Milestone Report 1: Capstone Project

What is the problem you want to solve?

Through the analysis of the phone user's features (information regarding age group, location, app usage, or gender) is it possible to determine the user's phone manufacturer? In addition through the knowledge of the user's phone features and app usage, can I determine extrapolate some of the user's details? In addition to these two questions, can we use the data to map out any correlations that may help a business make a more informed decision on advertising?

Who is your client and why do they care about this problem? In other words, what will your client do or decide based on your analysis that they wouldn't have done otherwise?

If we are capable of determining the user's phone manufacturer, we can propose a more targeted advertising direction to phone manufacturers looking to expand or enlarge sales in specific regions. It will also allow for targeted ads to be placed in apps that can be used as reinforcement of phone brand loyalty. If we are capable of extrapolating some of the user's data through phone features and app usage, then we can use that successfully predict future users and extend the same model to other populations, through that we can shift the advertising budget to appeal new users of less researched populations. An example of the former is that if we know that in Shanghai most Samsung consumers are also males in their twenties. Perhaps the client can leverage that information to purchase advertising specific only to that region (as a cost saving measure) or perhaps if we can deduce that the users of that region that have a Samsung phone typically play games, then they can cooperate with the game publishers to help create a solid brand loyalty campaign. In a different scenario, by looking at a completely new region to start a business the model can be leveraged with a less complete description of the region’s expected consumers. This may be a stretch for regions outside of China but it is mostly likely a better starting point than random guessing. The cleaned and analyzed data can also be used as a guide for entrepreneurs looking for locations to start a new business (e.g. an Apple store or Samsung store).

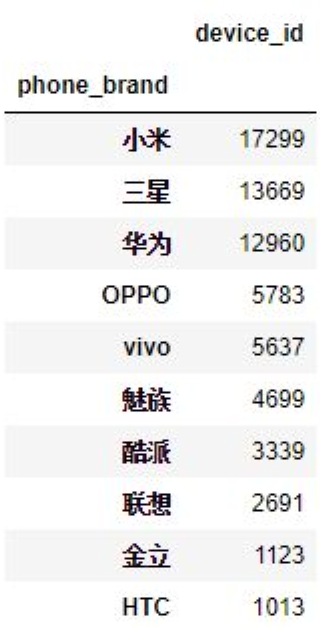
What data are you using? How will you acquire the data?

There is a kaggle dataset showing the chinese mobile user demographics which can be used with the GSMArena data (can be parsed through other people’s open sourced work).

Issues that occur with collected data

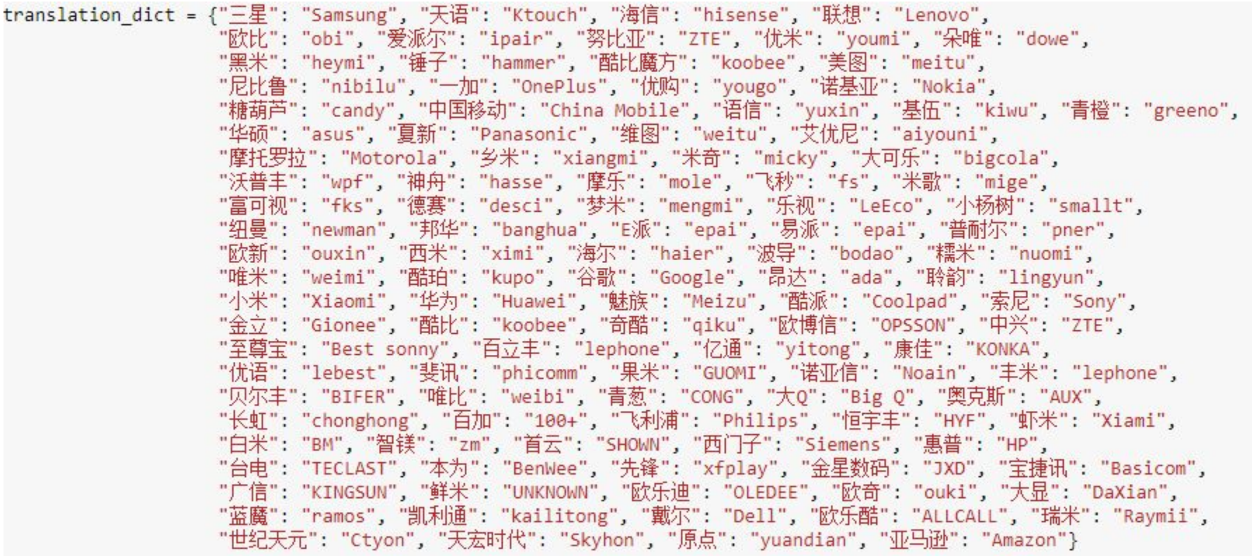
The collected data comes from TalkingData SDK and only a portion of the device_ids are matched with the event data. Also, the device_id to device_brand/model has translation problems and foreign-only releases. A lot of brands and devices cannot be matched with the gsmarena data so I had to eliminate a good portion of the available data. This analysis surprisingly does not include "Apple" as a brand which can be a result of the TalkingData SDK being unable to collect data from Apple devices or being a Android-only SDK. This analysis can only be used to give insight of the available data and not regarded as evidence to a conclusion.

