**Executive summary**

To assist firms, make more data-driven decisions, business intelligence includes business analytics, data mining, data visualization, data tools and infrastructure, and best practices. On this work, a well detailed analysis is don using the dataset provided. The two datasets are of different number of entries, but with the help of the R script the two datasets are been marge together. Irrelevant column like click was deleted from the dataset and all missing dataset are been dealt with. A summary statistic is performed on the dataset and we also checked the count of values of each of the categorical column on the dataset. Visualizations is done on each of the data column using bar chart and histogram. Finally, a performance dashboard is presented at the end of the work.

**Background**

We are been commissioned to analyse the customer database of a mobile phone provider and derive recommendations for actions to improve customer loyalty. The dataset provided is of two types which are been tagged *DDM 22 19 customer* and *DDM 22 19 score*. The goal of this analysis is to analyse the management problem using data analytics. We are to analyse the provided dataset and visualize the current performance of the company. In addition, a performance dashboard based on the analysis will be provided and the dashboard will be calibrated on the provided historical data.

**Data preparation**

Data restriction: from the work analysis done with the R script on the two dataset provided tagged *DDM 22 19 customer* and *DDM 22 19 score,* it’s been discovered that the *DDM 22 19 customer* dataset has 13 columns and 7513 entries (rows) which indicates data for 7513 customers with 13 data points (features) for each of the customers. The other dataset tagged *DDM 22 19 score* has just two data points (features) in it and entry values (rows) of 2078. This implies that if the two datasets are being magged together it will definitely result in missing data because some of the customers presented on the first dataset are not present on the second dataset.

The company current state with some graphical representation, table and important explanations of what is involved is stated in this report, also analytics-based recommendations to boost profitability and client retention is given.

|  |  |
| --- | --- |
| Variables | Metrics |
| Gender | The gender is categorical variable containing 3 different categorical values namely: female, male and others, having 7512 entries (rows) in total with a single missing data |
| Age | The age is an integer variable having 7513 entries (rows) with none missing entries |
| Citizenship | The citizenship is a categorical variable containing 7 different categorical values namely: British, Irish/NI residence permit, 0, other empty space and -1, having 7513 entries (rows) in total with no missing data. |
| Tenure | The tenure is a float type variable having 7476 entries having just 37 missing entries |
| Contract type | The contact type is a categorical variable containing 3 different categorical values namely: two years, pay-as-you-go and one year, having 7513 entries (rows) in total with no missing values |
| Data allowance | The data allowance is a categorical variable containing 5 different categorical values namely: unlimited, 10 GB, 5GB, 1GB and no data, having 7513 entries (rows) in total with no missing values |
| Payment method | The payment method is a categorical variable containing 3 different categorical values namely: direct debit, bank transfer and credit card, having 7513 entries (rows) in total with no missing values |
| Monthly charges | The monthly charges is an integer variable having 7513 entries (rows) with none missing entries |
| Payment arrears | The payment is an integer variable having 7513 entries (rows) with none missing entries |
| Contract terminated | The contact terminated is a categorical variable containing 3 different categorical values namely: no, yes and idiot, having 7375 entries (rows) in total with 138 missing values |
| Number of transferred | The number of transferred is an integer variable having 7513 entries (rows) with none missing entries |
| Click | The click is a categorical variable having just 15 entries with 7360 missing entries |
| cid | The cid is an integer variable having 7513 entries (rows) with none missing entries |

The table below shows the summary statistics which includes the mean, median, minimum and maximum etc. of the business intelligence of the dataset

