PROJECT

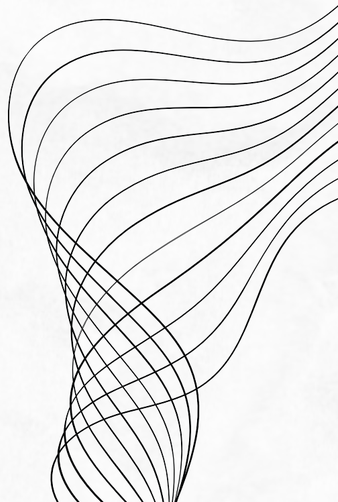
# PC COMPONENT PRICE ANALYSIS

20 JULY, 2023

Pro: CHAN SOPHAL

# AGENDA

1. INTRODUNCTION
2. DISCOVERY DATA
3. DATA PREPARATION
4. EDA
5. FEATURE ENGINEERING
6. MODEL SELECTION
7. MODEL COMPARISON
8. CONCLUSION



# INTRODUCTION

"Scraping, EDA, and Model: **PC Component Analysis**." In this session, we will delve into the process of web scraping, performing exploratory data analysis (EDA), and building a model using data collected from the website "https://www.khmer24.com/." Specifically, we will focus on *analyzing PC components* available on the website to gain insights into the market trends and make informed decisions.

# DISCOVERY DATA

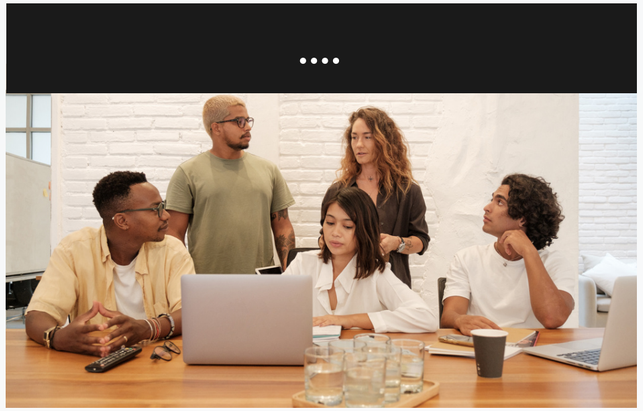
...

# UNDERSTANDING THE DATA

| | ID | Category | Location | Posted | Brand | Discription | Price | RAM | CPU | VGA | Price1 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 4235811 | Parts & Accessories | Battambang | 6-May-23 | PC Components | Used | 20.0 | 32GB | Intel Core i9 | 4GB | 1500.0 |
| 1 | 5005760 | Parts & Accessories | Phnom Penh | 6-May-23 | PC Components | Used | 200.0 | 8GB | Intel Core i5 | Integrated | 350.0 |
| 2 | 8541741 | Parts & Accessories | Phnom Penh | 6-May-23 | PC Components | Used | 40.0 | 8GB | Intel Core i5 | 2GB & Under | 360.0 |
| 3 | 3733350 | Parts & Accessories | Banteay Meanchey | 6-May-23 | PC Components | Used | 140.0 | 8GB | Intel Core i7 | 4GB | 440.0 |
| 4 | 8035739 | Parts & Accessories | Phnom Penh | 6-May-23 | PC Components | Used | 7.0 | 8GB | Intel Core i5 | 2GB & Under | 388.0 |

For data set above have 11 variables and all variables has detail in dataset and some variable not Important for predict. so we need to drop some columns.

>> Note: For This data we want to check size CPU and RAM to predict price.



# ASSESS THE DATA

Understand the dataset's structure and identify potential issues.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1055 entries, 0 to 1054
Data columns (total 5 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   Discription  1055 non-null   object 
 1   RAM          1055 non-null   object 
 2   CPU          1055 non-null   object 
 3   VGA          1055 non-null   object 
 4   TotalPrice   1054 non-null   float64
dtypes: float64(1), object(4)
memory usage: 41.3+ KB
```

```
category feature : ['Category', 'Location', 'Posted', 'Brand', 'Discription', 'RAM', 'CPU', 'VGA']
numerical features: ['ID', 'Price', 'Price1']
```

# DATA PREPARATION

HANDLE MISSING VALUES, OUTLIERS, AND INCONSISTENCIES IN THE DATA.

# HANDLE MISSING DATA

## Check Missing Value

Understand the dataset's structure and identify potential issues.

```
Discription     0
RAM             0
CPU             0
VGA             0
TotalPrice      1
N_CPU           0
dtype: int64
```

## Delete Missing Value

...

```
Discription     0
RAM             0
CPU             0
VGA             0
TotalPrice      0
N_CPU           0
N_Discription   0
N_VGA           0
dtype: int64
```

## Removing Duplicates

Identify and remove any duplicate records or observations.

**> Duplicate Data**

```
Data['ID'].duplicated().sum()
```

```
0
```

# DESCRIBE STATITICAL

```
df.describe(include = ['float64', 'int64'])
```

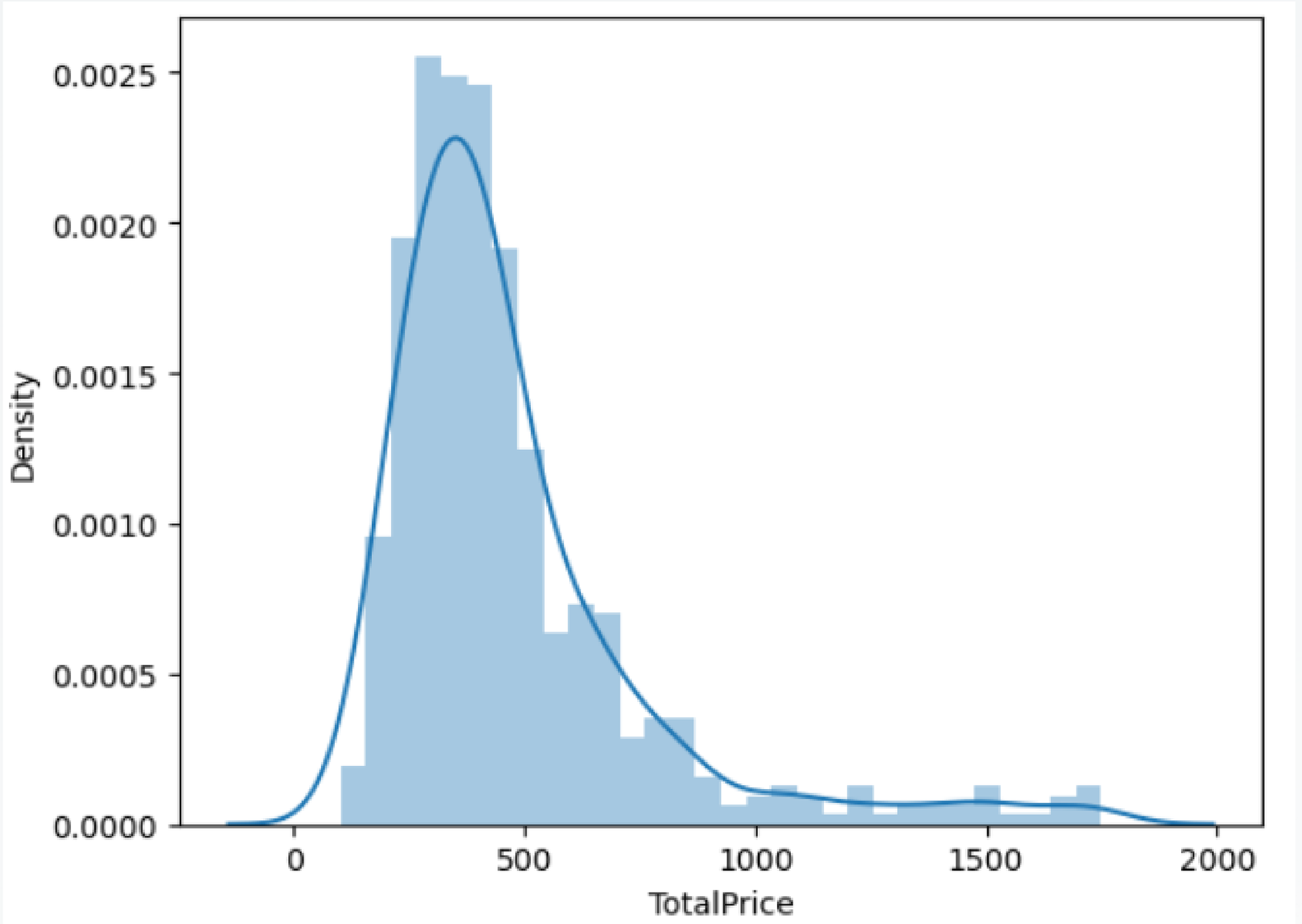
| | RAM | TotalPrice | N_CPU |
|---|---|---|---|
| **count** | 1055.000000 | 1054.000000 | 1055.000000 |
| **mean** | 8.536493 | 461.400380 | 7.048341 |
| **std** | 6.964558 | 280.977062 | 2.064550 |
| **min** | 4.000000 | 100.000000 | 0.000000 |
| **25%** | 4.000000 | 295.000000 | 6.000000 |
| **50%** | 8.000000 | 380.000000 | 6.000000 |
| **75%** | 8.000000 | 508.250000 | 8.000000 |
| **max** | 32.000000 | 1750.000000 | 14.000000 |

