# Predicting the rate of customer Churn in a telecom company

# Telecom customers Data Set

## Data set description

| | |
|---|---|
| Total number of Customers | 7043 |
| Total number of Variables | 21 |
| Total number of categorical variables | 18 |
| Total number of numerical variables | 3 |
| Total number of NULL values | 11 |
| Output Variable | “Churn” |

# Data Wrangling

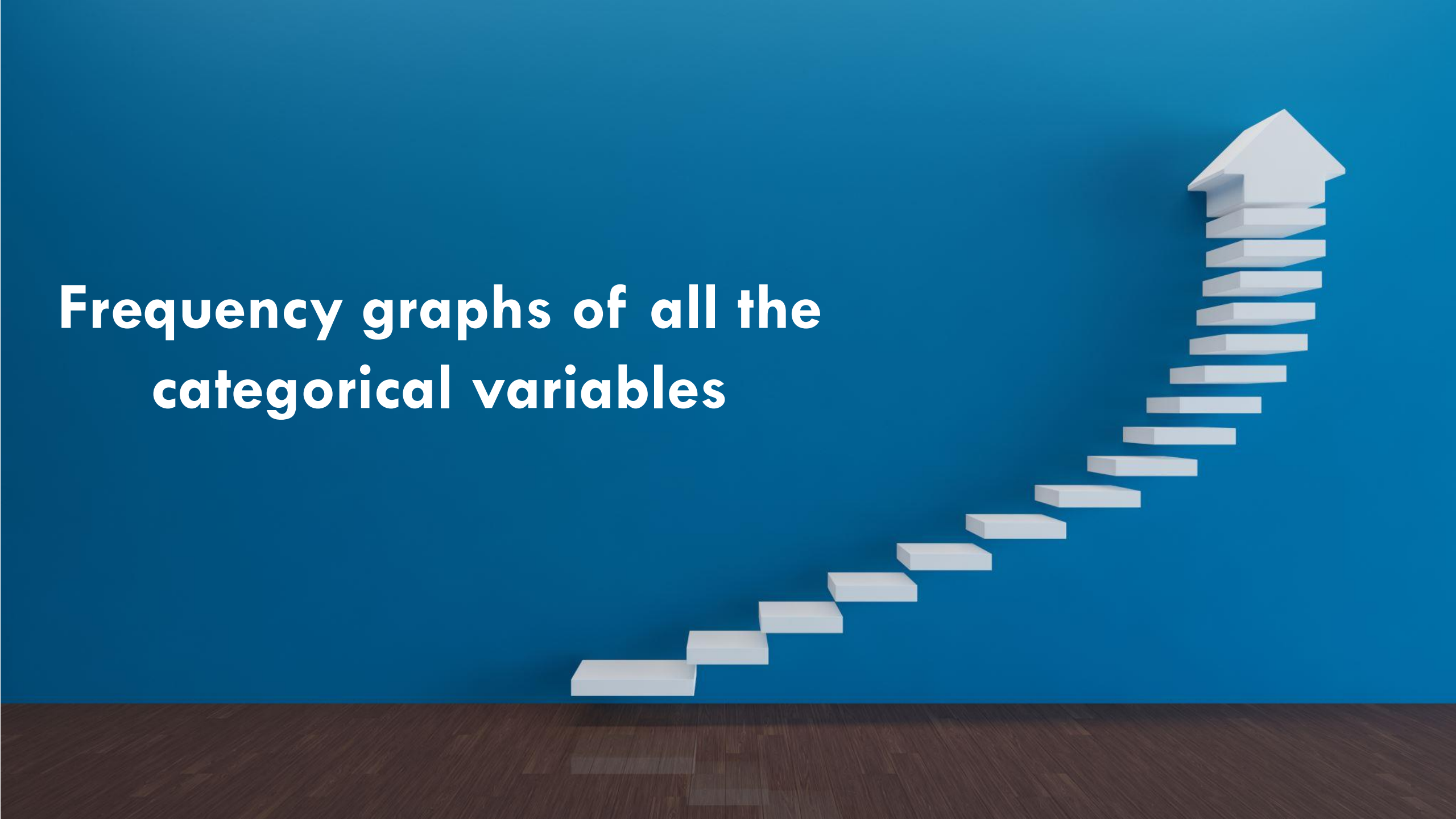
- OMITTED the missing/NULL values of the data frame, as it might interfere in the analysis
- DROPPED the Customer ID column, as it won't be useful in the analysis
- First, TRANSFOMED the data type of Senior citizen column to Factor, then changed the values from "1"& "0" to "YES" & "NO"
- In order to clean the data, changed the value "No internet service" to "NO" in all the columns.
- Other than this data was in tidy format.

