

# Capstone Project

## Mobile Price Range Prediction

Kiran Mamtani

# Content

- Summary of Data
- Exploratory Data Analysis
- Feature Engineering
- Machine Learning Models
- Conclusion
- Challenges

# Summary of Data

**About Data** - The dataset is based on sales data of mobile phones and factors which drive the prices

## Size of Data

**Rows** - 2000

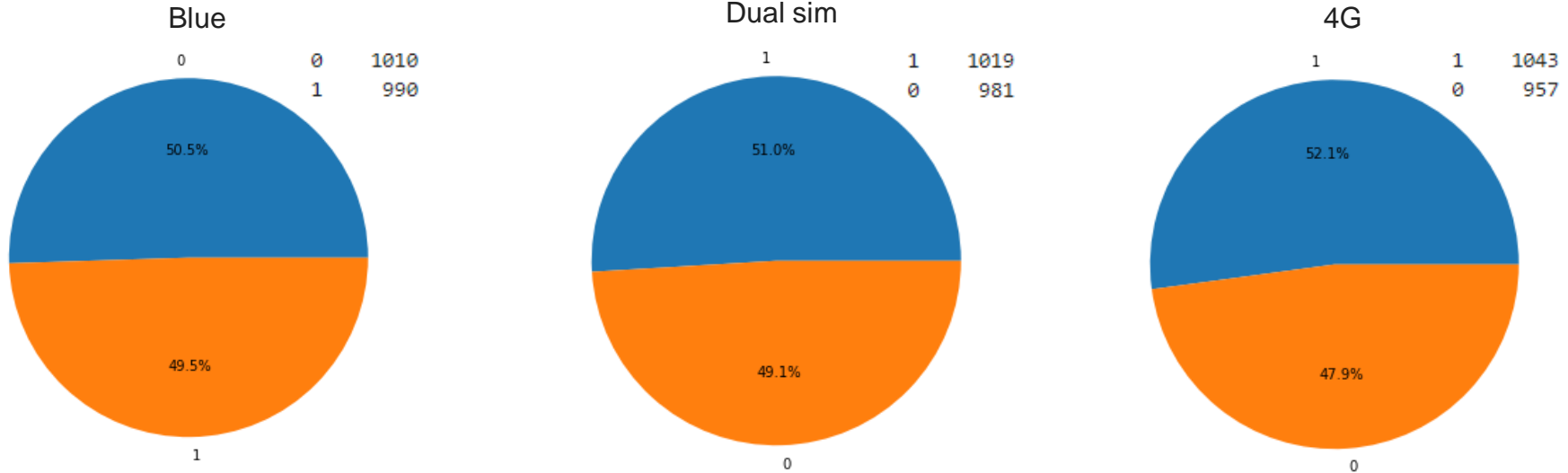
**Columns** - 21



Source: <https://www.freepik.com/>

# Exploratory Data Analysis

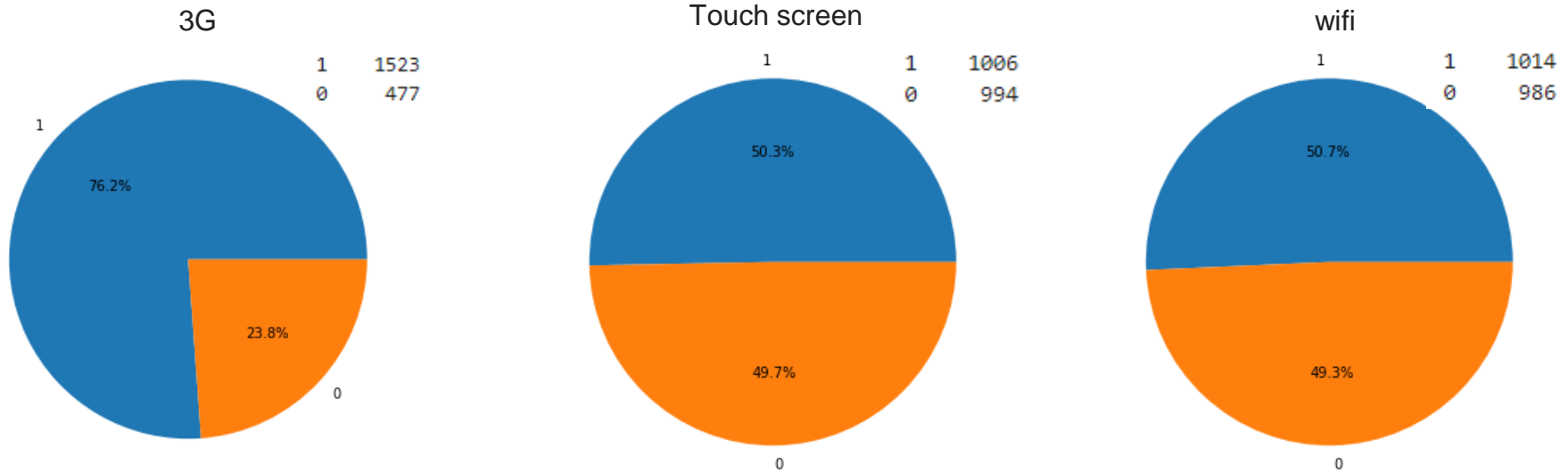
