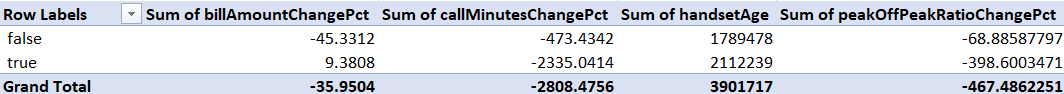
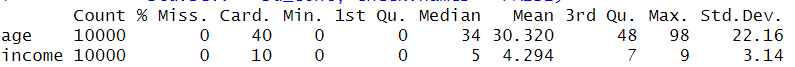
KSE521 Business Intelligence

Homework #4

1. **By one glance using Excel:** By merely using the pivot table to compare each variable with the Churn variable, it could be observed that there is some correlation between Churn variable and Peak Off-peak Ratio Change Percentage, Bill Amount Change Percentage, Handset Age, and Call Minutes Change Percentage respectively. The values of these variables show obvious difference between Churn and No Churn groups.

****

1. **Data Quality Report:** Prior to data preprocessing, a simple examination of the data quality of the variables Age, Income, Occupation and Churn is performed. The data quality report is as below.
   1. Age and Income

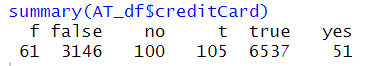


* 1. Occupation and Churn

From the data quality report, we can notice most of the numbers agree with the numbers in Table 9.2 except the two main differences below.

* + 1. The percentage of missing values for Age variable is 0% as compared to 11.47% in Table 9.2. Simple tabulation of count of values reveals that there are 29.42% (not 11.47%) of records with Age = 0 in the dataset. Further analysis might be required to determine whether the 11.47% obtained in Table 9.2 is a random mistake or a value derived based on some specific rules (which we need to refer to Acme Telephonica).
    2. The mode and 2nd mode (including frequency and percentage) of Occupation variable is different. From the data quality report above, a white space and ‘professional’ are found to be the mode and 2nd mode respectively for Occupation variable, while ‘professional’ and ‘crafts’ are determined as mode and 2nd mode respectively in Table 9.2. Further analysis shows that if we deem those white spaces as missing values and remove them, the mode and 2nd mode from our data quality report will match the ones in Table 9.2.

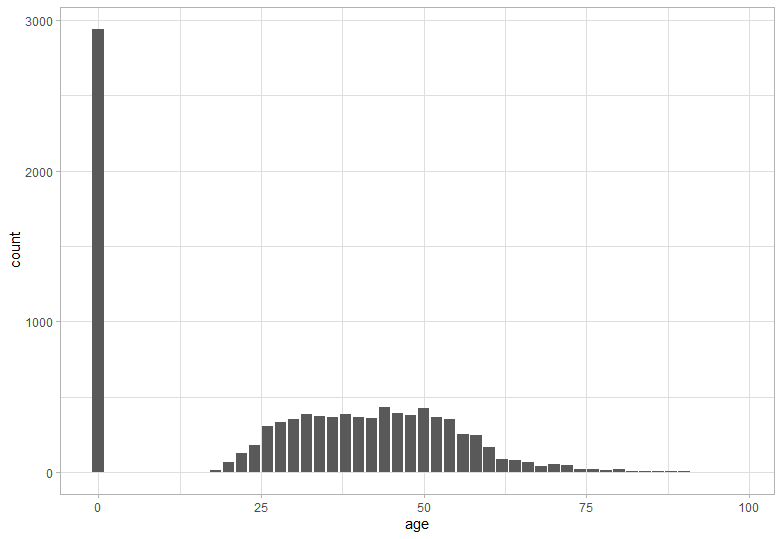
1. **Data Preprocessing:**
   1. Multiple representation for the same label
      1. Credit Card variable contains values ‘true’, ‘false’, ‘no’, ‘t’, ‘f’, and ‘yes’. We can safely assume that ‘true’, ‘t’ and ‘yes’ fall under the same category and can convert all of them to 1 while on the other hand 0 for the rest of the values.



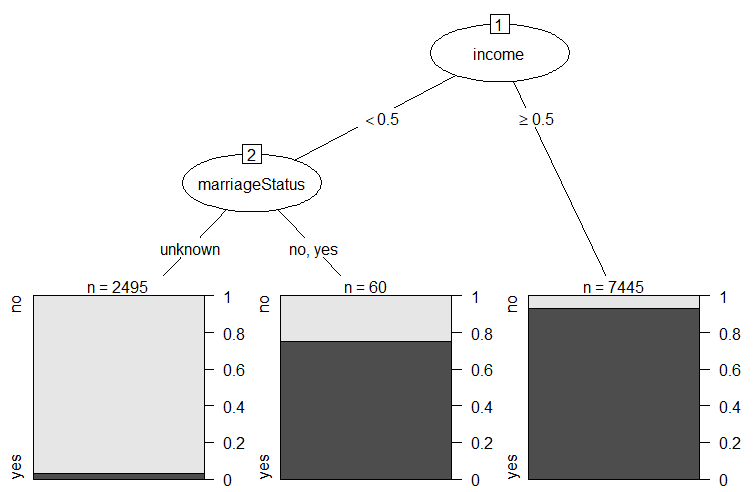
* + 1. Region Type variable contains values ‘r’, ‘rural’, ‘s’, ‘suburban’, ‘t’, ‘town’, ‘unknown’ and missing value (NA). ‘r’ will be converted to ‘rural’, ‘s’ will be converted to ‘suburban’, ‘t’ will be converted to ‘town’, and missing values will be converted ‘unknown’. However, the conversion to one-hot encoding which lead to more new variables will be on hold at this step as we will analyze further on the portion of missing values in the next step.

