NAME: FOLASEWA MARYAM ABDULSALAM
ID: B02294068
Link to my Github account: https://github.com/Folasewa/DS_Class_Project
MINI PROJECT DATA SCIENCE AND ADVANCED CONCEPTS

PART 1 - NUMPY, MATPLOTLIB/SEABORN

Q1 - CONVERT COVARIANCE MATRIX INTO CORRELATION MATRIX USING NUMPY

A. Briefly explain in 1-2 sentences the concepts of Covariance and Correlation, and describe their relationship. You may use equations for clarification.

Solution

Covariance and Correlation are two key statistical measures that help us analyze the relationship between two variables. **Covariance (Covri(x,y))** indicates the direction of the linear relationship between two variables by assessing how much they change together from their mean values. In comparison, **Correlation (Corr(x,y))** is a scaled version of covariance that provides a standardized measure of the strength and direction of the linear relationship between two variables. Its values lie between -1 and 1.

The relationship between them using equations where:

x_i and y_i represent individual sample sets

x' and y' represent the mean of a given sample set, and

n represents the total number of sample

$$Covri(x,y) = \frac{\sum_{i=1}^{n}(x_i - x')\,(y_i - y')}{n}$$

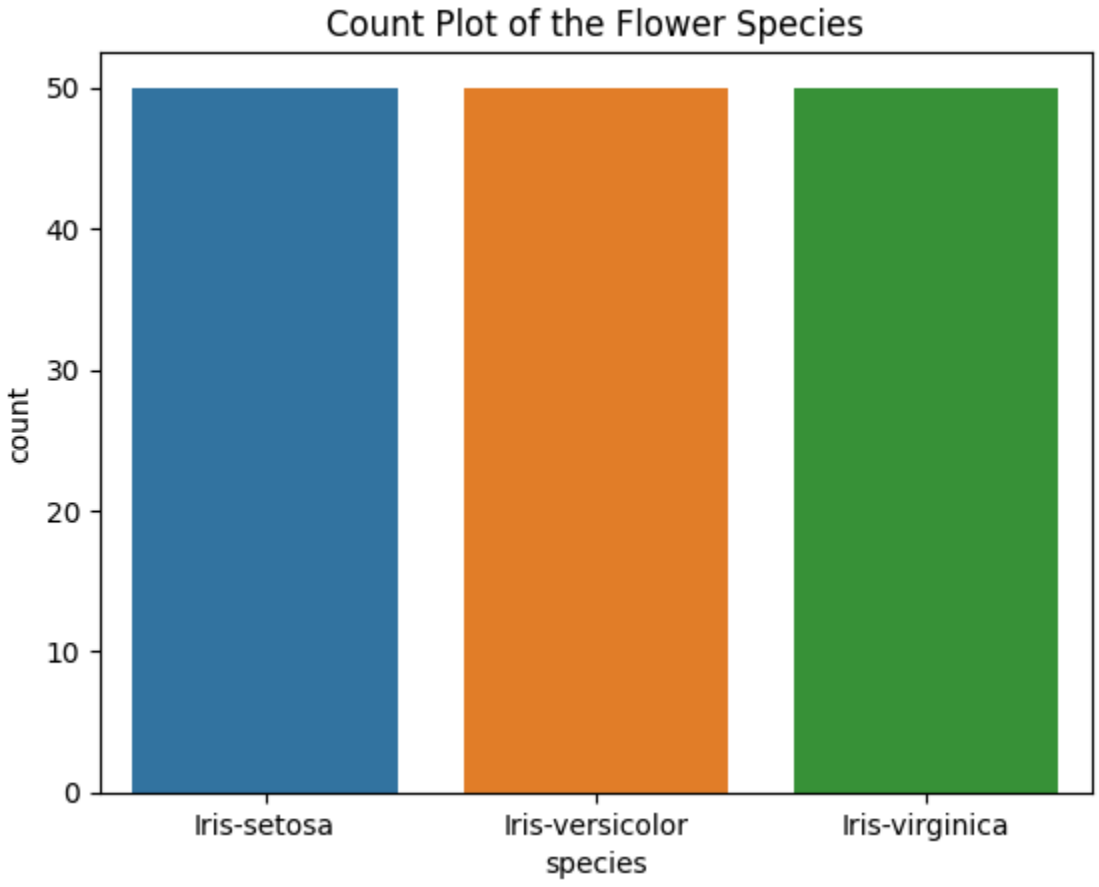
$$Corr(x, y) = \frac{\sum_{i=1}^{n}(x_i - x')\,(y_i - y')}{\sqrt{\sum_{i=1}^{n}(x_i - x')^2 \sum_{i=1}^{n}(y_i - y')^2}}$$

B. Load the Iris dataset, using any method of your choice (e.g.: pd.read_csv("iris.csv") in Pandas). Use visualization to explore the relationship between the different features.

Solution

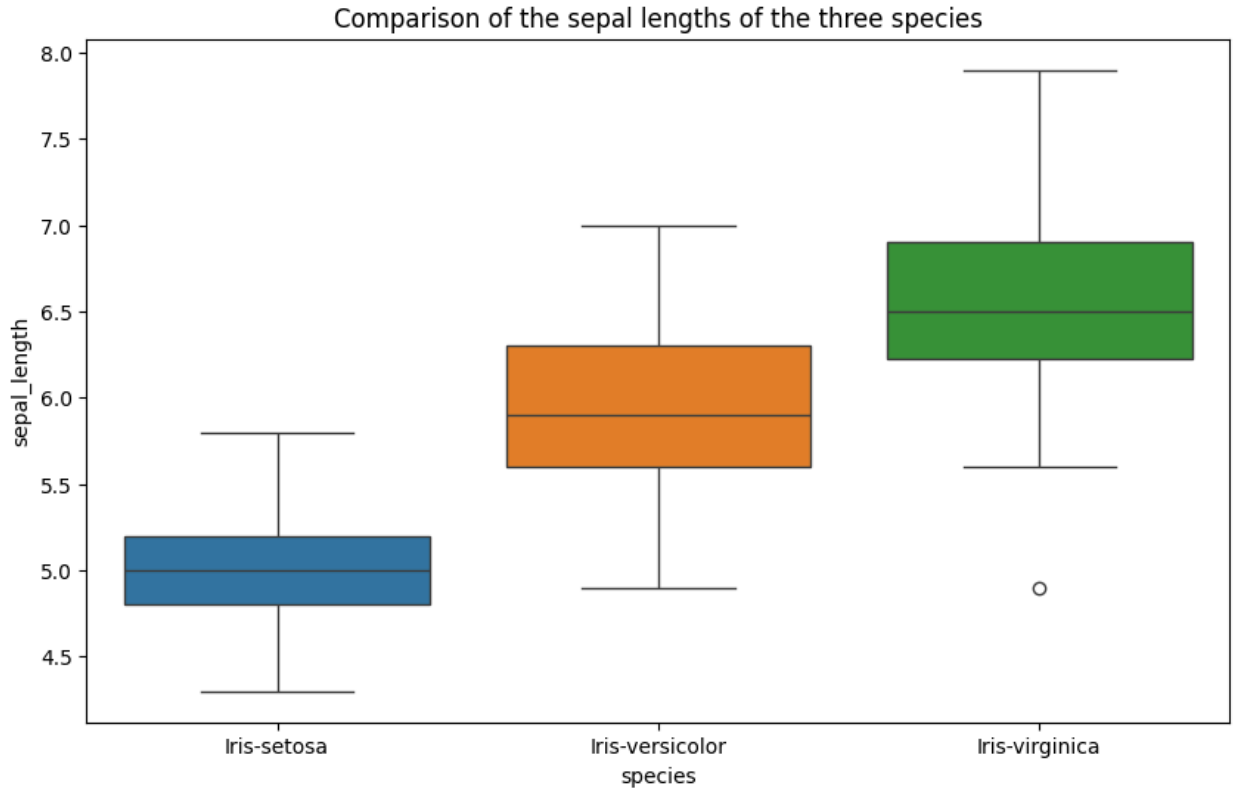
Visualization 1

This plot shows the total individual count of the flower species. It shows an equal distribution (50 each).



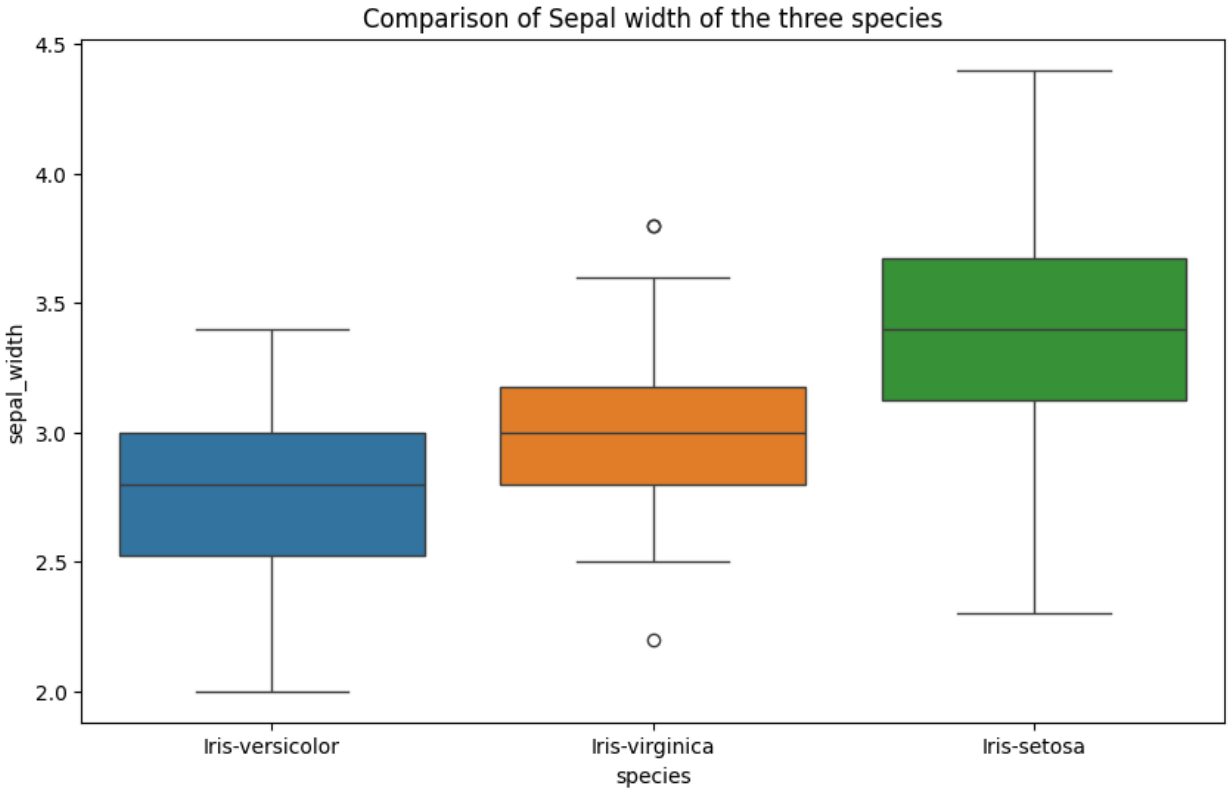
Visualization 2

Shows a boxplot distribution comparing the sepal lengths of the three species. Specie “Iris-Setosa” has the smallest sepal length range compared to the other species. The median sepal length of Iris-Setosa is around 5 cm, followed by Iris-Versicolor which has a median sepal length of around 6 cm, and Iris-Virginica having the largest sepal length range and median sepal length of around 6.6 cm. In the visualization, there is an outlier in the Iris-Virginica sepal length distribution.



