Homework Assignment 04 | Group Homework

Anna Tatinyan, Hovhannes Torosyan, Hovsep Avagyan, Knarik Manukyan, Rozi Tunyan

July 23, 2018

The problem/data description

In the scope of this project we decided to analyze the website https://www.gsmarena.com/makers.php3 and with web-scraping get the information about the characteristics of phones available there. The website contains 116 brands of phones with different models and characteristics for each. The initial scraped dataset's structure was the following: *10020 observations and 35 variables*. To match the dataset to our needs we modify the columns and clean the data from NAs and non-valid values.

In the first part of the assignment we perform multiple regression analysis to find out what are the factors that impact the phones' price. For that purpose we have subsetted the dataset of 900 observations and 11 features, which by intuition, can have a significant positive or negative impact on the price.
The resulting subsetted dataset has the following variables: *release_year, website_views, display_size, display_pixel_height, display_pixel_width, camera_megapixels, video_pixels, battery_mah_size, price, ram_mb_size, memory_mb_size*. Each variable is self-explanatory, so further on in regression analysis and statistical visualizations each of them will be used without explanations.
The regression analysis consists of three parts. In the first part we take all of the features that were formerly selected, afterwards we improve the model by picking and testing different independent variables. Finally, we take two of our independent variables and construct their interaction term and interpret the results.

Research questions

1.What are the factors that have significant impact on the phones' price.

Regression analysis

Before doing the analysis we examined some of our variables and put some limitations on them in order to avoid issues with outliars. The *price* of the phone was limited to 750 dollars as there were some phones that costed as much as 30000 dollars, *camera_megapixels* was limited to be maximum 20mp and *ram_mb_size* was limited to be at most 7500MB. Now let's have a look at our first model.

```
##
## Call:
## lm(formula = price ~ display_size + camera_megapixels + release_year +
##     website_views + battery_mah_size + ram_mb_size + memory_mb_size +
##     display_pixel_height + display_pixel_width + video_pixels,
##     data = model3Data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -178.91  -44.69  -10.12   33.19  363.64
```

```
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            5.536e+04  2.829e+03  19.567  < 2e-16 ***
## display_size           2.202e+00  4.034e+00   0.546 0.585284
## camera_megapixels      2.414e+00  9.599e-01   2.515 0.012077 *
## release_year          -2.749e+01  1.407e+00 -19.534  < 2e-16 ***
## website_views         -8.125e-07  3.651e-07  -2.225 0.026298 *
## battery_mah_size       1.075e-02  3.625e-03   2.967 0.003091 **
## ram_mb_size            1.659e-02  2.839e-03   5.843 7.18e-09 ***
## memory_mb_size         1.111e-03  1.672e-04   6.645 5.29e-11 ***
## display_pixel_height   2.399e-02  1.358e-02   1.767 0.077613 .
## display_pixel_width    4.184e-02  1.094e-02   3.825 0.000140 ***
## video_pixels           3.824e-02  9.992e-03   3.827 0.000139 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 71.67 on 888 degrees of freedom
## Multiple R-squared:  0.4896, Adjusted R-squared:  0.4838
## F-statistic: 85.17 on 10 and 888 DF,  p-value: < 2.2e-16
```

According to the results of the regression, the only statistically insignificant independent variable is the *display_size*. *display_pixel_height* is significant in 95% confidence level and one unit increase of that variable implies, on average, 2.399e-02 dollars increase in price, all other things fixed. *camera_megapixels* is significant in 99% confidence level and one unit increase implies, on average, 2.414e+00 dollars increase in price, all other things equal. *website_views* also is significant in 99% confidence level, but surprisingly the analysis shows, that one unit increase of the website views decreases the price of the phone by 8.125e-07 dollars. On 99.9% confidence level we can say that one unit increase of *battery_mah_size* increases the price by 1.075e-02 dollars. The rest of the variables e.g *release_year*, *ram_mb_size*, *memory_mb_size*, *display_pixel_with* and *video_pixels* have a significance level of 99.(9)%, and one unit increase of *release_year* decreases the price by 2.749e+01 dollars, one unit increase of *ram_mb_size* increases the price by 1.659e-02 dollars, one unit increase of *memory_mb_size* increases the price by 1.111e-03 dollars, one unit increase of *display_pixel_width* increases the price by 4.184e-02 dollars, one unit increase of *video_pixels* increases the price by 3.824e-02 dollars. The negative influence of the release year, website views and the fact that the intercept takes the value of 5, meaning that the phone price will start from 5 dollars if all of the explanatory variables take value 0 is not intuitive and cannot be explained properly, so the model does not accurately represent the reality. The standard error rates were also high for the explanatory variables which indicates that the data shown is not that accurate. Different approache was adopted in order to get better results, so some of the variables were excluded. The results of the improved regression model are illustrated below.

```
##
## Call:
## lm(formula = price ~ display_size + camera_megapixels + battery_mah_size +
##     ram_mb_size + memory_mb_size + video_pixels, data = model3Data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -196.78  -61.90  -14.76   46.92  436.44
##
## Coefficients:
##                      Estimate Std. Error t value     Pr(>|t|)
## (Intercept)       72.0460891 18.5424856   3.885     0.000110 ***
## display_size      -0.5225950  4.7047364  -0.111     0.911579
## camera_megapixels -0.2764611  1.0160900  -0.272     0.785621
```

```
## battery_mah_size   0.0141098  0.0044008   3.206     0.001393 **
## ram_mb_size        0.0074276  0.0034017   2.183     0.029261 *
## memory_mb_size     0.0006853  0.0001985   3.453     0.000581 ***
## video_pixels       0.0664453  0.0115608   5.747 0.0000000124 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 88.22 on 892 degrees of freedom
## Multiple R-squared:  0.2232, Adjusted R-squared:  0.2179
## F-statistic: 42.71 on 6 and 892 DF,  p-value: < 2.2e-16
```

