ATW115/3 INTRODUCTION TO DATA ANALYTICS

SEMESTER II, ACADEMIC SESSION 2021/2022

SCHOOL OF MANAGEMENT

INDIVIDUAL ASSIGNMENT
Data Analysis - Flipkart Mobile Dataset

LECTURER   **: DR. CHEANG PECK YENG SHARON**

NAME OF STUDENT **: NEO ZHAN YUAN**

NO. METRIC   **: 157254**

Table of content

Content	Page
1.0 Introduction	**1**
2.0 Methodology and Results	**2 - 32**

1.0 Introduction

Flipkart is an Indian online retailer with headquarters in Bangalore, Karnataka. It was created by Sachin and Binny Bansal and is India's biggest e-commerce start-up. The company offers a wide range of products, including electronics such as laptops, tablets, smartphones, and mobile accessories, as well as in-vogue fashion staples such as shoes, clothing, and lifestyle accessories; modern furniture such as sofa sets, dining tables, and wardrobes, as well as appliances that make life easier such as washing machines, TVs, air conditioners, mixer grinder juicers, and other time-saving kitchen and small appliances; and home furnishings such as cushion covers, mattresses, and bedsheets, as well as to toys and musical instruments.

The dataset chosen is the **Flipkart Mobile** Dataset. This dataset comes from Kaggle.com. The dataset contains descriptions of top 5 most popular mobile brands in India. There are 16 columns each having a title which is self-explanatory. Including information about brand, model, colour, processor, screen size, ROM, RAM, display size, number rear camera, number front camera, battery capacity, ratings, number of ratings, sales price, discount percent and sales in crore rupees. There are 430 rows each having a mobile with at least a distinct feature. The dataset did not provide a direct sales record that shows how many units of a certain mobile model were sold. To solve this, the new column "Unit Sold" was created with the formula [=[@[sales (crore rupees)]]*10000000/[@[sales_price]]]. All the data related to the currency in this report and dataset is the official currency of India (Indian rupee / ₹ / INR).

The problem statement is that a new seller is difficult to enter the Flipkart mobile phones market because it already has a lot of mobile phone brand sellers, which has become a huge competition. The objective of this report is to compare the products, brands and specifications of different mobile brands. The second objective is to determine the purchase point of consumers. The third objective is to predict phone sales price with specification with multiple regression models.

2.0 Methodology and Results

2.1 Logical function

Logical function is used in spreadsheets to design two different queries which are the "Rating Review" and "Price range". Furthermore, logical functions also help us on counting frequency and sum the cell with criteria we wanted. We will look at the IF function.

Firstly, for a better review for ratings, the column "Rating review" was created. We assume that rating greater or equal to 4.6 is considered Excellent, greater or equal to 4.4 considered Good, greater or equal to 4.2 considered Not Bad and those lower than 4.2 is considered Normal. The formula is [=IF([@ratings]>=4.6,"Excellent",IF([@ratings]>=4.4,"Good",IF([@ratings]>=4.2,"Not Bad","Normal")))] (**Figure 1**).

Moreover, The column 'Price range' is created to assume the price greater or equal to 60000 considered luxury, greater or equal to 40000 considered medium, greater or equal to 20000 considered affordable, lower than 20000 considered cheap. The formula is [=IF([@[sales_price]]>=60000,"Luxury",IF([@[sales_price]]>=40000,"Medium",IF([@[sales_price]]>=20000,"Affordable","Cheap")))] (**Figure 2**).

Figure 1 & 2: The Rating review and Price range

2

2.1.1 Result of the Logical Function

By using the if function, we can know that 236 out of 430 ratings are not bad, which is between 4.2 to 4.4, and 134 of them are good (4.4 - 4.6). Total 56, Apple gained 35 excellent and 21 good, while most of the phones of other brands are rated at not bad (below 4.2). Moreover, we know that most of the phones in India are cheap (below 20000). Realme and Samsung contribute the most cheap phone, while almost all of the 56 Apple phones are luxury (28) and medium (23).

2.2 Retrieval Application Table (INDEX and MATCH function)

Data retrieval is the process of identifying and extracting data from a database. There are three types of retrieval applications which are combination VLOOKUP and IF functions, VLOOKUP and MATCH function and INDEX and MATCH function. We will use the combination INDEX and MATCH functions as well as Define name, data validation, and IFERROR function. At the beginning, we created the new column "device name" to concentrate the model name & colour & ROM with formula [=A2&" "&B2&" "&C2&" "&G2&" "&"GB"] to better differentiate between different specifications of phones.

We define the name of column "Device name" to their brand name. Then we use data validation to create a list of brand names at cell X4. At cell Y4, we also create the list with data validation and fill in the formula [=INDIRECT($X$4)] at its source. After doing these steps, when we list out the brand at cell X4, the Y4 will also list out the device name of the particular brand name (**Figure 3**). Next we start to do INDEX and MATCH functions.

The INDEX function returns a value or the reference to a value from within a table or range. I will use it to return the value of position given by match function.

The MATCH function finds for a given item in a range of cells, and then returns the relative position of that item in the range. I used this to find the relative position of Y4 to column A1:A431 and match X6 - X14 & Z6 - Z14 to the row A1:D1 and used the INDEX function to return the value for their interaction cell. I also used the IFERROR function to avoid error when the Y4 is blank. The formula is:

[=IFERROR(INDEX($A$1:$R$431,MATCH($Y$4,$D$1:$D$431,0),MATCH(X6,$A$1:$R$1,0)),"")] (**Figure 4 & 5**)

- $A$1:$R$431 : The table array
- $Y$4 : The list cell of device names
- $D$1:$D$431: Look up array of column "Device names"
- X6 : Look up value of row
- $A$1:$R$1 : Look up array of row
- 0 : measure the value exactly equal to the lookup value.

In addition, I also create a retrieval applications of Count of distinct model with formula [=DCOUNTA($U$1:$V$121,V1,$X$3:$X$4)], $U$1:$V$121 is the database, V1 is the field, $X$3:$X$4 is the criteria. This function can tell us how many models for the particular brand.

=A2&" "&B2&" "&C2&" "&G2&" "&"GB"

A	D	W	X
nd	Device name		
ple	Apple iPhone SE Black 64 GB		
ple	Apple iPhone 12 Mini Red 64 GB		
ple	Apple iPhone SE Red 64 GB		
ple	Apple iPhone XR Others 64 GB		
ple	Apple iPhone 12 Red 128 GB		
ple	Apple iPhone 12 Blue 64 GB		
ple	Apple iPhone 12 White 128 GB		
ple	Apple iPhone 12 Green 64 GB		
ple	Apple iPhone 12 Blue 128 GB		
ple	Apple iPhone 12 Black 128 GB		
ple	Apple iPhone 12 Mini Black 64 GB		

Edit Name

Name: Apple
Scope: Workbook
Comment:
Refers to: ='INDEX + MATCH Table'!$D$2:$D$57

OK   Cancel

Data Validation

Settings | Input Message | Error Alert

Validation criteria
Allow: List
Ignore blank
In-cell dropdown
Data: between
Source: =$AG$2:$AG$6

Apply these changes to all other cells with the same settings

Clear All   OK   Cancel

Y4 — Realme C20 Blue 32 GB

W	X	Y	Z	AA
		Retrieval Application table of mobile phone		
	Brand		Device Name:	
	Realme		Realme C20 Blue 32 GB	
	Count of distinct model	35		
	model	C20	ROM	32
	base_color	Blue	battery_capacity	5000
	processor	MediaTek	ratings	4.4
	screen_size	Large	num_of_ratings	97955
	RAM	2	sales_price	INR 7,499.00
	display_size	6.5	discount_percent	0.06
	num_rear_camera	1	sales (crore rupees)	INR 73.46
	num_front_camera	1	Unit sold	97960
	Rating Review	Good	Price range	Cheap

Data Validation

Settings | Input Message | Error Alert

Validation criteria
Allow: List
Ignore blank
In-cell dropdown
Data: between
Source: =INDIRECT($X$4)

Apply these changes to all other cells with the same settings

Clear All   OK   Cancel

Figure 3: Define name & Data validation

=IFERROR(INDEX($A$1:$R$431,MATCH($Y$4,$D$1:$D$431,0),MATCH(X6,$A$1:$R$1,0)),"")

	W	X	Y	Z	AA	AB
4 GB			Retrieval Application table of mobile phone			
ed 64 GB		Brand		Device Name:		
GB		Apple		Apple iPhone SE Black 64 GB		
64 GB		Count of distinct model	6			
8 GB		model	iPhone SE	ROM	64	
4 GB		base_color	Black	battery_capacity	1800	
128 GB		processor	Water	ratings	4.5	
64 GB		screen_size	Very Small	num_of_ratings	38645	
8 GB		RAM	2	sales_price	INR 32,999.00	
28 GB		display_size	4.7	discount_percent	0.17	
ack 64 GB		num_rear_camera	1	sales (crore rupees)	INR 127.52	
128 GB		num_front_camera	1	Unit sold	38644	
hite 128 GB		Rating Review	Good	Price range	Luxury	
6 GB						
ack 128 GB						
128 GB						

=IFERROR(INDEX($A$1:$R$431,MATCH($Y$4,$D$1:$D$431,0),MATCH(X6,$A$1:$R$1,0)),"")

	W	X	Y	Z	AA	AB
GB						
64 GB						
iB						
4 GB						
GB						
iB						
8 GB						
4 GB						
GB						
8 GB						
ck 64 GB						
8 GB						
ite 128 GB						
GB						
k 128 GB						

Retrieval Application table of mobile phone

Brand	Device Name:	
Realme	Realme C20 Blue 32 GB	
Count of distinct model	35	
model	C20	ROM	32
base_color	Blue	battery_capacity	5000
processor	MediaTek	ratings	4.4
screen_size	Large	num_of_ratings	97955
RAM	2	sales_price	INR 7,499.00
display_size	6.5	discount_percent	0.06
num_rear_camera	1	sales (crore rupees)	INR 73.46
num_front_camera	1	Unit sold	97960
Rating Review	Good	Price range	Cheap

Figure 4 & 5: Result of Retrieval Application with different examples.

