Data Exploration Project

Mohit Sainani

April 23, 2019

Prologue

GSMArena is a gadget review website with a focus on cellular and mobile devices. I believe to have come across the site a long time ago through the now defunct iGoogle start-page. It was among the first few tech blogs among the likes of How-To Geek, XDA Developers, Android Police, and AnandTech that had invoked in me the sort of enthusiasm for gadgets and devices that I have carried on ever since.

Introduction

Over the years, GSMArena has compiled a rich database of mobile device specifications. The decision to scrape off their website - despite there being numerous data sets out in the wild to run exploratory analysis upon - was based upon multiple factors. Some of them are:

- an excuse to demonstrate (read : hone, improve) my data mining skills
- experiment with alternative looping techniques in R using the purrr library
- most importantly, the nostalgia attached with working with such a dataset

You can find the code used for scraping device specifications and the generated raw data here on my GitHub. The exploration code and rmarkdown file are also included in above repository.

The questions aimed to be answered at the end of this project are:

1. Observe the trend in screen-to-body ratio of devices over the years
2. Trends in battery capacity of devices vs overall weight of devices
3. How the CPU chipmakers are related through OEMs / device manufacturers
4. Trends in adoption of multiple camera lenses and fingerprint sensors
5. Number of devices launched with each Android over the years

This project has been entirely performed in the R (ver 3.5.3) programming language.

Data Wrangling

We start off with reading the exported json file (created after data scrape-ing) into R and converting it into R's bread and butter data.frame structure. Each data frame element from the list is changed into the wide format. After this operation, each row represents specifications of one device.

The initial dataset comprises of 8980 rows x 86 columns

Comparing this dataset with the GSMArena repository, we find that we managed to capture data for 93.5% (8980 out of 9601) devices listed on the website.

We restrict the dataset to about 24 columns which fall under the scope of the proposed questions.

Next, we'll clean some of the data columns that we know will be required for answering the suggested questions. Regex is used for pattern matching, and the R tidyr and dplyr libraries are used for wrangling purposes. The tasks performed are:

1. Device weight in grams
   The *body_weight* column is a string which has information in both grams and ounces. We extract the weight in grams and cast it as a numeric value
2. Body dimensioms
   Extract the length, breadth, and width (in millimeters) of each device from the the *body_dimensions* column
3. Display size
   Extract the display size (diagonal) of the device in inches
4. Battery capacity
   Extract the battery capacity (in mili ampere hours, mAh)
5. Display resolution
   Extract the screen resolution as horizontal and vertical pixel counts, and save them in respective fields
6. Price
   Separate the price amount and price currency into different variables
7. RAM
   Derive the RAM from *memory_internal* field and convert it into gigabytes of memory
8. Camera lense count
   Collpase the *main_camera* fields into a new column which indicates the count of main camera lense(s)
9. Year announced
   Obtain the year in which the device was first announced
10. Screen-to-body Ratio
    Although GSMArena provides this metric for devices announced in the last few years, we'd prefer to obtain this number for all devices. It may be calculated based on the diagnoal screen size, the display resolution of the screen, and length and width of the phone. Simple use of Pythagoras' theorem yields the desired value.

Data Checking

Our dataset consists of all sorts of devices including tablets and smartwatches. These were discovered using filters based on screen size, available cellular networks, and year of announcement. We know that the first modern smartwatch was announced in 2011[1]. Also, tablets usually tend to be over 7 inches in screen size[2].

Removing smartwatches and tablet devices

Since we are only concerned with smart (or dumb/feature) *phones* we can filter for devices that match certain criteria:

- Remove devices with *No cellular connectivity*
- Remove devices that have screens over 6.9 inches (predominantly tablets)
- Remove devices that were announced after 2011 and have screens smaller than 1.7 inches (mostly smartwatches)

[1] https://en.wikipedia.org/wiki/Smartwatch#2010s
[2] https://en.wikipedia.org/wiki/Tablet_computer#Mini_tablet

