# CFG MOOC Challenge: Intro to Python for Data

Sydney Humayun

## Overview:

The purpose of this project is to analyse business and sales data using Python. The libraries used to complete this task are: Pandas and Matplotlib. Pandas is used to complete the statistical analyses while Matplotlib is used to create the graphs.

This project is part of the CodeFirstGirls MOOC series: Introduction to Python for Data. A four week “sprint” to introduce the topic occurred before this challenge was released. The data set is supplied from CFG so all teams have the same starting field. The challenge is set to be completed within 13 days as an individual or as part of a team. I have opted to challenge myself by completing this as an individual.

## Specifications and Design:

The requirements of the challenge include:

- Use of Pandas to read sales data
- Use functions to calculate the following:
    - Total sales for each product
    - Average sale price for each product category
    - Month with the highest sales
    - Month with the lowest Sales
    - Total money spent by the customers who spent the most money
- Write the results of the analysis to a CSV file

Optional extras teams could choose to include:

- Use matplotlib to create graphs
- Calculate additional metrics

## File Organisation and Justification:

Each topic (monthly comparisons, customer comparisons, and product comparisons) have their own python file for ease of understanding and information. All csv files are in a separate folder, as are all the charts and graphs. This is to maintain a clean and understandable organisation.

## Implementation and Execution:

### Approach:

The first days of the challenge were devoted to exploration of the data set and research into how the outcomes outlined above would be achieved. Mainly, pivot tables, bar charts, and group

by functions. The following days were spent implementing the plan and troubleshooting as needed.

## Tools:

Libraries: pandas, matplotlib
Analysis tools: group by, pivot tables, mean, min, max, sum, bar charts, pie charts,

## Challenges:

The largest hurdle to overcome was learning new tools and knowing where to find the information that was needed. After finding resources, learning how to implement them was another struggle. These were solved by re-reading the resources and learning properly how to perform group by functions and how the pandas library works (aggfunc and to_csv).

## Analysis and Discussion:

The investigation found that electronics had the highest average sale price out of the categories[1], most likely due to the high cost of iPhones and Samsung Galaxies. Followed closely by the Home category which also has some high ticket items[2].

The month that the highest gross amount of Sales by £ was February, closely followed by December[3]. These are two very high gift giving times in the year which may have contributed to the higher amount of sales.

The highest spender by far was Jane Doe with £7,299.87 spent throughout the year.[4]

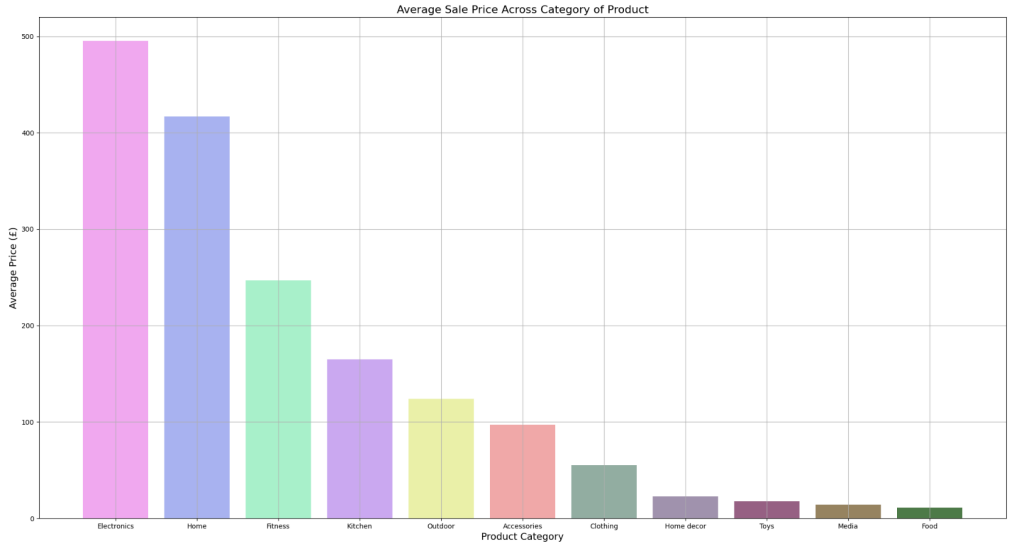
Further investigations and questions: what did Jane Doe spend the most on? What products were bought the most in which months? Which items are being bought the most often, and in what quantities? Could we determine statistical significance of the differences between variables?

