ReCell

Supervised Learning- Foundations Project, PGP DSBA

January 27th 2023

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Contents / Agenda

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Executive Summary

- The model we created accounts for about 83% of the variation in the test data and the MAPE predicts within 4.56% of the normalized used price on the test data, so we can be confident that this model is a good predictor.
- When the release year increases by one, meaning the device is a year older, then the used price decreases by 0.029 and when the ram increases by one, meaning more ram, then the used price increases by 0.021.
- When the weight of a device increases by one, the used price increases by 0.0017 and when the main camera increases by one, then the used price increases by 0.021.
- According to the statistical inferences stated above, it would be smart of ReCell to provide only devices with higher RAMs and higher resolution cameras.
- ReCell should aim for more variety in operating systems of devices as well as having devices with more current release years.
- Devices with an older release year, 5g networks, and iOS have a lower normalized used price, and therefore should not be stocked/sold as much.

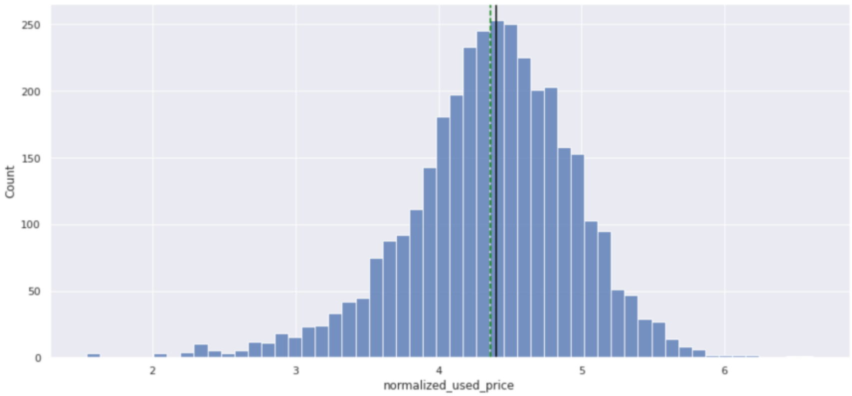
Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

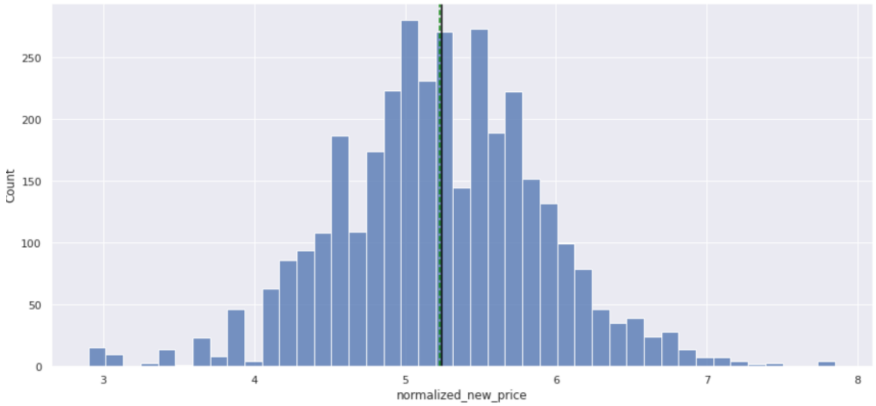
Business Problem Overview and Solution Approach

- Buying and selling used phones and tablets is a rapidly growing industry with technological advances happening so often and devices going in and out of style. Refurbished/used devices provide a cost-effective alternative to customers that are trying to save money while also having an up-to-date device. Increasing the longevity of devices through second-hand sales also reduces the environmental impact of e-waste and encourages recycling them. ReCell, a startup selling these devices, is trying to tap into the potential of this market. By creating a linear regression model to analyze which factors most significantly affect the price of a used device most, ReCell is attempting to determine how to best determine which products they should sell and how they should be priced.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

