Exploratory Data Analysis on Mobile Devices Data

Yifan Liu
Faculty of Engineering
University of Sydeney
Sydney, Australia
yliu2015@uni.sydney.edu.au

Ye Cai
Faculty of Engineering
University of Sydeney
Sydney, Australia
ycai6198@uni.sydney.edu.au

Niti Patel
Faculty of Engineering
University of Sydeney
Sydney, Australia
npat6703@uni.sydney.edu.au

Dian Shi
Faculty of Engineering
University of Sydeney
Sydney, Australia
dshi9534@uni.sydney.edu.au

Yanhong He
Faculty of Engineering
University of Sydeney
Sydney, Australia
yahe2102@uni.sydney.edu.au

I. Market Leading Devices

A. New Type of Mobile Devices

The new type of mobile device usually refers to devices that have new features. Since all devices all have the same category of features, our team's understanding is new feature means a significant increase or decrease in the value of features. For example, if a device has doubled CPU clock compared to other devices in a short period of time, and the trend of CPU of all devices is increasing, then this device is regarded as a new type of device. Subsequently, since the increase of CPU is align with the trend of all devices, it is appropriate to assume that this new type of device is leading the market since the following devices show the same trend in certain features. This method is derived from the visualization of the general trend of devices to answer question 1.

B. General Trend of Mobile Devices

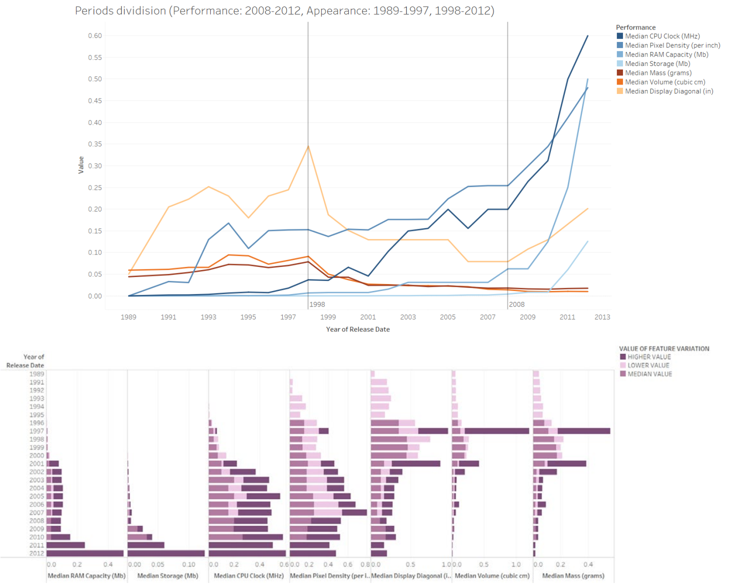


Figure 1

To achieve conceive visualization of multiple visual variables and considering the correlation between features, devices features is divided into two category (refer to table 1), each category is represent by same hue, and features in the same category is distinguished by different chroma. Line chart of temporal analysis is applied to analyze the trend of features in different time periods to explore which features can be regarded as new features. This visualization mainly uses two hues, yellow and blue [1]. They respectively represent the performance and appearance of the device. The reason why choosing blue and yellow hues is because it is easy to distinguish whether the feature belongs to the performance or appearance. Meanwhile, Each hue uses corresponding chromas to distinguish different features in it. The final visualization of line chart is shown in the top of Figure 1.

As shown in the bottom, we have also divided all the median values separately in the range of higher, median and lower to help gather more information through the visualization and displayed them using variations in chroma and hue over the period of time from1989-2012.

In Figure 1, our team used the median values in two graphs to show changes. The reason why choosing the median instead of the average is because the median is more stable than average [2]. Especially in this dataset that for all the values of features, there is a big difference between the maximum and the minimum. The average value is easily affected by both of them. However, median is insensitive neither to maximum nor minimum value. Thus it can reflect the common level of these two features better.