2.2.1 Result of Retrieval Application

When I choose the brand Apple at cell X4 with a list, then we will be able to see the device names are all listed at Y4, and we choose Apple iPhone SE Black 64 GB, lastly I can see all the specifications of phones (**Figure 4**). Another retrieval, we choose Realme at cell X4, cell Y4 will automatically renew the list to the device names of Realme, I choose Realme C20 Blue 32 GB, and I can get the specifications of Realme C20 Blue 32 GB (**Figure 5**).

2.3 Pivot Table and Pivot Chart

PivotTable is a tool for summarising, analysing, exploring, and presenting summary data. Pivot Charts supplement PivotTables by providing visualisations to the summary data in a PivotTable so that comparisons, patterns, and trends may be easily seen.

It is composed of four grids: Filter, Rows, Columns, and Values. By clicking the value field settings, the values grid's most effective feature allows us to count the total, average, maximum, and minimum values, among others. After entering all the desired elements, a Pivot Table will be generated.

I am using the Pivot Table to figure out the unit sold of phones based on their brand and processor. The elements that I put in the grids: (**Figure 6**)

- Filter: -
- Columns: processor
- Rows: brand
- Values: Unit sold

Lastly, I visualise the table by Pivot Chart.
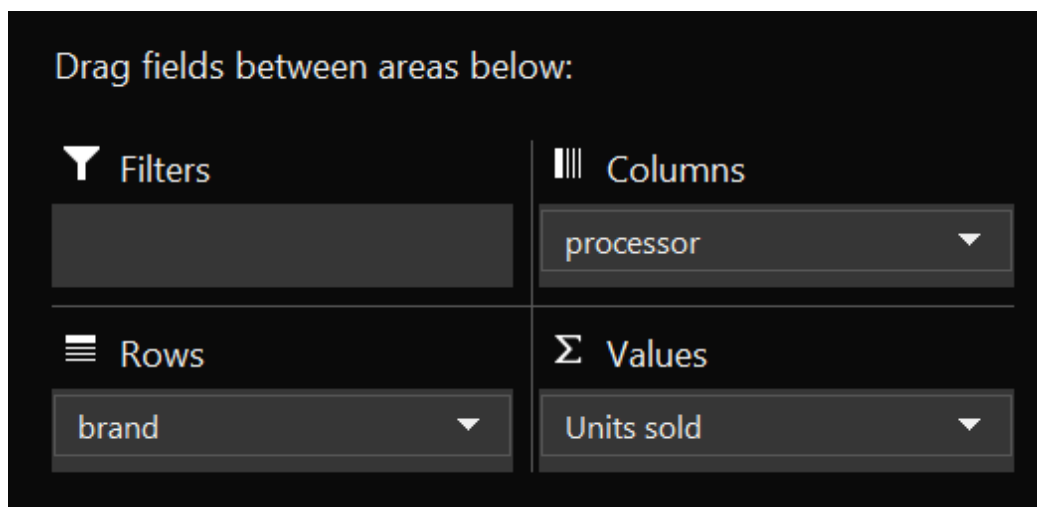
Figure 6: Options for Pivot Table

Units sold	Processors							
Brands	Ceramic	Exynos	iOS	MediaTek	Others	Qualcomm	Water	Total
Apple	16570		94654				158859	270083
Poco				720498		1087477		1807975
Realme				3635207	21703	376414		4033324
Samsung		732385		26866	18787	53136		831175
Xiaomi				84782		3106808		3191591
Total	16570	732385	94654	4467354	40490	4623835	158859	10134147

Figure 7 & 8: Pivot Table and Pivot Chart
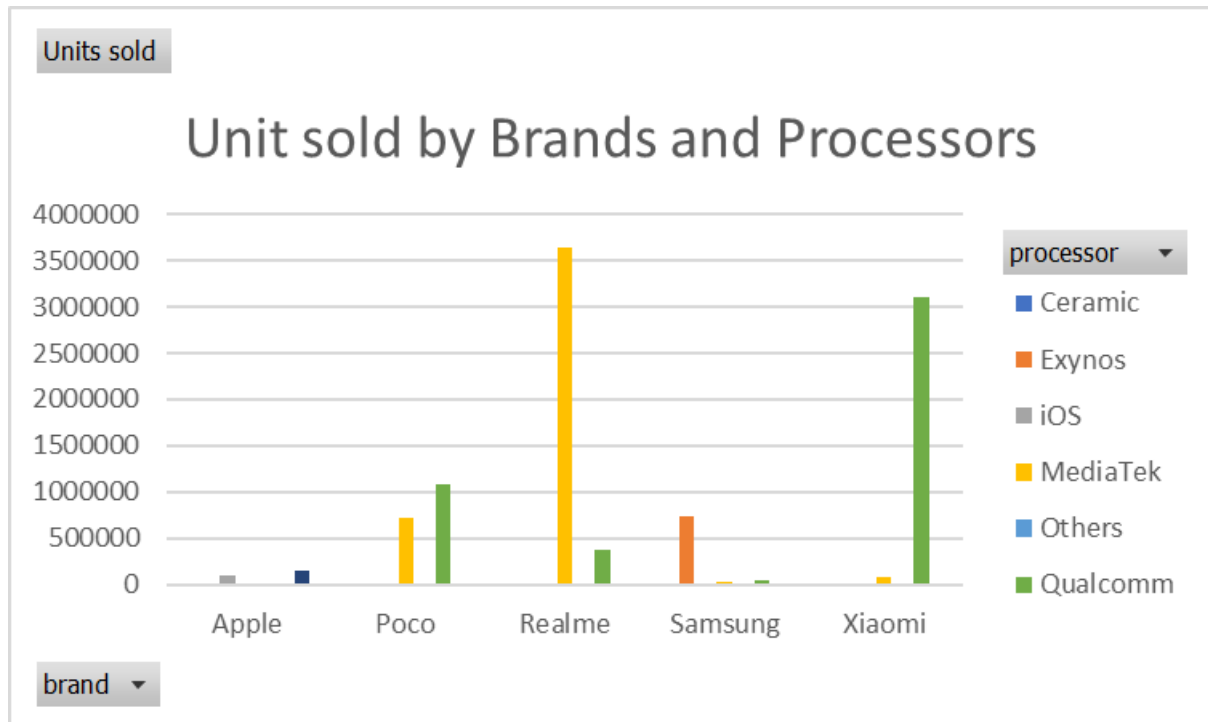
2.3.1 Result of Pivot Table and Pivot Chart

According to the Pivot table and chart, we can know the sum of the units sold of their phone based on brands and processor.vRealme looks to be doing well in terms of sales with over 4 millions smartphones sold on Flipkart. It is probably because Qualcomm and MediaTek are the preferred processor brands for Indian consumers as two of them sold around 9 millions out of the around 10 millions of total processors sold and Realme have 3635207 phones using Qualcomm and 376414 phones using MediaTek.

Xiaomi sold 3191591 units of phone, it is probably because their phones are fully using processor MediaTek and Qualcomm. Poco is also the same with Xiaomi but it only sold 1807975 units of phones. While Samsung and Apple are not doing so well, they sold 831175 and 270083 respectively.

2.4 Chart

2.4.1 Pie chart

Pie charts aid in understanding the parts-to-whole connection. A pie chart is a circular split into sections, or slices. Each slice reflects the number or percentage of observations of a particular level for the variable.

I am using pie charts to visualise the products distinctive by brand. Advanced filter and COUNTIF function will be used in products distinctive by brand. First we apply the advanced filter at the Data toolbar, the list range is A1:A431 (column Brand), and we tick the 'unique record only' so that we can get the 5 unique values in the column Brand (**Figure 9**). Next, I apply the Countif function to compute how many products are distinctive by brand. The formula is [=COUNTIF($A$2:$A$431,H4)]. $A$2:$A$431 is the column Brand, cell H4 is the unique value of the outcome of the advanced filter (**Figure 10**).

