This is a document containing the result of than analysis of company XYZ a telecommunication company for the country in question. The focus of this analysis is to explains some factor responsible for customer transferring from our network to our competition, our clients contract types and what is responsible for them having the contract type they are on, how we can get them to buy into our services, retain them, and also reduce the amount of contract terminations and arrears from our clients.

Two datasets are provided with one tagged DDM\_22\_26\_customer with 12 columns and 8337 rows indicating the data for 8337 clients and 12 data points er client. The second data set tagged DDM\_22\_26\_SCORE as 2 columns and 2350 meaning the joining of the two data sets will introduce missing data as some customers were not catered for in in the DDM\_22\_26\_score dataset. This analysis presents the companies current standing with some visualization, tables, explanation of what they entail and some suggestions from the analytics angle to make the company more profitable and increase customer retention.

We have 219 data points in total missing in the dataset which puts it at 0.002189037 ratio of missing data i.e., 0.2% of the data points missing from the DDM\_22\_26\_customer dataset which is very small compared to the whole dataset so we can remove all rows containing missing data as it’s effect on our analysis is minimal and negligible. We the joint our two data sets on the cid column

data <- left\_join(DDM22\_26\_customer, DDM22\_26\_score)

Joining, by = "cid"

There are 8118 distinct cid meaning we actually have 8118 unique customers on the table and some customers are repeated which we will work with like that without removing them from our dataset. The DDM\_22\_26\_score as 2350 unique cid which is the exact number of row the data as which mean we do not have repeated scoring for our clients.

Not all our data point are continuous variables as demonstrated below

summary(DDM22\_26\_customer)

gender age citizenship tenure

Length:8118 Min. : 15.00 Length:8118 Min. : 0.00

Class :character 1st Qu.: 31.00 Class :character 1st Qu.: 36.00

Mode :character Median : 41.00 Mode :character Median : 59.00

Mean : 42.81 Mean : 63.63

3rd Qu.: 53.00 3rd Qu.: 83.00

Max. :104.00 Max. :555.00

contract\_type data\_allowance payment\_method monthly\_charges

Length:8118 Length:8118 Length:8118 Min. : 0.00

Class :character Class :character Class :character 1st Qu.: 20.00

Mode :character Mode :character Mode :character Median : 33.00

Mean : 33.03

3rd Qu.: 50.00

Max. :103.00

payment\_arrears contract\_terminated number\_transferred cid

Min. : 0.000 Length:8118 Length:8118 Min. : 1

1st Qu.: 0.000 Class :character Class :character 1st Qu.:2086

Median : 0.000 Mode :character Mode :character Median :4170

Mean : 2.454 Mean :4169

3rd Qu.: 4.000 3rd Qu.:6255

Max. :26.000 Max. :8337

To cater for the categorical feature we check t=for the number of unique variables we have.

sapply(DDM22\_26\_customer, n\_distinct)

gender age citizenship tenure

3 85 6 170

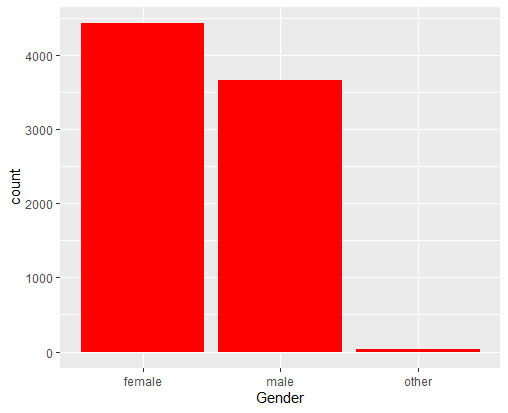
contract\_type data\_allowance payment\_method monthly\_charges

3 6 3 89

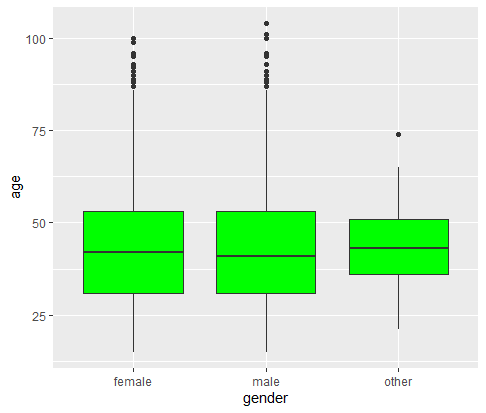
payment\_arrears contract\_terminated number\_transferred cid

27 2 2 8118

The tables above show a summary statistic of our entire table with min as minimum, 1st Qu as first quantile, Median, mean, 3rd Qua as third quantile and max as maximum for the continuous variables i.e., the variables that are number while the second table shows the number of unique variables in each variable to cater for the categorical variable i.e., the variables that are non-numbers.