Fix Android version number

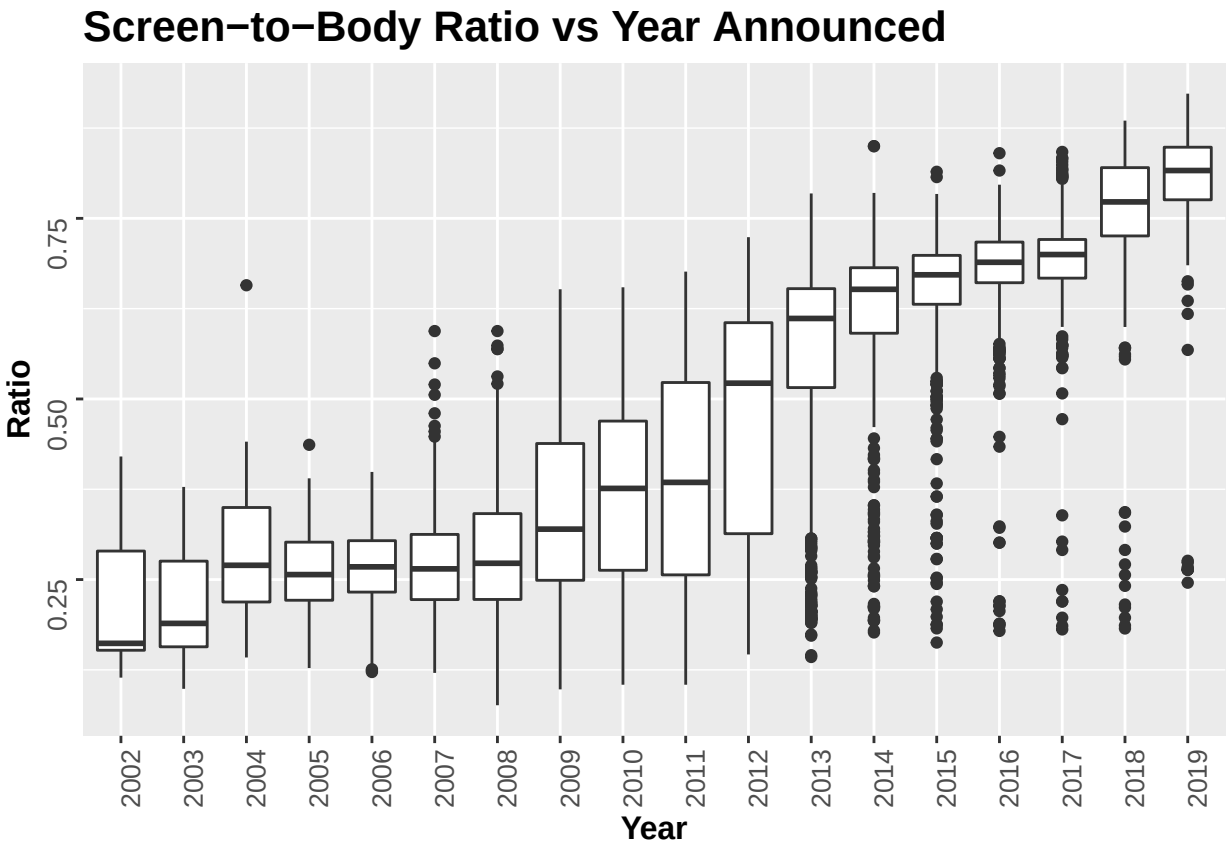
We also notice that the android versions for some of the devices are stripped down to an integer value. For example, Android 4.0 is written as Android 4 for a small subset of the devices. For all such values, we can fix them by appending a *0 (zero)* if the version number is an integer

Exploration

Part 1

Say No to Bezels

Bezel-less display is one of the many buzz words surrounding the mobile phone industry these days. The term bezel refers to the portion of a device's front surface that is not screen estate. Most mobile phones today look more or less the same with their ubiquitous rectangular profile and a display ingrained on the front. However this wasn't always the case when companies were trying to innovate on the design and usability fronts. Some of the popular designs that have gone extinct are flip-phones, phones with slide-out or front-facing hardware keyboards, devices with track-balls, and phones resembling portable gaming consoles.



The boxplot depicts a general trend of increasing screen estate over the years.

