

Business Presentation

Contents

Business Problem Overview and Approach

Data Overview

EDA

Building linear Regression Model

Model Performance Evaluation

Checking Linear Regression Assumptions

Business Problem Overview and Solution Approach

Core business idea

Analyze the data and build a linear regression model to predict price of used_phone develop a dynamic pricing strategy.

Problem to tackle

- Identify the attributes that effect the used_price
- Identify deficiencies if any in the current target segmentation.

Financial implications

- Increase the used_phone sales

use ML model to solve the problem to set appropriate price for used_phones – maximizing sales

Data Overview

Variable	Description
brand_name:	Name of manufacturing brand
os:	OS on which the phone runs
screen_size:	Size of the screen in cm
4g:	Whether 4G is available or not
5g:	Whether 5G is available or not
main_camera_mp:	Resolution of the rear camera in megapixels
selfie_camera_mp:	Resolution of the front camera in megapixels
int_memory:	Amount of internal memory (ROM) in GB
ram:	Amount of RAM in GB
battery:	Energy capacity of the phone battery in mAh
weight:	Weight of the phone in grams
release_year:	Year when the phone model was released
days_used:	Number of days the used/refurbished phone has been used
new_price:	Price of a new phone of the same model in euros
used_price:	Price of the used/refurbished phone in euros

```
df.shape
```

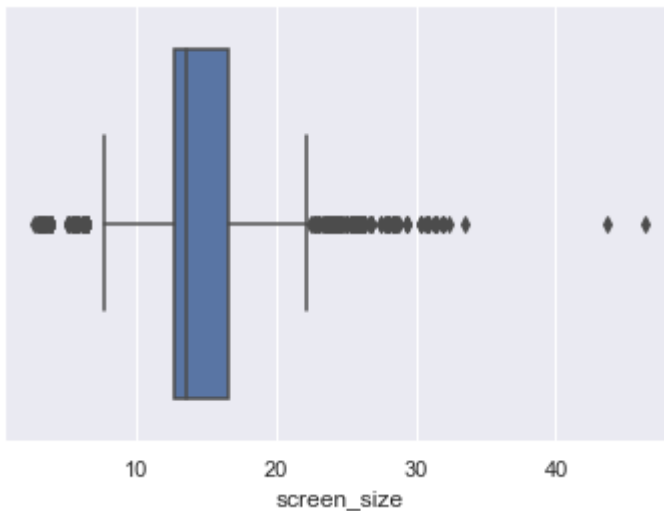
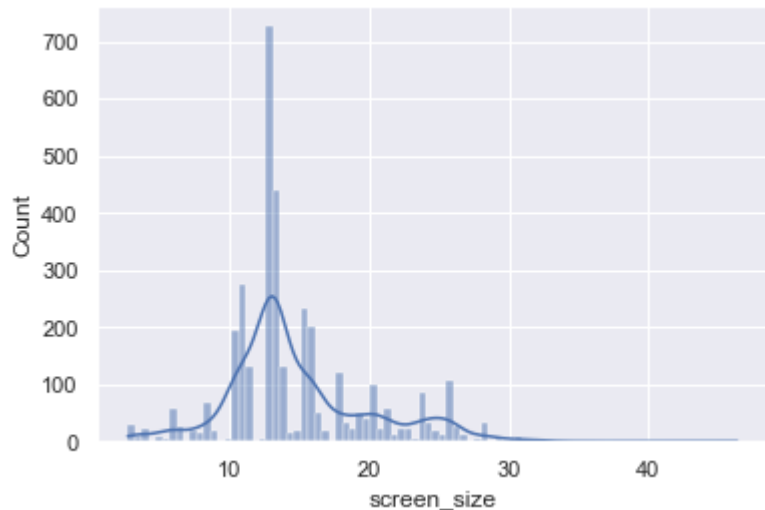
```
(3571, 15)
```

Observation

1. There are 3571 rows and 15 columns

- Brand_name, os, 4g and 5g are object variables. All Object variables should be converted to categorical variables.

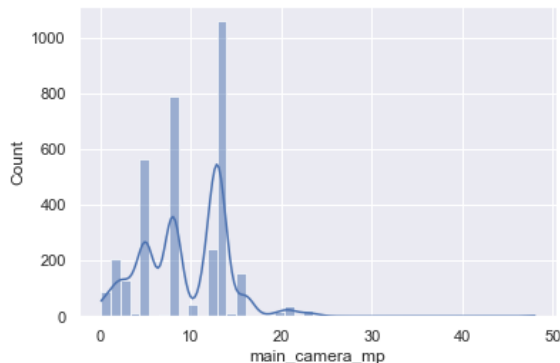
EDA - distribution of screen_size,



● Observations

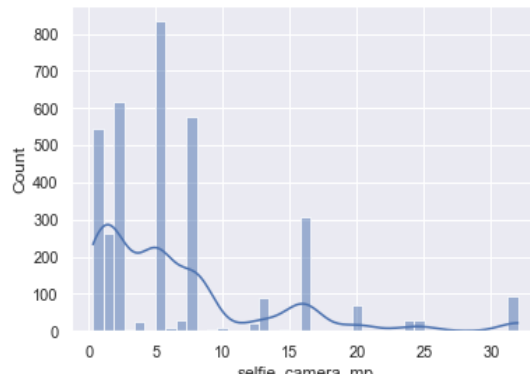
1. The distribution of screen_size is positively skewed.
2. The screen_size for most the used_phones fall between 10 and 20 cms.
3. Median is 14cm and mean is 13cm.
4. We have observatuons where the screen size is more than 40cm as well.

EDA – main_camera_mp, selfie_camera_mp, int_memory



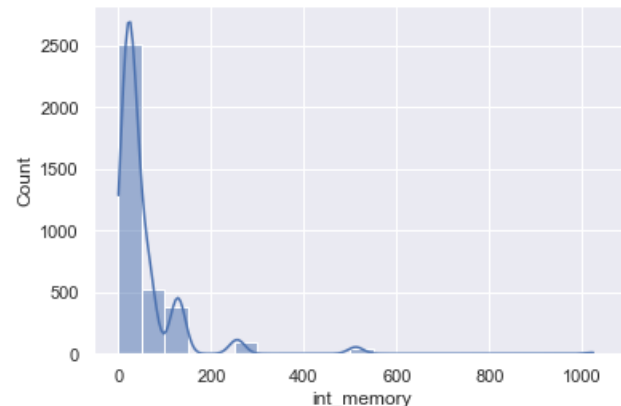
Observations

1. The resolution of main_camera_mp ranges from 0 to 48 megapixels.
2. Most of the used_phones have ~13 megapixels resolution.



Observations

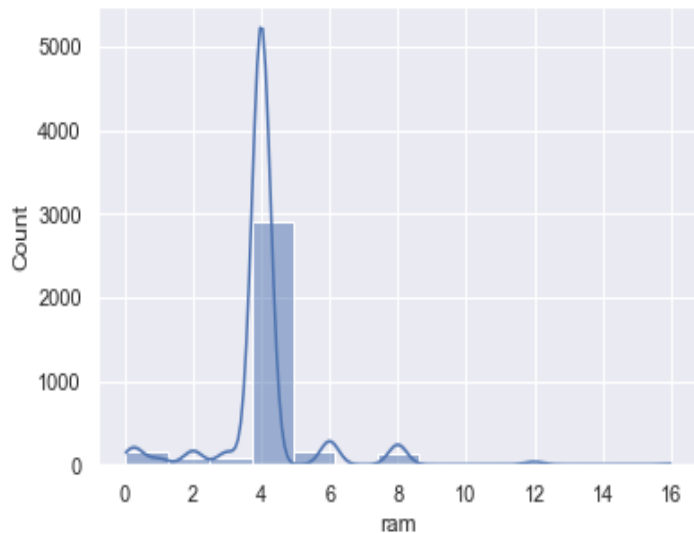
1. The distribution is positively skewed.
2. There are more used_phones with 5 megapixel resolution.
3. There are also some phones with 30 megapixel