```
df.describe(include = ['object'])
```

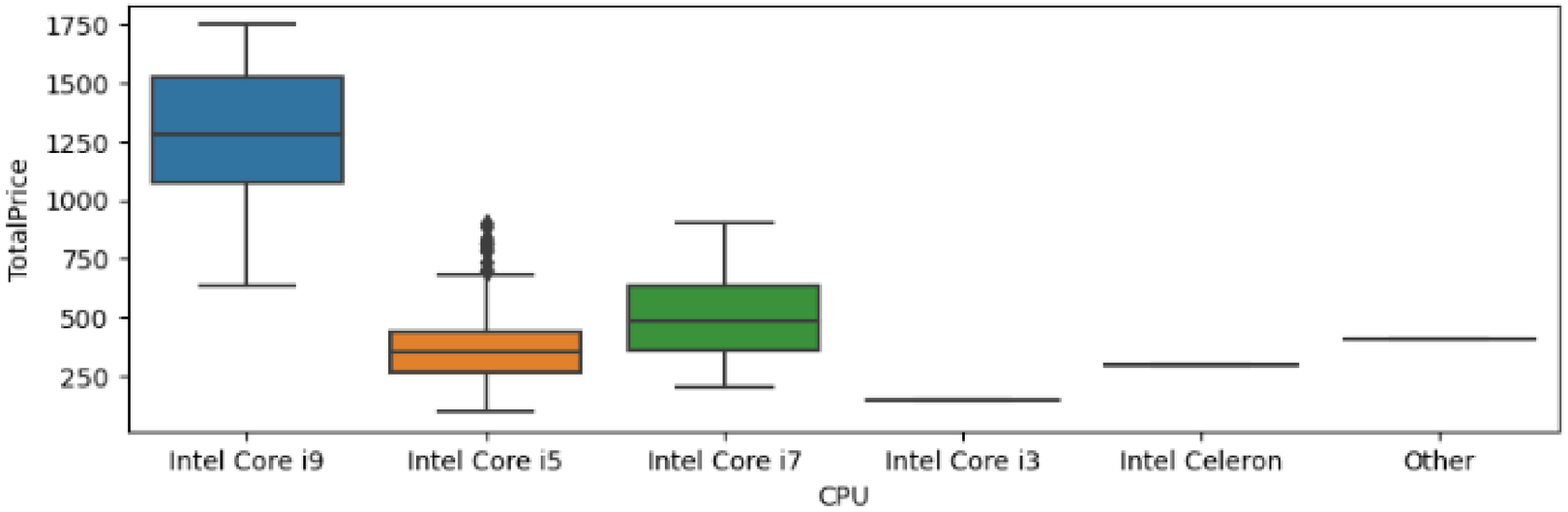
| | Discription | CPU | VGA |
|---|---|---|---|
| **count** | 1055 | 1055 | 1055 |
| **unique** | 2 | 6 | 3 |
| **top** | Used | Intel Core i5 | 4GB |
| **freq** | 614 | 703 | 973 |

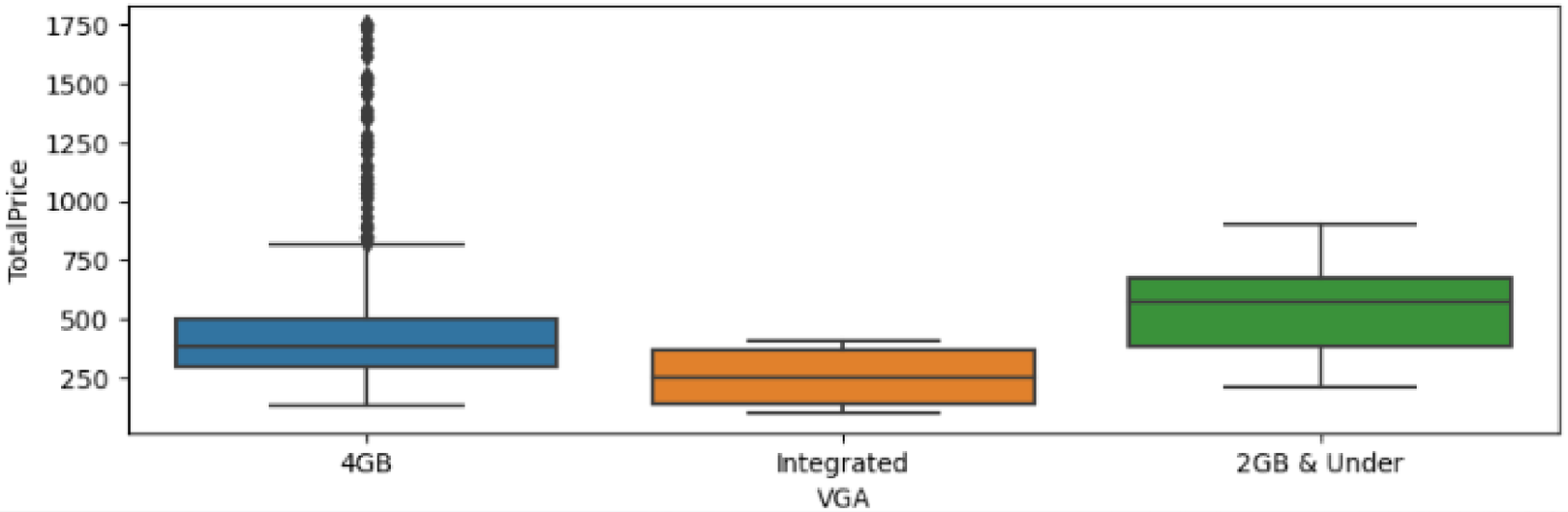
# EXPLORATORY DATA ANALYSIS (EDA)

# HISTORGRAM

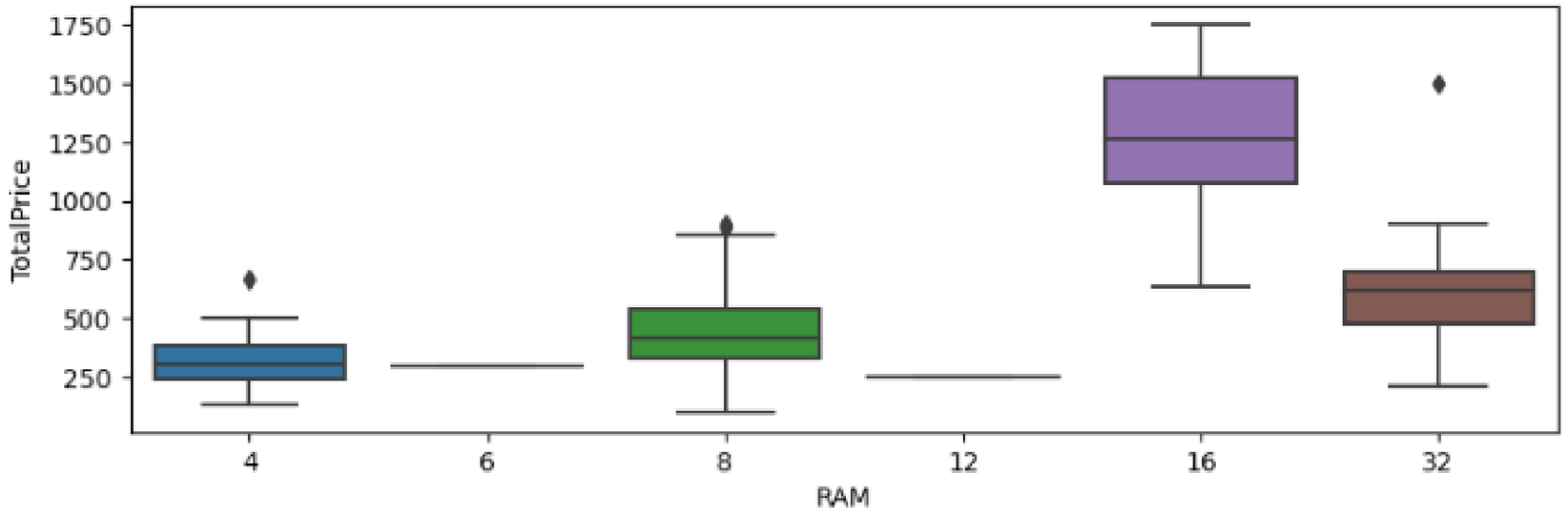

0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
Density
0
500
1000
1500
2000
TotalPrice

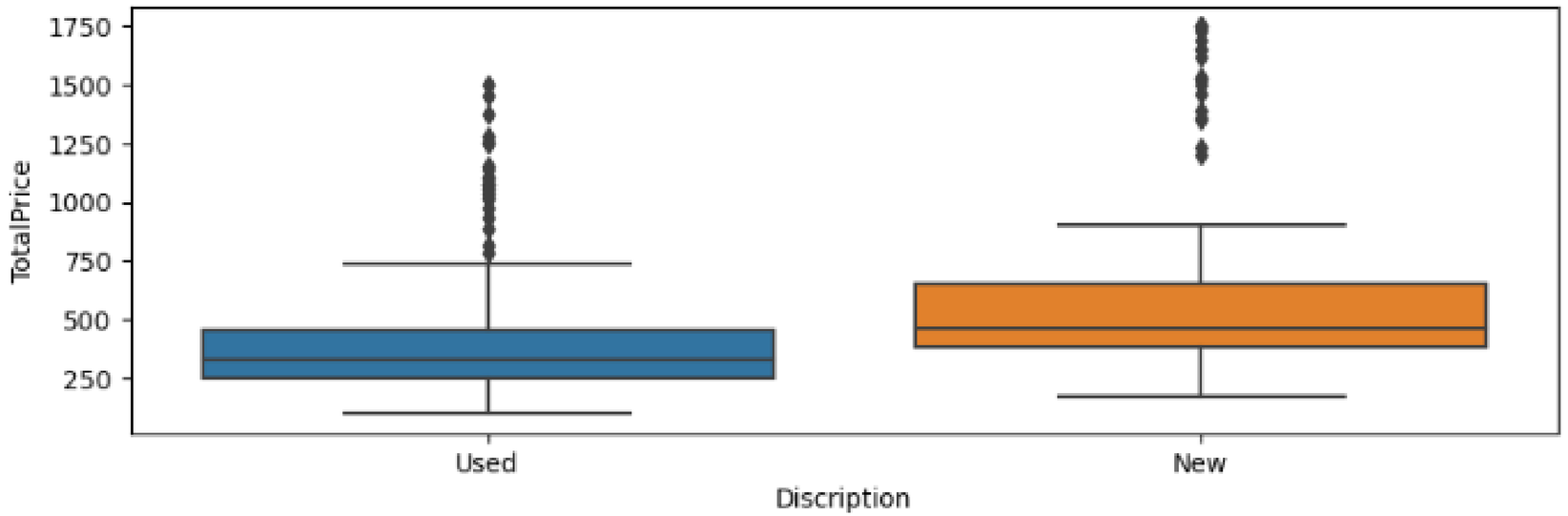
# BOX PLOT


1750
1500
1250
1000
750
500
250
TotalPrice
Intel Core i9
Intel Core i5
Intel Core i7
Intel Core i3
Intel Celeron
Other
CPU

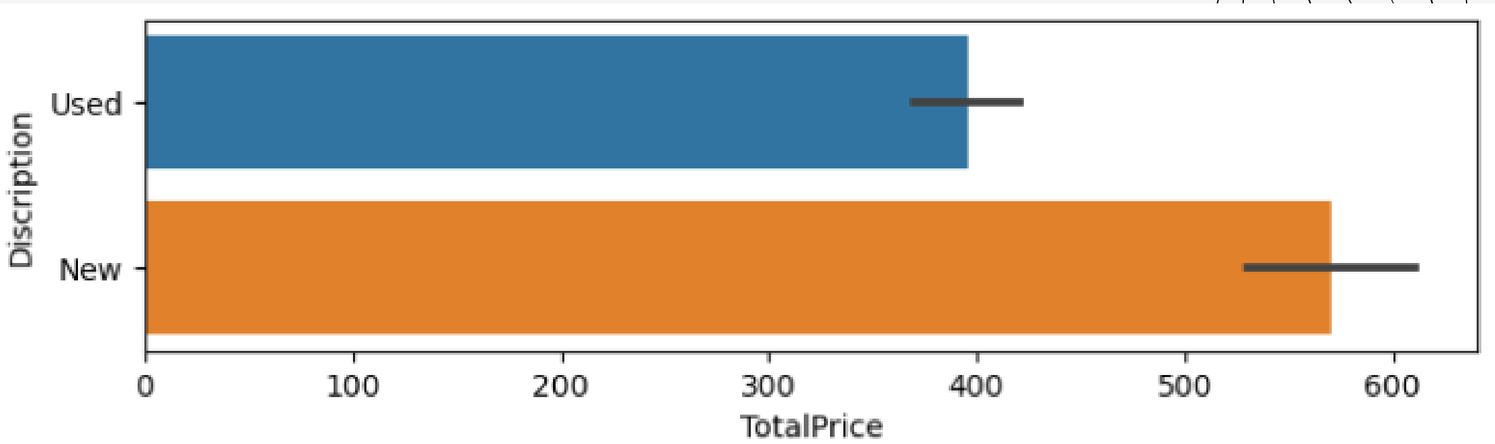

1750
1500
1250
1000
750
500
250
TotalPrice
4GB
Integrated
2GB & Under
VGA

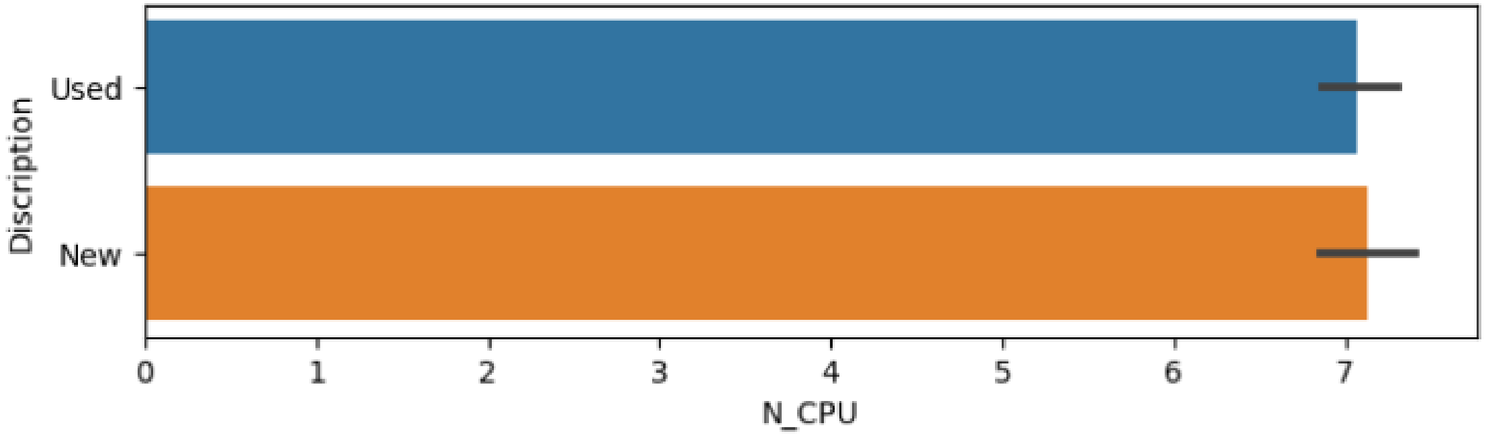
# BOX PLOT

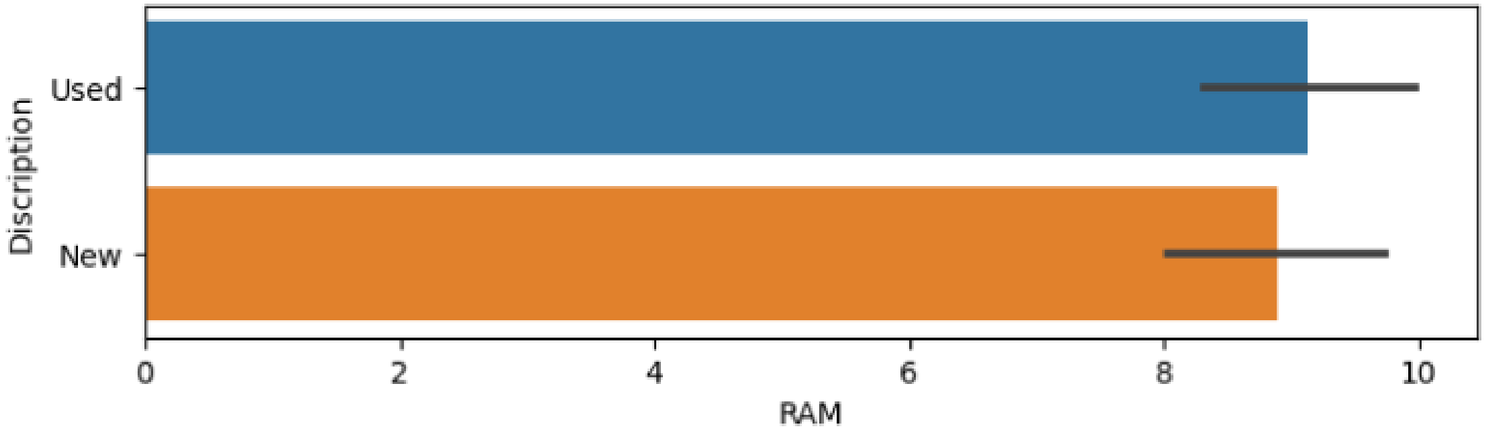

1750
1500
1250
1000
750
500
250
TotalPrice
4
6
8
12
16
32
RAM

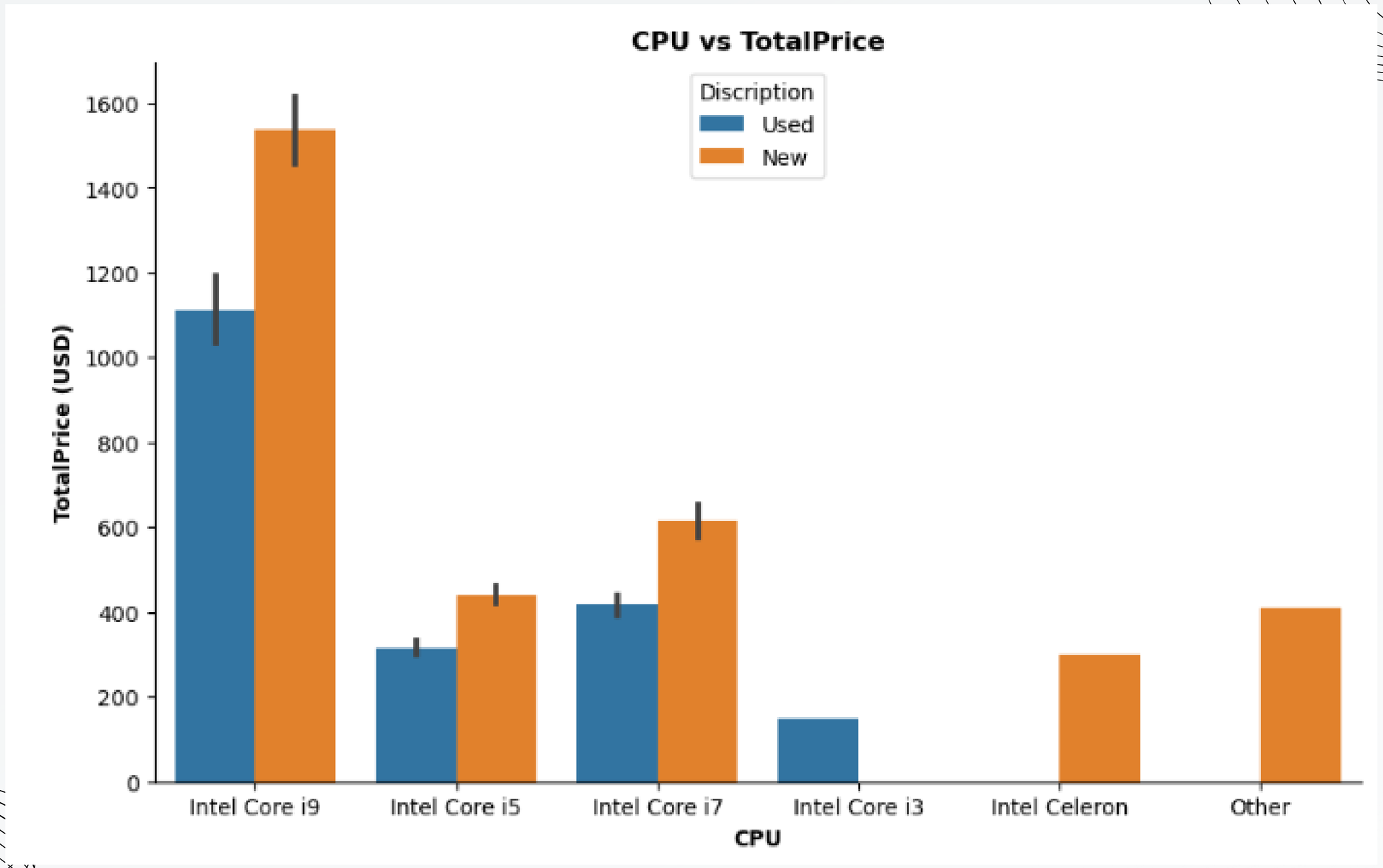

1750
1500
1250
1000
750
500
250
TotalPrice
Used
New
Discription

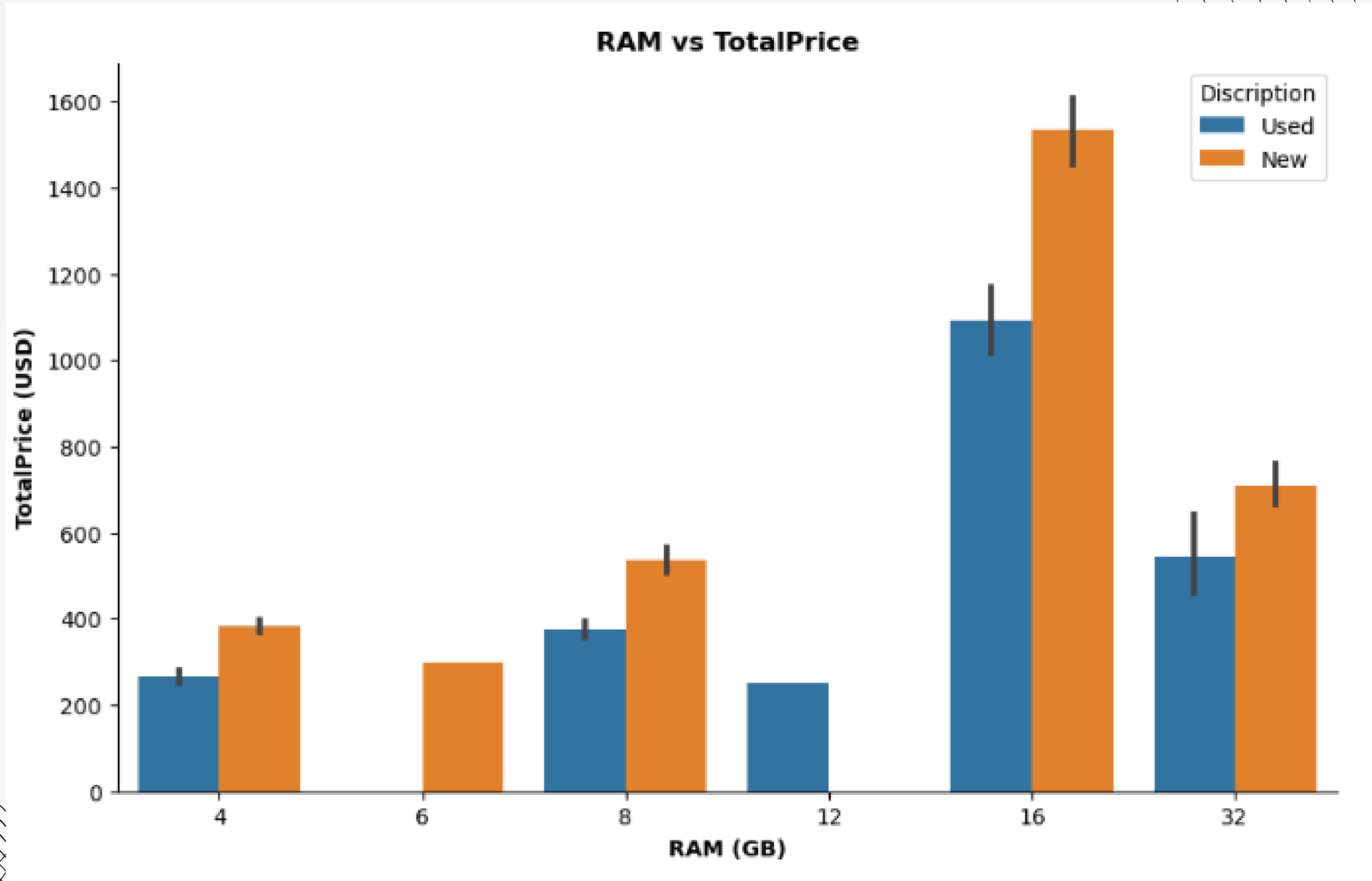
# BAR GRAPH

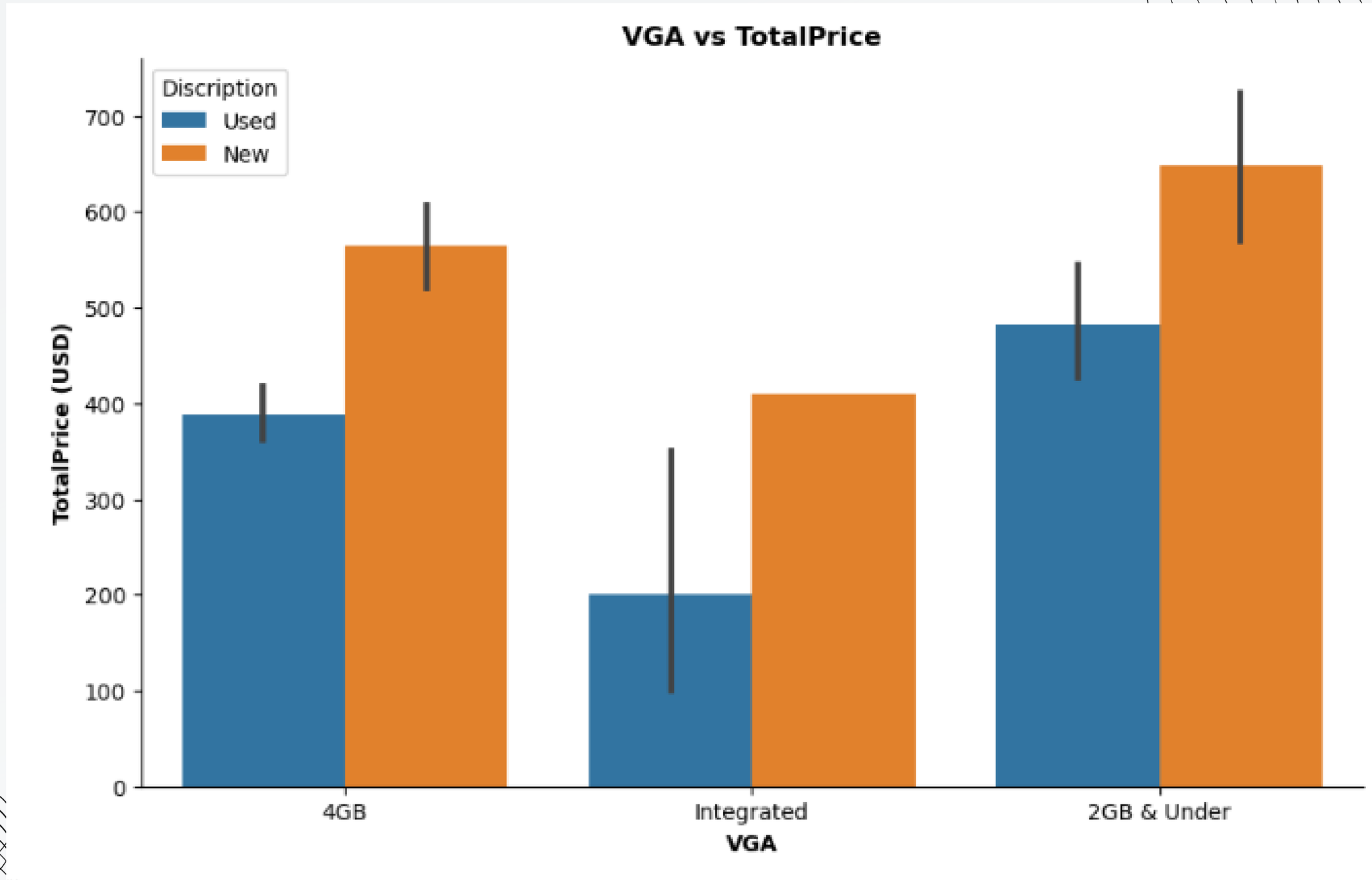

Discription
Used
New
0
100
200
300
400
500
600
TotalPrice


Discription
Used
New
0
1
2
3
4
5
6
7
N_CPU

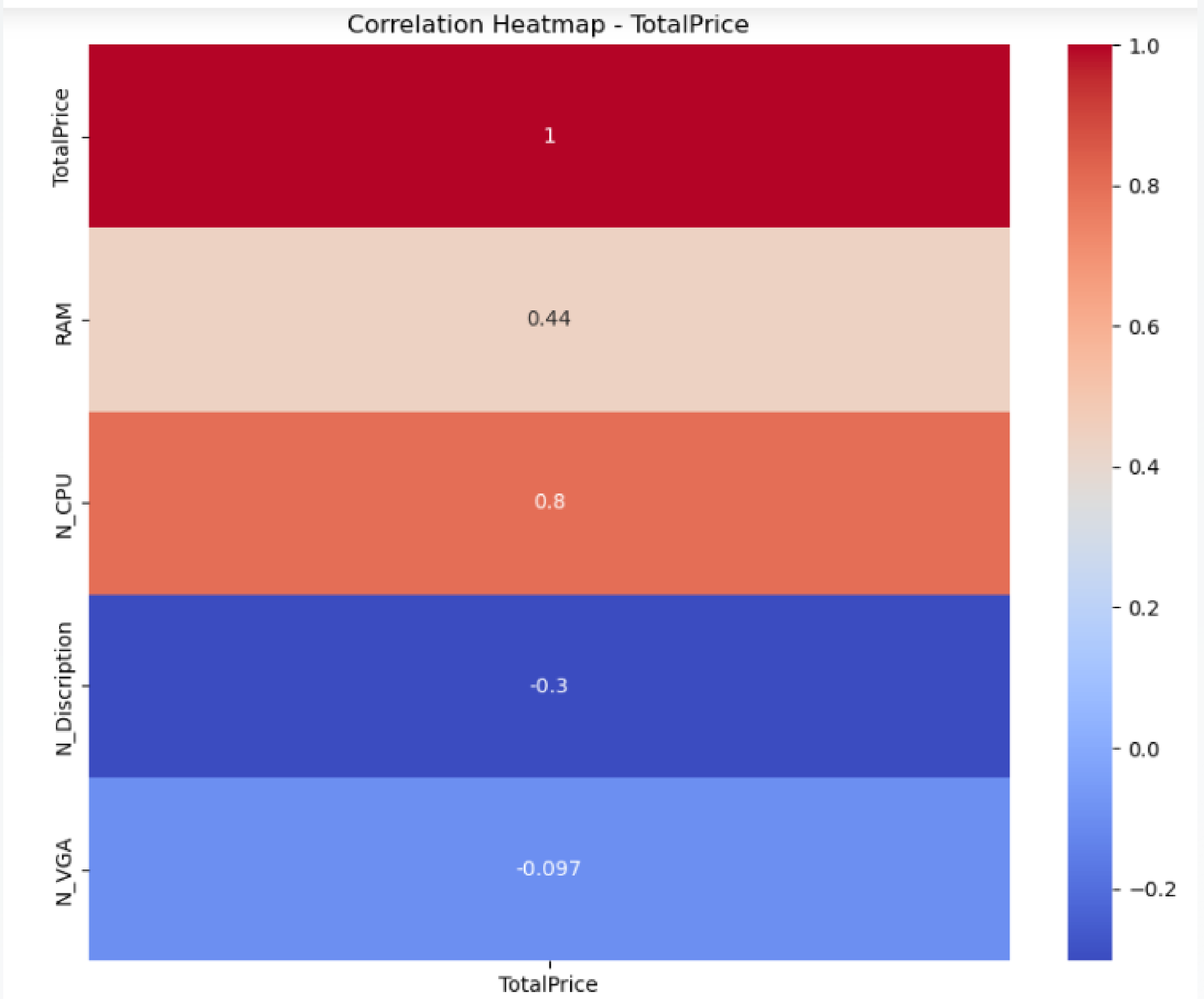

Discription
Used
New
0
2
4
6
8
10
RAM


CPU vs TotalPrice
Discription
Used
New
TotalPrice (USD)
0
200
400
600
800
1000
1200
1400
1600
Intel Core i9
