Business Update:
Used Cell Phone Price

Using data insights to predict the price of a used cell phone based on phone characteristics.
October 23, 2021

RE-CELL electronics

Today, we are

- **Presenting findings from data analysis of customer behavior on the new landing page**

Recommending whether to proceed with the implementation of the new landing page

The presentation is broken into the following areas:

- **Overview**
- **Executive Summary**
- **Data Analysis**

Executive Summary

The second-hand phone business/industry is poised for significant growth in the near to medium future with the IDC (Internation Data Corporation) predicting a whooping $52.7bn by 2023 with a compound annual growth rate (CAGR) of 13.6% from 2018 to 2023.

There are many advantages to using a used cell phone: i) significant cost savings with warranties, ii) increases the life of the cell phone

In order to take advantage of the market, the marketing team has tasked the Data group to develop a linear regression model that predicts the price of a used phone and identify factors that significantly influence the price

Prediction Model

```
The coefficient for screen_size is -0.5539539589279906
The coefficient for int_memory is 0.062497826962702945
The coefficient for ram is 3.6898285865112763
The coefficient for release_year is 1.6256236713451884
The coefficient for days_used is -0.11016940919607254
The coefficient for new_price is 0.5098751313475658
The coefficient for 4g_yes is -15.707796569849757
The coefficient for 5g_yes is 60.61976625334174
```

Factors that affect price

- **Connectivity (4g / 5g)**
- **Price of a new version of the phone**
- **Release year**
- **Days Used**
- **Memory (Internal and Ram)**

Original Data

Source Data

- The Data was provided by ReCell marketing group

Data Cleaning

- The data set was relatively intact but some data cleaning for required.
- A lot of of the missing data was filled from internet search information from a cell phone dealer.
- The cell phones were grouped by their brands and research was done based on brands and year of release.
- For example missing Realme data was filled by estimating features for the year of production.
- Missing data for things like Ram, the mean of the brand for that particular year was used.

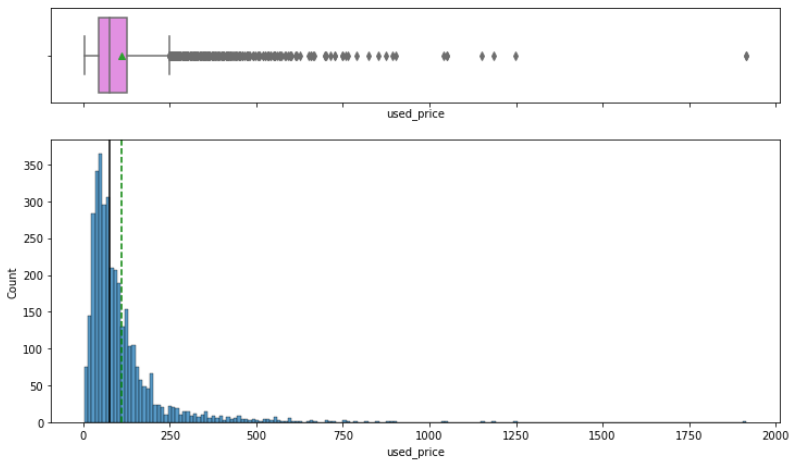
	brand_name	os	screen_size	4g	5g	main_camera_mp	selfie_camera_mp	int_memory	ram	battery	weight	release_year	days_used	new_price	used_price
2501	Samsung	Android	13.49	yes	no	13.0	13.0	32.0	4.00	3600.0	181.0	2017	683	198.680	79.47
2782	Sony	Android	13.81	yes	no	NaN	8.0	32.0	4.00	3300.0	156.0	2019	195	198.150	149.10
605	Others	Android	12.70	yes	no	8.0	5.0	16.0	4.00	2400.0	137.0	2015	1048	161.470	48.39
2923	Vivo	Android	19.37	yes	no	13.0	16.0	64.0	4.00	3260.0	149.3	2019	375	211.880	138.31
941	Others	Others	5.72	no	no	0.3	0.3	32.0	0.25	820.0	90.0	2013	883	29.810	8.92
1833	LG	Android	13.49	no	no	8.0	1.3	32.0	4.00	3140.0	161.0	2013	670	240.540	96.18
671	Apple	iOS	14.92	yes	no	12.0	7.0	64.0	4.00	5493.0	48.0	2018	403	700.150	350.08
1796	LG	Android	17.78	yes	no	5.0	0.3	16.0	4.00	4000.0	294.8	2014	708	189.300	75.94
757	Asus	Android	13.49	yes	no	13.0	8.0	32.0	4.00	5000.0	181.0	2017	612	270.500	108.13
3528	Realme	Android	15.72	yes	no	NaN	16.0	64.0	4.00	4035.0	184.0	2019	433	159.885	80.00