I will use two-dimensional (2-D) pie charts as it is easy to visualise. (**Figure 11**)
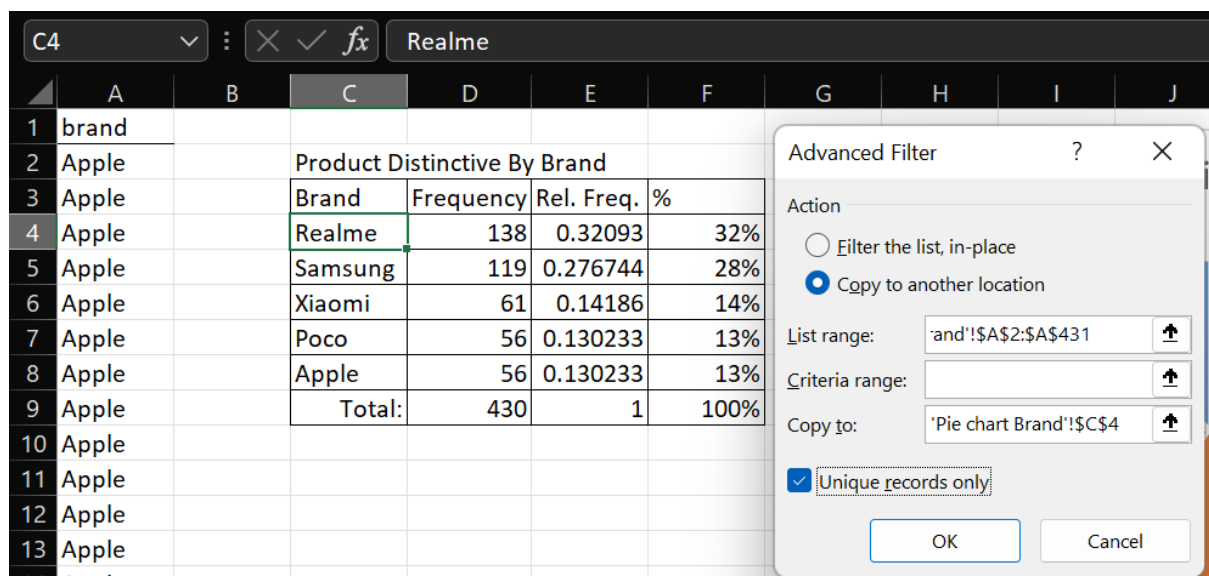
Figure 9: Advanced Filter to filter Brand name

Figure 10: COUNTIF Function to count frequency of Brand
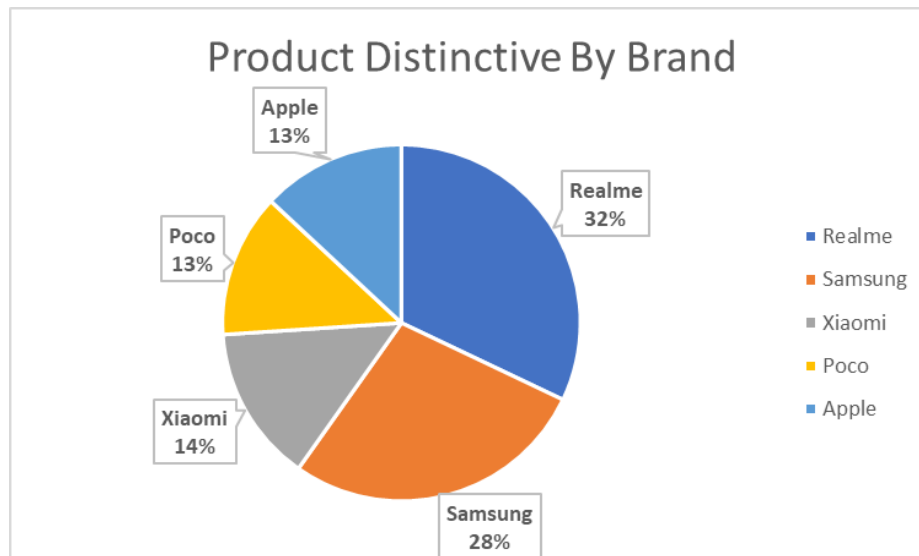
Figure 11: Product Distinctive by Brand

2.4.1.1 Result of pie chart of products distinctive by brand,

From the Pie chart of products distinctive by brand, Realme has the most alternatives (138) in terms of specifications, they have a wide choice of phones to choose from. On the other hand, both Poco and Apple have the most limited options (56). Samsung has 119 choices of phones and Xiaomi offers 61 options to choose from.

2.4.2 Column chart

A column chart, often known as a vertical bar chart, is a way of showing data with categories represented by a rectangle. They make it simple to compare a variety of goods and analyse patterns. In general, numbers and data provided in tables or textual format are difficult to comprehend. Column charts make things simple, quick, and easy to understand.

I will use this to visualise the products distinctive by models. Advanced filter and COUNTIF function will be used. First I apply the advanced filter at the Data toolbar, the list range is A2:B431 (column brand and model), and we tick the 'unique record only' so that we can get the unique values in the column brand and model (**Figure 12**). Next, I apply the Countif function to compute how many products are distinctive by model. The formula is [=COUNTIF($E$4:$E$123,H19)]. $E$4:$E$123 is the range, cell H19 is the criteria (**Figure 13**).

I will use clustered column charts as it is easy to visualise.
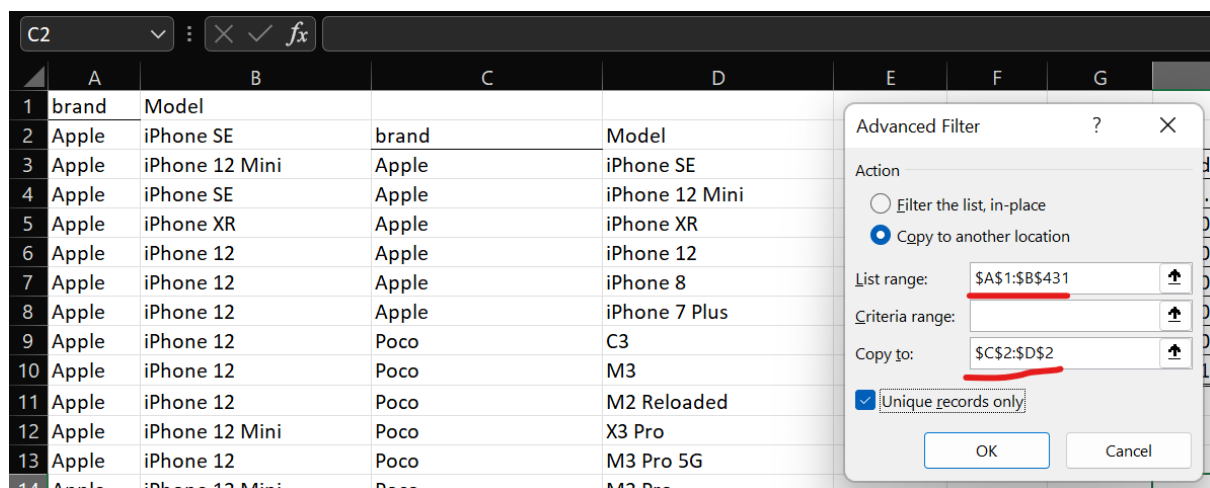
Figure 12: Advanced filter to filter Brand and Model
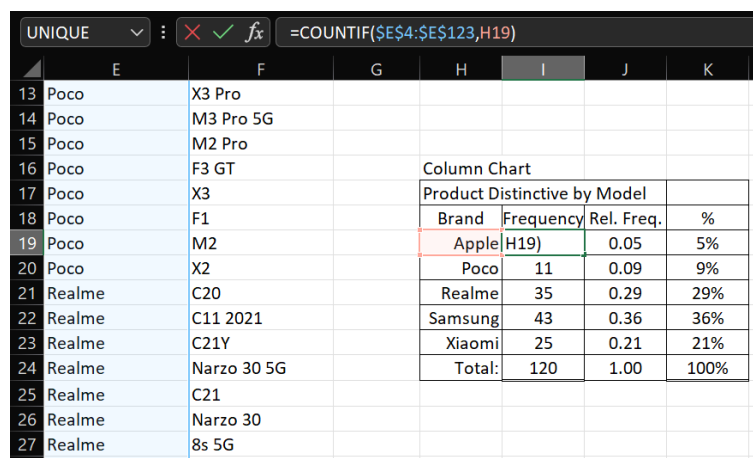
Figure 13: COUNTIF Function to count frequency for each brand

11
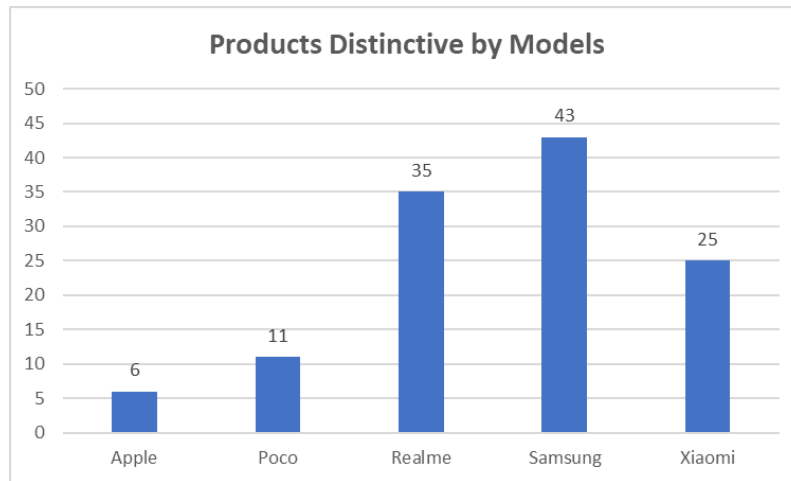
Figure 14: Column Chart of products distinctive by models

2.4.2.1 Result of column chart of products distinctive by models,

     From the column chart of products distinctive by models, Samsung offers a much more distinctive mobile phone model. Despite being a very well known brand, Apple offers the least number of model options. Regardless of the fact that Realme offers the most product options, it provides 35 model options, while Xiaomi and Poco provide 25 and 11 options respectively.

