Institute of Technology of Cambodia

Department OF Applied Mathematics and Statistics

Introduction to Data Science

Topic: Laptop Price Prediction

Team Member Group1:

Member's Name:	ID
1. **Heng Sophanha**	e20210931
2. **HENG Seaklong**	e20210329
3. **CHHIN Visal**	e20210742
4. **ENG Sive eu**	e20210914
5. **HAYSAVIN RONGRAVIDWIN**	e20211502

Professor: Dr. Phank Sokkhey(Course)
Dr. Pen Chetra(TP)

2022-2023

Contents

1. Introduction

We design this project for people who are looking for purchasing laptops. In this project, we will build a Laptop price prediction project and learn about the machine learning project lifecycle.it becomes difficult for laptop makers to sell their products and for customers to pick their laptops easily as because laptops have become one of the most essential and used items in our daily life. However, I am sure that you have struggled to choose the personal laptop that fits your needs? With the overwhelming amount of specifications and brand names on the market, it becomes difficult for laptop makers to sell their products and for customers to pick their laptops. Laptops will continue to rise in demand as the growing spending capacities of consumers and an upsurge in demand for technologically advanced products. As laptops are selling across the world, people and manufacturers alike will have to be knowledgeable and competitive that's why we need to make this project.

1.1. Brief overview of the dataset

- Company: The dataset includes computers manufactured by various companies, such as Dell, HP, Lenovo, Apple, Acer, Asus, and custom/local builders.
- TypeName: This column represents the name or type of computer, including Ultrabook, Notebook, Netbook, Gaming, 2-in-1 Convertible, and Workstation.
- Inches: This column denotes the screen size of the computer, measured in inches. It includes various sizes like 13.3, 15.6, 15.4, 14., 12., 11.6, 17.3, 10.1, 13.5, 12.5, 13., 18.4, 13.4, 13.9, 12.3, 17., 15., 14.1, and 11.3.
- Screen Resolution: This column describes the screen resolution or clarity of the computer, including different types such as IPS Panel Retina Display 2880x1800, 1366x768, IPS Panel Full HD 1920x1080, IPS Panel Retina Display 2304x1440, IPS Panel Full HD/Touchscreen 1920x1080, Full HD/Touchscreen 1920x1080, touchscreen/Quad HD+ 3200x1800, IPS Panel Touchscreen 1920x1200, touchscreen 2256x1504, etc.
- CPU: This column represents the Central Processing Unit (CPU) of each computer. It includes various CPU models from different manufacturers, such as Intel Core i5, Intel Core i7, AMD A9_series, and their respective clock speeds.
- RAM: This column denotes the Random Access Memory (RAM) of the computers, which is used to store data and machine code being processed by the CPU. It includes different RAM sizes like 2GB, 4GB, 6GB, 8GB, 12GB, 16GB, 24GB, 32GB, and 64GB.
- Memory: This column represents the storage capacity of the computers, including different types such as Flash Storage, SSD (Solid State Drive), and HDD (Hard Disk Drive), along with their respective capacities.
- OpSys: This column indicates the operating system installed on the computers, including macOS, Windows 10, Mac OS X, Linux, Android, Windows 10 S, Chrome OS, and Windows 7.
- Weight: This column specifies the weight of the computers, measured in kilograms (Kg).

- Price: This column represents the price of the computers. The prices can vary based on the type of computer, brand, specifications, and whether it's a pre-built system or custom-built. The dataset provides some general price ranges for different types of computers, including desktops, laptops, all-in-ones, gaming computers, ultrabooks/premium laptops, and custom-built PCs.

2. Understanding the dataset

- **Details of the particular col company** : Computers are manufactured by various companies, and the brand of a computer can vary based on the manufacturer. Some well-known companies that produce computers include:
- **Dell**: Dell is a multinational computer technology company that manufactures a wide range of computers, including desktops, laptops, and workstations.
- **HP (Hewlett-Packard):** HP is another major player in the computer industry, producing a variety of computers, printers, and other hardware.
- **Lenovo**: Lenovo is a Chinese multinational technology company that manufactures computers, laptops, tablets, and other electronic devices.
- **Apple**: Apple is known for its line of Macintosh (Mac) computers, including desktops (iMac) and laptops (MacBook).
- **Acer**: Acer is a Taiwanese multinational hardware and electronics corporation that produces a variety of computer products, including laptops and desktops.
- **Asus**: Asus is a Taiwanese multinational computer hardware and electronics company known for its laptops, desktops, and other components.
- **Custom/Local Builders**: Many computers are also assembled by local or custom builders using components from various manufacturers. These systems are often custom-built to meet specific user requirements.
- **Details of the particular col TypeName** : it mean name of utilization function of computer there are : Ultracbook, Notbook, Netbook, Gaming, 2 in 1 Convertible, Workstation.
- **Details of the particular col Inches** : The term "inches of computer" is a bit ambiguous, and it could refer to different aspects depending on the context. Here are a few possibilities:
- **Screen Size**: If you're asking about the size of a computer in inches, it might refer to the screen size of a monitor or laptop. For example, you might see laptops or monitors advertised with a 13-inch, 15-inch, 24-inch, etc., screen size.
- **Form Factor:** It could also refer to the physical size or form factor of a computer. For desktop computers, the term might be used to describe the dimensions of the case or the overall size of the system.
- **Storage Size**: In some cases, the term might be used to describe the storage capacity of a computer's hard drive or solid-state drive. For instance, you might see a computer with a 500GB or 1TB storage capacity.he screen size of computer there are : 13.3, 15.6, 15.4, 14., 12., 11.6, 17.3, 10.1, 13.5, 12.5, 13., 18.4, 13.4, 13.9, 12.3, 17., 15., 14.1, 11.3.