The first issue that I have encounterd when dealing with the Kaggle dataset (TalkingData data) is that the data is in chinese. The displayed table shows the phone brands with the count of the users for each one.



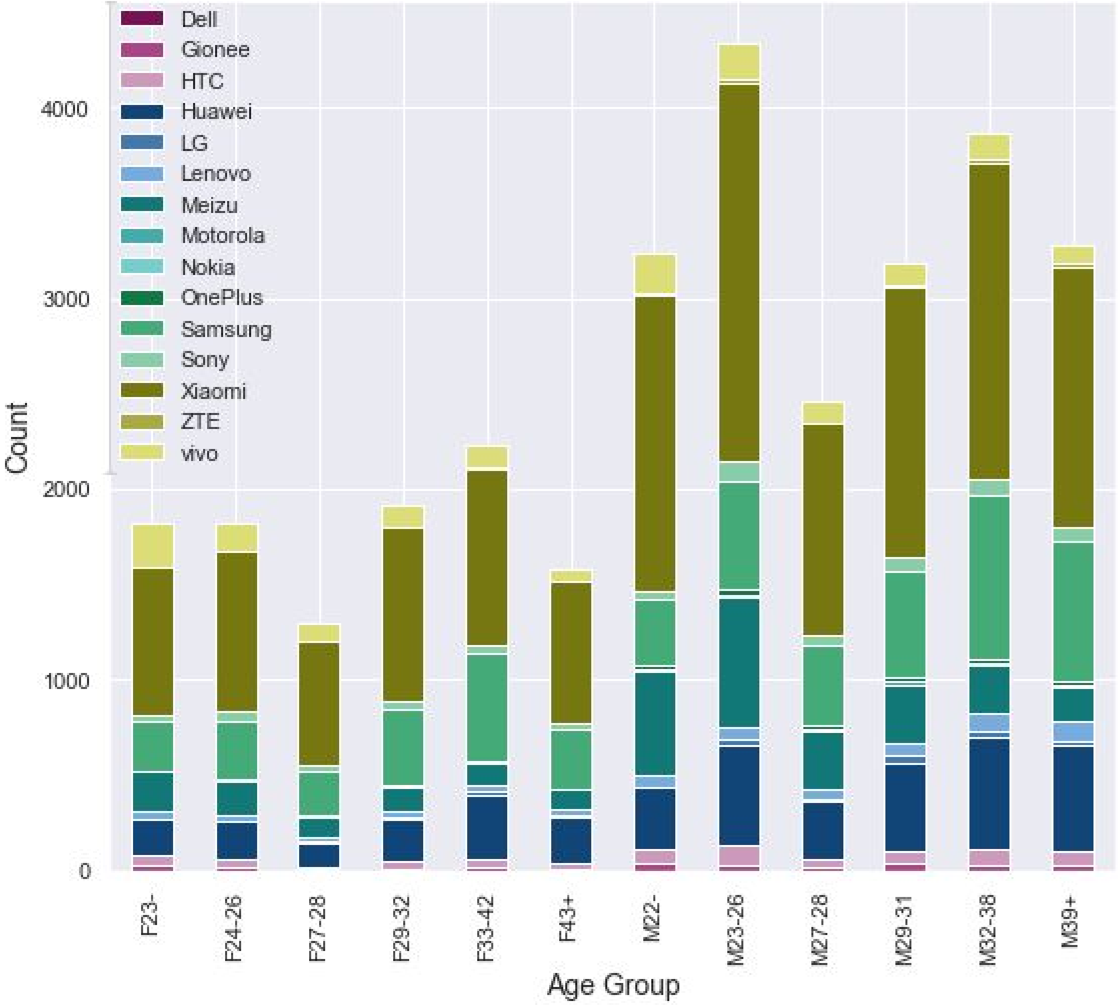
	device_id
phone_brand	
小米	17299
三星	13669
华为	12960
OPPO	5783
vivo	5637
魅族	4699
酷派	3339
联想	2691
金立	1123
HTC	1013

