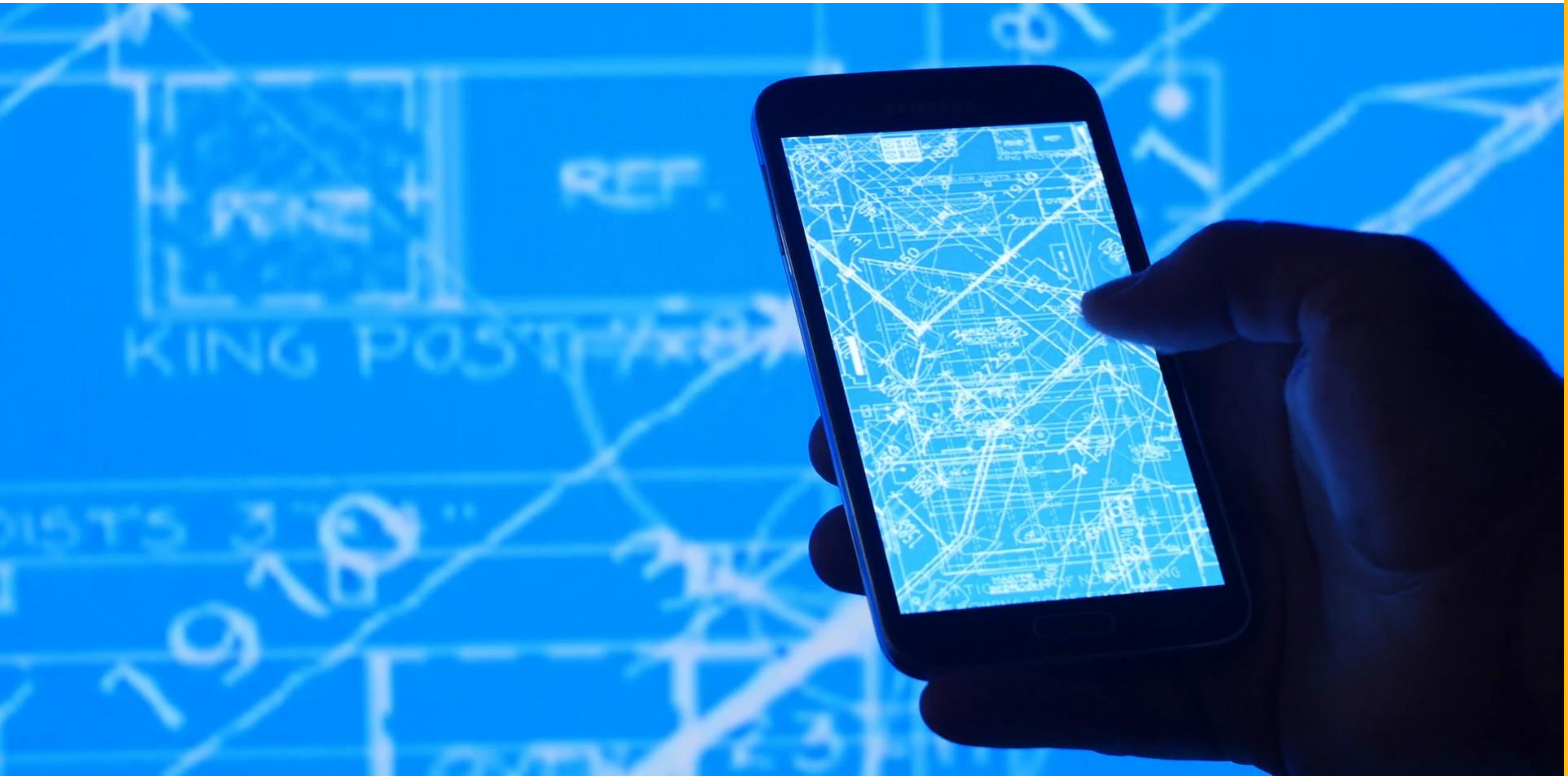
Smartphone Price Prediction



December 2, 2021

Agenda

I. Project Overview

II. Data Collection & Cleaning

III. Analysis & Visualization

IV. Results & Summary

Project Overview

- Develop a robust framework for predicting smartphone prices based on data from past smartphone prices.
- In this presentation, we will include a little explanation about the procedures to get our final dataset first. Because we know those processes could cause potential bias or human error, which may impact our results/observations in an unexpected way.
- Apply different techniques for understanding the relationship between the data and predicting the price of the smartphones
- The more high-quality data you have, the more confidence you can have in your decisions.

