Национальный исследовательский университет

«Высшая школа экономики»

Факультет мировой экономики и мировой политики

Образовательная программа «Мировая экономика»

Отчет о самостоятельной работе

по дисциплине «Эконометрика» на тему

«A Study of the Mobile Phone Market in India»

Выполнена:

Фам Тху Чанг, БМЭ204

Москва 2022

Table of Contents

1. General formulation of the research

Introduction

Over the last 30 years, the mobile phone market has drastically been evolving with the massive development of technology and the increase in demand for a digital phone among the worldwide population. The mobile phone market is undoubtedly one of the most profitable and rapidly-growing markets in the technology industry nowadays.

The mobile phone market in India has grown into one of the largest in the world, thanks to rising incomes, low internet prices and the need to always stay connected everywhere. Despite being affected by COVID-19, phone sales reached a peak in 2021, partly thanks to Chinese low-cost phone brands. In India, the phone brands that dominate the market are not Samsung or Apple but Chinese ones: Xiaomi, Vivo, Realme. In the future, the demand for smartphones in India will increase very quickly, the competition of phone brands will not only be about the features of the phone but also the price.

Formulation of the research's goal and tasks

Our project aims to be relevant for potential customers in India who are in the process of making a decision to purchase a mobile phone that suits their preferences and needs to purchase. At the same time, the result of our study will help mobile phone manufacturers who intend to enter the Indian mobile phone market and study necessary features to determine their prices in a reasonable way.

Hence, our task is to study the association between a phone's price with its most basic features that characterize the market of mobile phone in India as well as attract specifically Indian customers. In particular, in order to understand the Indian mobile phone market, we will consider price our first priority, since it is among the most important (and indeed the most for the majority of customers) criteria when it comes to choosing a suitable mobile phone to buy. Additionally we have chosen such features as internal storage, screen size, battery capacity, number of sim cards, cellular network, operating system, and brand's credibility in order to learn how price can be determined by them.

Applied task: Study the behaviors of a telephone's price on the Indian market under the following conditions: battery capacity of at least 2600 mAh (i.e. the phone can be used for at least 2.5 hours), size

ranging from 5 to 7 inches, and internal storge from 32GB, system operated by Android. In addition, phones from all brands, regardless of their reputation or experience in producing mobile phones, are included in the research. However, we would look distinctively at phones produced by companies with the lowest brand value (brand 1).

In order to achieve the goals, we structure our research based on these following guidelines:

1. Conduct a preliminary analysis of data characteristics and the statistical relationship between them;
2. Create a regression model based on preliminary data analysis and test its predictive properties;
3. Identify the optimal final model to test all formulated hypotheses and draw conclusions.

1.1. General description of data

The sample consists of 1358 observations and 1 dependent variable and 7 independent.

No.	Object's characteristic	Name of variable	Scale of measurement	Role
1	**Price** *(US dollars)*	price	relative	dependent
2	Brand *(ranked by brand value)* 1 – Others 2 – Motorola, Asus, Tecno, LG, ZTE, Nokie, HTC, Panasonic, Honor, Alcatel 3 – OnePlus, Oppo, Vivo, Sony, Realme 4 – Apple, Samsung, Google, Xiaomi, Huawei	brand	ordinal	independent
3	Battery capacity *(mAh)*	battery	relative	independent
4	Screen size *(inches)*	size	relative	independent
5	Internal storage (*GB)*	storage	relative	independent
6	Operating system 1 – Android 0 – Others	system	nominal	independent
7	Number of sim cards 1 – 2 SIMs 0 – 1 SIMs	sims	nominal	independent
8	The presence of mobile network 4G/LTE 1 – Yes 0 – No	4G/LTE	nominal	independent

1.2. Main hypotheses which are planned to be examined in the framework of solving the tasks

In the framework of our study, we will take into consideration these following hypotheses:

1. The price of a telephone positively depends on its internal storage
2. There exists an exponential dependence of a telephone's price on its battery capacity (i.e., a telephone's battery capacity has a positive effect on its price, and after a certain value, the growth of price increases)
3. There exists an exponential dependence of a telephone's price on its screen size (i.e., a telephone's screen size has a positive effect on its price, and after a certain value, the growth of price increases)
4. The brand value of a telephone has different effects on how its internal storage determines its price

   (i.e., it is not always the case that on the rise of the internal storage would the price increase, as this also depends on the brand value)

1.3. Main source of data

The main source of data is taken from Kaggle, which gathers information about the features of phones from 75 brands in the Indian mobile phone market.

> Mobile Phone Specifications and Prices
> URL: https://www.kaggle.com/datasets/pratikgarai/mobile-phone-specifications-and-prices

Throughout the research progress, we used the programming language of Python and its libraries including pandas, numpy, scipy, seaborn, matplotlib.

2. Preliminary analysis of the collected data

2.1. Analysis of data's characteristics: potential errors and missing values, groups and outliers.

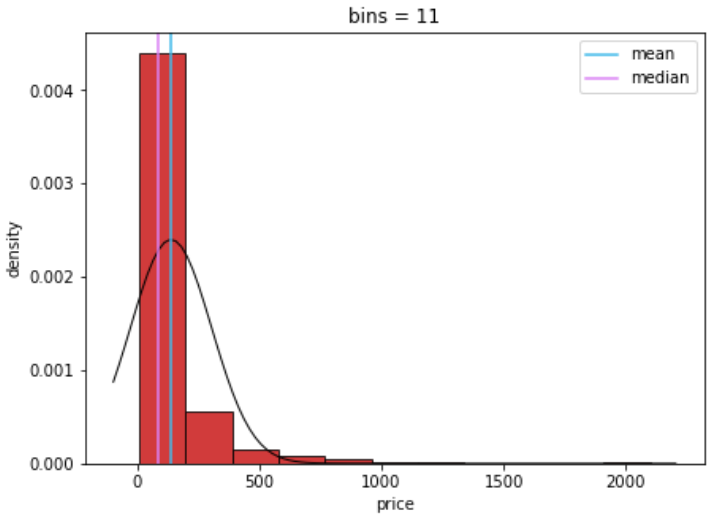
2.1.1. Analysis of quantitative variables

In the analysis of quantitative variables, we calculated the main statistics of the variables of built histograms based on the number of columns (bins) under the Sturge's rule and the Freedman–Diaconis rule. However, it should be taken into consideration that the Sturge's formula works best for samples with less than 200 observations, while our sample includes 1358. Therefore, it would be more precise to pay more attention to the histogram built upon the formula of Freedman–Diaconis.

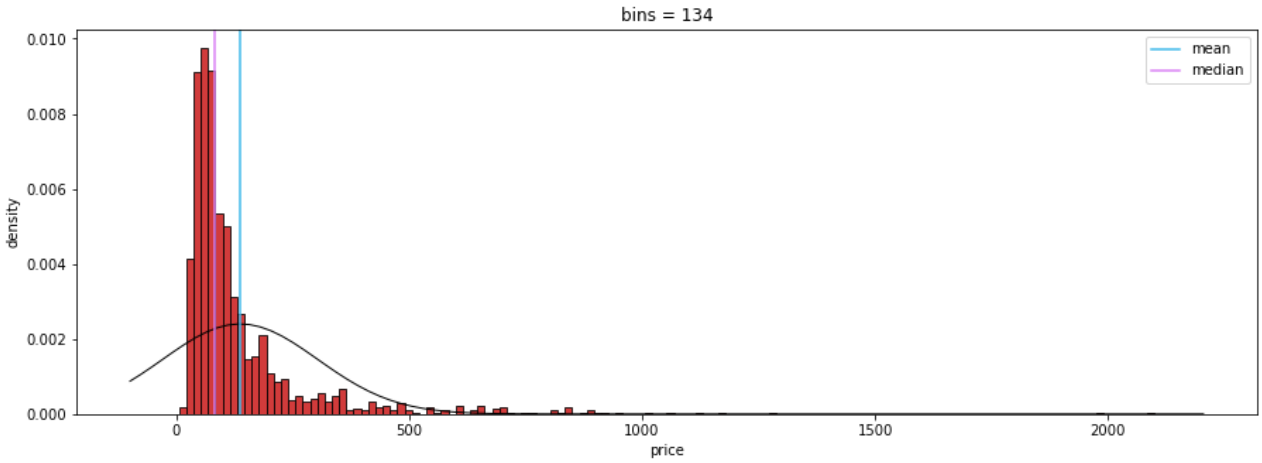
Price

Mean	137.65
Median	83.99
Standard deviation	166.34
IQR	86.63
Q3	143.99
Q1	57.36
Skewness	4.6
Kurtosis	33.33
Number of observations	1358
Number of missing observations	0

Table 1: Descriptive statistics of price.



Graph 1: Histogram of price (Sturges' rule)



Graph 2: Histogram of price (Freedman – Diaconis's rule)

From the graphs, we can see that the distribution of price is skewed to the right *(average > median)* and it is quite heavily skewed *(skewness = 4.606)*. The majority of mobile phones in the market has a price that is less than average, which is understandable because the demand for this product is huge from the lower class to the higher class. However, not everyone can afford or feel the need to purchase a highly priced phone, especially in India where there is an enormous imbalance of wealth between the rich and the poor as 77% of national wealth belongs to the top 10% of the population[1]. Hence, the share of low-price phones in the dataset corresponds with the share of people with a low budget in the population of India.

As can be seen, the graph is not a normal distribution and is sharp with many peaks *(kurtosis = 33.3)*, which implies heterogeneity. The peaks are distributed mostly to the right of average price, which means that there is a number of high-price mobile phones in the market with and the values of price differ a lot. This makes sense because the demand for owning such phones probably belongs to the wealthier part of population in India.

Based on our observation, there is no presence of polymodality and hence, clusters.

Left	Right	Number of outliers
-361.358	636.659	35

Table 2: Outliers for price (3-sigma rule)

[1] https://www.statista.com/statistics/482579/india-population-by-average-wealth/

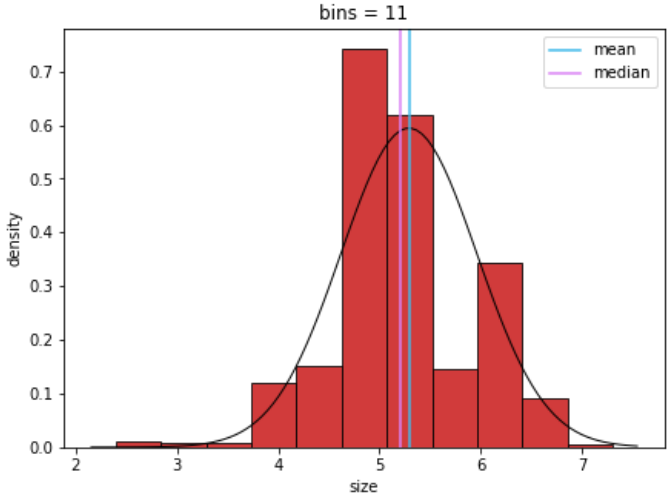
The result after searching for outliers based on the three-sigma rule showed 35 anomalous observations, whose range is from $647.8 to $2099.9. It is obvious that values in this range deviate far from the median *($83.99)*, which explains a fairly wide gap between the median and the average price *($137.65)*. Phones with price in this range mostly have 2 SIMs, a connection to the 4G/LTE network and the Android operating system. However, while more than half of these phones are from valuable brands *(Group 4),* there are still representatives of other groups. At the same time, the ranges of values for battery capacity *(2600 – 4500 mAh)*, screen size *(5.0 – 7.3 inches)* and internal storage *(32 – 512 GB)* are also quite wide. Therefore, we decided not to delete these outliers and keep them for further parts of our study.

Conclusion: Heavily right skewness, high positive kurtosis, non normal distribution.

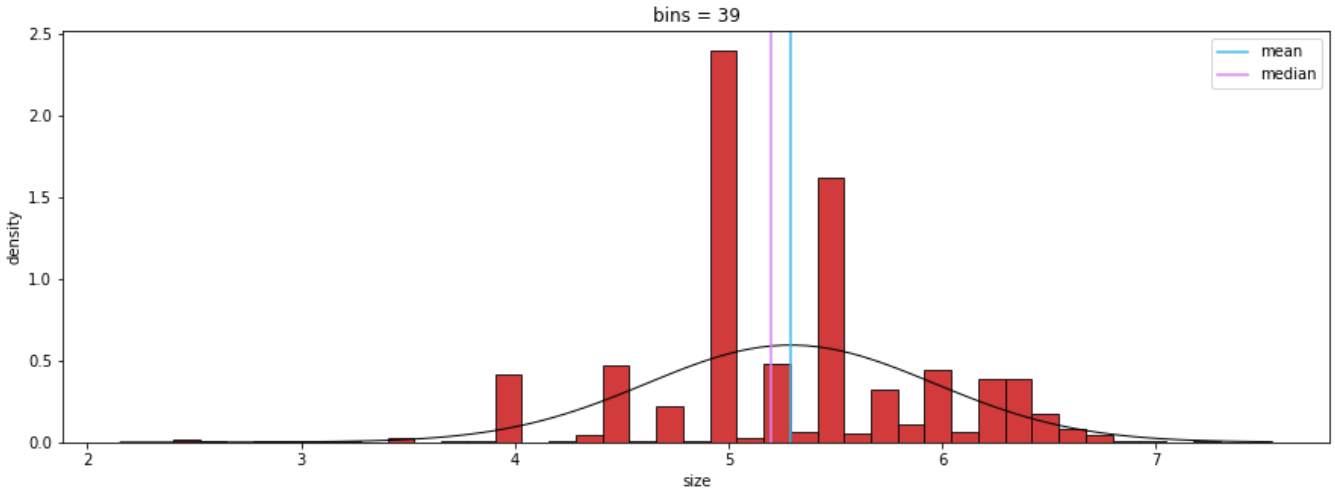
Screen size

Mean	5.29
Median	5.2
Standard deviation	0.67
IQR	0.7
Q3	5.7
Q1	5
Skewness	- 0.32
Kurtosis	0.94
Number of observations	1358
Number of missing observations	0

Table 3: Descriptive statistics of screen size



Graph 3: Histogram of screen size (Sturges' rule)



Graph 4: Histogram of screen size (Freedman – Diaconis's rule)

From the graphs, we can see that the distribution of screen size is almost a normal distribution as its skewness and kurtosis are close to 0 *(skewness = - 0.32, kurtosis = 0.9)* and the difference between median and average is subtle *(gap = 0.09 inches)*. Theoretically, the distribution is slightly skewed to the left.

The majority of mobile phones in the market has a screen size from 4.9 to 5.5 inches, which is understandable because phones of size in this range are convenient to bring everywhere. Smaller-sized phones (less than 4.9 inches) and bigger-sized phones (more than 6.5 inches) are less in number and are distributed with the same amount, as the graph is not sharp with many peaks. This indicates that the demands for both these types of phones are equal.

Based on our observation, there is no presence of polymodality and hence, clusters.

Left	Right	Number of outliers
3.276	7.306	11

Table 4: Outliers for size (3-sigma rule)

The result after searching for outliers based on the three-sigma rule showed 11 anomalous observations, whose range is from 2.4 to 3.2 inches. Values in this range deviate quite far from the median *(5.0 inches)*. Such phones are mostly from low-valued brands *(group 1 and 2)* and have low battery capacity *(1010 – 2180 mAh)*, no 4G/LTE network, a non-Android operating system, and of course, a low price *($15 – $129.6).* This is comprehensible as small-sized phones tend to have only 2G cellular

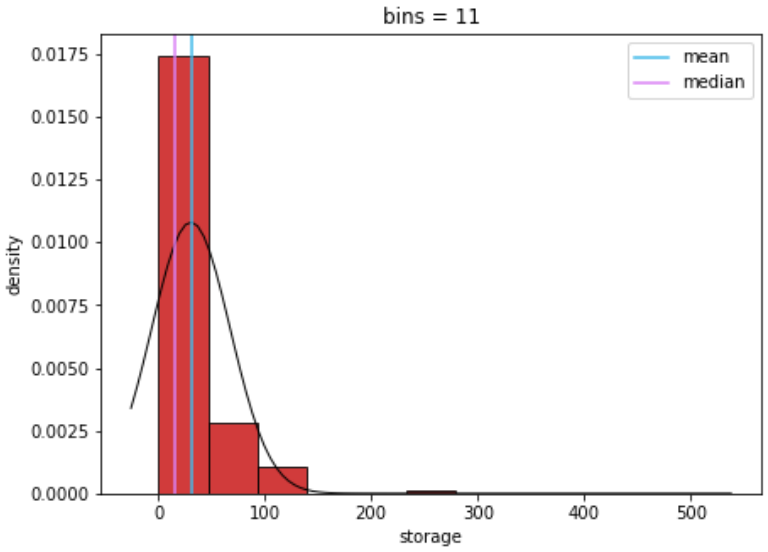
network that satisfy the need to call and send messages of users and belong to the old generation of mobile phones before the touch-screen evolution.

Overall: Slightly left skewness, low positive kurtosis, close to a normal distribution.