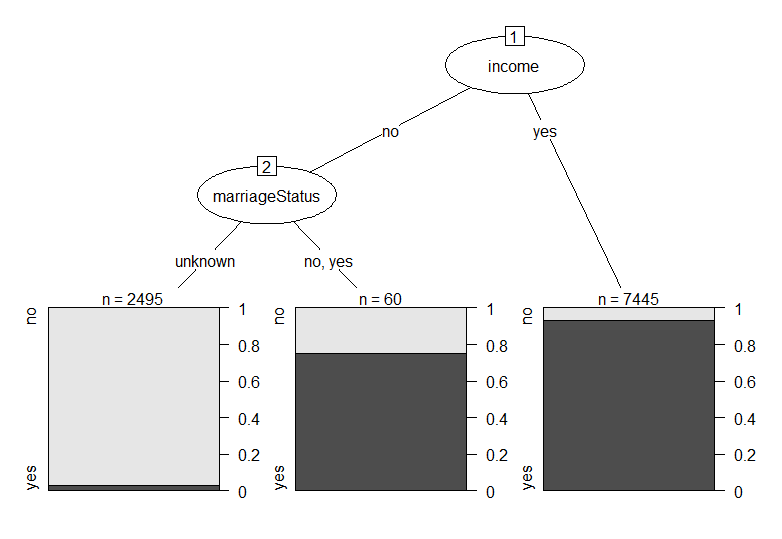


* 1. White spaces as missing values
     1. After first basic screening via Excel, Occupation and Region Type variables are found to contain cells which filled with a white space only, and all the non-numeric variables contain a leading white space.
     2. To handle this problem, during the import of dataset, we need to first strip the leading and trailing white spaces. The cells which contain only a white space will become blank cells.
     3. Subsequently, for R to be able to treat the blank cells as missing values, we need to replace the blank cells by ‘NA’ during import.
     4. Occupation variable contains 74% missing values. We will remove this variable as it is unlikely to provide useful insight for further analysis.
     5. Notice that there are 48.59% of the records contain ‘unknown’ or missing values (NA) for Region Type variables. Thus, we will also remove this variable as it is also unlikely to provide useful insight for further analysis.
  2. 0 as missing values
     1. Age variable contains 2942 records with value 0. Logically speaking, it is not normal and could be treated as missing values.



* + 1. We need to find out the type of missing values of Age. By using simple decision tree, it is discovered that among 2495 customers with Income Level <0.5 and Marriage Status = unknown, 2414 (96.7%) of them are also labelled as Age = 0. If we label the records with Income Level = 0 as ‘no’ and ‘yes’ for other incomes, we would obtain the same decision tree.





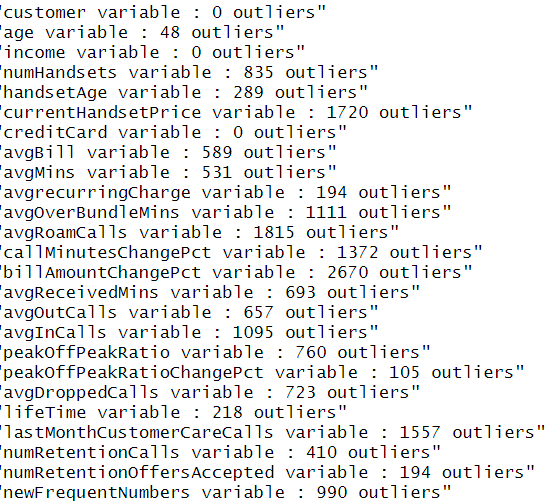
* + 1. From b(ii), we could also derive that **Income Level = 0** is highly possible to be **missing values** too. Note that Income Level variable contains 2555 records with value 0.
    2. This result suggests that customers who do not provide information of Income Level and Marriage Status are highly unlikely to disclose their Age information too. If we replace the values of each of these variables by 0 for information unavailable (missing value/unknown/0), and 1 for information available, we can see that the correlation of co-existence of Income, Marriage Status and Age is quite high.



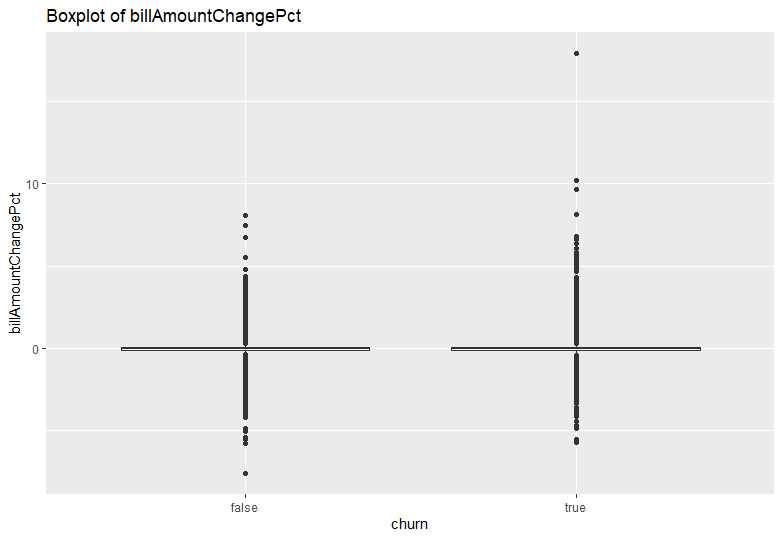
* + 1. This also suggests that the variables are not missing completely at random, in which we cannot simply impute by mean, median or mode. However, at the same time, we could not provide an estimation based on the other attributes or just remove those records because they make up more than 50% of the dataset. Hence the best we could do is to just leave them as it is (being assigned explicit value ‘0’ or ‘unknown’) at this stage, then in the later stage, we will look at the correlation for these variables between the target variable ‘Churn’ and decide whether to retain or remove them.
  1. Non-negative variable contains negative values
     1. Handset Age variable which is normally non-negative (larger or equal to 0) contains 12 records with negative values. This could possibly be a data entry error or error in deriving this value based on other values.
     2. Since only 0.12% records contain negative Handset Age value, we can safely remove them from the dataset.
  2. Transforming binary/ordinal variable into numerical variable
     1. For Marriage Status, Children, Smart Phone, Home Owner, and Credit Card variables, ‘true’/’yes’ values will be transformed into 1 whereas ‘false’/’no’ values will be transformed into 0.
     2. For Credit Rating, the following transformation will be used.

