STA217: Analysis on Smartphones

December, 2024

Hou Zheyan, Zhang Hangyuan, Zhao Xuanzhi

1 Introduction

This is the name of our data column, and our main analysis object are brand name, price, rating, processor brand, num cores, processor speed, ram capacity, screen size, resolution width, resolution height.

Figure 1: Data Content

This is the data size where there are no duplicate columns.

This is a description of the numerical data.

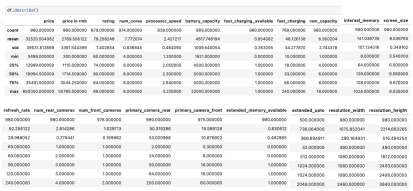

Figure 3: Range of numerical data

This is our most basic visualization, divided into qualitative and quantitative parts.

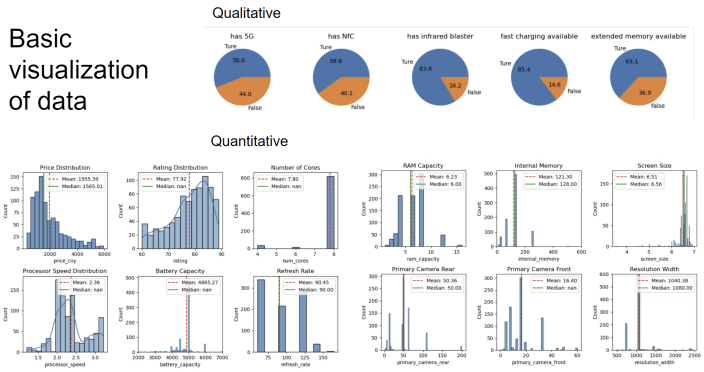

Figure 4: basic visualization of data

This is a heat map of the correlation coefficients between all of our numeric type variables, for Boolean types, we change true and false to 1 and 0 with a blue line inside, indicating that extend memory is negatively correlated with all of the other variables, because extended memory is being eliminated with the development of technology, This property also has ir blaster infrared remote control function.

