

ReCell - Case Study

Jesus Torres



A startup aiming to tap the potential of the used phones market



Provide cost-effective alternatives to both consumers and businesses



Forecast predicts that the used phone market would be worth \$52.7bn by 2023

OBJECTIVE

ReCell, a startup aiming to tap the potential in this market, has hired you as a data scientist. They want you to analyze the data provided and build a linear regression model to predict the price of a used phone and identify factors that significantly influence it.

- A few questions have been mentioned which will help you approach the analysis in the right manner and generate insights from the data.
- A thorough analysis of the data, in addition to the questions mentioned, should be done.

DATA OVERVIEW

The dataset contains the different data related to each user. The detailed data dictionary is given below:

- 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

DATA OVERVIEW (Contd.)

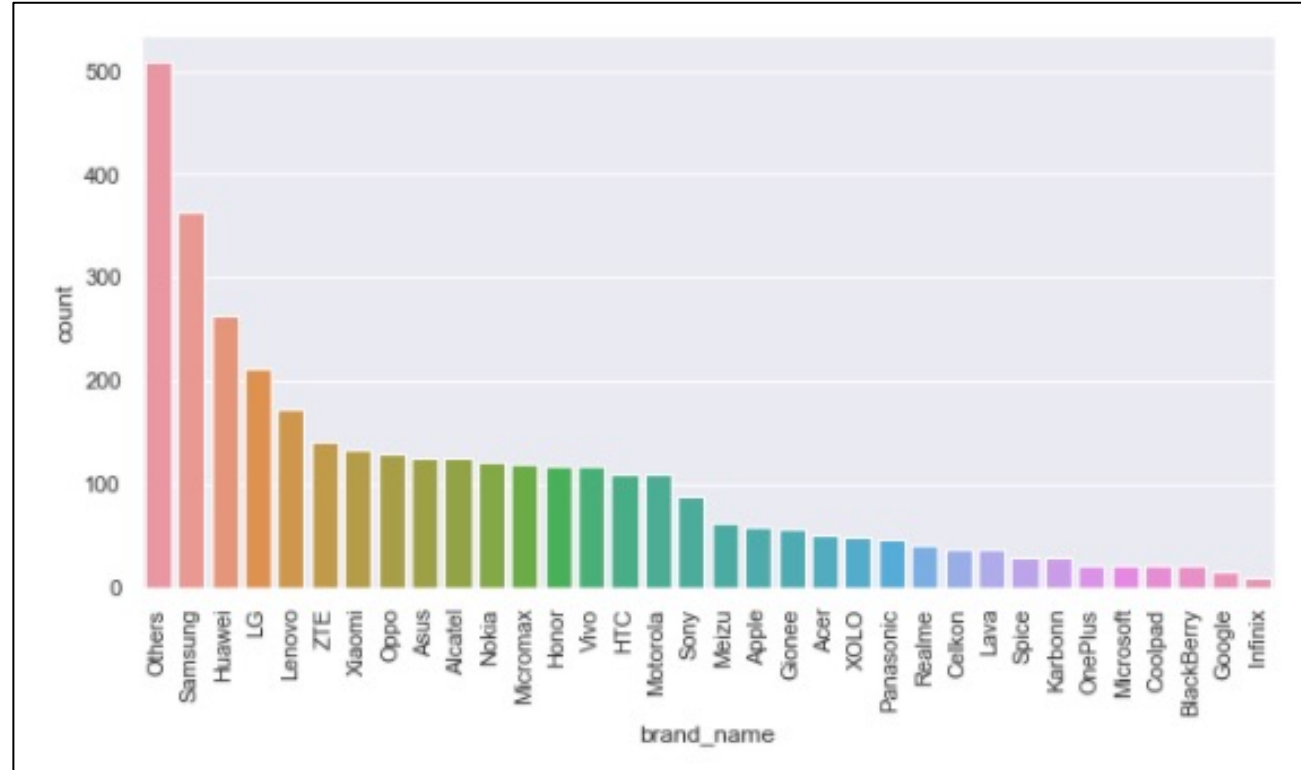
- ram: Amount of RAM in GB
- battery: Energy capacity of the phone battery in mAh
- weight: Weight of the phone in grams
- release_year: Year when the phone model was released
- days_used: Number of days the used/refurbished phone has been used
- new_price: Price of a new phone of the same model in euros
- used_price: Price of the used/refurbished phone in euros

Observations	Features
3571	15

- There are a total of 3571 observations.
- Each observation has 15 features.

EXPLORATORY DATA ANALYSIS

UNIVARIATE ANALYSIS – BRAND NAME

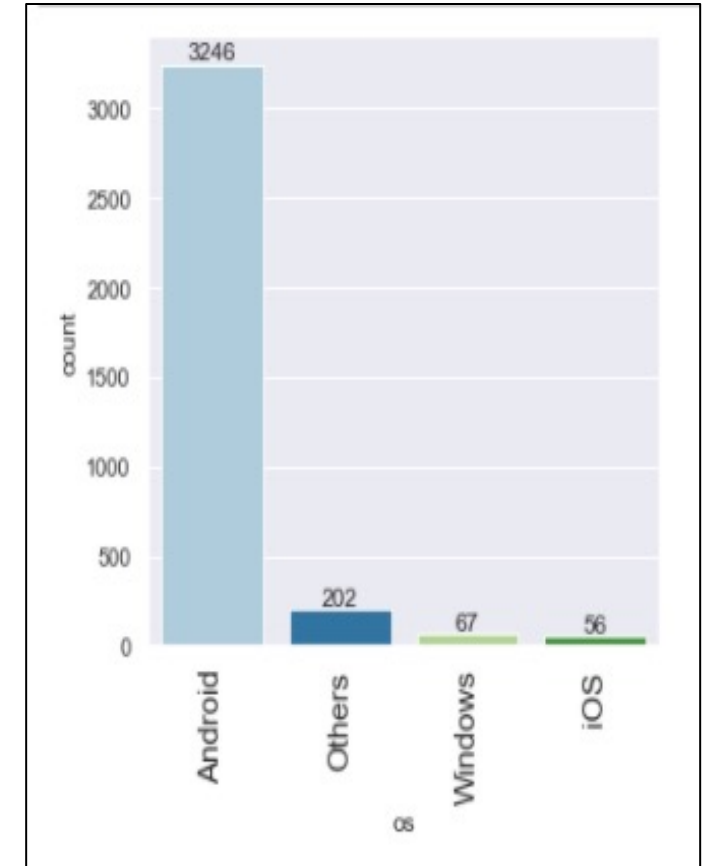


The category “Others” is the highest sold followed by brand names: Samsung, Huawei, LG, Lenovo,...

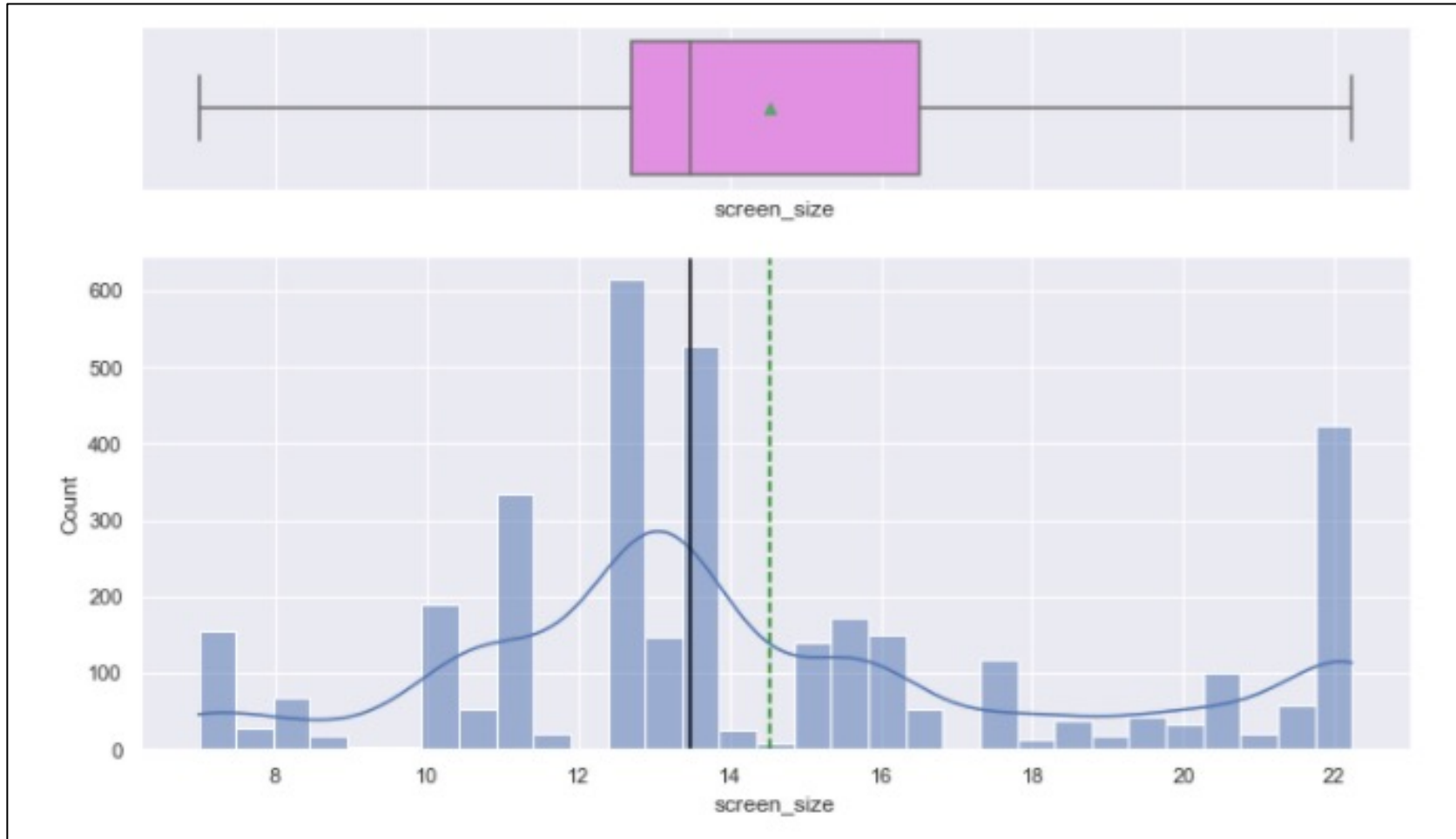
UNIVARIATE ANALYSIS – OS

Observations:

1. ReCell's market is mostly populated by Android OS
 - 3246 observations – 91%
2. Followed by:
 - Others – 6%
 - Windows – 2%
 - iOS – 1%



UNIVARIATE ANALYSIS – SCREEN SIZE



UNIVARIATE ANALYSIS – SCREEN SIZE

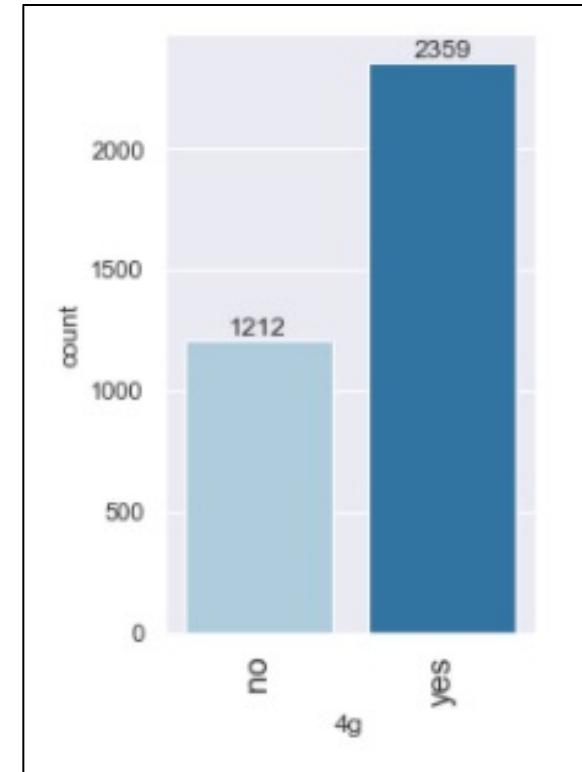
Observations:

