DANFRADAT

DAN FRANKLIN

DATA ANALYST

Rockbuster Stealth

Movie Rental Company

This project focused on a film rental company, **Rockbuster Stealth**, which is a fictional company that previously had stores around the world but faces stiff competition from the streaming giants.

Project Goal

To help with the launch strategy for the new online video service by conducting an analysis and making recommendations based on the results.

Key Objectives

- Which movies contributed the most/least to revenue gain?
- What was the average rental duration for all videos?
- Which countries are Rockbuster customers based in?
- Where are customers with a high lifetime value based?
- Do sales figures vary between geographic regions?

Datasets

Rockbuster Stealth LLC [data set.]

Tools

SQL

tableau

Skills & Procedures

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

Preparation

● Set up a database environment using the PostgreSQL client GUI tool to begin an analysis

● Analysing keys and indexes of the database

● Extract an entity relationship diagram and create a first draft of a data dictionary

Analysis

● SQL commands, basic CRUD operations

● Writing SQL queries to organise and sort data

● Filtering and ordering data

● Identifying dirty data cleaning it

● Creating a data profile of summary statistics using SQL

● Using SQL to join tables

● Writing subqueries and applying Common Table Expressions

● Create visualizations of SQL results

● Create a presentation of findings

Results & Recommendations

● Convert to a subscription model similar to rivals such as Netflix or Prime. This will generate steady income rather than sporadic.

● Create a reward programme for current customers inviting friends to join to help grow the customer base.

● Focus marketing campaigns on the current main markets initially – India, China and Japan in Asia, and USA and Mexico in North Amercia to gain traction and growth.

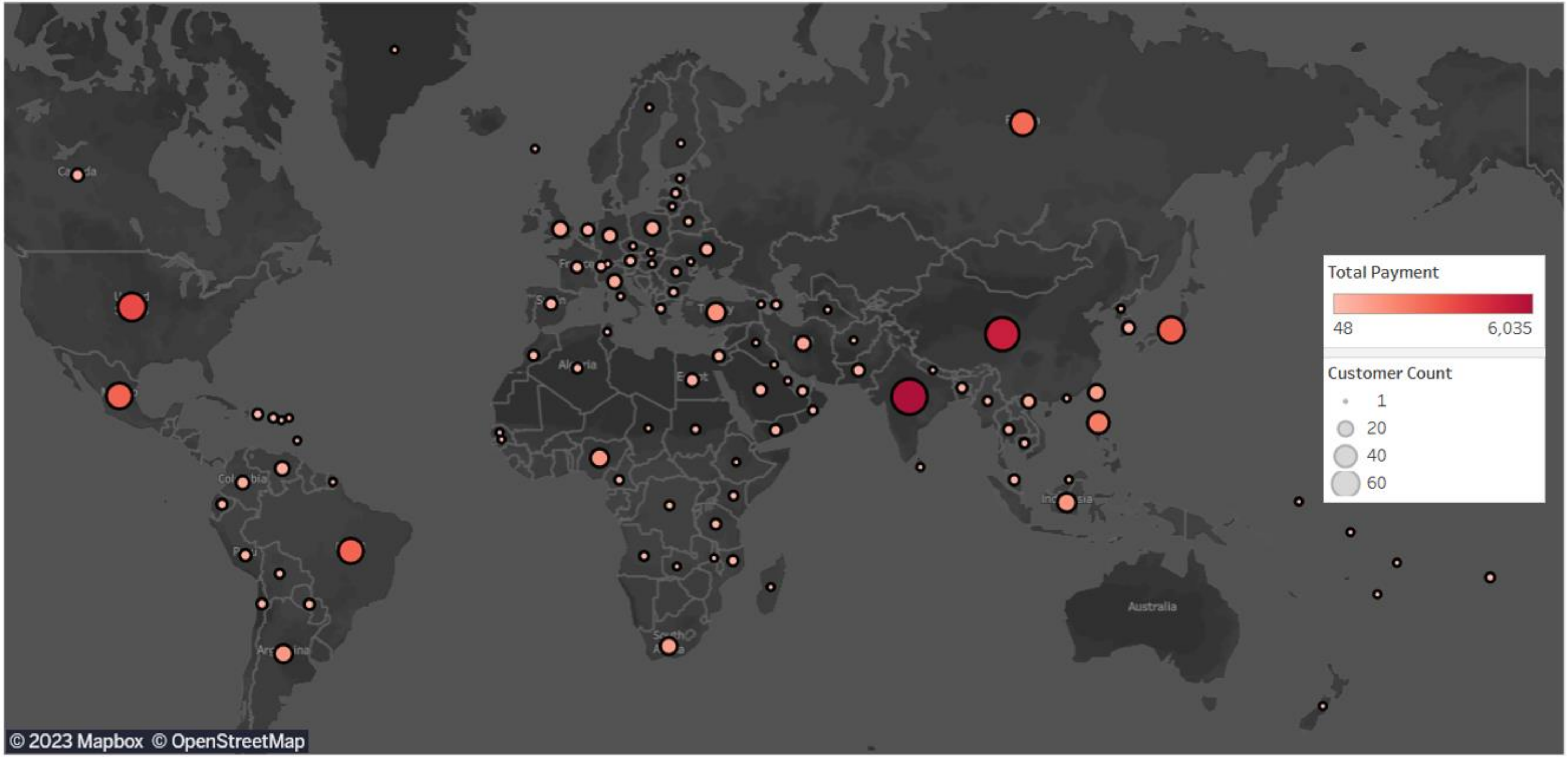
● Modernise the library of films – all current titles are from 2006.

● Offer series as well as films to attract a larger audience.

● Increase the offering of films in other languages to attract new customers.

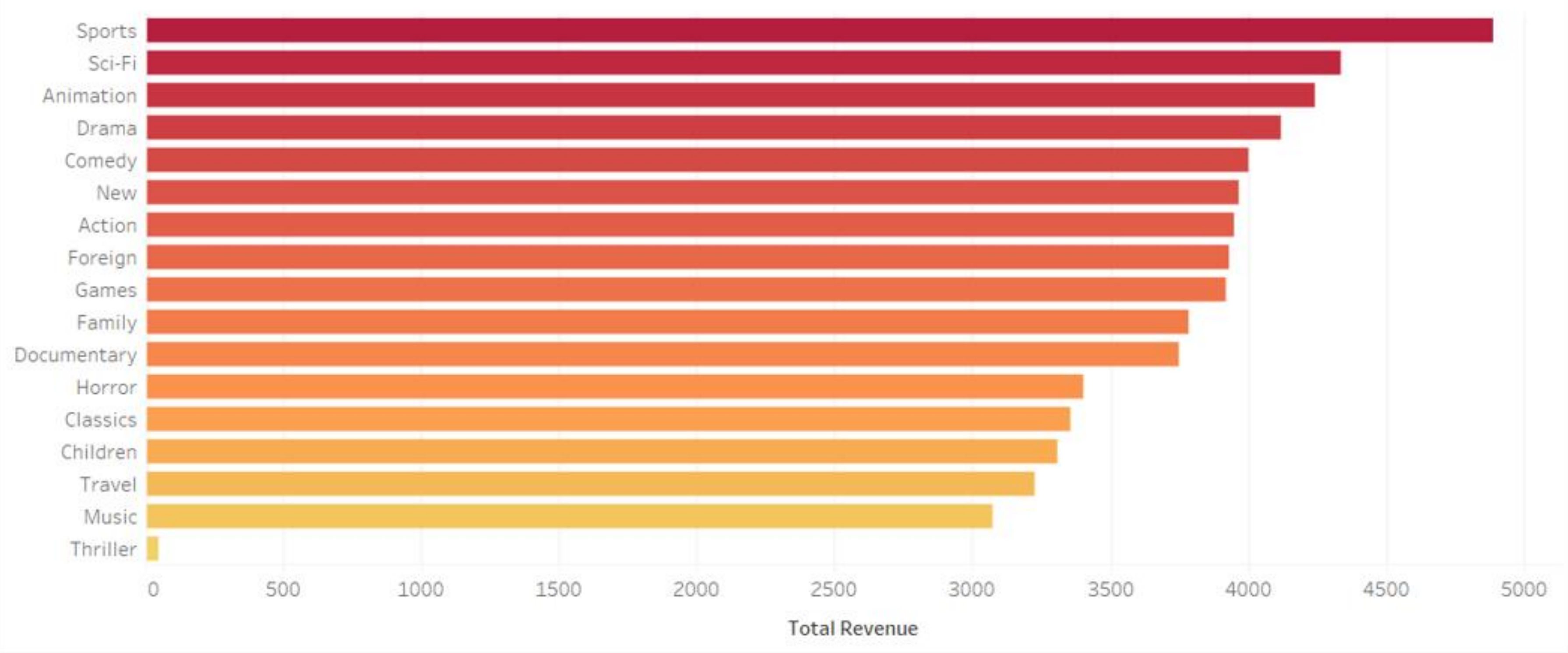
VISUALISATIONS

Which countries are Rockbuster customers based in?