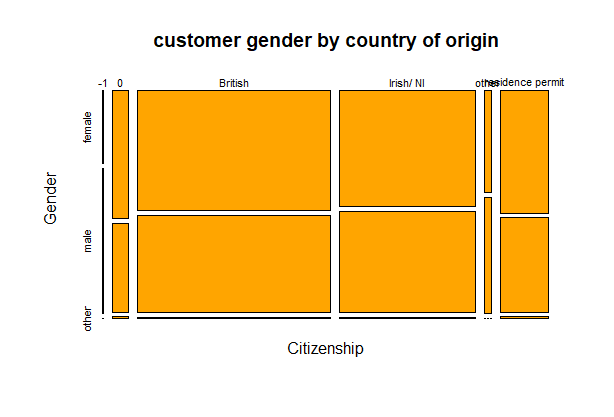
The plot above shows the bar chart of the different gender of our customers

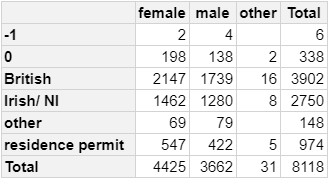


This show a box plot of the different genders against age which show a whole lot of outliers and since we have too many outliers we can’t just remove them from our dataset so we should probably not work with the age data for our analysis

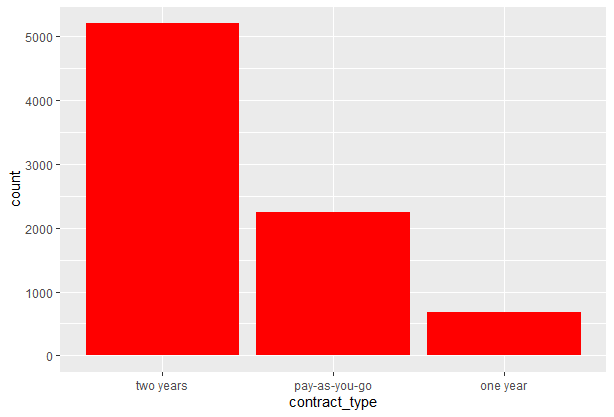
We did a pivot table of the ages as against the sex which shows the age data seems to follow the normal distribution as the there are more people with age 19 to 60 compared to other ages of our customers.

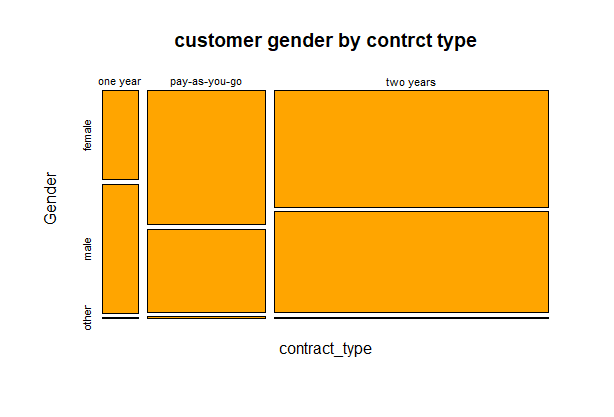
We removed the age and the cid variables majorly cos there are a whole lot ou=f outliers in the age column and the cid is just the customer id which we’ve already used to join our tables.

