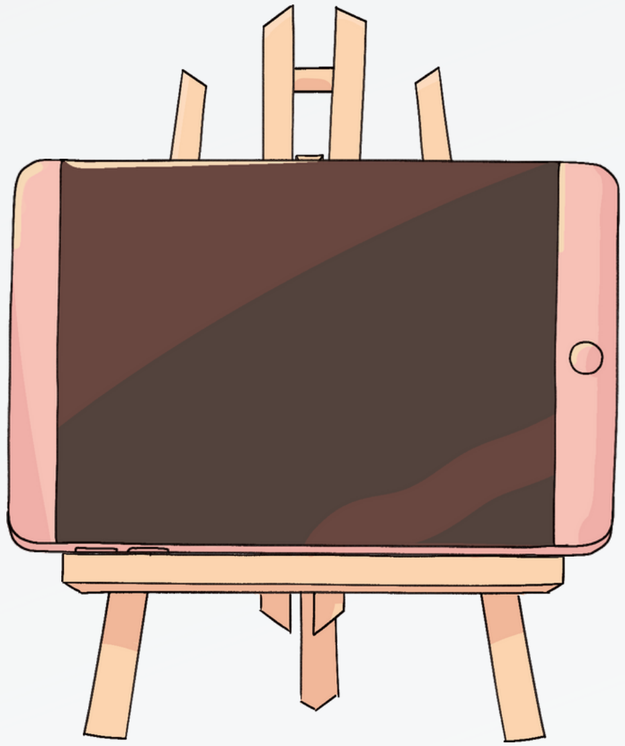
Institute of Technology of Cambodia

Programming for Data Science

3rd year Engineer's Degree in Data Science
Department of Applied Mathematics and Statistics



Lecturer: Chan Sophal

GROUP: 5

PROJECT: PHONE PRICE ANALYSIS

2022-2023

GROUP MEMBERS

PHAL DAVY	*NOEM KOEMHAK*	*NOR PHANIT*	*LAY CHHAY*	*NHAR RATANAK*
e20201437	*e20200808*	*e20200241*	*e20200054*	*e20190682*

Table Of Contents

01 INTRODUCTION

What is Phone Price Analysis?

Phone price analysis is the process of collecting and analyzing data on the prices of mobile phones. This data can be used to understand the factors that affect the price of a mobile phone, such as the brand, the features (Storage, Screen, Ram...), the specifications, and the market conditions.

Why is mobile phone price analysis important?

There are several reasons why mobile phone price analysis is important. It can help companies to set the right price for their mobile phones. By understanding the factors that affect the price of a mobile phone, companies can ensure that they are not overcharging or undercharging for their products.

• DATA SET

Data Loading

```
df = pd.read_csv('Final_Data.csv')
df.head()
```

There are 239 rows and 18 features...

	Title	Price	Id	Location	Mark	Model	Storage	Condition	screen_size	screen area	processor	rear_camera	front_camera	battery	operating_system	ram
0	Flip 4 99% full box	499.0	9419966	Phnom Penh	Samsung	Galaxy Z Fold4	128	Used	7.6	17.13	8	50	30.0	4400	12.0	8
1	I want sell phone galaxy Z fold 4 99%	1000.0	9423865	Phnom Penh	Samsung	Galaxy Z Fold4	256	Used	7.6	17.13	8	50	30.0	4400	12.0	8
2	Samsung ZFold4 (12g+512g)ស៊ីម២ក្រុមហ៊ុនស្អាតសំ...	880.0	9423258	Phnom Penh	Samsung	Galaxy Z Fold4	512	Used	7.6	17.13	8	50	30.0	4400	12.0	8
3	#GalaxyZfold4 256g company	1249.0	9021923	Phnom Penh	Samsung	Galaxy Z Fold4	256	New	7.6	17.13	8	50	30.0	4400	12.0	8
4	ចង់លក់-Sumsung-Z-Full-4,នៅស្អាត-(99%),របស់ក្រុ...	896.0	9024219	Phnom Penh	Samsung	Galaxy Z Fold4	256	Used	7.6	17.13	8	50	30.0	4400	12.0	8

- **Descriptive statistic**

Descriptive statistics are a set of tools used to summarize and describe data. They can be used to understand the data, identify trends, and make inferences about the population from which the data was collected.

	Price	Id	Storage	screen_size	screen area	processor	rear_camera	front_camera	battery	operating_system	ram
count	239.000000	2.390000e+02	239.000000	239.000000	239.000000	239.000000	239.000000	239.000000	239.000000	239.000000	239.000000
mean	480.937364	9.341881e+06	163.414226	6.558996	10.812385	11.891213	23.246862	14.185774	3370.121339	12.631381	5.029289
std	295.921836	1.120171e+05	129.809136	0.937272	2.982221	2.519814	27.965197	10.228283	824.797488	1.814061	1.988195
min	1.000000	9.001799e+06	64.000000	4.700000	6.900000	8.000000	8.000000	1.200000	1821.000000	8.000000	2.000000
25%	184.500000	9.318968e+06	64.000000	5.500000	8.500000	11.000000	12.000000	7.000000	2691.000000	11.000000	3.000000
50%	499.000000	9.394517e+06	128.000000	6.800000	10.800000	14.000000	12.000000	12.000000	3687.000000	14.000000	6.000000
75%	675.000000	9.417256e+06	256.000000	7.400000	10.800000	14.000000	12.000000	12.000000	3687.000000	14.100000	6.000000
max	1249.000000	9.424431e+06	1000.000000	7.600000	17.130000	14.000000	108.000000	40.000000	5000.000000	14.100000	8.000000

	Title	Location	Mark	Model	Condition
count	239	239	239	239	239
unique	220	9	2	7	2
top	Iphone12ProMax(256G)99%	Phnom Penh	Apple	iPhone 12 Pro Max	Used
freq	4	205	197	89	228

Categorical features

	Title	Location	Mark	Model	Condition
0	Flip 4 99% full box	Phnom Penh	Samsung	Galaxy Z Fold4	Used
1	I want sell phone galaxy Z fold 4 99%	Phnom Penh	Samsung	Galaxy Z Fold4	Used
2	Samsung ZFold4 (12g+512g)ស៊ីម២ក្រុមហ៊ុនស្អាតសំ...	Phnom Penh	Samsung	Galaxy Z Fold4	Used
3	#GalaxyZfold4 256g company	Phnom Penh	Samsung	Galaxy Z Fold4	New
4	ចង់លក់-Sumsung-Z-Full-4,នៅស្អាត-(99%),របស់ក្រុ...	Phnom Penh	Samsung	Galaxy Z Fold4	Used

Numerical features

	Price	Id	Storage	screen_size	screen area	processor	rear_camera	front_camera	battery	operating_system	ram
0	499.0	9419966	128	7.6	17.13	8	50	30.0	4400	12.0	8
1	1000.0	9423865	256	7.6	17.13	8	50	30.0	4400	12.0	8
2	880.0	9423258	512	7.6	17.13	8	50	30.0	4400	12.0	8
3	1249.0	9021923	256	7.6	17.13	8	50	30.0	4400	12.0	8
4	896.0	9024219	256	7.6	17.13	8	50	30.0	4400	12.0	8

02 DATA CLEANING

Missing Value

Missing values can be used to represent data that is not available or that has been intentionally omitted.