Observations

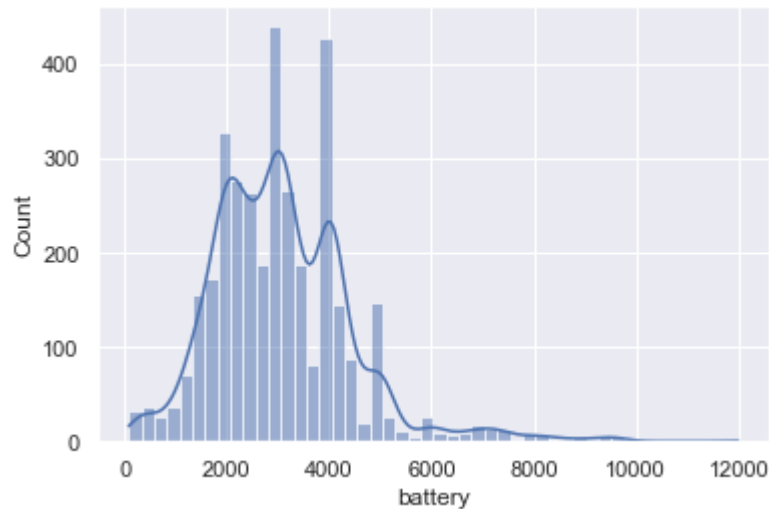
1. Most of the used_phones have int_memory between 0-1GB.
2. The distribution shows 3 outliers.
3. There are used_phones with 1024GB

EDA – RAM, Battery



Observations:

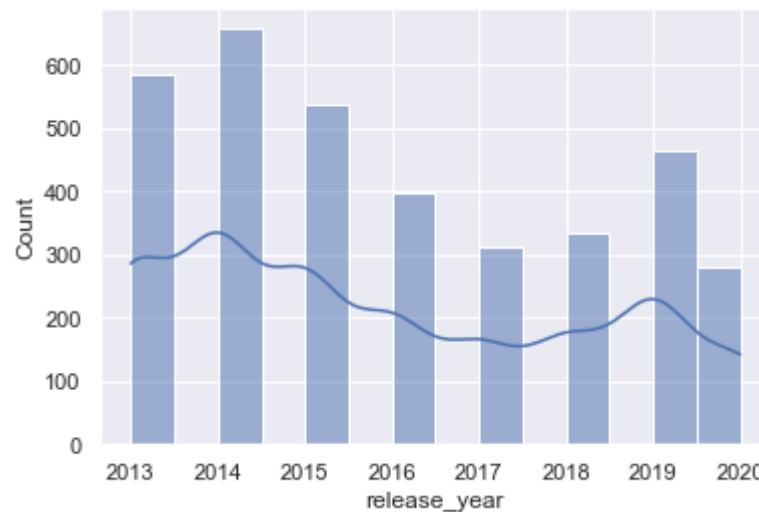
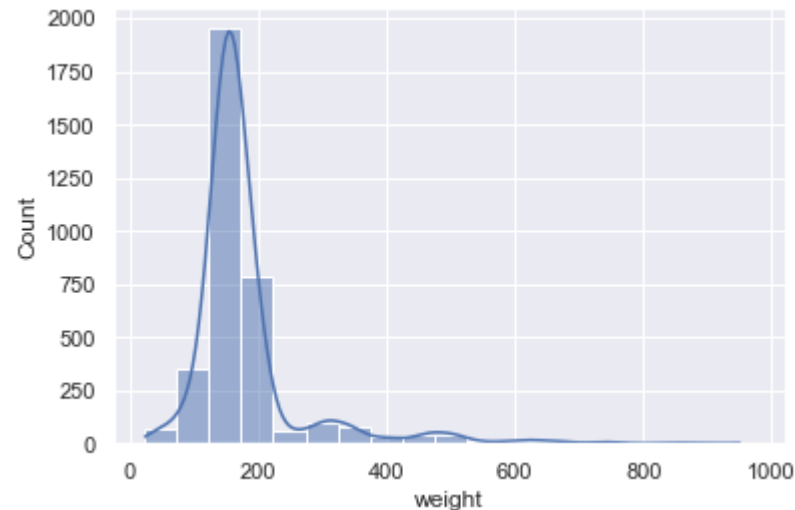
1. Maximum number of the used_phones have 4GB RAM.
2. The values other than 4GB are considered outliers.



Observations:

1. The Battery is right skewed.
2. There are many outliers.
3. There are many used_phones with ~3000mAh battery

EDA – Weight, Release_year,



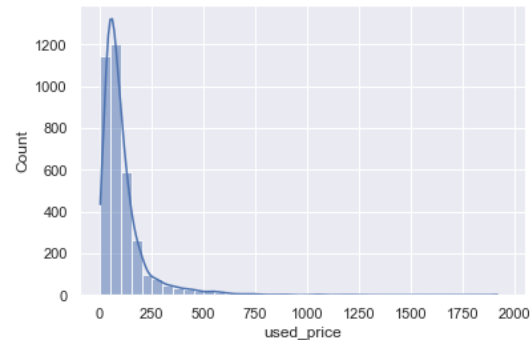
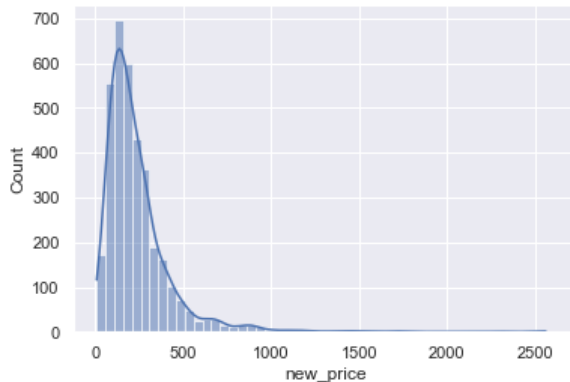
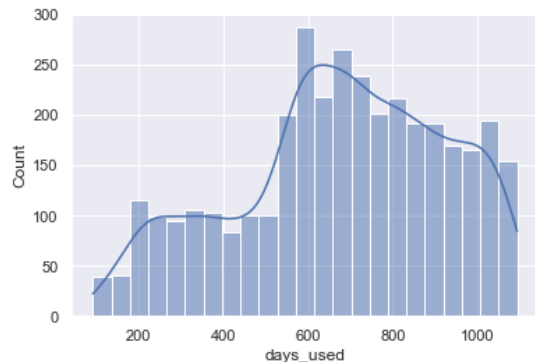
Observations

1. There are many phones whose weight is ~150 grams.
2. There are also some heavy phones(~900 grams)
3. There are some phones ~20 grams

Observations.

1. The data contains used phones released in the market between 2013 and 2020.
2. Maximum number of phones are released in 2014.

EDA- days_used,

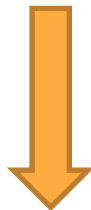


Observations

There are phones which are used for 600 - 1000 days

Observations.