EDA Results

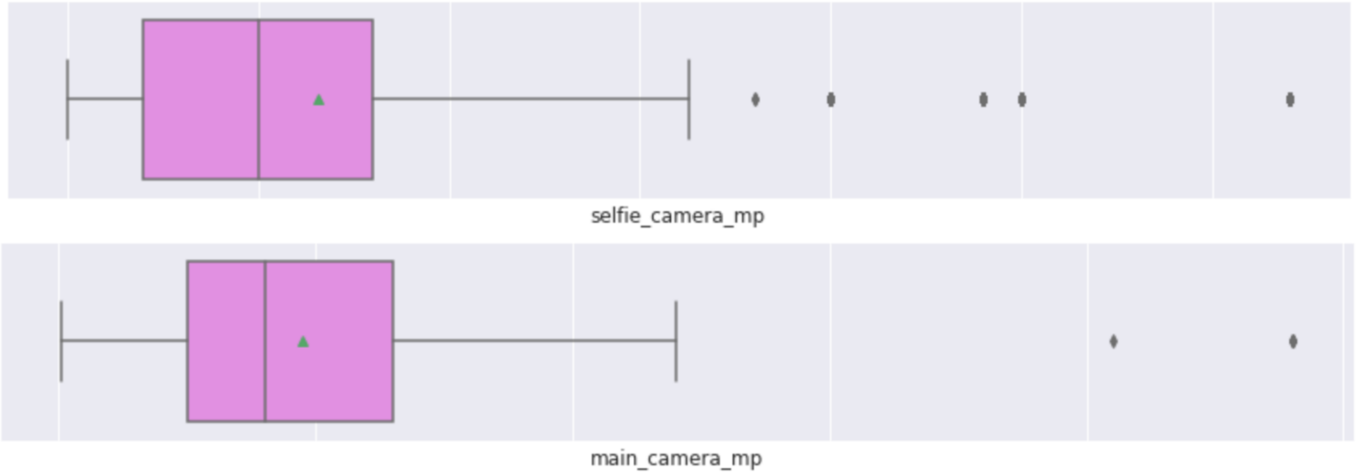




By looking at the two histograms above, we can see that the normalized price for used devices is significantly lower than that of new devices. Further, the distribution of normalized used price looks to be more normal than the normalized new price, meaning it might be simpler to predict/track fluctuations of.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

EDA Results

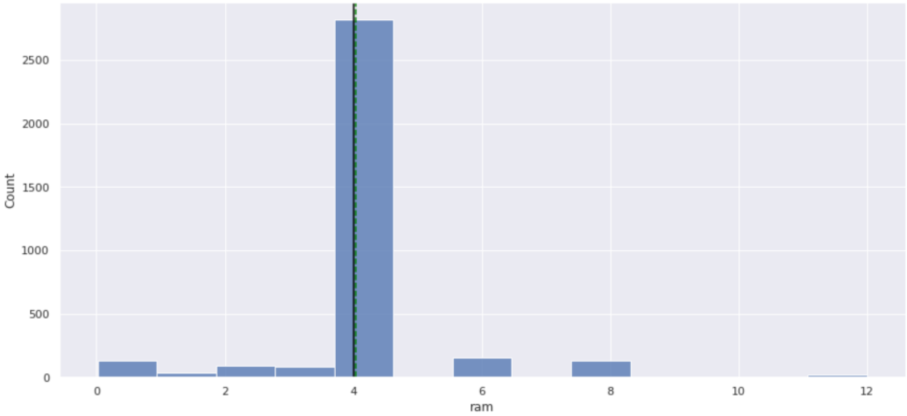


By looking at the boxplots above, we can see that usually the main cameras are of higher quality than selfie cameras, which makes sense because usually main cameras are bigger and more important than front cameras. Similarly, selfie cameras have more outliers than main cameras indicating that there is more variation in the lower quality cameras.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

EDA Results

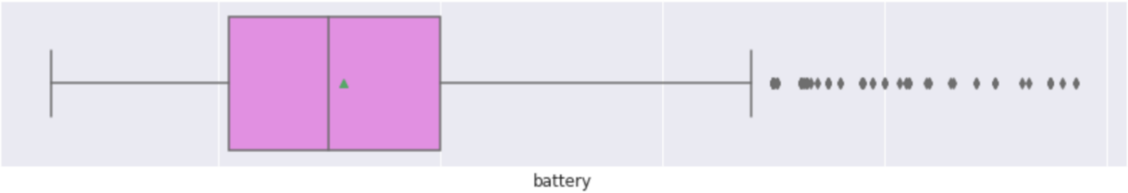
The distribution of device weights tell us that most devices have smaller weights, with several outliers which could simply be due to the variation of types of devices (tablets weigh more than phones).



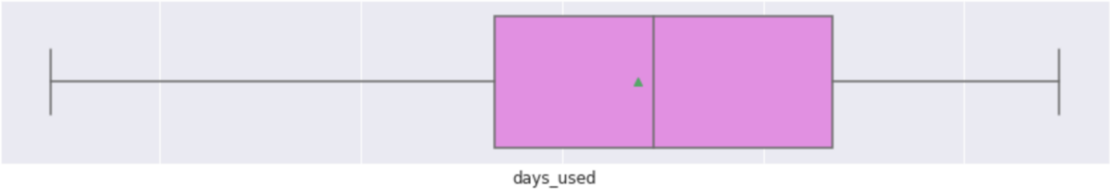
The distribution of RAM makes it clear that the most common one is 4, with a few outliers that are not quite significant enough to change the distribution.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