After improving the model it consists of 6 independent variables from which 4 have significant impact on the price. Those are *battery_mah_size*, *ram_mb_size*, *memory_mb_size* and *video_pixels*. The RAM size is significant on 99% confidence level and the data shows that one unit increase of the RAM size implies, on average, 0.0074276 dollars increase in price, all other things equal. Battery size is significant on 99.9% confidence level and one unit increase of it has, on average, a positive impact of 0.0141098 dollars on the price, all other things fixed. Memory size and video pixels are significant on 99.(9)% confidence level. One unit increase in memory size shows, on average, an increase of 0.0006853 dollars in price of the phone and one unit increase of video_pixels increases the price, on average, by 0.0664453 dollars, all other things equal. From the data we see that the value of the intercept is 72, with the error rate of 18, meaning that the price will differ between 65 to 90 dollars if all our independent variables are 0. With this model we have also decreased the error rates and can conclude that, in comparison with the previous model, this one shows more accurate results. The issue with this model is that it represents the *camera_megapixels* as a non-significant variable but we have that *video_pixels* has a high significance level. Besides this, we found out that camera megapixels and video pixels have dependancy, which means that here we have an interraction effect. To deal with this problem, in our next model the interaction term of the two independent variables is included.

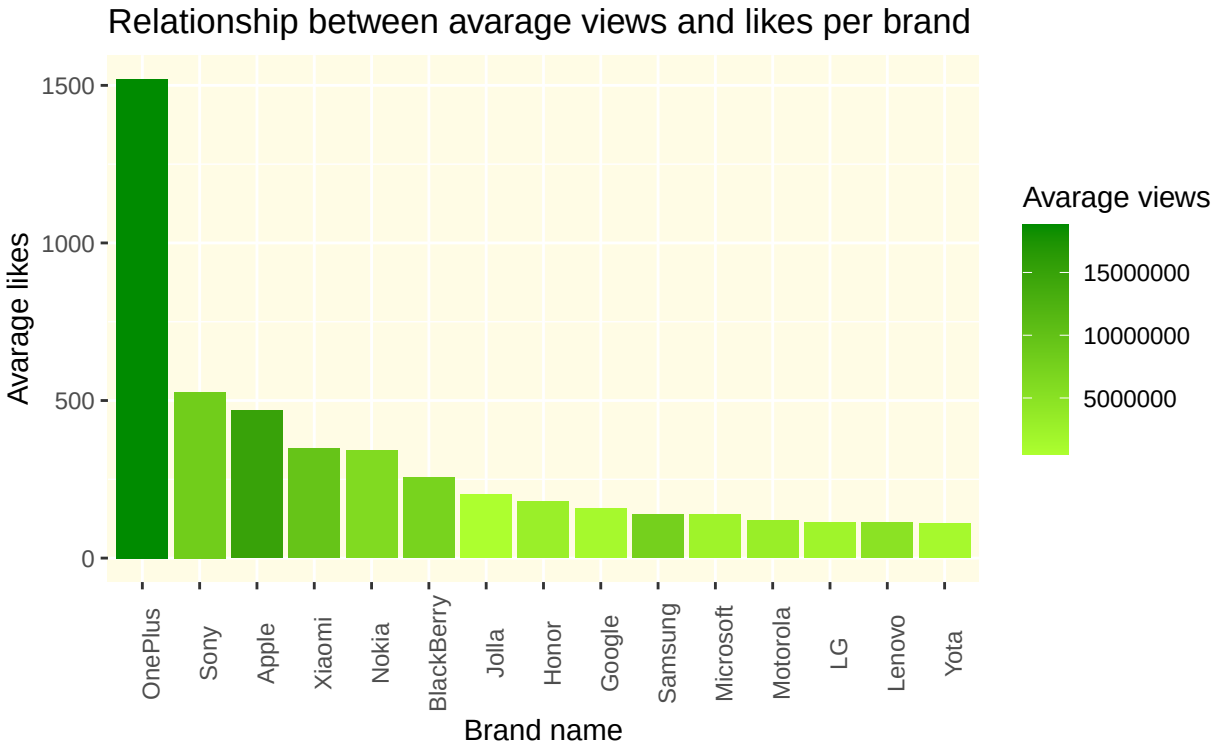
```
##
## Call:
## lm(formula = price ~ camera_megapixels * video_pixels + display_size +
##     battery_mah_size + ram_mb_size + memory_mb_size, data = model3Data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -194.85  -61.51  -16.28   47.19  437.49
##
## Coefficients:
##                                  Estimate  Std. Error t value Pr(>|t|)
## (Intercept)                   104.8850543  27.1542732   3.863 0.000120 ***
## camera_megapixels              -3.4842879   2.1890423  -1.592 0.111808
## video_pixels                    0.0264067   0.0268214   0.985 0.325120
## display_size                    0.5347553   4.7434407   0.113 0.910265
## battery_mah_size                0.0137632   0.0044015   3.127 0.001824 **
## ram_mb_size                     0.0073376   0.0033989   2.159 0.031129 *
## memory_mb_size                  0.0006615   0.0001988   3.327 0.000913 ***
## camera_megapixels:video_pixels  0.0033650   0.0020345   1.654 0.098482 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 88.13 on 891 degrees of freedom
## Multiple R-squared:  0.2255, Adjusted R-squared:  0.2195
## F-statistic: 37.07 on 7 and 891 DF,  p-value: < 2.2e-16
```

In the final model we observe the same independent variables and one interaction term of camera megapixels