- **Details of the particular col Screen Resolution** : the computer clarity there are:
  - IPS Panel Retina Display 2880x1800, 1366x768
  - IPS Panel Full HD 1920x1080, IPS Panel Retina Display 2304x1440
  - IPS Panel Full HD / Touchscreen 1920x1080
  - Full HD / Touchscreen 1920x1080, touchscreen / quad HD+ 3200x1800
  - IPS Panel Touchscreen 1920x1200, touchscreen 2256x1504
  - **Details of the particular col CPU:** the CPU in each computer. CPU stands for Center Processing Unit. It is oftern referred to as the brain of a computer because it is responsible for carrying out instuctions of a computer [rogram by performing basic arithmetic, logical, control, and input/ output operation. The CPU interprets and executes instuctions from the computer's memory, allowing it to perform tasks ranging from simple calculations to complex operations required by software application. there are
  - Intel core i5 2.3Ghz, intel core i5 1.8Ghz
  - Intel core i5 700U 2.5Ghz, Intel core i7 2.7Ghz
  - Intel core i5 3.1Ghz, AMD A9_series 9420 #Ghz, Intel core i7 2.2Ghz
  - Intel core i7 8550U 1.8Ghz, Intel core i5 8250U 1.6Ghz
  - Intel core i3 6006U 2Ghz, Intel Core i7 2.8Ghz
  - **Details of the particular col RAM :** RAM is stands for Random Access Memory. It is a type of computer that is used to store data and machine code currently being used and processed by a computer's Central processing unit (CPU). There are *GB, 16GB, 32GB, 64GB, 2GB, 4GB, 12GB, 6GB, 24GB.
  - **Details of the particular col Memory :** Memory it has in the context of computes generally refers to electronic components that store and retrieve data for to the purpose of computer processing. There are
  - 128GB Flash Storage, 256GB SSD, 512GB SSD, 500GB HDD
  - 256GB Flash Storage, 1TB HDD, 32GB Flash Storage
  - 128GB SSD + 1TB HDD, 256GB SSD + 256GB SSD, 64GB Flash Storage
  - 256GB SSD + 1TB HDD, 256GB SSD+ 2TB HDD, 32GB SSD, 2TB HDD
  - 64GB SSD, 1TB HYBRID, 512GB SSD+ 1TB HDD, 1TB SSD
  - **Details of the particular col OpSys :** OpSys is a common Abbreviation for "Operating System". The operating system is a crucial software computer that manages computer hardware and provides services for computer programs. There are: macOS, No OS, windows 10, mac OS X, Linux, Android, windows 10 S, Chrome OS, Windows 7.
  - **Details of the particular col Weight :** the weight of a computer can vary significantly based on the type and form factor of the device. There are : 1.37Kg, 1.34Kg, 1.86Kg,1.83Kg,2.9Kg, 3.3Kg, 4.4 Kg , etc.
  - **Details of the particular col Price of computer:** The price of a computer can vary widely depending on several factors, including the type of computer, brand, specifications, and whether it's a pre-built system or a custom-built one. Here are some general price ranges for different types of computers as of my last knowledge update in January 2022:

  - **Desktop Computers**:

- Basic desktops: $300 to $800
- Mid-range desktops: $800 to $1,500
- High-performance desktops: $1,500 and above

- **Laptop Computers**:
- Budget laptops: $200 to $600
- Mid-range laptops: $600 to $1,200
- High-performance laptops: $1,200 and above

- **All-in-One Computers:**
- Entry-level all-in-ones: $500 to $800
- Mid-range all-in-ones: $800 to $1,500
- Premium all-in-ones: $1,500 and above

- **Gaming Computers**:
- Entry-level gaming PCs: $700 to $1,200
- Mid-range gaming PCs: $1,200 to $2,000
- High-end gaming PCs: $2,000 and above

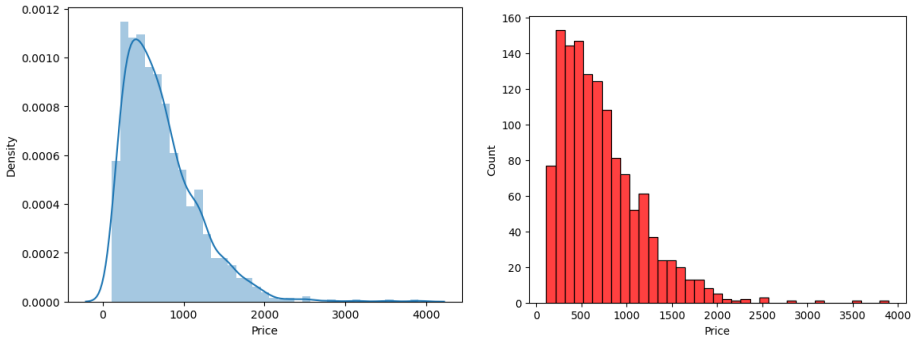
- **Ultrabooks and Premium Laptops:**
- Ultrabooks and premium laptops: $800 to $2,000 and higher

- **Custom-Built PCs:**
- Custom-built PCs can vary widely in price, depending on the components chosen. Prices can range from a few hundred dollars to several thousand dollars for high-end configurations.