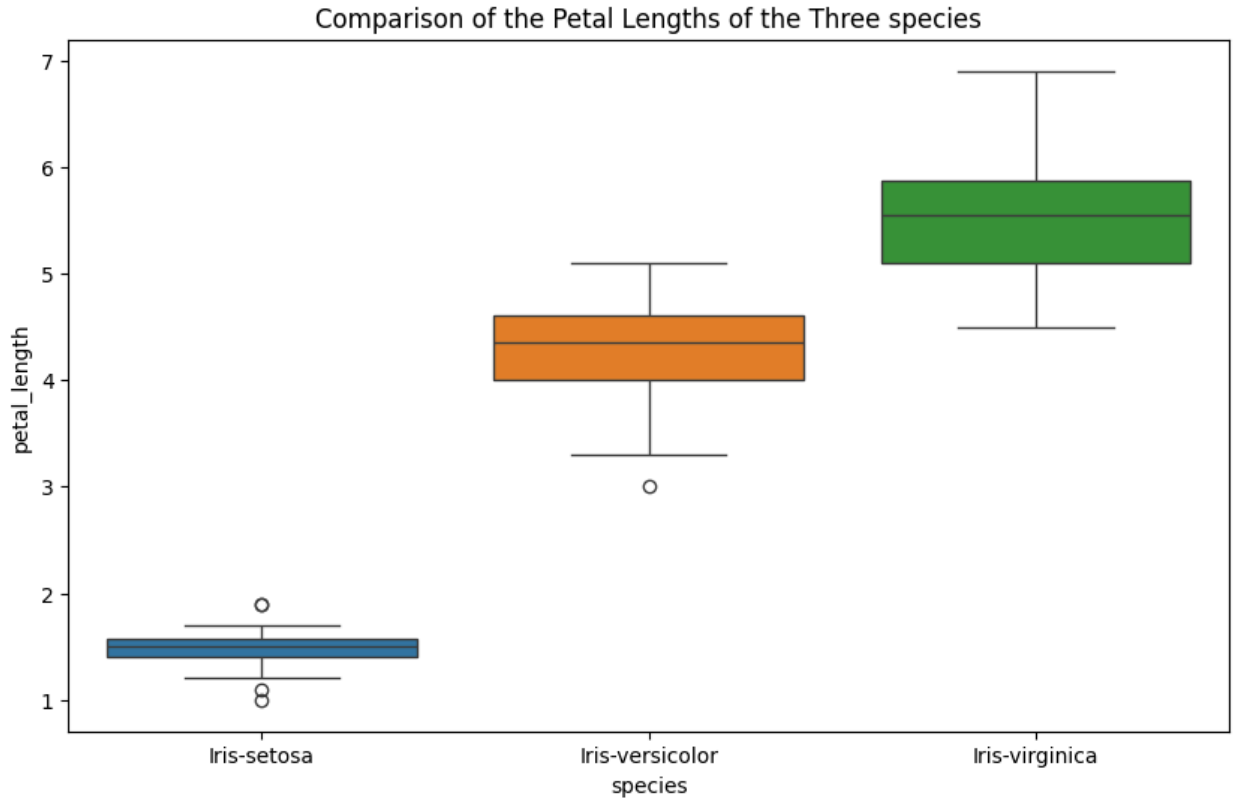
Visualization 3

Shows a boxplot distribution comparing the sepal widths of the three species. Iris-versicolor has the smallest sepal width range with a median value of around 2.7 cm. Followed closely is the Iris-Virginica which has a median sepal width of 3 cm and two outliers (one below 2.2 cm and one above 4.0 cm) in its distribution. Iris-Setosa has the largest sepal width and largest range with a median range of 3.3 cm.



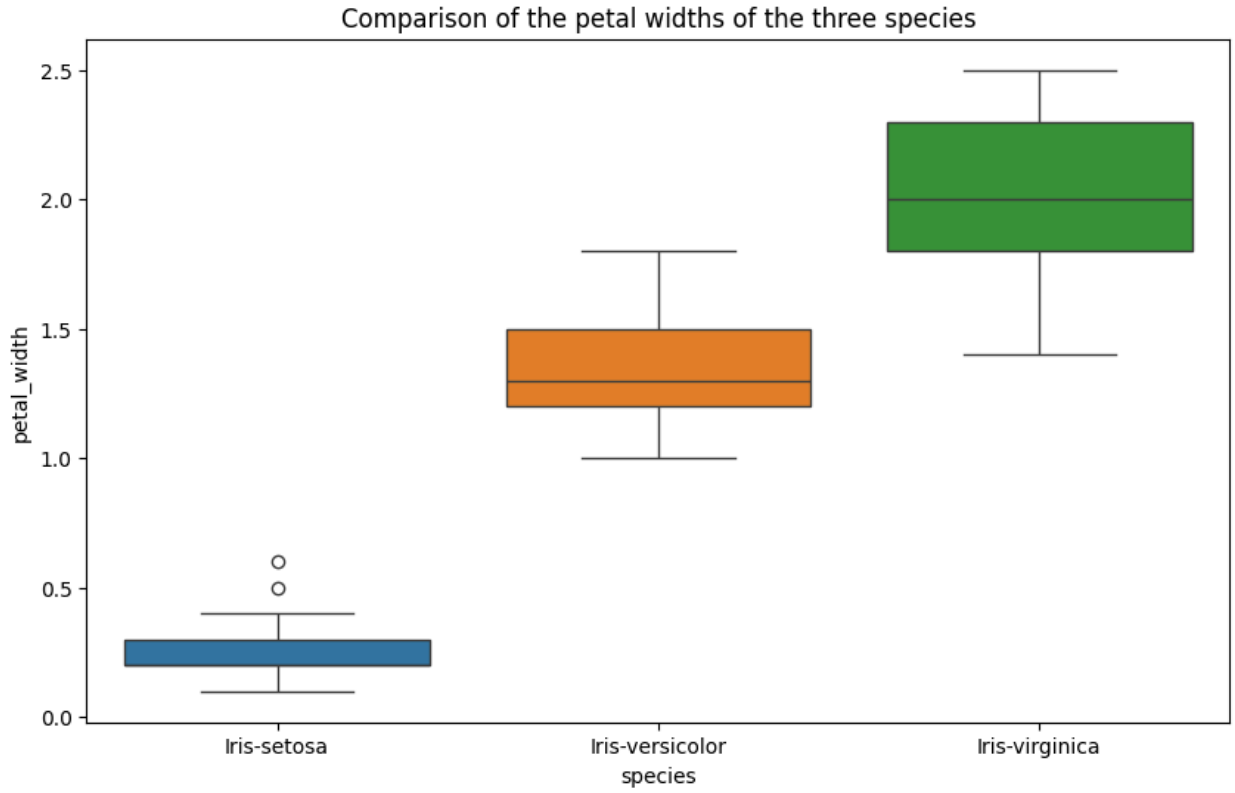
Visualization 4

Shows a boxplot distribution comparing the petal lengths of the three species. Iris-Setosa has the smallest petal length range with a median of around 1.5cm. The Iris-Setosa distribution contains some outliers two below 1cm and one around 2cm. Followed closely is the Iris-Versicolor having one outlier of around 3cm, and lastly is the species Iris-Virginica having the largest petal length range and median value of around 5.5cm.



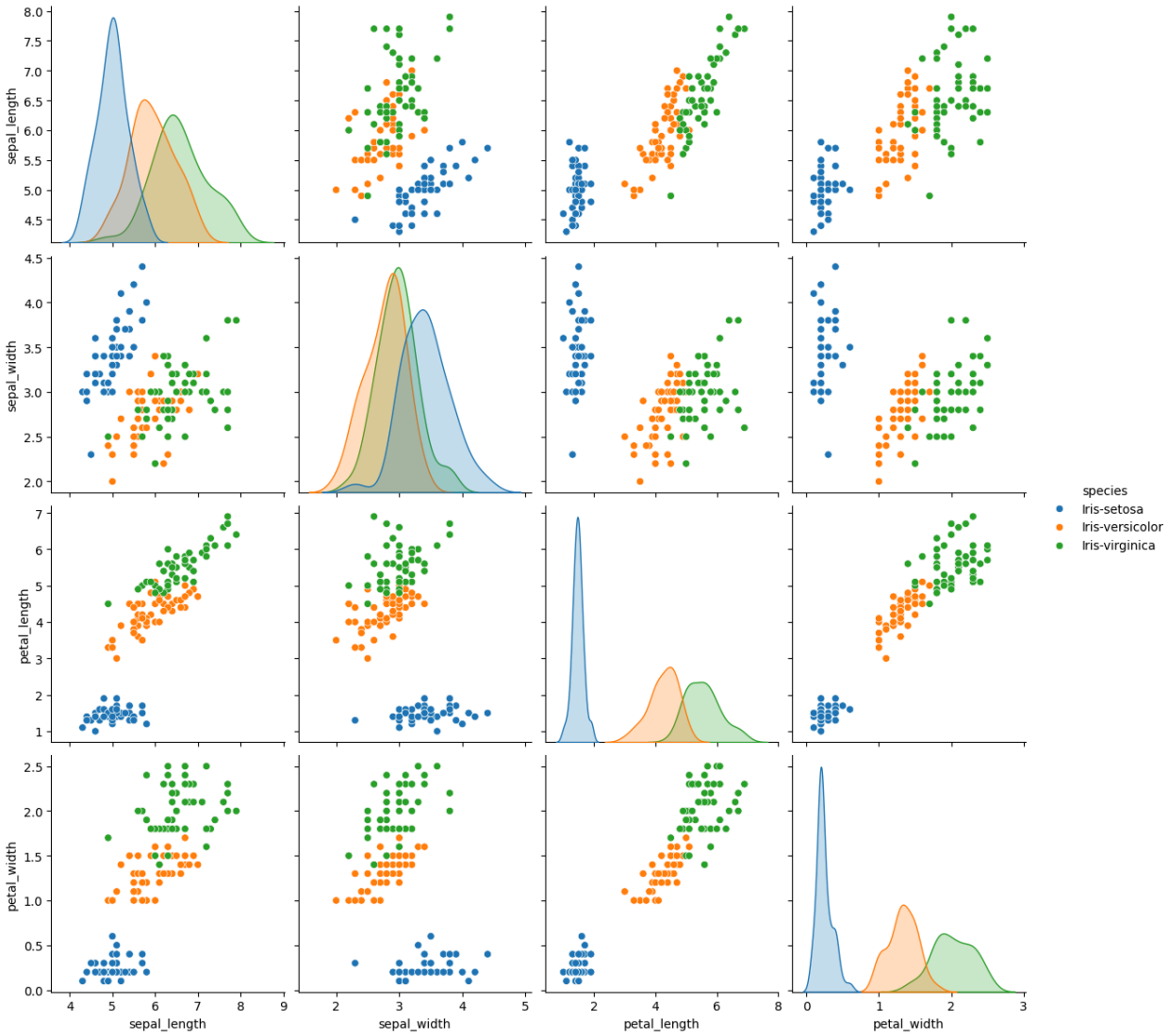
Visualization 5

A boxplot comparing the petal widths of the three species. Iris-Setosa's petal width is significantly smaller compared to the others with a median value of approximately 0.2cm. Followed closely is Iris-Versicolor with a median value of approximately 1.3 cm, and a range spanning from 1.0 cm to 1.8 cm. Lastly is the Iris-Virginica having the largest petal width ramge and median of 2.0 cm.



