RETAILER PROFITABILITY

A CASE STUDY

CONTENT

01 OVERVIEW

02 DATA CLEANING

03 DATA EXPLORATION

04 CORRELATION ANALYSIS

05 HYPOTHESIS ANALYSIS

06 CONCLUSIONS

07 DATA LIMITATION AND NEXT STEPS

08 FINAL THOUGHTS

01 OVERVIEW

BACKGROUND:

The Superstore **retailer** has been operational across the USA since 2014, offering a diverse range of products including **furniture**, **office supplies**, and **technology items**.

GOAL:

Explore the company's **performance** and uncover opportunities for **increased profitability**.

DATA:

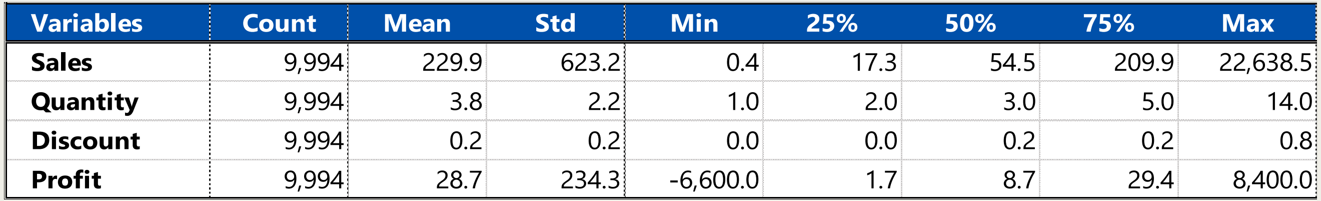
The Superstore Dataset (2017)
Source: Kaggle, accessed 09.12.2023

02 DATA CLEANING

As part of the data cleaning process:

- No **missing** nor **duplicate** values were found
- Some **column types** were updated to allow for correct analysis
- Column ‘Customer Name’ **dropped** due to **data privacy**
- Column ‘Country’ dropped as only one value (USA)

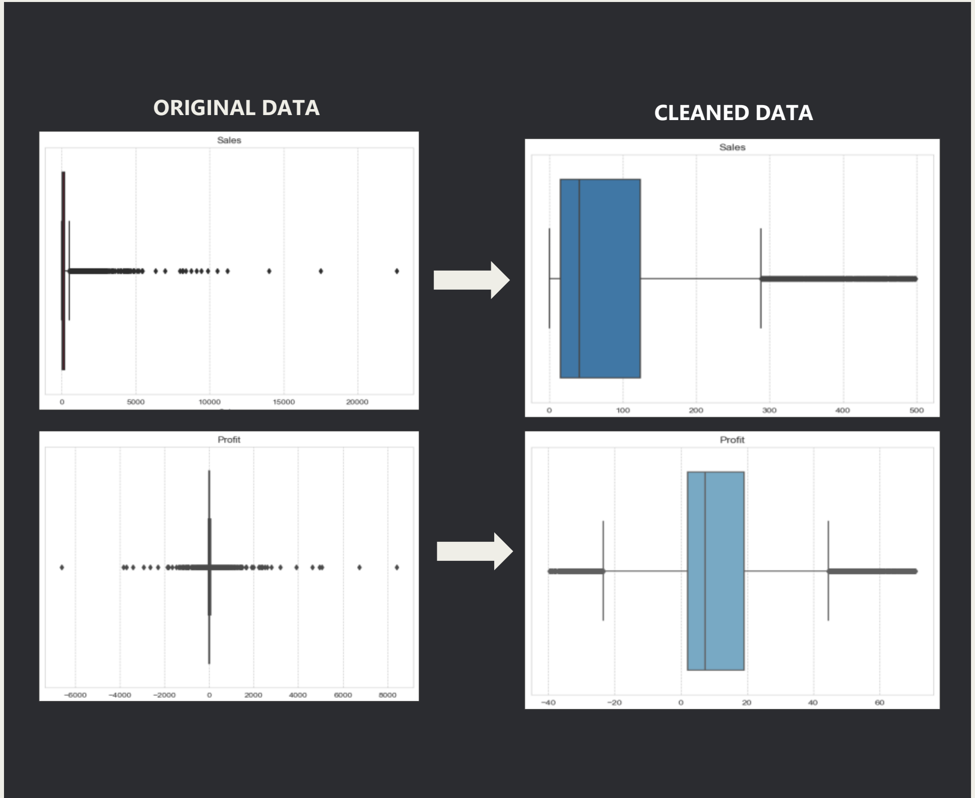
Some basic statistics showed me that the **Sales** and **Profit** variables needed further **investigation**, due to the possibility of **outliers** (for example, max. much higher than 75%)