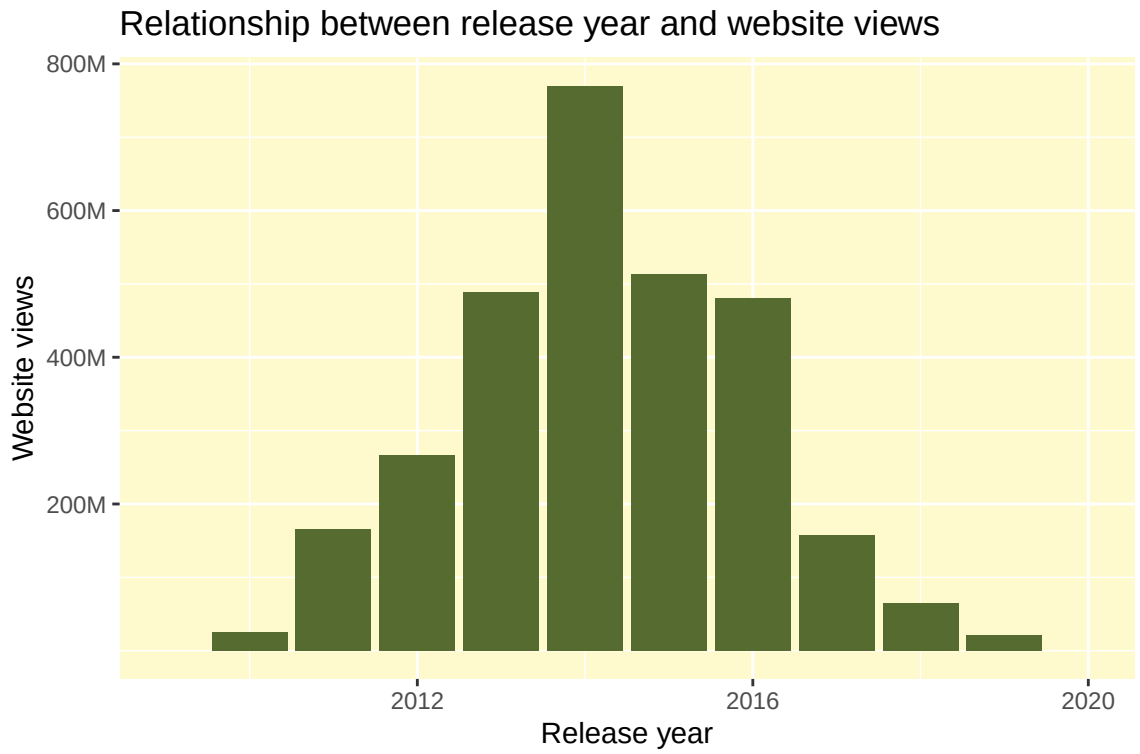
and video pixels. Here we have almost the same results with slight differences. On 99.9% confidence level we have that one unit increase of battery size shows 0.0137632 dollars increase in price, on 99% confidence level we have 0.0073376 dollars increase in price and on 99.(9)% confidence level we see that one unit increase of memory size increases the price by 0.0006615 dollars, on average, all other things equal. The interaction term is significant on 95% confidence level and one unit increase of it increases the price by 0.0033650 dollars, on average, all other things fixed. The Intercept here is 104 with the error rate of 27 which means that the price will differ between 77 and 131 dollars if our explanatory variables have 0 value. With this model we also made an attempt to minimize the error rates in order to have more accurate results. However, the model could be further improved and analized for getting better results.

Statistical plots/visualizations

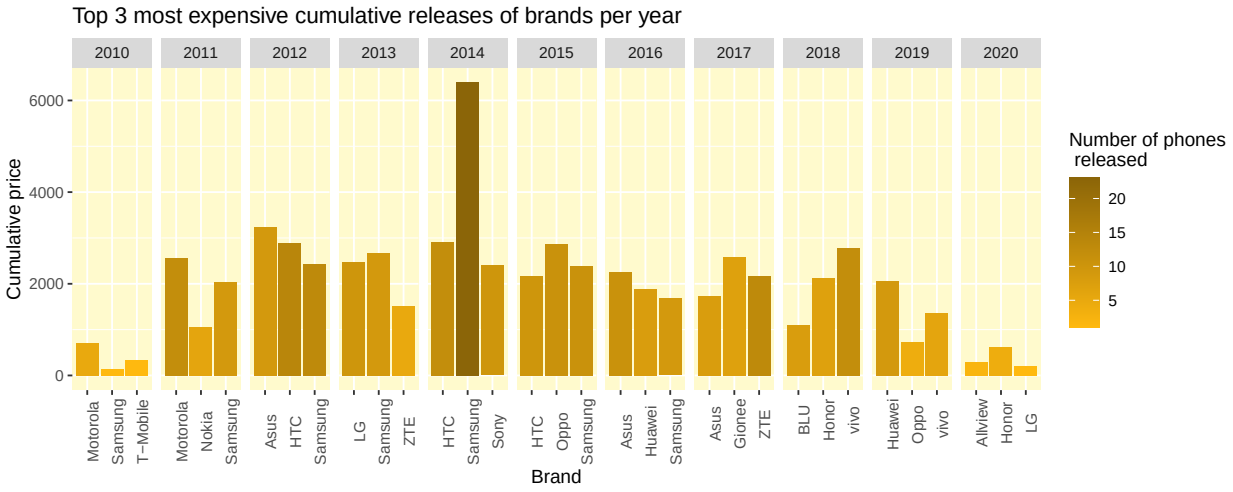
Now let's have a look at some plots and interesting visualizations regarding phones and their characteristics.



The graph shows the relationship between the average website views and likes of phones for the top 15 brands with the most likes. We analyzed to see whether the average number of likes and views are proportional to each other or not. The brand with the highest number of average website likes is OnePlus, which we can see has the highest average number of website views. In general, the fewer are the number of likes the fewer is the number of views, however, for brands with a close number of likes, this can differ. The second one by the number of likes is Sony and the third is Apple, however, Sony has fewer views than Apple, which can indicate that the users of SOny are more prone to liking the phones than Apple users. From the brands with fewer likes take Samsung, which is close to Google and Microsoft. In case of Microfost and Samsung the number of likes and views are proportional, however in case of Google and Samsung, again, we see a different pattern, as Google has more likes but fewer views.
So, we can say that in general scope the number of views and likes are proportional, but in smaller scopes there can be deviations.

