Exploratory Data Analysis for Part of Computer Prices

PHUONG BUNTHOEUN e20201854
MEN CHANCHHORPORN e20201146
LIM SUNHENG e20200807
LEAT SEANGLONG e20200971
MEACH SEAKLAV e20200683
NGEAV BONAT e20201691

Taught by: Professor Chan Sophal

I. Introduction

EDA stands for Exploratory Data Analysis. It is a process of analyzing and summarizing data sets to extract meaningful insights and identify patterns, trends, relationships, and anomalies in the data. EDA involves visualizing and interpreting data using statistical tools and techniques such as histograms, scatter plots, box plots, correlation matrices, and regression analysis.

The primary goal of EDA is to understand the underlying structure of the data and to uncover any hidden patterns or relationships that may exist. EDA can help identify data quality issues, missing values, outliers, and other anomalies that may affect the accuracy of subsequent analyses.

II. Data Cleaning

Data cleaning is an essential step in the process of Data Analysis. The goal of data cleaning is to ensure that data is accurate, complete, and consistent. Data cleaning techniques include handling missing value, handling outliers and handing with categorical variables. In our dataset, Before removing outliers we noticed that a column 'Dimensions' has complex values after loading the dataset, so we extract it to remove some charaters that complicated such that 'GB, Other ' and replace it and replace the object data by using label encoding to convert numerical data.

- Data After cleaning and drop some columns

	Discription	RAM(GB)	CPU	TotalPrice($)
0	Used	32	Intel Core i9	1150
1	Used	8	Intel Core i5	220
2	Used	8	Intel Core i5	599
3	Used	8	Intel Core i7	235
4	Used	8	Intel Core i5	388
5	Used	8	Intel Core i5	328
6	Used	8	Intel Core i5	399
7	Used	8	Intel Core i5	259

The steps of data cleaning about this data set is shown below:

A. *Handling missing values*

In our data set has no missing values.