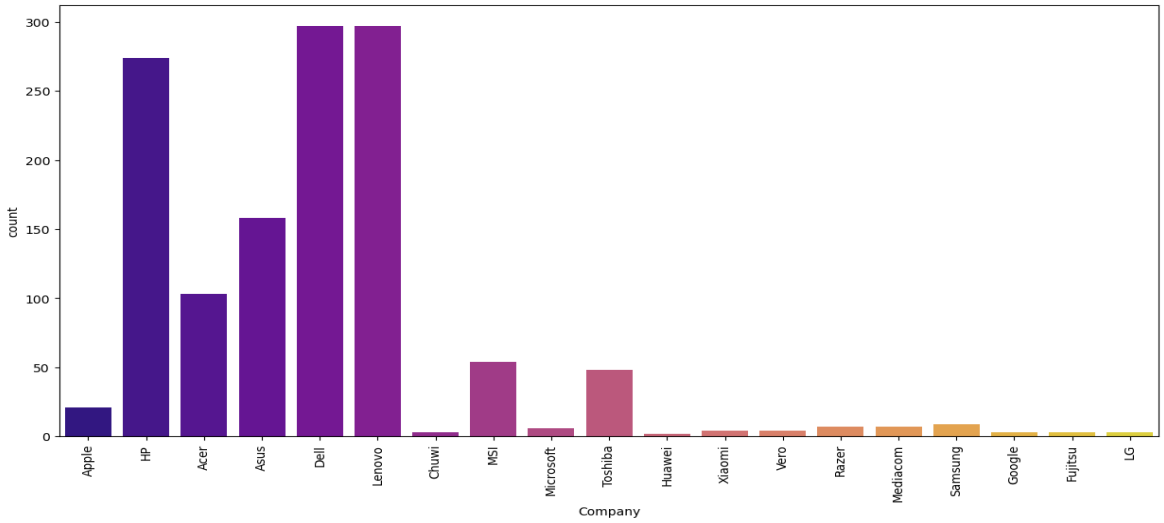
3. Preprocessing the Data

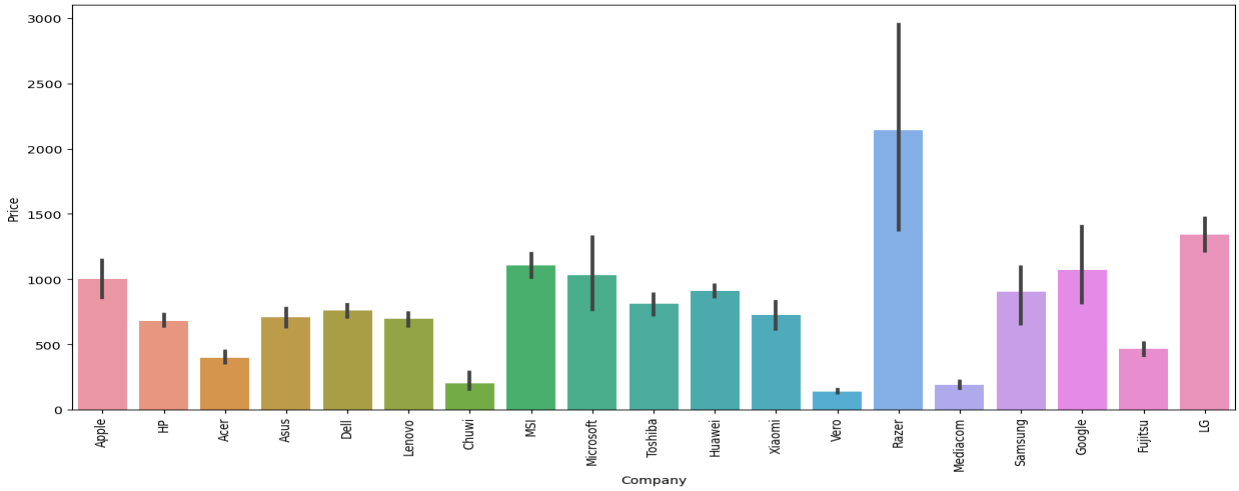
4. Data Visualization

Data visualization is the graphical representation of information and data. Use Visual elements such as: draw plot, hist plot, bar plot, scatter plot, dist plot and heatmap, and data visualization tools provide an easily accessible way to identify and understand trends in the data. Here are some distribution graphs of each variable. But the graph below shows the income distribution and us puts this income in relation to the normal distribution. It is desirable to have the income variable we want forecast according to a normally distributed. And why many statistical methods and model assume that the data analyzed is normally distributed.

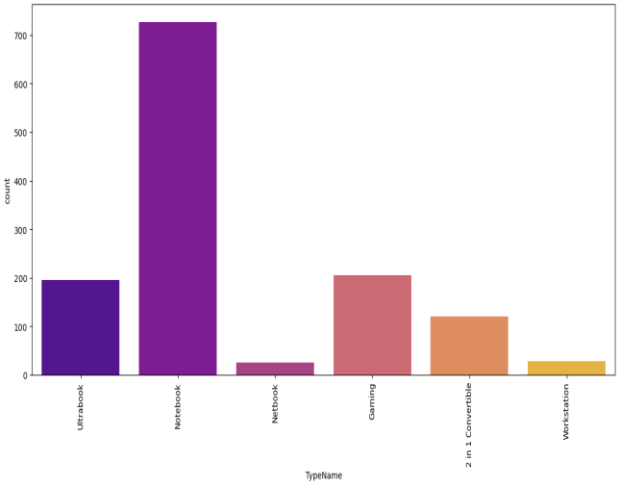


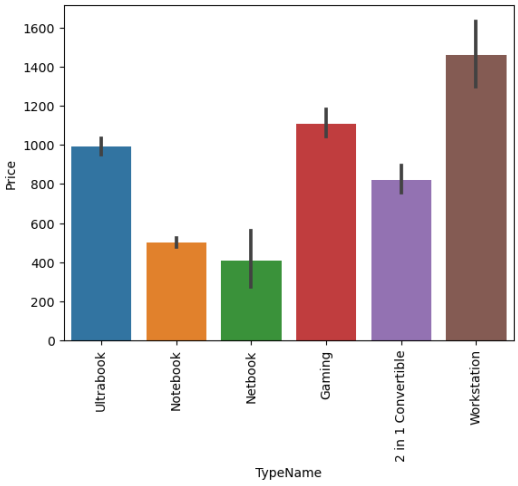
- After plorting the hisplot, we can see the top of the price of computer is 4000$ and the most of price is 300-400.



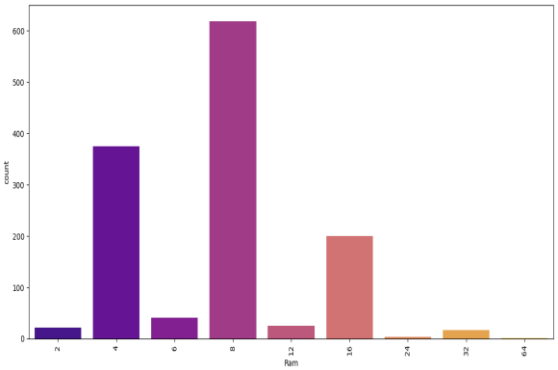


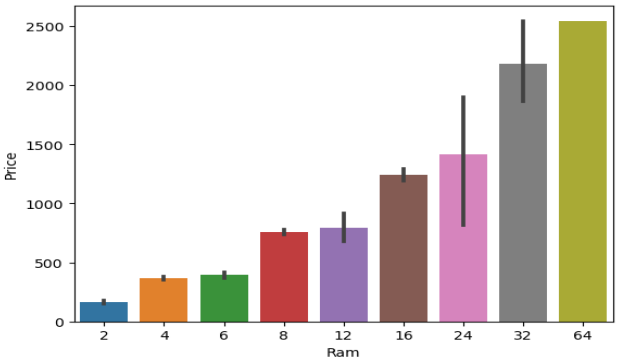
- After plotting the draw plot and bar plot is show which computer company that have many types of computers like: computer for gaming, computer that have touch screen...and which brand of computer was expensive.