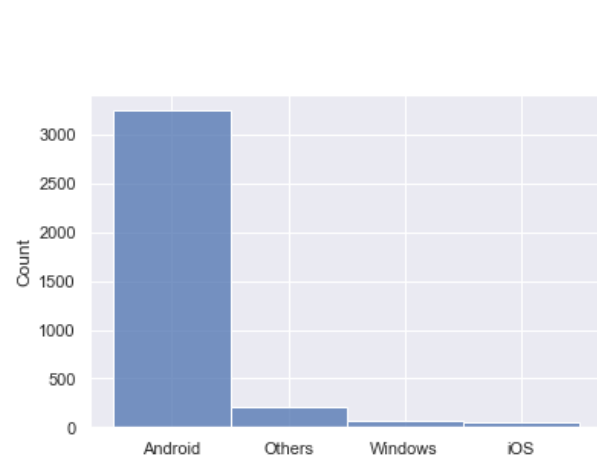
The graph is right skewed. There are many phones ~100 -150 Euros price range. There are also high priced phones



Observations

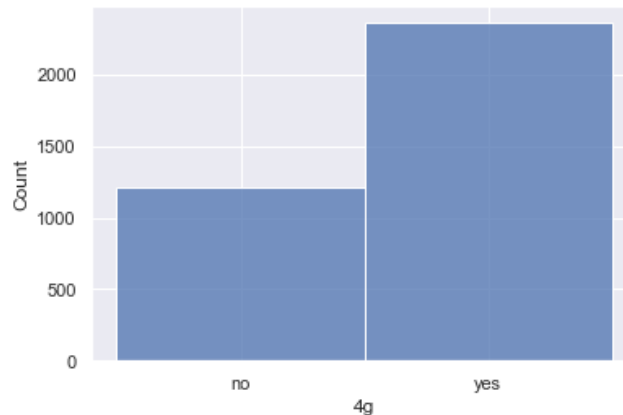
1. Maximum used_price is ~2000 Euros.
2. There are many phones used price is below 200 Euros.
3. There distribution is right skewed and there are many outliers.

EDA – Categorical Variables(OS, 4g,5g)



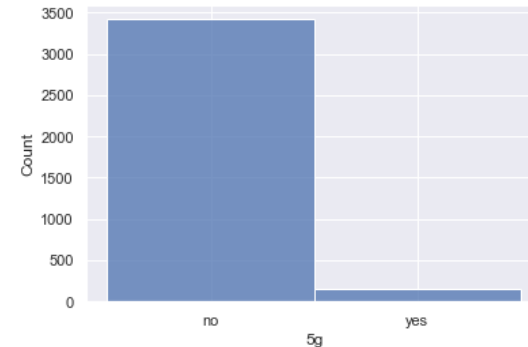
Observations

1. Many phones in the market are Android based.



Observations

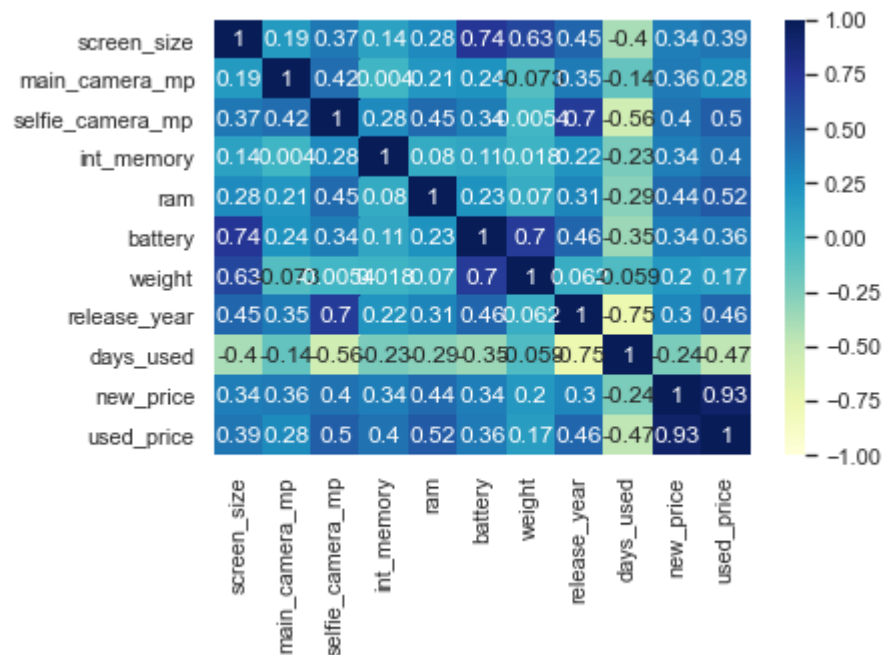
Many phones have 4g



Observation

Most of the used_phones donot have 5g

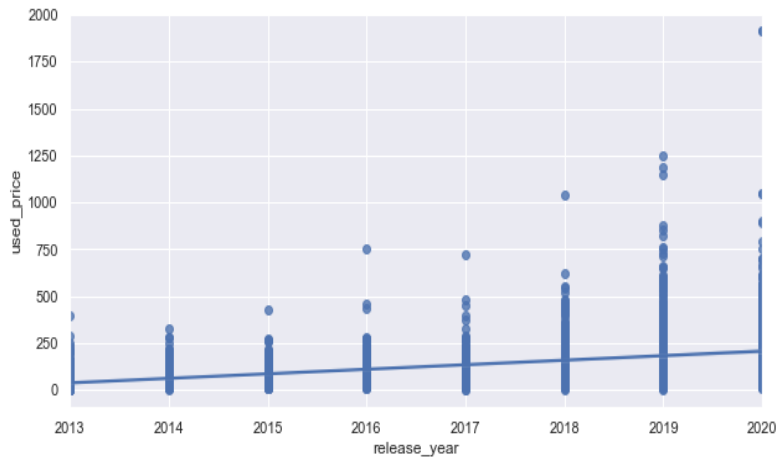
Bivariate Analysis



Observations

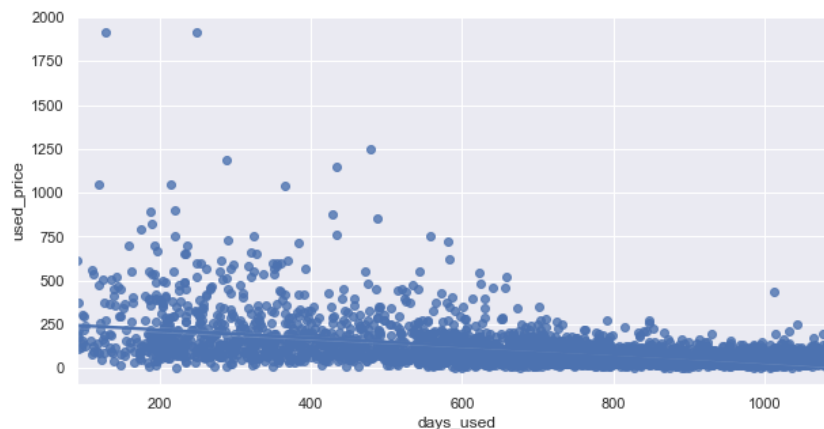
1. new_price is highly correlated with used_price with 0.93 correlation
2. ram and used_price have 0.52 correlation. selfie_camera_mp and used_price have 0.5 correlation

EDA - release_year and used_price, used_price and days_used



Observation

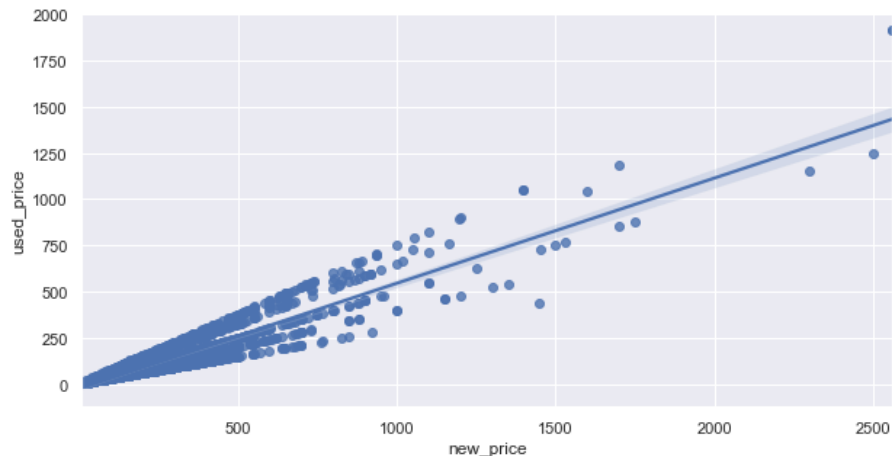
1. The latest phones have higher used_price



