Dr Elsa (Elizabeth) Ekevall

Data Analytics Portfolio

Dr Elsa (Elizabeth) Ekevall

Mission Statement

To become a data analyst in a humane organisation that invests in its employees and transforms lives. Working collaboratively with colleagues to interrogate the data, uncovering the patterns and trends, and creatively explaining the story.

Currently located near Frankfurt, I am a determined resilient person who enjoys variety and new challenges. Naturally curious with a love of learning, I would like to use my analytical and problem-solving skills in an energising focused environment that offers flexibility and allows some independence.

With my engineering background I offer critical thinking and logical abilities, along with transferable communication skills from an extensive career as an academic researcher and science, technology, engineering and mathematics communicator.

"I am naturally curious and love learning. Before moving into data analytics I worked as a researcher in academia; defining problems then delving deeper to understand the constituent parts before using my creative abilities to piece it all together to find a solution that overcomes the inevitable obstacles that crop-up. As well as a science communicator creating clear concise stories to explain complex information and difficult concepts.

And I have my own superhero character – Miz Ingenuity! Like me she enjoys solving puzzles. She uses her analytical and problem-solving skills to catch the villains, whereas I use mine to find patterns and trends in the data."

The knowledge, transferable skills and personal abilities I use most as a data analyst.

CAREER FOUNDRY PROJECTS

01

GameCo

- descriptive analysis of global video game sales

- Microsoft Excel
- Cleaning and wrangling data
- Exploratory data analysis
- Descriptive analysis
- Data visualisation
- Presentation
- Report

02

Planning for Influenza Season

- analysing seasonal trends and forecasting staffing ratios

- Microsoft Excel
- Tableau
- Sourcing and profiling data
- Cleaning and wrangling data
- Transforming and integrating data
- Statistical analysis
- Hypothesis design and testing
- Predictive analysis
- Data visualisation
- Forecasting
- Video Presentation

03

Rockbuster

- exploratory SQL data analysis to answer business questions

- SQL
- Tableau
- Summarising and cleaning data
- Database querying
- Filtering data
- Joining tables
- Performing subqueries
- Common table expressions
- Descriptive analysis
- Presentation

04

Instacart

- exploratory Python data analysis to derive insights and suggest strategies

- Python
- Profiling data
- Cleaning and wrangling data
- Merging and exporting data
- Grouping and aggregating variables
- Deriving new variables
- Prescriptive analysis
- Data visualisation
- Population flow report Microsoft Excel

05

UK household food shopping

- Advanced analytical analysis and dashboard design

- Python
- Sourcing open data
- Cleaning and wrangling data
- Exploratory data analysis
- Supervised Machine Learning
- Unsupervised Machine Learning
- Data visualisation
- Dashboard Presentation

01 GameCo

Descriptive analysis of global video game sales

Descriptive analysis of global video game sales

Descriptive Analysis | Microsoft Excel | Data Cleaning | Data Wrangling | Maths and Statistics | Statistical Analysis | Data Visualisation

Project Overview

GameCo, a new video company, have requested a descriptive analysis of the VGChartz data set to get a better understanding of how their new games might fare in the market.

Key Business Questions

- How have their sales figures varied between geographic regions over time?
- What other publishers will likely be the main competitors in certain markets?
- Are certain types of games more popular than others?
- Have any games decreased or increased in popularity over time?

Tools

The full analysis was carried out using Microsoft Excel.

GameCo Data Set

The GameCo data set covers historical sales of video games (sold 100,000+ copies) spanning different platforms, genres, and publishing studios from 1980 to 2016. There is no mention of how the software sales were produced and may be bias due to the recent increase in the digital market share. The data cleaning process is documented in the GameCo Report.

Original data set
- 12 Columns
- 16601 rows

→

Cleaned data set
- 11 Columns
- 16579 rows

Dealing with outliers

The decision was made not to remove outliers. They are the most popular games and removing them may impact some aspects of the analysis e.g. when looking at the most popular genres. It was highlighted where these outliers may have an influencing effect.

Outliers with Regional Sales Above 10 *Million Units Sold Between 2007 and 2017

Rank	Name	Platform	Year	Genre	Publisher	Global Sales *
3	Mario Kart Wii	Wii	2008	Racing	Nintendo	35.82
4	Wii Sports Resort	Wii	2009	Sports	Nintendo	33.00
9	New Super Mario Bros. Wii	Wii	2009	Platform	Nintendo	28.62
16	Kinect Adventures!	X360	2010	Misc	Microsoft Game Studios	21.82

Data Analysis Tasks

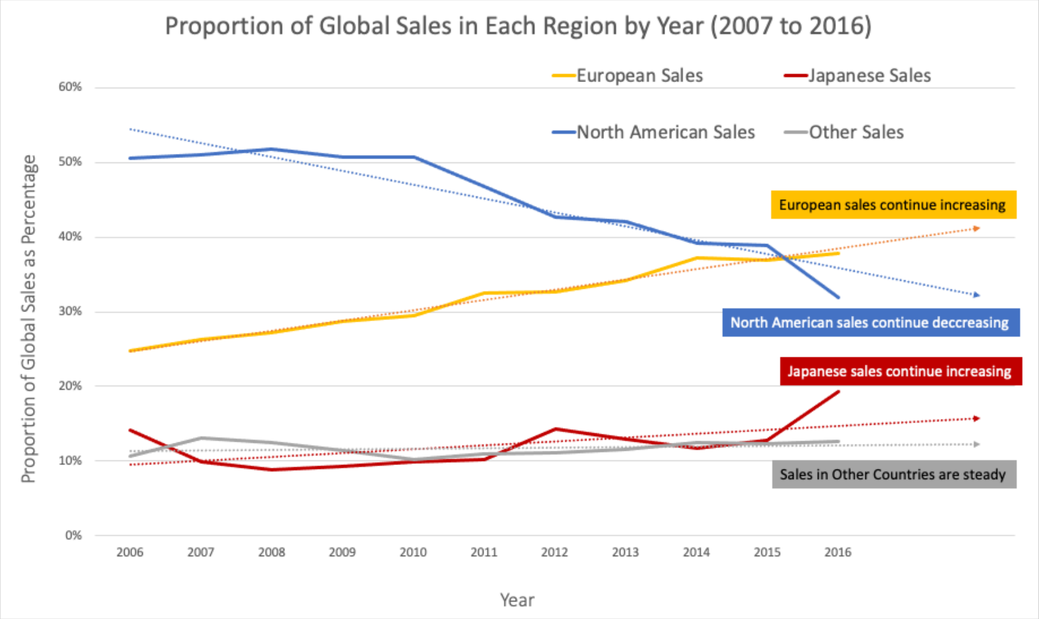
- ✓ Created a data dictionary (documented in the GameCo Report).
- ✓ Removed duplicates and missing values, and imputed values.
- ✓ Formatted the data and corrected values.
- ✓ Carried out an exploratory data analysis.
- ✓ Used pivot tables to group, filter and summarise the data.
- ✓ Created new variables to provide more meaningful data.
- ✓ Created bar and column charts, box and whisker plots, and scatter plots.
- ✓ Used consistent colour coding - Europe is yellow, Global is green, Japan is red, North American is blue and the Other Countries are grey.
- ✓ Created presentation communicating insights for stakeholders.

