Exploring Phone Prices in India: Descriptive Statistics, Probability Models, and Hypothesis Testing

STATISTICS FOR DATA ANALYSIS

MODULE CODE: B9DA101

LECTURER: DR Shahram Azizi

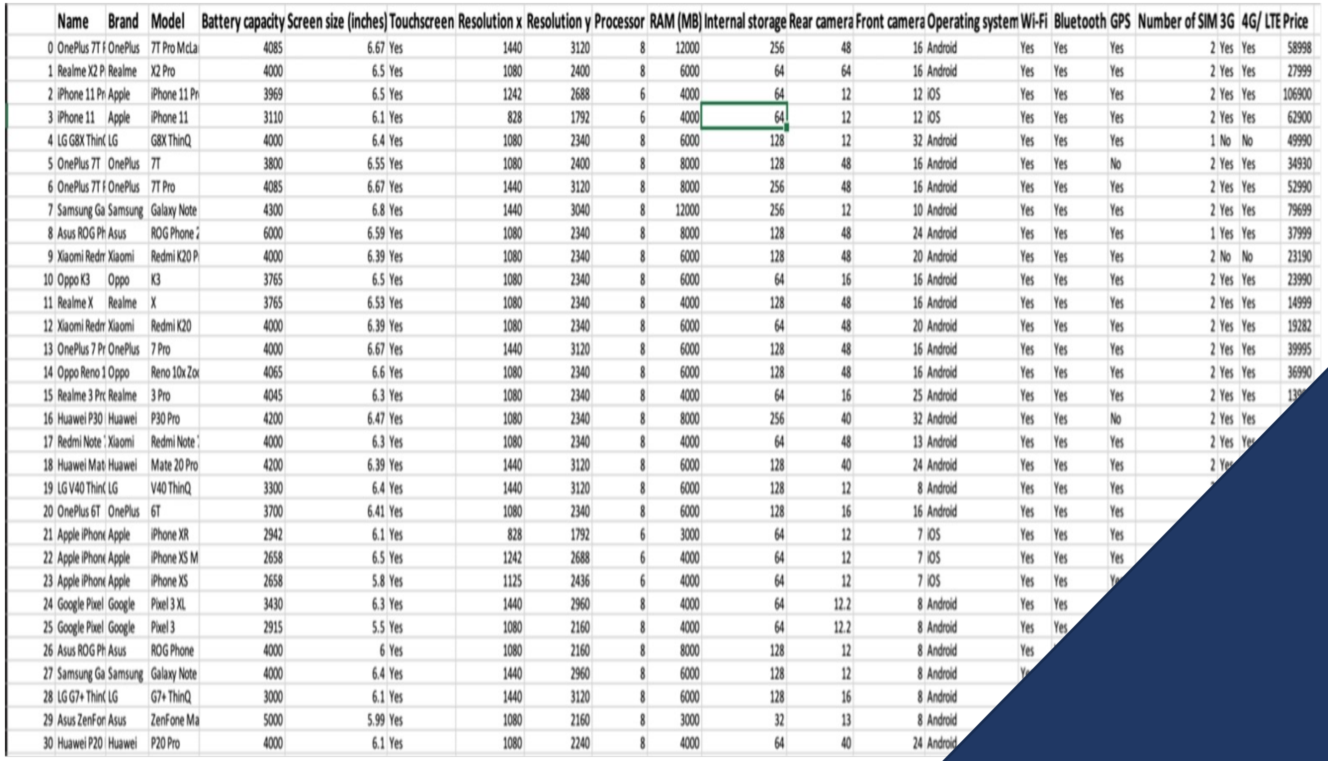
SUBMITTED BY: Peter Chukwuka Ibeabuchi

INTRODUCTION

In this assessment, we explore a comprehensive analysis of a phone distribution dataset. We looked at the descriptive statistics using various graphs and charts, as well as the central and variational measures. The Chebyshev's rule and the box plot technique is deployed to identify outliers, and various probability models are proposed to quantify uncertainty in various variables. Lastly, we carried out various statistical test to assess the independence of two categorical variables, a goodness of fit test for class frequency probabilities, and a test of the mean for a continuous variable.