From the line chart and bar chart, three time periods were selected where performance and appearance of devices present a significant variation. The result of period selection is shown in table 1. The rest of time periods are ignored since the trend of features is just fluctuating in a relatively small section, which indicates that there is no new feature is implemented. In this way, detailed analysis of what devices is trying to create new market can be implemented in the next step.

TABLE I. Results of Period and Features Selection

Categories	**Period Selection**	**Attributes**
Performance of devices	2008 to 2012	RAM (memory) capacity (Megabyte)
		Storage capacity (Megabyte)
		CPU clock (MHz)
		Pixel Density (per inch)
Appearance of devices	1989 to 1998	Display size (dimensions) (diagonal in inch, width, and length in pixels)
	1998 to 2012	Volume (width-length-depth in mm)
		Mass (grams)

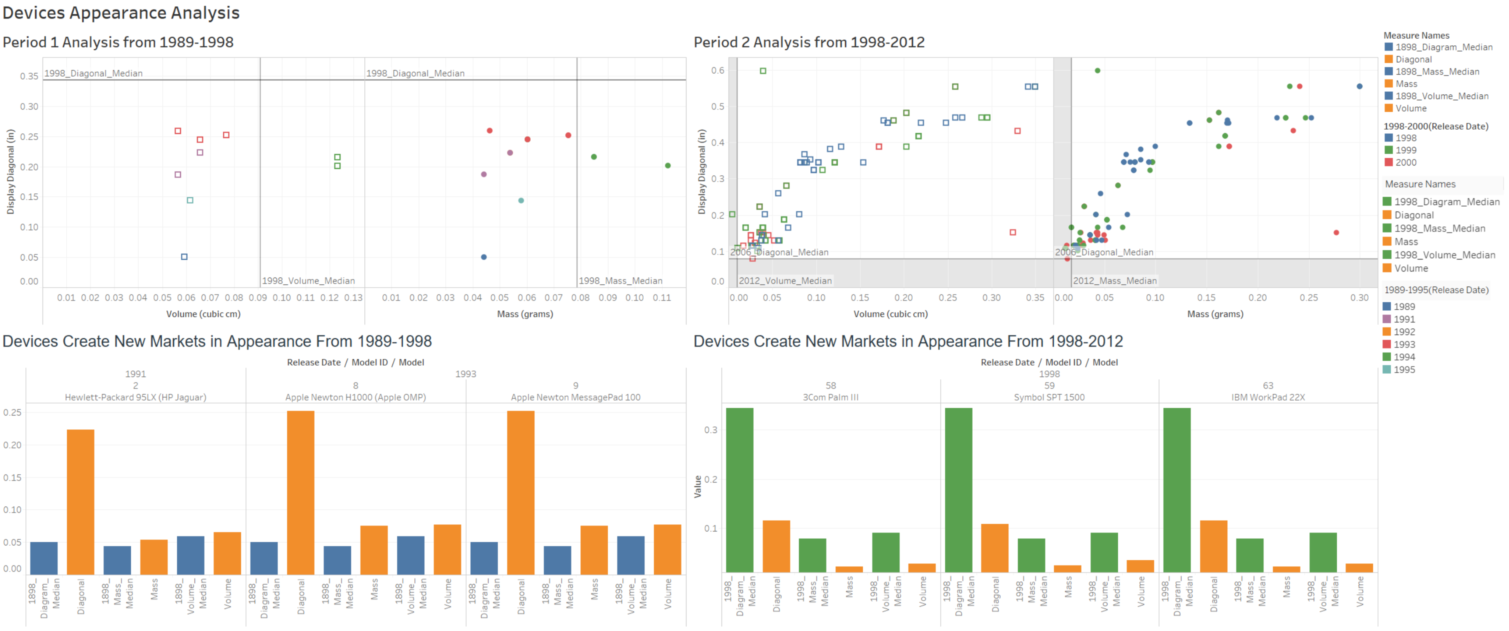


Figure 2

C. Appearance of devices analysis

To analyze the appearance of devices in two periods, scatter plots were used to show the distribution of different features in different years (Refer to Figure 2). The reason of using scatter plots is scatter plot is one of the best approaches which has advantages in distribution analysis and it can display more than two attributes on one diagram which means our team can analyze all aspects of the device on appearance at the same time and observe the significant points easier.