Data Collection and Cleaning

Data Collection

- Smartphone data were collected from **GSM Arena** in a scheduled manner
- Data was collected for the OEM : *Samsung, Realme, Huawei, Motorola*
- Created python project to **scrape** smartphone data from GSM Arena

Data Cleaning

- Major challenge in this problem was the data type. Every column was in their textual representation which must be cleansed for the prediction
- Using certain features were straightforward, needed simple parsing. However, there were certain features which needed aggregation of other features and similar data from other websites
- Examples: Number of bands, CPU score was a result of aggregation and mapping from websites like TechCenturion.

Data Collection and Cleaning

CPU Score & GPU Score

- **Centurion Mark** which is one of the industry-leading benchmarking techniques to evaluate the performance of a processor has been used as a feature in place of the CPU processor name
- Used fuzzy logic to map existing CPU and GPU name to the closest GPU name and the corresponding Centurian Mark (score) as illustrated in the table on right.

GPU	GPU_mapped
Mali-G57 MC2	Mali-G57 MC5
Adreno 642L	Adreno 640
Mali-G52 MC2	Mali-G52 MC2
Mali-G57 MC3	Mali-G57 MC3
PowerVR GE8320	PowerVR GE8320
Adreno 642L	Adreno 640
Adreno 660	Adreno 660