VISUALISATIONS

Total revenue by genre



The most profitable film genre is Sports.
Sci-Fi, Animation and Drama also break the $4000 barrier in terms of revenue
Sports has the shortest average rental duration but the highest total revenue.

Additional Sources

Data dictionary
https://github.com/danfradat/RockbusterSQL/blob/main/DF%20Ex.3.10%20Data%20Dictionary.pdf

Presentation
https://github.com/danfradat/RockbusterSQL/blob/main/DF%20Ex.3.10%20pres.pdf

Tableau visualisations
https://public.tableau.com/app/profile/dan.franklin

SQL code
https://github.com/danfradat/RockbusterSQL/blob/main/CTE%20Query%20Top%205%20customers
https://github.com/danfradat/RockbusterSQL/blob/main/JOIN%20QUERY%20customer%20spend

Instacart

Online Groceries

This project looked at online grocery store **Instacart**. The goal was to uncover more information about their sales patterns by performing an initial data and exploratory analysis using python in order to derive insights and suggest strategies for better segmentation based on the provided criteria.

Project Goal

Uncover more information about sales patterns. Perform an initial data and exploratory analysis in order to derive insights and suggest strategies for better segmentation based on the provided criteria

Key Objectives

● Busiest days of the week and hours of the Day for ad scheduling

● Establish times of the day when people spend the most money

● Which departments have the highest frequency of product orders

● Uncover customer ordering behaviours based on various features such as region, family status, loyalty, and other demographics

Datasets

Instacart Data Sets:

● Data Dictionary

● "The Instacart Online Grocery Shopping Dataset 2017", Accessed from www.instacart.com/datasets/grocery-shopping-2017 via Kaggle on 23.6.23

CareerFoundry Data Sets:

● Customers Data Set

Tools

python™

X

NumPy

jupyter

ANACONDA

pandas

Skills & Procedures

Python

Data wrangling

Data merging

Deriving variables

Grouping data

Aggregating data

Reporting in Excel

Population flows

Preparation

- Download data and import into notebook as a pandas dataframe

- Conduct basic descriptive exploratory tasks

- Change data types of identifier variables into more suitable types and rename
columns where needed

- Access values and determine their meaning using a data dictionary

Analysis

- Creating new dataframes based on a certain criteria

- Fixing mixed-type variables, missing values, and removing duplicates

- Merging sets of dataframes and exporting as pkl

- Creating new columns using conditional logic in the form of if-statements, user-defined functions, the loc() function, and for-loops

- Creating flags

- Creating summary columns of descriptive statistics using the groupby() function

- Creating histograms, bar charts, line charts, and scatterplots

- Putting together a report

Results & Recommendations

- Focus advertising high-end products during early hours targeting commuters and high-earners

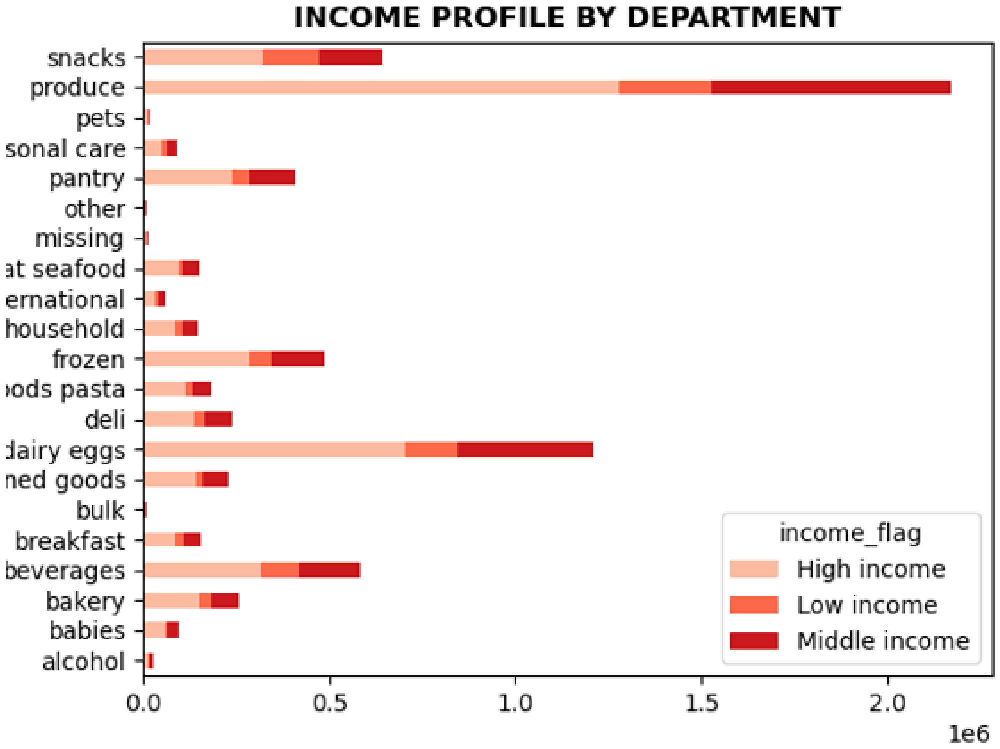
- Groupings are based just on purchase price. They could also group products by mark-up and attribute values to them that way.

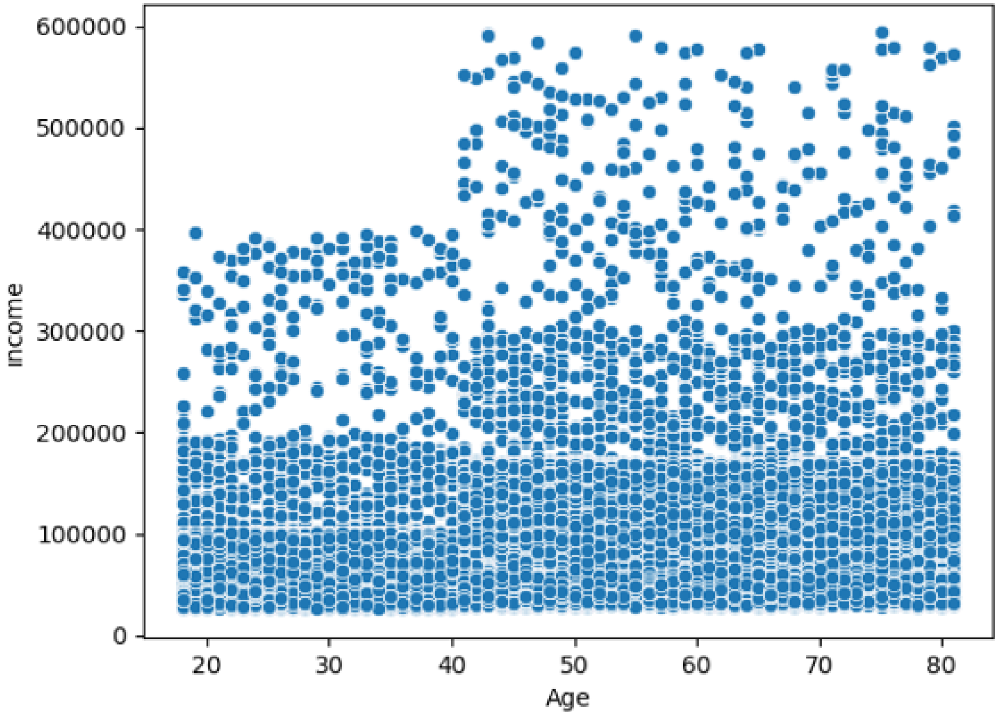
- Attach promotions to these best selling items as they seem to be staples, which then encourage customers to buy less well selling items.