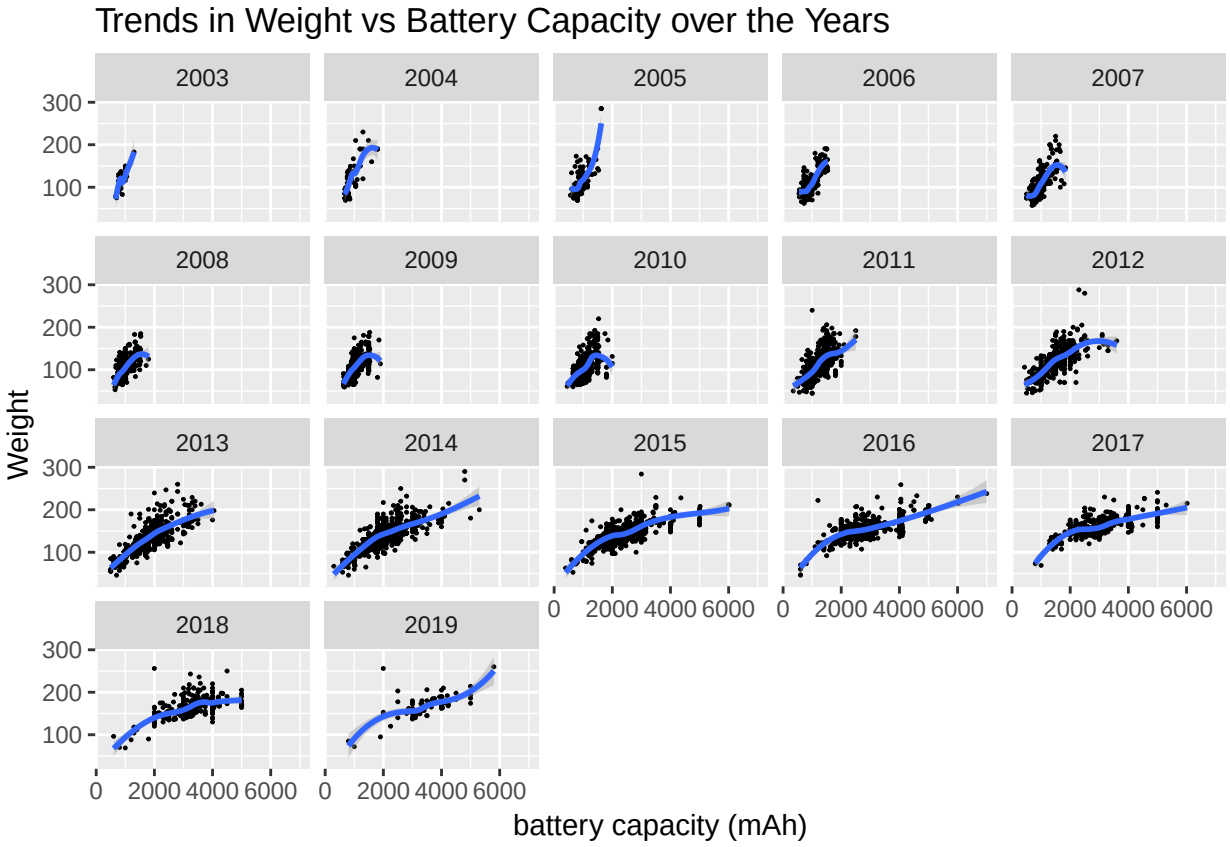
Notice how the rectangular boxes keep getting smaller in lenght as we move ahead the x-axis. This shift represents a homegenity in design which the industry seems to have embraced following in the footsteps of

the first monolithic iPhone[3]. Outliers below the first quartile also witness a drop in number, signalling a decrease in feature-phone announcements, as these are typically the devices that have both a display and hardware keypad, and thus a lower screen-body-ratio.

Part 2

Heavier doesn't always imply higher battery capacity

How long a phone can last on a single charge is an important choice factor for many people.
We explore the trends in battry capacity of devices and how it is affected by their overall weight.