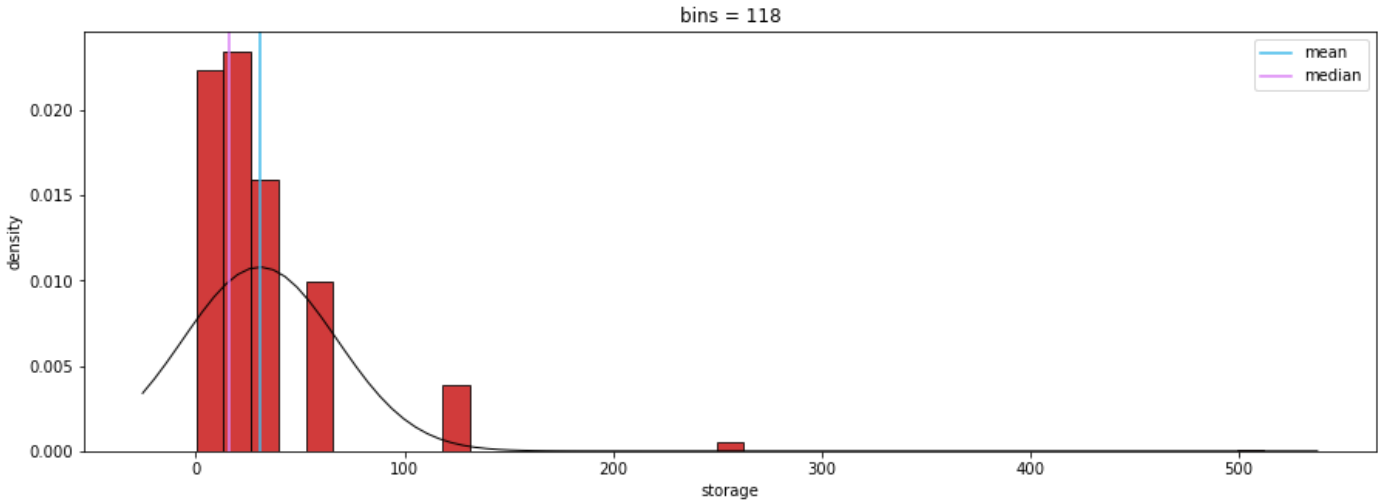
Internal storage

Mean	30.67
Median	16
Standard deviation	36.96
IQR	24
Q3	32
Q1	8
Skewness	4.06
Kurtosis	30.08
Number of observations	1358
Number of missing observations	0

Table 5: Descriptive statistics of internal storage



Graph 5: Histogram of internal storage (Sturges' rule)



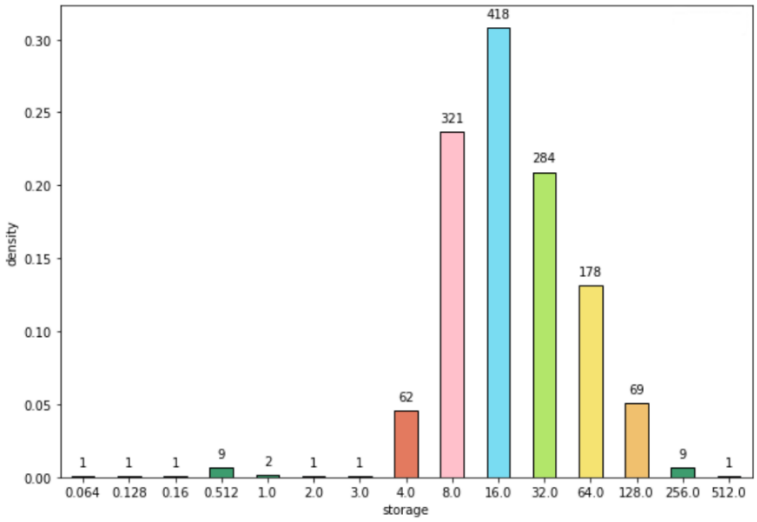
Graph 6: Histogram of battery capacity (Freedman – Diaconis's rule)

From the graphs, we can see that the distribution of price is skewed to the right *(average > median)* and it is quite heavily skewed *(skewness = 4.06).* The majority of mobile phones in the Indian market has internal storage that is equal to or less than 16 GB, which indicates that suppliers lay their focus on phones with just enough capacity. From this, it can be supposed that Indian customers do not have much demand for storing a lot of data in the device, or utilities that take up a lot of internal storage such as taking pictures, downloading games and storing media files such as photos and videos.

As can be seen, the graph is not a normal distribution and is sharp with different peaks (kurtosis = 45,6), which implies heterogeneity.

Based on our observation, there is no presence of polymodality and hence, clusters.

For further understanding, we also attempted to graph a bar chart for internal storage to observe its distribution more clearly. From our observation, mobile phones with internal storage of 8GB, 16GB, 32GB are the most popular, together making up about 75% of the market.



Graph 7: Bar chart of internal storage

Left	Right	Number of outliers
- 80.22	141.55	10

Table 6: Outliers for storage (3-sigma rule)

The result after searching for outliers based on the three-sigma rule showed 10 anomalous observations, only one of which has a value of 512GB while others have 256GB storage. In general, most phones that have either 256 or 512GB of storage have a large screen size *(5.7 – 7.3 inches)*, battery capacity above average *(3000 – 4380)*, two sim cards and are produced by high-value brands of groups 3 and 4. Therefore, it is quite obvious that their prices are also very high *($540 – $2099.9)*, with only one exception of $277.2 for a phone that does not provide a 4G/LTE cellular network.

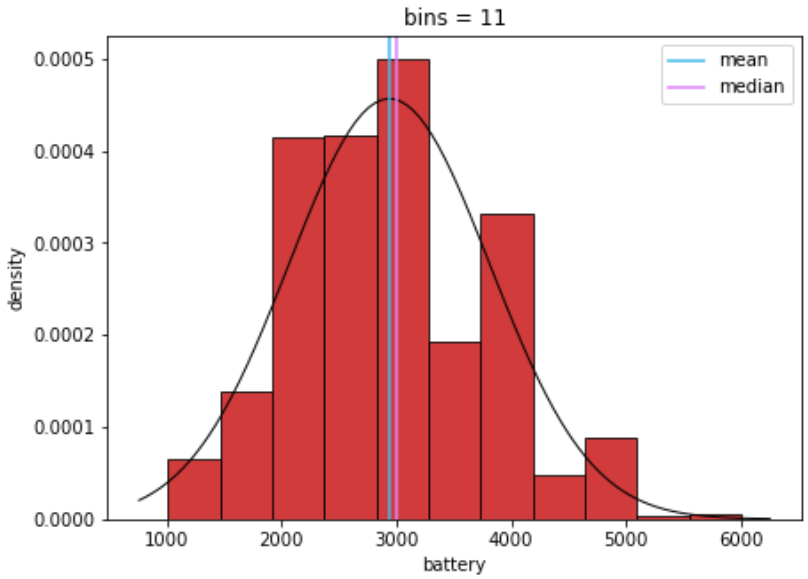
Overall: Heavily right skewness, high positive kurtosis, non normal distribution.

Battery capacity

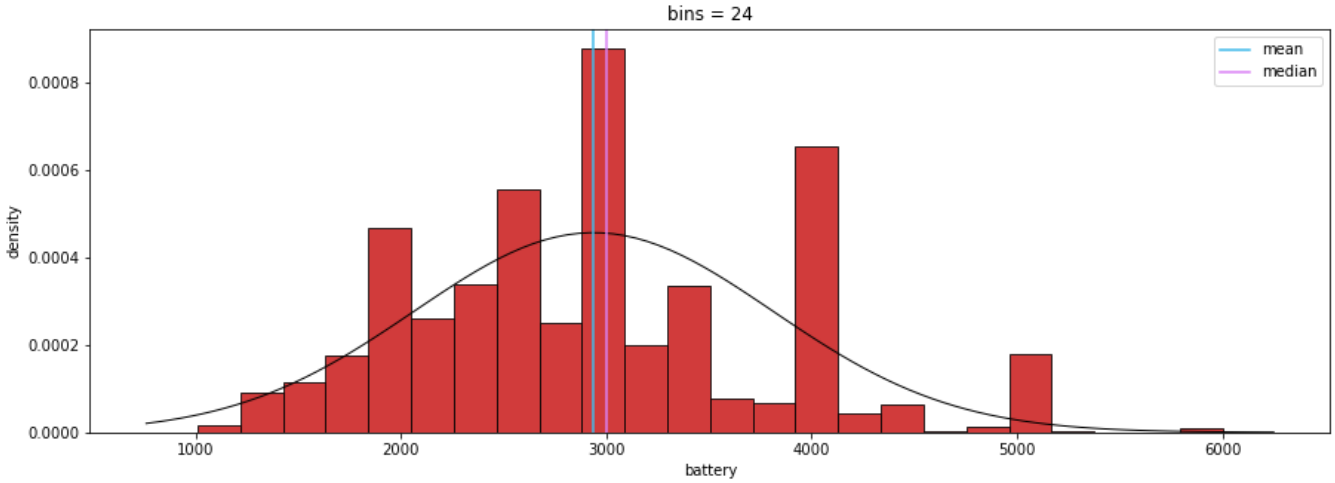
Mean	2938.407
Median	3000
Standard deviation	873.831
IQR	1200
Q3	3500
Q1	2300
Skewness	0.486
Kurtosis	-0.105
Number of observations	1358

Number of missing observations	0

Table 6: Descriptive statistics of battery capacity



Graph 8: Histogram of battery capacity (Sturges' rule)



Graph 9: Histogram of battery capacity (Freedman – Diaconis's rule)

From the graphs, we can see that the distribution of battery capacity is close to a normal distribution as its skewness and kurtosis are quite close to 0 *(skewness = 0.49, kurtosis = - 0.11)* and the difference between median and average is subtle (gap = 61.6 mAh). Theoretically, the distribution is slightly skewed to the right.

The majority of mobile phones in the market has storage capacity of between nearly 2000 and above 4000 mAh and those of about 3000 mAh take up the highest amount in the market. This shows

that mobile phone suppliers tend to sell their products with average battery capacity, which is about 8 – 12 hours of usage, and that corresponds to the need of most customers nowadays.

Based on our observation, there is no presence of polymodality and hence, clusters.

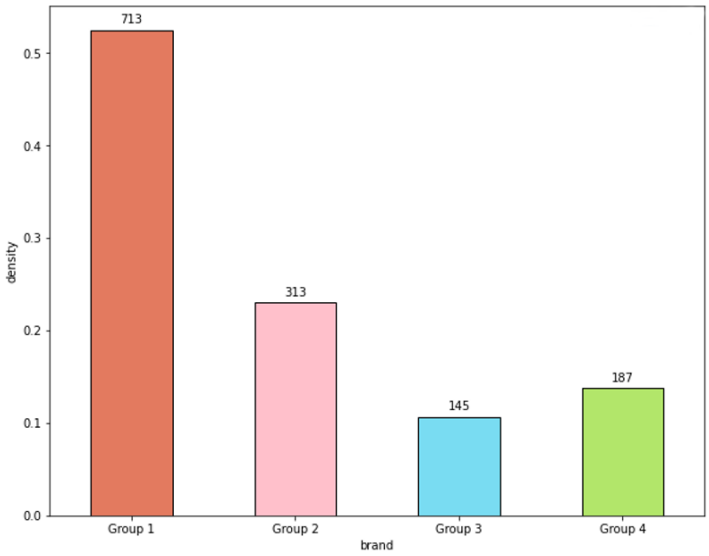
Left	Right	Number of outliers
1010	5300	3

Table 6: Outliers for battery (3-sigma rule)

The result after searching for outliers based on the three-sigma rule showed 3 anomalous observations and they all have storage capacity of 6000 mAh, which doubles the average. These phones with high battery capacity share a common feature, that is rather big screen size *(6.35 – 6.59 inches)*. This signifies that phones with high storage capacity tend to be big in terms of screen size. Apart from this, there is hardly any other common characteristic between these outliers.

2.1.2. Analysis of qualitative variables

Brand

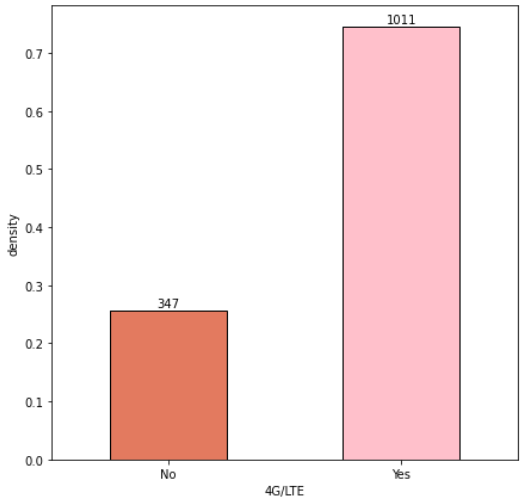


Graph 10: Bar chart of brand

As can be observed, the low-valued brands (group 1) offer the highest amount of mobile phones, which can be explained by the fact that the number of brands in this group is 55, whereas other groups only have 5 or 10 brands. However, on average, each brand of group 1 does not produce as many mobile phones as each brand of the other groups, which we suppose is because they do not have as many financial resources to research, develop and produce as the other groups.

Conclusion: Heterogeneity.

4G/LTE cellular network

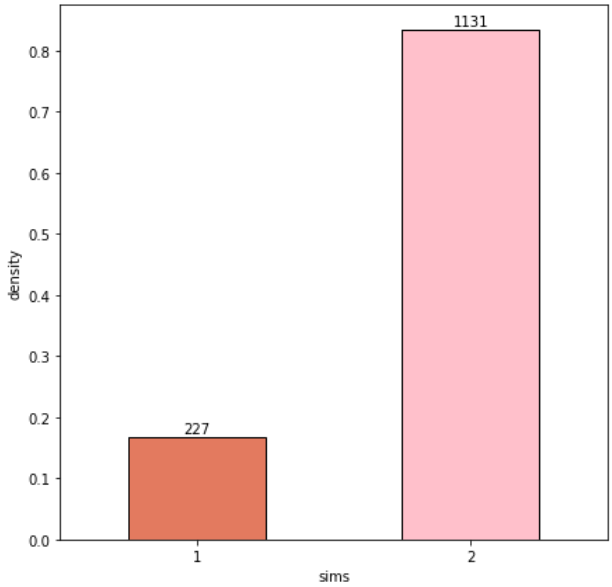


Graph 11: Bar chart of 4G/LTE

On the one hand, more than 70% of mobile phones in the Indian market allow users to access the Internet via 4G/LTE cellular network so that they can use the Internet everywhere. On the other hand, mobile phones without the 4G/LTE network will be suitable with customers who have no need to use the Internet via their mobile phone. These types of phones usually belong to the pre-touchscreen evolution and satisfy mostly the need to call and send messages.

Conclusion: Widespread heterogeneity

Number of sim cards

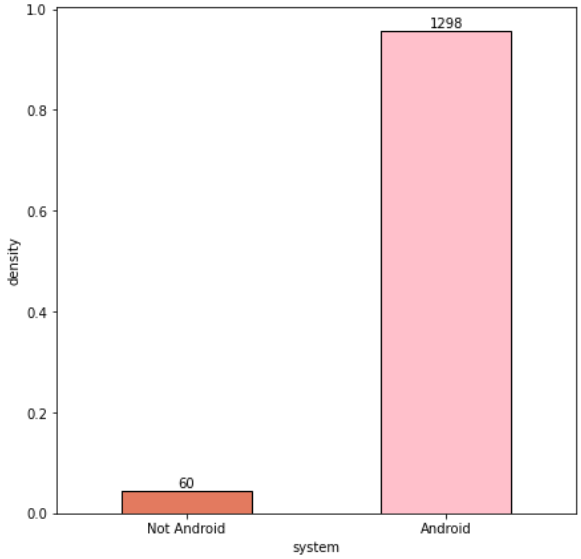


Graph 12: Bar chart of sim

As can be observed, most mobile phones in the Indian market (80%) allow users to install up to 2 sim cards. Therefore, we suppose demand for the function of communication in a mobile phone among Indian customers is high and they feel the necessity to use mobile phones as a tool to interact with their family, friends, colleagues and acquaintances. In addition, having 2 SIMs allows customers to separate between their personal and home contact. In India, people also prefer to have one cheap SIM and one expensive SIM, because usually the cheap SIM has poor connectivity, which forces users to switch to the expensive one in urgent situations.

Conclusion: Widespread heterogeneity

Operating system



Graph 13: Bar chart of operating system

More than 95% of mobile phones in the Indian market run with an Android operating system, while other types of operating system (including Window, iOS, Cyanogen, BlackBerry, Tizen and Sailfish) make up a very small share in the market. The fact that Android is popular in the mobile market of India can be explained by the fact that in general, Android and iOS are the biggest runners; however, while the Android system is installed within various types of phones with various ranges of price from low to high, iOS is exclusively belong to phones produced by Apple and they usually have a rather high price.

Conclusion: Widespread heterogeneity.