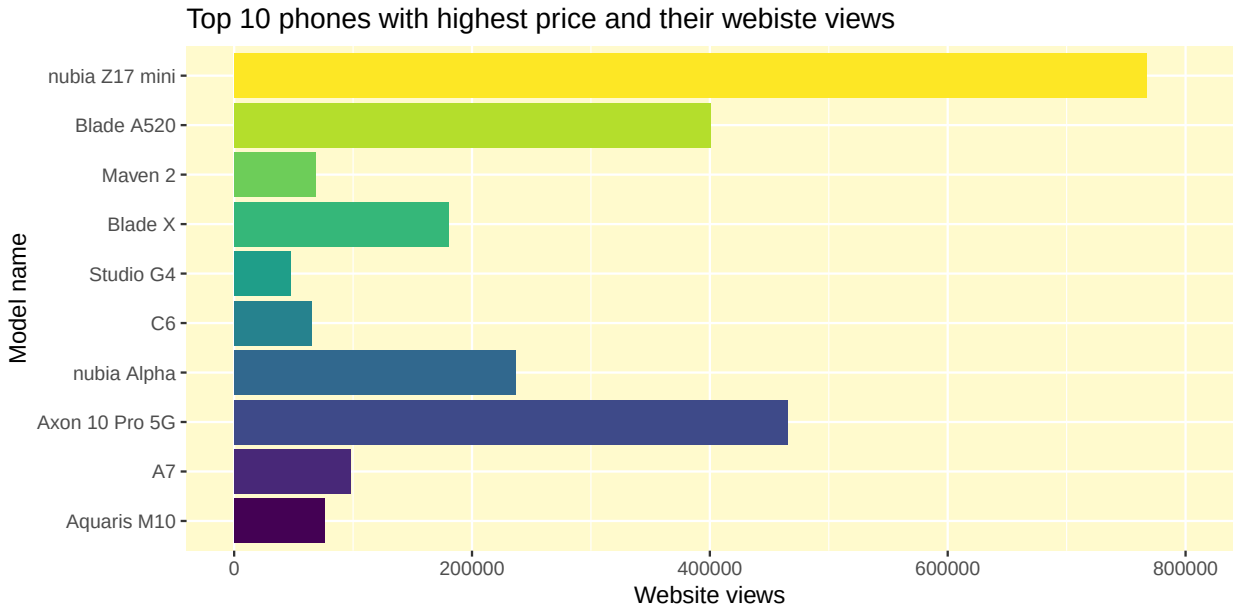


The plot represents the relationship of phones released between 2009 and 2020 and their website views. We see that in gradually the viwes of websites started to increase and in 2014 it reached its peak. Afterwads, it gradually started to decrease and the graph looks like almost a normal distribution.
Apparently in 2014 there was some kind of change or boom in the industry.
It was estimated that smartphones will capture almost 69% of global mobile phone sales in the market so it may be the reason of so massive phone website views.
The gradual decrease of the graph could be explained as following: smartphones became a common practice and people started to be less interested in the industry.

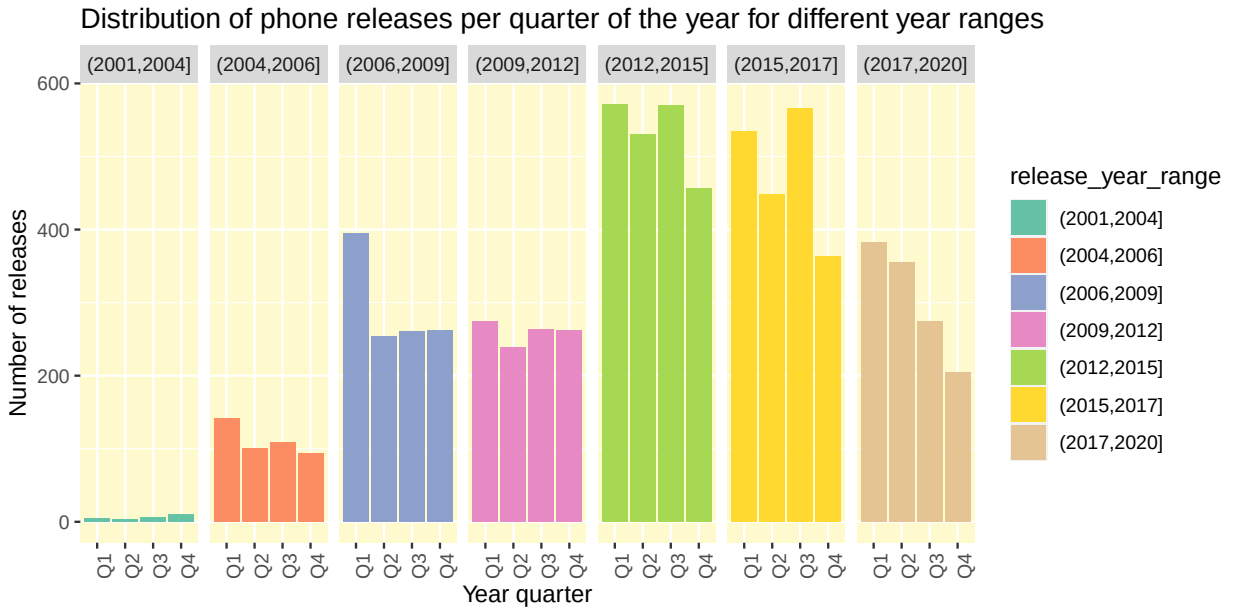


The plot shows the top 3 brands with the highest cumulative release price per year.The highest cumulative price means the overall price of all newly released phones. Also, the bars are colored according to the number of phones released. We can see that the cumulative price per brand increased from 2010 to 2014 and then decreased. Samsung is the only brand that appeared in top 3 for 7 years. ALso, it's the highest with

the cumulative price in 2014, and during all years from 2010 to 2020, in 2014 the number of released phones of Samsung is the highest. For the other years we can notice that in general the higher is the cumulative price the higher is the number of released phones.