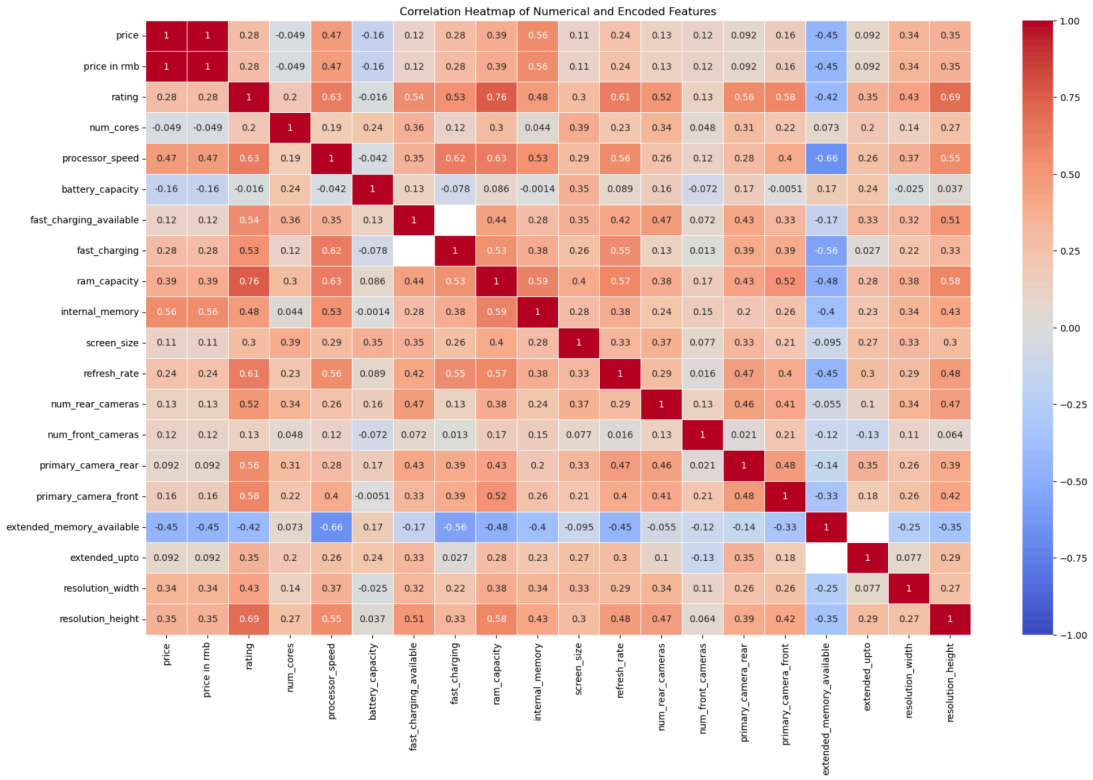

Figure 5: Correlation Heatmap of Numerical and Encoded Features

2 Data cleaning

This is our presentation of the vacancy values, and through this heat map, the vacancies are mainly concentrated in three columns.

```
df.isna().sum()
```

```
brand_name                  0
model                       0
price                       0
price in rmb                0
rating                    101
has_5g                      0
has_nfc                     0
has_ir_blaster              0
processor_brand            20
num_cores                   6
processor_speed            42
battery_capacity           11
fast_charging_available     0
fast_charging             211
ram_capacity                0
internal_memory             0
screen_size                 0
refresh_rate                0
num_rear_cameras            0
num_front_cameras           4
os                         14
primary_camera_rear         0
primary_camera_front        5
extended_memory_available   0
extended_upto             480
resolution_width            0
resolution_height           0
dtype: int64
```

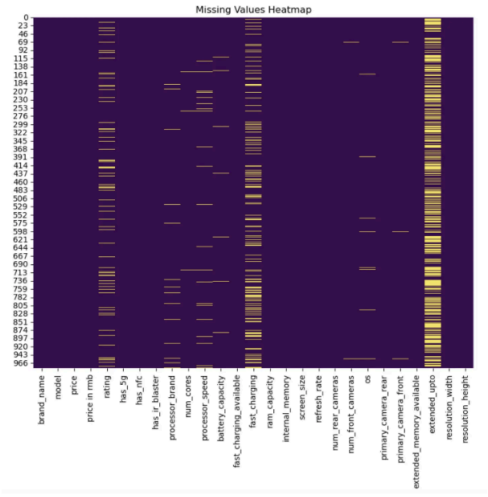

Figure 6: Distribution of null values

First of all, the columns that are not used in the following analysis are deleted, and then the empty values are filled with homogeneity to get the cleaned data. Next, remove the phones that are too expensive and too low, and mainly analyze the middle part.

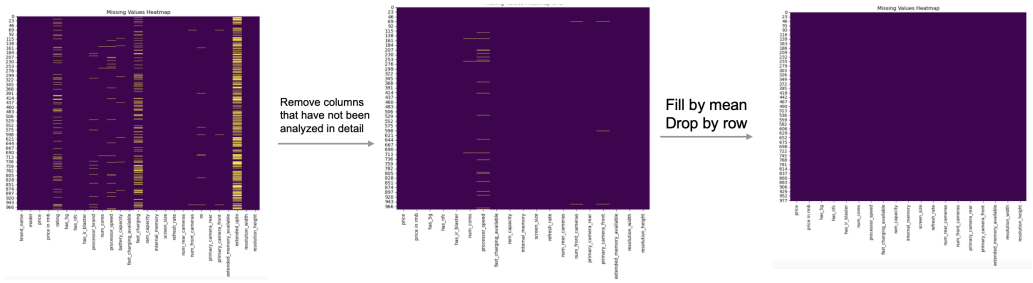

Figure 7: Data cleaning

3 Some details of the data

3.1 Correlation coefficient and causation

RAM capacity represents the amount of memory in your phone that can be used for temporary storage and quick access to data. The larger RAM capacity means that the device can run more applications at the same time and switch tasks more smoothly. The amount of RAM determines the performance of the phone under the demands of multitasking and high performance.