## Univariate Analysis



In all the above graphs, the values are almost equal. So we can say that Bluetooth, Dual sim, 4G is distributed almost equal in our data set

# Exploratory Data Analysis

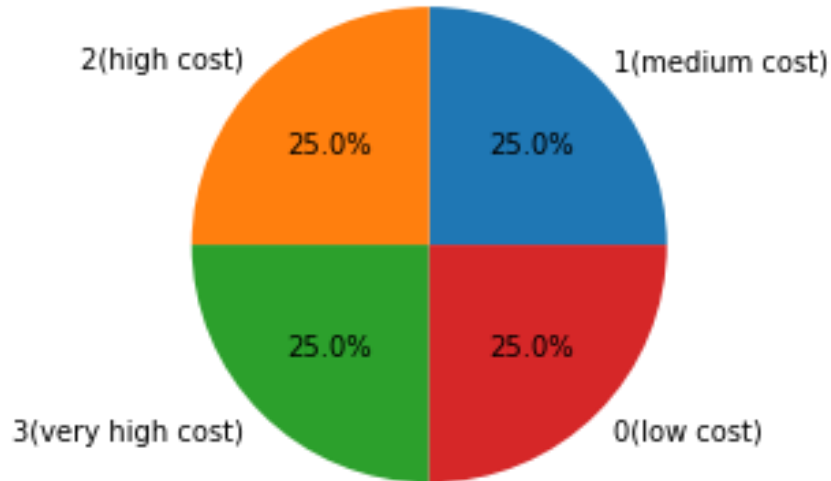
## Univariate Analysis



Touch screen & WIFI is distributed in almost equal values in our dataset whereas 76% mobile phone has 3G technology and 24% does not have 3G. We can assume that the remaining 24% who don't have 3G may have 2G