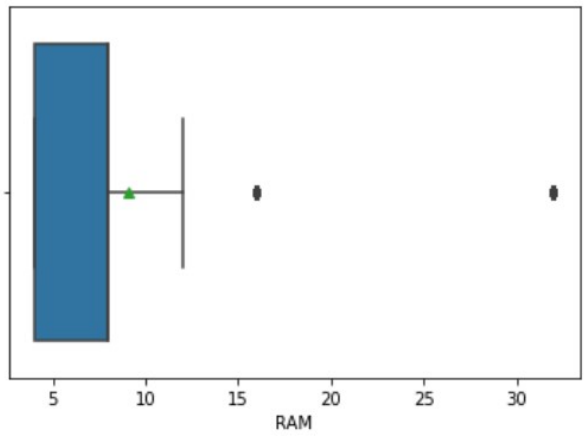
```
df.isnull().sum()

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

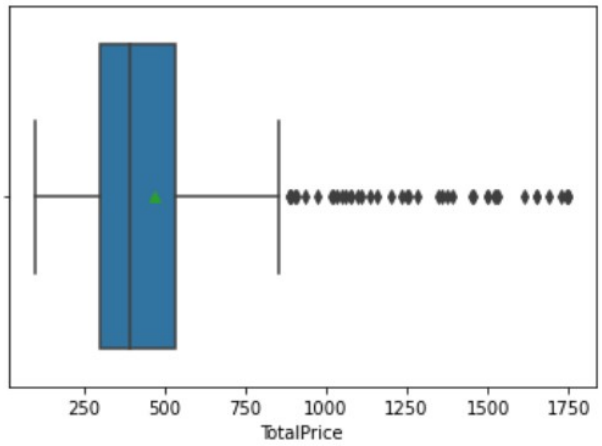
Fig. 1. Missing value

B. *Handling Outliers*

We use boxplot for checking outlier.



Check Outlier RAM



Check Outlier Prices

- The average price for a particular type of CPU is $350, but there is one listing for 1000$, this would be considered an outlier. Similarly, if the average price for a certain type of RAM module is $100, but there is one listing for 500$, this would also be considered an outlier. we don't need to drop Outlier beacuase of We want to analyse price for PCU components.

C. *Encoding Categorical Variable*

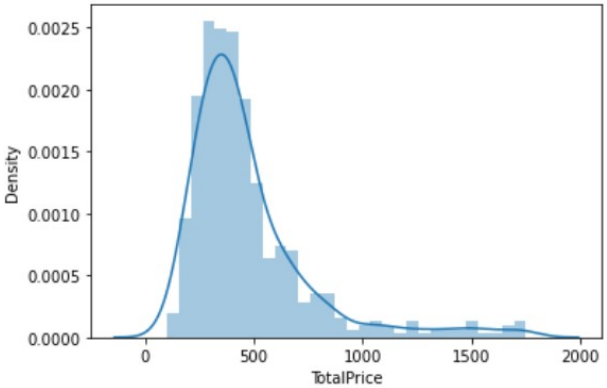
Since machine learning algorithm can understand only numerical variables, we need to label the categorical variables in our dataset with numbers. There are several ways for encoding categorical variables such as one-hot encoding and label encoding.

- One-hot encoding is the process of converting categorical data into numerical data by creating a new binary column for each category.
- Label encoding is the process of converting categorical data into numerical data by assigning a unique number to each category.

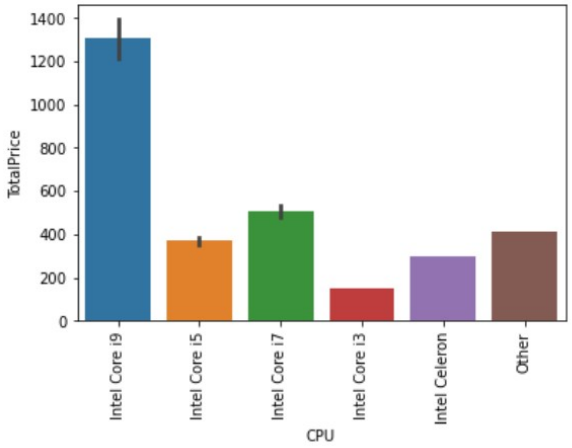
In our dataset, we decided to use one-hot encoding on the book format feature since there are only 3 unique values in format column. Also using using label encoding can make the machines misunderstand that there is a ranking between values, while the book format seems have ranking.

III. Data Exploration

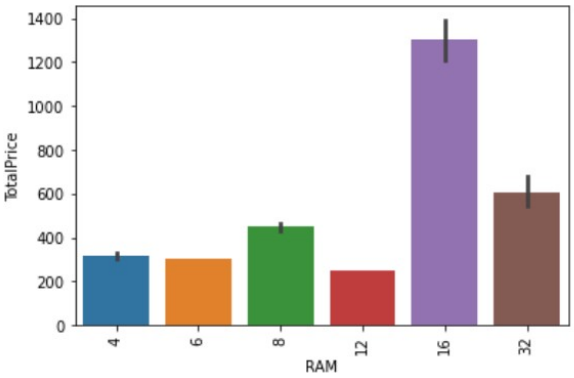
We plot data and histogram, Piechat, densityplot etc.