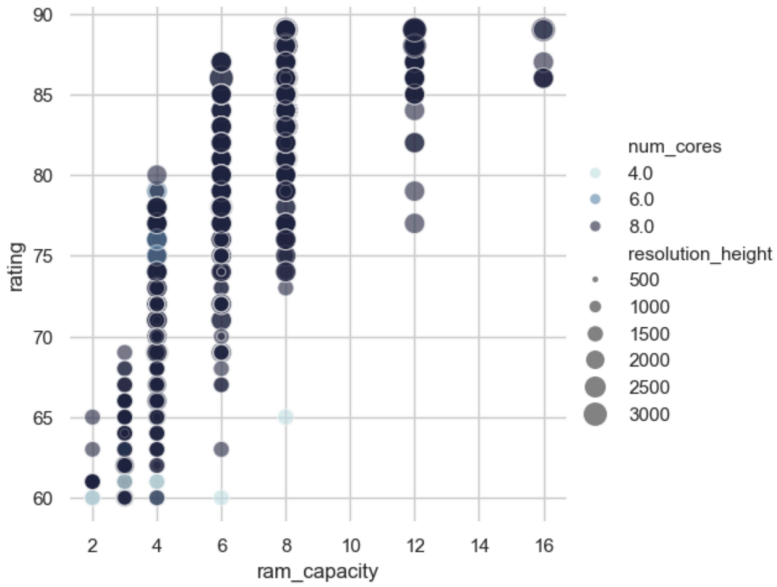

Figure 8: The relations of some causal features

As can be seen from the scatter plot, the mobile phone with a higher score has a larger ram capacity and more audits. Timely audit more, if the ram capacity is not enough, the score is not high. For example, if the phone has a powerful processor (such as a high-performance multi-core processor) and a small RAM capacity (such as 4 GB or 6 GB), it may cause bottlenecks when multitasking. Conversely, if the phone has a large RAM capacity (such as 8 GB or 12 GB) but a weak processor, it may still not be able to take full advantage of the RAM. The best performance is usually when the processor and RAM are matched. (Some positive and negative correlations are large but there is no causal relationship, such as the expanded memory mentioned earlier).

This graph also shows the distribution of the data by changing the transparency. When the screen resolution is higher, the larger RAM capacity can help better process high-resolution images and video data, especially when playing high-quality games or editing high-definition videos. Larger RAM can store more graphics data, reducing picture lag.

3.2 Relationship between processor brands and smartphone brands

Here shows the main frequency rate between different brands of processors, you can see that there is a big difference between different brands, some brands are uneven, some are maintained at a high level, and some are generally low speed.

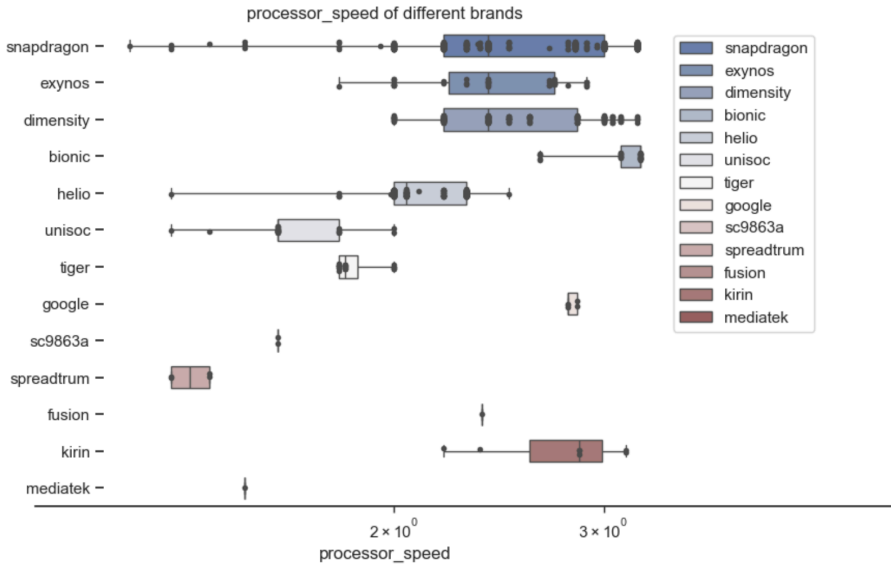

Figure 9: Processor speed of different brands

This graph shows the relationship between mobile phone brands and processor brands, some of which have fixed partnerships with each other. For example, Apple and helio and fusion, their cooperation is very fixed. You can also see overlap in some circles, indicating that there is also a big difference between the same processor brand bought by the same phone brand.