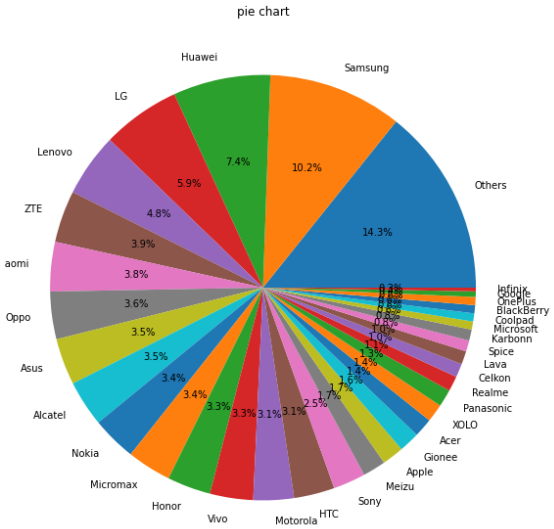
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3571 entries, 0 to 3570
Data columns (total 15 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   brand_name        3571 non-null   object
 1   os                3571 non-null   object
 2   screen_size       3571 non-null   float64
 3   4g                3571 non-null   object
 4   5g                3571 non-null   object
 5   main_camera_mp    3391 non-null   float64
 6   selfie_camera_mp  3569 non-null   float64
 7   int_memory        3561 non-null   float64
 8   ram               3561 non-null   float64
 9   battery           3565 non-null   float64
 10  weight            3564 non-null   float64
 11  release_year      3571 non-null   int64
 12  days_used         3571 non-null   int64
 13  new_price         3571 non-null   float64
 14  used_price        3571 non-null   float64
dtypes: float64(9), int64(2), object(4)
memory usage: 418.6+ KB
```

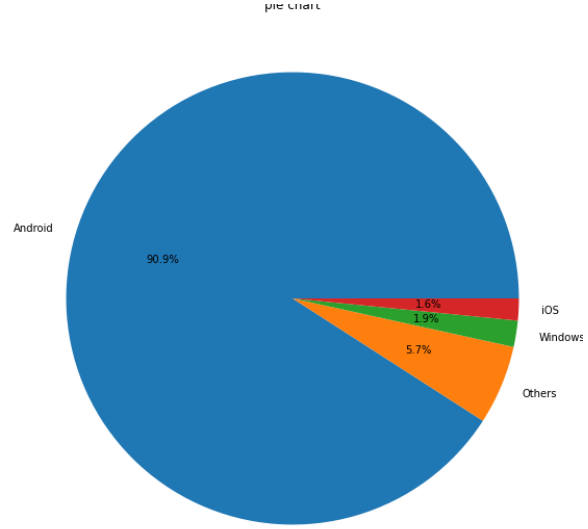
Univariate Analysis

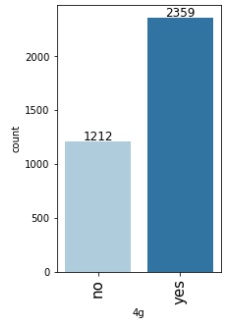
Unitariate

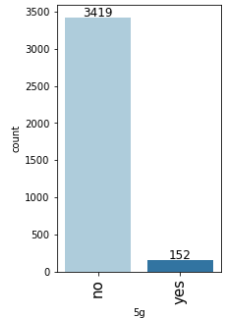
- The variables were explored individually to get a sense of their distribution.
- The target variable, used_price, looks normally distributed. With a slight right skew.
