SC1015 MINI PROJECT

Team 6

VEDA HO YONG QIAN (U2321709B)
TIFFANY MUN (U2330399H)
WANG LI LING KATIE (U2322083J)

INSPIRATION

Parents hunt for laptops, deal with distracted kids on first day of ongoing home-based learning

MOE lends 3,300 devices to students who need them for home-based learning amid Covid-19 pandemic

All secondary school students to get personal laptop or tablet for learning by 2021: Tharman

IT WAS REPORTED

98% of students need laptops in singapore!

PROBLEM STATEMENT

Students find it **difficult** to **determine the price of a laptop** with their wanted specifications given the large amount of information

AIM

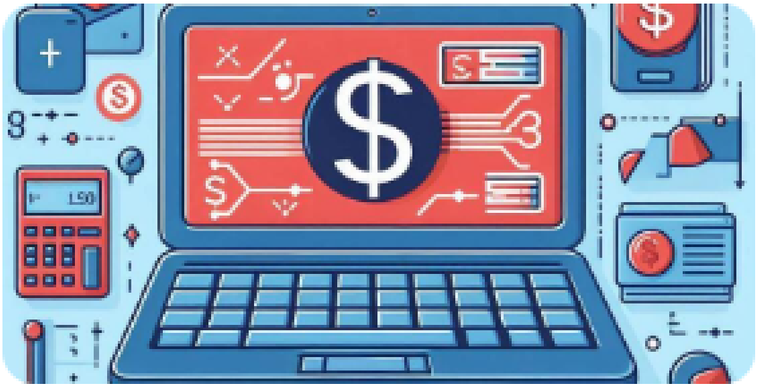
Aid decision-making for students by estimating the cost of laptops based on required specifications

DATA SET

kaggle

Laptop Prices Based on its specifications

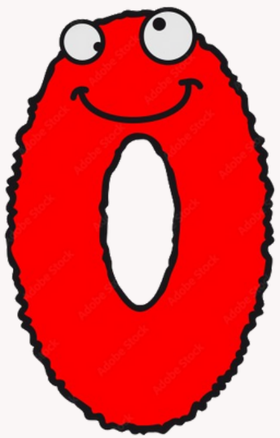
Exploring the Relationship Between Laptop Features and Pricing Trends



Variables laptop_ID, Company, Product, TypeName, Inches, ScreenResolution, Cpu, Ram, Memory, Gpu, OpSys, Weight, Price_euros, Price_SGD

STEP 1: DATA CLEANING

STEP 1: DATA CLEANING



Check if there are any null values

In [5]:
```
# Check if there is any NULL values in laptop data
laptop_data.isnull().sum()
```

Out[5]:
```
laptop_ID           0
Company             0
Product             0
TypeName            0
Inches              0
ScreenResolution    0
Cpu                 0
Ram                 0
Memory              0
Gpu                 0
OpSys               0
Weight              0
Price_euros         0
Price_SGD           0
dtype: int64
```

Check if there are duplicates

In [6]:
```
# Check if there is any duplicates in laptop data
laptop_data.duplicated().sum()
```

Out[6]: 0

STEP 1: DATA CLEANING

Raw data

Out[6]:

	laptop_ID	Company	Product	TypeName	Inches	ScreenResolution	Cpu	Ram	Memory	Gpu	OpSys	Weight	Price_euros	Price_SGD
0	1	Apple	MacBook Pro	Ultrabook	13.3	IPS Panel Retina Display 2560x1600	Intel Core i5 2.3GHz	8GB	128GB SSD	Intel Iris Plus Graphics 640	macOS	1.37kg	1339.69	1955.9474
1	2	Apple	Macbook Air	Ultrabook	13.3	1440x900	Intel Core i5 1.8GHz	8GB	128GB Flash Storage	Intel HD Graphics 6000	macOS	1.34kg	898.94	1312.4524
2	3	HP	250 G6	Notebook	15.6	Full HD 1920x1080	Intel Core i5 7200U 2.5GHz	8GB	256GB SSD	Intel HD Graphics 620	No OS	1.86kg	575.00	839.5000
3	4	Apple	MacBook Pro	Ultrabook	15.4	IPS Panel Retina Display 2880x1800	Intel Core i7 2.7GHz	16GB	512GB SSD	AMD Radeon Pro 455	macOS	1.83kg	2537.45	3704.6770
4	5	Apple	MacBook Pro	Ultrabook	13.3	IPS Panel Retina Display 2560x1600	Intel Core i5 3.1GHz	8GB	256GB SSD	Intel Iris Plus Graphics	macOS	1.37kg	1803.60	2633.2560

STEP 1: DATA CLEANING

Cleaned data

In [11]:
```
laptop_data.head()
```

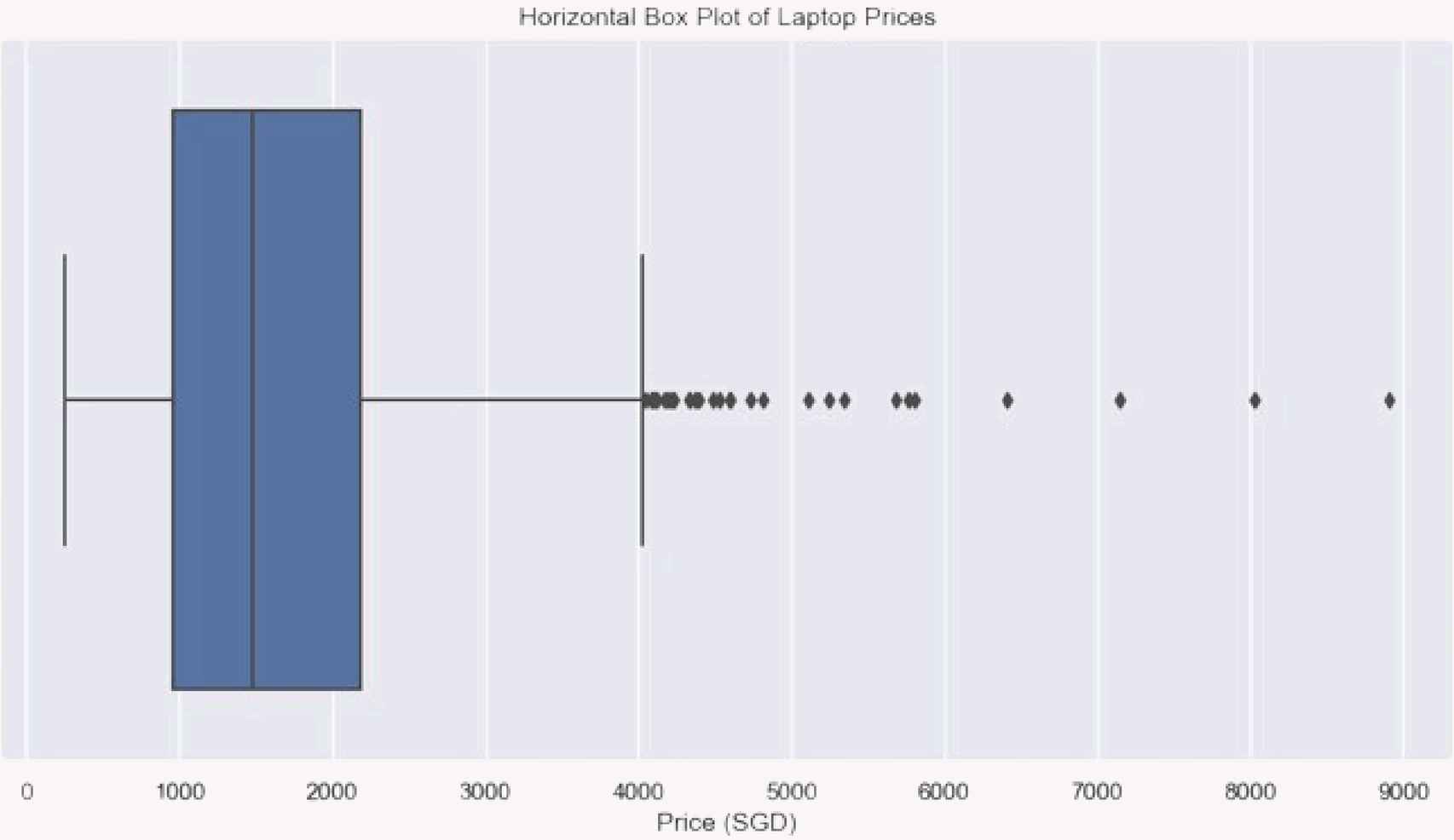
Out[11]:

	laptop_ID	Company	Inches	ScreenResolution	Cpu	Ram	Memory	Gpu	OpSys	Weight	Price_SGD	TouchScreen
0	1	Apple	13.3	2560x1600	2.3	8	128	Intel	macOS	1.37	1955.95	False
1	2	Apple	13.3	1440x900	1.8	8	128	Intel	macOS	1.34	1312.45	False
2	3	HP	15.6	1920x1080	2.5	8	256	Intel	No OS	1.86	839.50	False
3	4	Apple	15.4	2880x1800	2.7	16	512	AMD	macOS	1.83	3704.68	False
4	5	Apple	13.3	2560x1600	3.1	8	256	Intel	macOS	1.37	2633.26	False