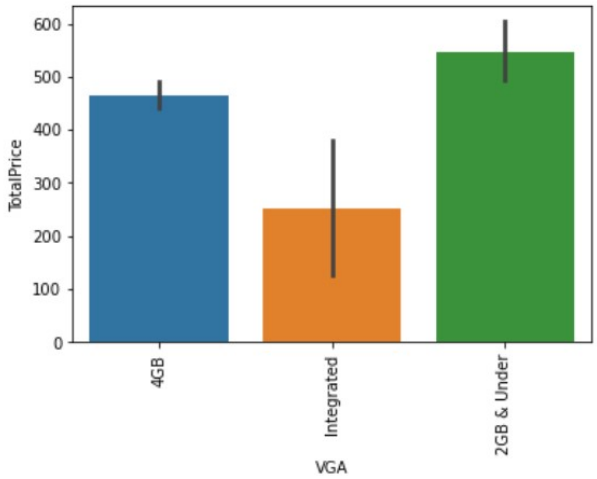
Check DesityPlot Price



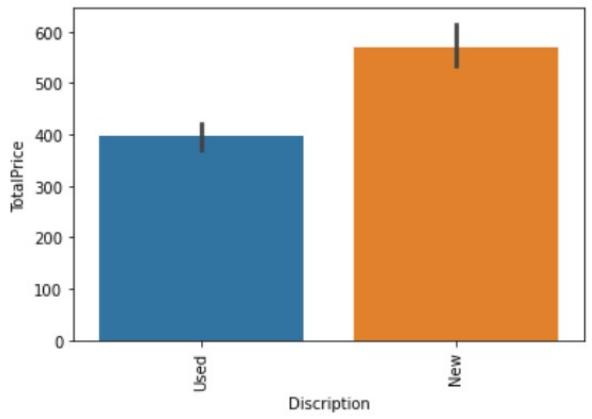
Histogram for CPU with Price



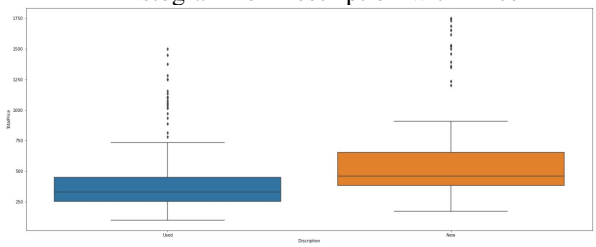
Histogram for RAM with Price



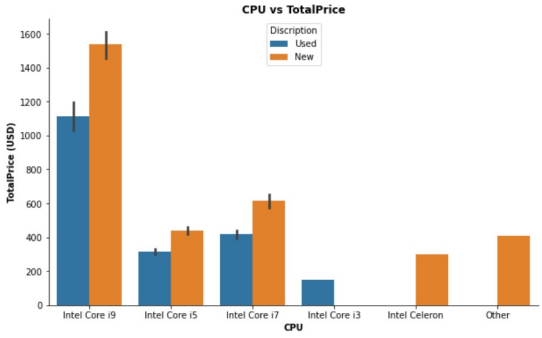
Histogram for VGA with Price



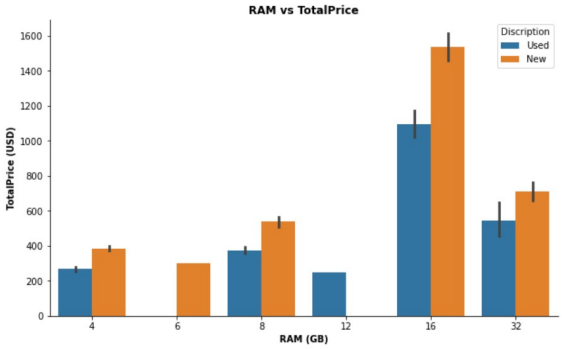
Histogram for Description with Price



Box plot for Description with Price



Histogram for Description(use,New) with CPU and Price



Histogram for Description(use,New) with RAM and Price

IV. Feature Engineering

Feature engineering is the process involves selecting, transforming, extracting, combining, and manipulating raw data to generate the desired variables for analysis or predictive modeling. The motivation is to use these extra features to improve the quality of results from a machine learning process,compared with supplying only the raw data to the machine learning process. This section will discuss how feature engineering techniques are being used the Project.

A. *Features Important for features*

Create variables that important in data.

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

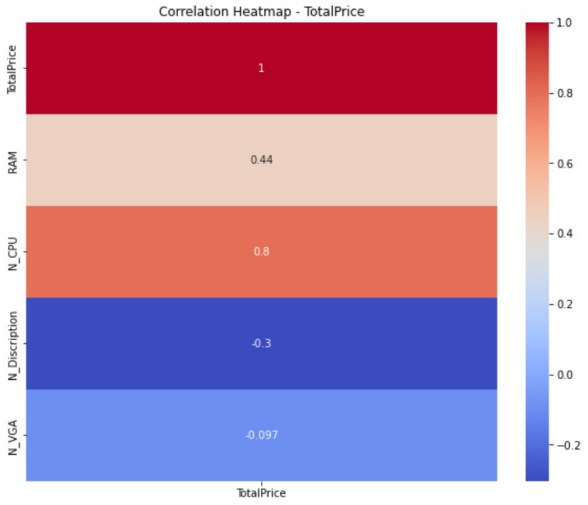
Feature Important

B. *Features creation and reduction*

Dimensionality reduction is the process of reducing the number of features (or dimensions) in a dataset while retaining as much information as possible. For instance, as in our dataset we got the features such as RAM, CPU, Price. We need tranform data to numerical data By using Label Encoding.

	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

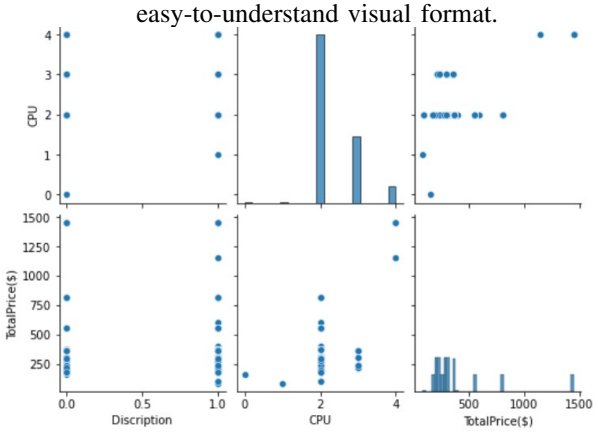
Feature are Important