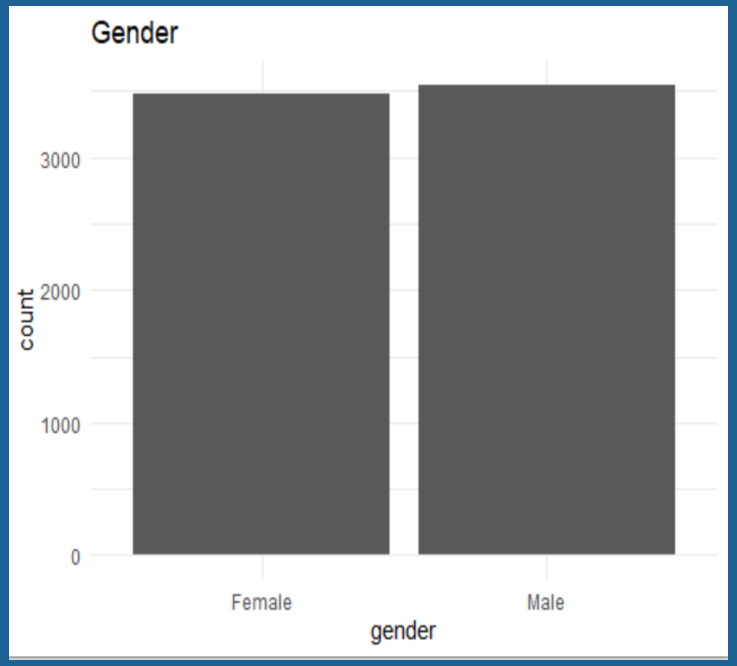
# Statistical information from summary of data set

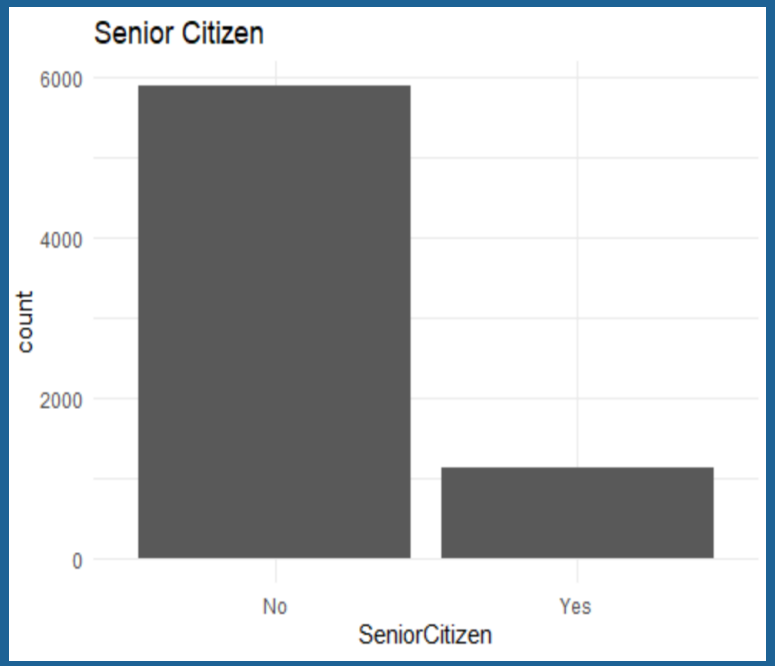
| | |
|---|---|
| Minimum Tenure | 1 Month |
| Maximum Tenure | 72 Months (6 years) |
| Avg. Tenure | 32.42 (approx 2 years 7 months) |
| Avg. Monthly Charges | 64.80 |
| Avg. Total Charges | 2283.3 |
| Max. & Min. Total charge | 18.8 & 8684.8 |
| Number of churned Customers | 1869 |
| Number of Customers who did not Churn | 5163 |

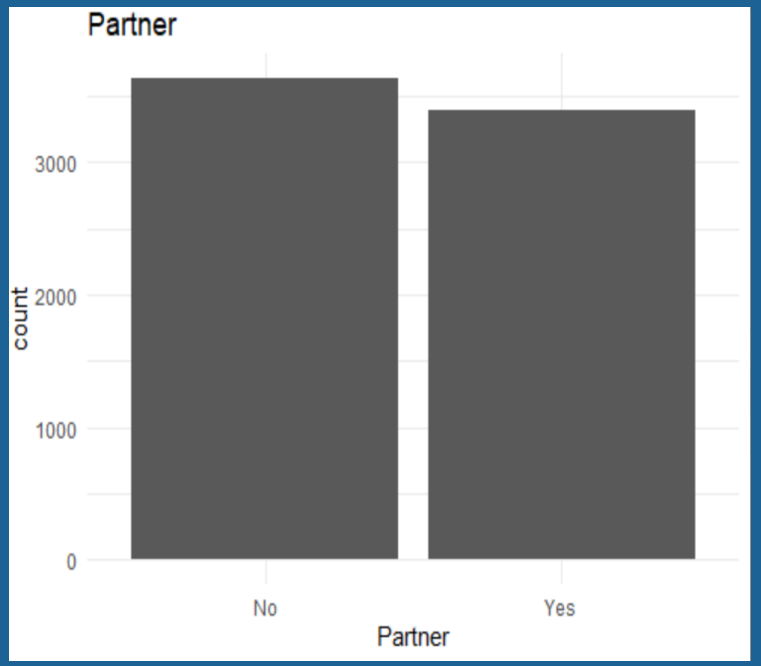
- **The data set is not balanced as the number of customers who churned are very less.**