2..4.4 Boxplot

Boxplot is a method for graphically demonstrating the locality, spread and skewness groups of numerical data through their quartiles.

I am using boxplot to visualise the display size, battery capacity and discount percentage provided by different brands. To create a boxplot, I highlight the brand and display size column and discount percent column respectively, then choose the boxplot in the Insert toolbar.
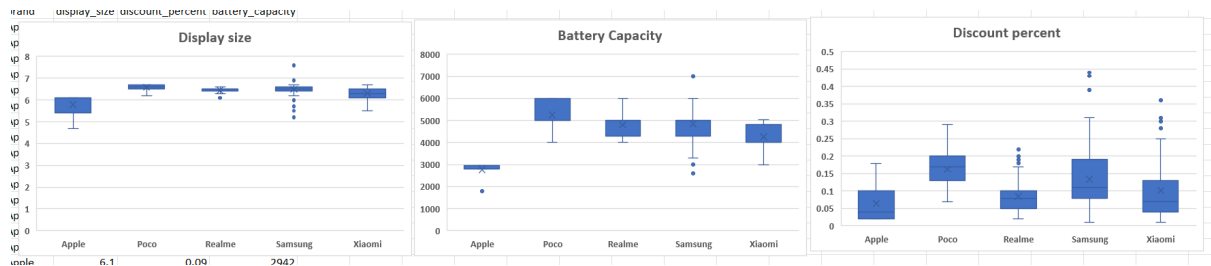
Figure 15: Boxplot of display size, battery capacity and discount percentage

2.4.4.1 Result of boxplot of display size

Most of the phone brands have 6 inch to 7 inch display size, excluding Apple phones that have 5 inch to 6 inch.

2.4.4.1 Result of boxplot of battery capacity

Poco phones have the largest battery capacity among the phone brands, while Apple phones have the smallest battery capacity.

2.4.4.2 Result of boxplot of discount percent

The mean of the Apple, Realme, Samsung and Xiaomi are greater than median, which is right skewed. Poco is left skewed. Realme, Samsung and Xiaomi have outliers.

2.4.5 Combination chart

The combination chart combines the properties of a bar chart and a line chart into one picture. The data is represented in the combination chart by a number of bars and/or lines, each of which represents a different category. When comparing numbers in various categories, combining bars and lines in the same picture may be advantageous since it makes it apparent whether a category is greater or lower. The Combination chart I used is side by side chart and pareto chart.

2.4.5.1 Side by side chart

I am using a side by side chart to visualise products distinctive by colour, processor, screen size ROM and RAM and number of front and rear cameras from different brands. Advanced filter and COUNTIF function will be used. I use advanced filters to find the unique value in the colour, processor, screen size ROM and RAM and number of front and rear cameras column. Then I create a contingency table with brands at column side, colour, processor, screen size ROM and RAM and number of front and rear cameras at the row side respectively. Next, I use the formula to count the phones with different specifications by their brands. The formula is [=COUNTIFS(Criteria Range 1, Criteria 1, Criteria Range 2, Criteria 2]. Criteria Range 1 and Criteria 1 used to target the brands and Criteria Range 2, Criteria 2 used to target the colour, processor, screen size ROM and RAM and number of front and rear cameras, thus counting relative value. (**Figure 16 & 17**)

I will use side by side bar chart to visualise them. (**Figure 18**)
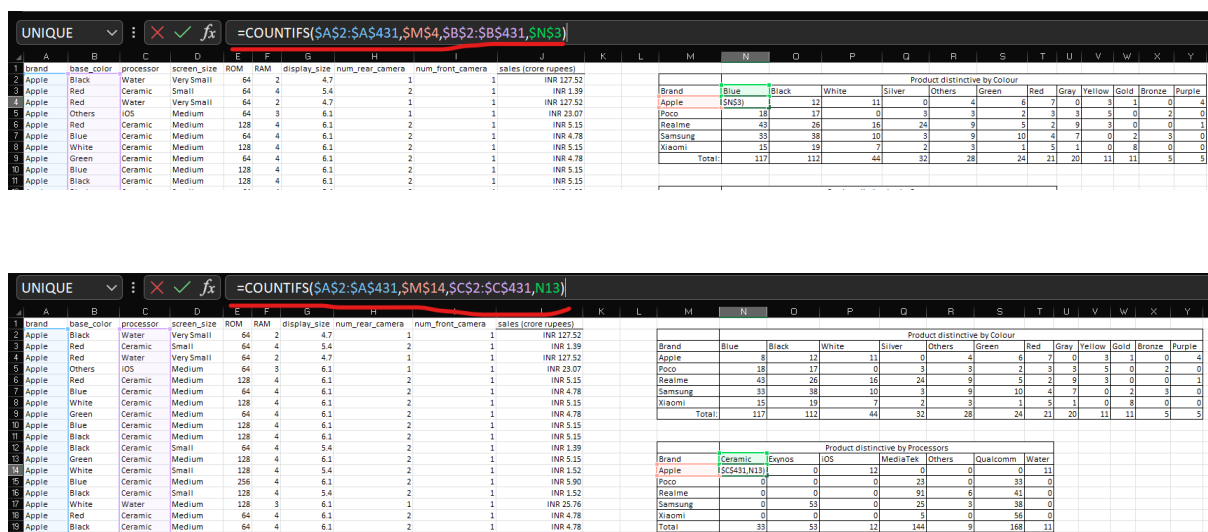

Figure 16 and 17: COUNTIFS Function to count frequency of different specifications brands
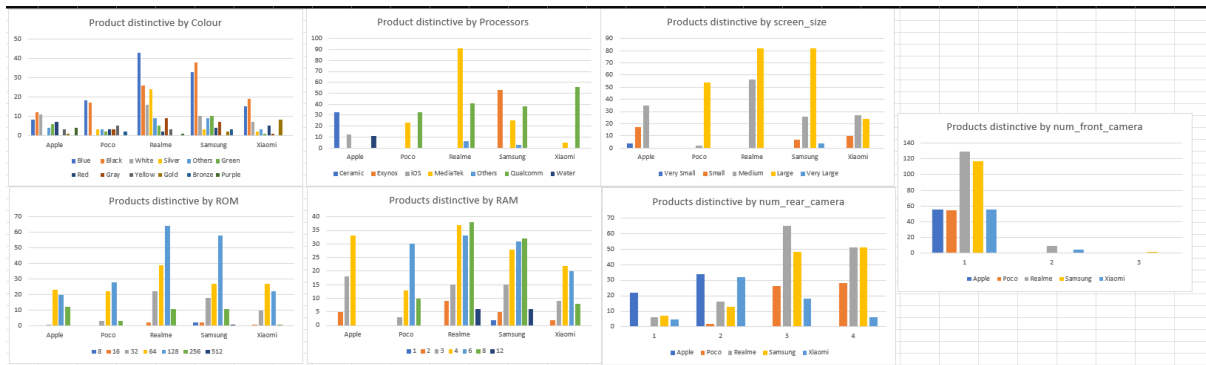
Figure 18: side by side bar charts

2.4.5.1.1 Result of side by side chart of product distinctive by colour

Apple provides 9 kinds of colours which are blue, black, white, green, red, yellow, gold, purple, and others on their product. Poco's products are available in nine colours, including blue, black, silver, green, red, grey, yellow, bronze and others. There are nine colour options for Realme's goods, including blue, black, white, silver, green, red, grey, yellow, purple and others. Samsung's products are available in a variety of colours, including blue, black, white, silver, green, red, grey, gold, bronze, and others. Lastly, blue, black, white, silver, green, red, grey, gold and other colours are available on Xiaomi phones.

2.4.5.1.2 Result of side by side chart of product distinctive by Processor

Ceramic, iOS and Water are the processors of Apple phones. Poco phone processors consist of MediaTek and Qualcomm. MediaTek, Qualcomm and Other make up the Realme phone's processors. The Samsung phone's processors are made up of Exynos, MediaTek, Qualcomm and Others. MediaTek, Qualcomm are among the Xiaomi phone's processors

2.4.5.1.3 Result of side by side chart of product distinctive by screen size

Apple targets the medium and small screen consumers. Poco phones are using a large screen. The screen size of Realme phones are medium and large. Samsung phones have a small, medium, large and very large screen size. Xiaomi phones come in a variety of screen sizes, including small, medium, and large.

