# Analysis & Visualisation Report

Employabilty and Transferable Skills

IN0008

Alexander Smolowitz
230065722

April 11, 2024

# Introduction

In this report I shall be. exploring the Company Sales Dataset, through analysing trends, observations and data visualisation. I aim to deliver an informed set of results, conclusions and recommendations by the end of this report, allowing one to gain an understanding of the dataset.

- Methodology for cleaning and analysing the data set, touching upon software and the resources consulted to achieve goals.
- Outputs and results of the above methodology, delving into the data and drawing meaning.
- Finally, discussing findings, as well as conclusions and observation, in order to outline recommendations for the company.

# Methodology

## 1.0 Python & Power BI

```python
import pandas as pd
csv = pd.read_csv('Company Sales Dataset.csv')
csv['Date'] = pd.to_datetime(csv['Date'], format='%d-%b-%y')
csv['Revenue'] = pd.to_numeric(csv['Revenue'].str.replace('$', ''))
# Standardising text data to lowercase and replacing blankspace
target1 = ['Product_Category', 'Sales_Channel']
for i in target1:
    csv[i] = csv[i].str.lower()
    csv[i] = csv[i].str.replace(' ', '-')
# Capping the data to counter extereme values
target2 = ['Sales_Amount', 'Revenue']
for i in target2:
    lower_bound = csv[i].quantile(0.01)
    upper_bound = csv[i].quantile(0.99)
    csv[i] = csv[i].clip(lower_bound, upper_bound)
# Interpolation
rating_av = round(csv['Customer_Rating'].mean(), 0)
csv['Customer_Rating'].fillna(rating_av, inplace=True)
csv.fillna('unknown',inplace=True)
# Summary Statistics
target3 = ['Sales_Amount', 'Units_Sold', 'Revenue', 'Customer_Rating']
measures = csv[target3].describe()
# Correlation
target4 = ['Units_Sold', 'Revenue']
cor = (csv[target4].corr('pearson'))
#  finding the mean and sum of the columns by product category
group_mean = csv.groupby('Product_Category')[target3].mean()
group_sum = csv.groupby('Product_Category')[target3].sum()
# Saving dataframe to csv for later use
csv.to_csv('data.csv')
measures.to_csv('measures.csv')
cor.to_csv('cor.csv')
group_mean.to_csv('group_mean.csv')
group_sum.to_csv('group_sum.csv')
```

Listing 1: *Python3 code used to clean data, main.py*

## 1.1 Data Cleansing

For this step I utilised python together with the panda library in order to cleanse the data. I began with correcting the datatypes in the CSV. The date was formatted using the panda to_datetime() function, however, another incorrect datatype was the revenue column in the CSV, which was in a string[1] format and could not be changed due to the presence of a $ sign. Thus, I removed the $ sign before revenue using the python replace() function, once the dollar sign was removed from the CSV, I was then able to use the panda to_numeric function.

Next, I aimed to standardise all the strings format in the CSV to have no whitespace and be in lower case. To achieve this, I targeted the two-remaining string-based columns in the data set, using the python lower() function which sets all letters to lowercase, and the replace() python function from earlier to replace the whitespace with a '-'.

Finally, I aimed to cleanse extreme scores as well as replace empty data in quantitative columns using interpolation[2]. To cleanse the extreme scores, but retain the data integrity, I used a panda function, called clip, which takes two inputs as a range and replaces any values which exceed or are lower than that range with the given values. To get these values I got the upper and lower bound of each column using the python quantile() function, which allowed me to use clip() on each individual column with their respective upper and lower bounds.

Next, before performing interpolation I first identified where it was appropriate to use, which ruled out anything qualitative/text based. So that left customer ratings, which I executed through finding the average using the mean() function and then used the fillna() function to replace all blank cells in the customer rating column with the mean in an attempt to prevent skewing the average. Lastly, for text-based data I used the same fillna() function but instead replaced the mean with the text 'unknown'.

## 1.2 Data Analysis

The first step in this data analysis was gathering the measures of dispersion and measures of central tendency. I accomplished this using the panda describe() method, targeting the sales amount, units sold, revenue and customer rating column. This function returns the count, mean, standard deviation, minimum, maximum and quartiles.

In order to perform a correlational analysis, we first need to choose what correlation we are testing. Due to wanting to find a relationship between co-variables that shows direction, Pearson's correlation analysis was used. I performed this using the corr('pearson') function on the units sold and revenue.