Missing Value

```
feature_nan = [feature for feature in df.columns if df[feature].isnull().sum() >= 1]
feature_nan
```

```
[]
```

Duplicate

Duplicates can be problematic and analysis because they can skew the results of the analysis.

Duplicate

```
df.duplicated().sum()
```

```
0
```

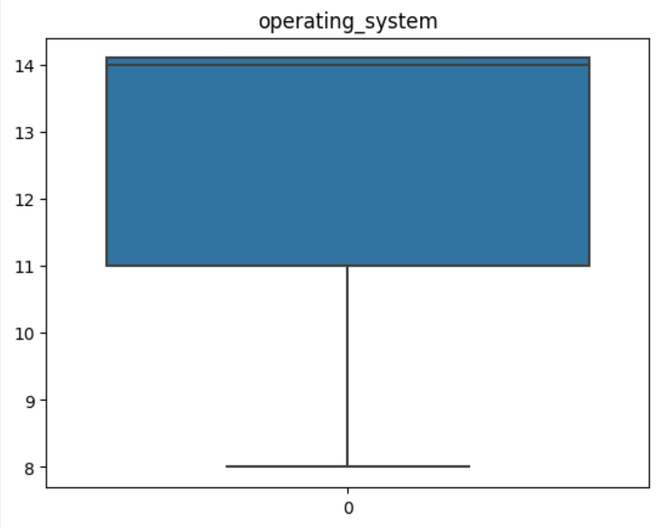
Outlier

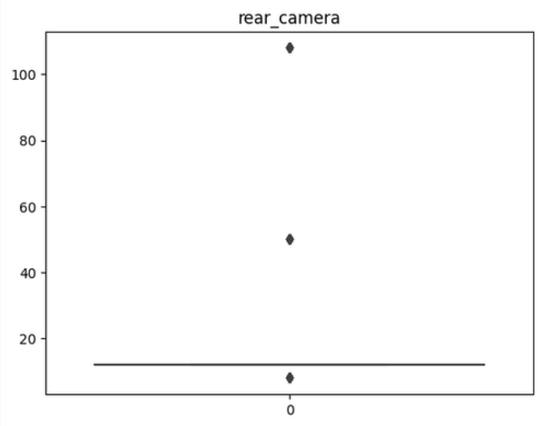
Outliers can also be problematic because they can mislead the analyst into thinking that there is a trend in the data that does not actually exist.

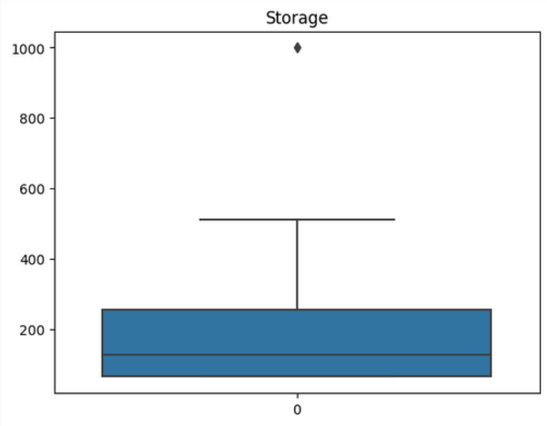
Outlier

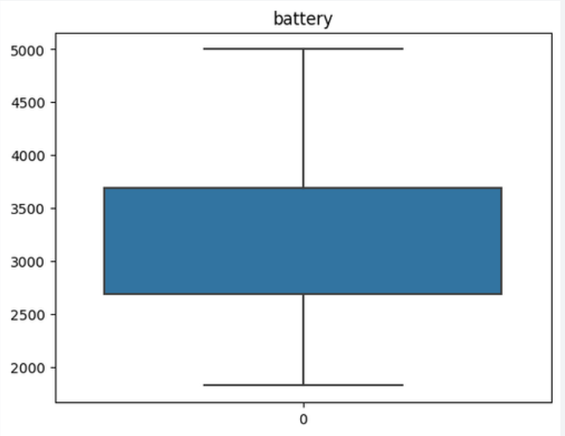
```
data = df.drop(columns = ['Mark', 'Model', 'Condition'])
for col in data.columns:
    plt.title(col)
    sns.boxplot(data[col])
    plt.show()
```

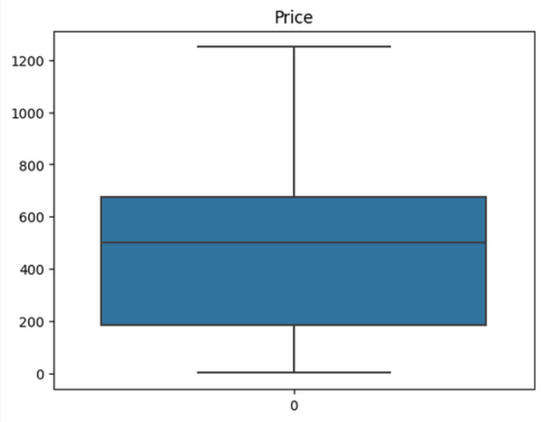
- **Outlier** when we drop columns Mark, Model and Condition.