EDA Results



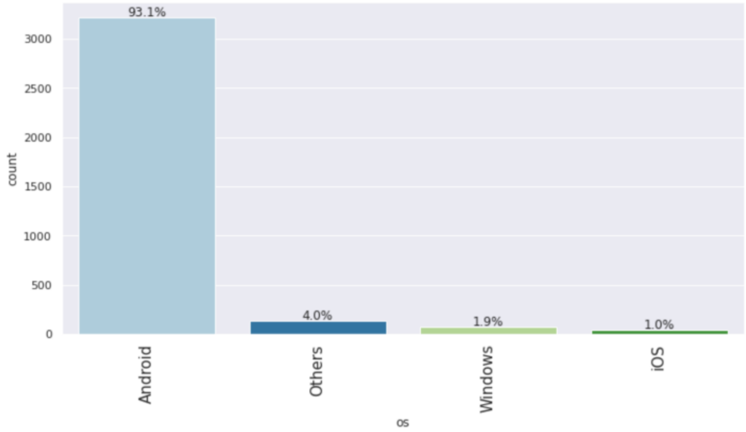
The boxplot of battery has a fairly low mean with many outliers on the upper side. This tells us that a lot of the devices being sold have higher battery capacity than the average, again could be pointing to a difference between tablets and phones.

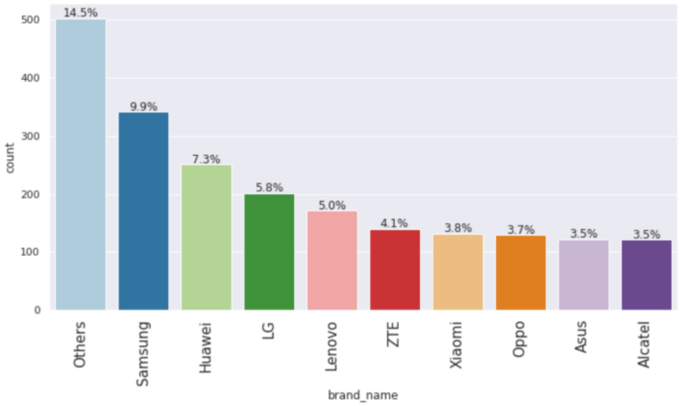


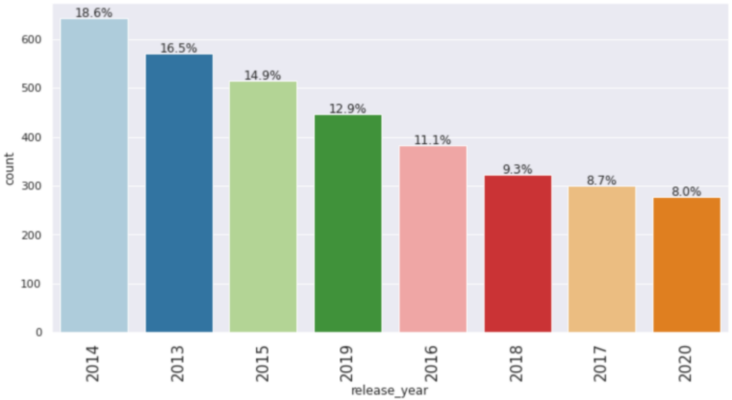
The boxplot of days used tells us that the average use time before a device is being sold is pretty high. This could have some consequences in the resale value of the devices, meaning they will probably be lower than the ones that have not been used as long.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

EDA Results



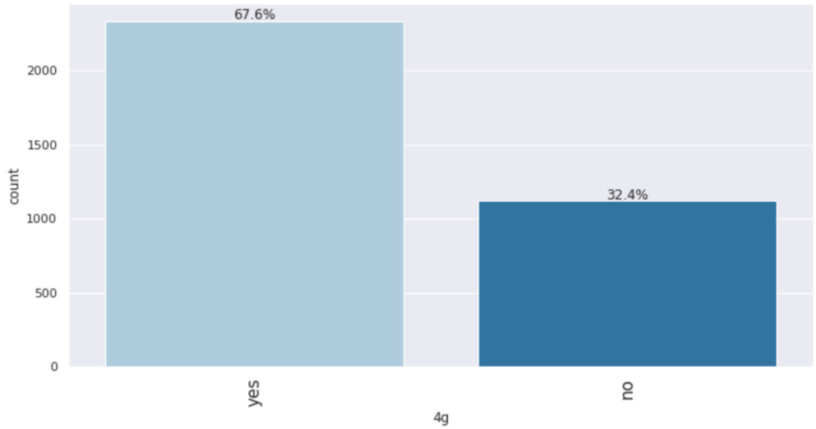


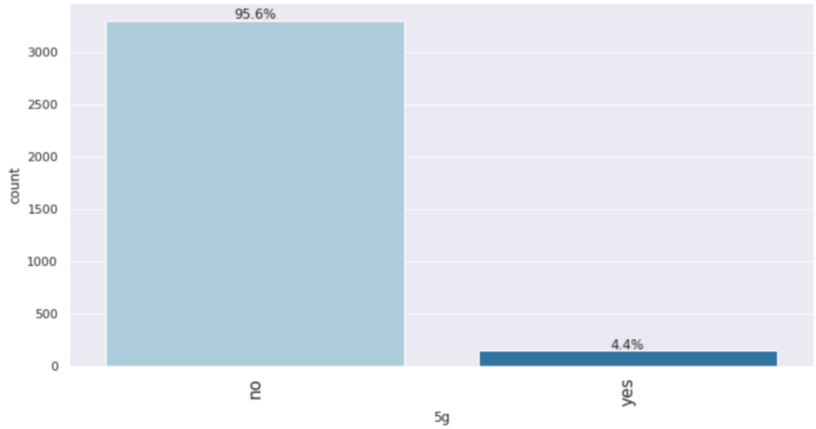


From these three graphs, we see the distributions of brand, operating system, and release year. Most devices are androids with a variety of brands. The most common brand is Samsung, but we probably need more information on the brands labeled as “other.” The most common release year is 2014, which is considerably old in terms of hand-held devices as they are always changing and evolving.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

