ReCell

Data Analysis Observations & Recommendation

John D Noble

October 2021

Business Problem Overview & Solution Approach

ReCell wants to launch a ML-based point of sale (web and mobile) solution that will use a real-time and dynamic pricing strategy for used and refurbished smartphones.

Key Questions We Will Answer!

1. *Does used phone price have a positive or negative correlation with the different factors (new prices, memory, size of selfie camera?*
2. *Can we build a linear model to predict used phone price? If yes, how accurate will the model be?*
3. *What are the predictive variables actually affecting used phone price?*

What are we trying to solve for?

How best to take advantage of the cheaper refurbished smartphone segment, as consumers cut back on discretionary spending and buy phones only for immediate needs.

Financial Implications

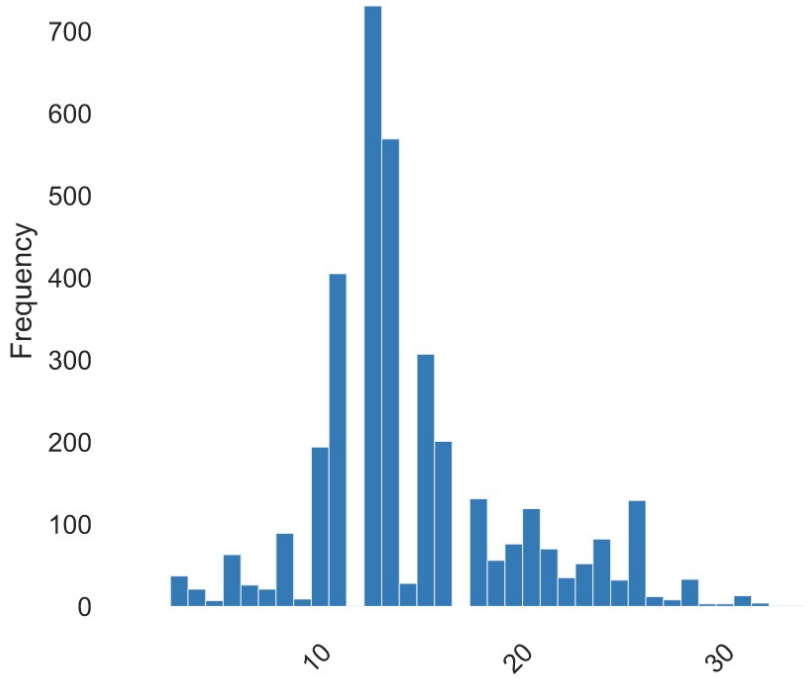
This is an entirely new, post-pandemic market opportunity. No one else is in the space and if we can quickly bring to market an attractive product backed by the latest ML that can grow a new subscriber segment to improve shareholder results by driving both bottom line revenue.

Exploratory Data Analysis (EDA): General Observations

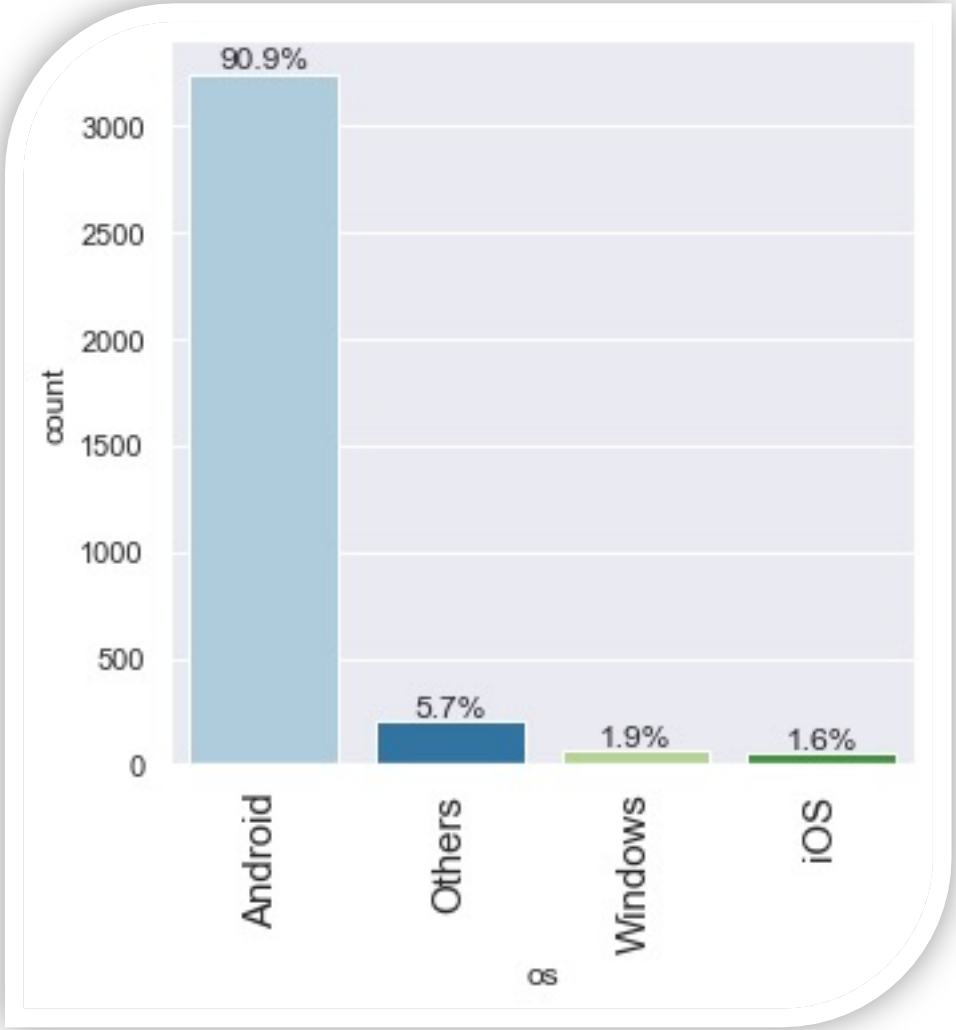
- There are 34 unique brands
- There are 4 different OS versions
- Years range from 2013 to 2020
- Avg screen size is ~15cm
- The avg main camera is ~9MP
- The avg selfie camera is ~6.5mp
- The avg ram is ~4 avg batterty capacity is ~3067 in mah
- The avg phone weight is ~179 grams
- The new price range is ~9 to 2560 (?) ;
- Avg new price is ~237 compared to a used price if ~109
- The used price range is ~2.51 *to* 1916; avg used price is ~$109 most of the used phones are 4g, not 5g

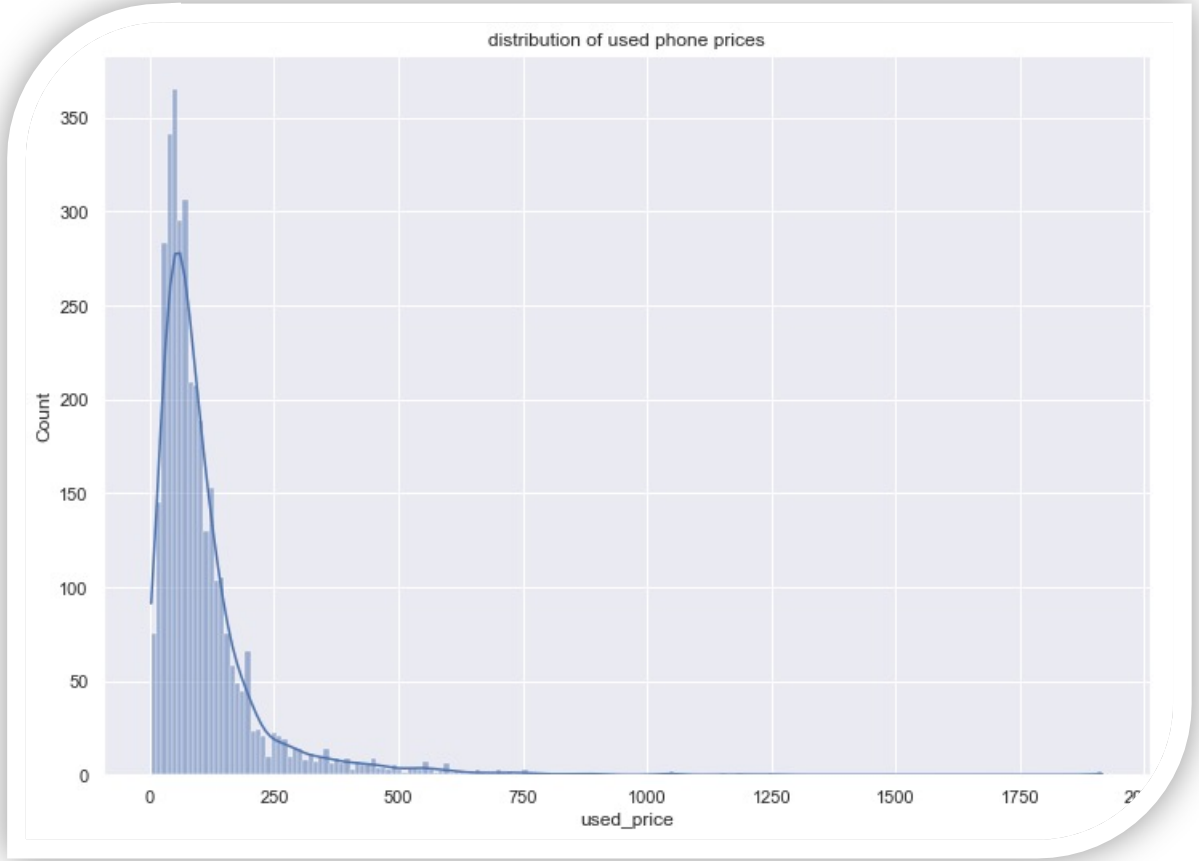
Note: Prices reflect Euro's, as this model is intended for Europe, Africa, and potentially Asia.