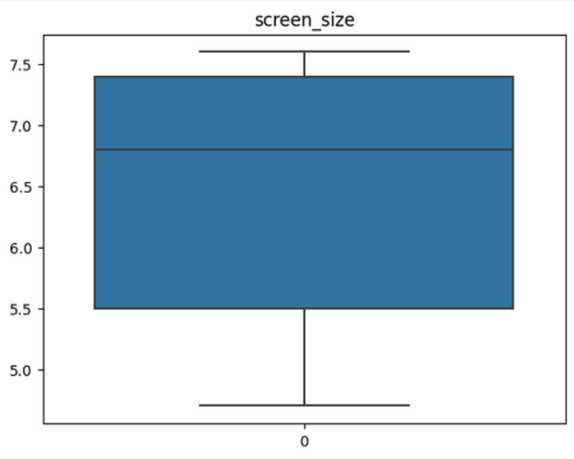


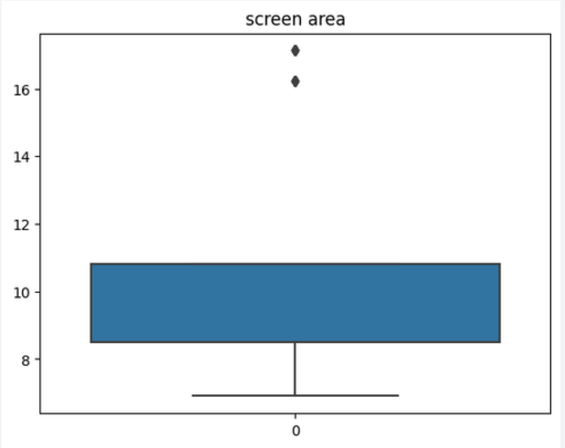


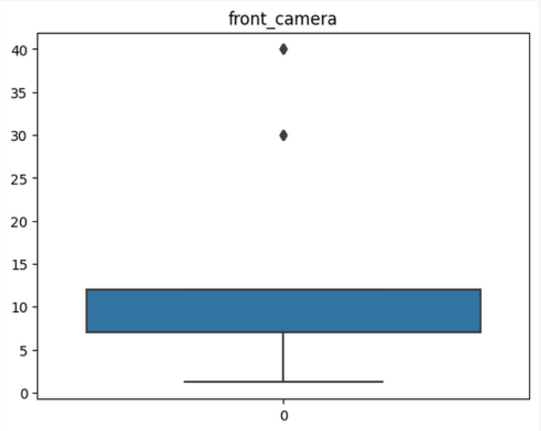


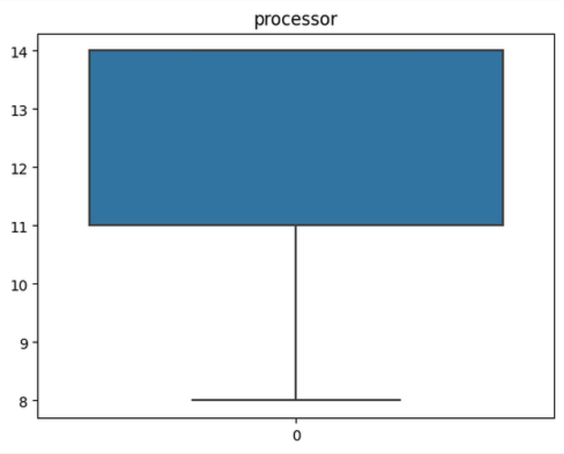


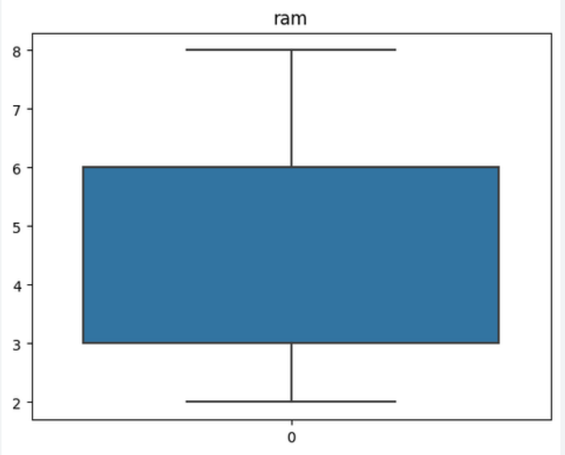






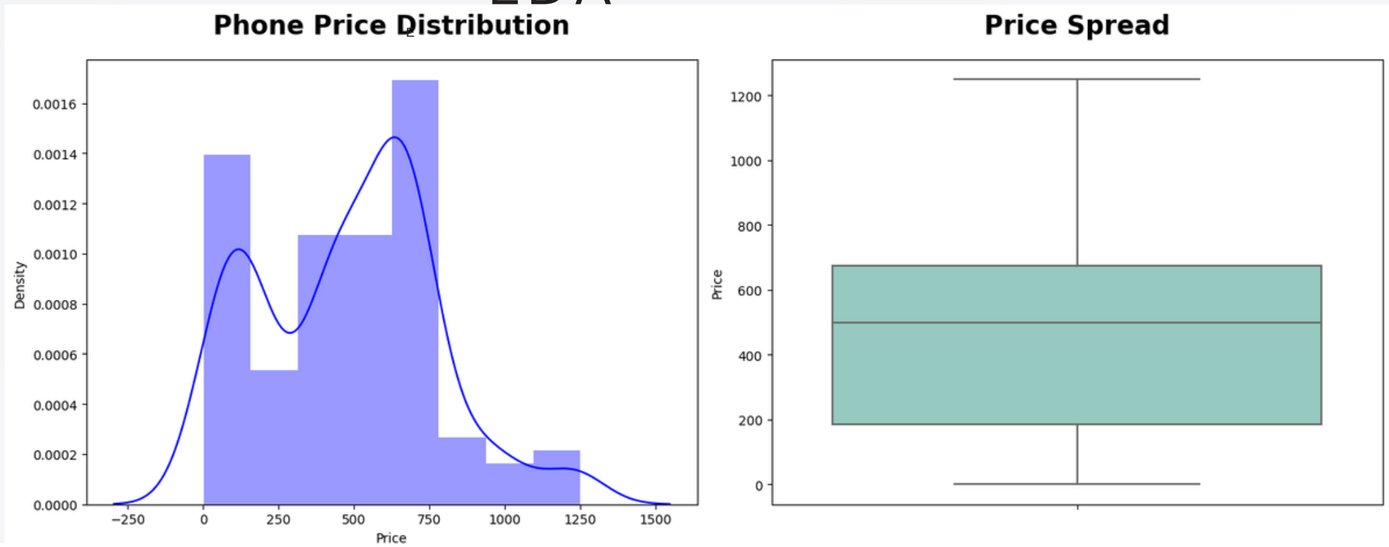






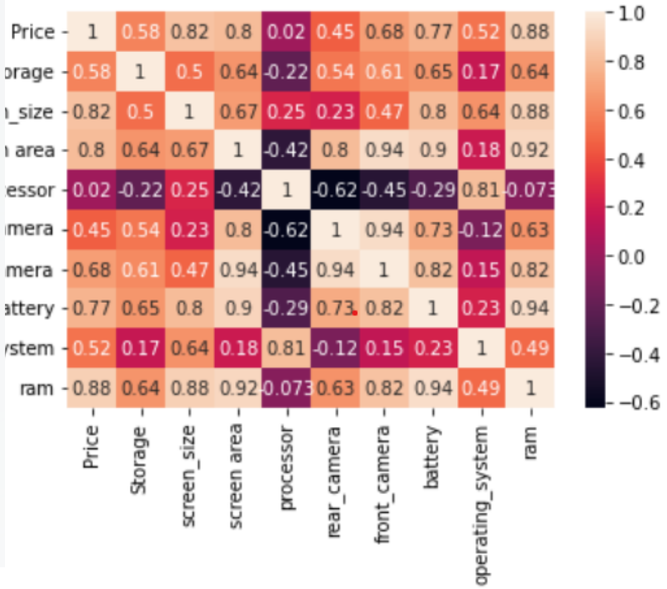
03 EXPLORATORY DATA ANALYSIS (EDA)