THE DATASET

The dataset used for this analysis is taken from Kaggle and can be downloaded here. The dataset contains both numerical and object variables. The dataset consists of 22 columns and 1,359 rows. Here is a summary of the columns in the dataset. **It is important to note that the price in this dataset is given in India Rupees (INR)**

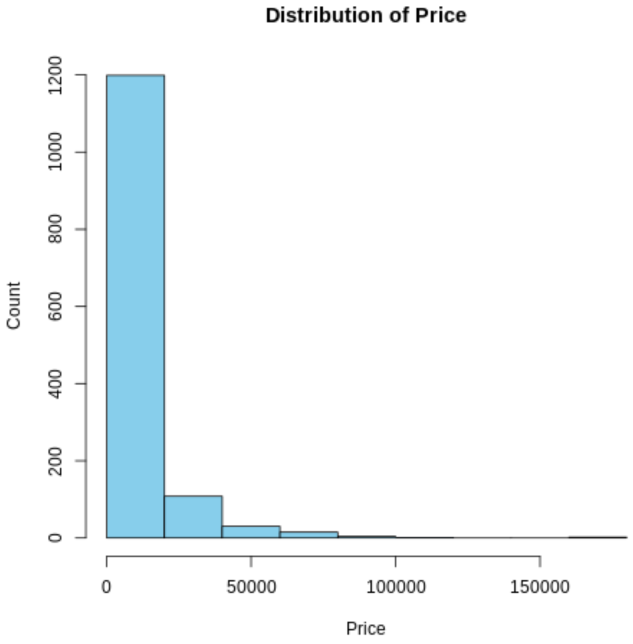


	Name	Brand	Model	Battery capacity	Screen size (inches)	Touchscreen	Resolution x	Resolution y	Processor	RAM (MB)	Internal storage	Rear camera	Front camera	Operating system	Wi-Fi	Bluetooth	GPS	Number of SIM	3G	4G/ LTE	Price
0	OnePlus 7T F	OnePlus	7T Pro McLa	4085	6.67	Yes	1440	3120	8	12000	256	48	16	Android	Yes	Yes	Yes	2	Yes	Yes	58998
1	Realme X2 P	Realme	X2 Pro	4000	6.5	Yes	1080	2400	8	6000	64	64	16	Android	Yes	Yes	Yes	2	Yes	Yes	27999
2	iPhone 11 Pr	Apple	iPhone 11 Pr	3969	6.5	Yes	1242	2688	6	4000	64	12	12	iOS	Yes	Yes	Yes	2	Yes	Yes	106900
3	iPhone 11	Apple	iPhone 11	3110	6.1	Yes	828	1792	6	4000	64	12	12	iOS	Yes	Yes	Yes	2	Yes	Yes	62900
4	LG G8X Thin	LG	G8X ThinQ	4000	6.4	Yes	1080	2340	8	6000	128	12	32	Android	Yes	Yes	Yes	1	No	No	49990
5	OnePlus 7T	OnePlus	7T	3800	6.55	Yes	1080	2400	8	8000	128	48	16	Android	Yes	Yes	No	2	Yes	Yes	34930
6	OnePlus 7T F	OnePlus	7T Pro	4085	6.67	Yes	1440	3120	8	8000	256	48	16	Android	Yes	Yes	Yes	2	Yes	Yes	52990
7	Samsung Ga	Samsung	Galaxy Note	4300	6.8	Yes	1440	3040	8	12000	256	12	10	Android	Yes	Yes	Yes	2	Yes	Yes	79699
8	Asus ROG Ph	Asus	ROG Phone 2	6000	6.59	Yes	1080	2340	8	8000	128	48	24	Android	Yes	Yes	Yes	1	Yes	Yes	37999
9	Xiaomi Redm	Xiaomi	Redmi K20 P	4000	6.39	Yes	1080	2340	8	6000	128	48	20	Android	Yes	Yes	Yes	2	No	No	23190
10	Oppo K3	Oppo	K3	3765	6.5	Yes	1080	2340	8	6000	64	16	16	Android	Yes	Yes	Yes	2	Yes	Yes	23990
11	Realme X	Realme	X	3765	6.53	Yes	1080	2340	8	4000	128	48	16	Android	Yes	Yes	Yes	2	Yes	Yes	14999
12	Xiaomi Redm	Xiaomi	Redmi K20	4000	6.39	Yes	1080	2340	8	6000	64	48	20	Android	Yes	Yes	Yes	2	Yes	Yes	19282
13	OnePlus 7 Pr	OnePlus	7 Pro	4000	6.67	Yes	1440	3120	8	6000	128	48	16	Android	Yes	Yes	Yes	2	Yes	Yes	39995
14	Oppo Reno 1	Oppo	Reno 10x Zo	4065	6.6	Yes	1080	2340	8	6000	128	48	16	Android	Yes	Yes	Yes	2	Yes	Yes	36990
15	Realme 3 Pr	Realme	3 Pro	4045	6.3	Yes	1080	2340	8	4000	64	16	25	Android	Yes	Yes	Yes	2	Yes	Yes	139
16	Huawei P30	Huawei	P30 Pro	4200	6.47	Yes	1080	2340	8	8000	256	40	32	Android	Yes	Yes	No	2	Yes	Yes	
17	Redmi Note	Xiaomi	Redmi Note	4000	6.3	Yes	1080	2340	8	4000	64	48	13	Android	Yes	Yes	Yes	2	Yes	Ye	
18	Huawei Mat	Huawei	Mate 20 Pro	4200	6.39	Yes	1440	3120	8	6000	128	40	24	Android	Yes	Yes	Yes	2	Ye		
19	LG V40 Thin	LG	V40 ThinQ	3300	6.4	Yes	1440	3120	8	6000	128	12	8	Android	Yes	Yes	Yes				
20	OnePlus 6T	OnePlus	6T	3700	6.41	Yes	1080	2340	8	6000	128	16	16	Android	Yes	Yes	Yes				
21	Apple iPhon	Apple	iPhone XR	2942	6.1	Yes	828	1792	6	3000	64	12	7	iOS	Yes	Yes	Yes				
22	Apple iPhon	Apple	iPhone XS M	2658	6.5	Yes	1242	2688	6	4000	64	12	7	iOS	Yes	Yes	Yes				
23	Apple iPhon	Apple	iPhone XS	2658	5.8	Yes	1125	2436	6	4000	64	12	7	iOS	Yes	Yes	Ye				
24	Google Pixel	Google	Pixel 3 XL	3430	6.3	Yes	1440	2960	8	4000	64	12.2	8	Android	Yes	Yes					
25	Google Pixel	Google	Pixel 3	2915	5.5	Yes	1080	2160	8	4000	64	12.2	8	Android	Yes	Yes					
26	Asus ROG Ph	Asus	ROG Phone	4000	6	Yes	1080	2160	8	8000	128	12	8	Android	Yes						
27	Samsung Ga	Samsung	Galaxy Note	4000	6.4	Yes	1440	2960	8	6000	128	12	8	Android	Ye						
28	LG G7+ Thin	LG	G7+ ThinQ	3000	6.1	Yes	1440	3120	8	6000	128	16	8	Android							
29	Asus ZenFon	Asus	ZenFone Ma	5000	5.99	Yes	1080	2160	8	3000	32	13	8	Android							
30	Huawei P20	Huawei	P20 Pro	4000	6.1	Yes	1080	2240	8	4000	64	40	24	Androi							