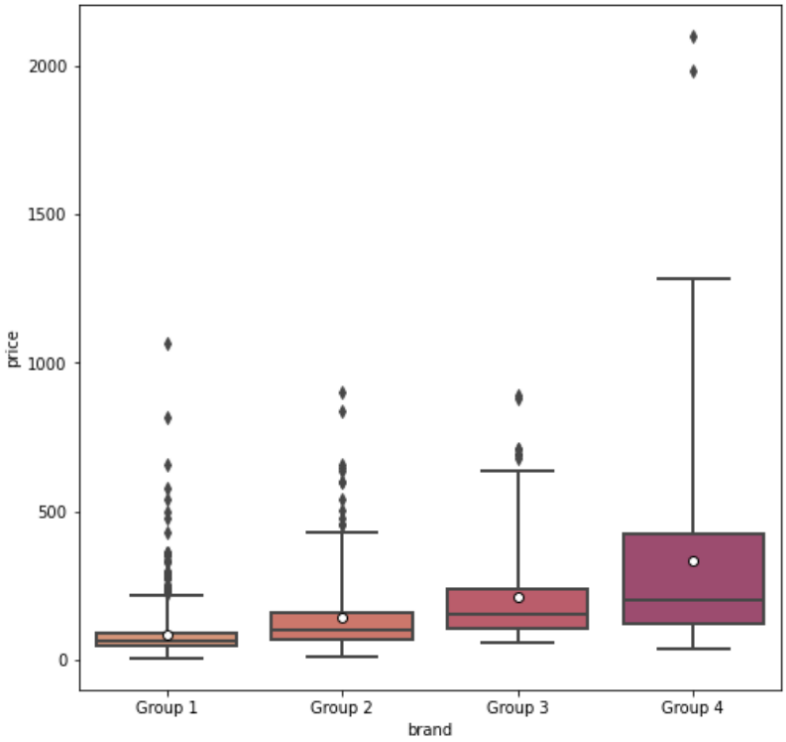
2.2. Analysis of statistical relationships

2.2.1. Graphical analysis of "numeric dependent variable – qualitative independent variable" pairs

Price – Brand

H_0: Price does not depend on brand, i.e. medians are equal

H_1: Price does depend on brand



Graph 14: Box plot of the price – brand pair

From the graph, it can be observed that all of the subsamples are right-skewed. The whiskers for group 4 stretch further than those of the other groups and its box is also the biggest, which implies that this group has the widest range of prices and its price values are most scattered. In addition, all groups of brands have outliers, but most outliers are from group 1, mainly because this group has 55 brands. However, the highest price of a phone belongs to group 4 and in general, its outliers are of higher prices than those of other groups. At the same time, there exist low-valued brands that produce phones with prices above the average price set by high-valued brands. This can be explained by the fact that some brands focus their resources on producing a certain amount of phones of the best quality they can, so sometimes these brands can charge their product at high prices, whereas top brands usually have more

resource to produce a big number of good-quality products, so the prices of these brands do not differ too much.

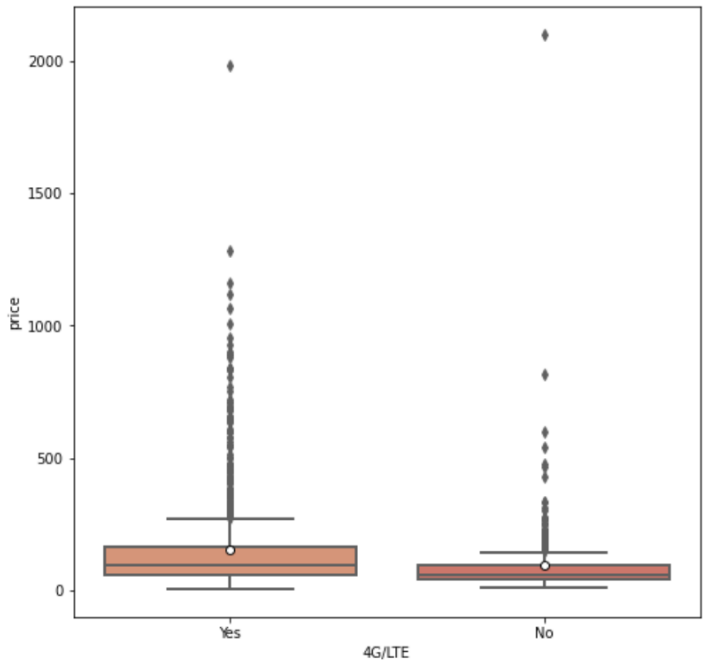
Value of Kruskal-Wallis test	426.389
p-value	0

Based on the received result, we conclude that hypothesis H_0 is rejected. There are significant differences in statistics between the price of a telephone depending on the type of the brand value

Price – 4G/LTE cellular network

H_0: Price does not depend on the presence of the 4G/LTE cellular network, i.e. medians are equal

H_1: Price does depend on the presence of the 4G/LTE cellular network.



Graph 15: Box plot of the price – 4G/LTE pair

From the graph, it can be observed that both subsamples are right-skewed, and the median price for phones with ability to connect to the Internet via 4G/LTE is a little bit higher. On average, price for mobile phones with the 4G/LTE network is higher and this difference accounts for the fact that customers are willing to pay for the convenience of having access to the Internet in the fastest way possible. At the same time, phones that allow the use of 4G/LTE network have a wider range of prices, their price values are more scattered and they have more outliers (which is probably because of the imbalance of the dataset).

Value of Kruskal-Wallis test 99.702

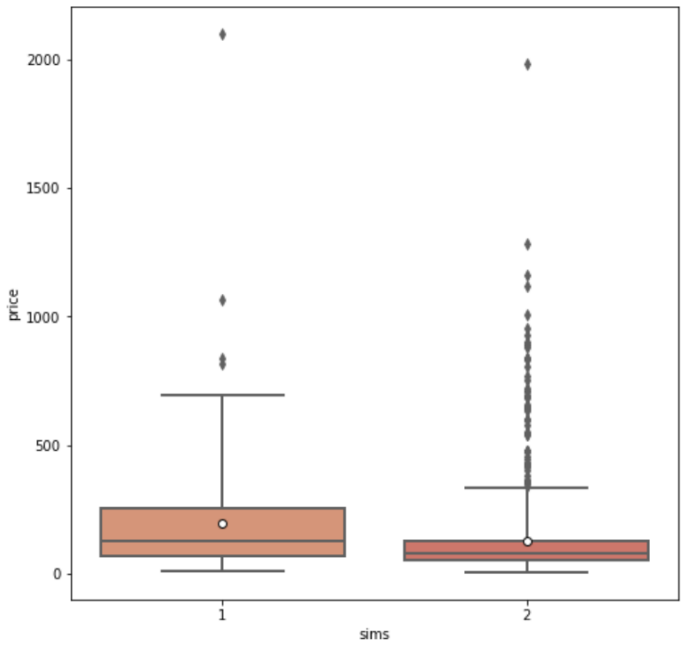
p-value 0

Based on the received result, we conclude that hypothesis H_0 is rejected. There are significant differences in statistics between the price of a telephone depending on the presence of the 4G/LTE network

Price – Number of sim cards

H_0: Price does not depend on the number of sim cards, i.e. medians are equal

H_1: Price does depend on number of sim cards



Graph 16: Box plot of the price – sim pair

From the graph, it can be observed that both subsamples are right-skewed, and the median price for phones 1 SIM is higher. On average, prices for mobile phones with 1 SIM are higher and more scattered, but there are more outliers for phones with 2 SIMs and prices of these outliers are generally higher than those of phones with 1 SIM. Another reason to account for this is that more than 80% of phones in the dataset have 2 sim cards.

Value of Kruskal-Wallis test 58.995

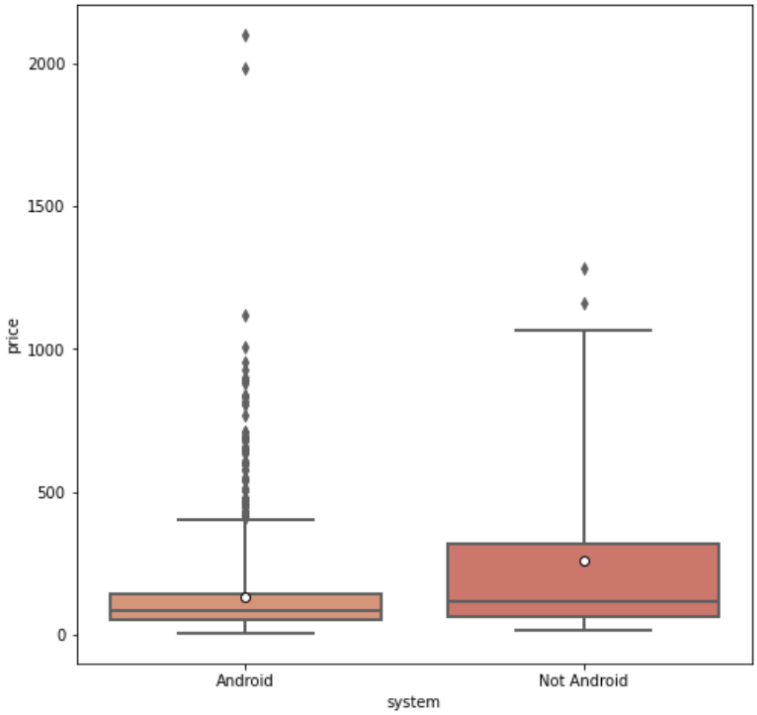
p-value 0

Based on the received result, we conclude that hypothesis H_0 is rejected. There are significant differences in statistics between the price of a telephone depending on the number of sim cards.

Price – Operating system

H_0: Price does not depend on operating system, i.e. medians are equal

H_1: Price does depend on operating system



Graph 17: Box plot of the price – system pair

From the graph, it can be observed that subsamples are right-skewed, and the median price for phones not using the Android operating system is a little bit higher. On average, price for mobile phones without Android is higher, their price range is also wider and their values are more scattered *(non-Android box is bigger, non-Android whiskers stretch further)*. As the majority of mobile phones in our study operate with the Android system, it is understandable that there are many outliers for such phones, which means that there exist many of such phones that have higher prices than on average. At the same time, the maximum price belongs to the one with an Android operating system, and it is significantly higher than the highest price of phones using non-Android operating systems

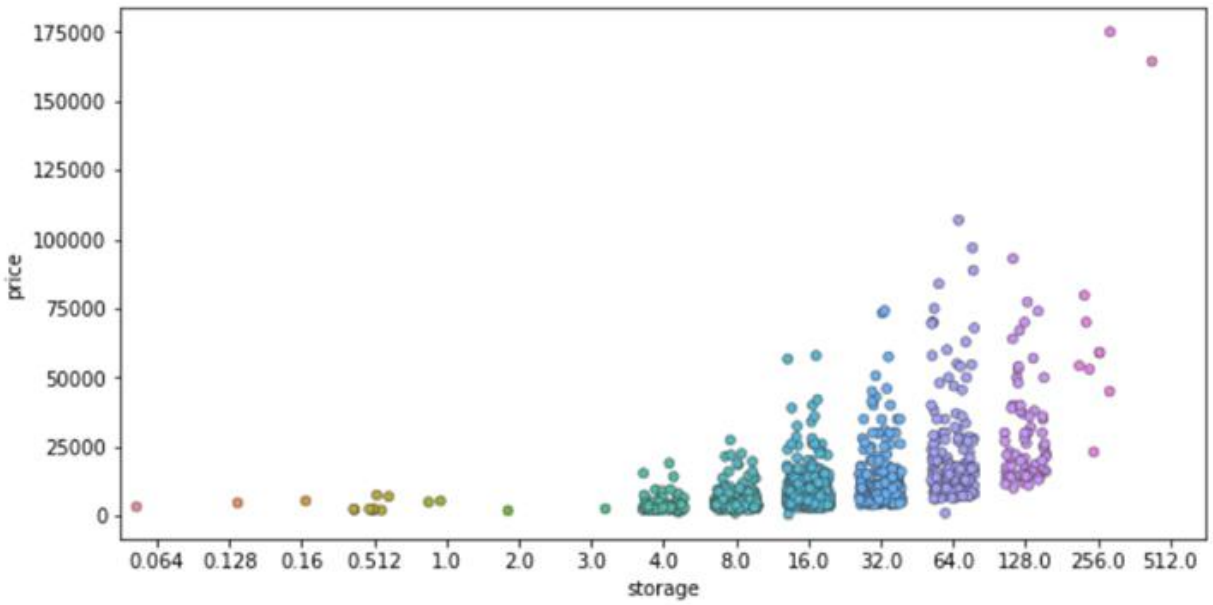
Value of Kruskal-Wallis test	13.716
p-value	0

Based on the received result, we conclude that hypothesis H_0 is rejected. There are significant differences in statistics between the price of a telephone depending on the operating system.

2.2.2. Graphical analysis of "numeric dependent variable – quantitative independent variable" pairs

Price – Internal storage

Battery capacity is a continuous variable that ranges from 1000 mAh to 6000 mAh



Graph 18: Strip plot of the price – storage pair

As observed from the graph, we conclude that there is a positive dependence of a mobile phone's price on how strong the battery capacity is, as price tends to increase with the increase of battery capacity. At the same time, phones that cost more than $500 generally have battery capacity above average, which equals 3000 mAh. These types of phones are not common in the Indian market, which can be explained by the fact that the majority of Indian population are not very wealthy, so the demand for expensive phones is not very high. We also mentioned the three outliers of 6000 mAh, which we can see that their prices are less than $500 and we suppose that strong battery capacity does not always allow suppliers to charge their mobile phones at a high price.

In addition, the graph also illustrates an area of density where battery capacity ranges from 2000 to 4000 mAh and the price for each mobile phone ranges from $6 to $2099.9, which are very wide.

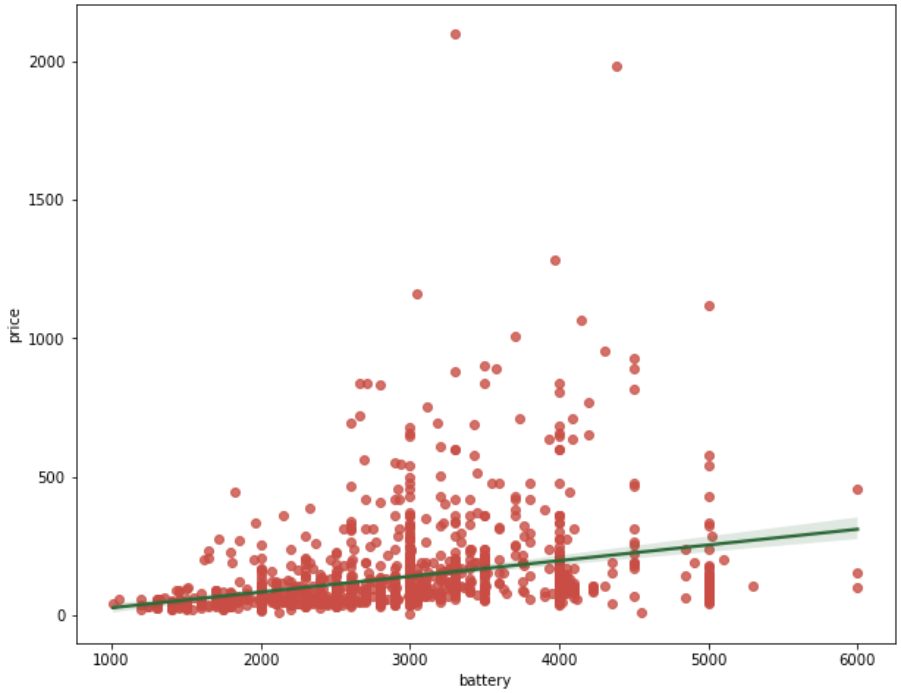
	Pearson	Spearman	Tau Kendall
Correlation coefficient	0.644	0.698	0.559
p-value	0.000	0.000	0.000

Table 8: Correlation coefficients for the price – storage pair

The Pearson, Spearman and Tau Kendall coefficients are significant on the level of 0.05 and they indicate a rather weak to average positive correlation between price and battery capacity.

Price – Battery capacity

Internal storage is a discrete variable that ranges from 0.064 GB to 512 GB



Graph 19: Scatter plot for the pair of price – battery

As observed from the graph, we conclude that there is a positive dependence of a mobile phone's price on the amount of internal storage, as it is obvious that the larger the internal storage is, the higher the price increases. As it stands, the price for a phone with internal storage of 512GB is the second highest of all. However, taking into consideration that the highest price is for a phone with 256GB storage, it can be said that there is another factor / there are other factors that also affect the price level.

In addition, the graph also illustrates an area of density where the amount of internal storage ranges from 4 to 128 GB and the price for each mobile phone ranges from $15 to $1116.

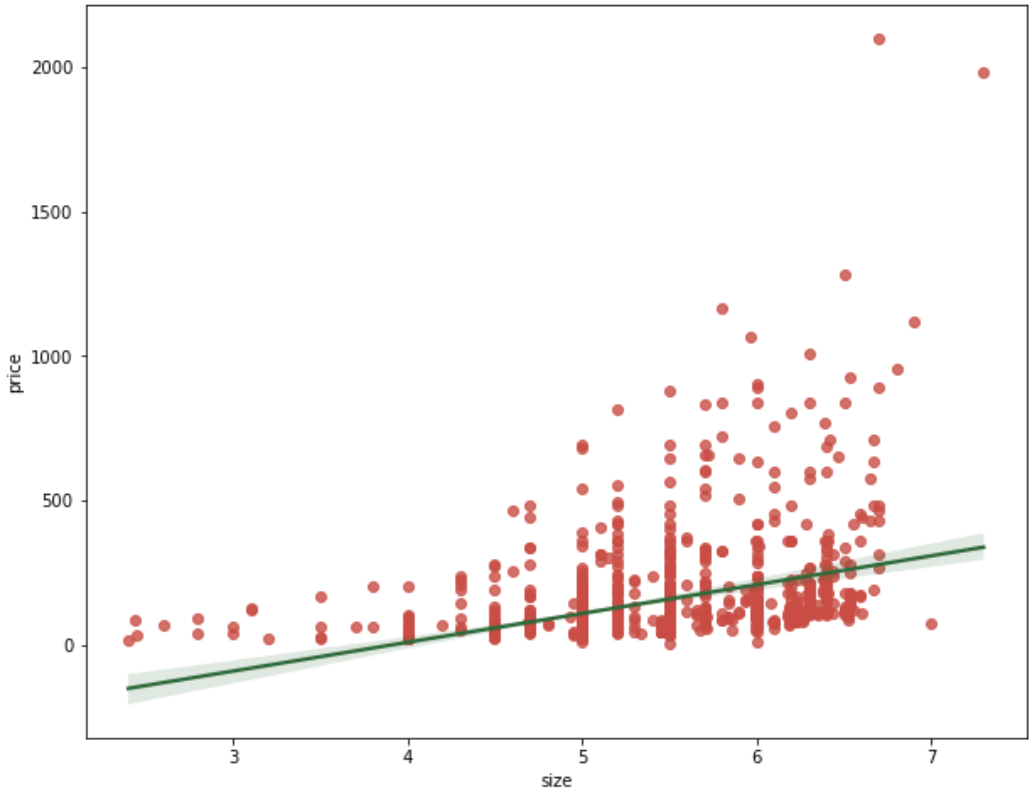
	Pearson	Spearman	Tau Kendall
Correlation coefficient	0.298	0.519	0.368
p-value	0.000	0.000	0.000

Table 9: Correlation coefficients for the price – battery pair

The Pearson, Spearman and Tau Kendall coefficients are significant on the level of 0.05 and they all indicate a moderate to fairly strong positive correlation between price and internal storage.