- The skew can be explained by a few relatively expensive phones in the inventory.
- Interestingly, android phone were responsible for about.90% of the market.
- Also Samsung seems to be the leading brand in the used phone market
- Also, a vast majority of the phones seem to be 5g connected





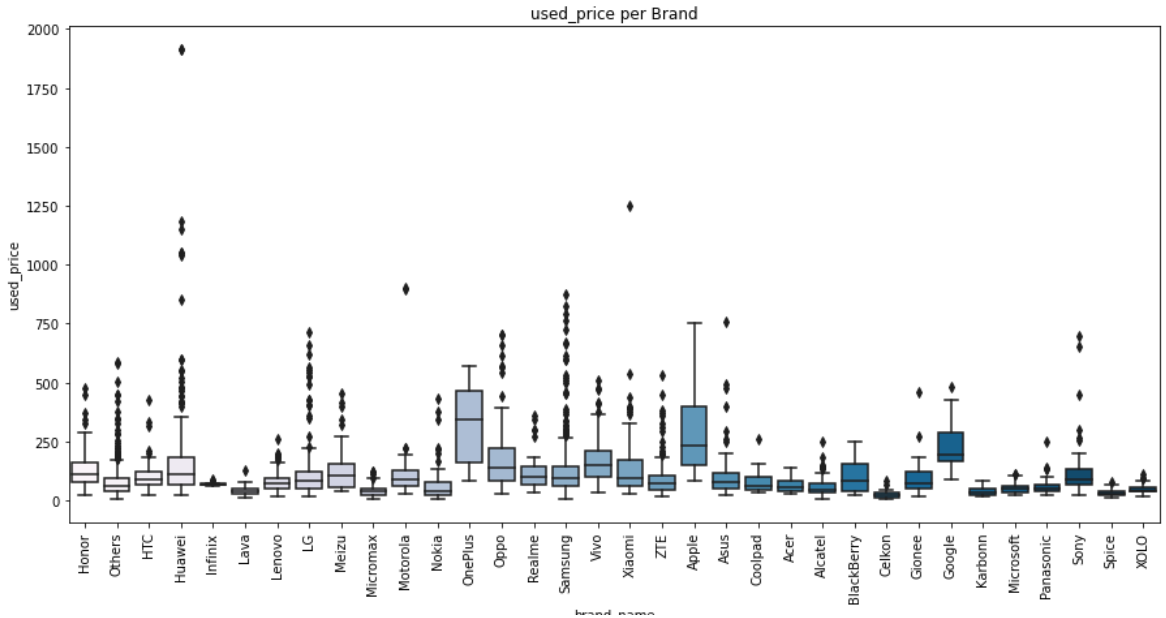






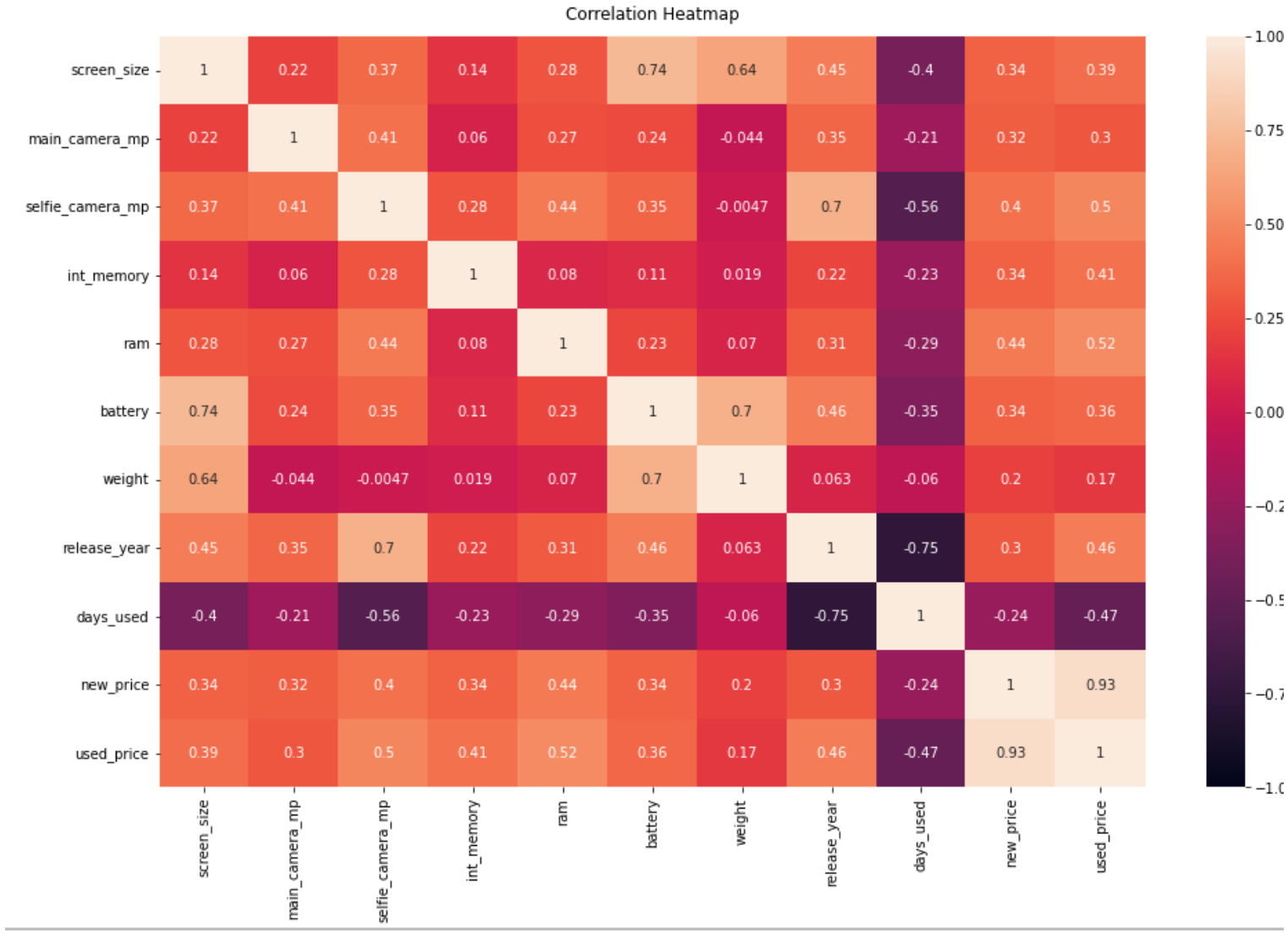
Brand Influences

- Looking at the used prices per brand, As expected, Apple products seem to be the most expensive products. Followed by One Plus and Google.
- From general knowledge, as expected the OnePlus brand also was leading the ram specification. When they launched, the aimed to rival top brand specifications with affordable pricing.
- On average a lot of the brands seemed to provide at least 4GB.



Multi-Variate

- Seems the target variable is most correlated with the price of the new version of the phone.
- It is also correlated with the ram of the phone. Seems people will pay more for higher ram
- All the variables are in line with the expected relationship with the target variable. For example it is expected that the number of days used should be negatively correlated with the used_price.



Model