In the scatter plots of Figure 2, the attributes about appearance are present as columns and rows. The reference lines are the median of these features in the last year of periods which used to compare and find out the points that first started to close to these lines. It is necessary to note that our team just used the first few years of each period like just 1998,1999 and 2000 were used in the second period of appearance analysis. The reason is that the question is to find the devices that tried to create the new markets which means the devices which achieve a better performance after referring to the experience of others need to be removed. Therefore, our team decided to just use the first few years of periods to avoid this phenomenon.

Colors in the scatter plots are used to distinguish different years and the card includes the category of years with different colors that can be used to highlight points in different years. Both could provide help for customers observe the changes of devices in different features and pick up the best points in each year. In addition, Square and round shapes are used to distinguish two scatter plots in Figure 2 to avoid misunderstanding or confusion.

After selecting the significant points in different years and storing them into different sets, out team found out the devices which tries to create new market in two periods after combining sets and using filters to remove useless data. To reconfirm the results, bar charts were utilized to compare the value of each feature of the device and the median value of the first year of the period. The model ID and name were also displayed on the graph (Refer to Figure 2). It is evident that the devices have obvious changes which means our team can ensure the results are correct and the devices created new markets in appearance aspect.

Consequently, in the period from 1989 to 1998, Model 2, 8, and 9 tried to create a new market in appearance. Model 58, 59, and 63 which have a rapid decrease from 1998 to 2012 created new markets in appearance in the second period.

D. Performance of devices analysis

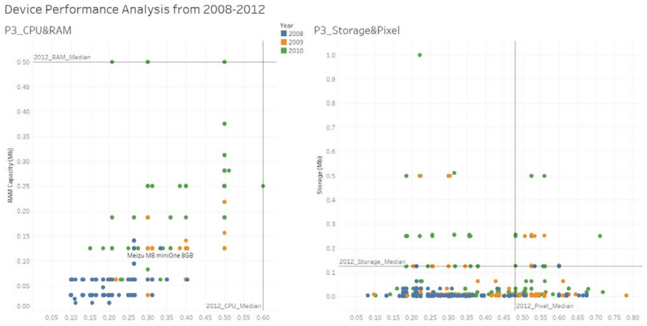


Figure 3

The process of performance devices analysis is similar to appearance analysis that using scatter plots to show the distribution and select the best points in each year firstly (Refer to Figure 3) and then use a bar chart in Figure 4 to show the result and ensure the accuracy after filtering data.

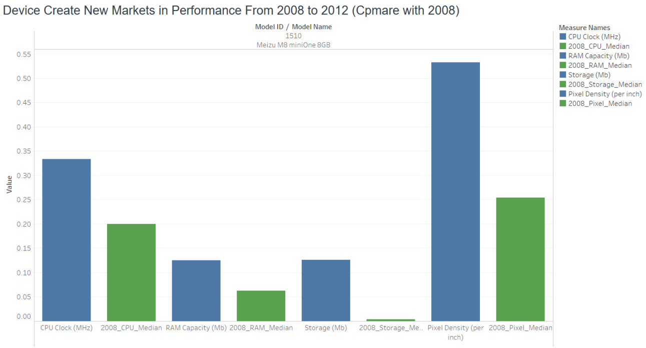


Figure 4

In Figure 3, colors are also used to distinguish the year but our team used two scatter plots to show four feature in performance and each scatter plot displays two attributes in the column and row. One category card can be used on both two scatter plots to highlight the points in the same year that would help customers better observe the distribution and changes of the trend of changes in different features (Refer to Figure 5).

In summary, model 1510 has been considered as the device which created new markets in the performance from 2008 to 2012.