Data Analysis

The analysis generated insights based on data from the last 10 years (2007 - 2016) concentrating mainly on Europe, Japan and North America.

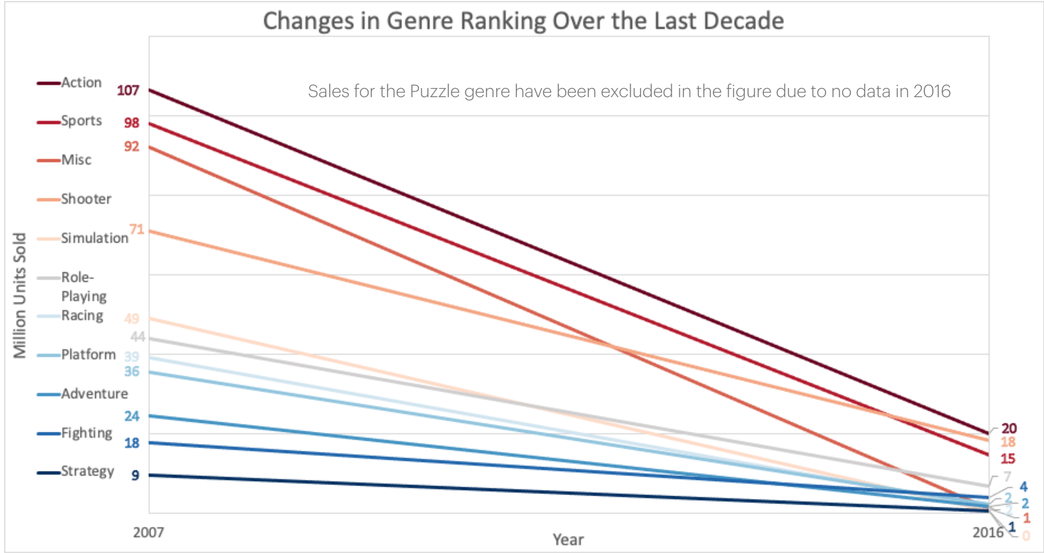
Key Findings

- Sales for the various geographic regions have changed over the years and the marketing budget should be revised to reflect the changes in market share. Linear forecasts predict these trends will continue.



- The top six publishers in the last four years are Nintendo (grew in Japan), Electronic Arts, Activision, Ubisoft, Sony Computer Entertainment (Grouped) - all had growth in Europe - and Take-Two Interactive (grew in North America).
- Focus on games for PS4 and XOne in Europe and North America and 3DS in Japan.

- Action, Shooter and Sports are the top three genres. The Role-Playing and Fighting genres have increased in popularity, while Misc and Simulation have decreased.



Main Recommendation

Since 2008 the number of million units sold globally and the number of games released has declined. The 2008 financial crisis may have had an impact, but the market has never recovered. Why? Is there something else lacking in recent years e.g. innovative games, groundbreaking consoles?

Further analysis of the digital market share for software, best selling games and market developments in consoles (e.g. will anything replace the WiiU) may offer an insight.

The GameCo Presentation and GameCo Report are archived in the GitHub project repository.

02 Planning for Influenza Season

Analysing seasonal trends and forecasting staffing ratios

Analysing seasonal trends and forecasting staffing ratios

Predictive Analysis | Descriptive Analysis | Microsoft Excel | Tableau | Data Cleaning | Data Wrangling | Maths and Statistics | Data Transformation | Hypothesis Testing | Statistical Analysis | Data Visualisation

Project Overview

The United States has an influenza season where more people than usual suffer from the flu. Some people, particularly those in vulnerable populations, develop serious complications and end up in hospital. Hospitals and clinics need additional staff to adequately treat these extra patients. The objective of this project is to determine when to send staff, and how many, to each state.

Formulated Research Hypothesis

If states have a high proportion of vaccinated health care workers (nurses, physician assistants and doctors), then fewer vulnerable people will develop serious complications and end up in the hospital.

Tools

The analysis was carried out using Microsoft Excel and Tableau.

Planning for Influenza Season Data Sets

The data sets only account for vulnerable populations under 5 years and 65+ years. 80% of the CDC Influenza data set was suppressed and the values imputed using the mean. Due to missing values New Population Data replaced Census Population Transformed. Hospitalisation Rates only cover approx 9% of the US population and not all 50 states (see Data Overview).

1. CDC Influenza Deaths
2. Census Population Transformed 202101
3. New Population Data 2010_17
4. CDC Influenza Visits
5. Vaccination Coverage among Health Care Personnel
6. Influenza Vaccination Coverage for All Ages 6 Months
7. CDC Hospitalisation Rates

→ Integrated data set

- 52 Columns
- 4896 rows

Mapping and Integrating the Data

Three mapping grids - Age Groups (as shown below), Months, Seasons and Years - were used to produce a VLOOKUP table for each data set. The data was integrated using two combined key variables; Combined Key 1: State and Year e.g. Alabama, 2010; and Combined Key 2: State and Year plus Month e.g. Alabama, 2010, Jan (see Data Transformations).