***However** … the data suggest we might have some tablet's in this base on the screen size that should be consider as part of another model.*



Exploratory Data Analysis (EDA)

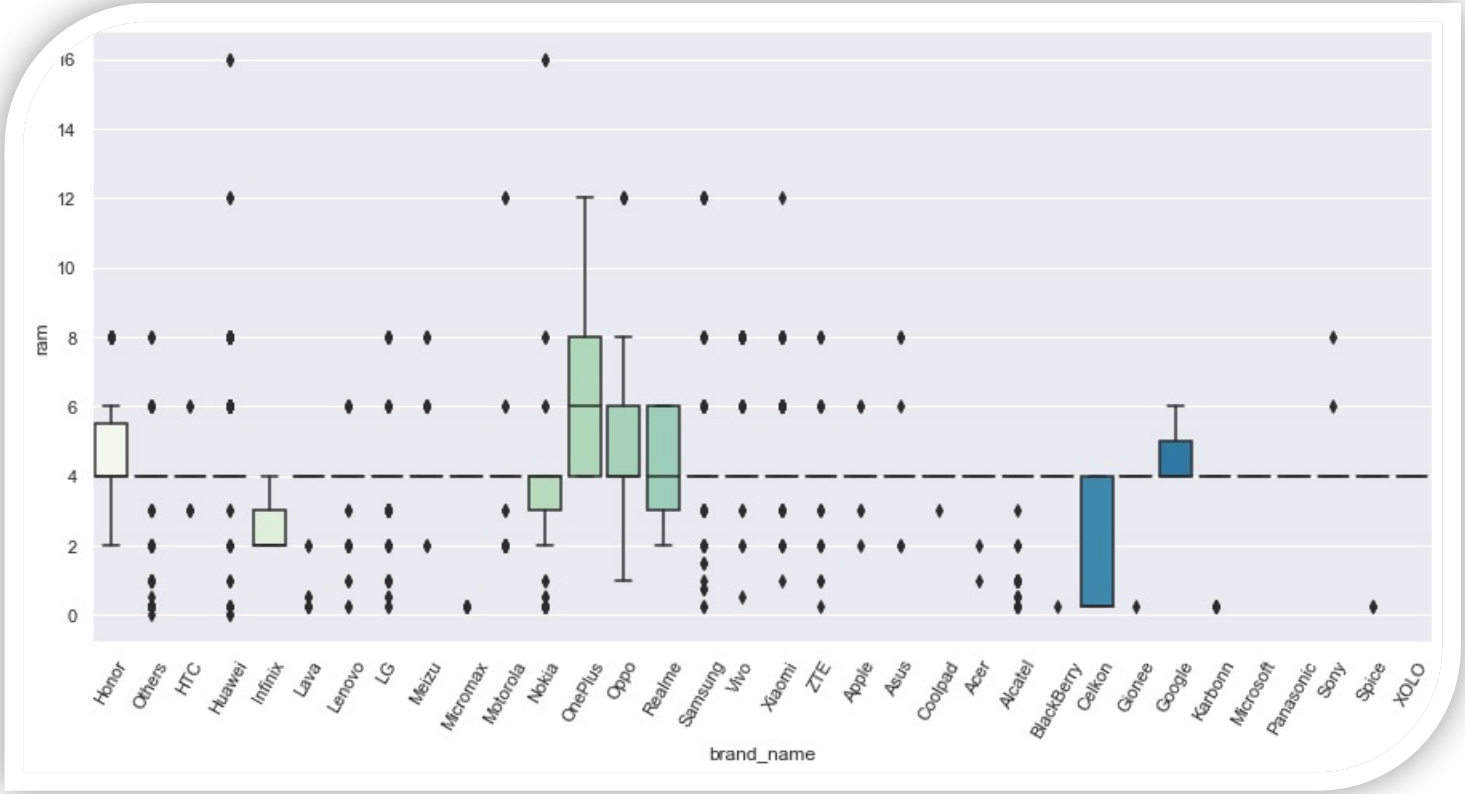




Android phones dominate the used market for phones <250.

Exploratory Data Analysis (EDA)

There are large # of entry level phones < 4GB RAM, but as you can see each maker has additional RAM offerings



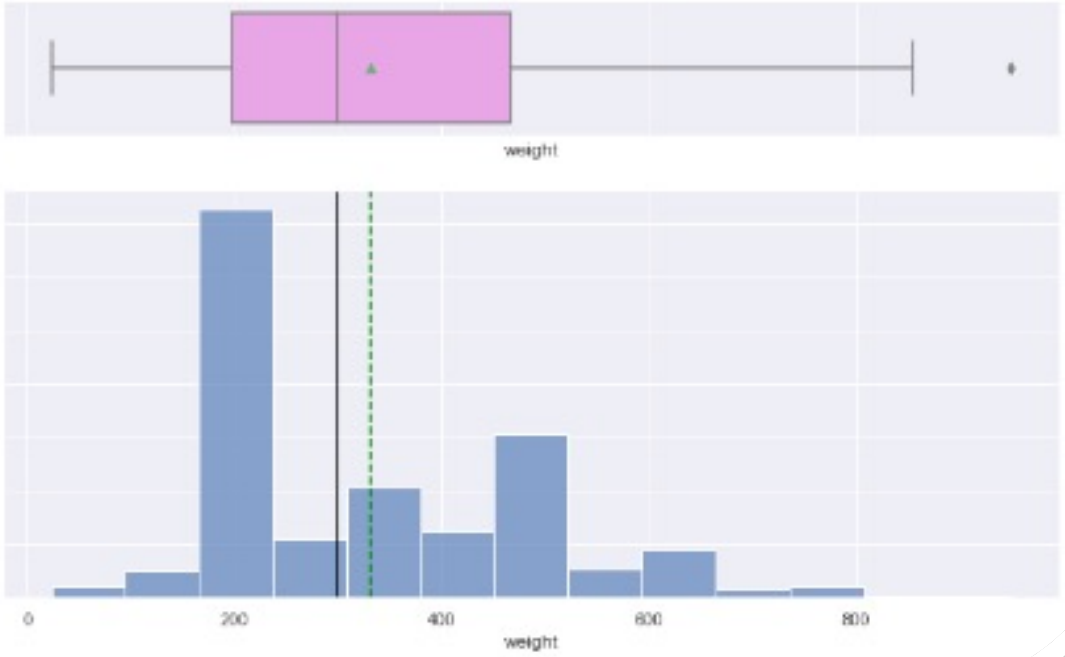
brand_name	amax	amin	len	mean
Acer	4.000	1.000	51.000	3.902
Alcatel	4.000	0.250	125.000	3.426
Apple	6.000	2.000	59.000	4.000
Asus	8.000	2.000	126.000	4.048
BlackBerry	4.000	0.250	22.000	3.830
Celkon	4.000	0.250	37.000	1.466
Coolpad	4.000	3.000	22.000	3.955
Gionee	4.000	0.250	56.000	3.933
Google	6.000	4.000	15.000	4.533
HTC	6.000	3.000	110.000	4.000
Honor	8.000	2.000	118.000	4.593
Huawei	16.000	0.030	264.000	4.641
Infinix	4.000	2.000	10.000	2.600
Karbonn	4.000	0.250	30.000	3.375
LG	8.000	0.250	212.000	3.894
Lava	4.000	0.250	36.000	3.278
Lenovo	6.000	0.250	172.000	3.887
Meizu	8.000	2.000	62.000	4.452
Micromax	4.000	0.250	120.000	3.750
Microsoft	4.000	4.000	22.000	4.000
Motorola	12.000	2.000	110.000	3.945
Nokia	16.000	0.250	121.000	3.601
OnePlus	12.000	4.000	22.000	6.364
Oppo	12.000	1.000	129.000	4.961
Others	8.000	0.030	509.000	3.751
Panasonic	4.000	4.000	47.000	4.000
Realme	6.000	2.000	41.000	4.195
Samsung	12.000	0.250	364.000	4.159
Sony	8.000	4.000	88.000	4.068
Spice	4.000	0.250	30.000	3.750
Vivo	8.000	0.500	117.000	4.756
XOLO	4.000	4.000	49.000	4.000
Xiaomi	12.000	1.000	134.000	4.567
ZTE	8.000	0.250	141.000	4.023

Exploratory Data Analysis (EDA)

1. The overall realtionship , as the battery mAH increase, so does the weight of the phone.
2. There are 346 phones that have batteries > than 4500 mAh
3. For phones offering large batteries > than 4500 mAh ... as measured in grams
    1. mean is 330.72
    2. min is 23
    3. 25% is 198
    4. 50% is 299
    5. 75% is 467
    6. Max is 950

	count	mean	std	min	25%	50%	75%	max
screen_size	346.000	21.487	5.170	10.160	16.350	20.960	25.560	46.360
main_camera_mp	292.000	9.186	4.900	0.300	5.000	8.000	13.000	48.000
selfie_camera_mp	346.000	7.206	6.254	0.300	2.000	5.000	8.000	32.000
int_memory	346.000	59.145	86.464	16.000	16.000	32.000	64.000	1024.000
ram	346.000	4.246	1.467	1.000	4.000	4.000	4.000	12.000
battery	346.000	5884.103	1327.656	4520.000	5000.000	5100.000	6890.000	12000.000
weight	346.000	330.719	160.953	23.000	198.000	299.000	467.000	950.000
release_year	346.000	2017.113	2.456	2013.000	2015.000	2017.500	2019.000	2020.000
days_used	346.000	568.760	278.010	92.000	319.250	581.500	778.750	1089.000
new_price	346.000	307.462	200.482	80.820	180.143	250.265	351.098	1200.850
used_price	346.000	155.901	134.678	33.090	79.958	111.460	167.787	901.270

```
1 histogram_boxplot(df_g, "weight")
```

