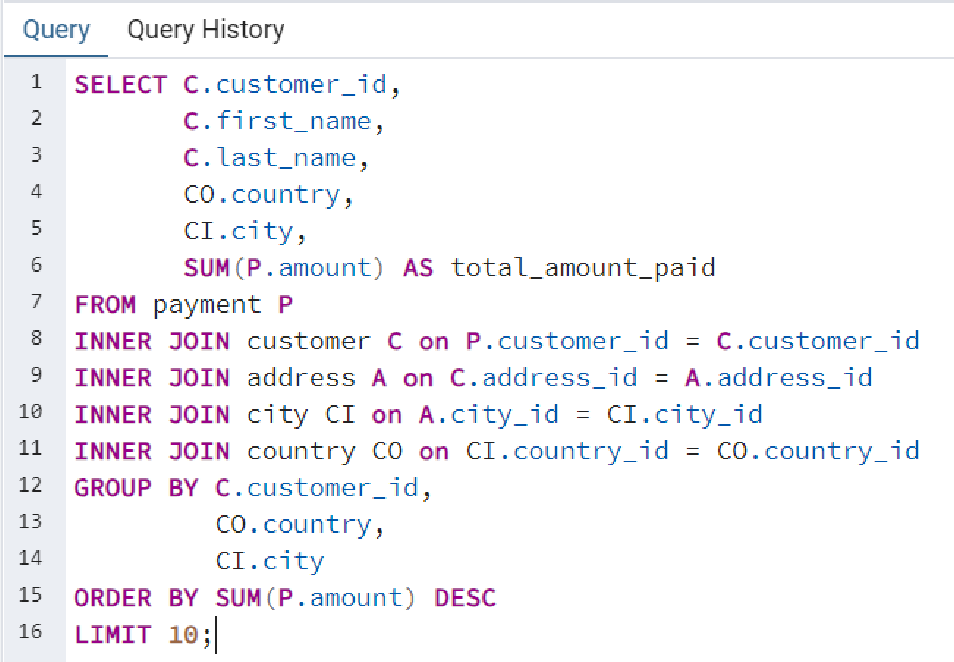
Visualisations and Presentation

1. Developed Tableau dashboards to visualise insights from SQL queries.
2. Created a PowerPoint presentation with business recommendations for stakeholders.

Challenges and Solutions

Complex Queries Across Tables

Data required to answer business questions often existed across multiple tables within the database. JOINS, GROUP BY, LIMIT, and other clauses used to query this data.

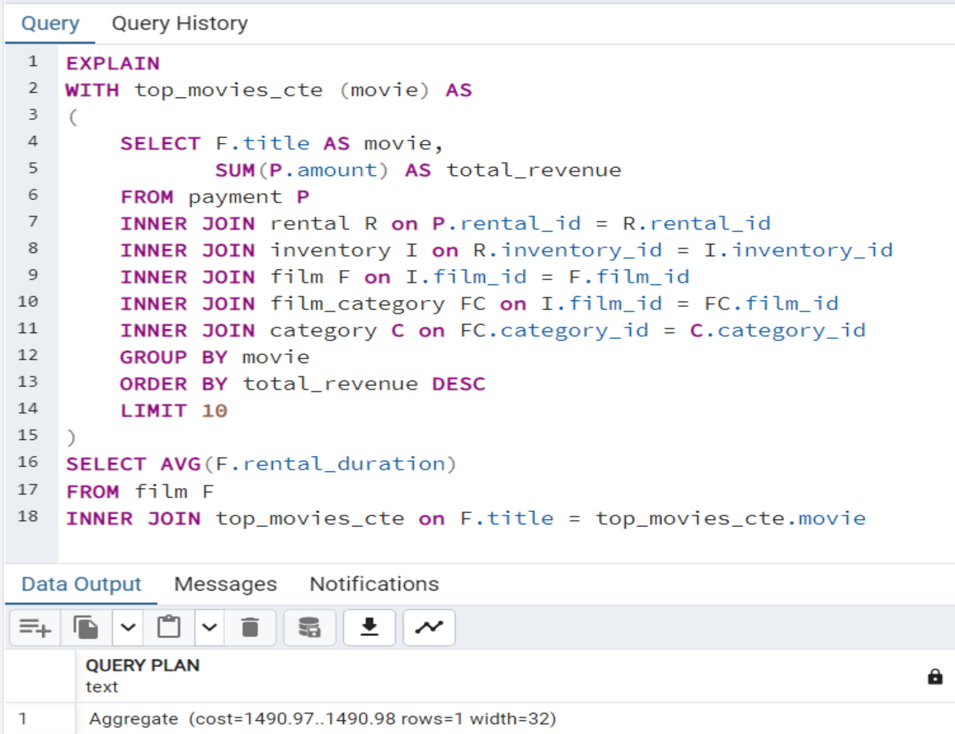


```sql
SELECT C.customer_id,
       C.first_name,
       C.last_name,
       CO.country,
       CI.city,
       SUM(P.amount) AS total_amount_paid
FROM payment P
INNER JOIN customer C on P.customer_id = C.customer_id
INNER JOIN address A on C.address_id = A.address_id
INNER JOIN city CI on A.city_id = CI.city_id
INNER JOIN country CO on CI.country_id = CO.country_id
GROUP BY C.customer_id,
         CO.country,
         CI.city
ORDER BY SUM(P.amount) DESC
LIMIT 10;
```

SQL query designed to answer the question “*Where are customers with a high lifetime value based*”. Result returns top 10 customers based on lifetime spending along with their country and city of residence.