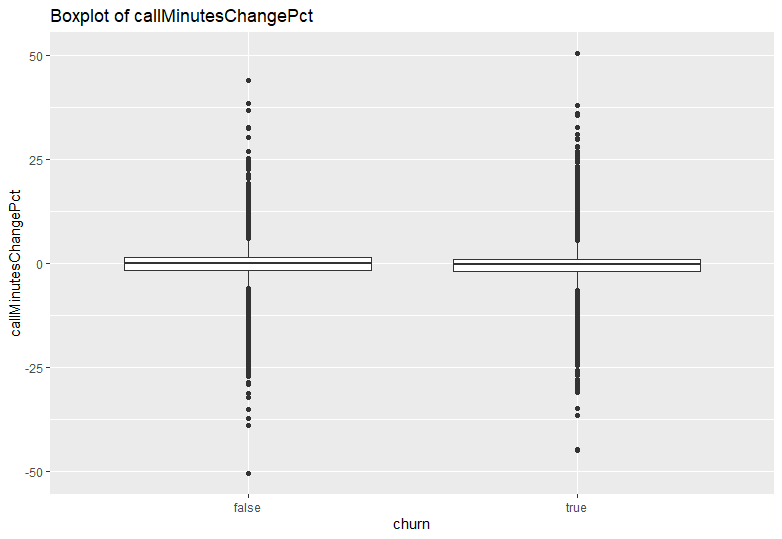
|  |  |
| --- | --- |
| **Original Value** | **New Value** |
| A | 1 |
| B | 2 |
| C | 3 |
| D | 4 |
| E | 5 |
| F | 6 |
| G | 7 |

* 1. Other outliers
     1. **Numerical variables:** By plotting the boxplot (disregard the missing values) for each numerical variable vs Churn variable, we can observe a lot of outliers for almost every variable. On the other hand, we can also observe that the distribution of each variable is quite similar for each Churn and Not Churn group. (Please refer to the Appendix for the boxplots)

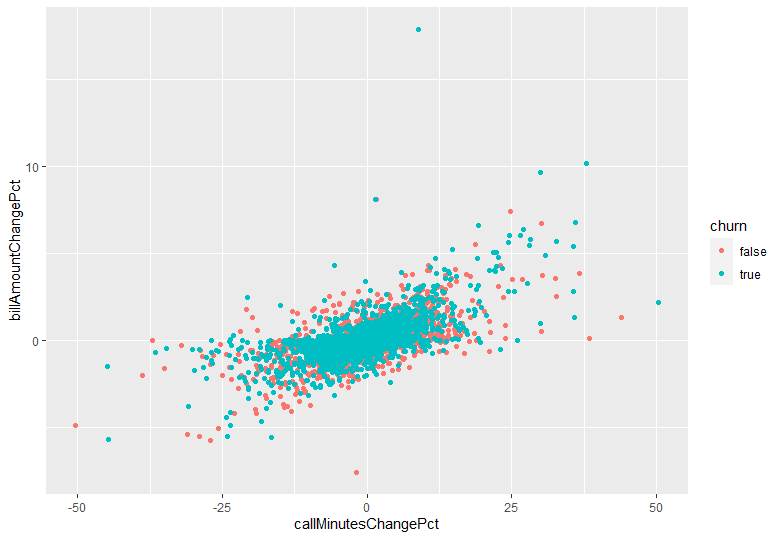


* + 1. Whether to retain or remove the outliers, we need to check each of them carefully. For example, Bill Amount Change Percentage with the most outliers, if we look at the boxplot grouped by Churn variable, it seems valid. From business perspective, it is possible that some customers suddenly make more or less calls on that month and hence result in the huge spread of change percentage of bill amount.



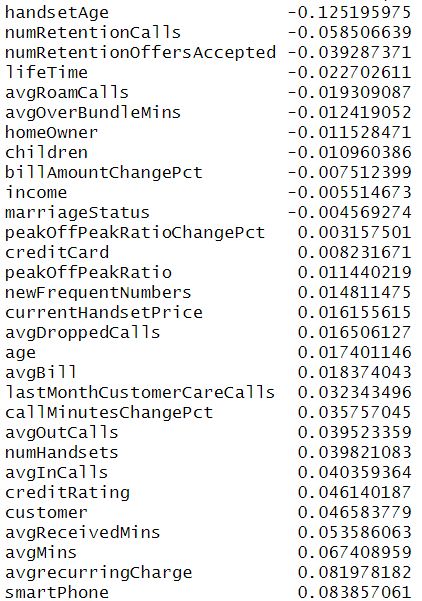


* + 1. If we compare the boxplots of Bill Amount Change Percentage and Call Minutes Change percentage, we can observe a somehow similar shape of distribution. Further checking the correlation between these two variables are 0.6185729.



* + 1. Based on correlation among those numerical variables, we can see some grouping patterns in which some variables are related to each other. For example, (1) Number of Current Handsets, Current Handset Price, and Current Handset Age, (2) Average Call Minutes, Average Out-of-bundle Minutes, Average Monthly Bill, and Average Monthly Recurring Charge, (3) Number of Retention Calls, and Number of Accepted Retention Offers.
    2. These numbers all are seems valid from business perspective at this stage, and none of them will be removed.
  1. We already handled all the categorical variables before this stage. Please refer to Appendix for barplots.

