Supermarket Sales – A Case study

Explorative Data Analytics

By Sharad Nalawade

Data Context

Data generated through business transactions is the reflection of the overall performance of the underlying business model and its operations. Every super-market generates terabytes of data daily that hide crucial patterns that provide key insights into their business's performance. A few examples are the patterns related to sales and revenue trends, customer buying habits, employee productivity gains over time, market share, etc.

But the data per se is dumb and carries no meaningful information unless it is collected, organized, explored, and analysed. Data Science is now an established engineering practice that unravels the hidden patterns from a vast amount of data. It's rightly said that data science is all about *torturing the data until it confesses*.

There are many tools like *Tableau, Power BI, R, Python* that help data scientists to apply algorithms to understand various aspects of what the data is

trying to communicate. But one of the strategic applications of the data science is business decision making. While AI/Machine Learning algorithms help business in predictions, associations, and clustering large amount of data to help them understand the overall trend and forecast the future, the decision makers need an easier and visual method to understand the insights from the data. One key discipline for this critical requirement is **Data Visualization**.

Explorative data analysis is all about applying various techniques to explore and characterise datasets. Several statistical models can be applied to learn about the data profile, bias or variance and other key useful stats that reveal about the nature of the data.

Data visualization also play an important role in analysing the data using graphical representations.

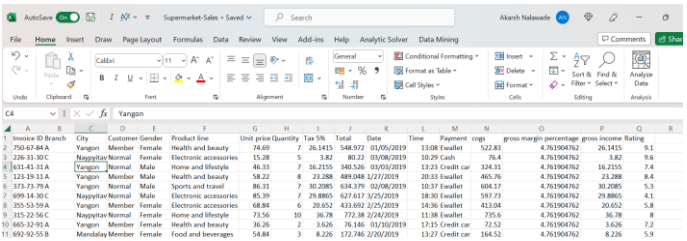
Reference:

https://en.wikipedia.org/wiki/Exploratory_data_analysis

Superstore Dataset:

Supermarket-Sales.xlsx

A typical dataset like **Supermarket-Sales** captures essential features that are useful for data analytics. This may be CVS or an Excel file. For better analysis, the dataset should be as large as possible in terms of rows.



Invoice ID	Branch	City	Customer	Gender	Product line	Unit price	Quantity	Tax 5%	Total	Date	Time	Payment	cogs	gross margin percentage	gross income	Rating
750-67-84	A	Yangon	Member	Female	Health and beauty	74.69	7	26.1415	548.972	01/05/2019	13:08	Ewallet	522.83	4.761904762	26.1415	9.1
226-31-30	C	Naypyitaw	Normal	Female	Electronic accessories	15.28	5	3.82	80.22	03/08/2019	10:29	Cash	76.4	4.761904762	3.82	9.6
631-41-31	A	Yangon	Normal	Male	Home and lifestyle	46.33	7	16.2155	340.526	03/03/2019	13:23	Credit card	324.31	4.761904762	16.2155	7.4
123-19-11	A	Yangon	Member	Male	Health and beauty	58.22	8	23.288	489.048	1/27/2019	20:33	Ewallet	465.76	4.761904762	23.288	8.4
373-73-79	A	Yangon	Normal	Male	Sports and travel	86.31	7	30.2085	634.379	02/08/2019	10:37	Ewallet	604.17	4.761904762	30.2085	5.3
699-14-30	C	Naypyitaw	Normal	Male	Electronic accessories	85.39	7	29.8865	627.617	3/25/2019	18:30	Ewallet	597.73	4.761904762	29.8865	4.1
355-53-59	A	Yangon	Member	Female	Electronic accessories	68.84	6	20.652	433.692	2/25/2019	14:36	Ewallet	413.04	4.761904762	20.652	5.8
315-22-56	C	Naypyitaw	Normal	Female	Home and lifestyle	73.56	10	36.78	772.38	2/24/2019	11:38	Ewallet	735.6	4.761904762	36.78	8
665-32-91	A	Yangon	Member	Female	Health and beauty	36.26	2	3.626	76.146	01/10/2019	17:15	Credit card	72.52	4.761904762	3.626	7.2
692-92-55	B	Mandalay	Member	Female	Food and beverages	54.84	3	8.226	172.746	2/20/2019	13:27	Credit card	164.52	4.761904762	8.226	5.9

The first step is to pre-process the dataset to eliminate the following anomalies:

- Missing values
- Empty rows
- Bias
- Variance

Once the dataset is cleaned, The second step is to learn its statistical profile like min, max, standard deviation of

each column. This is where tools like Python, R, Power BI, Tableau, etc. can be used.

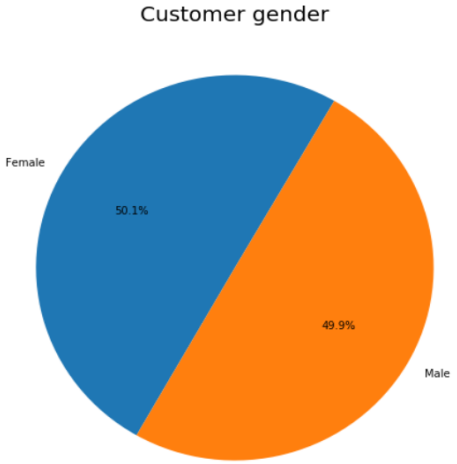
During the sessions, Supermaket-sales dataset is used to learn about its overall statistical profile.

	count	mean	std	min	25%	50%	75%	
Unit price	1000.0	55.672130	2.649463e+01	10.080000	32.875000	55.230000	77.935000	99
Quantity	1000.0	5.510000	2.923431e+00	1.000000	3.000000	5.000000	8.000000	10
Tax 5%	1000.0	15.379369	1.170883e+01	0.508500	5.924875	12.088000	22.445250	49
Total	1000.0	322.966749	2.458853e+02	10.678500	124.422375	253.848000	471.350250	1042
cogs	1000.0	307.587380	2.341765e+02	10.170000	118.497500	241.760000	448.905000	993
gross margin percentage	1000.0	4.761905	6.131498e-14	4.761905	4.761905	4.761905	4.761905	4
gross income	1000.0	15.379369	1.170883e+01	0.508500	5.924875	12.088000	22.445250	49
Rating	1000.0	6.972700	1.718580e+00	4.000000	5.500000	7.000000	8.500000	10

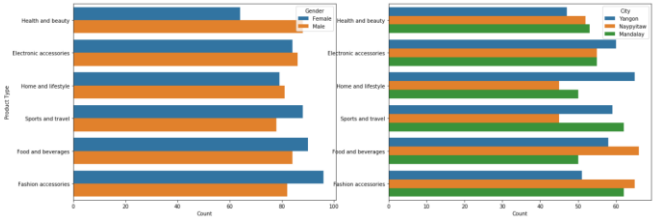
As seen above, one can see the min max, std values of each column of the dataset and derive some inference as to the nature of the dataset.

The next step is to visualize the data using visualisation techniques. Some of the key visualization outcomes are shown below:

For example, the gender break of customer is visualized using a pie chart as shown below:



One can also create a visual chart showing how product categories are bought gender-wise and city-wise.



More useful and powerful visualization can be carried out using several tools.

Several more explorative techniques are covered during the session.