Query Optimization

Optimized SQL queries for efficiency and accuracy. This was done by using the EXPLAIN clause to evaluate query costs.



```sql
EXPLAIN
WITH top_movies_cte (movie) AS
(
    SELECT F.title AS movie,
           SUM(P.amount) AS total_revenue
    FROM payment P
    INNER JOIN rental R on P.rental_id = R.rental_id
    INNER JOIN inventory I on R.inventory_id = I.inventory_id
    INNER JOIN film F on I.film_id = F.film_id
    INNER JOIN film_category FC on I.film_id = FC.film_id
    INNER JOIN category C on FC.category_id = C.category_id
    GROUP BY movie
    ORDER BY total_revenue DESC
    LIMIT 10
)
SELECT AVG(F.rental_duration)
FROM film F
INNER JOIN top_movies_cte on F.title = top_movies_cte.movie
```

```
   QUERY PLAN
   text
1  Aggregate  (cost=1490.97..1490.98 rows=1 width=32)
```

EXPLAIN query that returns the cost of a query that finds the average rental duration of Rockbuster’s top 10 movies.

Results and Deliverables

GitHub Project Repository Link

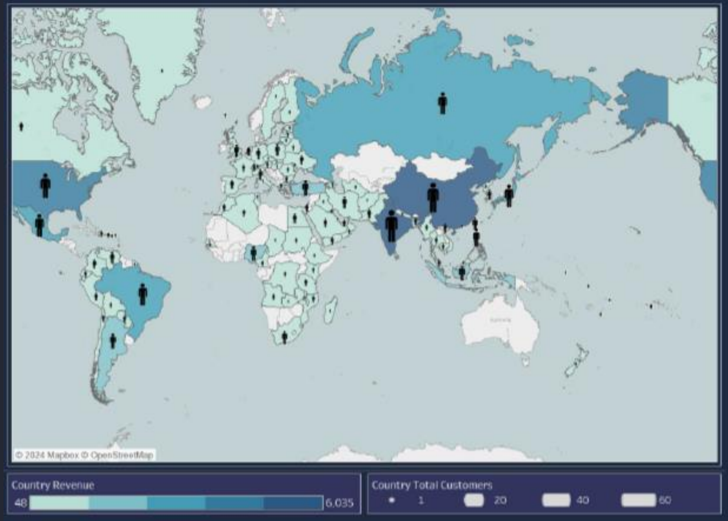
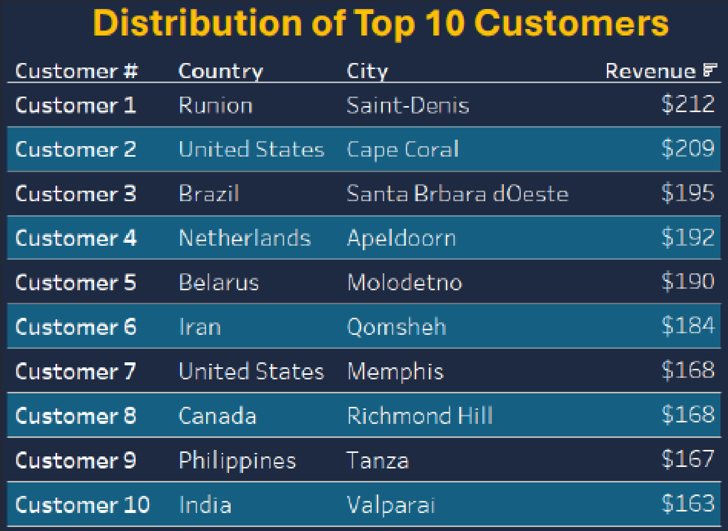


Chart showing revenue and total customers in each country



Distribution of Top 10 Customers

Customer #	Country	City	Revenue ⊽
Customer 1	Runion	Saint-Denis	$212
Customer 2	United States	Cape Coral	$209
Customer 3	Brazil	Santa Brbara dOeste	$195
Customer 4	Netherlands	Apeldoorn	$192
Customer 5	Belarus	Molodetno	$190
Customer 6	Iran	Qomsheh	$184
Customer 7	United States	Memphis	$168
Customer 8	Canada	Richmond Hill	$168
Customer 9	Philippines	Tanza	$167
Customer 10	India	Valparai	$163

Table showing top 10 customers and their locations determined by total revenue.

Results

- Identified Rockbuster's top and bottom performing movies by revenue (including unrented movies)
- Mapped global customer and revenue distribution.
- Determined high-value customers don't necessarily reside in countries with the most customers.

Deliverables

1. **Data Dictionary:** comprehensive document detailing the structure and relationship of Rockbuster's database.
2. **SQL Queries File:** Excel file storing SQL queries and their outputs.
3. **Final Presentation:** PowerPoint summarizing key findings and recommendations for Rockbuster.
4. **Tableau Dashboards:** visualisations used in presentation.

Link to Tableau Dashboards

Instacart Basket Analysis

Online Grocery Shop

Context

Instacart is an online grocery store operating through an app. They want to target different customers with applicable marketing campaigns.

Goal

Perform an initial data and exploratory analysis of Instacart's data to derive insights for better customer segmentation based on provided criteria.

Tools Used

Data

Instacart's customer data, product data, and orders data.

- The final dataset contained 32,399,732 rows

(Instacart is a real company, but the data used was fabricated by CareerFoundry for this project)

Technical Skills

- Python:
  - ➢ Data cleaning
  - ➢ Wrangling and merging
  - ➢ Deriving variables
  - ➢ Aggregations
  - ➢ Data visualisation
- Population flows
- Reporting in Excel

Approach and Process

Data Preparation

- Loaded and cleaned datasets using **Pandas.**
- Merged datasets to make a comprehensive dataframe.

Data Analysis

- Conducted descriptive statistics with **NumPy** and **Pandas.**
- Segmented customers based on purchasing habits and demographic information.
- Constructed visualisations using **Matplotlib** and **Seaborn** to answer business questions.