## Conclusion:

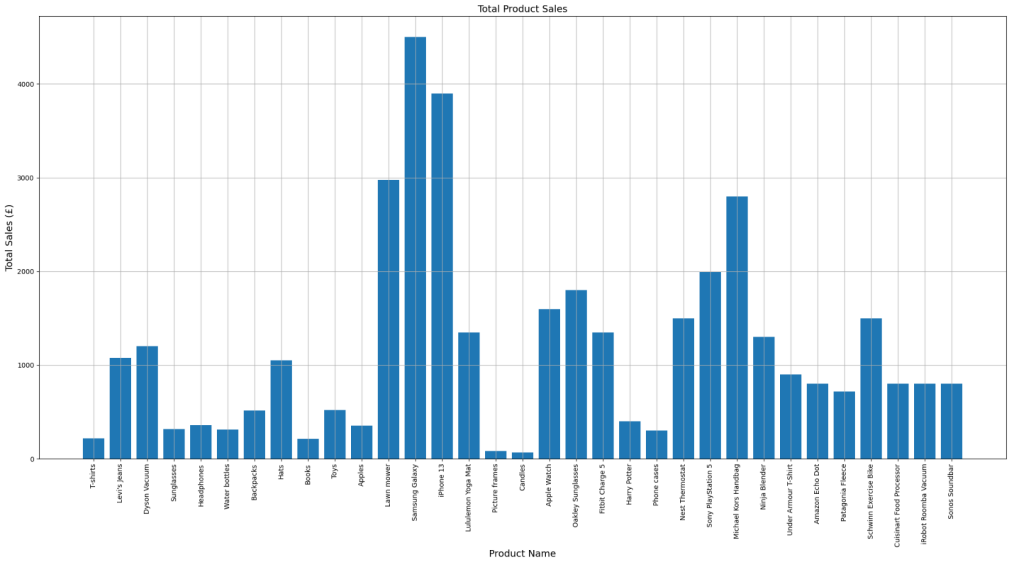
This was a fun introduction into data analysis using python. I learned many new tools and now have many new ideas to keep learning.

## Charts:

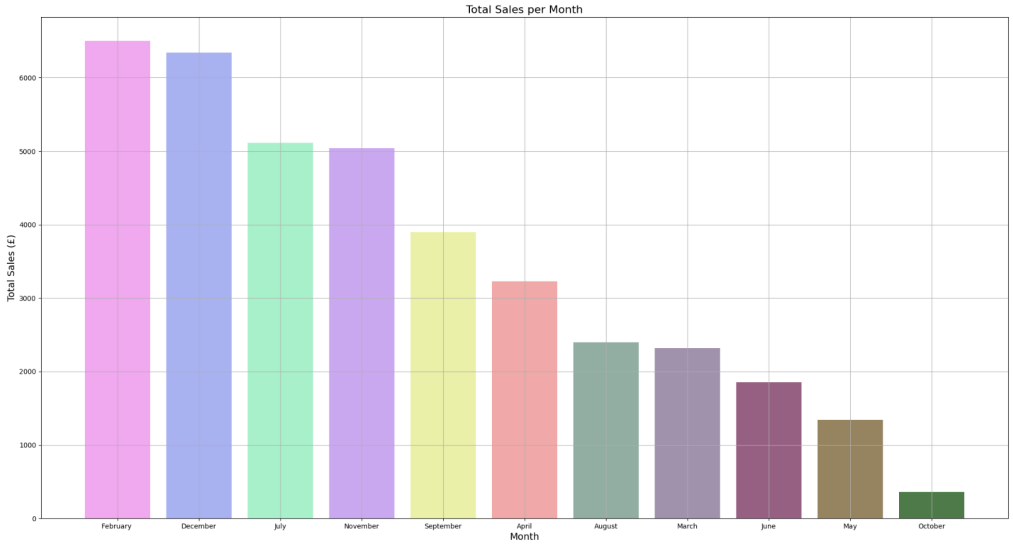
### [1] Average Sale Prices Across Category



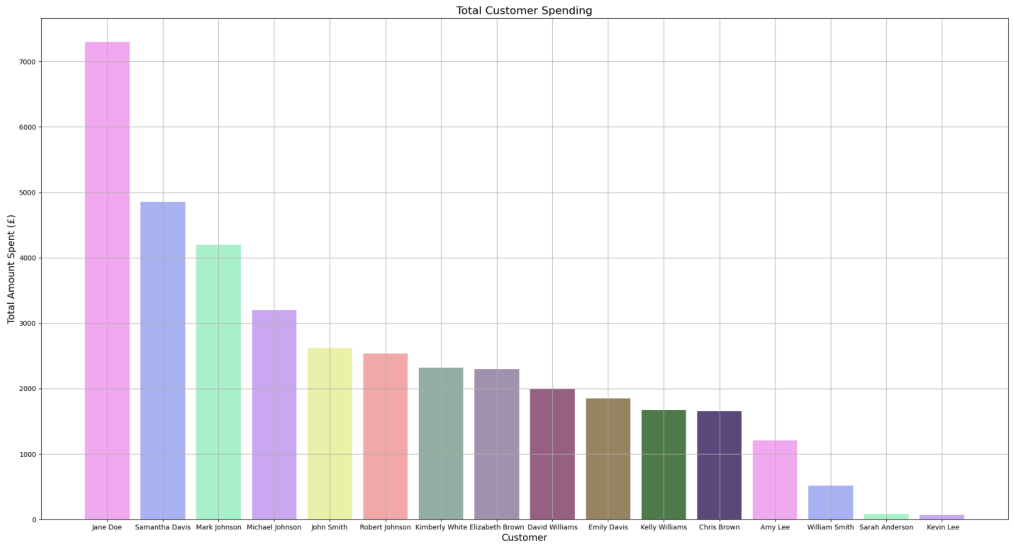

### [2] Total Product Sales



## [3] Total Sales per Month




## [4] Customer Spending



## [5, bonus] Percentage of Sales per Month