Intel Core i5
Intel Core i7
Intel Core i3
Intel Celeron
Other
CPU


RAM vs TotalPrice
Discription
Used
New
TotalPrice (USD)
0
200
400
600
800
1000
1200
1400
1600
4
6
8
12
16
32
RAM (GB)


VGA vs TotalPrice
Discription
Used
New
TotalPrice (USD)
0
100
200
300
400
500
600
700
4GB
Integrated
2GB & Under
VGA

# HEATMAP


Correlation Heatmap - TotalPrice
TotalPrice
RAM
N_CPU
N_Discription
N_VGA
1
0.44
0.8
-0.3
-0.097
TotalPrice
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

# FEATURE ENGINEERING

TRANSFORM AND CREATE NEW FEATURES TO ENHANCE THE PREDICTIVE POWER OF THE DATA.

# LABEL ENCODING

**N_CPU 14= CPU INTEL CORE I9**
**N_CPU 8= CPU INTEL CORE I7**
**N_CPU 6= CPU INTEL CORE I5**
**N_CPU 4= CPU INTEL CORE I2**
**N_CPU 2= CPU INTEL CELERON**
**N_CPU 0 = OTHER**

**N_VGA 1 = VGA 4G**
**N_VGA 2 = VGA INTEGRATED**
**N_VGA 0 = VGA 2G & UNDER**

| | TotalPrice | RAM | N_CPU | N_Discription | N_VGA |
|---|---|---|---|---|---|
| 0 | 1500.0 | 32 | 14 | 1 | 1 |
| 1 | 350.0 | 8 | 6 | 1 | 2 |
| 2 | 360.0 | 8 | 6 | 1 | 0 |
| 3 | 440.0 | 8 | 8 | 1 | 1 |
| 4 | 388.0 | 8 | 6 | 1 | 0 |

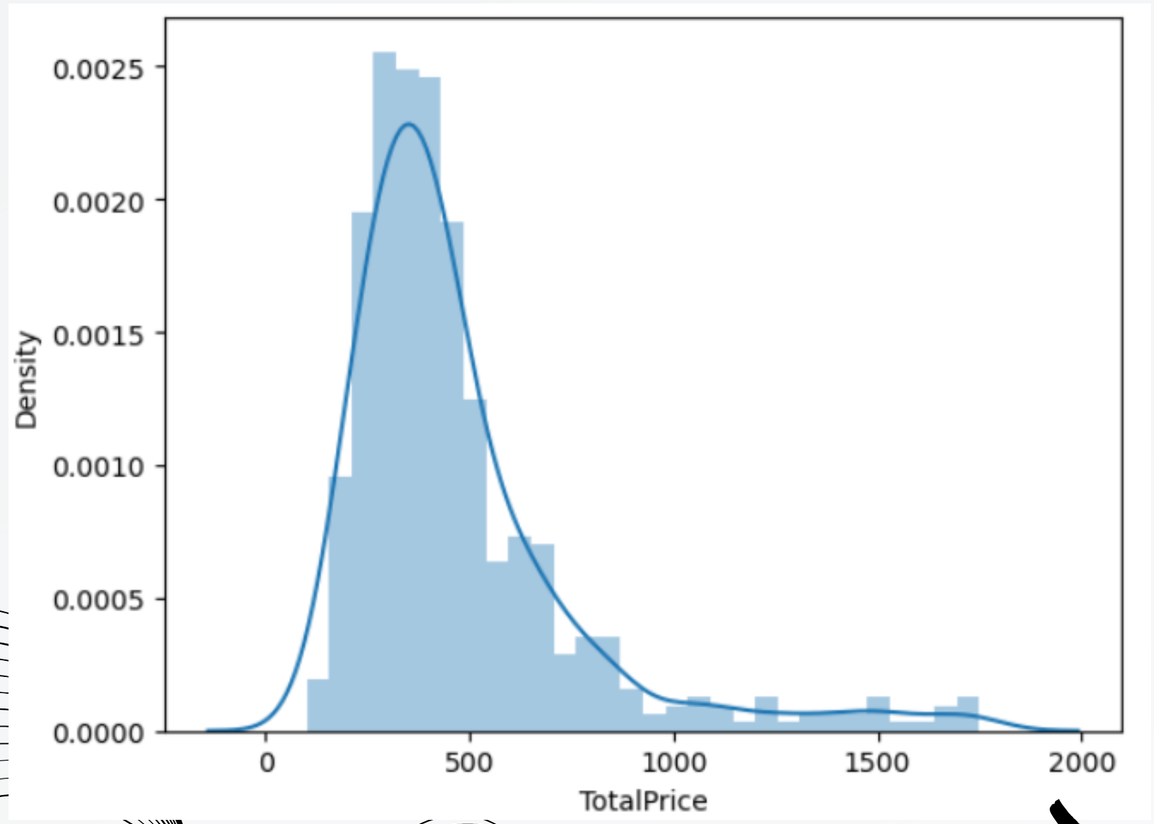
**DISCRIPTION 1 = USED**
**DISCRIPTION 0 = NEW**

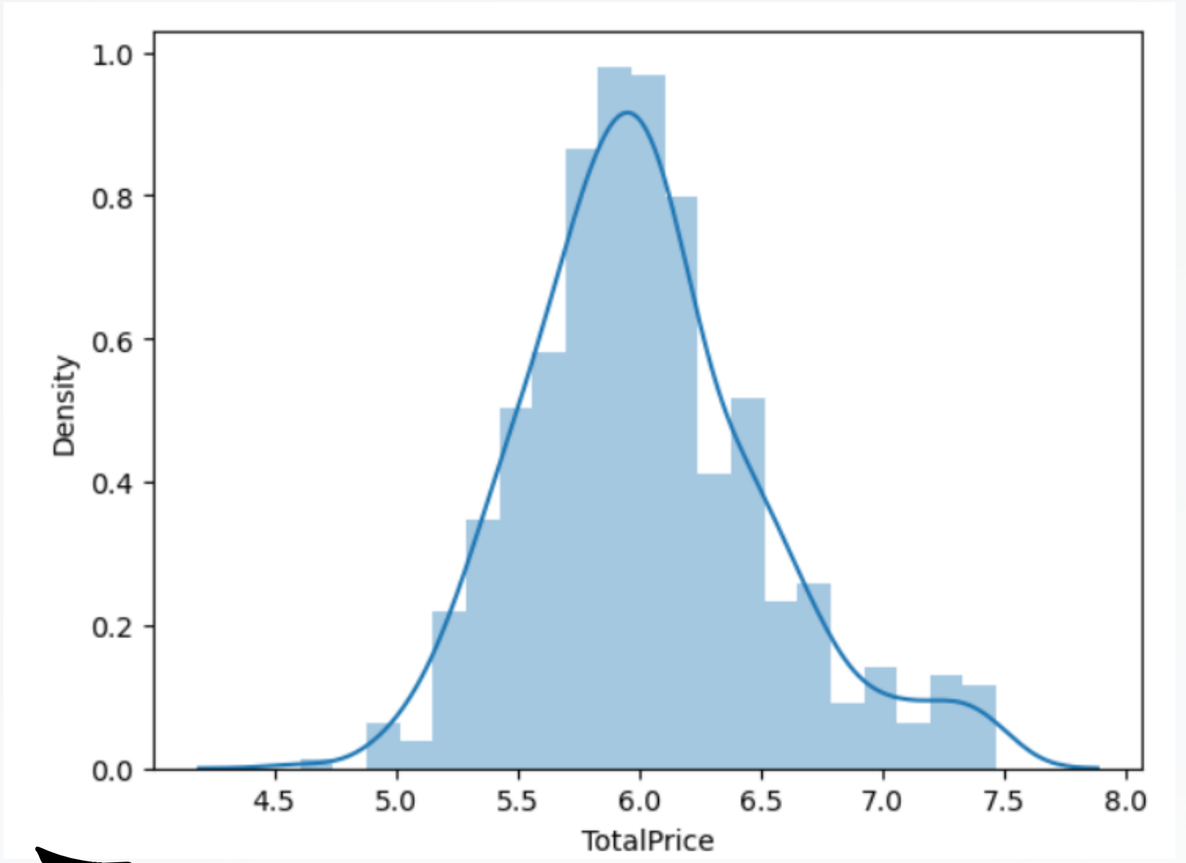
## Feature Importance using f_regression & SelectKBest

| | Feature | Importance Score |
|---|---|---|
| 1 | N_CPU | 1008.364786 |
| 0 | RAM | 132.602823 |
| 2 | N_Discription | 57.376478 |
| 3 | N_VGA | 5.411309 |