Variables	Count	Mean	Std	Min	25%	50%	75%	Max
Sales	9,994	229.9	623.2	0.4	17.3	54.5	209.9	22,638.5
Quantity	9,994	3.8	2.2	1.0	2.0	3.0	5.0	14.0
Discount	9,994	0.2	0.2	0.0	0.0	0.2	0.2	0.8
Profit	9,994	28.7	234.3	-6,600.0	1.7	8.7	29.4	8,400.0

02 DATA CLEANING

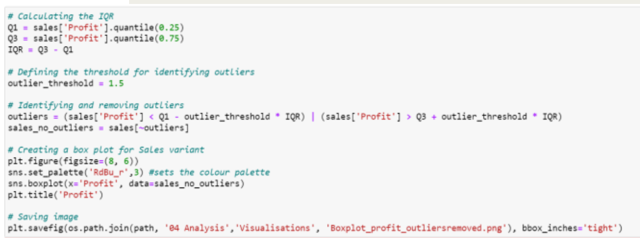
Sales and Profit **outliers** were removed (with +/- 1.5 IQR):



But even after outliers were removed, there were still a lot of **extreme values** for both **Sales** and **Profit** variables.

I've decided to **keep these**, as they were high but possible sales.
(ex: 160USD ergonomic office chairs)

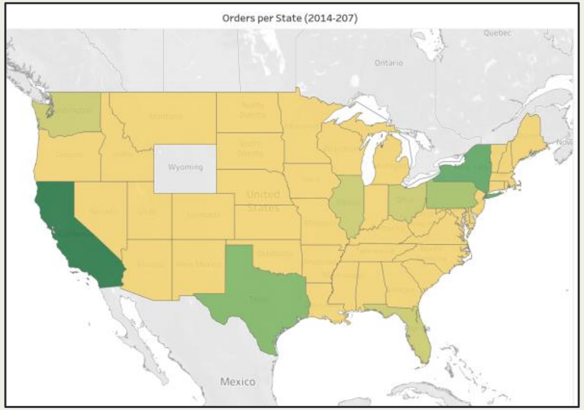
# just an example of the coding involved...



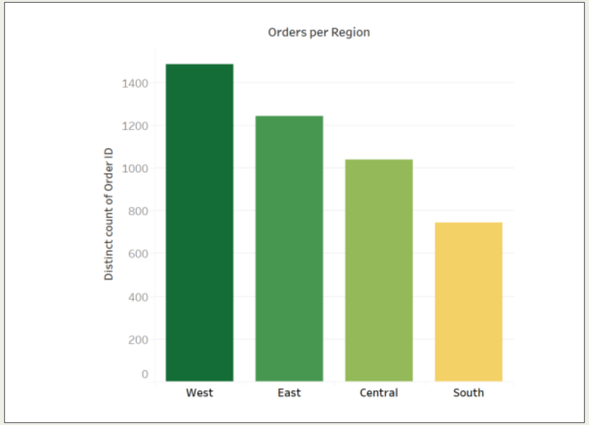
```
# Calculating the IQR
Q1 = sales['Profit'].quantile(0.25)
Q3 = sales['Profit'].quantile(0.75)
IQR = Q3 - Q1

# Defining the threshold for identifying outliers
outlier_threshold = 1.5

# Identifying and removing outliers
outliers = (sales['Profit'] < Q1 - outlier_threshold * IQR) | (sales['Profit'] > Q3 + outlier_threshold * IQR)
sales_no_outliers = sales[~outliers]

# Creating a box plot for Sales variant
plt.figure(figsize=(8, 6))
sns.set_palette('RdBu_r',3) #sets the colour palette
sns.boxplot(x='Profit', data=sales_no_outliers)
plt.title('Profit')

# Saving image
plt.savefig(os.path.join(path, '04 Analysis','Visualisations', 'Boxplot_profit_outliersremoved.png'), bbox_inches='tight')
```