Feature correlation

V. Data Visualization

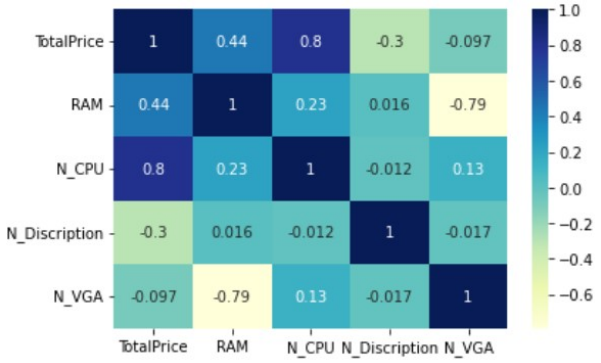
Data visualization is the process of representing data or information in a visual format such as charts, graphs, tables, maps, and diagrams. The goal of data visualization is to help people understand complex data by presenting it in an easy-to-understand visual format.



Pair Plot

	TotalPrice	RAM	N_CPU	N_Discription	N_VGA
count	569.000000	569.000000	569.000000	569.000000	569.000000
mean	471.639719	9.029877	7.096661	0.565905	0.917399
std	285.977472	7.303348	2.104524	0.496074	0.299993
min	100.000000	4.000000	0.000000	0.000000	0.000000
25%	300.000000	4.000000	6.000000	0.000000	1.000000
50%	395.000000	8.000000	6.000000	1.000000	1.000000
75%	533.000000	8.000000	8.000000	1.000000	1.000000
max	1750.000000	32.000000	14.000000	1.000000	2.000000

Describe Data



Check Correlation

VI. Model Building

Explain the concept of each model:

- Multiple Linear Regression: Multiple Linear Regression is a statistical model used to analyze the relationship between multiple independent variables and a dependent variable. It assumes a linear relationship between the predictors and the response variable. The model estimates the coefficients for each predictor, indicating the strength and direction of their influence on the response variable.
- Ridge Regression: Ridge Regression is a technique used in linear regression to mitigate the problem of multicollinearity. It adds a penalty term to the least squares objective function, which shrinks the coefficient estimates towards zero. This helps to reduce the impact of highly correlated predictors and improves the stability of the model.