- Incentivise brand loyalty by introducing repeat customer bonus schemes, a points system for example.

- Those over 40 have more spending power so targeting advertising to this age group could help boost income.

VISUALISATIONS





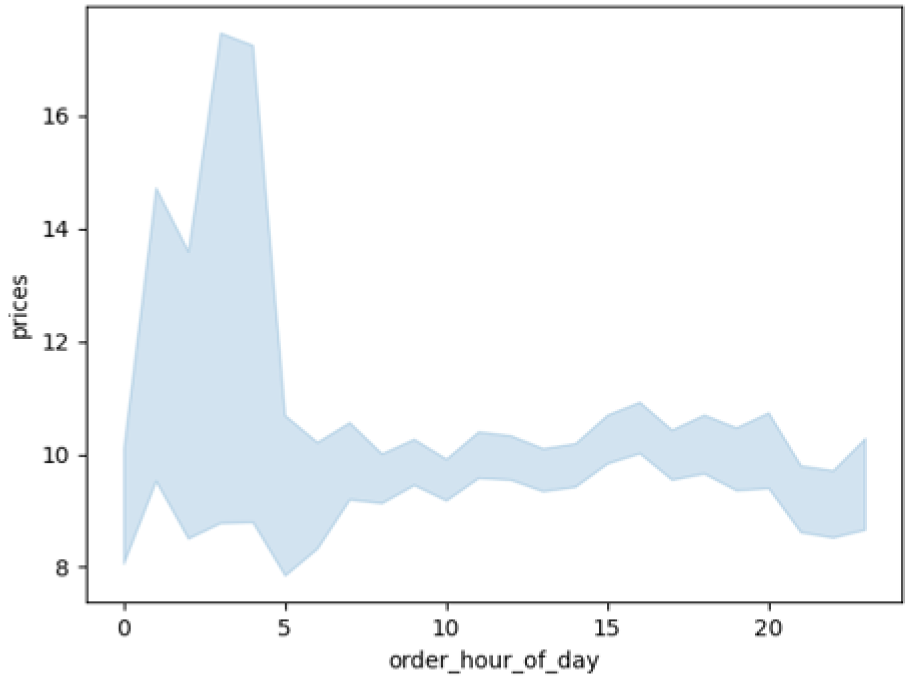
This shows that high income customers spend more per department than those earning less by roughly the same proportion as the income profile share (58%). An exception is snacks, where they only make up 50%, suggesting that those on lower incomes by more snacks.

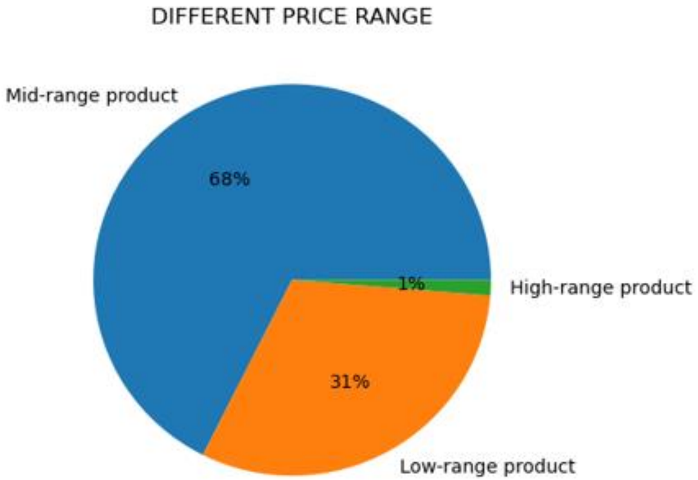
This scatterplot shows the relationship between customer age and income.
It shows that the majority of people earn under 200k. Very few 20-40 year olds earn more than 200k, with none earning more that 400k per year. People older than 40 start making over 200k, with some earning up to 600k.
From this chart we can see that those over 40 have more spending power.

VISUALISATIONS

2. Are there particular times of the day when people spend the most money?

This line plot chart shows the average price of items ordered by order hour of the day.
It seems people spend more between 3am and 5am.





The majority of products fall in to the mid-range price category. Only 1% of products are in the high-range category.

Additional Sources

Final Report
https://github.com/danfradat/Instacartpython/blob/main/Instacart%20Basket%20Analysis/05%20Sent%20to%20Client/DF%20Achievement%204%20Final%20Report.xlsx

Python Scripts
https://github.com/danfradat/Instacartpython/tree/main/Instacart%20Basket%20Analysis/03%20Scripts

Analysis & Visualisation
https://github.com/danfradat/Instacartpython/tree/main/Instacart%20Basket%20Analysis/04%20Analysis/Visualisations

Medical Staffing Agency

Preparing for Flu Season

This project was about analysing trends in influenza across the USA and determining where to send additional medical staff

Project Goal

To identify when and where to send extra medical staff across the USA to provide care during flu season

Key Objectives

- Determine whether influenza occurs seasonally or throughout the entire year

- Make a forecast for the coming year

- Examine influenza trends and how they can be used to plan for staffing needs across the country

- Determine which states need extra staff

Datasets

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

Population data by geography
Source: US Census Bureau

Tools

X

tableau

Skills & Procedures

Excel
Translating business requirements
Data cleaning
Data integration
Data transformation
Statistical hypothesis testing
Visual analysis
Forecasting
Storytelling in Tableau
Presenting results to an audience

Preparation

- Design data research project
- Formulate research hypothesis
- Creating a data profile for each of the data sets
- Implementing additional data quality measures
- Integrating data from two sources into one cohesive data set using data transformations.

Analysis

- Calculating variance and standard deviation for key variables
- Identifying variables and testing correlation
- Formulating a statistical hypothesis regarding an outcome of interest around two groups
- Conducting hypothesis testing
- Interim report consolidating the findings
- Creating a time forecast
- Creating visualizations that look at the distribution of a variable and the correlation between variables
- Publishing analysis as a Tableau Storyboard
- Recording video presentation for stakeholders

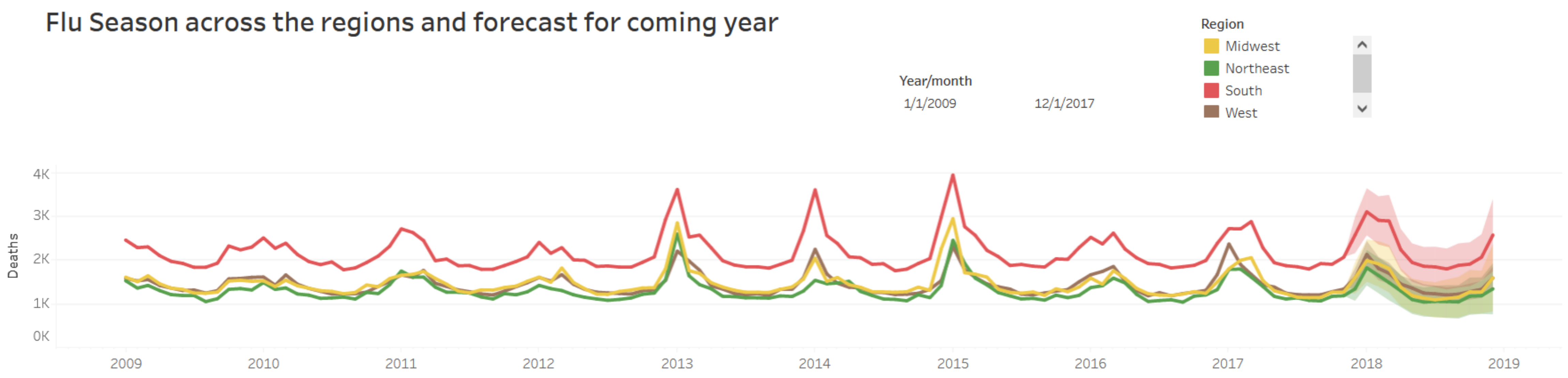
Results & Recommendations

When ? Flu season begins in November and lasts until April
Where? The worst affected states are California, New York, Texas, Pennsylvania and Florida
Who? People aged 65+ are the most vulnerable to flu and therefore carry a greater risk of fatality.