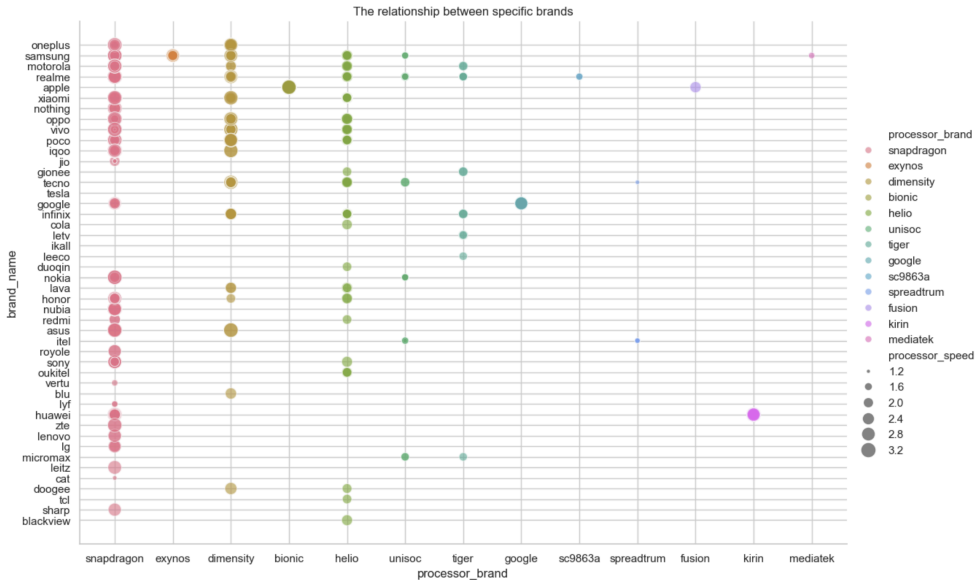

Figure 10: Cooperative relationship of processor brands and smartphone brands

4 Performance comparisons between various price ranges

We divides the smartphone price data into four quantile-based intervals and assigns them numerical labels (0, 1, 2, 3).

4.1 Screen features

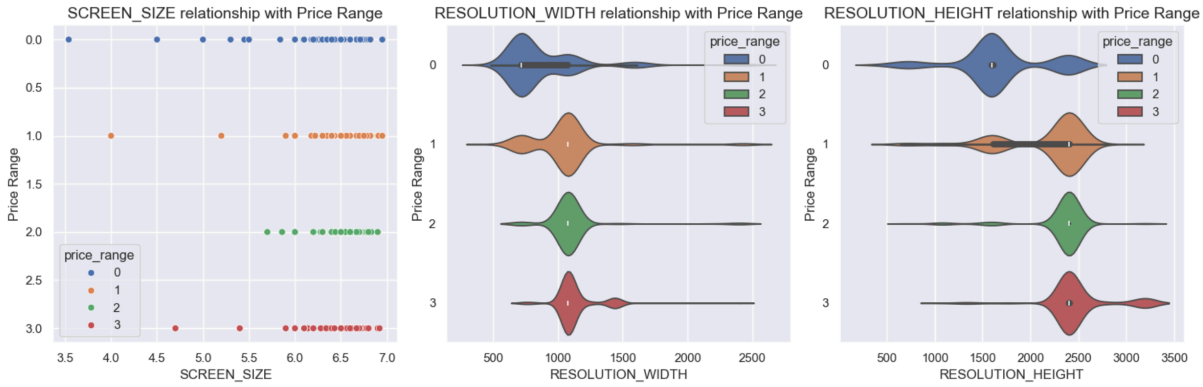

Figure 11: Relationship between screen features and the price range

The plots reveal the following key insights: Larger screens are linked to higher price ranges, with a clear division above 6.5 inches. There's a mix of price ranges across resolution widths, with higher ranges showing a broader distribution. Similar to width, a variety of price ranges are present across resolution heights, with a notable concentration around 2000 pixels in higher ranges.And only the cheapest mobile phones can not guarantee the quality of the screen, even the screens of cheaper mobile phones and mid-range phones have been able to do almost the same as the screen of the higher price.

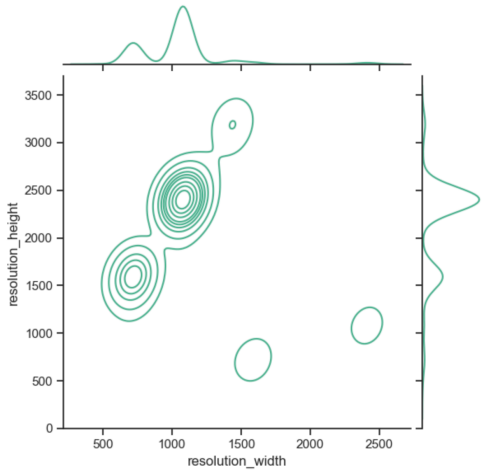

Figure 12: The relationship between resolution width and height

The aspect ratio of the resolution (such as 16:9, 4:3, etc.) determines the proportion of the screen image and usually does not change with the resolution, but the display can be improved by increasing the resolution (that is, increasing the number of pixels of width and height).Higher resolution results in a better display, while lower resolution can lead to blurring or distortion.But the higher the screen pixels, the better the screen quality is not necessarily.