Price – Screen size

Screen is a continuous variable that ranges from 2.4 to 7.3 inches



Graph 20: Scatter plot for the price – screen size pair

As observed from the graph, we conclude that there is a positive dependence of a mobile phone's price on the size of screen, as it is obvious that the larger the screen size is, the higher the price increases.

In addition, the graph also illustrates an area of density where the screen size ranges from 5 to roughly 6.6 inches and the price for each mobile phone ranges from $6 to $1283. This shows that the

majority of mobile phones in the Indian market have such screen sizes and are available at very low to rather high prices, which allow customers various options.

	Pearson	**Spearman**	**Tau Kendall**
Correlation coefficient	0.402	0.592	0.445
p-value	0.000	0.000	0.000

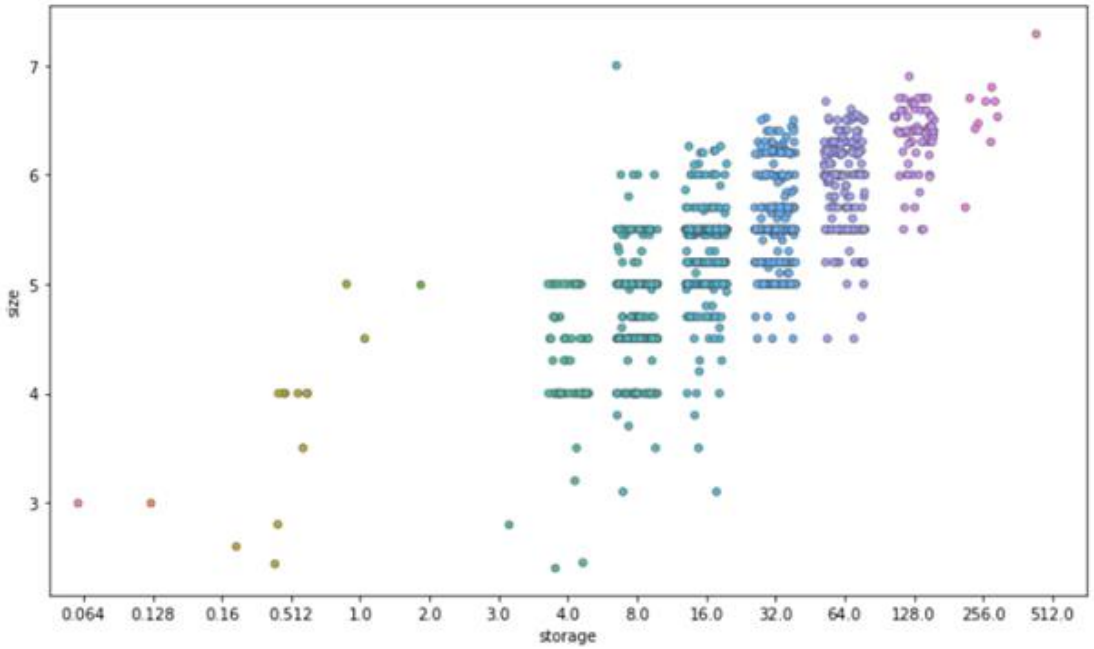
Table 10: Correlation coefficients for the price – screen size pair

The Pearson, Spearman and Tau Kendall coefficients are significant on the level of 0.05 and they all indicate an average positive correlation between price and battery capacity.

2.2.3. Analysis of the presence of correlation between independent variables

Quantitative variables – Quantitative variables

Internal storage – Screen size



Graph 21: Strip plot[2] of the storage – screen pair

	Pearson	Spearman	Tau Kendall
Correlation coefficient	0.613	0.768	0.660
p-value	0.000	0.000	0.000

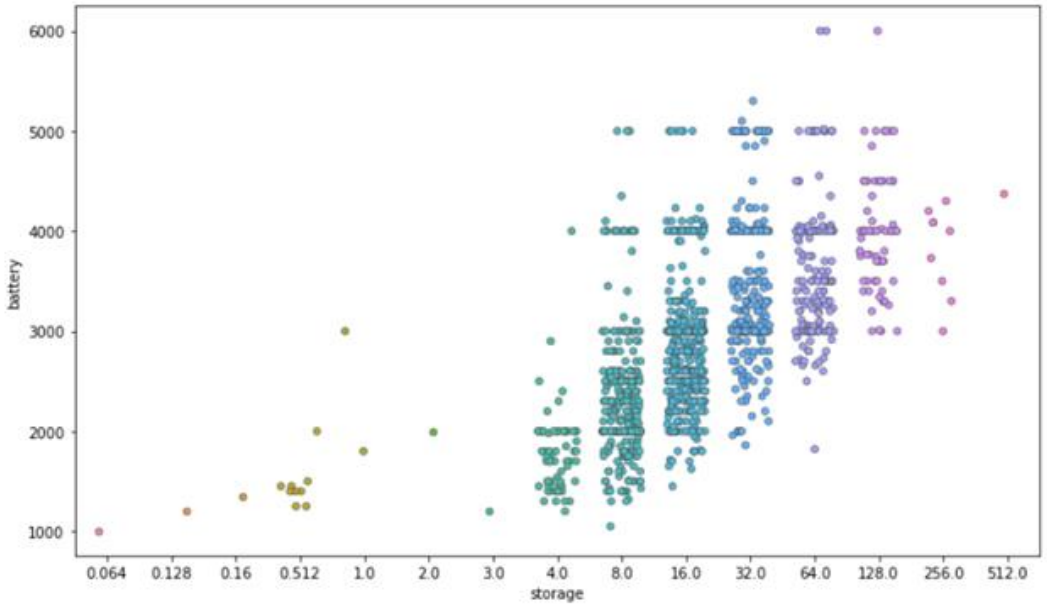
Table 11: Correlation coefficients for the pair storage – screen size

All three correlation coefficients are significant at the alpha level $\alpha = 0.05$, which means that there is a statistical relationship between these two variables. Based on the value of the correlation coefficients, we draw a conclusion that there is a moderate to fairly strong positive correlation between

[2] *Strip plot is a special type of scatter plot using jitter to visualize more effectively data especially when values overlap*.

internal storage and screen size. The area of density indeed indicates that as screen size increases, so does the amount of internal storage.

Internal storage – Battery capacity



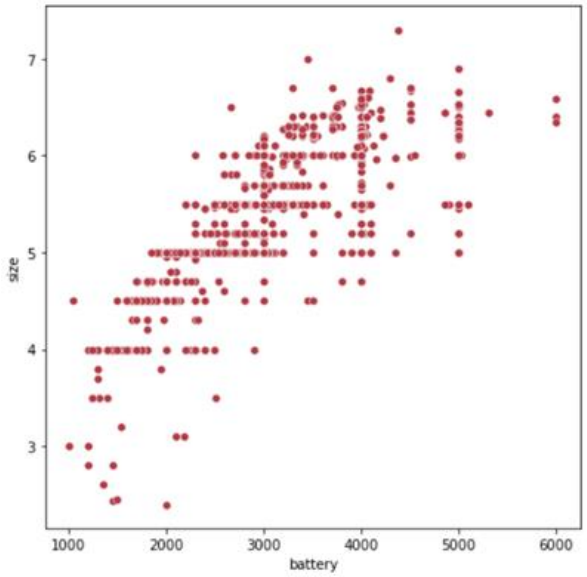
Graph 22: Strip plot of the storage – battery pair.

	Pearson	Spearman	Tau Kendall
Correlation coefficient	0.486	0.703	0.572
p-value	0.000	0.000	0.000

Table 12: Correlation coefficients for the pair storage – battery

All three correlation coefficients are significant at the alpha level $\alpha = 0.05$, which means that there is a statistical relationship between these two variables. Based on the value of the correlation coefficients, we draw a conclusion that there is a moderate positive correlation between internal storage and battery capacity. On the graph, the area of density also indicates that phones with larger internal storage tend to have strong battery capacity. This can be explained by the fact that using a telephone with high intensity tends to take up more space of internal storage and consume more battery at the same time.

Screen size – Battery capacity



Graph 23: Scatter plot for the pair of screen size - battery

	Pearson	Spearman	Tau Kendall
Correlation coefficient	0.747	0.793	0.660
p-value	0.000	0.000	0.000

Table 13: Correlation coefficients for the pair screen size – battery

All three correlation coefficients are significant at the alpha level $\alpha = 0.05$, which means that there is a statistical relationship between these two variables. Based on the value of the correlation coefficients, we draw a conclusion that there is a moderate to strong positive correlation between battery capacity and screen size. The graph also depicts that the area of density indeed indicates that as screen size tends to increase on the rise of battery capacity. In reality, this can be easily explained by the fact that the bigger a telephone is, the more energy it requires to perform the necessary functions, hence, the more battery it needs.

Qualitative variables - Qualitative variables

Brand – 4G/LTE cellular network

4G/LTE Brands	Yes	No	Total
Group 1	521	239	760
Group 2	243	70	313
Group 3	126	19	145
Group 4	121	19	140

Table 14: Cross table for the brand – 4G/LTE pair

Degrees of freedom	Critical value	Chi-square value (Test statistic)	p-value	Significance level (alpha)	Cramer's V
3	7.815	37.933	0	0.05	0.167

Table 15: Table of chi-squared test result for the brand – 4G/LTE pair

Result: critical value < chi-square value, p-value < alpha

From the result obtained, it can be concluded that the null hypothesis, stating that there is no relationship between the value of brand and the presence of 4G/LTE cellular network, is rejected. In addition, with 3 degrees of freedom and Cramer's V coefficient of 0.167, the strength of association between the type of brand and the 4G/LTE network is considered to be of a medium level.

Brand – Number of sim cards

Number of SIMs Brand	1	2	Total
Group 1	80	680	760
Group 2	68	245	313
Group 3	28	117	145
Group 4	51	89	140

Table 16: Cross table for the brand – sims pair

Degrees of freedom	Critical value	Chi-square value (Test statistic)	p-value	Significance level (alpha)	Cramer's V
3	7.815	66.335	0	0.05	0.221

Table 17: Table of chi-squared test result for the brand – sim pair

Result: critical value < chi-square value, p-value < alpha

From the result obtained, it can be concluded that the null hypothesis, stating that there is no relationship between the value of brand and the number of sim cards, is rejected. In addition, with 3 degrees of freedom and Cramer's V coefficient of 0.221, the strength of association between the type of brand and the 4G/LTE network is considered to be above medium, but not large.

Brand – Operating system

System Brand	Android	Not Android	Total
Group 1	736	24	760
Group 2	298	15	313
Group 3	144	1	145
Group 4	120	20	140

Table 18: Cross table for the brand – operating system pair

Degrees of freedom	Critical value	Chi-square value (Test statistic)	p-value	Significance level (alpha)	Cramer's V
3	7.815	40.014	0	0.05	0.172

Table 17: Table of chi-squared test result for the brand – operating system pair

Result: critical value < chi-square value, p-value < alpha

From the result obtained, it can be concluded that the null hypothesis, stating that there is no relationship between the value of brand and the type of operating system, is rejected. In addition, with 3 degrees of freedom and Cramer's V coefficient of 0.172, the strength of association between type of brand and operating system is considered to be of a medium level.

4G/LTE cellular network – Number of sim cards

Number of SIMs Network 4G/LTE	1	2	Total
Yes	143	868	1011
No	84	263	347

Table 19: Cross table for the 4G/LTE – sims pair

Degrees of freedom	Critical value	Chi-square value (Test statistic)	p-value	Significance level (alpha)	Cramer's V
1	3.841	18.075	0	0.05	0.118

Table 20: Table of chi-squared test result for the 4G/LTE – sims pair

Result: critical value < chi-square value, p-value < alpha

From the result obtained, it can be concluded that the null hypothesis, stating that there is no relationship between the presence of the presence 4G/LTE network and the number of sim cards, is rejected. In addition, with 1 degree of freedom and Cramer's V coefficient of 0.118, the strength of association between the 4G/LTE network and the number of SIMs is considered to be of a low level.

4G/LTE cellular network – Operating system

Operating system 4G/LTE	Android	Not Android	Total
Yes	967	44	1011
No	331	16	347

Table 21: Cross table for the 4G/LTE – operating system pair

Degrees of freedom	Critical value	Chi-square value (Test statistic)	p-value	Significance level (alpha)	Cramer's V
1	3.841	0.003	0.959	0.05	0.005

Table 22: Table of chi-squared test result for the 4G/LTE –operating system pair

Result: critical value > chi-square value, p-value > alpha

From the result obtained, it can be concluded that the null hypothesis, stating that there is no relationship between the presence of 4G/LTE network and type of the operating system , is not rejected. In addition, a degree of freedom of 1 and a Cramer's V coefficient of 0.005 also support the fact the statistical relationship between these two variables does not exist.

Number of sim cards – Operating system

Operating system Number of SIMs	Android	Not Android	Total
1	188	39	227
2	1110	21	1131

Table 23: Cross table for the sims – operating system pair

Degrees of freedom	Critical value	Chi-square value (Test statistic)	p-value	Significance level (alpha)	Cramer's V
1	3.841	101.526	0	0.05	0.278

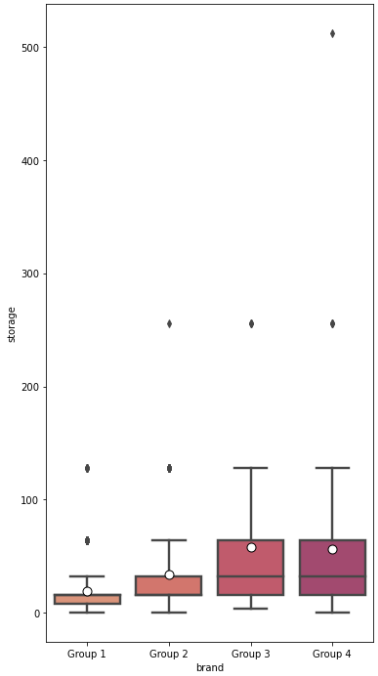
Table 24: Table of chi-squared test result for the sim – operating system pair

Result: critical value < chi-square value, p-value < alpha

From the result obtained, it can be concluded that the null hypothesis, stating that there is no relationship between the number of SIMs and type of the operating system, is rejected. In addition, with 1 degree of freedom and a Cramer's V coefficient of 0.278, the strength of association between the 4G/LTE network and the number of SIMs is considered to be of a medium level.

Quantitative variables – Qualitative variables

Internal storage – Brand



Graph 24: Box plot of the storage – brand pair

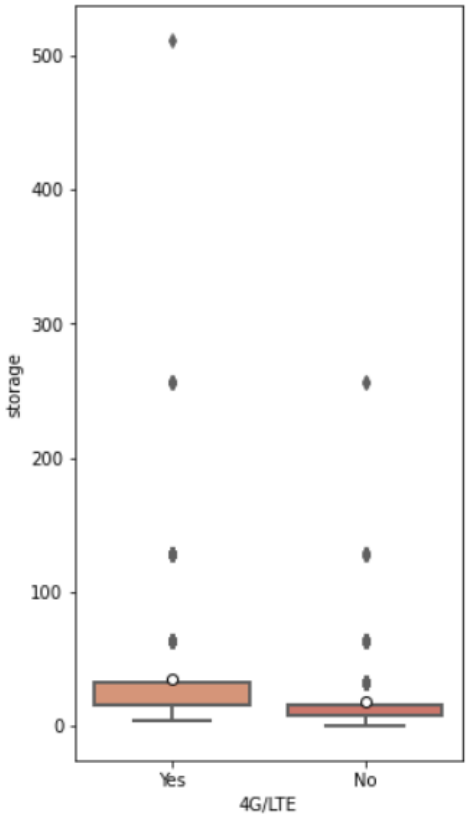
Null Hypothesis	p-value	Test	Statistical value	Result
No statistical relationship between variables	0	Kruskal-Wallis	1357	H_0 is rejected

Table 25: Kruskal-Wallis Test for the pair of storage – brand

From the test's statistics and the graph, we conclude that there exists a statistical relationship between internal storage and brand, as the amount of internal storage varies evidently among brands of high and low values.

Although the number of phones produced by more valuable brands (group 3 and 4) is less than 30% in total, their range of internal storage is wider than the ranges of group 1 and group 2. Top 10 brands fall into group 3 and 4, so it is comprehensible that their medians and averages of internal storage are almost equal. On average, phones produced by top brands have greater internal storage, whereas smaller brands tend to produce phones with less storage.