03 DATA EXPLORATION

As quick visualisation of the dataset revealed that between Jan-2014 and Dec-2017:


Orders per State (2014-207)

There were orders placed across the **all USA States**, except Wyoming,

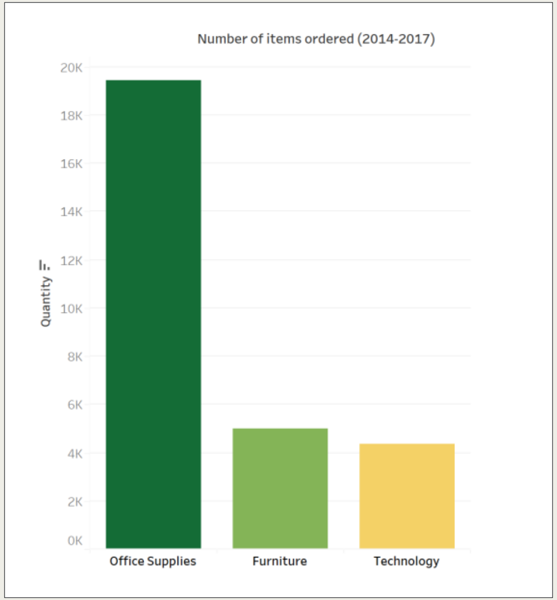

Orders per Region

with **West** and **East** placing most of the **orders**.

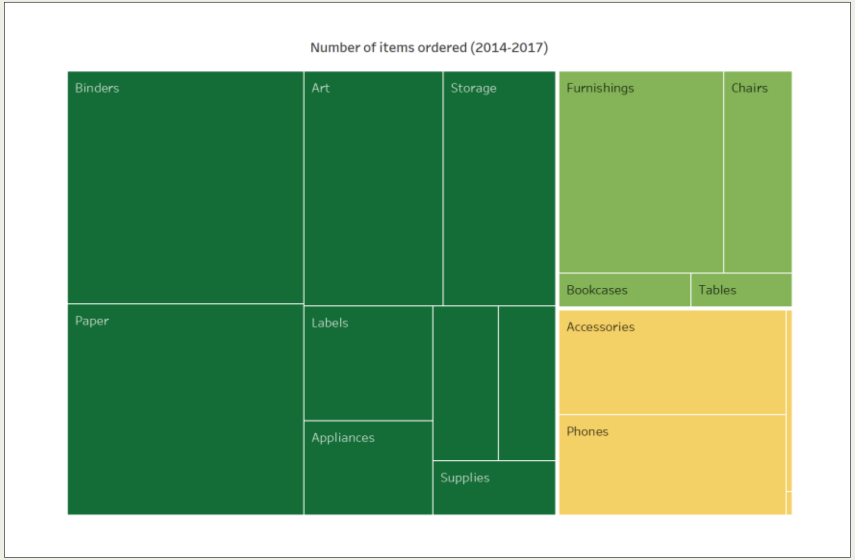
And the big majority of orders were **delivered** via **Standard Class**.

03 DATA EXPLORATION

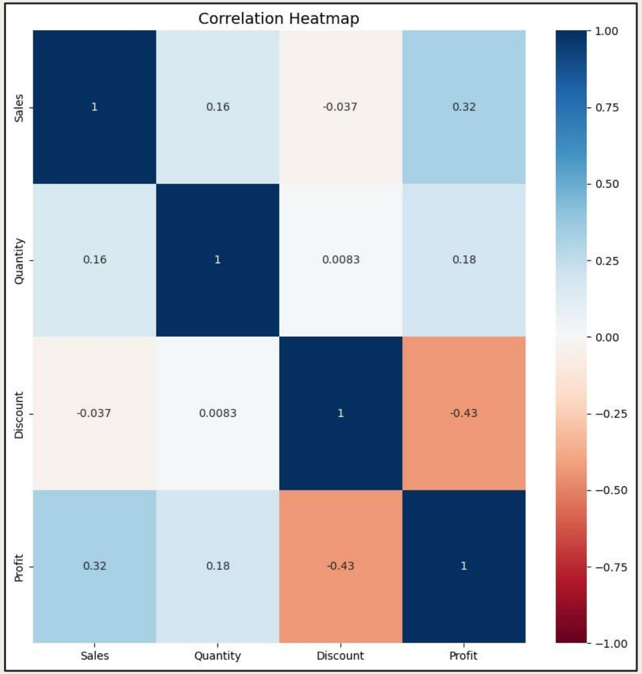
Office Supplies are by far the **biggest sellers**,