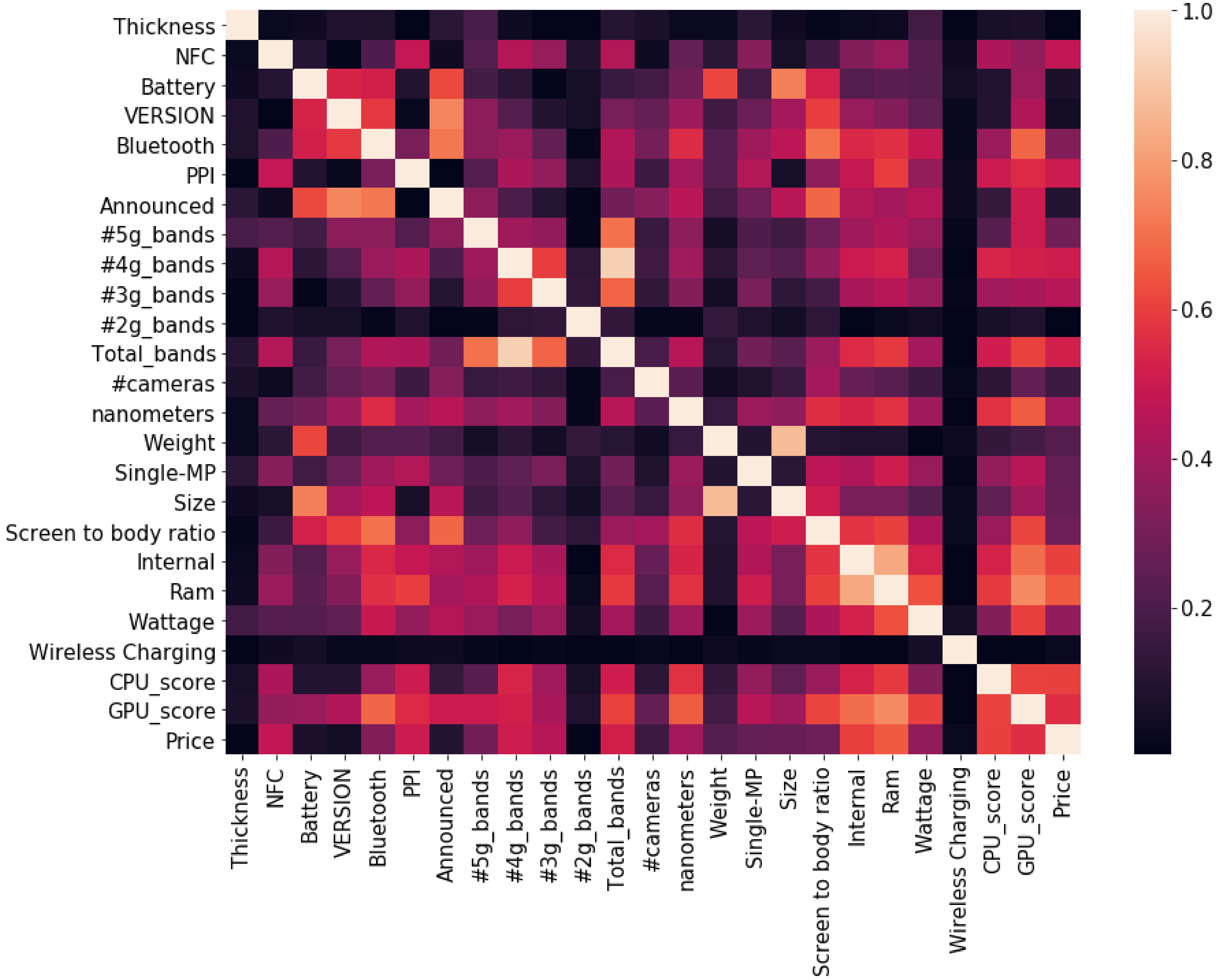
Other Features

- **Number of bands -** Extracted the number of frequency supported in each band and constructed the total no of bands
- **Maximum no of cores, Clock Speed & Frequency** – From the CPU description, it was possible to calculate the above features
  Ex: Octa-core (2x2.2 GHz Cortex-A76 & 6x2.0 GHz Cortex-A55)

Analysis

Correlation Heatmap

Average Correlation between features is **0.26** (absolute value).
So, feature reduction was not required