|  |  |  |
| --- | --- | --- |
| **Gender**  Length:7513  Class: character  Mode: character | **Age**  Min.: 13.00  1st Qu.: 31.00  Median: 41.00  Mean: 42.57  3rd Qu.: 52.00  Max. :101.00 | **Citizenship**  Length:7513  Class: character  Mode: character |
| **Tenure**  Min.: 39.00  1st Qu.: 80.00  Median: 89.00  Mean: 89.67  3rd Qu.: 100.00  Max. :1146.00  NA's :37 | **Contract** **type**  Length:7513  Class: character  Mode: character | **Data** **allowance**  Length:7513  Class: character  Mode: character |
| **Payment** **method**  Length:7513  Class: character  Mode: character | **Monthly** **charges**  Min.: 10.00  1st Qu.: 50.00  Median: 50.00  Mean: 54.76  3rd Qu.: 62.00  Max. :109.00 | **Payment** **arrears**  Min.: 0.0000  1st Qu.: 0.0000  Median: 0.0000  Mean: 0.1619  3rd Qu.: 0.0000  Max. :4.0000 |
| **Contract** **terminated**  Length:7513  Class: character  Mode: character | **Click**  Length:7513  Class: character  Mode: character | **cid**  Min.: 1  1st Qu.: 1879  Median: 3757  Mean: 3757  3rd Qu.: 55635  Max. :7513 |

For categorical variables, the table above indicates the data type and length, whereas for continuous variables, it shows Min as minimum, 1st Q as first quantile, median, mean, 3rd Qu as third quantile, and max (continuous variable are variables that are numbers).

7807 distinct missing data points was found, accounting for a ratio of 0.07993324, or 7.99% of the data. This indicates we need to address the missing data points, as eliminating the rows in question could limit our dataset to something we can get a useful knowledge. Further investigation revealed that click has the highest percentage of missing data with 7498 unique missing data points, accounting for 0.9980035 of our total number of rows, and that removing click from our dataset reduced our error to 0.009427393, indicating that we now have 0.943 percent of our data missing and that the row containing missing data can be removed.

|  |  |  |  |
| --- | --- | --- | --- |
| Gender  3 | Age  84 | Citizenship  6 | Tenure  103 |
| Contract type  3 | Data allowance  5 | Payment method  3 | Monthly charges  89 |
| Payment arrears  5 | Contract terminated  3 | Number transferred  2 | cid  7210 |

The table above shows the number of unique data each data point as. After this we joint the score data and the customer data on the customer id column.

*data <- left\_join(DDM22\_19\_customer, DDM22\_19\_score)*

*Joining, by = "cid"*

Number of distinct customer id we have is 7210 which is less than 7807 number of rows we have which indicates we have some repeated data point where we have a customer data repeated but as this will not necessarily affect our business analysis, we can neglect it at the moment.

Also from the visualizations it was discover on the contract\_terminated column that there is an outlier value “idiot” and it has been removed using the R script.

**Analysis**

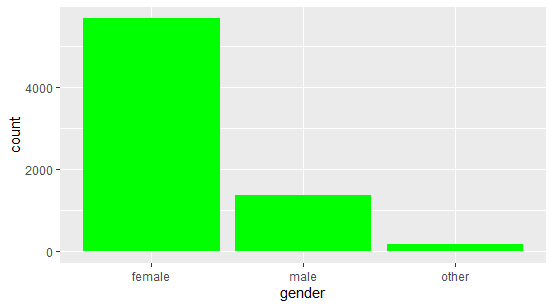


Figure 1: Visualizing the count of the gender using bar chat, the female categorical value having the highest count followed by the male and lastly others.

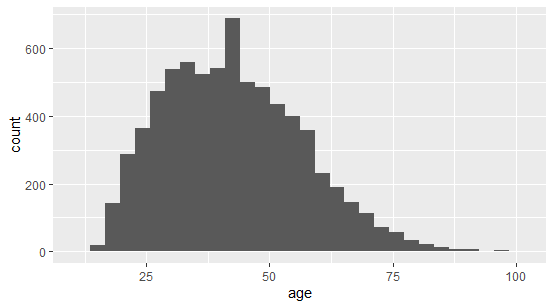


Figure 2: Visualization of the age column using histogram

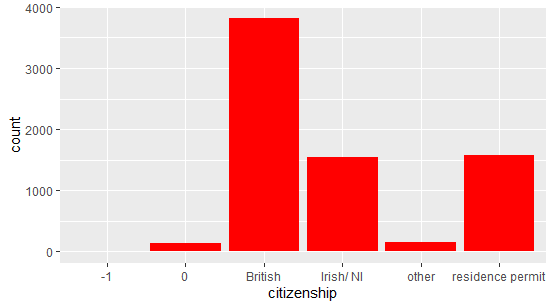


Figure 3: shows the plotting of the citizenship counting the number of the occurrence of each of the categorical values