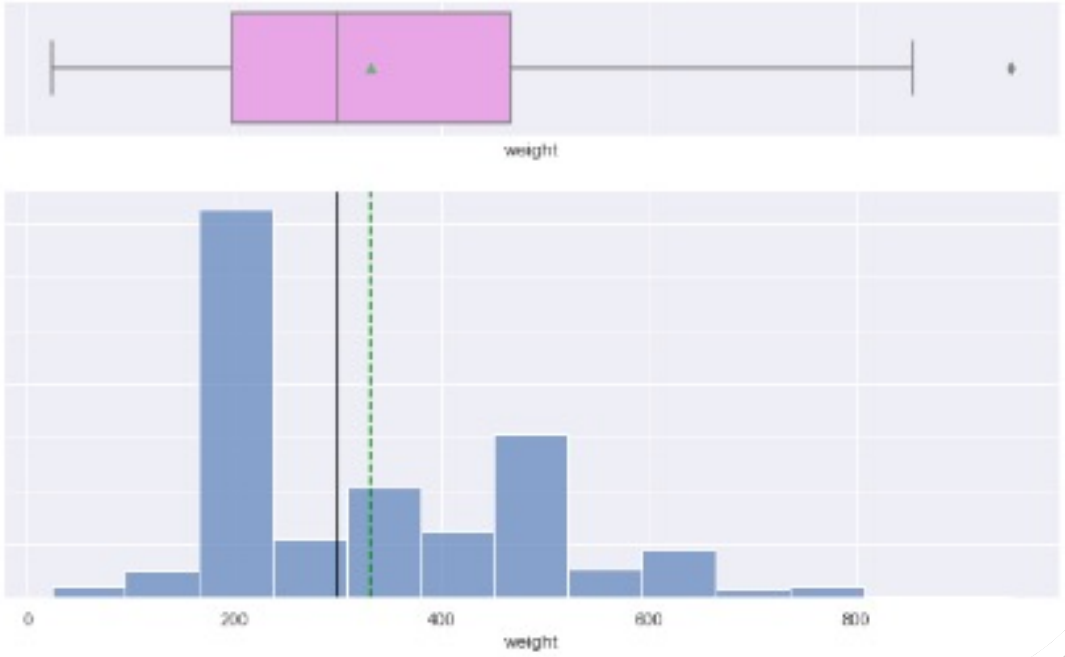


Exploratory Data Analysis (EDA)

1. The overall realtionship , as the battery mAH increase, so does the weight of the phone.
2. There are 346 phones that have batteries > than 4500 mAh
3. For phones offering large batteries > than 4500 mAh ... as measured in grams
    1. mean is 330.72
    2. min is 23
    3. 25% is 198
    4. 50% is 299
    5. 75% is 467
    6. Max is 950

	count	mean	std	min	25%	50%	75%	max
screen_size	346.000	21.487	5.170	10.160	16.350	20.960	25.560	46.360
main_camera_mp	292.000	9.186	4.900	0.300	5.000	8.000	13.000	48.000
selfie_camera_mp	346.000	7.206	6.254	0.300	2.000	5.000	8.000	32.000
int_memory	346.000	59.145	86.464	16.000	16.000	32.000	64.000	1024.000
ram	346.000	4.246	1.467	1.000	4.000	4.000	4.000	12.000
battery	346.000	5884.103	1327.656	4520.000	5000.000	5100.000	6890.000	12000.000
weight	346.000	330.719	160.953	23.000	198.000	299.000	467.000	950.000
release_year	346.000	2017.113	2.456	2013.000	2015.000	2017.500	2019.000	2020.000
days_used	346.000	568.760	278.010	92.000	319.250	581.500	778.750	1089.000
new_price	346.000	307.462	200.482	80.820	180.143	250.265	351.098	1200.850
used_price	346.000	155.901	134.678	33.090	79.958	111.460	167.787	901.270

```
1 histogram_boxplot(df_g, "weight")
```



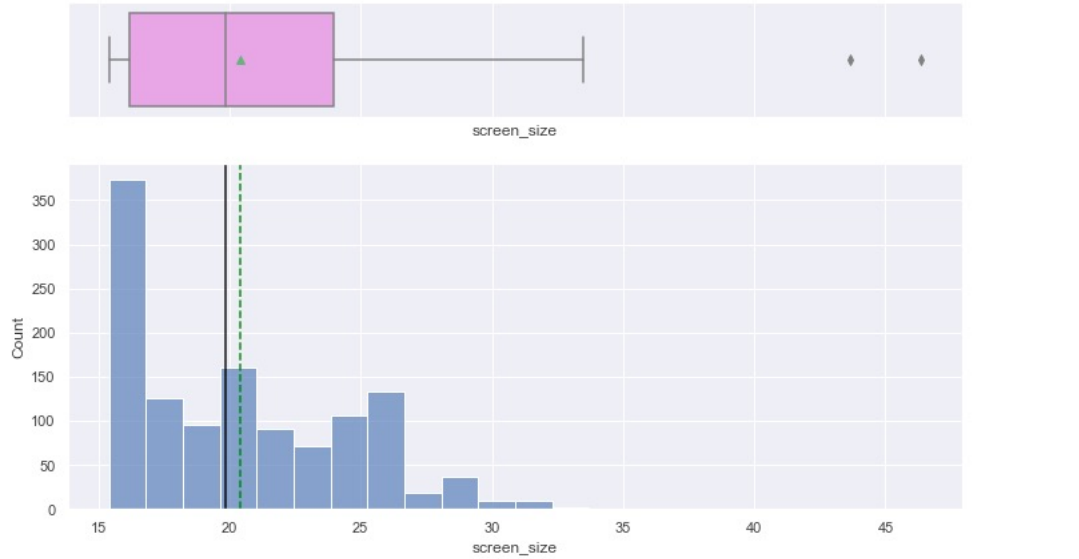
Exploratory Data Analysis (EDA)

Insight: There are 1235 phones with a screen size > 6 inches (15.24 cm)
mean 20.38
min 15.40
25% 16.19
50% 19.84
75% 23.97
Max 46.36

IMPORTANT: for the purposes of this exercise you will see latter that I will be removing the "tablets" based on the latest cell phone screen size research.

	count	mean	std	min	25%	50%	75%	max
screen_size	1235.000	20.384	4.220	15.400	16.190	19.840	23.970	46.360
main_camera_mp	1062.000	9.430	4.812	0.300	5.000	10.500	13.000	48.000
selfie_camera_mp	1234.000	10.812	8.969	0.300	2.200	8.000	16.000	32.000
int_memory	1235.000	80.094	90.459	0.200	32.000	64.000	128.000	1024.000
ram	1235.000	4.518	1.833	0.250	4.000	4.000	4.000	12.000
battery	1233.000	4245.503	1306.243	1200.000	3500.000	4000.000	4500.000	12000.000
weight	1235.000	244.604	123.042	23.000	171.000	192.000	291.000	950.000
release_year	1235.000	2017.672	2.324	2013.000	2016.000	2019.000	2019.000	2020.000
days_used	1235.000	493.794	250.634	91.000	285.500	452.000	667.500	1090.000
new_price	1235.000	310.519	256.334	40.080	150.825	241.810	381.504	2560.200
used_price	1235.000	173.820	169.874	15.190	74.980	122.830	200.700	1916.540

```
1 histogram_boxplot(df_ss, "screen_size")
```

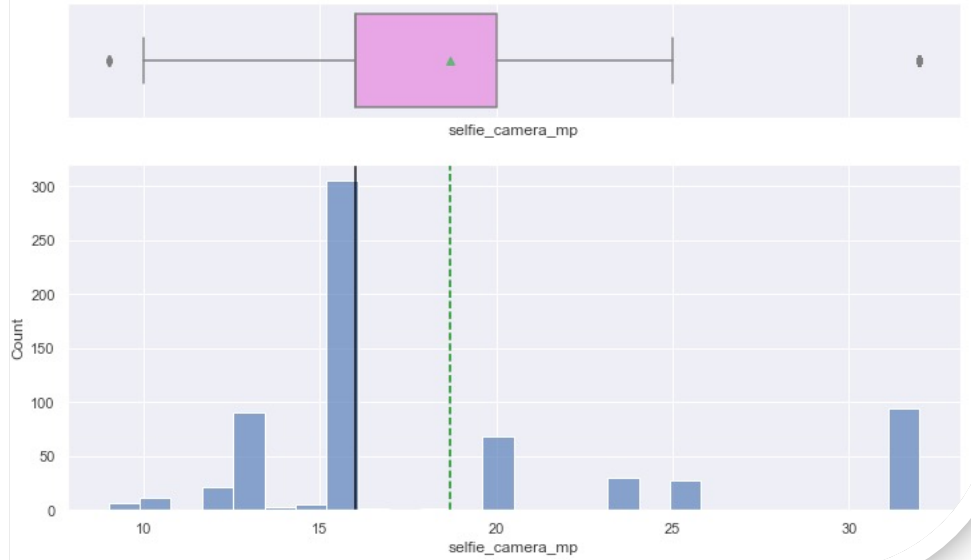


Exploratory Data Analysis (EDA)

