Dear Intern

Interim project report is an inherent component of your internship. We are enclosing a reference table of content for the interim project report.

The key objective of this report is for you to capture how far you have got in completing the internship work against milestones expected to be achieved within a specific duration and seek the mentor’s feedback. Depending on the internship project and your progress (IT/Non-IT, Technical/Business Domain), you may choose to include or exclude or rename sections or leave some sections blank from the table of content mentioned below. You can also add additional sections. You can refer the project presentation to view the milestones related to your internship project. Please populate milestone# (1 / 2 / 3) and the milestone description in the interim project report based on the milestone for which you are submitting the interim project report.

You can refer the project presentation to view the milestones related to your internship project.

|  |  |
| --- | --- |
| Internship Project Title | Rank features of a smartphone-build a python application to classify and rank dataset |
| Name of the Company | TCS iON |
| Name of the Industry Mentor | Debashis Roy |
| Name of the Institute | ICT Academy of Kerala |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Start Date | End Date | | Total Effort (hrs.) | | Project Environment | Tools used |
| 11/11/22 | 15/11/22 | | 12 | | python | Jupyter notebook |
| Milestone # | 1 | Milestone: | | Find out dataset, clean & sanitize it and classify it into hardware and software. | | |

**TABLE OF CONTENT**

* Acknowledgements
* Objective
* Introduction / Description of Internship
* Internship Activities
* Approach / Methodology
* Assumptions
* Exceptions / Exclusions
* Charts, Table, Diagrams
* Algorithms
* Challenges & Opportunities
* Risk Vs Reward
* Reflections on the Internship
* Recommendations
* Outcome / Conclusion
* Enhancement Scope
* Link to code and executable file
* Research questions and responses

**Acknowledgements**

First I would like to thank ICT Academy of Kerala, for giving me the opportunity to do an internship from TATA Consultancy Services.

I also would like to thank to the organization for giving an opportunity to do an internship of one month.

It is indeed with a great sense of pleasure and immense sense of gratitude that I acknowledge the help of the industry mentor Mr. Debashis Roy.

**Objective**

The objective of this project is to build a python application that ranks the features of a smartphone based on the requests received from various users.

**Introduction**

Smart phone is a mobile phone which offers advanced technologies with functionality similar as a personal computer. While offering a standardized platform for application developers a smart phone performs as complete operating system software. Secondly, there are also very advanced features in smart phones such as internet, instant messenger and e-mail and also built-in keyboard are very typical. Because of these reasons we can say a smart phone a miniature computer with the similarities of a simple phone.

Analyst house Gartner gives the definition of a smart phone as: “A large-screen, data-centric, handheld device designed to offer complete phone functions whilst simultaneously functioning as a personal digital assistant (PDA).” (Analyst House Gartner: 2009)

With the growing speed of technological advancement, Smart phones are now an essential part of our daily life routine. When we go for our convenience we also look for those devices which contain multiple features such as office work, mobility, networking and entertainment. As the world is getting advance our needs become sophisticated. Where we need quality, effectiveness and performance we also ask for these all in one single pocket device so we can take that to anywhere with us.

The main aim of the project is to build a python application to classify and rank dataset i.e. to rank features of a smartphone.  The dataset is obtained from the digital discussion room of TCSiON learning platform. This dataset contains information on battery power, clock speed, dual sim, four g, int memory, touch screen, Wi-Fi, price range, ram, talk time etc.

**Internship Activities**

**Milestone #1**: Create a dataset, clean the dataset and also sanitize it and classify the dataset into hardware and software request.

**Day 1 to Day 5**: Watch the vedios in TCS learning hub . Understand the topic , go through different reference link that are pinned in the learning hub page. Downloaded the dataset from digital discussion room and started working on it. Call the dataset into python environment , checked for the basic details ie shape,column names, data types ,etc.Next done the pre-processing ie checked for the null values,outliers . There was no null values present in the test and train data. There was some outliers in two columns but it is not considered .

Done the Exploratory data analysis. Identified the target column(price\_range) and plot graph for it. Plotted univariate graphs for other features in the dataset. Take cor-relation heatmap and analyse the relation between different features and drawn bivariate and multivariate graphs for different features.Next classify the data into hardware and software and then concate the data.

**Methodology**

**Source of data:** The dataset was downloaded from the digital discussion room of TCS iON Remote Internship 125.

**Columns in the dataset**: battery\_power: Total energy a battery can store in one time measured in mAh

blue: Has bluetooth or not

clock\_speed: speed at which microprocessor executes instructions

dual\_sim: Has dual sim support or not

fc: Front Camera mega pixels

four\_g: Has 4G or not

int\_memory: Internal Memory in Gigabytes

m\_dep: Mobile Depth in cm

mobile\_wt: Weight of mobile phone

n\_cores: Number of cores of processor

pc: Primary Camera mega pixels

px\_height: Pixel Resolution Height

px\_width: Pixel Resolution Width

ram: Random Access Memory in Mega Bytes

sc\_h: Screen Height of mobile in cm

sc\_w: Screen Width of mobile in cm

talk\_time: longest time that a single battery charge will last when you are

three\_g: Has 3G or not

touch\_screen: Has touch screen or not

wifi: Has wifi or not

price\_range: This is the target variable with value of 0(low cost), 1(medium cost), 2(high cost) and 3(very high cost).