Observations

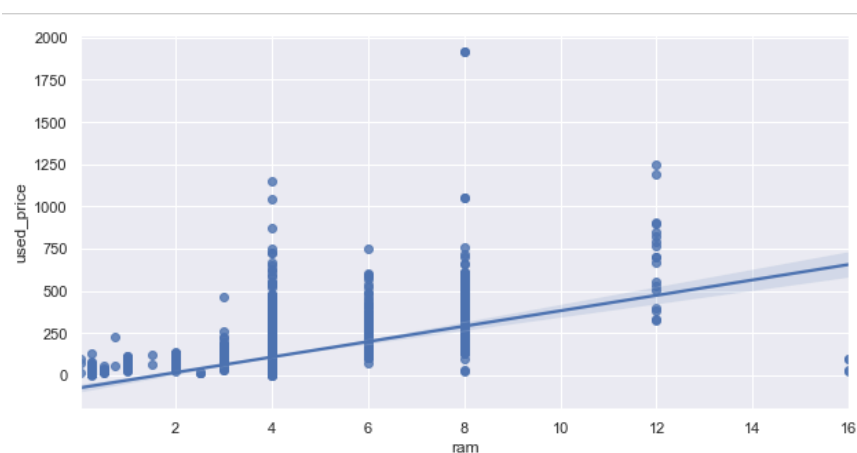
Most of the used_phones fall below 250 Euros price range

EDA - used_price and new_price, used_price and ram



Observations

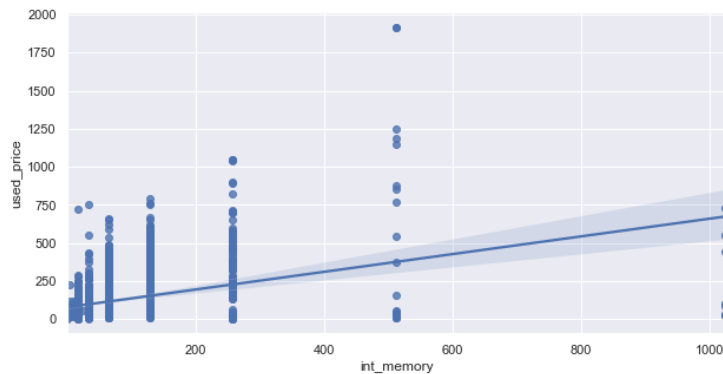
The used_price and new_price are linearly related.



Observations

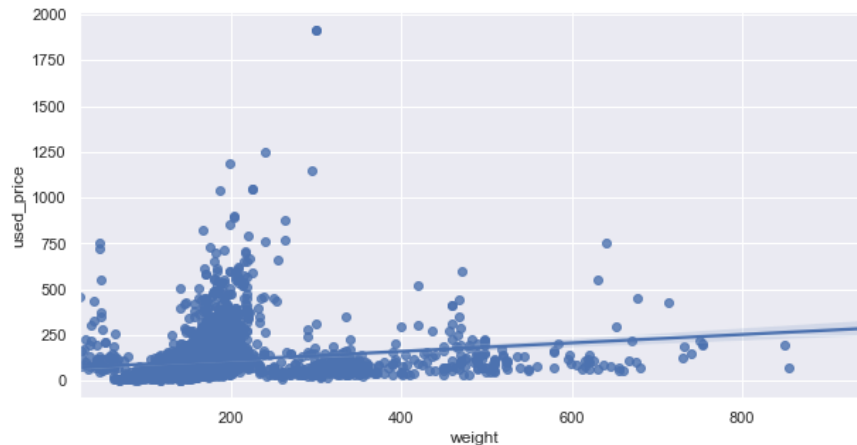
Most of the used_phones have Ram between 4-12GB

EDA - used_price and int_memory, used_price and weight



Observations

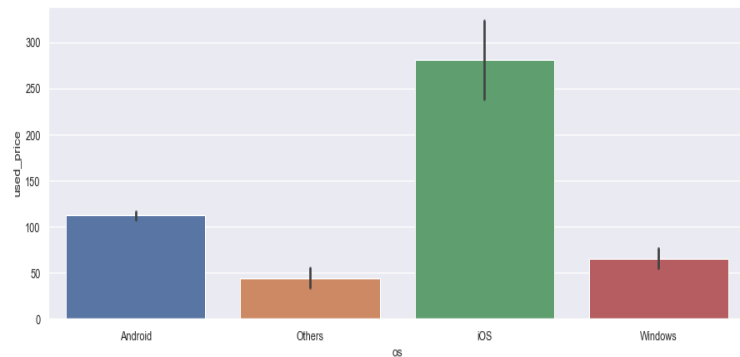
int_memory and used_price are linearly related.



Observations

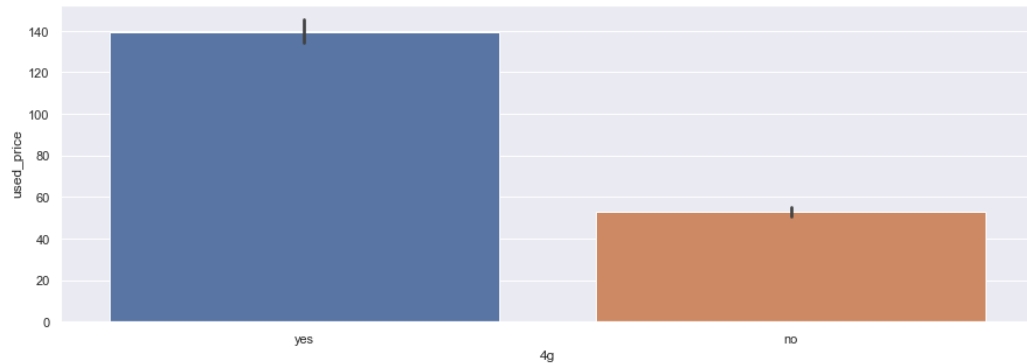
Most of the used_phones are below 200 grams and their price range is below 500 Euros.

EDA – Used_price vs os



Observations

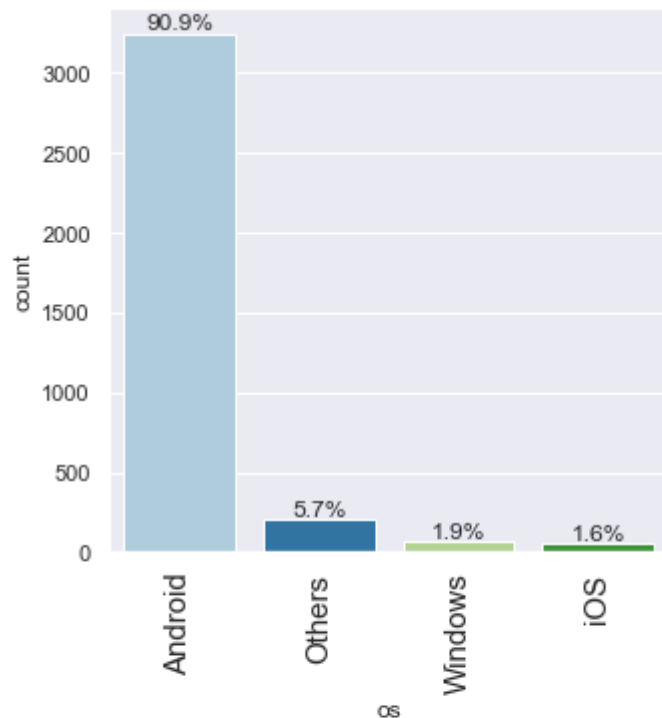
1. iOS based used_phones have high used_price.
2. Other os based and windows based used_phones have low used_price



Observations

1. The used_phones with 4g enables have high used_price

What percentage of the used phone market is dominated by Android devices?