Lastly, using the groupby() function to analyse the best performing category through methods such as mean() and sum(). The data for each of the prior analysis and cleaning were saved to csv using the to_csv() function for further analysis, as well as visualisation using power BI.

## 1.3 Power BI & Data Visualisation

The Power BI software was used to perform data visualisation, I imported the data from the CSV files which I had saved in the previous step and formatted the data into matrixes, tables and graphs. Power BI, to me, was the obvious choice due to the interactive reports which one can create, and it's ability to explore large data sets in an in depth manner, it also seamlessly allowed me to compare data across the different saved CSV.

[1]string - a data type used to represent text as opposed to numbers
[2]interpolation - the process of determining an unknown value within a sequence based on other points in that set

Thus, when it came to visualising, for example, the trend of sales amount over time and producing bar charts on the product categories, it was merely the actions of clicking new visual and selecting the relevant columns from the available data.

However, when it came to a histogram, there is no set method to creating/editing the bin[3] of the histogram in Power BI as a stand-alone. Thus, I used a Power BI app by PBIVizEdit.com, allowing me to create a histogram for customer ratings.

# Analysis

## 2.0 Measures

Taking a look at the measures of dispersion and central tendency saved to a csv, we can see that this data is not so meaningful to us as it is, however, we can draw some meaning. Firstly, Through the use of interpolation, we now have a complete mean and count for Customer Ratings, and the standard deviation of 1.3 seems to check out with the ratings being between 1 and 5, thus have little spread.

| Measure | Sales Amount | Units Sold | Revenue | Customer Rating |
|---|---|---|---|---|
| Mean | 10178 | 49.969 | 25558 | 3 |
| StD | 5605.9 | 28.4 | 14157 | 1.2 |
| Min | 690.7 | 1 | 1485 | 1 |
| LQ | 5348.7 | 25 | 13177 | 2 |
| MQ | 10110.5 | 50 | 25663 | 3 |
| UQ | 14949.2 | 75 | 37773.7 | 4 |
| Max | 19817 | 98.2 | 49510.2 | 5 |

Table 1: Measures (2 DP)

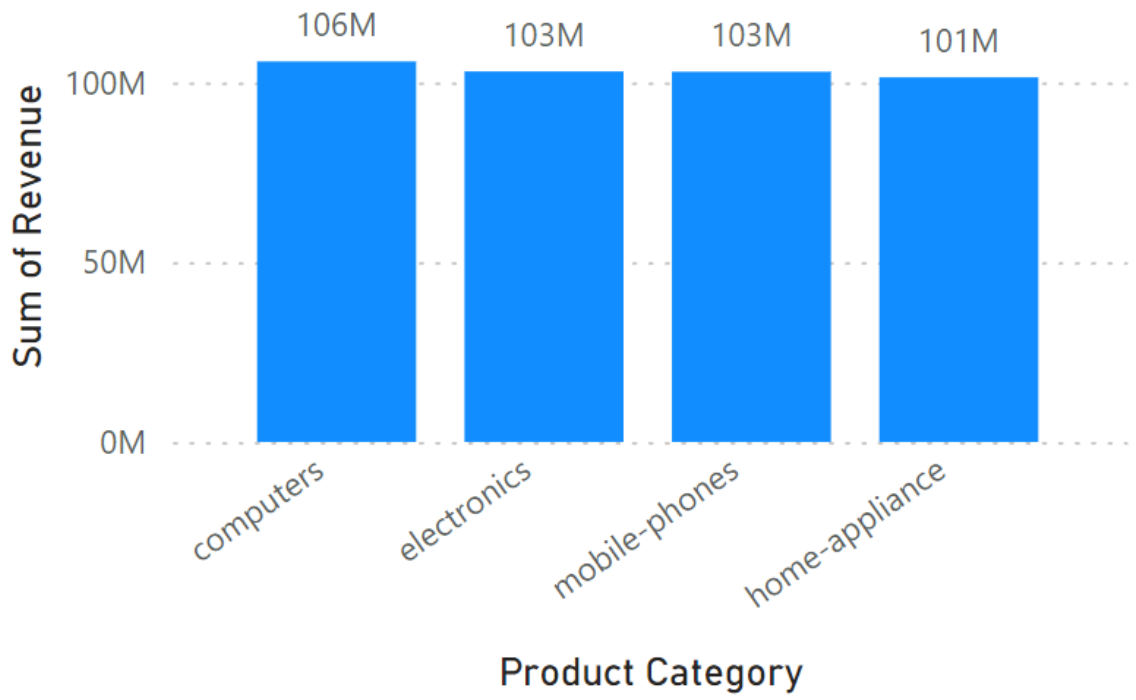
The mean units sold for the computer category is over an entire unit lower than that of home-appliances and mobile-phones, despite this, the mean value of revenue for the computer category is the largest, although only by a couple hundred, thus suggesting that computers have a higher turn over and are thus more profitable. However, this can also be as a result of trends in the use of computers and their demand over time.