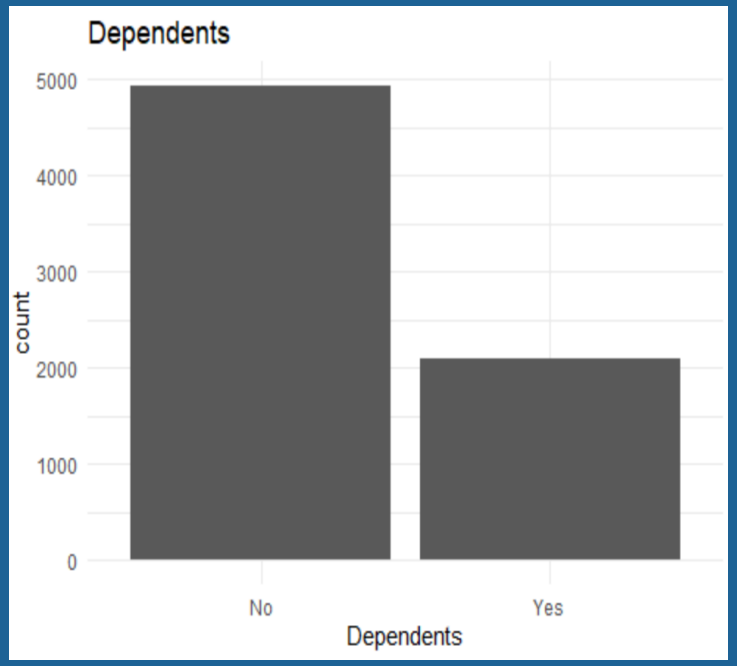
# Frequency graphs of all the categorical variables