- Lasso Regression: Lasso Regression, similar to Ridge Regression, is a regularization technique used in linear regression. It also adds a penalty term to the objective function but uses the absolute values of the coefficients instead of their squares. Lasso Regression not only addresses multicollinearity but also performs feature selection by driving some coefficients to exactly zero. This makes it useful for models with a large number of predictors.
- Random Forest Regression: Random Forest Regression is an ensemble learning method that combines multiple decision trees to make predictions. It constructs a multitude of decision trees using random subsets of the training data and features. The final prediction is obtained by averaging the predictions of all the individual trees. Random Forest Regression is known for its ability to handle complex relationships and capture non-linear patterns in the data.

A. *Multiple Linear Regression*

Using Summary For 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:	22:54:01	**Log-Likelihood:**	-56.576
No. Observations:	483	**AIC:**	121.2
Df Residuals:	479	**BIC:**	137.9
Df Model:	3		
Covariance Type:	nonrobust		

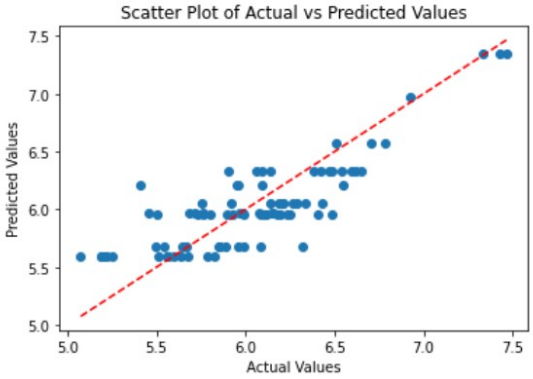
	coef	std err	t	P>|t|	[0.025	0.975]
const	5.0440	0.045	111.028	0.000	4.955	5.133
RAM	0.0217	0.002	12.237	0.000	0.018	0.025
N_CPU	0.1397	0.006	23.417	0.000	0.128	0.151
N_Discription	-0.3733	0.025	-14.877	0.000	-0.423	-0.324

Omnibus:	24.585	**Durbin-Watson:**	1.831
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	31.922
Skew:	-0.447	**Prob(JB):**	1.17e-07
Kurtosis:	3.886	**Cond. No.**	49.2

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Summaries OLS