Question 1(A)

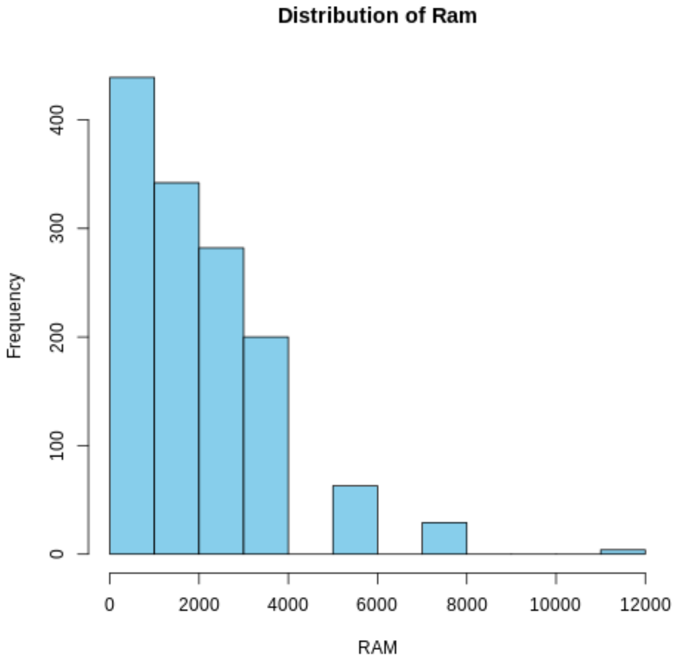
Descriptive Analysis

Question: What is the distribution of phone prices?



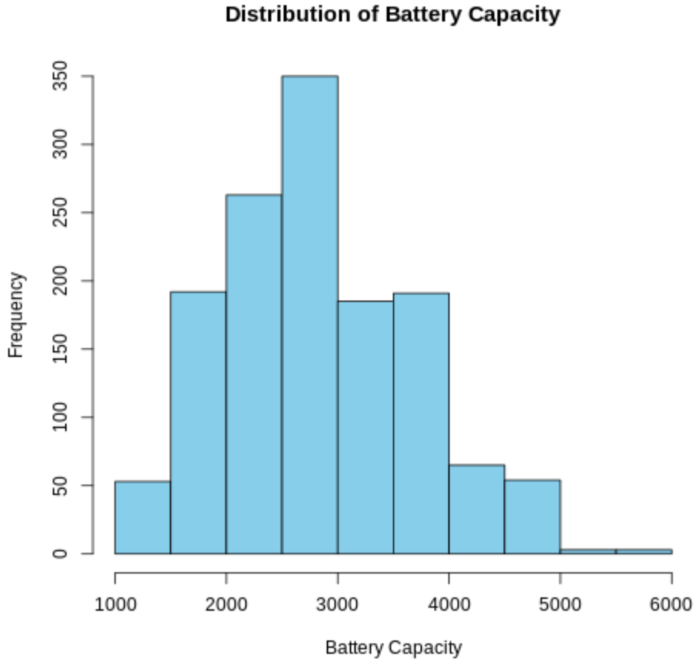
From this we can tell that most of the phones in the dataset cost more than 50,000 INR

Question: What is the distribution of RAM in the dataset?