**Data Analysis:** it is the technique to collect, transform, and organize data to make future predictions, and make informed data-driven decisions. It also helps to find possible solutions for a business problem. There are six steps for Data Analysis. They are:

* Ask or Specify Data Requirements
* Prepare or Collect Data
* Clean and Process
* Analyze
* Share
* Act or Report

Call the data into python environment and take the basic details about the data. Then done the pre-processing, checked for the null values but there was no null values present in the dataset. Checked for the outliers, there was few outliers in the columns ‘fc’ & ‘px\_height’ but is not considered as an outlier. Then done the Exploratory Data Visualization.

**Exploratory Data Analysis:** EDA is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods. It helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions.

**Univariate Visualization:** Univariate data visualization plots help us comprehend the enumerative properties as well as a descriptive summary of the particular data variable. These plots help in understanding the location/position of observations in the data variable, its distribution, and dispersion.

**Bivariate Visualization:** Bivariate analysis is an analysis that is performed to determine the relationship between 2 variables. In this analysis, two measurements were made for each observation.

**Multivariate Visualization:** Multivariate data visualization involves visualizing more than one data value in a single renderer.

**Count plot:** Show the counts of observations in each categorical bin using bars. A count plot can be thought of as a histogram across a categorical, instead of quantitative, variable. The basic API and options are identical to those for barplot() , so you can compare counts across nested variables.

**Pie plot:** A pie chart is a circular statistical graphic, which is divided into slices to illustrate numerical proportion. In a pie chart, the arc length of each slice is proportional to the quantity it represents.

**Dist plot:** A Distplot or distribution plot, depicts the variation in the data distribution. Seaborn Distplot represents the overall distribution of continuous data variables. The Seaborn module along with the Matplotlib module is used to depict the distplot with different variations in it.

**Correlation heat map:**  Correlation heatmaps are a type of plot that visualize the strength of relationships between numerical variables. Correlation plots are used to understand which variables are related to each other and the strength of this relationship.

**Bar plot:** A bar chart or bar graph is a chart or graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent. The bars can be plotted vertically or horizontally. A vertical bar chart is sometimes called a column chart.

**Swarm plot:** A swarm plot is a type of scatter plot that is used for representing categoricalvalues. It is very similar to the strip plot, but it avoids the overlapping of points. We can use the seaborn. swarmplot() to create such graphs. It is not advisable to use this type of graph when the sample size is large.

**Reg plot:** This method is used to plot data and a linear regression model fit. There are a number of mutually exclusive options for estimating the regression model.

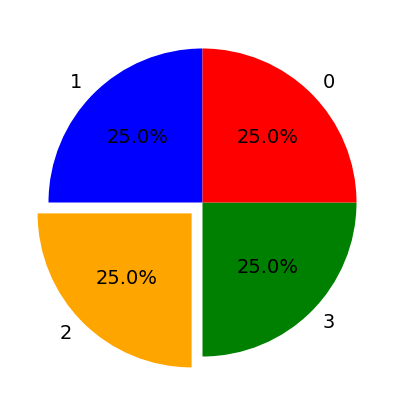
**Joint plot:** A Jointplot comprises three plots. Out of the three, one plot displays a bivariate graph which shows how the dependent variable(Y) varies with the independent variable(X). Another plot is placed horizontally at the top of the bivariate graph and it shows the distribution of the independent variable(X).

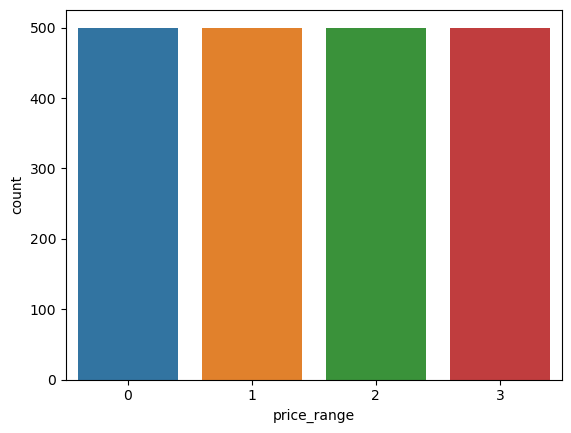
**Box plot:** In descriptive statistics, a box plot or boxplot is a method for graphically demonstrating the locality, spread and skewness groups of numerical data through their quartiles.