Insight: There are 666 phones with selfie camera > 8MP
mean 18.68
min 9.00
25% 16.00
50% 16.00
75% 20.00
max 32.00

	count	mean	std	min	25%	50%	75%	max
screen_size	666.000	18.334	4.593	2.700	15.560	16.110	22.030	32.390
nain_camera_mp	545.000	12.138	4.410	0.300	10.500	13.000	13.000	48.000
selfie_camera_mp	666.000	18.680	6.312	9.000	16.000	16.000	20.000	32.000
int_memory	666.000	108.240	83.033	4.000	64.000	128.000	128.000	1024.000
ram	666.000	5.266	2.089	0.030	4.000	4.000	6.000	16.000
battery	666.000	3828.559	776.013	230.000	3400.000	4000.000	4200.000	6000.000
weight	666.000	180.390	28.711	25.000	165.550	181.000	196.375	300.000
release_year	666.000	2018.631	1.234	2013.000	2018.000	2019.000	2019.000	2020.000
days_used	666.000	424.200	206.416	91.000	264.000	385.500	545.250	1091.000
new_price	666.000	387.318	291.621	99.700	218.923	299.340	459.870	2560.200
used_price	666.000	228.294	195.126	35.740	118.400	165.020	268.502	1916.540

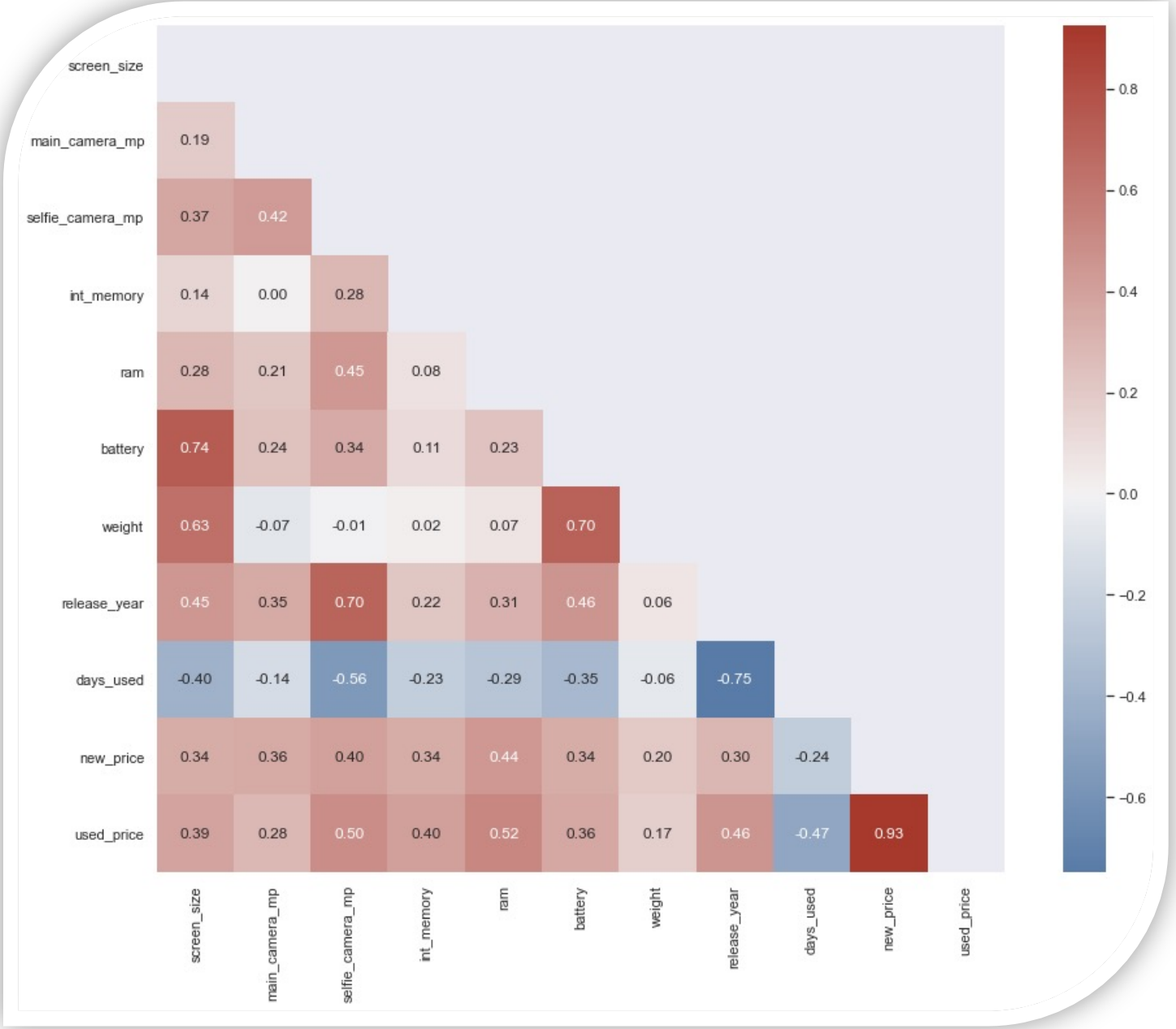
```
1 histogram_boxplot(df_sc, "selfie_camera_mp")
```



Exploratory Data Analysis (EDA)

Insight: The variables that are correlated the most with used_price:

- Selfie_camera_MP
- RAM
- Release_Year
- New_Price
- Days_Used (Negative)



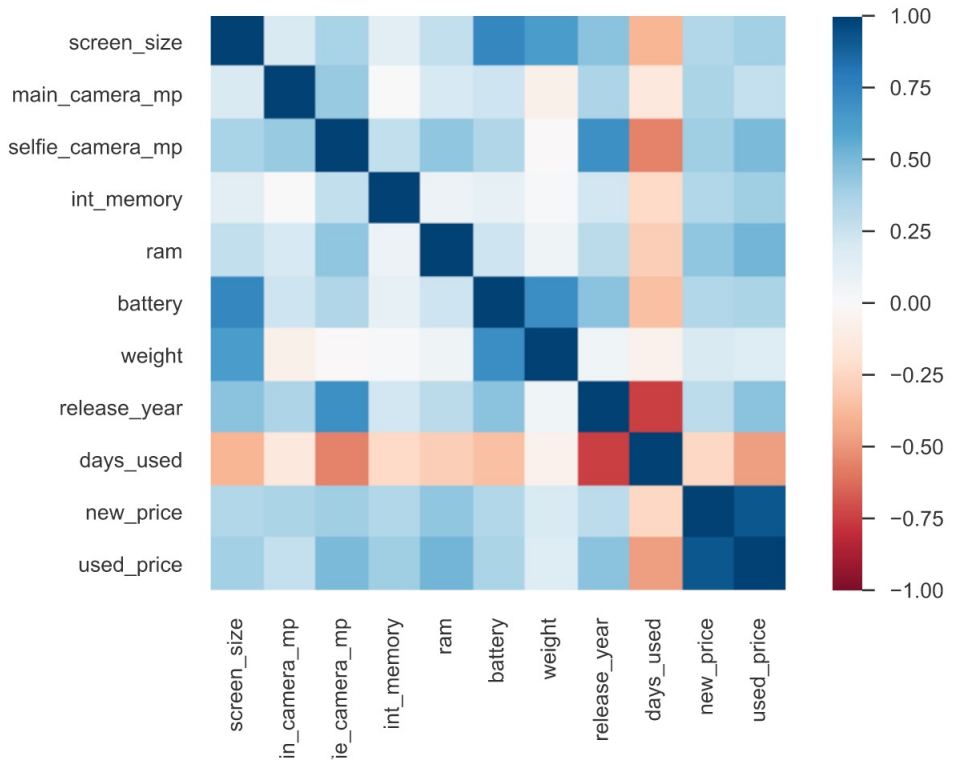
Initial Exploratory Data Observations

Brief description of the raw data.

Key Variables

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

1. Number of variables 15
2. Number of observations 3571
3. Missing cells 215
4. Missing cells (%) 0.4%
5. Duplicate rows 0
6. Duplicate rows (%) 0.0%
7. Total size in memory 418.6 KiB
8. Average record size in memory 120.0 B
9. Categorical 2
10. Numeric 11
11. Boolean 2

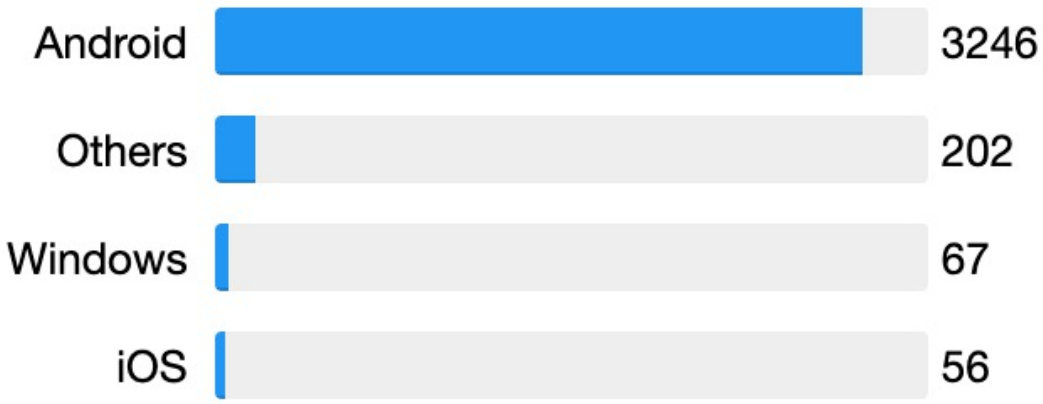


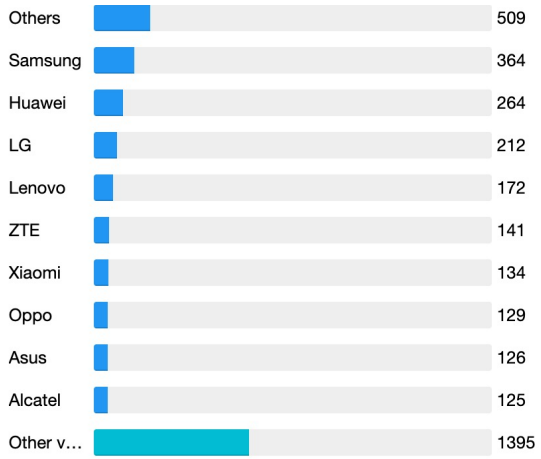
A number of strong +/- relationships jump out:

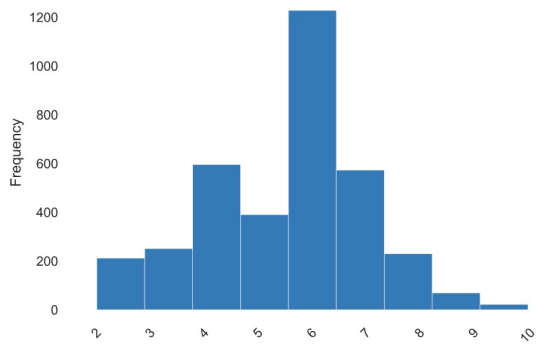
1. The number of days used neatively impacts used price
2. Screen size and selfie camera MP both have a positive effect on used prices

❖ *Full Details in Appendix*

Exploratory Data Analysis (EDA)



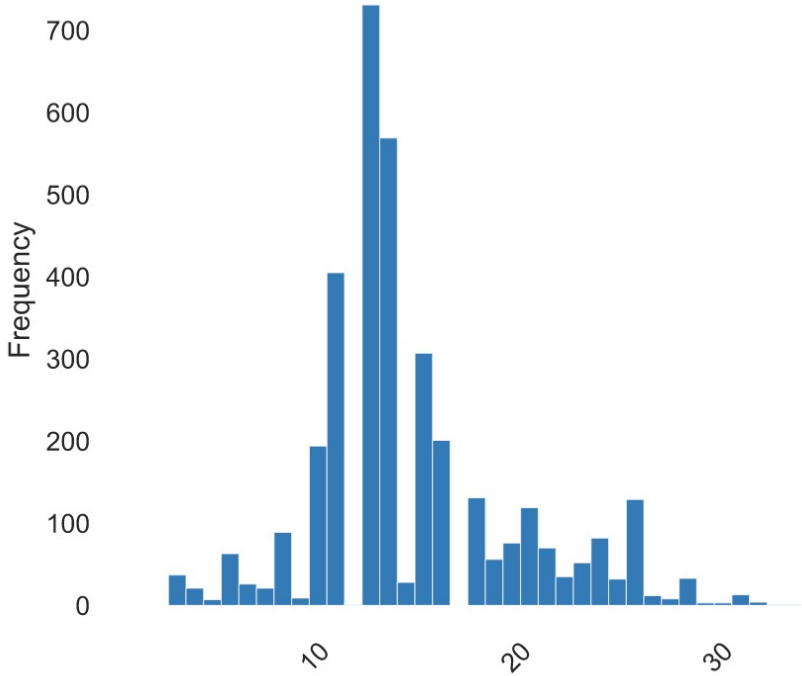




Histogram of lengths of the category

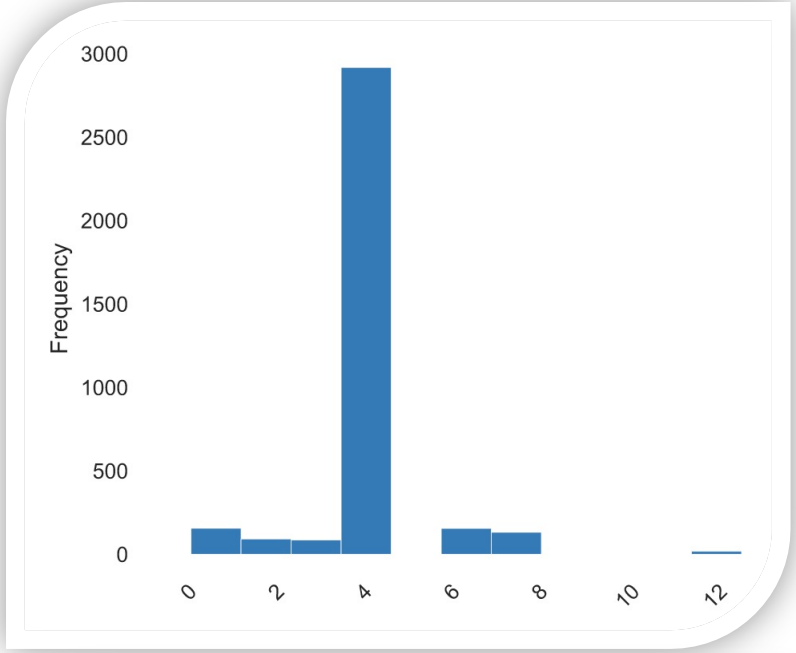
Android phones represent the majority with many known brands like Samsung, Hauwei, and LG represented.

***However** … the data suggest we might have some tablet's in this base on the screen size that should be consider as part of another model.*

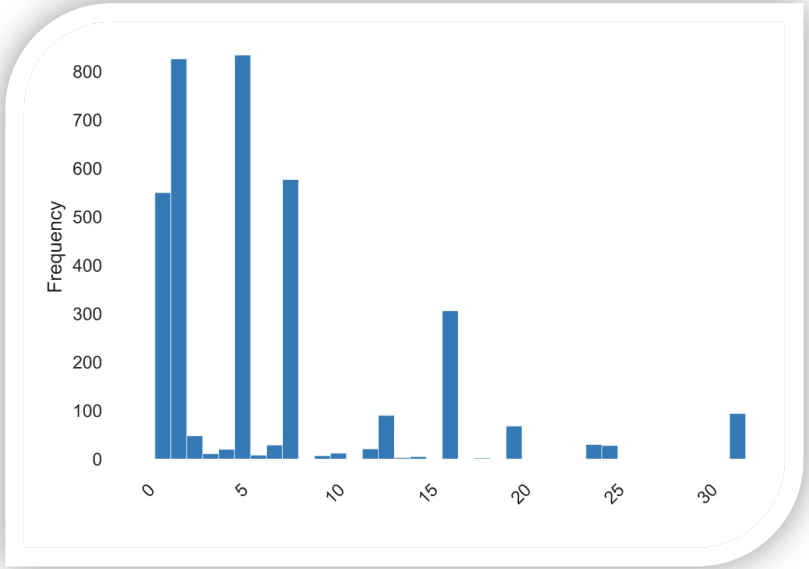


Exploratory Data Analysis (EDA)

Users expect a minimum of RAM to be available so the can run multiple apps on the phone. Most of our phones have @ least 4GB of RAM.



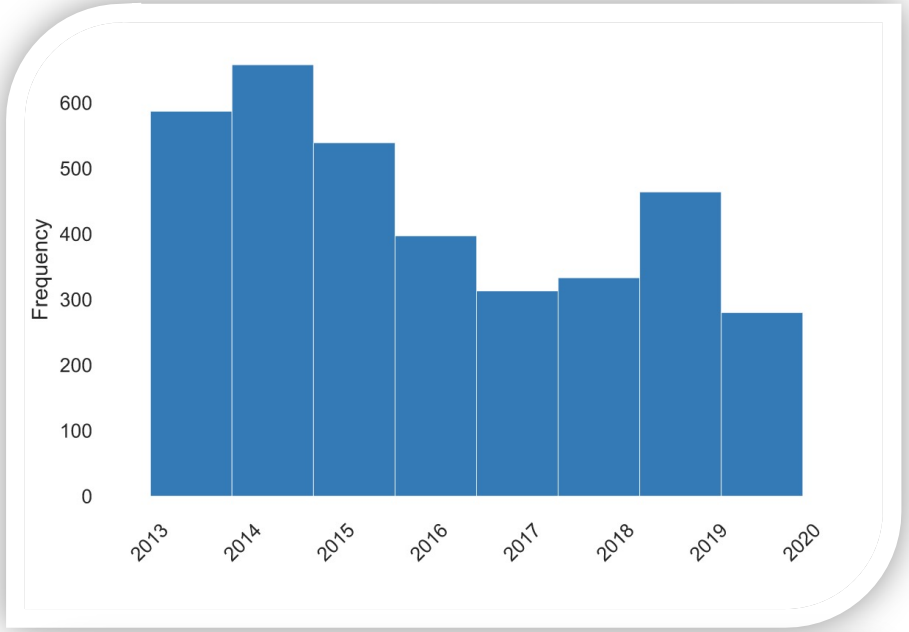
This is very important! With the advent of social media users want and will pay for larger MP selfie cameras, so that should help our overall used price @ market.

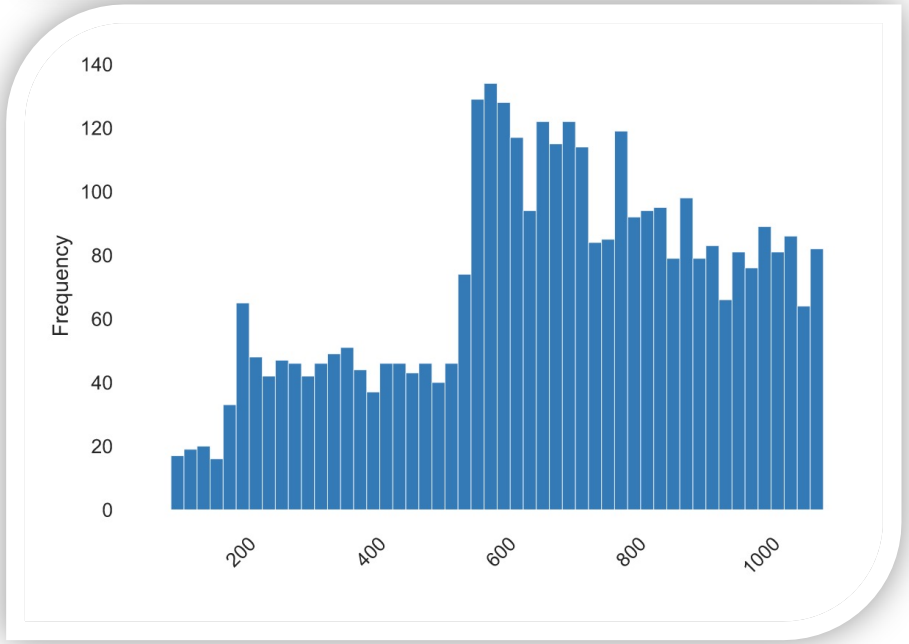


Exploratory Data Analysis (EDA)

As this chart shows, we have phones going back to 2013-2017. One thing our model will have to be smart about, is most phone prices have come down and have more features, so we will liklely have to account for the in pricing.