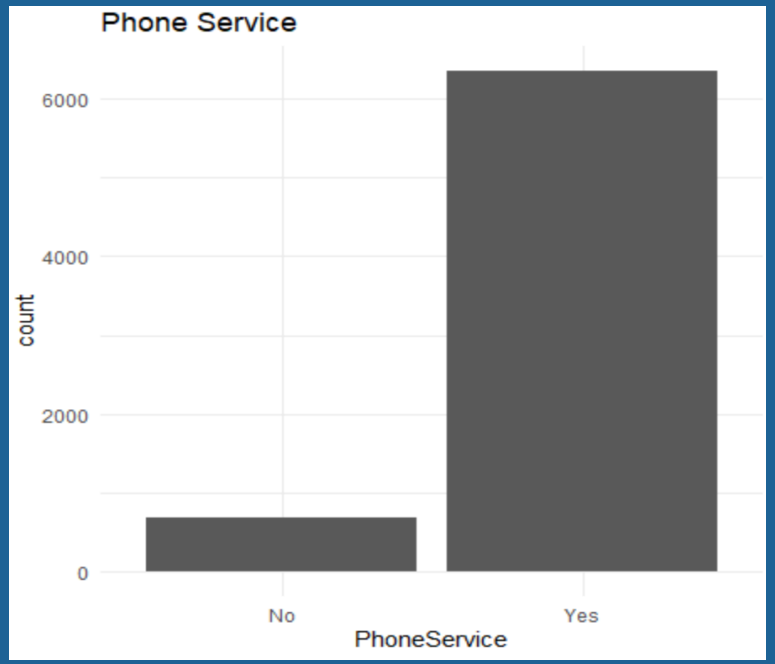


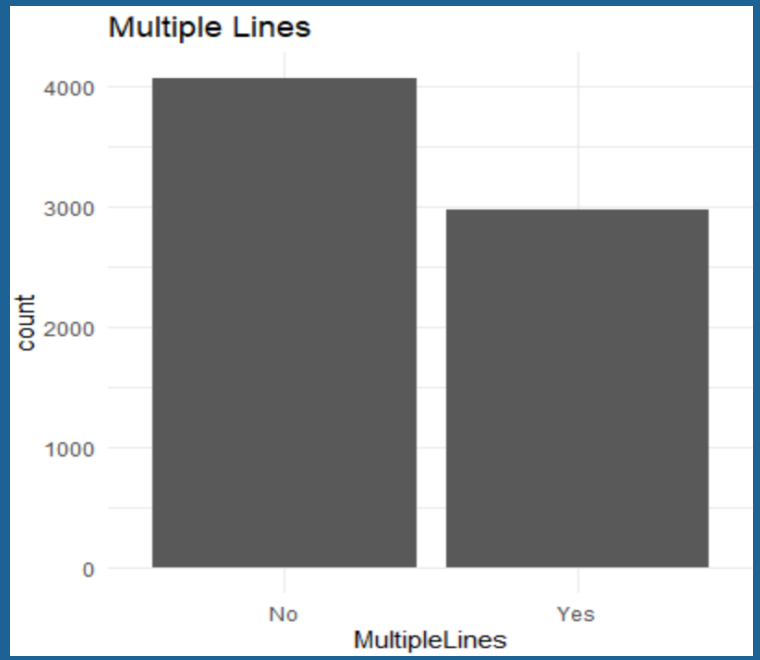

Gender
count
3000
2000
1000
0
Female
Male
gender

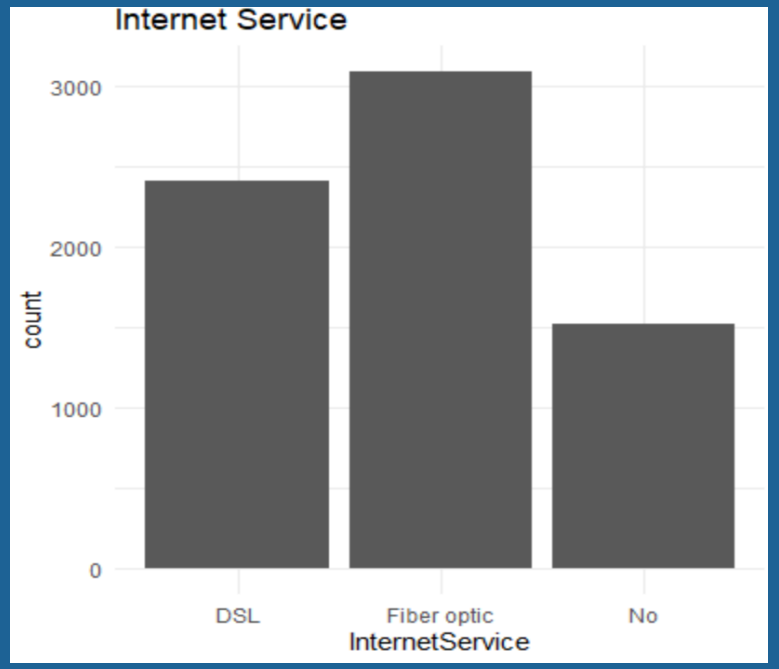

Senior Citizen
count
6000
4000
2000
0
No
Yes
SeniorCitizen

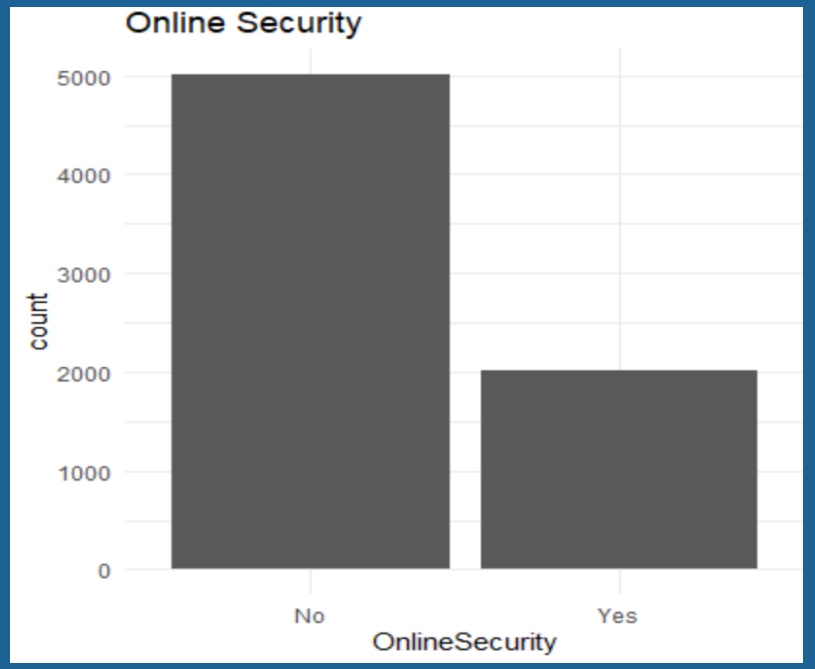