STEP 2: EXPLORATORY DATA ANALYSIS

STEP 2: EXPLORATORY DATA ANALYSIS

UNI-VARIATE BOX-PLOT OF LAPTOP PRICES

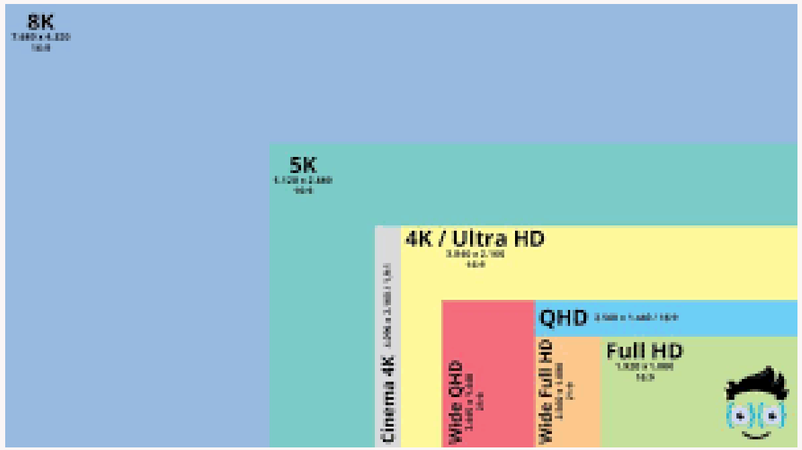


MEDIAN PRICE OF LAPTOPS = $1400

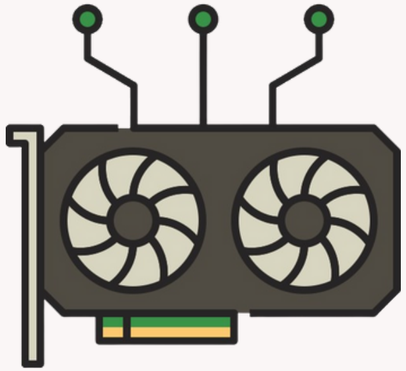
STEP 2: EXPLORATORY DATA ANALYSIS

CATEGORICAL PREDICTORS AND PRICE

COMPANY

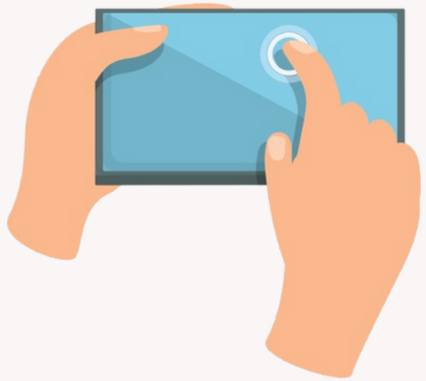


SCREEN RESOLUTION



GPU

OPERATING SYSTEMS

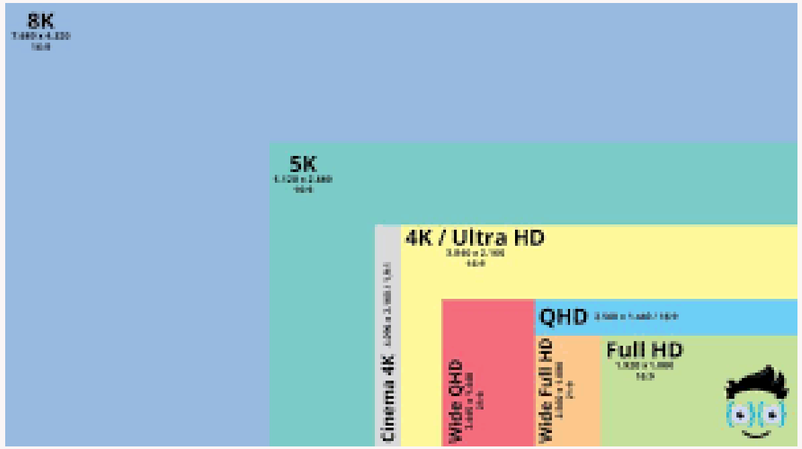


TOUCHSCREEN

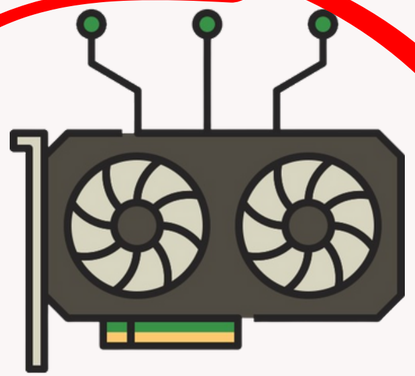
STEP 2: EXPLORATORY DATA ANALYSIS

CATEGORICAL PREDICTORS AND PRICE