1. The average "screen size" is about 14.5 centimeters
2. The standard deviation is about 4.2 centimeters
3. Inter-quartile range includes:
 - Min = 6.99
 - 25% quartile = 12.7
 - 50% quartile = 13.49
 - 75% quartile = 16.51
 - Max = 22.23
4. The data seems to follow a right-skewed distribution

UNIVARIATE ANALYSIS – 4G Capability

Observations:

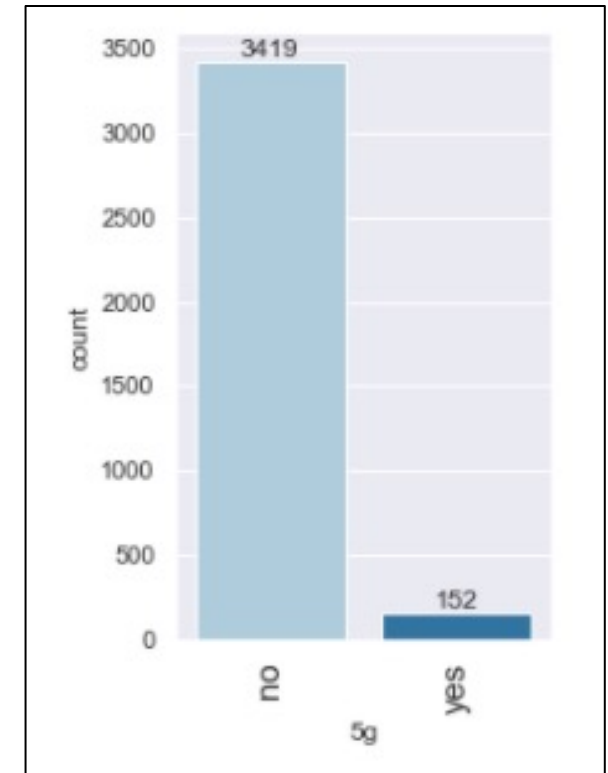
1. Two unique values, “No” and “Yes”
 - No – 1212 observations
 - Yes – 2359 observations
2. 66% of ReCell’s phone market has 4G capability



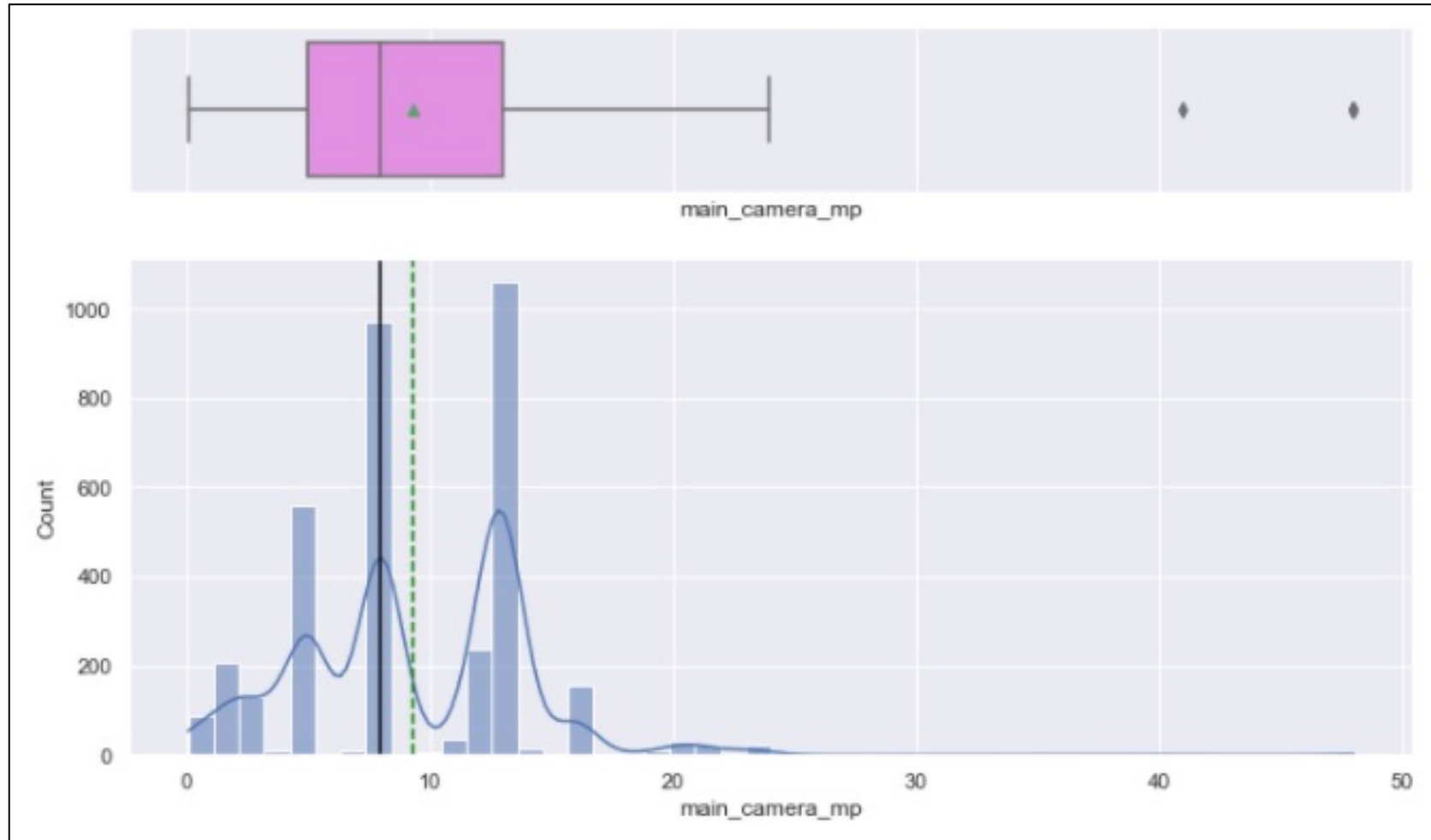
UNIVARIATE ANALYSIS – 5G Capability

Observations:

1. Two unique values, “No” and “Yes”
 - No – 3419 observations
 - Yes – 152 observations
2. Only 4% of ReCell’s phone market has 5G capability



UNIVARIATE ANALYSIS – MAIN CAMERA MP

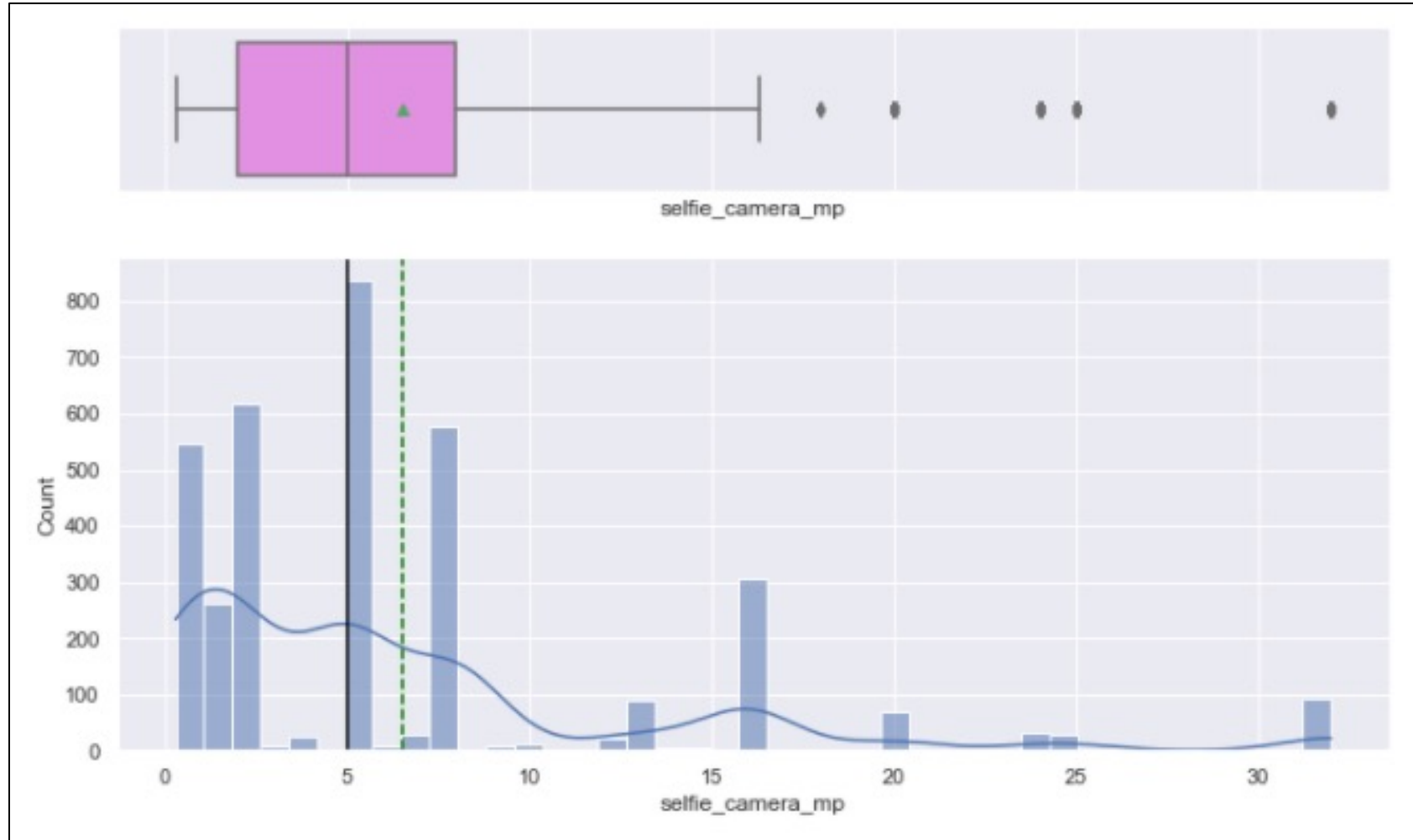


UNIVARIATE ANALYSIS – MAIN CAMERA MP

Observations:

1. The average is about 9.3 megapixels
2. The standard deviation is about 4.7 megapixels
3. Inter-quartile range includes:
 - Min = 0.08
 - 25% quartile = 5
 - 50% quartile = 8
 - 75% quartile = 13
 - Max = 48
4. The data seems to follow a right-skewed distribution