# Exploratory Data Analysis

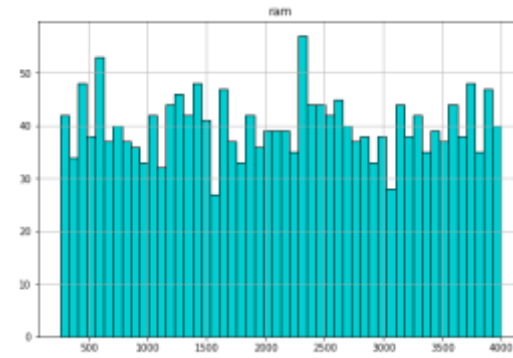
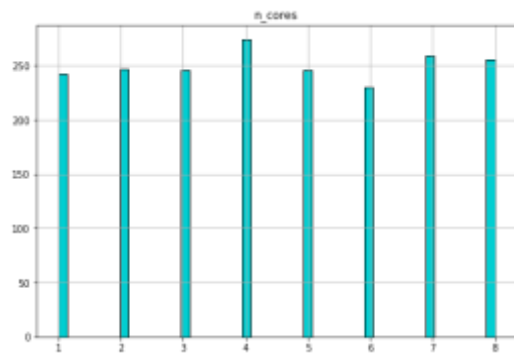
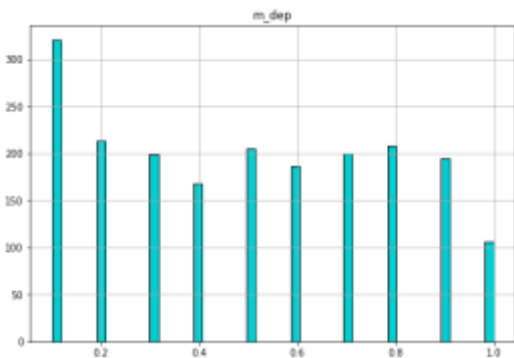
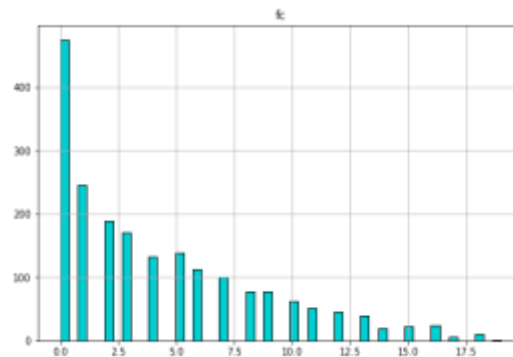
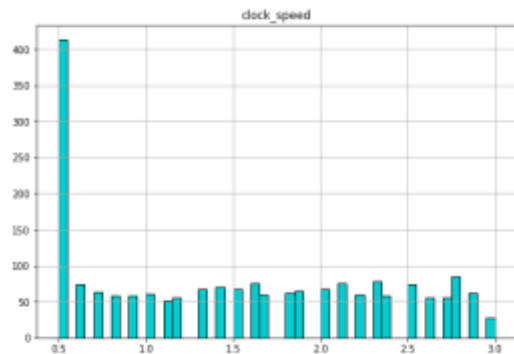
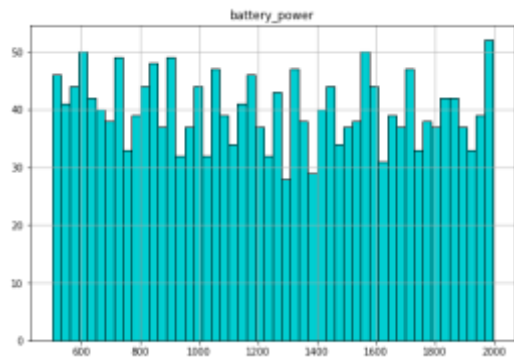
## Univariate Analysis



As seen in graph, our target variables i.e. price range is distributed equally. It means in our data, price range column as equal number of values (0, 1, 2 & 3)