Most of the phones we have identified for resale, have been used for at least two years or more, so buyers will expect them to be heavily discounted.





Data Transformations

As noted initially, a number of clues were given up front that we were going to need to do some data preprocessing and transformations:

1. Missing data
2. outliers in the data do exist
3. skewness exists in a number of variables preprocessing/feature creation.

TRANSFORMATIONS

1. DROP the os column - 1) ios as it is perfectly correlated with brand APPLE.
2. Fill the missing values with for all all numeric columns - we are going to use the median value as so few values are missing from each column.
3. Feature creation - create a "totalmem'= int + ram since they are on the same scale df["totalmem"] = df["ram"] + df["int_memory"]; and drop the ram and int memory from the data.
   - Bin the new “totalmem” into ‘bronze", "silver", "gold", "platnium” which corresponds to levels of memory.
4. There are tablets and/or pc that mistakeningly got added to the used "cell phone" list. Drop anything larger than > 8inches or 20.32 centimeters ** research note (iphone 13 pro max 6.7-inch screen, samsung galaxy S21 ultra 6.8-inch screen, galaxy Z fold 3- 7.6-inch screen)
5. Feature creation - change the used and new prices to log versions of the euro price
6. Floor & cap all other variables to address skewness by eliminating the outliers.
7. Change categorical & object values to 1 or zero using pd.Dummies to make it easier to interpret their impact on price.

Linear Model Building – *Initial Results*

Step 1: We want to predict the used price.

- Before we proceed to build a model, we'll have to encode categorical features.
- We'll split the data into train and test to be able to evaluate the model that we build on the train data.
- We will build a Linear Regression model using the train data and then check it's performance.

Number of rows in train data = 2143
Number of rows in test data = 919

Training Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.091	0.074	0.986	0.985	1.847

Test Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.097	0.078	0.984	0.983	1.922

1. The training R^2 is 98.4%, indicating that the model explains 98.4% of the variation in the train data. So, the model is not underfitting.
2. MAE and RMSE on the train and test sets are comparable, which shows that the model is not overfitting.
3. MAE indicates that our current model is able to predict used price within a mean error of .078 on the test data.
4. MAPE on the test set suggests we can predict within 1.922% of the used price.
5. Negative values of the coefficient show that *Used Price* decreases with the increase of corresponding attribute value.
6. Positive values of the coefficient show that *Used Price* increases with the increase of corresponding attribute value.

Linear Model Building – *Coefficients*

	Coefficients
screen_size	0.002
main_camera_mp	0.000
selfie_camera_mp	0.001
battery	-0.000
weight	0.000
release_year	0.004
days_used	-0.001
new_price_log	0.991
brand_name_Alcatel	0.002
brand_name_Apple	0.010
brand_name_Asus	0.022
brand_name_BlackBerry	-0.002
brand_name_Celkon	0.043
brand_name_Coolpad	0.023
brand_name_Gionee	-0.023
brand_name_Google	0.009
brand_name_HTC	0.016
brand_name_Honor	0.028
brand_name_Huawei	0.018
brand_name_Infinix	0.029
brand_name_Karbonn	-0.016
brand_name_LG	0.014
brand_name_Meizu	0.012
brand_name_Micromax	0.001
brand_name_Microsoft	-0.018
brand_name_Motorola	0.010
brand_name_Nokia	-0.010
brand_name_OnePlus	-0.029
brand_name_Oppo	0.006
brand_name_Others	0.006
brand_name_Panasonic	-0.009
brand_name_Realme	0.058
brand_name_Samsung	0.012
brand_name_Sony	0.037
brand_name_Spice	0.014
brand_name_Vivo	-0.005
brand_name_XOLO	-0.001
brand_name_Xiaomi	0.016
brand_name_ZTE	0.007
4g_yes	-0.005
5g_yes	-0.048
bintotalmem_silver	0.002
bintotalmem_gold	-0.168

Step 2: Checking Linear Regression Assumptions

	Test	Results	Remediation
1. No Multicollinearity	• VIF • *usually grater than 5 needs to be addressed	• "brand_name_Huawei", "brand_name_LG", "brand_name_Others", "brand_name_Samsung", • have VIF greater than 5.	• After dropping brand_name_Others and rerunning the VIF, ALL values are now under 5. Having met the criteria, we can proceed with no additional changes.
2. Linearity of variables	• Make a plot of fitted values vs residuals.	• The errors seem generally consistent, except for the tail where they increase and there is that odd left/right line. This possibly suggests that there is a still optimization that can be done.	• Test Passes … SEE APPENDIX
3. Independence of error terms	• Make a plot of fitted values vs residuals.	• SAME	• Test Passes … SEE APPENDIX
4. Normality of error terms	• The shape of the histogram of residuals can give an initial idea about the normality. • Shapiro-Wilk	• The distribution appears to be normal • Since p-value < 0.05, the residuals are not normal as per the Shapiro-Wilk test. Strictly speaking, the residuals are not normal. However, as an approximation, we can accept this distribution as close to being normal.	• Test Passes … SEE APPENDIX
5. No Heteroscedasticity	• goldfeldquandt test	• Since p-value > 0.05, we can say that the residuals are homoscedastic. So, this assumption is satisfied.	• Test Passes … SEE APPENDIX

Step 3: Predicting Used Price

Training Performance

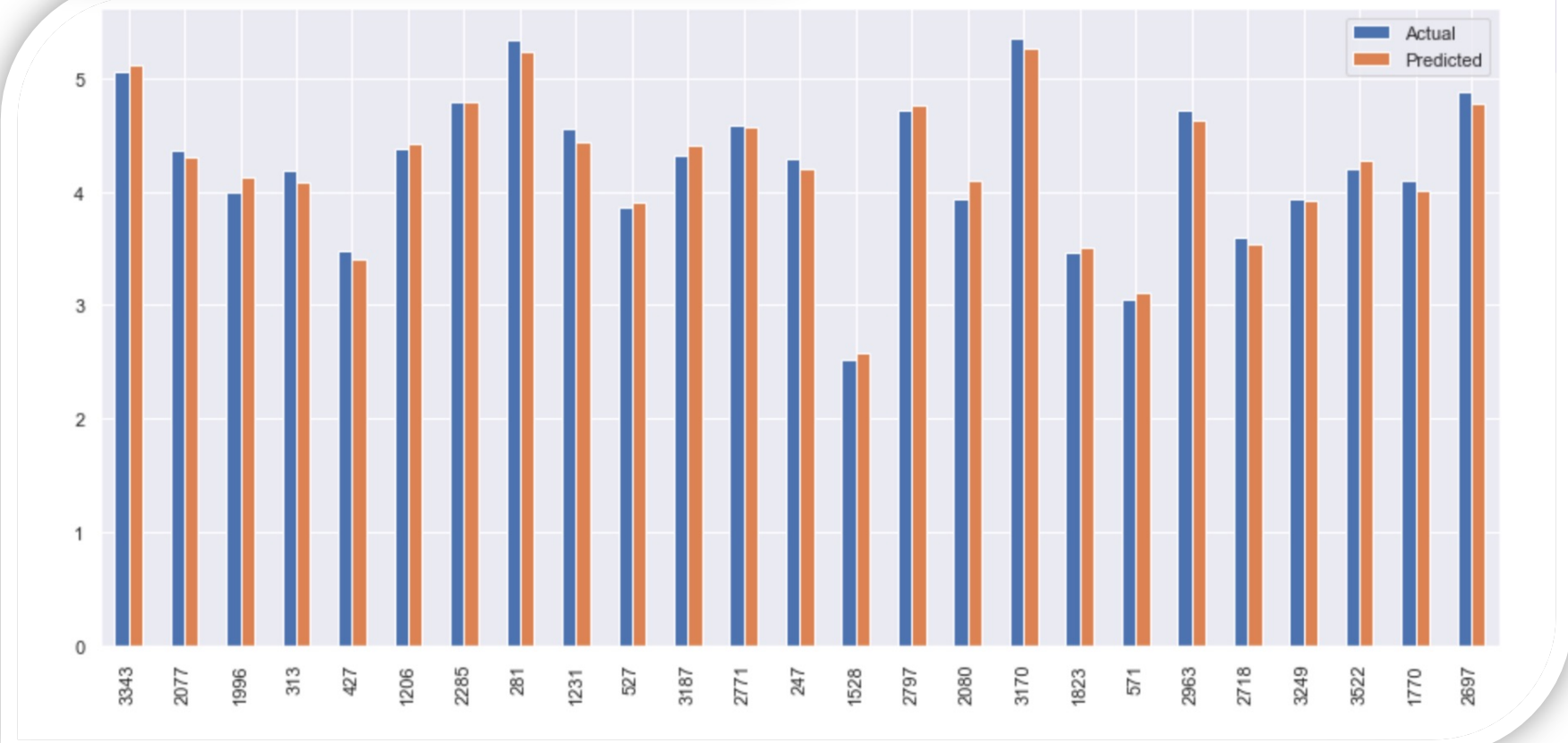
click to scroll output; double click to hide

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.091	0.074	0.985	0.985	1.859

Test Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.097	0.078	0.984	0.984	1.921

	Actual	Predicted
3343	5.059	5.123
2077	4.359	4.305
1996	3.988	4.132
313	4.194	4.079
427	3.481	3.409
1206	4.386	4.423
2285	4.788	4.787
281	5.343	5.240
1231	4.551	4.441
527	3.863	3.910



- The model is able to explain ~98% of the variation in the data, which is very good.
- The train and test RMSE and MAE are low and comparable. So, our model is not suffering from overfitting.
- The MAPE on the test set suggests we can predict within 1.92% of the used price.
- Hence, we can conclude the model *olsmod2* is good for prediction as well as inference purposes.