```
SelectKBest:
RAM
N_CPU
N_Discription
```

# TRANFORMING INDEPENDENT VARIBLE



# MODEL SELECTION

**CREATE VARIABLE WITH OUTCOME VARAIBLE.**

```
X = df1.drop(columns=['TotalPrice','N_VGA'])
y = np.log(df1['TotalPrice'])
```

# MULTIPLE LINEAR REGRESSION

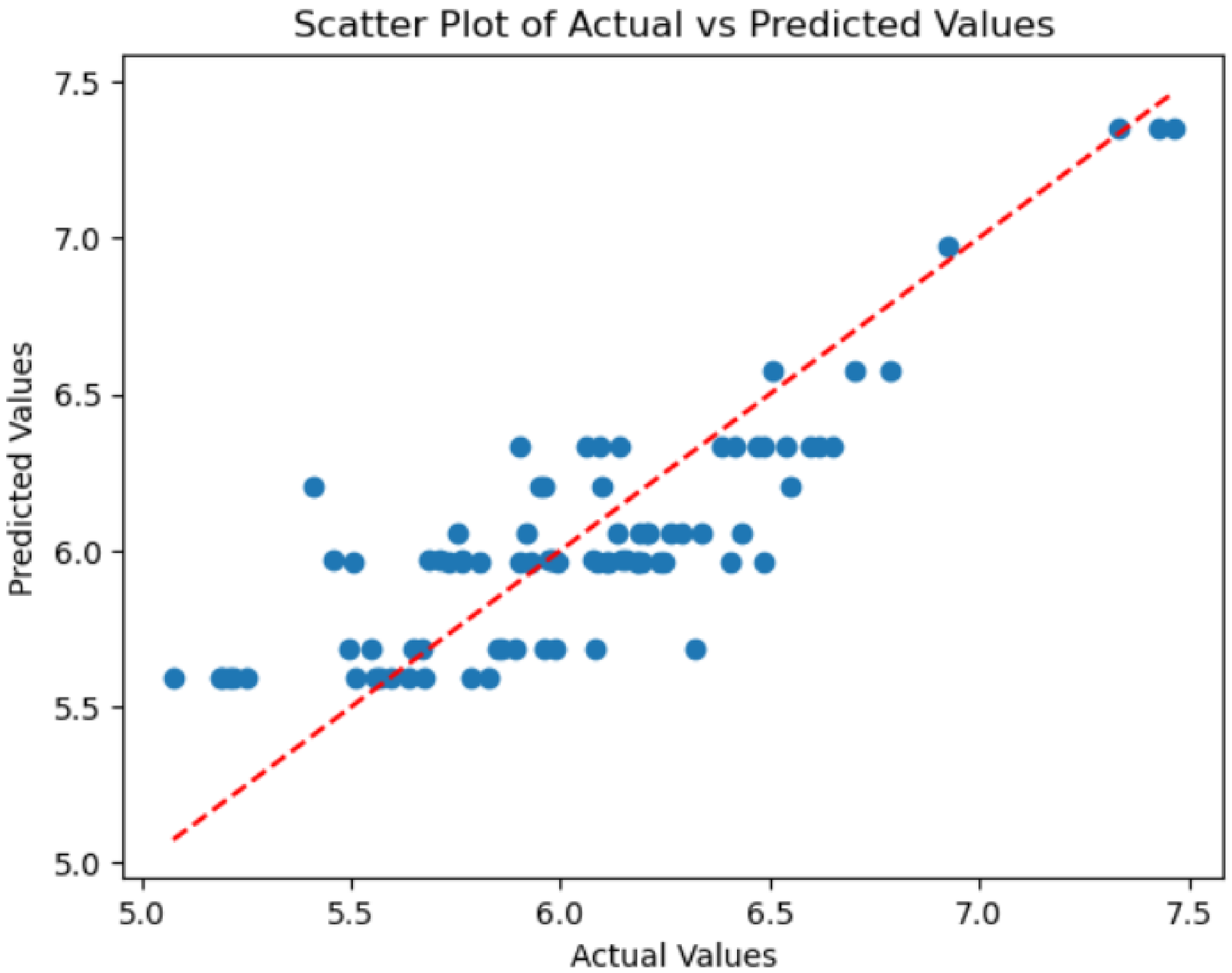
OLS Regression Results

| | | | |
|---|---|---|---|
| **Dep. Variable:** | TotalPrice | **R-squared:** | 0.699 |
| **Model:** | OLS | **Adj. R-squared:** | 0.697 |
| **Method:** | Least Squares | **F-statistic:** | 370.1 |
| **Date:** | Sat, 15 Jul 2023 | **Prob (F-statistic):** | 2.56e-124 |
| **Time:** | 20:39:42 | **Log-Likelihood:** | -56.576 |
| **No. Observations:** | 483 | **AIC:** | 121.2 |
| **Df Residuals:** | 479 | **BIC:** | 137.9 |
| **Df Model:** | 3 | | |
| **Covariance Type:** | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **const** | 5.0440 | 0.045 | 111.028 | 0.000 | 4.955 | 5.133 |
| **RAM** | 0.0217 | 0.002 | 12.237 | 0.000 | 0.018 | 0.025 |
| **N_CPU** | 0.1397 | 0.006 | 23.417 | 0.000 | 0.128 | 0.151 |
| **N_Discription** | -0.3733 | 0.025 | -14.877 | 0.000 | -0.423 | -0.324 |

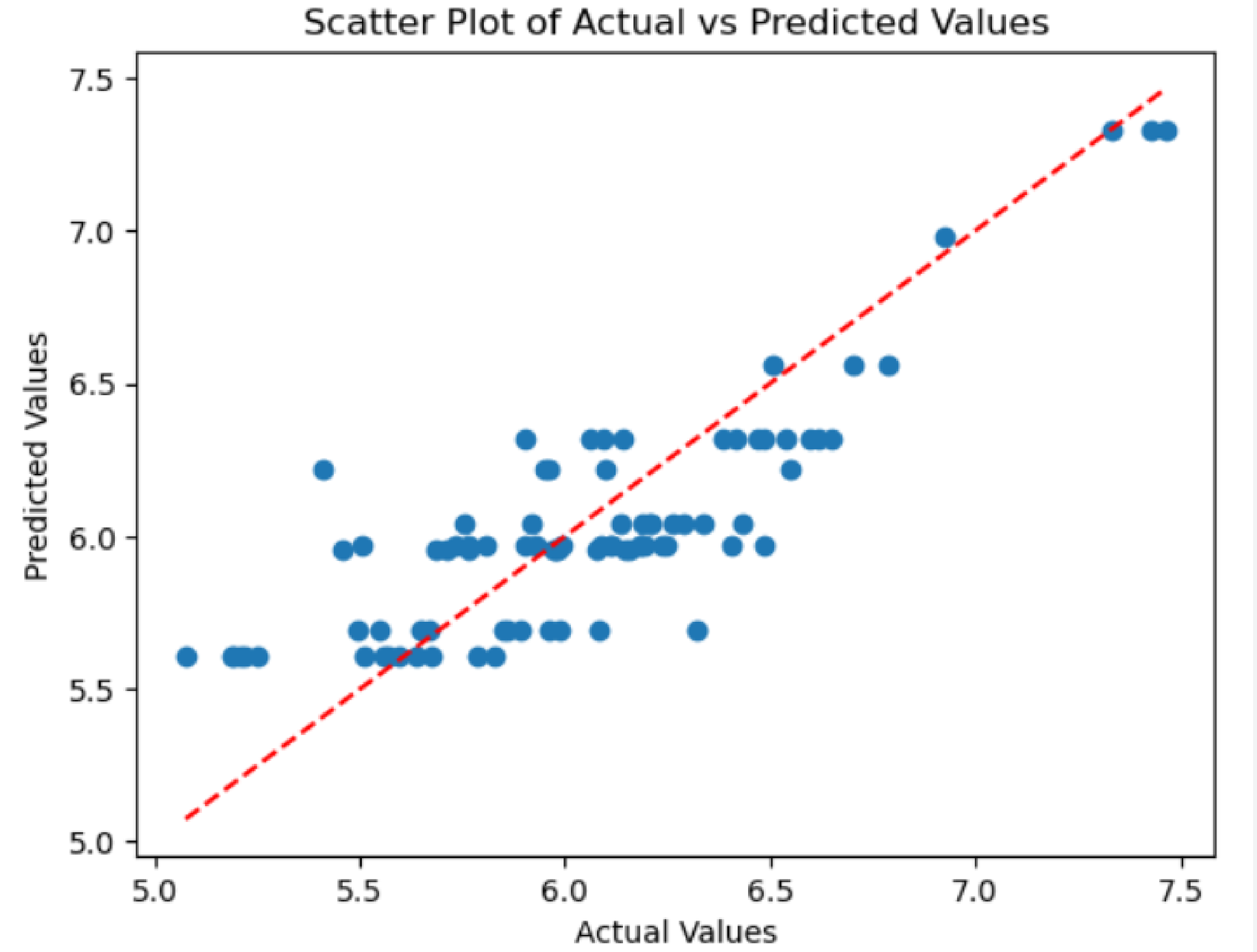
| | | | |
|---|---|---|---|
| **Omnibus:** | 24.585 | **Durbin-Watson:** | 1.831 |