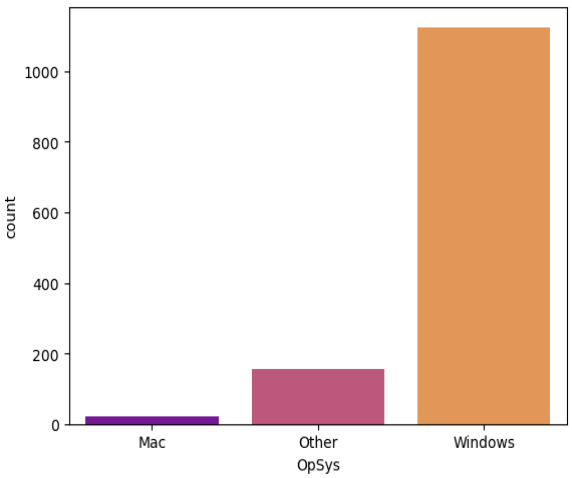


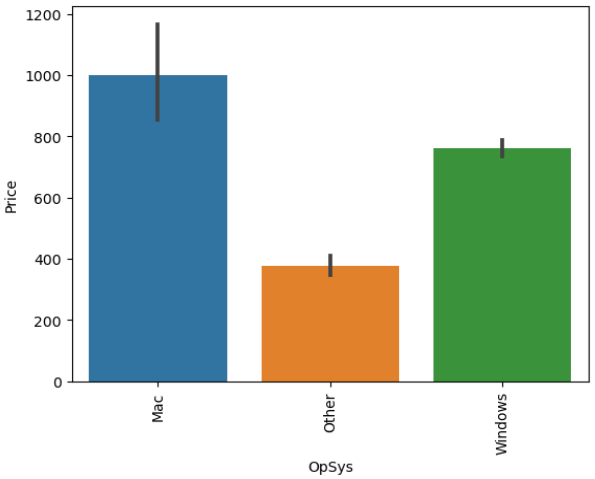
- This is showing all Type of computers that have in all brand of computer which one that was popular for user and which type of computer was expensive for user.

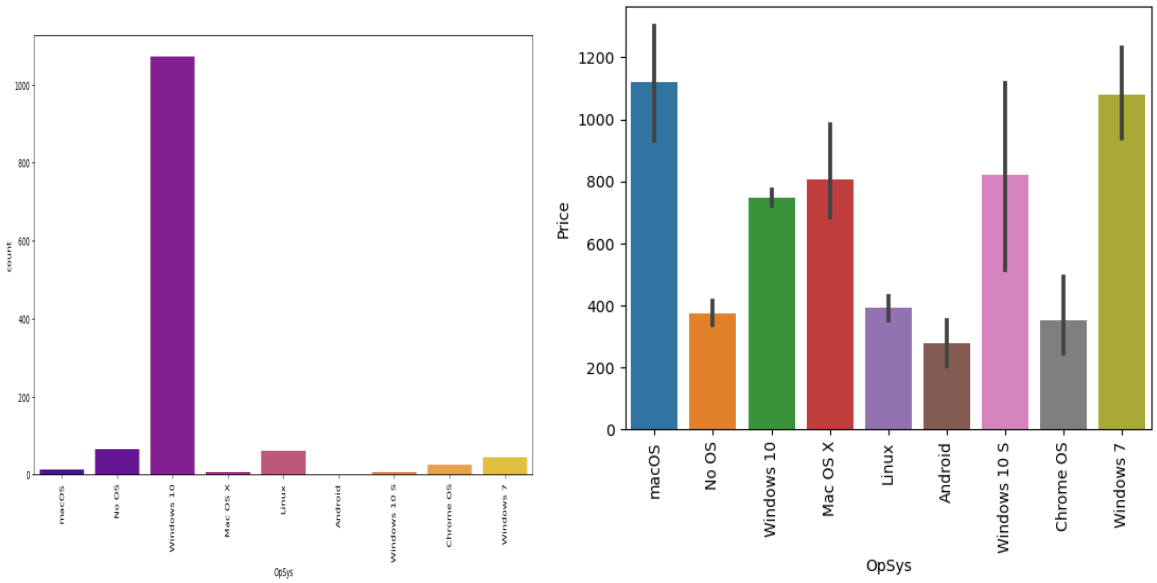




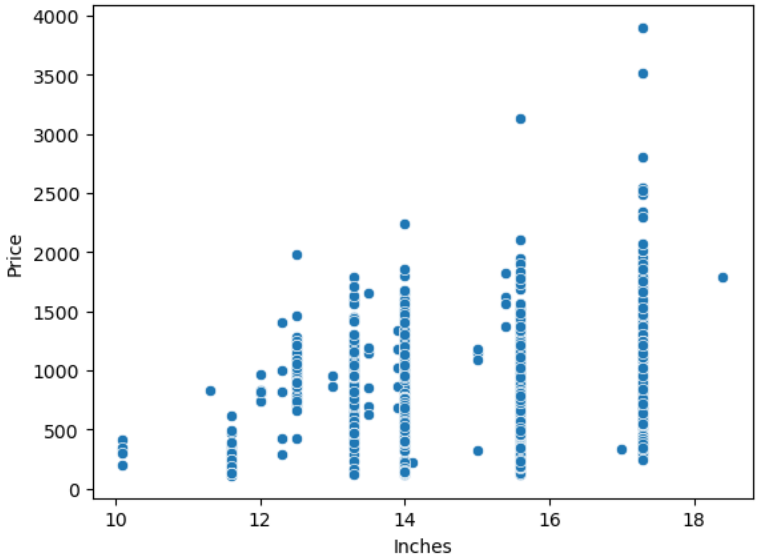
- This is showing all RAM of computers that have in all brand of computer which one that was popular for user and which type RAM was expensive for user.



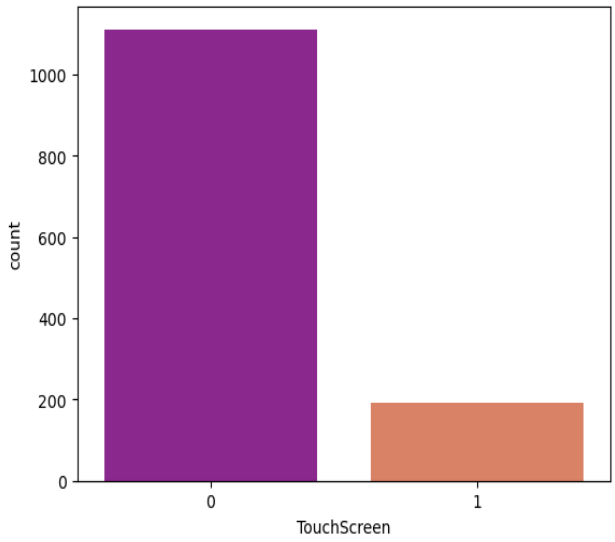


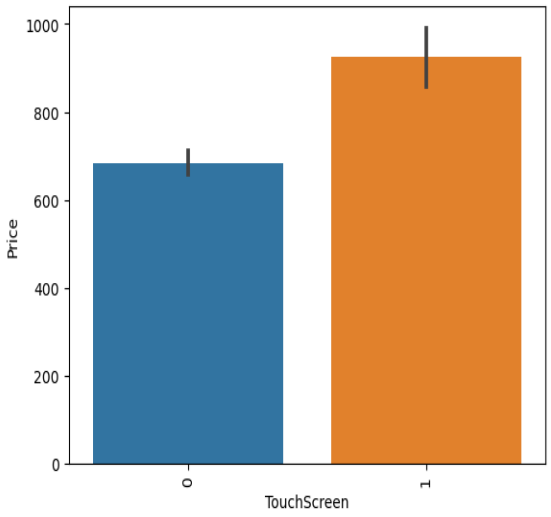


- This is showing all of Operated System that most popular, which one that was more expensive for user.