Step 4: Let's compare the initial model created with sklearn and the final statsmodels model.

Training performance comparison:

	Linear Regression sklearn	Linear Regression statsmodels
RMSE	0.091	0.091
MAE	0.074	0.074
R-squared	0.986	0.985
Adj. R-squared	0.985	0.985
MAPE	1.847	1.859

Test performance comparison:

	Linear Regression sklearn	Linear Regression statsmodels
RMSE	0.097	0.097
MAE	0.078	0.078
R-squared	0.984	0.984
Adj. R-squared	0.983	0.984
MAPE	1.922	1.921

- The performance of the two models is close to each other.

Step 5: Let's compare the initial model created with sklearn and the final statsmodels model.

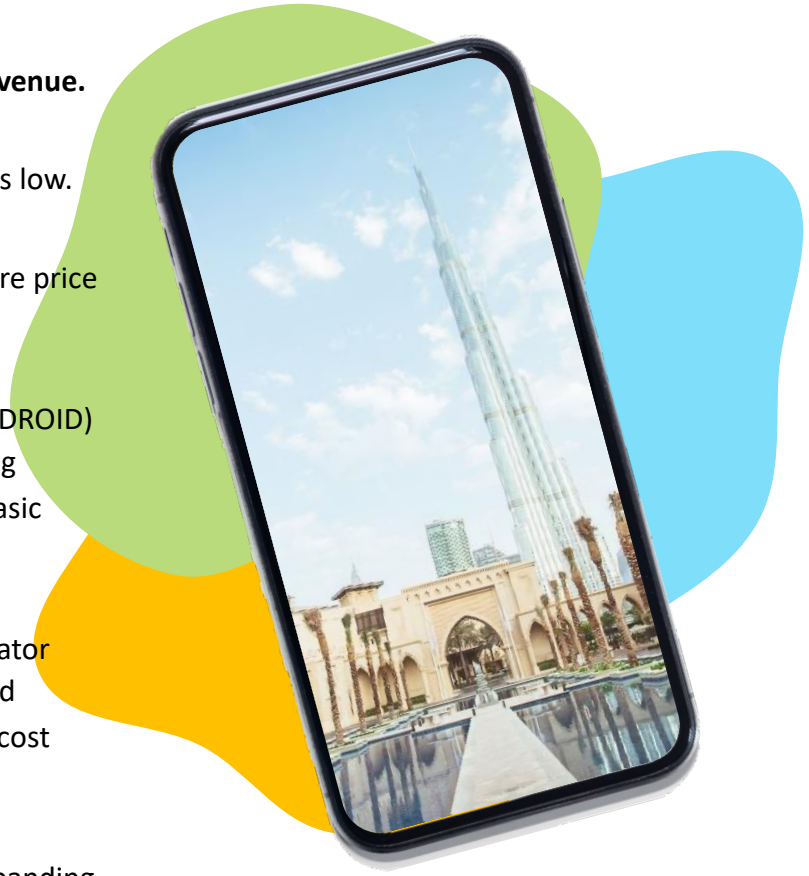
```
                            OLS Regression Results
==============================================================================
Dep. Variable:         used_price_log   R-squared:                       0.985
Model:                            OLS   Adj. R-squared:                  0.985
Method:                 Least Squares   F-statistic:                 1.312e+04
Date:                Thu, 21 Oct 2021   Prob (F-statistic):               0.00
Time:                        18:29:16   Log-Likelihood:                 2093.7
No. Observations:                2143   AIC:                            -4163.
Df Residuals:                    2131   BIC:                            -4095.
Df Model:                          11
Covariance Type:            nonrobust
==============================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const             -8.3329      3.450     -2.415      0.016     -15.099      -1.567
screen_size        0.0017      0.001      2.270      0.023       0.000       0.003
selfie_camera_mp   0.0014      0.001      2.083      0.037     8.3e-05       0.003
release_year       0.0041      0.002      2.372      0.018       0.001       0.007
days_used         -0.0011    1.2e-05    -88.620      0.000      -0.001      -0.001
new_price_log      0.9929      0.004    265.860      0.000       0.986       1.000
brand_name_Celkon  0.0410      0.019      2.173      0.030       0.004       0.078
brand_name_Gionee -0.0308      0.015     -2.026      0.043      -0.061      -0.001
brand_name_Realme  0.0476      0.021      2.321      0.020       0.007       0.088
brand_name_Sony    0.0282      0.013      2.147      0.032       0.002       0.054
5g_yes            -0.0488      0.017     -2.811      0.005      -0.083      -0.015
bintotalmem_gold  -0.1655      0.042     -3.971      0.000      -0.247      -0.084
==============================================================================
Omnibus:                      287.881   Durbin-Watson:                   2.044
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              717.980
Skew:                          -0.751   Prob(JB):                    1.24e-156
Kurtosis:                       5.406   Cond. No.                     3.74e+06
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.74e+06. This might indicate that there are
strong multicollinearity or other numerical problems.
```

1. Screen Size, Selfie Camera MP, Release year all have positive impacts on used price.
2. Days Used as expected has a negative imnpact on Used Price
3. 5g/Gold Memory (a high memory category) both take away from the price, as suggested before the market is not for the US or developed countries.
4. Brand Name works both ways (Celkon, Realme, and Sony are positive factors) and Gionee is negative impcat to used price.

Business Insights and Recommendations

Recommendations we can implement now to drive ReCell revenue.

- ✓ We should focus on markets where cell phone ownership is low.
- ✓ We can offer good value and performance phones for where price is the deciding factor.
- ✓ We should focus on the minimum requirements (RAM, ANDROID) and possible test create two “packages” based on brand e.g “Premium = Song, LG, Samsung, Huawei” and a second “Basic Value – which the brand isn’t disclosed”.
- ✓ A further feature would be to develop a real-time configurator that allows the user to select from all of the “variables” and compare that to what the Premium/Basic offerings would cost and what features would be included.
- ✓ Explore in markets where “Apple” is a dominant brand, expanding the offering in those targeted markets based on the same principles.



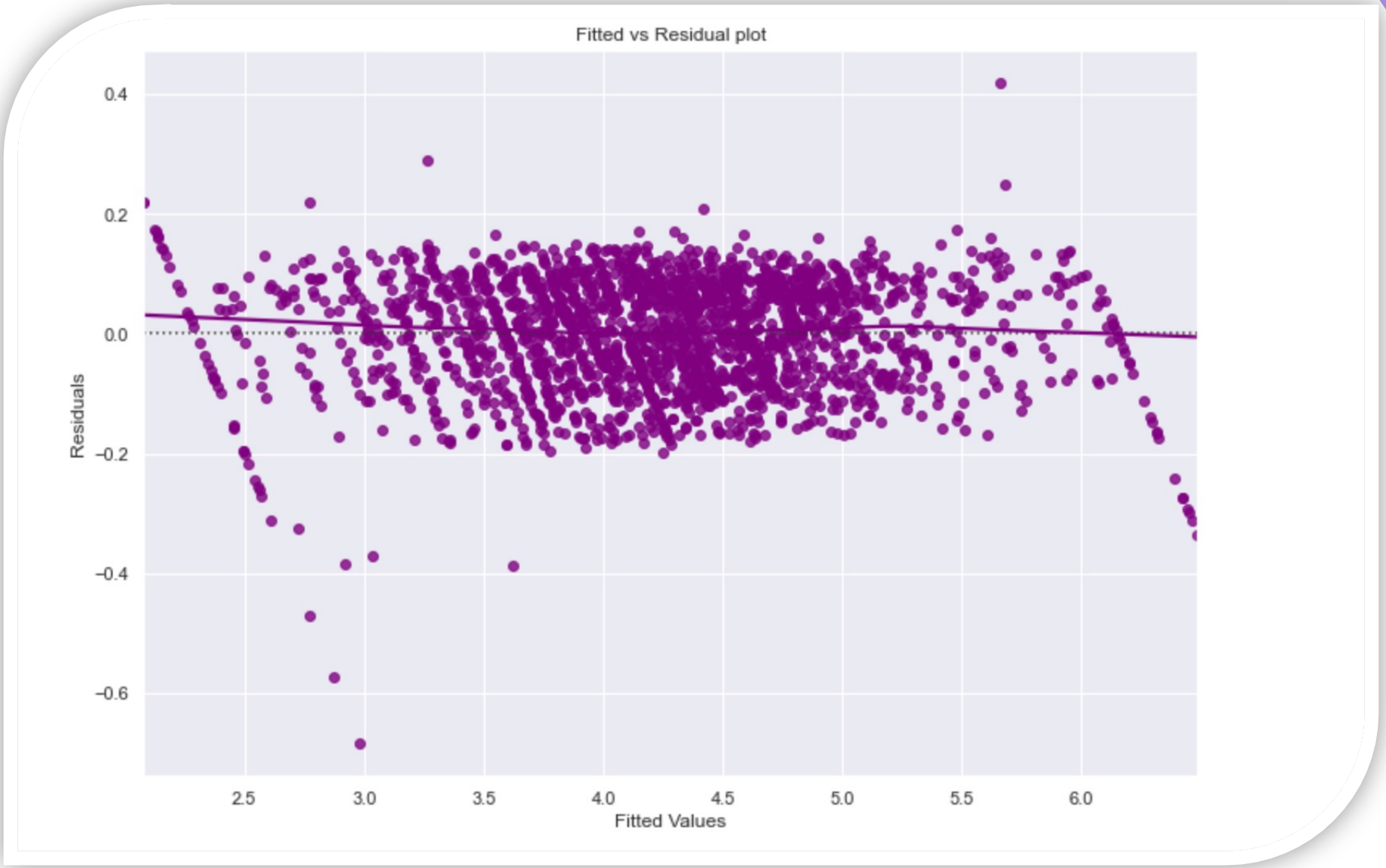
Continue to enhance actionable insights based on the results of ongoing analysis!