COMPANY

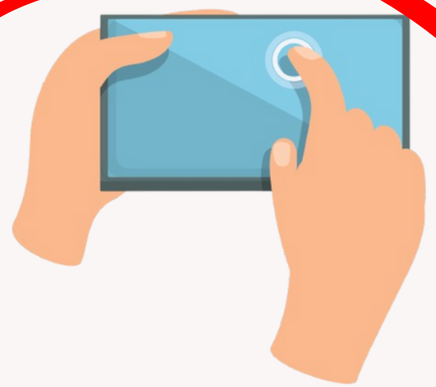


SCREEN RESOLUTION



GPU

OPERATING SYSTEMS



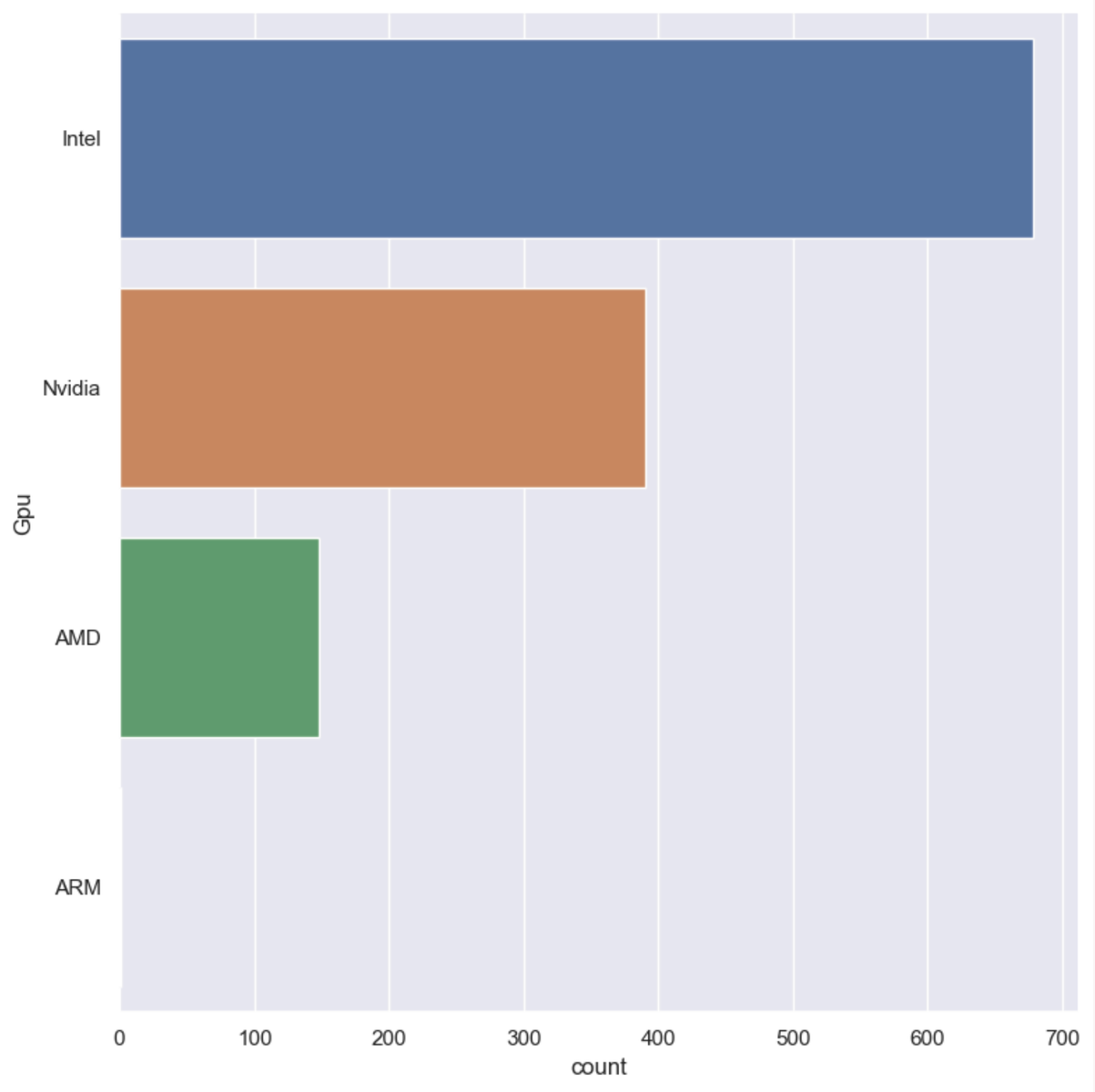
TOUCHSCREEN

GPU VS PRICE

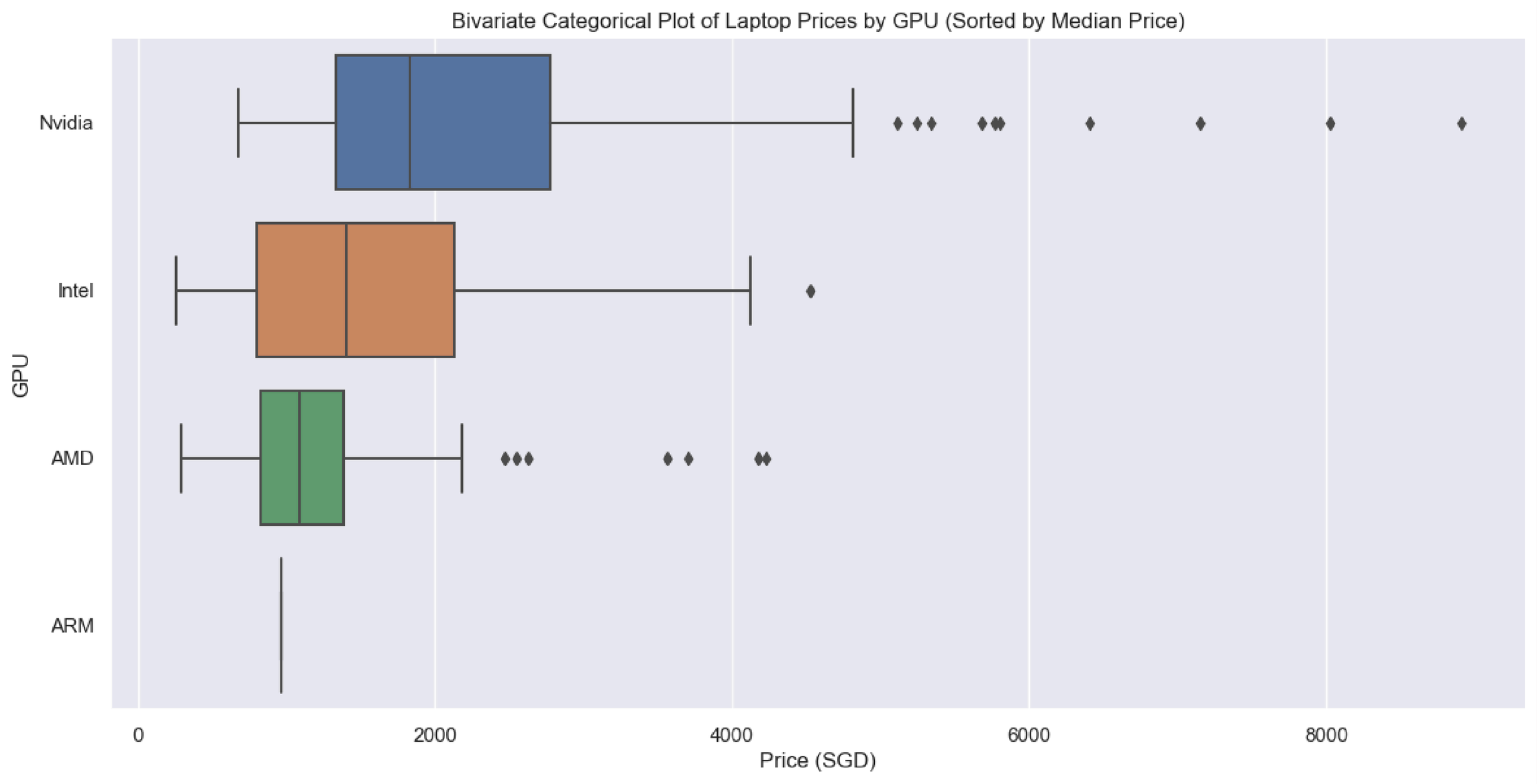
NO. OF UNIQUE GPU:4

Gpu Count

Intel	678
Nvidia	390
AMD	148
ARM	1

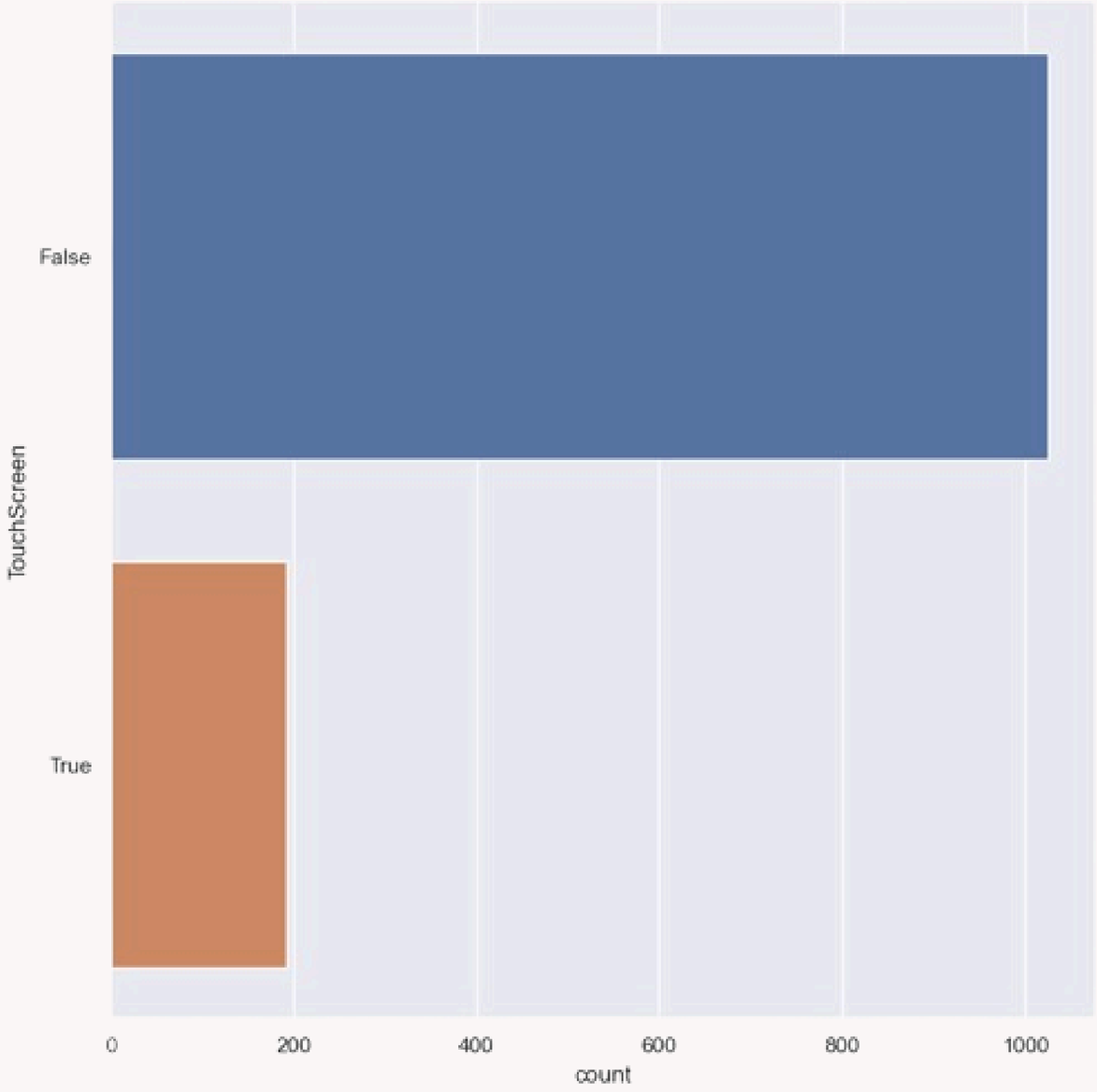


GPU VS PRICE

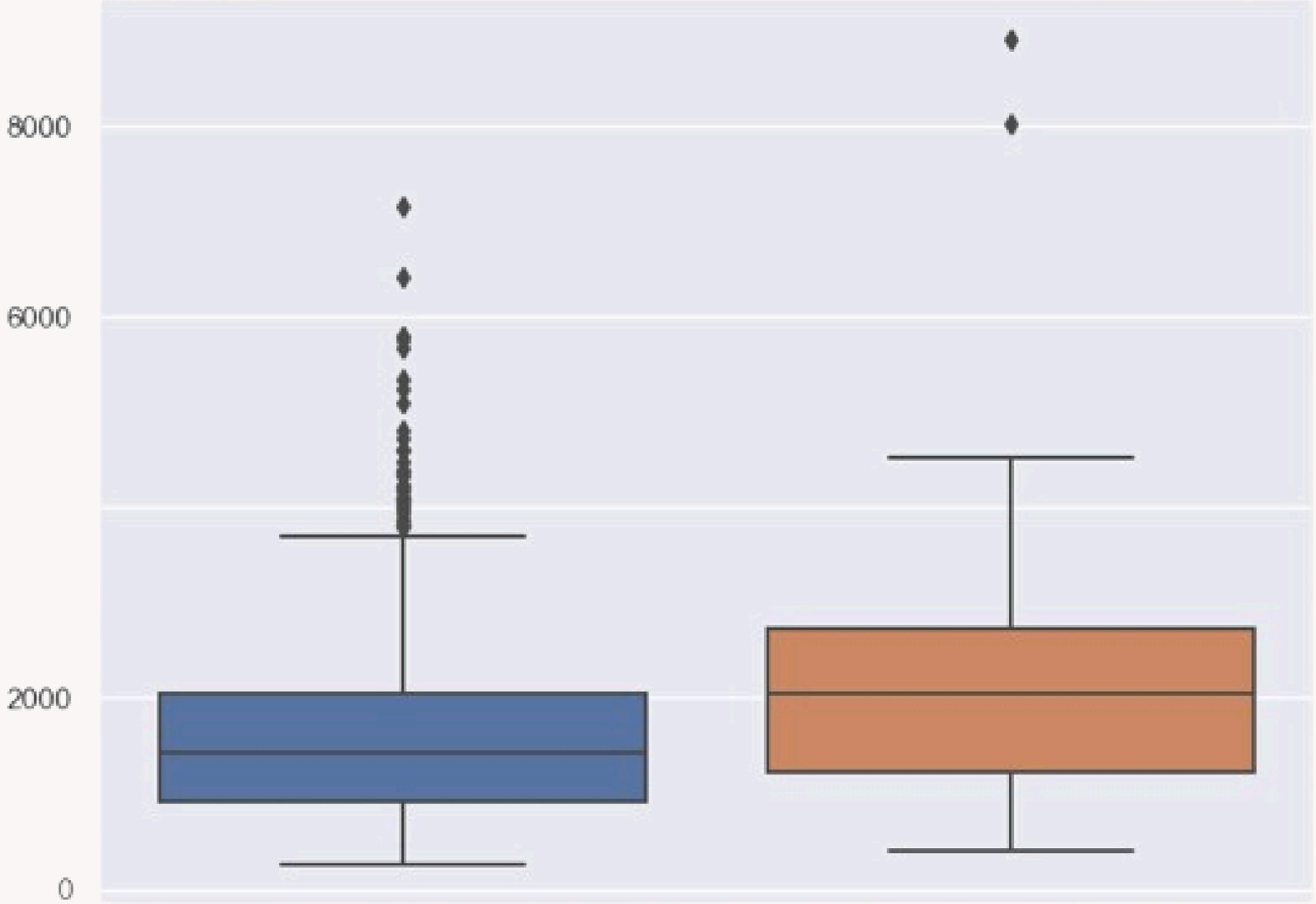

Bivariate Categorical Plot of Laptop Prices by GPU (Sorted by Median Price)
Nvidia
Intel
AMD
ARM
GPU
0
2000
4000
6000
8000
Price (SGD)

TOUCHSCREEN VS PRICE