UNIVARIATE ANALYSIS – SELFIE CAMERA MP

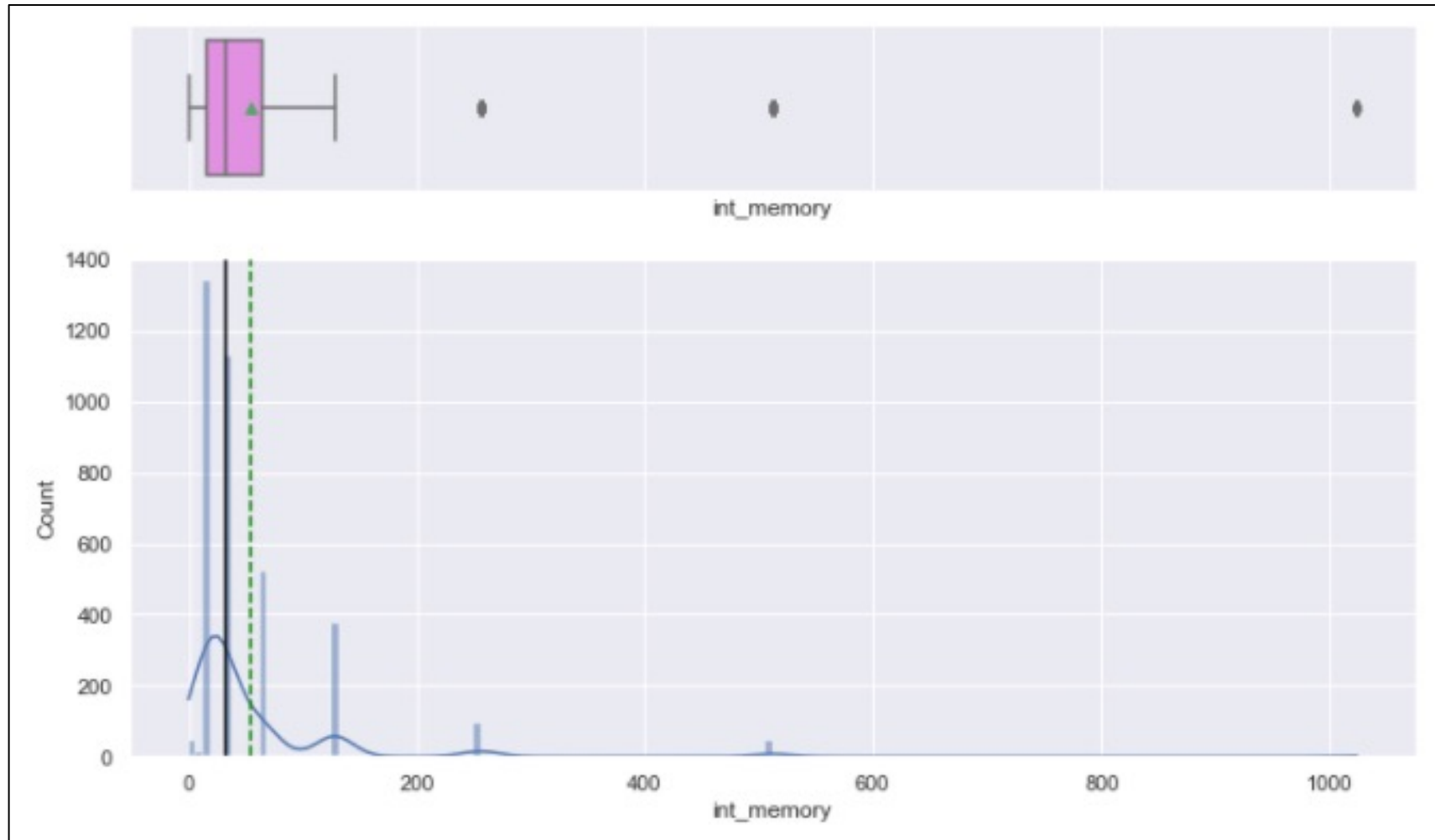


UNIVARIATE ANALYSIS – SELFIE CAMERA MP

Observations:

1. The average is about 6.5 megapixels
2. The standard deviation is about 6.9 megapixels
3. Inter-quartile range includes:
 - Min = 0.3
 - 25% quartile = 2
 - 50% quartile = 5
 - 75% quartile = 8
 - Max = 32
4. The data seems to follow a right-skewed distribution

UNIVARIATE ANALYSIS – INTERNAL MEMORY

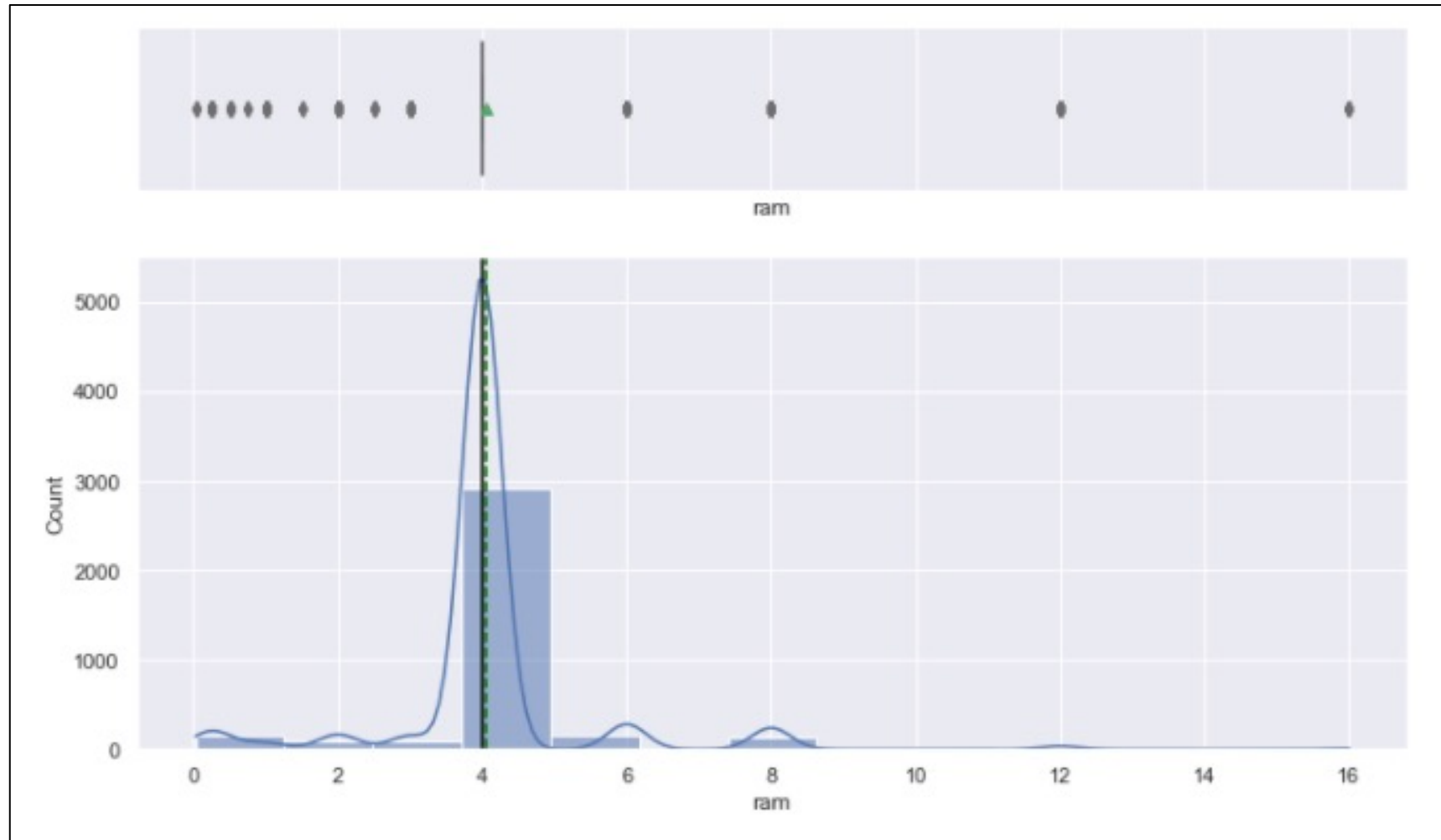


UNIVARIATE ANALYSIS – INTERNAL MEMORY

Observations:

1. The average is about 54.5 GB
2. The standard deviation is about 84.6 GB
3. Inter-quartile range includes:
 - Min = 0.005
 - 25% quartile = 16
 - 50% quartile = 32
 - 75% quartile = 64
 - Max = 1024
4. The data seems to follow a right-skewed distribution

UNIVARIATE ANALYSIS – RAM

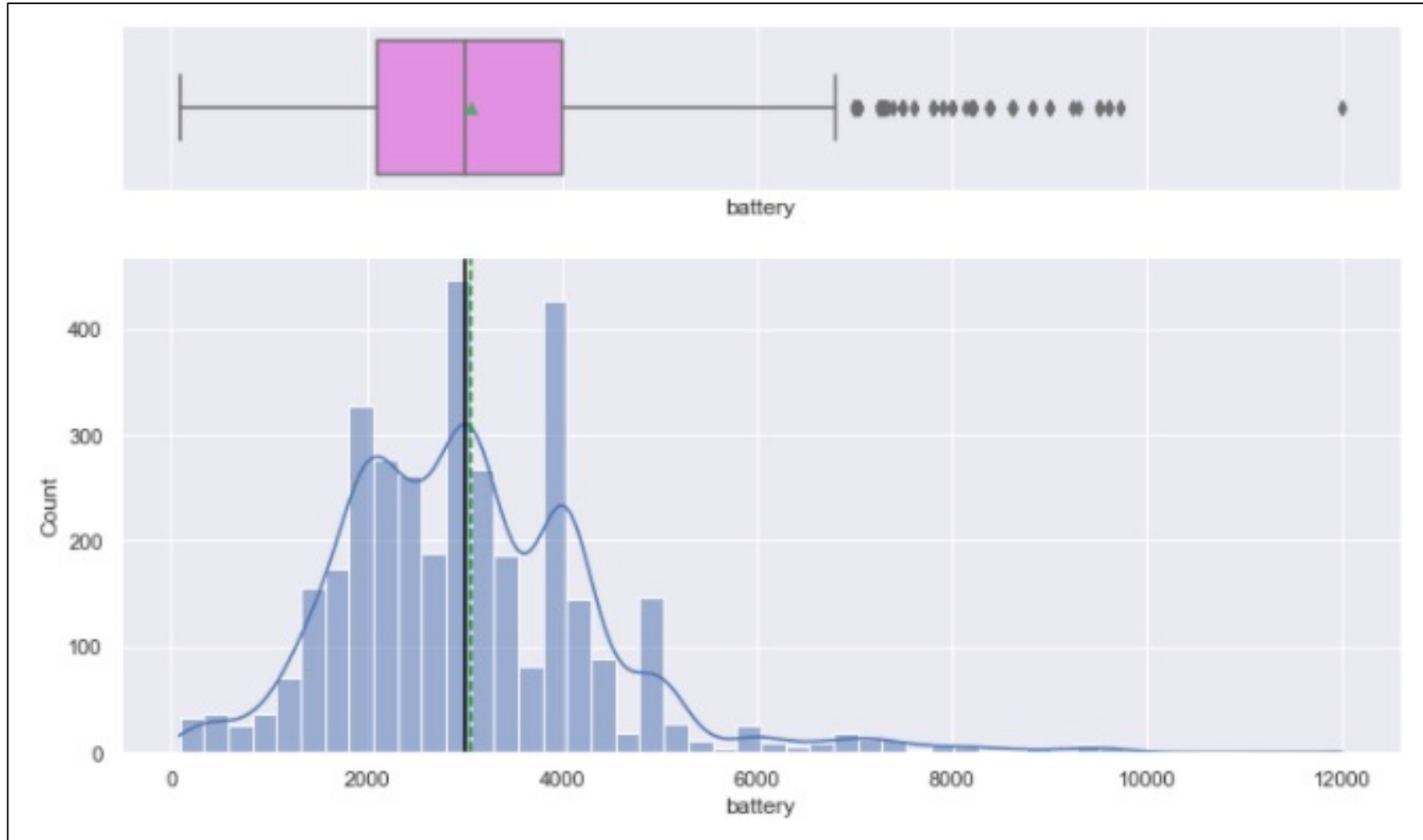


UNIVARIATE ANALYSIS – RAM

Observations:

1. The average is about 4.1 GB
2. The standard deviation is about 1.4 GB
3. Inter-quartile range includes:
 - Min = 0.03
 - 25% quartile = 4
 - 50% quartile = 4
 - 75% quartile = 4
 - Max = 16
4. The data seems to follow a normal distribution

UNIVARIATE ANALYSIS – BATTERY

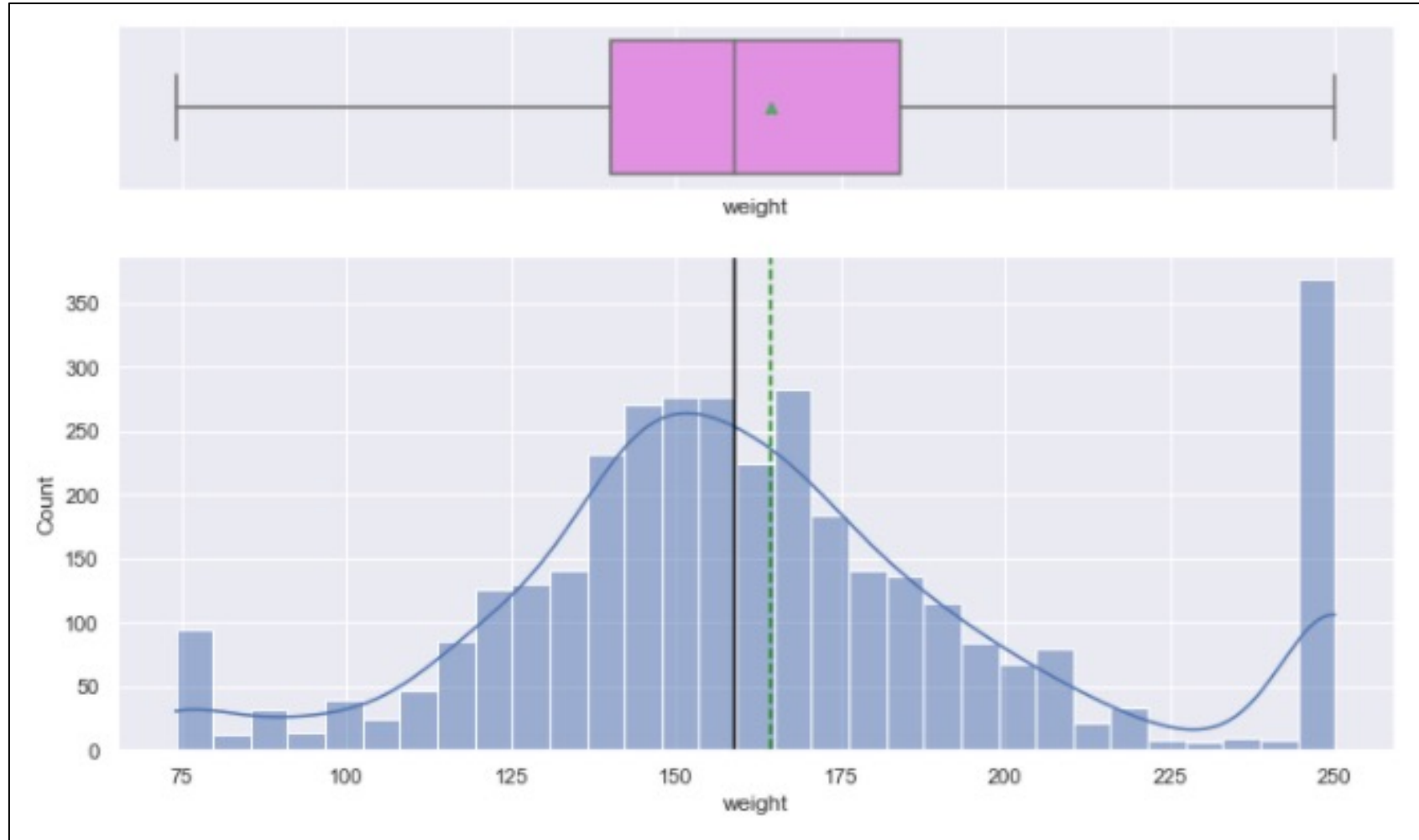


UNIVARIATE ANALYSIS – BATTERY

Observations:

1. The average is about 3067 mAh
2. The standard deviation is about 1363 mAh
3. Inter-quartile range includes:
 - Min = 80
 - 25% quartile = 2100
 - 50% quartile = 3000
 - 75% quartile = 4000
 - Max = 12000
4. The data seems to follow a slight right-skewed distribution

UNIVARIATE ANALYSIS – WEIGHT

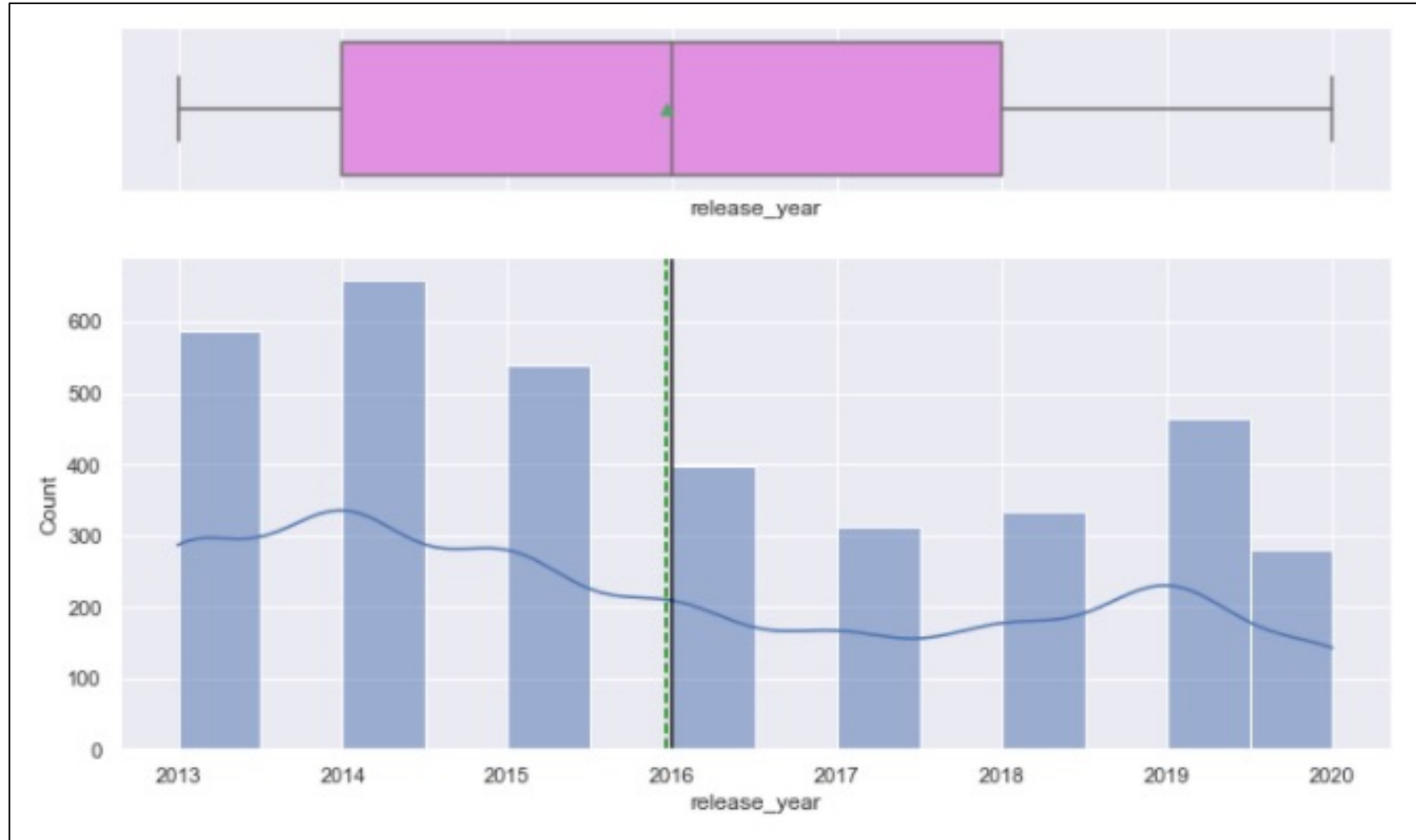


UNIVARIATE ANALYSIS – WEIGHT

Observations:

1. The average is about 164 grams
2. The standard deviation is about 41 grams
3. Inter-quartile range includes:
 - Min = 74
 - 25% quartile = 140
 - 50% quartile = 159
 - 75% quartile = 184
 - Max = 250
4. The data appears to follow a normal distribution

UNIVARIATE ANALYSIS – RELEASE YEAR

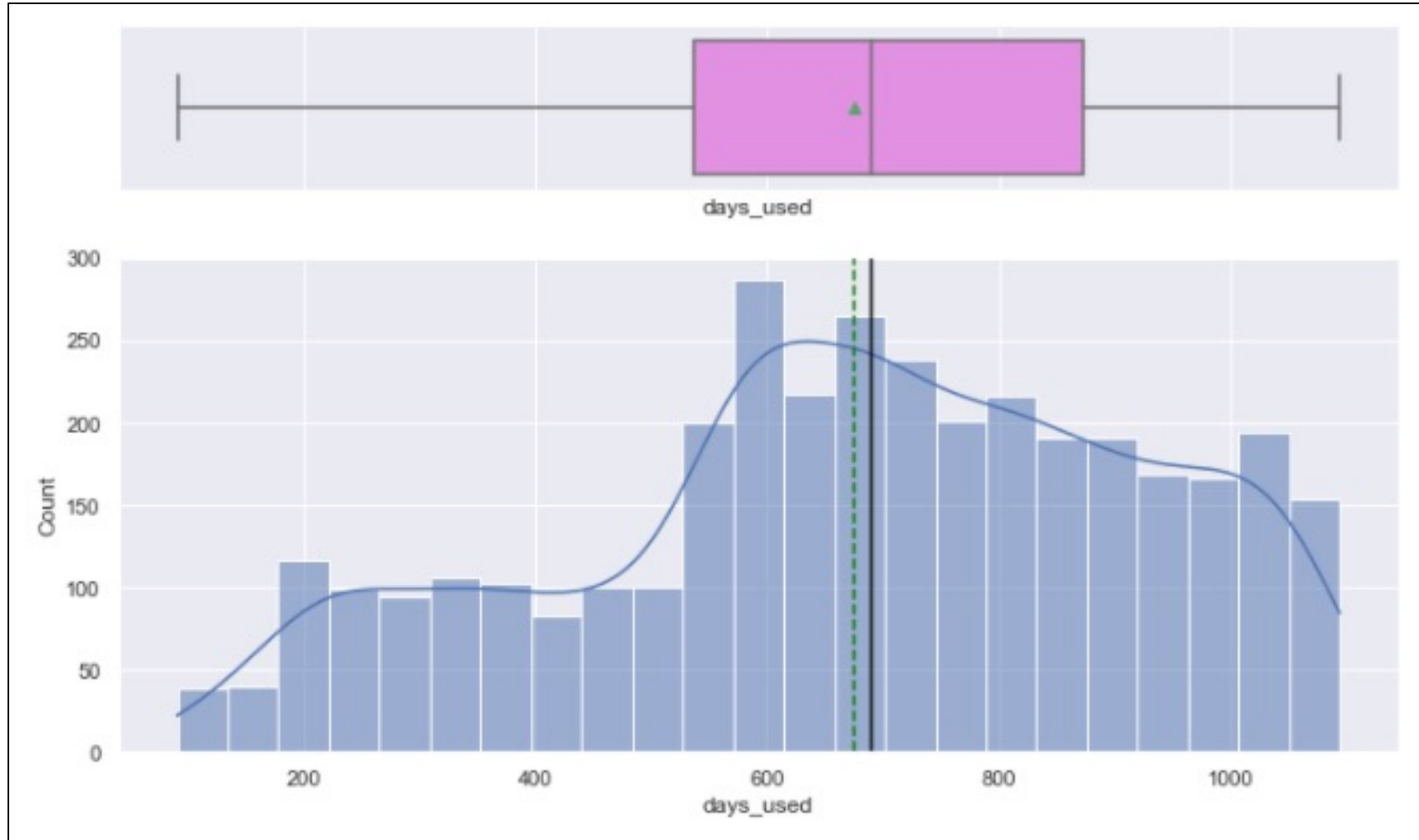


UNIVARIATE ANALYSIS – RELEASE YEAR

Observations:

1. The average is about 2016 (year)
2. The standard deviation is about 2.3 years
3. Inter-quartile range includes:
 - Min = 2013
 - 25% quartile = 2014
 - 50% quartile = 2016
 - 75% quartile = 2018
 - Max = 2020
4. The data seems slightly right-skewed with newer phones coming to the market

UNIVARIATE ANALYSIS – DAYS USED

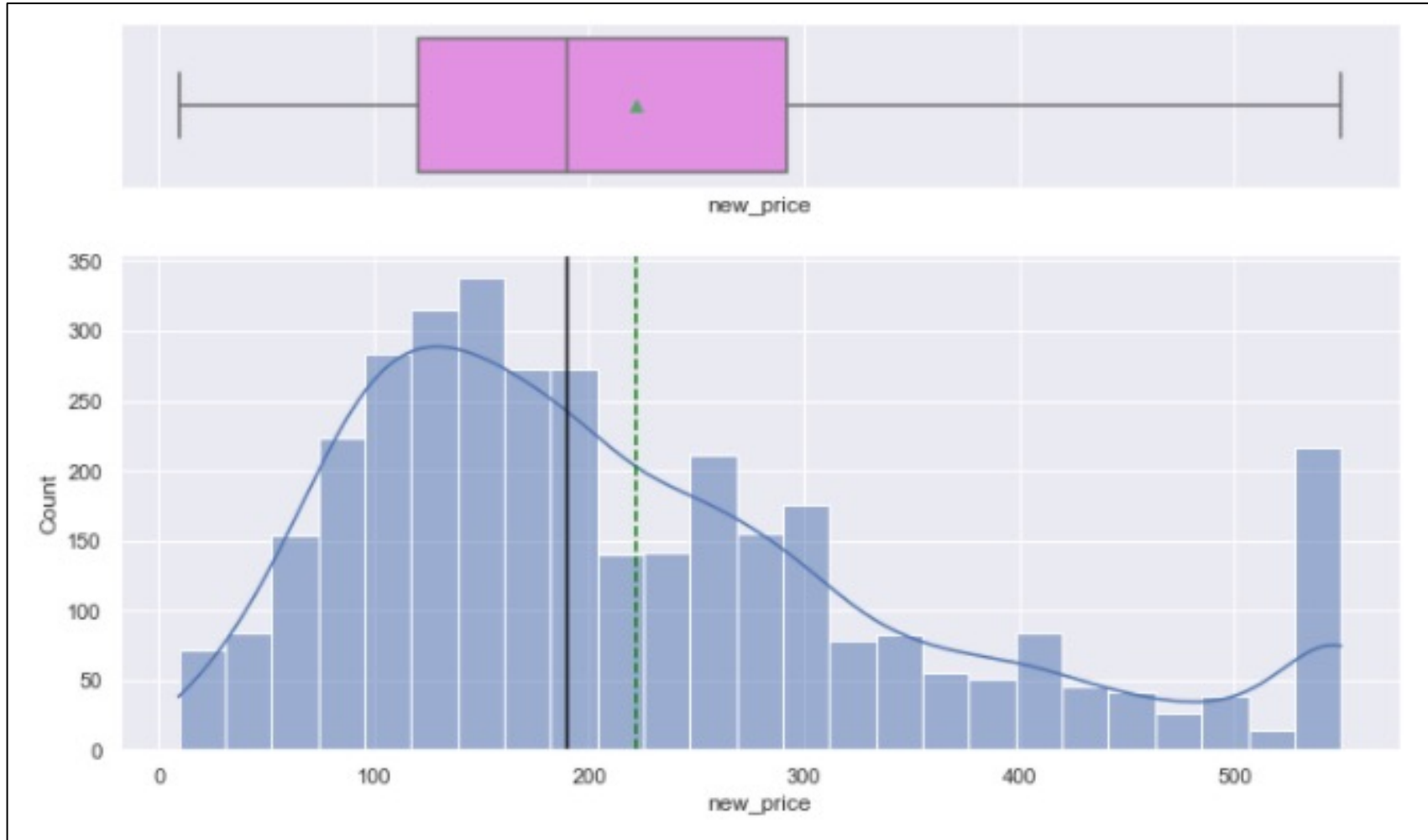


UNIVARIATE ANALYSIS – DAYS USED

Observations:

1. The average is about 675 days
2. The standard deviation is about 248 days
3. Inter-quartile range includes:
 - Min = 91
 - 25% quartile = 536
 - 50% quartile = 690
 - 75% quartile = 872
 - Max = 1094
4. The data seems to follow a left-skewed distribution indicating a high population of phones with more used days on them

UNIVARIATE ANALYSIS – NEW PRICE

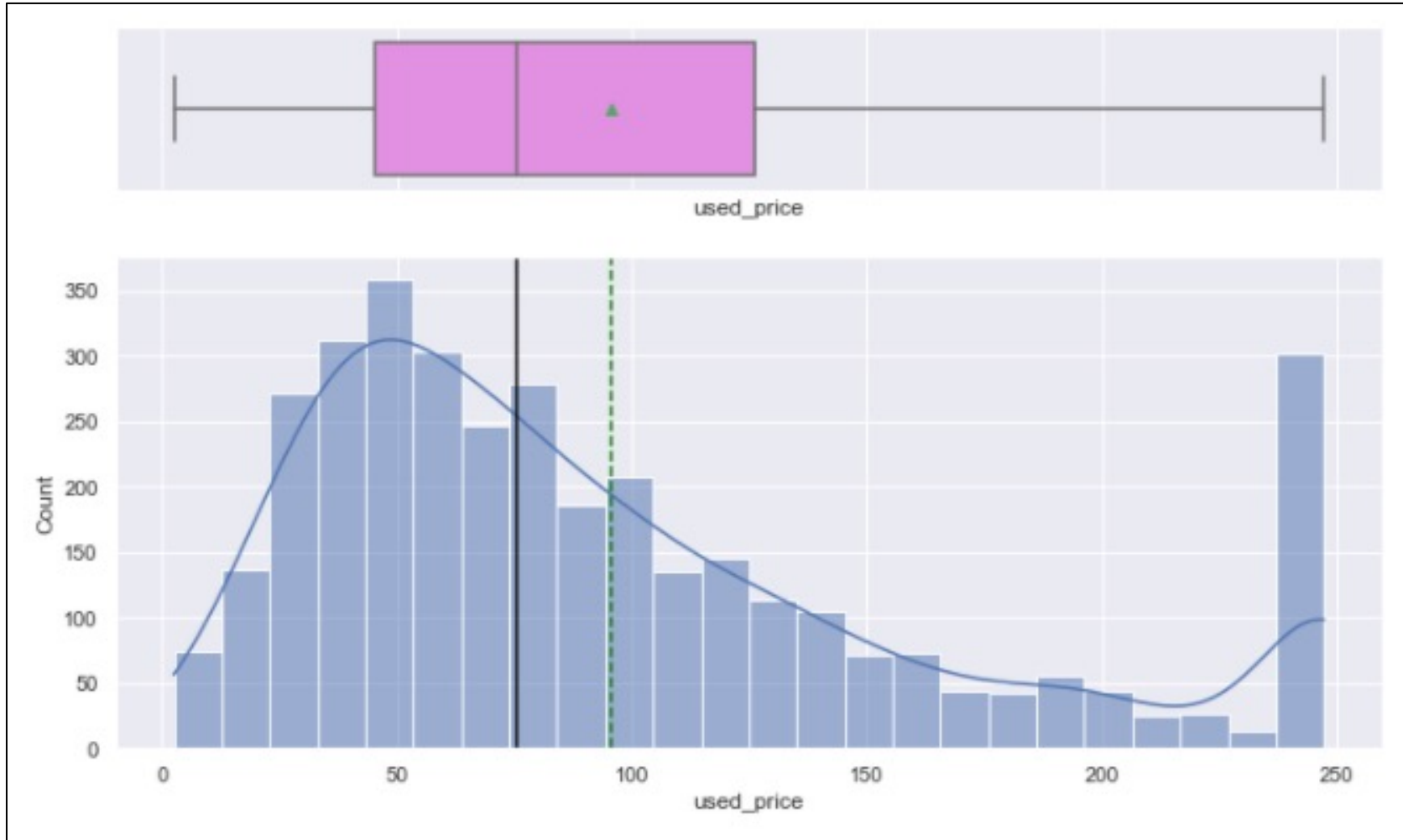


UNIVARIATE ANALYSIS – NEW PRICE

Observations:

1. The average is about 222 euros
2. The standard deviation is about 135 euros
3. Inter-quartile range includes:
 - Min = 9
 - 25% quartile = 120
 - 50% quartile = 190
 - 75% quartile = 292
 - Max = 550
4. The data seems to follow a slight right-skewed distribution