[3]bin - grouping continuous data into a set of discrete intervals.

## 2.1 Trends

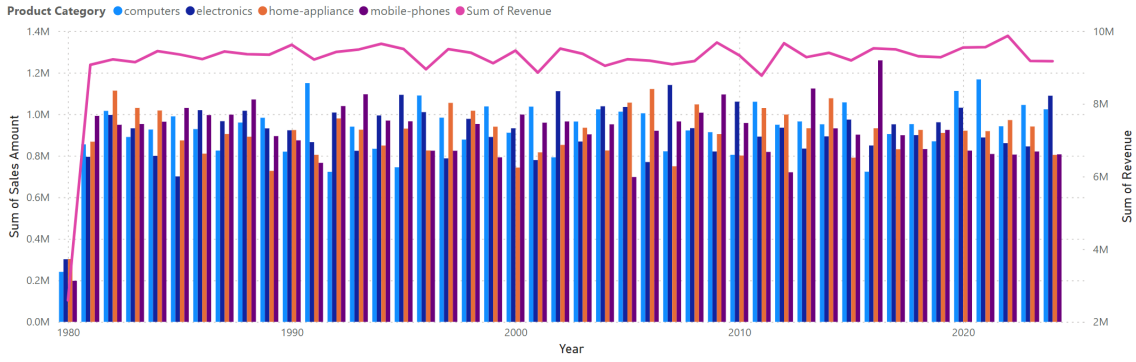
The disparity between the computer category and others widens as we take a look at the summed data. To further explore this, we shall take a look at the data visualised. In figure 1 we can see that computers have produced 3 million greater revenues than electronics and mobile-phones, as well as 5 million more revenue than home-appliances.

Figure 1: Sum of Revenue by Product Category



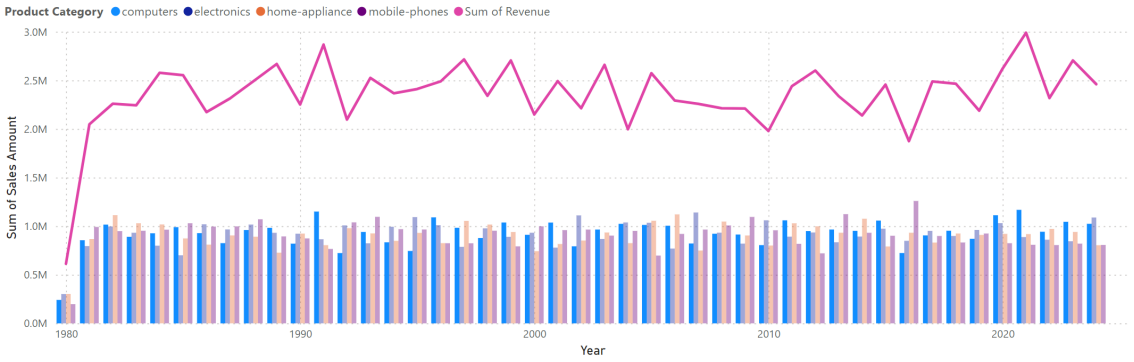

In figure 2 we can see that computers contribute to a large incline in profits between 2021 and 2023