FALSE: 1025
TRUE: 192



TOUCHSCREEN VS PRICE



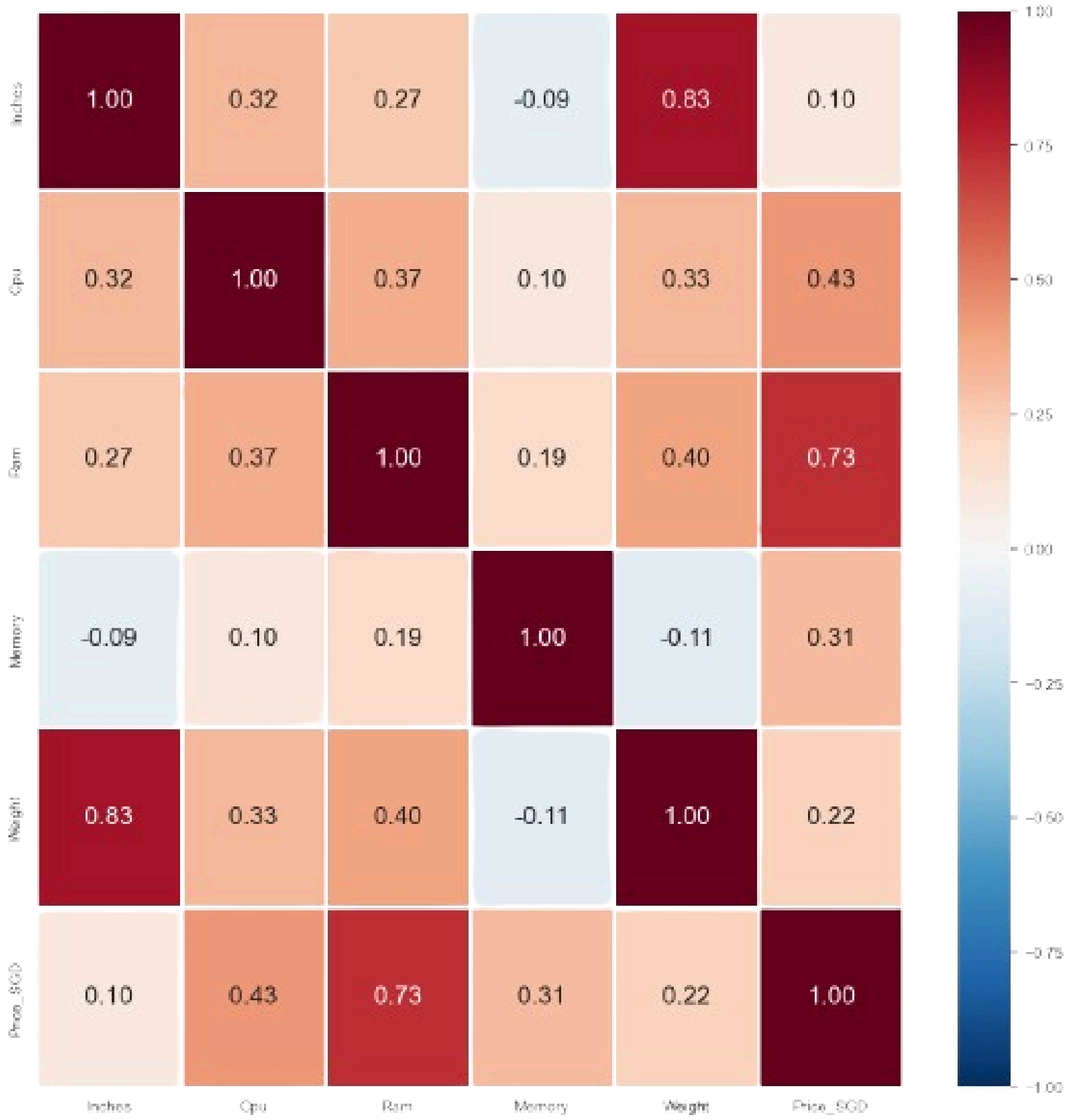
NUMERICAL PREDICTORS

Out[32]:

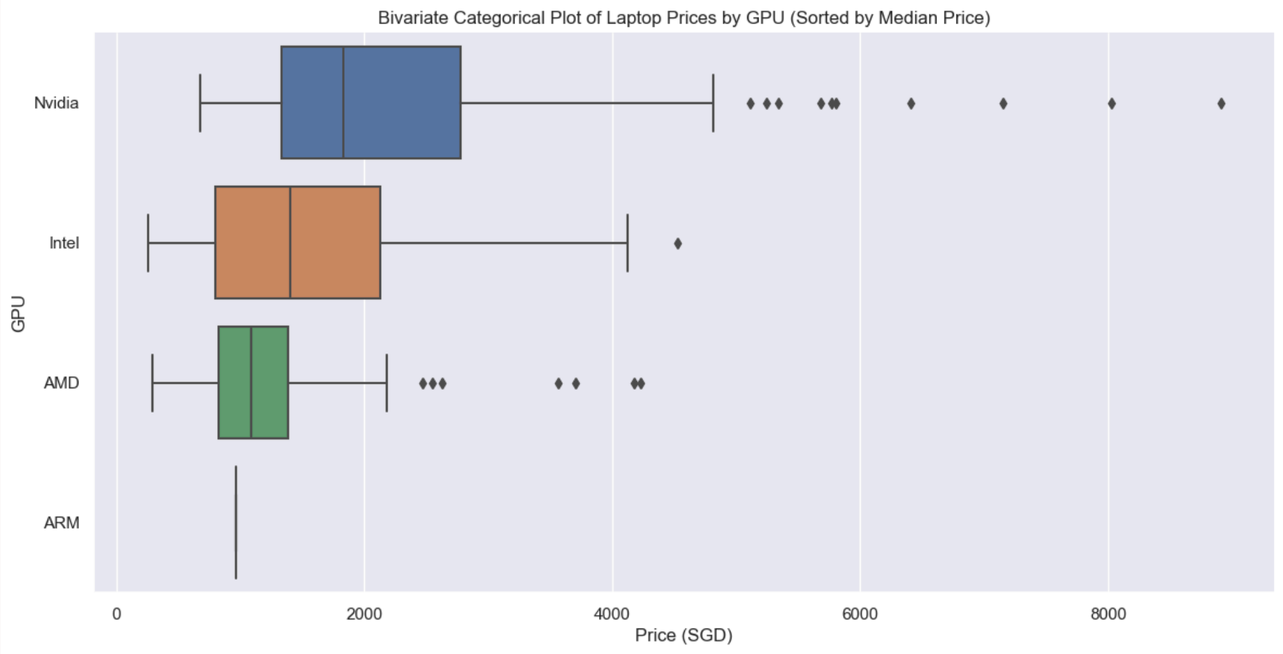
	Inches	Cpu	Ram	Memory	Weight	Price_SGD
count	1217.00	1217.00	1217.00	1217.00	1217.00	1217.00
mean	14.97	2.30	8.61	229.37	2.03	1696.93
std	1.45	0.51	5.14	172.82	0.68	1021.17
min	10.10	0.90	2.00	1.00	0.69	254.04
25%	14.00	1.80	4.00	128.00	1.49	959.22
50%	15.60	2.50	8.00	256.00	2.02	1477.51
75%	15.60	2.70	8.00	256.00	2.31	2188.54
max	18.40	3.60	64.00	512.00	4.70	8904.54