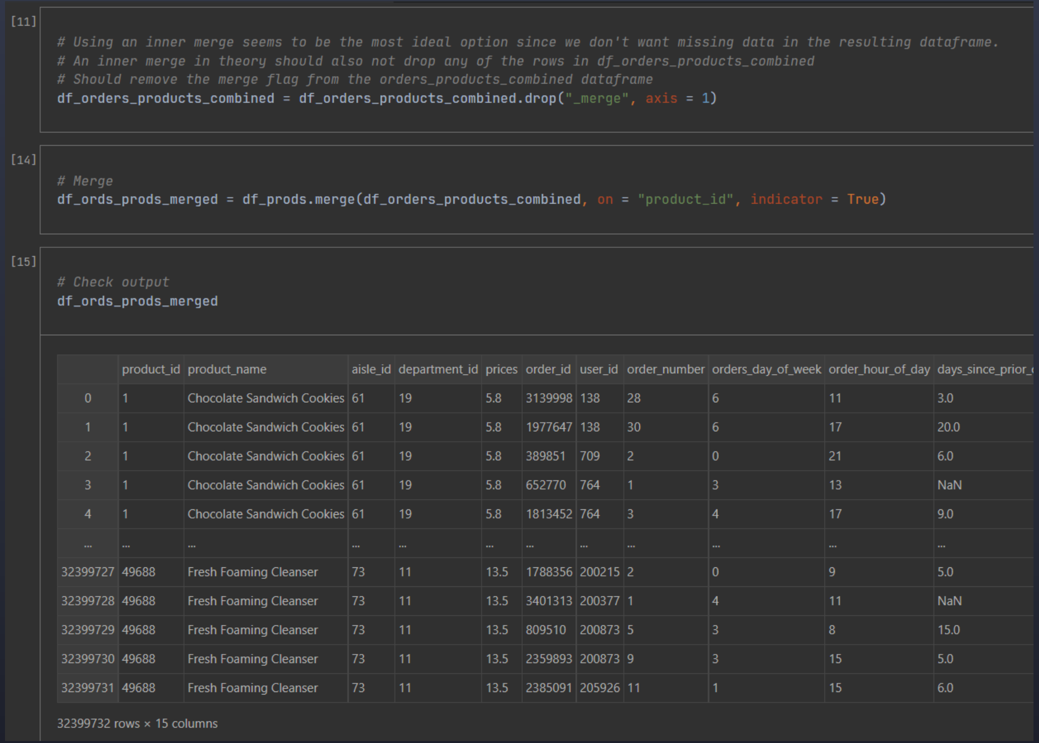
Reporting

- **Population flow** refined.
- Data processes documented.
- Visualisations selected and recommendations made.

Challenges and Solutions

Data Integration

Managed the integration of multiple datasets ensuring data integrity and consistency.



```
[11]
# Using an inner merge seems to be the most ideal option since we don't want missing data in the resulting dataframe.
# An inner merge in theory should also not drop any of the rows in df_orders_products_combined
# Should remove the merge flag from the orders_products_combined dataframe
df_orders_products_combined = df_orders_products_combined.drop("_merge", axis = 1)
```

```
[14]
# Merge
df_ords_prods_merged = df_prods.merge(df_orders_products_combined, on = "product_id", indicator = True)
```

```
[15]
# Check output
df_ords_prods_merged
```

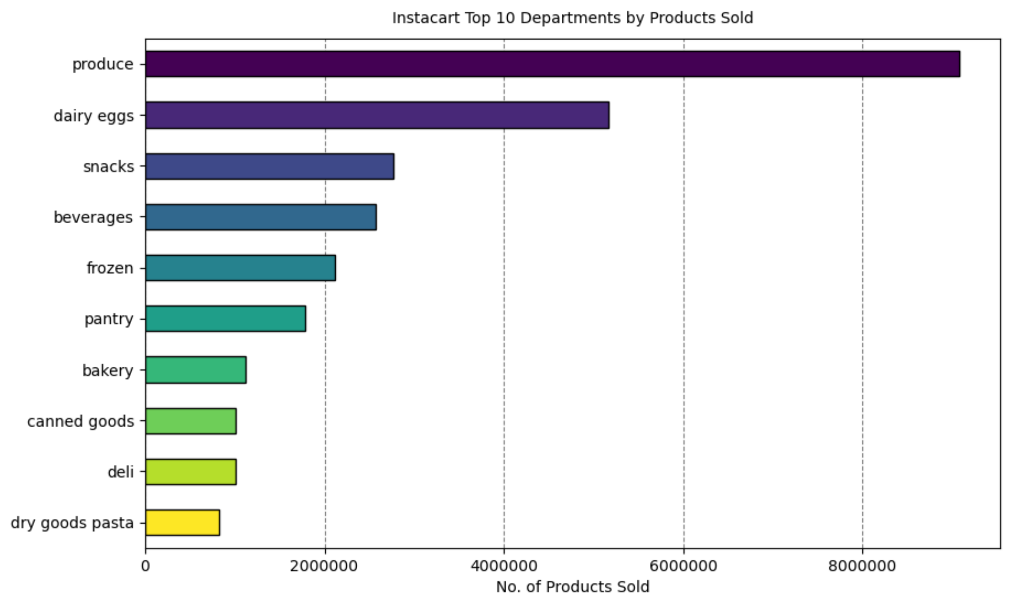
	product_id	product_name	aisle_id	department_id	prices	order_id	user_id	order_number	orders_day_of_week	order_hour_of_day	days_since_prior_
0	1	Chocolate Sandwich Cookies	61	19	5.8	3139998	138	28	6	11	3.0
1	1	Chocolate Sandwich Cookies	61	19	5.8	1977647	138	30	6	17	20.0
2	1	Chocolate Sandwich Cookies	61	19	5.8	389851	709	2	0	21	6.0
3	1	Chocolate Sandwich Cookies	61	19	5.8	652770	764	1	3	13	NaN
4	1	Chocolate Sandwich Cookies	61	19	5.8	1813452	764	3	4	17	9.0
...	...	...	...	...	...	...	...	...	...	...	...
32399727	49688	Fresh Foaming Cleanser	73	11	13.5	1788356	200215	2	0	9	5.0
32399728	49688	Fresh Foaming Cleanser	73	11	13.5	3401313	200377	1	4	11	NaN
32399729	49688	Fresh Foaming Cleanser	73	11	13.5	809510	200873	5	3	8	15.0
32399730	49688	Fresh Foaming Cleanser	73	11	13.5	2359893	200873	9	3	15	5.0
32399731	49688	Fresh Foaming Cleanser	73	11	13.5	2385091	205926	11	1	15	6.0

32399732 rows × 15 columns

Code snippet showing inner merge of 2 datasets with checks and comments.