# Exploratory Data Analysis

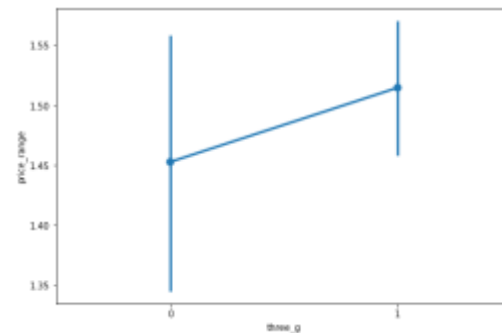
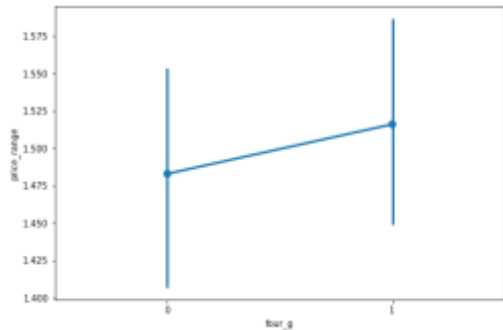
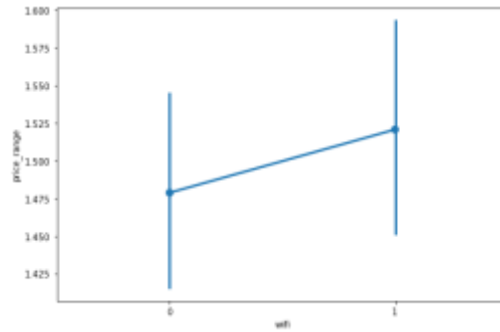
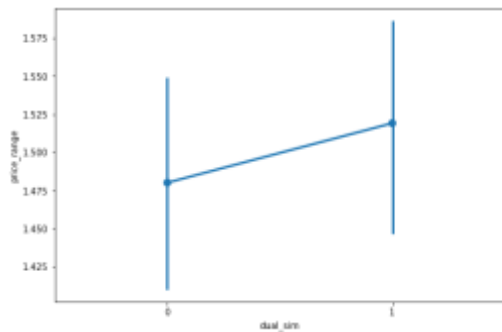
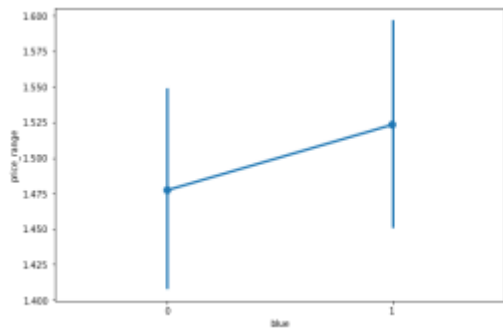
## Univariate Analysis



# Exploratory Data Analysis

## Bivariate Analysis

Let's visualize all the features with our target feature price range



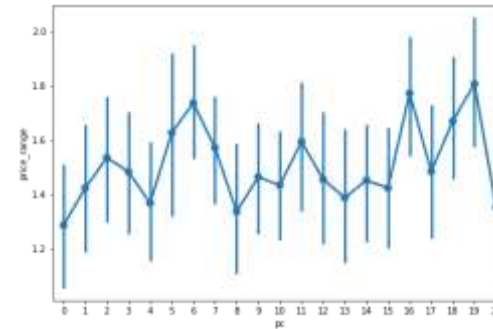
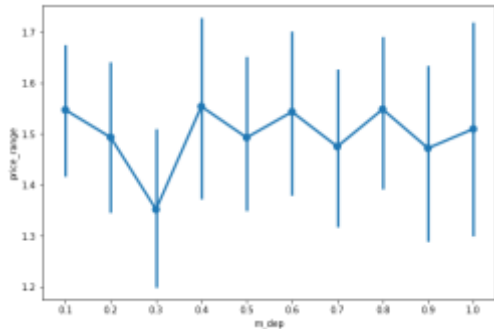
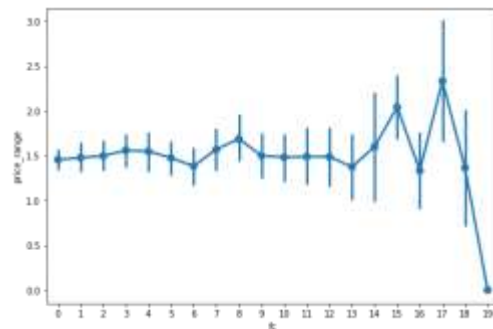
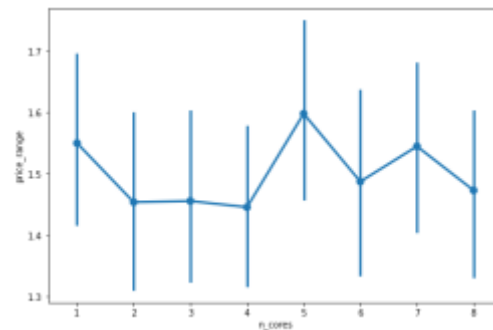
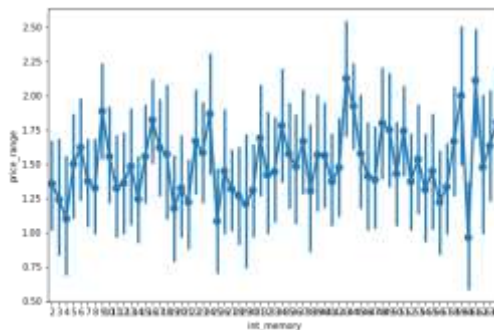
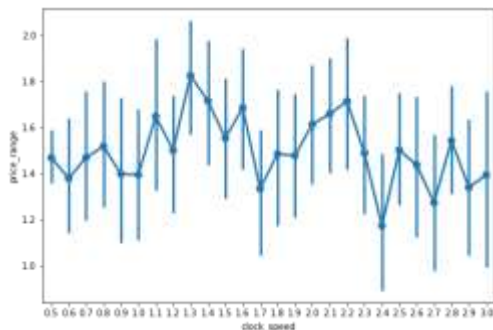
All these 5 graphs are almost same like price range is directly proportional to these features