The chart shows that a good percentage of the phones have 4000 RAM and below.

Question: What is the distribution of the Battery capacity

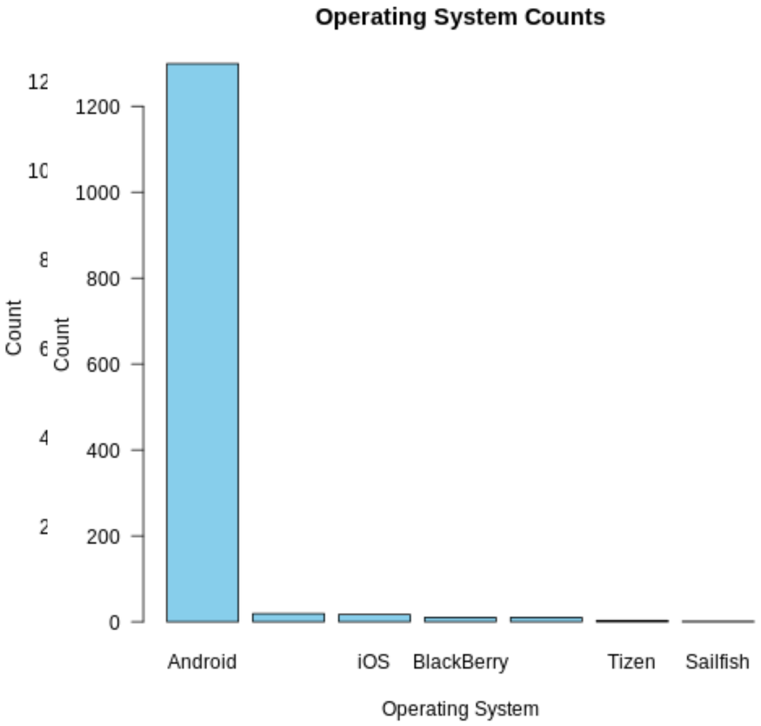


The distribution of Batter capacity has somewhat of a near normal distribution. Majority of the phones have within 2000 and 40000 mAh capacity.

Question 1(A)

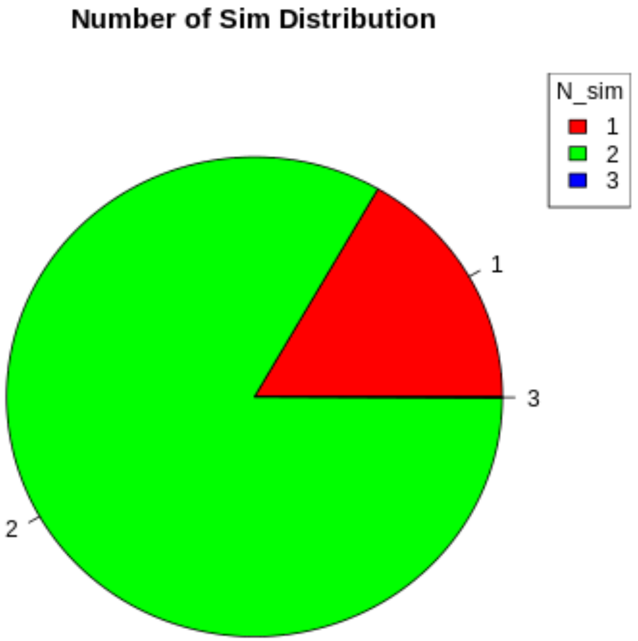
Descriptive Analysis(CONTD)

Question: What is the most common operating system?