EDA



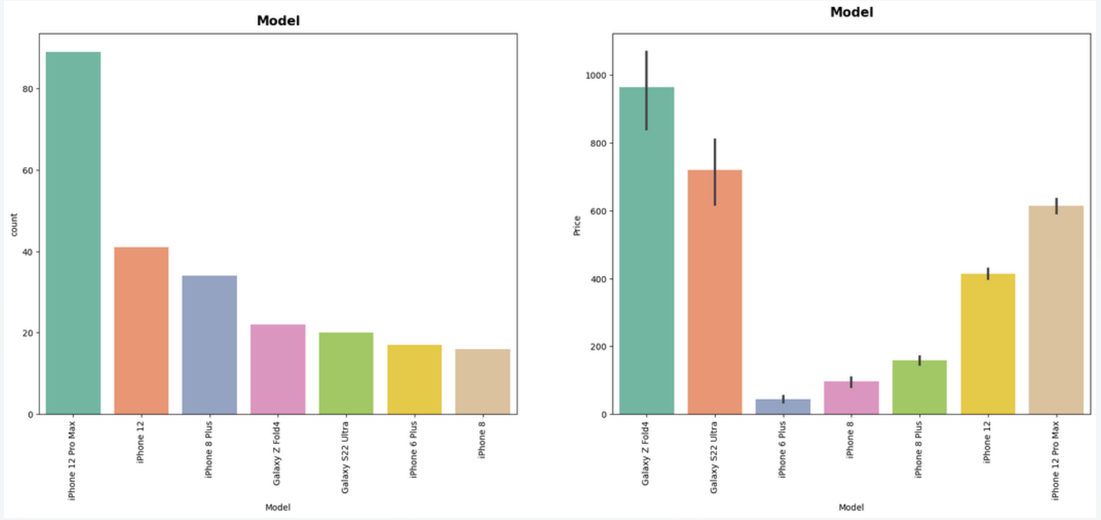
	min	mean	median	max	std	skew
Price	1.0	480.937364	499.0	1249.0	295.921836	0.25602