2.4.5.1.4 Result of side by side chart of product distinctive by ROM and RAM

Apple phones have ROM and RAM size available from 32GB to 256GB and 2GB to 4GB respectively. ROM and RAM storage capacities range from 32GB to 256GB and 3GB to 8GB, respectively for poco phones. Samsung phones have ROM and RAM sizes vary between 16GB and 256GB and between 2GB and 12GB, respectively. While 16GB to 256GB of ROM and 2GB to 8GB of RAM are available for Xiaomi phones.

<u>2.4.5.1.5 Result of side by side chart of product distinctive by front camera and rear cameras</u>

Apple phones have a single front camera and one or two rear cameras are available. Poco phones use either a single and double front-facing camera and two to four rear-facing cameras. There are models of Realme phones with a single or double front-facing camera and one to four rear-facing cameras. Samsung phones come with a single or triple front-facing camera, as well as one to four rear-facing cameras. Xiaomi phones have one to four rear-facing cameras, as well as a single or triple front-facing camera.

<u>2.4.5.1.6 Result of side by side chart of product distinctive by Battery capacity</u>

Apple phones have a battery capacity ranging 1800mAH to 2942mAH, depending on the model. The battery capacity of Poco phones ranges from 3000mAH to 4000mAH. Realme phones have a variety of battery capacities, ranging from 4000mAH to 5000mAH. The battery capacity of Sansung phones range from 5000mAH to 5160mAH. Xiaomi phones have a battery capacity ranging from 5160mAH to 7000mAH.

2.4.5.2 Pareto chart

Pareto chart used to visualise the sales of each brand and the cumulative percentage of them. Pareto charts provide the ordered frequency counts of values for a categorical or nominal variable's distinct levels. The charts follow the "80/20" rule. According to this rule, about 80% of problems are the result of 20% of the causes. This concept is also known as the "vital few and trivial many" rule.

To create a Pareto chart, I create a table with variables 'brand', 'sales', 'percentage', 'cumulative percentage'. Variable 'brands' include the 5 brands of mobile phone. For variable 'sales', we compute the total sales of each brand by the formula [=SUMIF($A$2:$A$431,D4,$B$2:$B$431)]. $A$2:$A$431 (brand) is the range of criteria, D4 is the criteria as known as brand names, $B$2:$B$431 (sales in crore rupees) is the sum range (**Figure 19**). Then calculate their percentage with formula [=E4/$E$9], E4 is the total sales of each brand, $E$9 is the sum of sales. At variable 'cumulative percentage', we use formula [=G4+F5) which is adding the last percentage with current percentage.

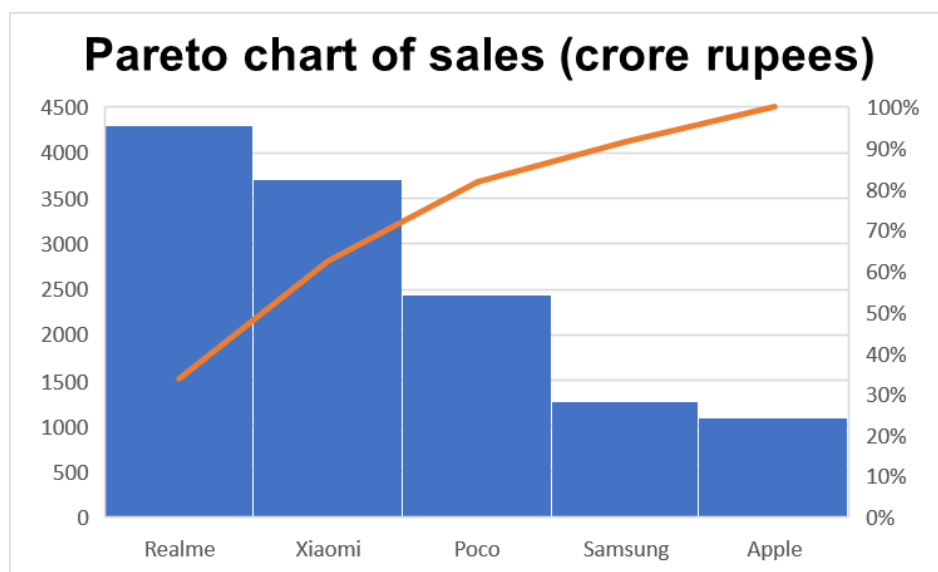Figure 19: SUMIF Function to count sales for each brand

Figure 20: Pareto Chart of Sales (crore rupees)

2.4.5.2.1 Result of pareto chart

From the Pareto chart, it had identified the "vital few" (Poco, Samsung and Apple) and "trivial many" (Realme and Xiaomi). The Pareto Chart has successfully helped to know 80% of sales come from Realme and Xiaomi. While Poco, Samsung and Apple contribute 20% of sales.

2.5 Data Visualisation with Conditional Formatting

Data bars, colour scales, and icon sets are conditional formats that create visual effects in your data. Conditional formatting makes it easier to compare the values of a range of cells at the same time. I am using data bars and colour scales to visualise the table of the sum of the unit sold based on brands and their specification.

We are going to sum up the units sold of phones from different specifications. Pivot Table will be used to summarise the data we need. the brand will put at the row grid, while the unit sold will be put at the values grid and set it as sum. Different specifications such as base colour, screen size, ROM, RAM, display size, num rear camera, num front camera, battery capacity, sales price, and discount percent will take turns putting in the column grid respectively, Then we copy the data that summarise by pivot table to another cell.

Next, we select the Conditional Formatting at the group styles in the Home toolbar, we use a data bar to visualise the sum of the total unit sold based on the different specifications (**Figure 22**) and use colour scales to visualise the sum of the total unit sold based on the different specifications by brand (**Figure 21**). For colour scales, green colour represents lowest value, while red colour represents highest value.

Sum of Unit sold	Screen Size					
Brands	Large	Medium	Small	Very Large	Very Small	Grand Total
Apple		109255	33617		127211	270083
Poco	1671951	136025				1807975
Realme	2798009	1235314				4033324
Samsung	180306	505561	145278	30		831175
Xiaomi	58369	1949008	1184214			3191591
Grand Total	4708635	3935162	1363109	30	127211	10134147

Figure 21: Colour Scales

screen_size	
Large	4708635
Medium	3935162
Small	1363109
Very Large	30
Very Small	127211

Figure 22: Data Bar

2.5.1 Result of Data Visualisation with Conditional Formatting

Generally, Realme is the one that sold almost 40% which is 4033324 units out of the 10134147 units of phone, followed by Xiaomi (3191591), Poco (1807975), Samsung (831175) and Apple (270083).

Based on the colour, we can see black is the most popular (2990091) that sold in India. Followed by blue (2488196), two of them colours accounted for more than 50% of total units sold. Xiaomi has 1394345 units of black colour phone, and almost 98% of gold colour phones from Xiaomi. While most of the phones from Realme are blue colour.

Based on the screen size, large and medium size phones are most popular in India with 85% of total units sold. The largest brand of large and medium phones is Realme. While Apple is not targeting the large phone user.

Based on the ROM, 64GB is enough for most Indians. In this ROM size market segment, Realme is the greatest.

Based on the RAM, 60% of the Indians require 4GB and 6GB. Xiaomi and Realme are doing well in selling phones with 4GB RAM, while Poco is doing well in 6GB phones.

Based on the number of rear cameras, most Indians prefer at least two rear cameras. All brands are doing well in multi rear cameras. Especially all the Poco Phones with more than one rear camera.

Based on the number of front cameras, single front camera are more common in indians. All the brands are doing well in this feature.

Based on the battery capacity, most Indians prefer 3800mAH - 4799mAH. Xiaomi has been doing well in selling the phone in this segment.

Based on the sales price, Indians prefer prices between 5742 to 15741. We can see Realme, Xiaomi, Poco are inexpensive phones while Apple and Samsung are premium phones.

Based on the discount percentage, although some brands offer the highest discount until 46%, most Indians take the percentage below 21%.

2.6 Frequency Distribution and Histogram

A frequency distribution is a graphical or tabular representation that shows the number of observations within a certain interval. The interval size is determined by the data being examined and the analyst's objectives. The intervals must be mutually exclusive as well as comprehensive. Histograms aid in determining the centre, spread, and shape of a collection of data. Histograms are an excellent tool for analysing data. They may be used to evaluate data for outliers and to better understand the distribution of your data.

2.6.1 Frequency Distribution for Categorical Variable and Histogram

Categorical data is a collection of information that is divided into groups. First, I use processors as categorical variables to analyse the frequency of products distinctive by processors. There are 7 groups which are Water, Ceramic, iOS, MediaTek, Qualcomm, Exynos, and Others. I used the formula COUNTIF [=COUNTIF($A$1:$A$431,C6)]. $A$1:$A$431 is processor column, C6 is a cell referring to the category of data (**Figure 23**). And I compute the relative frequency and percentage by dividing the frequency of each brand by the total frequency of the processor. Lastly, I insert the histogram with the function in the insect toolbar.