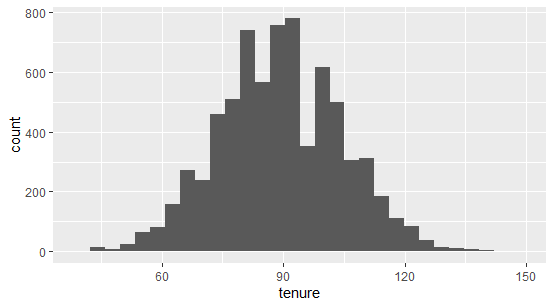


Figure 4: shows the plotting of the tenure column using histogram chart

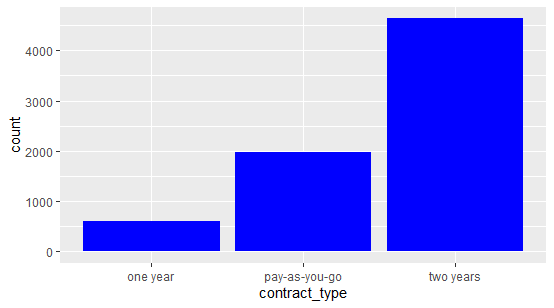


Figure 5: plotting the count of the contract type column using bar chart

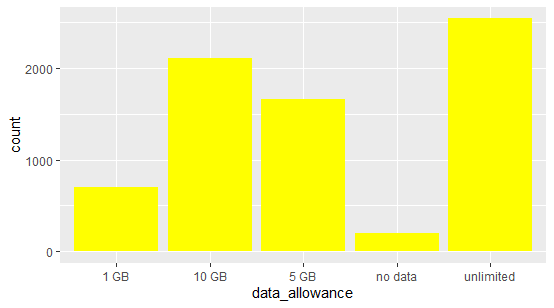


Figure 6: visualization of the data allowance column using bar chart

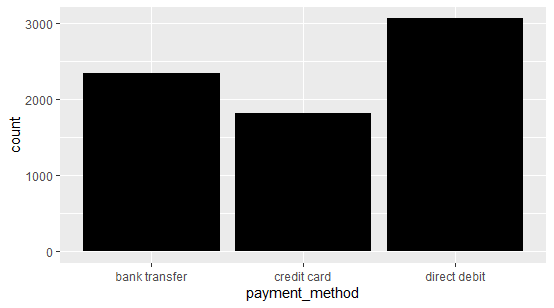


Figure 7: Visualization of the payment method column using bar chart showing each count of each of the categorical data entries on the column