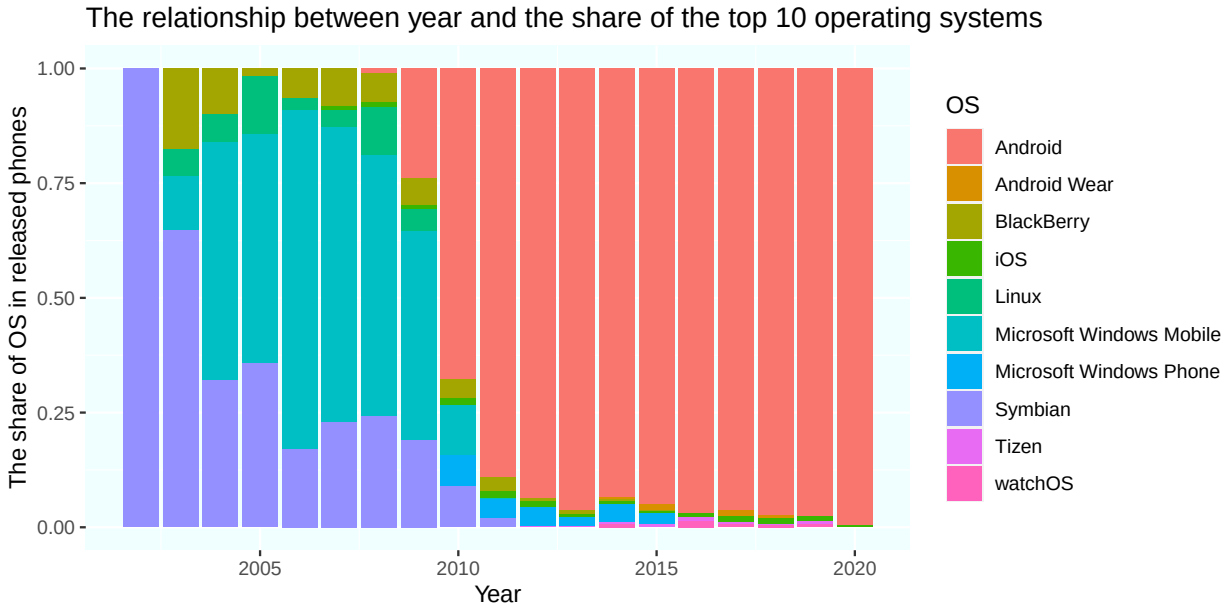


This plot represents the top 10 phones with the highest prices. It is ordered in descending order, so the phone with the highest price is at the bottom (*Aquaris M10*) and the phone with 10th highest price is at the top (*nubia Z17 mini*). From this graph we can see that there isn't a big correspondence between the price and the website views. Even though, *Aquaris M10* has the hgihest price it has low website views record; whereas *nubia Z17 mini* has the biggest number of website views in our list of top 10 pricy phones.

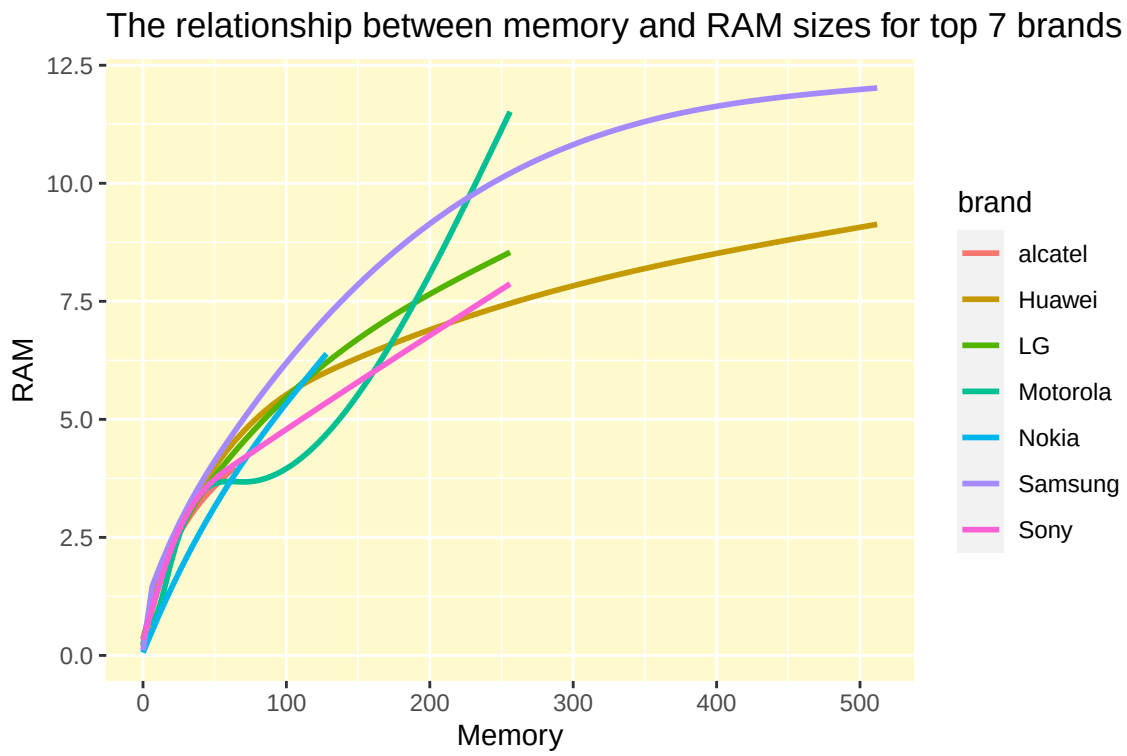


The plot shows how many phones were released per quarter of a year for each 3 year range starting from 2001 to 2020.We can notice that in general the number of releases increased from 2001 to 2012 years. The number of releases per quarter year is approximately the same for each range, except for the first quarter of