The method in which I deal with the problem is to through manually creating a translation dictionary to map the problematic brands to their respective english branding. The translation dictionary can be shown in the image below.



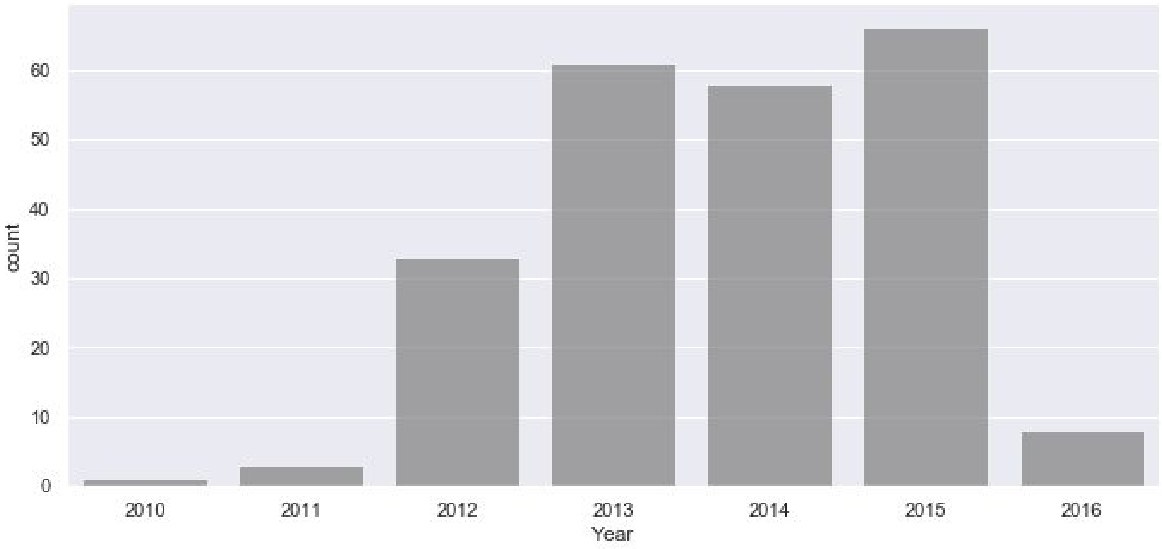
```
translation_dict = {"三星": "Samsung", "天语": "Ktouch", "海信": "hisense", "联想": "Lenovo",
                    "欧比": "obi", "爱派尔": "ipair", "努比亚": "ZTE", "优米": "youmi", "朵唯": "dowe",
                    "黑米": "heymi", "锤子": "hammer", "酷比魔方": "koobee", "美图": "meitu",
                    "尼比鲁": "nibilu", "一加": "OnePlus", "优购": "yougo", "诺基亚": "Nokia",
                    "糖葫芦": "candy", "中国移动": "China Mobile", "语信": "yuxin", "基伍": "kiwu", "青橙": "greeno",
                    "华硕": "asus", "夏新": "Panasonic", "维图": "weitu", "艾优尼": "aiyouni",
                    "摩托罗拉": "Motorola", "乡米": "xiangmi", "米奇": "micky", "大可乐": "bigcola",
                    "沃普丰": "wpf", "神舟": "hasse", "摩乐": "mole", "飞秒": "fs", "米歌": "mige",
                    "富可视": "fks", "德赛": "desci", "梦米": "mengmi", "乐视": "LeEco", "小杨树": "smallt",
                    "纽曼": "newman", "邦华": "banghua", "E派": "epai", "易派": "epai", "普耐尔": "pner",