Send additional medical staff to the high-need states in time for the coming flu season. Those with historically highest number of deaths and larger populations of 65+ should be prioritised in order to provide more effective care to patients.

Further evaluate states in terms of death rate in order to plan in more detail where to send staff.

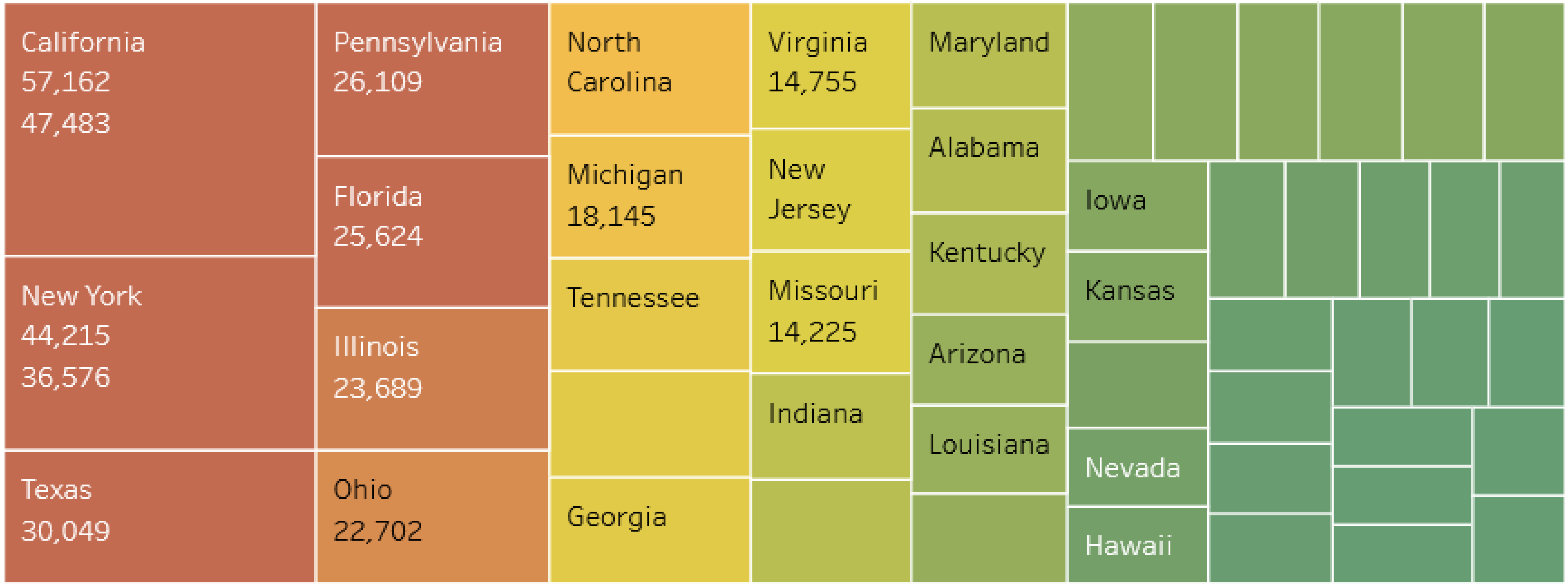
VISUALISATIONS



Flu Season across the regions and forecast for coming year

This graph shows that the deaths from influenza follow a seasonal trend - the flu season. Across all areas of the USA we can see that this is constant. The deaths typically peak from November til April. Use the slider to zoom in on specific periods.

VISUALISATIONS



State rankings

High Need - California, New York, Texas, Pennsylvania, Florida

Medium - Illinois, Georgia, Virginia, NJ, Missouri, Ohio, North Carolina, Michigan, Massachusetts, Tennesee

Low - All others

Additional Sources

Tableau Storyboard
https://public.tableau.com/app/profile/dan.franklin/viz/FluseasoninUSA/FluseasonintheUSA

Presentation
https://www.youtube.com/watch?v=HDHFNkEUScs

Visualisations
https://public.tableau.com/app/profile/dan.franklin

GameCo

Video Game Company

This project focused on a fictitious video game company called GameCo that wishes to use data to inform the development of new games

Project Goal

To perform a descriptive analysis of a video game data set to form a better understanding of how GameCo's new games might fare in the market place

Key Objectives

- To determine global and regional trends in video game sales
- To uncover which regions are currently experiencing rapid growth
- To establish which the most popular genres and platforms for games

Datasets

Historical game sales data

Source: VGChartz

Tools

Skills & Procedures

Excel

Grouping data

Summarizing data

Descriptive analysis

Visualizing results in Excel

Presenting results

Preparation

- Categorising data
- Removing duplicates
- Removing data errors
- Imputing missing values
- Correcting inconsistencies in formatting

Analysis

- Using pivot table to gain general data insights
- Grouping data
- Applying filters to look at specific segments
- Deriving new variables
- Analysing mean, median, mode
- Analyzing data distribution andskew
- Identifying outliers
- Creating visualisations
- Consolidation of project deliverables

Results & Recommendations

EU market –should continue receiving a large portion of the marketing budget as it has proven over a long period that it is an important, and stable, market.

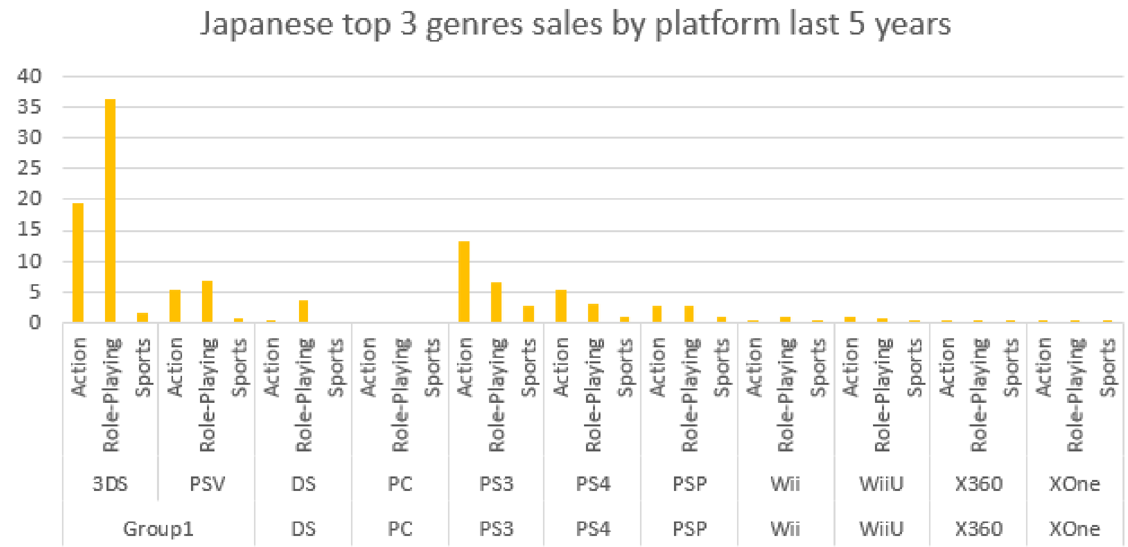
North America –Merits a large marketing budget in order to try and push NA sales and Global Sales up.

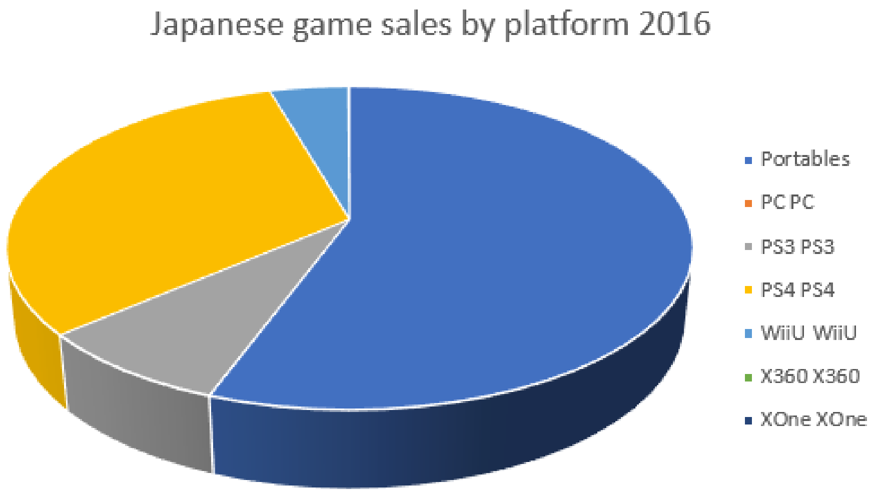
GameCo should focus on producing and marketing **Action, Shooter, and Sports** games for the EU and North American market, and publish these games to be sold on the latest Playstation and Xbox consoles.