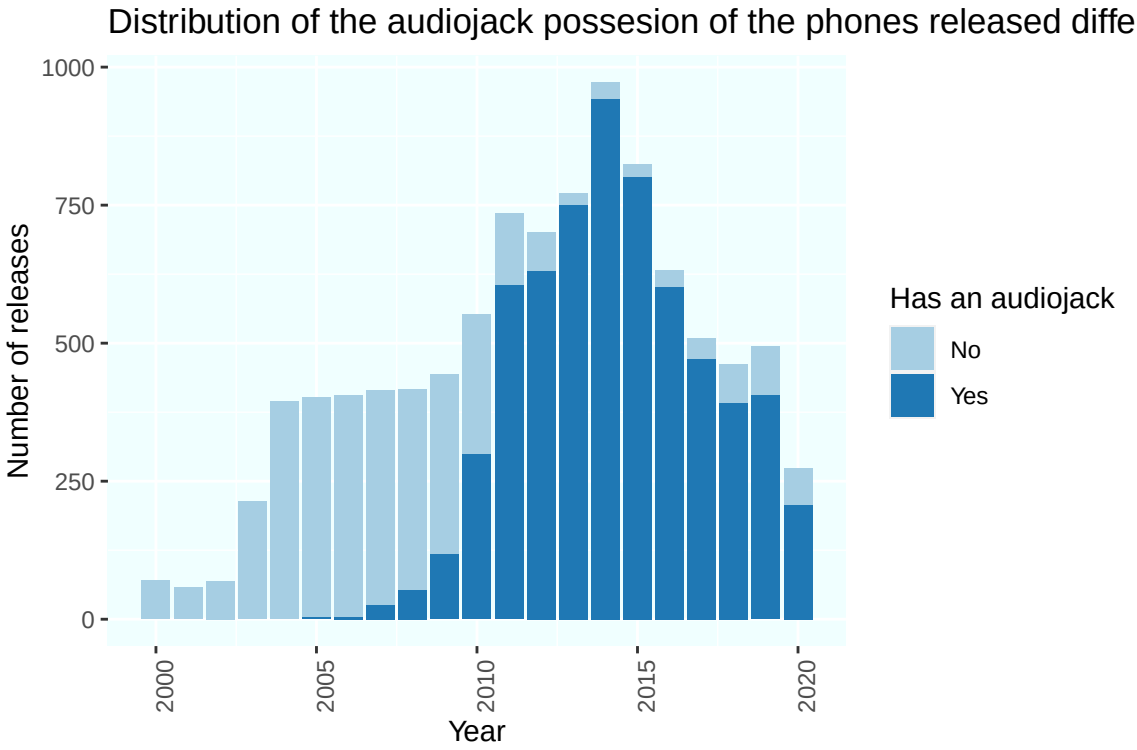
years 2006-2009, there was a peak of phone releases. The number of phone releases extremely increased in years 2012-2017. We can notice that for both ranges (2012-2015 and 2015-2017) the first and third quarters of years are more active in terms of phone releases, which means that most of the phone releases were at the begging of the years and beginning of the second half of the years.Second quarter for the range (2012-2015) is less active, but the least releases were by the end of the years. For the range (2015-2017) again the least releases were by the end of the years, however, there were about as many releases during the second quarter as for the fouth quarter of the range (2012-2015). So, this can indicate that during (2015-2017) years number of releases were tending to be dropped , which is proved by the last range of years, because obviously the number of releases started dropping by the end of the 2017 year and the overall number of releases dropped dramatically during (2017-2020) years. Again, we can see that least active for phone releases is the end of the years. However for the range (2017-2020) we can say that the number of releases gradually decreases from the beginning of the year, that is more phone releases are done during the first half of the year than the second. For the years 2001-2020 we can say, that until 2012 the number of releases increases, and reaches it's peak during 2012-2015 years, after that it again gradually decreases getting closer to the number of phone releases during 2006-2012 years.



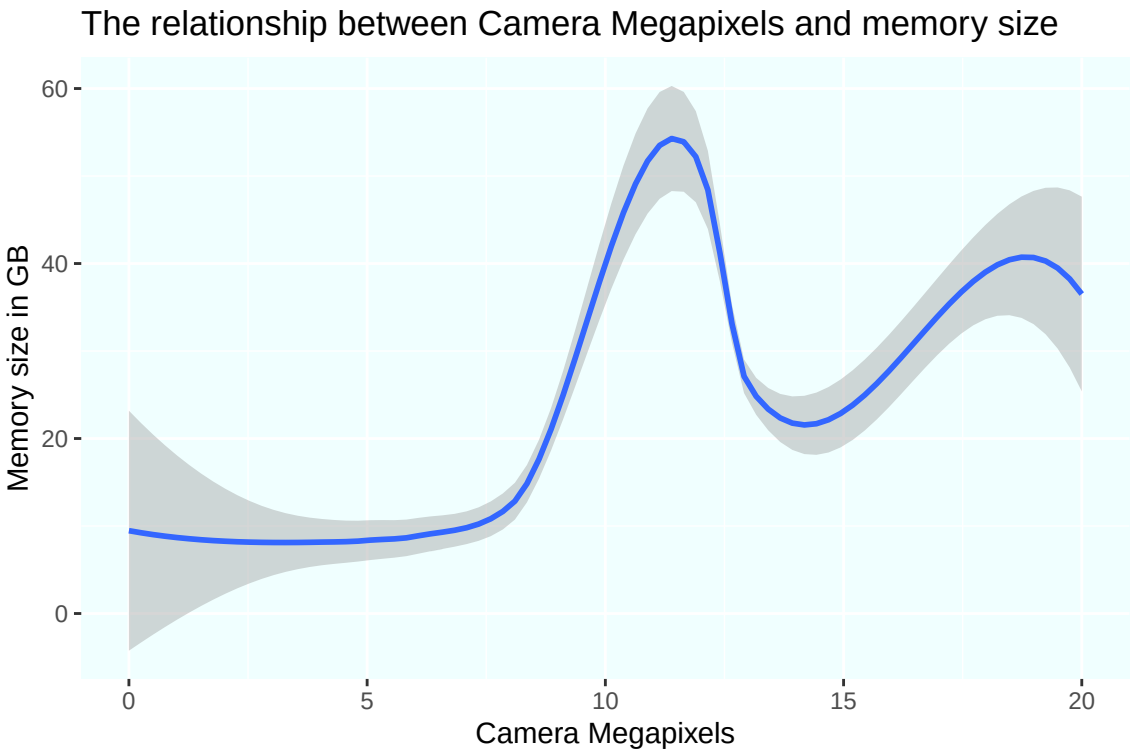
This plot represents the relationship between year and the share of the top 10 operating system, which released phones use. In 2002 almost all devices used Symbian, which was mostly the operating system of Nokia. After 2002 Symbian loses the dominance in the sphere and new operating systems like BlackBerry, Linux and Microsoft Windows Mobile appear. In 2009 there are more phones running Microsoft Windows Mobile than Symbian. Symbian runs on less than 25% of all devices of that year. First Android phones were released in 2009 and starting from 2010 it started to dominate more and more. In 2008 phones with IOS first start to be released, but the share of IOS doesn't increase in the future, as there is only one company using it and it releases few phones in a year. In 2020 there are only two main operating systems - Android and IOS. In this year Android reached it's maximum share of devices which is almost 100 percent.