with **binders**, **paper** and **furnishing** being the **top 3** types of items **ordered**.



04 CORRELATION ANALYSIS



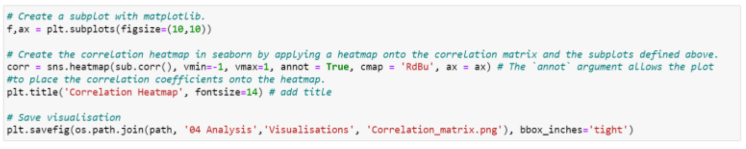
A heatmap showed a:

- **positive correlation** between Sales and Profit
- **negative correlation** between Discount and Profit
- **weak positive correlation** between Quantity and Sales
- **weak positive correlation** between Quantity and Profit

The positive correlation between **Sales** and **Profit** is understandable: more sales tend to lead to bigger profits

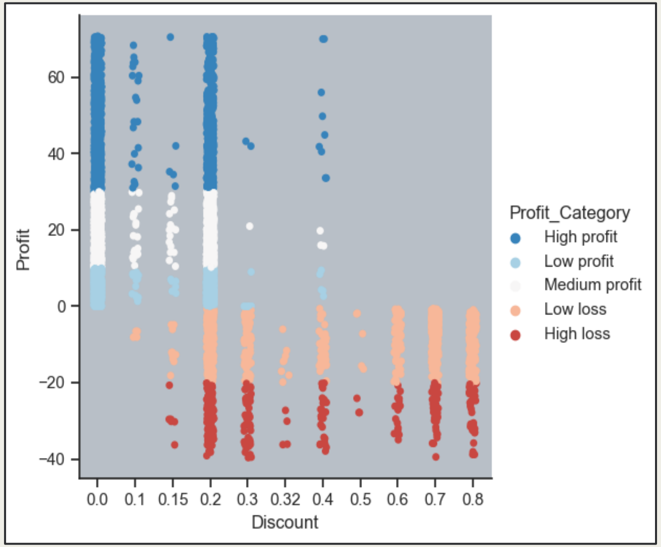
Based on the results, I was interested in understanding better how the variables **Discount** and **Profit** are interconnected and how discounts might be related to losses.

# some matplotlib:



```
# Create a subplot with matplotlib.
f,ax = plt.subplots(figsize=(10,10))

# Create the correlation heatmap in seaborn by applying a heatmap onto the correlation matrix and the subplots defined above.
corr = sns.heatmap(sub.corr(), vmin=-1, vmax=1, annot = True, cmap = 'RdBu', ax = ax) # The `annot` argument allows the plot
#to place the correlation coefficients onto the heatmap.
plt.title('Correlation Heatmap', fontsize=14) # add title

# Save visualisation
plt.savefig(os.path.join(path, '04 Analysis','Visualisations', 'Correlation_matrix.png'), bbox_inches='tight')
```

04 CORRELATION ANALYSIS

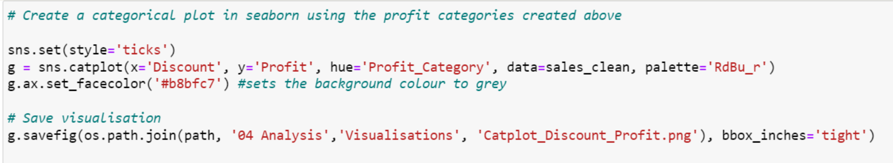


To better visualise the relationship between discount and profit, I've derived a new 'Profit Category' column based on the Profit distribution.