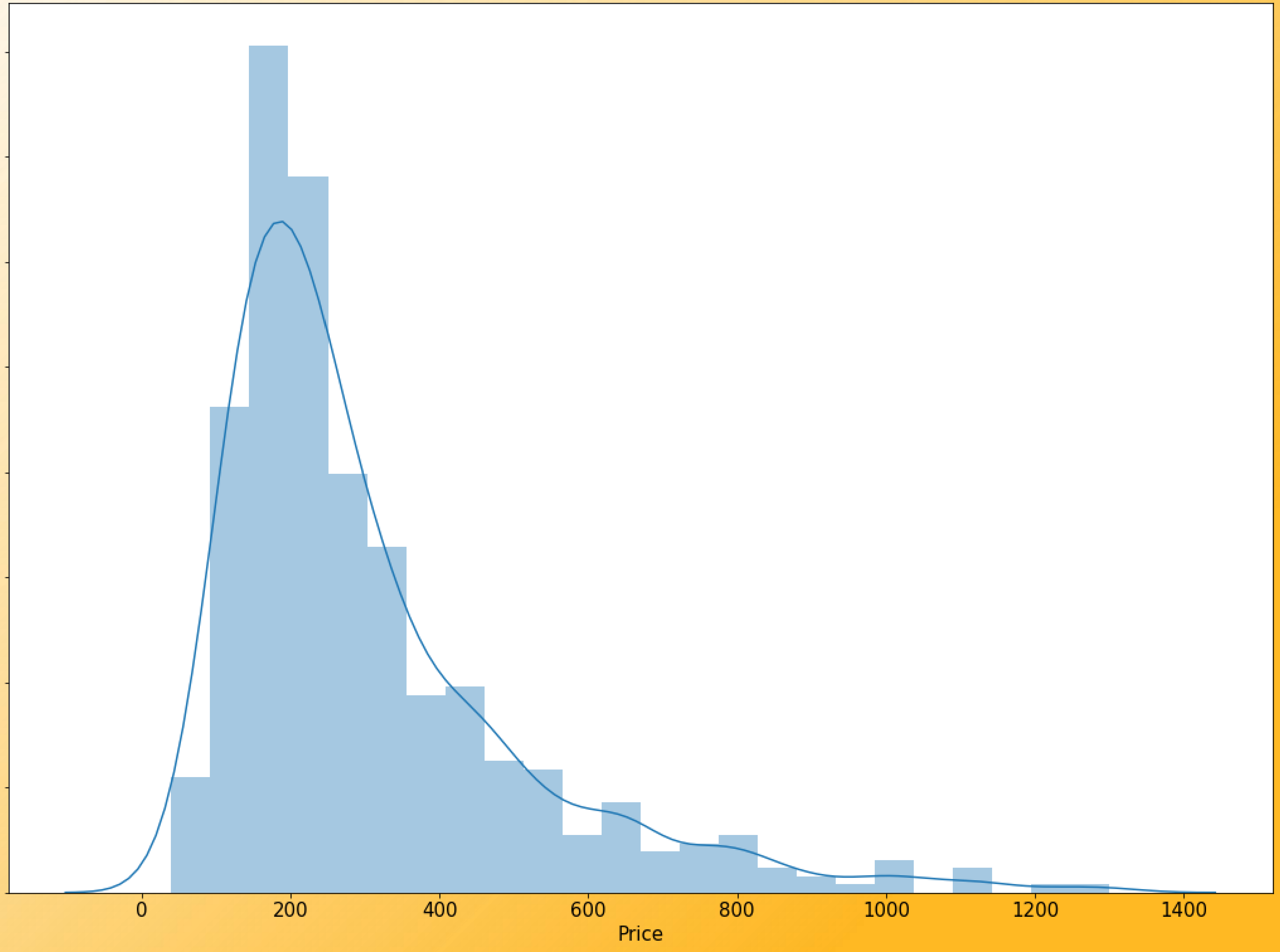


Analysis

Price Distribution

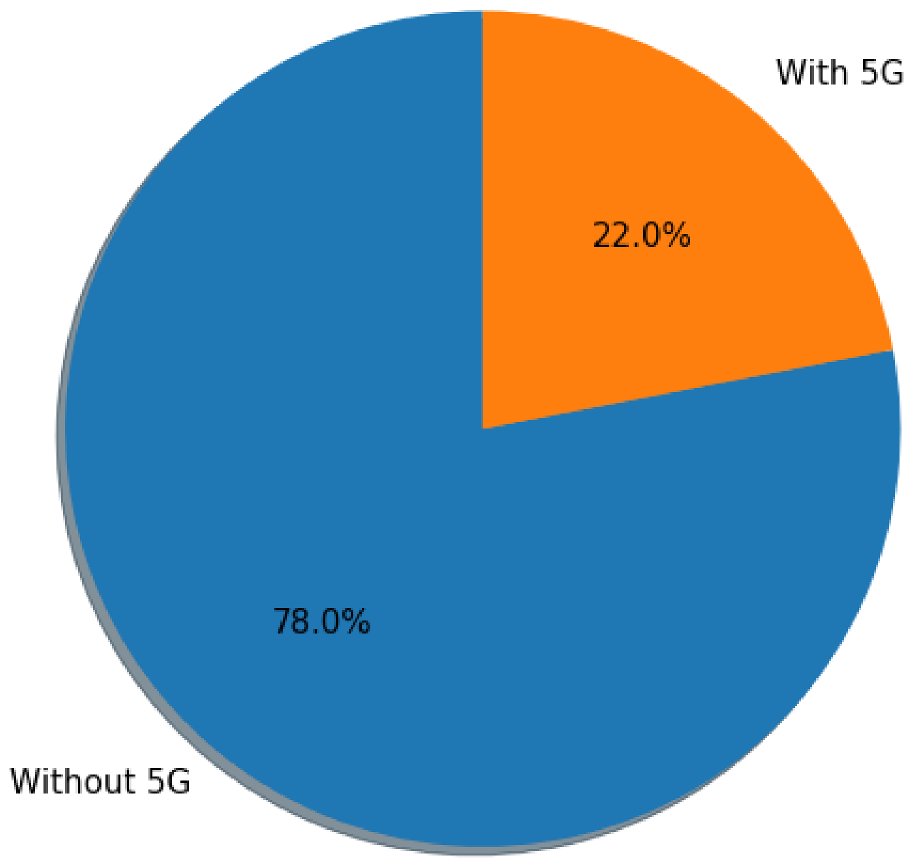
Price Bucket	Min	Median	Max
Base	$40	$146	$198
Low	$202	$258	$350
Mid range	$350	$460	$672
Flagship	$706	$850	$1,300

Dataset consists of Android phones from Samsung, Huawei, Motorola, and Realme with an average price of ~300$.