Internal storage – 4G/LTE cellular network



Graph 25: Box plot of the storage – 4G/LTE pair

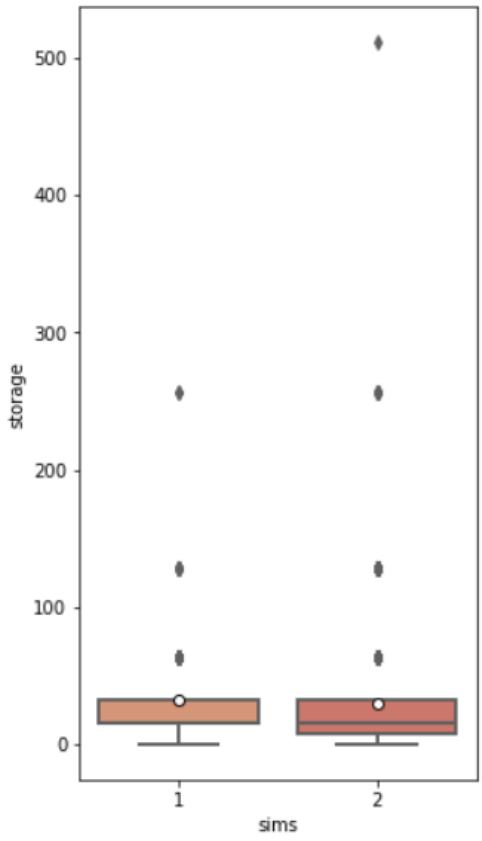
Null Hypothesis	p-value	Test	Statistical value	Result
There is a statistical relationship between variables	0	Kruskal-Wallis	208.516	H0 is rejected

Table 26: Kruskal-Wallis Test for the pair of internal storage – 4G/LTE

From the test's statistics and the graph, we conclude that there exists a statistical relationship between internal storage and the existence of 4G/LTE cellular network, as the ranges of internal storage for phones with and without 4G/LTE network are not equal.

On average, phones provided with cellular network have more internal storage than those without. This can be explained by the fact that when users access the Internet with 4G/LTE network too often, this requires the phone to use more data to process activities; therefore, more internal storage is needed.

Internal storage – Number of sim cards



Graph 26: Box plot of the storage – sims pair

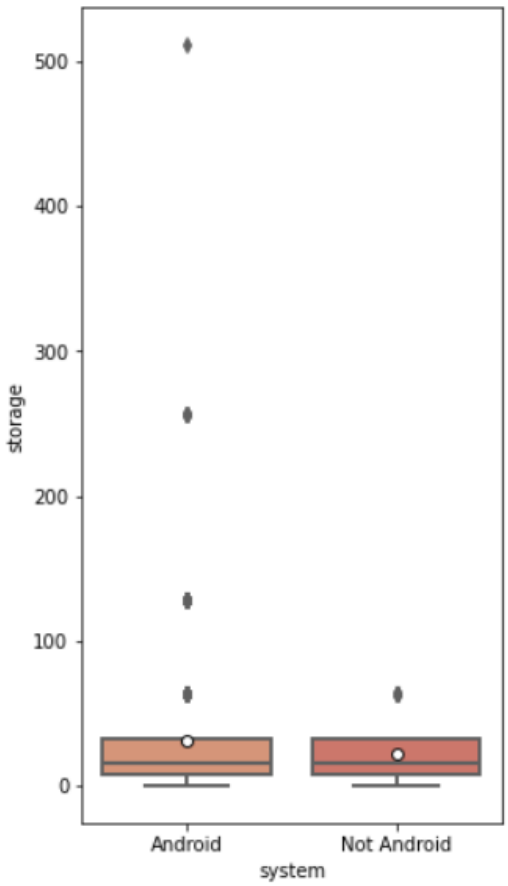
Null Hypothesis	p-value	Test	Statistical value	Result
There is a statistical relationship between variables	0	Kruskal-Wallis	564.81	H0 is rejected

Table 27: Kruskal-Wallis Test for the pair of internal storage – sims

From the test's statistics and the graph, we conclude that there exists a statistical relationship between internal storage and the number of sim cards.

Since in our study the number of phones with 2 SIMs already takes up more than 80% in total, it is understandable that the range of internal storage for phones with 2 SIMs is wider and the values are more dispersed, which indicates that internal storage for phones with 2 SIMs tends to be a bit larger. In addition, the average amounts of internal storage for phones with 1 SIM and 2 SIMs are almost equal.

Internal storage – Operating system



Graph 27: Box plot of the storage – system pair

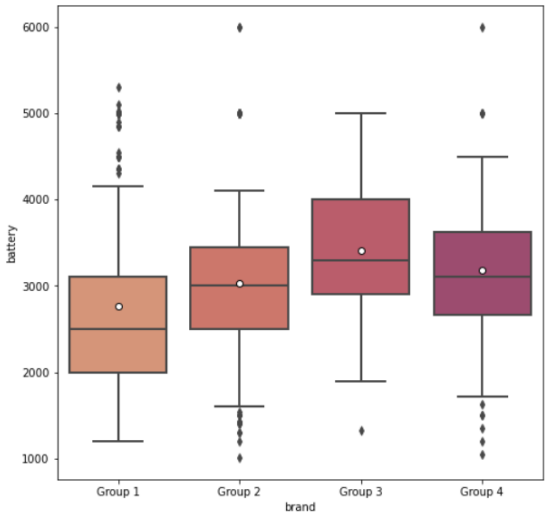
Null Hypothesis	p-value	Test	Statistical value	Result
No statistical relationship between variables	0.050038	Kruskal-Wallis	3.84	H0 is not rejected

Table 28: Kruskal-Wallis Test for the pair of internal storage – operating system

From the test's statistics and the graph, we conclude that there is no statistical relationship between internal storage and the type of operating system.

Except for several outliers for phones with Android operating system, the amounts of internal storage for phones operated by Android and non-Android systems have the same median, the same range of values and their values are similarly dispersed.

Battery capacity – Brand



Graph 28: Box plot of the battery – brand pair

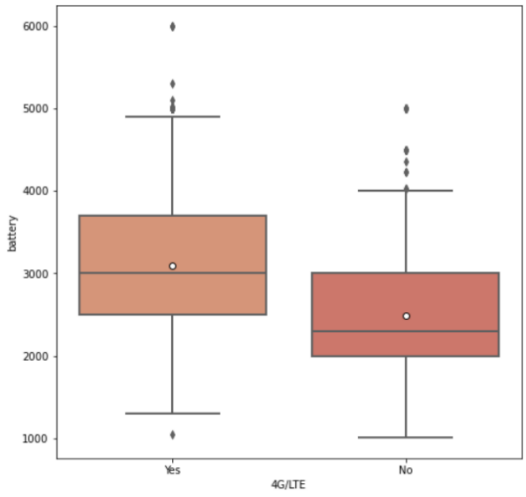
Null Hypothesis	p-value	Test	Statistical value	Result
There is a statistical relationship between variables	0	Kruskal-Wallis	110.366	H0 is rejected

Table 29: Kruskal-Wallis Test for the pair of battery – brand

From the test's statistics and the graph, we conclude that there exists a statistical relationship between battery capacity and brand, as the ranges of battery capacity vary evidently among brands of high and low values.

On average, battery capacity is the highest among phones from brands of group 3 and group 4. This can be explained by the fact that brands from these groups are the top brands, producing phones with better qualities, including battery capacity. However, it should be noticed that some of low-valued brands also have the ability to produce phones with quite strong battery capacity.

Battery capacity – 4G/LTE cellular network



Graph 29: Box plot of the battery – 4G/LTE pair

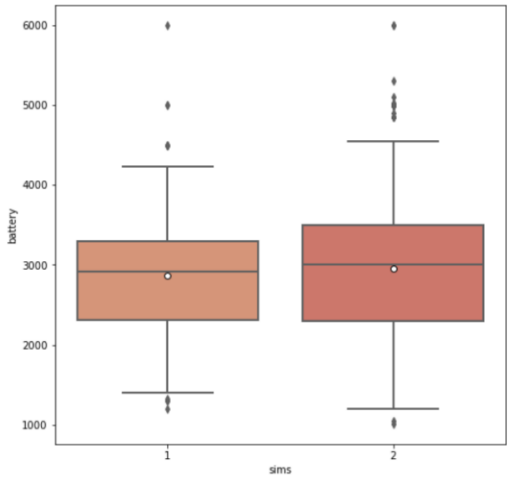
Null Hypothesis	p-value	Test	Statistical value	Result
There is a statistical relationship between variables	0	Kruskal-Wallis	142.244	H0 is rejected

Table 30: Kruskal-Wallis Test for the pair of battery – 4G/LTE

From the test's statistics and the graph, we conclude that there exists a statistical relationship between battery capacity and the existence of 4G/LTE cellular network, as the ranges of battery capacity vary evidently among phones with and without 4G/LTE.

On average, battery capacity is higher among phones with 4G/LTE. This can be explained by demand from users. Phone with 4G/LTE users use phones more frequently, therefore require more battery capacity.

Battery capacity – Number of sim cards



Graph 30: Box plot of the battery – sims pair

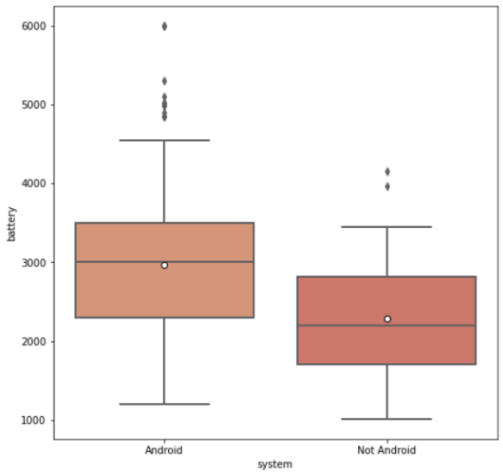
Null Hypothesis	p-value	Test	Statistical value	Result
No statistical relationship between variables	0.354	Kruskal-Wallis	0.857	H0 is not rejected

Table 31: Kruskal-Wallis Test for the pair of battery – sims

From the test's statistics and the graph, we conclude that there's no statistical relationship between battery capacity and number of sim cards.

On average, battery capacity is insignificantly higher among phones with 2 SIMs. This can be explained by the fact that 1 extra SIM is only used to call, receive calls and text, consuming not much battery. Therefore phones with 2 SIMs do not need more battery capacity than phones with 1 SIM.

Battery capacity – Operating system



Graph 31: Box plot of the battery – operating system pair

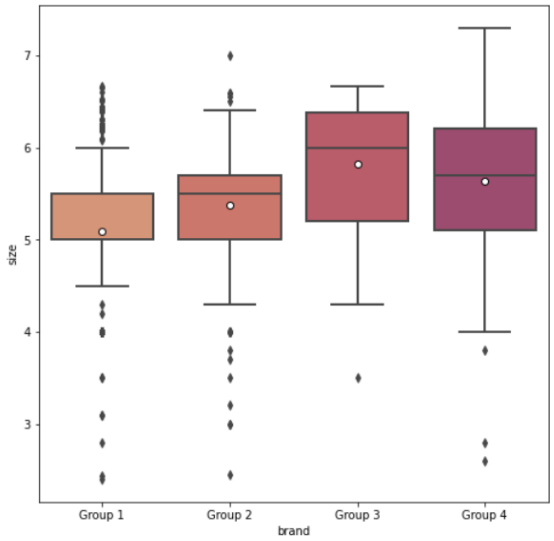
Null Hypothesis	p-value	Test	Statistical value	Result
There is a statistical relationship between variables	0	Kruskal-Wallis	30.015	H0 is rejected

Table 32: Kruskal-Wallis Test for the pair of battery – operating system

From the test's statistics and the graph, we conclude that there exists a statistical relationship between battery capacity and the type of operating system, as the amount of battery capacity varies evidently among phones operated by Android and by other systems.

On average, battery capacity is higher among phones using the Android system. Android is known for having better battery capacity than other operating systems, because its phones manage energy consumption well and use various low-power modes to cut down power consumption.

Screen size – Brand



Graph 32: Box plot of the size – brand pair

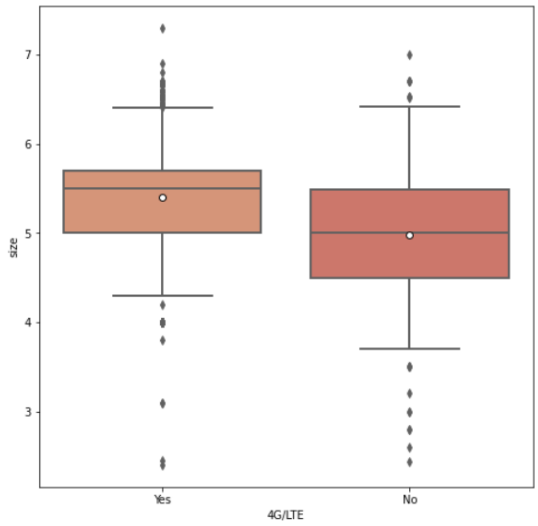
Null Hypothesis	p-value	Test	Statistical value	Result
There is a statistical relationship between variables	0	Kruskal-Wallis	225.31	H0 is rejected

Table 33: Kruskal-Wallis Test for the pair of size – brand

From the test's statistics and the graph, we conclude that there exist a statistical relationship between screen size and brand, as the ranges of screen size vary evidently among brands of high and low values.

Although the number of phones produced by more valuable brands (group 3 and 4) is less than 30% in total, their ranges of screen size are wider than the ranges of group 1 and group 2. On average, screen size is the highest among phones from brands of group 3 and group 4. This can be explained by the fact that brands from these groups are the top brands, producing phones with high quality, in order to provide the best customer experience.

Screen size – 4G/LTE cellular network



Graph 33: Box plot of the size – 4G/LTE pair

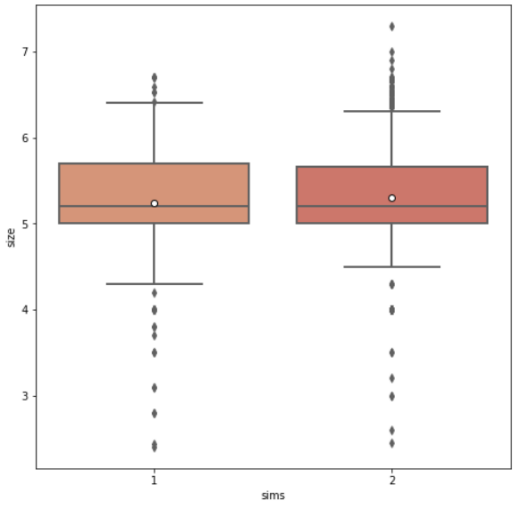
Null Hypothesis	p-value	Test	Statistical value	Result
There is a statistical relationship between variables	0	Kruskal-Wallis	112.57	H0 is rejected

Table 33: Kruskal-Wallis Test for the pair of size – 4G/LTE

From the test's statistics and the graph, we conclude that there exists a statistical relationship between screen size and 4G/LTE cellular network accessibility.

On average, phones provided with cellular networks have higher screen size than those without. This can be explained that users who use phones with 4G/LTE tend to watch videos on the Internet and play videogames, which are the utilities that require higher screen size.

Screen size – Number of sim cards



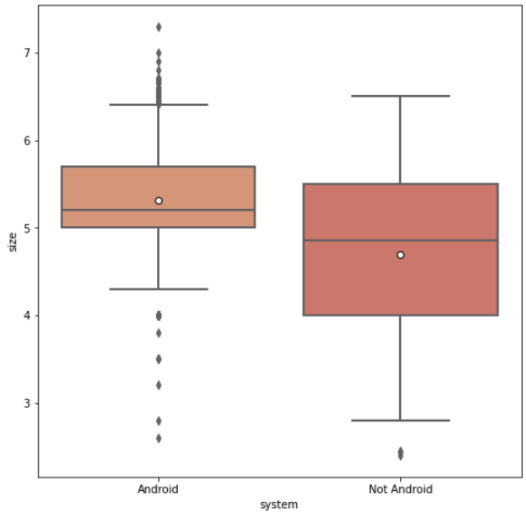
Graph 34: Box plot of the size – sims pair

Null Hypothesis	p-value	Test	Statistical value	Result
No statistical relationship between variables	0.774	Kruskal-Wallis	0.082	H0 is not rejected

Table 34: Kruskal-Wallis Test for the pair of size – sims

From the test's statistics and the graph, we conclude that there is not a relationship between screen size and the number of sim cards. This can be explained by the fact that having more SIMs does not affect the screen size. The main functions of a sim card is to make phone calls and send messages, which do not really require a change in screen size.