**Hardware and Software Classification:** train data is classified into hardware and software data.

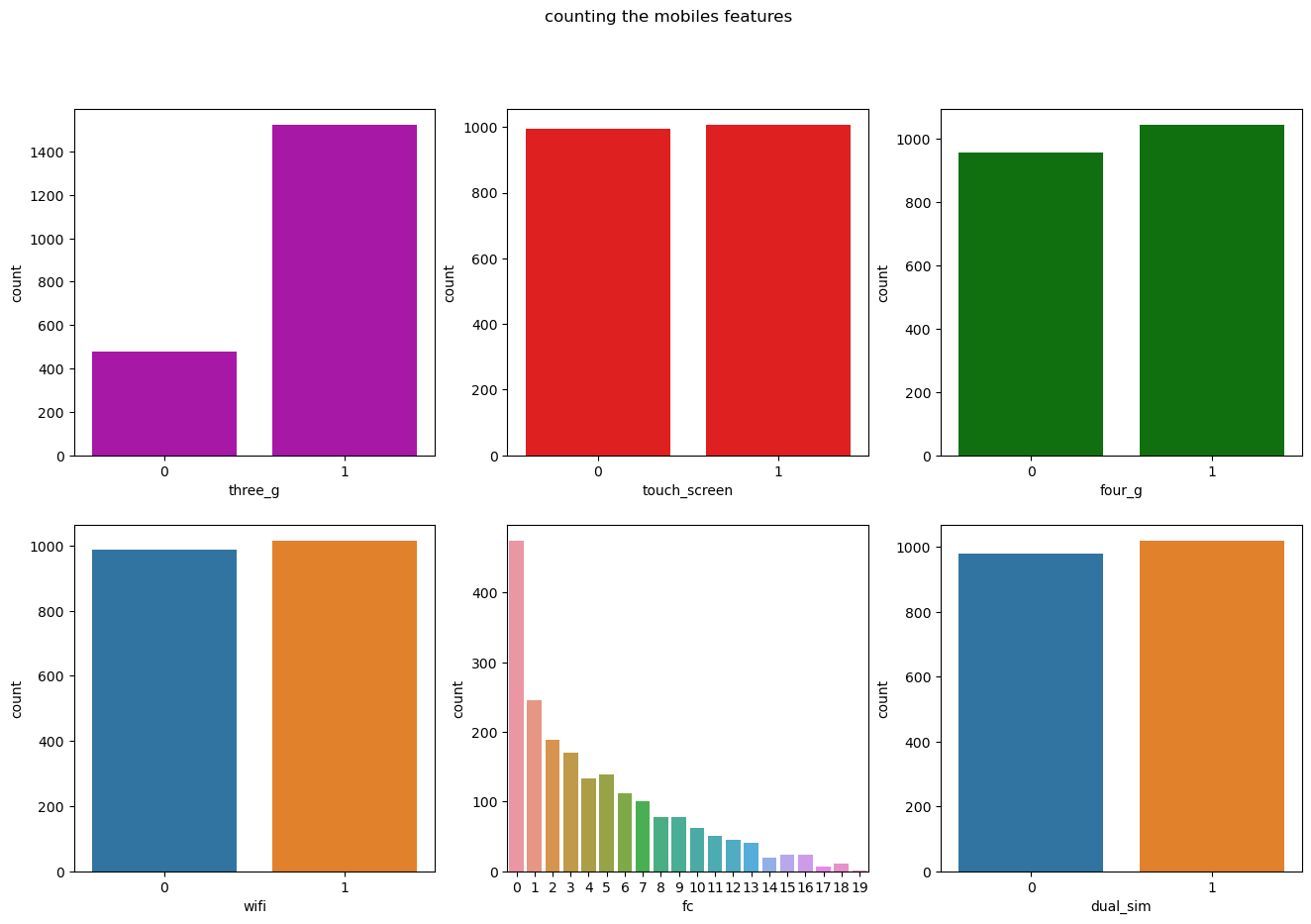
**Graphs**

**Target column:** ‘price\_range’ is the dependent column in the dataset. When plotted a count plot and pie chat , it’s seen that the data is balanced.