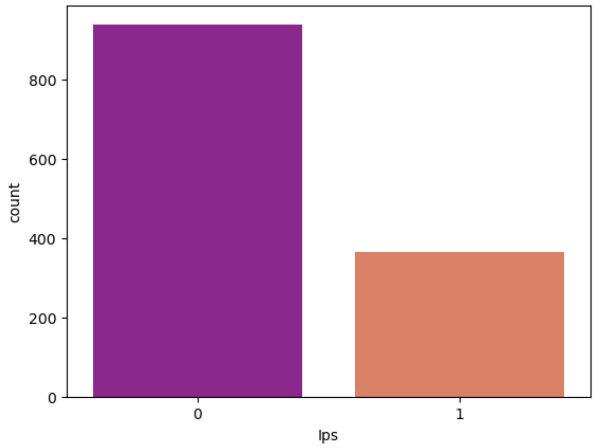


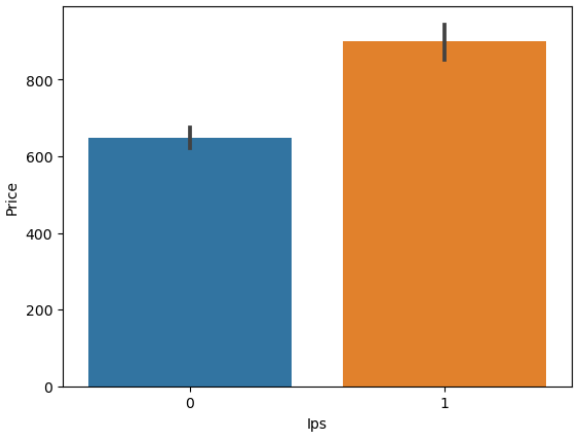
- After plotting the scatter plot, it was show price of size screen computer who it was expensive and which one that was popular for user.



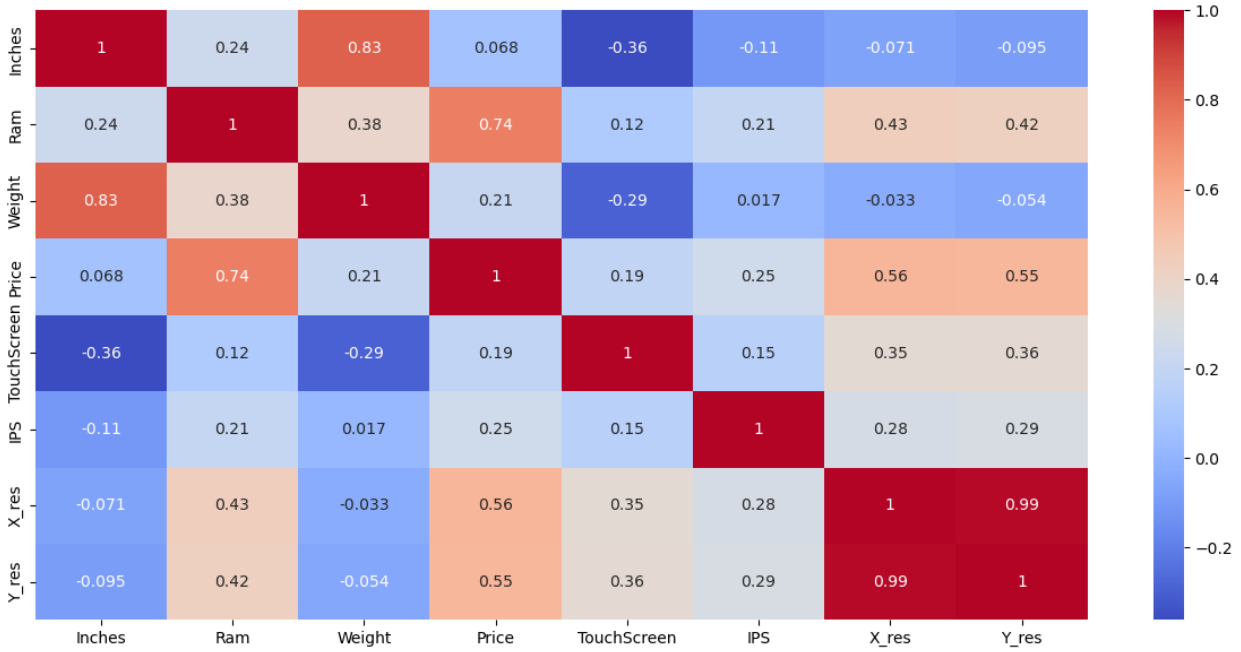


- This is bar plot for show that have many of computer that can use touchscreen and which one was more expensive. 0 it's mean computer can't use touchscreen and 1 that can use touchscreen.