Observation

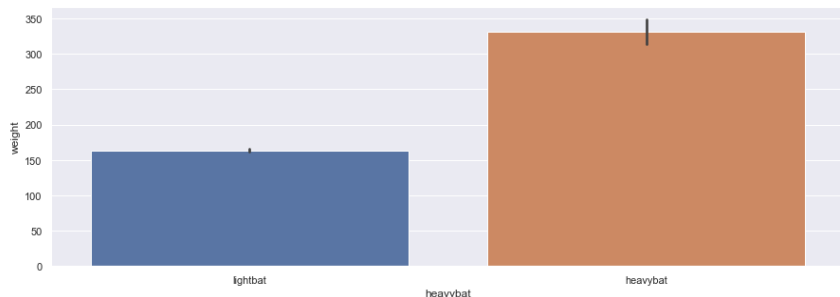
About 90.9% of the used_phones are dominated by Android devices

The amount of RAM is important for the smooth functioning of a phone. How does the amount of RAM vary with the brand?

Observation

1. RAM seems to be discrete in nature.
2. Most of the phones have 4GB RAM.
3. Vivo, Samsung, OnePlusOne, Lenovo, LG, Huawei, Honor brand phones have RAM more than 5GB
4. There seems to be very little relation or no relation between RAM and brand_name.

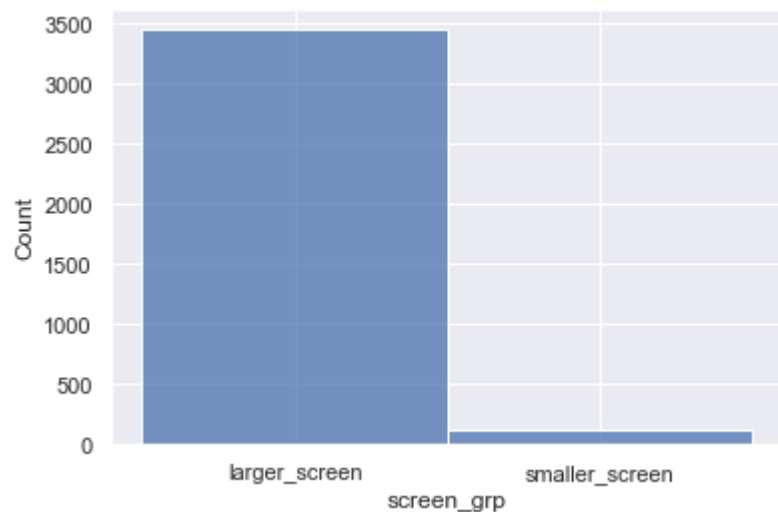
A large battery often increases a phone's weight, making it feel uncomfortable in the hands. How does the weight vary for phones offering large batteries (more than 4500 mAh)?



Observations

1. Heavy Battery phones seems to be heavy than the light battery phones.

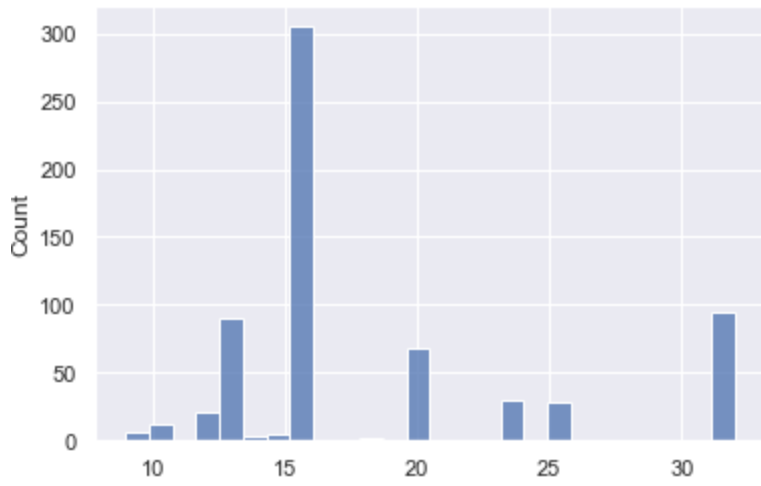
Bigger screens are desirable for entertainment purposes as they offer a better viewing experience. How many phones are available across different brands with a screen size larger than 6 inches?



Observations

There are 3450 used_phones that have scree_size greater than 6 inches.

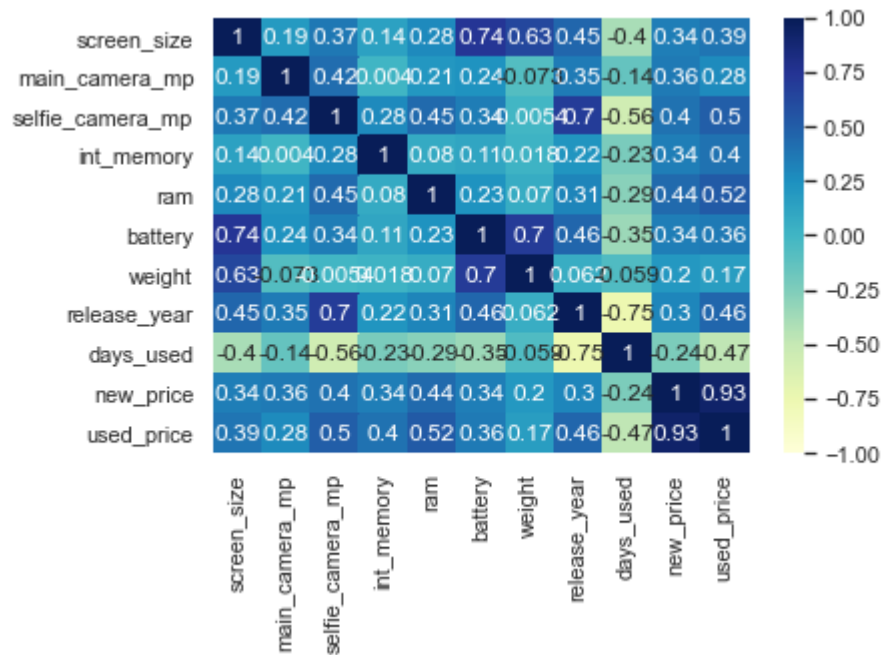
Budget phones nowadays offer great selfie cameras, allowing us to capture our favorite moments with loved ones. What is the distribution of budget phones offering greater than 8MP selfie cameras across brands?



Observations

1. The graph is right skewed.
2. There are many used_phones whose selfie_camera_mp value is 16Megapixel value.
3. There are also some ohne with 32megapixel selfie_camera_mp.

Which attributes are highly correlated with the used phone price?



Observations

1. new_price is highly correlated with used_price with 0.93 correlation
2. ram and used_price have 0.52 correlation. selfie_camera_mp and used_price have 0.5 correlation

Data Preprocessing

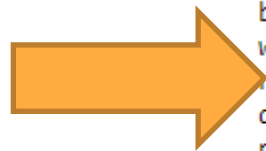
- This Process involves:

- Missing value treatment
- Feature engineering (if needed)
- Outlier detection and treatment (if needed)
- Preparing data for modeling
- Any other preprocessing steps