EDA Results

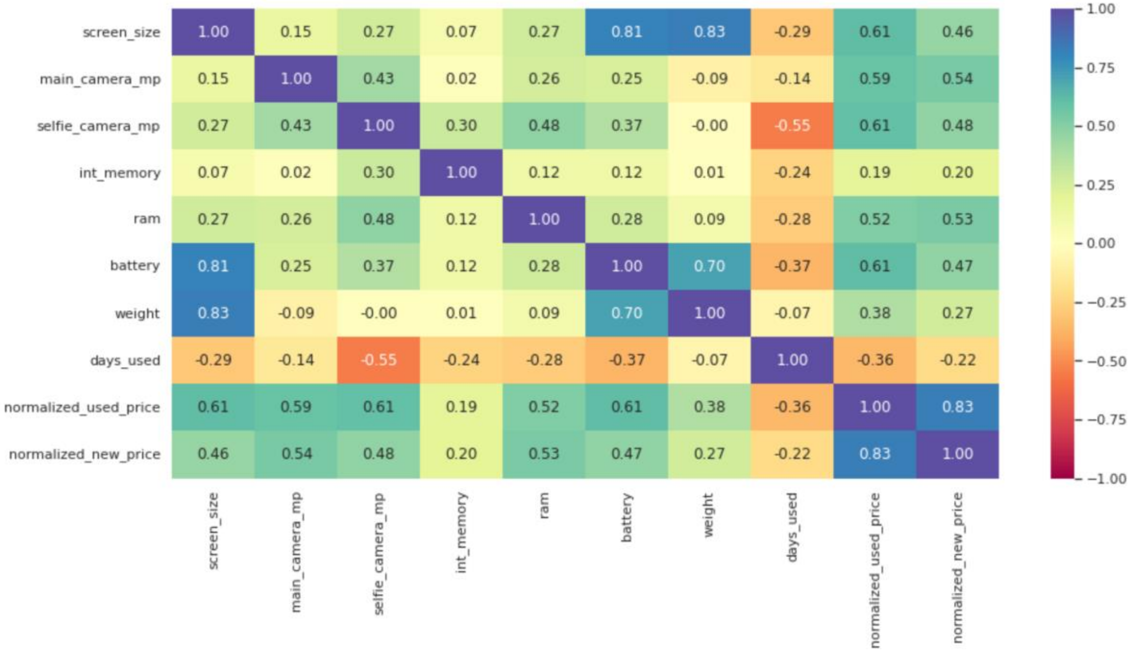




Most devices are 4g and fewer have 5g capabilities. This information will help when we analyze the cost trends in 4g vs. 5g trends so we can come to a conclusion on which would have better resale value