Complex Visualisation Coding

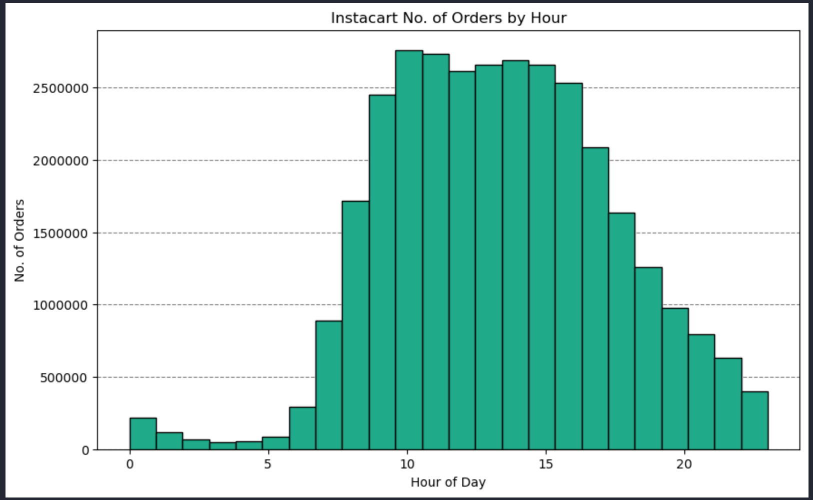
Created concise and impactful visualisations by leveraging ChatGPT to edit my previous code.



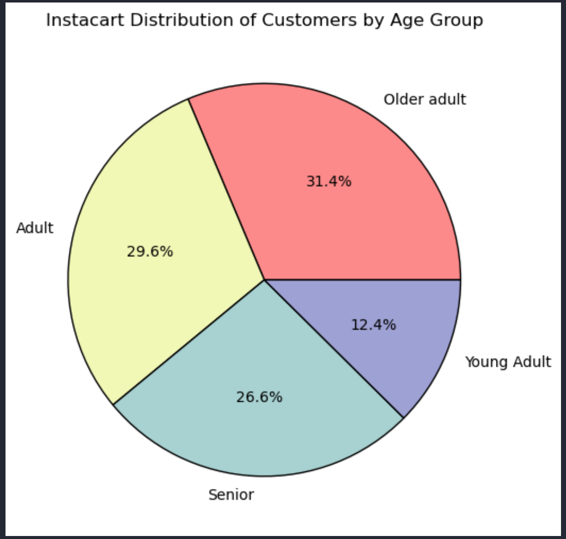
Horizontal bar chart showing Instacart's top 10 departments based on total product sales. ChatGPT was used to modify some visual aesthetics and limit the visualisation to only the top 10 departments.

Results and Deliverables

GitHub Project Repository Link



Histogram showing the number of products ordered across each hour of the day (24 bins used)



Pie Chart showing the distribution of customers based on assigned age groupings.

Results

- Identified peak operational hours and days of the week to optimize ad scheduling.
- Segmented customers based on loyalty, region, age, and family status to help direct marketing efforts.
- Identified most of Instacart's customer base consists of new customers, who may be enticed to return with loyalty programs.

Deliverables

1. **Jupyter Notebooks:** All python code used throughout project, documenting data cleaning, analysis, and visualisations.
2. **Excel Report:** File containing population flow, data cleaning and wrangling steps, key findings, Python visualisations, and recommendations for Instacart.

Pig E. Bank

Finance Industry

Context

Pig E. Bank is a European financial institution which is seeking to understand factors contributing to customer attrition.

Goal

Identify key factors influencing customer churn and develop a model to assess the likelihood a customer will leave Pig E. Bank.

Tools Used

Data

Pig E. Bank customer data

- Includes account information, demographic data, and other variables.

Technical Skills

- Data Mining
  - ➢ Data Preparation
  - ➢ Analysis (Pivot Tables)
  - ➢ Classification Modelling (Decision Trees)
- Understanding of Data Ethics

Approach and Process

Data Preparation

- Cleansed data from missing values and inconsistent formatting.
- Removed **Personally Identifiable Information.**
- Separated data into “exited customers” table and “current customers” table.

Data Analysis

- Created **pivot tables** for all tables.
- Compiled pivot tables into large comparison table.
- Identified and recorded variables associated with high customer attrition.