Partner
count
3000
2000
1000
0
No
Yes
Partner

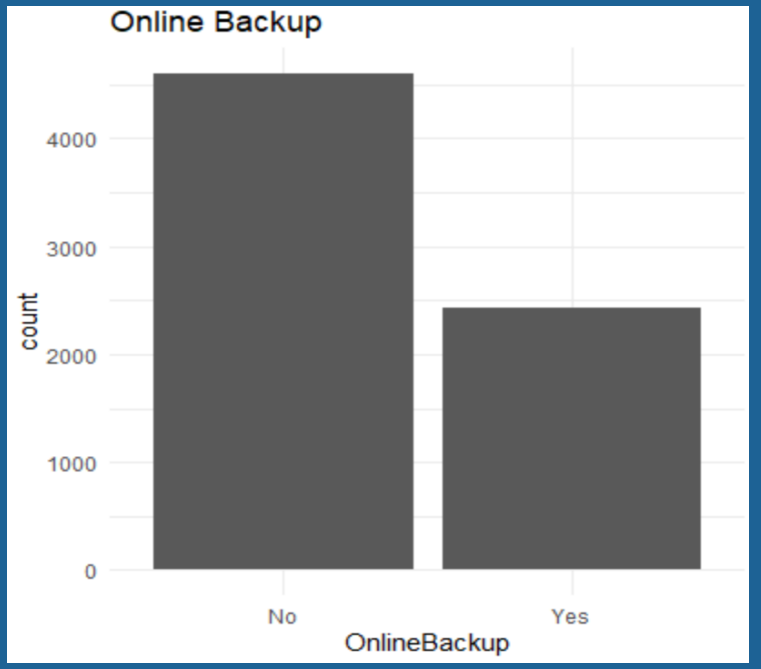

Dependents
count
5000
4000
3000
2000
1000
0
No
Yes
Dependents

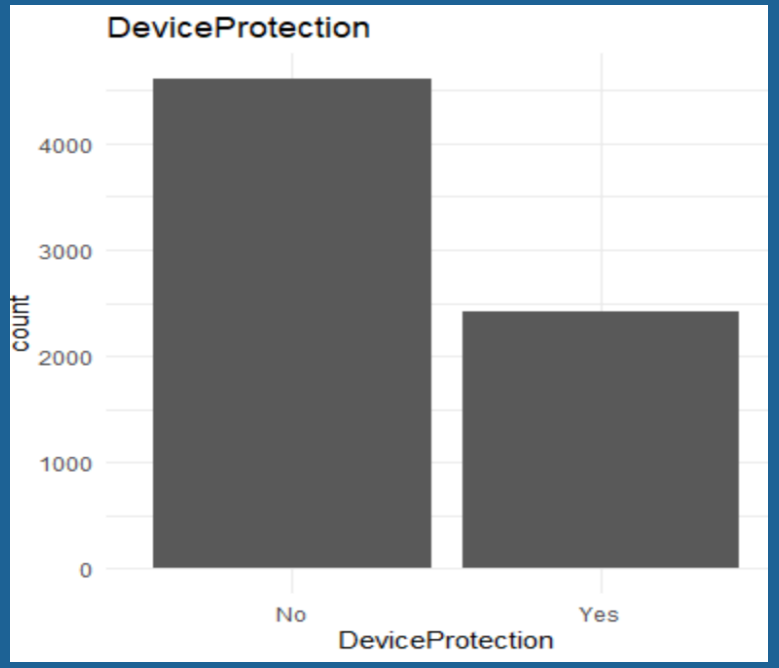

Phone Service
count
6000
4000
2000
0
No
Yes
PhoneService

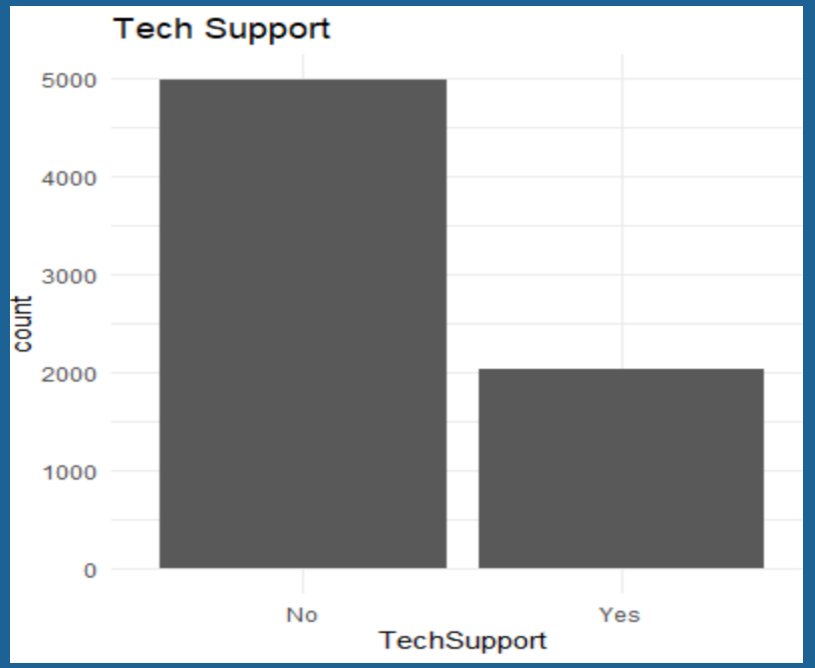