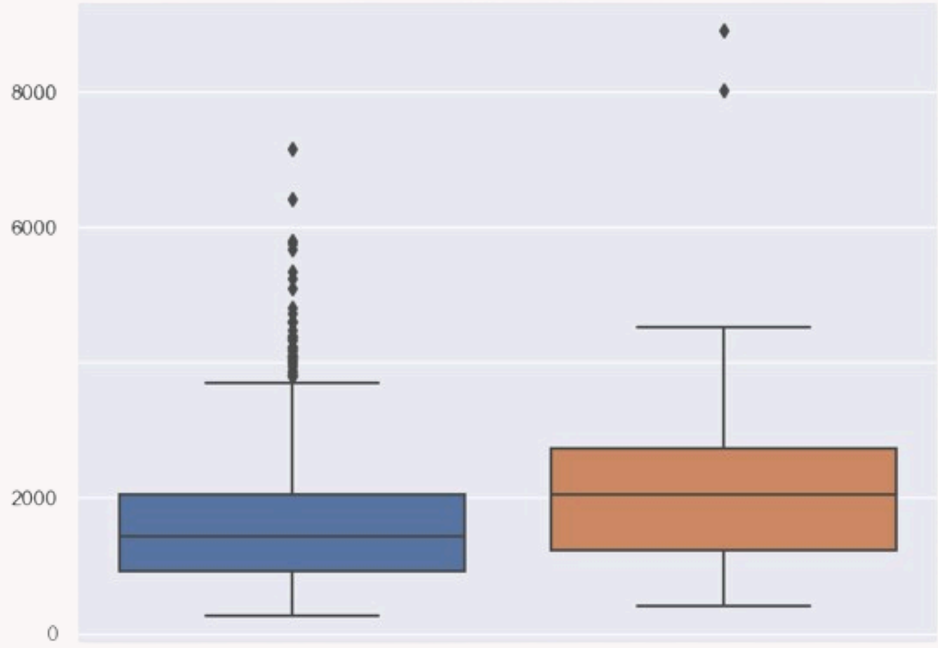
CORRELATION MATRIX

- ALL NUMERICAL VARIABLES ARE POSITIVELY RELATED TO THE PRICE OF LAPTOPS.
- RAM (0.73) HAS THE HIGHEST CORRELATION TO PRICE.