Modelling

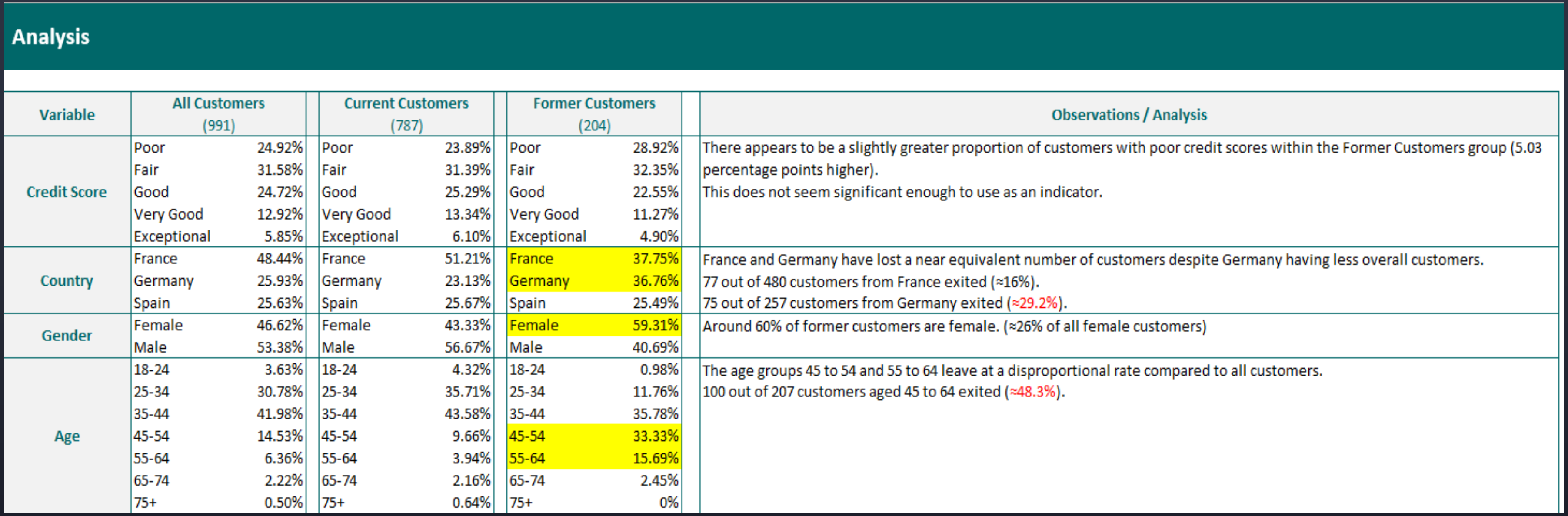
- Selected and ranked top 4 variables associated with customer attrition.
- Created decision tree based on these variables.

Challenges and Solutions

Deciding Variable Importance for the Model

Variables were considered and ranked for implementation in the decision tree model based on relative attrition rates and overall customers left.

This ensured the model would not unfairly favour categories with more total customers (such as binary categories like gender) over more segmented categories with higher attrition rates (such as age groupings).

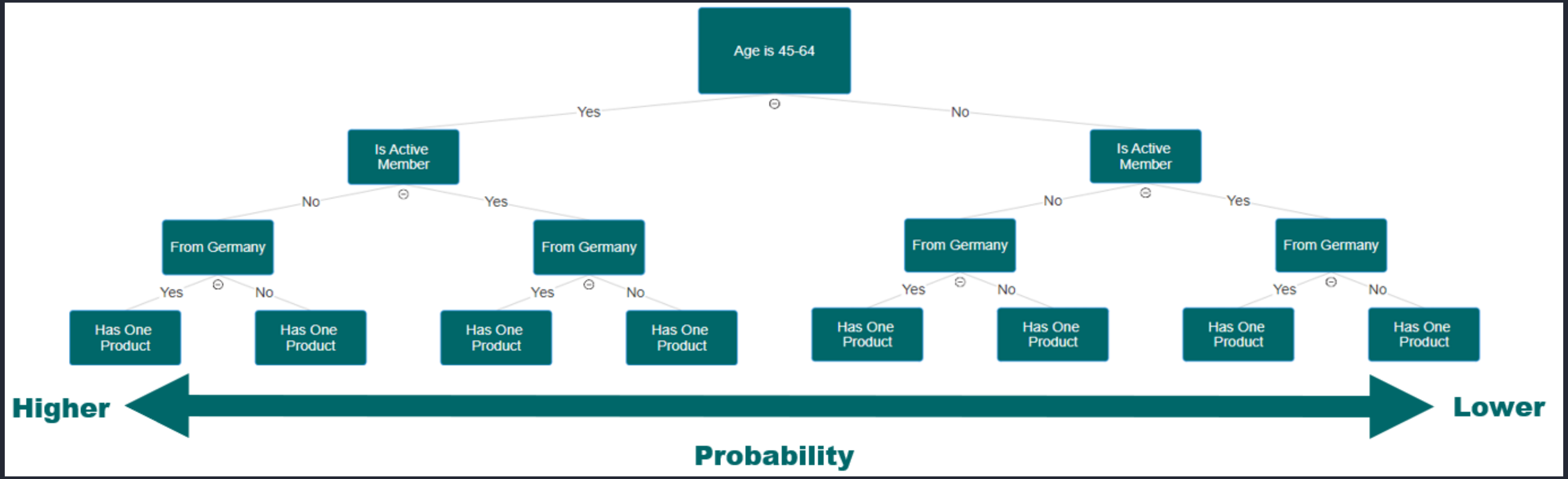


Analysis

Variable	All Customers		Current Customers		Former Customers		Observations / Analysis
	(991)		(787)		(204)		
Credit Score	Poor	24.92%	Poor	23.89%	Poor	28.92%	There appears to be a slightly greater proportion of customers with poor credit scores within the Former Customers group (5.03 percentage points higher). This does not seem significant enough to use as an indicator.
	Fair	31.58%	Fair	31.39%	Fair	32.35%	
	Good	24.72%	Good	25.29%	Good	22.55%	
	Very Good	12.92%	Very Good	13.34%	Very Good	11.27%	
	Exceptional	5.85%	Exceptional	6.10%	Exceptional	4.90%	
Country	France	48.44%	France	51.21%	France	37.75%	France and Germany have lost a near equivalent number of customers despite Germany having less overall customers.
	Germany	25.93%	Germany	23.13%	Germany	36.76%	77 out of 480 customers from France exited (≈16%).
	Spain	25.63%	Spain	25.67%	Spain	25.49%	75 out of 257 customers from Germany exited (≈29.2%).
Gender	Female	46.62%	Female	43.33%	Female	59.31%	Around 60% of former customers are female. (≈26% of all female customers)
	Male	53.38%	Male	56.67%	Male	40.69%	
Age	18-24	3.63%	18-24	4.32%	18-24	0.98%	The age groups 45 to 54 and 55 to 64 leave at a disproportional rate compared to all customers.
	25-34	30.78%	25-34	35.71%	25-34	11.76%	100 out of 207 customers aged 45 to 64 exited (≈48.3%).
	35-44	41.98%	35-44	43.58%	35-44	35.78%	
	45-54	14.53%	45-54	9.66%	45-54	33.33%	
	55-64	6.36%	55-64	3.94%	55-64	15.69%	
	65-74	2.22%	65-74	2.16%	65-74	2.45%	
	75+	0.50%	75+	0.64%	75+	0%	

Analysis containing multiple pivot table results – variables of interest highlighted, and relative attrition rates calculated (in red)

Results and Deliverables



Decision Tree model created to determine the risk that a customer will exit the bank.