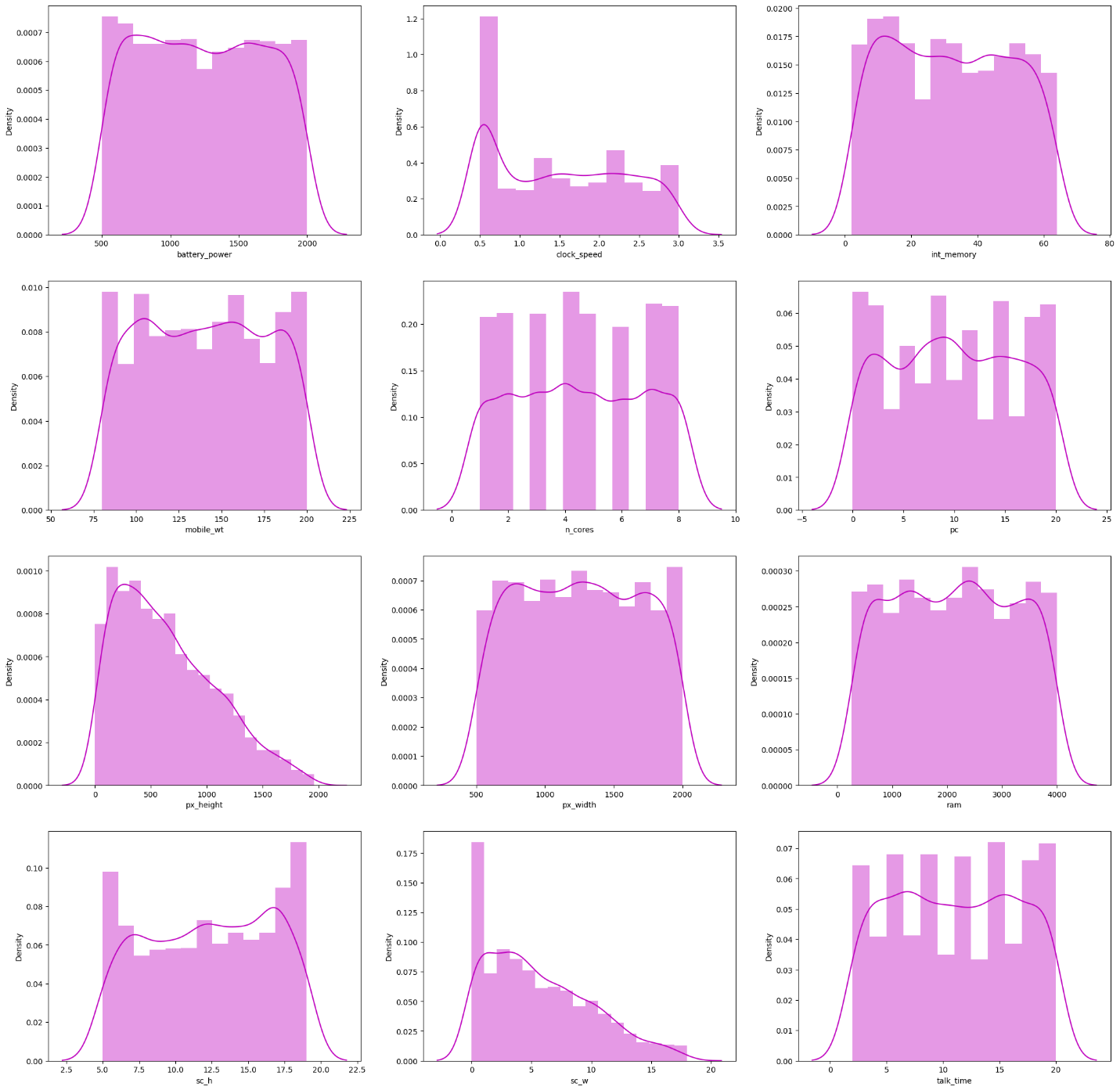
**Univariate visualization:** using countplot , drawn graphs of some features like 3G , 4G, fc, dual sim , wifi , touch screen.



**Dist plot:** drawn the dist plot of some features like "battery\_power","clock\_speed","int\_memory","mobile\_wt","n\_cores","pc","px\_height","px\_width",