Figure 2: Sales Amount and Revenue by Year and Product Category



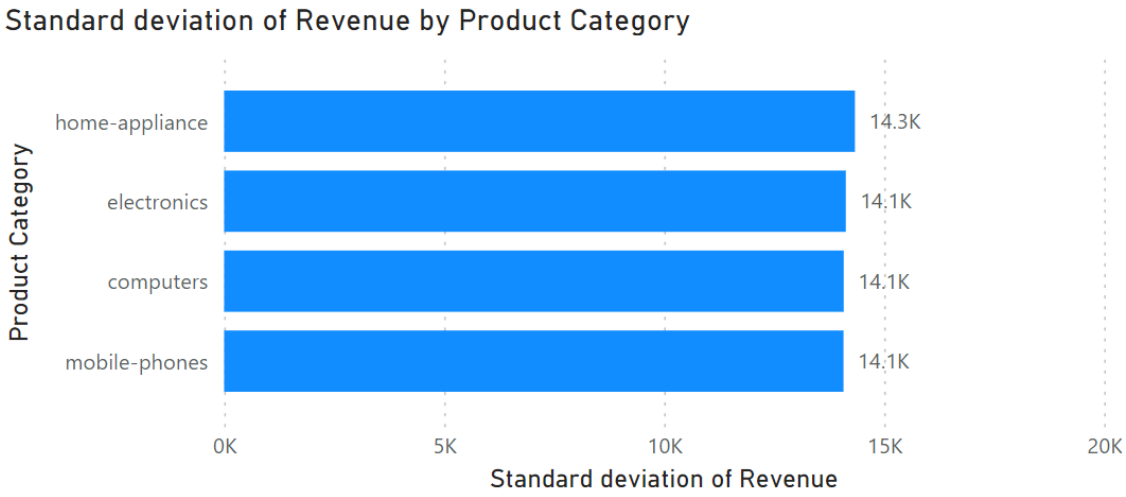
However, in figure 3, when looking at computers as a whole over a wide range of data we can see that it is quite volatile in its revenue, changing by a million each year, this is as a result of much lower or higher sales in such years, likely due to a decrease or increase in demand for computers.

Figure 3: Sales Amount and Revenue by Year and Product Category




In figure 4, when looking at the standard deviation (SD) of revenue by category, we can see that despite such observations, computers have a similar standard deviation to its sibling categories. This suggests that computers have a similar amount in fluctuation in it's profit as other categories focused on by the company, thus with its much larger potential for revenue is a more beneficial category to focus on as opposed to house appliances with their much greater standard deviation.

Figure 4




Moving on, we shall point out one of the drawbacks of using the mean for interpolation within customer ratings, and that is there is a distortion in the quantity of distinct ratings at the mean, which in turn skews data that aims to look at quantity of occurrences in relation to each rating.

The correlation between revenue and units sold is so low that it may as well be zero. This suggests several things, one which is the data integrity, but it also could be as a result of multiple factors contributing to this within the sales model. For example, large quantities of units could be sold at very low profit margins, such is the idea of selling in bulk, however such a method if done correctly would produce a large amount of overall revenue, thus suggesting that the profit margins could have been too low (Something we shall explore later on).

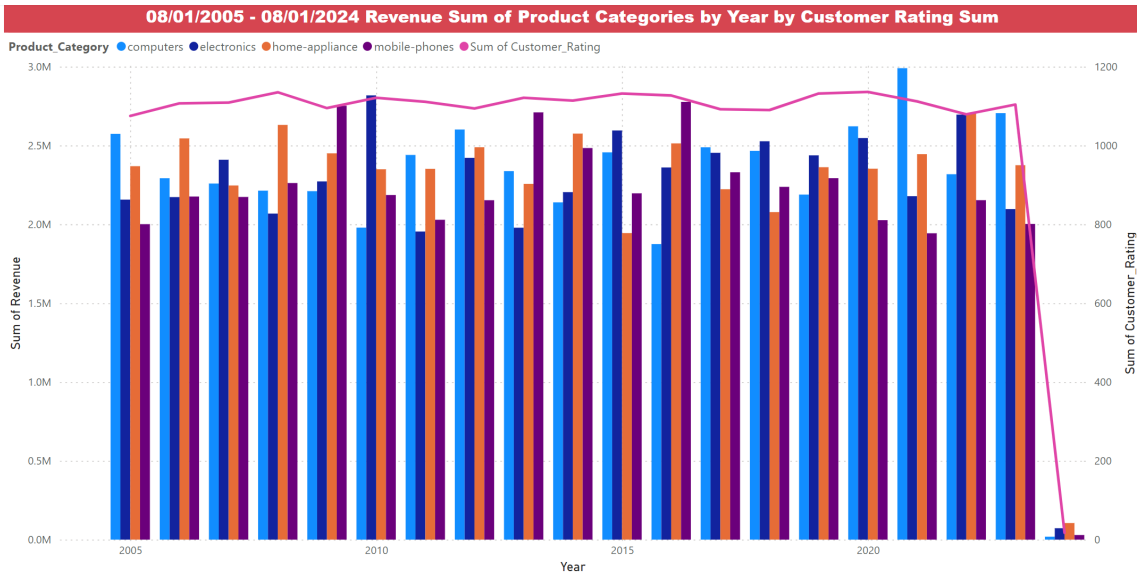
Another possibility is that such large amounts of units sold were not recorded at one data point but rather spread across multiple, meaning the revenue is not representative.