Results

- Identified top risk factors that a customer will exit the bank
- Produced decision tree model

Deliverables

1. **Excel Report:** Comprehensive Excel report containing raw data, data cleaning practices, pivot tables, descriptive statistics, analysis and model.

Lung Cancer Survival

Public Health Sector - EU

Tools Used

Context

Lung cancer is the leading cause of cancer death in men and second in women. Predictive models can help determine patient chance of survival.

Goal

Analyse health indicators, demographic data, and treatment-related variables of lung cancer patients to determine which factors increase survival rates.

Data

Lung Cancer Mortality Dataset – Kaggle

Custom shapefile containing EU countries – Vector Maps

Country Development Indicators – World Bank

Technical Skills

- Sourcing Open Data
- Correlation Heatmaps and Scatterplots
- Geospatial Analysis with JSON files
- Linear Regression Analysis in Python
- Cluster Analysis (k-means)
- Tableau Dashboard Creation

Guiding Questions

Who is this for?

This project could be utilised by public health and research agencies across EU countries.

Why is it being built?

The project is being built to explore factors that may be used to assess lung cancer patients' chance of survival. It may also serve as a jumping-off point for further research pending its results.

Where will it be hosted?

The project will be hosted as a storyboard on Tableau Public

What will it consist of?

The project will start with the aim to analyse patient survival rates based on demographic factors, health indicators, and treatment received. This may shift pending the results of the various analyses.

When will it be used?

In theory, the results of the project could be used when planning further research into lung cancer mortality. The results may highlight areas worth further exploration.

Approach and Process

1. Data Sourcing & Preparation

- Sourced main dataset from **Kaggle**.
- Checked and addressed missing values, duplicates and outliers.
- Sourced supplementary data on EU countries from **World Bank** and reformatted in Excel.
- Merged World Bank data with main dataset.

2. Exploratory Data Analysis

- Created correlation heatmap, histograms, and pair plots using **SciPy** and **Matplotlib**.
- Liaised with supervisor/mentor about potential issues within the data.
- Focused efforts on weakly correlated treatment-specific variables.

3. Linear Regression

- Conducted a linear regression in Python using **Scikit-Learn** library.
- As the model explained less than 1% (r-squared < 0.01) of the variance in the data, a linear model was deemed unsuitable.

Approach and Process

4. Cluster Analysis

- Used the **k-means algorithm** to find clusters in the data with the aim to find unexpected patterns.
- This resulted in patients being grouped in two clusters determined mostly by the only two strongly correlated variables in the data.
- Adding more clusters was trialled unsuccessfully.

5. Geospatial Analysis

- Conducted geospatial analysis with custom **shapefile** of EU countries.
- Used **Folium** library in Python to analyse lung cancer distribution, survival rates, and treatment types by country.

6. Further Analysis & Presentation

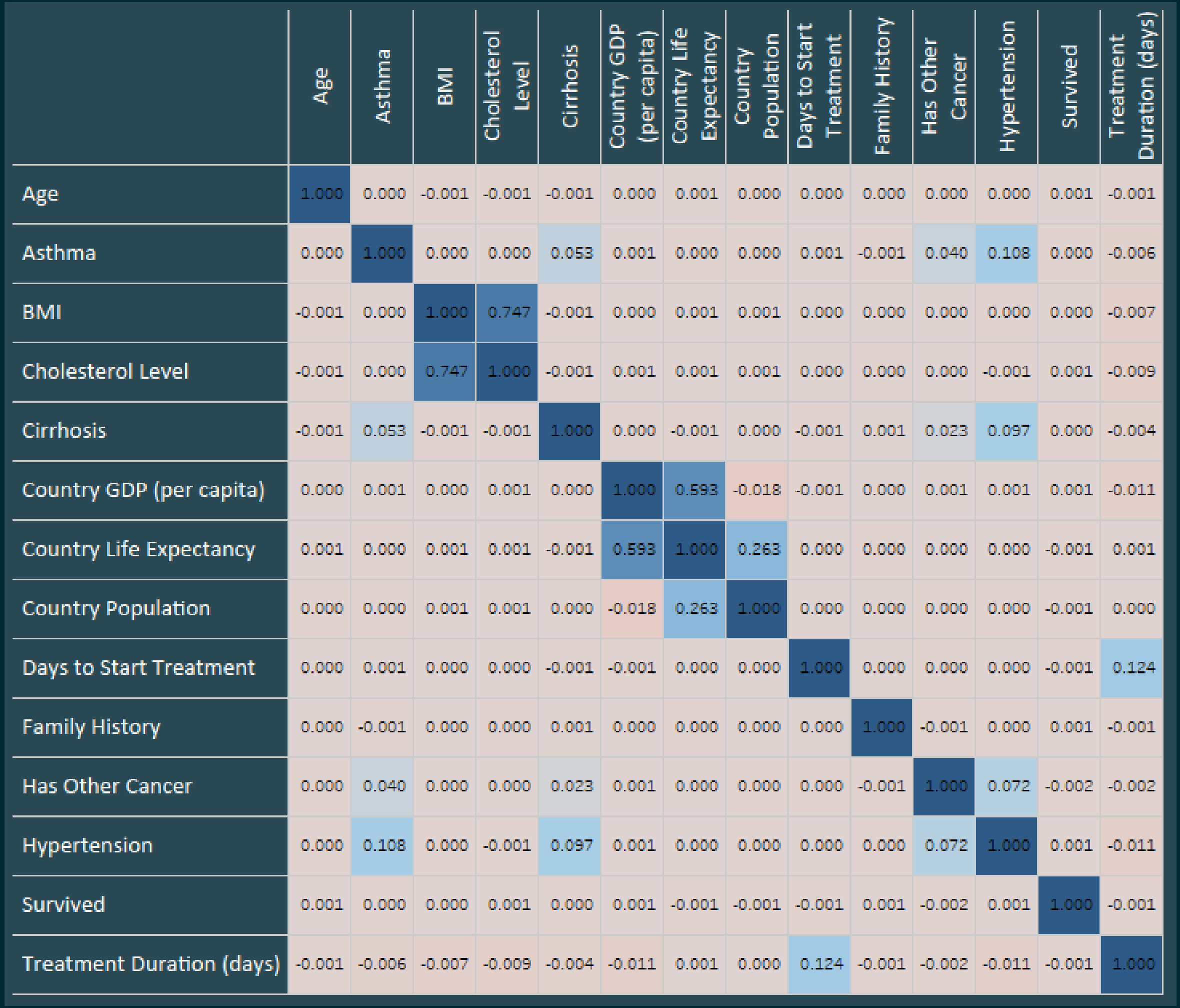
- Further explored relationships between survival rates and health indicators and smoking status.
- Delved further into linear regression by subdividing data by cancer stage and treatment received.
- Document and presented all analyses in **Tableau Storyboard**.