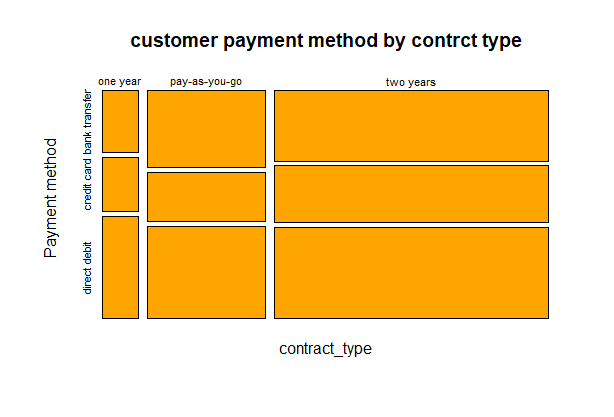


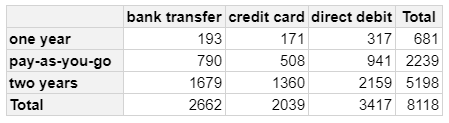


The plot and table above show our customer distribution by gender and country

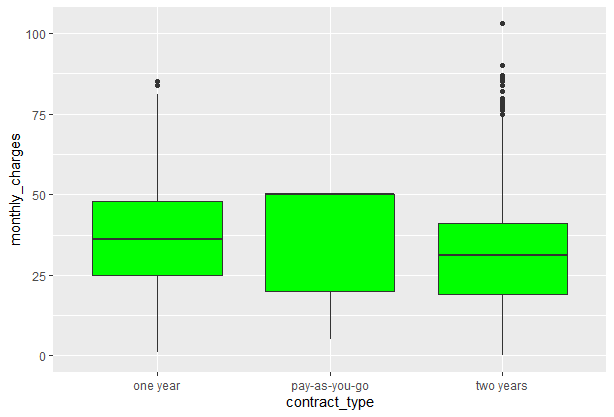




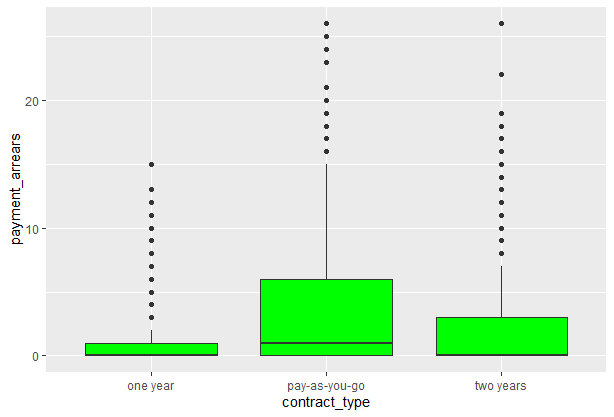


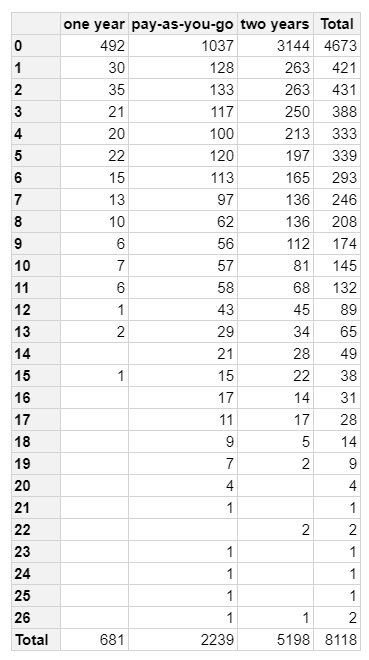


The plot above shows the distribution of the different contract types and those with 2 years contract type are the more which is a good sign. Then we have the distribution by gender and payment method as well where we see we have more direct credit then bank transfer and credit card being the least as can be seen the table picture.