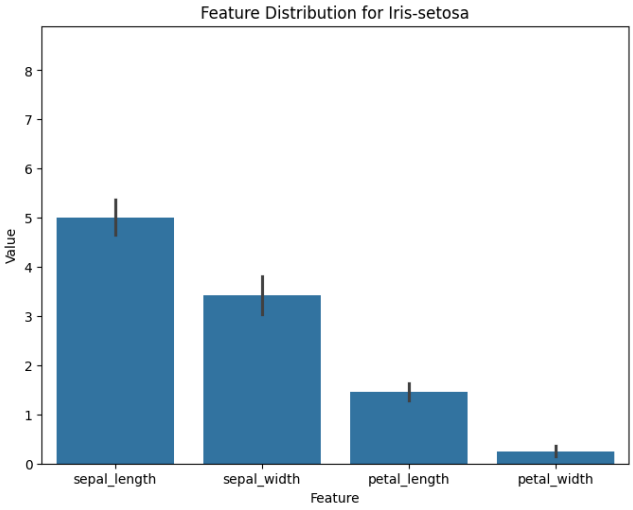
Visualization 6

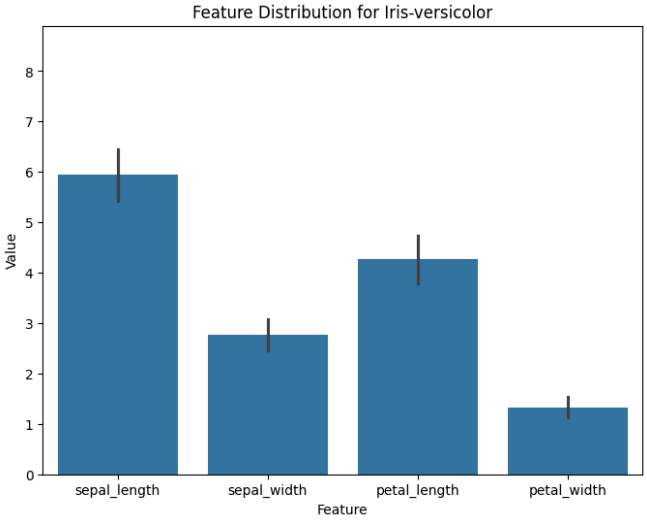
Pairplot showing the pair-wise relationships between the features and their distribution across the three species. Petal_length and petal_width show strong separation between species, especially for distinguishing Iris-setosa from the others. Iris-setosa forms separate clusters in almost all feature combinations, making it easy to distinguish. Finally, Iris-versicolor and Iris-virginica overlap in several feature combinations, indicating they are more challenging to separate.

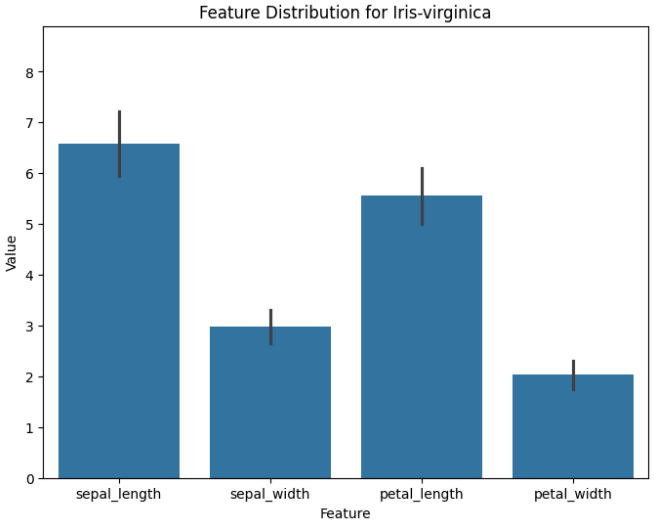


Visualization 7

A feature distribution of each species against their features. The following are the observations made: petal features (both length and width) are the most reliable for identifying Iris-virginica, Iris-versicolor is characterized by moderate-sized petals and sepals, making it distinct from the smaller petals of Iris-setosa and the larger petals of Iris-virginica.


Feature Distribution for Iris-setosa
Value
Feature
sepal_length
sepal_width
petal_length
petal_width
0
1
2
3
4
5
6
7
8


Feature Distribution for Iris-versicolor
Value
Feature
sepal_length
sepal_width
petal_length
petal_width
0
1
2
3
4
5
6
7
8


Feature Distribution for Iris-virginica
Value
Feature
sepal_length
sepal_width
petal_length
petal_width
0
1
2
3
4
5
6
7
8

C. Implement the following functions:

i. A function to calculate the covariance between two variables:

Solution

```
def cov(x, y):
    mean_x = sum(x) / float(len(x)) #to calculate the mean of x
    mean_y = sum(y) / float(len(y)) #to calculate the mean of y

    sub_x = [i - mean_x for i in x ] #subtracting the mean from each x variable
    sub_y = [i - mean_y for i in y] #subtracting the mean from each y variable
    sum_value =  sum([sub_x[i] * sub_y[i] for i in range (len(x))]) # sum of all the product
difference

    denom = float (len(x) - 1)

    covariance_value = sum_value / denom # divide the sum by the number of samples

    return covariance_value
```

ii. A function to compute the Covariance matrix:

Solution

```
def covMat(arr):
    c = [[cov(a,b) for a in arr.T] for b in arr.T]
    return c
```

D. Test1: compare the results of your function with NumPy's np.cov(data,rowvar=False) using the iris dataset.

Solution

Using np.cov()

Covariance Matrix:

```
[[ 0.68569351 -0.03926846  1.27368233  0.5169038 ]
 [-0.03926846  0.18800403 -0.32171275 -0.11798121]
 [ 1.27368233 -0.32171275  3.11317942  1.29638747]
 [ 0.5169038  -0.11798121  1.29638747  0.58241432]]
```

Using the computed covariance matrix function

Computed Covariance matrix [[0.68569351 -0.03926846 1.27368233 0.5169038]
[-0.03926846 0.18800403 -0.32171275 -0.11798121]
[1.27368233 -0.32171275 3.11317942 1.29638747]
[0.5169038 -0.11798121 1.29638747 0.58241432]]

The results are the same!

E. Using your covariance function, implement a function to calculate the correlation matrix:
Solution

```
def corrMat(arr):
  # Get the covariance matrix
  covariance_matrix = covMat(arr)

  # Calculate standard deviations for each feature
  std_devs = np.std(arr, axis=0)

  # Calculate the correlation matrix
  correlation_matrix = np.array([
    [covariance_matrix[i, j] / (std_devs[i] * std_devs[j]) for j in range(len(std_devs))]
    for i in range(len(std_devs))
  ])

  return correlation_matrix
```

F. Test2: validate your correlation matrix implementation by comparing it with the results of NumPy's np.corrcoef(data,rowvar=False), using the iris dataset.
Solution
Using np.corrcoef:
Correlation Coefficient:
[[1. -0.10936925 0.87175416 0.81795363]
[-0.10936925 1. -0.4205161 -0.35654409]
[0.87175416 -0.4205161 1. 0.9627571]
[0.81795363 -0.35654409 0.9627571 1.]]

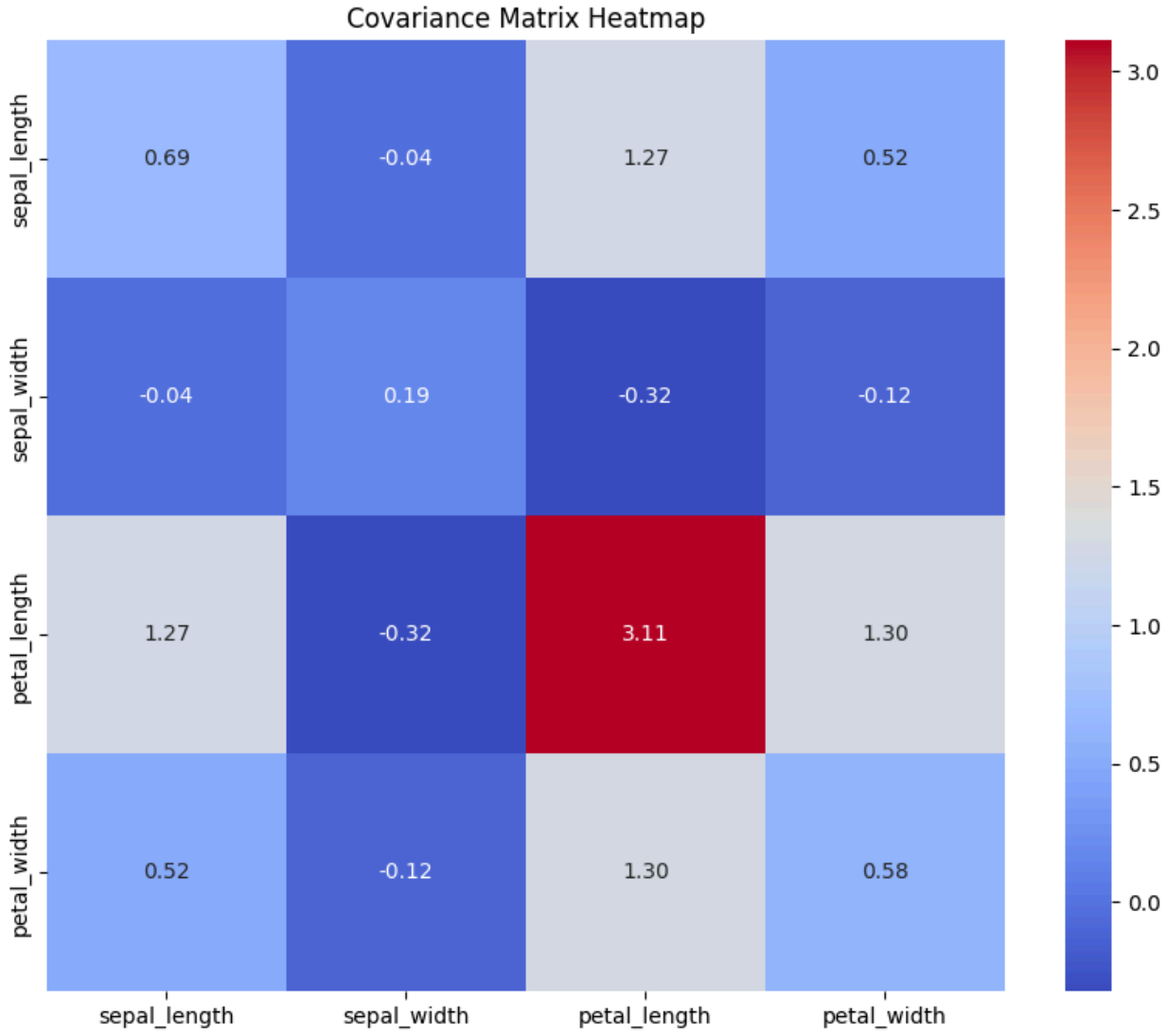
Using the computed correlation coefficient