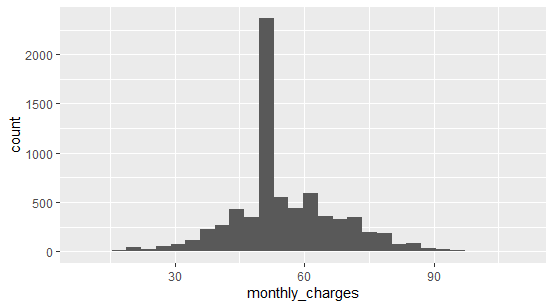


Figure 8: shows the visualization of the monthly charges using histogram plot

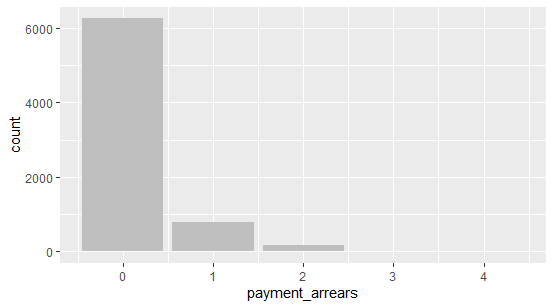


Figure 9: Visualization of the count of the payment arrears categorical values

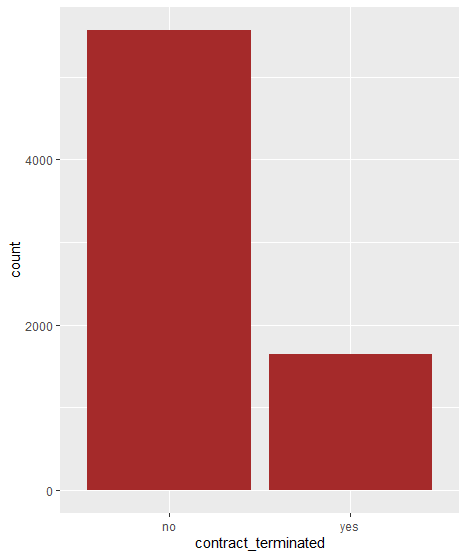


Figure 10: Visualization of the count of the contract terminated categorical values

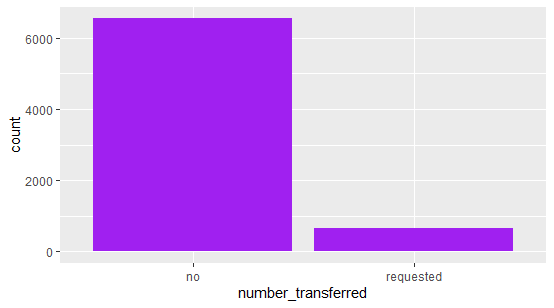
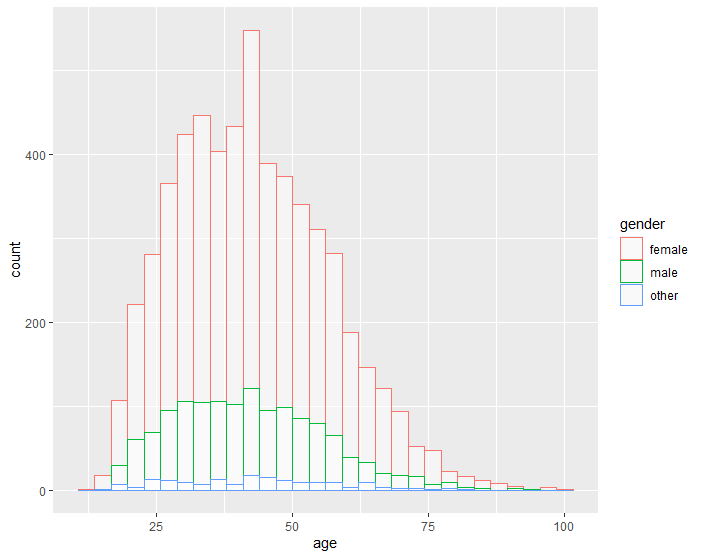


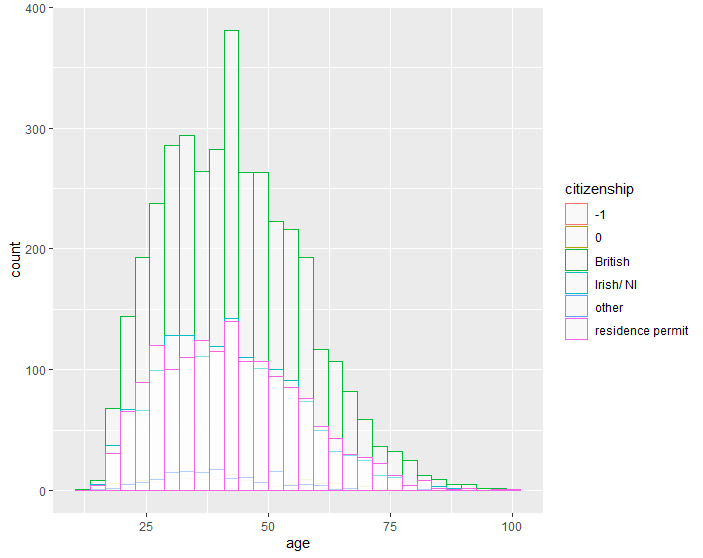
Figure 11: Visualization of the count of the number transferred categorical values