# Exploratory Data Analysis

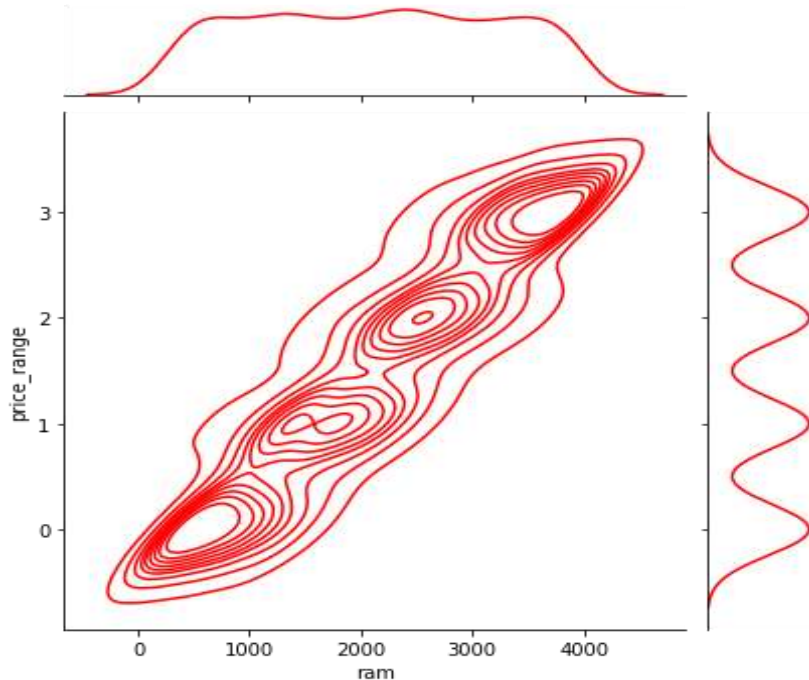
## Bivariate Analysis

Here the price range is not changing vastly



# Exploratory Data Analysis

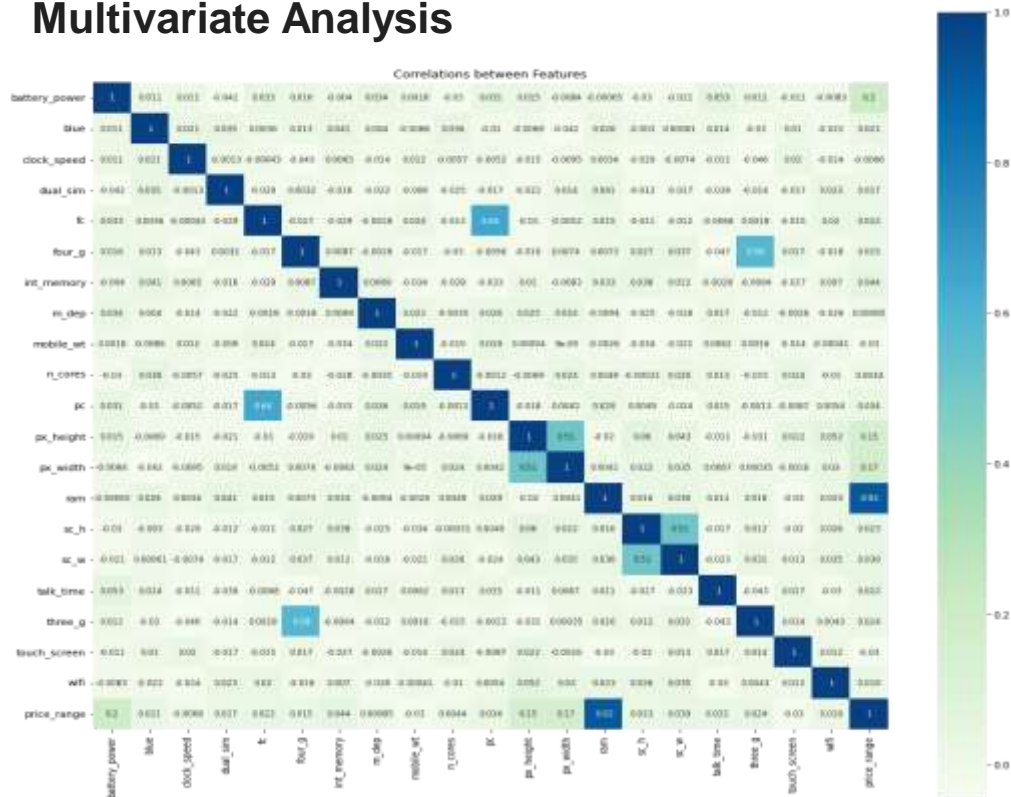
## Bivariate Analysis



The line trend is increasing order. When ram is increasing the price