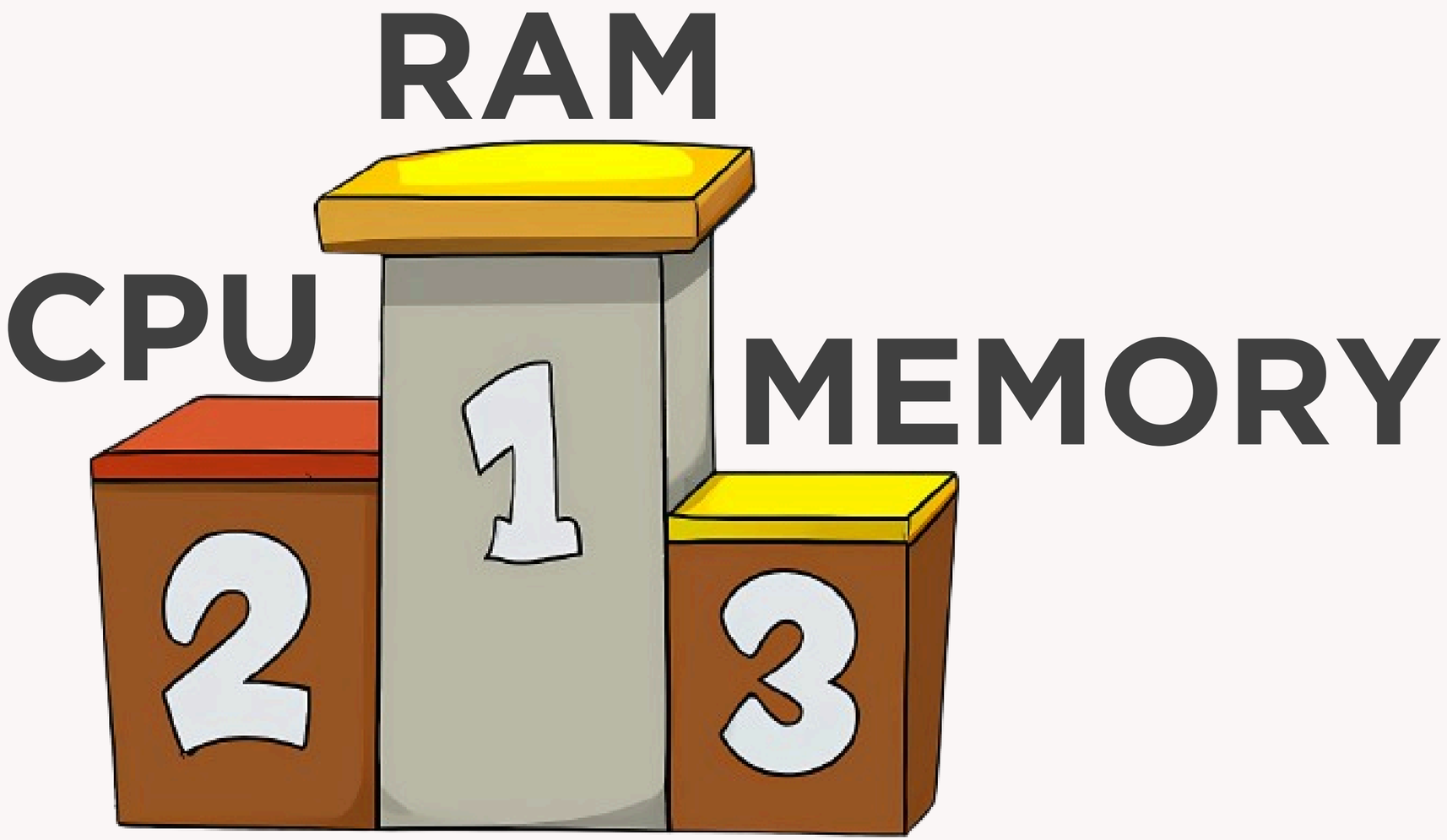
INSIGHTS FOR CATEGORICAL PREDICTORS





GPU AND TOUCHSCREEN HAVE GREATEST CORRELATION TO PRICE

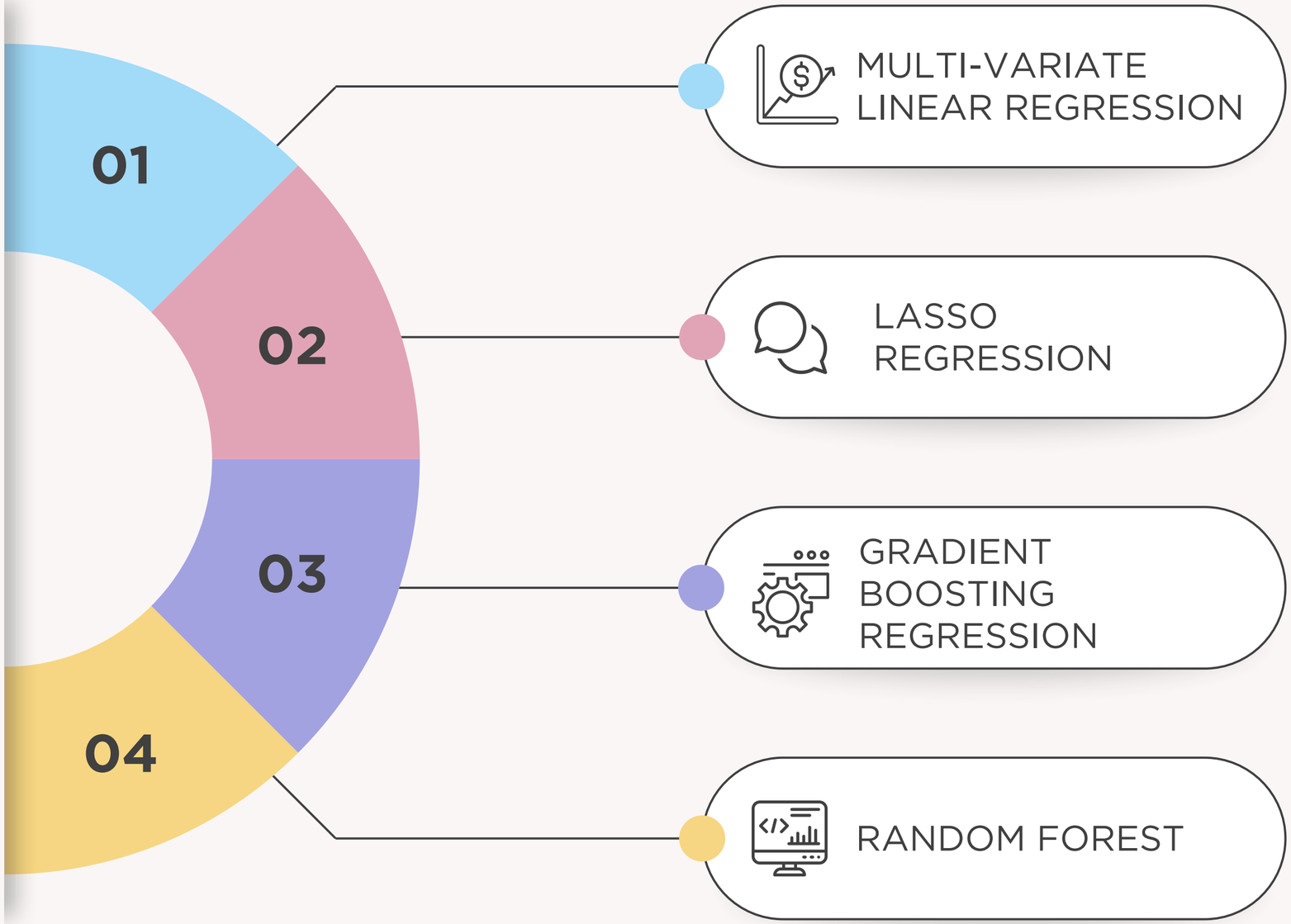
INSIGHTS FOR NUMERICAL PREDICTORS



STEP 3: PREDICTION OF PRICE

MACHINE LEARNING MODELS

train - test data split: 75-25



ENCODING CATEGORAL VARIABLES

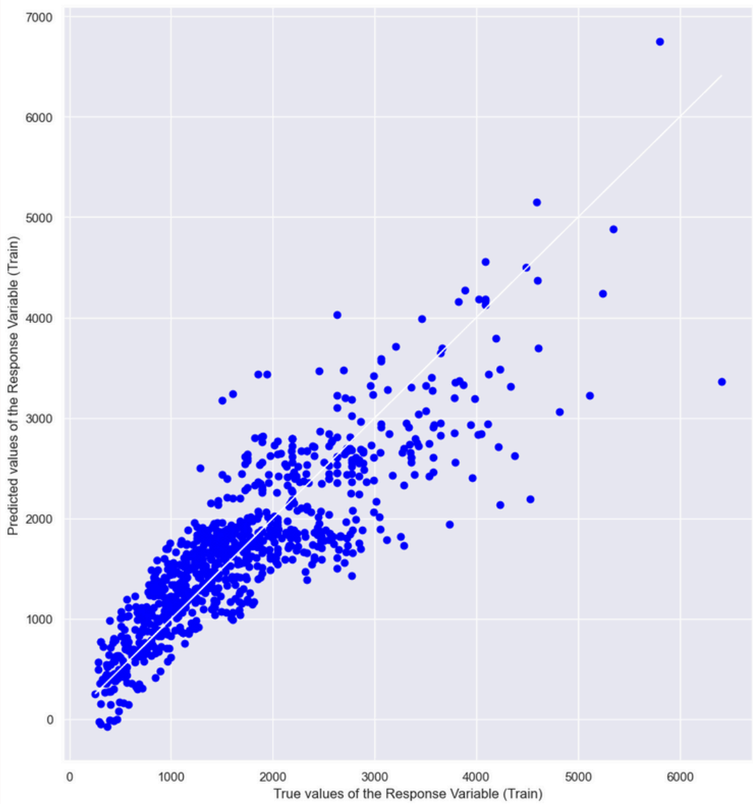
BEFORE:

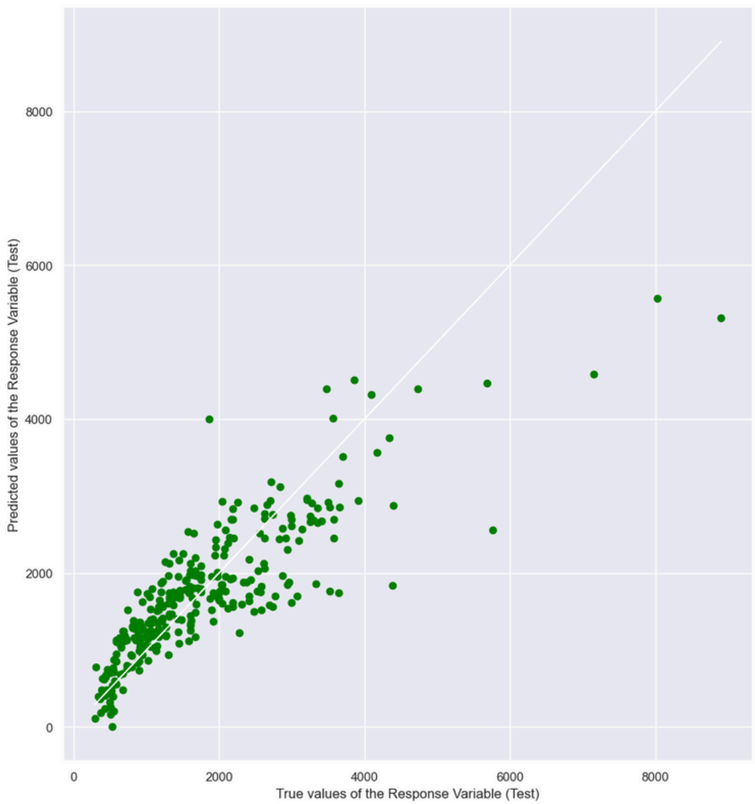
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1217 entries, 0 to 1216
Data columns (total 12 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   laptop_ID         1217 non-null   int64
 1   Company           1217 non-null   object
 2   Inches            1217 non-null   float64
 3   ScreenResolution  1217 non-null   object
 4   Cpu               1217 non-null   float64
 5   Ram               1217 non-null   int32
 6   Memory            1217 non-null   int32
 7   Gpu               1217 non-null   object
 8   OpSys             1217 non-null   object
 9   Weight            1217 non-null   float64
 10  Price_SGD         1217 non-null   float64
 11  TouchScreen       1217 non-null   object
dtypes: float64(4), int32(2), int64(1), object(5)
memory usage: 104.7+ KB
```

AFTER:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1217 entries, 0 to 1216
Data columns (total 55 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   Inches                      1217 non-null   float64
 1   Cpu                         1217 non-null   float64
 2   Ram                         1217 non-null   int32
 3   Memory                      1217 non-null   int32
 4   Weight                      1217 non-null   float64
 5   Company_Acer                1217 non-null   float64
 6   Company_Apple               1217 non-null   float64
 7   Company_Asus                1217 non-null   float64
 8   Company_Chuwi               1217 non-null   float64
 9   Company_Dell                1217 non-null   float64
 10  Company_Fujitsu             1217 non-null   float64
 11  Company_Google              1217 non-null   float64
 12  Company_HP                  1217 non-null   float64
 13  Company_Huawei              1217 non-null   float64
 14  Company_LG                  1217 non-null   float64
 15  Company_Lenovo              1217 non-null   float64
 16  Company_MSI                 1217 non-null   float64
 17  Company_Mediacom            1217 non-null   float64
 18  Company_Microsoft           1217 non-null   float64
 19  Company_Razer               1217 non-null   float64
 20  Company_Samsung             1217 non-null   float64
 21  Company_Toshiba             1217 non-null   float64
 22  Company_Vero                1217 non-null   float64
 23  Company_Xiaomi              1217 non-null   float64
 24  ScreenResolution_1366x768   1217 non-null   float64
 25  ScreenResolution_1440x900   1217 non-null   float64
 26  ScreenResolution_1600x900   1217 non-null   float64
 27  ScreenResolution_1920x1080  1217 non-null   float64
 28  ScreenResolution_1920x1200  1217 non-null   float64
 29  ScreenResolution_2160x1440  1217 non-null   float64
 30  ScreenResolution_2256x1504  1217 non-null   float64
 31  ScreenResolution_2304x1440  1217 non-null   float64
 32  ScreenResolution_2400x1600  1217 non-null   float64
 33  ScreenResolution_2560x1440  1217 non-null   float64
 34  ScreenResolution_2560x1600  1217 non-null   float64
```