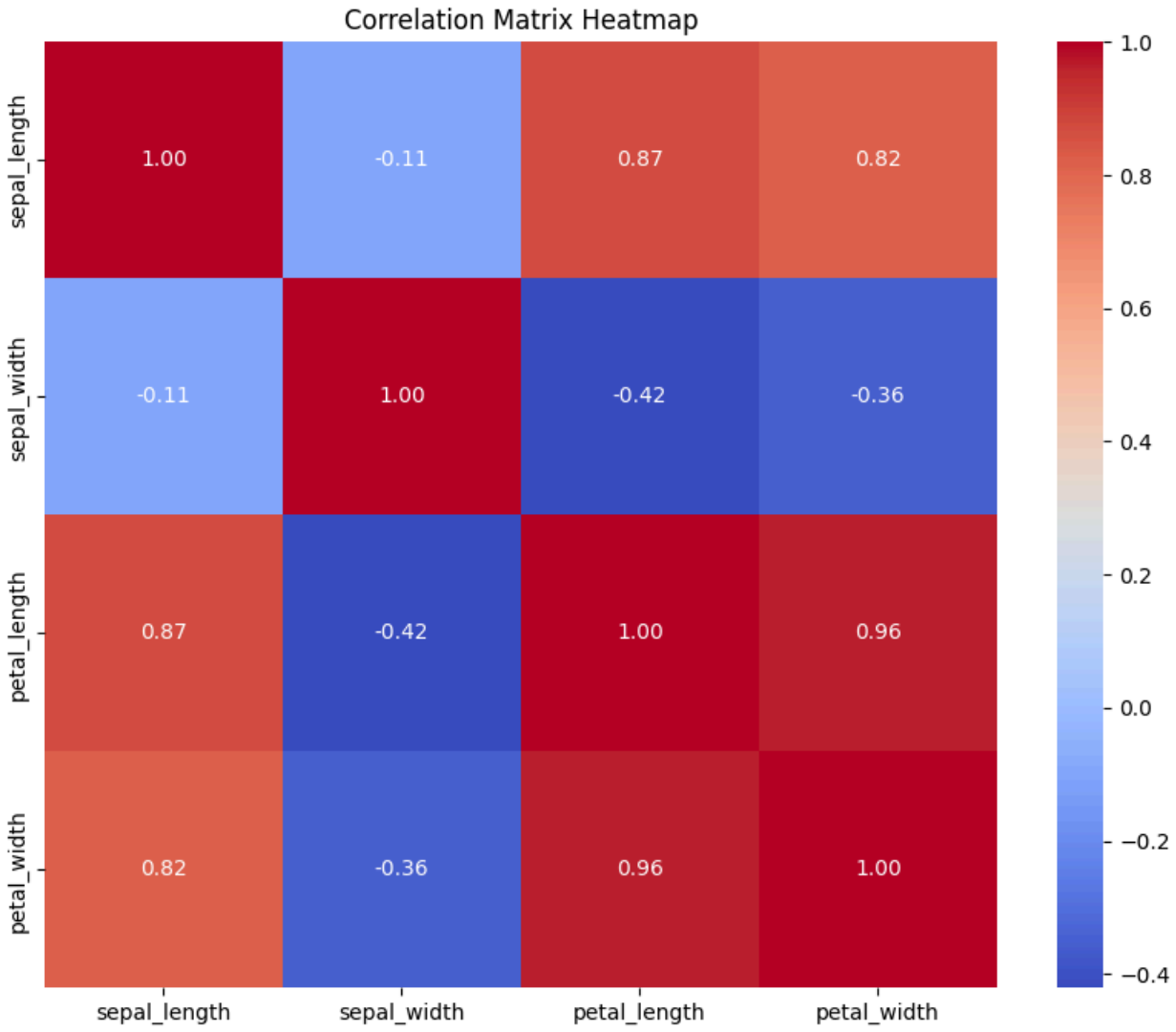
Computed Correlation matrix [[1.00671141 -0.11010327 0.87760486 0.82344326]
 [-0.11010327 1.00671141 -0.42333835 -0.358937]
 [0.87760486 -0.42333835 1.00671141 0.96921855]
 [0.82344326 -0.358937   0.96921855 1.00671141]]

The results are almost similar

G. Use visualizations to communicate the test results. Include appropriate titles, axis labels, and color bars where relevant.

Solution




Correlation Matrix Heatmap
sepal_length
sepal_width
petal_length
petal_width
1.00
-0.11
0.87
0.82
-0.11
1.00
-0.42
-0.36
0.87
-0.42
1.00
0.96
0.82
-0.36
0.96
1.00
sepal_length
sepal_width
petal_length
petal_width
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

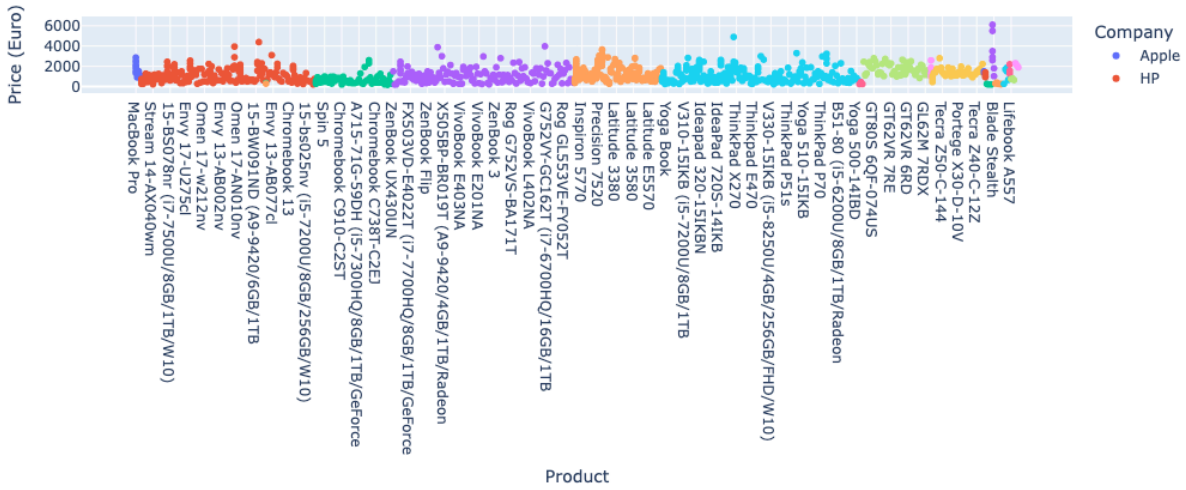
Q2

Please write code to complete the following tasks with the dataset "laptop-price – dataset.csv":

• Plot the price of all the laptops

Solution

An interactive plotly plot was plotted of all the prices of the laptops indicating their product names, Type name and RAM



• Which company has on average the most expensive laptop? What is the average laptop price for each company?

Solution

The **company** with the most expensive laptops on average is **Razer** with an average price of 3346.14 Euro.

Average laptop price for each company:

Company

Razer 3346.142857

LG 2099.000000

MSI 1728.908148

Google 1677.666667

Microsoft 1612.308333

Apple 1564.198571

Huawei 1424.000000

Samsung 1413.444444

Toshiba 1267.812500

Dell 1199.225120
Xiaomi 1133.462500
Asus 1123.829737
Lenovo 1093.862215
HP 1080.314664
Fujitsu 729.000000
Acer 633.464455
Chuwi 314.296667
Mediacom 295.000000
Vero 217.425000

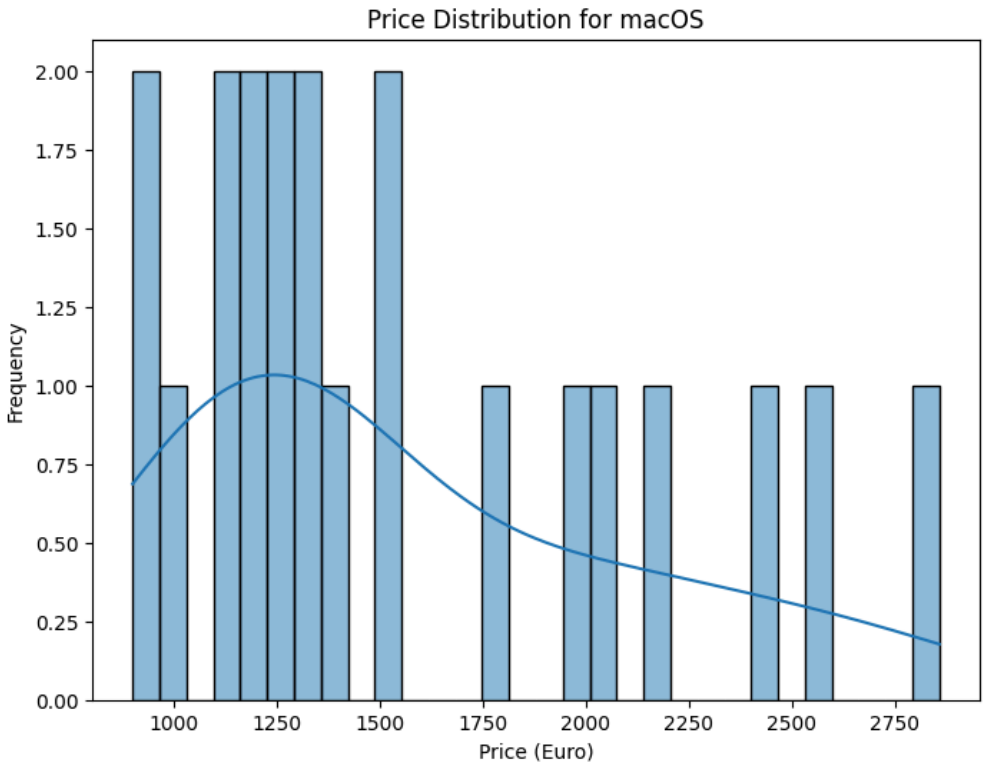
• Find the different types of Operating systems present in the data - under the column name "OpSys".
o Please note - there are operating systems that are the same systems and just written differently in the column - please fix them to be uniform.