In the above plot, a few trends stand out:

- The battery capacity of devices was limited to 2000 mAh until the year 2010.
- In the earlier years - until 2012 - the weight of a device rose rapidly along with the battry capacity. That is to say, the icline of the model-fit is much steeper
- From 2015 onwards, a major chunk of the devices habe battery capacities between 2000 and 4000 mAh
- Nowadays, there is a much more gradual increase in weigh as the battery capacity rises
- A majority of the devices weigh under 200 grams

[3]Citation Pending

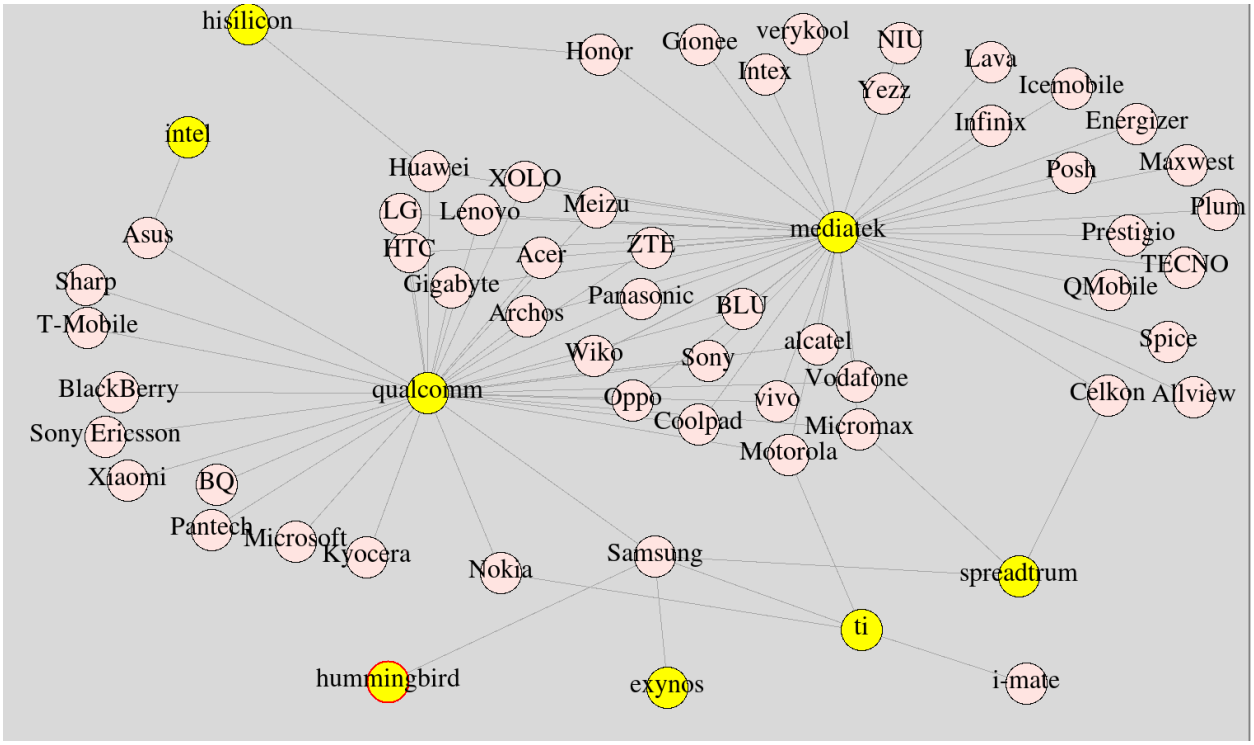


Figure 1: OEM - SoC Provider Network

Part 3

SoC Provider Network

In this section, we try to identify the relationship amongst mobile SoC (system-on-chip) manufacturers based on the their clients: viz the OEMs. While some OEMs are large enough to design their own CPUs (like Samsung and Apple), most other OEMs buy their CPU hardware from other companies. Notable major mobile SoC manufacturers are Qualcomm, Texas Instruments (TI) and Mediatek[4].

The above graph (Figure 1) was generated using the igraph R library.

Here we see the relationship between SoC providers and OEMs:

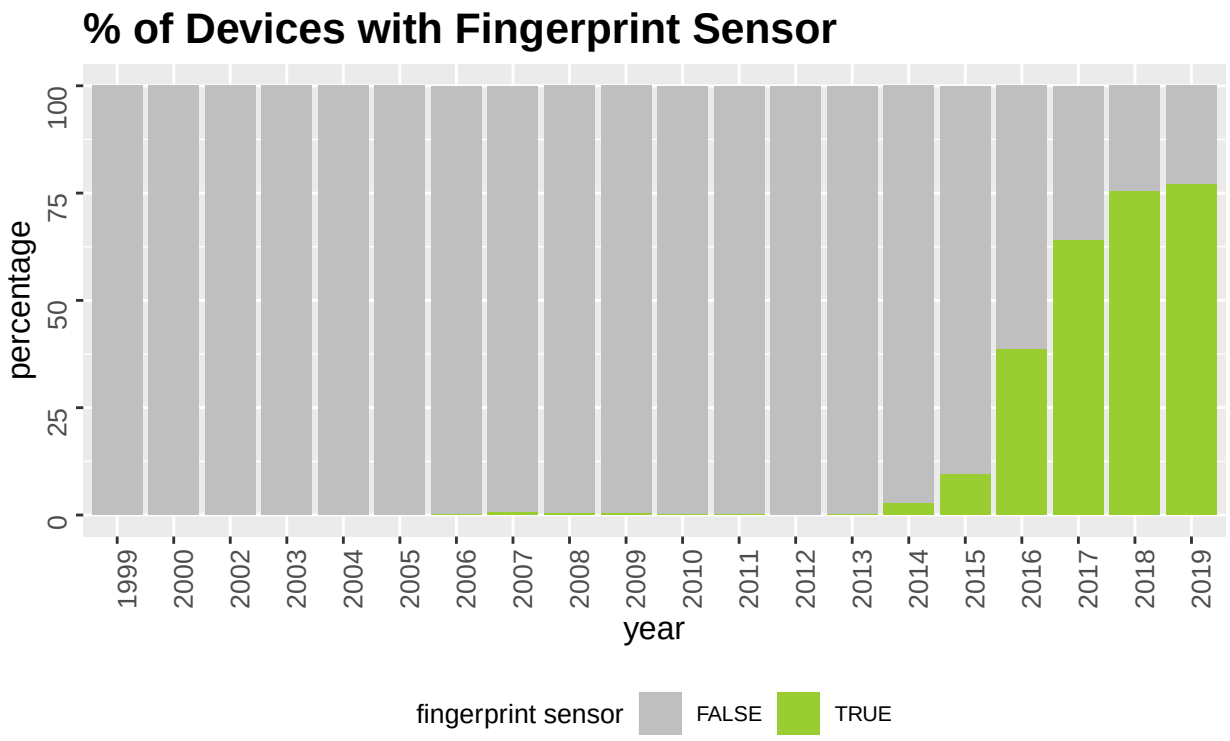
- Exynos being Samsung's in-house brand is only (yet) used by Samsung itself
- Asus is the only OEM to have ever used Intel chipsets
- Mediatek and Qualcomm are the SoC providers for a majority of OEMs. While some OEMs source their chipsets from both of them, some are loyal to either of the two.