UNIVARIATE ANALYSIS – USED PRICE

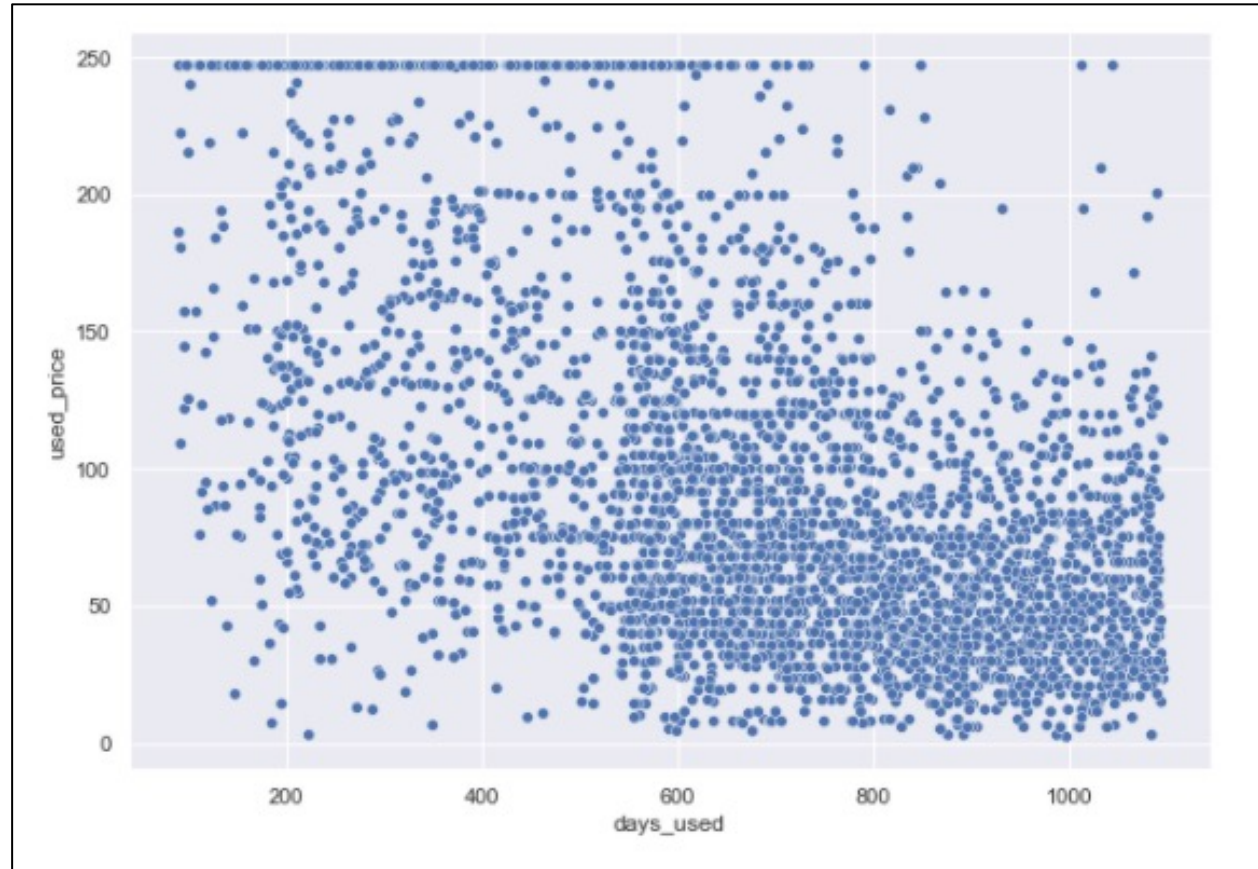


UNIVARIATE ANALYSIS – USED PRICE

Observations:

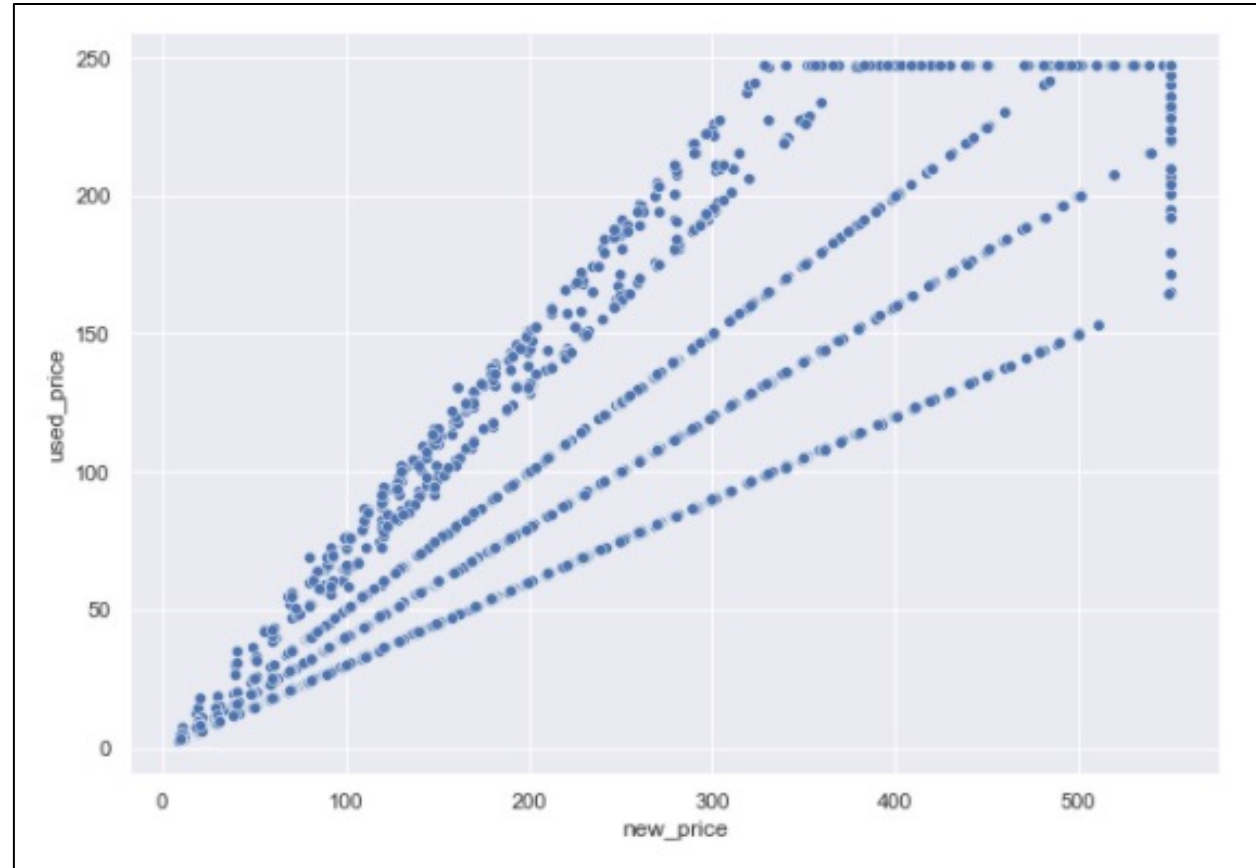
1. The average is about 96 euros
2. The standard deviation is about 66 euros
3. Inter-quartile range includes:
 - Min = 3
 - 25% quartile = 45
 - 50% quartile = 76
 - 75% quartile = 126
 - Max = 247
4. The data seems to follow a slight right-skewed distribution

BIVARIATE ANALYSIS – USED PRICE vs DAYS USED



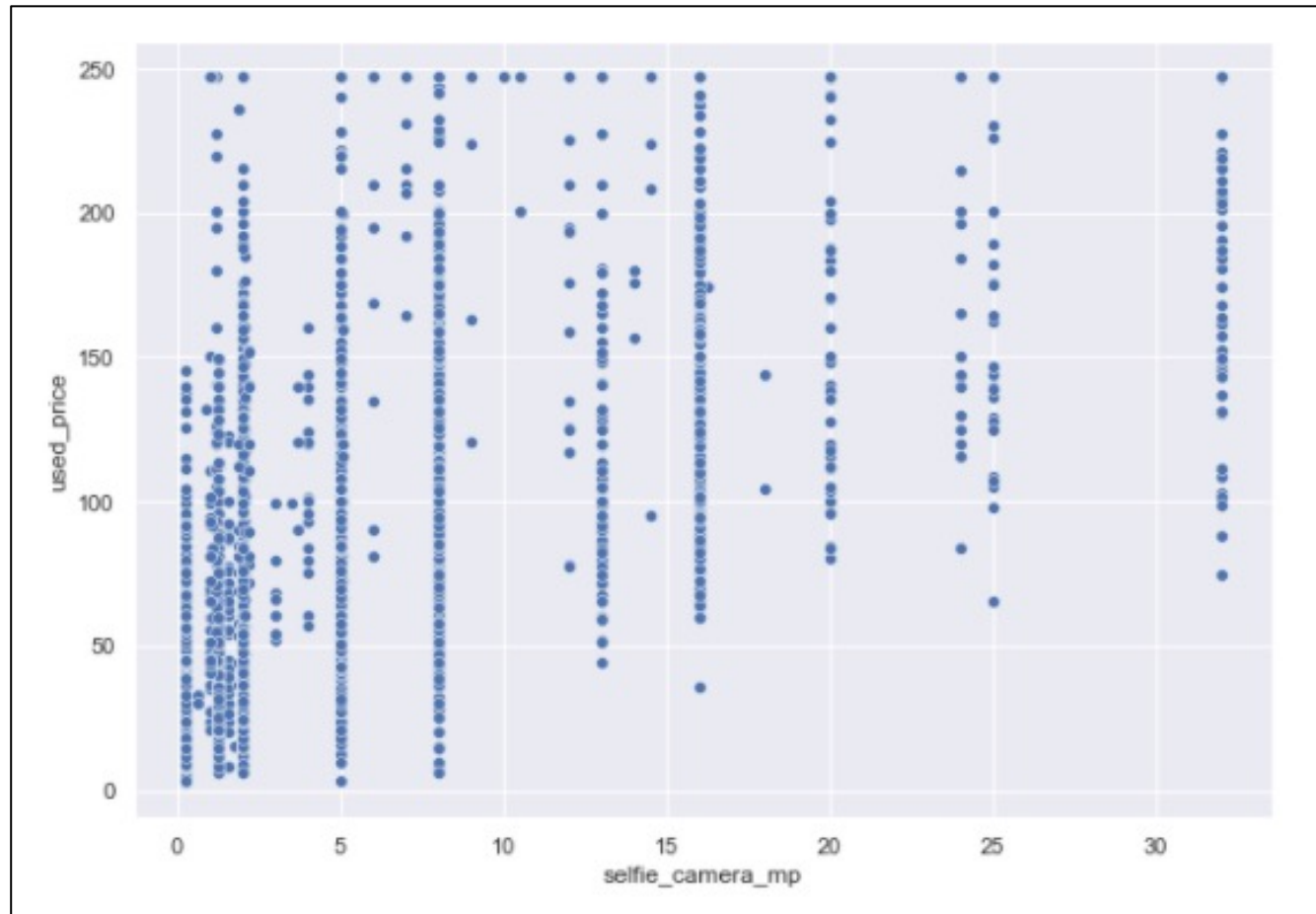
A negative trend can be seen, as the number of days a phone is used increases its used price decreases.

BIVARIATE ANALYSIS – USED PRICE vs NEW PRICE



A positive trend is clearly defined in the graph, as a phone's new price increases so does the used price.