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

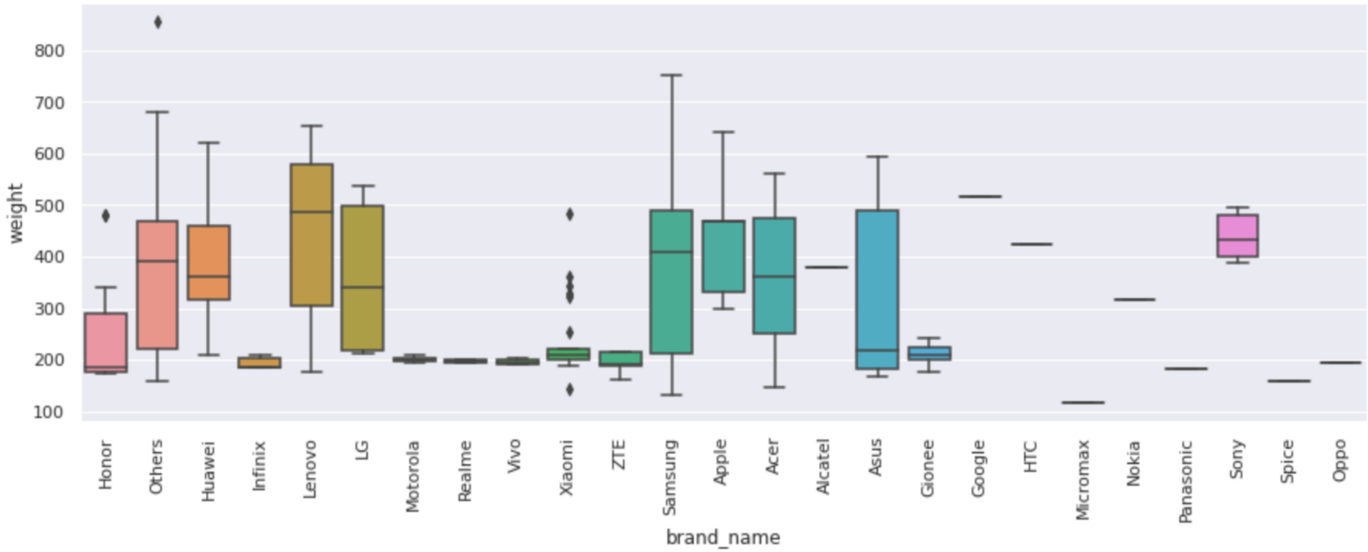
EDA Results



To summarize the heat map above, there are a few key points we can extract. Normalized prices for used and new devices are both positively correlated with screen size, camera strength, internal memory, RAM, battery life, and weight. They are both also negatively correlated with days used.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

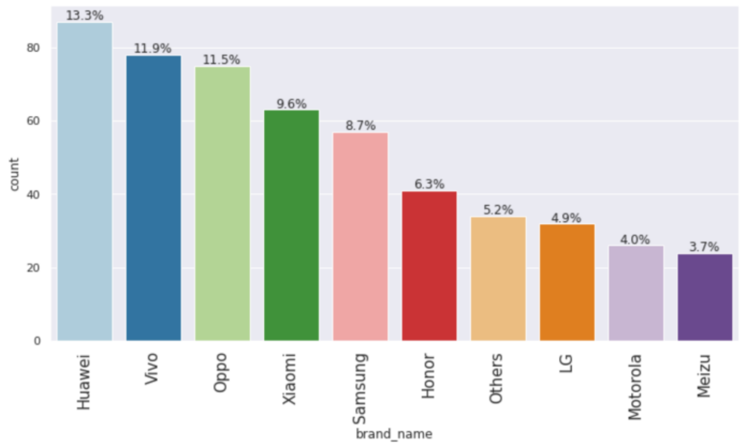
EDA Results

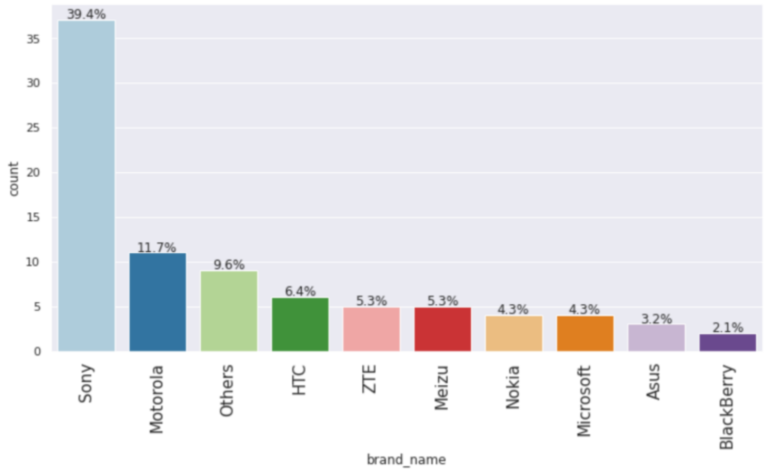


In this graph, we analyzed the brands with only large battery sizes compared to their weights. It would be smarter to keep more devices from brands with large battery and lower weights as it is more user-friendly. This means brands to avoid having a lot of devices from are Lenovo and Samsung.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