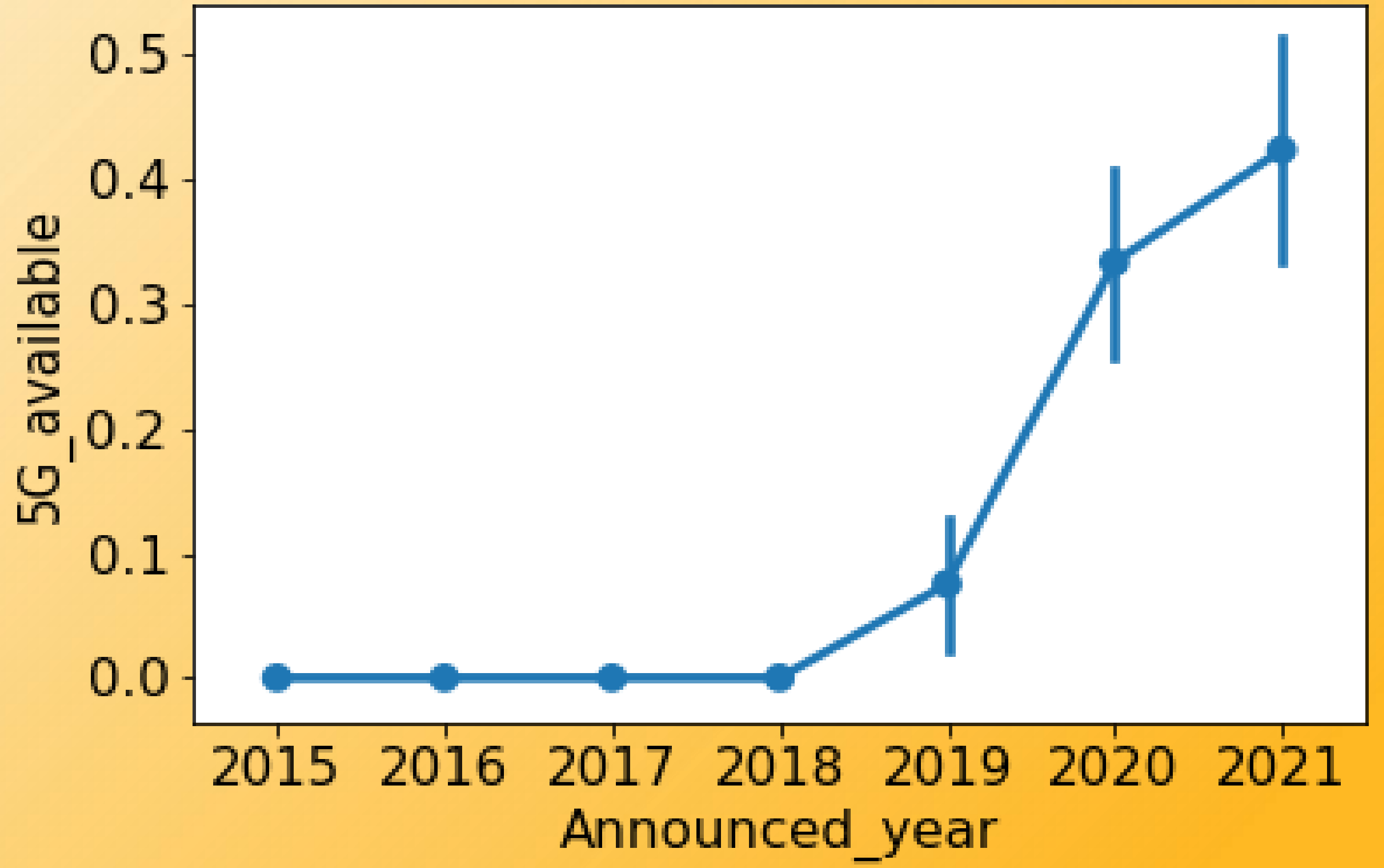


Analysis

5G Adoption



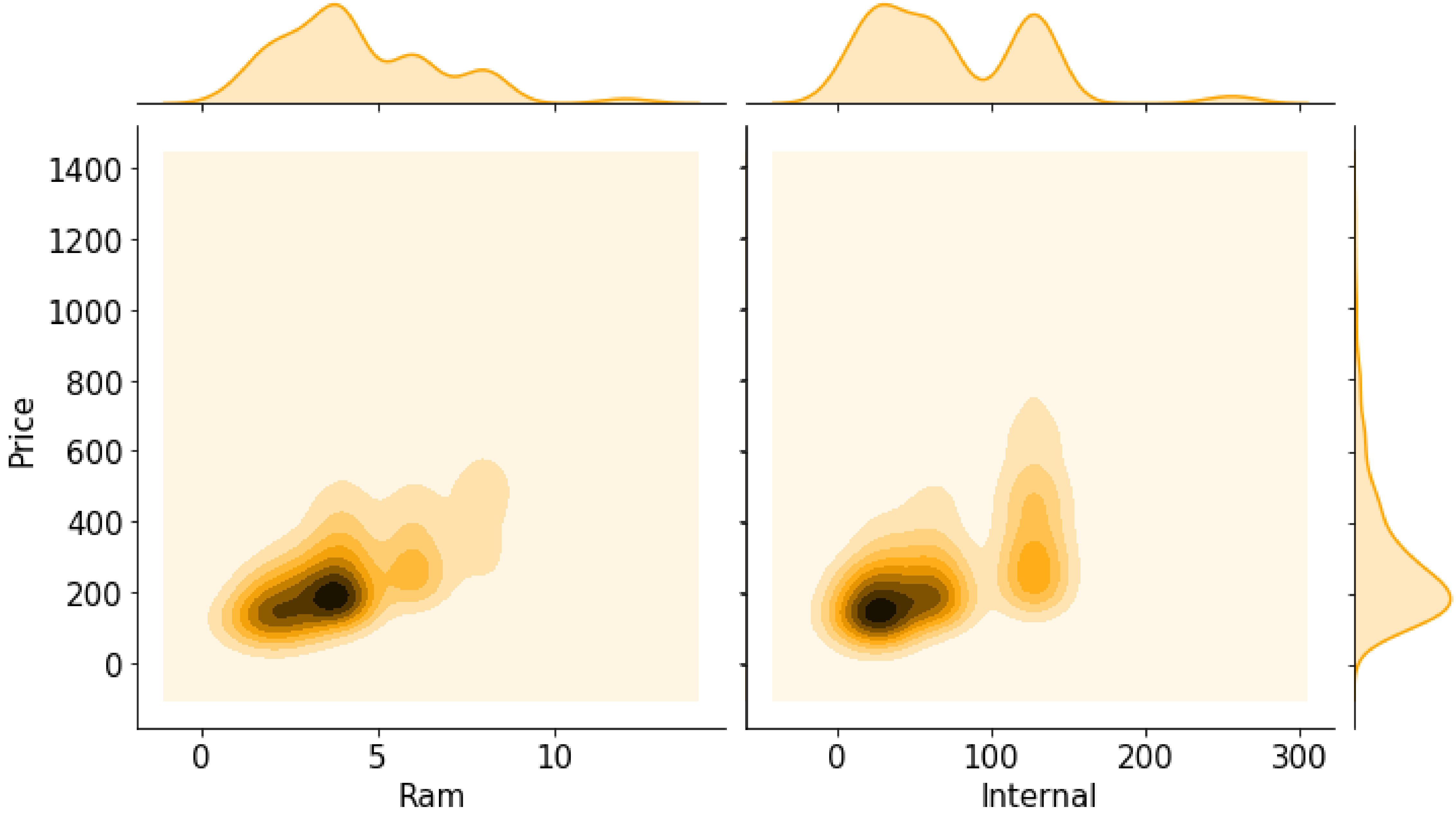
5G Enabled Smartphone percentage



Analysis