```
R2 score 0.6941456750648598
MSE 0.0683585932392593
MAE 0.21307632554362313
```

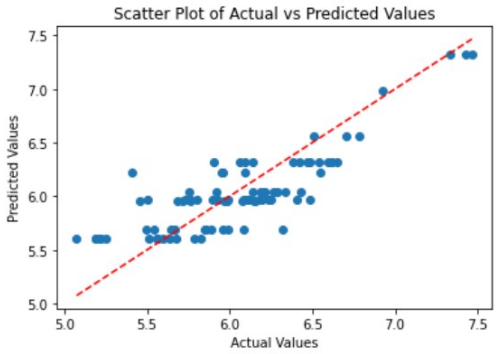


Graph for Linear Regression

B. *Ridge Regression*

Using plot line Ridge Regression.

```
R2 score 0.6911118921256407
MSE 0.06903664523006328
MAE 0.21623357186850206
```

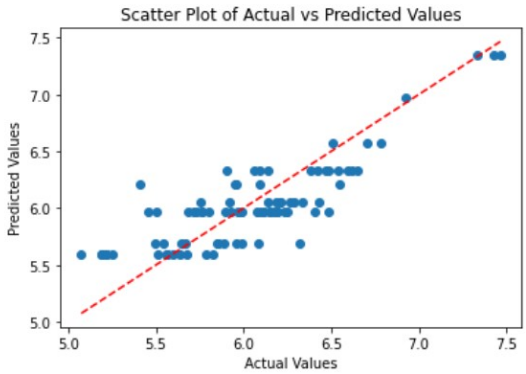


Graph for Ridge

C. *Lasso Regression*

Using Plot line Lasso Regression.

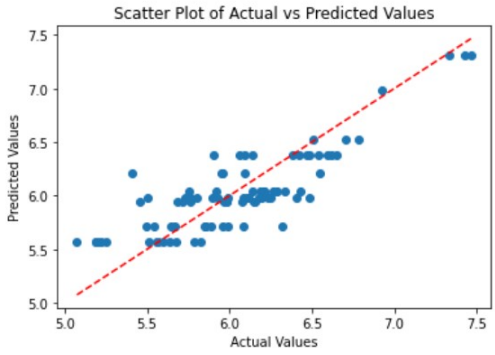
```
R2 score 0.6937996430889991
MSE 0.06843593155739705
MAE 0.2134947453414031
```



D. *Decision Tree*

Using Plot line Decision Tree and accuracy .

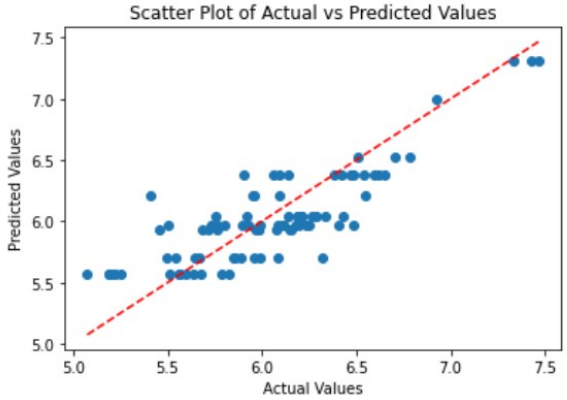
```
R2 score 0.7038093486507847
MSE 0.06619875740238729
MAE 0.2109959689709734
```



E. *Random Forest*

Using Plot line Random forest and accuracy .

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



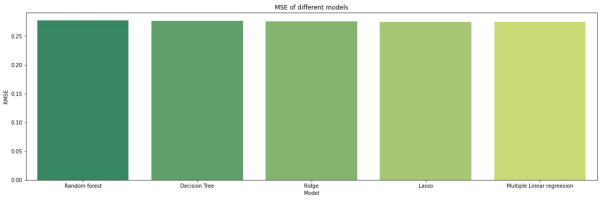
F. *Evaluate Model*

Evaluate all model.

```
Multiple Linear regreesion : 0.275
Ridge : 0.275
Lasso : 0.275
Decision Tree : 0.277
Random forest : 0.277
```

G. *Compare Model*

Compare all model.



VII. RESULT

By using all Model for Predict we see that all model it the same good and can predict and trust

for predicting price for PCcomponents. We need too use Multiple Linear Regression and Random Forest is the best model for predict in this project to be best. So, Model is the best all.

VIII. Conclusion

Based on the exploratory data analysis (EDA) and data analysis of CPU, RAM, and price, it can be concluded that: There is a positive correlation between CPU and price, indicating that higher CPU specifications generally result in greater pricing.There is no significant correlation between the number of cores and price. However, this may depend on the specific use case and processor architecture. The pricing of CPUs and RAM has been consistently decreasing over time due to competition in the market and advancements in technology. And also can help us to know about EDA and statistic. Model Can help us for predict CPU, RAM to be good.

IX. Reference

- https://www.kaggle.com
- Data Mining for Genomics and Proteomics: Analysis of Gene and Protein Expression Data" by Dimitrios Vlachakis and Andreas Zoumboulis, published by John Wiley Sons in 2020.