EDA Results



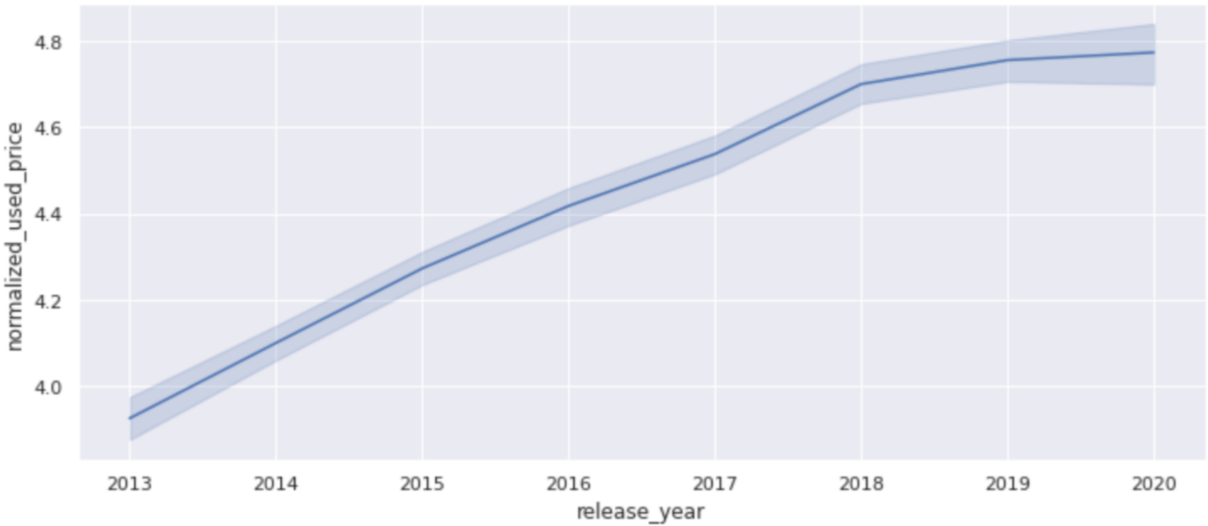


When shopping for a new phone, many users look for a good front camera and to analyze which brands offer this, we created a graph of the best selfie camera quality and which brands have them. It would be a good idea to have more of these brands in stock.

This second graph is the same but for back camera quality, as the main camera is also important to customers. This shows us Sony is the most available brand with good quality main camera.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

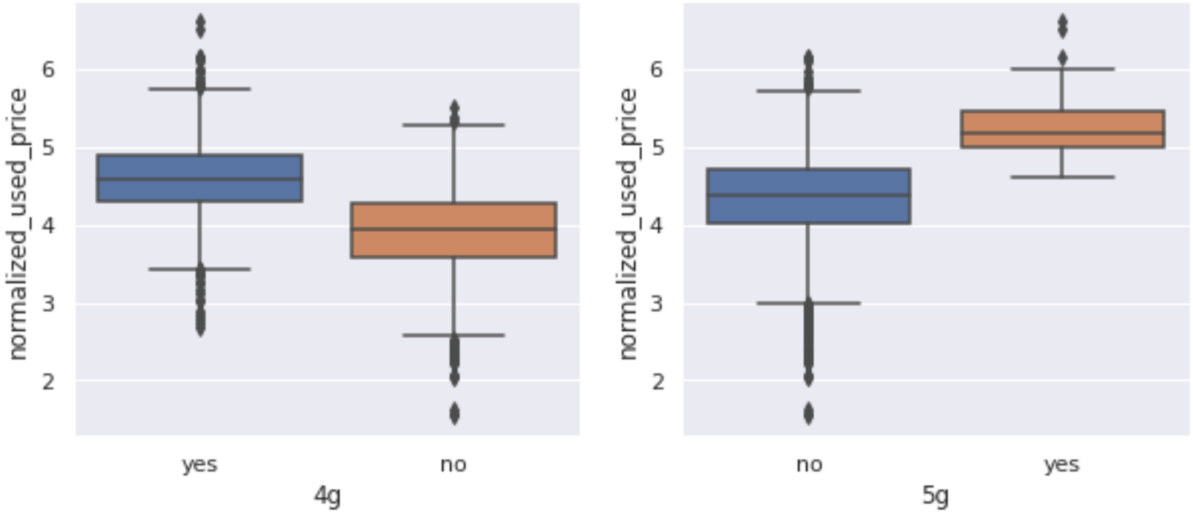
EDA Results



This graph shows us that as a device's release year gets more recent, the normalized used price increases. So this confirms that newer devices, even though they are used, are worth more than older devices.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

EDA Results



These two boxplots show that normalized used prices are higher for devices that are 4g than those that aren't. Further, those that are 5g are even higher usually.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Data Preprocessing

- In the dataset, there were several missing values for main camera quality and a few missing in selfie camera mp, internal memory, RAM, battery, and weight.
    - These missing values were all imputed with their respective medians
- There were outliers in all the data except for days used and years since release, however not treatment was needed for these.
- A new column was added for years since release to replace the dropped release year column, as shown below:

```
In [ ]: df1["years_since_release"] = 2021 - df1["release_year"]
        df1.drop("release_year", axis=1, inplace=True)
        df1["years_since_release"].describe()

Out[ ]: count    3454.000000
        mean        5.034742
        std         2.298455
        min         1.000000
        25%         3.000000
        50%         5.500000
        75%         7.000000
        max         8.000000
        Name: years_since_release, dtype: float64
```

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Data Preprocessing

- To prepare the data for modeling, we added the intercept to the data as well as dummy variables

Out[ ]:

	const	screen_size	main_camera_mp	selfie_camera_mp	int_memory	ram	battery	weight	days_used	normalized_new_price	...	brand_name_Spice	brand_name_Vivo	brand
0	1.0	14.50	13.0	5.0	64.0	3.0	3020.0	146.0	127	4.715100	...	0	0	
1	1.0	17.30	13.0	16.0	128.0	8.0	4300.0	213.0	325	5.519018	...	0	0	
2	1.0	16.69	13.0	8.0	128.0	8.0	4200.0	213.0	162	5.884631	...	0	0	
3	1.0	25.50	13.0	8.0	64.0	6.0	7250.0	480.0	345	5.630961	...	0	0	
4	1.0	15.32	13.0	8.0	64.0	3.0	5000.0	185.0	293	4.947837	...	0	0	

5 rows × 49 columns

- Then, we split the data in a 7:3 ratio for train : test
  - The result was 2417 rows in training data and 1037 rows in testing data
- Now we are ready to start building our model

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Model Performance Summary

- The R-square of training data is 0.84 and both the train and test RMSE and MSE are comparable, so we can be confident that the model is not overfit.
- The MAE tells us that the model can predict the price of used devices within a mean actual error of 0.18 on the testing data.
- The MAPE of 4.50 for the test data means that we can predict within 4.50% of the price of used devices.