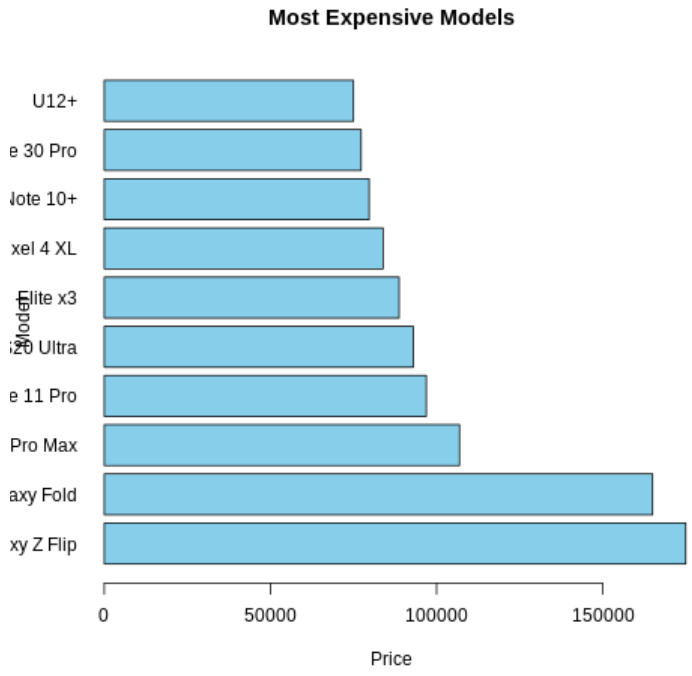
From the bar chart, we can see that the most common OS in the data set is the andriod OS.

Question: What is the distribution of SIMs in the dataset?



The pie chart shows that a good percentage of the Phones have 2 sim ports, very few phones have 3 sims.

Question: What is the most expensive phone in the dataset?



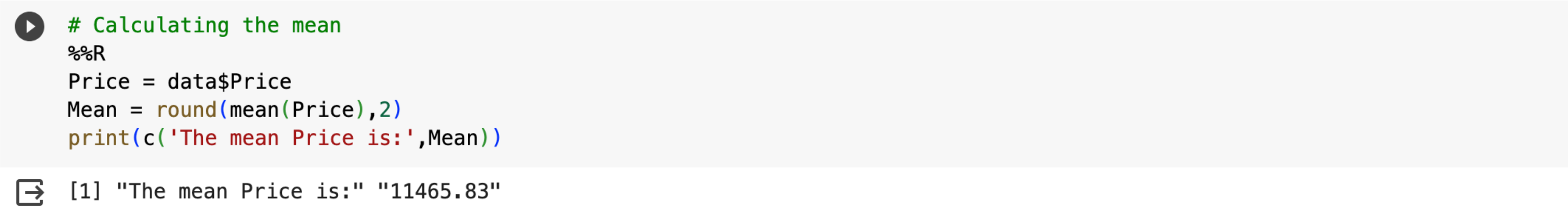
The samsung galaxy z flip and smasung galaxy fold are the most expensive phones in the dataset, costing over 150,000 INR **.**

Question 1(B)

Central And Variational Measures

THE MEAN

In our analysis, the mean price serves as a key metric to understand the central tendency of the dataset. The mean price provides insight into the average cost of phones in the dataset.



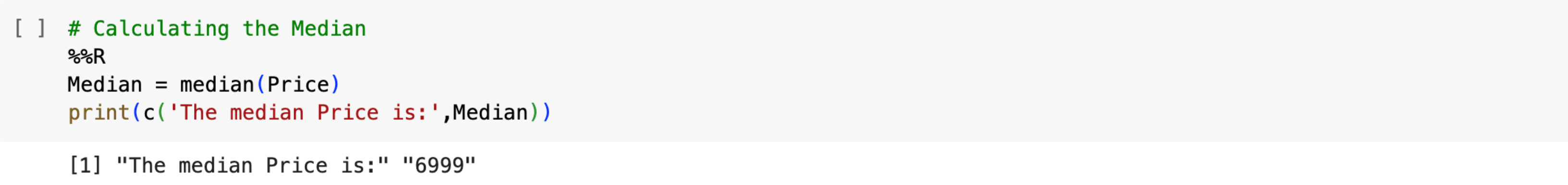
```
# Calculating the mean
%%R
Price = data$Price
Mean = round(mean(Price),2)
print(c('The mean Price is:',Mean))
```

```
[1] "The mean Price is:" "11465.83"
```

From the image above, we can see that the average cost of phones in the dataset is 11,465.83 INR.

THE MEDIAN

The median price represents the middle price in our dataset, indicating the price at which half of the phones are priced below and half are priced above.



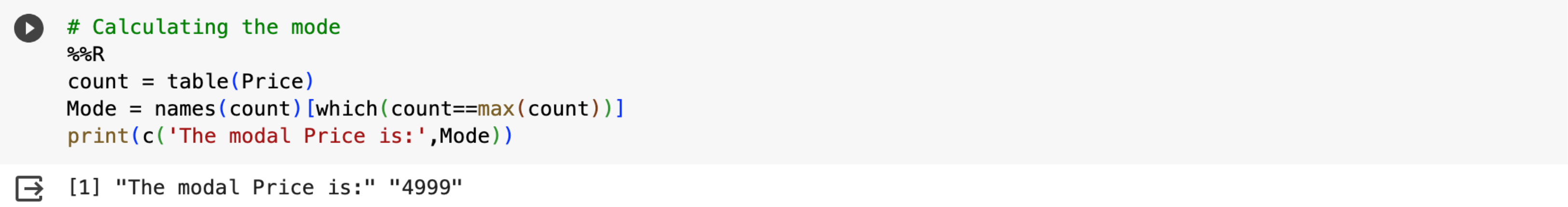
```
[ ]  # Calculating the Median
%%R
Median = median(Price)
print(c('The median Price is:',Median))
```

```
[1] "The median Price is:" "6999"
```

From the table, we can see that the median price for phones in the dataset is 6,999 INR.