Mapping Grid 1: Age Groups

Age Group	Month and Age	New Population	All Vaccinated	Hospitalisation Rate
A Under 5 years	< 1 year	Under 5 years	6 Months - 4 Years	0-4 yr
	1-4 years		*NB - no vaccinations under 6 months*	
B 5-14 years	5-14 years	5 to 9 years		
		10 to 14 years		
C 15-24 years	15-24 years	15 to 19 years		
		20 to 24 years		
D 25-34 years	25-34 years	25 to 34 years		
E 35-44 years	35-44 years	35 to 44 years		
F 45-54 years	45-54 years	45 to 54 years		
G 55-64 years	55-64 years	55 to 59 years		
		60 to 64 years		
H 65-74 years	65-74 years	65 to 74 years		65-74 yr
I 75-84 years	75-84 years	75 to 84 years		75-84 yr
J 85+ years	85+ years	85 years and over		85+
K All Age Groups	Row Grand Total	Total Population	*≥6 Months*	Row Grand Total
L 65+ years	65-74 years	65 to 74 years	≥65 Years	65+ yr
	75-84 years	75 to 84 years		
	85+ years	85 years and over		

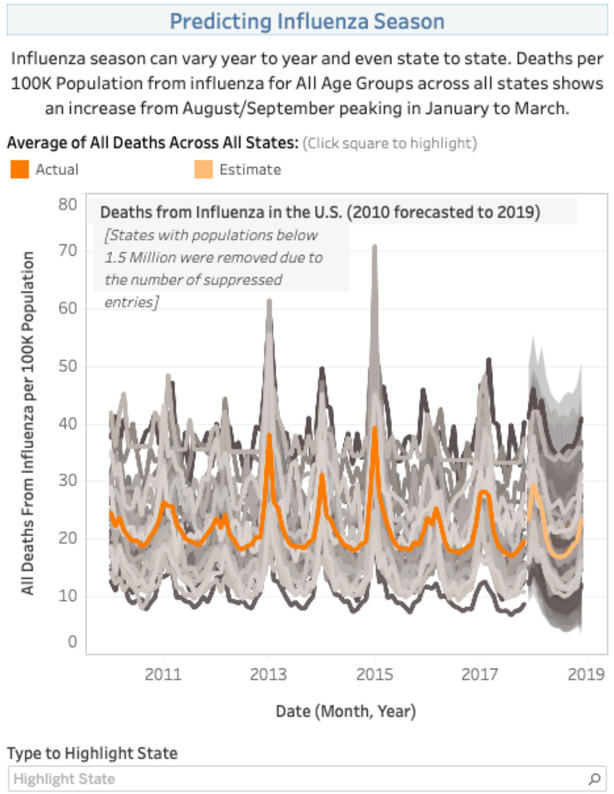
Data Analysis Tasks

- ✓ Cleaned the data - corrected entries, removed duplicates, dealt with missing values (see Data Profiles) and used consistent colour coding for the data sets.
- ✓ Used data mapping and transformations to integrate data from seven sources into one unified data set.
- ✓ Created a data dictionary for the integrated data set.
- ✓ Conducted statistical analysis and hypothesis testing, then consolidated the analytical insights in an interim analytical report.
- ✓ Created exploratory visualisations in Tableau; composition and comparison charts, temporal visualisations and forecasts, statistical visualisations and spatial analysis.
- ✓ Presented findings in a recorded video presentation of the Tableau Storyboard.

Data Analysis

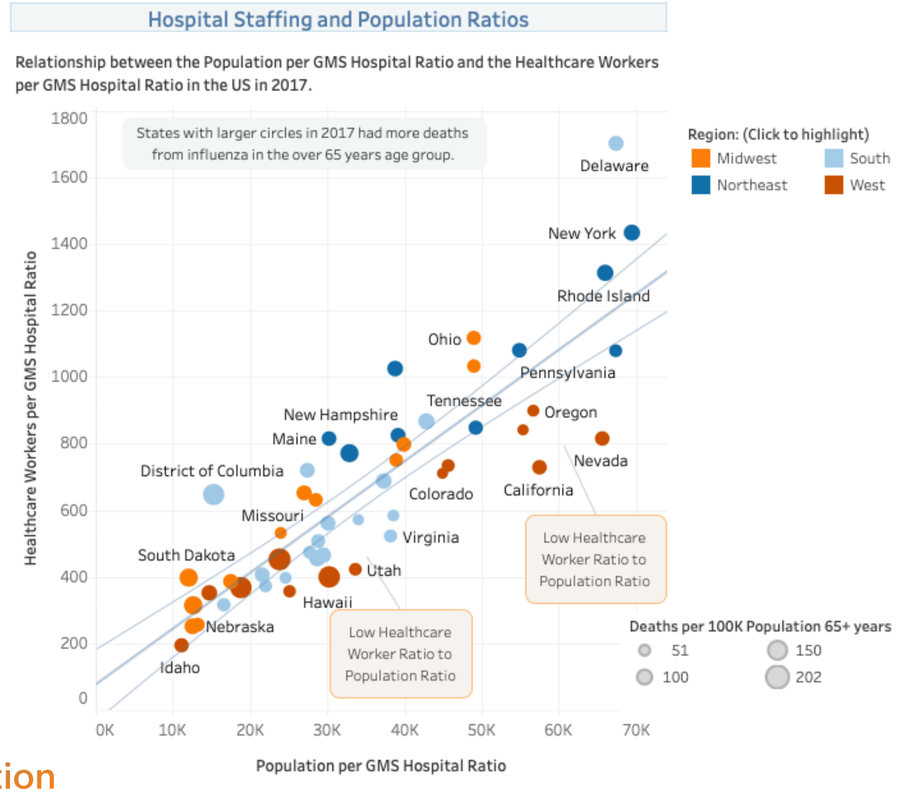
Statistical Results

The average vaccination rate for health care workers is high (86.6%) and this may contribute to the lack of statistical difference between states with higher and lower HW vaccination rates.



Key Findings

- The 2018 forecast predicts the busiest period will be between November 2017 to April 2018 with additional staff required from September 2018.
- Due to the variability of influenza season it is very difficult to accurately predict when and where to send staff; and this analysis could not determine how many.
- States were categorised depending on the anticipated likelihood of hospitals requiring additional staff.



Main Action

States in the West generally have lower Healthcare Worker per General Medical and Surgical (GMS) Hospital ratios and are thus more likely to require more healthcare staff during influenza season.

The Planning for Influenza Video Presentation and Interactive Story in Tableau. The project is archived in the GitHub project repository.