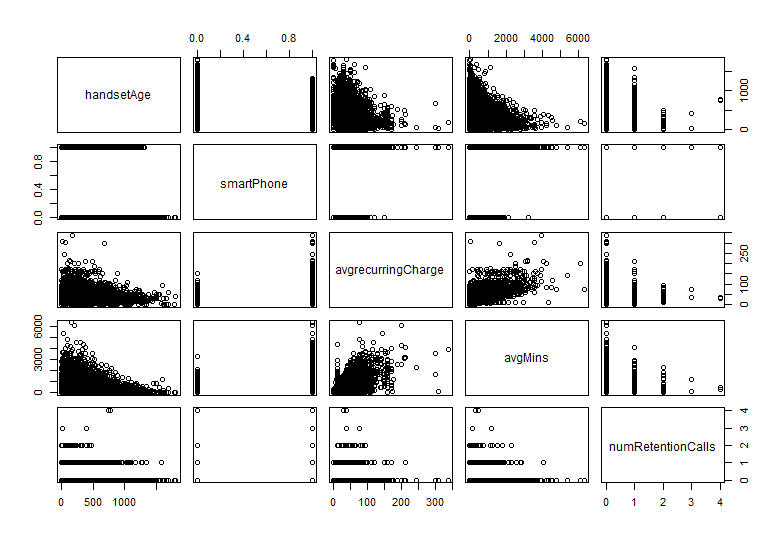
1. By using point-biserial correlation measure, below are the results for correlation between each variable with Churn variable (our target variable).



* 1. By sorting their absolute values, 5 variables with the highest correlations are as follow.

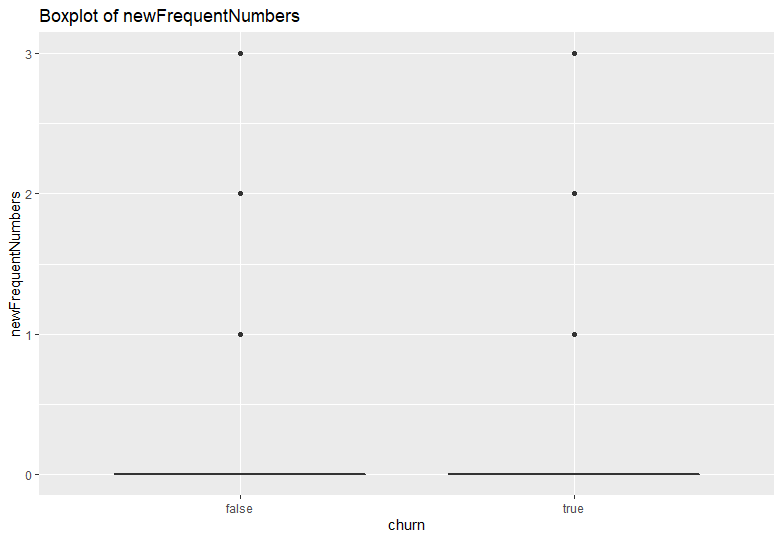
|  |  |  |
| --- | --- | --- |
| Variable | Correlation (Absolute Value) | Correlation (Actual Value) |
| Number of Retention Calls | 0.0585 | -0.0585 |
| Average Call Minutes | 0.0674 | 0.0674 |
| Average Recurring Charge | 0.0820 | 0.0820 |
| Smart Phone | 0.0839 | 0.0839 |
| Handset Age | 0.1252 | -0.1252 |

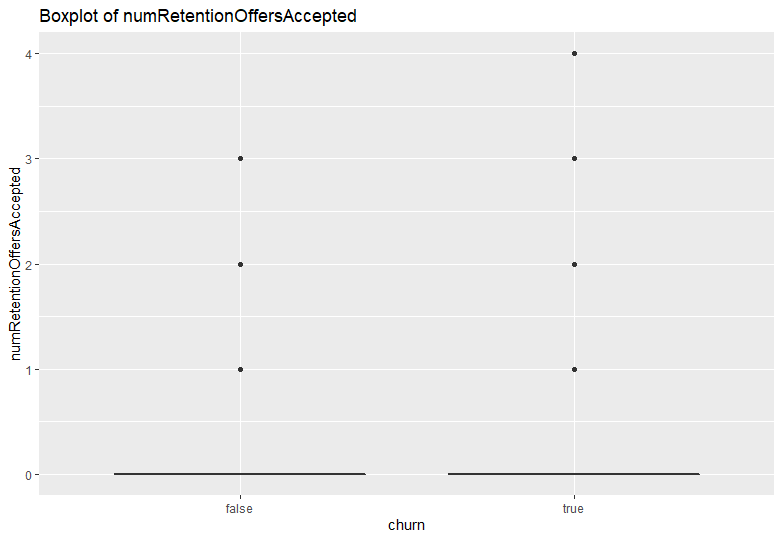
* 1. Below is the scatterplot between these 5 variables. There is no obvious relationship between them, but there is a mild positive correlation between Average Recurring Charge and Average Call Minutes.

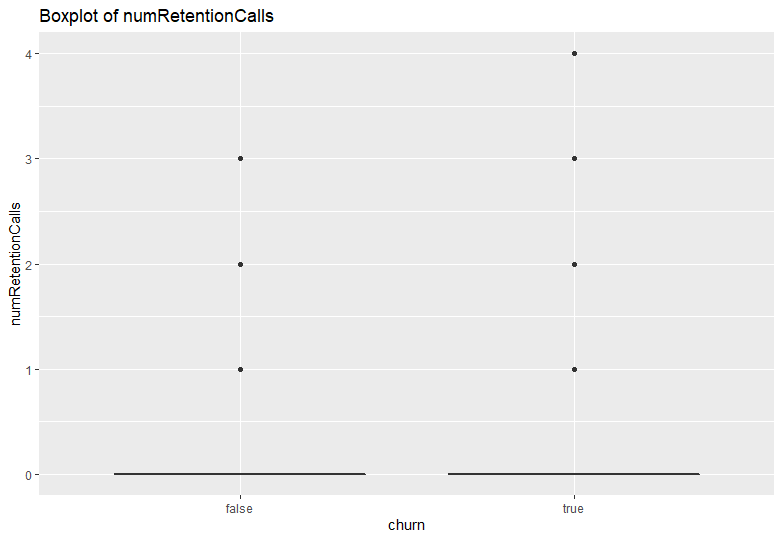


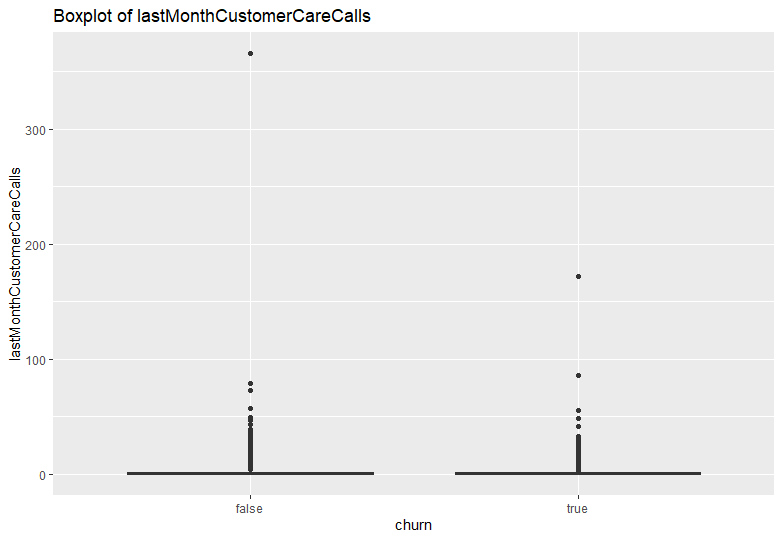
Appendix

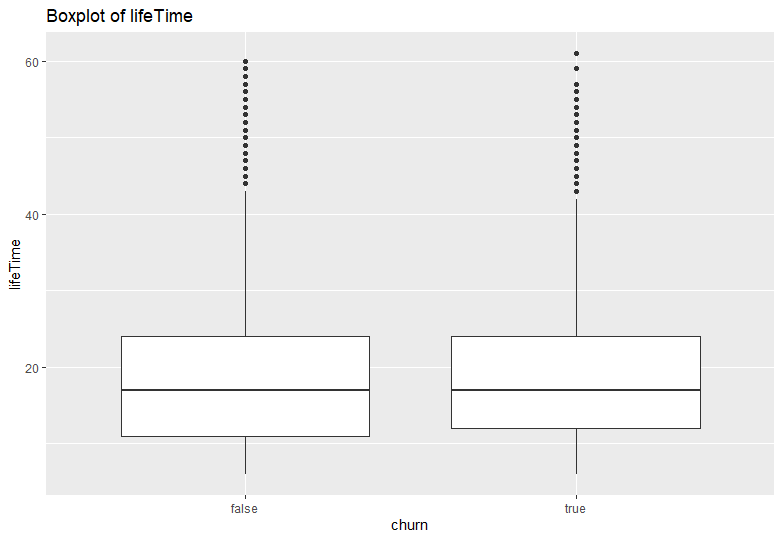
1. Boxplots for continuous numerical variables

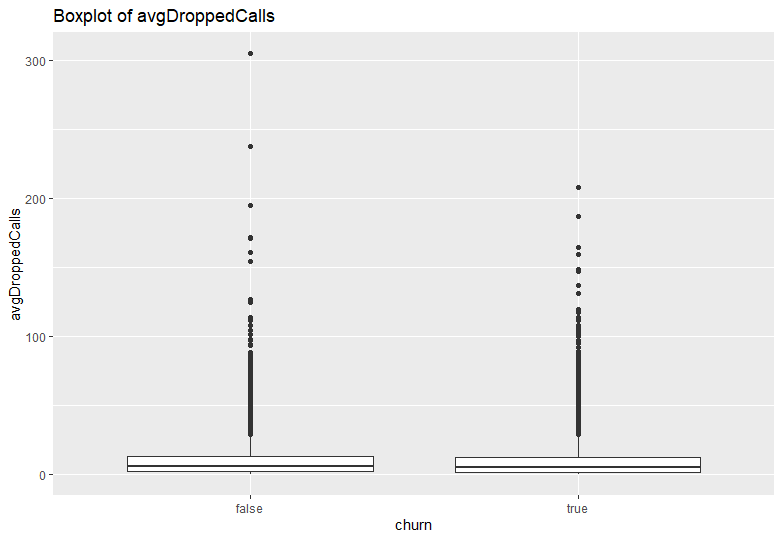


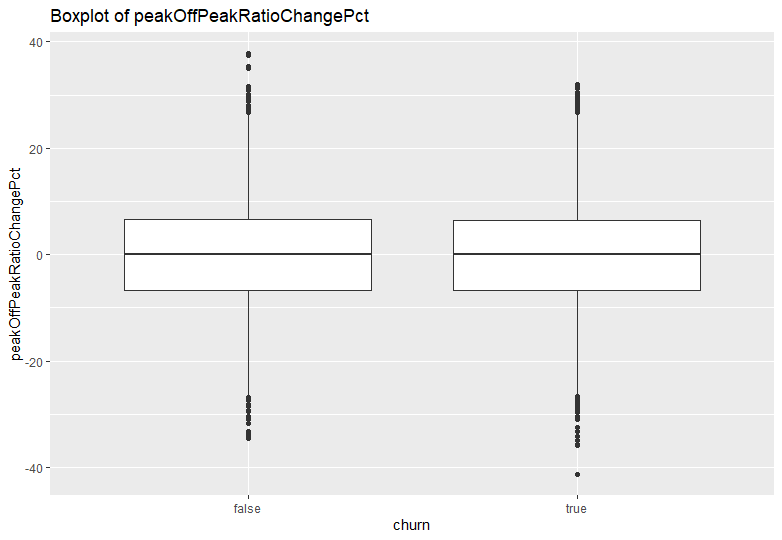


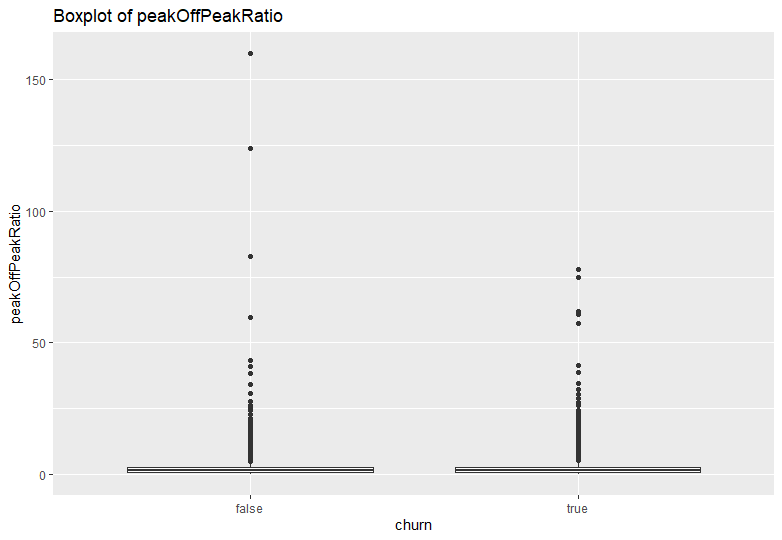


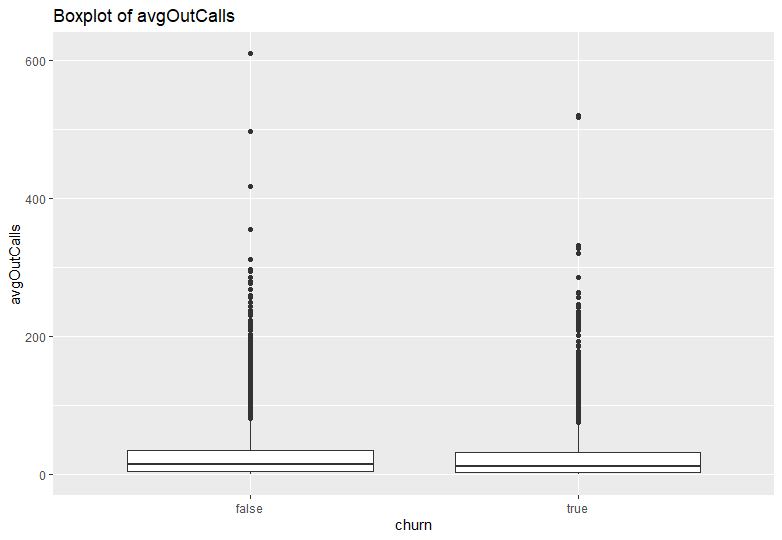


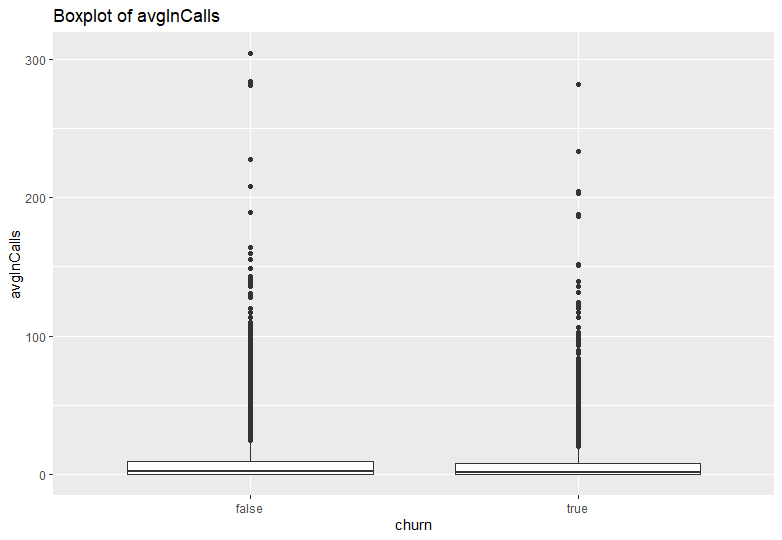


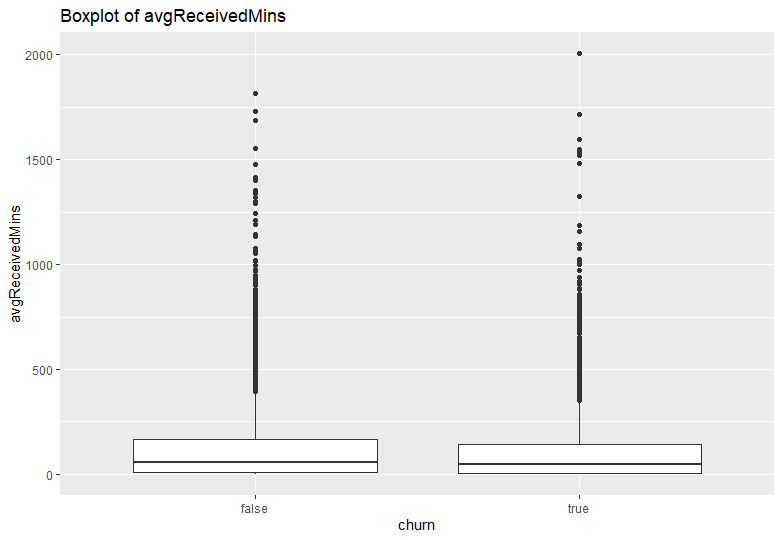


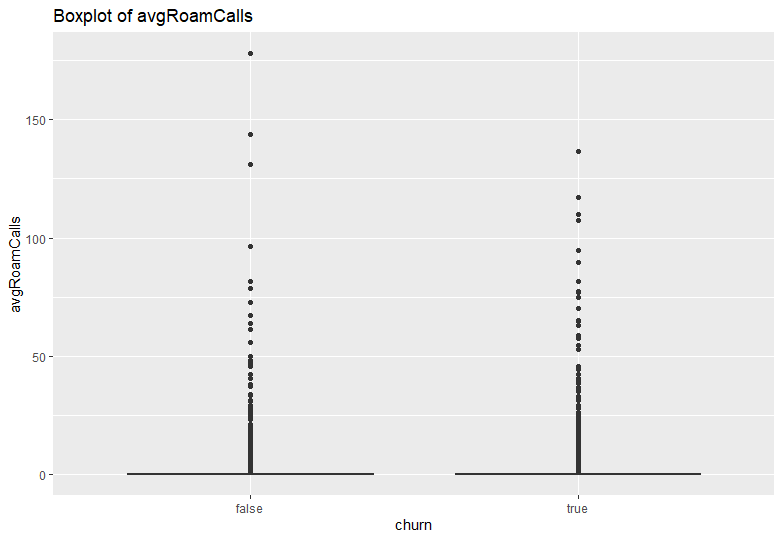


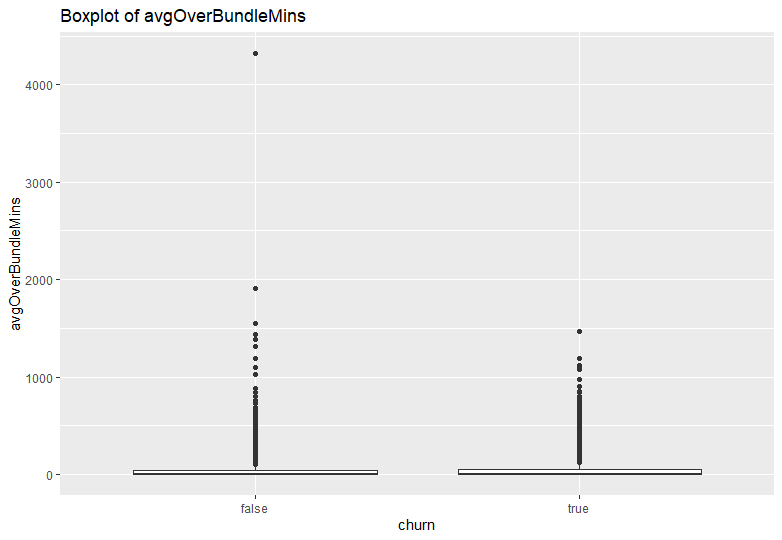


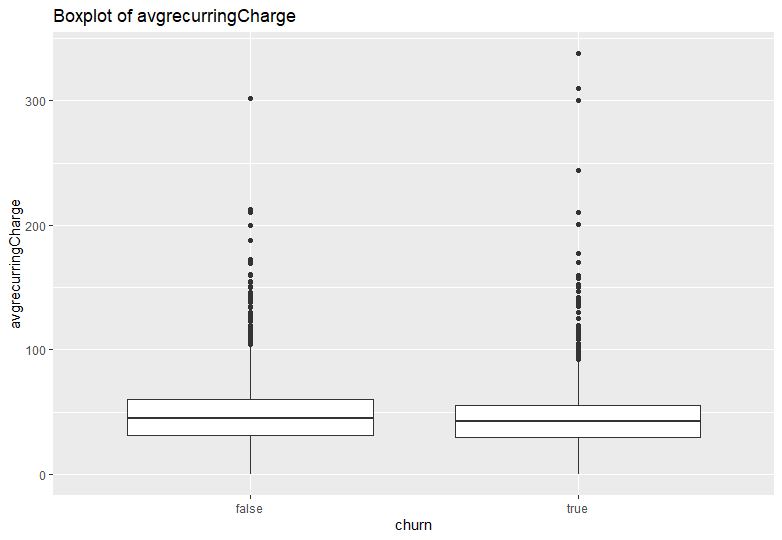


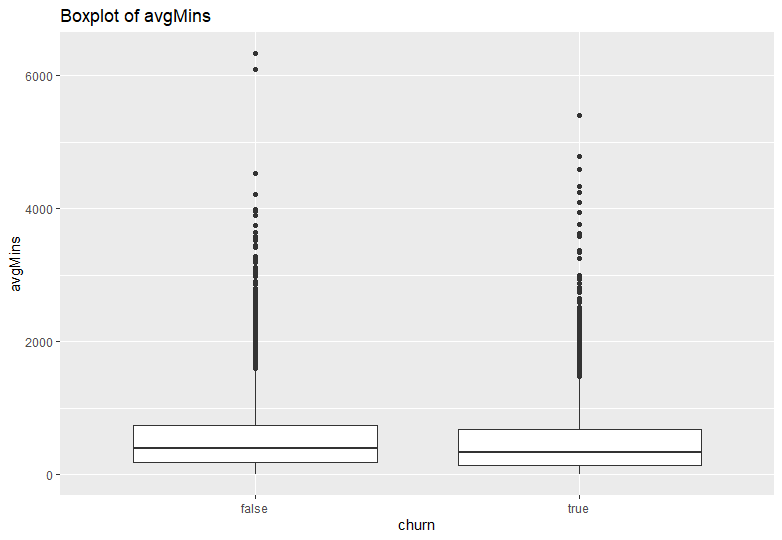


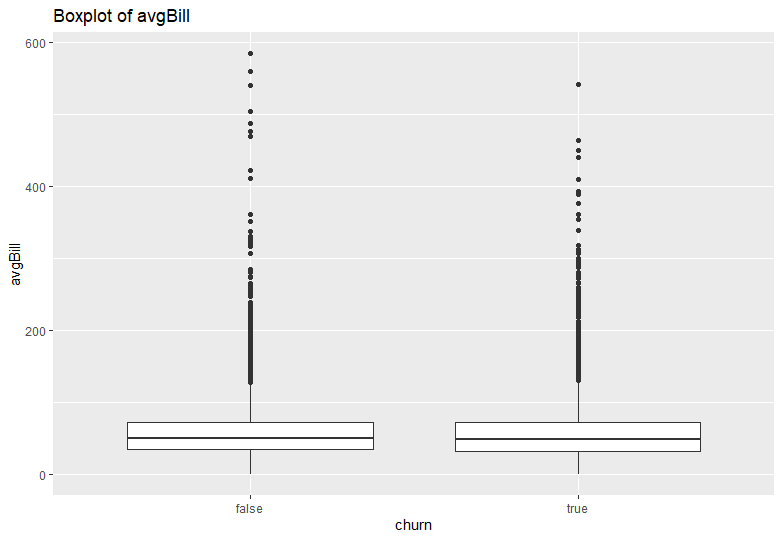


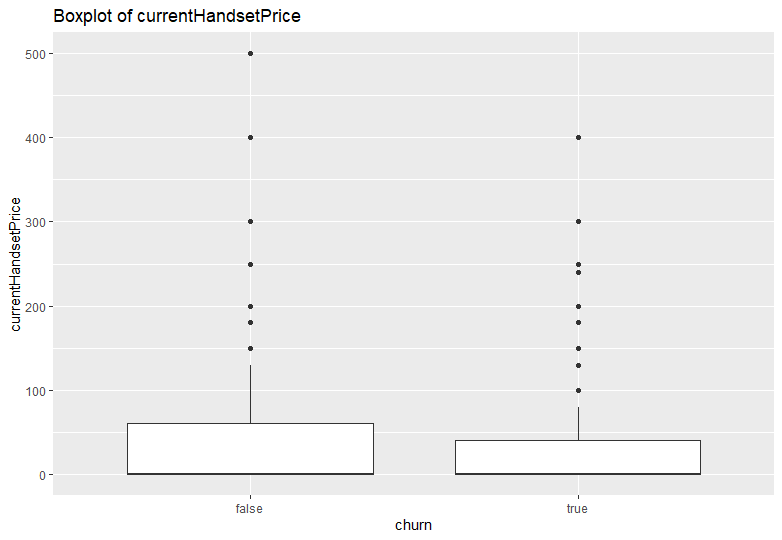


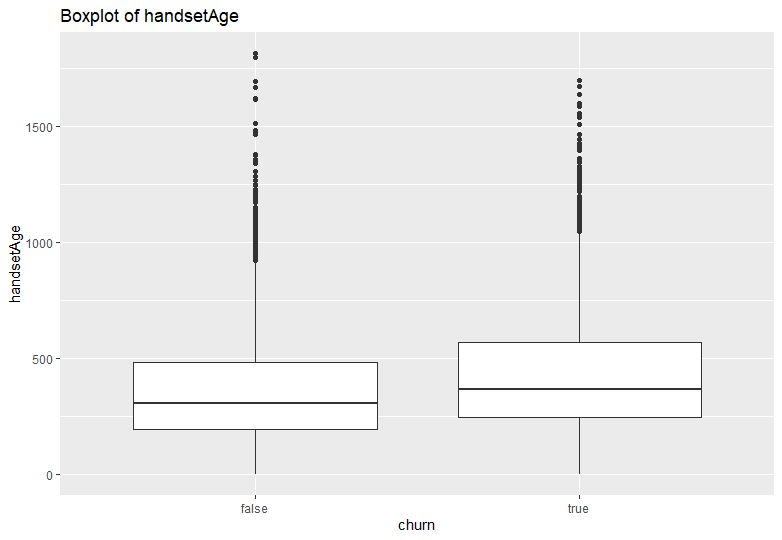


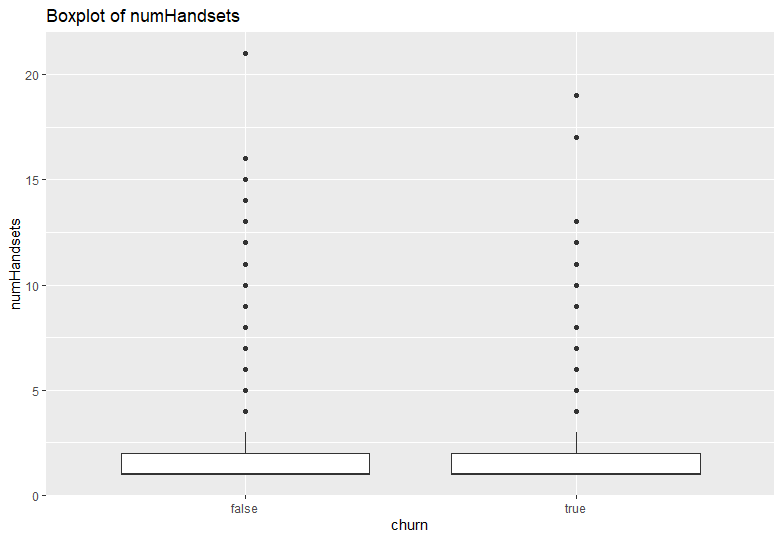


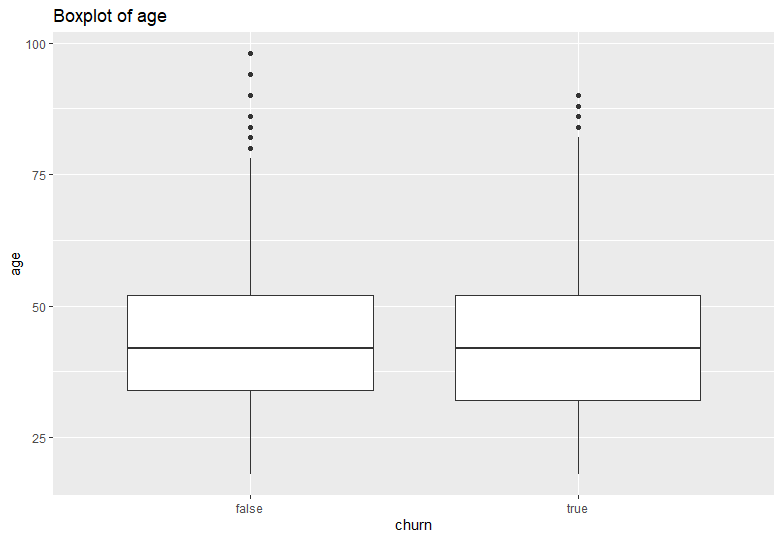












1. **Bar plots for categorical / ordinal variable**