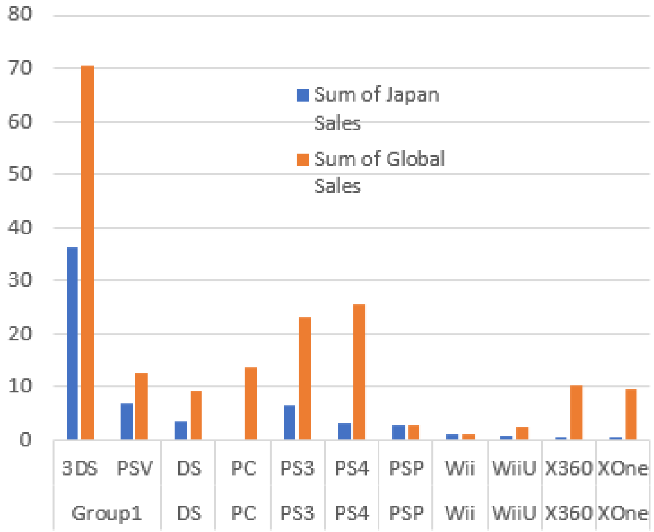
Japan –In order to maximise sales in this market **GameCo** should produce **Role-Playing** games for 3DS and PSV hand-held devices and also release the Action titles made for NA and EU on hand-held devices in this market.

VISUALISATIONS

Importance of Role-Playing games on hand-held devices in Japanese market





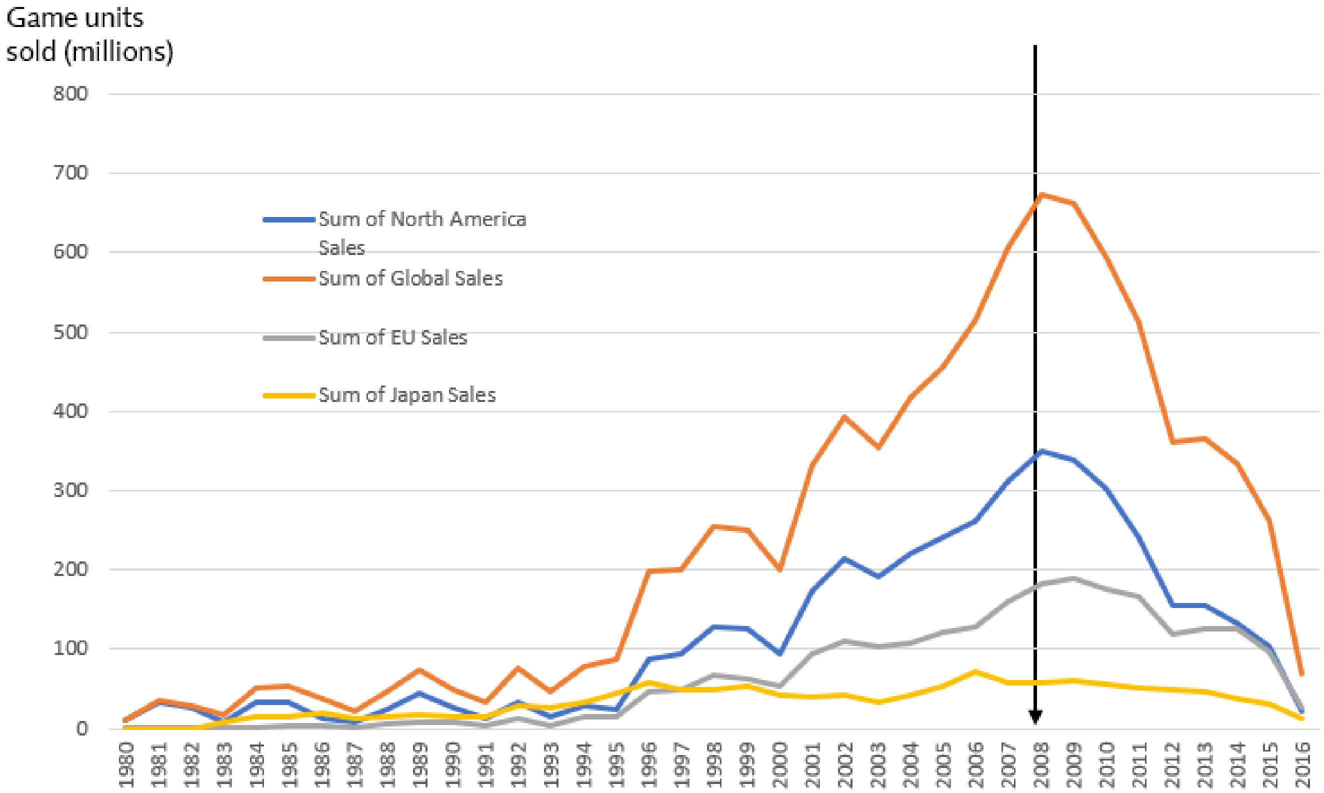


Above and to the left are 2 charts showing game sales in Japan over the last 5 years. Above looks at genre sales by platform and shows that role-playing and Action games are the best sellers, and on portable devices such as 3DS and PSV. On the left is a comparison between Role-playing game sales in Japan and globally, by platform. Again, you can see that half the role-playing game sales globally are in the Japanese market, and that the majority of these are either for the 3DS or PSV.

Above is a pie chart representing all Japanese game sales by platform. The large blue section represents the hand-held devices 3DS and PSV, and is labelled Portables. As you can see it dominates Japanese game sales, followed by PS4. Thiswould suggest that investing in the development of Role-playing games for hand-held devices is of utmost importance in the Japanese market.

VISUALISATIONS

Financial crisis of 2008



The 2008 financial crisis resulted in a negative shift in sales and a downwards trend ever since. Disposable income levels are lower than they were pre-recession hence less game sales.

Year	Japan	North America	EU
2008	60.36m	351.44m	184.7m
2016	13.7m = 23% of 2008 total	22.66m = 6% of 2008 total	26.76m = 14% of 2008 total

The traditionally largest market, North America, has taken the biggest hit, 2016 sales were 6% of those in 2008, this has of course pulled down global sales significantly. In the chart it is visible that the plot of Global Sales closely resembles that of NA sales.

Additional Sources

Clean dataset and workings
https://github.com/danfradat/GameCo/blob/main/DF%20Exercise%201.10.xlsx

Presentation
https://github.com/danfradat/GameCo/blob/main/Final%20Project%20Presentation.pptx

Project Reflections
https://github.com/danfradat/GameCo/blob/main/DF%20Project%20Reflections%20Exercise%201.10.pdf

Pig E. Bank

Bank

This project focused on Pig E. Bank, a fictitious bank hoping to integrate data mining algorithms to identify customers most likely to leave the bank

Project Goal

Build a decision tree using data mining mechanisms, in order to predict which customers are most likely to leave the bank

Key Objectives

- Calculate basic descriptive statistics to understand the data
- Using pivot tables to identify factors behind customers leaving the bank
- Analyse data on remaining customers and those that have left
- Build a decision tree

Datasets

Client Dataset
Source: Career Foundry

Tools

Skills & Procedures

Big data

Data ethics

Data mining

Predictive analysis

Time series analysis and forecasting

Using GitHub

Preparation

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

Analysis

- Calculating basic descriptive statistics
- Separating clients into existing and former , then analysing statistics using pivot tables
- Determining the leading contributing factors that lead to client loss
- Establishing the weight / importance of the leading factors and building a decision tree
- Documenting the analysis in Excel