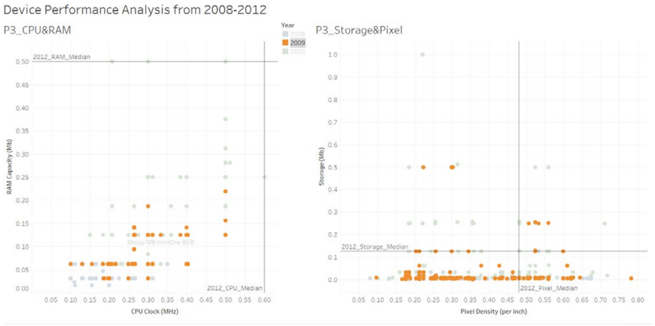


Figure 5

II. Market Leading Companies

A. Scatter plots between median values of the four performance-based features

As observed in the first analysis we can say that the appearance-based features do not seem to vary a lot over the period of time from 1989-2012, and it is extremely difficult to derive any information by using these features. Hence to perform our second analysis we have only considered performance related features since they vary a lot and therefore a lot of information can be obtained by analysing these features further. We have also considered the time period between 2008-2012 the reason being the features that we have selected to analyse vary mainly between this time frame and therefore it is critical to analyse them during this time frame.

The data considered for further analysis: Median values of RAM Capacity, CPU Clock, Storage and Pixel density for all mobile devices released within the time period of 2008-2012. The final visualization is displayed in Figure 6.

1) Axes:

– X axis and Y axis take turns in representing one the four performance features that is CPU Clock, RAM, Pixel or Storage in their median values. Since each of these values takes turns on the scatter plot there is no specific reason for selecting any axis for these features.

– The features on both the axes increase bottom to top left to right with an interval of 0.2. The Value starts from 0.0 to 1.0, the higher the position, the greater the value. This is done to find the points of correlation between these features above the median values.

2) Visual variables

– Scatter plot: This visualization selects the scatter plot. The reason being that the median values of the features used for the mobile devices exist at multiple points between 2008-2012. And to capture all these varying values, scatter plot is the right fit. It is also essential to help find the co-relating points between these features.

– This visualization mainly uses one hue Blue to plot the points the feature values for each feature.

– Each hue uses corresponding chromas to distinguish between the points that are selected for further analysis the darker points on the scatter plots which indicate the points above median for each feature in that plot.