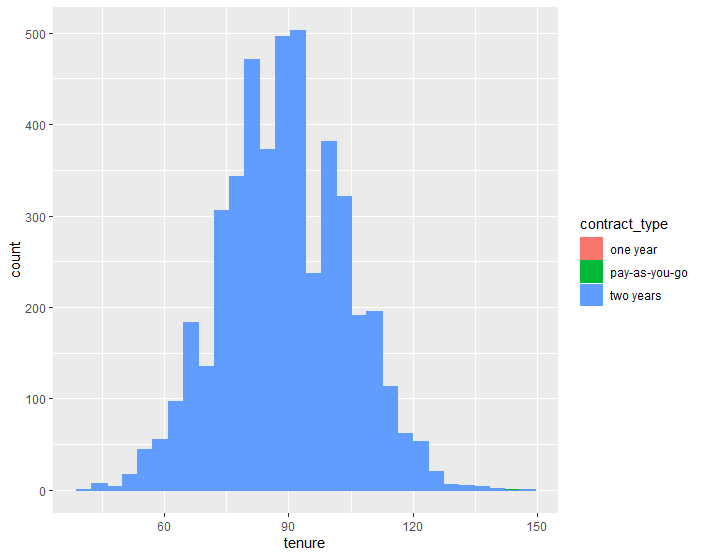
The figures below show the visualization of certain continuous columns values against selected categorical columns values to deduce some important information out of them.



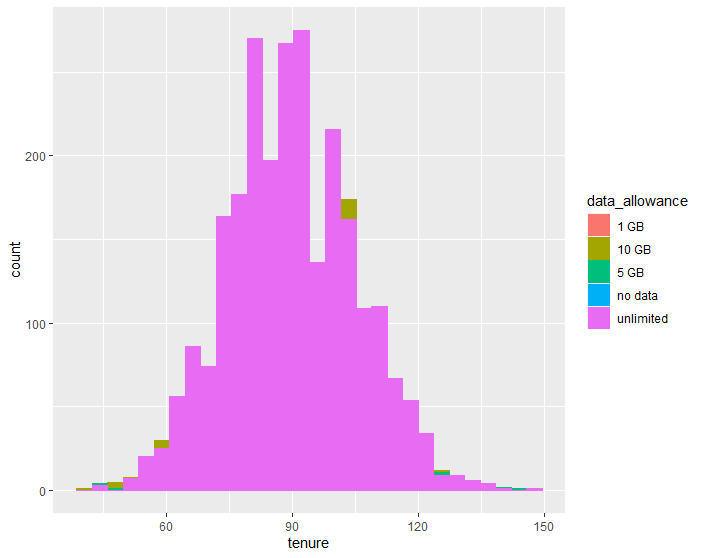
The figure above shows the visualization of the plotting of the age column against the gender, looking at the plot it obviously showed that the female caries the larger portion of the count. Taking a further step by making a pivot table of the age against the gender, the tabulated result shows that age 32 has the highest number of counts, on the other part the total count shows that the female, male and others has total count of 5665, 1374 and 171 respectively.

The figure below shows the plotting of the age vs the citizenship, as it’s been known to us that the citizenship comprises of 6 categories namely -1, 0, British, Irish/NL, others and residence permit. The most important categories are the British, Irish/NL, residence permit and others, the -1 and 0 categories have no tangible meaning in the entries which probably might have been due to an error in the entry, better still should have been categorized as others. From the visualization it shows that the British citizenship has the highest number of entries. The pivot table shows a total result for each of the category’s statin that the British, Irish/NL, residence permit and others has the total values of 3810, 1547,1573 and 148 respectively.