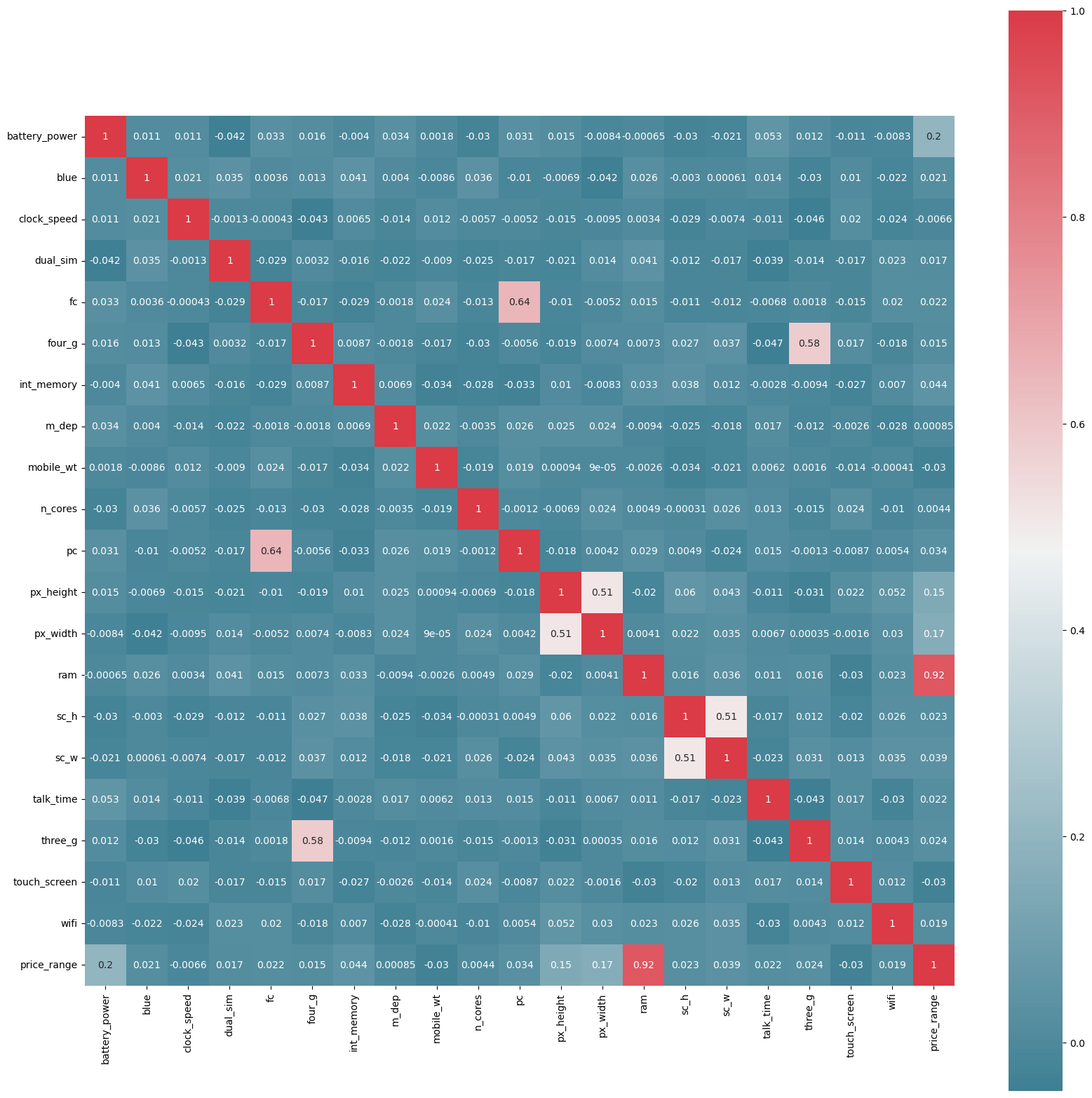
"ram","sc\_h","sc\_w","talk\_time","price\_range".

This shows the distribution of continues variables present in the dataset.



**Correlation heatmap**: from the heat map we can understand that there is an high positive correlation between ram and price range of smartphone. There is a correlation between fc and pc and also a correlation between px\_height and px\_width and between three g and four g.

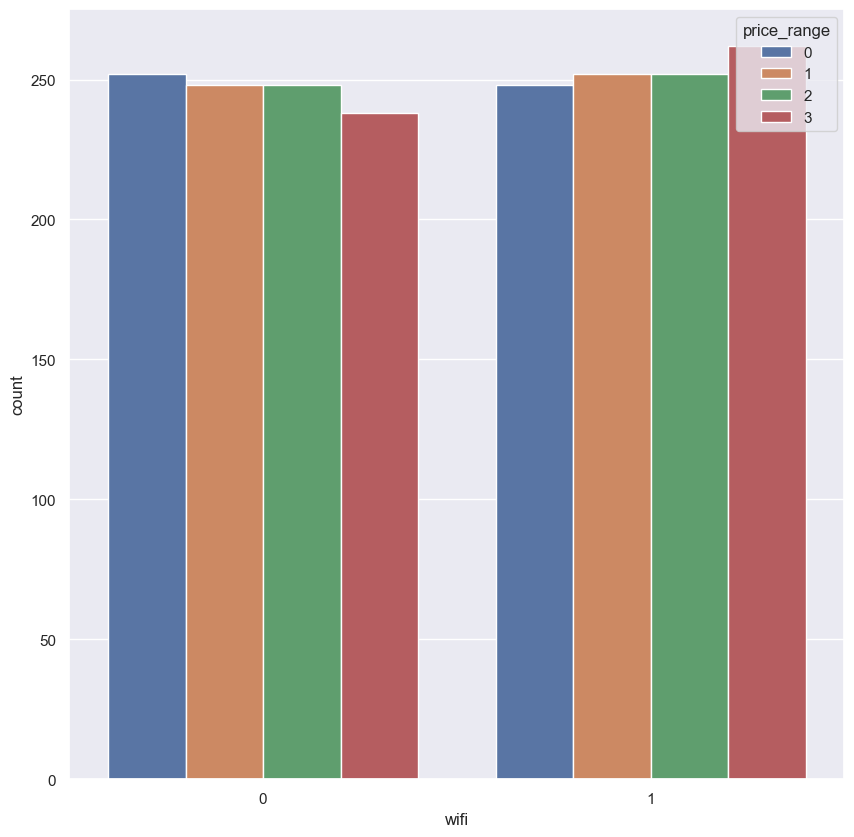
By observing the correlation we have drawn bivariate and multivariate graphs.