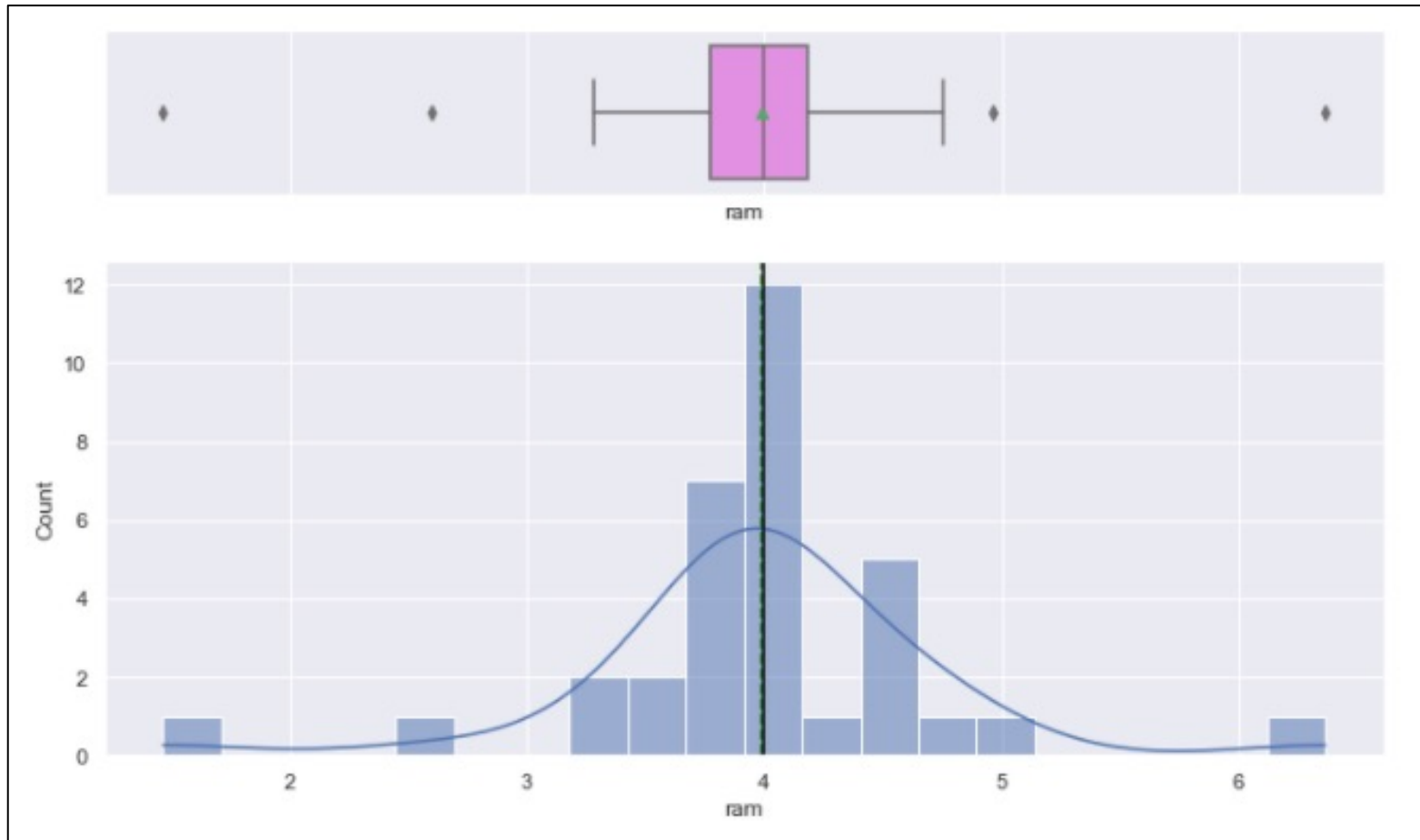
BIVARIATE ANALYSIS – USED PRICE vs SELFIE CAMERA MP



There seems to be a positive trend, as the number of megapixels in the selfie camera increases so does the used price.

BIVARIATE ANALYSIS – RAM vs BRAND NAME

Q: How does the amount of RAM vary with the brand?



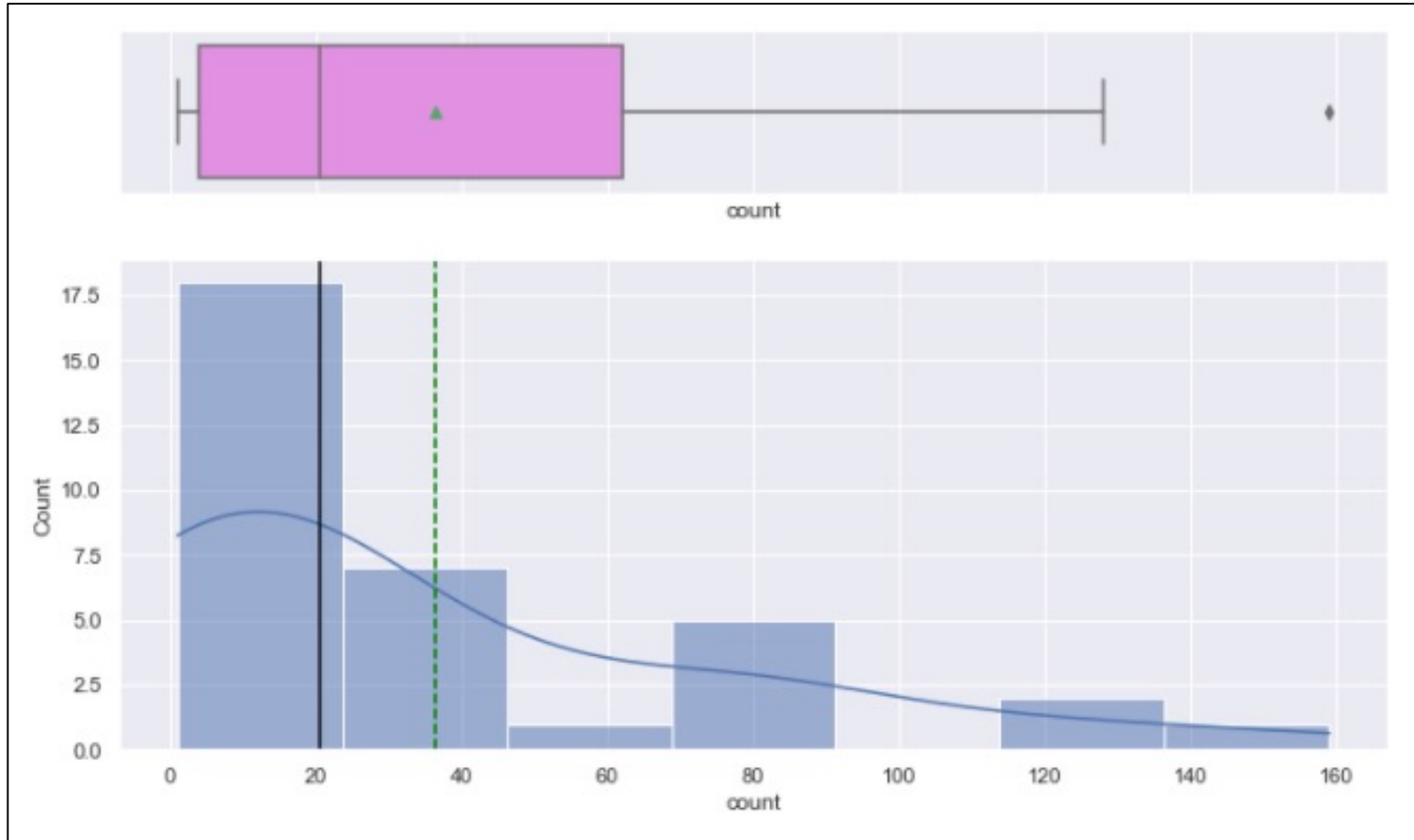
BIVARIATE ANALYSIS – RAM vs BRAND NAME

Observations:

1. The average is about 4 GB
2. The standard deviation is about 0.8 GB
3. Inter-quartile range includes:
 - Min = 1.5
 - 25% quartile = 3.8
 - 50% quartile = 4
 - 75% quartile = 4.2
 - Max = 6.4
4. The data seems to follow a normal distribution

BIVARIATE ANALYSIS – SCREEN SIZE vs BRAND NAME

Q: How many phones are available across different brands with a screen size larger than 6 inches?



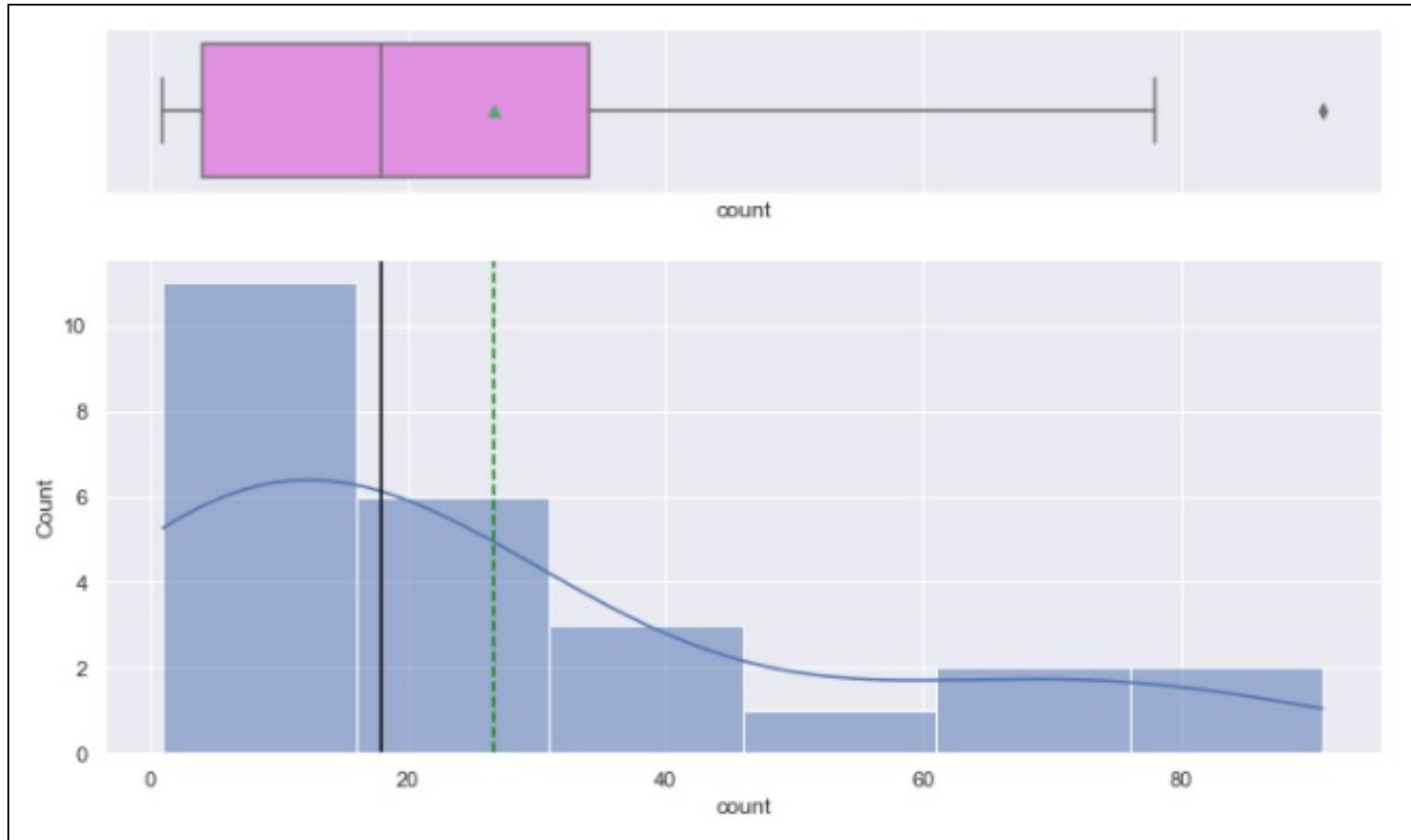
BIVARIATE ANALYSIS – SCREEN SIZE vs BRAND NAME

Observations:

1. The average is about 36 phones per brand
2. The standard deviation is about 42 phones per brand
3. Inter-quartile range includes:
 - Min = 1
 - 25% quartile = 4
 - 50% quartile = 21
 - 75% quartile = 62
 - Max = 159
4. The data seems to follow a right-skewed distribution

BIVARIATE ANALYSIS – SELFIE CAMERA MP vs BRAND NAME

Q:What is the distribution of budget phones offering greater than 8MP selfie cameras across brands?