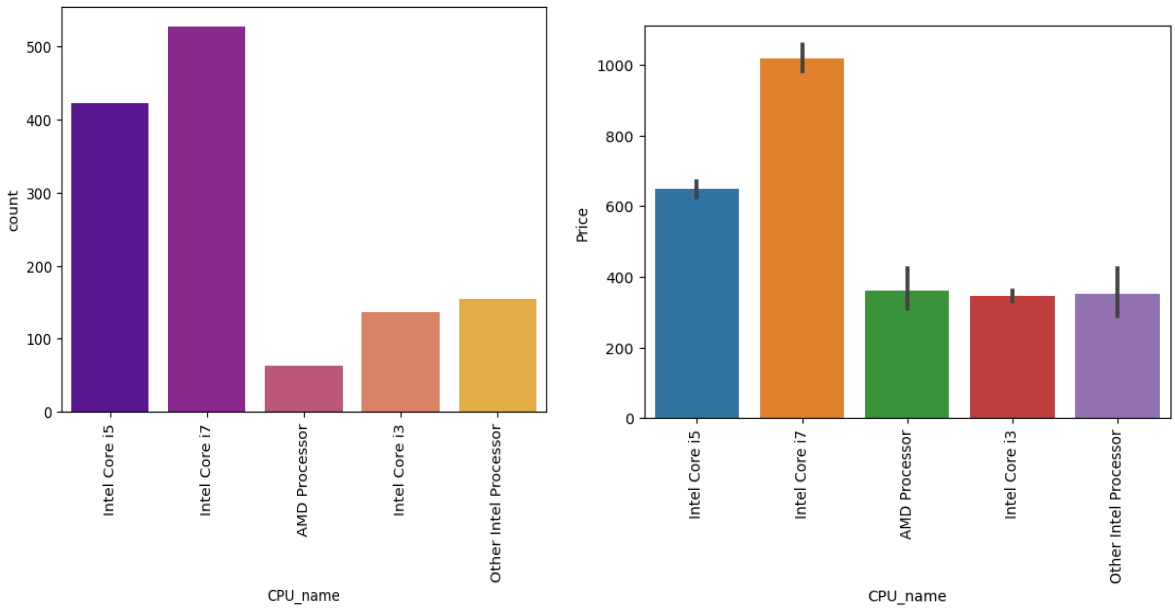




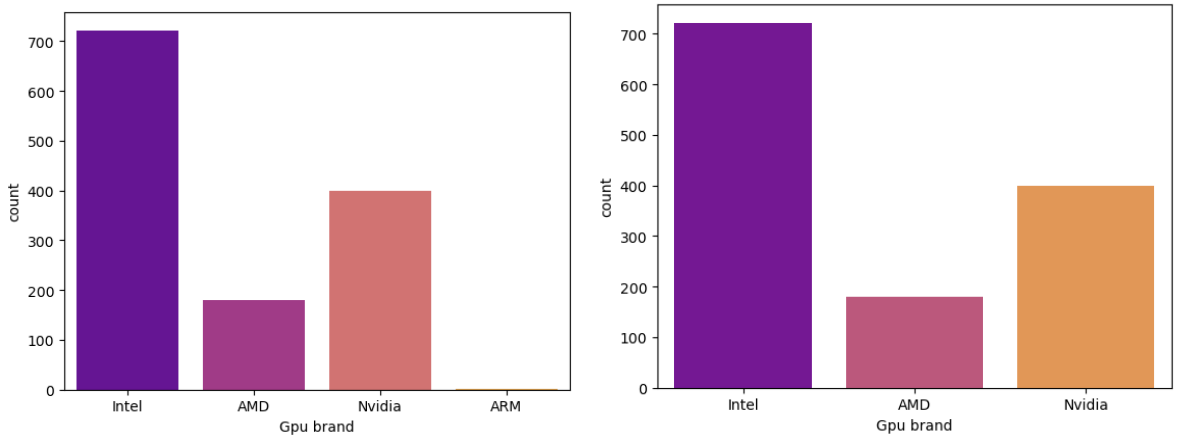
- This is bar plot for show that have many of computer that can use Screen Resolution and which one was more expensive. 0 it's mean computer can't use Screen Resolution and 1 that can use touchscreen.



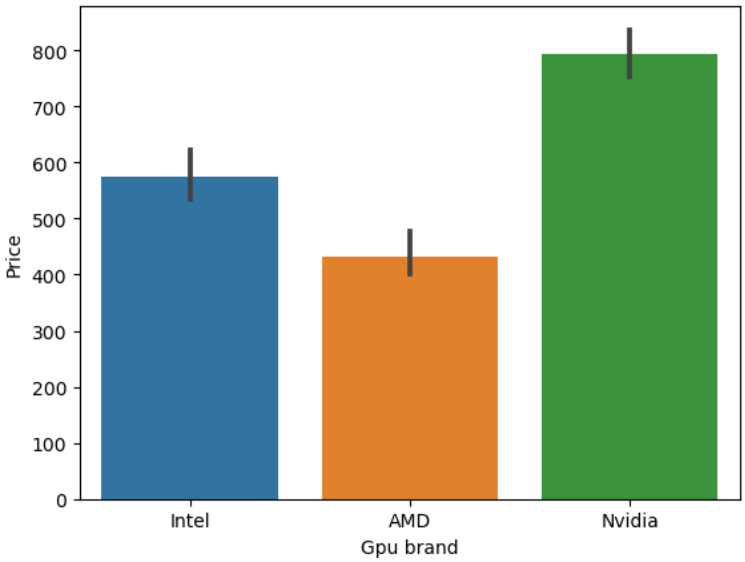
- The correlation matrix can be a useful tool for understanding the relationships between different laptop features. However, it is important to interpret the matrix carefully and to be aware that correlation does not equal causation.



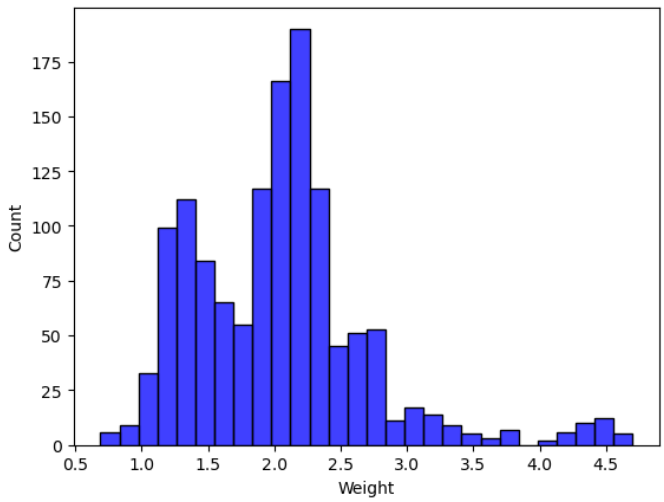
- This is showing all CPU type of computers that have in all brand of computer which one that was popular for user and which type of CPU was expensive for user.