The MODE

In our analysis of the phone price dataset, understanding the mode price is crucial for identifying the most common cost for phone.



```
# Calculating the mode
%%R
count = table(Price)
Mode = names(count)[which(count==max(count))]
print(c('The modal Price is:',Mode))
```

```
[1] "The modal Price is:" "4999"
```

From the image above, we see that most phones cost 4999 INR

DEFINITIONS

The mean, often referred to as the average, is the sum of all values divided by the number of observations.

The median price is a statistical measure of central tendency that represents the middle value of a dataset when arranged in ascending order. Specifically, it is the value that separates the higher half from the lower half of the data.

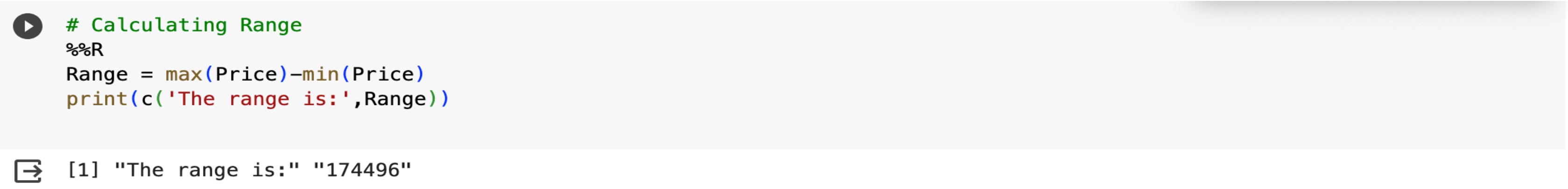
The mode is the most occuring frequency or value in a set of data values

Question 1(B)

Central And Variational Measures

THE RANGE

In our analysis, understanding the range of prices is crucial for gauging the overall spread and variability in pricing.

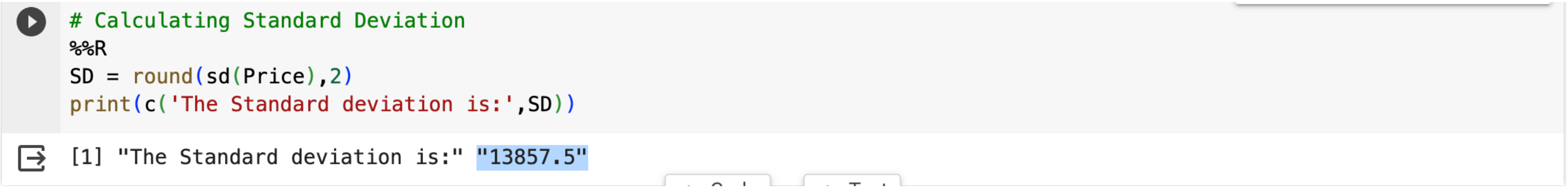


```
# Calculating Range
%%R
Range = max(Price)-min(Price)
print(c('The range is:',Range))
```

```
[1] "The range is:" "174496"
```

STANDARD DEVIATION

In our analysis, the standard deviation provides insights into how individual prices deviate from the average, giving us a better understanding on the distribution of prices.

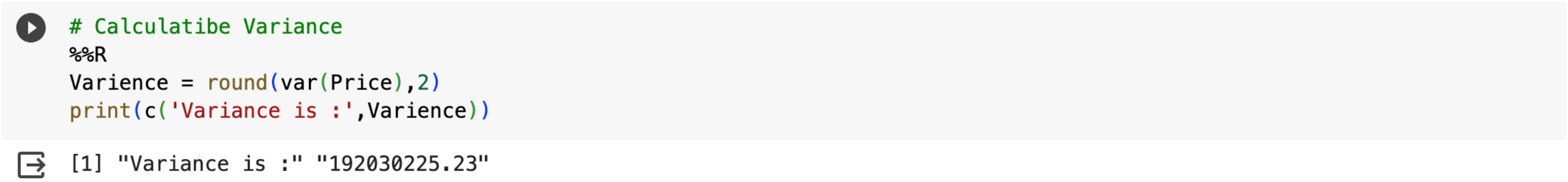


```
# Calculating Standard Deviation
%%R
SD = round(sd(Price),2)
print(c('The Standard deviation is:',SD))
```

```
[1] "The Standard deviation is:" "13857.5"
```

THE VARIANCE

In our analysis,, understanding the variance in prices is essential for gaining insights into how much individual phone prices deviate from the average cost(mean).



```
# Calculatibe Variance
%%R
Varience = round(var(Price),2)
print(c('Variance is :',Varience))
```

```
[1] "Variance is :" "192030225.23"
```

DEFINITIONS

The range of a dataset is the difference between the maximum value and the minimum value within a collection of number.