LINEAR REGRESSION MODEL

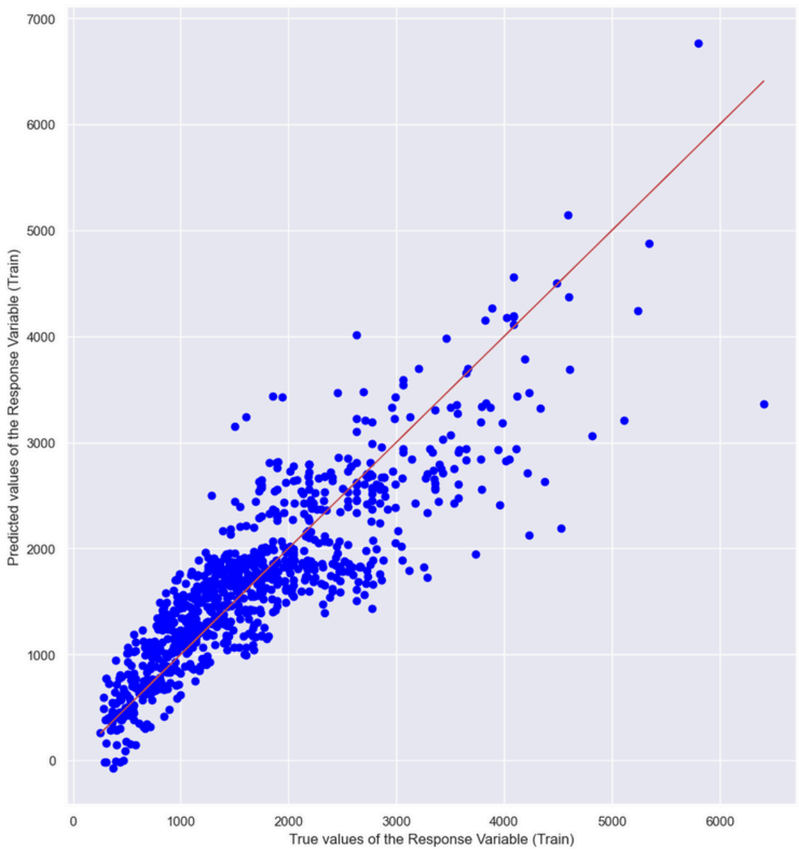


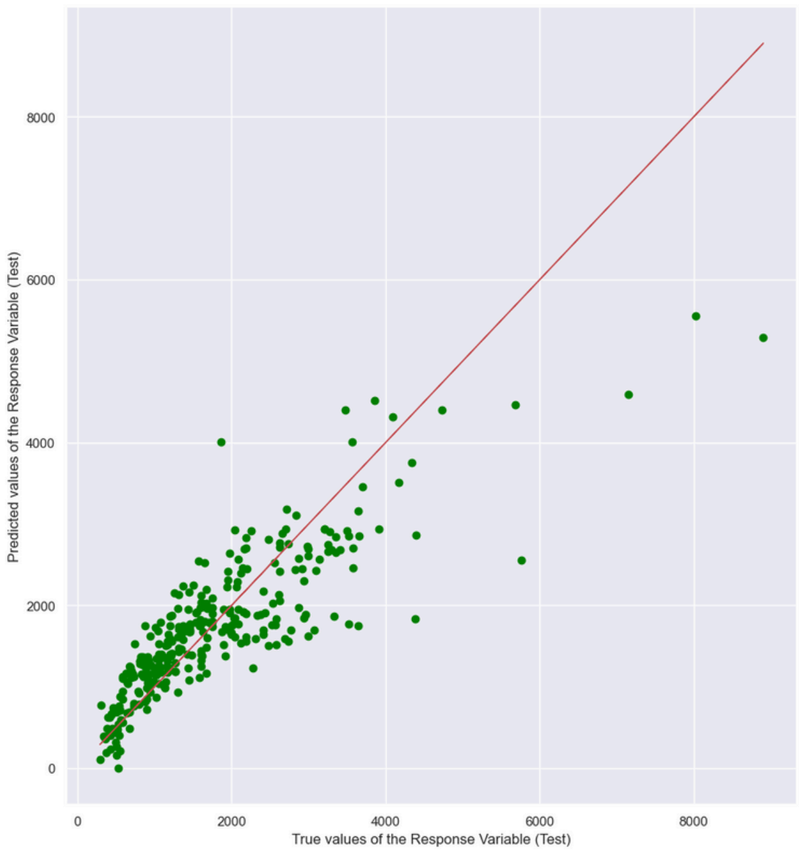



```
Goodness of Fit of Model        Train Dataset
Explained Variance (R^2)        : 0.7456667335456566
Mean Squared Error (MSE)        : 233895.502910667
Root Mean Squared Error (RMSE)  : 483.62744226384325

Goodness of Fit of Model        Test Dataset
Explained Variance (R^2)        : 0.716789044543215
Mean Squared Error (MSE)        : 396928.8821022894
Root Mean Squared Error (RMSE)  : 630.0229218864099
```

LASSO REGRESSION MODEL

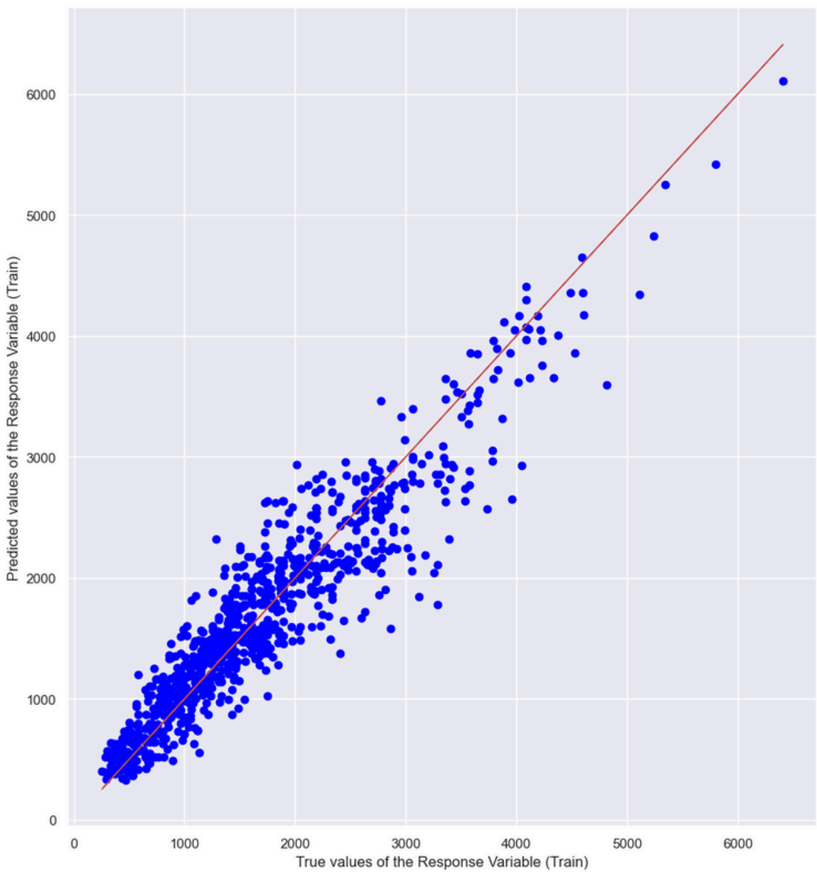


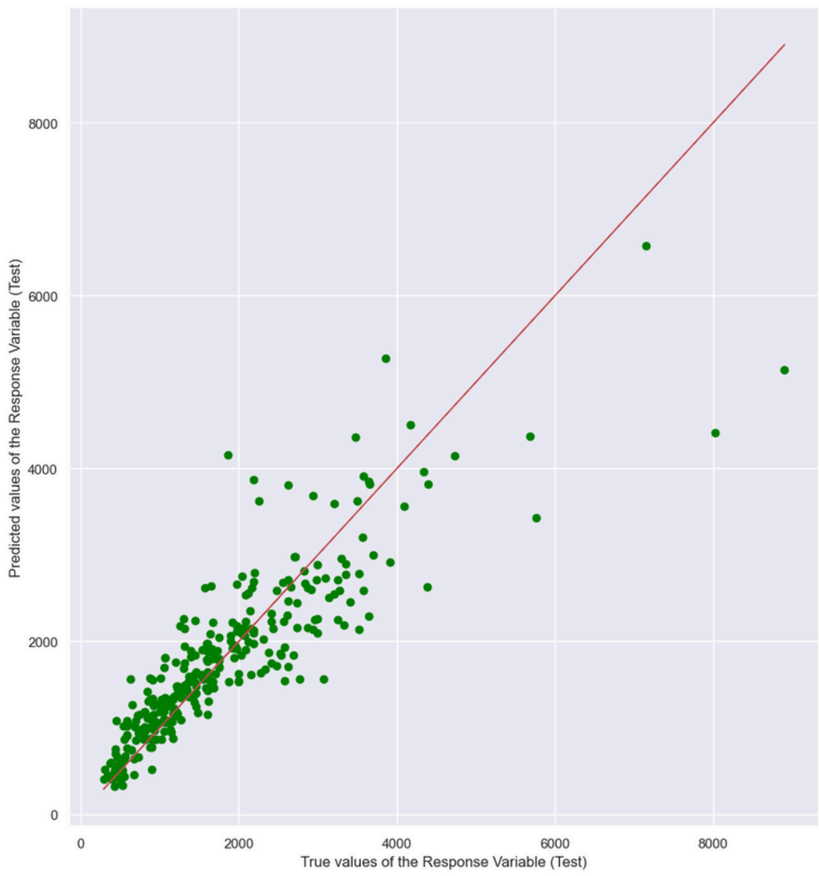



```
Goodness of Fit of Model        Train Dataset
Explained Variance (R^2)        : 0.745555813824818
Mean Squared Error (MSE)        : 233997.5093223721
Root Mean Squared Error (RMSE)  : 483.73289046990806

Goodness of Fit of Model        Test Dataset
Explained Variance (R^2)        : 0.7161956985325821
Mean Squared Error (MSE)        : 397760.4748220006
Root Mean Squared Error (RMSE)  : 630.6825467872094
```

GRADIENT BOOSTING REGRESSION






```
Goodness of Fit of Model        Train Dataset
Explained Variance (R^2)        : 0.8808627685801075
Mean Squared Error (MSE)        : 109563.5779259831
Root Mean Squared Error (RMSE)  : 331.00389412510407

Goodness of Fit of Model        Test Dataset
Explained Variance (R^2)        : 0.7576485112568716
Mean Squared Error (MSE)        : 339663.0802910943
Root Mean Squared Error (RMSE)  : 582.8062116099093
```