                    "欧新": "ouxin", "西米": "ximi", "海尔": "haier", "波导": "bodao", "糯米": "nuomi",
                    "唯米": "weimi", "酷珀": "kupo", "谷歌": "Google", "昂达": "ada", "聆韵": "lingyun",
                    "小米": "Xiaomi", "华为": "Huawei", "魅族": "Meizu", "酷派": "Coolpad", "索尼": "Sony",
                    "金立": "Gionee", "酷比": "koobee", "奇酷": "qiku", "欧博信": "OPSSON", "中兴": "ZTE",
                    "至尊宝": "Best sonny", "百立丰": "lephone", "亿通": "yitong", "康佳": "KONKA",
                    "优语": "lebest", "斐讯": "phicomm", "果米": "GUOMI", "诺亚信": "Noain", "丰米": "lephone",
                    "贝尔丰": "BIFER", "唯比": "weibi", "青葱": "CONG", "大Q": "Big Q", "奥克斯": "AUX",
                    "长虹": "chonghong", "百加": "100+", "飞利浦": "Philips", "恒宇丰": "HYF", "虾米": "Xiami",
                    "白米": "BM", "智镁": "zm", "首云": "SHOWN", "西门子": "Siemens", "惠普": "HP",
                    "台电": "TECLAST", "本为": "BenWee", "先锋": "xfplay", "金星数码": "JXD", "宝捷讯": "Basicom",
                    "广信": "KINGSUN", "鲜米": "UNKNOWN", "欧乐迪": "OLEDEE", "欧奇": "ouki", "大显": "DaXian",
                    "蓝魔": "ramos", "凯利通": "kailitong", "戴尔": "Dell", "欧乐酷": "ALLCALL", "瑞米": "Raymii",
                    "世纪天元": "Ctyon", "天宏时代": "Skyhon", "原点": "yuandian", "亚马逊": "Amazon"}
```