The standard deviation is a measure of the amount of variation from the mean. A low standard deviation indicates that the data points tend to be close to the mean, while a high standard deviation indicates that the data points are spread out over a wider range of values.

The variance is a measure of how much the values in a set of data vary from the mean. A low variance indicates that the data points tend to be close to the mean, while a high variance indicates that the data points are spread out over a wider range of values

Question 1(C,D)

OUTLIER DETECTION

CHEBYSHEV’S RULE AND PROPOSED ONE-SIGMA INTERVAL (C)

Chebyshev's rule is a statistical principle that states that for any dataset with a finite mean and standard deviation, a certain proportion of the data points will fall within a specified number of standard deviations from the mean. The formula for Chebyshev's rule is:

$P(\mu - k\sigma \leq X \leq \mu + k\sigma) \geq 1 - K\hat{}2$

Where:

•P =probability; .μ =mean; σ= standard deviation ; k represents the number of SD from the mean.

One sigma interval proposal indicates that K is = 1, which implies that all data points should be within one standard deviation of the mean. Data points outside this is are identified as outliers. Thus, the ranges for the data is given as follows

Lower_Range = Mean - k * SD

Upper Range = Mean + k * SD

In our analysis, we identified 126 outliers in the price column using the Chebyshev's Rule parameters, with a one-sigma interval

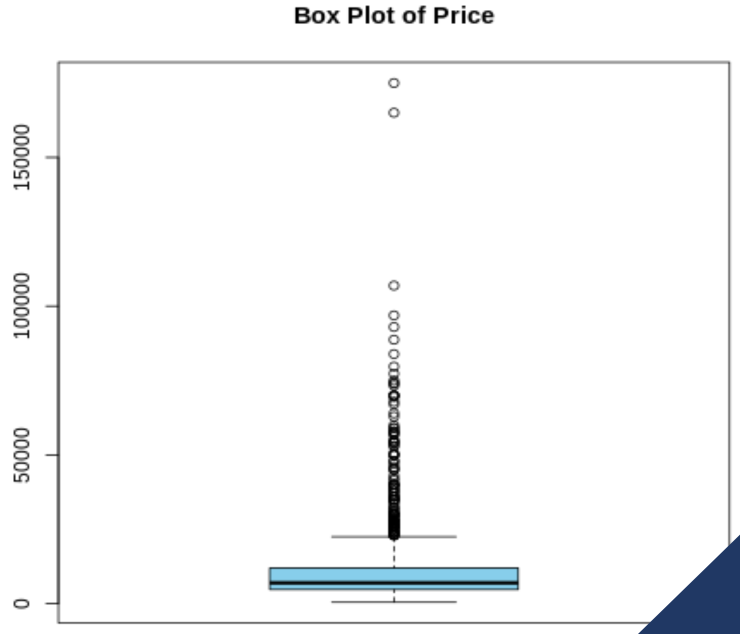
THE BOX PLOT TECHNIQUE FOR OUTLIER DETECTION (D)

A box plot is a graphical representation of data that summarizes key features of a data set. It consists of a box, whiskers, and outliers. The box represents the interquartile range (IQR), which is the 50% of the data that falls between the 25th and 75th percentiles. All data points that fall outside of the whiskers are considered outliers.

To calculate the number of outliers in a dataset, we define the 1st and 3rd quantile of the data, as well as the IQR(Q1-QR). A formula to workout the upper and lower range is as follows:

- Lower Range = Q1 - 1.5 * IQR
- Upper Range = Q3 + 1.5 * IQR

In our analysis, we identified 22 outliers in the price column using the box plot technique.