Results & Recommendations

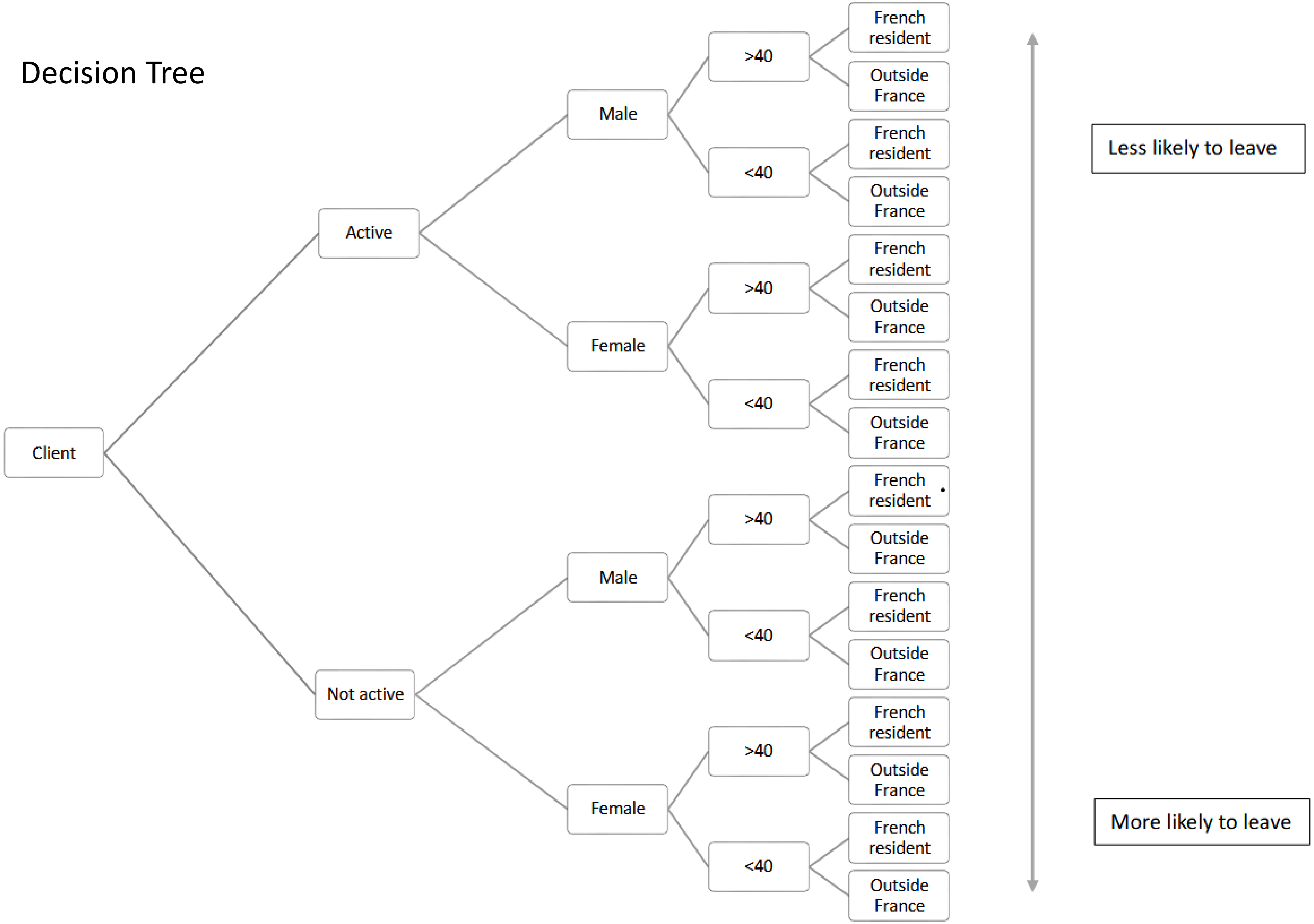
I saw patterns that suggested active customers are more likely to stay, females more like to leave, as well as people aged <40.

Interestingly, a much smaller % of French customers leave than the Germans or Spanish do.

I also looked at salaries, balances, and tenure What varied most was balances, with those remaining having on average a lower balance than those that leave – perhaps they take their money elsewhere to try and make more from it suggesting interest rates at Pig E. Bank might be lower than competitors.
Average tenure suggests that those who have banked with them for over 5 years are more likely to stay, but the average tenure of those that left was 4.7 years, so not much different.