Training Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.229884	0.180326	0.844886	0.841675	4.326841

Test Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.238358	0.184749	0.842479	0.834659	4.501651

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Model Performance Summary

- To the right is the OLS Regression Results
- The condition number is large, 1.78e+05, which indicates that there might be strong multicollinearity in some data

OLS Regression Results

Dep. Variable:	normalized_used_price	R-squared:	0.845
Model:	OLS	Adj. R-squared:	0.842
Method:	Least Squares	F-statistic:	268.7
Date:	Fri, 26 Aug 2022	Prob (F-statistic):	0.00
Time:	07:39:53	Log-Likelihood:	123.85
No. Observations:	2417	AIC:	-149.7
Df Residuals:	2368	BIC:	134.0
Df Model:	48		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.3156	0.071	18.454	0.000	1.176	1.455
screen_size	0.0244	0.003	7.163	0.000	0.018	0.031
main_camera_mp	0.0208	0.002	13.848	0.000	0.018	0.024
selfie_camera_mp	0.0135	0.001	11.997	0.000	0.011	0.016
int_memory	0.0001	6.97e-05	1.651	0.099	-2.16e-05	0.000
ram	0.0230	0.005	4.451	0.000	0.013	0.033
battery	-1.689e-05	7.27e-06	-2.321	0.020	-3.12e-05	-2.62e-06
weight	0.0010	0.000	7.480	0.000	0.001	0.001
days_used	4.216e-05	3.09e-05	1.366	0.172	-1.84e-05	0.000
normalized_new_price	0.4311	0.012	35.147	0.000	0.407	0.455
years_since_release	-0.0237	0.005	-5.193	0.000	-0.033	-0.015
brand_name_Alcatel	0.0154	0.048	0.323	0.747	-0.078	0.109
brand_name_Apple	-0.0038	0.147	-0.026	0.980	-0.292	0.285
brand_name_Asus	0.0151	0.048	0.314	0.753	-0.079	0.109
brand_name_BlackBerry	-0.0300	0.070	-0.427	0.669	-0.168	0.108
brand_name_Celkon	-0.0468	0.066	-0.707	0.480	-0.177	0.083
brand_name_Coolpad	0.0209	0.073	0.287	0.774	-0.122	0.164
brand_name_Gionee	0.0448	0.058	0.775	0.438	-0.068	0.158
brand_name_Google	-0.0326	0.085	-0.385	0.700	-0.199	0.133
brand_name_HTC	-0.0130	0.048	-0.270	0.787	-0.108	0.081
brand_name_Honor	0.0317	0.049	0.644	0.520	-0.065	0.128
brand_name_Huawei	-0.0020	0.044	-0.046	0.964	-0.089	0.085
brand_name_Infinix	0.1633	0.093	1.752	0.080	-0.019	0.346
brand_name_Karbonn	0.0943	0.067	1.405	0.160	-0.037	0.226
brand_name_LG	-0.0132	0.045	-0.291	0.771	-0.102	0.076
brand_name_Lava	0.0332	0.062	0.533	0.594	-0.089	0.155
brand_name_Lenovo	0.0454	0.045	1.004	0.316	-0.043	0.134
brand_name_Meizu	-0.0129	0.056	-0.230	0.818	-0.123	0.097
brand_name_Micromax	-0.0337	0.048	-0.704	0.481	-0.128	0.060
brand_name_Microsoft	0.0952	0.088	1.078	0.281	-0.078	0.268
brand_name_Motorola	-0.0112	0.050	-0.226	0.821	-0.109	0.086
brand_name_Nokia	0.0719	0.052	1.387	0.166	-0.030	0.174
brand_name_OnePlus	0.0709	0.077	0.916	0.360	-0.081	0.223
brand_name_Oppo	0.0124	0.048	0.261	0.794	-0.081	0.106
brand_name_Others	-0.0080	0.042	-0.190	0.849	-0.091	0.075
brand_name_Panasonic	0.0563	0.056	1.008	0.314	-0.053	0.166
brand_name_Realme	0.0319	0.062	0.518	0.605	-0.089	0.153
brand_name_Samsung	-0.0313	0.043	-0.725	0.469	-0.116	0.053
brand_name_Sony	-0.0616	0.050	-1.220	0.223	-0.161	0.037
brand_name_Spice	-0.0147	0.063	-0.233	0.816	-0.139	0.109
brand_name_Vivo	-0.0154	0.048	-0.318	0.750	-0.110	0.080
brand_name_XOLO	0.0152	0.055	0.277	0.782	-0.092	0.123
brand_name_Xiaomi	0.0869	0.048	1.806	0.071	-0.007	0.181
brand_name_ZTE	-0.0057	0.047	-0.121	0.904	-0.099	0.087
os_Others	-0.0510	0.033	-1.555	0.120	-0.115	0.013
os_Windows	-0.0207	0.045	-0.459	0.646	-0.109	0.068
os_iOS	-0.0663	0.146	-0.453	0.651	-0.354	0.221
4g_yes	0.0528	0.016	3.326	0.001	0.022	0.084
5g_yes	-0.0714	0.031	-2.268	0.023	-0.133	-0.010