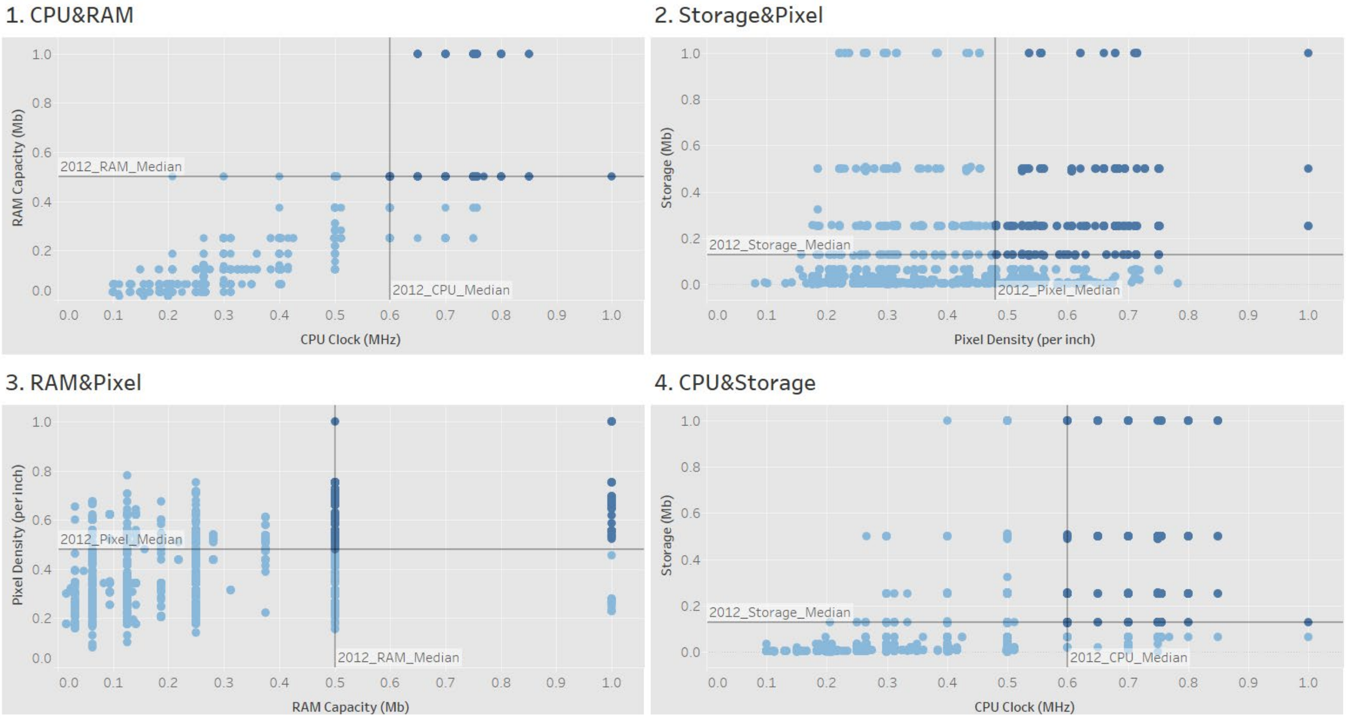


Figure 6

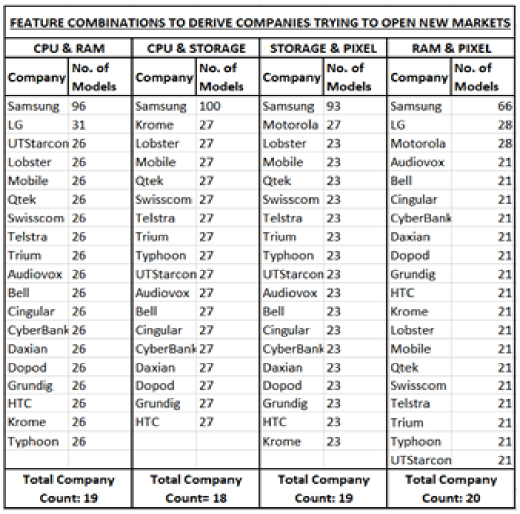
Meaning for further analysis we only consider the common above median points for each plot.

- The reason on plotting the median lines in each of the plots is because it is observed that these features have an increase in value over meaning new performance based are mainly high values therefore to find co-relating points between these features, median values about a certain range indicate better performing devices as they have better features.

3)Inferences from the scatter plots:

Based the darker blue hue points on the scatter plots the following data is collected by us for further analysis:

TABLE II. FEATURE COMBINATIONS TO DETRIVE COMPANIES TRING TO OPEN NEW MARKET



FEATURE COMBINATIONS TO DERIVE COMPANIES TRYING TO OPEN NEW MARKETS							
CPU & RAM		CPU & STORAGE		STORAGE & PIXEL		RAM & PIXEL	
Company	No. of Models	Company	No. of Models	Company	No. of Models	Company	No. of Models
Samsung	96	Samsung	100	Samsung	93	Samsung	66
LG	31	Krome	27	Motorola	27	LG	28
UTStarcon	26	Lobster	27	Lobster	23	Motorola	28
Lobster	26	Mobile	27	Mobile	23	Audiovox	21
Mobile	26	Qtek	27	Qtek	23	Bell	21
Qtek	26	Swisscom	27	Swisscom	23	Cingular	21
Swisscom	26	Telstra	27	Telstra	23	CyberBank	21
Telstra	26	Trium	27	Trium	23	Daxian	21
Trium	26	Typhoon	27	Typhoon	23	Dopod	21
Audiovox	26	UTStarcon	27	UTStarcon	23	Grundig	21
Bell	26	Audiovox	27	Audiovox	23	HTC	21
Cingular	26	Bell	27	Bell	23	Krome	21
CyberBank	26	Cingular	27	Cingular	23	Lobster	21
Daxian	26	CyberBank	27	CyberBank	23	Mobile	21
Dopod	26	Daxian	27	Daxian	23	Qtek	21
Grundig	26	Dopod	27	Dopod	23	Swisscom	21
HTC	26	Grundig	27	Grundig	23	Telstra	21
Krome	26	HTC	27	HTC	23	Trium	21
Typhoon	26			Krome	23	Typhoon	21
						UTStarcon	21
Total Company Count: 19		Total Company Count= 18		Total Company Count: 19		Total Company Count: 20	