- Regular linear regression and Ordinary Least Squares method
- With all the variables, both methods produced similar metrics

	Linear Regression	OLS
R^2	0.941	0.941
RMSE	26.617	26.617

OLS Regression Results

Dep. Variable:	used_price	R-squared:	0.941
Model:	OLS	Adj. R-squared:	0.941
Method:	Least Squares	F-statistic:	5000.
Date:	Fri, 22 Oct 2021	Prob (F-statistic):	0.00
Time:	15:56:46	Log-Likelihood:	-12028.
No. Observations:	2499	AIC:	2.407e+04
Df Residuals:	2490	BIC:	2.413e+04
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-3216.1147	967.083	-3.326	0.001	-5112.485	-1319.744
screen_size	-0.5540	0.134	-4.130	0.000	-0.817	-0.291
int_memory	0.0625	0.008	7.830	0.000	0.047	0.078
ram	3.6898	0.547	6.741	0.000	2.616	4.763
release_year	1.6256	0.479	3.391	0.001	0.686	2.566
days_used	-0.1102	0.004	-29.188	0.000	-0.118	-0.103
new_price	0.5099	0.004	136.224	0.000	0.503	0.517
4g_yes	-15.7078	1.671	-9.402	0.000	-18.984	-12.432
5g_yes	60.6198	3.793	15.982	0.000	53.182	68.058

Omnibus:	2051.279	Durbin-Watson:	1.982
Prob(Omnibus):	0.000	Jarque-Bera (JB):	337606.213
Skew:	3.098	Prob(JB):	0.00
Kurtosis:	59.603	Cond. No.	3.47e+06

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.47e+06. This might indicate that there are strong multicollinearity or other numerical problems.