In order to better describe the quality of the screen, we introduced a new concept called PPI (Pixels Per Inch) . PPI is a measure that indicates the density of pixels in a given space, and a higher PPI value means more pixels per inch, leading to a sharper and more detailed image.

The text also provides a formula to calculate PPI based on screen resolution (width and height in pixels) and screen size (in inches):

$$PPI = \frac{(width^2 + height^2)^{\frac{1}{2}}}{size}$$

In essence, for a fixed screen size, a higher resolution (more pixels) results in a higher PPI, leading to a clearer display. Conversely, for a fixed resolution, a smaller screen size will have a higher PPI, again leading to a clearer display

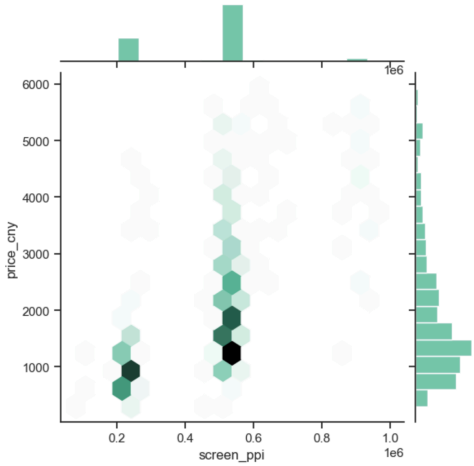

Figure 13: Hexbin plot of price vs PPI for smartphones

This graph shows the relationship between phone price and screen pixel density (PPI), and the distribution is generally consistent with the previous violin graph.only the cheapest mobile phones have low qualities of the screen.The hexbin plot highlights a concentration of smartphones with PPIs around 500, mostly priced under 3000 CNY, and a few high-end models with PPIs over 500 and prices exceeding 6000 CNY, indicating a diverse price range for higher PPIs.

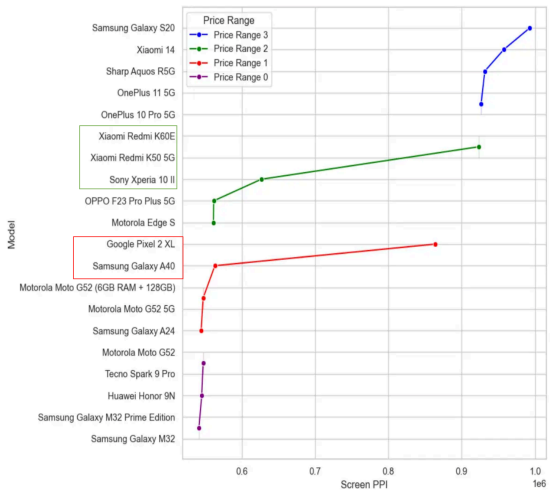

Figure 14: Top 5 models by PPI in each price range

Here we have selected some models with outstanding screen quality in different price ranges to recommend.

4.2 Camera features

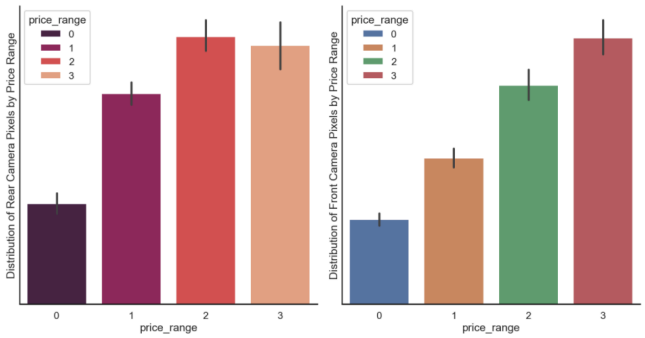

Figure 15: Distribution of rear and front camera pixel counts across price range

The bar charts reveal that higher price ranges tend to have rear cameras with more pixels, with the most significant increase seen in price range 2. For front cameras, there's a notable rise in pixel count from price range 1 to 2, with price range 3 maintaining a high pixel count, indicating a correlation between camera quality and smartphone price.