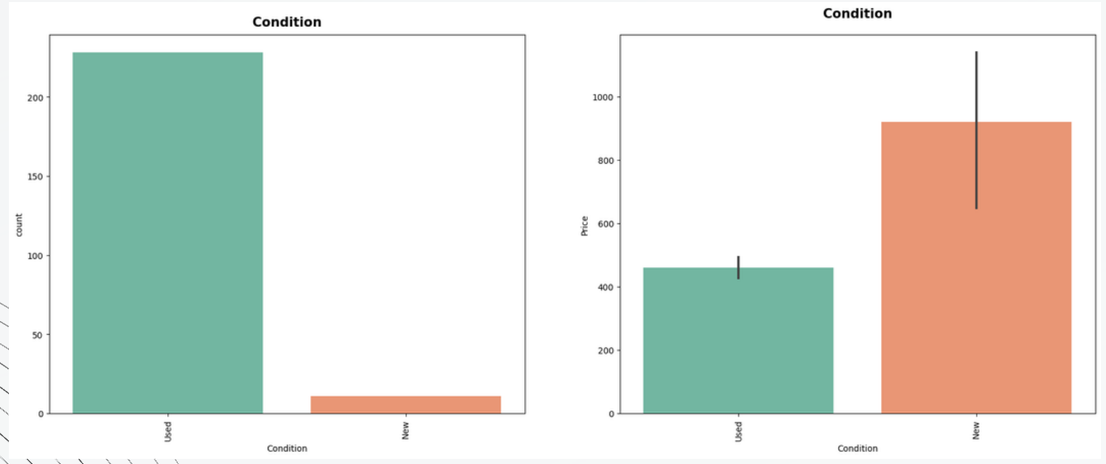
CORRELATION

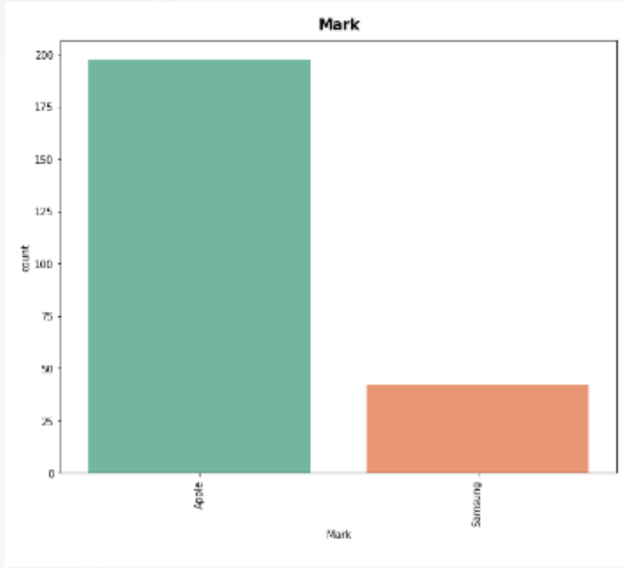


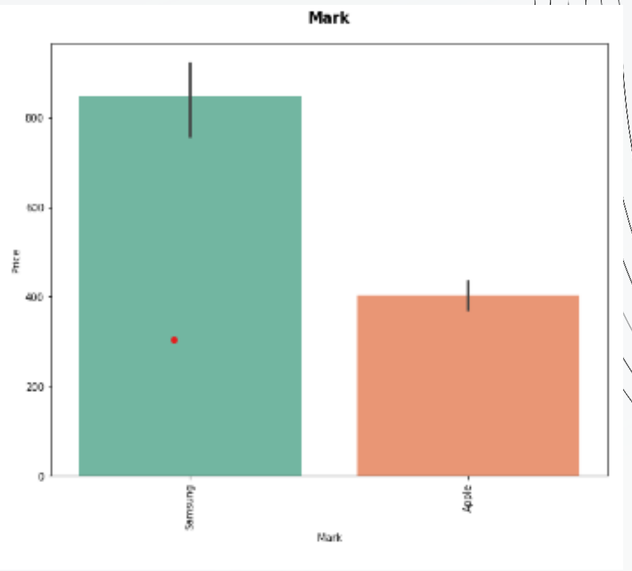
• Data Visualization

FOR CATEGORICAL FEATUERS

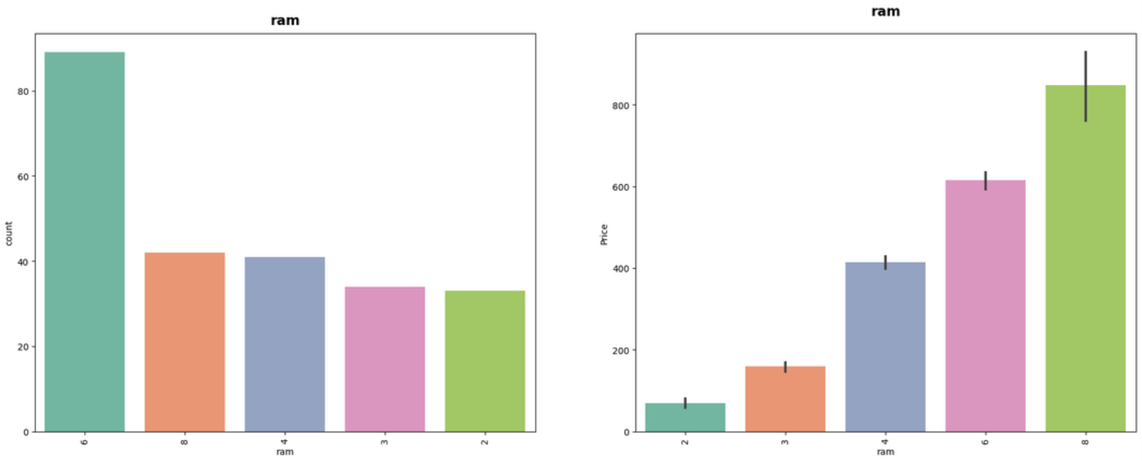


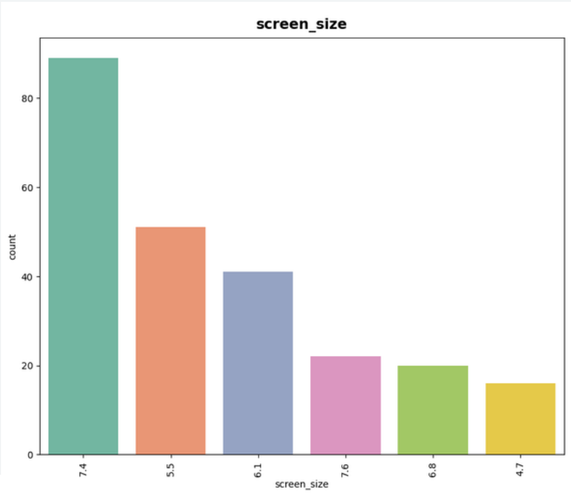


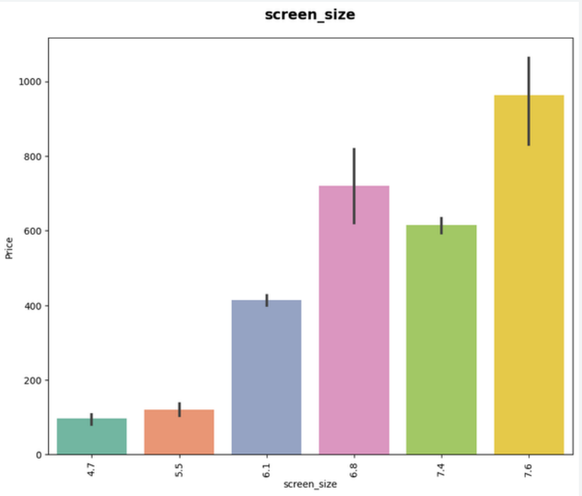


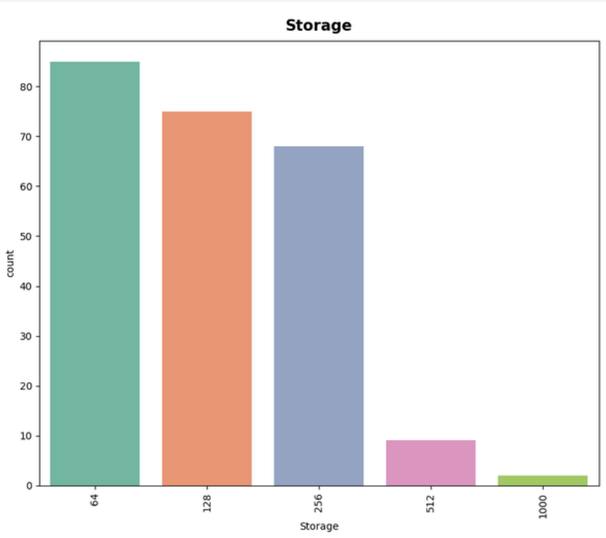


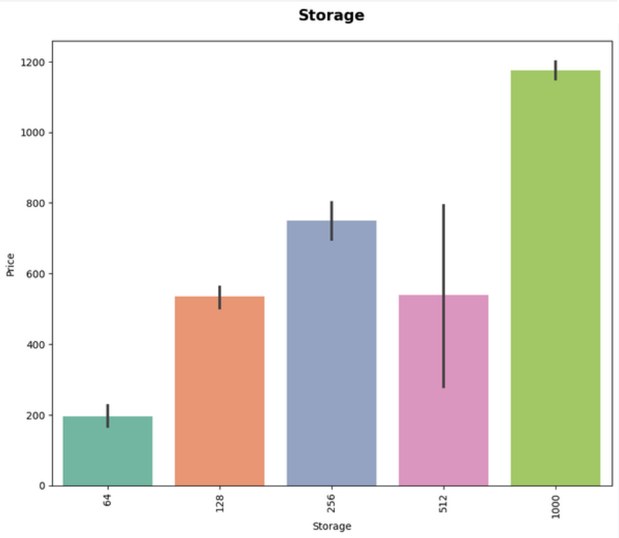
FOR NUMERICAL FEATURES

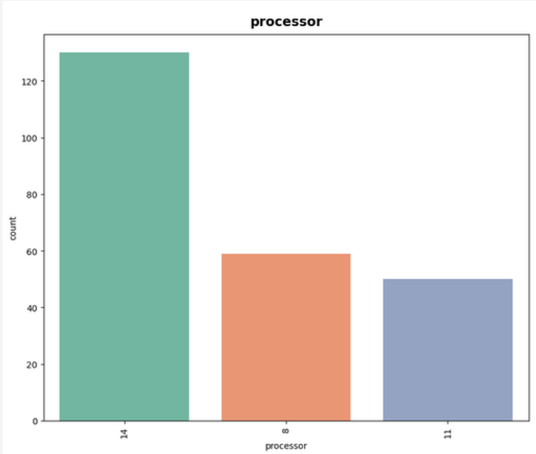


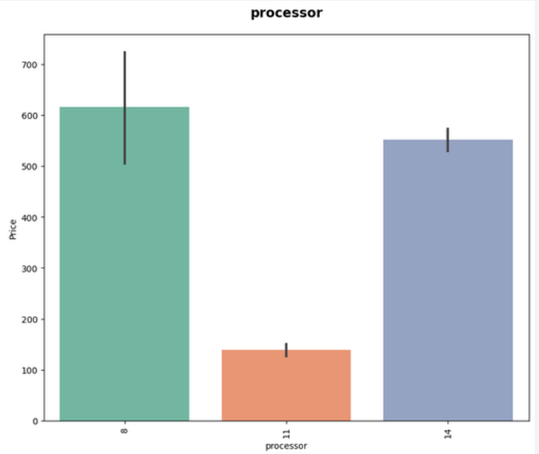












04 DATA PREPROCESSING

Based on Khmer Market, we can see that there is no phone cost lower that $10, so we decided to drop.....

```
df = df.drop(df[df['Price'] <= 10].index)
df.shape
```

```
(234, 13)
```

```
df.shape
```

```
(234, 13)
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 234 entries, 0 to 238
Data columns (total 13 columns):
 #   Column            Non-Null Count  Dtype  
---  ------            --------------  -----  
 0   Price             234 non-null    float64
 1   Mark              234 non-null    object 
 2   Model             234 non-null    object 
 3   Storage           234 non-null    int64  
 4   Condition         234 non-null    object 
 5   screen_size       234 non-null    float64
 6   screen area       234 non-null    float64
 7   processor         234 non-null    int64  
 8   rear_camera       234 non-null    int64  
 9   front_camera      234 non-null    float64
 10  battery           234 non-null    int64  
 11  operating_system  234 non-null    float64
 12  ram               234 non-null    int64  
dtypes: float64(5), int64(5), object(3)
memory usage: 25.6+ KB
```

- Low = 1 (>10 - 499)
- Medium = 2 (500 - 999)
- High = 3 (1000 - 1300)

	Model	Storage	Condition	screen_size	screen area	processor	rear_camera	front_camera	battery	operating_system	ram	price_category
0	Galaxy Z Fold4	128	0	7.6	17.13	8	50	30.0	4400	12.0	8	1
1	Galaxy Z Fold4	256	0	7.6	17.13	8	50	30.0	4400	12.0	8	2
2	Galaxy Z Fold4	512	0	7.6	17.13	8	50	30.0	4400	12.0	8	2
3	Galaxy Z Fold4	256	1	7.6	17.13	8	50	30.0	4400	12.0	8	2
4	Galaxy Z Fold4	256	0	7.6	17.13	8	50	30.0	4400	12.0	8	2

- **Dummy**

	Storage	Condition	screen_size	screen area	processor	rear_camera	front_camera	battery	operating_system	ram	price_category
count	234.000000	234.000000	234.000000	234.000000	234.000000	234.000000	234.000000	234.000000	234.000000	234.000000	234.000000