Figure 23: COUNTIF Function to calculate frequency of processor

Figure 24: Frequency distribution of processor

2.6.1.1 Result of Frequency Distribution for Categorical Variable and Histogram

According to the Histogram of frequency distribution of processors, the frequency distribution of processors by mobile brands showed they are not equally divided across the 7 categories. The most popular processor is Qualcomm (168), followed by MediaTek (144), Exyno (53), Ceramic (33), IOS (12), Water (11), and Others (9). Qualcomm is the most popular CPU brand, accounting for 168 of the 430 phones. MediaTek and Qualcomm supply processors for more than half of the entire mobile brands.

2.6.2 Frequency Distribution for Numerical Variable and Histogram

For numerical variables, it is a little more complex than categorical variables. We cannot simply use the COUNTIF function. Firstly, we need to find the minimum and maximum value from the data range. The excel formula is [=MAX($A$2:$A$431)] and [=MIN($A$2:$A$431)].

The lower limit battery capacity that I get is 1800mAH and the upper limit battery capacity is 7000mAH. To simplify it, I will take 0mAH as the lower limit and 7000mAH as upper limit. The range between them is 7000mAH. I will divide it into 7 classes which are easy to visualise and not messy. The width of classes can be determined by using the formula (range/ number of classes). Each classes' upper limit is 1000mAH (**Figure 25**).

Next, we are going to use the Excel Histogram Tool at the data toolbar and select the Data Analysis button and choose the Histogram option. We can use the Excel Histogram Tool by selecting the Data Tab and Data Analysis button and choosing the Histogram option. I select the processor column as my input range and the bin range is my upper limit. Since I put the header inside the input range and bin range, I tick the Labels and tick the chart output.

Battery capacity	Upper Limit	Frequency
1000-2000	2000	5
2001-3000	3000	55
3001-4000	4000	62
4001-5000	5000	232
5001-6000	6000	70
6001-7000	7000	6

bin	Bin	Frequency
2000	2000	5
3000	3000	55
4000	4000	62
5000	5000	232
6000	6000	70
7000	7000	6
	more	0

max	7000
min	1800
Median	4500
Range	5200
Mean	4529
#Class	7
lower limit	0
upper limit	7000
range	7000
class witdh	1000

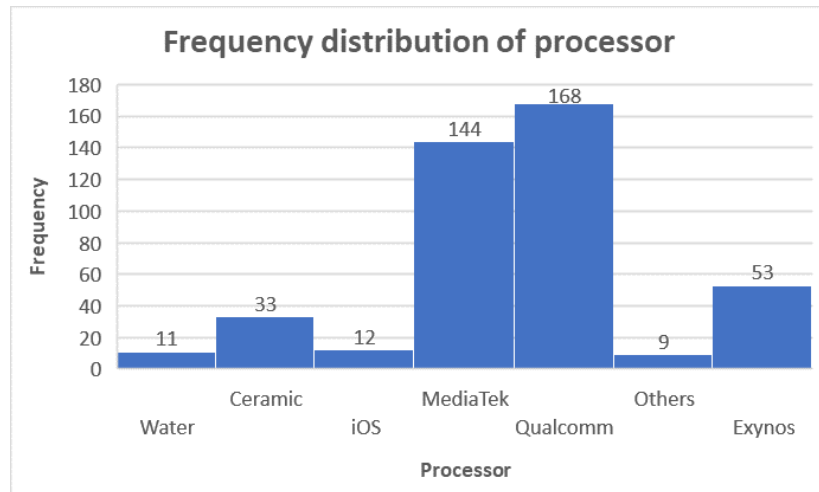**Frequency Distribution Battery Capacity**

Figure 25: Histogram of frequency distribution of battery capacity

2.6.2.1 Result of Frequency Distribution for Numerical Variable and Histogram

According to the Histogram of frequency distribution of battery capacity, the frequency distribution of the battery capacity by the mobile brands is roughly symmetric. The median of the battery capacity is 4500mAH. The range of battery capacity is 5200mAH. There are no outliers. Most mobile phones have a battery backup of 4000 to 5000 mAH.

2.7 Excel Descriptive Statistics Tool

Descriptive statistics are brief descriptive coefficients that summarise a given data set, which can be either a representation of the entire population or a sample of a population.

I am going to do an Excel Descriptive Statistics Tool with the sales price variable. First I am going to the Data toolbar, select Data Analysis from the Analysis group. From the list of tools, select the Descriptive Statistics. I enter input ranges for sales prices variables by different brands which I repeat the step for 5 times to output the summary descriptive statistics for each brand, then I create a table that consists of all summary descriptive statistics for each brand (**Figure 26 & 27**). Moreover I add the calculation of First Quartile andThird Quartile with excel formulas [=QUARTILE($B$2:$B$57,1)] and [=QUARTILE($B$2:$B$57,3)].

Figure 26: Descriptive Statistics for Apple

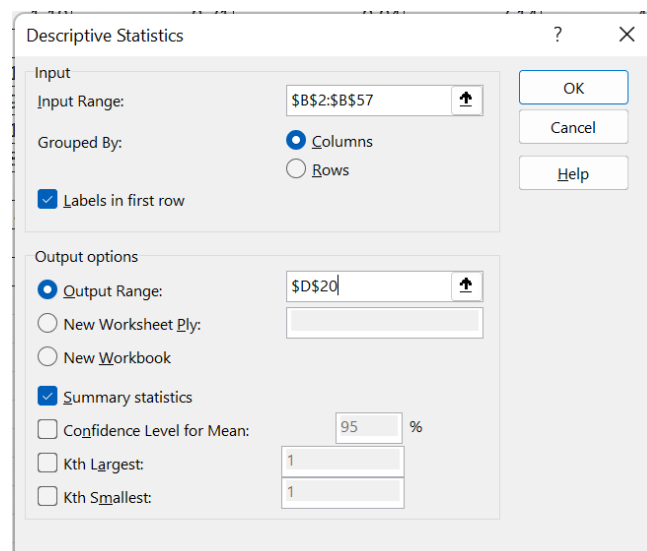| Descriptive Statistics	Apple	Poco	Realme	Samsung	Xiaomi
Mean	57748.11	15936.50	17105.52	28823.54	16711.31
Standard Error	1847.12	778.99	666.03	2821.57	1166.01
Median	59649.00	14999.00	15749.00	16499.00	13400.00
Mode	42999.00	18999.00	14999.00	14999.00	12990.00
Standard Deviation	13822.59	5829.44	7824.11	30779.71	9106.86
Sample Variance	191063953.73	33982386.36	61216745.48	947390716.15	82934972.45
Kurtosis	-1.19	0.71	0.04	7.14	4.42
Skewness	-0.08	1.03	0.81	2.65	1.84
Range	46150.00	23000.00	35500.00	150009.00	49257.00
Minimum	32999.00	7999.00	6499.00	7990.00	5742.00
Maximum	79149.00	30999.00	41999.00	157999.00	54999.00
Sum	3233894.00	892444.00	2360562.00	3430001.00	1019390.00
Count	56	56	138	119	61
First Quartile	46749	11499	10499	12499	10990
Third Quartile	69149	18999	20149	27499	21999
IQR	22400	7500	9650	15000	11009

Figure 27: summary descriptive statistics of sales price variable for each brand

2.7.1 Result of Excel Descriptive Statistics Tool of sales price variable

For 56 units of Apple phones, it has a Mean 57748.11, which is the average price. The Median is 59649 which is the half position of the phone price. The Mode of sales price is 42999 which has the highest frequency. Maximum price is 79149 and Minimum is 32999, price Range is 46150. 75% of Apple phones are below 69149 and 25% are below 46749. The Mean and Median of Apple is nearly, so we can imply that it is approximately Symmetric. It can be proved by the Skewness -0.08 is near the expected value of 0. Kurtosis -1.19 is less than expected value of 3, so it is a Platykurtic. Because the Platykurtic has a shorter tail and is stretched around the centre tails, the majority of the data points are close to the mean. When compared to the normal distribution, a platykurtic distribution is flatter (less peaked).

For 56 units of Poco phones, its mean is 15936.50 (average price), Minimum price is 30999 and Maximum price is 7999, price range 23000. The sales price of the Poco phone has a Skewness of 1.03, greater than 0, so it is a Positive Skew with a long right tail. Kurtosis of 0.71, lower than 3 indicates a Platykurtic. First Quartile is 11499 and Third Quartile is 18999.

For 138 units of Realme phone, its average sales price is 17105.52 (Mean), Minimum price is 6499 and Maximum price of 61999, price Range 35500. It has a Positive Skew since the Skewness 0.81 is greater than 0, and Platykurtic as the Kurtosis 0.04 lower than 3. The First Quartile is 10499 and Third Quartile is 20149.