03 Rockbuster

Exploratory SQL data analysis to answer business questions

Exploratory SQL data analysis to answer business questions

Descriptive Analysis | SQL | Tableau | Data Cleaning | Data Wrangling | Maths and Statistics | Statistical Analysis | Data Visualisation

Project Overview

Rockbuster Stealth LLC is a movie rental company that used to have stores around the world. Facing stiff competition from streaming services the management team is planning to use its existing movie licenses to launch an online video rental service in order to stay competitive.

Key Business Questions

- 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?

Tools

The SQL analysis was carried out using PostgreSQL in pgAdmin 4.

The results of the queries were exported to Excel and analysed in Tableau.

Data Source and Limitations

The Rockbuster database supports standard online payment transaction processing for the companies' video rental business. It contains information about Rockbuster's film inventory, payments, rentals, customers, staff and stores.

After loading the Rockbuster data set into the PostgreSQL relational database management system (RDBMS) the entity relationship diagram was extracted using DbVisualizer. There are two fact tables - payment (which can be linked to all tables) and rental.

Interrogating the Data

1 Language	2 Stores	2 Employees	5 Ratings	17 Genres	109 Countries

```
--To find the payment and rental transaction periods
SELECT MAX(payment_date) AS max_payment_date,
       MIN(payment_date) AS min_payment_date,
       MAX(rental_date) AS max_rental_date,
       MIN(rental_date) AS min_rental_date
FROM payment A
INNER JOIN rental B ON A.rental_id = B.rental_id
```

Early on in the analysis I discovered that there were discrepancies in the dates of the data provided. All the films in this analysis were released in 2006, but the data covers rental transactions during the period 2005 to 2006, and all the payment transactions were made in 2007.

599 Customers	1000 Film Titles	4581 Videos	14,596 Transactions	16,044 Rentals	$61,312 Revenue

Data Analysis Tasks

- ✓ Created a data dictionary defining the structure of the database and containing the SQL queries.
- ✓ Wrote SQL statements to check for duplicates, identify dirty data and missing values.
- ✓ Created a data profile of summary statistics.
- ✓ Joined tables to write subqueries and common table expressions.
- ✓ Exported the SQL query results as csv files and created visualisations in Tableau.
- ✓ Created a technical document containing the descriptive statistics, and SQL queries with the extracted results and presentation for stakeholders.

Data Analysis

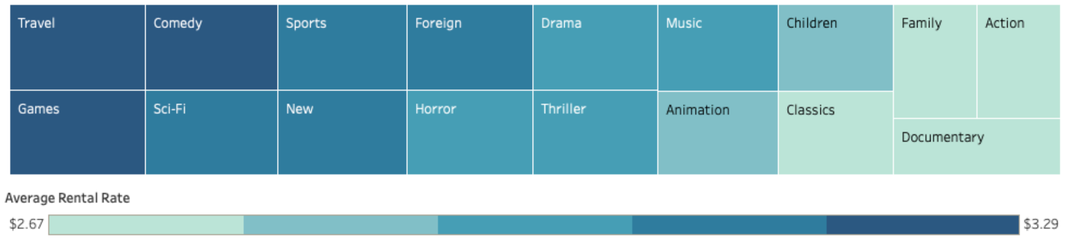
Key Findings

- ➡ The average rental duration was 5 days.
- ➡ Higher rental rates generate higher revenues.
- ➡ The rating PG-13 generates the most revenue.
- ➡ The most popular genres in terms of revenue are Sports and Sci-Fi.

Top 5 customers within the top 10 cities of top 10 countries

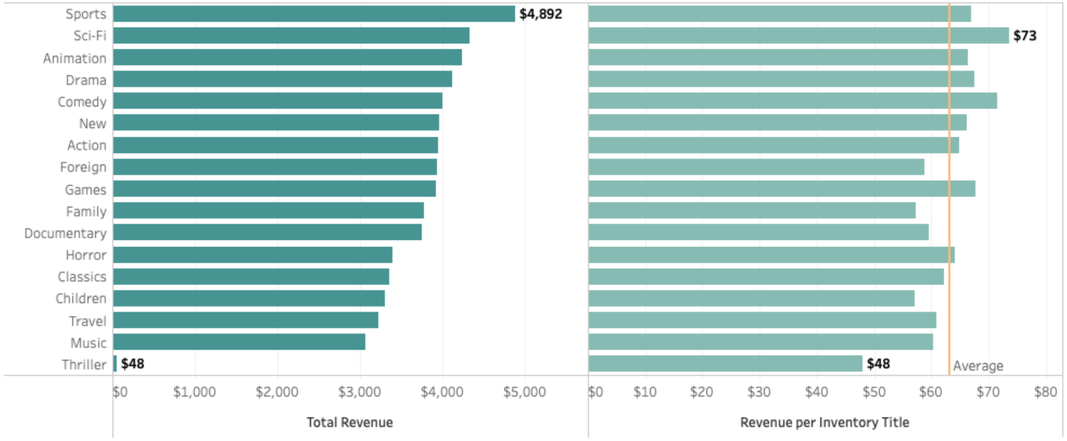
Customer ID	Customer First Name and Last Name	Country	City	Total Amount Paid
84	Sara Perry	Mexico	Atlixco	128.70
518	Gabriel Harder	Turkey	Sivas	108.75
587	Sergio Stanfield	Mexico	Celaya	102.76
537	Clinton Buford	United States	Aurora	98.76
367	Adam Gooch	India	Adoni	97.80