mean	0.102345	0.042735	0.642205	0.380438	0.655983	0.149316	0.332562	0.485991	0.763486	0.504274	0.271368
std	0.133224	0.202693	0.321585	0.288030	0.417208	0.276179	0.260241	0.256963	0.295292	0.328260	0.293112
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.275862	0.156403	0.500000	0.040000	0.149485	0.273671	0.491803	0.166667	0.000000
50%	0.068376	0.000000	0.724138	0.381232	1.000000	0.040000	0.278351	0.586977	0.983607	0.666667	0.000000
75%	0.205128	0.000000	0.931034	0.381232	1.000000	0.040000	0.278351	0.586977	1.000000	0.666667	0.500000
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

Galaxy S22 Ultra	Galaxy Z Fold4	iPhone 12	iPhone 12 Pro Max	iPhone 6 Plus	iPhone 8	iPhone 8 Plus
234.000000	234.000000	234.000000	234.000000	234.000000	234.000000	234.000000
0.081197	0.089744	0.175214	0.376068	0.068376	0.064103	0.145299
0.273722	0.286427	0.380965	0.485436	0.252931	0.245461	0.353158
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000
1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

- **KBest**

Feature_scores & Names

	Feature_Scores	Feature_Names
0	0.416882	Storage
1	0.038211	Condition
2	0.585401	screen_size
3	0.607792	screen area
4	0.224578	processor
5	0.259710	rear_camera
6	0.395589	front_camera
7	0.612622	battery
8	0.603419	operating_system
9	0.580142	ram
10	0.055909	Galaxy S22 Ultra
11	0.051296	Galaxy Z Fold4
12	0.181332	iPhone 12
13	0.286484	iPhone 12 Pro Max
14	0.081703	iPhone 6 Plus
15	0.112534	iPhone 8
16	0.088692	iPhone 8 Plus

The important features

	Feature_Scores	Feature_Names
7	0.612622	battery
3	0.607792	screen area
8	0.603419	operating_system
2	0.585401	screen_size
9	0.580142	ram
0	0.416882	Storage
6	0.395589	front_camera
13	0.286484	iPhone 12 Pro Max
5	0.259710	rear_camera
4	0.224578	processor

05 MODEL BUILDING

Model building can be used in various ways to predict or estimate phone prices in a project such as Model Training, Model Evaluation, Model Optimization, Model Deployment

Split data into training and testing sets

```
print(f'Data train: {X_train.shape, y_train.shape}')
print(f'Data test: {X_test.shape, y_test.shape}')
```

```
Data train: ((163, 17), (163,))
Data test: ((71, 17), (71,))
```

5.1. Linear regression

Ordinary Least Squares regression (OLS) is a common technique for estimating coefficients of linear regression equations which describe the relationship between one or more independent quantitative variables and a dependent variable .