Screen size – Operating system



Graph 35: Box plot for the pair of screen size – operating system

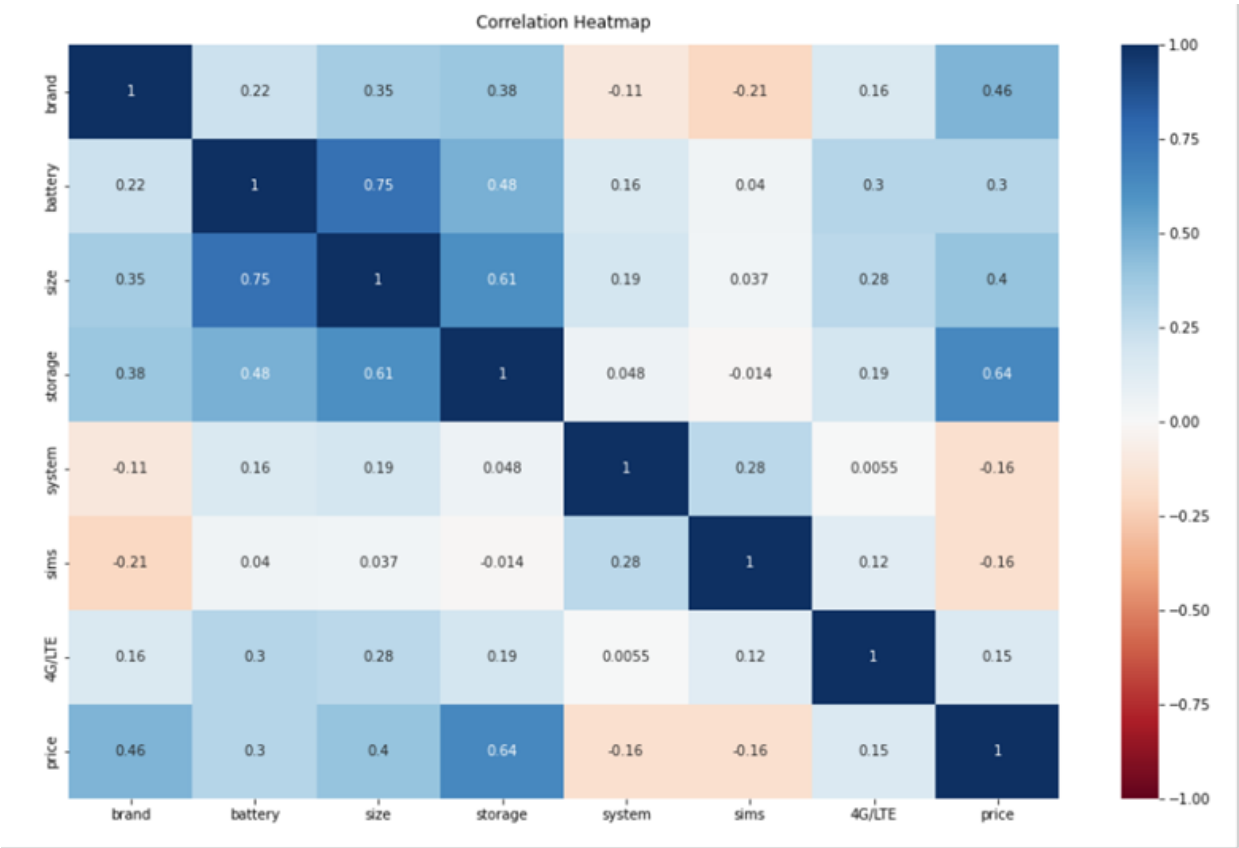
Null Hypothesis	p-value	Test	Statistical value	Result
There is a statistical relationship between variables	0	Kruskal-Wallis	22.532	H0 is rejected

Table 35: Kruskal-Wallis Test for the pair of size – 4G/LTE

From the test's statistics and the graph, we conclude that there exists a statistical relationship between screen size and operating system , as the lengths of screen size are different between phones with and without an Android operating system.

As can be seen from the graph, Android phones have higher screen size on average. This can be explained by the fact that Android phones are produced to have bigger screens, which make them really stand out from its giant competitor - iOS phone.

`



Graph 36: Pearson correlation for all variables

In general, with regard to the dependent variable, variables that are associated with price the most are internal storage, brand, size, then battery at a lower level. Taking into account the correlations among independent variables, the strongest correlated pairs are battery – size and size – storage with coefficients of 0.75 and 0.61 accordingly. As explained above, phones functioning at an intensive basis will require more space (storage) for loading data, more battery and hence, a bigger screen size to operate properly. The levels of dependency among independent variables also indicate that there is likely the presence of multicollinearity in the model.

2.2.4. Preliminary testing of hypotheses

01 The price of a telephone positively depends on its internal storage **confirmed**

Our preliminary analysis approves this hypothesis on the basis of the received scatterplot between price and internal storage as well as their correlation coefficients.

02 There exists an exponential dependence of a telephone's price on its battery capacity (i.e., a telephone's battery capacity has a positive effect on its price, and after a certain value, the growth of price increases) **rejected**

Our preliminary analysis does not approve this hypothesis, as the received scatterplot between price and battery capacity does not show a clear positive relationship between them after a certain value of battery capacity. In addition, their correlation coefficients are quite weak. Although the coefficient of Spearman signifies a moderate correlation, it should be noted that Spearman correlation works best in evaluating relationships involving ordinal variables, whereas both of these variables are continuous.

03 There exists an exponential dependence of a telephone's price on its screen size (i.e., a telephone's screen size has a positive effect on its price, and after a certain value, the growth of price increases) **confirmed**

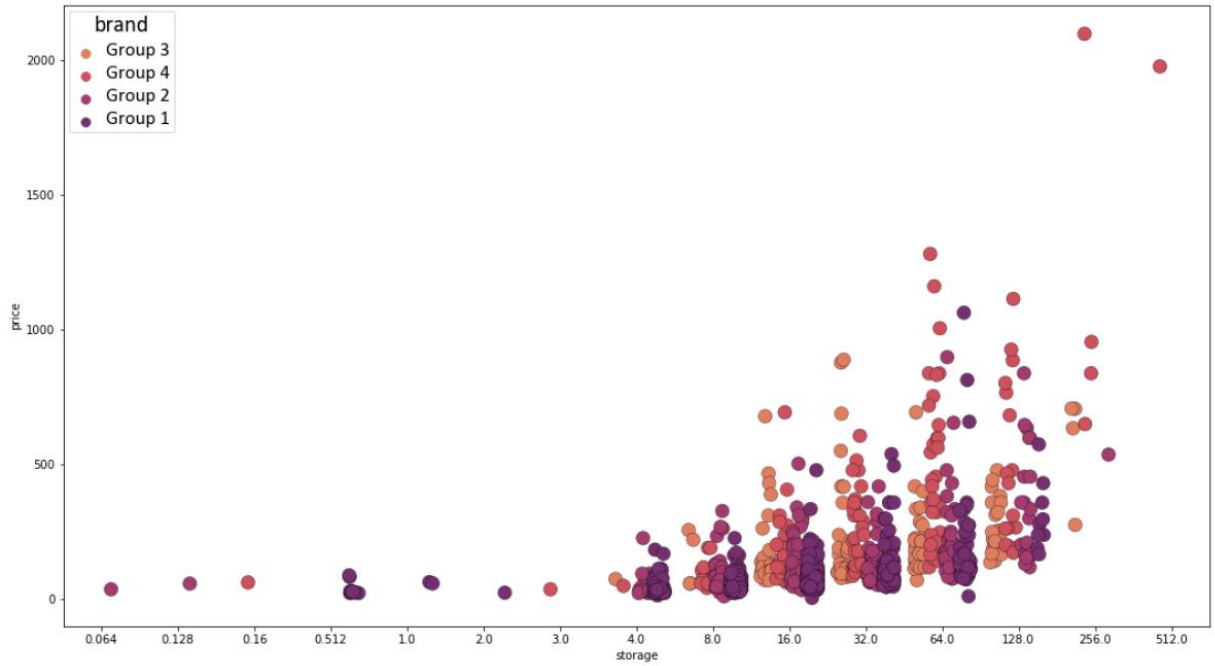
Our preliminary analysis approves this hypothesis, as the received scatterplot between price and screen size indicates that there is an increase in the growth of price after a certain value of screen size. In addition, their correlation coefficients are from moderate to fairly strong.

04 The brand value of a telephone has different effects on how its internal storage determines its price (i.e., it is not always the case that on the rise of the internal storage would the price increase, as this also depends on the brand value) **confirmed**

The first clause of this hypothesis is approved by our preliminary analysis, as the received strip plot between price and internal storage indicates that price also increases on the rise of internal storage. In addition, their correlation coefficients are above average.

The second clause of this hypothesis is examined with the strip plot down below. It can be observed that, although there are representatives of the low-valued brands from group 1 that have the same amount of internal storage as high-valued brands, their prices are relatively lower than those of phones from top brands.

Therefore, we do not reject this hypothesis.



Graph 37: Strip plot of the price - storage pair sorted by brand

3. Specification, evaluation and optimization of the model

In this section, we once again take into account the applied task: Study the behaviors of a telephone's price on the Indian market under the following conditions: battery capacity of at least 2600 mAh (i.e. the phone can be used for at least 2.5 hours), size ranging from 5 to 7 inches, and internal storge from 32GB, system operated by Android. In addition, phones from all brands, regardless of their reputation or experience in producing mobile phones, are included in the research. However, we would look distinctively at phones produced by companies with the lowest brand value (brand 1).

From the above preliminary analysis, we see that p-values of all variables with regards to their relationship with the target variable (price) are all lower than 0.05, indicating that they are significant. At the same time, correlations coefficients are fairly weak to moderate between independent qualitative variables and moderate – fairly strong between independent numeric variables.

However, it is also important to note that the correlations between *battery – size* and *size – storage* pairs (with coefficients of 0.75 and 0.61 respectively) signify the possible presence of multicollinearity.

The sample was divided into by ratio 90:10 train sets and test sets for independent variables and dependent variable. A large part of the sample (training) will be used to build the model, a smaller part (test) will be used to test the predictive properties

3.1. Evaluation of base model and results of hypothesis testing

According to part 2.2.4, our preliminary showed that hypotheses 1 3 4 could be confirmed, and hypothesis 2 could be rejected.

Base model

Observation:

On our first attempt of building the base regression model, we detected some signs of data-based multicollinearity based on very high VIFs of some independent variables, namely screen size, number of sims, battery capacity, operating system and brand value. Optimally, VIFs should be close to 1 as much as possible, as high VIFs signify multicollinearity that can spoil our model.

Variable	VIF
brand	5.371439
battery	25.279718
size	60.979058
storage	3.245435
system	6.786363
sims	25.383305
network	24.462913

Explanation:

According to our preliminary analysis in part 2.2.3, the strongest correlation between independent variables belong to the *battery – size* (0.75). Because battery and size not only have high correlation with each other, but also with other variables such as *storage*, it is not unpredicted that the VIFs of *battery* and *size* are quite high. As for qualitative independent variables, *sims* has the highest VIF – 25.383, while the VIFs of *brand* and *system* are smaller but above 5.

Amendments:

After a few attempts at correcting the problem of multicollinearity, we decided to:

- remove variables *sims* and *network* because their VIFs are too high (and almost impossible to reduce to desirable values after many attempts)
- use the centering variables method to reduce the VIFs of *storage*, *size*, *battery*, *price*

For qualitative independent variables, removing them from the base model does not affect the model's score by much. However, for quantitative independent variables, it is impossible to simply remove them out of the model because the hypotheses revolve around them. Instead, we can reduce this multicollinearity by centering the variables, a simple method of standardization. This method is used because by standardizing the variables only by subtracting the mean, the centered variables are easier to interpret than by using Standard Scaler / Min – Max Scaler. In addition, we did not choose to use the logarithmic transformation, because this still resulted in very high multicollinearity.

Result:

Variable	Coefficients estimators	Standard error	t-test	p-value	95% confidence interval (lower bound)	95% confidence interval (upper bound)	VIF
const	74.891	18.990	3.944	0.000	37.633	142.149	
battery_mean	0.007	0.006	-2.124	0.034	-0.018	0.005	2.271
size_mean	5.297	8.723	2.307	0.021	-10.104	22.410	2.819
storage_mean	2.715	0.118	23.031	0.000	-11.817	2.947	2.304
system	-36.763	17.159	-7.971	0.000	-170.919	-103.099	3.986
brand	30.751	3.591	8.563	0.000	23.706	37.797	4.218

Table 36: Results of regression for base model

So now, our base model will include 5 independent variables instead of 7, and all VIFs are less than 5, which indicates slight but not significant multicollinearity. In addition, all regressors that are significant at the alpha level of 0.05 entered the model.

As such, the base model takes the following form:

price_mean = 74.891 + 0.007 * *battery_mean* + 5.297 * *size_mean*
+ 2.715 * *storage_mean* – 36.763 * system + 30.751 * brand

Test of significance: F-statistics has a p-value of 3.99e-184, which is extremely low. Therefore, we reject the null hypothesis and assume that there is a linear relationship between the target variable and independent variables.

Variables' functions of their entries into the model:

const = 74.891. When all other features equal to 0 then the price is $80.758[3], which is about $56.99 lower than the average price.

The coefficient of battery_mean is 0.007. Holding other factors constant, when battery capacity increases by 1mAh, price will increase by $0.007.

The coefficient of size_mean is 5.297. Holding other factors constant, when screen size increases by 1 inch, price will increase by $5.297.

The coefficient of storage_mean is 2.715. Holding other factors constant, when internal storage increases by 1GB, price will increase by $2.715.

The coefficient of system is -36.763. Holding other factors constant, if the system is Android, price will decrease by $36.763

The coefficient of brand is 30.751. Holding other factors constant, price will increase by $30.751.

Hypothesis 2: There exists an exponential dependence of a telephone's price on its battery capacity (i.e., a telephone's battery capacity has a positive effect on its price, and after a certain value, the growth of price increases)

Modification

We modified the (already standardized) base model in this following way: We created a new variable called *battery_ind*, which takes the value of 1 if *battery* >= 3000mAh, and 0 if *battery* < 3000mAh.

As such, the general form of the regression model for hypothesis 2:

price_mean = const + a_1 * battery_mean + a_2 * size_mean + a_3 * brand + a_4 * system + a_5 * storage_mean

in which $a_1 = b_1$ * battery_ind + c_1

⇨ a_1 * battery_mean = b_1 * battery_ind * battery_mean + c_1 * battery_mean

⇨ price_mean = const + (b_1 * battery_ind * battery_mean + c_1 * battery_mean) + a_2 * size_mean + a_3 * brand + a_4 * system + a_5 * storage_mean

Explanation:

The idea of this regression equation is that:

- if battery_ind = 0, then a_1 * battery_mean = c_1 * battery_mean; but
- if battery_ind = 1, then a_1 * battery_mean = c_1 * battery_mean + b_1 * battery_ind * battery_mean.

[3] = 137.748 + 74.89 - 0.007*2938.362 - 5.297*5.291 - 2.715*30.676 (the average price + const – sum of products of average battery, size, storage and their coefficients)

This means that the difference between phones' price with a 1 mAh difference in battery capacity is b_1 * battery_ind * battery_mean. Therefore:

- It is necessary that c_1 should be > 0 so that an increase of 1 mAh in battery capacity will increase phone's price by c_1 * battery_mean. If c_1 < 0, the price will decrease by c_1 * battery_mean.
- If battery >= 3000 mAh, b_1 should be > 0 so that the price will increase by the value of b_1 * battery_ind * battery_mean.
- If b_1 is smaller than 0, then when battery < 3000, the price will decrease by the value of b_1 * battery_ind * battery_mean.

Hence, in order to confirm the main hypothesis, we need the coefficients b_1 and c_1 to be greater than 0.

H_0: b_1 =< 0, c_1 =< 0, i.e. a phone of battery capacity >= 3000 mAh will not increase the phone's price.
H_1: b_1 > 0, c_1 > 0, i.e. a phone of battery capacity >= 3000 mAh will increase the phone's price.

Variable	Coefficients estimators	Standard error	t-test	p-value	95% confidence interval (lower bound)	95% confidence interval (upper bound)	VIF
const	92.877	19.795	4.692	0.000	54.040	131.714	
battery_mean	0.025	0.012	2.110	0.035	0.002	0.047	8.493
size_mean	6.189	9.451	-2.655	0.01	-24.731	12.352	3.523
storage_mean	2.777	0.119	23.304	0.000	2.543	3.011	2.366
system	-38.688	17.110	-8.106	0.000	-170.255	-105.120	5.291
brand	30.453	3.580	8.507	0.000	23.429	37.476	4.302
battery_ind *battery	-0.045	0.014	-3.096	0.002	-0.073	-0.016	7.346

Table 37: Results of regression for hypothesis 2

All regressors that are significant at the alpha level of 0.05 entered the model.

As can be seen, the coefficient of battery_ind * battery_mean (b1) = -0.045 < 0 and the coefficient of battery_mean (c1) = 0.025 > 0. Therefore, H_0 is not rejected.

Result: Hypothesis 2 is rejected.

As such, the regression model for hypothesis 2 takes the following form:

price_mean = 92.877 + (-0.045 * battery_ind * battery_mean + 0.025 * battery_mean)
+ 6.189 * size_mean + 30.453 * brand - 38.688 * system + 2.777 * storage_mean

Test of significance: F-statistics has a p-value of 5.69e-185, which is extremely low. Therefore, we reject the null hypothesis and assume that there is a linear relationship between the target variable and independent variables. In other words, the regression model is meaningful.

Variables' functions of their entries into the model:

- The phone price in the absence of these features will be $39.232, which is the expected price of a phone, which can be the costs for administration, facility and equipment, …

- The coefficient of the brand is 30.453. On average, when the phone brand's valued increase by 1 level, it would increase the price of the phone by $30.453.
- The coefficient of the battery capacity is 0.025 if battery capacity < 3000 mAh. On average, an increase of battery capacity by 1 mAh (but still < 3000 mAh) would increase the price of the phone by $0.025.
- The coefficient of the battery capacity is -0.07 if battery capacity >= 3000 mAh. On average, an increase of battery capacity by 1 mAh (battery capacity from 3000 mAh) would decrease the price of the phone by $0.07.
- The coefficient of the screen size is -6.189. On average, an increase of screen size by 1 inch would decrease the price of the phone by $6.189.
- The coefficient of internal storage is 2.777. On average, an increase of internal storage by 1GB would increase the price of the phone by $2.777.
- The coefficient of the operating system is -38.688. On average, the price would decrease by $38.688 if the phone is operated by Android.

VIFs interpretation: VIFs for price, size, storage, brand are higher than 2 but lower than 5, the rest higher than 5 but lower than 10. This means that there is a sign of slight multicollinearity in the model. We can also observe that the highest VIFs are battery_mean and battery_ind * battery, but this structural multicollinearity does not spoil the model as we created the new variables on purpose.

In conclusion, an exponential relationship between a phone's price and its battery capacity does not exist.