1. ***More data is need to further understand the drives behind E-News’s business:***
2. ***Date and time of each visitor***
3. ***What other news sites do they user***
4. ***More competitive pricing analysis***

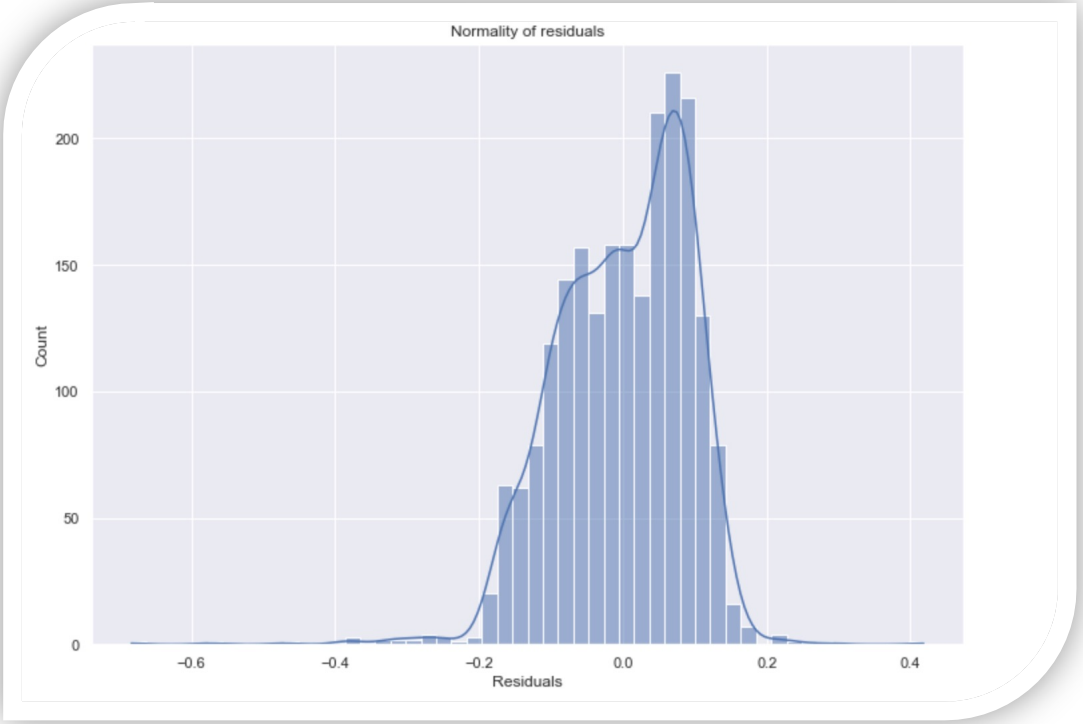
Appendix: EDA Initial Relationship Between Variables

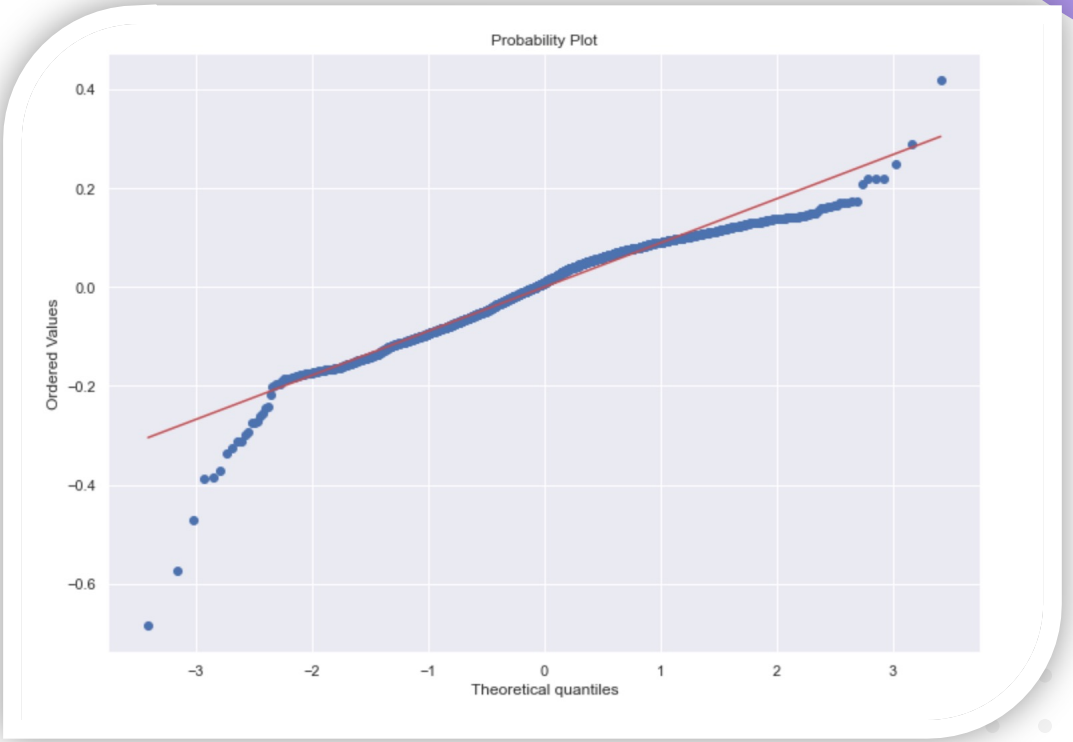
1. screen_size is highly correlated with battery and 1 other fields
2. selfie_camera_mp is highly correlated with release_year and 1 other fields
3. ram is highly correlated with used_price
4. battery is highly correlated with screen_size and 1 other fields
5. weight is highly correlated with screen_size and 1 other fields
6. release_year is highly correlated with selfie_camera_mp and 1 other fields
7. days_used is highly correlated with selfie_camera_mp and 1 other fields
8. new_price is highly correlated with used_price
9. used_price is highly correlated with ram and 1 other fields
10. screen_size is highly correlated with selfie_camera_mp and 4 other fields
11. main_camera_mp is highly correlated with selfie_camera_mp and 2 other fields
12. selfie_camera_mp is highly correlated with screen_size and 7 other fields
13. int_memory is highly correlated with selfie_camera_mp and 1 other fields
14. battery is highly correlated with screen_size and 4 other fields
15. weight is highly correlated with screen_size and 1 other fields
16. release_year is highly correlated with screen_size and 5 other fields
17. days_used is highly correlated with selfie_camera_mp and 2 other fields
18. new_price is highly correlated with main_camera_mp and 2 other fields
19. used_price is highly correlated with screen_size and 6 other fields
20. screen_size is highly correlated with battery and 1 other fields
21. main_camera_mp is highly correlated with selfie_camera_mp
22. selfie_camera_mp is highly correlated with main_camera_mp and 1 other fields
23. battery is highly correlated with screen_size and 1 other fields
24. weight is highly correlated with screen_size and 1 other fields
25. release_year is highly correlated with selfie_camera_mp and 1 other fields
26. days_used is highly correlated with release_year
27. new_price is highly correlated with used_price
28. used_price is highly correlated with new_price
29. weight is highly correlated with screen_size and 1 other fields
30. new_price is highly correlated with used_price
31. used_price is highly correlated with new_price and 2 other fields
32. selfie_camera_mp is highly correlated with release_year and 3 other fields
33. main_camera_mp is highly correlated with brand_name and 1 other fields
34. release_year is highly correlated with selfie_camera_mp and 6 other fields
35. brand_name is highly correlated with selfie_camera_mp and 8 other fields
36. os is highly correlated with brand_name and 2 other fields
37. screen_size is highly correlated with weight and 4 other fields
38. days_used is highly correlated with release_year and 4 other fields
39. 5g is highly correlated with used_price and 3 other fields
40. int_memory is highly correlated with brand_name and 3 other fields
41. 4g is highly correlated with selfie_camera_mp and 4 other fields
42. battery is highly correlated with weight and 7 other fields
43. ram is highly correlated with used_price and 5 other fields
44. 4g is highly correlated with brand_name
45. os is highly correlated with brand_name
46. brand_name is highly correlated with 4g and 1 other fields
47. main_camera_mp has 180 (5.0%) missing values

Appendix: EDA Initial Relationship Between Variables



Appendix: TEST FOR NORMALITY






```
ShapiroResult(statistic=0.9583384990692139, pvalue=2.593337430209968e-24)
```

Since p-value < 0.05, the residuals are not normal as per the Shapiro-Wilk test. Strictly speaking, the residuals are not normal. However, as an approximation, we can accept this distribution as close to being normal.

```
[('F statistic', 1.0169981039356226), ('p-value', 0.3919203930762724 7)]
```

Since p-value > 0.05, we can say that the residuals are homoscedastic. So, this assumption is satisfied.

So, this assumption is satisfied.

Thank You

John Noble

617.519.9065