VISUALISATIONS

Decision Tree



Client

Active

Male

>40

French resident

Outside France

<40

French resident

Outside France

Female

>40

French resident

Outside France

<40

French resident

Outside France

Not active

Male

>40

French resident

Outside France

<40

French resident

Outside France

Female

>40

French resident

Outside France

<40

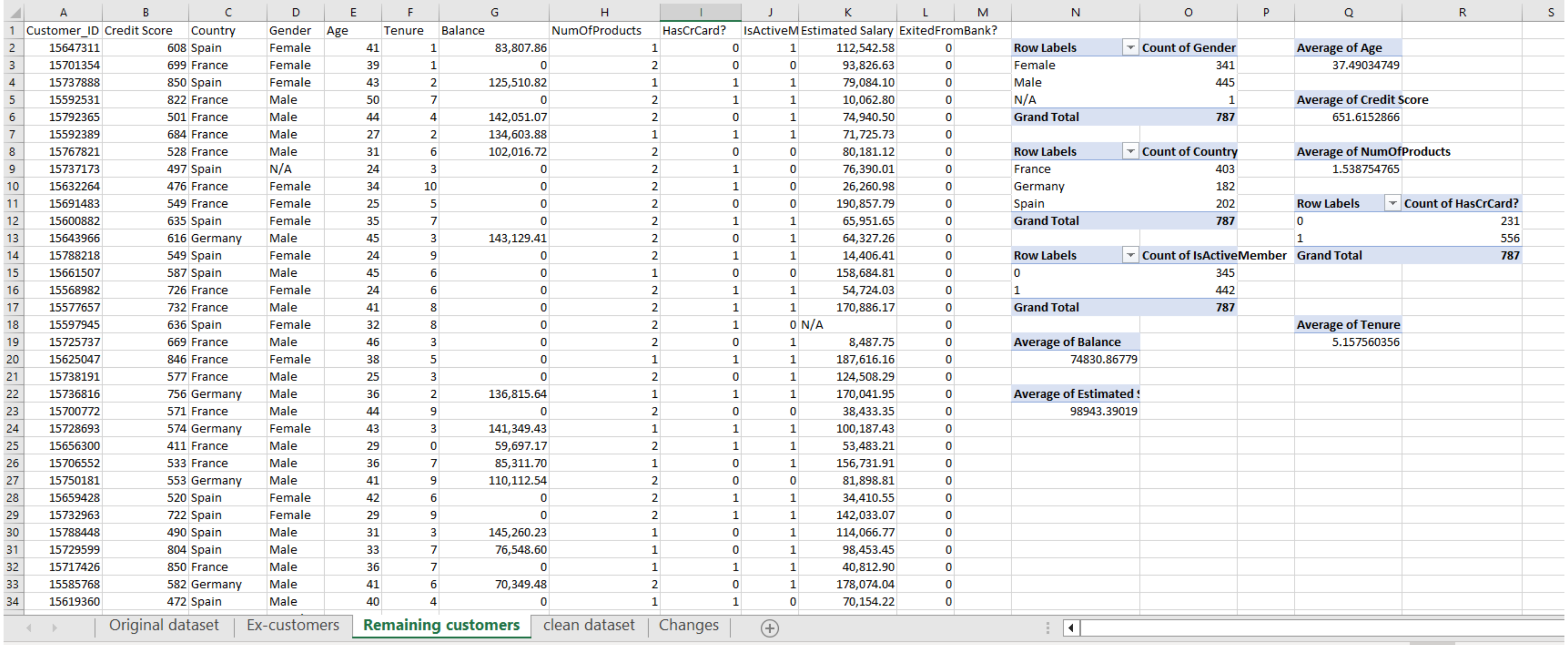
French resident

Outside France

Less likely to leave

More likely to leave

VISUALISATIONS



	A	B	C	D	E	F	G	H	I	J	K	L
1	Customer_ID	Credit Score	Country	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard?	IsActiveM	Estimated Salary	ExitedFromBank?
2	15647311	608	Spain	Female	41	1	83,807.86	1	0	1	112,542.58	0
3	15701354	699	France	Female	39	1	0	2	0	0	93,826.63	0
4	15737888	850	Spain	Female	43	2	125,510.82	1	1	1	79,084.10	0
5	15592531	822	France	Male	50	7	0	2	1	1	10,062.80	0
6	15792365	501	France	Male	44	4	142,051.07	2	0	1	74,940.50	0
7	15592389	684	France	Male	27	2	134,603.88	1	1	1	71,725.73	0
8	15767821	528	France	Male	31	6	102,016.72	2	0	0	80,181.12	0
9	15737173	497	Spain	N/A	24	3	0	2	1	0	76,390.01	0
10	15632264	476	France	Female	34	10	0	2	1	0	26,260.98	0
11	15691483	549	France	Female	25	5	0	2	0	0	190,857.79	0
12	15600882	635	Spain	Female	35	7	0	2	1	1	65,951.65	0
13	15643966	616	Germany	Male	45	3	143,129.41	2	0	1	64,327.26	0
14	15788218	549	Spain	Female	24	9	0	2	1	1	14,406.41	0
15	15661507	587	Spain	Male	45	6	0	1	0	0	158,684.81	0
16	15568982	726	France	Female	24	6	0	2	1	1	54,724.03	0
17	15577657	732	France	Male	41	8	0	2	1	1	170,886.17	0
18	15597945	636	Spain	Female	32	8	0	2	1	0	N/A	0
19	15725737	669	France	Male	46	3	0	2	0	1	8,487.75	0
20	15625047	846	France	Female	38	5	0	1	1	1	187,616.16	0
21	15738191	577	France	Male	25	3	0	2	0	1	124,508.29	0
22	15736816	756	Germany	Male	36	2	136,815.64	1	1	1	170,041.95	0
23	15700772	571	France	Male	44	9	0	2	0	0	38,433.35	0
24	15728693	574	Germany	Female	43	3	141,349.43	1	1	1	100,187.43	0
25	15656300	411	France	Male	29	0	59,697.17	2	1	1	53,483.21	0
26	15706552	533	France	Male	36	7	85,311.70	1	0	1	156,731.91	0
27	15750181	553	Germany	Male	41	9	110,112.54	2	0	0	81,898.81	0
28	15659428	520	Spain	Female	42	6	0	2	1	1	34,410.55	0
29	15732963	722	Spain	Female	29	9	0	2	1	1	142,033.07	0
30	15788448	490	Spain	Male	31	3	145,260.23	1	0	1	114,066.77	0
31	15729599	804	Spain	Male	33	7	76,548.60	1	0	1	98,453.45	0
32	15717426	850	France	Male	36	7	0	1	1	1	40,812.90	0
33	15585768	582	Germany	Male	41	6	70,349.48	2	0	1	178,074.04	0
34	15619360	472	Spain	Male	40	4	0	1	1	0	70,154.22	0

Row Labels	Count of Gender
Female	341
Male	445
N/A	1
Grand Total	**787**

Row Labels	Count of Country
France	403
Germany	182
Spain	202
Grand Total	**787**

Row Labels	Count of IsActiveMember
0	345
1	442
Grand Total	**787**

Average of Balance
74830.86779

Average of Estimated S
98943.39019

Average of Age
37.49034749

Average of Credit Score
651.6152866

Average of NumOfProducts
1.538754765

Row Labels	Count of HasCrCard?
0	231
1	556
Grand Total	**787**

Average of Tenure
5.157560356

Additional Sources

Findings and Decision Tree
https://github.com/danfradat/Pig-E.-Bank/blob/main/DF%20Exercise%205.4.pdf

Datasets and Workings
https://github.com/danfradat/Pig-E.-Bank/blob/main/DF%20exercise%205.4%20excel.xlsx

Emissions and Life Expectancy

Looking into the relationship
between emissions per capita and
life expectancy

This project set about looking into the relationship between countries **emissions per capita and life expectancy**

Project Goal

To look at relationships between emissions and life expectancy. I was interested to see if there is a correlation between the two. Increased emissions over the years can suggest a country is developing and therefore such things as education and medical care would likely improve.

Key Objectives

To explore:

- a country's average life expectancy as emission levels increase.
- Correlation between emissions and life expectancy.
- Time after the initial increase in a country's emissions and a rise in life expectancy.
- Any decline in the rate of increasing life expectancy or a decrease once emission levels get to a certain amount per capita?
- How emissions and life expectancy compare across the world?

Datasets

https://www.kaggle.com/datasets/ulrikthygepedersen/life-expectancy

https://www.kaggle.com/datasets/thedevastator/global-fossil-co2-emissions-by-country-2002-2022

https://www.statista.com/statistics/1040159/life-expectancy-united-kingdom-all-time/#statisticContainer

Tools

X

P

jupyter

python

tableau

quandl

NumPy

Skills & Procedures

-Exploratory analysis through visualizations (scatterplots, correlation heatmaps, pair plots and categorical plots)

- Geospatial analysis using a shapefile
- Regression analysis
- Cluster analysis
- Time-series analysis
- Analysis narrative and final results

Preparation

- Sourcing data
- Conduct exploratory visual analysis using relevant Python libraries
- Define questions to explore based on understanding of what the data contains

Analysis

- Source a shapefile containing location data that corresponds to the location data
- Data cleaning and wrangling in Python
- Creating a data profile of summary statistics
- Conduct a geospatial analysis by creating a choropleth map using Python libraries
- Create visualizations using Python
- Prepare data for a regression analysis
- Split data into a training set and a test set
- Run a linear regression on the data and analyse the model performance statistics
- Prepare data for a cluster analysis
- Use the elbow technique to determine the optimal number of clusters
- Run the k-means algorithm
- Source time-series data relevant to project data via an API
- Conduct a Dickey-Fuller test and plot autocorrelations to test for stationarity.
- Perform differencing to stationarize non-stationary data
- Create dashboard of findings on Tableau

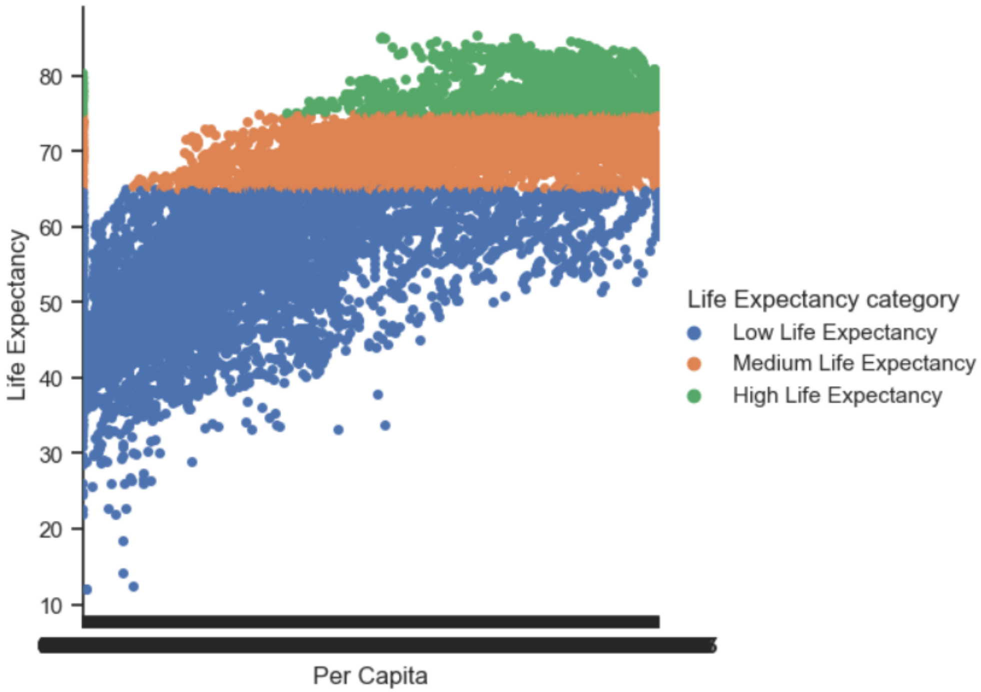
Results & Recommendations

- To further this analysis it would be suggested to look into data on a more individual basis, as I did with the UK, from developed, emerging markets, and developing countries.