For 119 units of Samsung, the mean is 28823.54, Minimum price of 7990 and Maximum price of 157999, price Range 150009. Samsung has a Positive Skew because of

Skewness 2.65 greater than 0 and Leptokurtic as Kurtosis 7.14 greater than 3. Outliers are more likely in Leptokurtic because its tails are quite long and slender. Positive kurtosis values suggest that the distribution is peaked with thick tails. Samsung's Median price is 16499, First Quartile is 12499 and Third Quartile is 27499.

For 61 units of Xiaomi, it has Mean 16711.31, Median of 13400, First Quartile is 10990 and Third Quartile is 21999. Maximum price of Xiaomi phone is 54999, Minimum price of 5742. Skewness 1.84 greater than 0, identify it is a Positive Skew. It is a Leptokurtic as Kurtosis 4.42 greater than 3.

2.8 Correlation between two variables

Correlation analysis is a statistical method used in research to calculate the strength of a linear connection between two variables. A high correlation between two variables implies a strong relationship, while a low correlation indicates a weak relationship.

I will set a discount percent as X and units sold as Y. There are three ways to find the correlation value which are mathematical method, Excel function and Excel Tool. I am going to use the Excel tool for correlation of all brands and Excel formula for correlation of each brand.

For the Excel tool, we are going to use the Excel Correlation Tool at the data toolbar and select the Data Analysis button and choose the Correlation option. We input the range for column discount percent and unit sold, grouped by column, tick the label in the first row and output to the $E$3 (**Figure 28 & Figure 29**).

Next, we are using the excel formula to find correlation between discount percent and unit sold for each brand. I add the filter at the title row and I tick Apple to filter out the Apple brand, then I use excel formula to count correlation. The formula is [=CORREL(B2:B57,C2:C57)] (**Figure 30**). I repeat the step for calculation correlation for Poco, Realme, Samsung, and Xiaomi. (**Figure 31**)
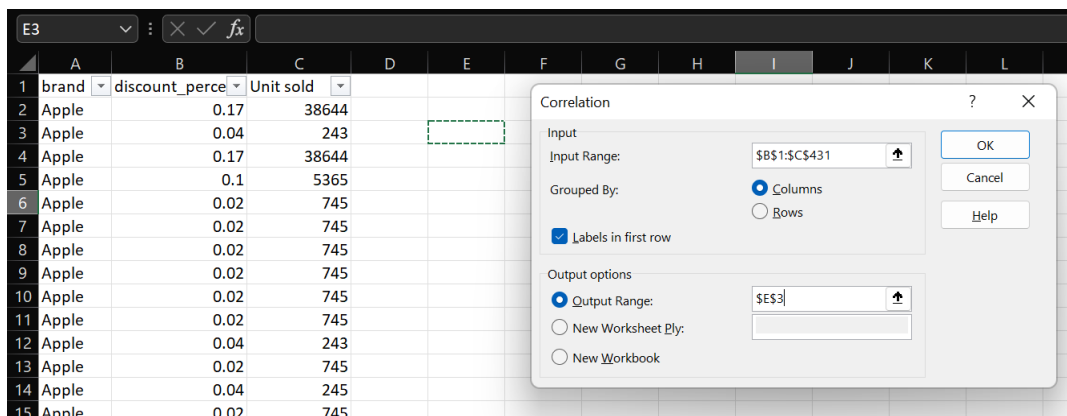
Figure 28: Excel Correlation Tool

	discount_percent	Unit sold
discount_percent	1	
Unit sold	0.130273125	1

Figure 29: Outcome of Excel Correlation Tool
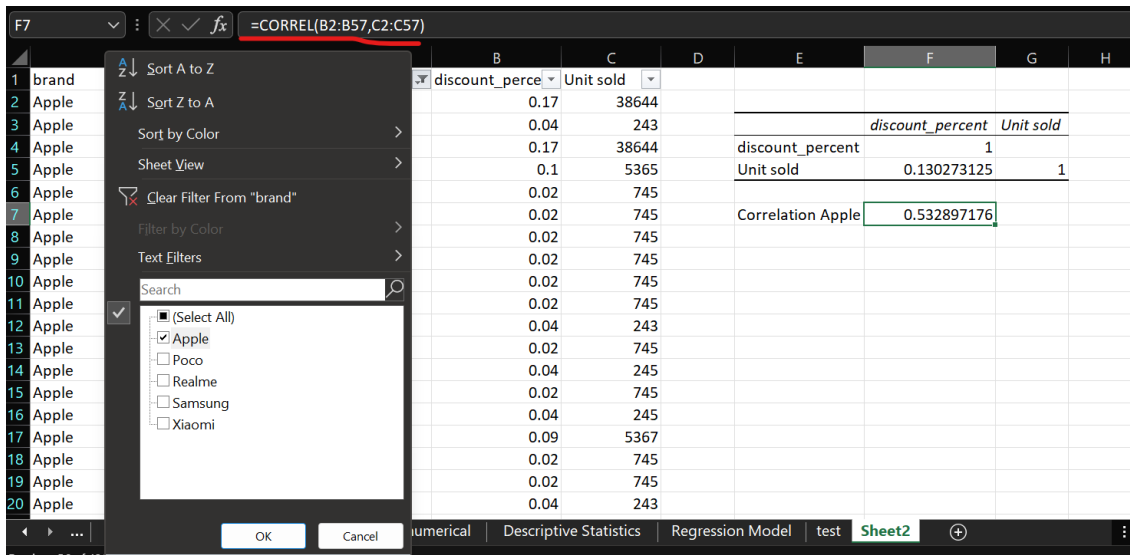
Figure 30: Outcome of Excel CORREL Function Formula

Correlation Apple	0.532897176
Correlation Poco	0.220680224
Correlation Realme	0.465185222
Correlation Samsung	0.386106503
Correlation Xiaomi	0.001379964

Figure 31: Outcome of correlation for each brand

2.8.1 Result for Investigate the correlation between discount percent and unit sold

From the outcome of correlation between discount percent and unit sold, we can know that there is no linear relation (0.13) between the discount percent and unit sold of all brands. While Poco suggests a weak positive linear relationship (0.22) through a wobbly linear rule. Apple (0.53), Realme (0.47) and Samsung (0.39) using a fuzzy-firm linear rule, suggest a somewhat positive linear connection. This means that when discount percent provided go higher, the unit phones sold of Apple, Ralme, and Samsung will also increase.

2.9 Regression model for Prediction

Regression analysis is a type of predictive modelling that examines the connection between a dependent variable (Y) and independent variable (s) (X). This method is used for predicting, modelling time series, and determining the cause-and-effect relationship between variables.

We are going to use multiple regression models. First I separate the different brands to different sheets to predict each brand with their accurate data. Then I set sales price as a dependent variable and ROM, RAM, display_size, num_rear_camera, num_front_camera, battery_capacity, ratings, and discount_percent as independent variables. Then I run the Excel regression model at the data analytic tool.

For the first time in the regression model, we found that some variables have p-value less than 0.05, we remove that variable and rerun the regression model. We continuously repeat the steps and remove the variable with less than 0.05 p-value or negative slope.

After we get the final multiple regression model for each brand, we identify that the R 2 is positive which means sales price can be explained by those variables. All multiple regression models are the statistically significant variables when it comes to explaining the sales price. (**Figure 32 - 36**)

Lastly, with the prediction formula [Y=Constant + B1(X1) + B2(X2) +….BnXn], we identify these formulas to predict the sales price for each brand phone.

- Apple phones: [-3075.617+79.184(ROM)+14495.253(RAM)]
- Poco phones: [-94091.562+58.919(ROM)+1141.075(RAM)+14770.345(display size)]
- Samsung phones: [-19579.365+236.430(ROM)+3897.166(RAM)]
- Xiaomi phones: [-3091.307+ 115.414(ROM)+2030.334(RAM)]
- Realme phones: [-3091.307+ 115.414(ROM)+2030.334(RAM)]

I used these formulas to create a table to predict the sales price.(**Figure 37 & 38**)

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.966954
R Square	0.934999
Adjusted R	0.932546
Standard E	3589.978
Observatic	56

ANOVA

	df	SS	MS	F	ignificance F
Regressior	2	9.83E+09	4.91E+09	381.188	3.49E-32
Residual	53	6.83E+08	12887940		
Total	55	1.05E+10			

	Coefficients	andard Err	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	pper 95.0%
Intercept	-3075.62	2643.319	-1.16354	0.249821	-8377.44	2226.208	-8377.44	2226.208
ROM	79.18429	7.430213	10.65707	8.67E-15	64.28117	94.08741	64.28117	94.08741
RAM	14495.25	830.7421	17.44856	1.27E-23	12828.99	16161.51	12828.99	16161.51

Figure 32: Multiple Regression model of Apple

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.8525
R Square	0.72676
Adjusted R	0.711
Standard E	3133.85
Observatic	56

ANOVA

	df	SS	MS	F	ignificance F
Regressior	3	1.4E+09	4.5E+08	46.1032	1.1E-14
Residual	52	5.1E+08	9821000		
Total	55	1.9E+09			