4.3 Memory features

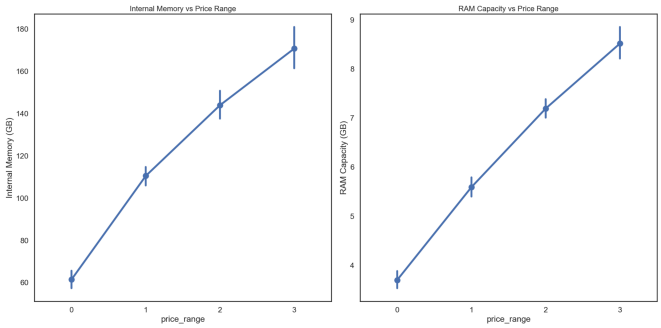

Figure 16: The relationship between price range and memory

The line graphs demonstrate a clear positive correlation between price range and both internal memory (GB) and RAM capacity (GB), with both increasing consistently as the price range goes up, indicating that higher-priced smartphones generally offer more memory and better performance.

5 Research on smartphone rating

When observing raw data, we found a distinctive column — Rating: The rating of the smartphone, typically based on user reviews or expert evaluations. Rating does not belong to the properties of smartphone, but is an important parameter for users to evaluate these smartphones. But there is a certain degree of missing data in this column, so we tried to fill it in.

5.1 Data cleaning based on 'rating'

Because we need data with 'rating' parameter to infer the relationship between other parameters and 'rating' parameter, we first remove empty rows from the 'rating' column. Referring to the correlation coefficients between various parameters and 'rating' obtained from the beginning, for columns with high correlation with rating and many null values ('processor steed', 'fast charging'), directly filling the null values with the mean for estimation will result in significant bias. Therefore, null values are removed for model fitting. For columns with low

correlation with rating and a high number of null values (‘battery capacity’, ‘extended upto’), choose to remove the columns. For columns with fewer missing data (g), choose to directly fill the gaps with its mean. Check the data again, the cleaned data has no empty values, and use it for model learning and prediction.

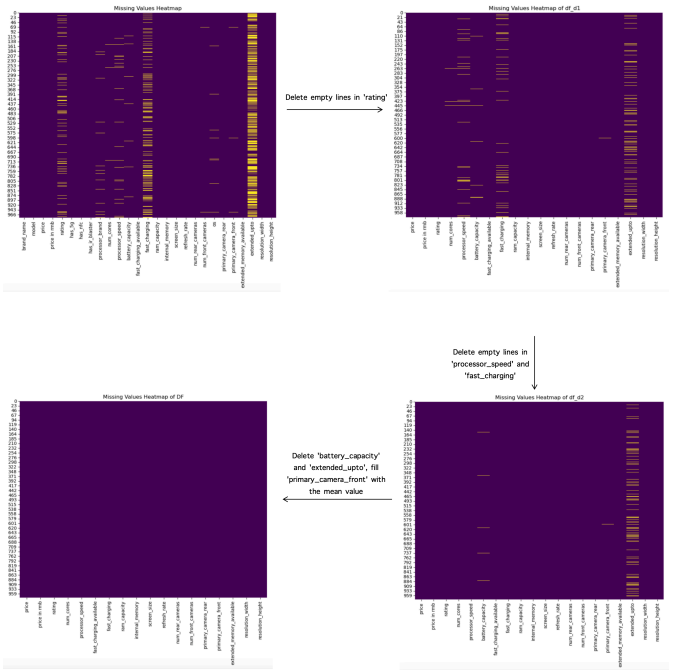

Figure 17: Data cleaning based on ‘rating’

5.2 Model prediction

We first divide the data into various smartphone parameters X and ‘rating’ data y, and divide the dataset into training and testing sets in a ratio of 17:3 for model training and evaluation. Afterwards, models were trained and predicted using Linear Regression algorithm based on linear relationships and Decision Tree Regression algorithm that captures nonlinear relationships, respectively. The tree structure method was obtained by calculating the coefficient of determination and mean absolute error to better fit the model. Subsequently, the Random Forest algorithm based on multiple decision tree prediction results was further used, and it was found that this algorithm could best fit. Finally, the random forest algorithm was used to calculate the ratings of all smartphones, and the parameter ‘test rating’ was obtained. Linear regression and visualization were used to examine its fitting accuracy with the original data, and it was found that both the trend and accuracy were satisfactory.