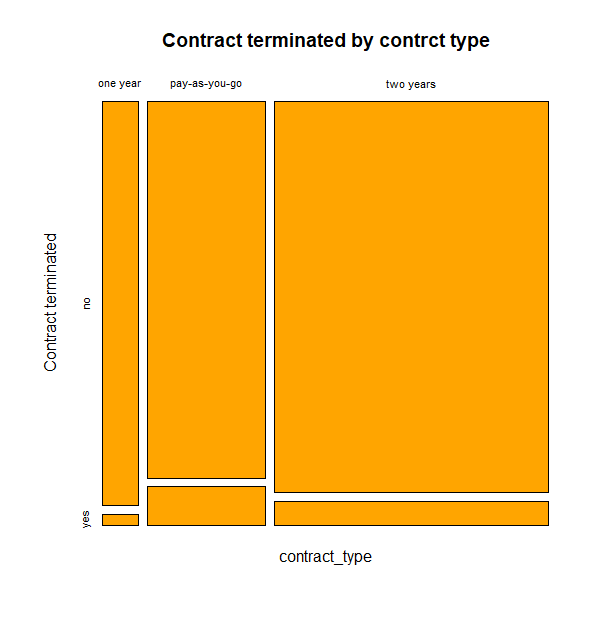


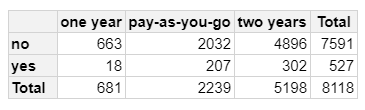
The plot above shows we a whole lot of our most paying customers in the 2 year contract type which is a really good sign but it is not worthy to say the we have a whole lot of customers in the oner year contract type and it will be nice have them to subscribe for a longer period of time and the least of all is our pay as you go customer who might just be trying our services to see if they should subscribe or not, it will be advisable our sales representative recommend the two years contract type when the eventually want to subscribe and this plot shows we have a really great customer retention rate. The information on the number of customers can be seen on the pivot table available on the R script.



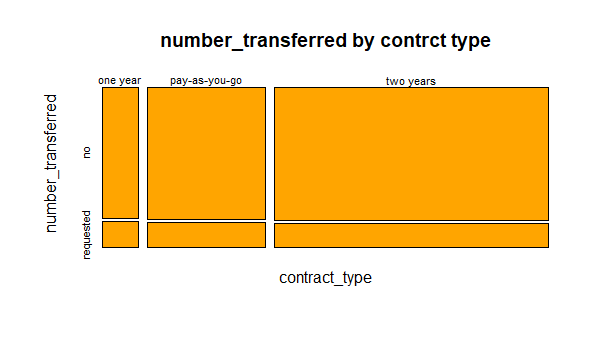


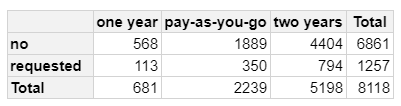
The plot above shows payment arears of each our contract types and as it can be seen all the categories have areas but the intriguing part is the pay as you go customer have a whole lot of areas which might be indicative of the fact that the accounting department needs to seat up and take there responsibilities more seriously.



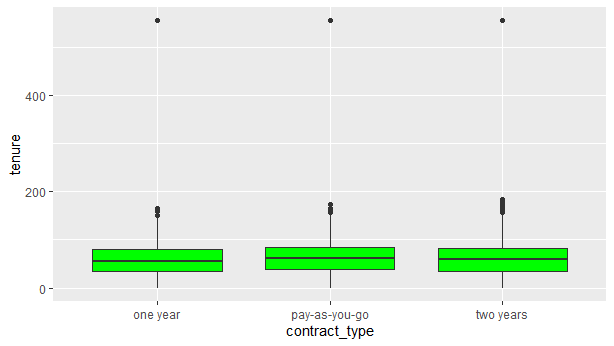


The plot ad table above shows the distribution of contract terminated by contract type and it can be seen less than 10% of our paying clients terminated there contracts but they are more than 5% which be indicative of the fact that some of our competitors have something we might need to work on to retain more of our clients.