Multiple Lines
count
4000
3000
2000
1000
0
No
Yes
MultipleLines

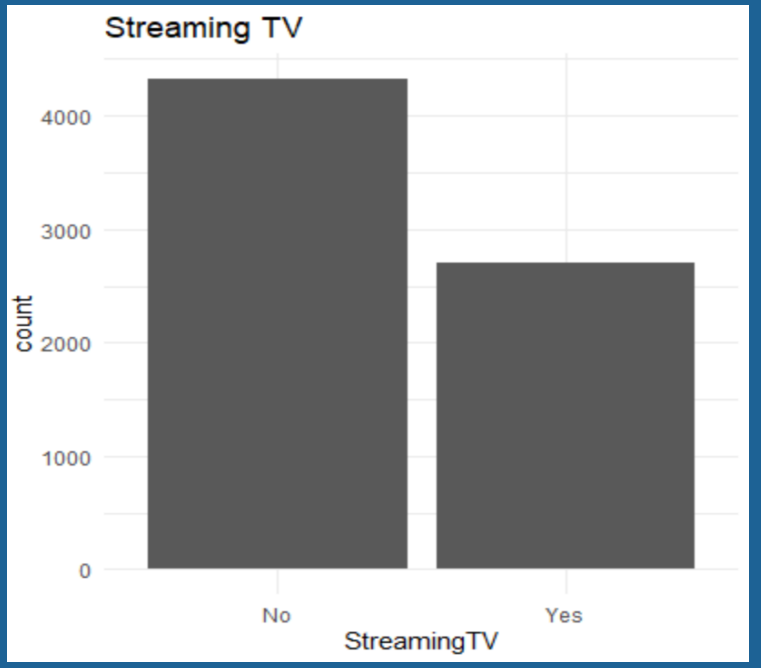

Internet Service
count
3000
2000
1000
0
DSL
Fiber optic
No
InternetService

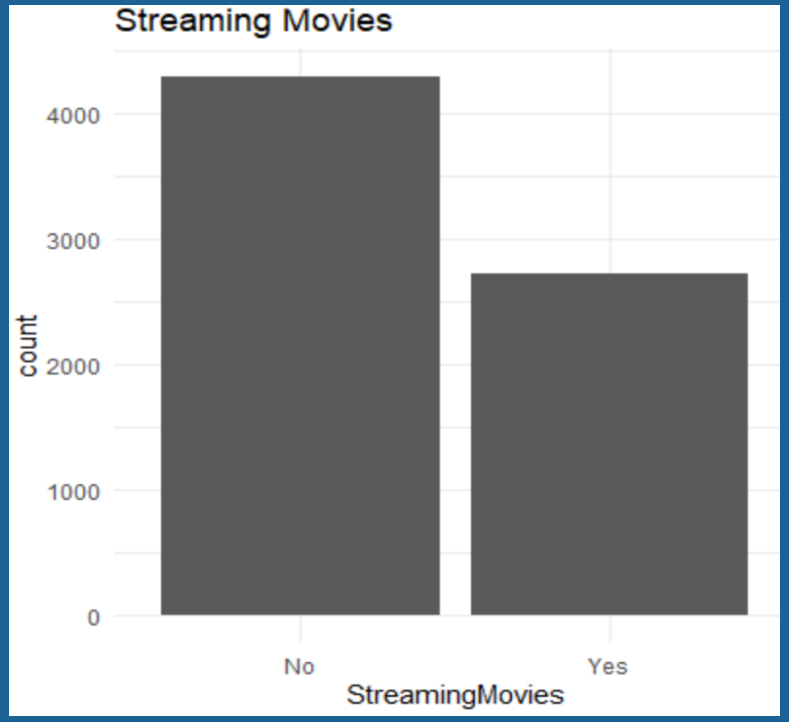

Online Security
count
5000
4000
3000
2000
1000