```
R2 score 0.6941456750648596
MSE 0.06835859323925936
MAE 0.21307632554362332
```



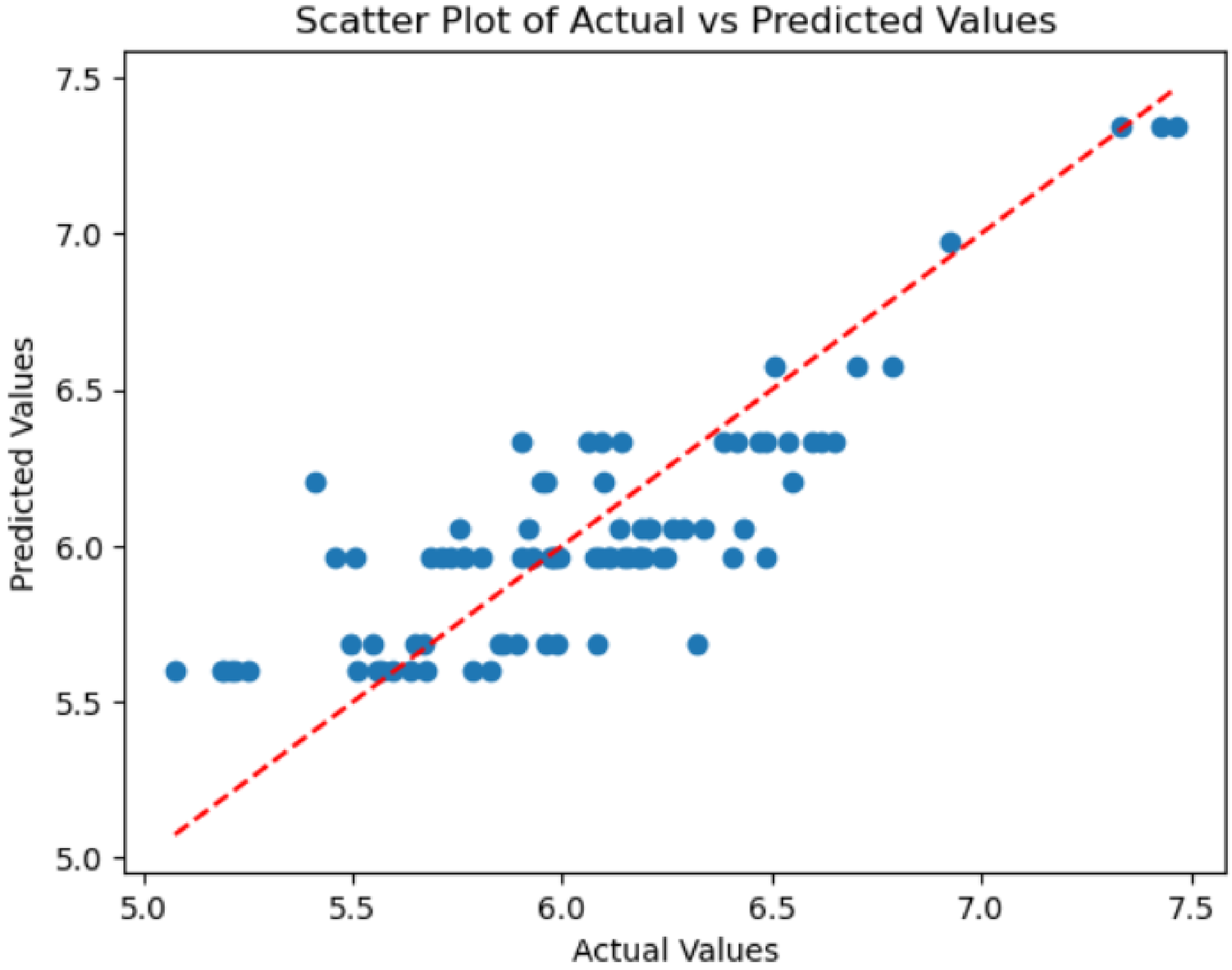
# RIGDE REGRESSION