range is also increasing. We can say that both are directly related to each other

# Exploratory Data Analysis

## Multivariate Analysis



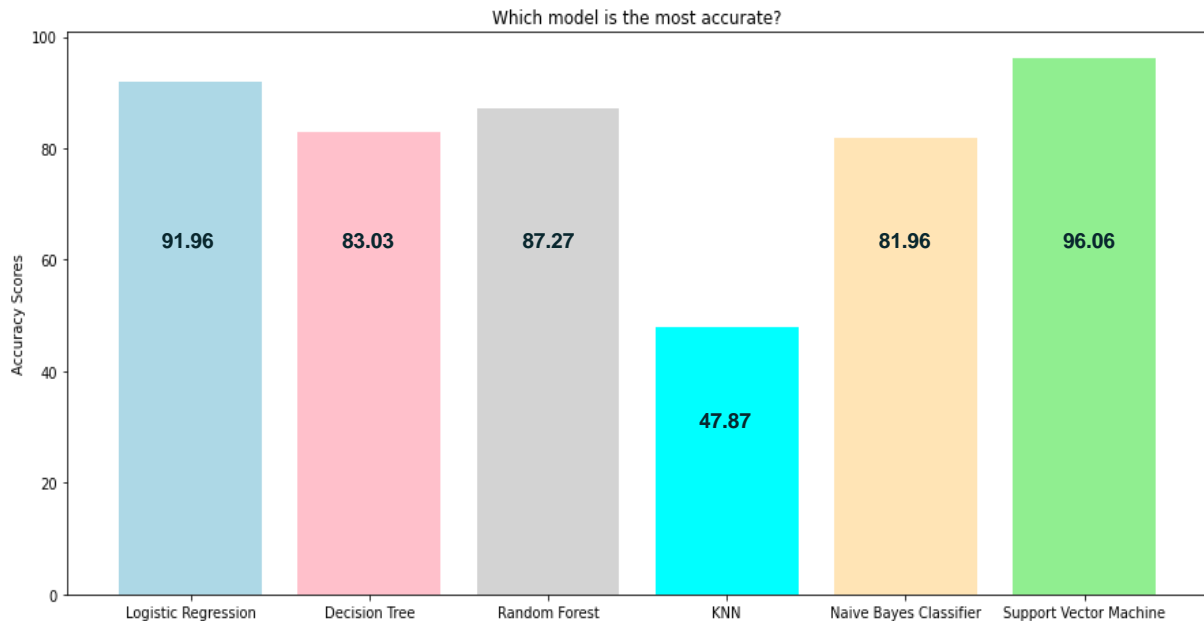
We can see the highest correlation between ram and price range. The two other correlations are between 3G and 4G, pc & fc