Challenges and Solutions

Working with Poorly Correlated Variables

A correlation heatmap showed that patient survival was poorly correlated with all other variable. After liaising with a mentor, it was decided that:

- Proceeding analyses could focus on relationships between treatment related variables.
- Data would be wrangled in later steps to obtain and compare survival rates.



	Age	Asthma	BMI	Cholesterol Level	Cirrhosis	Country GDP (per capita)	Country Life Expectancy	Country Population	Days to Start Treatment	Family History	Has Other Cancer	Hypertension	Survived	Treatment Duration (days)
Age	1.000	0.000	-0.001	-0.001	-0.001	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.001	-0.001
Asthma	0.000	1.000	0.000	0.000	0.053	0.001	0.000	0.000	0.001	-0.001	0.040	0.108	0.000	-0.006
BMI	-0.001	0.000	1.000	0.747	-0.001	0.000	0.001	0.001	0.000	0.000	0.000	0.000	0.000	-0.007
Cholesterol Level	-0.001	0.000	0.747	1.000	-0.001	0.001	0.001	0.001	0.000	0.000	0.000	-0.001	0.001	-0.009
Cirrhosis	-0.001	0.053	-0.001	-0.001	1.000	0.000	-0.001	0.000	-0.001	0.001	0.023	0.097	0.000	-0.004
Country GDP (per capita)	0.000	0.001	0.000	0.001	0.000	1.000	0.593	-0.018	-0.001	0.000	0.001	0.001	0.001	-0.011
Country Life Expectancy	0.001	0.000	0.001	0.001	-0.001	0.593	1.000	0.263	0.000	0.000	0.000	0.000	-0.001	0.001
Country Population	0.000	0.000	0.001	0.001	0.000	-0.018	0.263	1.000	0.000	0.000	0.000	0.000	-0.001	0.000
Days to Start Treatment	0.000	0.001	0.000	0.000	-0.001	-0.001	0.000	0.000	1.000	0.000	0.000	0.000	-0.001	0.124
Family History	0.000	-0.001	0.000	0.000	0.001	0.000	0.000	0.000	0.000	1.000	-0.001	0.000	0.001	-0.001
Has Other Cancer	0.000	0.040	0.000	0.000	0.023	0.001	0.000	0.000	0.000	-0.001	1.000	0.072	-0.002	-0.002
Hypertension	0.000	0.108	0.000	-0.001	0.097	0.001	0.000	0.000	0.000	0.000	0.072	1.000	0.001	-0.011
Survived	0.001	0.000	0.000	0.001	0.000	0.001	-0.001	-0.001	-0.001	0.001	-0.002	0.001	1.000	-0.001
Treatment Duration (days)	-0.001	-0.006	-0.007	-0.009	-0.004	-0.011	0.001	0.000	0.124	-0.001	-0.002	-0.011	-0.001	1.000

Correlation heatmap showing “survived” column having no correlation with other variables

Challenges and Solutions

Analyses Yielding Insignificant Result

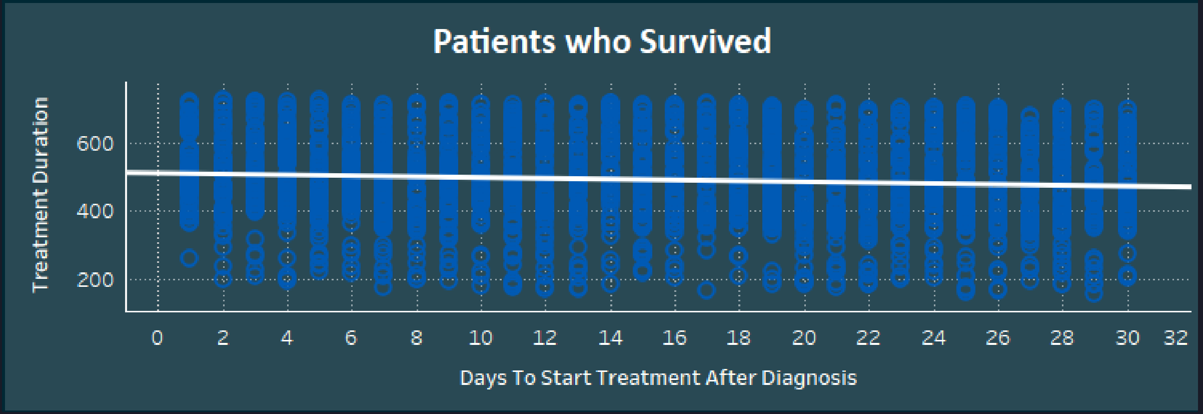
Linear Regression Analysis:

- More than 99% of the data’s variance could not be explained by the model.
- Data was subdivided based on categorical variables aiming to reduce the variance and yield new insights.