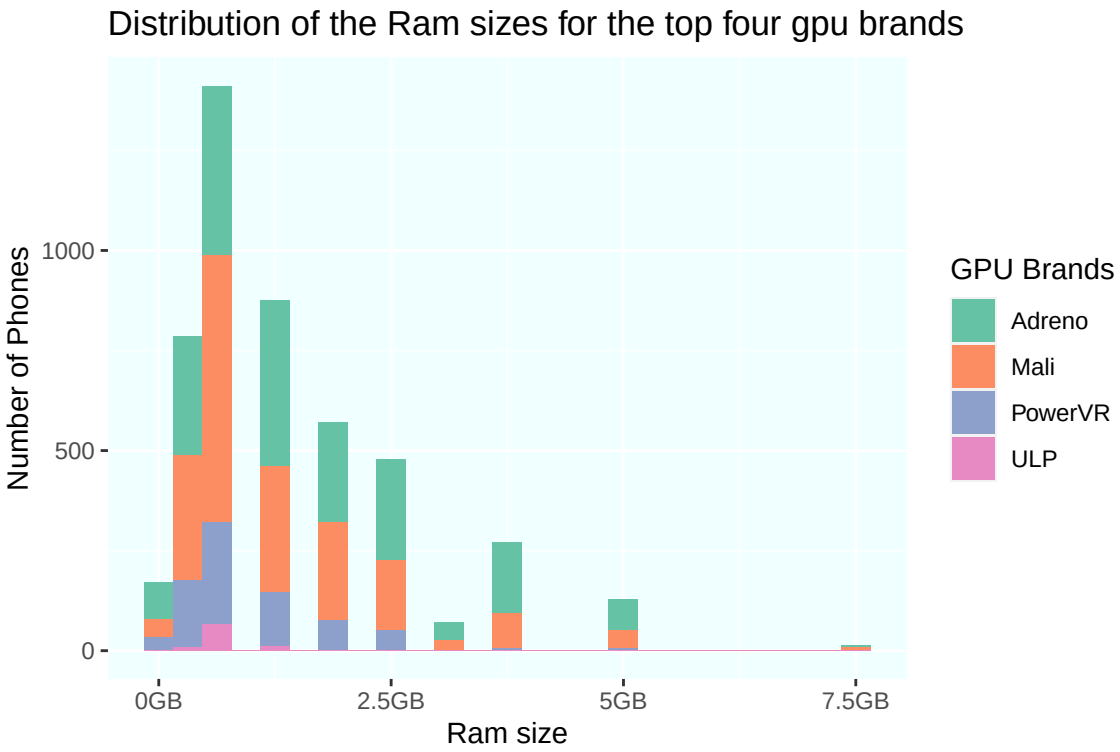
This plot shows the relationship between memory and RAM sizes for top 7 brands by the number of their devices. All companies for their new model increase both the size of RAM and Memory In this plot we can identify if memory and RAM sizes increase by the same pace. For Nokia the graph is almost linear. So the RAM and memory size increase with the same pace. However Nokia by the maximum sizes of RAM and memory is behind 5 brands. For LG and Sony it's easier to increase the memory than RAM. Samsung and Huawei have the maximum memory size, which is over 500 GB. However, they struggle with increasing the RAM. Their logarithmic graphs show that there is a limit of RAM that they still can not pass. The most interesting graph has Motorola. In this case approximately from 50 to 100 GB devices have the same size of RAM. After that point it increases exponentially. At approximately 250 GB it passes all the brands by having more RAM for less memory.



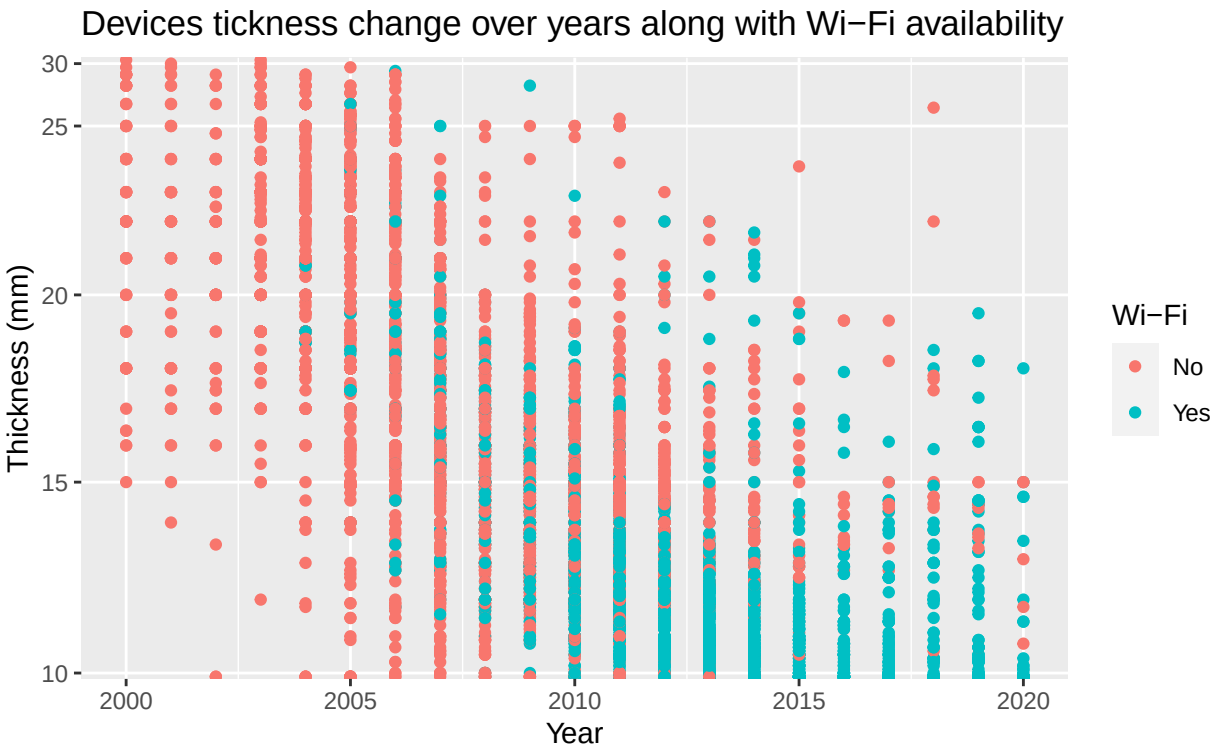
The plot shows how many of the released phones for different years have an audio jack. From the graph, we see that audio jacks appeared in the phone industry in 2005. After that, the percent of the phones with audio jacks started to gradually increase each year.Most of the released phones during 2013-2015 had audio jacks. However, after 2015 the percentage starts to decrease gradually. If before 2013, fewer phones were with audio jacks, because the technology wasn't advanced enough, after 2015 the number decreases because other new technologies started to be used instead of audio jacks. Also, from the graph we can notice that number of releases phones increases starting from 2000 year and reaches its peak in 2014, after which it decreases again. So 2014 is the year with the highest number of phone releases.