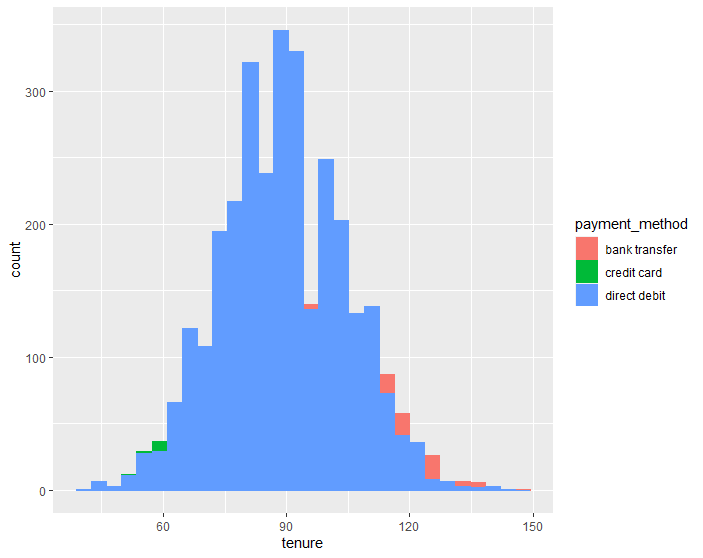


The figure above shows the visualization of the tenure against the contract type which obviously shown that most people are on a contract type of two years. The pivoted table result shows that contract\_type two years has total value of 4637, pay\_as\_you\_go has total value of 1969 and lastly the one-year contract type has 607 entries. On the other side the tenue has the highest frequency when it’s at 93.

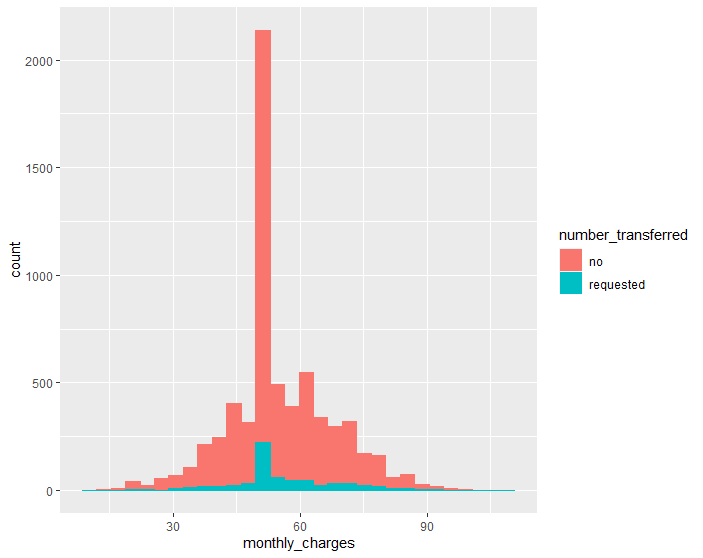


The figure above shows the plotted count of the tenure and data allowance, as it’s shown on the visualization, most customers opt in for the unlimited data allowance compared to other categories of data allowance, also the unlimited data allowance spread across the tenue. The pivoted table shows that the unlimited data allowance has the highest total of 2539, follow by the 10gb having 2103, 5gb with 1664, 1gb 699 and lastly no data with 209.

The figure below represents the visualization of tenure vs payment method, most of the customers make use of the direct debit across the tenure having the highest frequency at the point where tenure is 93. The pivot table shows that direct debit has total values of 3056, credit card with total value of 1811 and 2343 making use of the payment method bank transfer.



The figure below shows the visualization of monthly charges vs number transferred, there are very few transfers’ request. The requested number of transfer as shown on the pivot table is 666 while that of no request which have the highest value count has 6544.