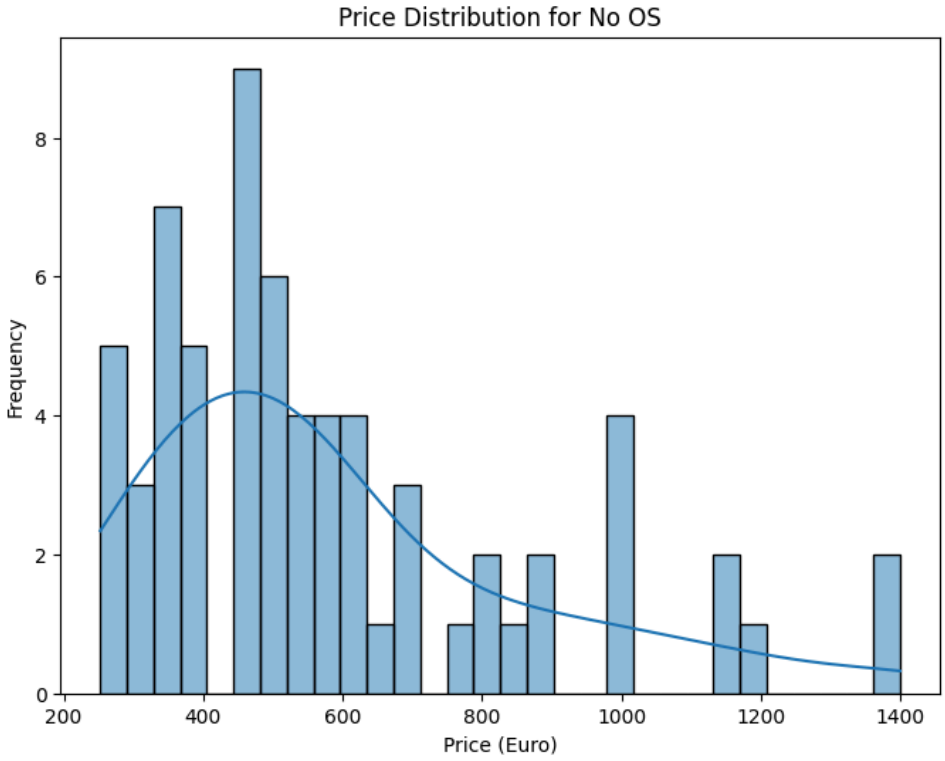
Solution

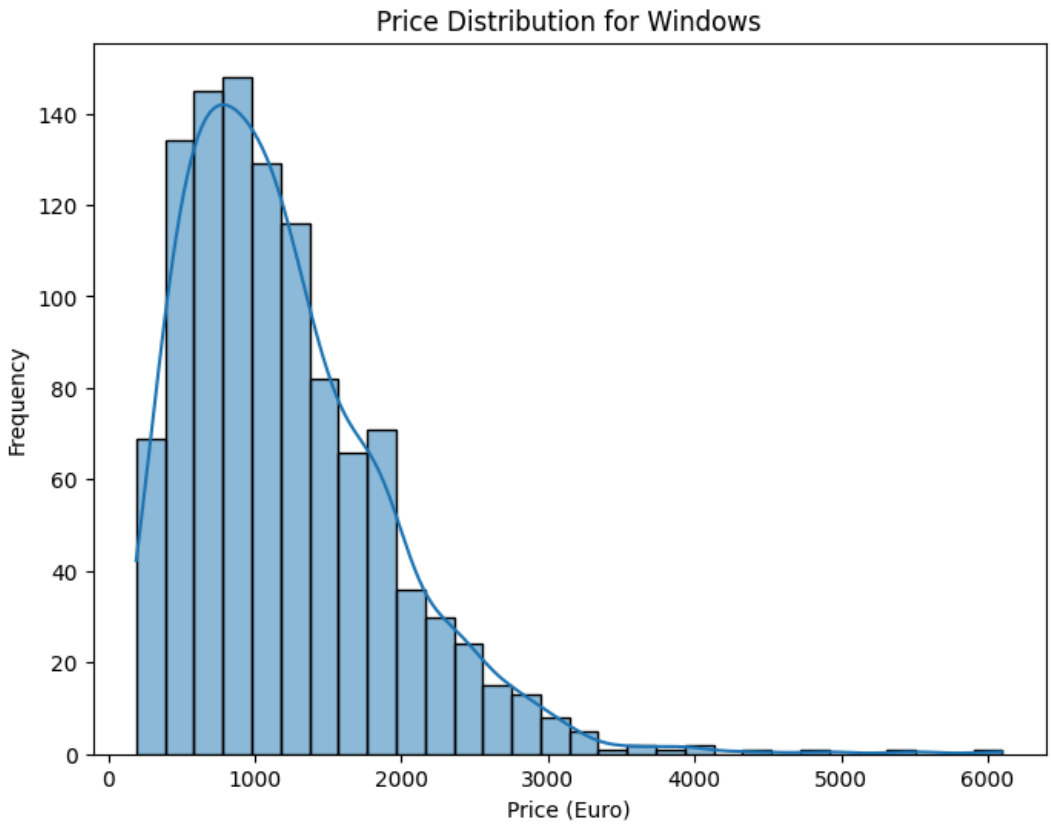
The different types of Operating Systems present are:
['macOS', 'No OS', 'Windows', 'Linux', 'Android']

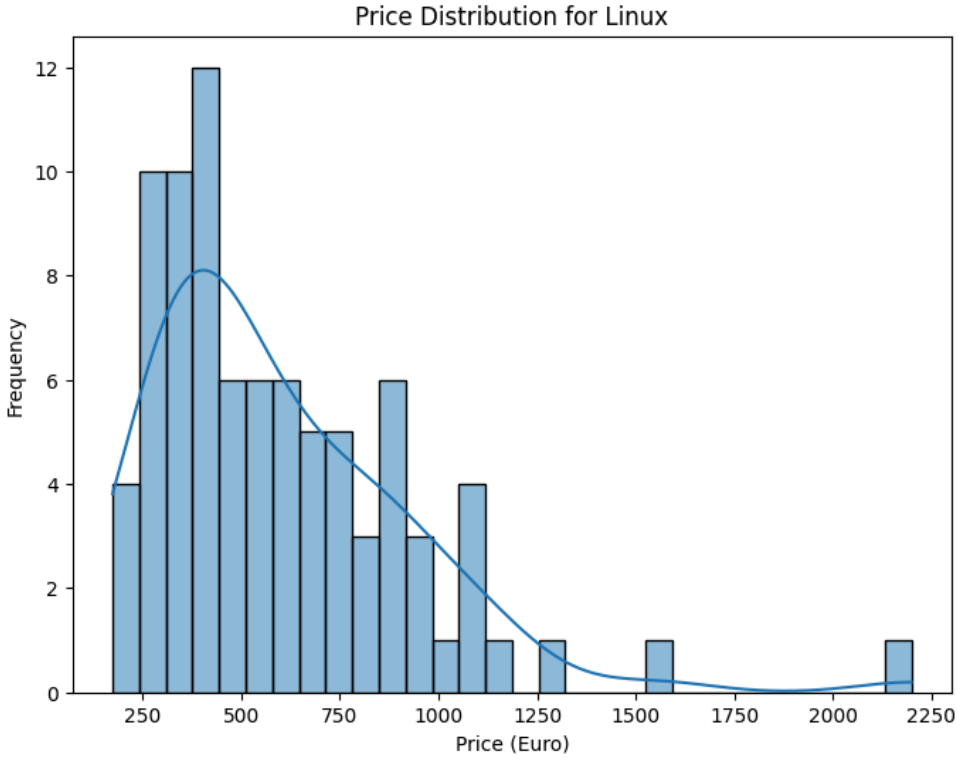
• Plot for each of the operating system types the distribution of the prices, so that the number of plots equals to the number of unique operating systems.

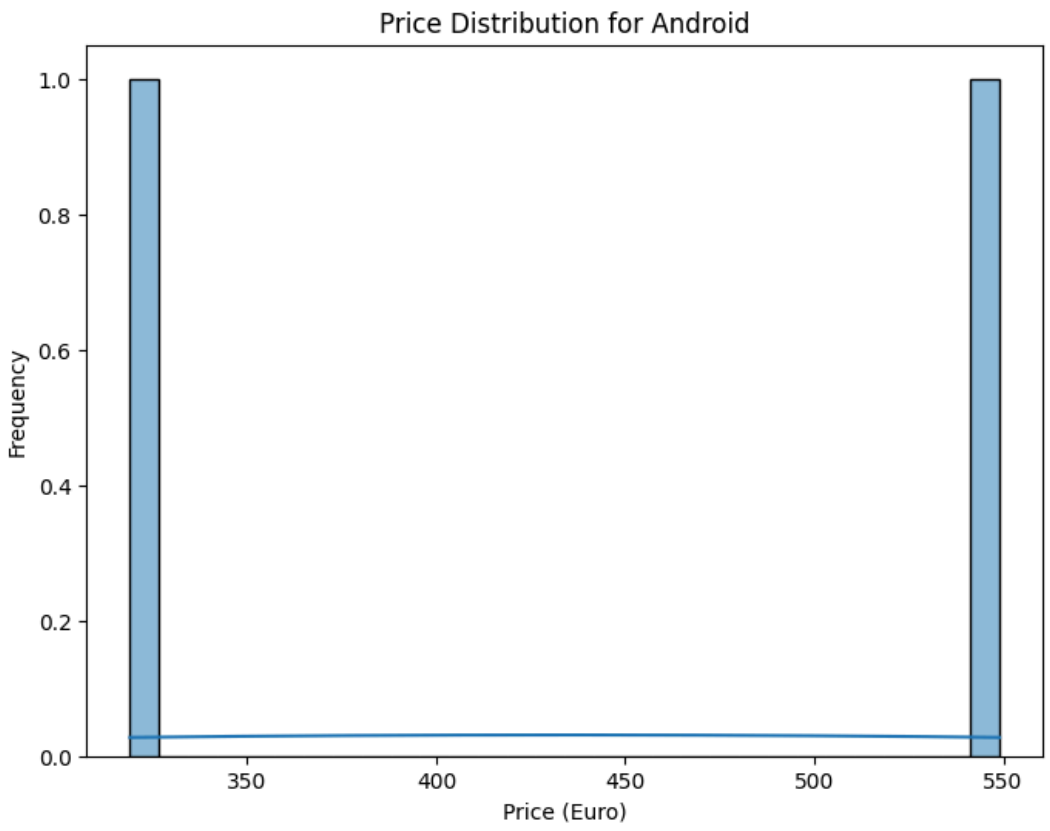
Solution


Price Distribution for macOS
Frequency
Price (Euro)
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
1000
1250
1500
1750
2000
2250
2500
2750


Price Distribution for No OS
Frequency
Price (Euro)
0
2
4
6
8
200
400
600
800
1000
1200
1400


Price Distribution for Windows
Frequency
Price (Euro)
0
20
40
60
80
100
120
140
0
1000
2000
3000
4000
5000
6000


Price Distribution for Linux
Frequency
Price (Euro)
0
2
4
6
8
10
12
250
500
750
1000
1250
1500
1750
2000
2250



• What is the relationship between RAM and computer price? add an adequate plot to support your findings.

Solution

Correlation between RAM and Price: 0.74028652716227

This implies that RAM has a positive linear relationship with Price which implies that laptops with higher RAMs relatively have higher prices although there are some other things contribute to the increase in price. If RAM was fully linear, the correlation would have been 1.

The plot is shown below:

• Create a new column for the dataframe called "Storage type" that extracts the storage type from the column "Memory".

o For example, in the first row in the column "Memory" it states "128GB SSD", the new column will have just "SSD" in its first row.

Solution

0 SSD

1 Flash Storage

2 SSD
3 SSD
4 SSD
Name: Storage Type, dtype: object

Q3 Think of additional questions related to this data. What types of analyses and visualizations would you use to address them? Select two questions from your list and implement. Submit your list of questions, suggested analyses and visualizations and the implementation.

Solution

Additional Questions are:

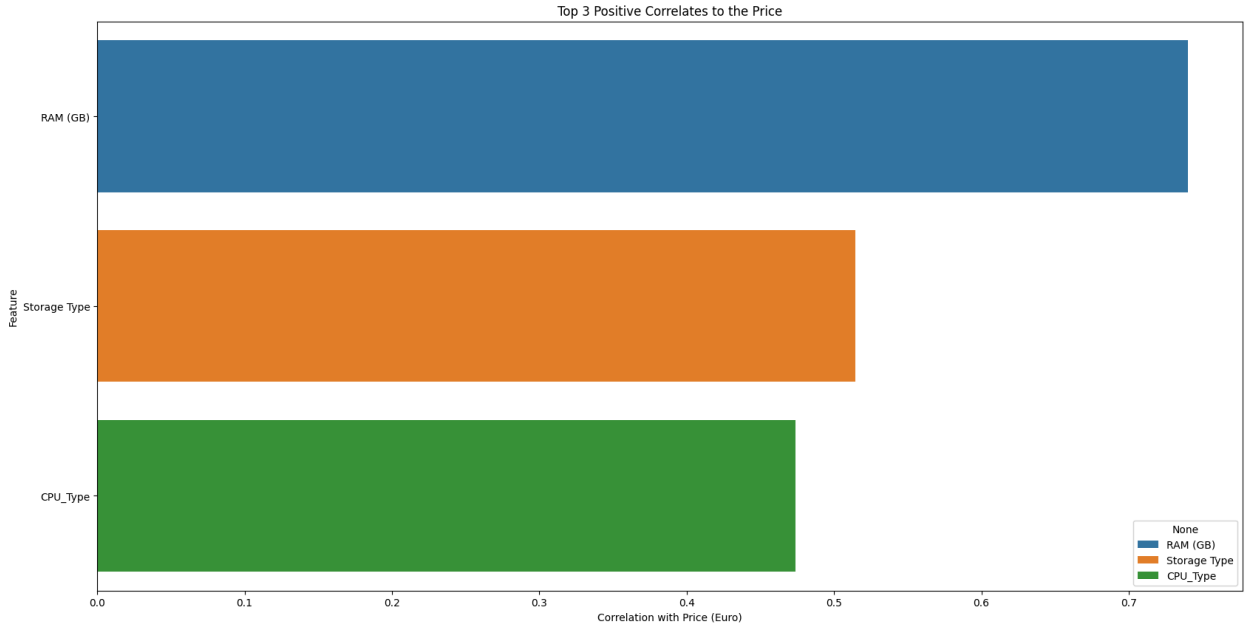
1. What top features positively influence the price of the laptop?
2. Which laptop feature is each of the company known for? - this allows us to know which company dominnates which laptop feature market.
3. What brands produce the best value for money laptops? (budget friendly)
4. Does hybrid storage (SSD + HDD) impact the price signifiantly?
5. What is the price difference between gaming laptops and other types?
6. What combination of features offers the best balance between price and performance?