0
No
Yes
OnlineSecurity

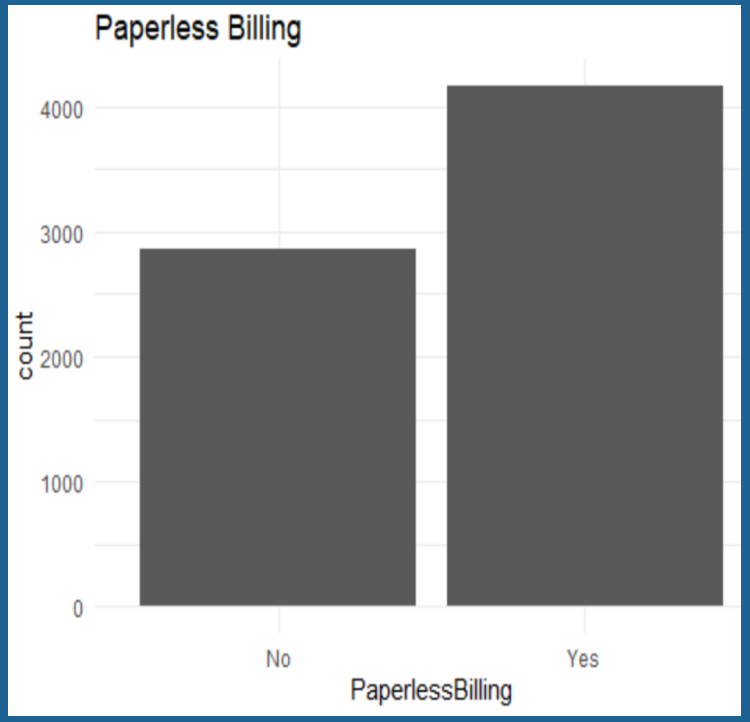

Online Backup
count
4000
3000
2000
1000
0
No
Yes
OnlineBackup

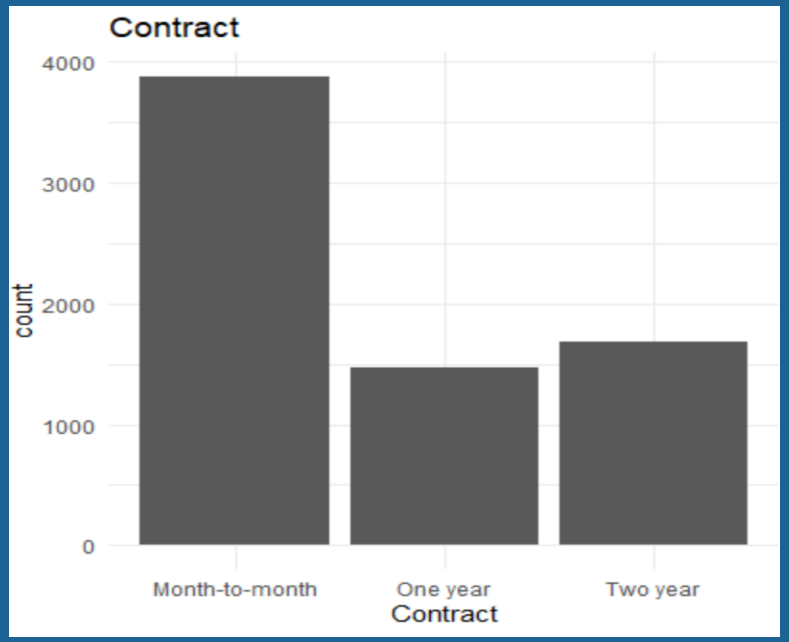

DeviceProtection
count
4000
3000
2000
1000
0
No
Yes
DeviceProtection

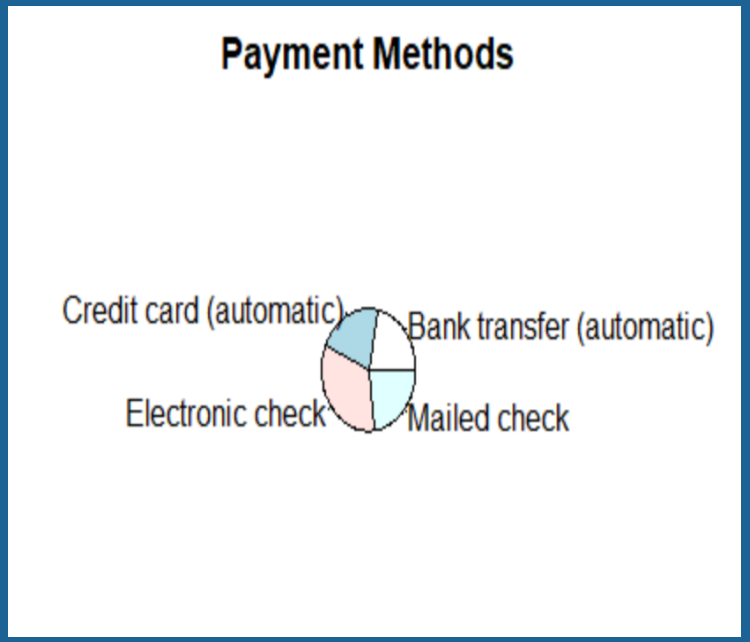