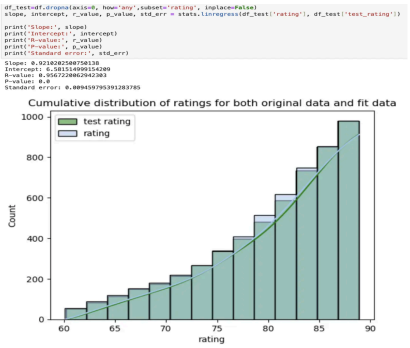

Figure 18: The accuracy of prediction

In order to make 'rating' parameter more accurate and facilitate subsequent analysis, we filled in the parts of the original 'rating' that were empty with 'test rating', and replaced the parts of the original 'rating' that were not empty with the mean of 'test rating' and 'rating'. Thus, we obtained the ratings for all smartphones, which also means that we can sort all phones in the data based on the ratings.

5.3 Recommendations for sample phones

We selected several commonly used and easy to compare smartphone parameters from the data, normalized the data in these columns by the following equation:

$$x^* = \frac{x - \min}{\max - \min}$$

Classified the daily smartphone brands OnePlus, Xiaomi, Huawei, Vivo, Samsung, and Apple, sorted and selected the top three rated phones for each brand. We then drew a performance radar chart with colors from dark to light representing the first to third.

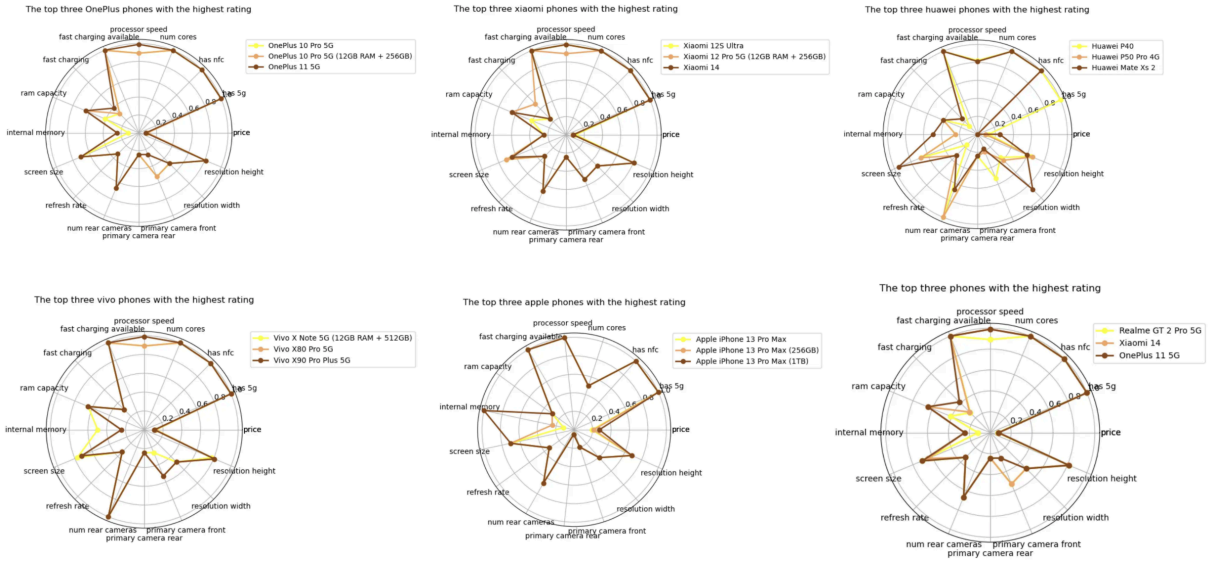

Figure 19: Radar chart of the top 3 smartphones rated by various brands

From these graphs, it can be seen that the advantages of OnePlus are reflected in their high cost-effectiveness, low price but good performance, and product improvements mainly reflected in processor speed, ram capacity, and fast charging rate; Xiaomi have good performance in all aspects and are not expensive; For the radar chart of Huawei, it is worth noting that the Huawei Mate Xs 2 is a foldable screen phone that does not support 5G, so its screen size, resolution width, and price are higher than Huawei's other two phones; The camera function of Vivo phones is outstanding, and its product improvement is mainly reflected in the processor speed; Iphone also have good performance in terms of processor speed and ram capacity. Finally, we also found the three highest rated phones under this rating scale, namely Xiaomi 14, Realme GT 2 Pro 5G, OnePlus 11 5G, and obtained their performance radar chart.