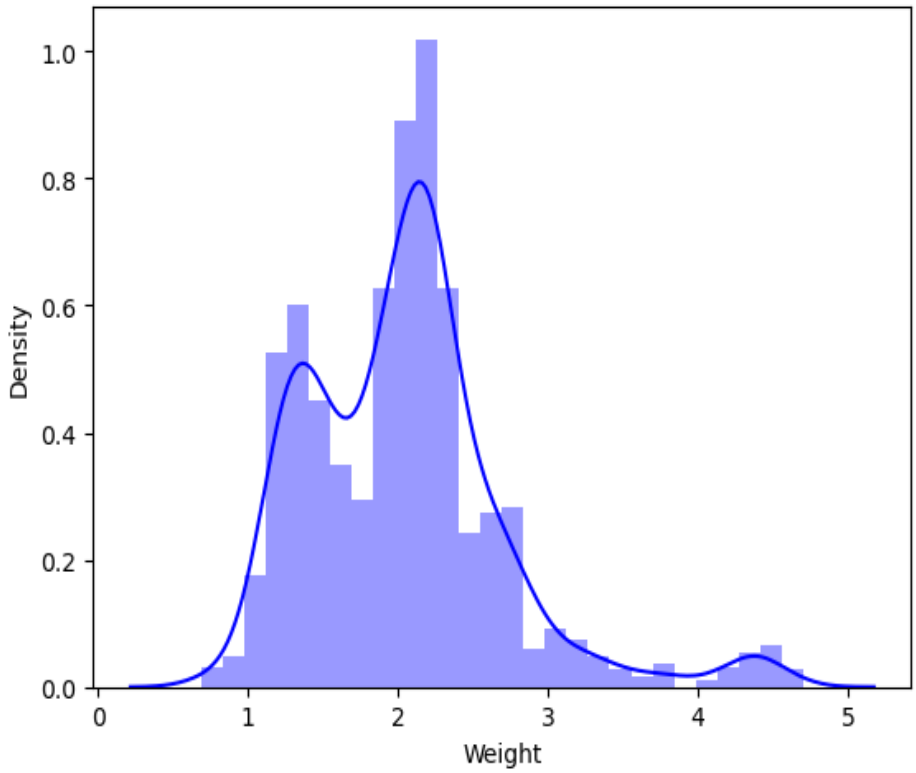
- This graph it was show all of GPU brand that the most popular for all computer, and after I was dropped one type is ARM because it is not popular to use until now, we also never hear about this brand in computer.



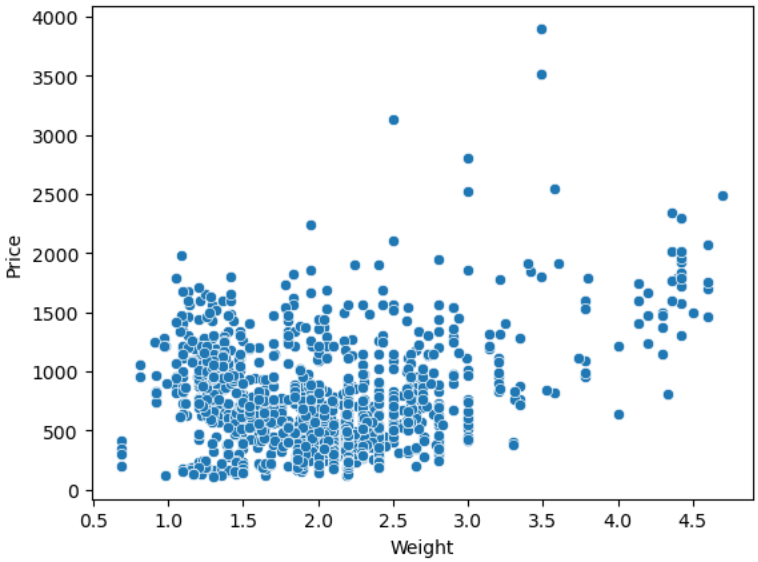
- This graph it was show which one of GPU brand who that most expensive in all computer.



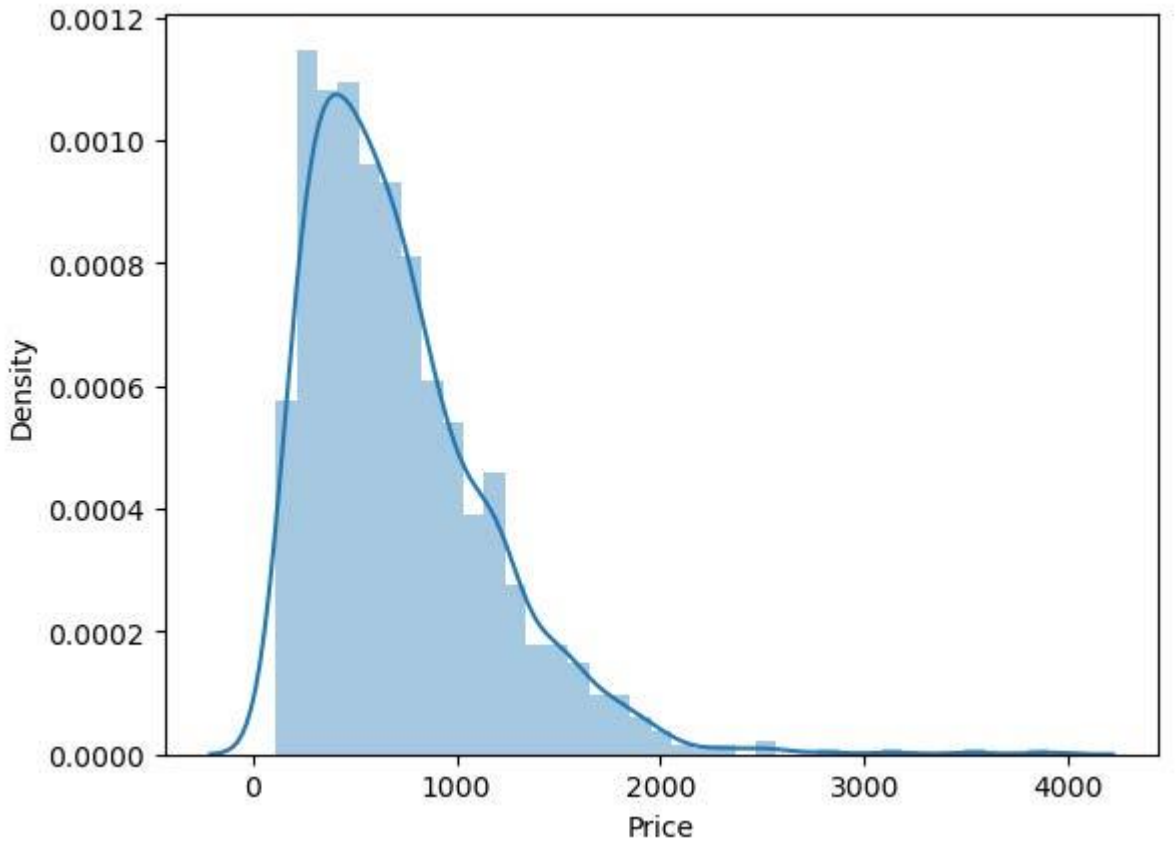
- After histplot, this graph was showing the weight of computer that have the most popular use in all of computer company.



- After distplot, the graph shows that the density of the population is highest at a weight of around 2, and then decreases gradually as weight increases or decreases.



- This graph is scatter plot, the scatter plot was show which one of weigh was had the most expensive for user and show which weight was popular that we use nowadays.