Omnibus:	223.612	Durbin-Watson:	1.910
Prob(Omnibus):	0.000	Jarque-Bera (JB):	422.275
Skew:	-0.620	Prob(JB):	2.01e-92
Kurtosis:	4.630	Cond. No.	1.78e+05

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Model Performance Summary

- To check for multicollinearity, we first checked the VIF of the features from training data, ignoring dummy variables and the intercept, and dropped those that were above 5.
- After doing this, we got the following adjusted R-squared and RMSE values for screen size and weight:

	col	Adj. R-squared after_dropping col	RMSE after dropping col
0	screen_size	0.838381	0.234703
1	weight	0.838071	0.234928

- There is no longer multicollinearity affected our model.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Model Performance Summary

- With our adjusted data, we did a new regression and got these values:

Training Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.23403	0.182751	0.83924	0.838235	4.395407

Test Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	0.241434	0.186649	0.838387	0.836013	4.556349

- Now, the adjusted R-square is 0.839, which means our model explains roughly 84% of the variance and the adjusted R-square of the first model was nearly the same so the variables we dropped were not affecting the model.
- We can accept this as our final model as it is not overfit.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Model Assumptions

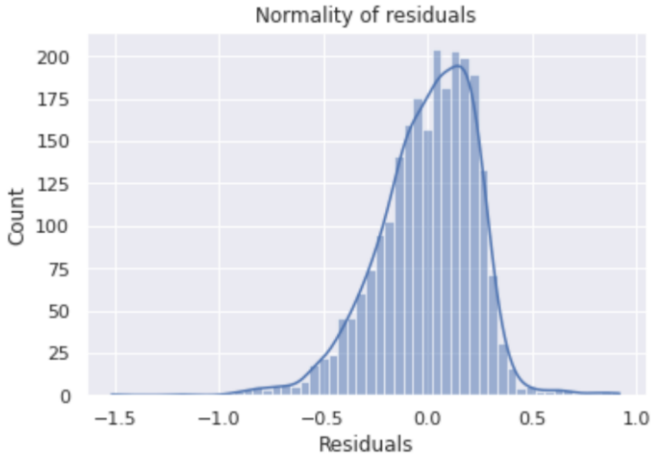
- Test for Linearity and Independence:

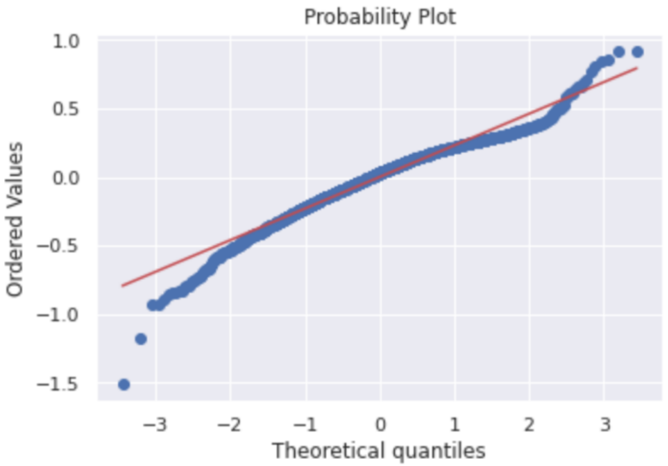
- Seeing as how there is no pattern in the residual plot above, we can assume that linearity and independence are satisfied for this model.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Model Assumptions

- Test for Normality:





- The normality of residuals plot appears to follow a normal distribution but could be slightly skewed to the left.
- The probability plot shows that the residuals follow a mostly straight line.
- The p value that we get from the Shapiro-Wilkes test is 6.995328206686811e-23, less than 0.5, so we can say the data follows a normal distribution.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Model Assumptions

- Test for Homoscedasticity:

```
Out[ ]:  [('F statistic', 1.008750419910676), ('p-value', 0.4401970650667301)]
```

- Since the p-value is less than 0.5, we can say that the residuals are homoscedastic.
- All of the assumptions for our model have been satisfied and it is the final model.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Happy Learning !