BIVARIATE ANALYSIS – SELFIE CAMERA MP vs BRAND NAME

Observations:

1. The average is about 27 MP per brand
2. The standard deviation is about 27 MP per brand
3. Inter-quartile range includes:
 - Min = 1
 - 25% quartile = 4
 - 50% quartile = 18
 - 75% quartile = 34
 - Max = 91
4. The data seems to follow a right-skewed distribution

STATISTICAL MODELING

INITIAL OLS REGRESSION MODEL

- Dependent variable:
used_price
- Number of variables: 47
- R-squared: 0.957
- Adj. R-squared: 0.956

OLS Regression Results						
Dep. Variable:	used_price	R-squared:	0.957			
Model:	OLS	Adj. R-squared:	0.956			
Method:	Least Squares	F-statistic:	1166.			
Date:	Thu, 21 Oct 2021	Prob (F-statistic):	0.00			
Time:	09:39:05	Log-Likelihood:	-10084.			
No. Observations:	2499	AIC:	2.026e+04			
Df Residuals:	2451	BIC:	2.054e+04			
Df Model:	47					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
screen_size	0.4038	0.127	3.186	0.001	0.155	0.652
main_camera_mp	-0.4386	0.091	-4.801	0.000	-0.618	-0.259
selfie_camera_mp	0.8373	0.102	8.227	0.000	0.638	1.037
int_memory	0.1063	0.011	9.862	0.000	0.085	0.127
ram	162.6711	136.515	1.192	0.234	-105.025	430.367
battery	-0.0001	0.000	-0.272	0.786	-0.001	0.001
weight	-0.0159	0.012	-1.342	0.180	-0.039	0.007
release_year	-0.2928	0.271	-1.081	0.280	-0.824	0.238
days_used	-0.0845	0.002	-46.952	0.000	-0.088	-0.081
new_price	0.3799	0.003	113.361	0.000	0.373	0.386
brand_name_Alcatel	1.8448	2.856	0.646	0.518	-3.757	7.446
brand_name_Apple	-13.7682	10.303	-1.336	0.182	-33.971	6.434
brand_name_Asus	2.4908	2.779	0.896	0.370	-2.958	7.940
brand_name_BlackBerry	10.7808	4.281	2.518	0.012	2.385	19.176
brand_name_Celkon	-3.5464	3.845	-0.922	0.356	-11.086	3.993
brand_name_Coolpad	2.4051	4.535	0.530	0.596	-6.488	11.298
brand_name_Gionee	-3.8041	3.221	-1.181	0.238	-10.120	2.512
brand_name_Google	16.3764	5.229	3.132	0.002	6.124	26.629
brand_name_HTC	0.7516	2.844	0.264	0.792	-4.825	6.328
brand_name_Honor	-0.4021	2.904	-0.138	0.890	-6.097	5.292
brand_name_Huawei	0.3240	2.601	0.125	0.901	-4.777	5.425
brand_name_Infinix	-15.9799	5.506	-2.902	0.004	-26.776	-5.184
brand_name_Karbonn	-0.6277	3.746	-0.168	0.867	-7.973	6.717
brand_name_LG	2.8820	2.639	1.092	0.275	-2.292	8.056
brand_name_Lava	0.8713	3.689	0.236	0.813	-6.363	8.106
brand_name_Lenovo	-2.0348	2.678	-0.760	0.447	-7.287	3.217
brand_name_Meizu	-0.3518	3.235	-0.109	0.913	-6.696	5.992
brand_name_Micromax	3.4311	2.822	1.216	0.224	-2.102	8.965
brand_name_Microsoft	1.8578	5.047	0.368	0.713	-8.040	11.755
brand_name_Motorola	-1.7627	2.912	-0.605	0.545	-7.473	3.948
brand_name_Nokia	-7.4163	2.883	-2.572	0.010	-13.070	-1.763
brand_name_OnePlus	-15.3973	4.077	-3.776	0.000	-23.392	-7.402
brand_name_Oppo	-1.7509	2.823	-0.620	0.535	-7.286	3.784
brand_name_Others	0.9510	2.461	0.386	0.699	-3.876	5.778
brand_name_Panasonic	-1.4156	3.389	-0.418	0.676	-8.061	5.230
brand_name_Realme	-2.0660	3.573	-0.578	0.563	-9.072	4.940
brand_name_Samsung	1.3967	2.531	0.552	0.581	-3.566	6.359
brand_name_Sony	4.1467	2.980	1.392	0.164	-1.696	9.990
brand_name_Spice	4.9620	3.607	1.376	0.169	-2.111	12.035
brand_name_Vivo	1.0232	2.896	0.353	0.724	-4.656	6.703
brand_name_XOLO	3.2719	3.294	0.993	0.321	-3.188	9.731
brand_name_Xiaomi	-1.3916	2.775	-0.501	0.616	-6.833	4.050
brand_name_ZTE	0.9376	2.795	0.335	0.737	-4.542	6.418
os_Others	-4.1035	1.528	-2.686	0.007	-7.100	-1.107
os_Windows	0.9183	2.653	0.346	0.729	-4.284	6.121
os_iOS	24.7206	10.221	2.419	0.016	4.679	44.763
4g_yes	-1.6862	0.909	-1.856	0.064	-3.468	0.096
5g_yes	2.4689	1.680	1.470	0.142	-0.825	5.763
Omnibus:	233.732	Durbin-Watson:	1.988			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	519.521			
Skew:	0.575	Prob(JB):	1.54e-113			
Kurtosis:	4.915	Cond. No.	1.92e+06			

FINAL OLS REGRESSION MODEL

- Dependent variable:
used_price
- Number of variables: 17
 - All with a p -value < 0.05
- R-squared: 0.956
- Adj. R-squared: 0.956

OLS Regression Results						
=====						
Dep. Variable:	used_price	R-squared:	0.956			
Model:	OLS	Adj. R-squared:	0.956			
Method:	Least Squares	F-statistic:	3192.			
Date:	Wed, 20 Oct 2021	Prob (F-statistic):	0.00			
Time:	18:22:38	Log-Likelihood:	-10110.			
No. Observations:	2499	AIC:	2.026e+04			
Df Residuals:	2481	BIC:	2.036e+04			
Df Model:	17					
Covariance Type:	nonrobust					
=====						
	coef	std err	t	P> t	[0.025	0.975]

const	58.8565	2.042	28.826	0.000	54.853	62.859
screen_size	0.3066	0.082	3.734	0.000	0.146	0.467
main_camera_mp	-0.4130	0.078	-5.300	0.000	-0.566	-0.260
selfie_camera_mp	0.8346	0.060	13.859	0.000	0.717	0.952
int_memory	0.0141	0.004	3.868	0.000	0.007	0.021
ram	0.7765	0.243	3.197	0.001	0.300	1.253
days_used	-0.0842	0.001	-57.611	0.000	-0.087	-0.081
new_price	0.3855	0.003	132.535	0.000	0.380	0.391
brand_name_BlackBerry	8.2550	3.502	2.357	0.018	1.387	15.123
brand_name_Gionee	-5.0398	2.224	-2.266	0.024	-9.402	-0.677
brand_name_Google	15.7137	4.658	3.374	0.001	6.581	24.846
brand_name_Infinix	-16.1283	4.966	-3.248	0.001	-25.867	-6.389
brand_name_Lenovo	-3.3736	1.317	-2.561	0.010	-5.957	-0.790
brand_name_Nokia	-9.1880	1.560	-5.890	0.000	-12.247	-6.129
brand_name_OnePlus	-12.5614	3.276	-3.834	0.000	-18.986	-6.136
brand_name_Oppo	-3.3216	1.509	-2.201	0.028	-6.280	-0.363
os_iOS	8.5067	2.333	3.646	0.000	3.931	13.082
4g_yes	-2.1445	0.781	-2.746	0.006	-3.676	-0.612
=====						
Omnibus:	240.425	Durbin-Watson:	1.980			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	565.793			
Skew:	0.572	Prob(JB):	1.38e-123			
Kurtosis:	5.031	Cond. No.	1.34e+04			
=====						

PREDICTIVE MODELING

LINEAR REGRESSION – TRAINING PERFORMANCE COMPARISON

	Linear Regression sklearn	Linear Regression statsmodels
RMSE	13.755476	13.827565
MAE	10.145893	10.195100
R-squared	0.956739	0.956284
Adj. R-squared	0.955891	0.955967
MAPE	17.988634	17.999734

LINEAR REGRESSION – TESTING PERFORMANCE COMPARISON

	Linear Regression sklearn	Linear Regression statsmodels
RMSE	14.507547	14.279404
MAE	10.414655	10.292618
R-squared	0.951889	0.953390
Adj. R-squared	0.949631	0.952593
MAPE	18.448138	18.231850

LINEAR REGRESSION (Contd.)

The performance of both models are significantly comparable.

- The model can explain ~95% of the variation in the data.
- The training and testing RMSE and MAE are low and comparable. This indicates that the model is neither underfitted nor overfitted.
- The MAPE on the testing set suggests we can predict within 18.2 euros of a used phone's price.
- Hence, we can conclude the model `olsmod2` is good for prediction as well as inference purposes.

FINAL MODEL SUMMARY

- screen_size is statistically significant; it has a positive correlation. Therefore, as the screen size of a used phone increases so does the used price.
- Camera megapixels for used phones is split.
 - main_camera_mp is negatively correlated indicating that as the megapixels of the main camera setup increase by one unit, the used price decreases by a factor of ~ 0.413 euros.
 - On the other hand, selfie_camera_setup is positively correlated. For every one unit increase in megapixels of the selfie camera setup, the used price increases by a factor of ~ 0.8346 euros.
- int_memory and ram are both positively correlated however a used phone's ram size is valued more than its internal memory.
 - One unit increase in ram increases the used price by about 0.7765 euros.
 - One unit increase in int_memory increases the used price by about 0.0141 euros.

FINAL MODEL SUMMARY (Contd.)

- new_price is positively correlated so as a used phone's original price increases, used price also increases.
- A used phone's brand name has both negative and positive correlations.
 - Positively correlated brand names in descending order:
 - brand_name_Google valued at ~15.71 euros followed by os_iOS which are Apple iPhones valued at ~8.51 euros and brand_name_BlackBerry equal to ~8.26 euros.
 - Negatively correlated brand names in descending order:
 - brand_name_Infinix decreases used price by a factor of ~16.13 euros. Followed by brand names OnePlus, Nokia, Gionee, Lenovo, and Oppo.
- Finally, 4g_yes or 4g capability decreases the used price by a factor of ~2.14 euros

RECOMMENDATIONS

Suggestion A: To increase monetary for POSITIVELY correlated brand names, the number of sales must increase and ideally outperform those of NEGATIVELY correlated brand names.

Method 1: Identify the supply and demand factor of the positively correlated brand names.

If the customer does demand these specific brands, then increase supply.

If customers are not in high demand of these brands, then develop some marketing strategy to encourage customers to buy from these high profit brands.

Method 2: Increase the customer base already buying from the low profit brand names.

Increase the number of sales of the lowest negatively correlated brands.

Decrease maybe even discontinue sales of the highest negatively correlated brands if they are not bringing in a substantial profit.

RECOMMENDATIONS

Suggestion B: Increase the frequency of customers buying a phone with a screen size greater than 14.5 cm.

Method 1: Develop an ad campaign encouraging customers to buy a bigger screen size phone.

Method 2: For customers who do want a bigger screen size phone, increase monetary by offering them some accessories to buy along with the phone.

Suggestion C: Maximize the net profit of a used phones' camera system.

Method 1: Build up a bigger supply of used phones with more megapixels in the selfie camera.

Suggestion D: Maximize the net profit of a used phones' storage capacity.

Method 1: Branch out to other markets, used phones with more ram storage, since they significantly increase profit.

RECOMMENDATIONS

Suggestion E: The relationship between new price and days used of a phone is simple, depreciation.

Method 1: Increase the supply of newer phones and develop ad campaigns encouraging customers to buy a newer phone model.

Suggestion F: Most people expect 4g capability a standard now, which could explain the decrease in value.

Method 1: Initiate a campaign to supply more 5g enabled used phones and encourage customers to make the switch to 5g connectivity.