Hypothesis 3: There exists an exponential dependence of a telephone's price on its screen size (i.e., a telephone's screen size has a positive effect on its price, and after a certain value, the growth of price increases)

Modification

We modified the (already standardized) base model in this following way: We created a new variable called *size_ind*, which takes the value of 1 if *size* >= 5.2 inches, and 0 if *size* < 5.2 .

As such, the general form of the regression model for hypothesis 3:

$$\text{price\_mean} = \text{const} + a_1 * \text{battery\_mean} + a_2 * \text{size\_mean} + a_3 * \text{brand} + a_4 * \text{system} + a_5 * \text{storage\_mean}$$

in which $a_2 = b_2 * \text{size\_ind} + c_2$

⇨ $a_2 * \text{size\_mean} = b_2 * \text{size\_ind} * \text{size\_mean} + c_2 * \text{size\_mean}$

⇨ $\text{price\_mean} = \text{const} + a_1 * \text{battery\_mean} + (b_2 * \text{size\_ind} * \text{size\_mean} + c_2 * \text{size\_mean}) + a_2 * \text{size\_mean} + a_3 * \text{brand} + a_4 * \text{system} + a_5 * \text{storage\_mean}$

Explanation:
The idea of this regression equation is that:

- if *size_ind* = 0, then a_2* size_mean = c_2 * size_mean; but
- if size_ind = 1, then a_2 * size_mean = c_2 * size_mean + b_2 * size_ind * size_mean.

This means that the difference between phones' price with a 1-inch difference in screen size is b2 * size_ind * size_mean. Therefore:

- It is necessary that c_2 should be > 0 so that an increase of 1 inch in screen size will increase phone's price by c_2 * size_mean. If c_2 < 0, the price will decrease by c_2* size_mean.
- If size >= 5.2, b_2 should be > 0 so that the price will increase by the value of b_2 * size_ind * size_mean.
- If b_2 is smaller than 0, then when size < 5.2, the price will decrease by the value of b_2 * size_ind * size_mean.

Hence, in order to confirm the main hypothesis, we need the coefficients b_2 and c_2 to be greater than 0.

H_0: b_2 =< 0, c_2 =< 0, i.e. a battery capacity >= 5.2 inches will not increase the phone's price.
H_1: b_2 > 0, c_2 > 0, i.e. a battery capacity >= 5.2 inches will increase the phone's price.

Variable	Coefficients estimators	Standard error	t-test	p-value	95% confidence interval (lower bound)	95% confidence interval (upper bound)	VIF
const	105.172	19.823	5.305	0.000	66.280	144.063	
battery_mean	0.006	0.006	-2.127	0.260	-0.018	0.005	2.271
size_mean	21.812	10.934	3.458	0.001	-6.360	9.262	4.174
storage_mean	2.944	0.126	23.374	0.000	-2.697	3.191	2.628
system	-50.319	17.229	-8.725	0.000	-184.122	-116.518	4.382
brand	33.155	3.593	9.229	0.000	26.106	40.203	4.703
size_ind * size	-84.588	17.422	-4.855	0.000	-118.768	-50.408	5.378

Table 38: Results of regression for hypothesis 3

All regressors that are significant at the alpha level of 0.05 entered the model.

As can be seen that the coefficient of size_ind * size (b_2) = -84.588 < 0 and the coefficient of size_mean (c_2) = 37.812 > 0. Therefore, H_0 is not rejected.

Result: Hypothesis 3 is rejected.

As such, the regression model for hypothesis 2 takes the following form:

price_mean = 105.172 + 0.006 * battery_mean + (-84.588 * size_ind * size_mean + 21.812 * size_mean) + 2.944 * storage_mean - 50.319 * system + 33.155 * brand

Test of significance: F-statistics has a p-value of 5.86e-188, which is extremely low. Therefore, we reject the null hypothesis and assume that there is a linear relationship between the target variable and independent variables. In other words, the regression model is meaningful.

Variables' functions of their entries into the model:

- The phone price in the absence of these features will be $105.172, which is the expected price of a phone, which can be the costs of administration, facility and equipment, …
- The coefficient of the battery is 0.006. On average, an increase of battery capacity by 1 mAh would increase the price of the phone by $0.006.
- The coefficient of the screen size is 21.812 if battery capacity < 5,2 inches. On average, an increase of screen size by 1 inch (but still < 5.2 inch) would increase the price of the phone by $37.811.
  The coefficient of the battery capacity is -46.777[4] if screen size >= 5,2 inch. On average, an increase of screen size by 1 inch (screen size from 5.2 inch) would decrease the price of the phone by $46.777.
- The coefficient of internal storage is 2.944. On average, an increase of internal storage by 1GB would increase the price of the phone by $2.944.
- The coefficient of the operating system is -50.319. On average, the price would decrease by $50.319 if the phone is operated by Android
- The coefficient of the brand is 33.155. On average, when the phone brand's valued increase by 1 level, it would increase the price of the phone by $33.155.

VIFs interpretation: VIFs for *price, battery, size, storage, brand, system* are higher than 2 but lower than 5. This means that there is a sign of slight multicollinearity in the model. We can also observe that the highest VIF is *size_ind * size* (which is greater than 5), but this structural multicollinearity does not spoil the model as we created the new variable on purpose.

In conclusion, the exponential relationship between a phone's price and its screen size does not exist. This can be explained by the fact that in recent years, the trend to produce mobile phones with a big screen size of mobiles have dominated the market, and the Indian one is not an exception. Although we found out that a bigger screen size might increase the price, it does not mean that after a certain value, the price will increase faster, as big-screen mobile phones have already become a norm on the market.

Hypothesis 4: The brand value of a telephone has different effects on how its internal storage determines its price
(i.e., it is not always the case that on the rise of the internal storage would the price increase, as this also depends on the brand value)

Modification

In this model, we divided the brand value into only 2 categories: low-valued brands and high-valued brands. We created a new variable called *brand_id*, where top 15 brands (i.e. those of levels 3 and 4) receive the value of 1, and the rest (i.e. those of levels 3 and 4) receive the value of 0.

There are two sides to this hypothesis: one, internal storage increases the price; two, even if the amounts of internal storage are the same, phones that are from high-valued brands tend to have higher price.

[4] -46.777 = -84.588 + 37.812

As such, the general form of the regression model for hypothesis 4:

price_mean = const + a_1 * battery_mean + a_2 * size_mean + a_3 * brand_id + a_4 * system + a_5 * storage_mean

in which $a_5 = b_5 * brand\_id + c_5$

⇨ a_5 * storage_mean = b_5 * brand_id * storage_mean + c_5* storage_mean

⇨ price_mean = const + a_1 * battery_mean + a_2 * size_mean + a_3 * brand_id + a_4 * system + (b_5 * brand_id * storage_mean + c_5* storage_mean)

Explanation

The idea of this regression equation is that

- if brand_id = 0, then a_5 * storage_mean = c_5 * storage_mean; but
- if brand_id = 1, then a_5 * storage_mean = c_5 * storage_mean + b_5 * brand_id * storage_mean.

This means that the difference between a phone's price of a low-valued brand and a phone's price of a high-valued is b5 * brand_id * storage_mean. Therefore:

- If the brand's value is high, with the same internal storage, b_5 should be > 0 so that the price will increase (by the value of b_5 * brand_id * storage_mean)
- If b_5 is smaller than 0, then when the brand's value is high, with the same internal storage, the price will decrease (by the value of b_5 * brand_id * storage_mean).

Hence, in order to confirm the main hypothesis, we need the coefficients b_5 and c_5 to be greater than 0.

H_0: $b_5 =< 0$, $c_5 =< 0$, i.e. with the same internal storage, a high-valued brand will decrease (or does not have any effect on) the phone's price.

H_1: $b_5 > 0$, $c_5 > 0$, i.e. with the same internal storage, a high-valued brand will increase the phone's price.

Variable	Coefficients estimators	Standard error	t-test	p-value	95% confidence interval (lower bound)	95% confidence interval (upper bound)	VIF
const	124.424	17.108	7.273	0.000	90.860	157.988	
battery_mean	0.005	0.006	-2.770	0.005	-0.016	0.007	2.307
size_mean	9.762	8.772	2.113	0.035	-7.449	26.972	2.869
storage_mean	2.411	0.184	13.129	0.000	2.050	2.771	4.547
system	-147.986	17.143	-8.633	0.000	-181.618	-114.354	1.438
brand	62.336	9.064	6.877	0.000	44.553	80.119	1.643
brand_storage	0.493	0.207	2.382	0.017	0.087	0.900	3.214

Table 39: Results of regression for hypothesis 4

All regressors that are significant at the alpha level of 0.05 entered the model.

As can be seen, the coefficient of brand_storage (b_5) = 0.493 > 0 and the coefficient of storage_mean = 2.415 > 0. Therefore, H_0 is rejected.

Result: Hypothesis 4 is confirmed.

As such, the regression model for hypothesis 4 takes the following form:

price_mean = 124.424 + 0.005 * battery_mean + 9.762 * size_mean + 62.336 * brand_id
– 147.986 * system + (0.494 * brand_id * storage_mean + 2.411 * storage_mean)

Test of significance: F-statistics has a p-value of 3.46e-180, which is extremely low. Therefore, we reject the null hypothesis and assume that there is a linear relationship between the target variable and independent variables. In other words, the regression model is meaningful.

Variables' functions of their entries into the model:

const = 124.424. When all other features equal to 0 then the price is $121.878 , which is about $16 lower than the average price

The coefficient of battery_mean is 0.005. Holding other factors constant, when battery capacity increases by 1mAh, price will increase by $0.005.

The coefficient of size_mean is 9.762. Holding other factors constant, when screen size increases by 1 inch, price will increase by $9.762.

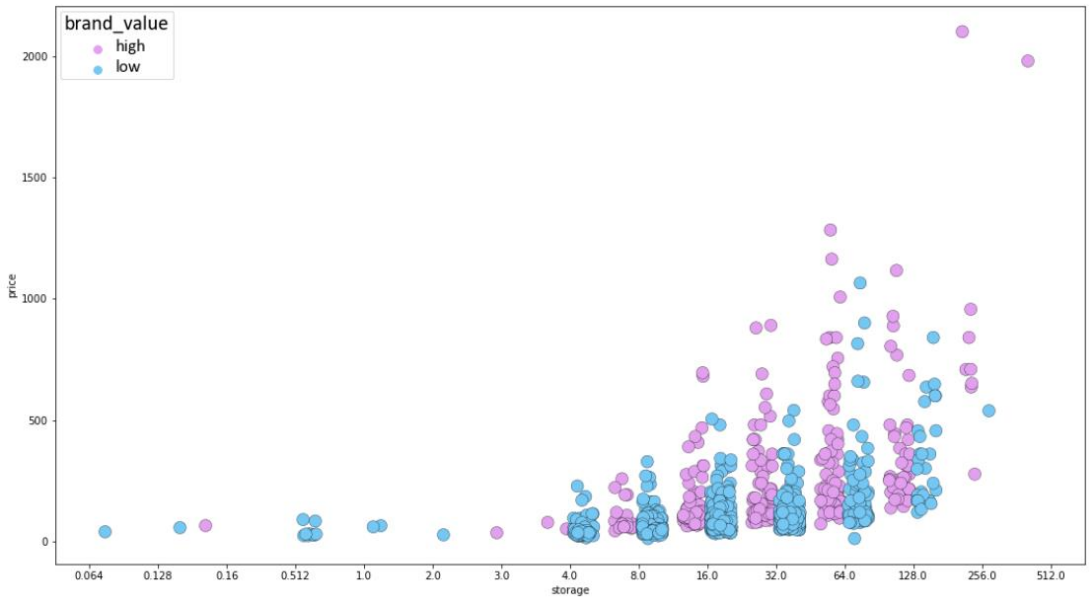
The coefficient of brand_id is 62.336. Holding other factors constant, when the brand's value is high, price will increase by $62.336.

The coefficient of system is -147.986. Holding other factors constant, if the system is Android, price will decrease by $147.986

The coefficient of storage_mean is 2.411. Holding other factors constant, when internal storage increases by 1GB, price will increase by $2.411.

VIFs interpretation: VIFs for price, brand and system are lower than 2, and the rest higher than 2 but lower than 5. This means that there is a slight sign of multicollinearity in the model. We can also observe that the highest VIFs are brand_storage and storage_mean, but this structural multicollinearity does not spoil the model as we created the new variables on purpose.

Graph: From the graph, it can be seen that there is a tendency of the price to rise when internal storage increases. At the same time, when we divided the brands into only two categories, we could observe that while the number of phones is a little bit higher for low-valued brands than the number of phones that have the same internal storage but for high-valued brands, the price for phones from high-valued brands tend to be higher, even though they have the same internal storage as many of phones from low-valued brands.



Graph 38: Strip plot of the price - storage pair sorted by brand

Hypothesis 1: The price of a telephone positively depends on its internal storage.

For this simple hypothesis, we can immediately confirm that the internal storage has a positive effect on a telephone's price. Based on our 4 built regression models, the increase of internal storage by 1GB will all result in the rise of price by roughly $2.4 - $2.9.

In conclusion, internal storage does have a positive effect on the price as it goes up, it does not mean that as long as internal storage is high then so is the price, because this also depends on whether the company that produces the phone has high-valued position on the market or not.

3.2. Analysis of outliers

	brand	battery	size	storage	system	sims	network	price
2	4	3969	6.50	64.0	0	1	1	1282.800
22	4	2658	6.50	64.0	0	1	1	839.988
26	2	4000	6.00	128.0	1	1	1	839.988
41	4	2716	5.80	64.0	0	0	1	839.988
169	2	3000	5.70	64.0	1	1	1	655.980
242	4	3200	5.70	32.0	1	0	1	607.800
243	3	3000	5.00	16.0	1	0	1	680.388
248	4	2600	5.00	16.0	1	0	1	694.980
338	1	3800	4.70	16.0	1	1	1	479.988
390	3	2900	5.20	32.0	1	1	1	551.988
439	2	3200	5.90	16.0	1	0	1	504.252
613	4	4500	6.70	128.0	1	1	1	887.988
614	4	5000	6.90	128.0	1	1	1	1115.988
617	4	3300	6.70	256.0	1	0	0	2099.880
626	4	3700	6.30	64.0	1	1	1	1006.800
627	4	2800	5.70	64.0	1	1	1	834.000
630	4	3046	5.80	64.0	0	1	1	1162.800
637	4	4500	6.53	128.0	1	1	1	927.588
651	4	4380	7.30	512.0	1	1	1	1979.988
677	1	4000	5.72	64.0	1	0	1	659.880
753	3	3180	5.70	64.0	1	1	1	694.800
757	1	4500	5.20	64.0	1	0	0	814.800
772	3	3300	5.50	32.0	1	1	1	879.600
846	1	5000	5.00	32.0	1	1	0	539.880
1186	1	4150	5.96	64.0	0	0	1	1064.628
1309	3	3430	5.50	32.0	1	1	1	690.000

Figure 1: Outliers

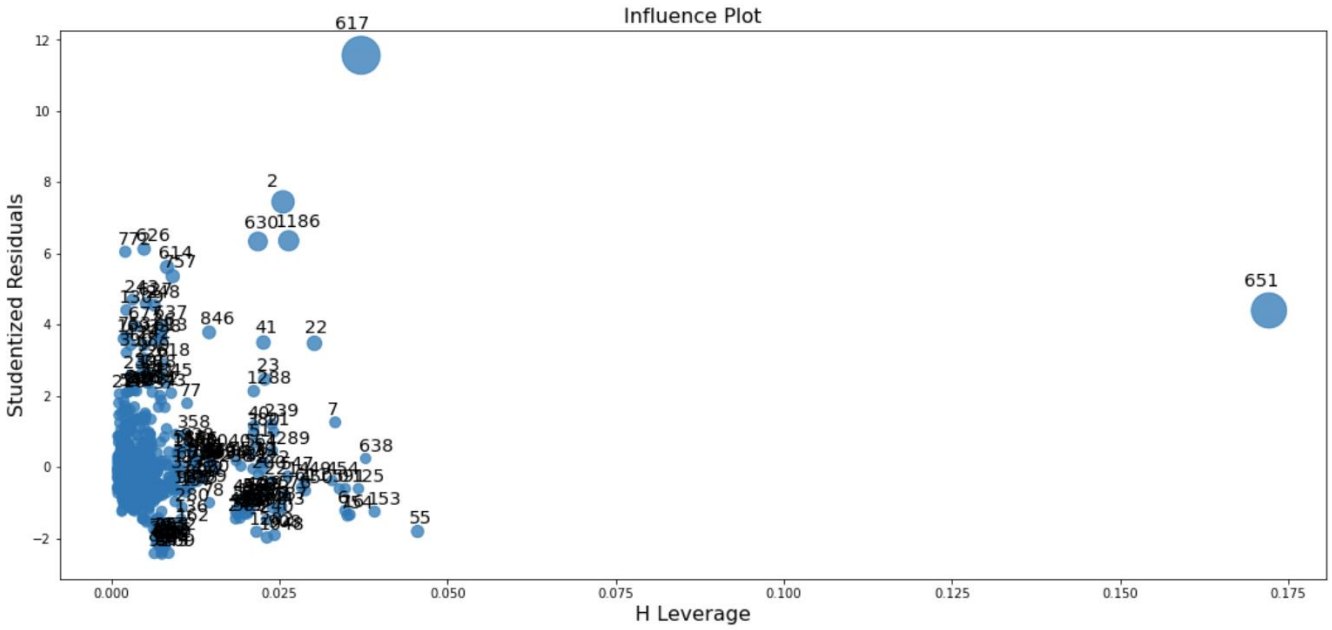
Analyzing the 26 detected outliers, these following points were found out:

- All mobile phones are priced over $479 (whereas average price is about $138)
- All mobile phones from brand 4 are priced over $800 (whereas average price is about $138) and have at least 64GB of internal storage, with only 2 exceptions of observations 242 and 248, which can be explained by their lower amount of internal storage, less battery capacity and smaller screen size than the rest.
- All mobile phones that are priced over $1000 have at least 64GB of internal storage, screen size and battery capacity exceeding their average values, and are from brand 4. However, there is one exception – observation 1186, which is a mobile phone from a low-valued brand (brand 1) but is priced at over $1000.
- A rather "strange" observation: 617 and 651 – while both are from brand 4, whereas phone^{651} has an absolute advantage in almost all studied features (the highest internal storage and screen size, having 4G/LTE network and 2 SIMs, highest battery capacity), phone^{651} is still second-placed in terms of the highest price as the first place goes to phone^{617}.