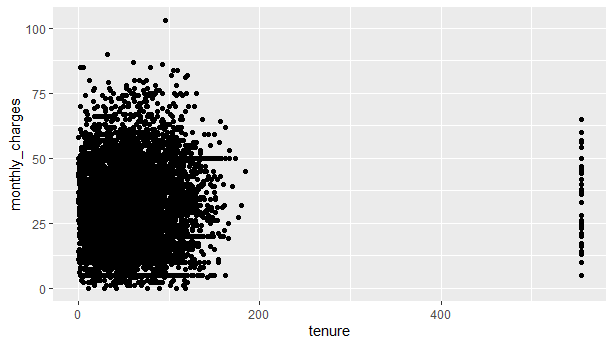




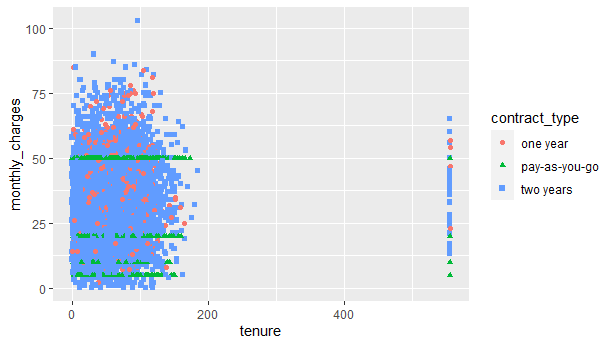
The table above shows the distribution of number of transferred by contract type and shows over 20% of our clients have requested to be transferred which is indicative of what we already detected in the previous post which is that our competition as some features we need to do something better than.

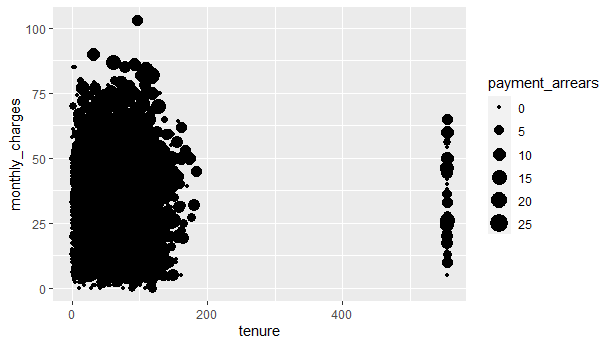


The tenure plot against contract type is a little confusing as it shows the company as existed for over 500 months but went out of operation for over 300 months.

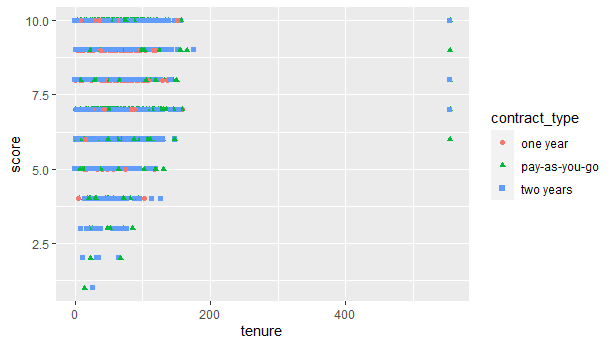


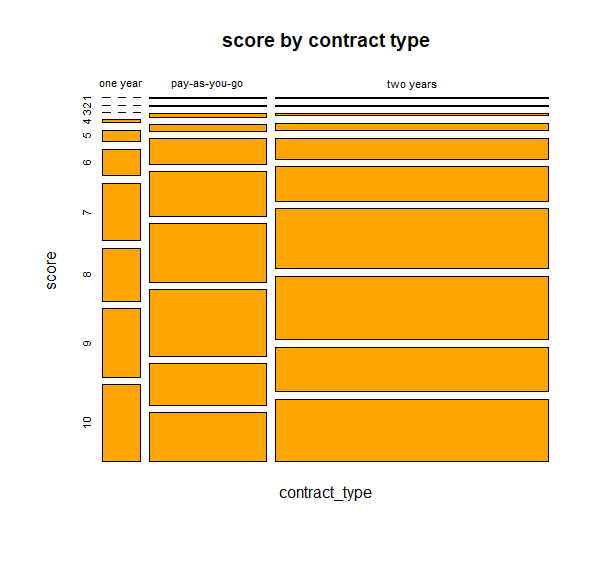
This plot shows the distribution of tenure as against monthly charges and still emphasizes the fact the company does not have data for over 300 months of operation.

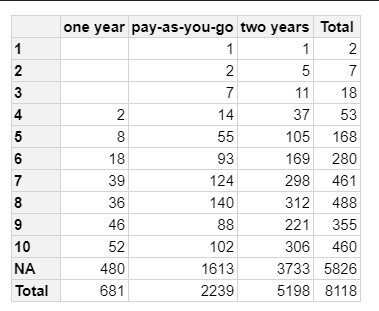




The plot above shows the distribution of monthly charges against tenure and hued by contract type.







The table above shows the comparison of tenure against score and hued by contract type which shows the higher the score the more the number of people subscribed on the network and also increases from one year to pay as you go to two years.