Missing Values

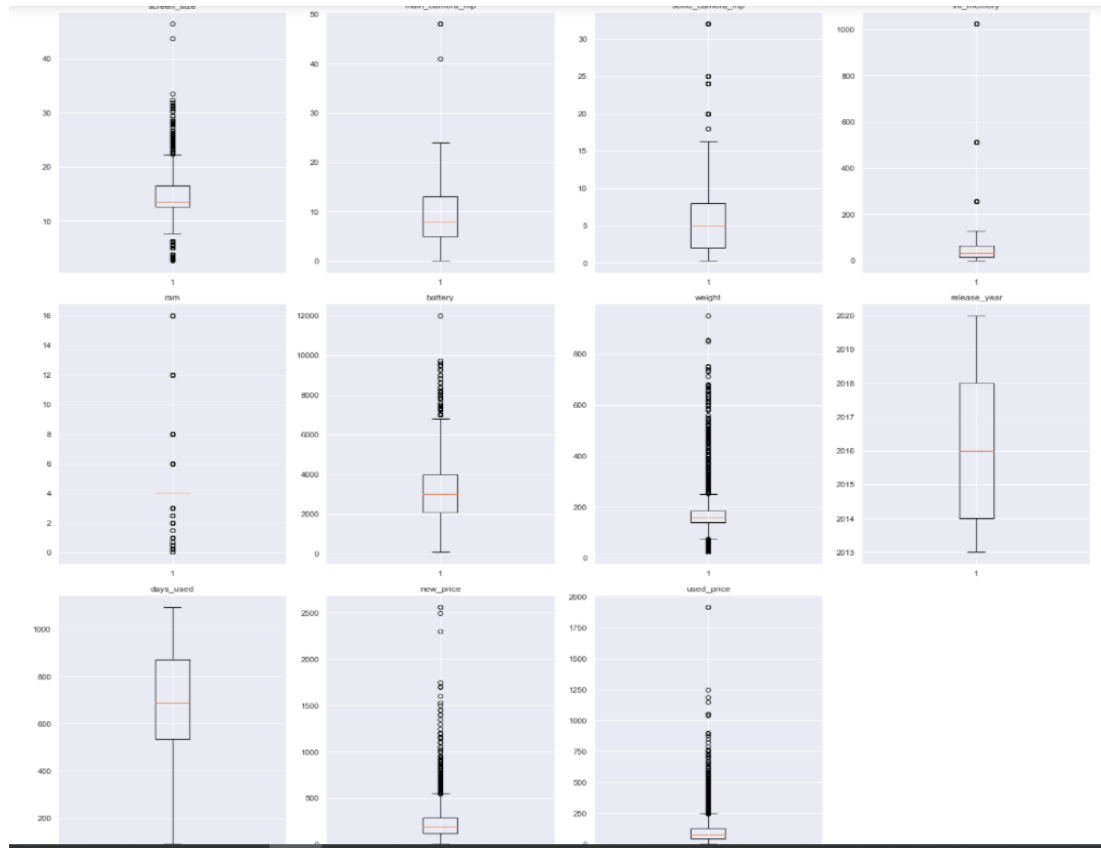
main_camera_mp	180
int_memory	10
ram	10
weight	7
battery	6
selfie_camera_mp	2
brand_name	0
os	0
screen_size	0
4g	0
5g	0
release_year	0
days_used	0
new_price	0
used_price	0
dtype:	int64

fill using
median values



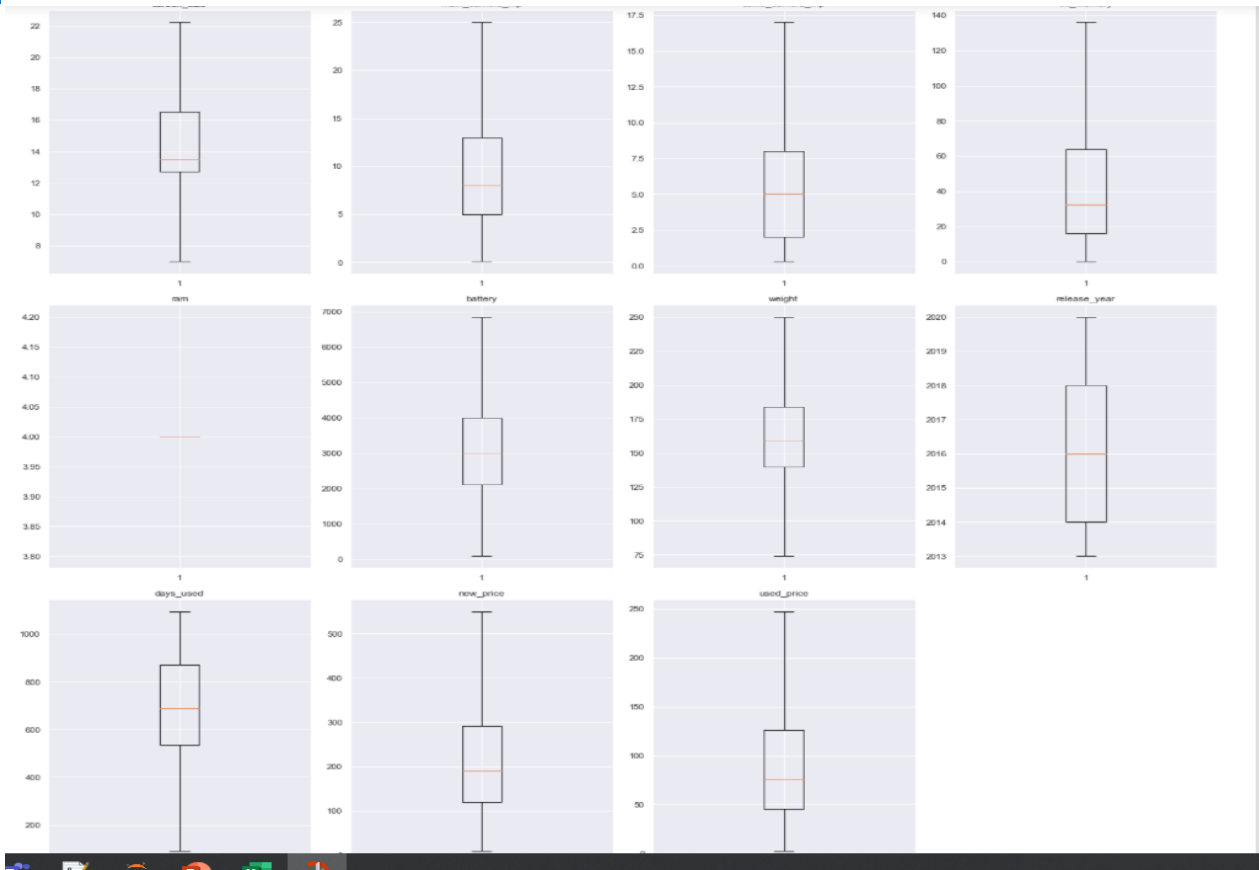
brand_name	0
os	0
screen_size	0
4g	0
5g	0
main_camera_mp	0
selfie_camera_mp	0
int_memory	0
ram	0
battery	0
weight	0
release_year	0
days_used	0
new_price	0
used_price	0
dtype:	int64

Outlier detection and treatment



1. There are lower outliers in screen_size and weight.
2. There are no outliers in days_used, release_year.
3. The other numerical columns have upper outliers.
4. We will treat these outliers as these might adversely affect the predictive power of linear model.

Outlier detection and treatment



Observations

1. Now, the outliers are all treated

Model Performance Summary

- Splitting the data into test and train data where test data is 30% and train data is 70%

- Sklearn linear_model**

Train set Performance

	RMSE	MAE	R-Squared	Adj. R-Squared	MAPE
0	13.980191	10.344831	0.955009	0.954718	19.115553

```
x_train.shape[0]
```

```
2499
```

```
x_test.shape[0]
```

```
1072
```

- Test set**

	RMSE	MAE	R-Squared	Adj. R-Squared	MAPE
0	13.617565	10.161106	0.95823	0.957597	16.589589

Observations

1. The training R^2 is 95.5%, indicating the model explained 95.5% of the variation in the train data. So, model is not underfitting.
2. MAE and RMSE are comparable indicating model is not underfitting.
3. MAE indicates that the current model is able to predict used_price within a mean error of 10.16 Euros on the test data.
4. MAPE on the test data suggests we can predict within 16.6% of the used_price.