Question 2(A,B)

PROBABILITY MODEL

IDENTIFYING COLUMNS AND THE APPROPRIATE PROBABILITY MODELS (A)

1.Price: This column contains a continuous distribution that represents the cost for phones and thus we will be using the **Normal Distribution**.
2. Number of sims: This column contains the count of the number of sims present in each device. For this, the appropriate probability model is the **Poisson distribution**
3. Operating Systems: This column contains categorical variables that shows various operating systems of each devices. The appropriate probability model for this distribution is the **Multinomial distribution**
4. Wi-fi: This column contains categorical variables of two options, yes and no, signifying the presence or absent of wi-fi in a device. For this column, the appropriate probability model is the the **Bernoulli Distribution**

ESTIMATE THE PARAMETER FOR EACH MODEL (B)

1.Normal Distribution: In a normal distribution, the parameters typically refer to the **mean (μ)** and **standard deviation (σ)**. These parameters define the shape and location of the distribution. In our analysis, the parameter for the price column **is mean = 11465.83** and **SD= 13857.5**

2. Poisson Distribution: In a Poisson distribution, the parameter is often denoted by the symbol λ (lambda), which represents the average rate of occurrence or the average number of events in a fixed interval of time or space, thus the mean. In our analysis, the parameter for number of sims columns is **Lamda = 1.83**

3.Multinomial Distribution: In a multinomial distribution, the parameters are probabilities assigned to different outcomes, and they are typically represented by the vector **P = (p1,p2,...,pk)** where k is the number of possibilities. In our analysis, the probability for each operating system is **Android: 0.956; BlackBerry: 0.007; Cyanogen: 0.007; iOS: 0.013;Sailfish: 0.001; Tizen: 0.002; Windows: 0.014**

4. Bernoulli Distribution. In a Bernoulli distribution, there is a single parameter denoted by **p**, which represents the probability of success. The probability of failure is = **1 – p.** In our analysis, the probability for **success is = 0.99**

Question 2(C)

PROBABILITY MODEL(CONTD)

MAKING PREDICTIONS WITH EACH MODELS

1.Normal Distribution: In R, we use the 'pnorm' function to calculate the CDF of a normal distribution. It is given as

pnorm(x, mean, sd) ,

-where x represents the value for which we want to calculate the probability.

Using the price column, the probability if randomly selected Phone costs less than or equal to 10,000 is = 0.457879 and the probability that price is greater than 10,000 is = 0.542121

2.Poisson Distribution: In R, the PMF for Poisson Distribution is given as

'dpois(x, lambda)',

- where x represents the number of events or occurrences, lambda is the average rate of occurrence

Using the number of sim column, the probability of a phone having 2 SIMs, i.e P(X=2) is = **0.2686958**

3. Multinomial Distribution: In R, we use the 'dmultinom' function to calculate the PMF of a multinomial distribution. It is denoted as

dmultinom(x,n,p)

- Where x = the observed counts for each category, n = number of trials, p = Probabilities for each category

Using the Operating system column, the probability of getting 3 Andriod, 2 blackbery, 2 Cyanogen, 1 ios and 1 windows is = **6.77028e-09**

4. Bernoulli distribution: In R, we use the 'dbinom' function to calculate the PMF for a Bernoulli distribution. It is denoted as

dbinom(j,1,p)

- where; j is the possible outcome and p is the probability of getting one

Using the Wifi column, the probability of getting a phone with a wifi? P(X=1) = `0.99`

Question 3(A,B,C)

HYPOTHESIS TEST

TEST OF INDEPENDENCE (A)

For this we compare the Wi_fi and Bluetooth columns.
alpha =0.01

Step 1: State the hypothesis
H0: Wifi is independent of Bluetooth
H1: Wifi is dependent of Bluetooth

Step 2: Find the statistic and critical values
The test value is : 97.65977
The critcal value is : 13.2767

Step 3: Decision rule
if the test value is greater than the c value reject H0, else accept H0.

Since the test.value is greater than c.value then we accept H0. This implies that on a siginificance level of 0.01, the two categorical variable Bluetooth and Wifi are dependent on each other

GOODNESS OF FIT (B)

Here we use the Operating system column.
alpha =0.05

Step 1: State the Hypothesis
H0: p1 = p1=p2=p3=p4=p5=p6=p7
H1: NOT H0

Step 2: Find the statistic and critical values
The test value is : 7336.973
The critcal value is : 1444.844

Step 3 Decision rule
If the test value is greater than the c value reject H0, else accept H0.

Since the test.value is greater than c.value then we reject H0

TEST OF MEAN (C)

Here, Using the price columns, let MuO = 10,000.
alpha =0.05

Step 1: State the Hypothesis
H0: mu == 10,000
H1: mu =/=10,000

Step 2: Find the statistic and critical values
The test value is : 3.899485
The Critcal value is : -1.959964

Step 3: Decision rule
if the test value is greater than the c value reject H0, else accept H0.

Thus, since the test.value is greater than c.value then we reject H0. This implies that the average price of phone in the dataset is not 10,000.

References:

- Frost, J. (2021, May 10). Measures of Central Tendency: Mean, Median, and Mode. Statistics by Jim. Retrieved from https://statisticsbyjim.com/basics/measures-central-tendency-mean-median-mode/
- Dr Shahram lecture notes