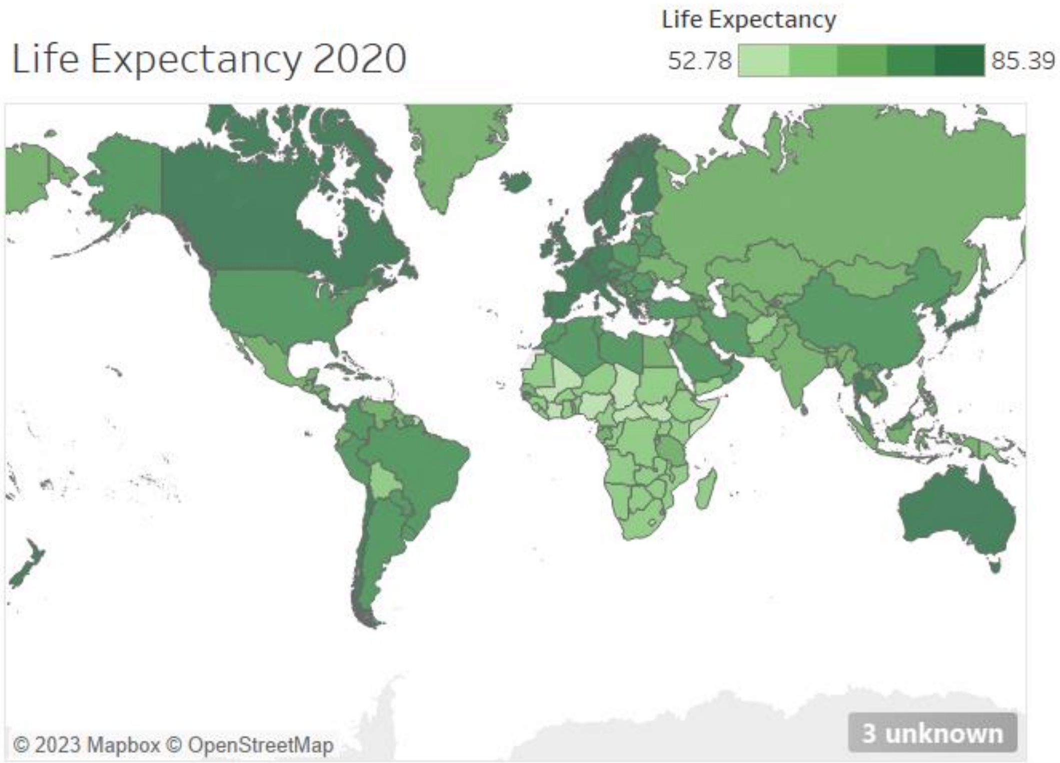
- Once having analysed a handful of countries from each different category, it will be interesting to see what patterns emerge regarding life expectancy and emissions per capita. For developed countries there is data out there for life expectancy going back 200 years such as that used for the UK in this analysis.

- As the world moves towards more green energy, it might be of benefit to not only look at emissions but also productivity or GDP as extra markers.

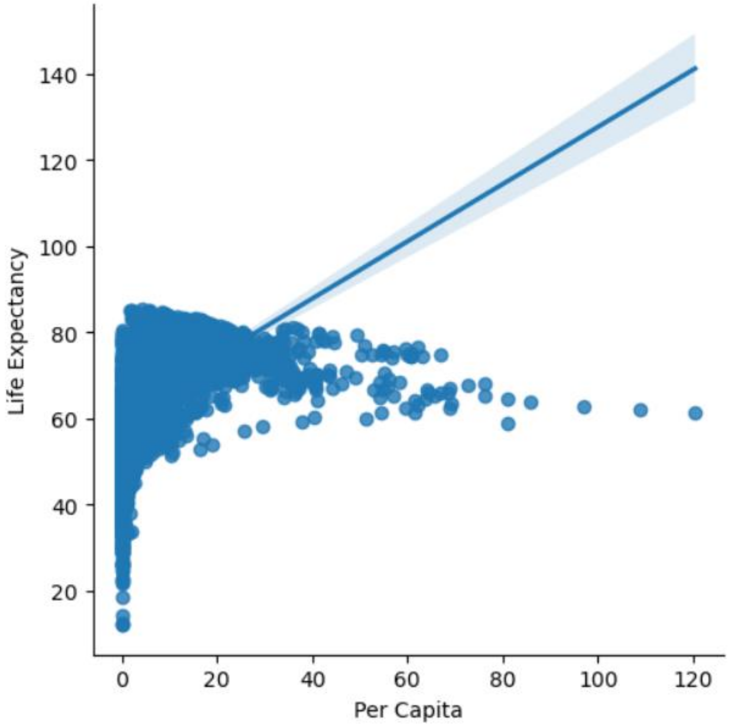
VISUALISATIONS



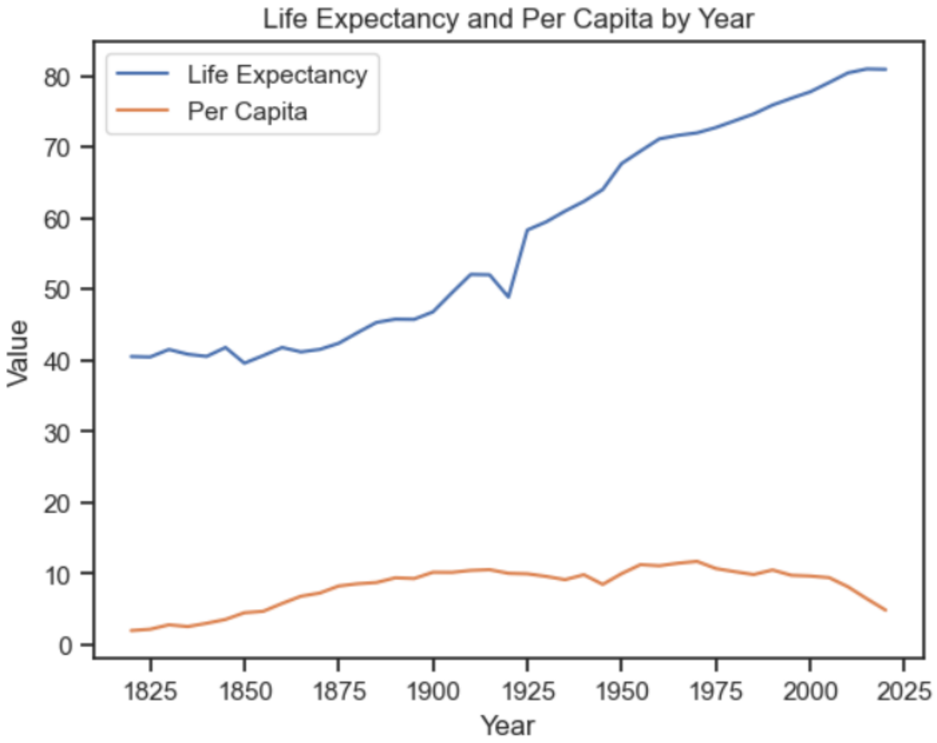
In this categorical plot, the most dense collection of High Life Expectancy appears to coincide with the highest emissions, and the largest collection of Low Life Expectancy with low emissions per capita. Interestingly, it appears that the highest life expectancies are in the mid-range of emissions per capita, but only marginally.


Life Expectancy 2020

VISUALISATIONS



This scatterplot shows a positive association between the variables, but not a strong relationship, and the data is non-linear



This is an interesting representation of life expectancy and emissions per capita over time in the UK. They both increase simultaneously, emissions level off in the 1900's whilst life expectancy continues to rise apart from a dip following ww1, possibly due to improved infrastructure and living conditions, healthcare etc, by 2020 life expectancy is at is highest level ever, whilst emissions per capita have reduced to their lowest levels since the 1800's. The y-axis is representing age in years and metric tons of emissions.

Additional Sources

Storyboard
https://public.tableau.com/views/EmissionsandLifeExpectancy/Story1?:language=en-US&publish=yes&:display_count=n&:origin=viz_share_link

Analysis and Visualisations
https://github.com/danfradat/EmissionsLifeExpectancy

Python code
https://github.com/danfradat/EmissionsLifeExpectancy