This plot shows the relationship between Camera Megapixels and memory size in GegaBytes. All brand tend to have higher and higher megapixels for cameras and sometimes have more than on camera. This also effects on the price of the device. If the camera Megapixels are high, than the memory size should be bigger to be able to store pictures and videos with high quality. As we can see from this plot from 0 to 7.5 megapixels the memory is near 10 GB. after 8 megapixels the memory size increases reaching it's maximum 60 GB but then decreases until 14 megapixels. After that we see an increase and a decrease again till the maximum 20 megapixels. It's intuitive that you get more memory space to take better pictures, but companies do not always do.

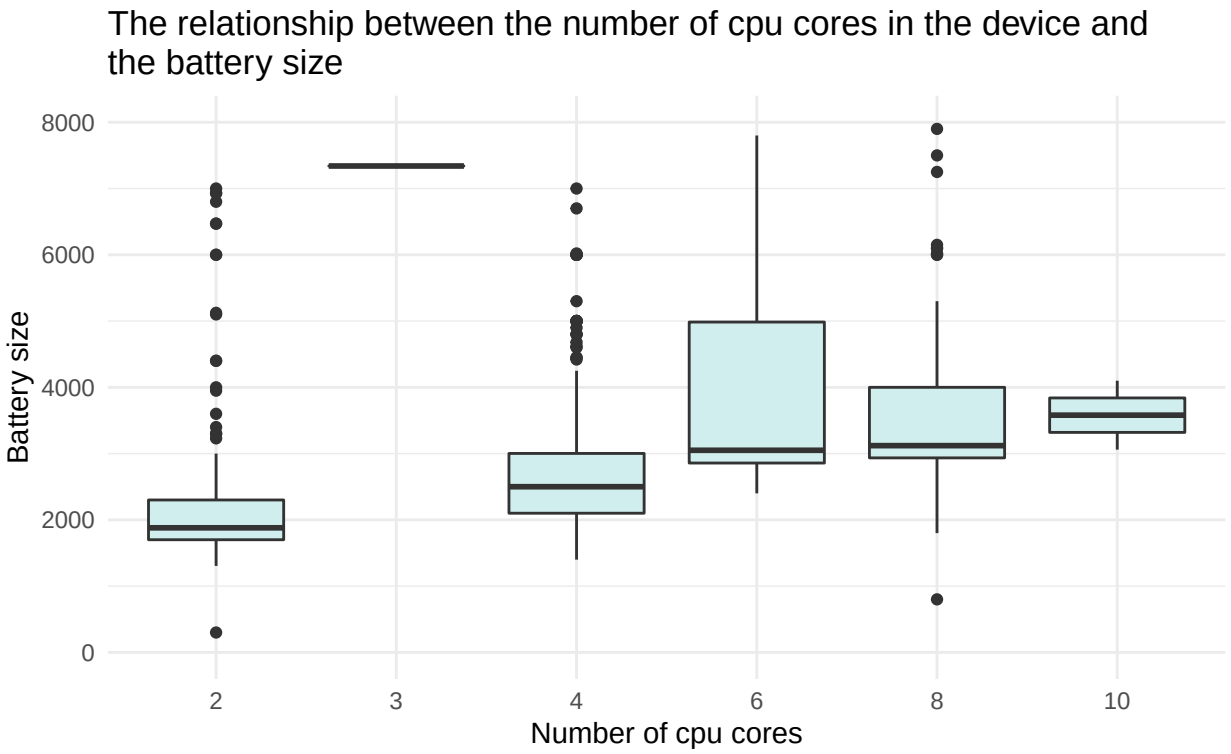


The plot shows the distribution of the ram sizes of the released phones that have GPUs of the most used 4 brands. From the graph we can see that the most widespread ones are Adreno and Mali. PowerVR and ULP are used mainly for phones with a small RAM size. So we can say that Adreno and Mali are more powerfull as they require more RAM. Also, the majority of phones have RAM up to 2.5 GB, only small part of the phones have more RAM size.



The above scatter plot is showing the change of devices' thickness/thinness over years with the marking of wi-fi availability/unavailability. The devices considered for the visualization are taken starting from year 2000 and with thickness less than 35mm (3.5cm, pretty big though), as there are some extraordinary exceptions like big black-boxes with a wi-fi receiver inside. The graph helps us prove 3 different assumptions:
- Mobile devices get thinner year by year
- Mobile devices got wi-fi chipsets starting from about 2008
- Thinner devices are more expected to have Wi-Fi connectivty rather than thicker ones*

* Even though in 2010-2015 there were many devices with Wi-Fi, but they always represented the thinner group (see the period of time and Wi-Fi colored marking on the plot).



This plot represents the relationship between the number of cpu cores in the device and the battery size. As the number of cores increases there is more and more tasks that the device can complete concurrently, so it will use more battery and the phone with multiple cores should have bigger battery size. Let's take a look at this boxplot. We can notice that the least battery size have devices with 2 core cpus. It's weird that 3 core cpu has the highest battery size, so this is an outliar. after 4 to 10 core cpus the battery size increases. The 8 core cpu is the double of 4 core cpu, but the difference between the mean of those two battery sizes is not significant. The device with 10 core cpu have 5 times more cores than 2 core cpu ones, however the battery size is not even the double.