[4] https://www.anandtech.com/show/8389/state-of-the-part-soc-manufacturers

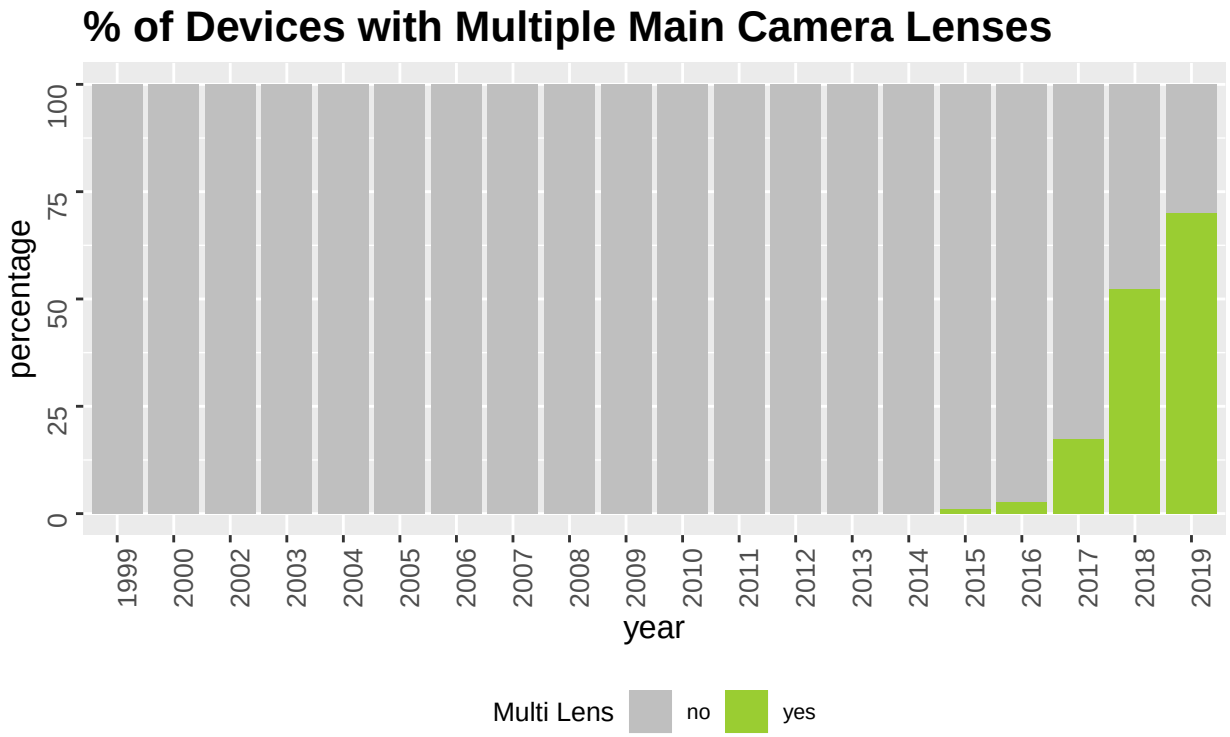
Part 4

In terms of senosrs and features, mobile phones have become extremely versatile devices. What started as a device meant for serving the purpose of calling people remotely and sending text messages, has evolved into a feature-packed gadget that comes with camera(s) for photography, GPS for navigation, fingerprint readers for secure access, and accelerometers and gyro that help in motion sense gaming. A few mobile phones even come equipped with health-related sensors like heart-rate monitors, or are IP certified to be used in extereme conditions, such as underwater.

Some of these features (or trends) have become more-or-less standardized and we will take look at two of them: multiple main camera lenses, and fingerprint readers



While fingerprint sensors first appeared on mobile devices in the year 2006, they weren't particularly popular. The technology behind them wasn't much involved either. Adoption of fingerprint sensors started gaining momentum from the year 2014.
In 2016, almost half the devices had a fingerprint reader. In 2018 and 2019, over 75% mobile phones shipped with the sensor.