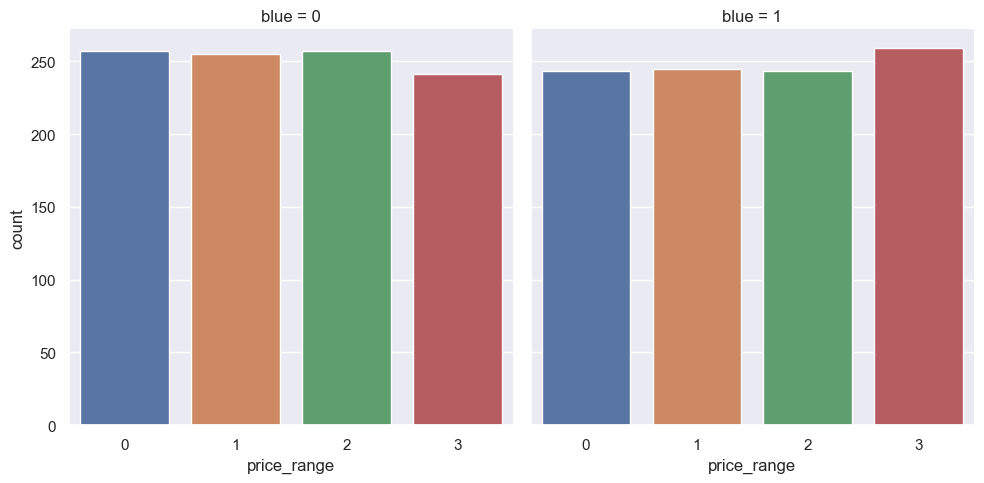


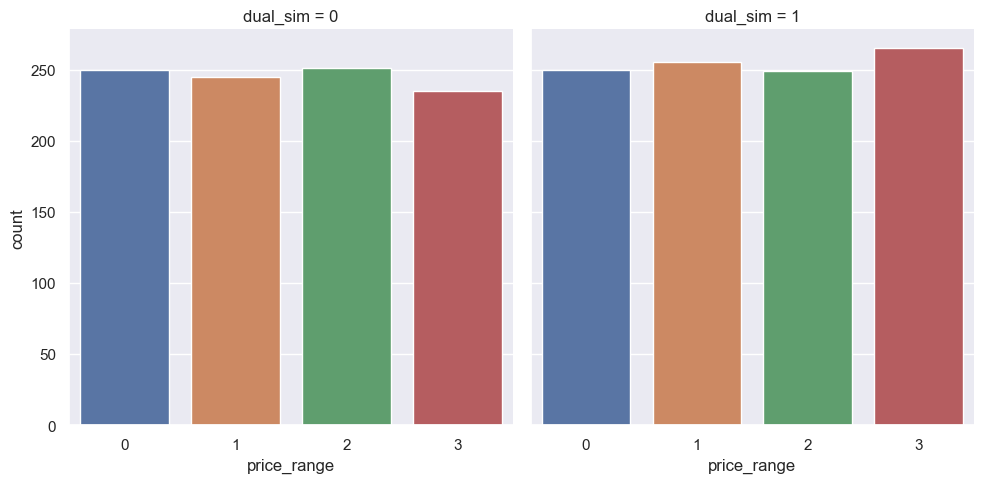
**Bivariate visualization:** since there is an high correlation between ram and price range, we have drawn a swarm plot .The smartphone having high ram capacity has high price range.



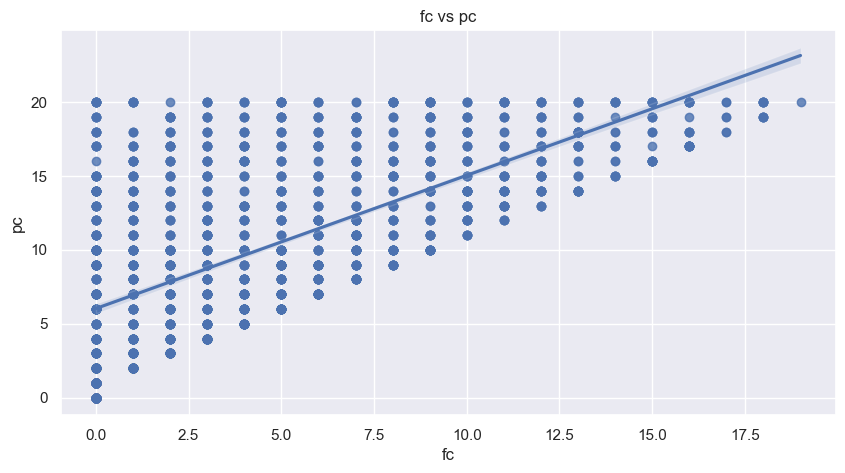
Drawn different bivariate graphs of dependent variable vs independent variable ie price range vs wifi, blue, dual sim.