I've then further looked at the correlation between Profit and Discount:

- **Higher discounts** are linked to **lower profits**/bigger losses.
- Items with **any profits** are items with discounts mostly **up to 20%**
- Items with **discounts of 30%** or more **almost never** result in profits.
- Items with **discount of 50%** of more **never** result in profits.

# and another example, now with seaborn:



```
# Create a categorical plot in seaborn using the profit categories created above

sns.set(style='ticks')
g = sns.catplot(x='Discount', y='Profit', hue='Profit_Category', data=sales_clean, palette='RdBu_r')
g.ax.set_facecolor('#b8bfc7') #sets the background colour to grey

# Save visualisation
g.savefig(os.path.join(path, '04 Analysis','Visualisations', 'Catplot_Discount_Profit.png'), bbox_inches='tight')
```

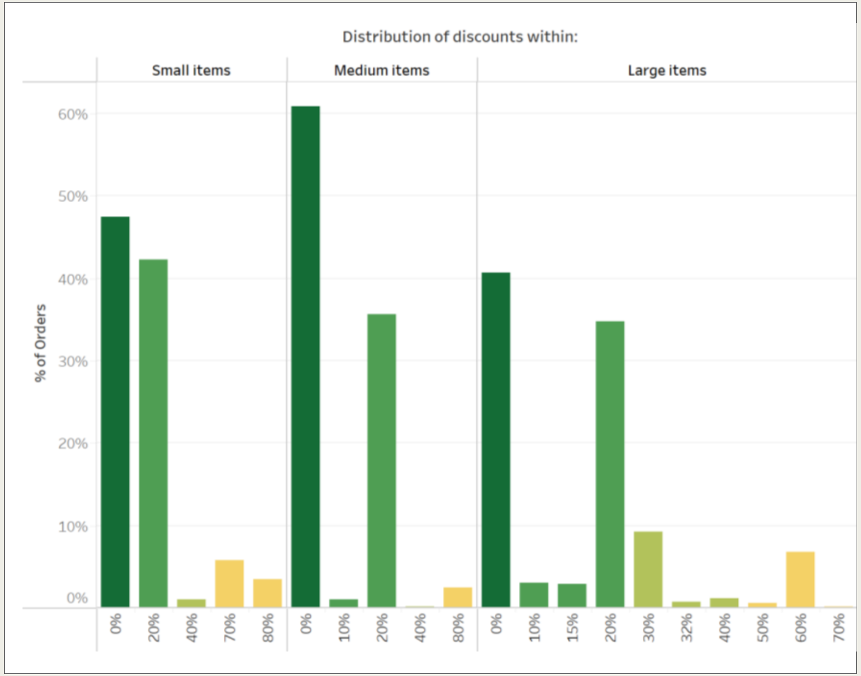
05 HYPOTHESIS ANALYSIS

One hypothesis was that **larger discounts** were being offered to alleviate **warehouse capacity** issues by selling **larger items**.

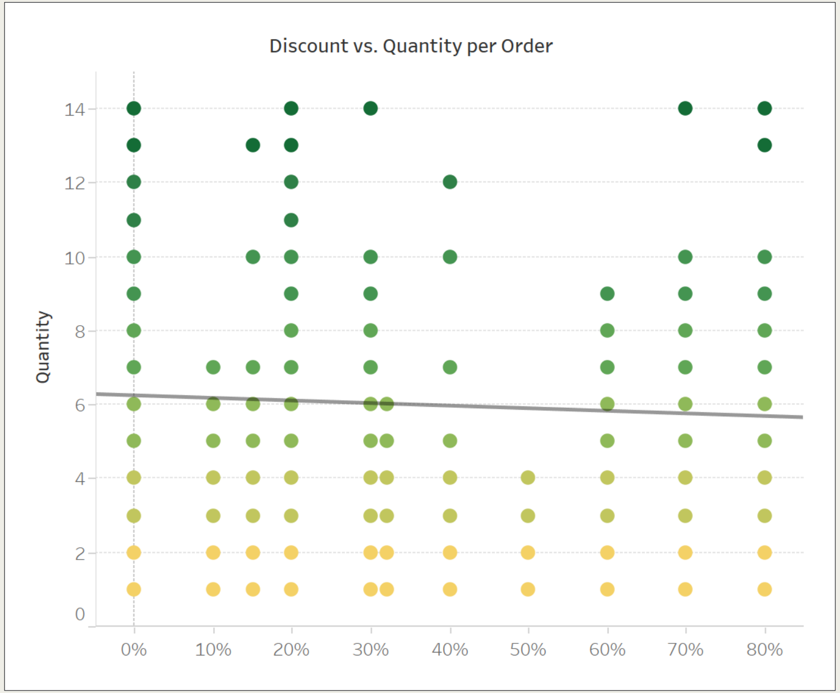
To explore this, I **segmented items into subgroups** based on their **typical size**, using **Tableau**.