OLS Regression Results

Dep. Variable:	price_category	R-squared:	0.825
Model:	OLS	Adj. R-squared:	0.816
Method:	Least Squares	F-statistic:	90.86
Date:	Mon, 17 Jul 2023	Prob (F-statistic):	2.23e-54
Time:	09:21:10	Log-Likelihood:	-3.0559
No. Observations:	163	AIC:	24.11
Df Residuals:	154	BIC:	51.96
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0140	0.004	3.133	0.002	0.005	0.023
Storage	0.0010	0.000	5.082	0.000	0.001	0.001
Condition	0.5288	0.106	5.009	0.000	0.320	0.737
screen_size	0.0723	0.016	4.591	0.000	0.041	0.103
screen area	0.0137	0.018	0.755	0.451	-0.022	0.049
processor	-0.0287	0.008	-3.736	0.000	-0.044	-0.014
rear_camera	2.773e-06	0.003	0.001	0.999	-0.006	0.006
front_camera	-0.0077	0.009	-0.863	0.390	-0.025	0.010
battery	8.432e-05	8.68e-05	0.972	0.333	-8.71e-05	0.000
operating_system	0.0263	0.008	3.319	0.001	0.011	0.042
ram	0.1080	0.015	7.271	0.000	0.079	0.137
Galaxy S22 Ultra	-0.0118	0.003	-3.896	0.000	-0.018	-0.006
Galaxy Z Fold4	0.0222	0.006	3.833	0.000	0.011	0.034
iPhone 12	-0.2110	0.037	-5.683	0.000	-0.284	-0.138
iPhone 12 Pro Max	0.1313	0.018	7.152	0.000	0.095	0.168
iPhone 6 Plus	-0.0294	0.019	-1.528	0.129	-0.067	0.009
iPhone 8	0.1983	0.035	5.720	0.000	0.130	0.267
iPhone 8 Plus	-0.0858	0.047	-1.837	0.068	-0.178	0.006

Omnibus:	63.593	Durbin-Watson:	2.004
Prob(Omnibus):	0.000	Jarque-Bera (JB):	262.749
Skew:	-1.420	Prob(JB):	8.81e-58
Kurtosis:	8.534	Cond. No.	2.07e+20