Linear Regression using StatsModel

OLS Model Summary

```

OLS Regression Results
=====
Dep. Variable:      used_price      R-squared:      0.955
Model:              OLS              Adj. R-squared:  0.955
Method:             Least Squares    F-statistic:    3293.
Date:               Fri, 20 Aug 2021  Prob (F-statistic): 0.00
Time:               22:01:53         Log-Likelihood: -10137.
No. Observations:   2499            AIC:            2.031e+04
Df Residuals:       2482            BIC:            2.041e+04
Df Model:           16
Covariance Type:    nonrobust
=====
                    coef    std err          t      P>|t|      [0.025      0.975]
-----
const             -314.9593    540.308      -0.583    0.560    -1374.460     744.541
screen_size         0.0944      0.121      0.779    0.436     -0.143      0.332
main_camera_mp     -0.3094      0.087     -3.561    0.000     -0.480     -0.139
selfie_camera_mp    0.6163      0.100      6.165    0.000      0.420      0.812
int_memory          0.0949      0.010      9.445    0.000      0.075      0.115
int_memory          0.0949      0.010      9.445    0.000      0.075      0.115
battery            -0.0001      0.000     -0.305    0.760     -0.001      0.001
weight            -0.0052      0.012     -0.432    0.666     -0.029      0.018
release_year        0.1868      0.268      0.697    0.486     -0.339      0.712
days_used          -0.0854      0.002    -47.774    0.000     -0.089     -0.082
new_price           0.4056      0.004    106.127    0.000      0.398      0.413
os_Others           -4.7890      1.355     -3.533    0.000     -7.447     -2.131
os_Windows          -1.2333      2.066     -0.597    0.551     -5.285      2.819
os_iOS              8.6583      2.473      3.501    0.000      3.809     13.508
4g_yes              -3.7920      0.874     -4.338    0.000     -5.506     -2.078
5g_yes              2.7110      1.693      1.601    0.109     -0.609      6.030
brand_type_Medium   -14.3824      1.627     -8.842    0.000    -17.572    -11.193
brand_type_High     -26.3773      6.428     -4.104    0.000    -38.981    -13.773
=====
Omnibus:           227.169    Durbin-Watson:      1.978
Prob(Omnibus):     0.000    Jarque-Bera (JB):    491.543
Skew:              0.569    Prob(JB):            1.83e-107
Kurtosis:          4.851    Cond. No.            7.39e+06
=====

```

Observations

1. Negative value of coefficients show used_price decreases with increase of corresponding attribute value. Positive values of coefficients show used_price increases with increase of corresponding values
2. p-value of the variable indicates if the variable is significant or not. Lets consider significance level to be 5%, then anyvalue with a p-value less than 5% would be considered significant. But, these variables might contain multicollinearity which might affect the p-values.
3. Lets deal with the multicollinearity and verify other assumptions of linear regression followed by p-values

- **Checking Linear Regression Assumptions**

In order to make statistical inferences from a linear regression model, it is important to ensure that the assumptions of linear regression are satisfied.

1. **No Multicollinearity**
2. **Linearity of variables**
3. **Independence of error terms**
4. **Normality of error terms**
5. **No Heteroscedasticity**

Test for MultiCollinearity

- Observations**

1. The VIF of the attributes are mostly between 1-5
2. Looking at the VIF, there seem to be low multicollinearity between the attributes.
3. But, there are some attributes with $p\text{-value} > 0.05$. Hence, looping through the code and dropping all the columns with $p\text{-value} > 0.05$

	feature	VIF
0	const	3.707300e+06
1	screen_size	3.231782e+00
2	main_camera_mp	2.009739e+00
3	selfie_camera_mp	3.556360e+00
4	int_memory	1.934750e+00
5	battery	3.597409e+00
6	weight	3.070146e+00
7	release_year	4.818425e+00
8	days_used	2.548222e+00
9	new_price	3.377643e+00
10	os_Others	1.332737e+00
11	os_Windows	1.021507e+00
12	os_iOS	1.132857e+00
13	4g_yes	2.194265e+00
14	5g_yes	1.438091e+00
15	brand_type_Medium	2.025408e+00
16	brand_type_High	1.047651e+00

OLS Regression Results

```

=====
                    OLS Regression Results
=====
Dep. Variable:          used_price    R-squared (uncentered):          0.985
Model:                  OLS          Adj. R-squared (uncentered):        0.985
Method:                 Least Squares    F-statistic:                  1.517e+04
Date:                   Fri, 20 Aug 2021    Prob (F-statistic):          0.00
Time:                   22:02:06          Log-Likelihood:              -10140.
No. Observations:      2499             AIC:                        2.030e+04
Df Residuals:          2488             BIC:                        2.037e+04
Df Model:              11
Covariance Type:       nonrobust
=====
                    coef    std err          t      P>|t|      [0.025    0.975]
-----
main_camera_mp        -0.3342     0.083     -4.041     0.000     -0.496     -0.172
selfie_camera_mp       0.6697     0.088     7.621     0.000     0.497     0.842
int_memory             0.0974     0.010     9.781     0.000     0.078     0.117
release_year           0.0309     0.001    45.263     0.000     0.030     0.032
days_used             -0.0865     0.001   -59.610     0.000     -0.089     -0.084
days_used              -0.0865     0.001   -59.610     0.000     -0.089     -0.084
new_price              0.4059     0.003   118.789     0.000     0.399     0.413
os_others              -4.9837     1.253    -3.977     0.000     -7.441    -2.526
os_iOS                 8.6099     2.421     3.557     0.000     3.863    13.356
4g_yes                -3.6670     0.792    -4.629     0.000     -5.220    -2.114
brand_type_Medium     -14.0870     1.580    -8.918     0.000    -17.185   -10.989
brand_type_High       -25.5026     6.387    -3.993     0.000    -38.027   -12.978
=====
Omnibus:                230.585    Durbin-Watson:              1.976
Prob(Omnibus):           0.000    Jarque-Bera (JB):           492.220
Skew:                    0.581    Prob(JB):                   1.31e-107
Kurtosis:                4.838    Cond. No.                    4.87e+04
=====

```

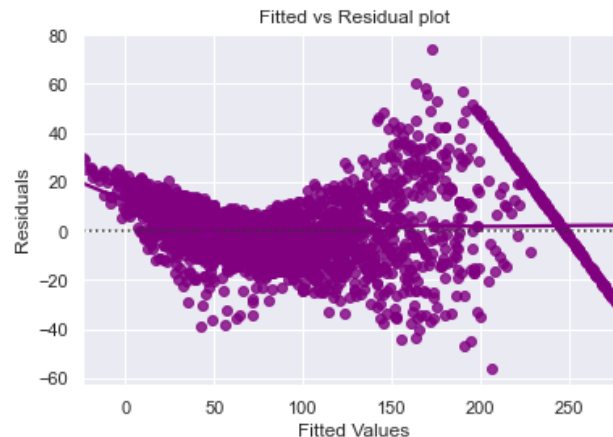
Notes:

- [1] R^2 is computed without centering (uncentered) since the model does not contain a constant.
- [2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
- [3] The condition number is large, 4.87e+04. This might indicate that there are strong multicollinearity or other numerical problems.

- Now no feature had p-value greater than 0.05. So, let's consider `x_train2` as the final data and `olsmod2` as final model
- **Observations**
 1. Now adjusted R-squared is 0.985 ie, the model is able to explain 98.5% of the variance.
 2. The adjusted R-squared of `olsmod0` was 0.955. This shows that the variables we dropped effected by 3%.