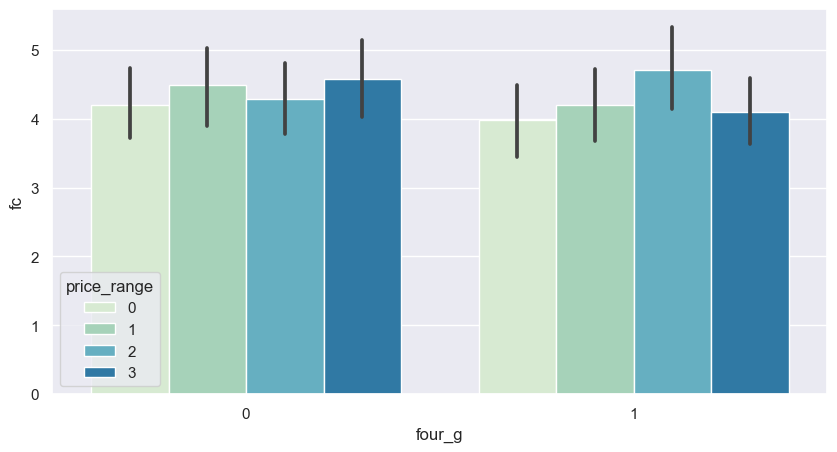


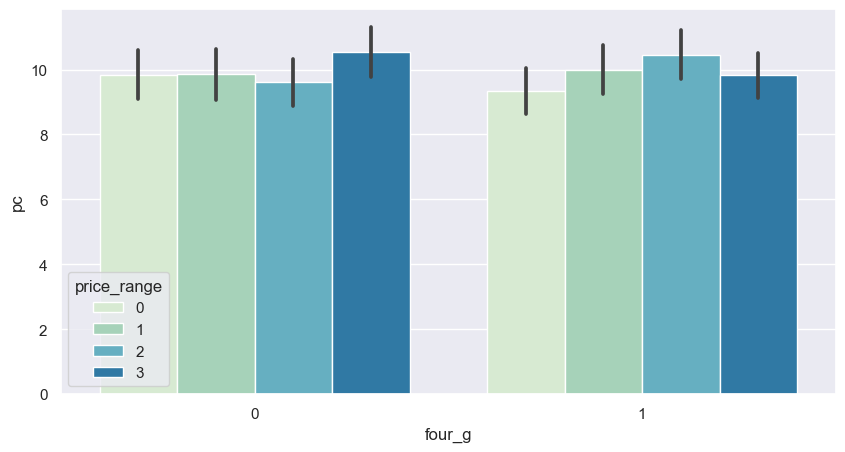


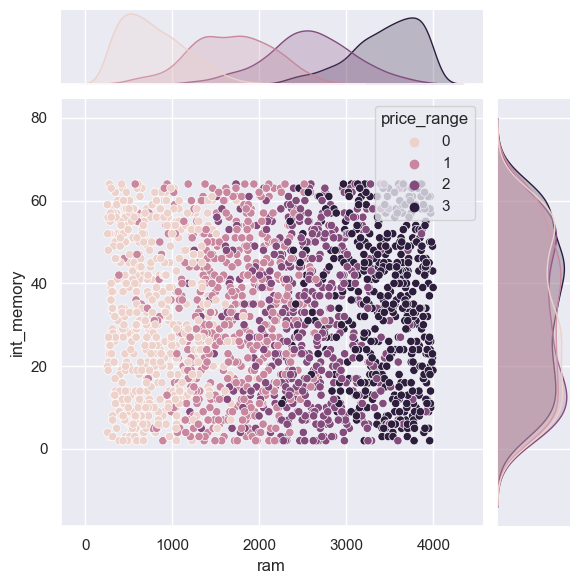
Drawn regression plot between fc and pc. It shows that when fc increases pc also increases.