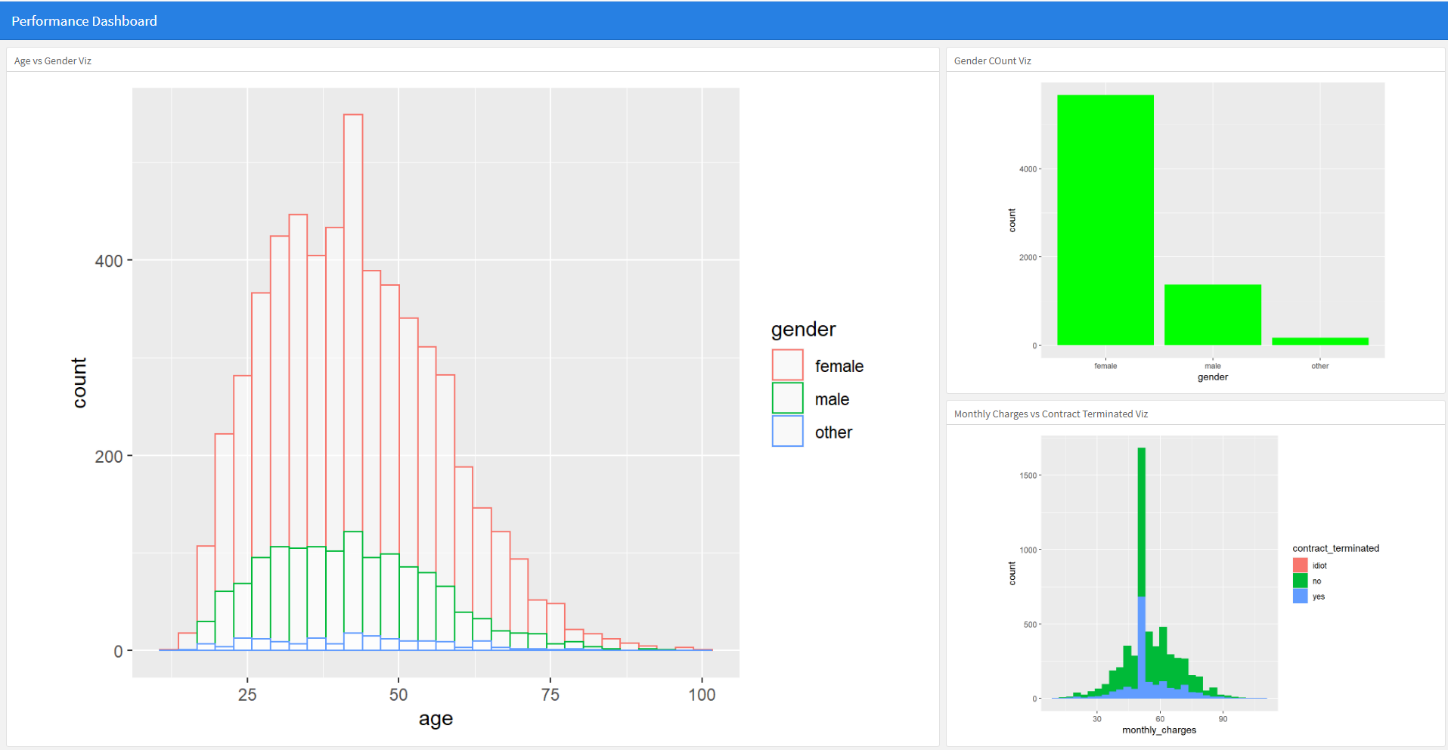
**Recommendations**

From the distribution of contracts terminated there seems to be a misclassification as we have a single idiot which was removed using the R script but that said we have a higher percentage of our clients not terminating their contract but the ratio terminating their contract is quite alarming as around 1 of every 5 customers seems to be terminating their contract. And this must be due to probably poor service delivery or better service from our competitors.

Also, most people considered the two years contract type compared to other contract categorical values, this could be a good idea for the company to propose to it customers. Likewise for data allowance most customers consider going for the unlimited or larger data like 10gb and 5gb compare to small data or no data allowance. The company should push more for the direct debit payment method because it seems convenient for the customers compared to other method. I believe these recommendations will help in moving the company forward.

Majorly the poor performance is due to the competitions from other competitors, so the company might need to improve on how they render their services and create a better ways to serve the customers better that will attract them from going to other competitors.

**Performance Dashboard**



The performance dashboard for the analysis is depicted in the image above, which was created using a R script and the Flexboard library. It's worth noting that the dashboard doesn't show all of our plotted charts. The most important ones have been chosen and are now displayed on the dash board.