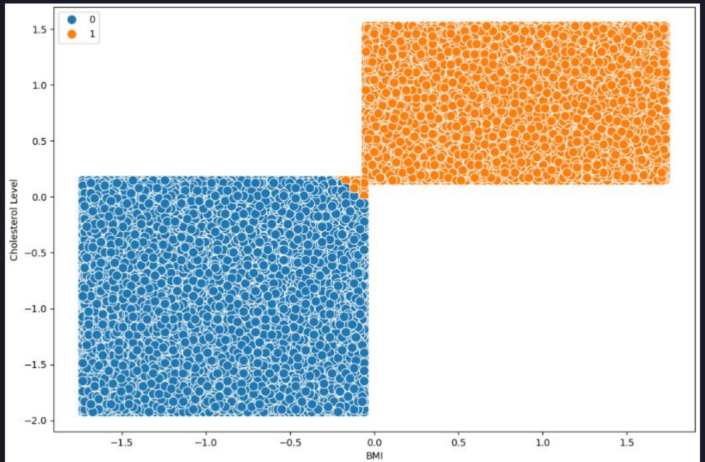
Cluster Analysis:

- Clusters determined by algorithm were heavily determined by only two variables. Analysis on scatterplots yielded insignificant results.
- Additional clusters were added to the algorithm but still yielded insignificant results.

Overall, these approaches were considered unsuitable and alternative analyses were conducted to further the project.



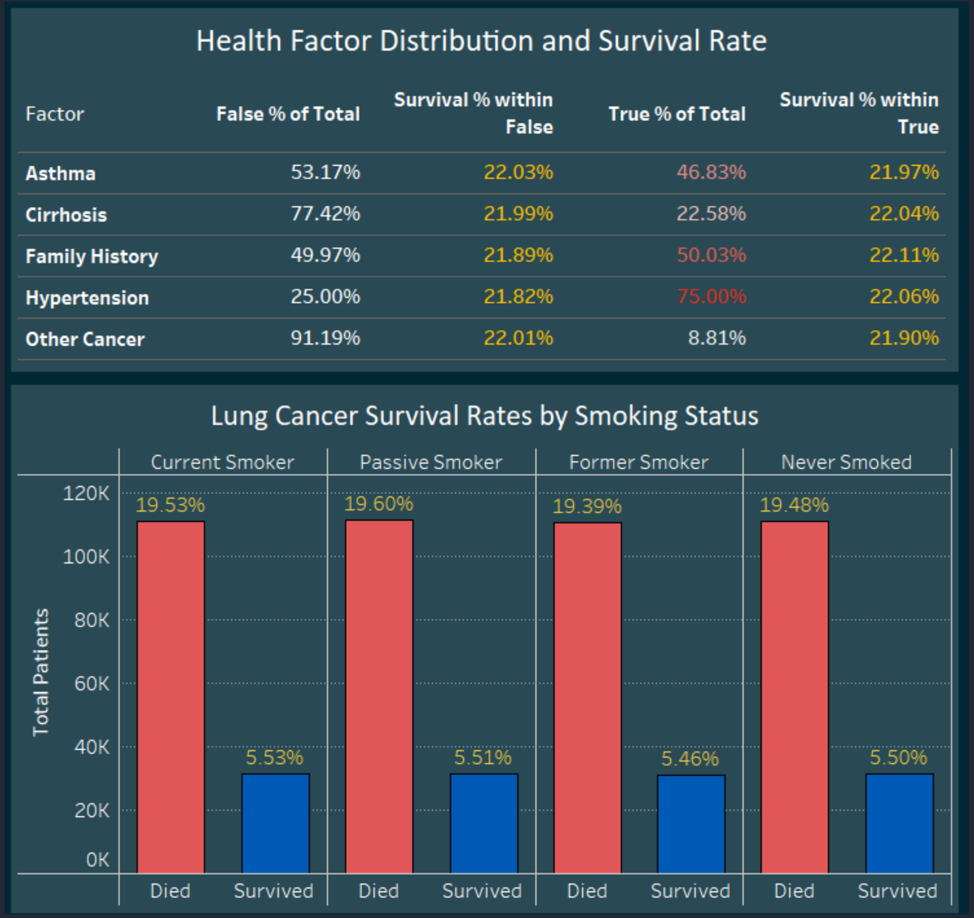
Linear regression on only patients who survived treatment with additional filters (treatment, cancer stage, and smoking status)



Scatterplot showing how clusters were mostly determined by the relationship between patients’ BMI and cholesterol levels.

Results and Deliverables

[GitHub Project Repository Link]



Health Factor Distribution and Survival Rate

Factor	False % of Total	Survival % within False	True % of Total	Survival % within True
Asthma	53.17%	22.03%	46.83%	21.97%
Cirrhosis	77.42%	21.99%	22.58%	22.04%
Family History	49.97%	21.89%	50.03%	22.11%
Hypertension	25.00%	21.82%	75.00%	22.06%
Other Cancer	91.19%	22.01%	8.81%	21.90%

Lung Cancer Survival Rates by Smoking Status

Snippet of Tableau Dashboard analysing distributions and survival rates of patients amongst groups based on health indicators.

Key Findings

- Lung cancer survival did not appear affected by any health indicator, treatment received, or demographic factors.
- By cross checking data with Eurostat publications, lung cancer does seem more prevalent in people with hypertension and asthma.
- Survival rates did not vary significantly across countries in the EU.

Project Limitations:

- The dataset is artificially generated, meaning it may not capture full variability and complexity of real-world data.
- If the data is not based on accurate distribution or contains inherent biases, results may be misleading.
- Models trained on this data have limited transferability to real-world applications.

Deliverables:

1. **Jupyter Notebooks:** All python code used throughout project, documenting data cleaning and all analyses.
2. **Tableau Storyboard:** Presentation of project analyses, results, limitations, and next steps.

(Link to Tableau Storyboard)

Thank you!

Kyle Stanford

Let’s connect: