Exploratory Data Analysis on churn data

Project Report

Arnab Mukherjee

12/25/22

Programming for Data Science

ACKNOWLEDGEMENT

Apart from the efforts of me, the success of any project depends largely on the encouragement and guidelines of many others. I take this opportunity to express my gratitude to the people who have been instrumental in the successful completion of this project.

I would like to express my greatest appreciation to our professor Dr Sudeep Mallick. I feel motivated and encouraged every time I attend his meeting. Without his encouragement and guidance this project would not have materialized.

EXECUTIVE SUMMARY

- Data has been described and objective is mentioned
- At first, the data has been cleaned in a manner to omit variables having many NA values.
- We considered continuous variables first in EDA, and drawn insights for churn non-churn differences.
- We used histograms and frequency polygons to compare among churn status, area, ethnicity, etc and draw insights.
- We considered categorical divisions like area, ethnicity, income level, profession, etc and used bar plots and multiple bar plots.
- We then considered some plots to give other inferences like rug plot, area plot and we used count plot for relating income level and no. of cars.
- Then to find which variables contribute to churn status we used boxplots. We also tried the same for some other status like having children or dwelling type.
- We considered two variable plots to check any association.
- We then went for some new features here we took months of service variable.
- We also missing value analysis by missing value substitution by mean, modal class (for categorical case)
- Then we concluded about some results that we got.
- Appendix part can be used further inferences

TABLE OF CONTENTS

DATA DESCRIPTION AND OBJECTIVE

Here in the churn dataset, we have 100000 customers of a telecom company. And there are 173 columns/variables per customer, where 128 variables are numeric and 45 are of character type. On the other hand, out of the 100000 customers there are 49562 customers who have churned and 50438 customers who still are current customers i.e., non-churned. Here churned is given as 1 and non-churned as 0 Here we want to make an Exploratory Data Analysis on this data. So that we can see what information we get from the data before performing any predictive analysis or statistical modelling. And compare our results after the other analyses. Primary objective of this EDA is to gain some information about which variables effect the churn status. And further also explore other parts of the data.

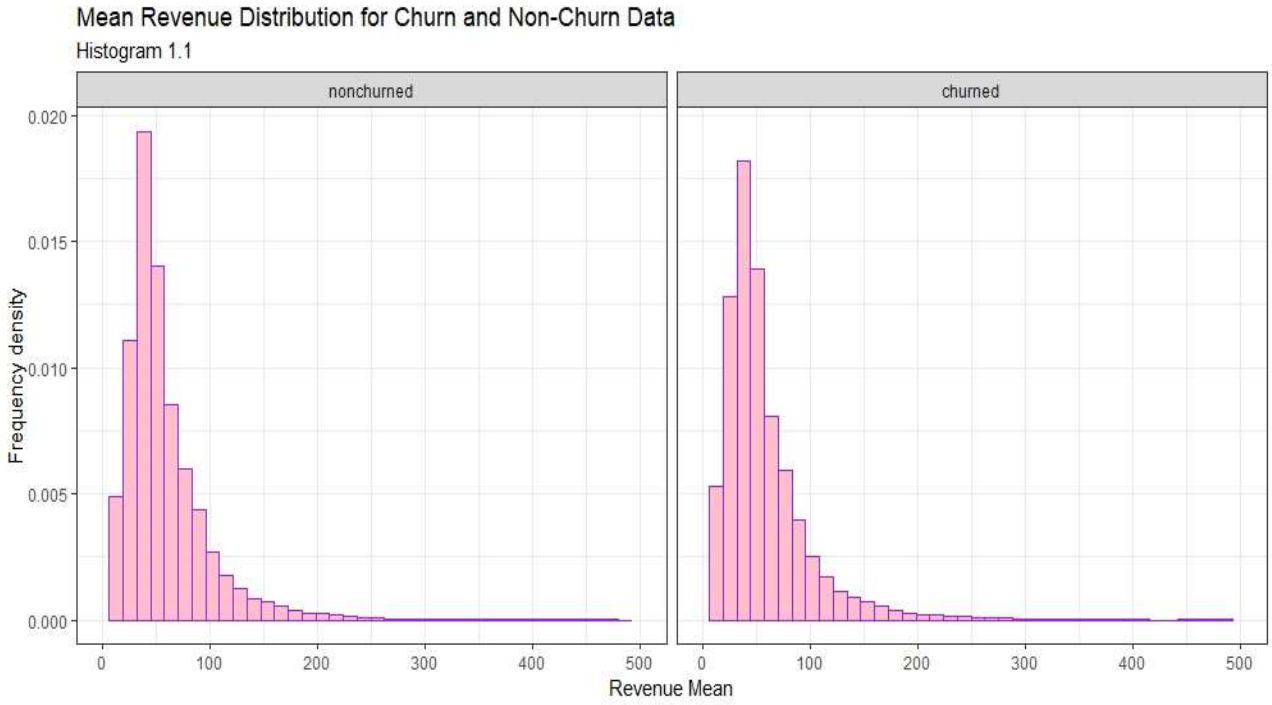
EDA WITH VISUALIZATION

DATA CLEANING

In big datasets there exists many missing values which interrupt us to get some analysis done. Here we have removed those columns which have NA values more than 50 percent. So, from 173 we were left with 163 variables/columns.

Visualization and Analysis

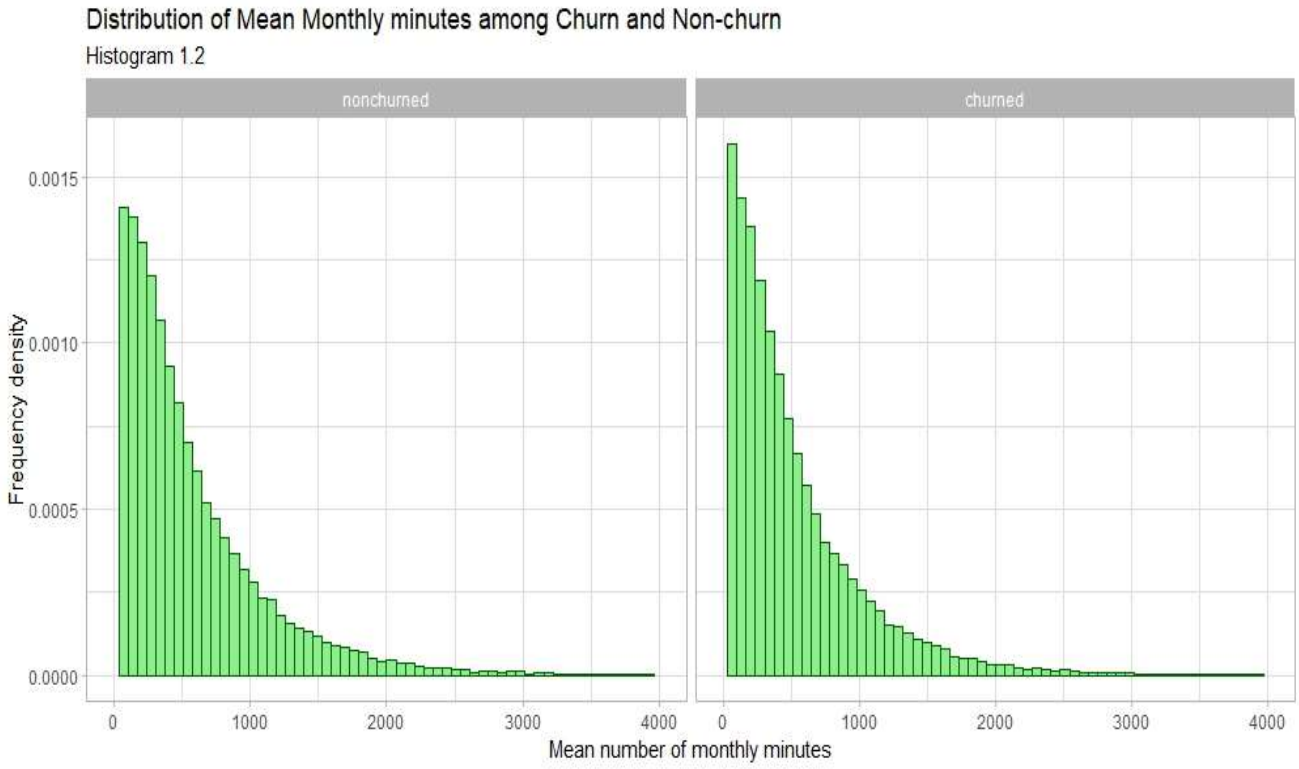
At first, we have taken Mean Revenue for Churn and Non-Churn Data and tried to get some information.



We have separately taken the histograms having frequency density of Mean Revenue as height and we can see that they look almost similar but in the third bin roughly in the interval 30-45, we see that the churned

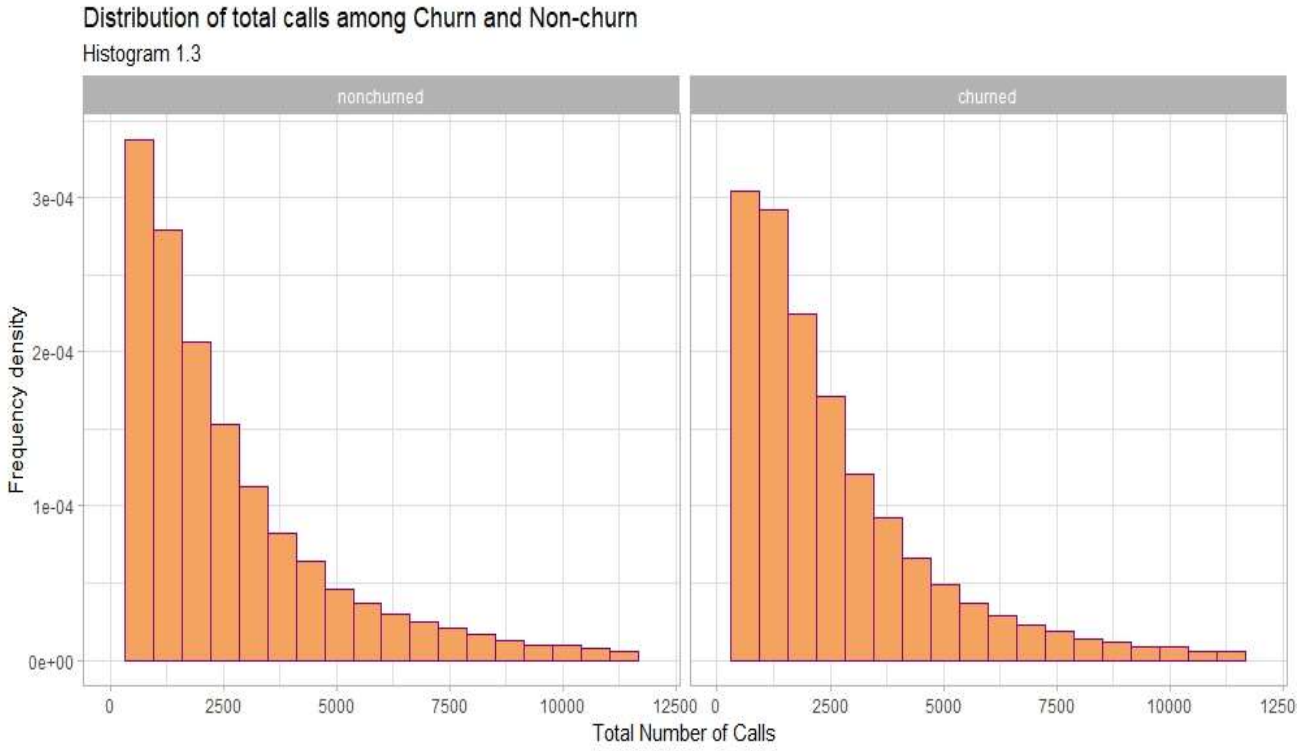
has a lower proportion of people compared to non-churned.

Now, we have taken Mean number of monthly minutes for Churn and Non-Churn Data and tried to get some information.



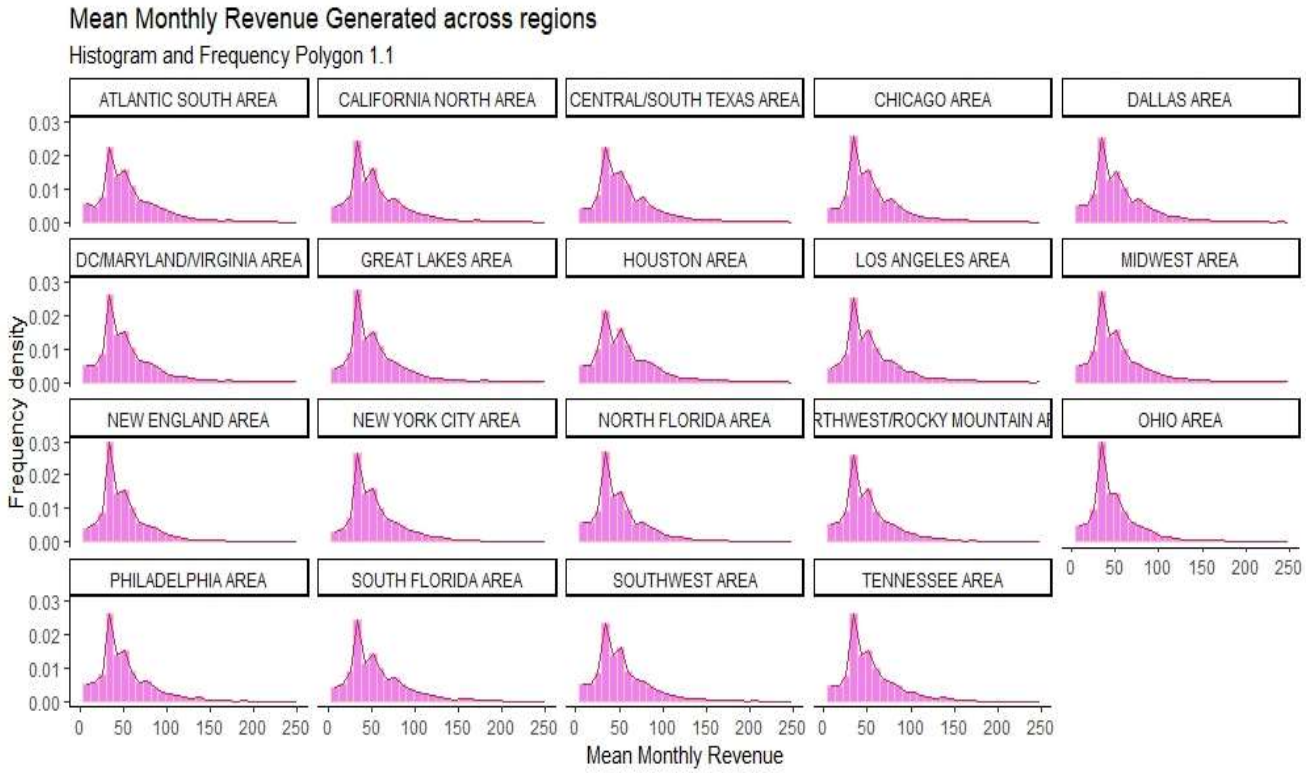
We have separately taken the histograms having frequency density of Mean number of monthly minutes as height and we can see that they look similar as the proportion of people decreases in both as Mean number of monthly minutes increases. But in the initial bins roughly in the interval 0-145, we see that the churned has a higher proportion of people compared to non-churned.

Now, we have taken Total Number of Calls for Churn and Non-Churn Data and tried to get some information.



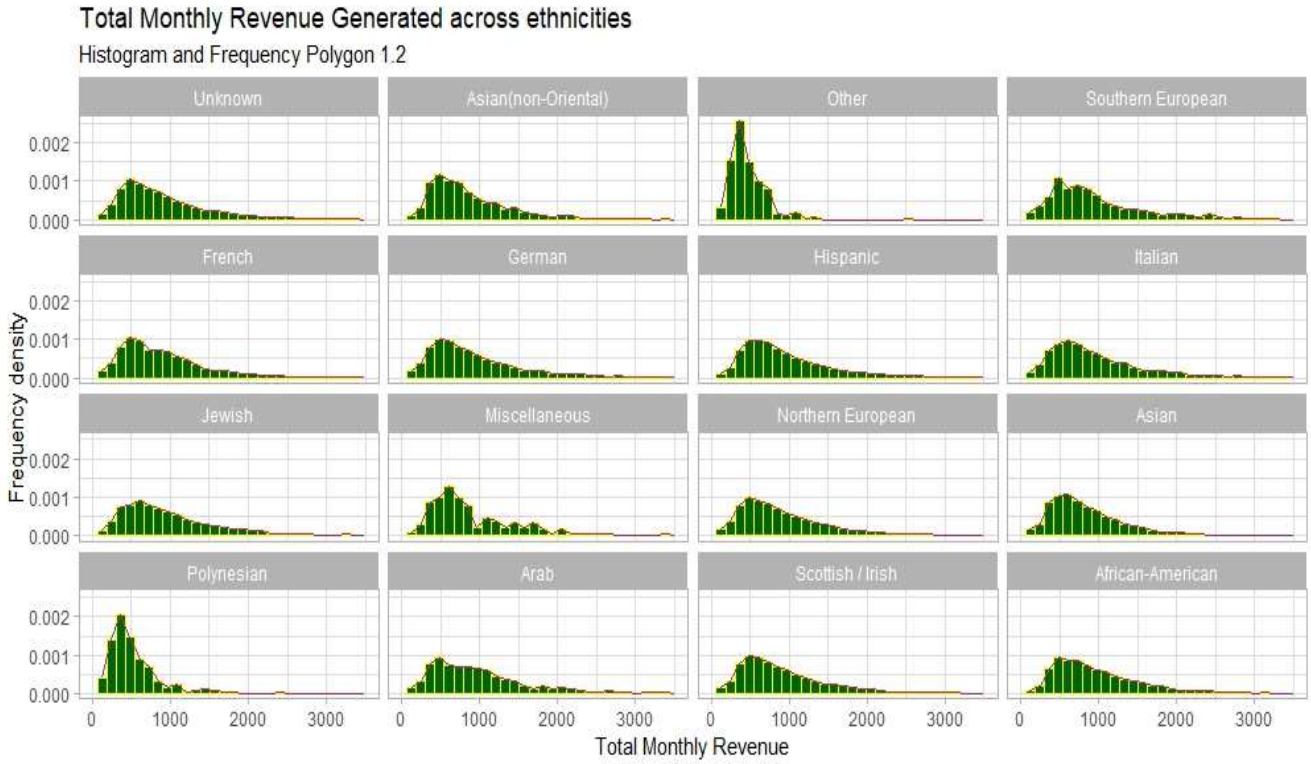
We have separately taken the histograms having frequency density of Total Number of Calls as height and we can see that they look similar as the proportion of people decreases in both as Total Number of Calls increases which is trivial. But in the initial bins roughly in the interval 0-2150, we see that the churned has a lesser proportion of people compared to non-churned. And in the next 3 bins roughly in the interval 2850-4290, we see that the churned has a higher proportion of people compared to non-churned.

Now, we have taken Mean Monthly Revenue Generated across regions and tried to get some information.



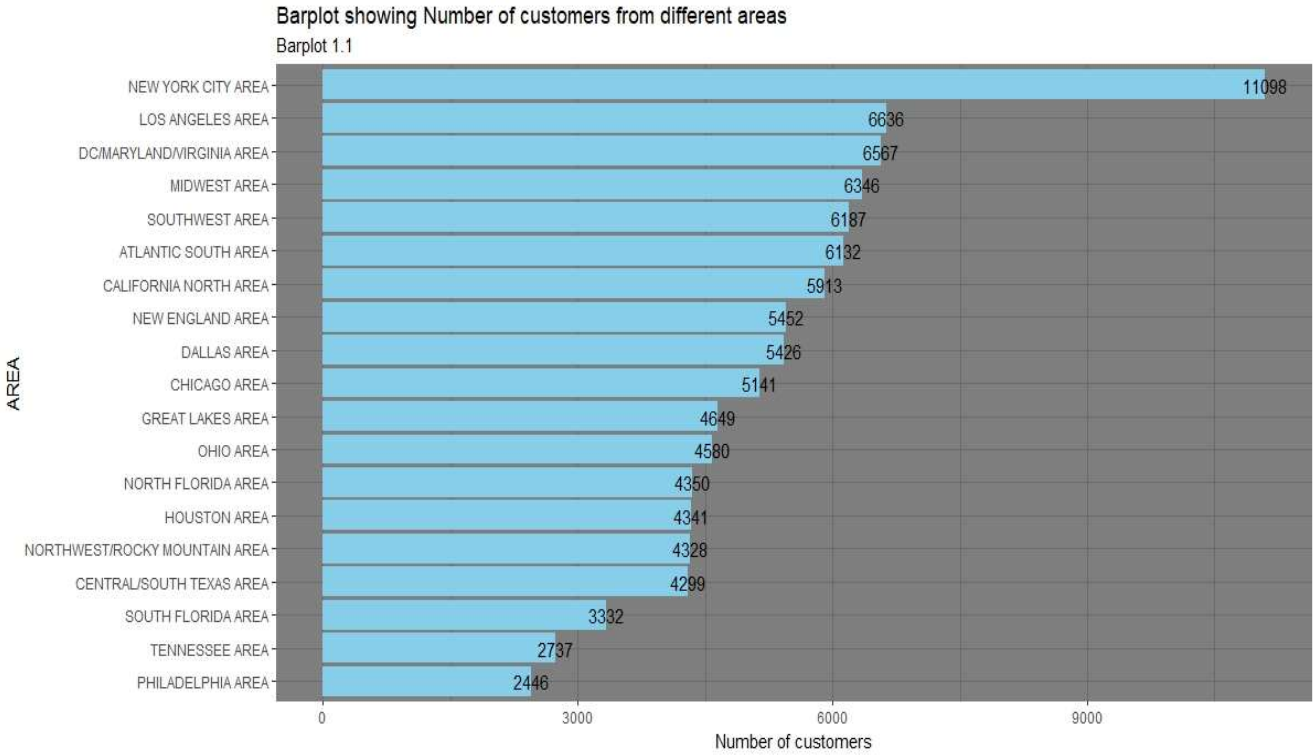
We have taken the histograms and frequency polygons having frequency density of Mean Monthly Revenue as height and we can see that they look similar across regions except Ohio area as it seems to have a unimodal kind of frequency polygon and other areas have bimodal kind i.e., the proportion of people roughly in the interval 50-60 (2nd modal class) is higher in other areas than Ohio whereas Ohio has highest proportion in the 1st modal class or 30-40 among other areas. And all of the areas have positively skewed frequency polygon as we can see.

Now, we have taken Total Monthly Revenue Generated across ethnicities and tried to get some information.



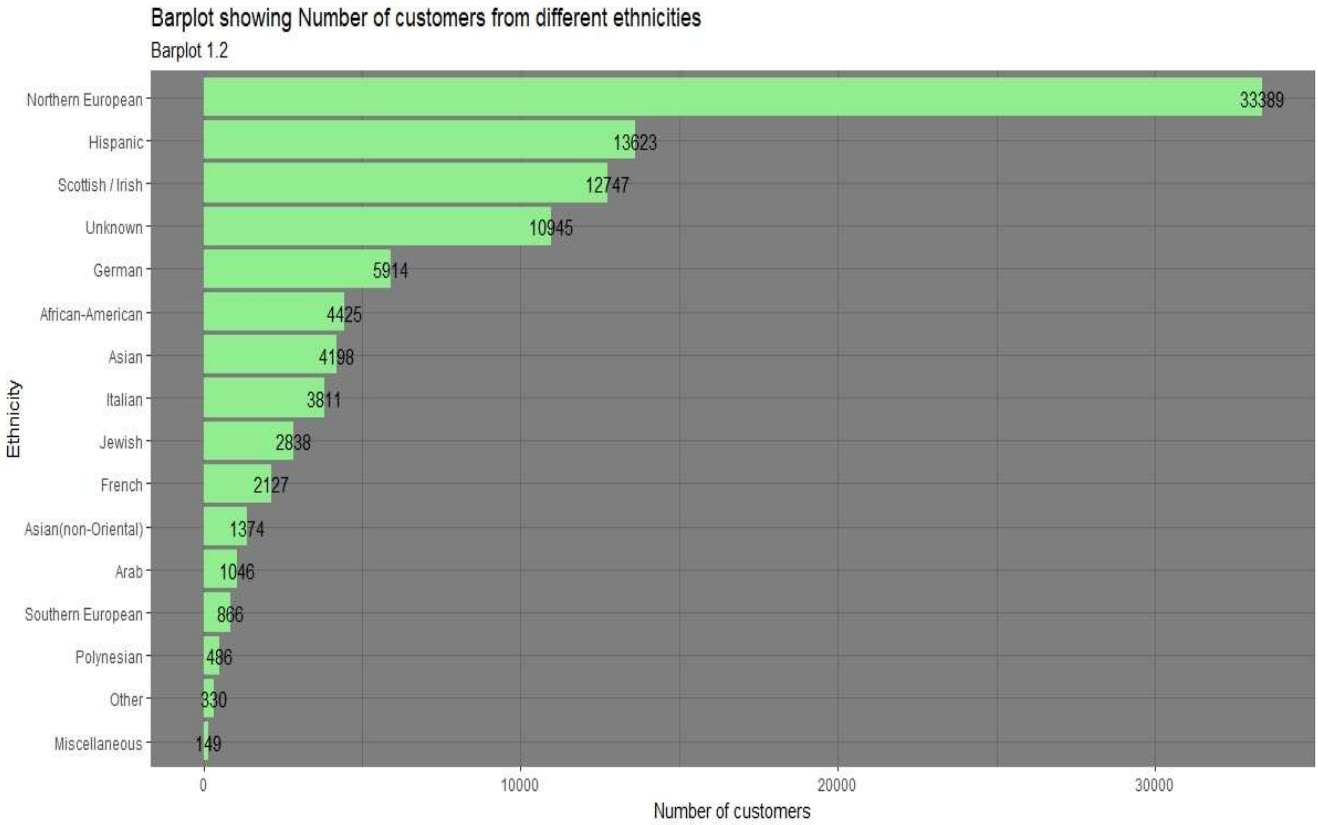
We have taken the histograms and frequency polygons having frequency density of Total Monthly Revenue as height and we can see that they look similar across ethnicities except Other, Southern European, Miscellaneous and Polynesian. In Other and Polynesian, the proportion of people is higher in 0-1000 than the rest. Miscellaneous has a disturbed one as it is a mixture. Southern European has a high proportion in roughly 375-400 than neighbouring intervals. And all of the ethnicities have positively skewed frequency polygon as we can see.

Now, we want to get some information on number of customers from different areas.



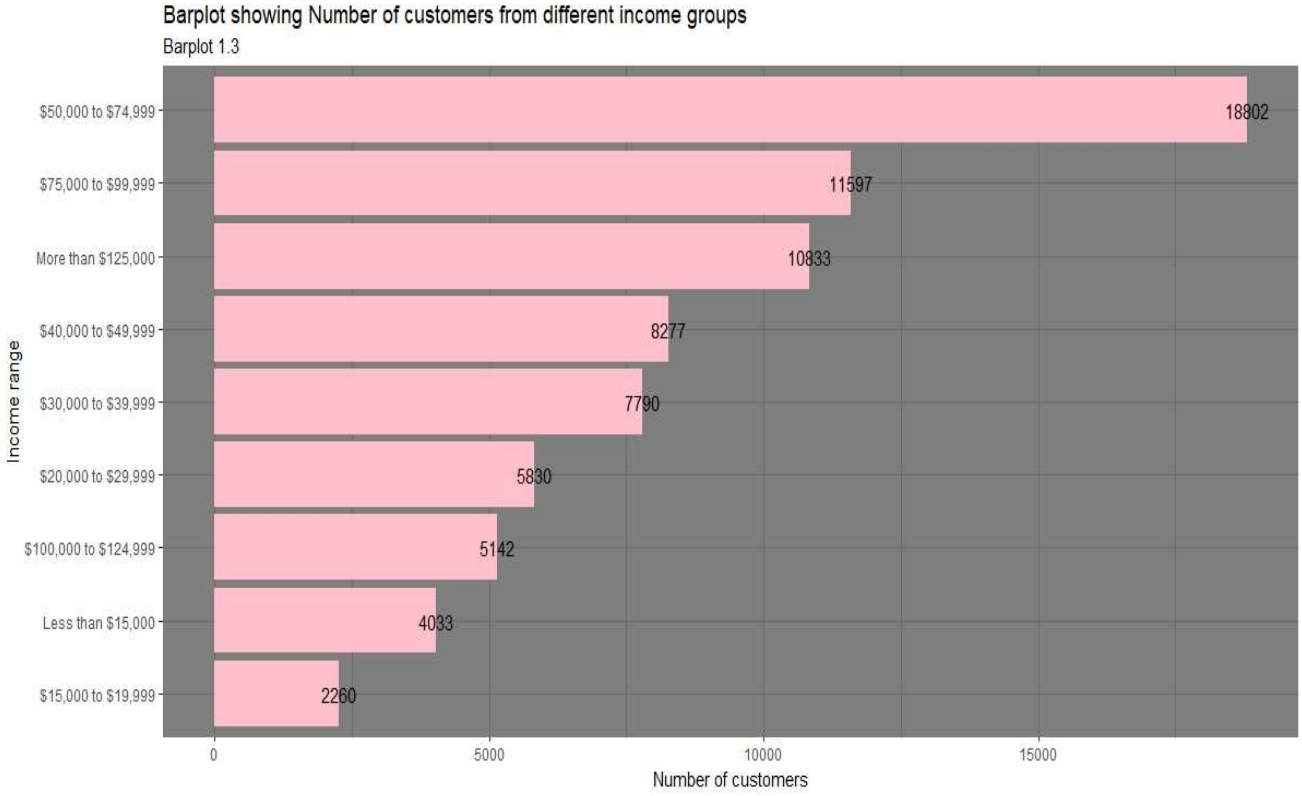
The horizontal bar plot is done in ascending order from bottom. We can see that most customers belong to New York City Area and least to Philadelphia area. Population may have taken a part here but that information we do not have.

Now, we want to get some information on number of customers from different ethnicities.



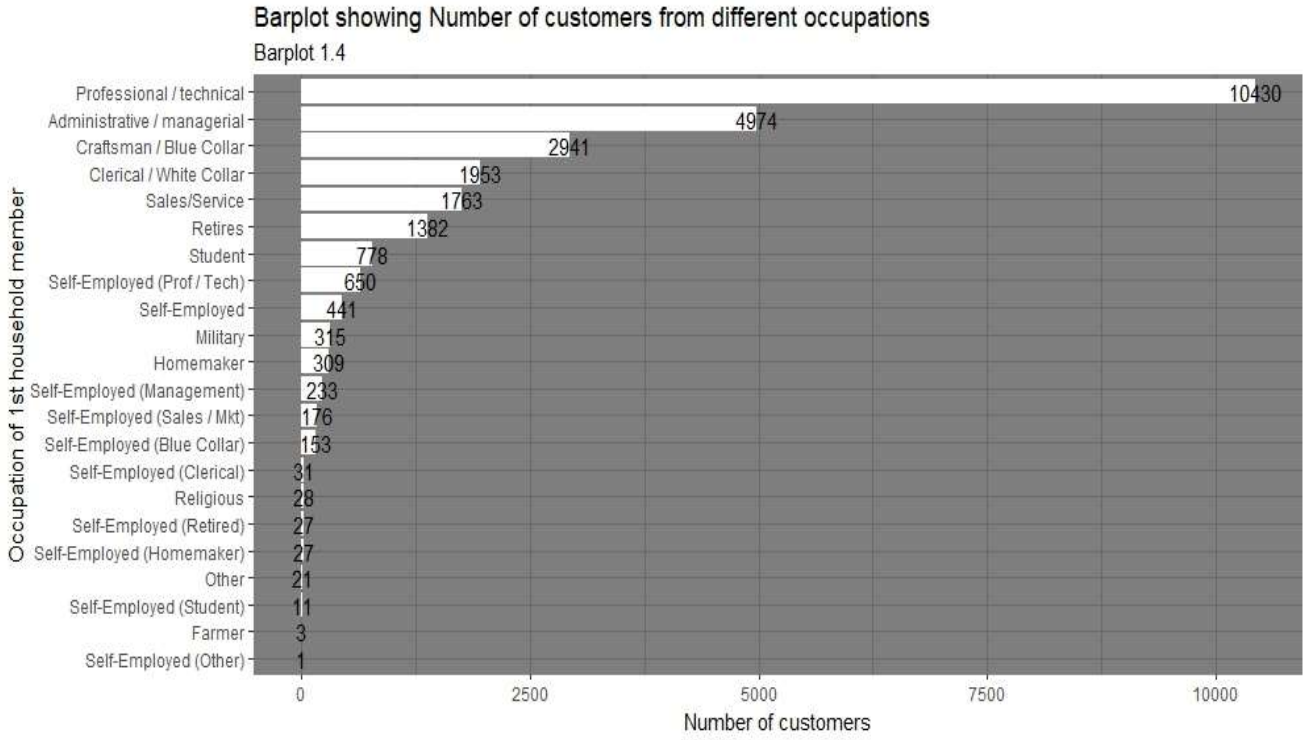
The horizontal bar plot is done in ascending order from bottom. We can see that most customers belong to Northern European ethnicity and least to Polynesian ethnicity barring other and miscellaneous ethnicity. Population may have taken a part here also but that information we do not have.

Now, we want to get some information on Number of customers from different income groups.



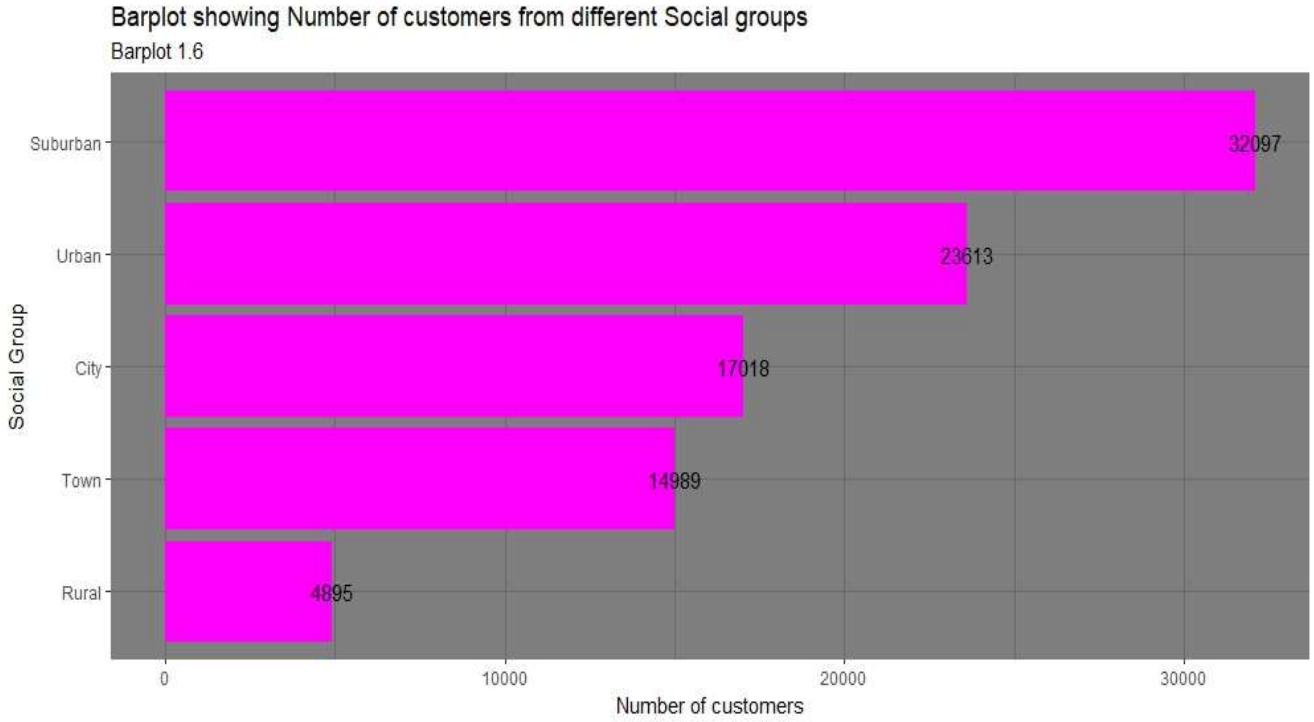
The horizontal bar plot is done in ascending order from bottom. We can see that most customers belong to $50000-$74999 and least to $15000-$19999.

Now, we want to get some information on Number of customers from different occupations



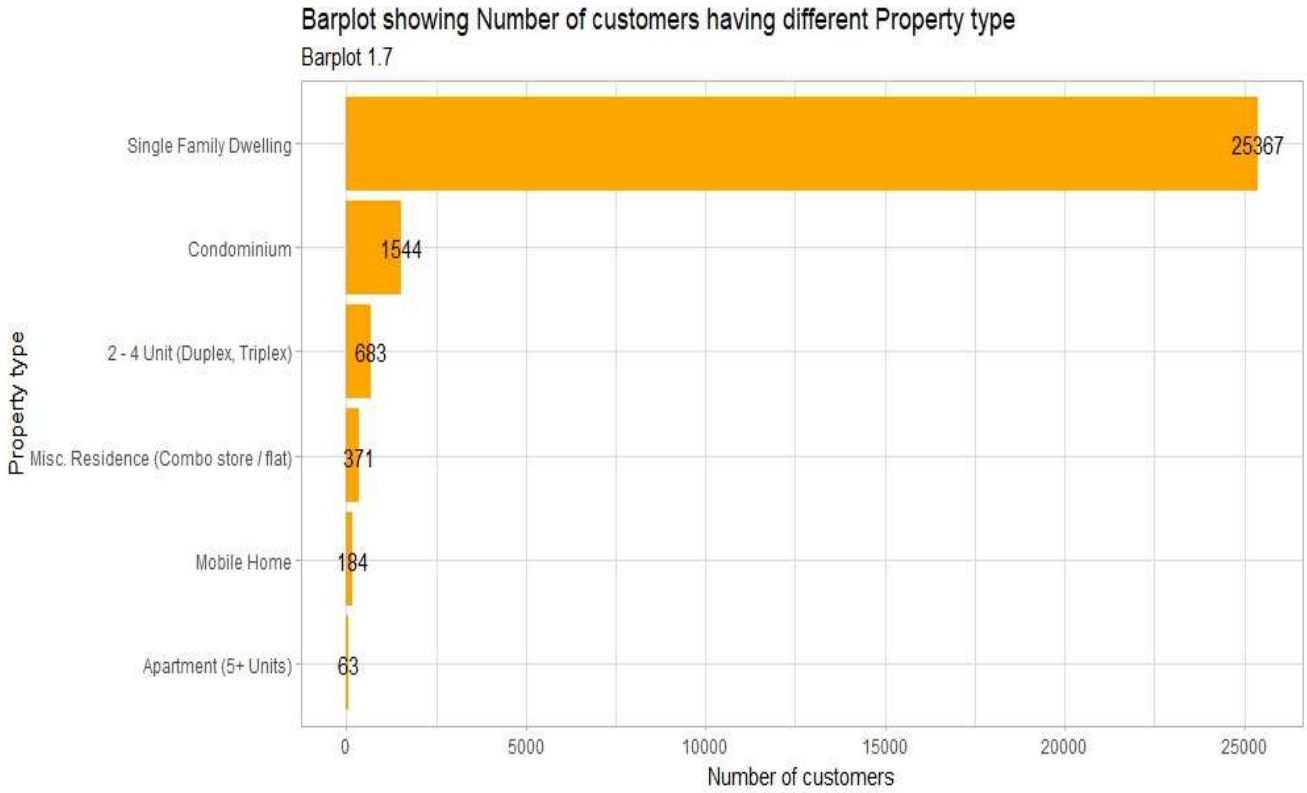
The horizontal bar plot is done in ascending order from bottom. We can see that most customers belong to Professional/technical and least to self-employed (other). Population division as per profession may have taken a part here but that information we do not have.

Now, we want to get some information on Number of customers from different social groups.



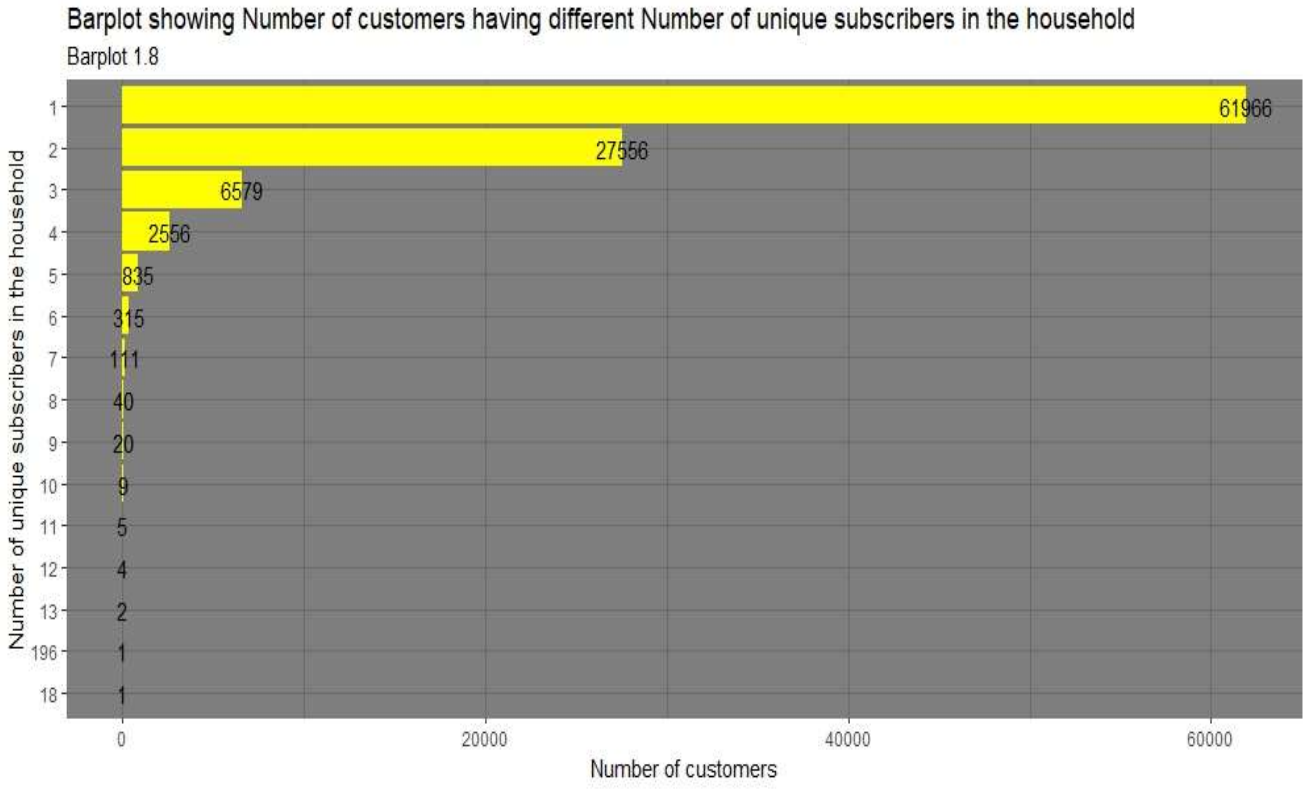
The horizontal bar plot is done in ascending order from bottom. We can see that most customers belong to sub urban and least to rural. Population division as per social groups may have taken a part here but that information we do not have.

Now, we want to get some information on Number of customers having different Property type



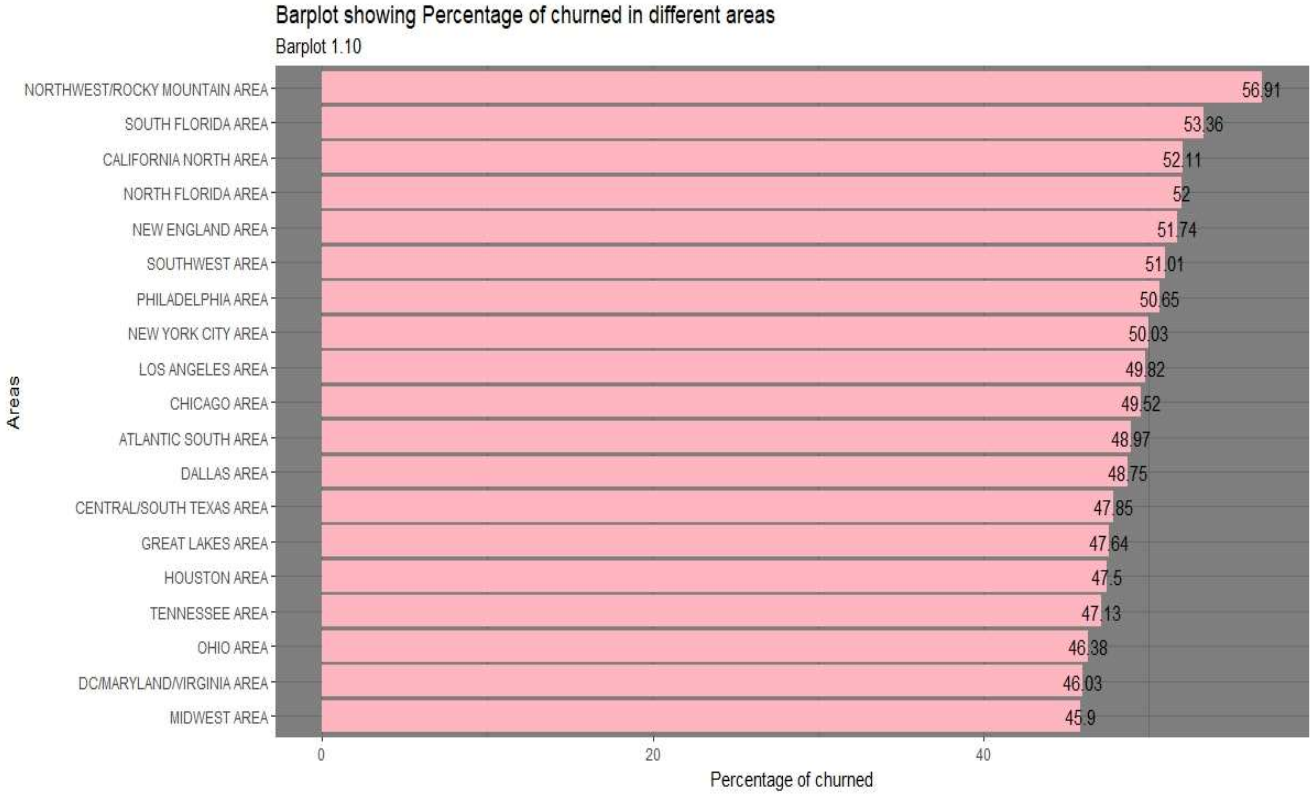
The horizontal bar plot is done in ascending order from bottom. We can see that most customers belong to Single Family Dwelling and least to Apartment (5+ Units). Population division as per Property type may have taken a part here but that information we do not have.

Now, we want to get some information on Number of unique subscribers in the household



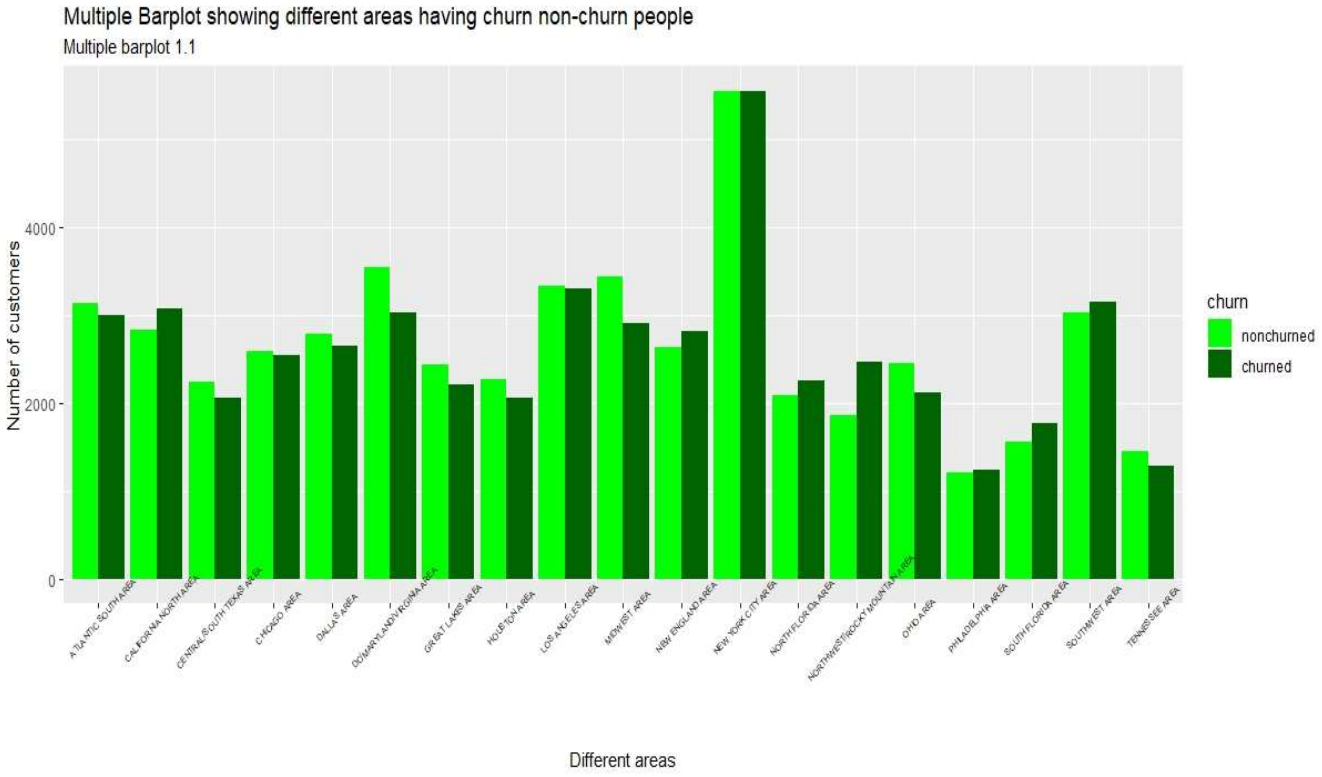
The horizontal bar plot is done in ascending order from bottom. We can see that most customers belong to Single unique subscriber in the household and least to 18 unique subscribers in the household. This seems quite trivial but we noticed that there is '196' unique subscribers in the household which is quite absurd.

Now, we want to get some information on Percentage of churned in different areas



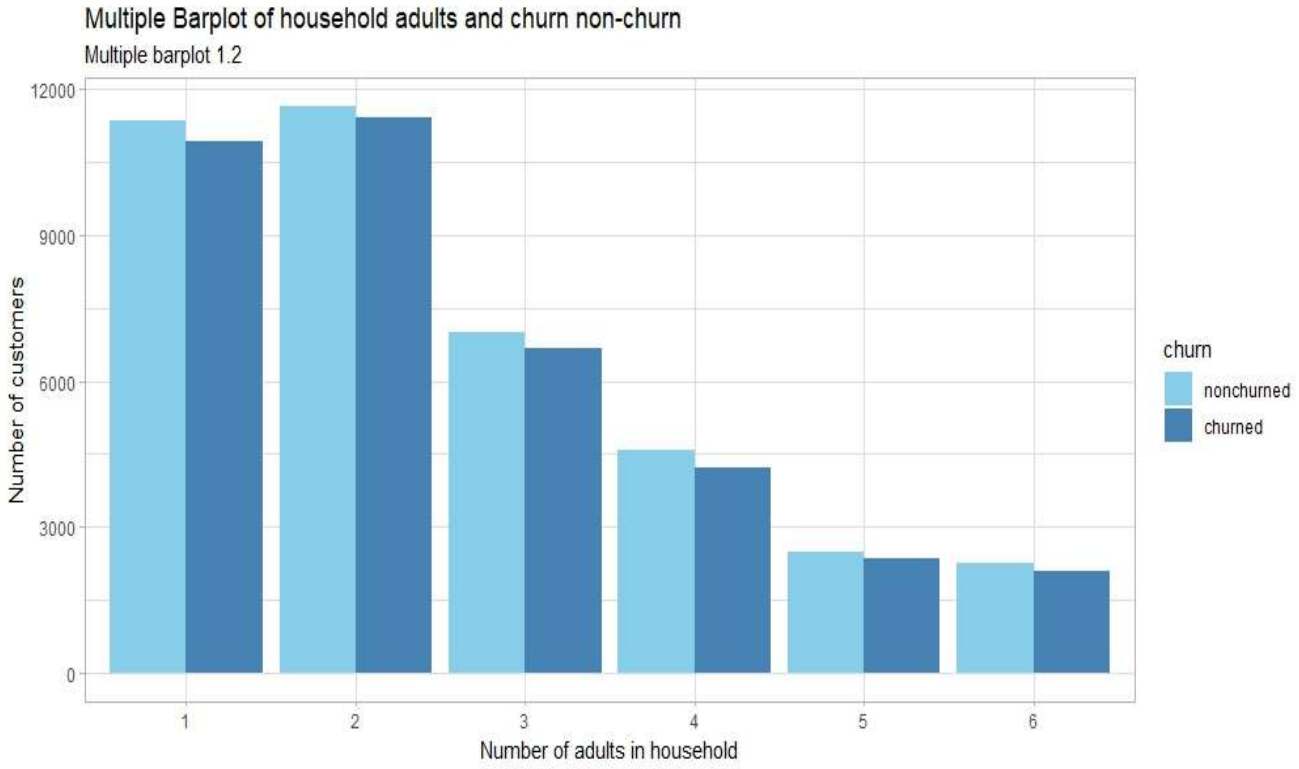
The horizontal bar plot is done in ascending order from bottom. We can see that highest proportion of churned customers belong to Northwest/Rocky Mountain area and lowest to Midwest Area.

Now, we want to get some information on different areas having churn non-churn people



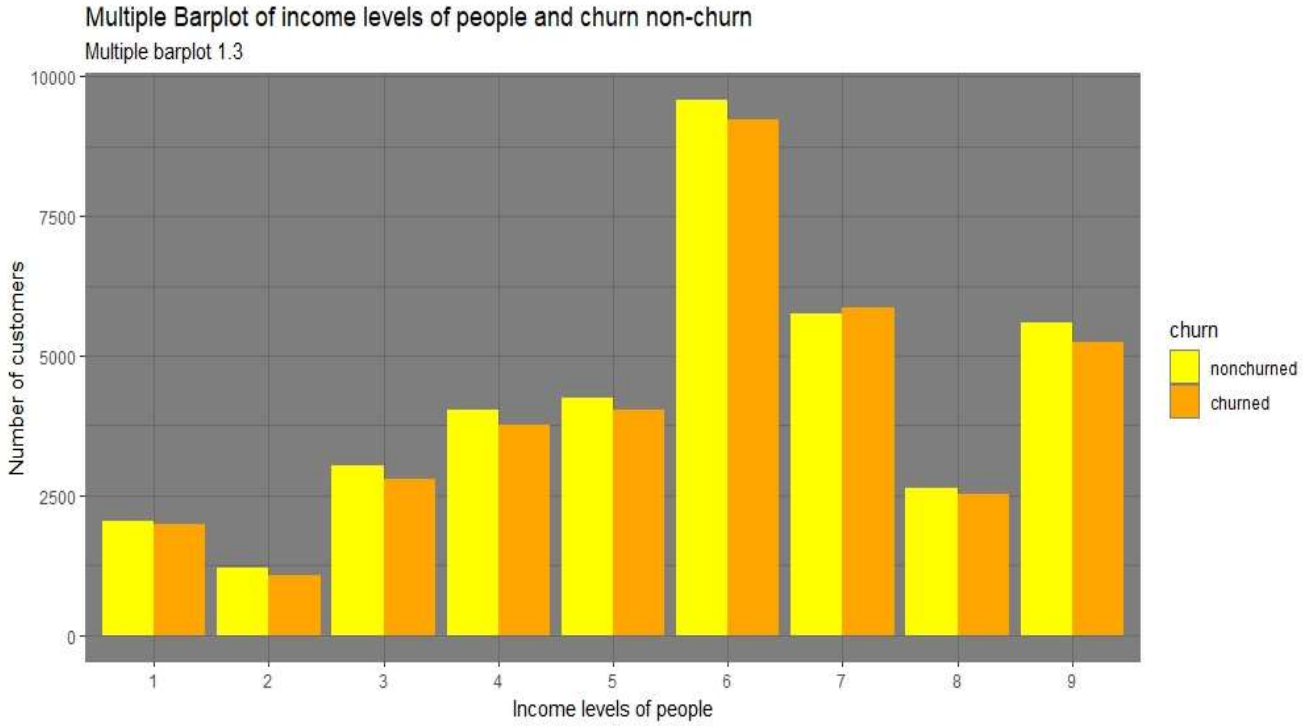
From this plot, we can see that there are 11 areas having higher number of non-churned customers than churned. And 7 have the opposite. For New York city, it is almost equal.

Now, we want to get some information on number of household adults and churn non-churn



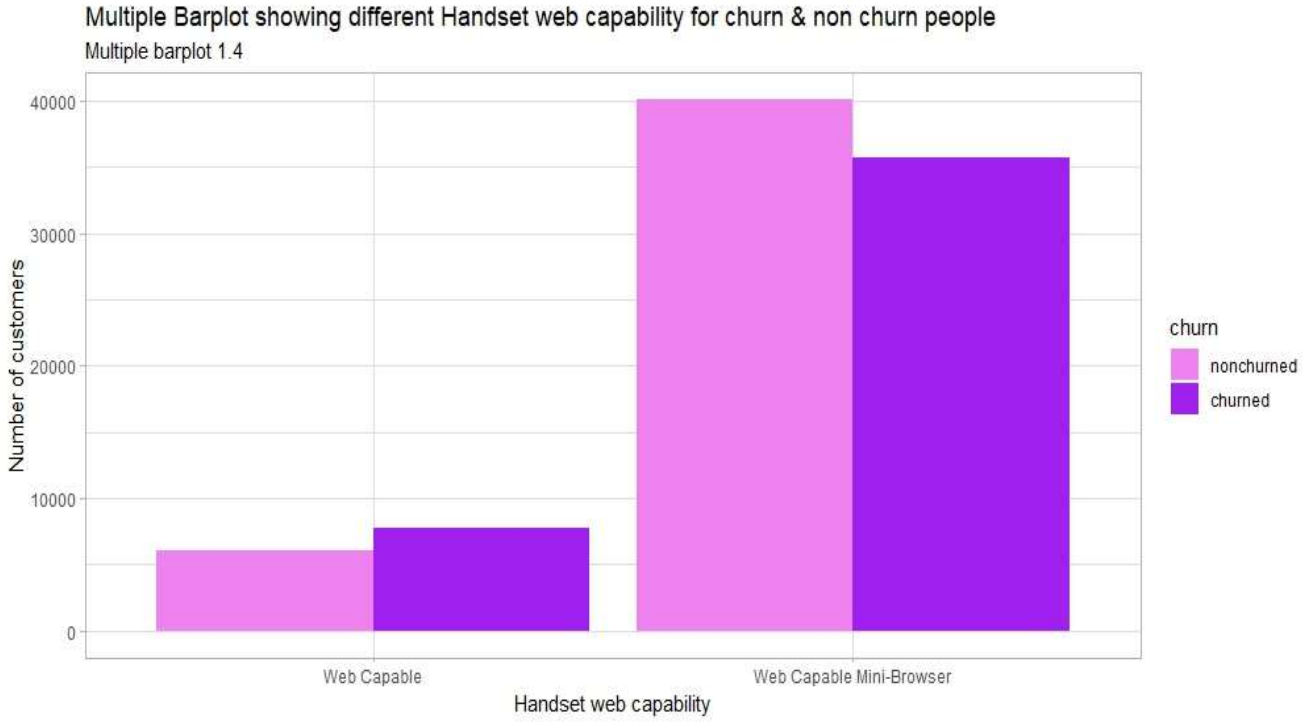
From this plot, we can see that all number of household adults have higher number of non-churned customers than churned. And most customer households have 2 adults.

Now, we want to get some information on income levels of people and churn non-churn



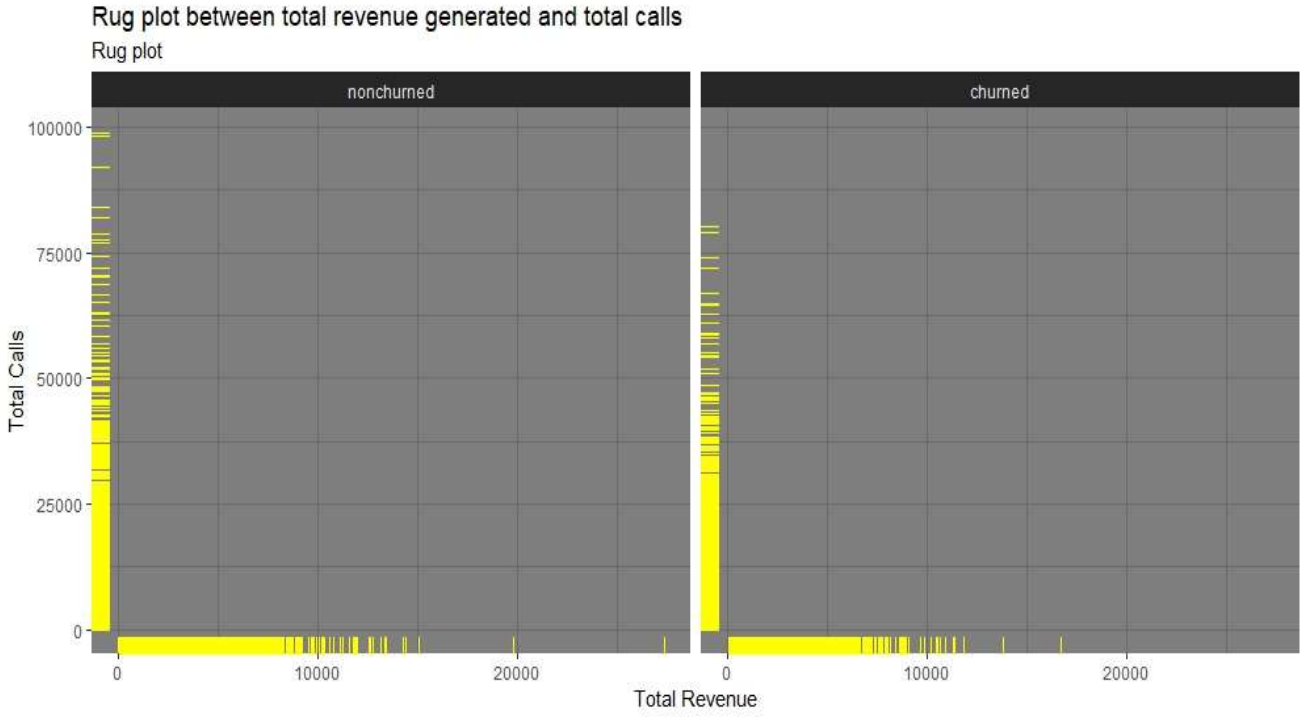
From this plot, we can see that all income levels except income level 7 have higher number of non-churned customers than churned.

Now, we want to get some information on different Handset web capability for churn & non churn people



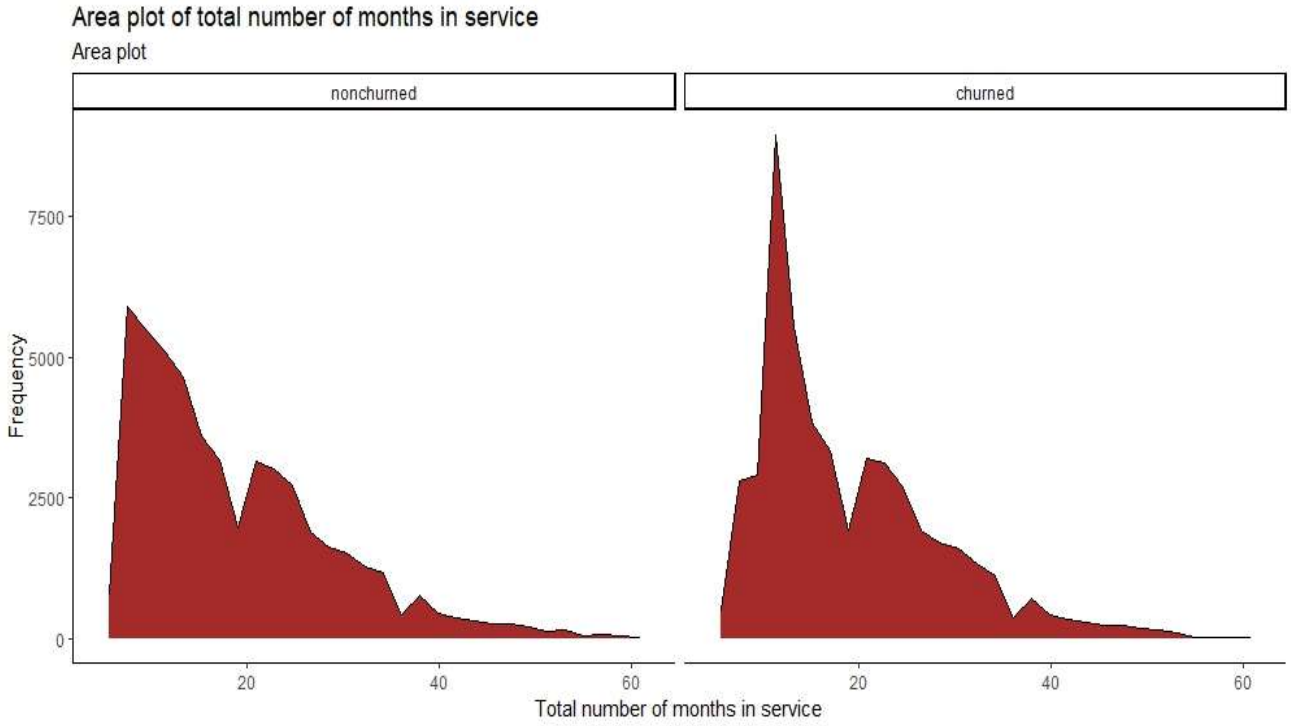
From this plot, we can see that Web Capable Mini-browser handset have higher number of non-churned customers than churned. And Web Capable have the opposite.

Now, we want to get some information between total revenue generated and total calls.



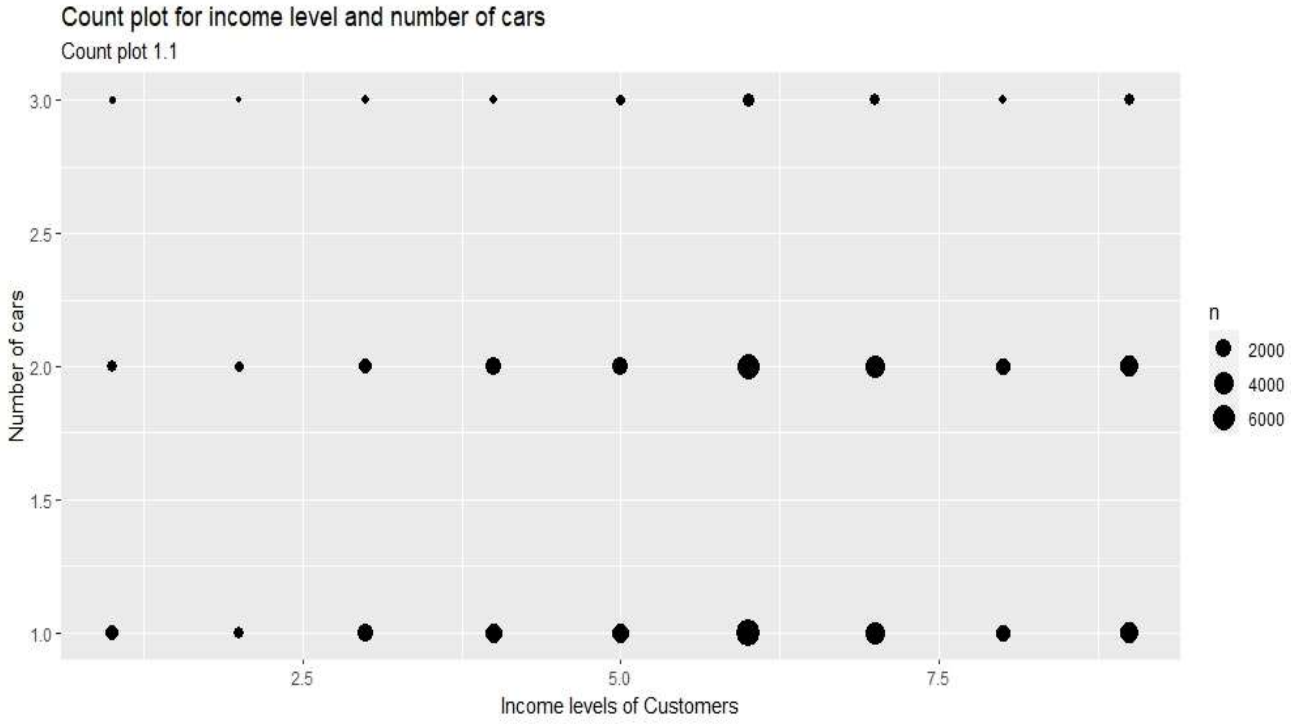
From this plot, we can see that total revenue generated have higher concentration in 0-10000 and total calls have higher concentration in 0-25000.

Now, we want to get some information on total number of months in service churn and non-churn



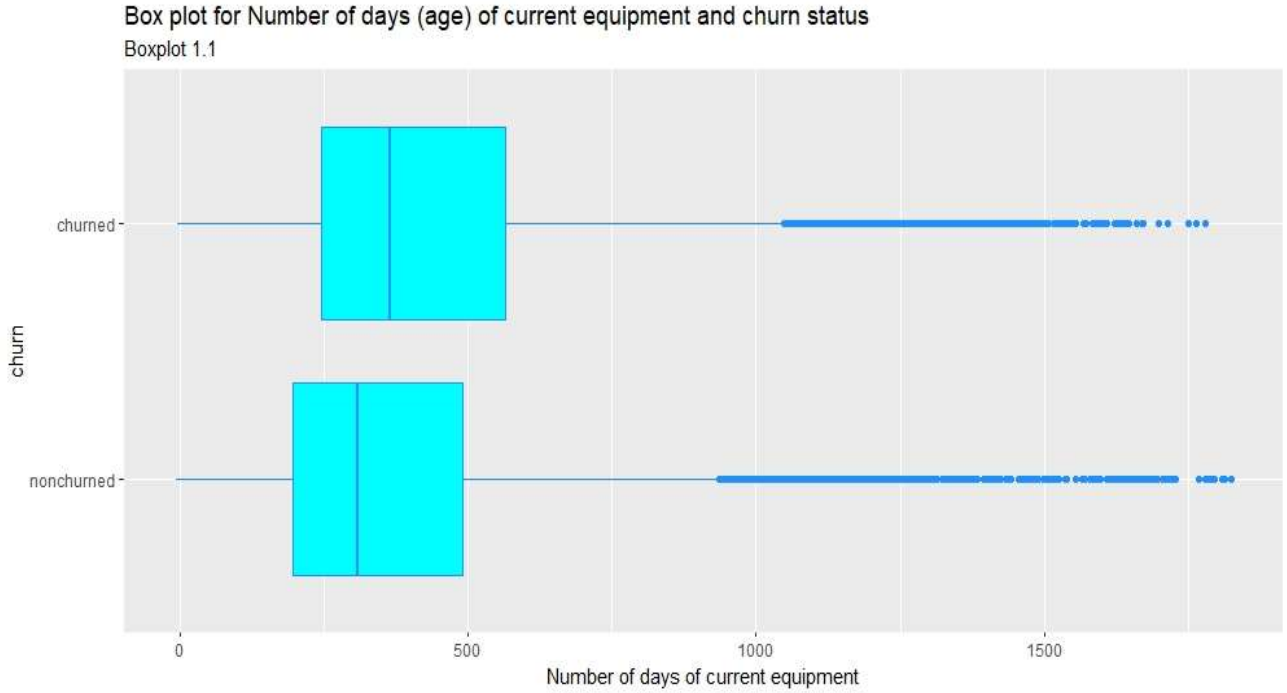
From this plot, we can see that total number of months in service for churn and non-churn have almost same graph after 20 and churned have higher frequency on some points in 0-20 than non-churned.

Now, we want to get some information on income level and number of cars



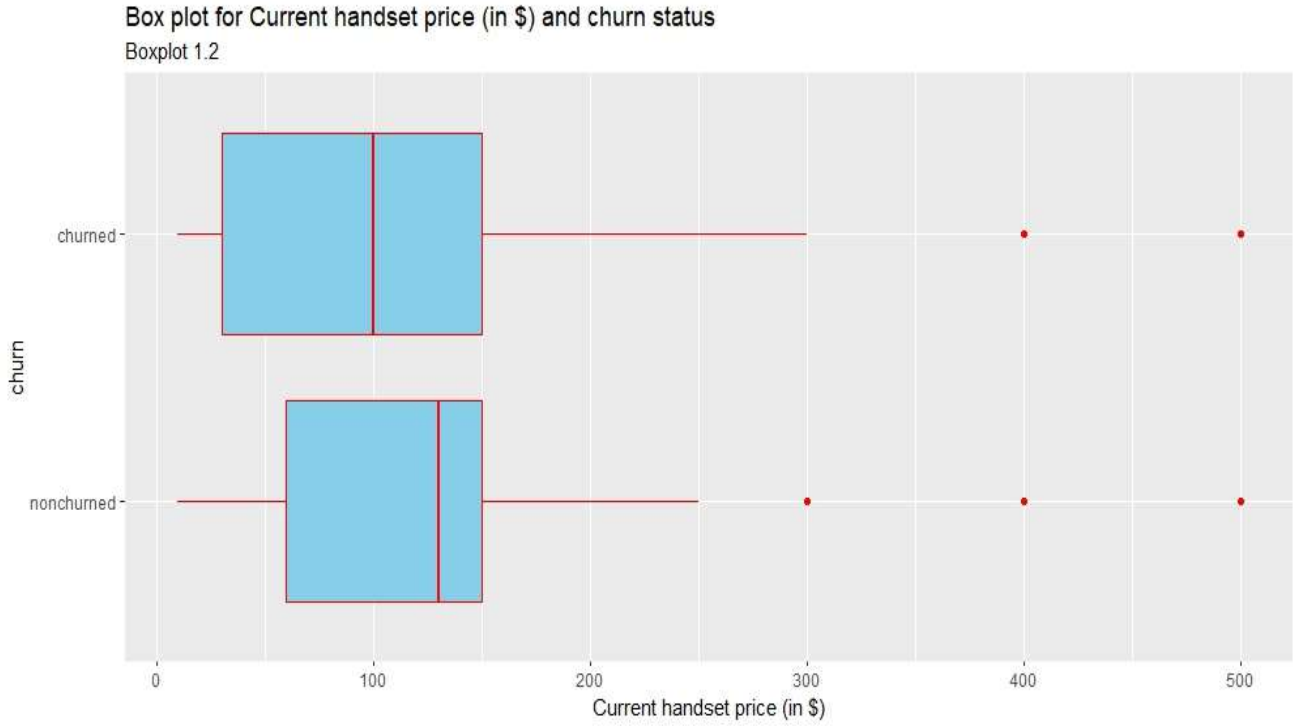
From this Count plot, we can see as the income level increases the frequencies (size of dot) of having any no. of cars increases. And frequencies of having more no. car cars (in a particular income level) are lesser. We can also notice that most people who have car (of any number) are of income level 6 i.e., $50,000 to $74,999 yearly

Now, we want to get some information on Number of days (age) of current equipment and churn status



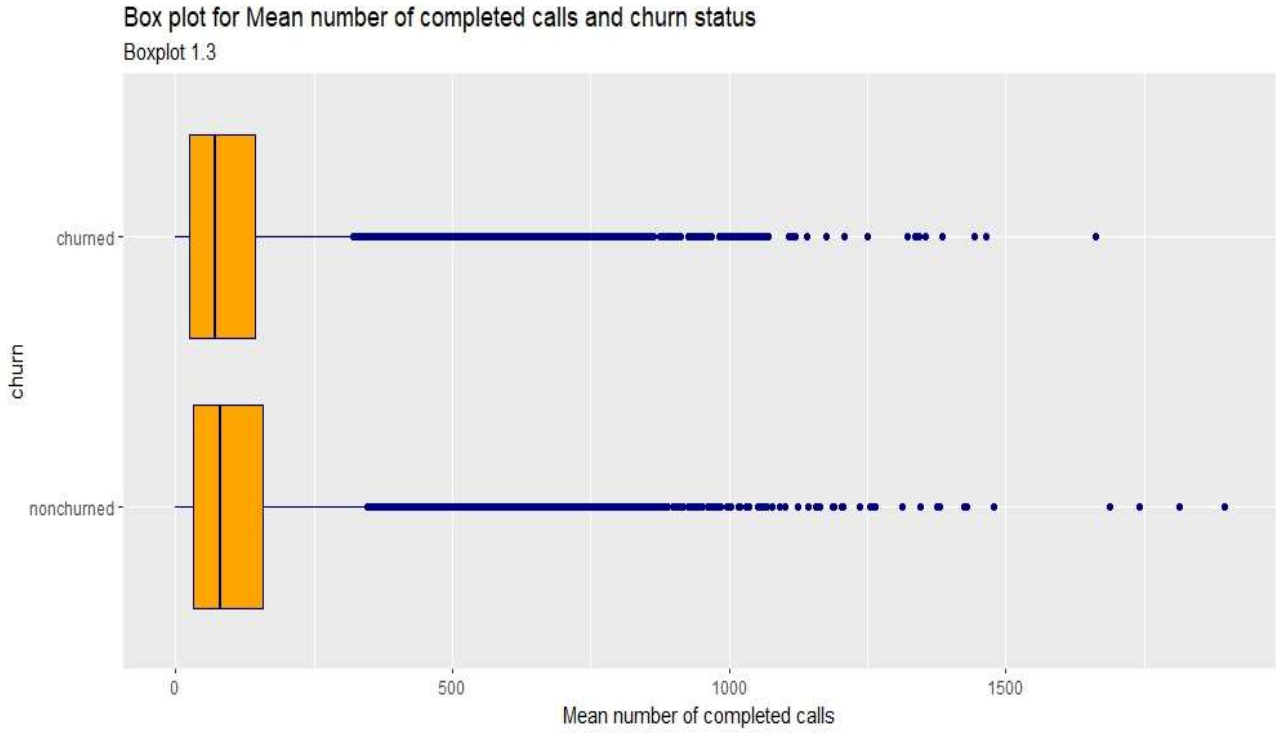
From this box plot, we can see that median of Number of days (age) of current equipment for churn (366) is higher than that of non-churn (310). So, we can say churned have positive association with Number of days (age) of current equipment. For both churn status it is positively skewed.

Now, we want to get some information on Current handset price (in $) and churn status



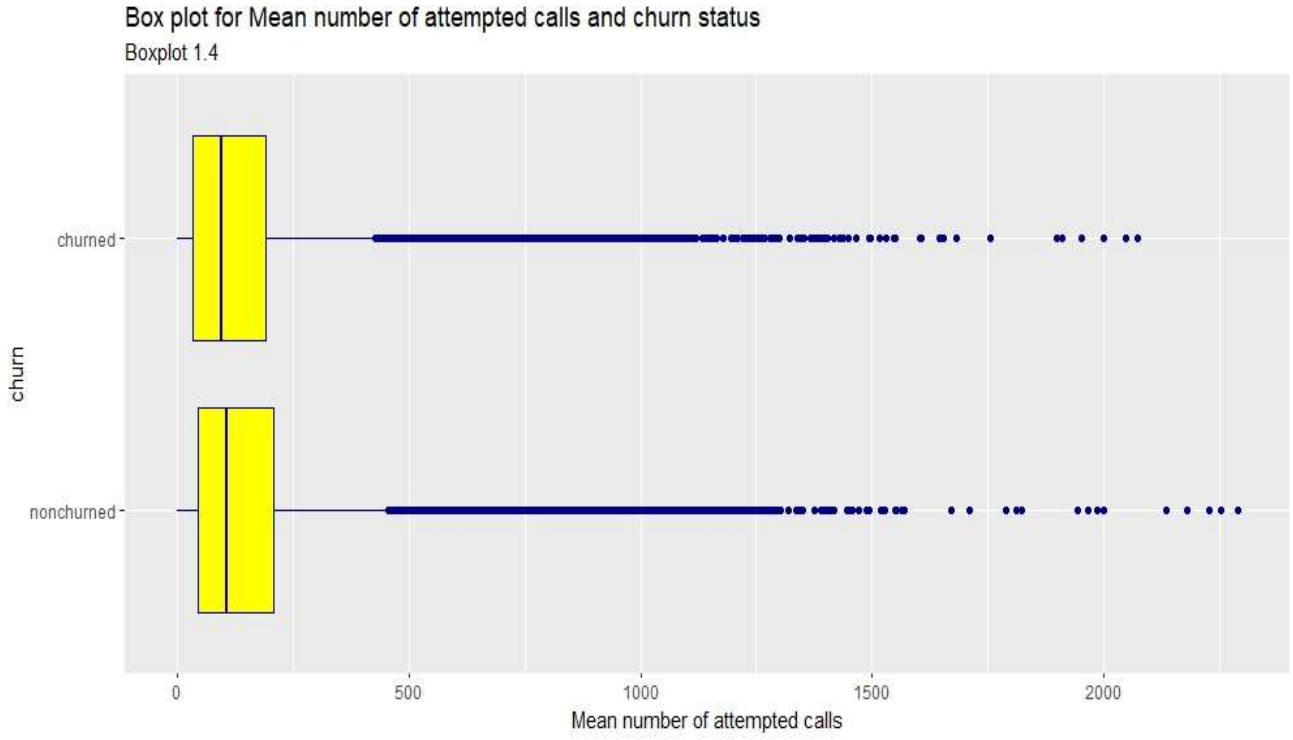
From this box plot, we can see that median of Current handset price (in $) for churn (99.98999) is lesser than that of non-churn (129.99). So, we can say churned have negative association with Current handset price. For both churn status it is negatively skewed.

Now, we want to get some information on Mean number of completed calls and churn status



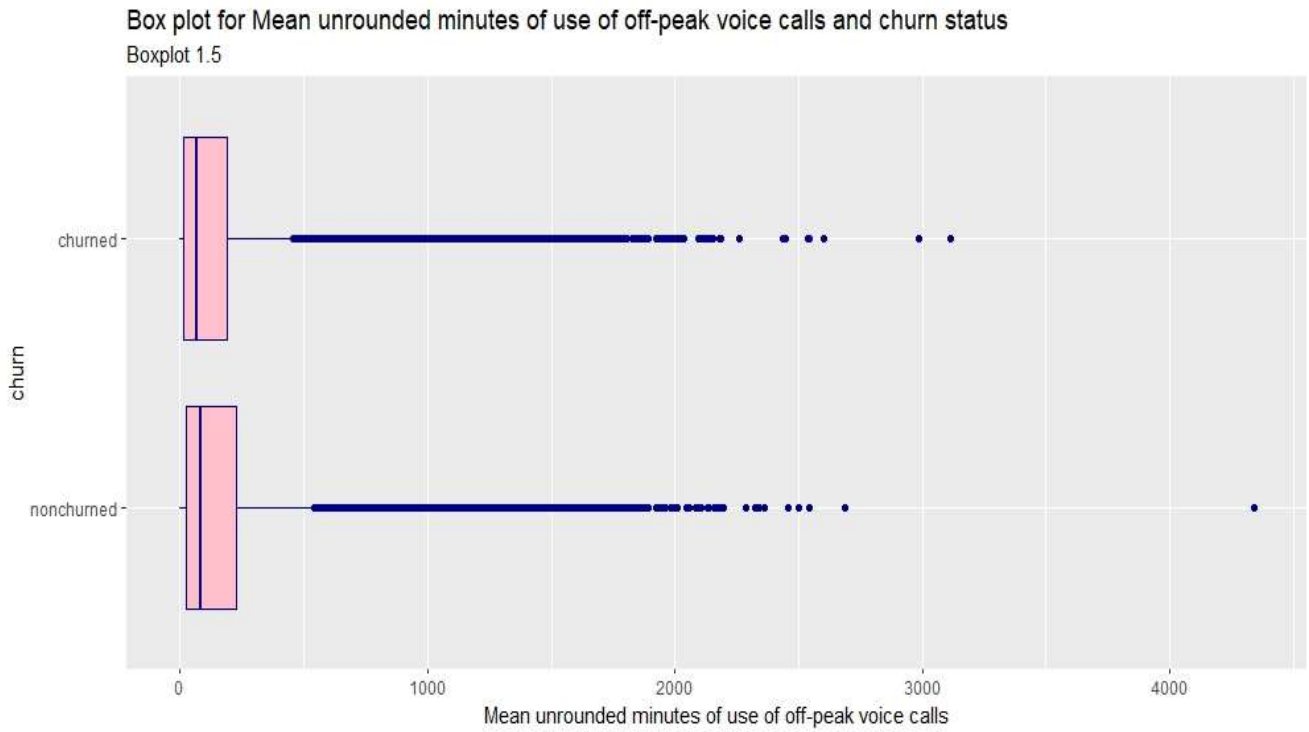
From this box plot, we can see that median of Mean number of completed calls for churn (71.33333) is lesser than that of non-churn (81.33333). So, we can say churned have negative association with Mean number of completed calls. For both churn status it is positively skewed.

Now, we want to get some information on Mean number of attempted calls and churn status



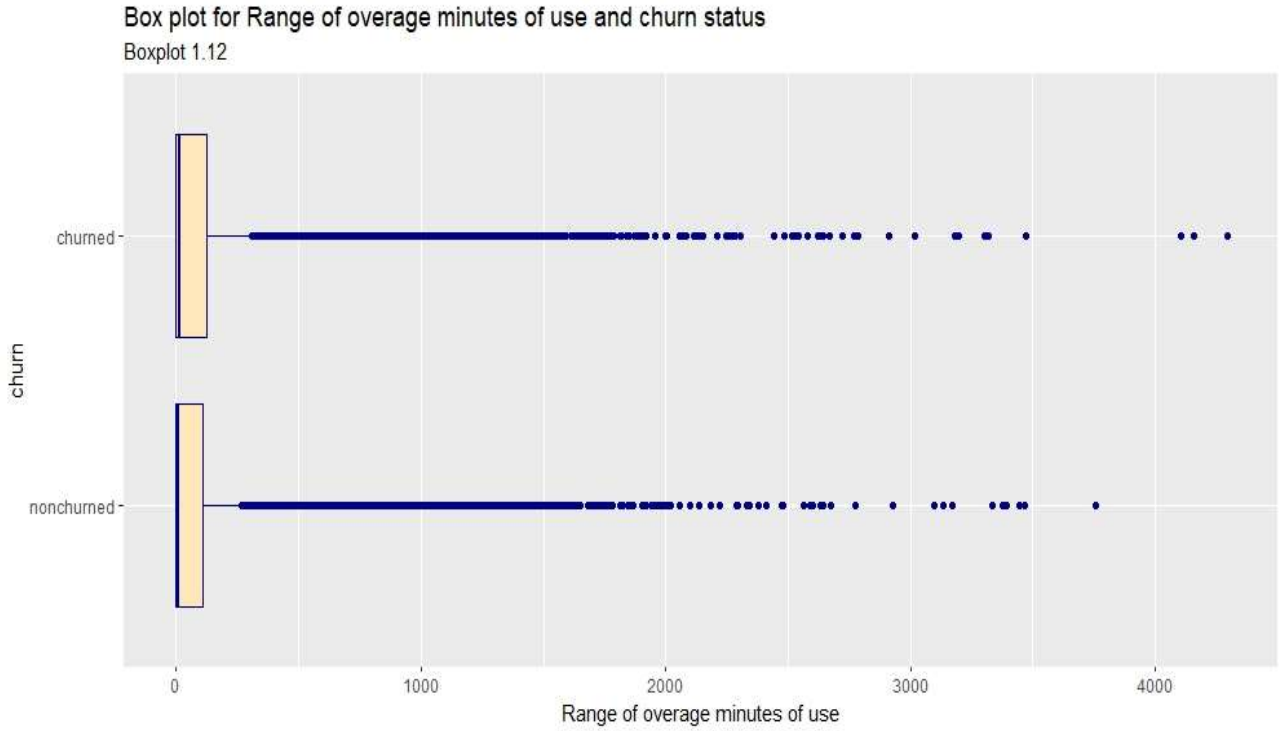
From this box plot, we can see that median of Mean number of attempted calls for churn (94.66667) is lesser than that of non-churn (107.3333). So, we can say churned have negative association with Mean number of completed calls. For both churn status it is positively skewed.

Now, we want to get some information on Mean unrounded minutes of use of off-peak voice calls and churn status



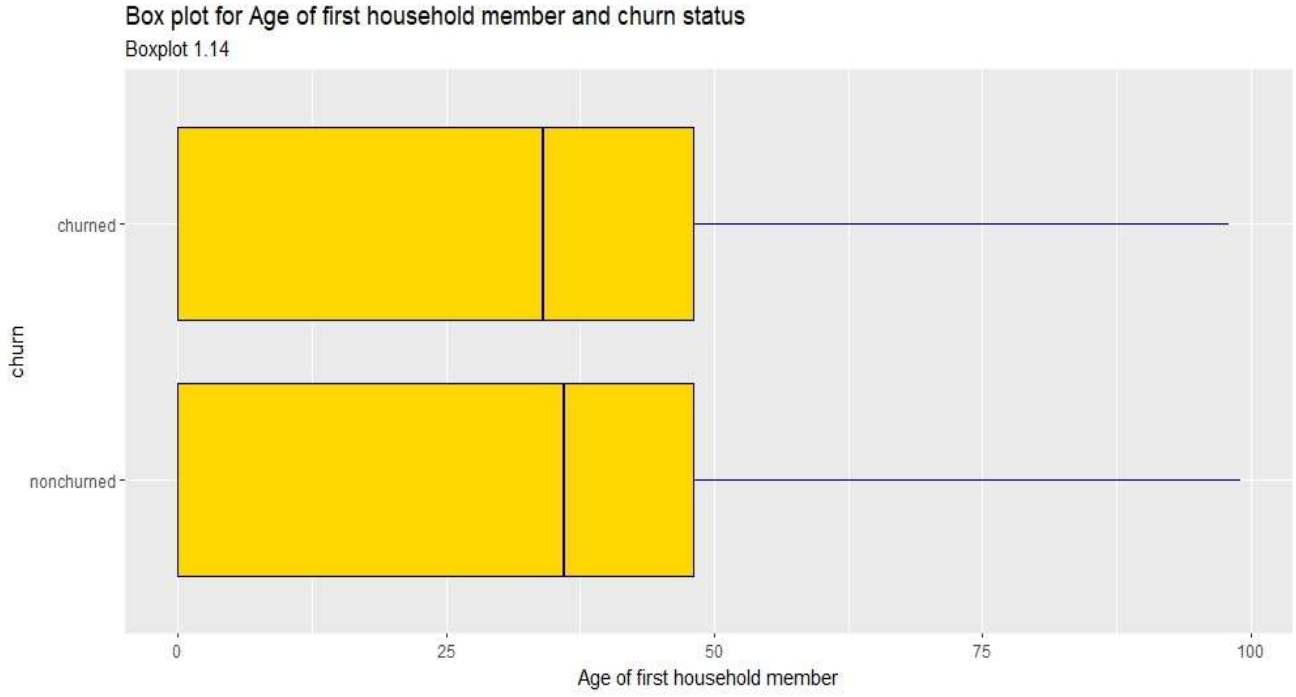
From this box plot, we can see that median of Mean unrounded minutes of use of off-peak voice calls for churn (66.40167) is lesser than that of non-churn (85.89667). So, we can say churned have negative association with Mean unrounded minutes of use of off-peak voice calls. For both churn status it is positively skewed.

Now, we want to get some information on Range of overage minutes of use and churn status



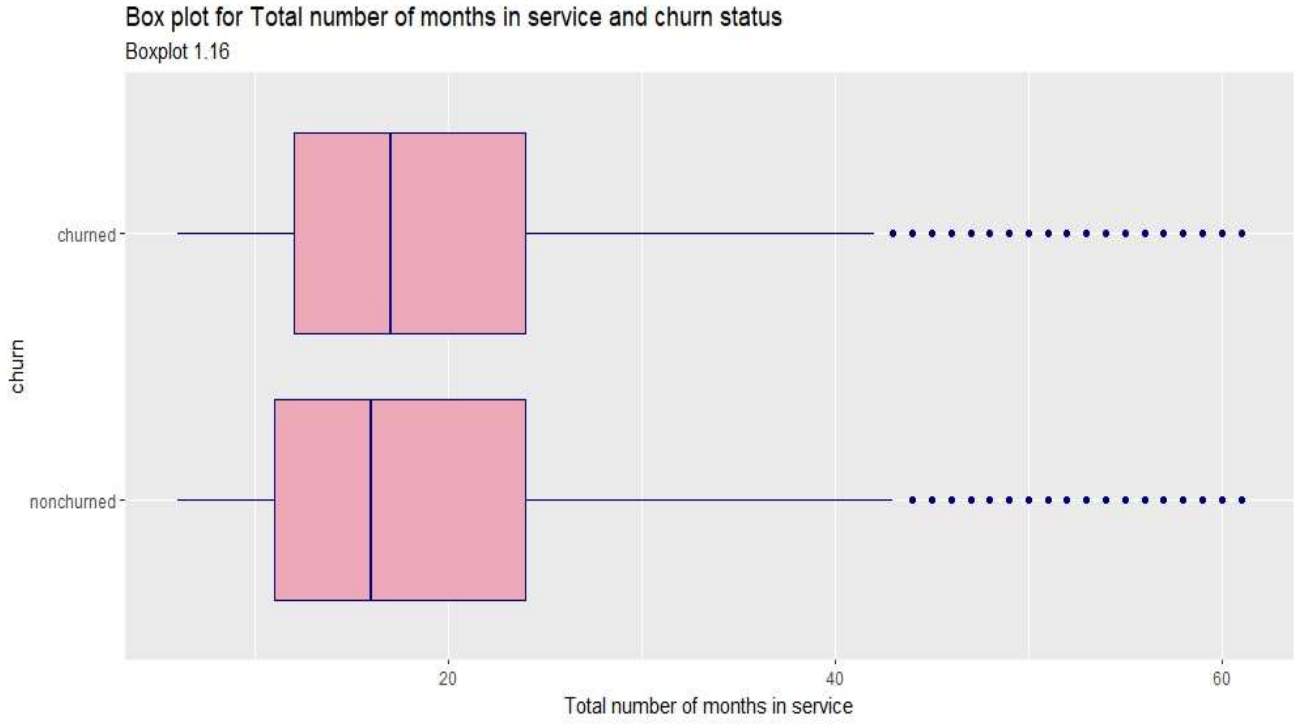
From this box plot, we can see that median of Range of overage minutes of use for churn (12) is higher than that of non-churn (8). So, we can say churned have positive association with Range of overage minutes of use. For both churn status it is positively skewed.

Now, we want to get some information on Age of first household member and churn status



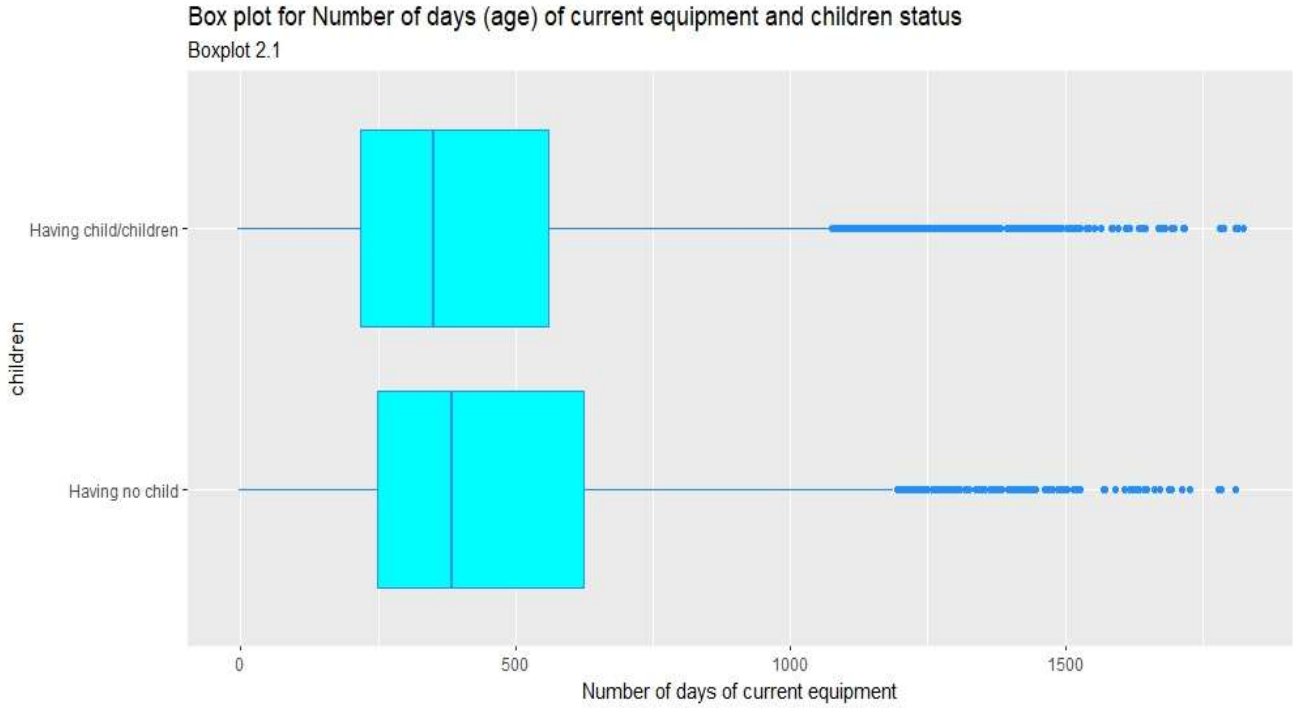
From this box plot, we can see that median of Age of first household member for churn (34) is lesser than that of non-churn (36). So, we can say churned have negative association with Age of first household member. For both churn status it is negatively skewed.

Now, we want to get some information on Total number of months in service and churn status



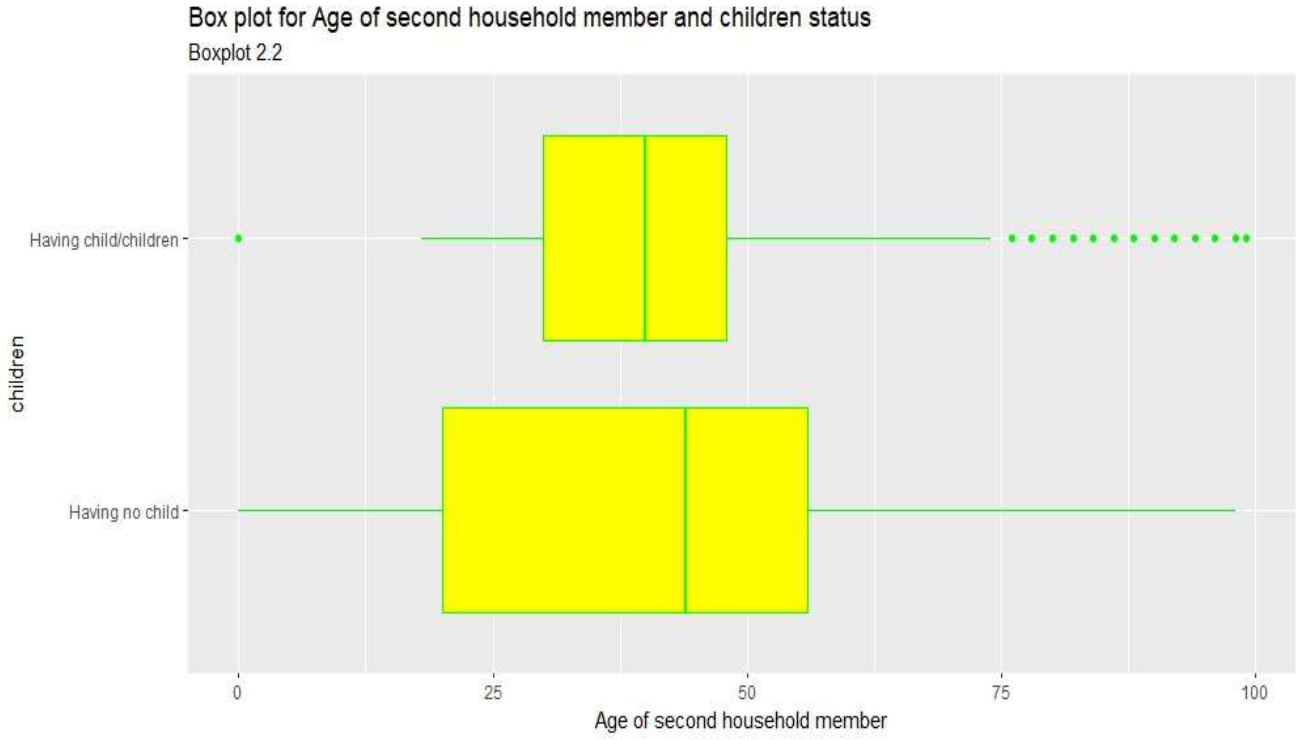
From this box plot, we can see that median of Total number of months in service for churn (17) is higher than that of non-churn (16). So, we can say churned have positive association with Total number of months in service. For both churn status it is positively skewed.

Now, we want to get some information on Number of days (age) of current equipment and children status



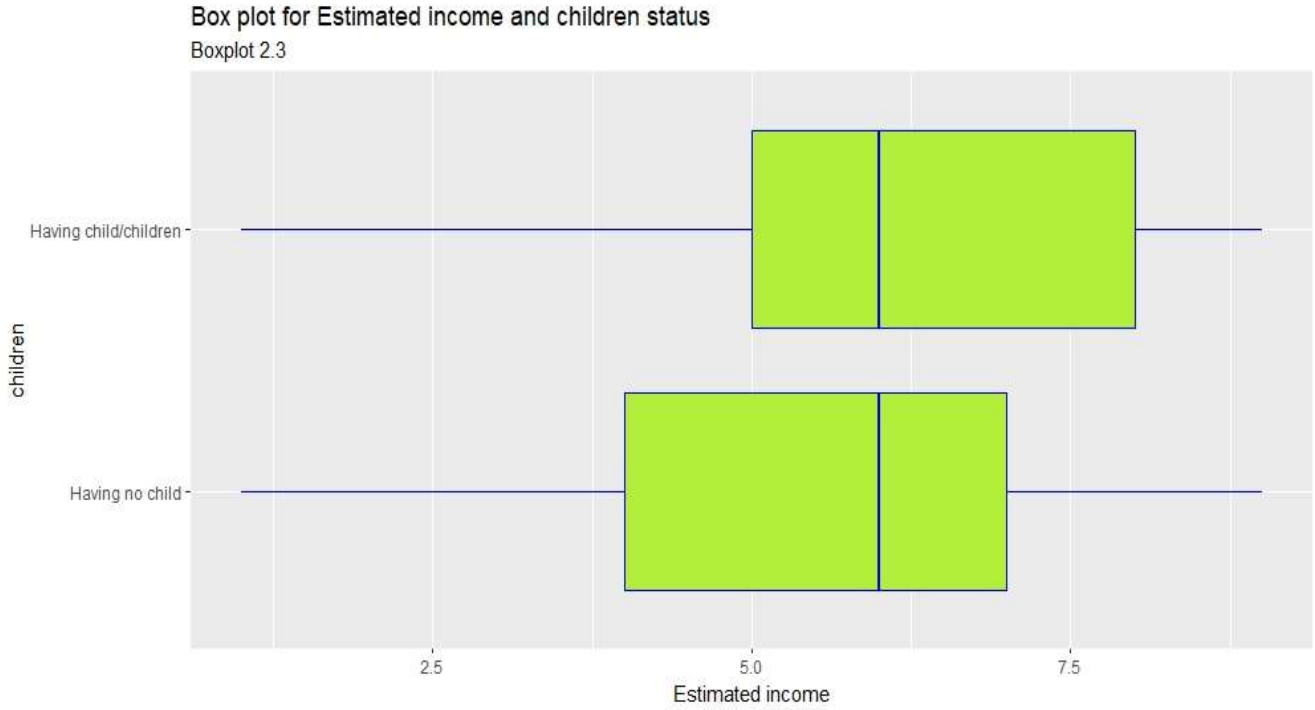
From this box plot, we can see that median of Number of days (age) of current equipment for Having child/children (352) is lesser than that of Having no child (384.5). So, we can say Having child/children have negative association with Number of days (age) of current equipment. For both children status it is positively skewed.

Now, we want to get some information on Age of second household member and children status



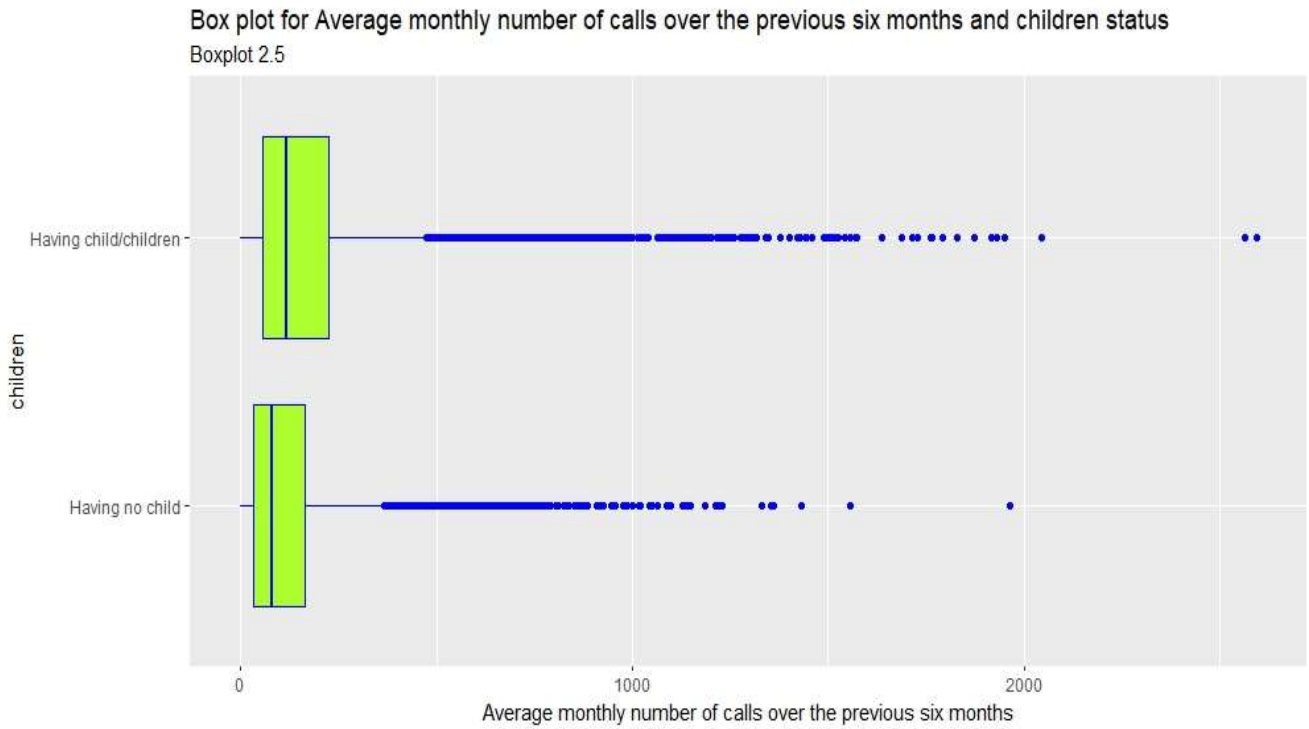
From this box plot, we can see that median of Age of second household member for Having child/children (40) is lesser than that of Having no child (44). So, we can say Having child/children have negative association with Age of second household member. For both children status it is negatively skewed.

Now, we want to get some information on Estimated income and children status



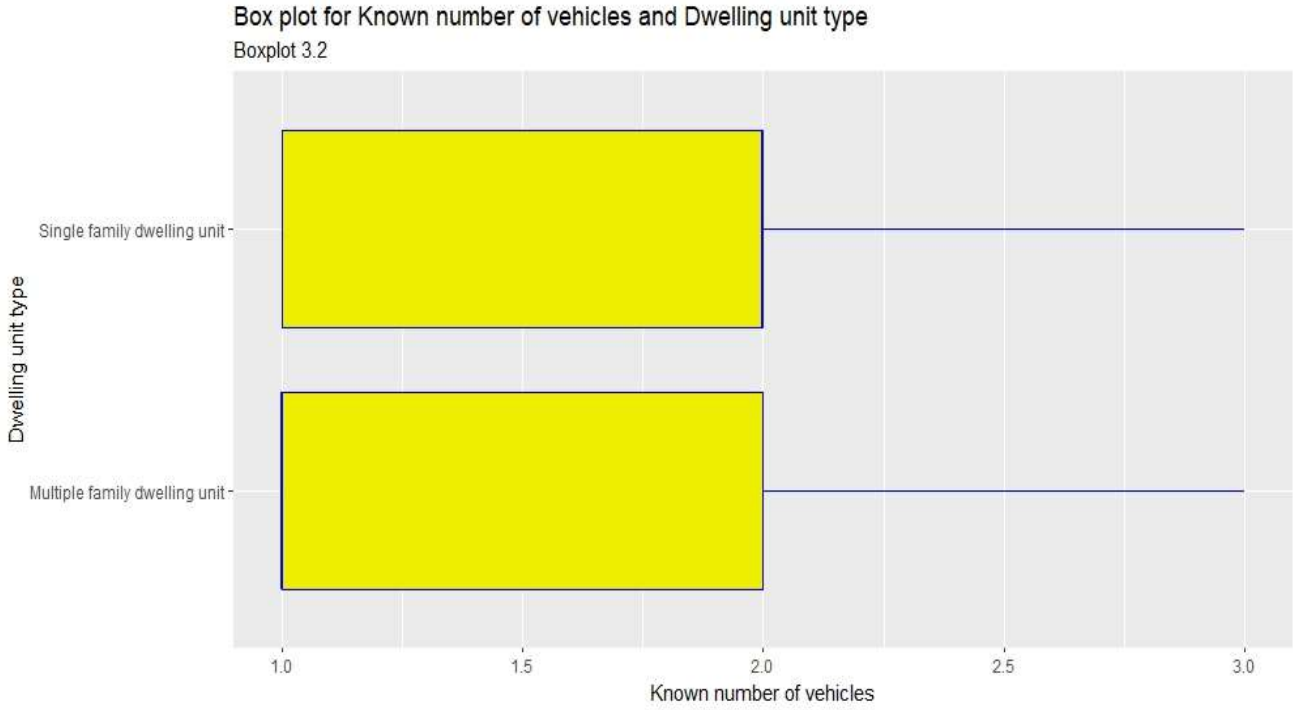
From this box plot, we can see that 1st and 3rd quartile of Estimated income for Having child/children (5.000, 8.000) is higher than that of Having no child (4.000, 7.000). So, we can say Having child/children have positive association with Estimated income. For Having child/children it is positively skewed. And negatively skewed for Having no child

Now, we want to get some information on Average monthly number of calls over the previous six months and children status



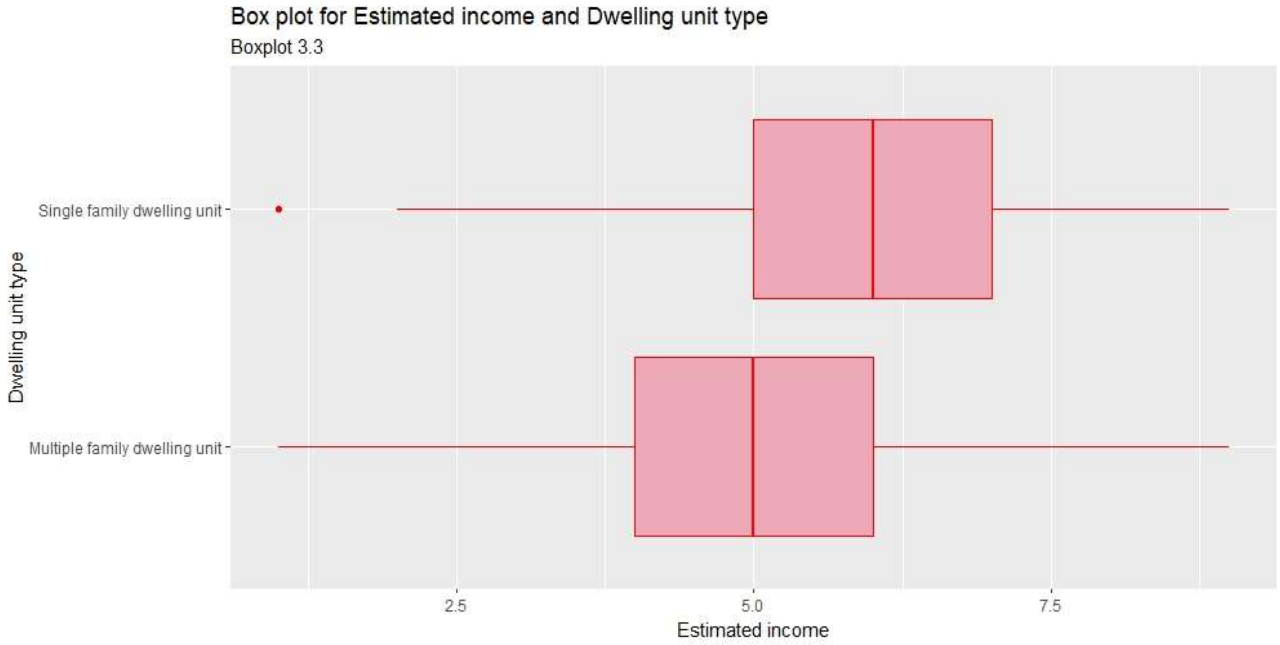
From this box plot, we can see that median of Average monthly number of calls over the previous six months for Having child/children (319) is higher than that of Having no child (247.0). So, we can say Having child/children have positive association with Average monthly number of calls over the previous six months. For both children status it is positively skewed.

Now, we want to get some information on Known number of vehicles and Dwelling unit type



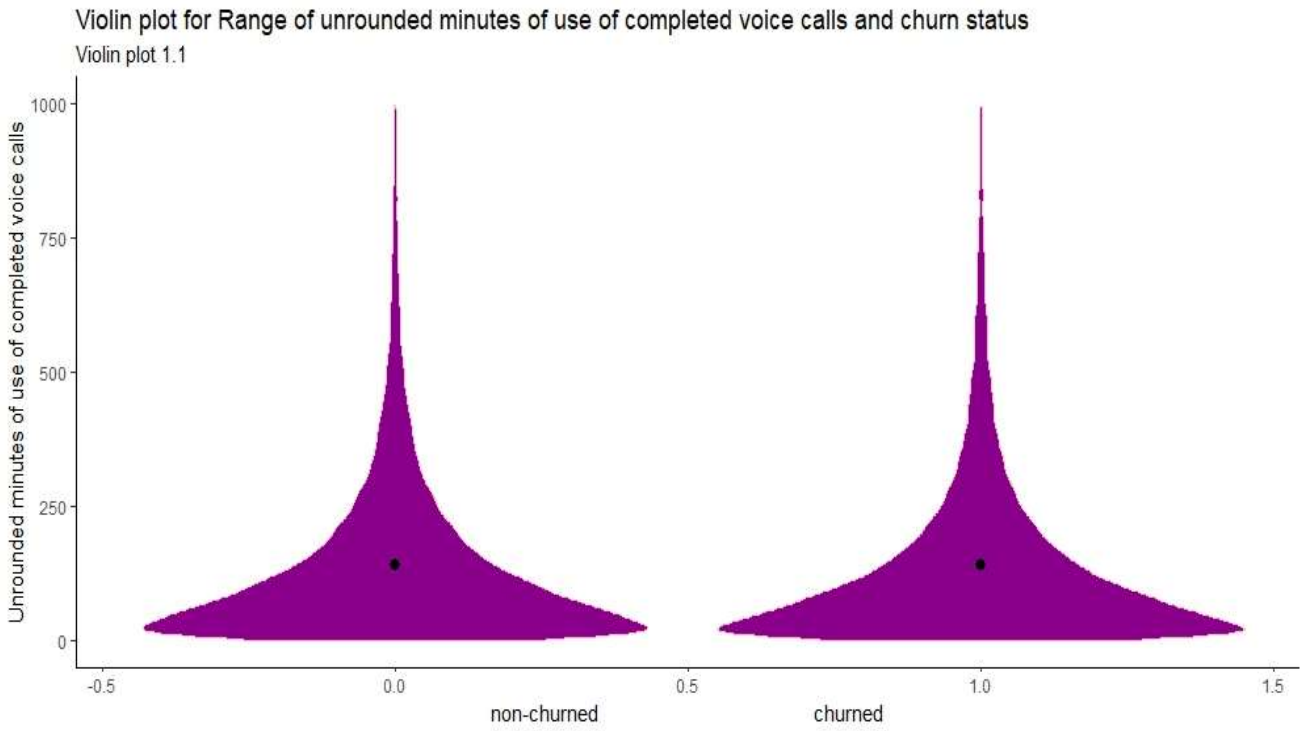
From this box plot, we can see that median of Known number of vehicles for Single family dwelling unit (2) is higher than that of Multiple family dwelling unit (1). So, we can say Single family dwelling have positive association with Known number of vehicles.

Now, we want to get some information on Estimated income and Dwelling unit type



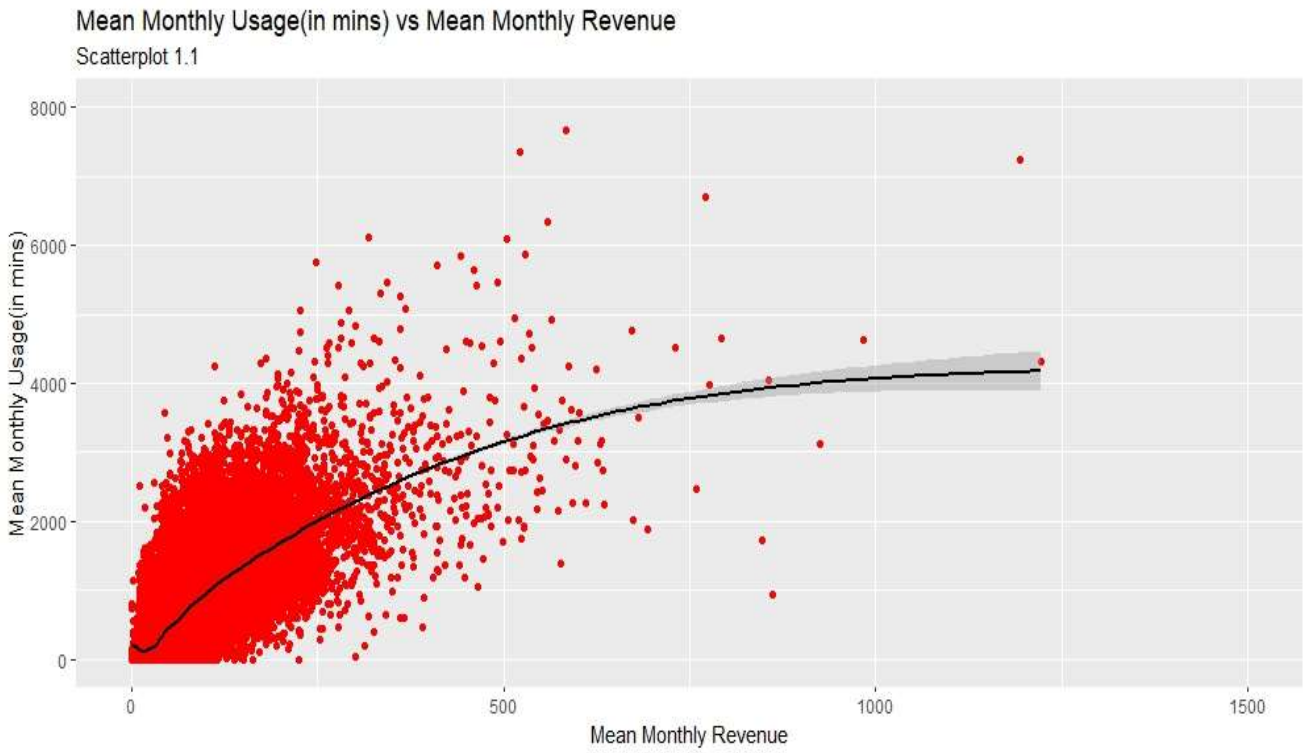
From this box plot, we can see that median of Estimated income for Single family dwelling unit (6) is higher than that of Multiple family dwelling unit (5). So, we can say Single family dwelling have positive association with Estimated income. Seems to be some odd thing

Now, we want to get some information on Range of unrounded minutes of use of completed voice calls and churn status



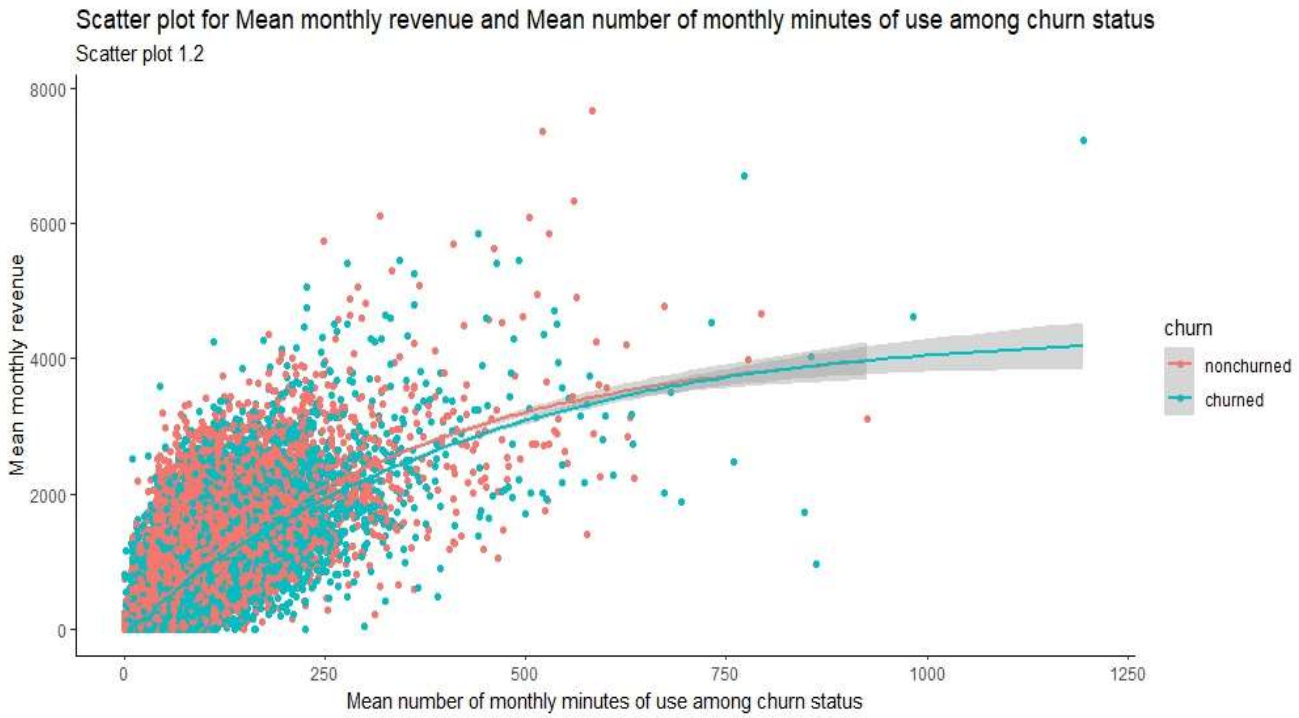
From this violin plot, we can see that for any churn status mean and median of Range of unrounded minutes of use of completed voice calls are almost same and positively skewed

Now, we want to get some information on relation of Mean Monthly Usage (in mins) and Mean Monthly Revenue



From this scatter plot, we can see that there is a positive association between Mean Monthly Usage (in mins) and Mean Monthly Revenue. And the smooth curve also states the same

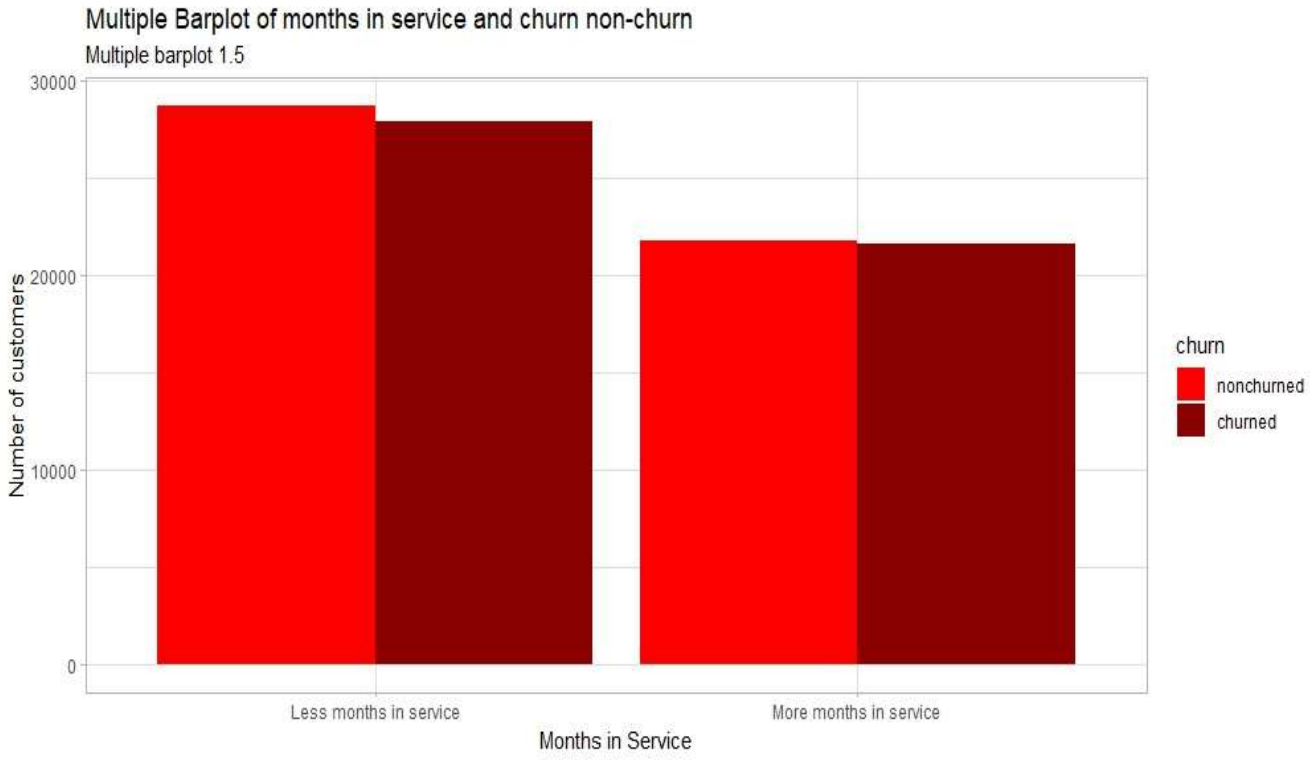
Now, we want to get some information on relation of Mean Monthly Usage (in mins) and Mean Monthly Revenue along with churn status



From this scatter plot, we can see that there is a positive association between Mean Monthly Usage (in mins) and Mean Monthly Revenue for any churn status. And the smooth curve also states the same. But we can see from 100 mins of Mean Monthly Usage, the smooth curve for non-churned is higher than churned one.

Building new features:

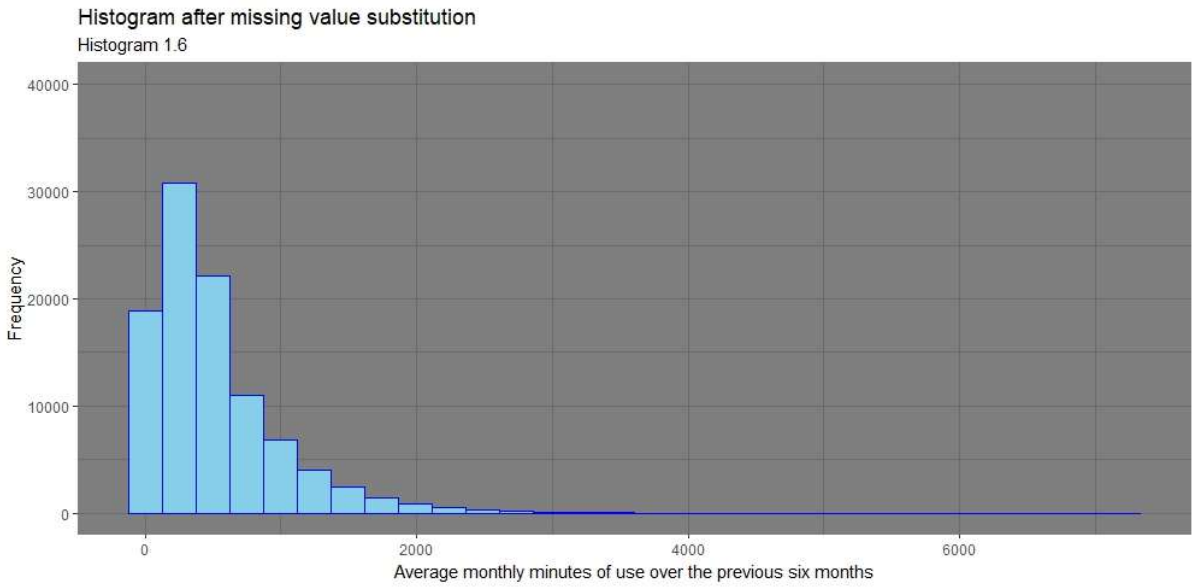
manipulation — addition, deletion, combination, mutation — of your data set to improve machine learning model training, leading to better performance and greater accuracy. Here, we have made a new feature by transferring numerical to categorical. Months in service greater than 18 is given as More months in service and less than equal to 18 is given as Less months in service.

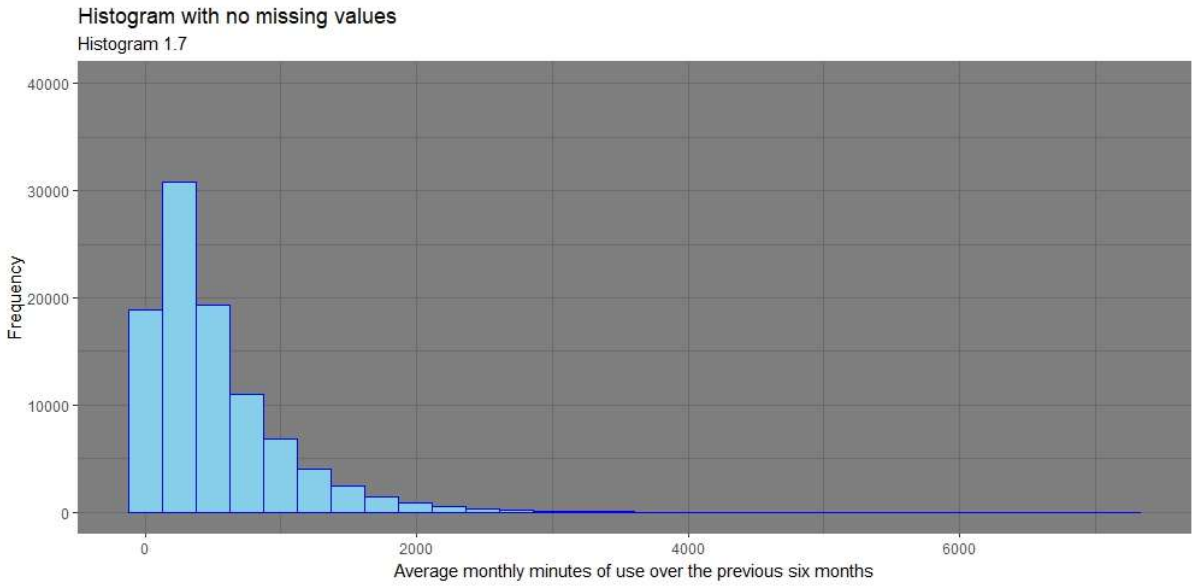


Here we can see more non-churned people in both the categories.

Identifying missing values and substituting with mean or median:

Here we have assigned the NA values with mean of non-NA values taking a continuous variable.

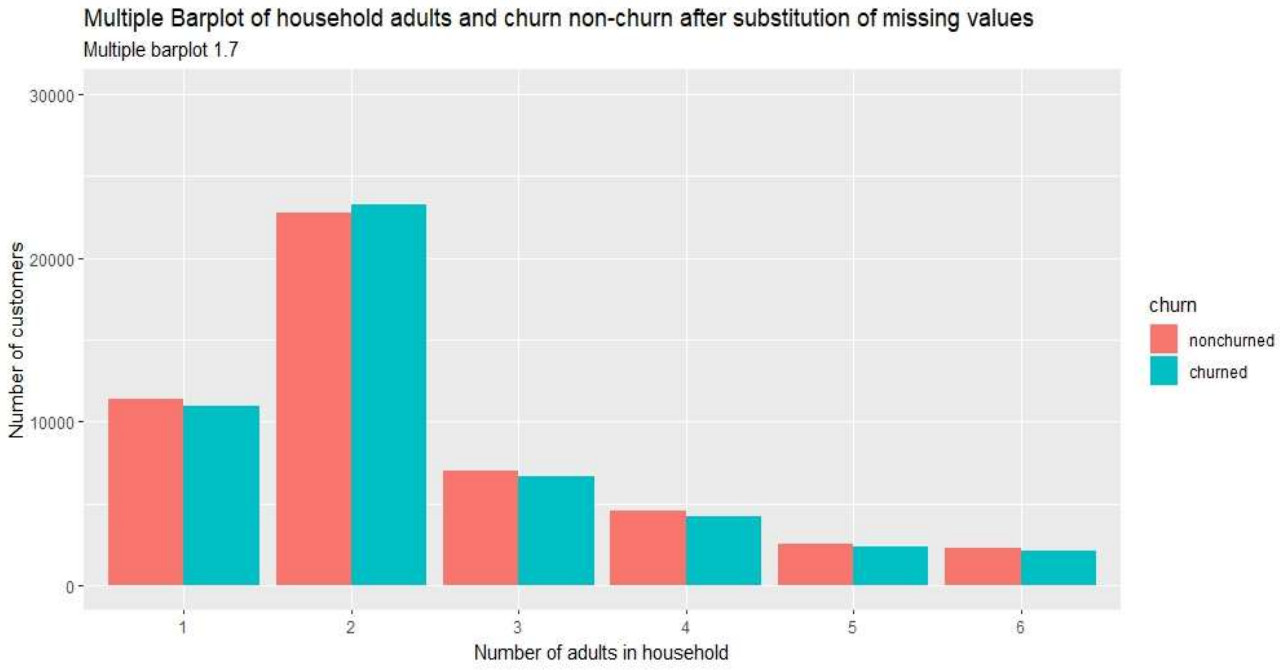


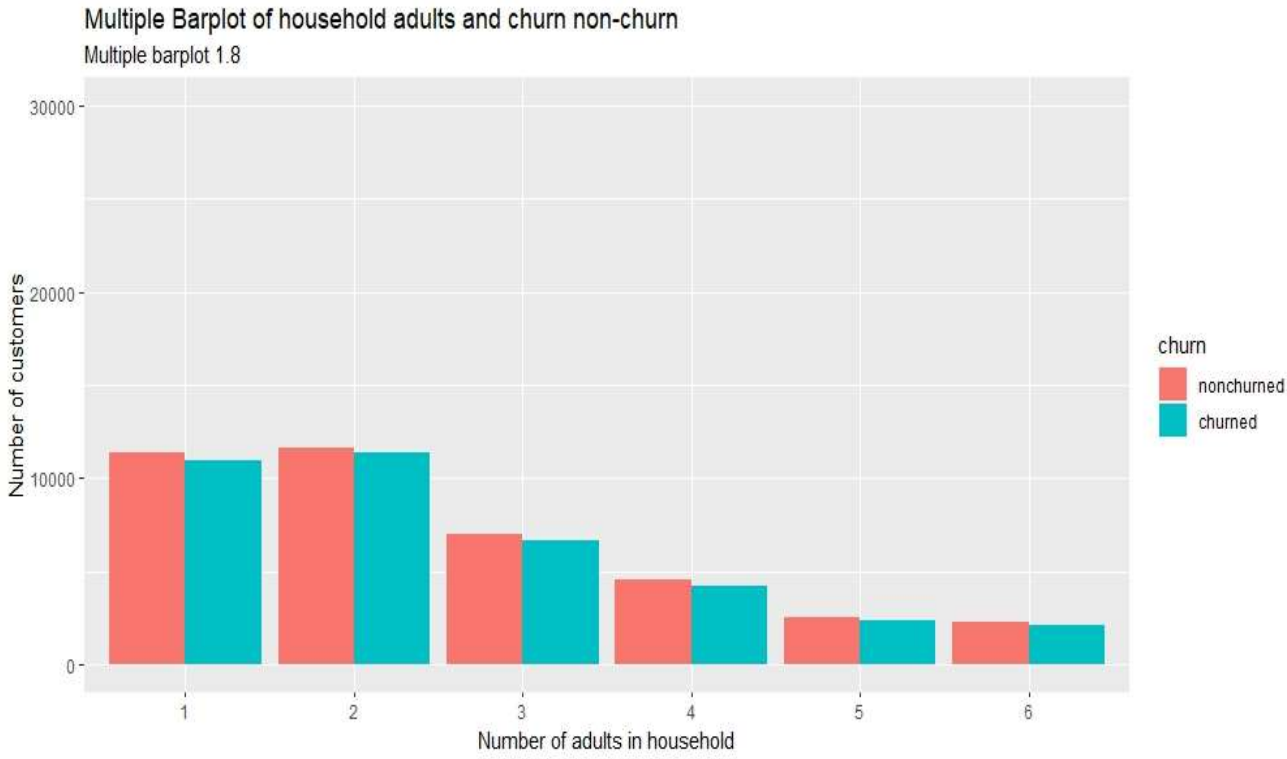


Here we can see that the third bin having roughly the interval 500-750 i.e., the interval in which the mean belongs has increased its frequency after substitution.

Substitution with modal level or class:

Here we have assigned the NA values with modal class.



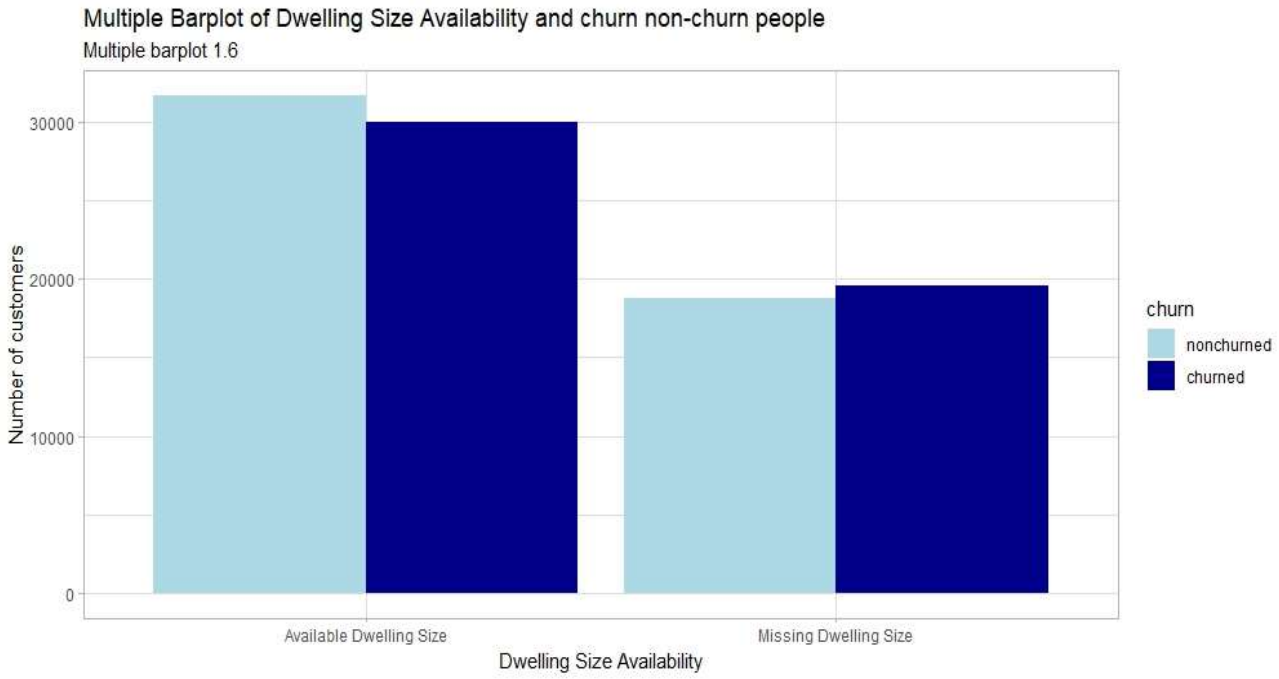


Here we can see that the third bar has increased its height after substitution. The modal class is preserved. Also we notice that before substitution non-churned

had dominated over all classes but after substitution in modal class churn dominates i.e., more people are churned in the modal class.

Missing values for a category and comparing them with available values respect to churn status.

Missing vs non-missing



Here the category is dwelling size. We can notice that values which have this category missing are more churned and the available ones are more non-churned

CONCLUSION

Variables like Number of days (age) of current equipment, Range of overage minutes of use, Total number of months in service are positively associated with Churning.

Whereas, Current handset price (in $), Mean number of completed calls, Mean number of attempted calls, Mean unrounded minutes of use of off-peak voice calls, Mean number of off-peak voice calls, Mean unrounded minutes of use of peak voice calls, Mean number of inbound and outbound peak voice calls, Mean unrounded minutes of use of received voice calls, Mean number of completed voice calls, Mean number of attempted voice calls placed, Mean number of monthly minutes of use, Average monthly minutes of use over the previous six months: are negatively associated with Churning.

Many more inferences are also found.

REFERENCES

R Cookbook

Slides and notes shared to us

geeksforgeeks

FUTURE SCOPE OF WORK

There are many columns. There is always scope to get more insights from the data. There is not only churn status but many other, we can get inferences about them from other variables which are not used for churn purpose.

TOOLS USED

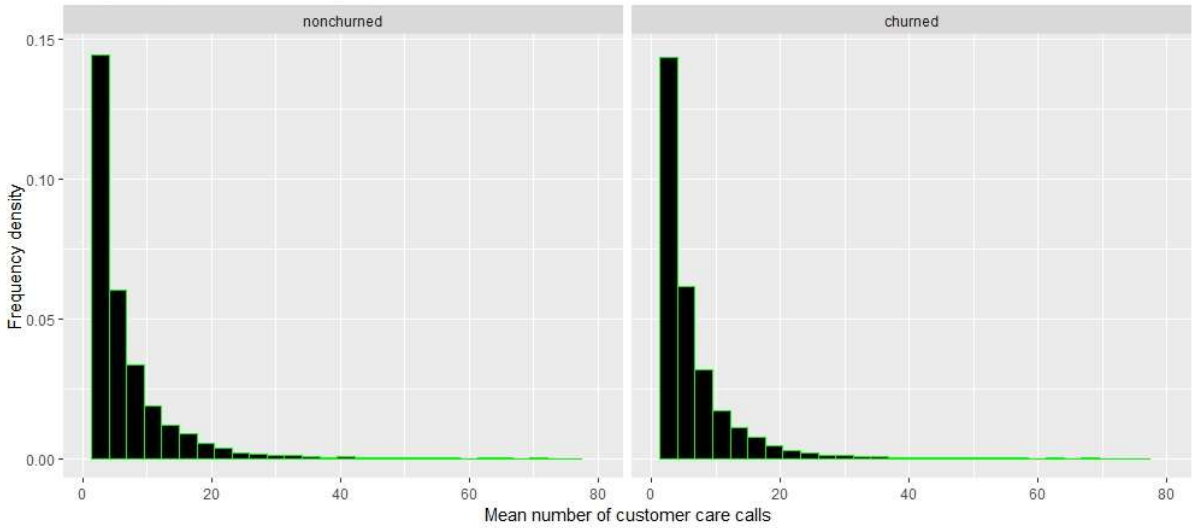
RStudio

ggplot2 package

APPENDIX

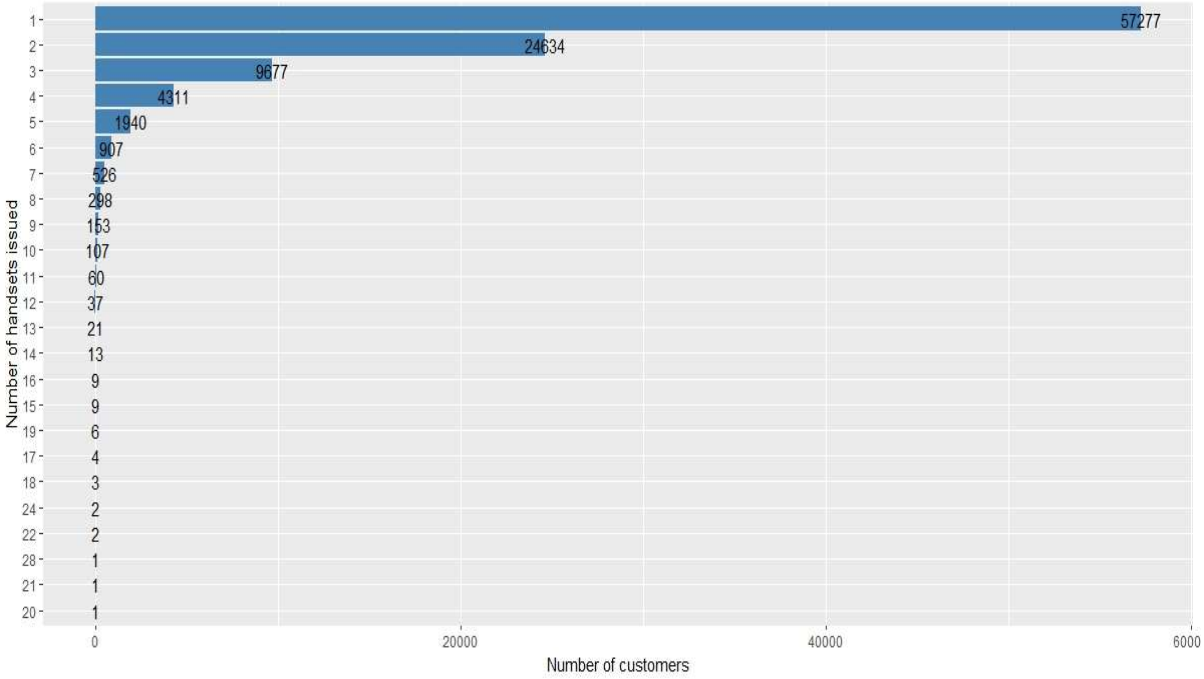
Mean number of customer care calls among Churn and Non-churn data

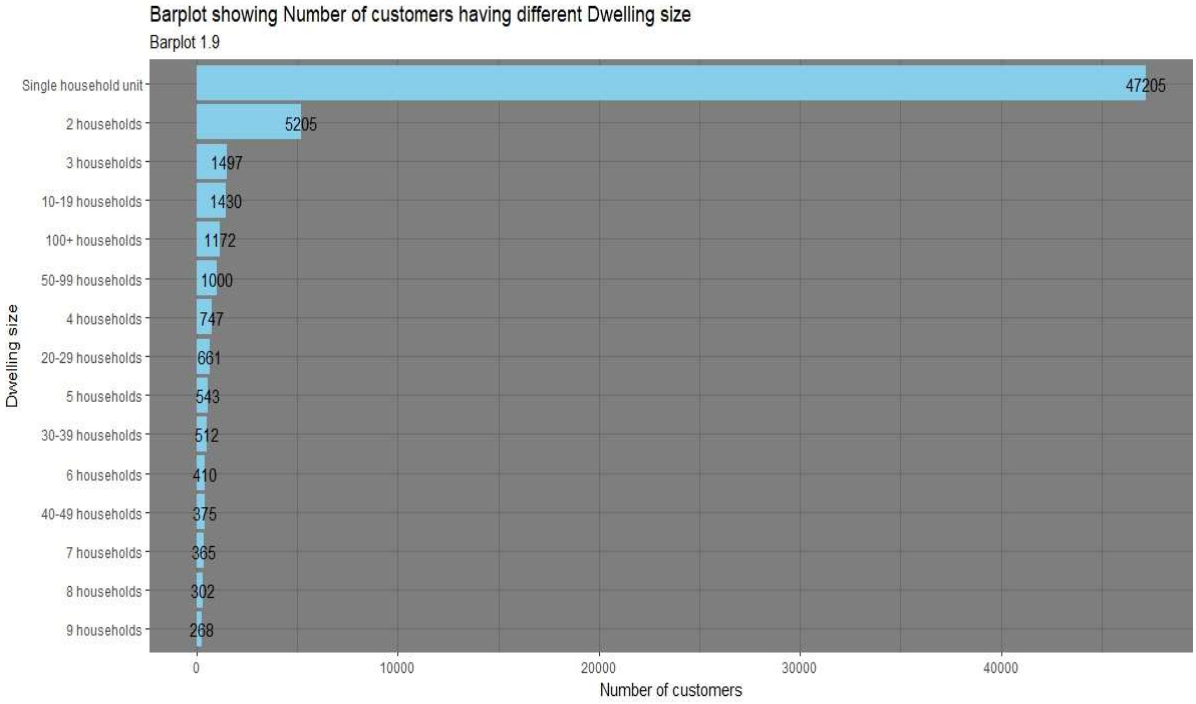
Histogram 1.4

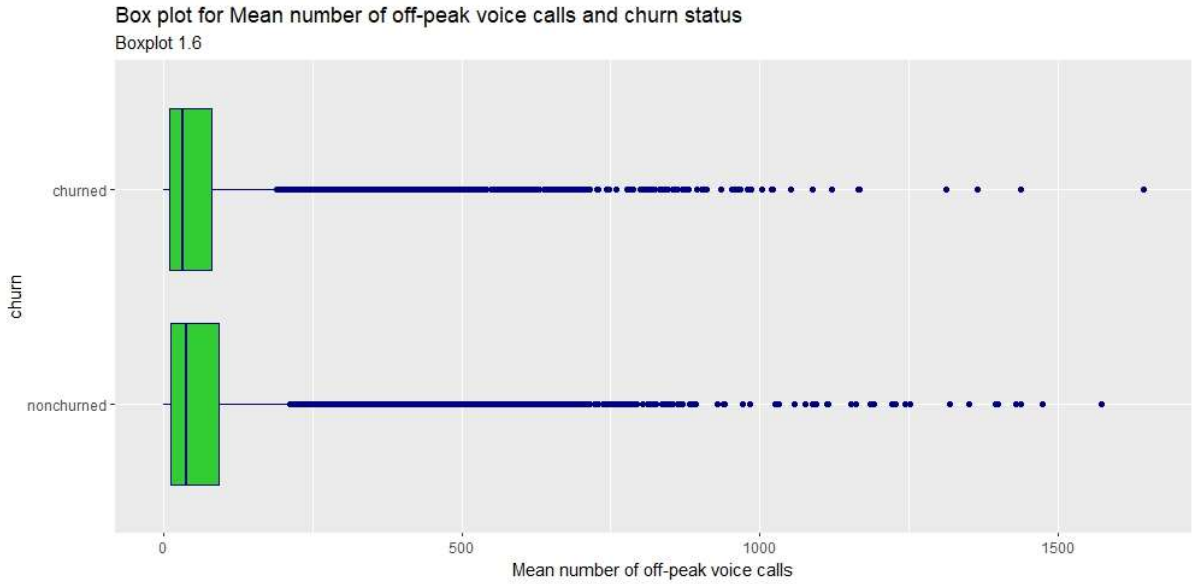


Barplot showing Number of customers with different number of handsets issued

Barplot 1.5

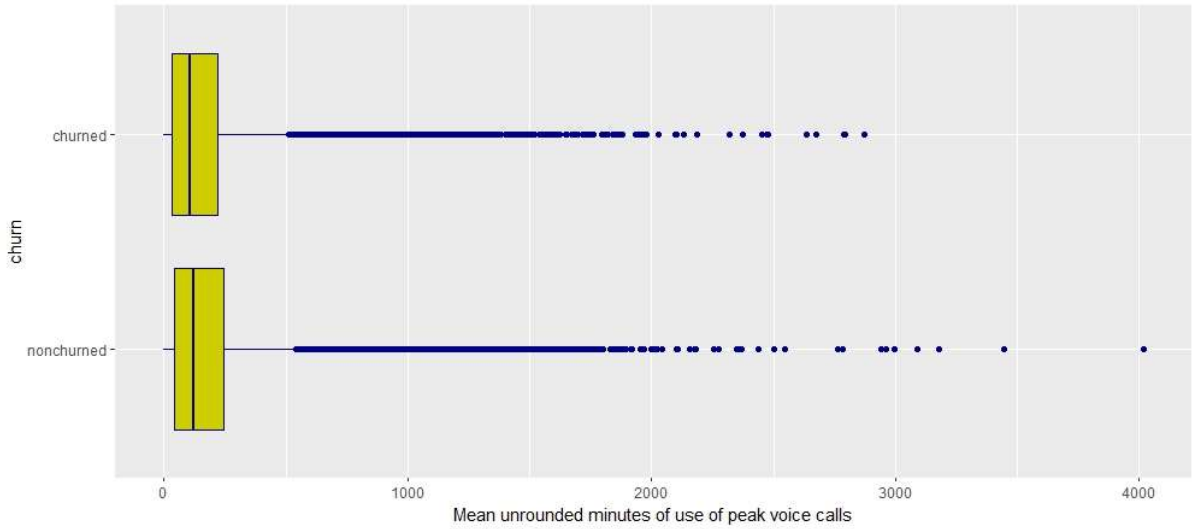



Barplot showing Number of customers having different Dwelling size
Barplot 1.9
Single household unit
2 households
3 households
10-19 households
100+ households
50-99 households
4 households
20-29 households
5 households
30-39 households
6 households
40-49 households
7 households
8 households
9 households
47205
5205
1497
1430
1172
1000
747
661
543
512
410
375
365
302
268
Dwelling size
0
10000
20000
30000
40000
Number of customers


Box plot for Mean number of off-peak voice calls and churn status
Boxplot 1.6
churned
nonchurned
churn
0
500
1000
1500
Mean number of off-peak voice calls

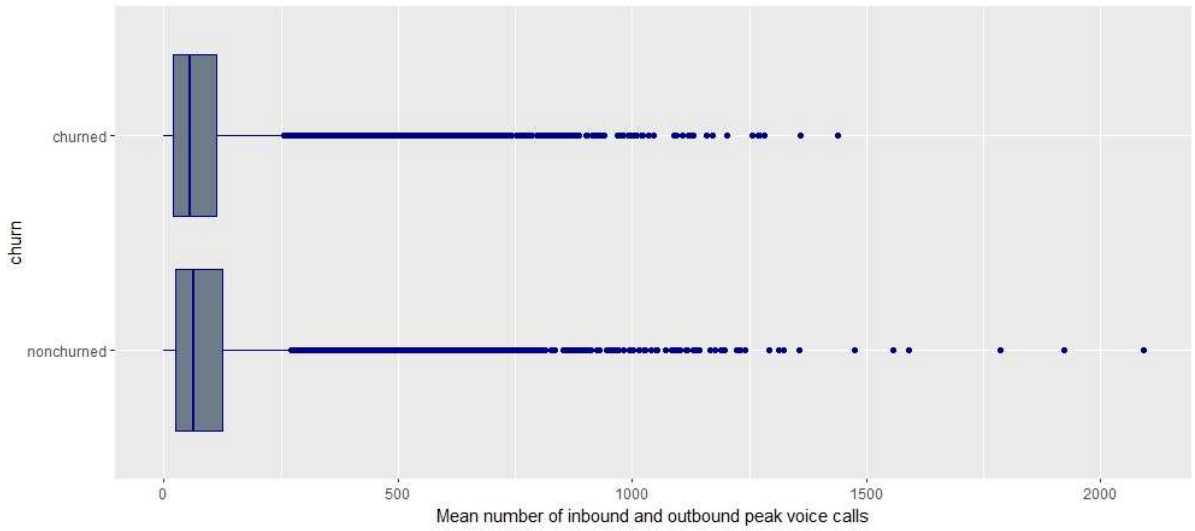
Box plot for Mean unrounded minutes of use of peak voice calls and churn status

Boxplot 1.7



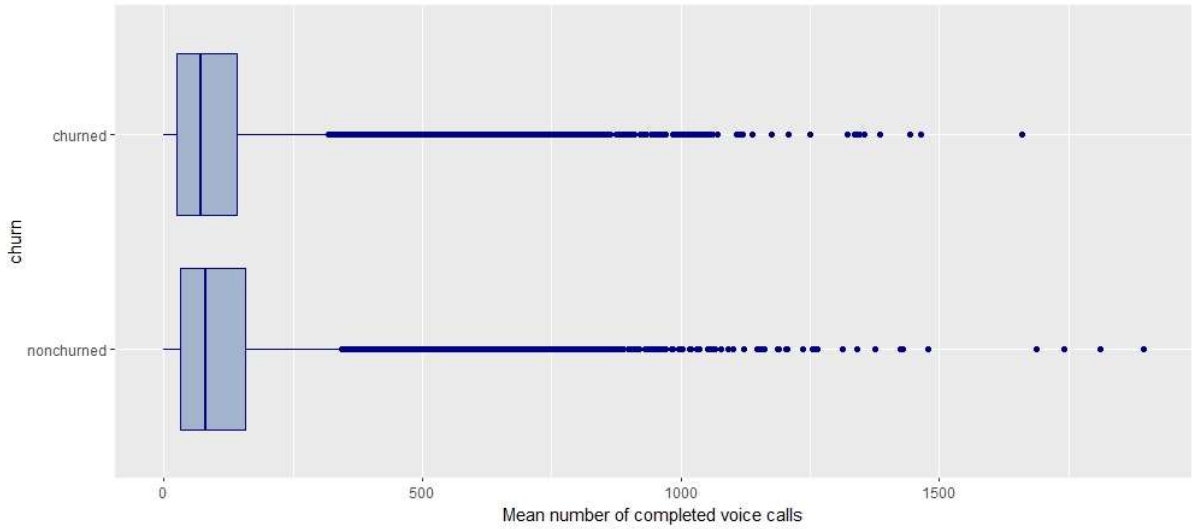
Box plot for Mean number of inbound and outbound peak voice calls and churn status

Boxplot 1.8



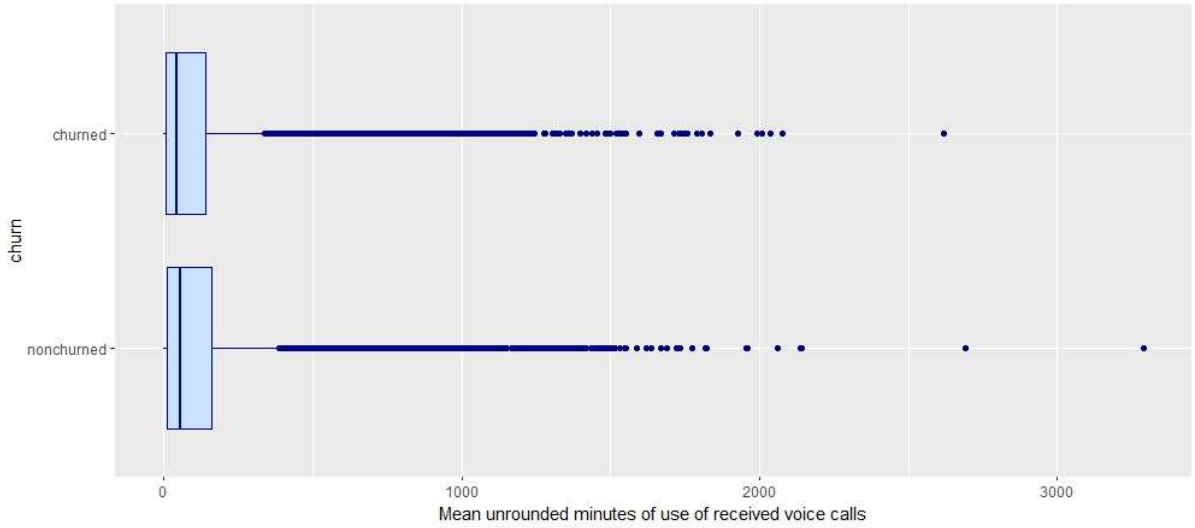
Box plot for Mean number of completed voice calls and churn status

Boxplot 1.10



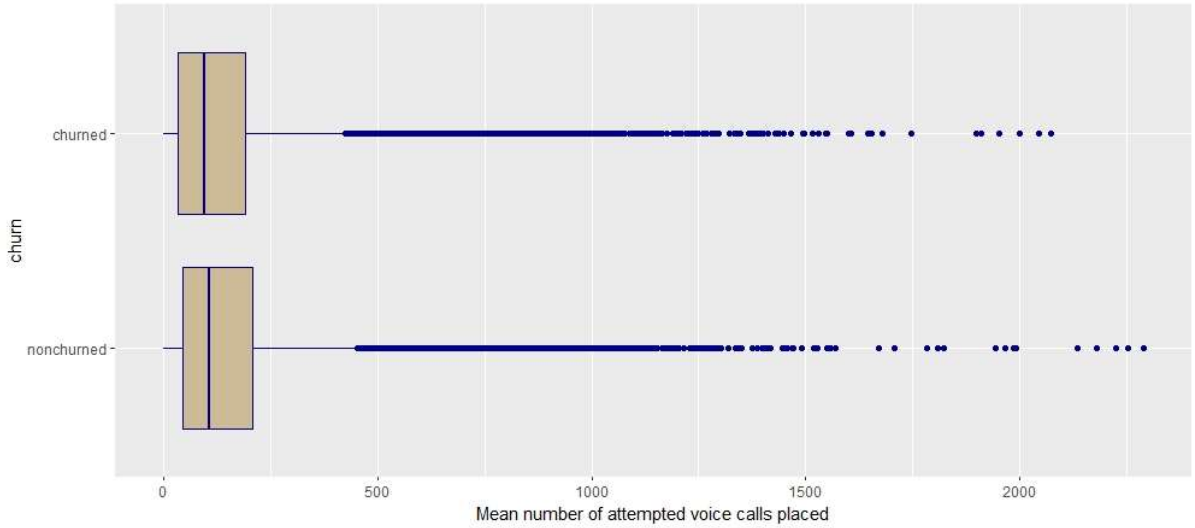
Box plot for Mean unrounded minutes of use of received voice calls and churn status

Boxplot 1.9



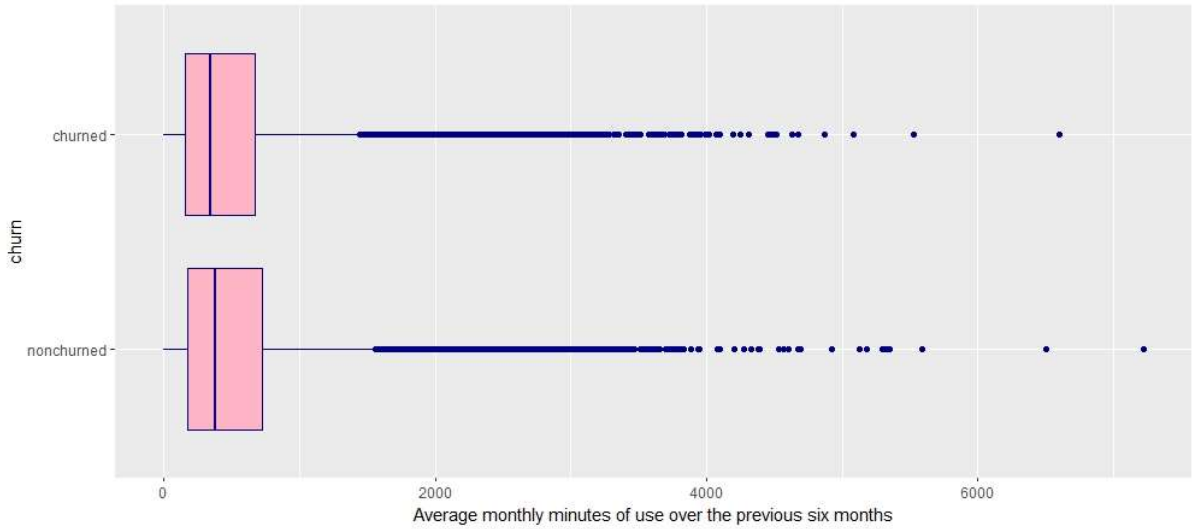
Box plot for Mean number of attempted voice calls placed and churn status

Boxplot 1.11



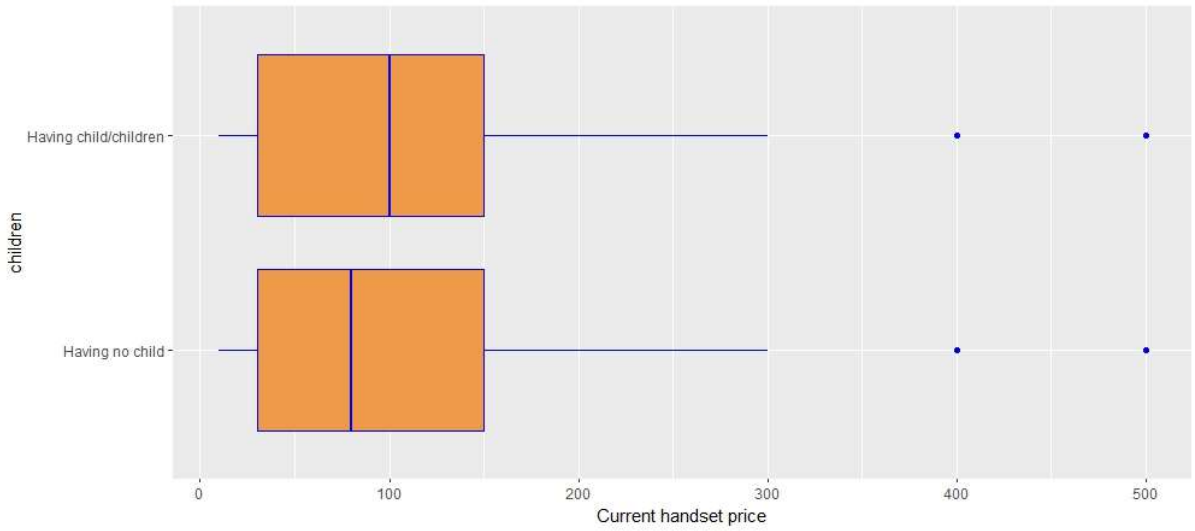
Box plot for Average monthly minutes of use over the previous six months and churn status

Boxplot 1.15



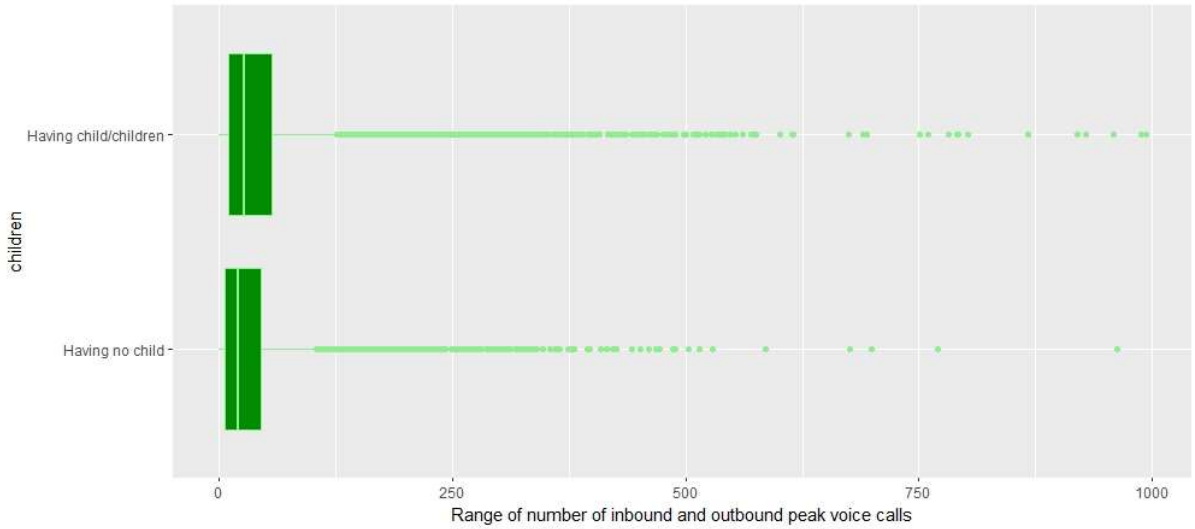
Box plot for Current handset price and children status

Boxplot 2.4



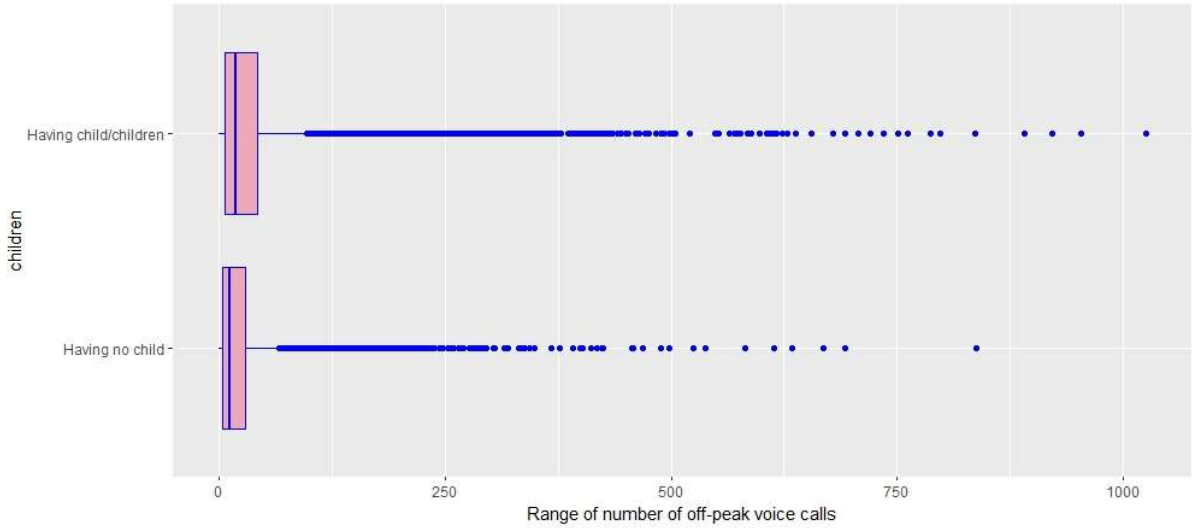
Box plot for Range of number of inbound and outbound peak voice calls and children status

Boxplot 2.11



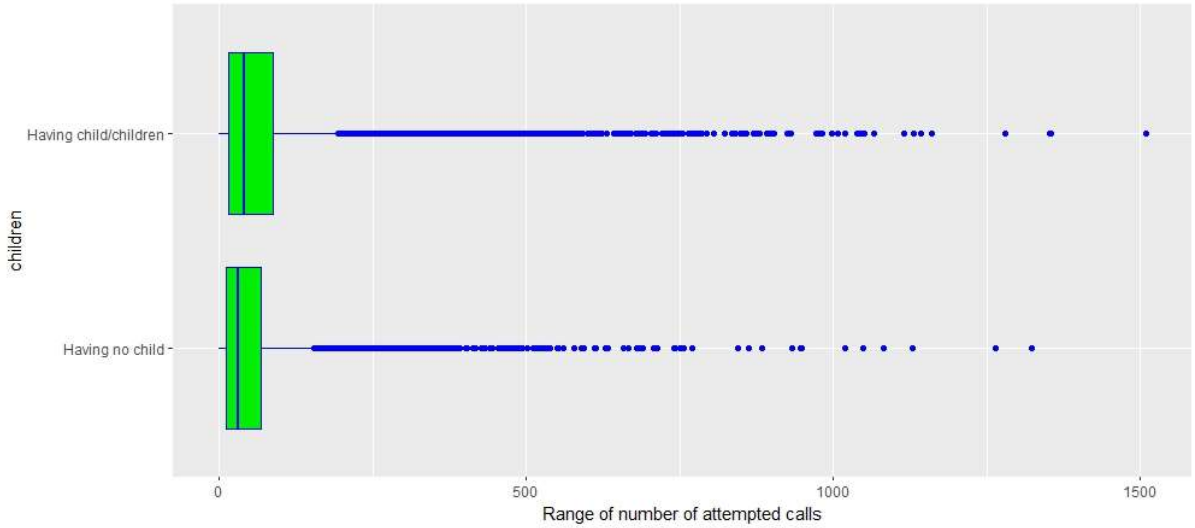
Box plot for Range of number of off-peak voice calls and children status

Boxplot 2.10



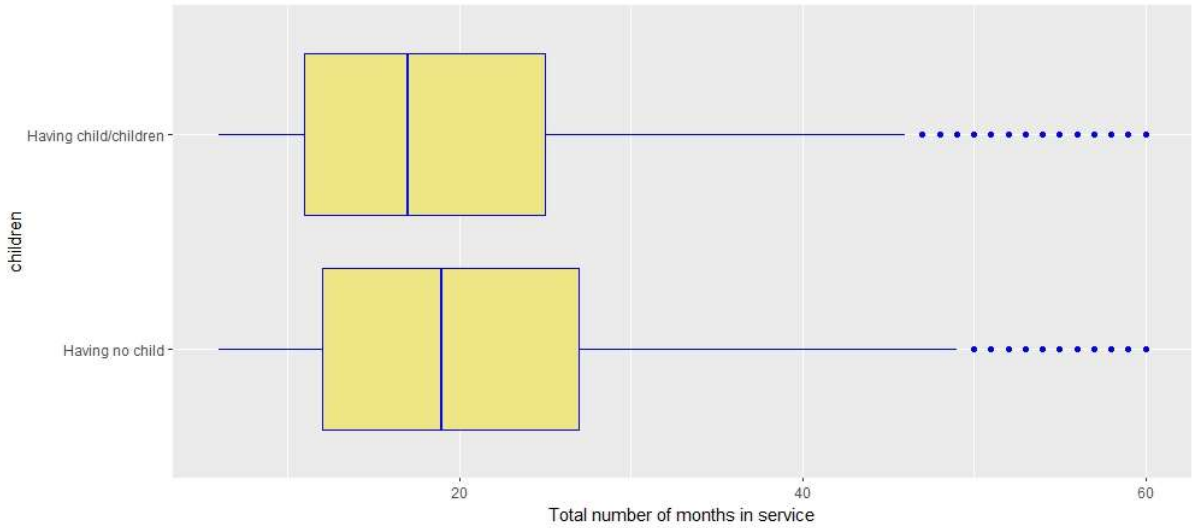
Box plot for Range of number of attempted calls and children status

Boxplot 2.9



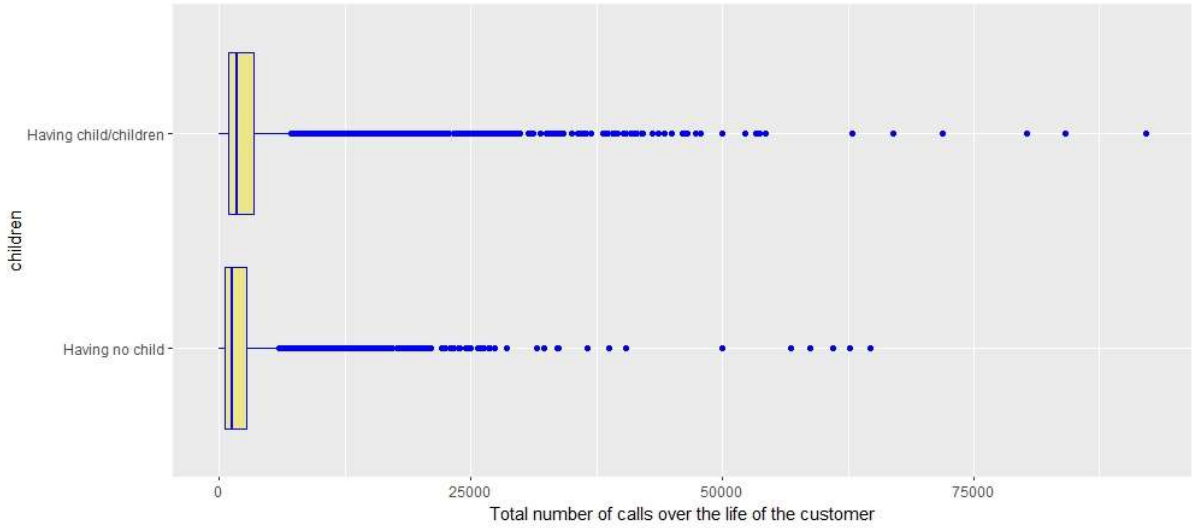
Box plot for Total number of months in service and children status

Boxplot 2.8



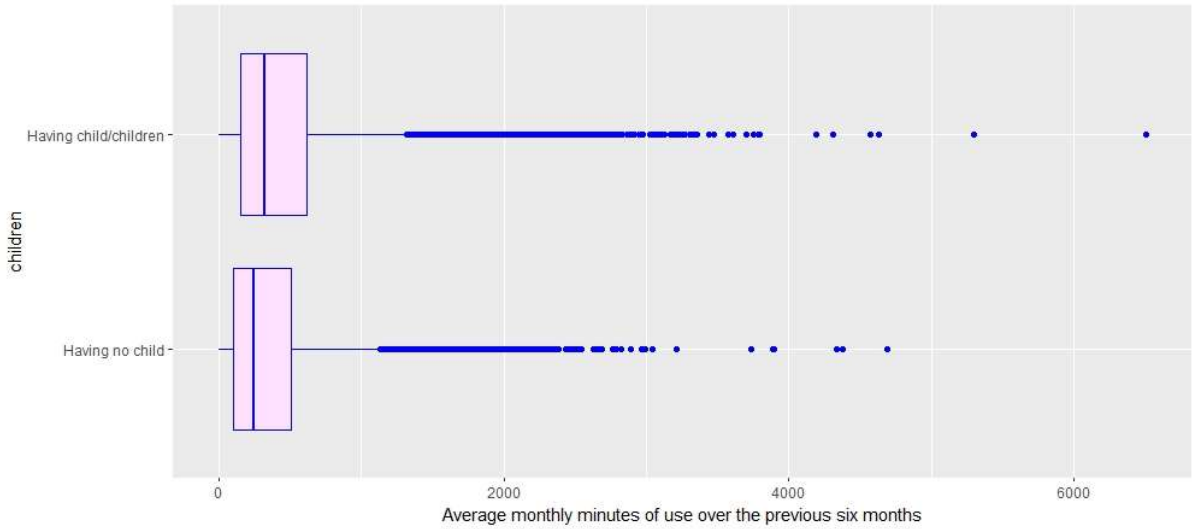
Box plot for Total number of calls over the life of the customer and children status

Boxplot 2.7



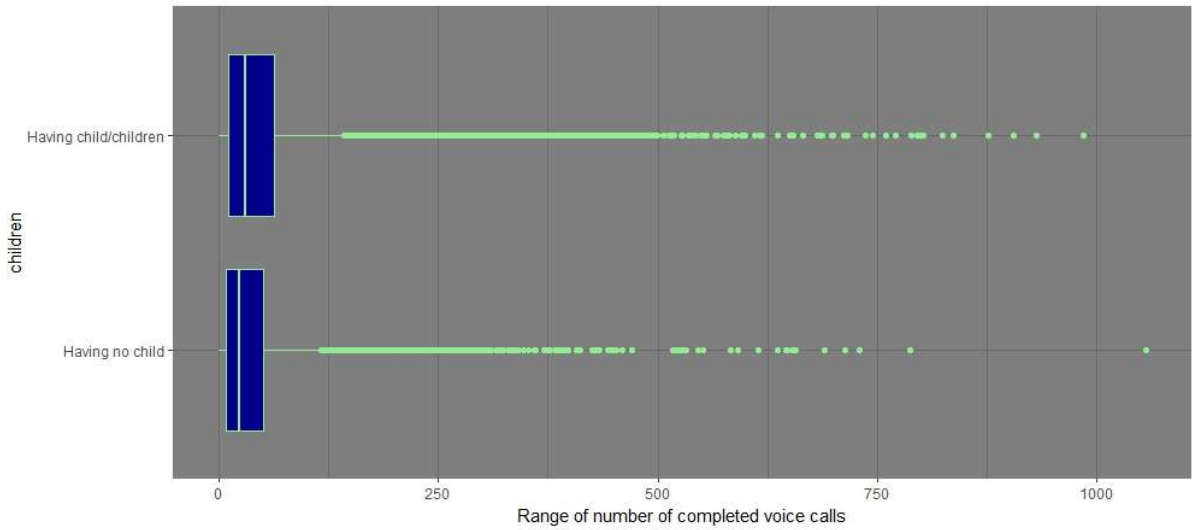
Box plot for Average monthly minutes of use over the previous six months and children status

Boxplot 2.6



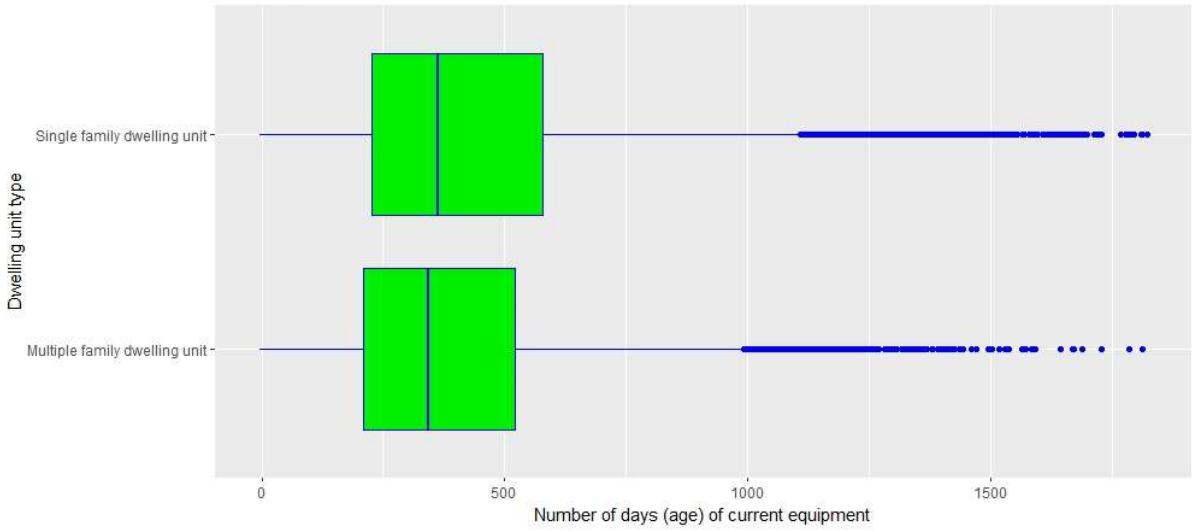
Box plot for Range of number of completed voice calls and children status

Boxplot 2.12



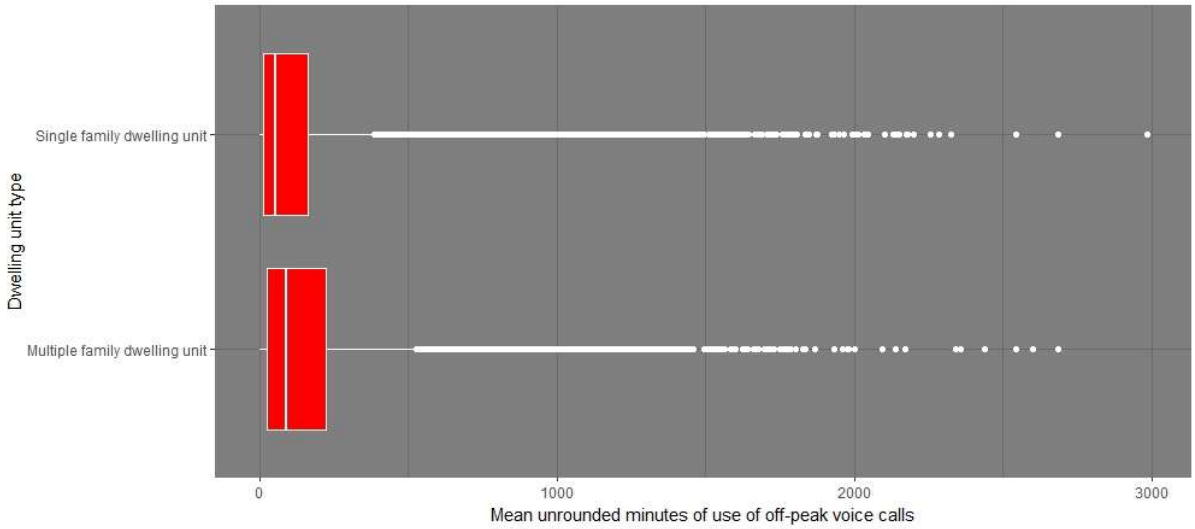
Box plot for Number of days (age) of current equipment and Dwelling unit type

Boxplot 3.1



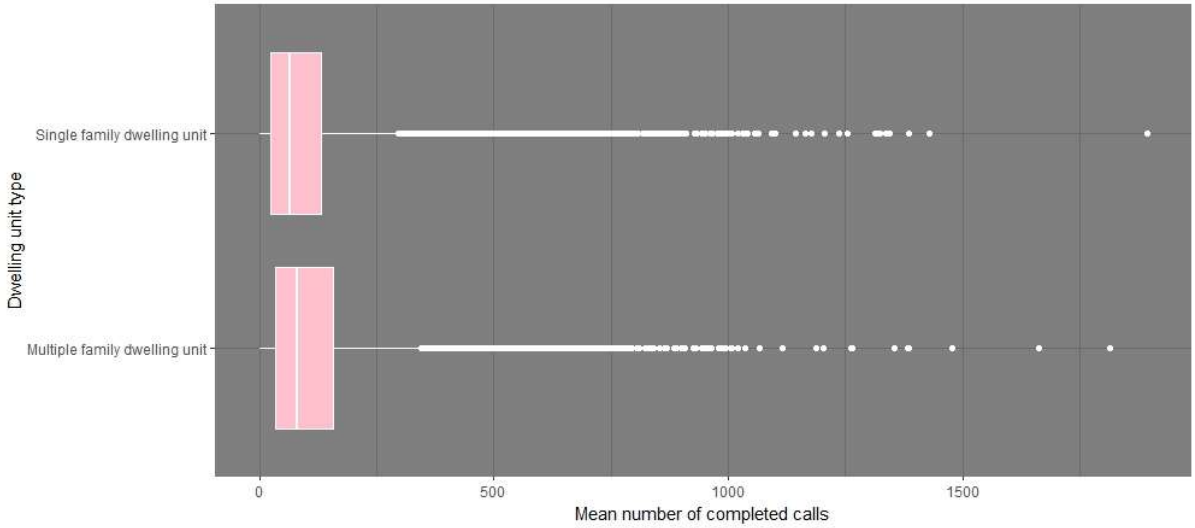
Box plot for Mean unrounded minutes of use of off-peak voice calls and Dwelling unit type

Boxplot 3.11



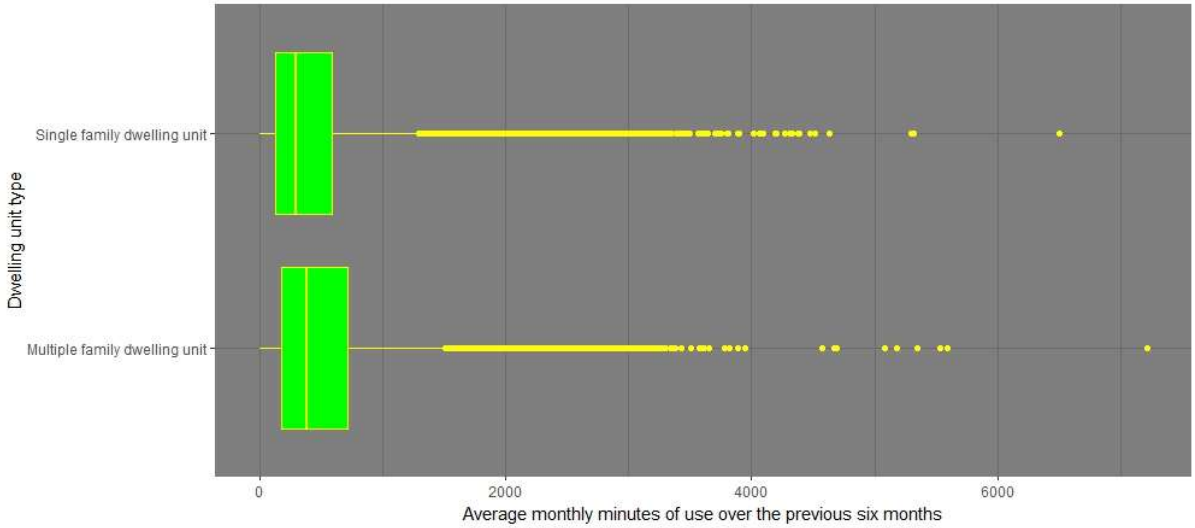
Box plot for Mean number of completed calls and Dwelling unit type

Boxplot 3.10



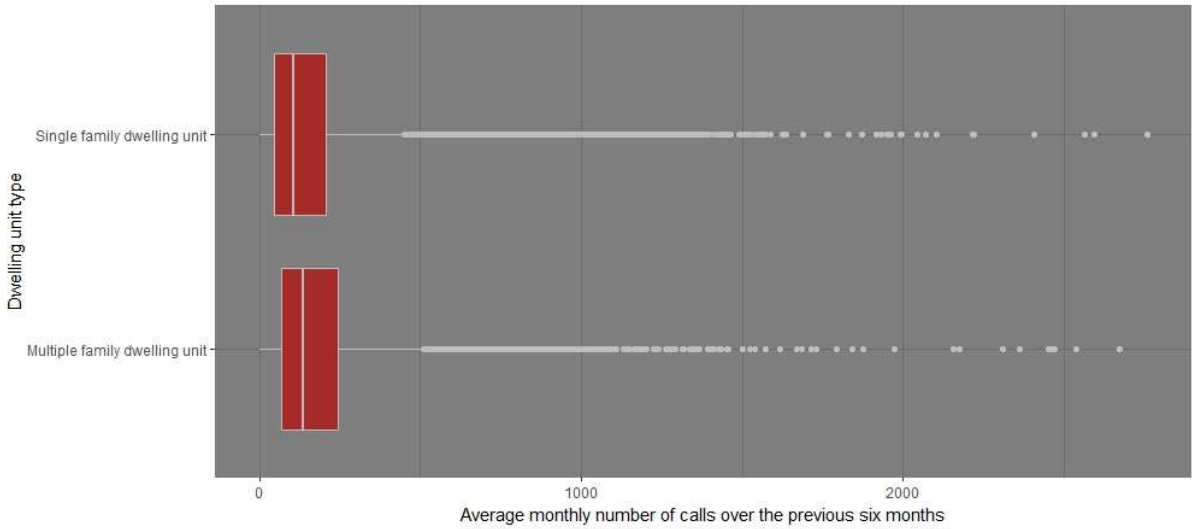
Box plot for Average monthly minutes of use over the previous six months and Dwelling unit type

Boxplot 3.9



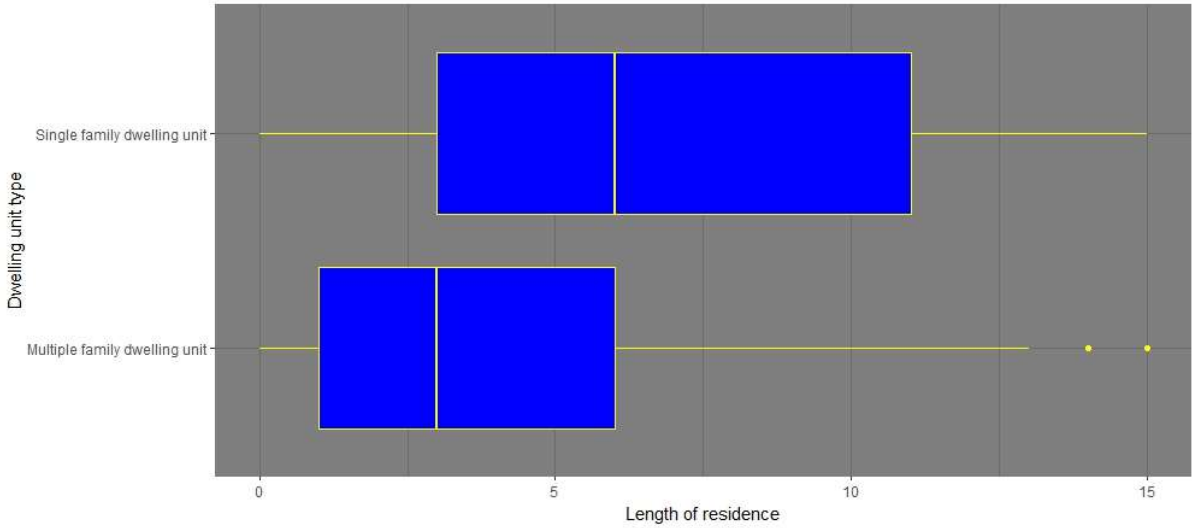
Box plot for Average monthly number of calls over the previous six months and Dwelling unit type

Boxplot 3.8



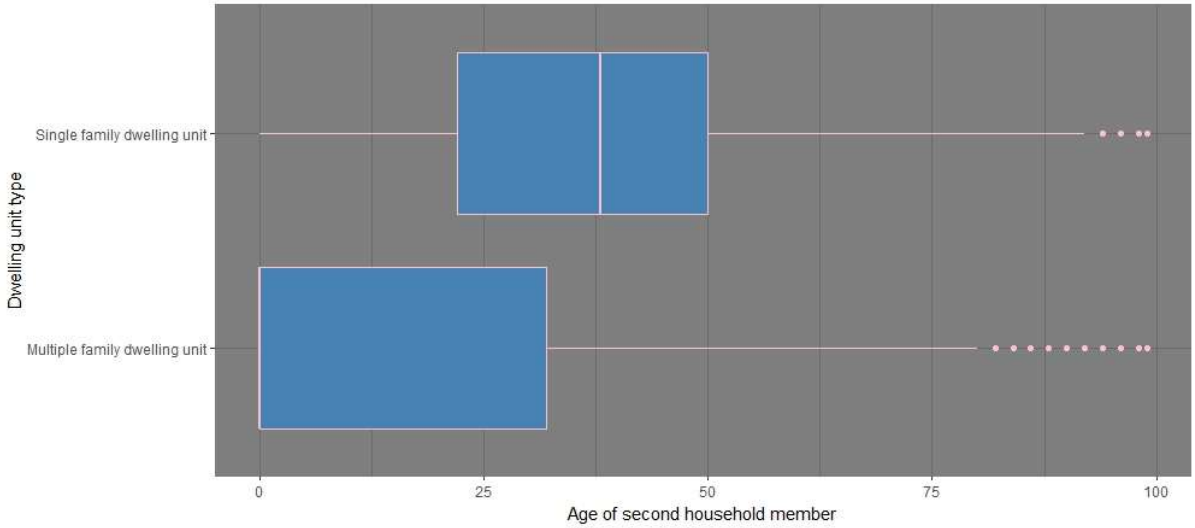
Box plot for Length of residence and Dwelling unit type

Boxplot 3.7



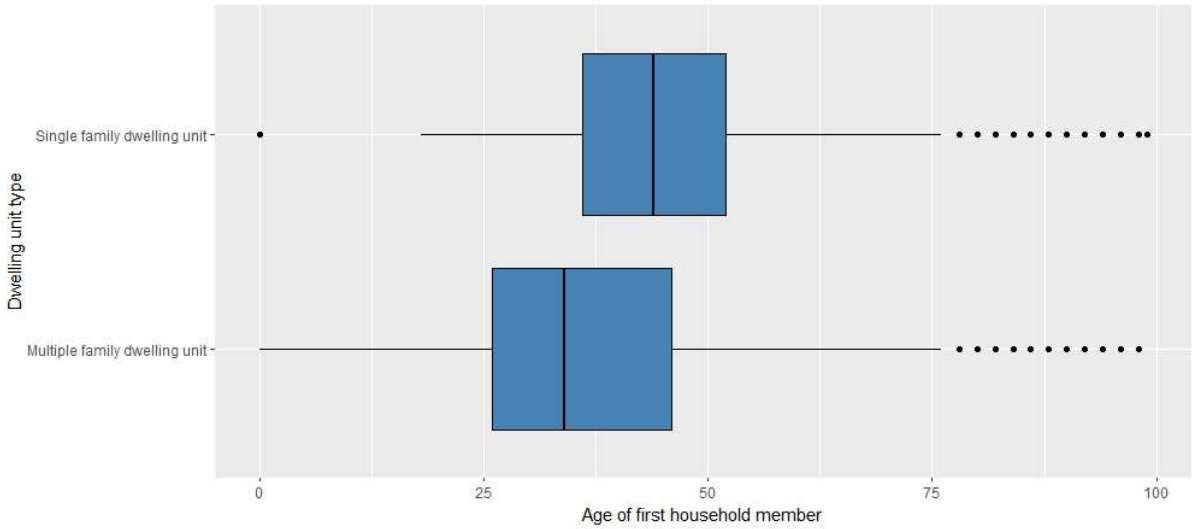
Box plot for Age of second household member and Dwelling unit type

Boxplot 3.6



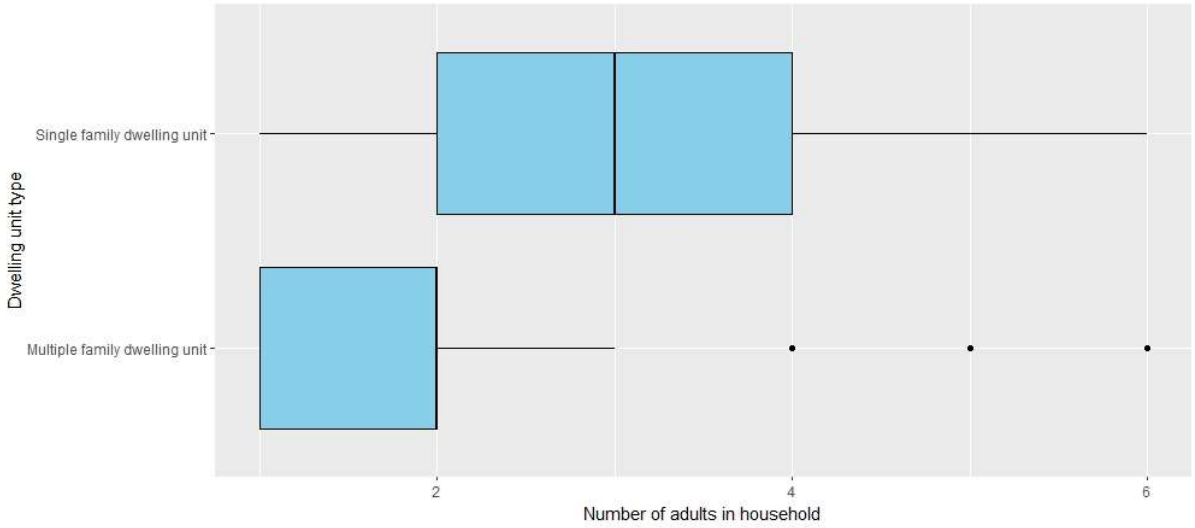
Box plot for Age of first household member and Dwelling unit type

Boxplot 3.5



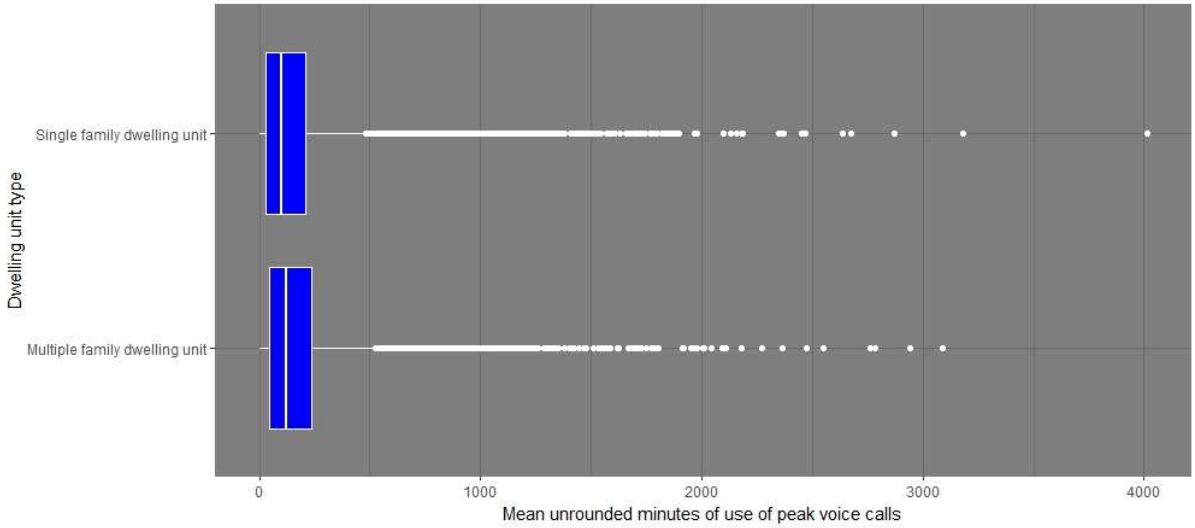
Box plot for Number of adults in household and Dwelling unit type

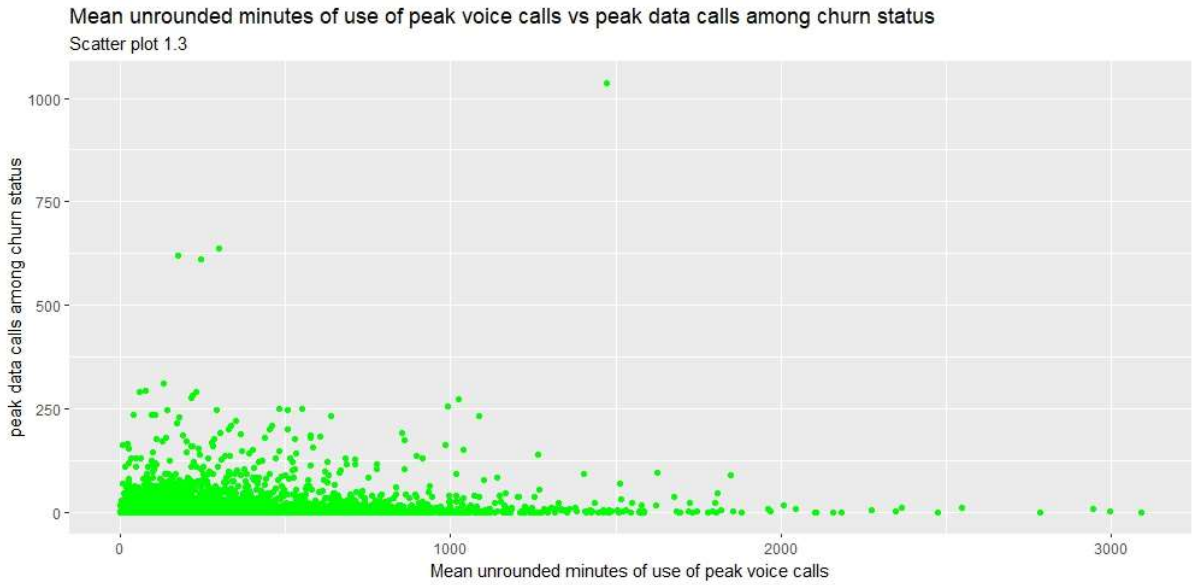
Boxplot 3.4



Box plot for Mean unrounded minutes of use of peak voice calls and Dwelling unit type

Boxplot 3.12




Mean unrounded minutes of use of peak voice calls vs peak data calls among churn status
Scatter plot 1.3
peak data calls among churn status
1000
750
500
250
0
0
1000
2000
3000
Mean unrounded minutes of use of peak voice calls