Question 1

What top features positively influence the price of the laptop?

From the plot below, it shows the top three laptop features that influence the prices of laptops are

Top 3 features most correlated with Price:
RAM (GB) 0.740287
Storage Type 0.514201
CPU_Type 0.473860



Question 2

Which market does each of these companies dominate with respect to the laptop features?

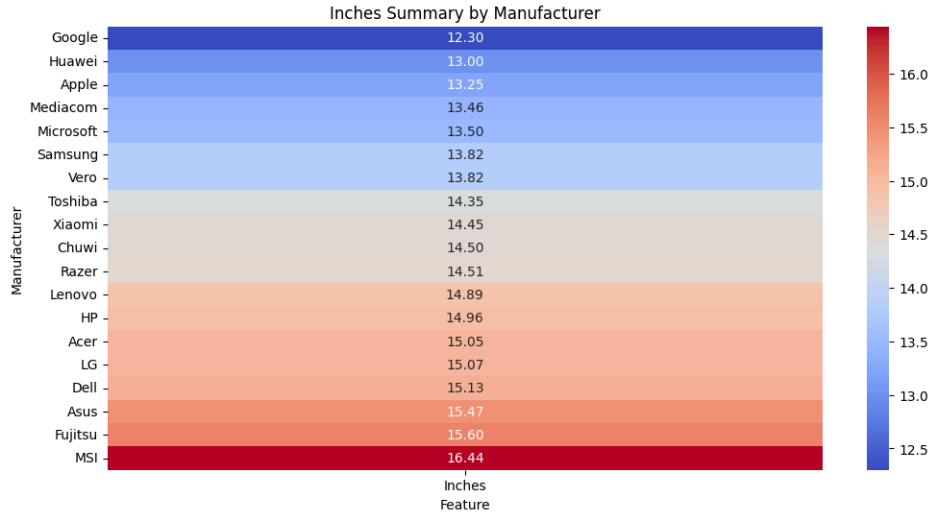
Solution

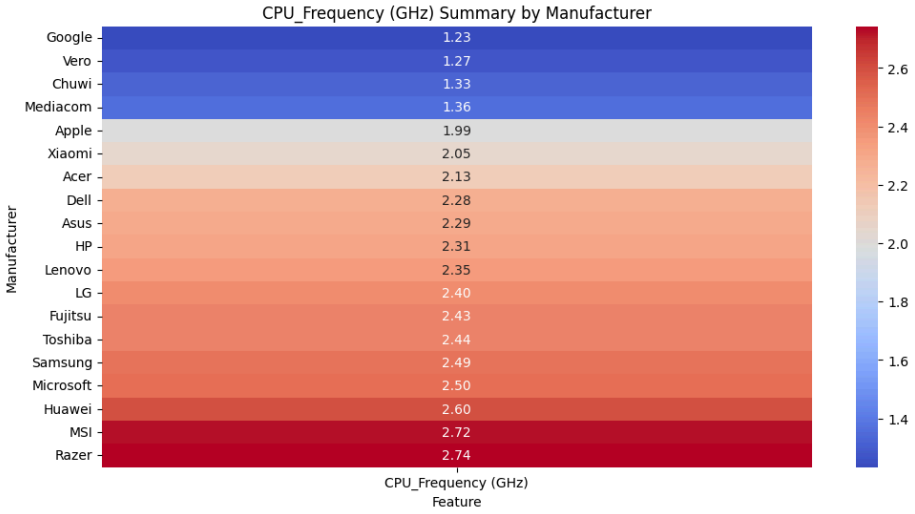
I divided the laptop features into two: Numerical and Categorical Features.

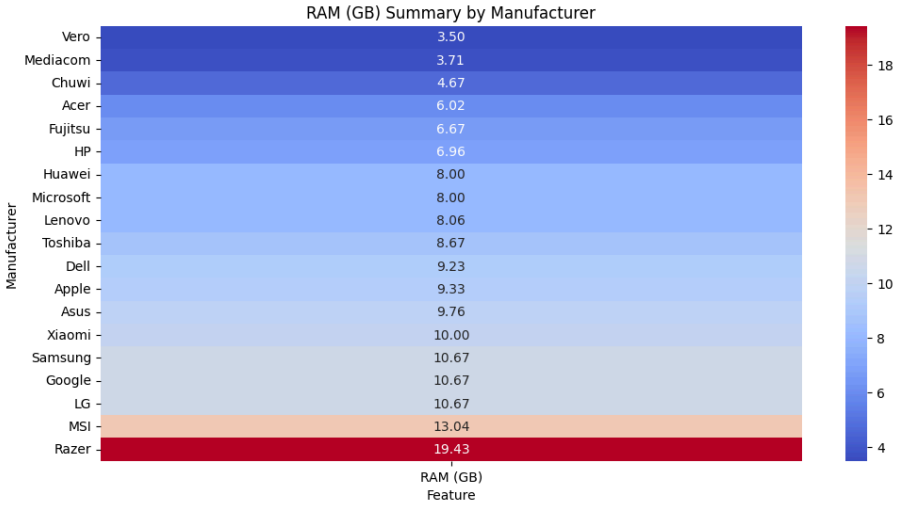
For the Numerical Features like:

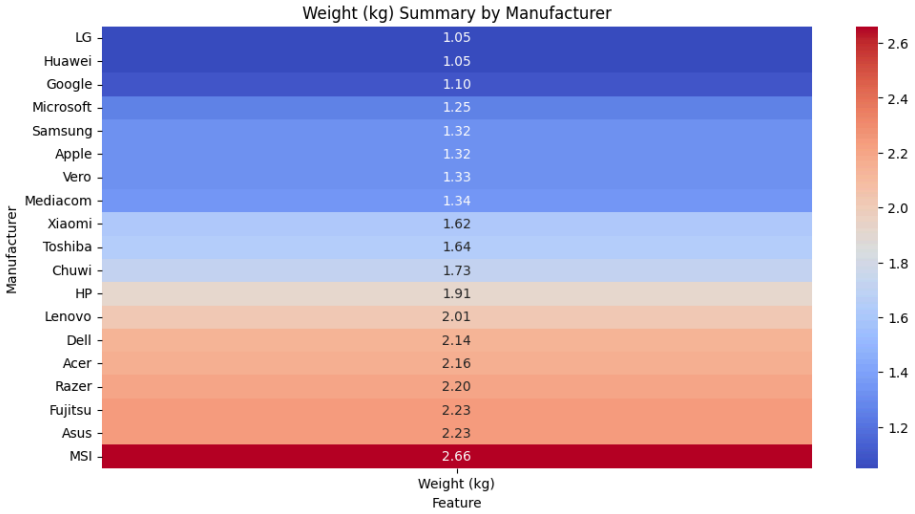
`Inches', 'CPU_Frequency (GHz)', 'RAM (GB)', 'Weight (kg)', 'Price (Euro)'

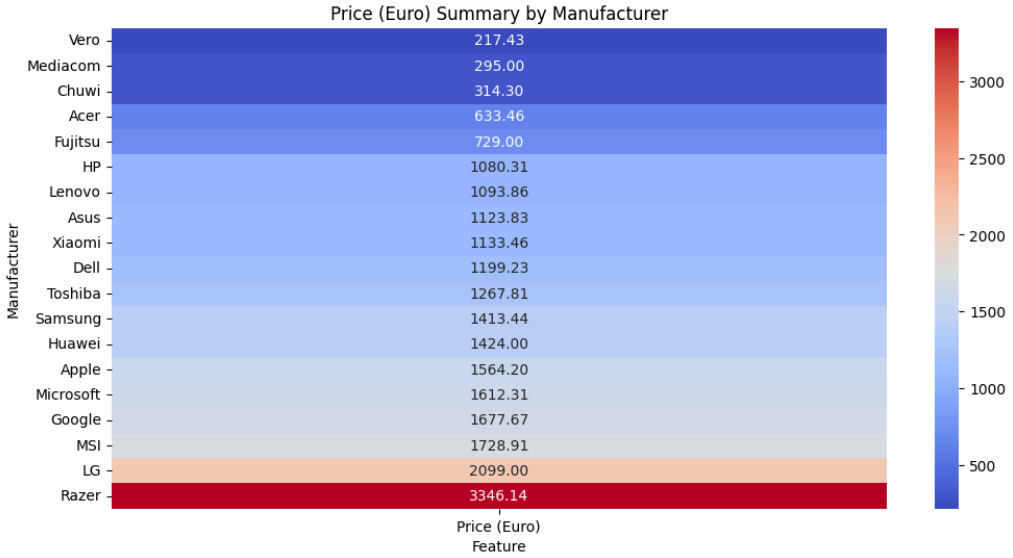
I plotted separate plots to see what each company is known for in terms of laptop designing.




CPU_Frequency (GHz) Summary by Manufacturer
Google 1.23
Vero 1.27
Chuwi 1.33
Mediacom 1.36
Apple 1.99
Xiaomi 2.05
Acer 2.13
Dell 2.28
Asus 2.29
HP 2.31
Lenovo 2.35
LG 2.40
Fujitsu 2.43
Toshiba 2.44
Samsung 2.49
Microsoft 2.50
Huawei 2.60
MSI 2.72
Razer 2.74
Manufacturer
CPU_Frequency (GHz)
Feature
2.6
2.4
2.2
2.0
1.8
1.6
1.4


RAM (GB) Summary by Manufacturer
Vero 3.50
Mediacom 3.71
Chuwi 4.67
Acer 6.02
Fujitsu 6.67
HP 6.96
Huawei 8.00
Microsoft 8.00
Lenovo 8.06
Toshiba 8.67
Dell 9.23
Apple 9.33
Asus 9.76
Xiaomi 10.00
Samsung 10.67
Google 10.67
LG 10.67
MSI 13.04
Razer 19.43
Manufacturer
RAM (GB)
Feature
18
16
14
12
10
8
6
4


Weight (kg) Summary by Manufacturer
LG 1.05
Huawei 1.05
Google 1.10
Microsoft 1.25
Samsung 1.32
Apple 1.32
Vero 1.33
Mediacom 1.34
Xiaomi 1.62
Toshiba 1.64
Chuwi 1.73
HP 1.91
Lenovo 2.01
Dell 2.14
Acer 2.16
Razer 2.20
Fujitsu 2.23
Asus 2.23
MSI 2.66
Manufacturer
Weight (kg)
Feature
2.6
2.4
2.2
2.0
1.8
1.6
1.4
1.2



Key Insights

MSI is known to produce large screen sized and heavy weights-laptops. One could infer their target market is the gaming industry. Professionals who use laptops for gaming. Razer, being the most expensive of them all focuses exclusively on high RAM and moderately sized laptops with very high CPUs also targeted towards the gamers.