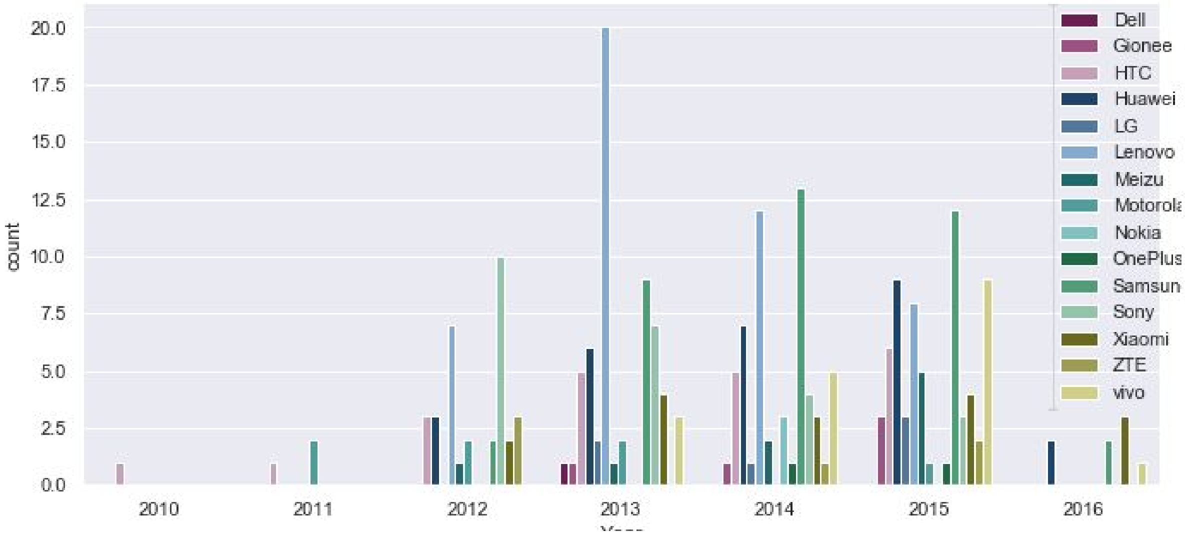
After that mapping, I quickly discovered that many of the brands that are listed do not sell in America. This causes me to dump all of the brands that I cannot get the equivalent phone data with on GSM Arena. I then use the remainder of the data to to display graphs showing the market share that each phone manufacturer had by age group and gender.



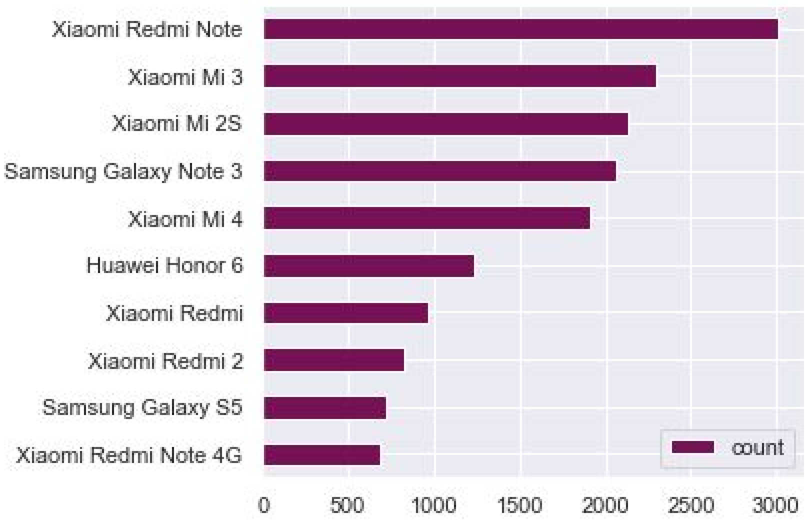
Xiaomi seems extremely popular with male adults but less so than female adults. This graph displays two inherent issues with my data. The first being that a lot of the phone brands hold very little of the market. This will create a large class imbalance making it extremely difficult to create an accurate model for the lesser known brands. The second that there is an unexplained deficiency in female users.

After merging the data together with the GSM dataset we can determine the age of the devices that each user has and aggregate the data to create a plot displaying the year in which the devices were released. This will give us insight to when our data was taken and the timeframe in which our data was collected.





Another important aspect of the data we want to see is which devices are the most popular in China. The following graph displays the top ten most popular android devices based off our data:


Xiaomi Redmi Note
Xiaomi Mi 3
Xiaomi Mi 2S
Samsung Galaxy Note 3
Xiaomi Mi 4
Huawei Honor 6
Xiaomi Redmi
Xiaomi Redmi 2
Samsung Galaxy S5
Xiaomi Redmi Note 4G
count
0
500
1000
1500
2000
2500
3000