```
R2 score 0.6911118921256408
MSE 0.06903664523006327
MAE 0.21623357186850203
```



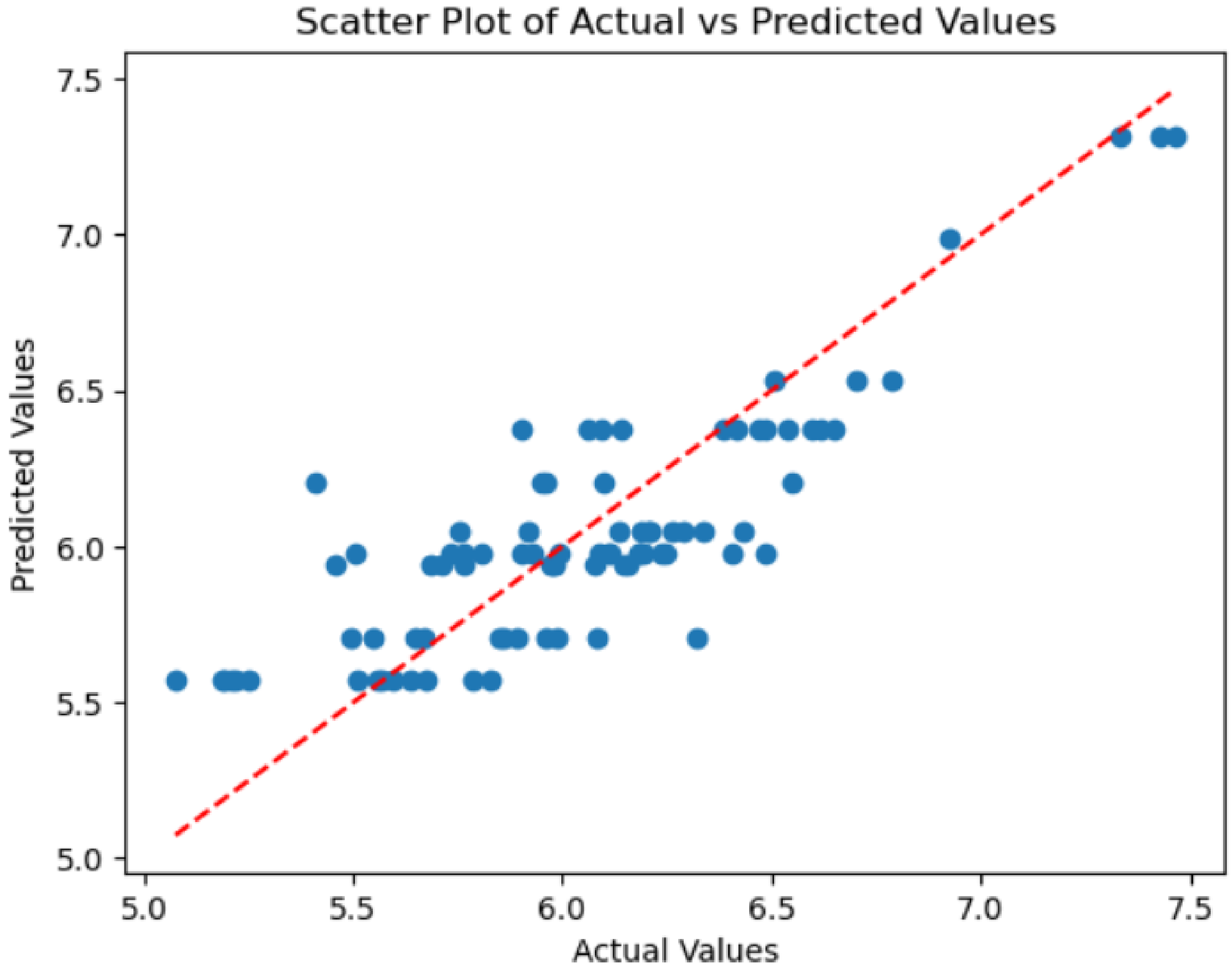
# LASSO REGRESSION