	Coefficients	andard Err	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	pper 95.0%
Intercept	-94091.6	23660.6	-3.97672	0.00022	-141570	-46613.1	-141570	-46613.1
ROM	58.9189	13.1115	4.4937	3.9E-05	32.6088	85.229	32.6088	85.229
RAM	1141.08	474.175	2.40644	0.0197	189.573	2092.58	189.573	2092.58
display_siz	14770.3	3688.42	4.00451	0.0002	7368.98	22171.7	7368.98	22171.7

Figure 33: Multiple Regression model of Poco

SUMMARY OUTPUT

Regression Statistics

Multiple R	0.83523
R Square	0.69762
Adjusted	0.6924
Standard	17070.9
Observati	119

ANOVA

	df	SS	MS	F	gnificance F
Regressio	2	7.8E+10	3.9E+10	133.809	7.5E-31
Residual	116	3.4E+10	2.9E+08		
Total	118	1.1E+11			

	Coefficient	andard Err	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	pper 95.0%
Intercept	-19579.4	4186.48	-4.67681	7.9E-06	-27871.2	-11287.5	-27871.2	-11287.5
ROM	236.43	39.8355	5.93516	3.1E-08	157.531	315.329	157.531	315.329
RAM	3897.17	1174.74	3.31746	0.00121	1570.44	6223.89	1570.44	6223.89

Figure 34: Multiple Regression model of Samsung

SUMMARY OUTPUT

Regression Statistics

Multiple R	0.87965
R Square	0.77378
Adjusted F	0.76598
Standard E	4405.48
Observati	61

ANOVA

	df	SS	MS	F	gnificance F
Regressio	2	3.9E+09	1.9E+09	99.1953	1.9E-19
Residual	58	1.1E+09	1.9E+07		
Total	60	5E+09			

	Coefficient	andard Err	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	pper 95.0%
Intercept	-3091.3	1824.87	-1.694	0.09563	-6744.2	561.559	-6744.2	561.559
ROM	115.414	21.3678	5.40131	1.3E-06	72.6421	158.187	72.6421	158.187
RAM	2030.33	558.283	3.63675	0.00059	912.809	3147.86	912.809	3147.86

Figure 35: Multiple regression of Xiaomi

SUMMARY OUTPUT

Regression Statistics

Multiple R	0.88897
R Square	0.79027
Adjusted	0.78716
Standard	3609.64
Observati	138

ANOVA

	df	SS	MS	F	gnificance F
Regressio	2	6.6E+09	3.3E+09	254.335	1.6E-46
Residual	135	1.8E+09	1.3E+07		
Total	137	8.4E+09			

	Coefficients	andard Err	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	pper 95.0%
Intercept	1523.91	807.807	1.88648	0.06138	-73.6823	3121.51	-73.6823	3121.51
ROM	43.3541	9.38713	4.61847	8.9E-06	24.7893	61.919	24.7893	61.919
RAM	1952.74	233.318	8.36944	6.5E-14	1491.31	2414.17	1491.31	2414.17

Figure 36: Multiple regression model of Realme

Multiple Regression Model Predict Formula
Y=Constant + B1(X1) + B2(X2) +....BnXn

			ROM	RAM
			8	1
			16	2
			32	3
			64	4
			128	6
			256	8
			512	12

Predict **Apple** Phones Sales Price
Predition Formula=-3075.617+79.184(ROM)+14495.253(RAM)

Apple	
	Coefficients
Intercept	-3075.617
ROM	79.184
RAM	14495.253

Enter Specification
ROM: 256
RAM: 4

Predicted Sales price: 75176.57

Predict **Xiaomi** Phones Sales Price
Formula= -3091.307+ 115.414(ROM)+2030.334(RAM)

Xiaomi	
	Coefficients
Intercept	########
ROM	115.414
RAM	2030.334

Enter Specification
ROM: 256
RAM: 4

Predicted Sales price 34576.13

Predict **Poco** Phones Sales Price
Predition Formula=-94091.562+58.919(ROM)+1141.075(RAM)+14770.345(display size)

Poco	
	Coefficients
Intercept	-94091.562
ROM	58.919
RAM	1141.075
display_size	14770.345

Enter Specification
ROM: 256
RAM: 4
Display size: 6.8

Predicted Sales price: 25994.32

Predict **Realme** Phones Sales Price
Formula= -3091.307+ 115.414(ROM)+2030.334(RAM)

Realme	
	Coefficients
Intercept	1523.912
ROM	43.35415
RAM	1952.738

Enter Specification
ROM: 256
RAM: 4

Predicted Sales price 20433.52

Predict **Samsung** Phones Sales Price
Predition Formula=-19579.365+236.430(ROM)+3897.166(RAM)

Samsung	
	Coefficients
Intercept	-19579.365
ROM	236.430
RAM	3897.166

Enter Specification
ROM: 256
RAM: 4

Predicted Sales price: 56535.42682

Predict **Realme** Phones Sales Price
Formula= -3091.307+ 115.414(ROM)+2030.334(RAM)

Enter Specification
ROM: 256
RAM: 4

Predicted Sales price *L19

Realme	
	Coefficients
Intercept	1523.912
ROM	43.35415
RAM	1952.738

Figure 37 & 38: Sales price prediction table and formula

Predict **Realme** Phones Sales Price		Realme	
Formula= -3091.307+ 115.414(ROM)+2030.334(RAM)			*Coefficients*
		Intercept	1523.912
Enter Specification		ROM	43.35415
ROM: 256		RAM	1952.738
RAM: 4			
Predicted Sales price: 20433.52			

Figure 39: Predict the Realme phone with 256 ROM and 4 RAM

Predict **Xiaomi** Phones Sales Price		Xiaomi	
Formula= -3091.307+ 115.414(ROM)+2030.334(RAM)			*Coefficients*
		Intercept	-3091.307
Enter Specification		ROM	115.414
ROM: 256		RAM	2030.334
RAM: 4			
Predicted Sales price: 34576.13			

Figure 39: Predict the Xiaomi phone with 256 ROM and 4 RAM

2.9.1 Result of sales price prediction table

We try to predict the sales price of Realme and Xiaomi phones with 256 ROM and 4 RAM. We can see that Realme will sell at 20433.52 (**Figure 39**) and Xiaomi will sell at 34576.13 (**Figure 39**)

3.0 DISCUSSION

In the literature review, the problem statement shows that it is difficult for new sellers to compete on the Flipkart mobile marketplace, here I discuss the objective of this report and how it solves the problem.

Firstly, I used the result of different charts, combination charts, and descriptive statistics of price to compare the products, brands and specifications of different mobile brands. There is a significant difference between Apple phones and other Android phones. Based on processors, screen size and display size, only the Apple phones are using Ceramic, IOS and Water processors and medium screen display size in 5 inch to 6 inch. Apple phones also have a luxury average sales price at INR 57748.11, this is a double mean price than other android phones. This is probably because their unique specification cannot easily be replaced by others. However, all the android phones provide similar specifications, so they have huge competition and it reflects on their sales price.

Secondly, I used the result of Excel data visualisation tool Conditional Formatting to determine the purchase point of consumers. In India mobile phone market, Realme and Xiaomi contribute more than 70% of the total units sold of phones (10134147). It is because their specifications are doing well to catch the purchase point of the consumer. Although blue is the most common colour among the phones, consumers prefer to buy black followed by blue. Large screen and display size is a consumer purchase preference. 32GB to 128GB of ROM and 2GB to 6GB of RAM is good enough for consumers. Consumers are more likely to choose the phone with more than two rear cameras while they prefer a single front camera. Consumers desire a battery size that is between 3800 mAH to 6799 mAH. Consumers will be satisfied with sales price from 5742 to 25741 and discount below 21%.

After estimating the first and second objective, Realme and Xiaomi brand are the first choice for the new sellers to start their business since they have high market share in the Flipkart mobile market with exceeding 70% unit sold of the whole market. If new sellers feel that Market Xiaomi and Realme are already saturated, they can try to sell Poco phones since they currently have less volume but it has potential to rise as its specification also satisfies the consumers' needs. They also can use the sales price prediction table to know how much to set the competition price for their product. Therefore, they can develop a good marketing plan for their business to capture the market share at the Flipkart.

4.0 Conclusion

In conclusion, I have accomplished my report's objectives. The problem statement is nicely addressed by the Excel result of this report. Apple phones have the most distinctive specifications in the Flipkart mobile market. So it also let the Apple phones be located at a luxury price phone. In contrast, other android phones which are Poco, Realme, and Xiaomi have a cheap price range. 75% of their products are around INR20000. It is possible because of the huge competition for android phones due to the similar specification. Although they are in a cheap price range, they are able to match the pochase point of their consumers. The Indian consumers prefer the phone's price between INR15000 to INR20000 category. Poco, Xiaomi, and Realme have done a good job of capturing this market. Last but not least, new sellers are able to use the sales prediction table to predict the competitiveness sales price for the phone. Therefore, Xiaomi, and Realme are the first choices for new sellers to begin a mobile phone business at the Flipkart. While Poco is the second choice. It is because Xiaomi and Realme currently have relatively high market share in the Flipkart mobile market while Poco has relatively low market share but it has potential to rise up.