- The graph you sent shows the distribution of the price of a certain item. The x-axis represents the price, and the y-axis represents the density, which is the number of items at that price.

- This means that there is no strong relationship between these features. For example, a computer with a lot of RAMS may or may not have a touchscreen, and a computer with a high price may or may not have a high PPI. Overall, the correlation matrix provides a useful overview of the relationships between the different features of a

computer. This information can be used to make informed decisions about which features are important to consider when choosing a computer.

5. Building Model

Model building, in the context of statistical modeling and machine learning, refers to the process of creating a mathematical representation or algorithm that captures the relationship between input variables (features) and the target variable (response) in a dataset. The goal of model building is to develop a predictive or descriptive tool that can be used to make informed decisions, gain insights, or make predictions on new, unseen data.

We have 5 model:

1 / Linear Regression

2 / Ridge Regression

3 / Lasso Regression

4 / Decision Tree

5 / Random Forest

But the best model is Random Forest because R2 score = 0.8839667676972109

This import library in sklearn

In [ ]:
```
test = np.log(df['Price'])
train = df.drop(['Price'],axis = 1)
```

In [ ]:
```
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import MinMaxScaler,StandardScaler
from sklearn.pipeline import Pipeline
from sklearn.compose import ColumnTransformer
from sklearn.preprocessing import LabelEncoder,OneHotEncoder
from sklearn import metrics
from sklearn.model_selection import RandomizedSearchCV
from sklearn.linear_model import LinearRegression,Lasso,Ridge
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestRegressor,GradientBoostingRegressor
from xgboost import XGBRegressor
from sklearn.svm import SVR
from sklearn.neighbors import KNeighborsRegressor
from sklearn import tree
```

Linear Regression

```
# we will apply one hot encoding on the columns with this indices-->[0,1,3,8,11]
# the remainder we keep as passthrough i.e no other col must get effected
# except the ones undergoing the transformation!

step1 = ColumnTransformer(transformers=[
    ('col_tnf',OneHotEncoder(sparse=False,drop='first'),[0,1,3,8,11])
],remainder='passthrough')

step2 = LinearRegression()

pipe = Pipeline([
    ('step1',step1),
    ('step2',step2)
])

pipe.fit(X_train,y_train)

y_pred = pipe.predict(X_test)

print('R2 score',metrics.r2_score(y_test,y_pred))
print('MAE',metrics.mean_absolute_error(y_test,y_pred))
```

```
R2 score 0.8073277669771588
MAE 0.21017828458766966
```

Ridge Regression

```
# we will apply one hot encoding on the columns with this indices-->[0,1,3,8,11]
# the remainder we keep as passthrough i.e no other col must get effected
# except the ones undergoing the transformation!

step1 = ColumnTransformer(transformers=[
    ('col_tnf',OneHotEncoder(sparse=False,drop='first'),[0,1,3,8,11])
],remainder='passthrough')

step2 = Ridge(alpha=10)

pipe = Pipeline([
    ('step1',step1),
    ('step2',step2)
])

pipe.fit(X_train,y_train)

y_pred = pipe.predict(X_test)

print('R2 score',metrics.r2_score(y_test,y_pred))
print('MAE',metrics.mean_absolute_error(y_test,y_pred))
```

```
R2 score 0.8127331198140741
MAE 0.20926803398249494
```

LassoRegression

```
# we will apply one hot encoding on the columns with this indices-->[0,1,3,8,11]
# the remainder we keep as passthrough i.e no other col must get effected
# except the ones undergoing the transformation!

step1 = ColumnTransformer(transformers=[
    ('col_tnf',OneHotEncoder(sparse=False,drop='first'),[0,1,3,8,11])
],remainder='passthrough')

step2 = Lasso(alpha=0.001)

pipe = Pipeline([
    ('step1',step1),
    ('step2',step2)
])

pipe.fit(X_train,y_train)

y_pred = pipe.predict(X_test)

print('R2 score',metrics.r2_score(y_test,y_pred))
print('MAE',metrics.mean_absolute_error(y_test,y_pred))
```

```
R2 score 0.8071857421978025
MAE 0.21114351839770898
```

Decision Tree

In [ ]:

```
# we will apply one hot encoding on the columns with this indices-->[0,1,3,8,11]
# the remainder we keep as passthrough i.e no other col must get effected
# except the ones undergoing the transformation!

step1 = ColumnTransformer(transformers=[
    ('col_tnf',OneHotEncoder(sparse=False,drop='first'),[0,1,3,8,11])
],remainder='passthrough')

step2 = DecisionTreeRegressor(max_depth=8)

pipe = Pipeline([
    ('step1',step1),
    ('step2',step2)
])

pipe.fit(X_train,y_train)

y_pred = pipe.predict(X_test)

print('R2 score',metrics.r2_score(y_test,y_pred))
print('MAE',metrics.mean_absolute_error(y_test,y_pred))
```

```
R2 score 0.8386774048320025
MAE 0.18348876732269775
```

Random Forest

In [ ]:

```
step1 = ColumnTransformer(transformers=[
    ('col_tnf',OneHotEncoder(sparse=False,drop='first'),[0,1,3,8,11])
],remainder='passthrough')

step2 = RandomForestRegressor(n_estimators=100,
                              random_state=3,
                              max_samples=0.5,
                              max_features=0.75,
                              max_depth=15)

pipe = Pipeline([
    ('step1',step1),
    ('step2',step2)
])

pipe.fit(X_train,y_train)

y_pred = pipe.predict(X_test)

print('R2 score',metrics.r2_score(y_test,y_pred))
print('MAE',metrics.mean_absolute_error(y_test,y_pred))
```

```
/usr/local/lib/python3.10/dist-packages/sklearn/preprocessing/_encoders.py:868: FutureWarning: `sparse` was renamed to `sparse
_output` in version 1.2 and will be removed in 1.4. `sparse_output` is ignored unless you leave `sparse` to its default value.
  warnings.warn(
R2 score 0.8839667676972109
MAE 0.1604176841506877
```

Now we see that the best model building is Random forest because we have R2 score = 0.88 and MAE = 0.16

6. Web application

7. Conclusion

Laptop price prediction is a use full tool for consumers and businesses, aiding in informed decision-making. It helps consumers plan budgets, compare prices, and time their purchases. For businesses, it assists in inventory management, pricing strategies, and marketing efforts. While not infallible, continuous data collection and model refinement improve prediction accuracy. Overall, laptop price prediction provides valuable insights for navigating the market effectively. For this project, the model that we are going to use for our laptop's price prediction is Support Vector Regressor. Because this model overall shown the perfect accuracy and its prediction is close to the actual values.