I then examined how **discounts were distributed across different sized items**.

The visualisation revealed a **consistent application of discounts across items of varying dimensions**, which challenged the hypothesis proposed.



05 HYPOTHESIS ANALYSIS



Another hypothesis was that Superstore applies **discounts** to **boost the basket size**.

However, there appears to be **no correlation** between **discount and quantity ordered**.

Offering larger discounts does not result in larger order quantity.

06 CONCLUSIONS

Although **number of customers**, **number of orders**, **sales** and **profit** has been **increasing year-on-year**, there is a clear **potential for better performance**.

- The **median profit per order** is only 11.60 USD.
- 16% of all orders **incur in loss**.
- **Larger discounts** are positively correlated with **lower profits**.

Even though **discounts** can be used as a **business strategy** (attraction of new customers, increase of basket size, etc.), this strategy should be **re-evaluated** due to the **high losses** the company is incurring on discounted items.

LINKS:
GitHub
Tableau

07 DATA LIMITATION AND NEXT STEPS

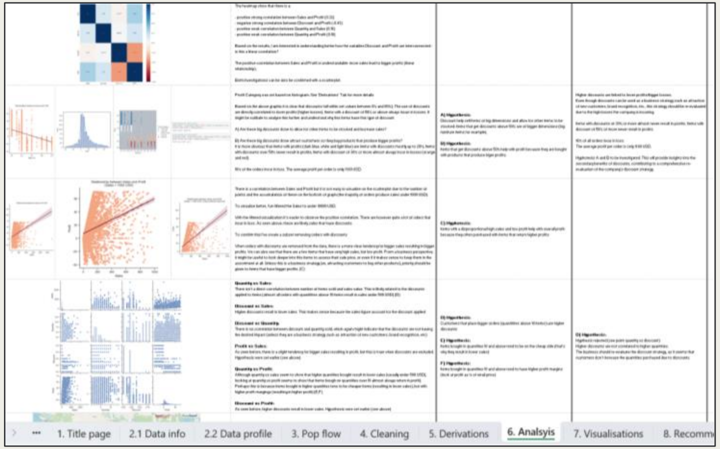
There have been numerous **external interventions**, such as alterations in the **discount strategy** over the period under analysis. To extract **valuable insights** from the data, it is crucial to **comprehend the business decisions** undertaken during distinct timeframes.

Additionally, it's worth noting that the available data only extends up to **December 2017**, potentially limiting its relevance to **more recent years**.

Next steps in the analysis could include:

- **Customer segmentation**.
- **Sales trends** by region, segment, or product category.
- **Profit margin** analysis by product category or sub-category.
- **Demand forecasting** for different product categories.
- **Basket analysis** to identify items frequently bought together.

08 FINAL THOUGHTS



Because I wasn't working on this project full-time, one of the biggest challenges was constantly remembering what I had already worked on, where I left off, and where to pick it up from.
To tackle this, I started keeping a project log with notes on the steps taken, results obtained, questions that arose, and the next steps to follow. It's been incredibly helpful, and I'll definitely be using this approach for future projects! (see a snippet on the side)

I had to create multiple derived columns for the analysis, both in Python and later in Tableau. It was a great exercise that helped me sharpen my skills.

Throughout the project, I relied on resources like Stack Overflow, the Tableau Forum, various Medium articles, and an AI bot to push myself to explore new types of analysis, visualisations, and aesthetics.