Google has the smallest average screen size (12.30 inches), indicating a focus on portability including Huawei which produces the lightest laptops on average (1.05 kg), focusing on portability. This implies that Huawei and Google lead in portability with lightweight laptops and smaller screens, suitable for professionals on the go.

Vero, Chuwi, and Mediacom produce low-cost laptops with smaller RAM and lower performance, which implies they cater to budget-conscious buyers.

Apple focuses on lightweight laptops with higher RAM and premium prices.

In summary,

Manufacturers specialize in different aspects:

Razer in performance, Apple in premium lightweight models, and Huawei in portability.

Razer and MSI target gamers and professionals needing performance.

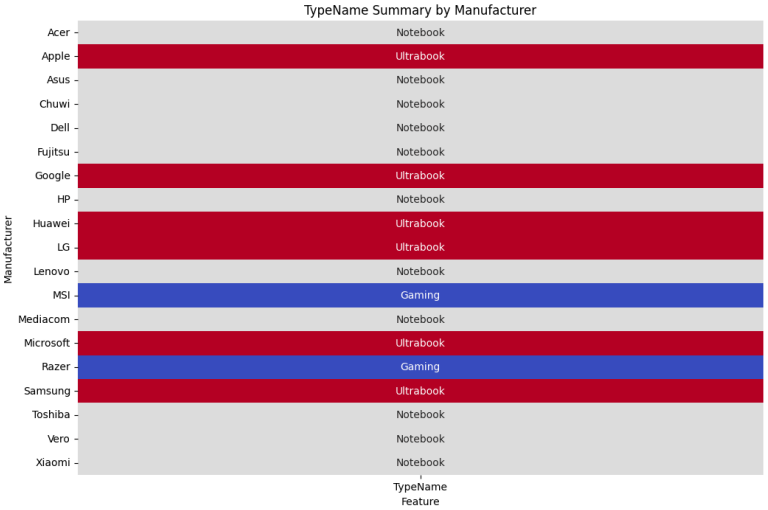
Vero and Chuwi are budget-focused manufacturers who cater to cost-sensitive users.

For the Categorical Features like:

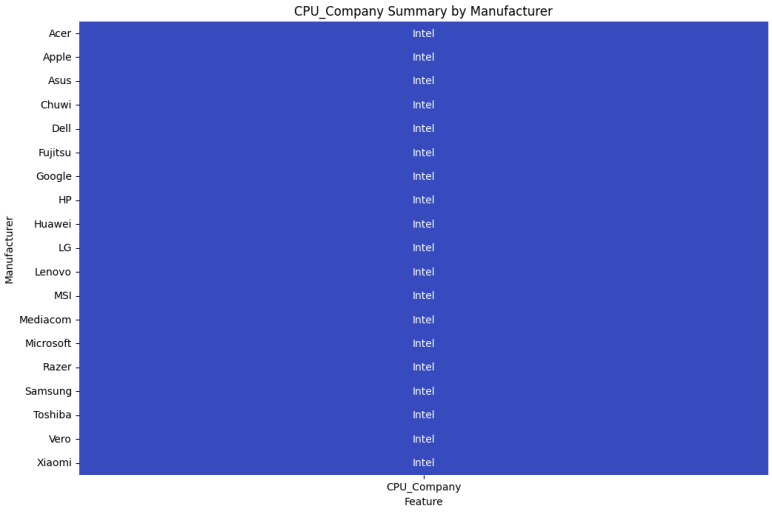
'Company', 'Product', 'TypeName', 'ScreenResolution', 'CPU_Company', 'CPU_Type', 'Memory', 'GPU_Company', 'GPU_Type', 'OpSys', 'Storage Type'

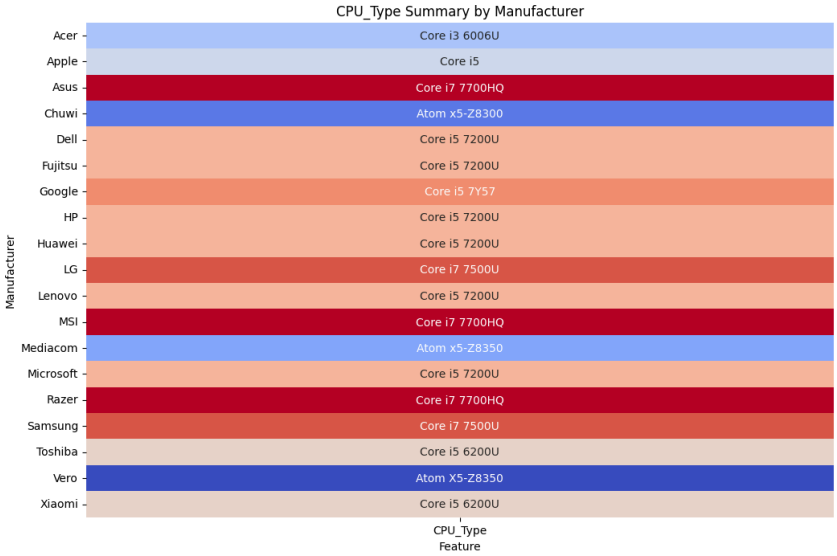
I plotted each feature per the manufacturer

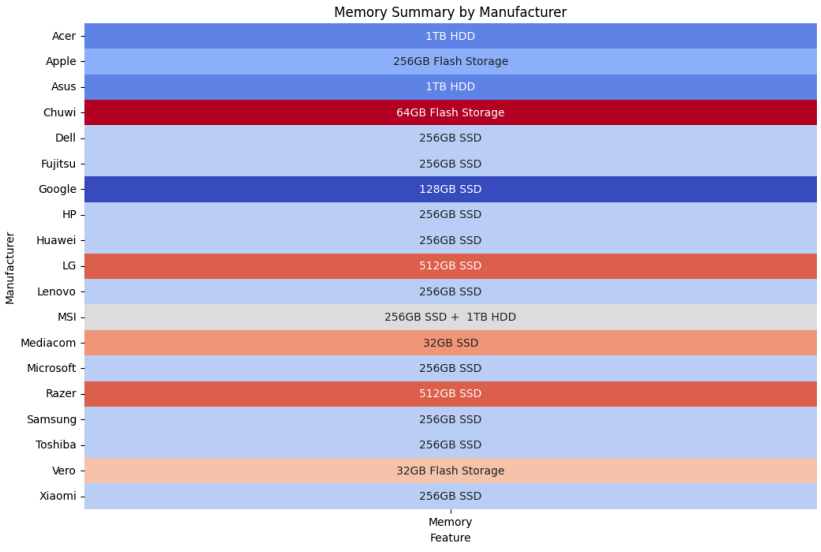

Product Summary by Manufacturer
Acer
Apple
Asus
Chuwi
Dell
Fujitsu
Google
HP
Huawei
LG
Lenovo
MSI
Mediacom
Microsoft
Razer
Samsung
Toshiba
Vero
Xiaomi
Manufacturer
Aspire 3
MacBook Pro
Rog Strix
LapBook 12.3
XPS 13
LifeBook A556
Pixelbook (Core
250 G6
MateBook X
Gram 14Z970
Legion Y520-15IKBN
GE72MVR 7RG
SmartBook Edge
Surface Laptop
Blade Pro
Notebook 9
Satellite Pro
K146 (N3350/4GB/32GB/W10)
Mi Notebook
Product
Feature


TypeName Summary by Manufacturer
Acer
Apple
Asus
Chuwi
Dell
Fujitsu
Google
HP
Huawei
LG
Lenovo
MSI
Mediacom
Microsoft
Razer
Samsung
Toshiba
Vero
Xiaomi
Manufacturer
Notebook
Ultrabook
Notebook
Notebook
Notebook
Notebook
Ultrabook
Notebook
Ultrabook
Ultrabook
Notebook
Gaming
Notebook
Ultrabook
Gaming
Ultrabook
Notebook
Notebook
Notebook
TypeName
Feature

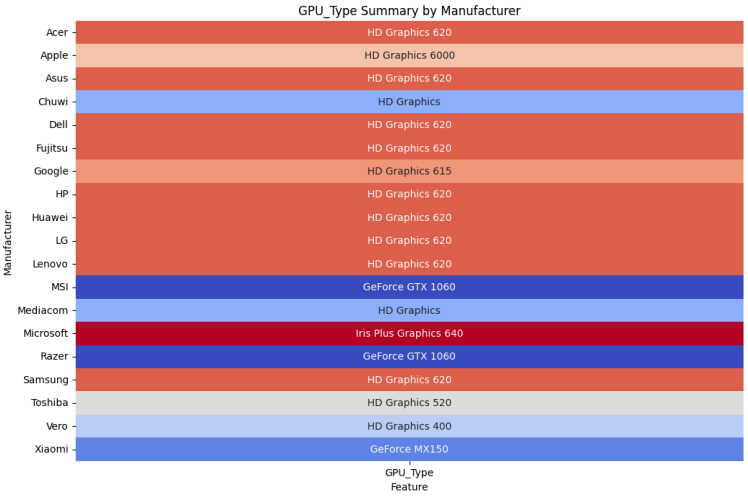

ScreenResolution Summary by Manufacturer
Acer
Apple
Asus
Chuwi
Dell
Fujitsu
Google
HP
Huawei
LG
Lenovo
MSI
Mediacom
Microsoft
Razer
Samsung
Toshiba
Vero
Xiaomi
Manufacturer
1366x768
IPS Panel Retina Display 2304x1440
Full HD 1920x1080
Full HD 1920x1080
Full HD 1920x1080
1366x768
Touchscreen 2400x1600
Full HD 1920x1080
IPS Panel Full HD 2160x1440
IPS Panel Full HD / Touchscreen 1920x1080
Full HD 1920x1080
Full HD 1920x1080
IPS Panel Full HD 1920x1080
Touchscreen 2256x1504
Full HD 1920x1080
Full HD 1920x1080
IPS Panel Full HD 1920x1080
1366x768
IPS Panel Full HD 1920x1080
ScreenResolution
Feature

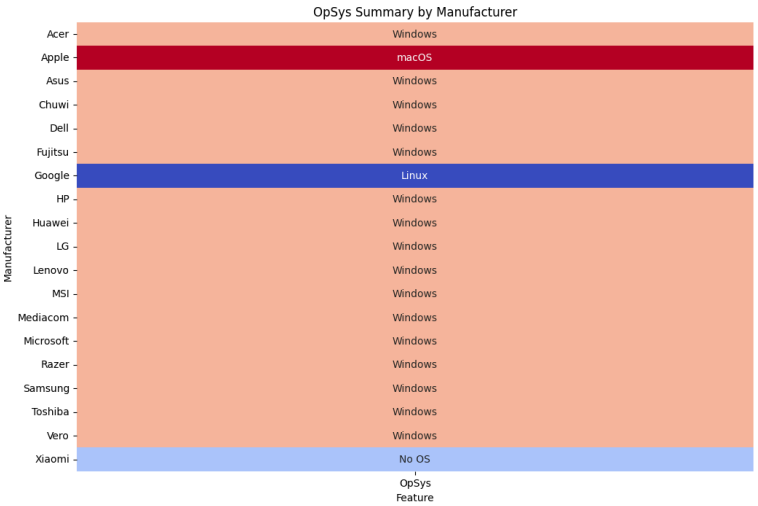

CPU_Company Summary by Manufacturer
Acer
Apple
Asus
Chuwi
Dell
Fujitsu
Google
HP
Huawei
LG
Lenovo
MSI
Mediacom
Microsoft
Razer
Samsung
Toshiba
Vero
Xiaomi
Intel
Manufacturer
CPU_Company
Feature