Test for Linearity and Interdependence

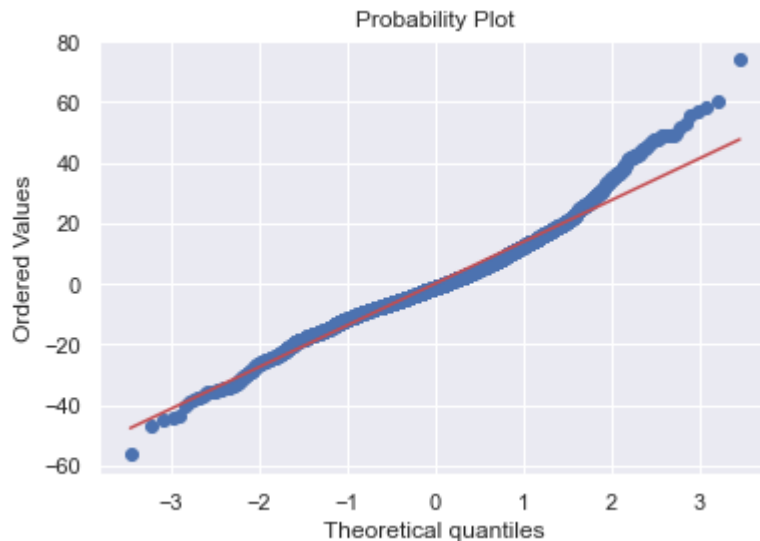
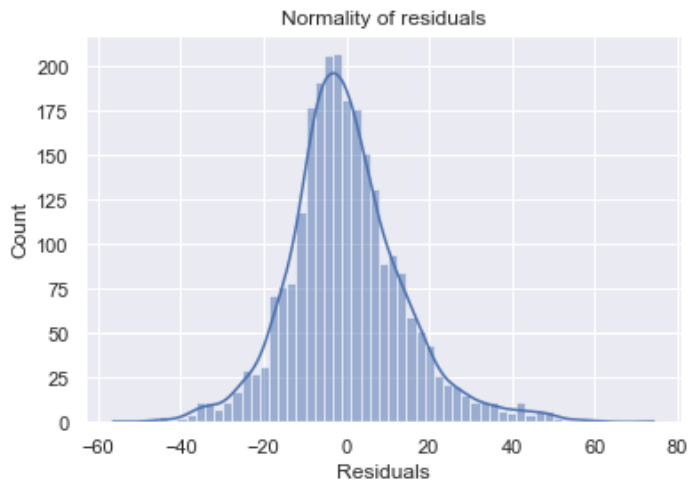
	Actual Values	Fitted Values	Residuals
844	100.48	102.074477	-1.594477
1539	111.68	118.123664	-6.443664
3452	113.89	110.904820	2.985180
1727	64.09	69.434766	-5.344766
1926	67.95	68.864293	-0.914293



- **Observation**

1. The above plot shows the distribution of residuals vs fitted values.
2. Since, we don't see any pattern, the assumptions of linearity and independence are satisfied.

Test for Normality



Observations

The histogram of residuals does have bell shape

Shapiro- Walk Test

```
ShapiroResult(statistic=0.9706025719642639, pvalue=3.120287435108585e-22)
```

- **Observation**

1. $p\text{-value} < 0.05$, the residuals are not as per Shapiro-Walk test
2. The residuals are not normal.
3. However, as an approximation, we accept the distribution as normal.
4. Hence, assumption is satisfied.

Test for HOMOSCEDASTICITY

- After performing goldfeldquant test:

```
[('F statistic', 1.0468462382983947), ('p-value', 0.2102941600456584)]
```

- We observe that the actual and the predicted values are comparable

	Actual	Predicted
2098	30.5200	21.695159
278	195.6700	192.070230
26	247.1925	229.477471
2910	89.9700	91.721518
2631	69.2000	64.044985
1582	89.5800	109.812760
2110	247.1925	264.449155
3160	65.3400	65.208595
2817	115.7700	106.862134
549	39.2900	47.590057



Checking OLS Model Again

Training Performance

	RMSE	MAE	R-Squared	Adj. R-Squared	MAPE
0	13.992121	10.363142	0.954932	0.954732	19.24802

Test Performance

	RMSE	MAE	R-Squared	Adj. R-Squared	MAPE
0	13.617565	10.161106	0.95823	0.957557	16.589589

- **Observations**

1. The model is able to see 95% of the variation in the data
2. The MAPE on the test set suggests we predict 16.58% of the used_price.
3. Hence, we can conclude that olsmod2 is a good model.

Comparing initial sklearn and final statsmodel

Training performance comparison:

	Linear Regression sklearn	Linear Regression statsmodels
RMSE	13.980191	13.992121
MAE	10.344831	10.363142
R-Squared	0.955009	0.954932
Adj. R-Squared	0.954718	0.954732
MAPE	19.115553	19.248020

Test performance comparison:

	Linear Regression sklearn	Linear Regression statsmodels
RMSE	13.617565	13.617565
MAE	10.161106	10.161106
R-Squared	0.958230	0.958230
Adj. R-Squared	0.957597	0.957557
MAPE	16.589589	16.589589

- **Observations**
- The performance of both the models seem to be close to each other

Final Summary Model

OLS Regression Results						
Dep. Variable:	used_price	R-squared (uncentered):	0.985			
Model:	OLS	Adj. R-squared (uncentered):	0.985			
Method:	Least Squares	F-statistic:	1.517e+04			
Date:	Fri, 20 Aug 2021	Prob (F-statistic):	0.00			
Time:	22:03:25	Log-Likelihood:	-10140.			
No. Observations:	2499	AIC:	2.030e+04			
Df Residuals:	2488	BIC:	2.037e+04			
Df Model:	11					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
main_camera_mp	-0.3342	0.083	-4.041	0.000	-0.496	-0.172
selfie_camera_mp	0.6697	0.088	7.621	0.000	0.497	0.842
int_memory	0.0974	0.010	9.781	0.000	0.078	0.117
release_year	0.0309	0.001	45.263	0.000	0.030	0.032
days_used	-0.0865	0.001	-59.610	0.000	-0.089	-0.084
new_price	0.4059	0.003	118.789	0.000	0.399	0.413
os_Others	-4.9837	1.253	-3.977	0.000	-7.441	-2.526
os_iOS	8.6099	2.421	3.557	0.000	3.863	13.356
4g_yes	-3.6670	0.792	-4.629	0.000	-5.220	-2.114
brand_type_Medium	-14.0870	1.580	-8.918	0.000	-17.185	-10.989
brand_type_High	-25.5026	6.387	-3.993	0.000	-38.027	-12.978
Omnibus:	230.585	Durbin-Watson:	1.976			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	492.220			
Skew:	0.581	Prob(JB):	1.31e-107			
Kurtosis:	4.838	Cond. No.	4.87e+04			

Notes:

- [1] R^2 is computed without centering (uncentered) since the model does not contain a constant.
- [2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
- [3] The condition number is large, 4.87e+04. This might indicate that there are strong multicollinearity or other numerical problems.

Business Insights and Recommendations

Selfie camera resolution, internal memory, release year, new price have a positive effect on Used price.

- Every unit increase in selfie camera resolution results in 0.67 times increase in used price.
- A unit increase in new price effects a 0.4 units increase in used price.
- An iOS device has 8.6 units increase in used price.
- Every unit increase in internal memory results in 0.1 units increase in used price.
- A unit increase in release year causes a 0.03 unit increase in used price.

Business Insights and Recommendations

Main camera resolution, days used, Operating system other than Android, iOS, Windows, 4g capability and brand type of medium/high has a negative effect on Used price.

- Every unit increase in Main camera resolution results in 0.33 units decrease in used price.
- Every unit increase in days used causes a 0.08 unit decrease in used price.
- A non iOS/Android/Windows device has 5 units decrease in used price.
- A 4g enabled device has 3.66 units less price.
- Brands of type Medium (I.e with New price between 500 and 1500) have 14 units decrease in used price.
- Brands of type High (I.e with New price between 1500 and 3000) have 26 units decrease in used price.

Business Insights and Recommendations

To Maximize revenue, concentrate on

- ❖ Phones that have better selfie camera resolution
- ❖ Phones that have higher new price.
- ❖ iOS phones
- ❖ Phones with higher internal memory.
- ❖ More recent phones

Following categories may not fetch much revenues:

- Non iOS/Android/Windows phones
- Phones with only high Main camera resolution.
- Phones with new prices between 500 and 3000

Business Insights and Recommendations

Additionally,

- Given there are very less iOS phones, higher used prices can be set for such phones. But they may cater only to a niche segment.
- Even though Non iOS/Android/Windows phones are less, their predicted prices are also less. Hence don't invest in sales of these models.
- Most of the used phones fall below 250 Euros price range. Hence this will be sector to concentrate for mass sales.
- Higher prices can be set for phones with higher selfie camera resolution which seems to be the latest trend.
- Some incentives / discounts can be given on phones with just higher main camera resolution as prices are not inclined towards such phones.

greatlearning
Power Ahead

Happy Learning !