```
R2 score 0.6937996430889993
MSE 0.06843593155739701
MAE 0.21349474534140303
```



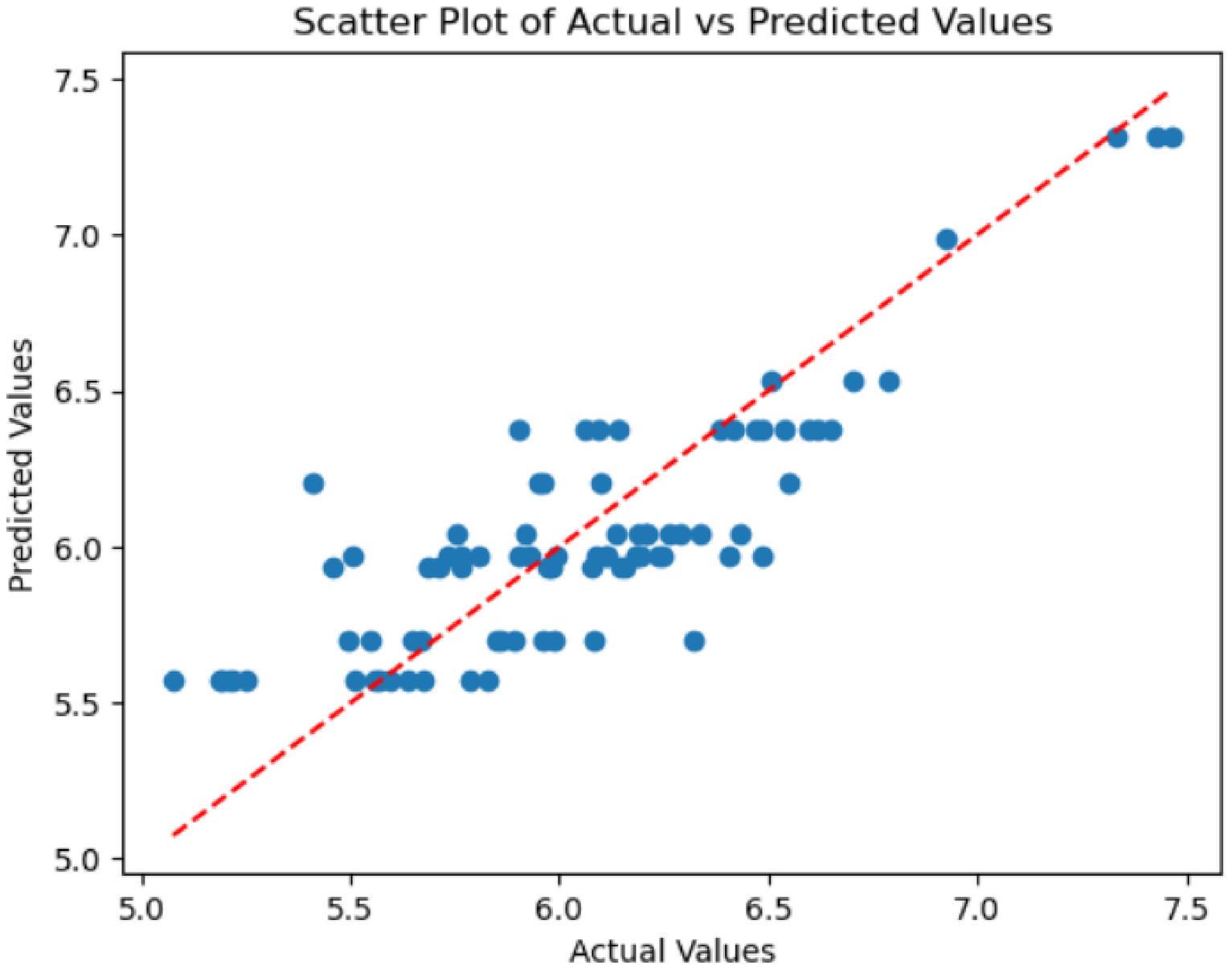
# DECISION TREE