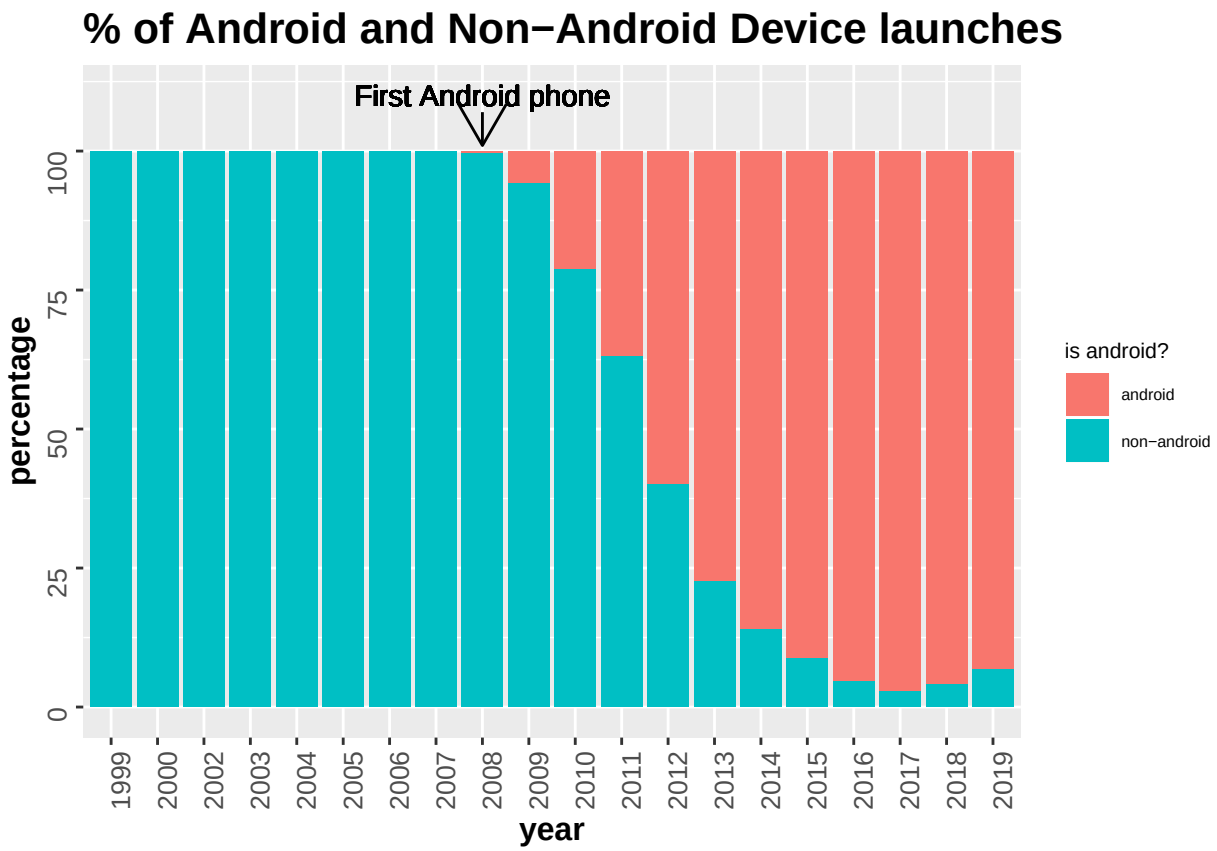


The momentum around multiple camera lenses witnesses a similar exponential growth: within 4 years of the first multi-camera setups being announced in 2015, nearly 75% devices boast of the feature in 2019. The benefits of multi-lens systems aren't particularly well laid-out, but that's just how the industry seems to operate. No OEM wants to feel left out of the trends.

Part 5

The first android device (according GSMArena's database) was released in the year 2008. It was T-Mobile's G1 running Android 1.6 (Donut). It probably marked the beginning of a new era for smartphone operating systems. Since then, Android has become wildly dominant and perhaps the only OS (apart from Apple's iOS) to survive in today's market. The likes of Symbian, Microsoft's Windows Mobile, BlackBerry OS, Palms OS, and Samsung's Bada OS have all become extinct now.

We take a look at the ratio of Android and non-Android devices announced in different years:



The above chart clearly displays the sort of dominance that Android has managed to gain in the smartphone market. In the last few of years, an overwhelming majority of mobile phones were running Android. The other devices are either running iOS, or some community-driven/ enthusiast software such as the Kai OS.

Conclusion

The questions proposed initially were tweaked slightly. This was done in order to allow for experimenting with more interesting plots (such as the SoC network graph) rather than the typical bars and lines. All the modified questions have been reasonably explored in this report.

I learnt from this project that web-scraping is a challenging task, esp. when we're dealing with thousands of iterations over web pages.

I learnt that trend adoption happens very quickly in the mobile world - take fingerprint readers and multiple camera lenses for example. It must be acknowledged that Apple may not be the first to introduce a particular technology, but it has been the trend setter when it comes to touchscreens, fingerprint sensors, and notched displays.

I also realized how dominant Android has become as a smartphone OS. The relationship between mobile SoC manufacturers is also worth pondering upon.

Overall, the mobile device industry has come a long way in the roughly 2 decades of time that it has existed for. Trends are dropped as easily as they are adopted, and staying a pioneer is difficult. This is evident from Nokia's demise and re-surgence[5].

[5]Note: Nokia's case wasn't explored in the report, but serves as a reference

Reflection

While I managed to obtain the tabular-like data from the website, I could also have included more information, such as the device poplarity and number of page hits of each device. It would have allowed for some more interesting results involving only the most popular devices.

The 93.5% success rate of scraping is nice to have but it would have been better to store the error logs for failed scraping instances. A majority of these errors, I suspect, were due to page encoding (or XML) related issues.

The dataset itself is pretty huge, and deeper analysis was possible. Most of the exploration performed here is at an year (time) level. Other grouping variables such as OEMs could have been used. I also realized that the information about a device having notched display is not available in the GSMArena tables.

References

1. Gsmarena.com. (2019). GSMArena.com - mobile phone reviews, news, specifications and more.... [online] Available at: https://www.gsmarena.com/
2. Technocracy. (2019). Mobile Device Specifications from GSMArena. [online] Available at: http://cigarplug.ml/blog/mobile-device-specifications-from-gsmarena/