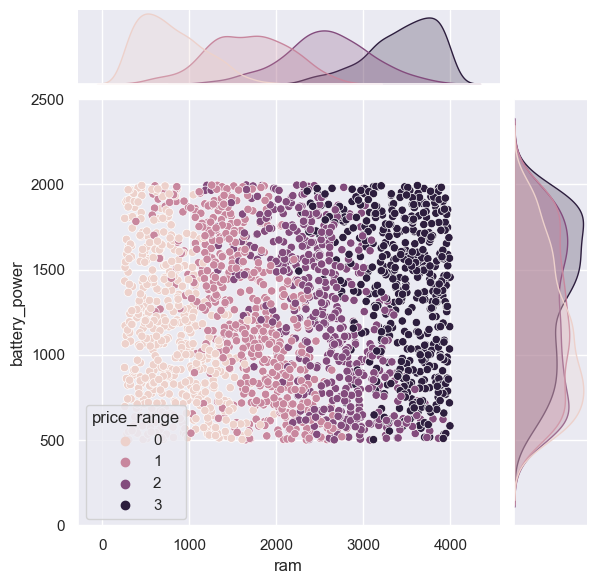


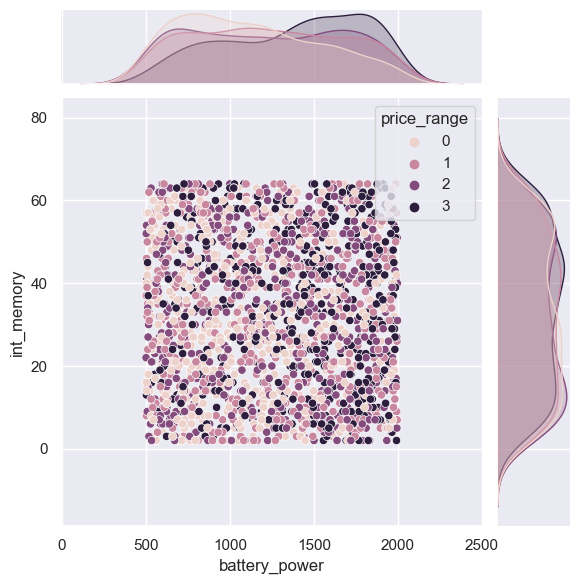
**Multivariate visualization:** drawn different multivariate graph between price range,fc,4G,pc. And between ram , internal memory,price range,battery power,dual sim ,talk time,n cores and clock speed.

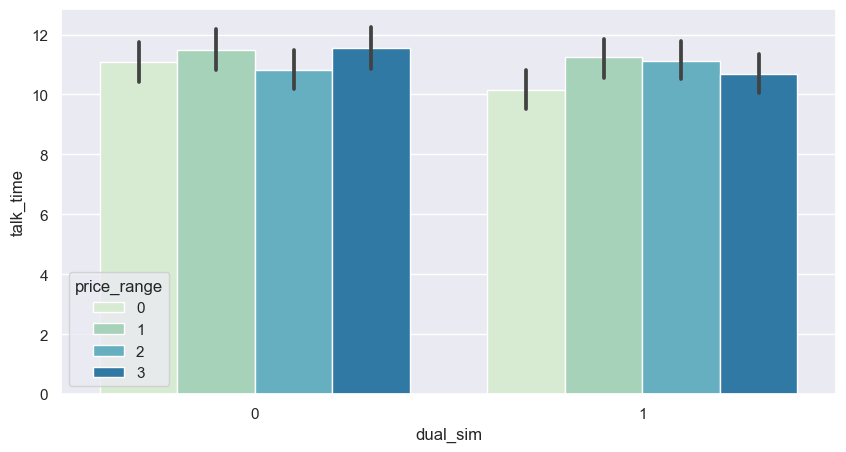


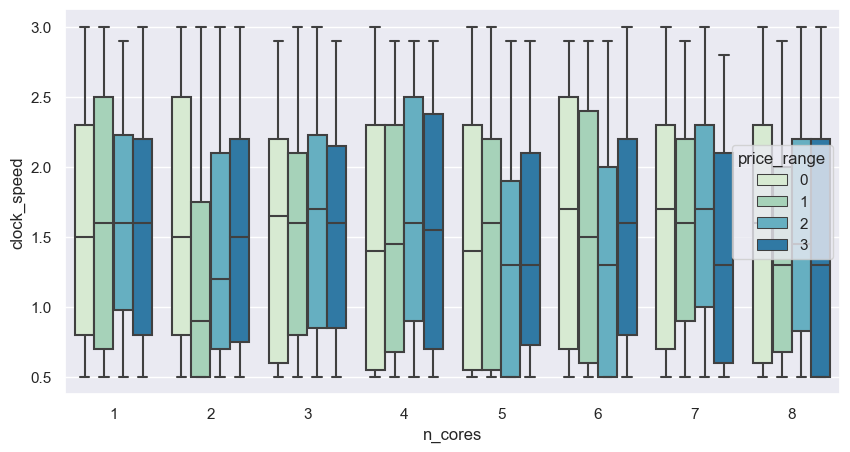












Classify the data into hardware and software and then by concat () function concatenated the data.

Github link: https://github.com/LINTA-STEPHEN/TCS\_iON\_RIO\_125/blob/main/TCS%20Internship.ipynb