Genre Rental Rate



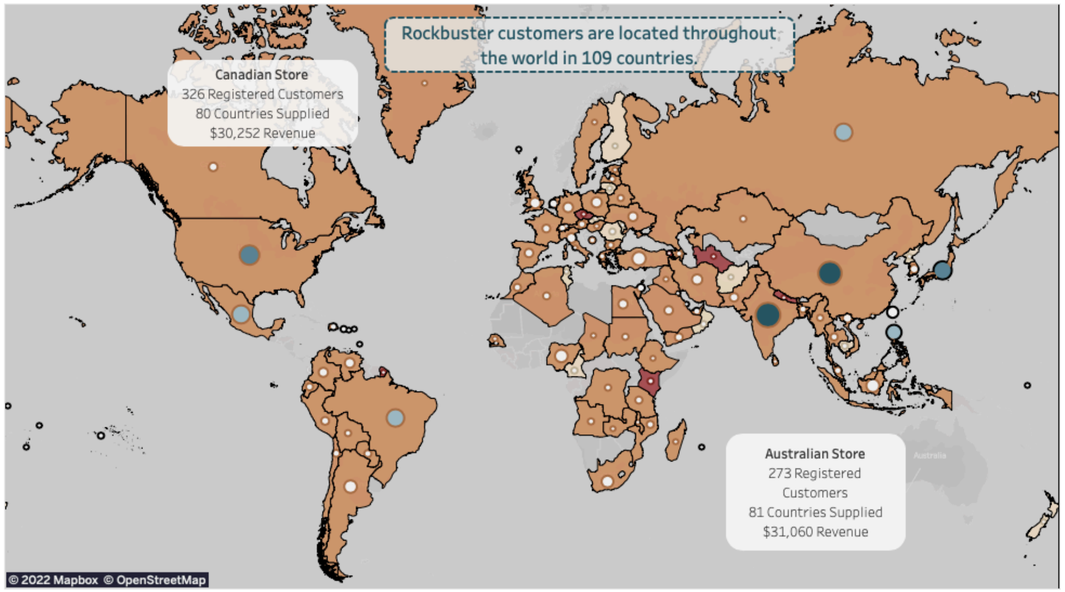
Documentaries, Action, Family and Classics have lower average rental rates. Out of these genres only Family has above average revenue per inventory title. **The Thriller genre performs very poorly, but that is not surprising with only one title in the inventory to rent.**

Genre Revenue

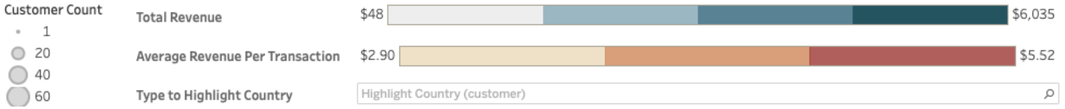


- ➡ The top ten countries in terms of customers are also the countries with the highest revenues and largest number of transactions.

Customer Count, Location and Revenue Map



The countries with the highest revenue are not the countries whose customers generally spend more. **These countries (China, India, Japan and the USA) have more customers and transactions.** The asian market is the biggest in terms of revenue and customer numbers. The Australian, European and African markets are much smaller.



Key Actions

Increase the $0.99 rental rate and review the: number of languages, licences for films (42 not used), rating G, genre Thriller (only one title), and rental rates.

The Rockbuster Presentation and Interactive Story in Tableau. The project is archived in the GitHub project repository.

04 Instacart

Exploratory Python data analysis to derive insights and suggest strategies

Exploratory Python data analysis to derive insights and suggest strategies

Prescriptive Analysis | Descriptive Analysis | Python | Microsoft Excel | Data Cleaning | Data Wrangling | Maths and Statistics | Data Transformation | Statistical Analysis | Data Visualisation

Project Overview

Instacart, an online grocery store that operates through an app is considering a targeted marketing strategy. They have requested an initial exploratory analysis. The stakeholders want to know more information about their sales patterns, especially the variety of customers in their database along with their purchasing behaviours.

Tools

The full analysis was carried out using the following **Python libraries** in Jupyter Notebook with the results exported or copied into Excel:

- pandas | numpy | os | matplotlib pyplot | seaborn | scipy | plotly express

Data Analysis Tasks

✓ Considered data ethics when dealing with customer information.

✓ Organised project folders logically.

✓ Conducted descriptive and consistency checks after importing each data set.

✓ Cleaned and wrangled the data - corrected entries, removed duplicates, renamed columns, changed data types, dealt with missing values and mixed-type variables.

✓ Merged four data sets into one and worked out how to map data to the dataframe.

✓ Derived new variables using if-statements, user-defined functions, the loc() function and for loops.

✓ Created flags and summary columns of descriptive statistics using the groupby() function.

✓ Created new columns and flags using customer data to inform customer profiling.

✓ Produced and exported histograms, bar charts, line charts, scatterplots and

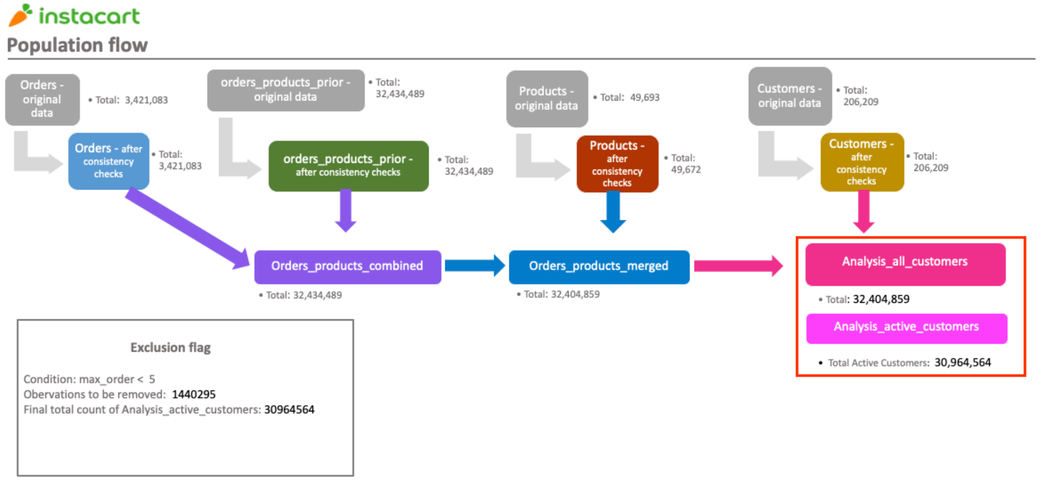
Data Source and Limitations

The project is based on an open-source data set from the Instacart Online Grocery Shopping Dataset 2017: orders_products_prior and orders with over 32 million rows; products and departments (the Instacart data dictionary). An anonymised sample of over 3 million grocery orders from more than 200,000 Instacart users.

Career Foundry created a fabricated customer data set, therefore parts of this analysis are fictitious.

The Instacart Population Flow Diagram

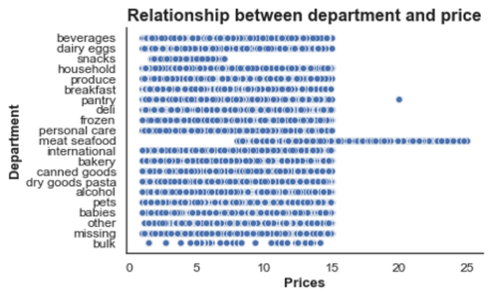
The grey boxes in the first row of the population flow diagram represent the four original data sets that were eventually merged into the final dataset Analysis_all_customers. The second row of coloured boxes represent the data sets after the cleaning process and the final row represents them after each merge.