The potential factor that profit margins may have been quite low for large quantities of sales can be further solidified when comparing sales amount over time and revenue over time. We can see, in figure 3, the largest quantity of total sales, occurring in 1982, did not result in an exponential, or rather equivalent increase in total revenue for 1982.

# 2005 - 2024

For this personal exploration, I am going to focus on two large spikes, one in revenue between 2005 and 2024, as well as the category which I am interested in delving into, computers. Firstly,

Figure 5




in 2021 there is a very large increase in computer revenue where the sum revenue for the year is $2,990,525.29 and the total sales amount in the year equals to 1167319. We can find the supposed profit margin per product if we assume that the revenue in this data refers to net revenue after costs, so 2990525.29 divided by 1167319 which would give us 2.5656.

If we follow the industry standard of between 5 and 20 percent profit for a computer retailer, and the average price of computers which falls at around $630, then one should make at least a $31.5 profit at 5% and $126 profit at 20% margin. Now let's assume the company is currently selling the lowest end computer, which could at best run chrome operating system on, at what is likely around $300, then the profit should be at least at $15. If we consider that the current price is average ($630) and that the profit is $2.5656, then the profit margin would be 0.407%.

The is potential for extremely larger quantities of profit, in fact even if we go at a 10% profit margin for an average price of 630% the profit is still 25.54 times larger than the current profit per computer. Now let's look at this in the context of 2021 sales, 1167319 computer sales times $63 dollars profit per sale of computer which would equal $73,541,097. This means that the potential for profit, even at 10% is around 2353% of $2990525.29, At the lowest percentage the profit will still be at around $37.77 million.

This extremely low profit margin of 0.407% for computers can explain a lot about the lack of correlation between sales and revenue. If we take the total number of sales between 2005 and 2024 for computers, we get $45,175,894.51 in revenue and 18,182,179 sales. If the profit margin was at 5%, the lowest average, and the average computer price was again $630 then then the lost potential profit over the past 18 and a bit year is around $528 million dollars. That is 11.73 times larger than the current total profit for computers alone.

# Key Take Away

In summary, our data on customer ratings is skewed, however, when taking into account this we can see that better customer ratings shows a positive trend towards sales. Secondly, computers have a larger capacity for revenue than other categories as of recent years, this trend can be seen from 2020 to present.

Observations seem to show that external factors are a large contributor towards sales, for example the large increase in mobile-phones sales in 2016, which occurred at the same time the Chinese phone manufacturers, OnePlus, Xiaomi and Huawei, gained traction into the global market and all released multiple flagship models. This could have potential have played a large hand in the increase in smart phone sales for the company.

The profit margins of computers is astronomically low, at around 0.4% per sale, I would recommend that the company increase this profit margin, even if the sales model is in quantity of sales, as even a 1% increase would significantly increase net revenue by a large quantity.

As pointed out earlier, according to LinkUp Digital and other various sources, the current lowest average profit margins in the computer retail sector is between five and twenty percent, meaning that even if the company decides to increase the profit margin by 3 – 10 percent they will still be considered a lower price option by potential customers.

Furthermore, if the profit margin is within this average range and the net revenue does not increase exponentially, the company should consider increasing the price of products to the customer, as such a trend would suggest that the cost of goods sold (COGS)[4] is much too high compared to the price at which they are being sold.

Finally, I would suggest researching the average retail price of other categories as well as the average profit margin within the market to make an informed decision on the final selling price. The procedure should take these factors into account into the mark up of products, by first researching the cost of goods sold for the company (COGS), and then separately adding on the desired revenue to that purchase price.

# Refrences

R.Goldsborough, 'Profit Margins of the Makers of PCs and Handheld Devices', linK:www.infotoday.com
Statista, 'Average selling price of personal computers (PCs) worldwide', link:www.statista.com
Pydata, User guide, link:pandas.pydata.org

[4]cost of goods sold (COGS) - the sum of all direct costs associated with making a product. It appears on an income statement and typically includes money mainly spent on raw materials and labour