We are drop_high_pvalue_features

```
                            OLS Regression Results
==============================================================================
Dep. Variable:         price_category   R-squared:                       0.825
Model:                            OLS   Adj. R-squared:                  0.817
Method:                 Least Squares   F-statistic:                     104.2
Date:                Mon, 17 Jul 2023   Prob (F-statistic):           2.55e-55
Time:                        09:21:10   Log-Likelihood:                -3.2610
No. Observations:                 163   AIC:                             22.52
Df Residuals:                     155   BIC:                             47.27
Df Model:                           7
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.4694      0.526      0.893      0.373      -0.569       1.508
Storage        0.0010      0.000      5.060      0.000       0.001       0.001
Condition      0.5380      0.104      5.157      0.000       0.332       0.744
screen_size    0.1240      0.108      1.149      0.252      -0.089       0.337
processor     -0.0551      0.011     -4.841      0.000      -0.078      -0.033
ram            0.1159      0.037      3.103      0.002       0.042       0.190
iPhone 12 Pro Max  0.3548      0.103      3.430      0.001       0.150       0.559
iPhone 8       0.2592      0.115      2.250      0.026       0.032       0.487
==============================================================================
Omnibus:                       61.442   Durbin-Watson:                   2.020
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              255.980
Skew:                          -1.360   Prob(JB):                     2.60e-56
Kurtosis:                       8.504   Cond. No.                     5.81e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.81e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Variance Inflation Factor

	Features	VIF
0	screen_size	110.53
1	processor	46.16
2	ram	33.94
3	Storage	4.13
4	iPhone 12 Pro Max	2.23
5	iPhone 8	1.18
6	Condition	1.12

Drop_high_VIF_features

	Features	VIF
0	processor	3.17
1	Storage	2.15
2	iPhone 12 Pro Max	1.96
3	iPhone 8	1.12
4	Condition	1.06

Last checking

OLS Regression Results

Dep. Variable:	price_category	R-squared:	0.699
Model:	OLS	Adj. R-squared:	0.691
Method:	Least Squares	F-statistic:	91.57
Date:	Mon, 17 Jul 2023	Prob (F-statistic):	3.96e-40
Time:	09:21:10	Log-Likelihood:	-47.435
No. Observations:	163	AIC:	104.9
Df Residuals:	158	BIC:	120.3
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.8613	0.170	10.937	0.000	1.525	2.197
Storage	0.0021	0.000	9.796	0.000	0.002	0.002
Condition	0.7637	0.129	5.919	0.000	0.509	1.019
processor	-0.0810	0.014	-5.648	0.000	-0.109	-0.053
iPhone 12 Pro Max	0.7302	0.075	9.708	0.000	0.582	0.879

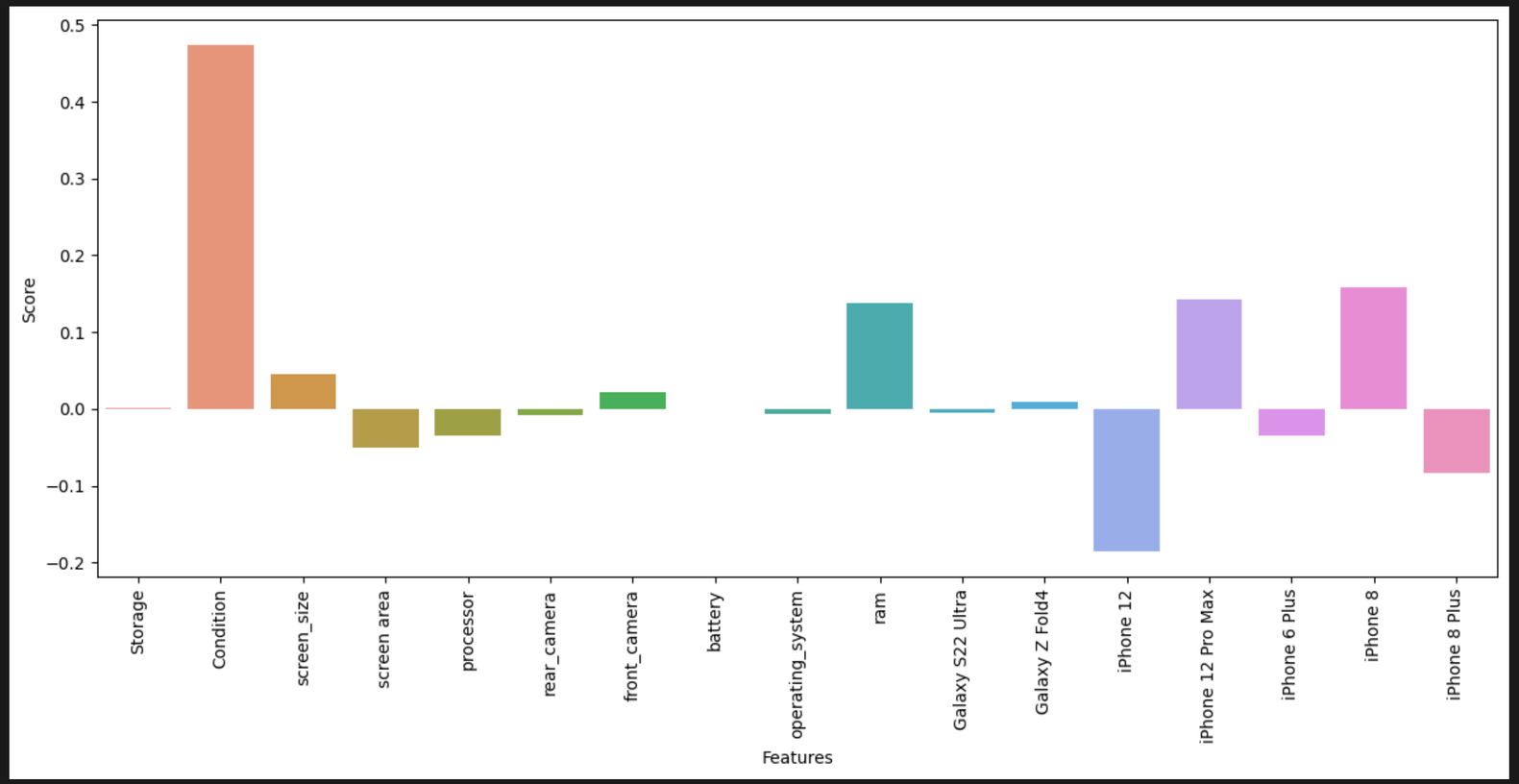
Omnibus:	10.613	Durbin-Watson:	1.778
Prob(Omnibus):	0.005	Jarque-Bera (JB):	25.250
Skew:	0.040	Prob(JB):	3.29e-06
Kurtosis:	4.927	Cond. No.	1.46e+03

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.46e+03. This might indicate that there are strong multicollinearity or other numerical problems.

5.2. Ridge Regressor

Ridge regression can be used to improve the accuracy of linear regression models when there is multicollinearity. it can make the coefficients of the independent variables unstable.

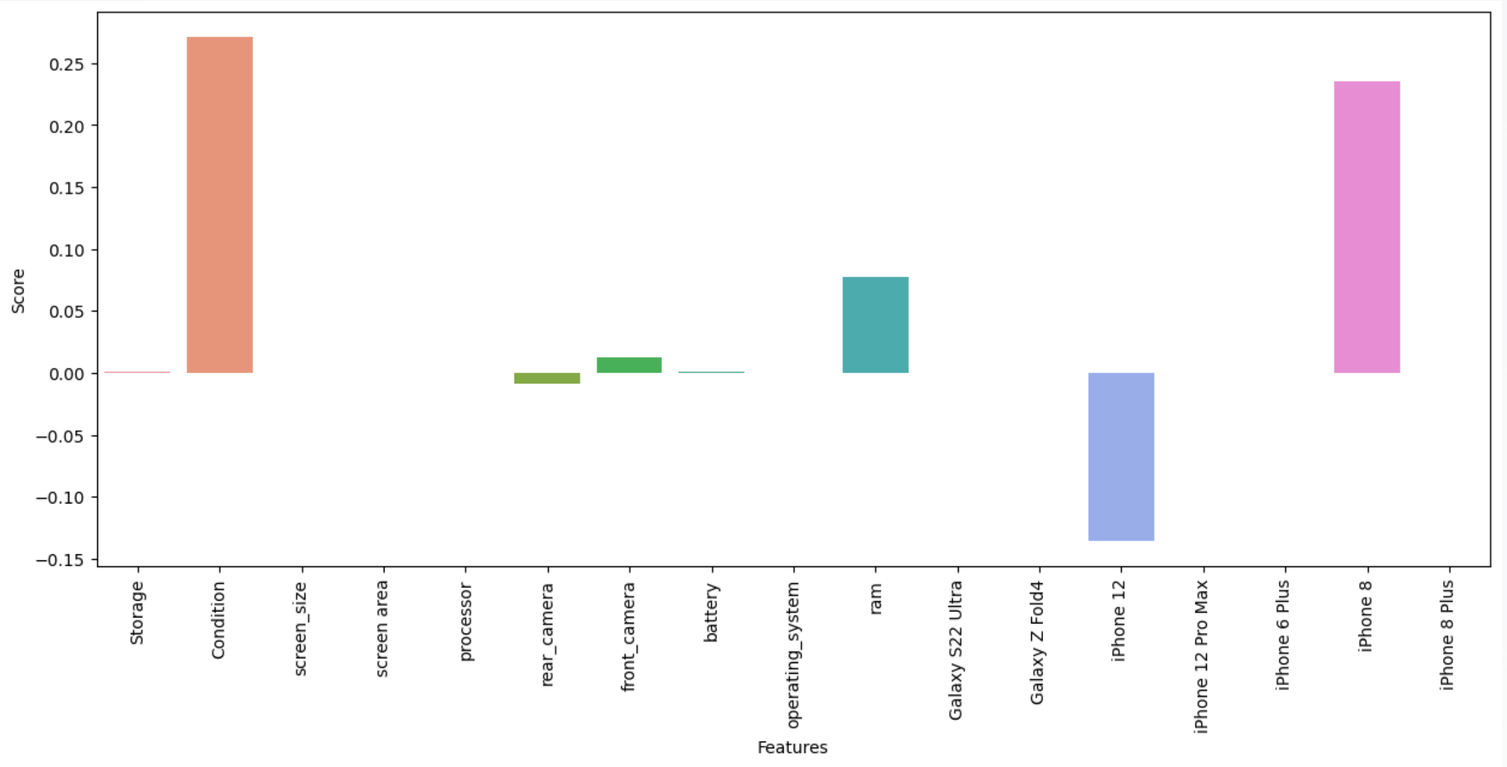
```
r2_Score of Ridge(alpha=0.6866488450043002) model on Training Data is: 82.48266273685213
r2_Score of Ridge(alpha=0.6866488450043002) model on Testing Data is: 80.63860961236784
MSE of Ridge(alpha=0.6866488450043002) model on Testing Data is: 0.06314248323203182
```



5.3. LASSO REGRESSOR

Lasso regression can be used to improve the accuracy of linear regression models when there is multicollinearity and to identify the most important independent variables.

```
r2_Score of Lasso(alpha=0.01) model on Training Data is: 81.02079867161308
r2_Score of Lasso(alpha=0.01) model on Testing Data is: 78.8036855306337
MSE of Lasso(alpha=0.01) model on Testing Data is: 0.06912664349858796
```



06 COMPARISON

We are use comparison with models to evaluate, improve, and understand algorithms.

```
df_model = pd.DataFrame.from_dict(Algorithms)
df_model
```

	Training Score	Testing Score	Algorithms
0	82.517366	80.723352	Linear Regression
1	82.482663	80.638610	Ridge Regression Model
2	81.020799	78.803686	Lasso Regresion Model

Thank's For Watching

Don't hesitate to ask any questions!