Influence plot

Technically, we received 26 outliers (as shown above), but in this graph it is quite difficult to visually detect them all. Nevertheless, the most obvious outliers are observations 617, 651, 2, 630, and 1186. They share a few similar characteristics such as being priced over $1000, battery capacity and screen size above average, internal storage at least 64GB. As mentioned above, except for the "rather strange" observation 1186, they are all from the highest-valued brands (brand 4).

In addition, observations 617 and 652 deviate significantly from the remaining observations. Their prices are the highest (approximately $2000), they have the largest screen sizes (7.3 inches and 6.7 inches) and very large amounts of internal storage (216GB and 512GB).

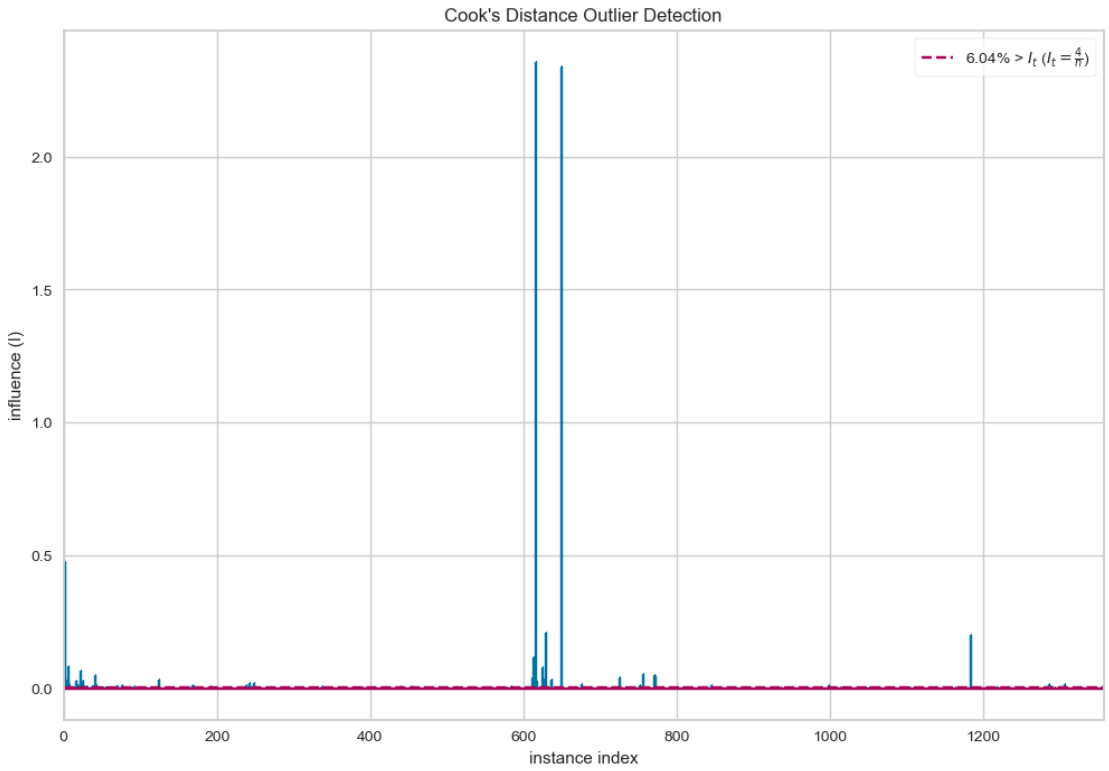


Graph 39: Influence plot

Cook's distance

By visualizing Cook's Distance, we can find out what observations can be considered as outliers by choosing those points that are above the horizontal line, where their Cook's distance (a combination of the leverage and residuals) is high.

Although it is quite difficult to detect the exact indices of observations that are outliers from the Cook's distance, it still shows that the range of indices correspond to the real index of outliers received above. Those observations that stand out from the rest are 2, 617, 652, 1186, which are exhibited evidently in the graph below.

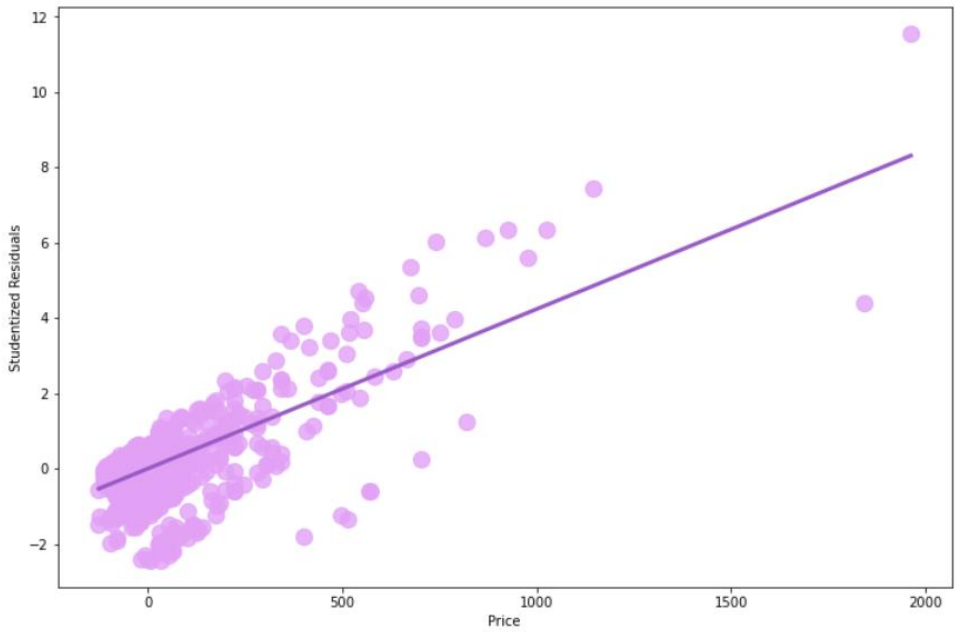


Graph 40: Cook's distance

Studentized residuals

By visualizing studentized residuals, we can find out what observations can be considered as outliers by choosing those points that have noticeable distance from the diagonal line.

Similarly, although it is nearly impossible to detect the exact indices of observations that are outliers from this graph, it still shows a number of observations that stand out from the rest. In addition, we can also observe that the "possible" outliers in this graph have their prices similar with the detected outliers above, which range from roughly $480 to $2100.



Graph 41: Studentized residuals

3.3. Analysis of the presence of heteroscedasticity

The presence of heteroscedasticity will cause our ordinary least square method to become less efficient and reliable.

H0: Homoscedasticity (the error term is the same across all values of the independent variables)

H1: Heteroscedasticity (the size of the error term differs across values of an independent variable)

White test

After we had performed the White test in order to detect heteroscedasticity, we received the F-Test p-value < 0.05, which means H_0 is rejected. Hence, *heteroscedasticity does exist in our model*.

Therefore, we applied two methods to fix this problem.

- First, we used the logarithmic transformation for the dependent variable. However, this method did not bring us the desired result of not having heteroscedasticity.
- Second, we used the generalized least square method. After applying this method, the new F-Test p-value = 0.057 > 0.05, which means we cannot reject the H_0 hypothesis.

In conclusion, using the generalized least square method helped fix our problem with the presence of heteroskedasticity.

3.4. Optimization of model

Model	R-squared	Adjusted R-squared	AIC	BIC
Pre-base model	0.595	0.592	15130	15700
Base model (standardized)	0.658	0.656	15146.4	15177.05
Model for hypothesis 2	0.662	0.659	15140	15170
Model for hypothesis 3	0.667	0.665	15120	15160
Model for hypothesis 4	0.660	0.658	15160	15520
Modified model after removing outliers and heteroskedasticity (with GLS)	0.693	0.691	1015	1045

Table 40: Evaluation of the quality of models

Based on the indicators in the table above it can be seen that the optimal model that the modified model after removing outliers, and then adjusting to eliminate heteroskedasticity with the generalized least square method. The GLS estimation is an excellent alternative for the OLS estimation because OLS tends to assume homoskedasticity while in many cases this is not correct and can lead to inaccuracy while building the model. Because our base model using the OLS estimation exhibited the presence of heteroskedasticity, it is necessary to use the GLS estimation as the best way to deal with this drawback.

As a result, we received the highest R-squared and Adjusted R-squared scores, which show that this GLS-estimated model is the best fit out of all regression models that were built. At the same time, the lowest values of the Akaike and Bayesian information criteria were also depicted in this model. Therefore, we chose the GLS-estimated model as our optimum regression model.

In addition, we also noted that compared with the pre-base model, the standardized base model we built by centering variables is better. Although removing variables (removed *sims* and *network*) tends to decrease the values of R-squared and Adjusted R-squared, the centering method in the end did improve the model.

3.5. Checking the predictive properties of the model

After optimizing the regression model, we carried out the prediction process for the price of mobile phones based on the remaining 10% of our dataset.

Having finished predicting the prices, we calculated MSE and RMSE to check how close our forecasts of the mobile phones' prices are to the actual prices on the market. In essence, the lower these values are, the better our model fits the dataset. As a result, we received MSE = 1496.336 (dollars^2) and RMSE = 38.682 (dollars), which is generally acceptable for our dataset, considering the average price is about $138.

Next, in order to test the quality of our model again, we built the 95% confidence intervals for the values of the dependent variable to see if there is a 95% probability that our predicted prices lie within this range. After this, we calculated the proportion of real values that are covered by the predicted values, and received a result of 76.34%. Also this number might not be the scenario we are looking for, it is understandable because it turns out that our confidence interval is quite narrow.

Returning to the applied task in the beginning, we set the task of predicting the price for phones with these following features: battery >= 2600mAh, SIMs = 2, size = 5 - 7 inches, storage >= 32GB, brand = 1, system = Android. From our predictions, we found 7 out of 134 observations that came very close to this, and the difference between the predicted prices and their real prices ranges from $0.979 to $18.149. Again, this is an acceptable range considering the fact that our predictions tend to deviate from the real values by about $38. In addition, the reason why we focused only on phones produced by brand 1 is because we would like to see if we can predict their prices given the characteristics that we considered challenging for such low-valued brands.

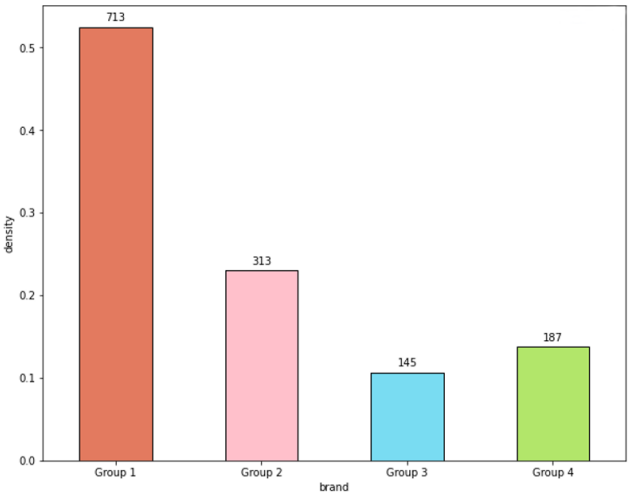
3.6. Extra-findings

These tasks have actually been applied in the previous parts, which were noted above. However, since the results of these findings did not change but only reinforce our decision or result earlier, we decided to place them near the end of this section.

Stratification

Our goal for using stratification is to reinforce our confirmation of hypothesis 4, which is about the fact that the brand value of a phone can strongly affect its price.

We stratified our sample into 2 stratas, each consists of 250 observations: one strata is for phones from low-valued brands, another is for those from high-valued brands.



Graph 10: Bar chart of brand

As we were trying to look closely into how the brand value differs from each other in terms of determining their prices, we considered between stratified sampling and clustering, and decided to adopt stratification because:

- Our sample has been found to be heterogeneous.
- There's an imbalance in the number of phones produced in different brands, therefore it is necessary to try to choose the same number of observations from each strata.
- Stratification ensures intricacy and accuracy.
- Stratification ensures more "fairness" as both strata are represented by an equal number of observations.

After stratifying our sample and perform thoroughly every stage of building regression models (from the base model and the models for hypotheses to the final model adjusted for outliers and heteroskedasticity), our result is that:

- In general, the coefficients for *const*, *size* and *system* have changed significantly.
  - The coefficient of *const* has increased. This positive contribution is highly likely because of the higher prices from high-valued brands, and the fact that it increased in spite of the reduction in the number of phones from low-valued brands also enforces this point.
  - The coefficient of *size* has increased. Similarly, this positive contribution is highly likely because of the reduction in the number of phones from low-valued brands, because high-valued brands usually have more capacity to produce phones with larger screen size. This is also connected with the fact that screen size of these phones has more impact on their prices.
  - The coefficient for system has decreased, which means if the system is Android, price will decrease even more than before stratification.
- The high-valued brands after stratification increased phone price by $35.751. This can be explained by the fact that because the number of phones from low-valued brands is not dominant anymore, and phones' prices for high-valued brands are more expensive. This supports our conclusion that phone brand does have an impact on its price.
- However, we decided not to build our model based on the stratified sampling, because after performing all necessary processes, we realized that because of the reduced number of observations, our models' the goodness of fit have decreased and this also make our "training" stage for independent variables less profound than with a large number of observations. Instead, we only drew a conclusion to support hypothesis 4.

Likelihood ratio rest for nested models

In general, the likelihood ratio test (LRT) is used to examine more precisely the goodness-of-fit between a more complex and a simpler model to see which one fits the studied dataset significantly better. Theoretically, the addition of more variables into the model should result in a higher likelihood score, and vice versa, removing variables should result in a lower score.

While building the base regression model, because we have decided to remove two variables *sims* and *network*, because of the high multicollinearity and insignificance, as well as the fact that they

make no difference to the functioning of our hypotheses. Therefore, we want to use the LRT to answer again the question of whether removing outliers would make our regression perform worse.

The complex model is the (not yet standardized) base model, including *sims* and *network.*
The nested model is the (not yet standardized) base model, having been removed *sims* and *network.*

H_0: The complex model does not fit the data as well as the nested model.
H_1: The complex model fits the data as well as the nested model.

Result:

Likelihood ratio statistics	18.388
p-value	0.0001

Because p-value < 0.05, H_0 is rejected, which means that the complex base model with more independent variables fits the dataset better than the nested model.

However, the LRT does not take into account the multicollinearity. As our VIFs for variables *sims* and *network* are relatively high (25.383305 and 24.462913 accordingly), we have to remove them, knowing that this would decrease the R-squared value.

In this chapter, we managed to build 5 models.

For the base model, we detected multicollinearity and decided to remove 2 qualitative independent variables. By using the LRT for the nested model, we confirmed that this decision would lower the R-square value. However, we improved this drawback by using the centering variables method, which also helped us to interprest coefficients more easily. We also performed a parallel task, in which we stratified the sample into 2 statas.

For the models of hypotheses 2,3 and 4, we modified accordingly to the independent variables that involves directly with the hypotheses, and manged to confirm hypothesis 4 and reject hypotheses 2 and 3. Model 1 did not need modifications, and was confirmed easily.

For the optimal model, we removed 26 outliers and adjusted for heteroskedasticity by using the GLS estimation.

Finally, we managed to predict prices based on the applied task, and found their deviation of about $38 from the real prices quite acceptable.

4. CONCLUSION

In the end, we have reached the goal of studying the mobile phone's market in India and predicting prices based on the factos in our applied task.

In the preliminary analysis of the data, we recognized the imbalance of qualitative variables (*brand, network, sims, system*) and skewness of quantitative variables. We also kept in mind that that there may exist multicollinearity in our dataset, as the correlations between some independent variables are relatively high. As expected, in the next section, we had to perform an additional test called LRT before removing variables that caused multicollinearity, then we finished reducing multicollineartity and improved R-squared caused by the removal of variables by centering variables.

Compared to the preliminary result of testing of hypotheses in part 2.2.4, the results were similar to our predictions, with the exception of hypothesis 3: hypothesis 1 – confirmed, hypothesis 2 – rejected, hypothesis 3 – rejected, hypothesis 4 – confirmed. In essence, we concluded that after a certain value of battery capacity or screen size, the increase in the price does not necessarily rice. Also, the amount of internal storage does increase the price, but the value of the brand also plays a critical part in determing the price, which we also double-confirmed by performing stratification for *brand.*

We also detected 26 outliers that could spoil our model and the presence of heteroskedasticity, and after removing all the outliers and using the GLS estimation to deal with this problem, we succeeded in raising the R-squared and adjusted R-squared value to 0.693 and 0.691, and chose this GLS-estimated model as the optimal one.

Finally, we tested the predictive ability of our model on a test sample and found a reasonable error, even though not all of our forecasts fall into the 95% confidence interval. We also completed the ultimate goal, which is to predict the price of phones with some specific features and got an error of below $20.