RANDOM FOREST

```
params = {
    'max_depth': [2, 3, 5, 10, 20],
    'min_samples_leaf': [5, 10, 20, 50, 100, 200],
    'n_estimators': [10, 25, 30, 50, 100, 200]
}
```

grid search

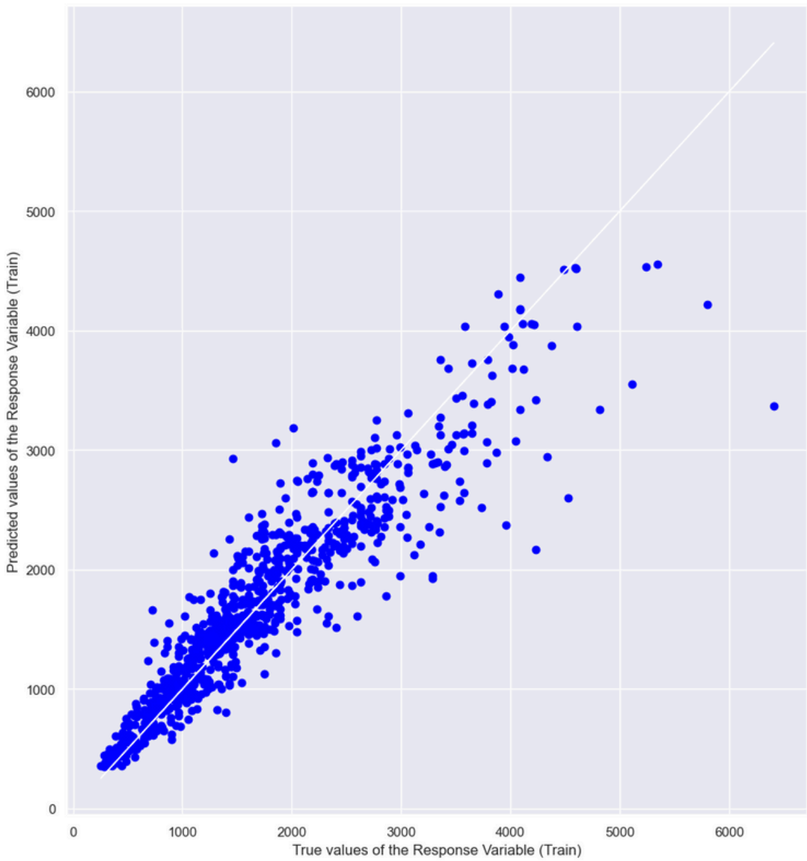
- test all combinations of the parameters
- optimise model for predicting price

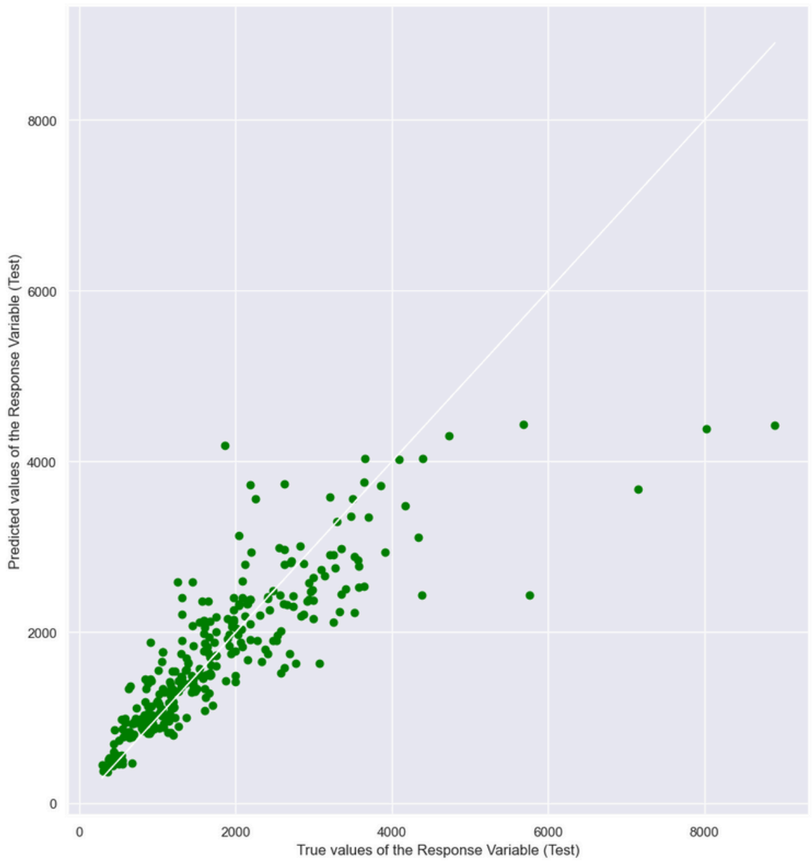
RANDOM FOREST

	Varname	Imp
2	Ram	0.608318
4	Weight	0.150154
1	Cpu	0.097437
3	Memory	0.031591
0	Inches	0.022639
27	ScreenResolution_1920x1080	0.016083
50	OpSys_Windows 7	0.014256
39	Gpu_AMD	0.007895
12	Company_HP	0.005899
24	ScreenResolution_1366x768	0.005823
9	Company_Dell	0.005401
7	Company_Asus	0.004982

features ranked based on importance!

RANDOM FOREST






```
Goodness of Fit of Model        Train Dataset
Explained Variance (R^2)        : 0.8611577463016726
Mean Squared Error (MSE)        : 127685.14007918959
Root Mean Squared Error (RMSE)  : 357.3305753489191

Goodness of Fit of Model        Test Dataset
Explained Variance (R^2)        : 0.7065071971355441
Mean Squared Error (MSE)        : 411339.2081113616
Root Mean Squared Error (RMSE)  : 641.3573170326831
```

STEP 4: INSIGHTS OF ALL MODELS

PREDICTED PRICE

ACTUAL PRICE: $1955.95

Models	Predicted Price	Percentage Error
linear regression	2335.84	19.42%
Lasso	2309.87	18.09%
GBR	1948.64	0.37%
Random Forest	2063.90	5.52%

COMPARISON

	MSE	R2	RMSE
linear regression	396928.88	0.7168	630.02
Lasso	397760.47	0.7161	630.68
GBR	339663.08	0.7576	582.81
Random forest	411339.21	0.707	641.36

0.757

INSIGHTS: BEST MODEL

	MSE	R2	RMSE
GBR	339663.08	0.7576	582.81
	LOWEST	HIGHEST	LOWEST

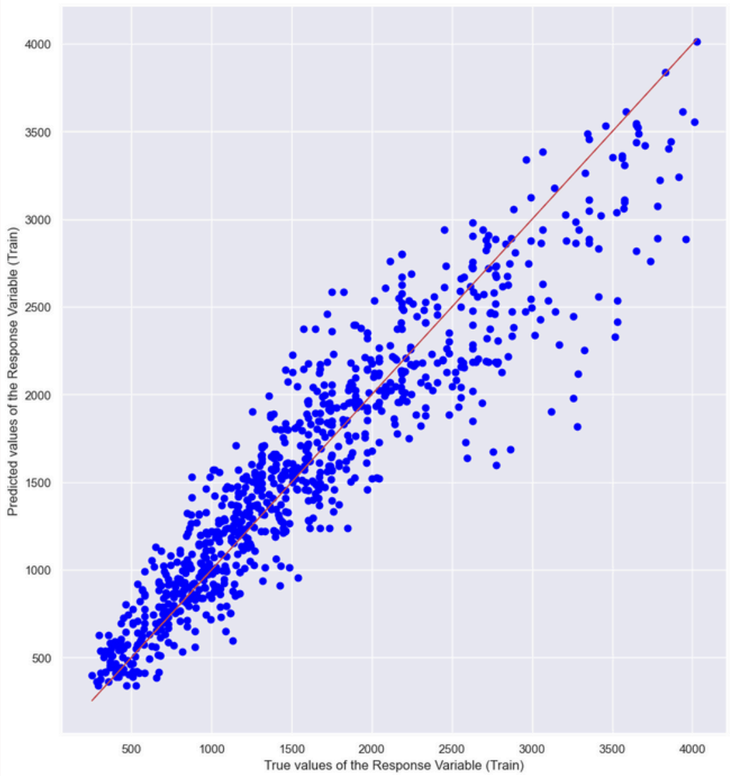
STEP 5: REMOVING OUTLIERS TO TEST GBR MODEL

REMOVING OUTLIERS IN DATASET

```
Shape of DataFrame before removing Outliers: (1217, 55)
Shape of DataFrame after removing Outliers: (1182, 55)
```

REDUCED BY 35 ROWS

GRADIENT BOOSTING REGRESSION - REMOVED OUTLIERS




```
Goodness of Fit of Model        Train Dataset
Explained Variance (R^2)        : 0.8644329008584486
Mean Squared Error (MSE)        : 95731.18347040856
Root Mean Squared Error (RMSE)  : 309.4045627821422

Goodness of Fit of Model        Test Dataset
Explained Variance (R^2)        : 0.8043662584351455
Mean Squared Error (MSE)        : 150573.37672425146
Root Mean Squared Error (RMSE)  : 388.0378547567897
```

INSIGHTS

	MSE ↓	R2 ↑	RMSE ↓
BEFORE	339663.08	0.7576	582.81
AFTER	150573.38	0.8044	388.04

ACCURACY IS IMPROVED

CONCLUSION

- FROM OUR EXPLORATORY ANALYSIS, GPU, PRESENCE OF TOUCHSCREEN AND RAM HAVE THE GREATEST CORRELATION TO PRICE.
- IN THE PREDICTION OF PRICE, THE GRADIENT BOOSTING REGRESSION MODEL IS THE BEST MODEL

ANNEX