```
R2 score 0.7038093486507853
MSE 0.06619875740238718
MAE 0.2109959689709732
```



# RANDOM FOREST

```
R2 score 0.7011022862950059
MSE 0.06680378718083242
MAE 0.2124421711219747
```

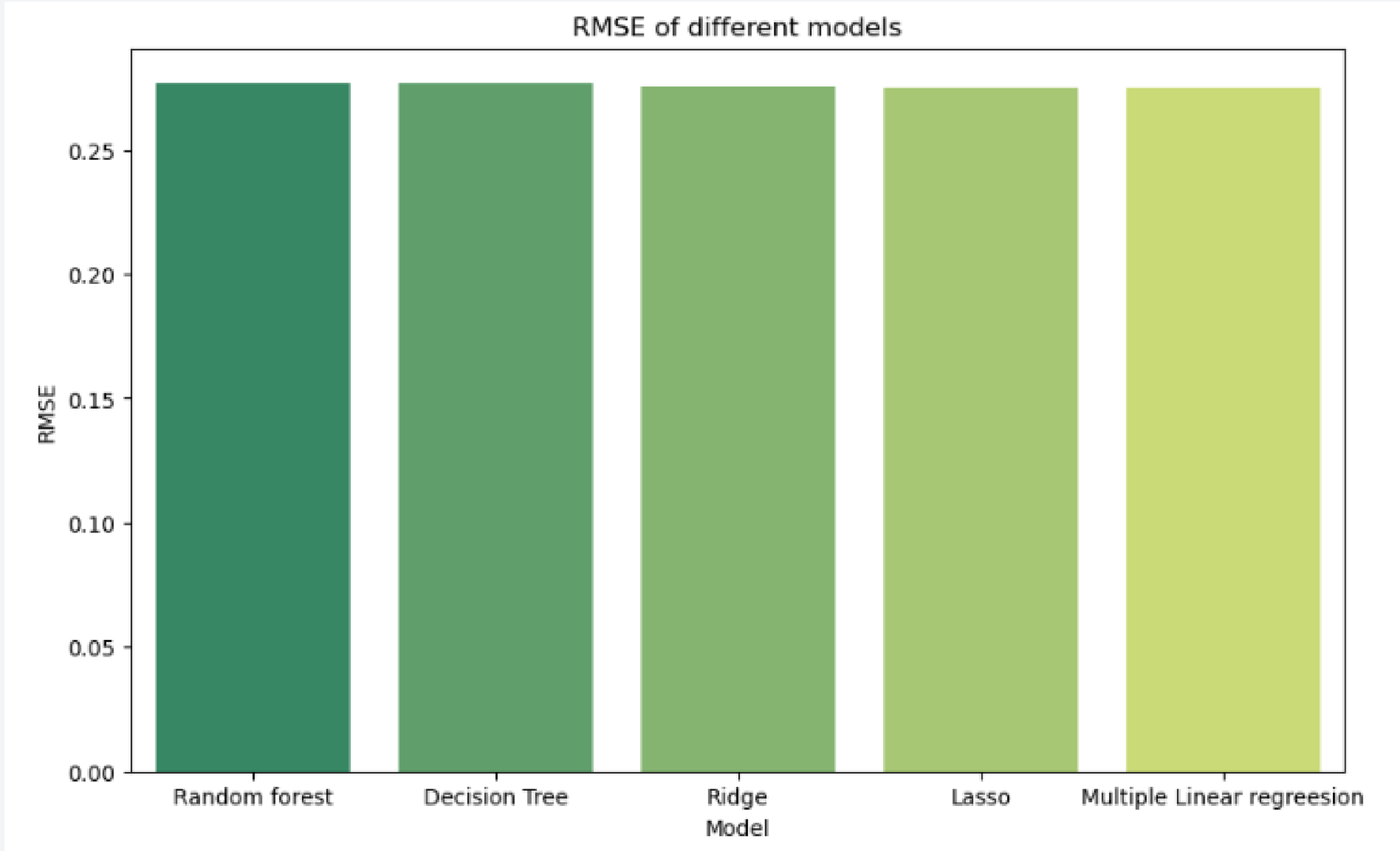


# MODEL COMPARASION

COMPARE ALL SELECTED MODEL BY THEIR ROOT MEAN SQUARED ERROR(RMSE)

| | Random forest | Decision Tree | Ridge | Lasso | Multiple Linear regreesion |
|---|---|---|---|---|---|
| **RMSE** | 0.277045 | 0.276639 | 0.275308 | 0.274814 | 0.274811 |

**WE CAN SEE ALL MODELS HAVE RMSE ARE LITTLE BIT DEFFERENCE. HOWEVER, WE STILL CAN SEE MULTIPLE LINEAR REGRESSION IS LOWEREST THAN OTHERS.**



# CONCLUSION

IMPORTANT FACTORS THAT AFFECT PC COMPONENT PRICES:

- CPU
- RAM
- DESCRIPTION( USED OR NEW)

WE BELIEVE THAT THE MODELS WE HAVE CHOSEN WILL ASSIST YOU TO PREDICT THE PC COMPONENT THAN THE TRADITIONAL METHODS.