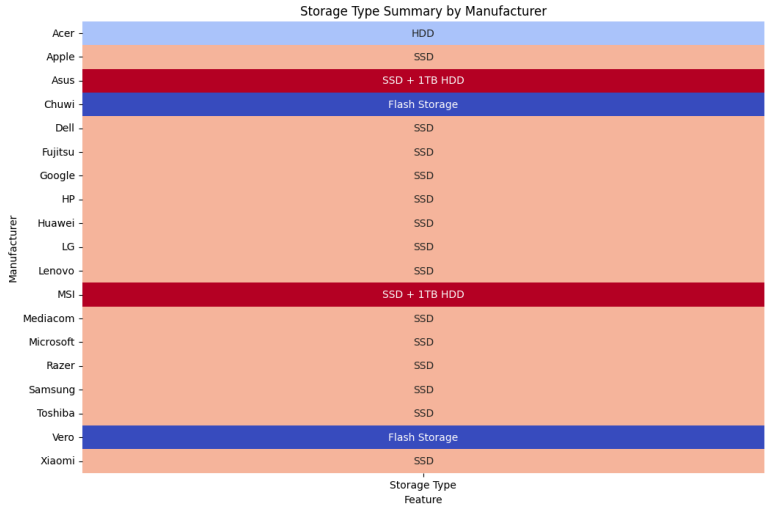

CPU_Type Summary by Manufacturer
Acer
Core i3 6006U
Apple
Core i5
Asus
Core i7 7700HQ
Chuwi
Atom x5-Z8300
Dell
Core i5 7200U
Fujitsu
Core i5 7200U
Google
Core i5 7Y57
HP
Core i5 7200U
Huawei
Core i5 7200U
LG
Core i7 7500U
Lenovo
Core i5 7200U
MSI
Core i7 7700HQ
Mediacom
Atom x5-Z8350
Microsoft
Core i5 7200U
Razer
Core i7 7700HQ
Samsung
Core i7 7500U
Toshiba
Core i5 6200U
Vero
Atom X5-Z8350
Xiaomi
Core i5 6200U
Manufacturer
CPU_Type
Feature


Memory Summary by Manufacturer
Acer
1TB HDD
Apple
256GB Flash Storage
Asus
1TB HDD
Chuwi
64GB Flash Storage
Dell
256GB SSD
Fujitsu
256GB SSD
Google
128GB SSD
HP
256GB SSD
Huawei
256GB SSD
LG
512GB SSD
Lenovo
256GB SSD
MSI
256GB SSD + 1TB HDD
Mediacom
32GB SSD
Microsoft
256GB SSD
Razer
512GB SSD
Samsung
256GB SSD
Toshiba
256GB SSD
Vero
32GB Flash Storage
Xiaomi
256GB SSD
Manufacturer
Memory
Feature


GPU_Company Summary by Manufacturer
Acer
Apple
Asus
Chuwi
Dell
Fujitsu
Google
HP
Huawei
LG
Lenovo
MSI
Mediacom
Microsoft
Razer
Samsung
Toshiba
Vero
Xiaomi
Intel
Nvidia
Manufacturer
GPU_Company
Feature


GPU_Type Summary by Manufacturer
Acer
Apple
Asus
Chuwi
Dell
Fujitsu
Google
HP
Huawei
LG
Lenovo
MSI
Mediacom
Microsoft
Razer
Samsung
Toshiba
Vero
Xiaomi
HD Graphics 620
HD Graphics 6000
HD Graphics
HD Graphics 615
GeForce GTX 1060
Iris Plus Graphics 640
HD Graphics 520
HD Graphics 400
GeForce MX150
Manufacturer
GPU_Type
Feature


OpSys Summary by Manufacturer
Acer
Apple
Asus
Chuwi
Dell
Fujitsu
Google
HP
Huawei
LG
Lenovo
MSI
Mediacom
Microsoft
Razer
Samsung
Toshiba
Vero
Xiaomi
Windows
macOS
Linux
No OS
Manufacturer
OpSys
Feature



Key Insights

Apple exclusively uses macOS. Dell, HP, Lenovo primarily manufacture Notebook laptops, catering to a wide range of users. Most manufacturers use SSD as the dominant storage type, except: Acer which still has significant use of HDD and MSI that offers high-performance storage configurations like SSD + 1TB HDD. In terms of screen resolution, Apple is known for high-resolution displays (IPS Panel Retina Display 2304x1440). Google and Microsoft use advanced touchscreen resolutions like 2256x1504 or 2400x1600, while other manufacturers like HP, Dell, and Lenovo primarily use Full HD (1920x1080). Finally, Intel CPUs dominate across most manufacturers.

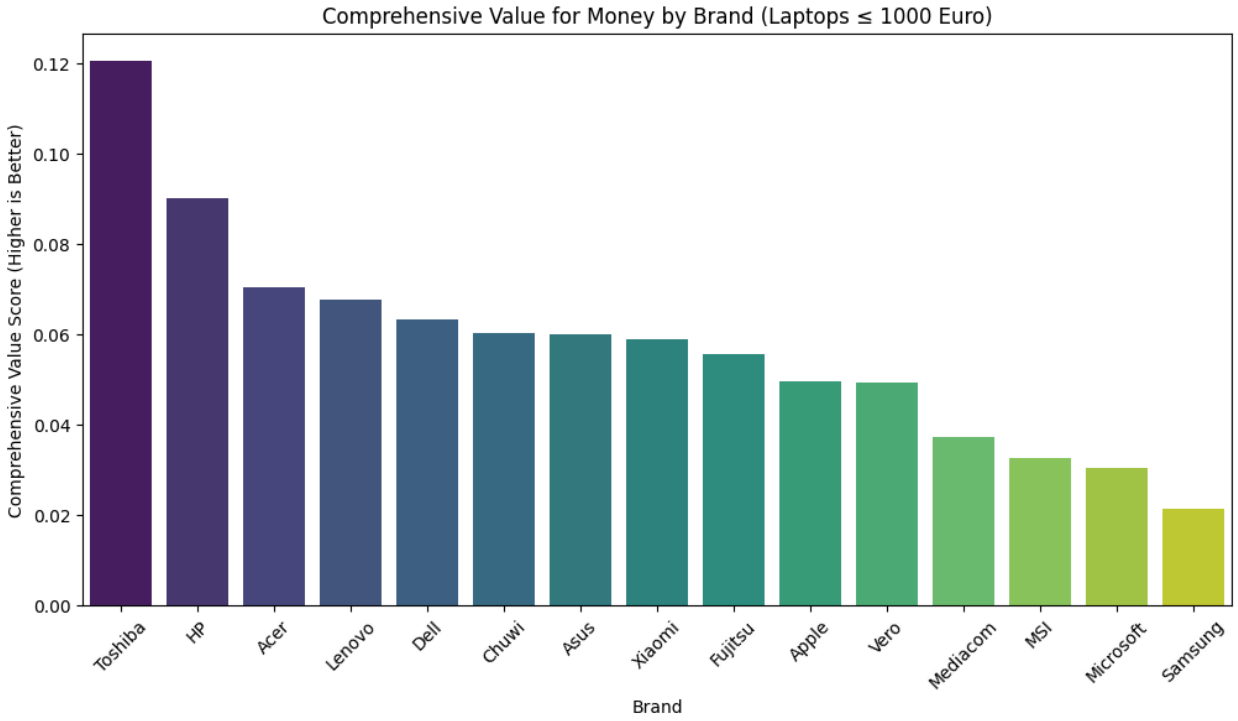
In summary,

- Apple: Focuses on premium, lightweight Ultrabooks with macOS, SSD storage, and high-resolution displays.
- MSI: Targets gamers with gaming laptops, high-end GPUs, and hybrid storage configurations.
- Google and Microsoft: Cater to professionals with touchscreens and specialized designs.
- Mainstream Manufacturers (HP, Dell, Lenovo): Offer affordable notebooks with standard Windows, Intel CPUs, and Full HD screens.

Question 3

What brands produce the best value for money laptops (budget friendly)?

I want to focus on brands that provide value for lower prices priced at 1000 euro and below



This visualization implies that buyers looking for a good mix of RAM, CPU, storage, and screen size within 1000 euros should consider brands like Toshiba, HP, or Acer.

Question 4

Does hybrid storage (SSD + HDD) impact the price signifiantly?

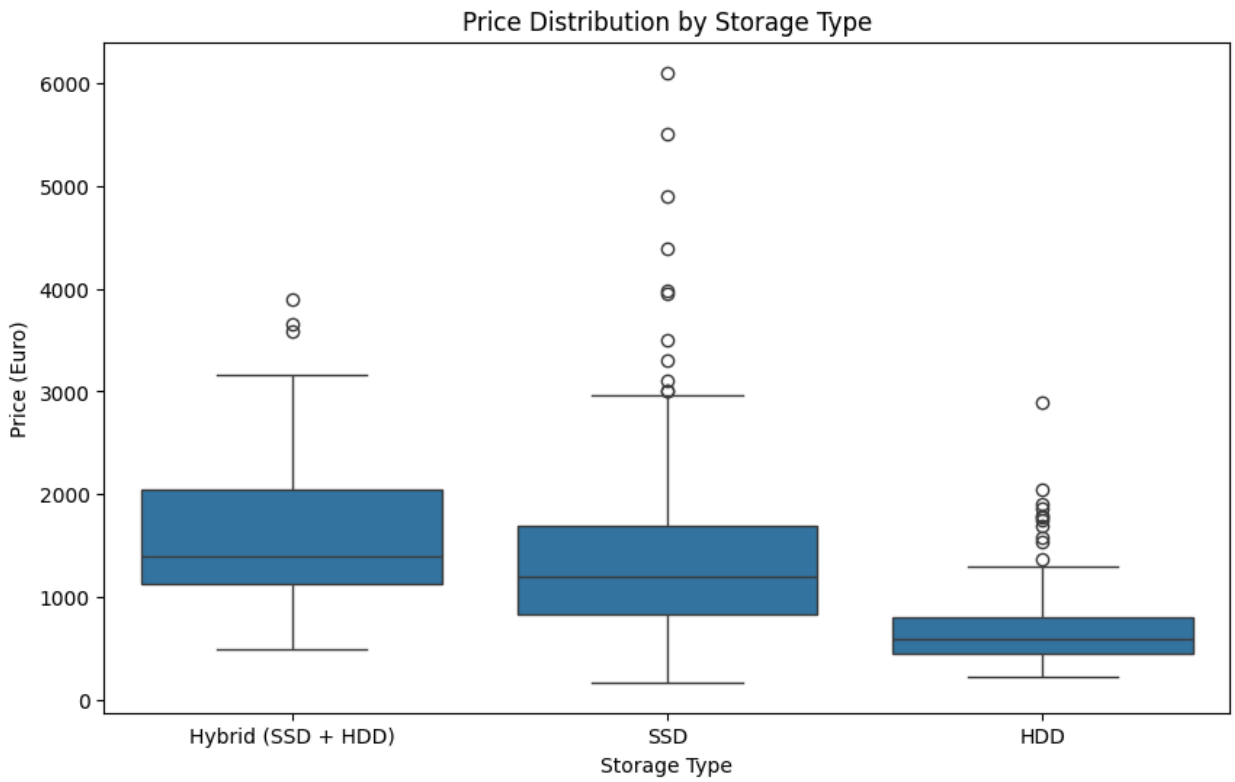
Average Prices by Storage Type:

Hybrid (SSD + HDD): €1613.46

SSD Only: €1316.41

HDD Only: €658.50

The visualization below shows hybrid storage significantly impact prices of laptops.


Price Distribution by Storage Type
Price (Euro)
0
1000
2000
3000
4000
5000
6000
Hybrid (SSD + HDD)
SSD
HDD
Storage Type