We can also place the rating within the brand scope, as there are many brands and smartphone models, we will divide the data into three intervals based on ratings: above 80, 70 to 80, and below 70. Next, plot a bar chart of the number of phones from each brand in the three intervals in descending order.

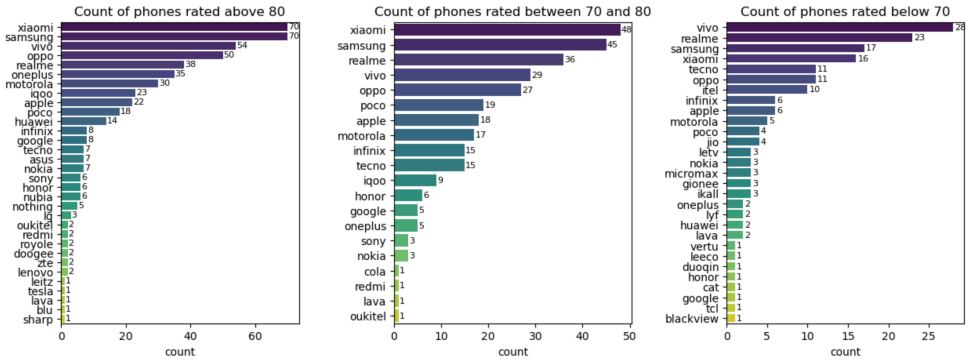

Figure 20: Number of smartphones in different rating ranges

To make it more intuitive, we plotted three graphs on one and used colors from light to dark to correspond to the number of phones for each brand in different rating ranges. Above the column, we marked the proportion of phones with a score of 80 or above to the total number of phones for the brand (i.e., the excellence rate).

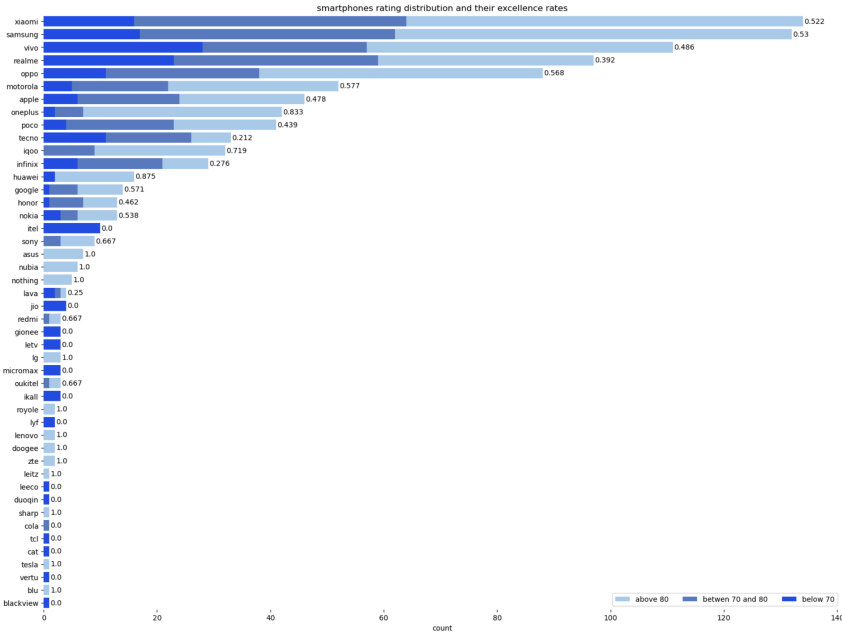

Figure 21: Smartphones rating distribution and their excellence rates

We found that for brands with sufficient statistical quantity, OnePlus and Huawei have a large proportion of high rated phones.

6 Summary

In conclusion, our analysis provides valuable insights into the key features of smartphones, such as screen quality, camera performance, processor speed, and memory, and their relationships with price and brand. By leveraging data cleaning techniques, correlation analysis, and predictive modeling, we were able to fill missing ratings and rank smartphones according to their performance. Additionally, the performance radar charts and rating distributions offer useful recommendations for consumers looking for high-performance phones within their budget. This study underscores the importance of balancing different smartphone attributes to achieve optimal value and satisfaction. Future work can expand on these findings by incorporating more granular data or exploring additional factors influencing consumer preferences.