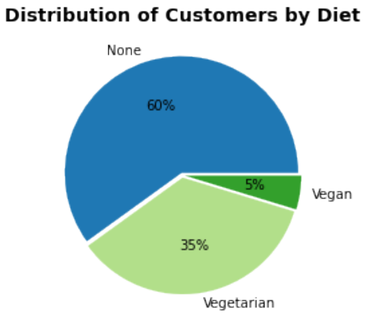


Data Analysis

Customer Profiling

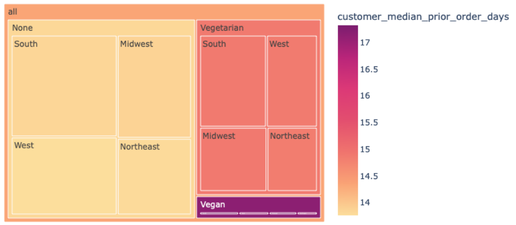
The diet flag classification was derived by grouping customers based on which departments they ordered from. The vegan group excludes the department with the highest prices 'meat seafood' and one of the most popular departments 'dairy eggs'. The Vegetarian group only excludes the 'meat seafood' department.

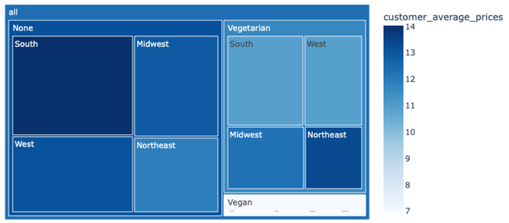




Within the deaprtments **only 'pantry' and 'meat seafood' have products above $15**. Prices of products in most departments range from $1.00 to $15.00 with the exception of the 'snacks' department $1.60 to $7.00 and bulk $1.4 to $14.

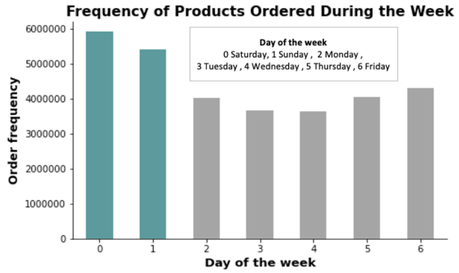
Download and then view the interactive treemaps showing the relationship between diet and region with customer median prior order days (orange) and average prices (blue).



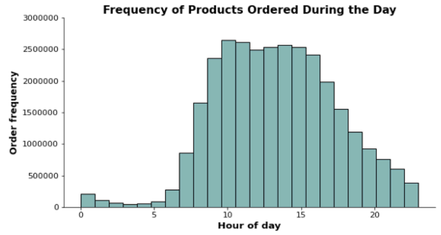


➡ There were clear differences between the diet classification groups. The Vegan group on average wait more than 2 weeks (17 days) to place an order, they also spend less ($7) and earn less ($73, 518) than the other groups in this classification.

Sample of Key Findings and Recommendations

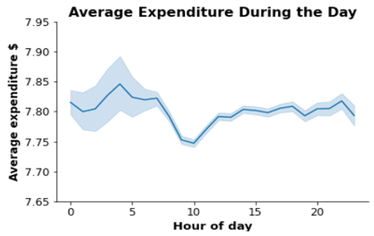


The weekend, **Saturday (day 0) and Sunday (day 1), are the busiest days of the week**. the least busy days are around the middle of the week Tuesday (day 2) and Wednesday (day 3).

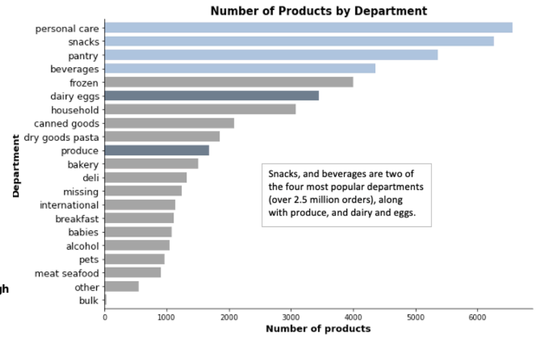


The histogram shows that **most of the orders are placed between 9 am and 4 pm** (around 2.5 million orders per hour). This peak tails off around 5 pm and there are fewer orders (below 5 million orders) between 11 pm and 6 am.

➡ Tuesday and Wednesday are the least busy days with fewer orders between 6 pm and 9 am, which could be considered a good time to schedule ads.



The average expenditure during the day is around 7.80 dollars. There is a slight decrease from **the high (7.85 dollars) around 4 am** to the low point (7.75 dollars) around 9 am. *[NB this chart was produced using a representative sample (70%) of the data.]*



➡ At 4 am customers spend slightly more money $7.85 (as opposed to $7.80). The lighter blue band shows uncertainty around this estimate ($7.79 to $7.89).

➡ Try to increase orders or reduce the number of items stocked in departments such as personal care, with low orders, but a large range of products.

The Instacart Analysis Final Report and Recommendations are archived in the GitHub project repository.

05 UK household food shopping

Advanced analytical analysis and dashboard design

Advanced analytical analysis and dashboard design

Machine Learning | Descriptive Analysis | Python | Tableau | Data Cleaning | Data Wrangling | Maths and Statistics | Data Transformation | Statistical Analysis | Data Visualisation

Project Overview

Poor diet within the UK is a public health issue and it is one of the worst in Europe for levels of obesity. Families buy more 'ultra-processed' food than other European countries; with these products accounting for more than half of all UK family food purchases. What are any relationships are there between 'ultra-processed' food, household expenditure and obesity.