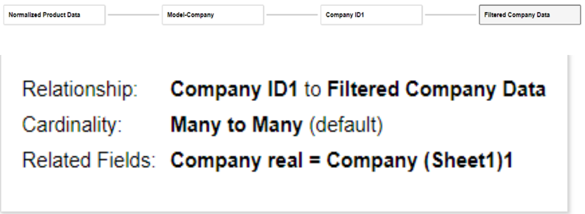
The above data in TABLE 2 is used to create another datasheet to attach to the existing database. This data represents the top companies from co-relating points on the scatter plots based on different combinations of performance features.

4)Conclusions and Observations:

- We were successfully able to derive the necessary information from the scatter plots and filter out companies necessary to identify the top companies trying to lead the market.
- We also observe that a lot of companies are common amongst this data derived dataset alone can stand for the companies trying to lead the market with best performance-based features. However, using this data, we will try to conduct further analysis and derive the top most companies amongst these listed in the above figure.

B. *Derive the final top companies that tried to lead the market with new features*

- The dataset used for further visualization includes the list of companies obtained in the previous step we have combined it with the existing dataset using the following relationship and cardinality.



Relationship: **Company ID1** to **Filtered Company Data**
Cardinality: **Many to Many** (default)
Related Fields: **Company real = Company (Sheet1)1**

Figure 7

- This is done to derive the companies to answer our second question, the top companies from the previous analysis represents companies that were able to produce devices with the best features.
- However, to answer the second question to identify new markets, we need to count the number of new features released by the list of companies we derived in the previous step. By using a filter in the visualization and arranging the companies in descending order based on the count the mobile device features released by the companies we derive the following visualization representing the companies in descending order to obtain top 3 companies which were able to deliver devices with best performance and also lead the market by constantly releasing new devices (Refer to Figure 8).

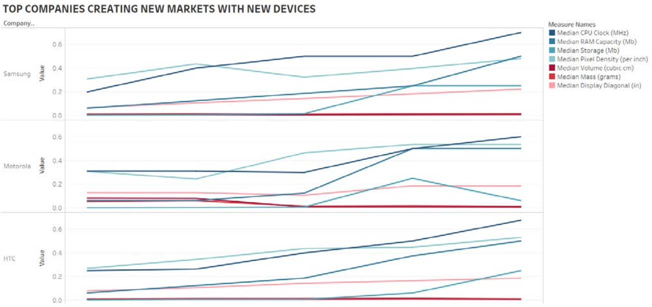


Figure 8

1) Axes

- X axis represents the year of release for the features of all the devices in the dataset, Y axis represents the median values for all the features again similar to figure 1 y axis is the value from 0 to 0.8. Since all the data is normalized in the original dataset, the difference between units could be ignored. The reason for choosing years as the x axis and value as the y axis is because the value of the corresponding feature can be found through different years. The Y axis also represents the top companies that tried to lead the market. This is so because we can easily observe the new features these companies tried to release from 2008-2012, and easily compare them to learn the trends based on the differences and similarities in the new features and new devices these top companies release.
- The X-axis represents the years ranging from 2008-2012 with an interval of 1 year increasing from left to right. The value on the Y-axis starts from 0.00 to 0.6 at the higher the position, increasing gradually in bottom-up direction with the greater the value at the top. This will enable us to compare and contrast feature values of different companies over the period of time. Y axis also

shows the top 3 companies in descending order from top to bottom with Samsung leading the way to the top.