# Machine Learning Models

## Applied 6 classifications models

1. Logistic Regression
2. Decision Tree Classifier
3. Random Forest Classifier
4. K-Nearest Neighbours (KNN)
5. Naive Bayes Classifier
6. Support Vector Machines

# Machine Learning Models



As we can see support vector machine model is best for our dataset with 96% of accuracy

# Support Vector Machines (SVM)

Applied Grid Search to find out the best hyperparameter

Result of test dataset					Result of train dataset				
	precision	recall	f1-score	support		precision	recall	f1-score	support
0	0.97	0.99	0.98	166	0	0.99	0.99	0.99	334
1	0.94	0.93	0.94	147	1	0.97	0.98	0.97	353
2	0.95	0.94	0.94	165	2	0.98	0.97	0.97	335
3	0.98	0.98	0.98	182	3	0.99	0.98	0.99	318
accuracy			0.96	660	accuracy			0.98	1340
macro avg	0.96	0.96	0.96	660	macro avg	0.98	0.98	0.98	1340
weighted avg	0.96	0.96	0.96	660	weighted avg	0.98	0.98	0.98	1340

Test and train, in both we are getting almost similar result and SVM has highest accuracy

# Conclusion

- All the important libraries was imported
- Data set was imported and understood all the features
- Exploratory Data Analysis was done to investigate dataset patterns and outliers
- From EDA, we got to know that there is zero null values
- Visualizes each feature in Univariate analysis.
- In Bivariate analysis, compared target feature (price\_range) with all the features and got to know that RAM is directly proportional to price range. As RAM increases, price range also increases
- Also formed multivariate graphs. In correlation of every columns with each other, we can say that RAM is highly corelated to each other
- So our data is linear

# Conclusion

- Applied different classification models like 'Logistic Regression', 'Decision Tree', 'Random Forest', 'KNN', 'Naive Bayes Classifier' and 'Support Vector Machine'
- The accuracy score of Logistic Regression is 91.96%
- The accuracy score of Decision Tree is 83.03%%
- The accuracy score of Logistic Regression is 87.27%
- The accuracy score of Logistic Regression is 47.87%
- The accuracy score of Logistic Regression is 81.96%
- The accuracy score of Logistic Regression is 96.06%
- **We conclude that SVM (Support Vector Machine) is best model for our dataset (with the highest accuracy score = 96%)**



# Challenges

- The most difficult thing I faced in this project is to choose the hyperparameter
- Also got afraid of if I am not missing any other classification technique which can give me more better accuracy
- Some of my coding was taking a lot of time to run so with the help of few websites I have revised my code to get implemented faster

**Thank You!**