Tools

The full analysis was carried out using the following **Python libraries** in Jupyter Notebook.

pandas | numpy | os | pyreadstat | gc | matplotlib pyplot | seaborn | scipy | | folium | json | sklearn | statsmodels api

Data Analysis Tasks

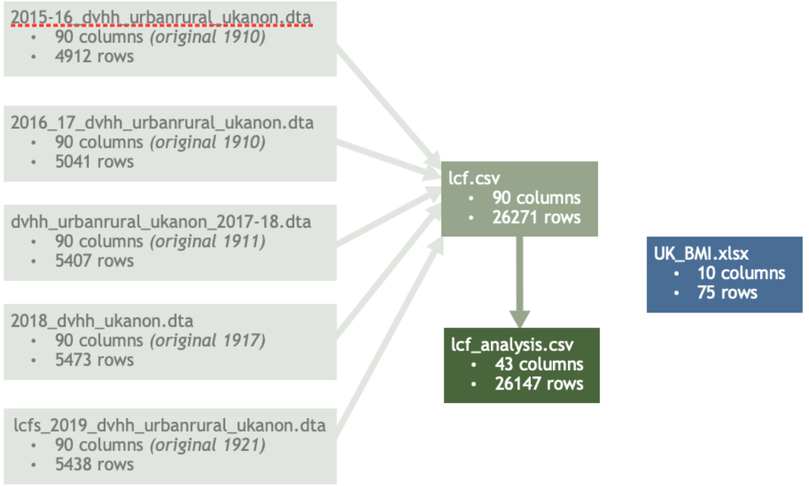
- ✓ Considered data ethics when sourcing, analysing and presenting.
- ✓ Organised project folders logically.
- ✓ Merged, cleaned and wrangled the data and conducted consistency checks after importing each data set.
- ✓ Performed an exploratory data analysis and formulated a hypothesis.
- ✓ Conducted a geospatial analysis using folium.
- ✓ Carried out machine learning regression and clustering analyses. Implemented a time series analysis.
- ✓ Produced and exported maps, box plots, pairplots, correlation matrices, scatterplots and stripplots using matplotlib pyplot and seaborn.
- ✓ Undertook an advanced analytical analysis for an interactive dashboard in tableau.

Data Source and Limitations

The project is based on data sets from voluntary surveys, the Living Costs and Food Survey and BMI data sourced from the UK regions. The majority of expenditure information collected relates to the fortnight immediately following the interview and may not accurately represent household spending in the population (see UK Household Food Shopping).

Data Sourcing and Merging

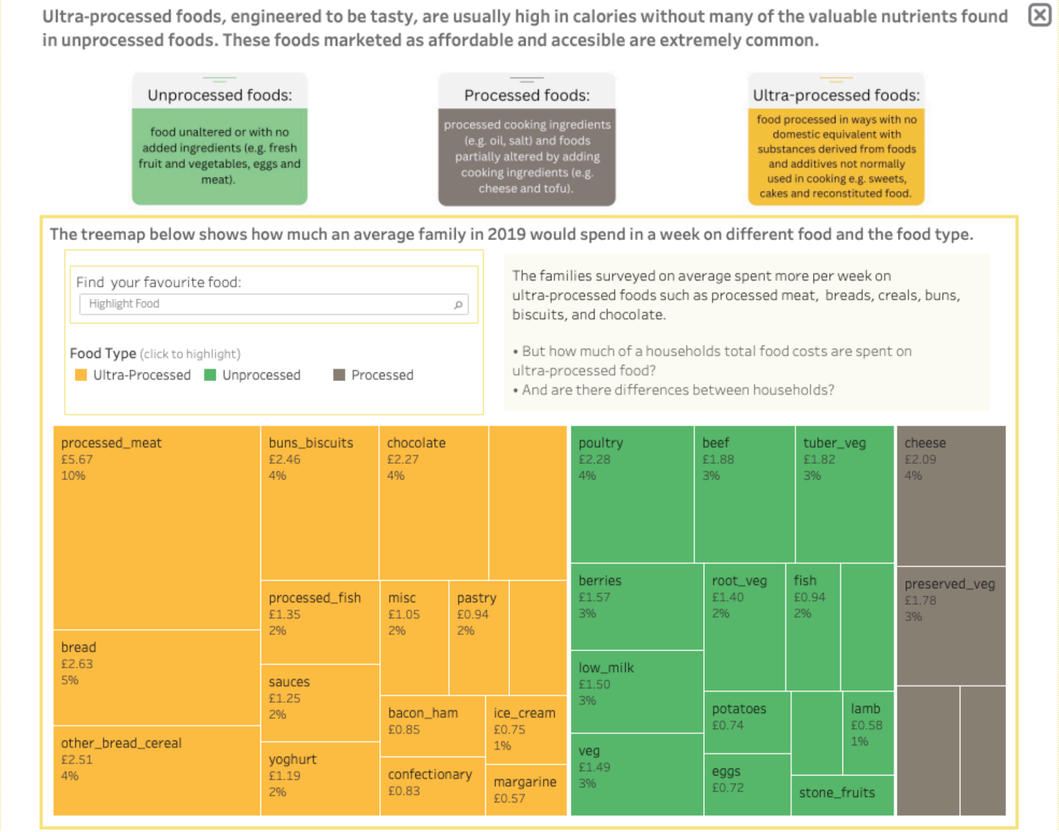
The five original data sets were merged into one dataset which was cleaned and wrangled. A second data set UK BMI contained the regional obesity levels. The UK Household Food Shopping document provides an overview of the data sets, the cleaning, wrangling and consistency checking process.