2)Visual variables

- Line Graph: This visualization selects the line graph. Because the abscissa is the year and the ordinate is the value, the change trend can be seen through the increase of the year. This will help us learn about the features for these top companies.
- Hue: This visualization mainly uses two hues, red and blue. They respectively represent the performance and appearance of the device. The reason why choosing blue and red hues is because it is easy to distinguish whether the feature belongs to the performance or appearance-based features respectively.
- Chroma: Each hue uses corresponding chromas to distinguish different features in it [1]. Blue hue the performance. Red hue the appearance-based features.

3) Conclusion and Observation:

- We can observe that all the top 3 companies Samsung, Motorola and HTC showed similar trends in terms of their features which makes it easy for us to conclude that these companies had similar trend patterns.
- We can also observe that there is a constant variation at least in the performance-based features represented by blue hues in these companies which confirms that are findings are correct and all these companies did try to constantly release new devices to create new markets.
- We can also observe that the appearance-based features represented by red hues although do not vary much do show similar behaviour for all the three companies.
- In conclusion Samsung, Motorola and HTC are the top companies that tried to lead the market delivering best performing devices and constantly releasing new features trying to create a new market.

III. The Most Successful Market-Leading Company

This task does not specify what does it imply by the company being the most successful one meaning we cannot conduct analysis unless we define the success of a company. Therefore, to answer this question, we first need to give a definition indicating the successfulness of a company. For this task we have come up with 2 definitions based on the dataset provided for analyses they are:

1) Definition 1: The company that performed the best by delivering the best mobile device. This definition is answered by the answers obtained by the first question.
2) Definition 2: The company that tried to create a new market by constantly trying to release new devices with new features. This definition was answered by the analysis performed in the second question.

Therefore, based on the previous two analysis we can conclude that the 2 companies that performed the best are (Refer to Figure 9):

1) **Meizu Technology Co., Ltd.-** Because it satisfies our first definition since it has the best performing mobile device with the model ID 1510 called Model Meizu M8 miniOne.
2) **Samsung-** The reason being that Samsung turned out to be the company answering the second definition, by leading the market by constantly releasing new devices with new features.

A. Most Successful companies leading the market.

1) Axes: Similar to the Previous visualization only difference is that the Y axis shows the top two most successful companies Samsung and Meizu Technology.

2) Visual variables: Identical to the previous visualization.

3) Conclusions and Observations: We can easily observe that both companies have many similarities in the patterns of the median feature values. In conclusion Samsung and Meizu Technology Co. Ltd. Are the 2 most successful companies based on the success KPIs defined by us.

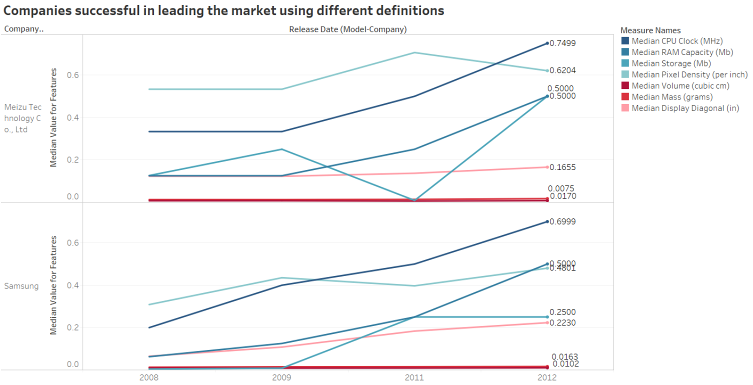


Figure 9

References

[1] Munsell Color System. (2012, March 31). Retrieved May 28,2021,from https://en.wikipedia.org/wiki/Munsell_color_syste m

[2] The Tradition of Submitting the Median in Data Mining Competitions as a Baseline. (2019, August 16). Retrieved May 28, 2021, from https://zhuanlan.zhihu.com/p/78357730