RAM & Internal Memory

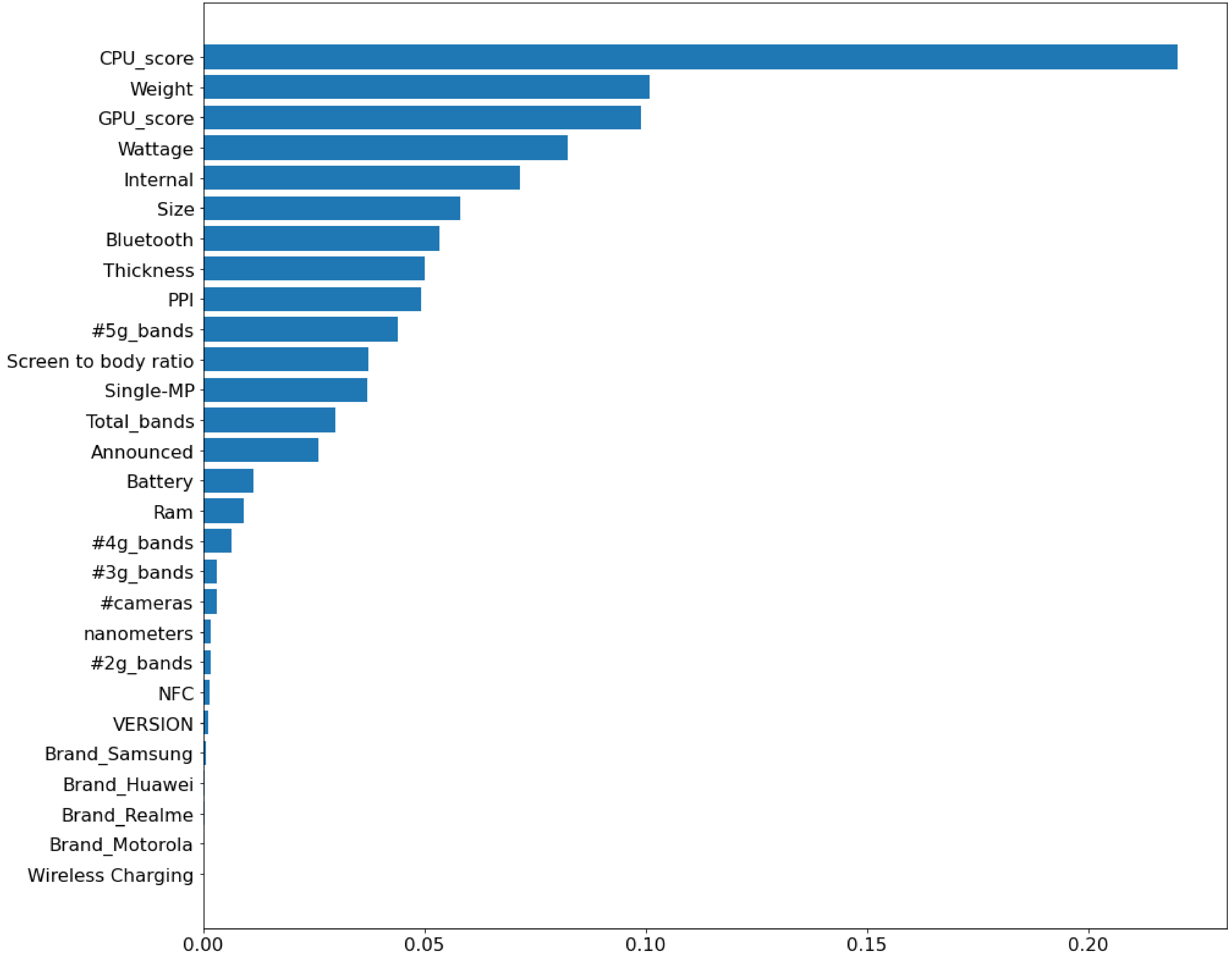
Trend of RAM and Internal Memory vs Price.



Analysis

Feature Importance

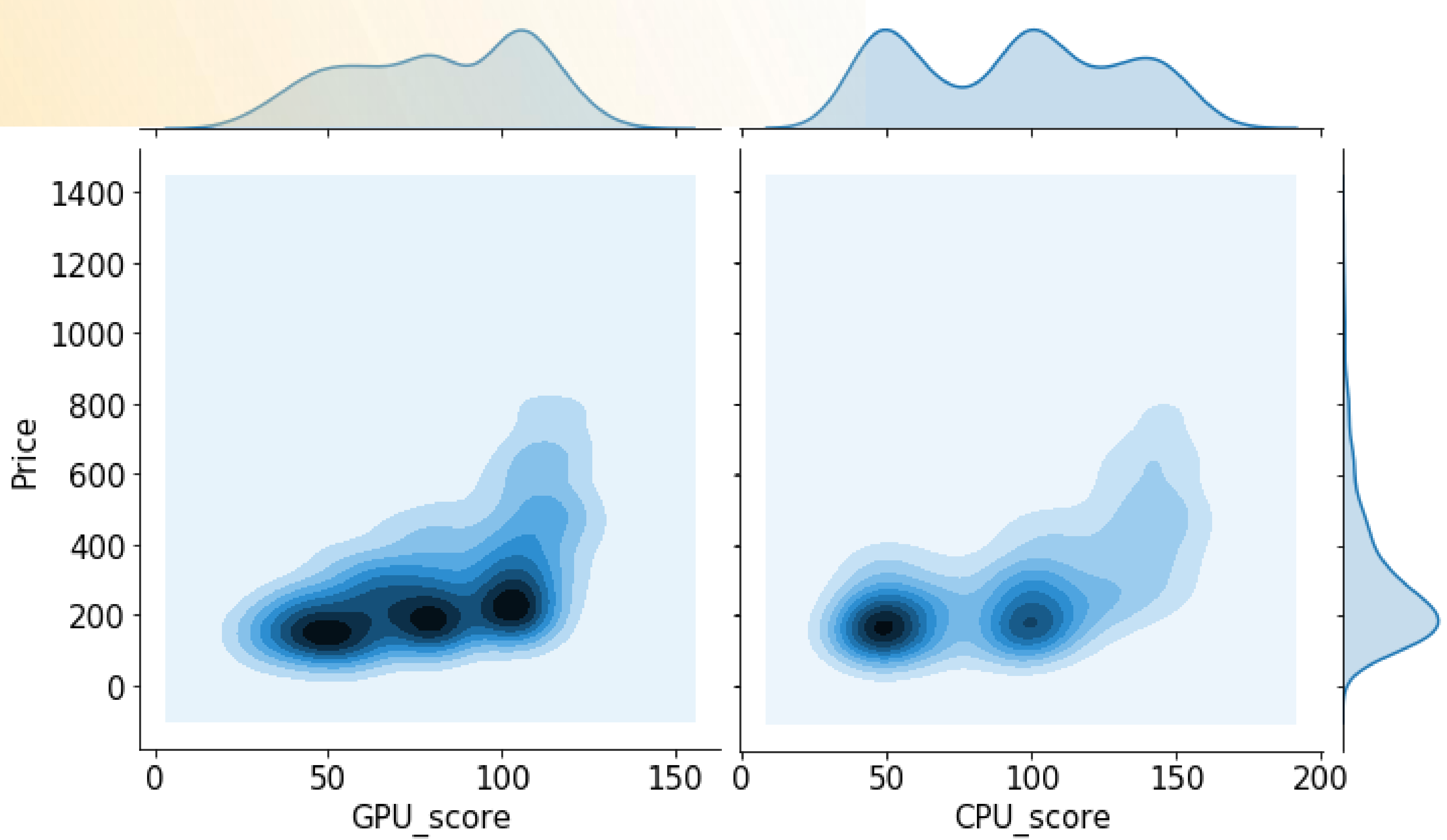
- **CPU_score** consistently pop up as the most important feature for Price Prediction.
- **GPU_score** as consistently in top 3 features.