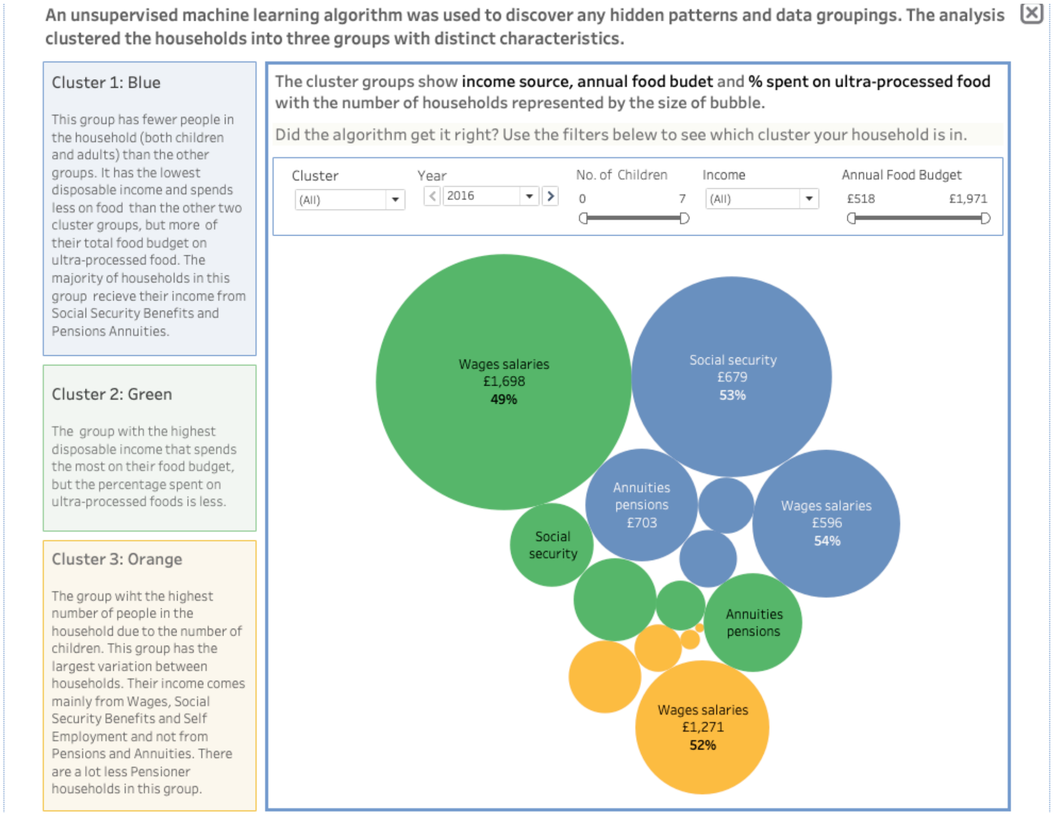
Data Analysis

Key Findings

- The average family spends over half their food budget on ultra-processed foods and the gap between spending on ultra processed and unprocessed is growing.
- The group with the highest disposable income spends the most on their food budget, but the percentage spent on ultra-processed foods is less.

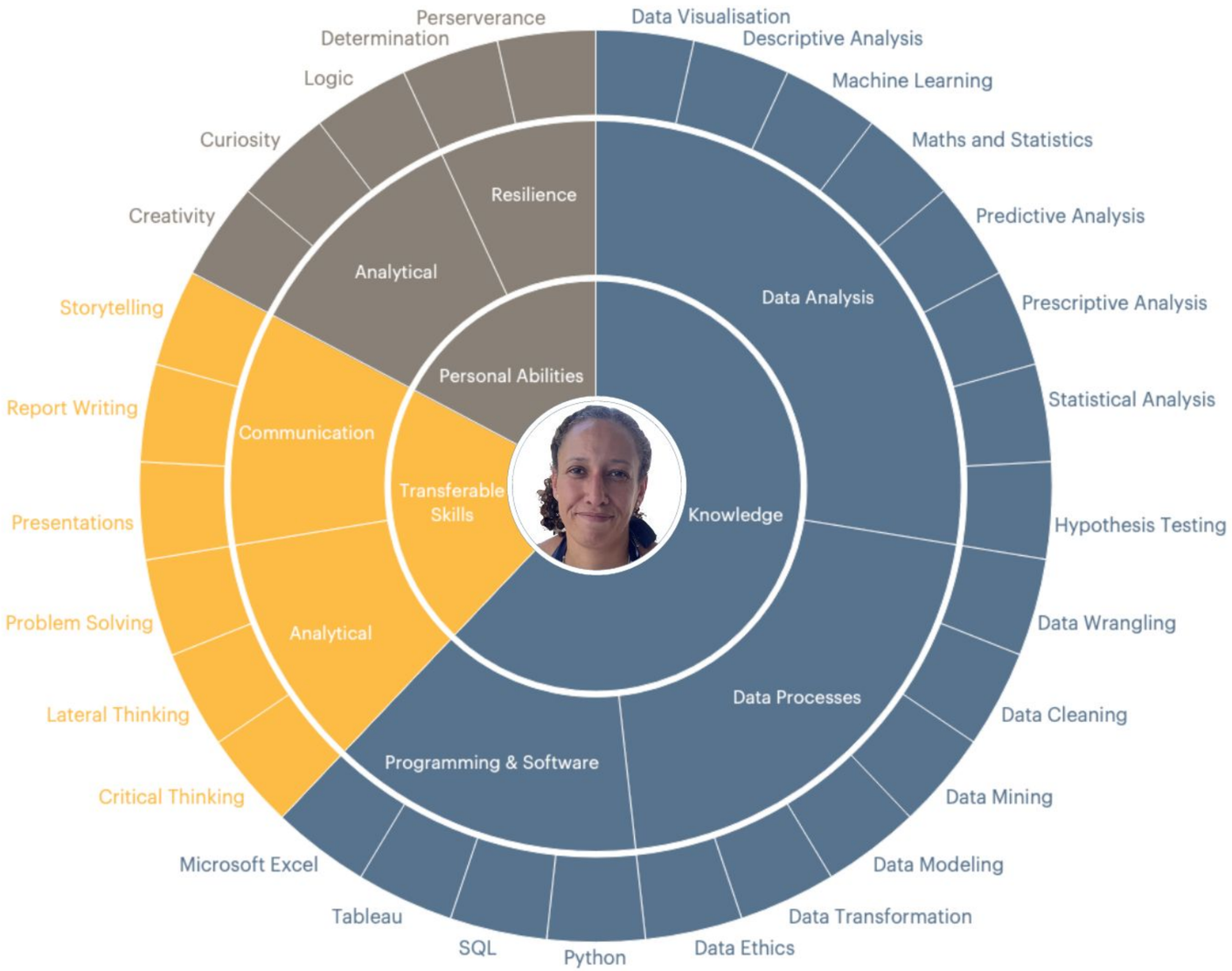


- The amount spent on ultra-processed food increases more in households on Social Security Benefits as disposable income increases. Another factor affecting this relationship is the number of children in the household.
- The group with the highest disposable income spends the most on their food budget, but the percentage spent on ultra-processed foods is less.
- The data is seasonal with the amount spent increasing in the run up to Christmas.



The UK Household Food Shopping Interactive Story in Tableau. The project is archived in the GitHub project repository.

Areas of Interest

Naturally curious data analyst who enjoys interrogating data, uncovering patterns and trends, and creatively explaining the story.

- Data Cleaning, Transformation and Wrangling
- Data Visualisation
- Storytelling
- Python