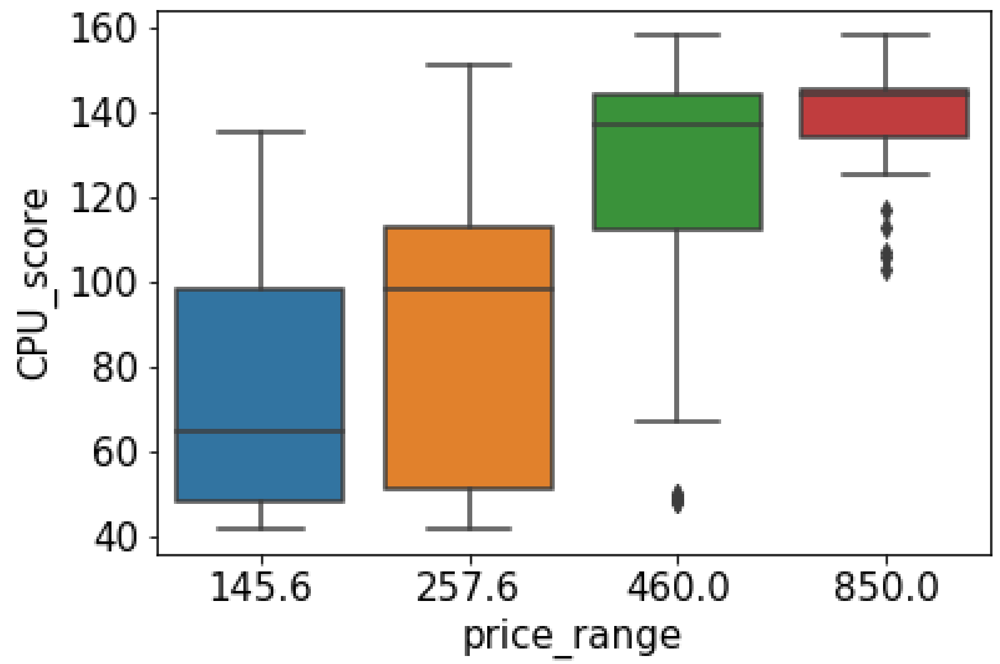


Analysis

CPU & GPU performance

We can see positive trend of Price vs CPU and GPU Scores





Results & Summary

Impact :

We can use our models to test if a smartphone is valuable to buy

Conclusion:

Random Forest performed best w.r.t. Median Absolute Percent Error with **16%** error

	Median Absolute Percent Error		R2 Score	
Model Name	Test	Train	Test	Train
Random Forest	16.25%	7.66%	0.68	0.89
Multiple Linear Regression	35.07%	36.26%	0.73	0.53
Support Vector Regression	32.77%	31.11%	0.03	0.29
Decision Tree	27.58%	0.01%	-0.49	0.99

Thank you