Tech Support
count
5000
4000
3000
2000
1000
0
No
Yes
TechSupport


Streaming TV
count
4000
3000
2000
1000
0
No
Yes
StreamingTV



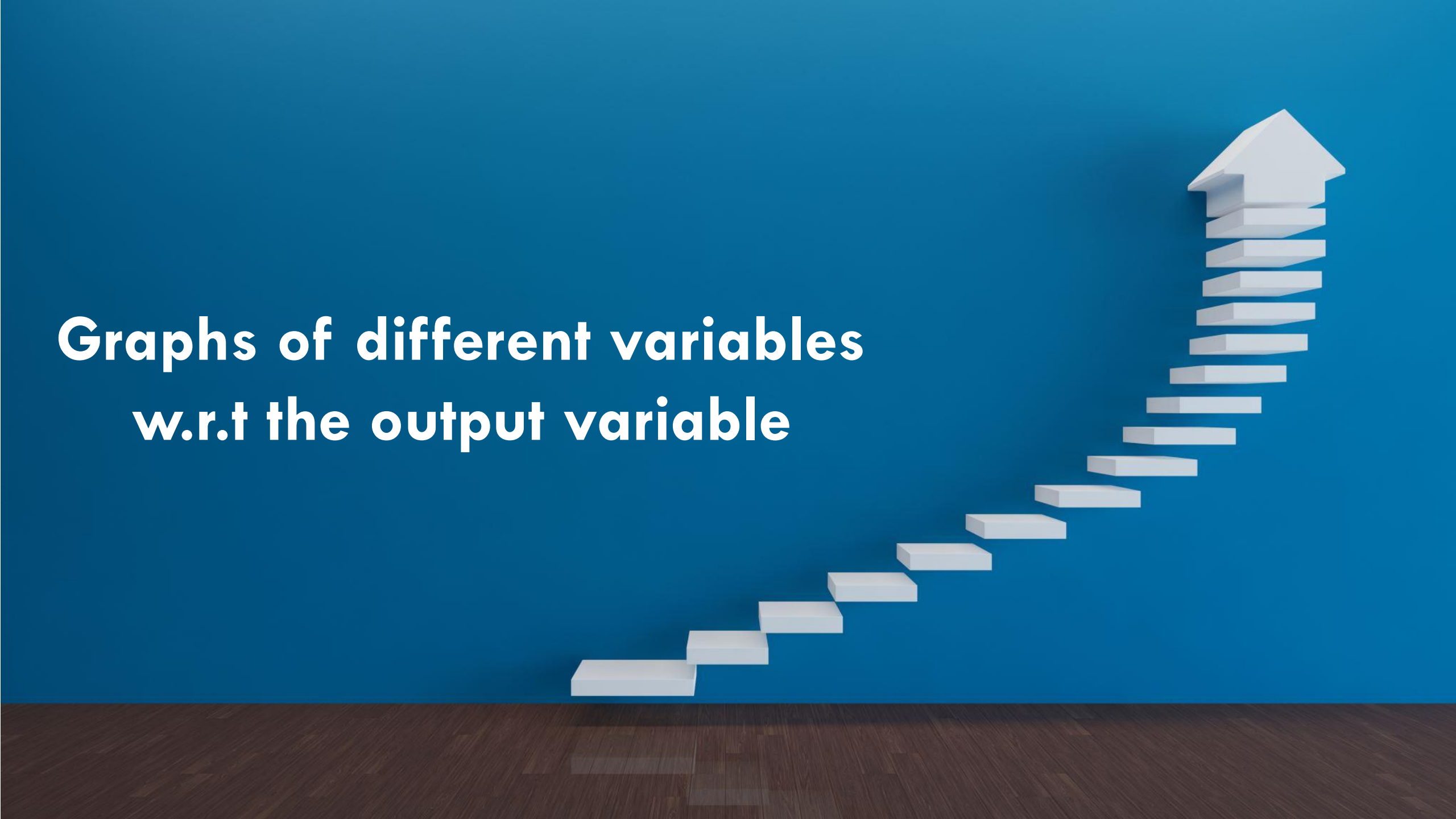



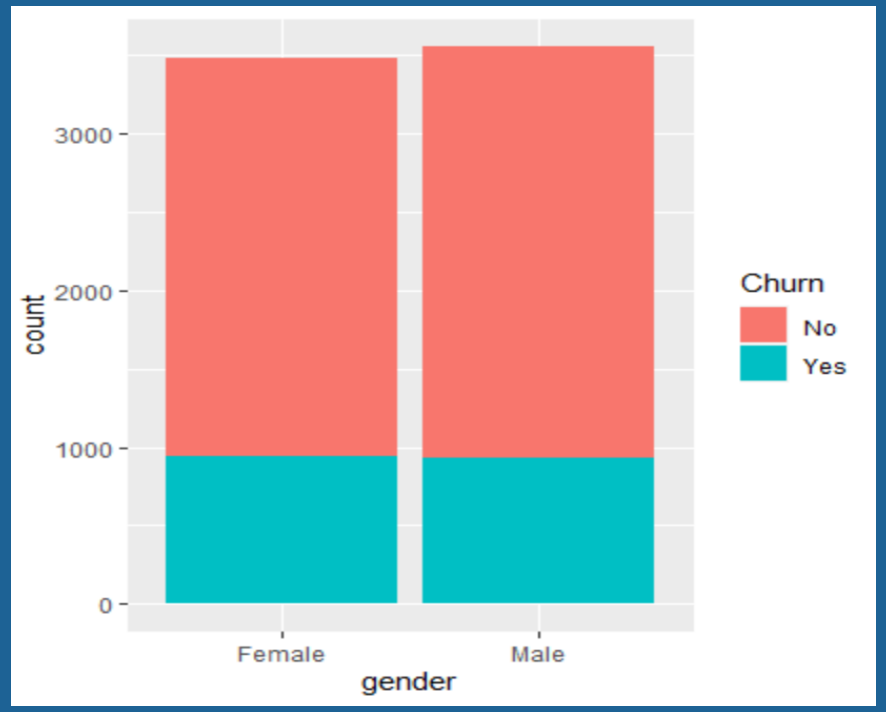



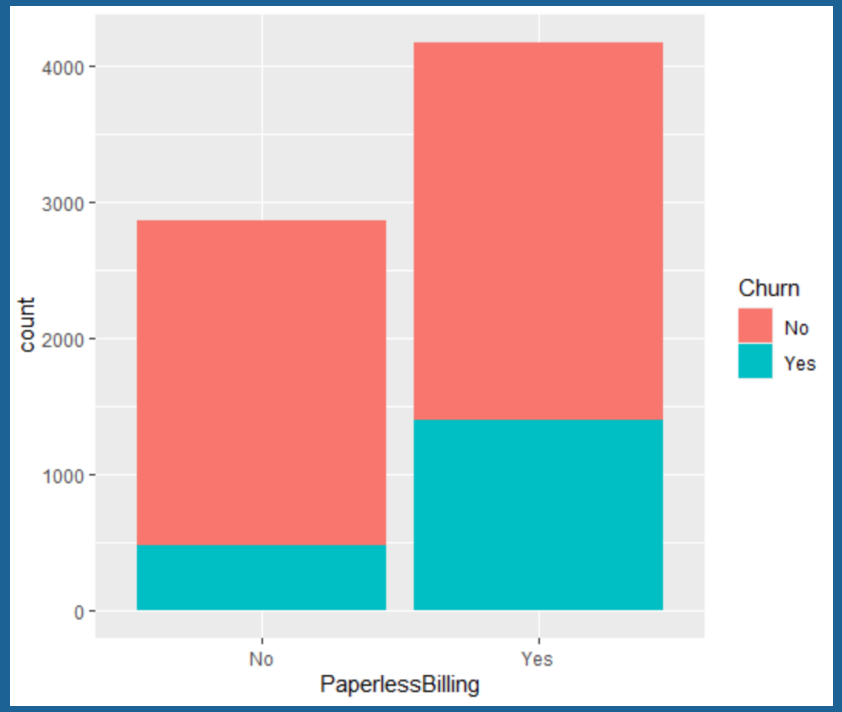

Some inferences:-

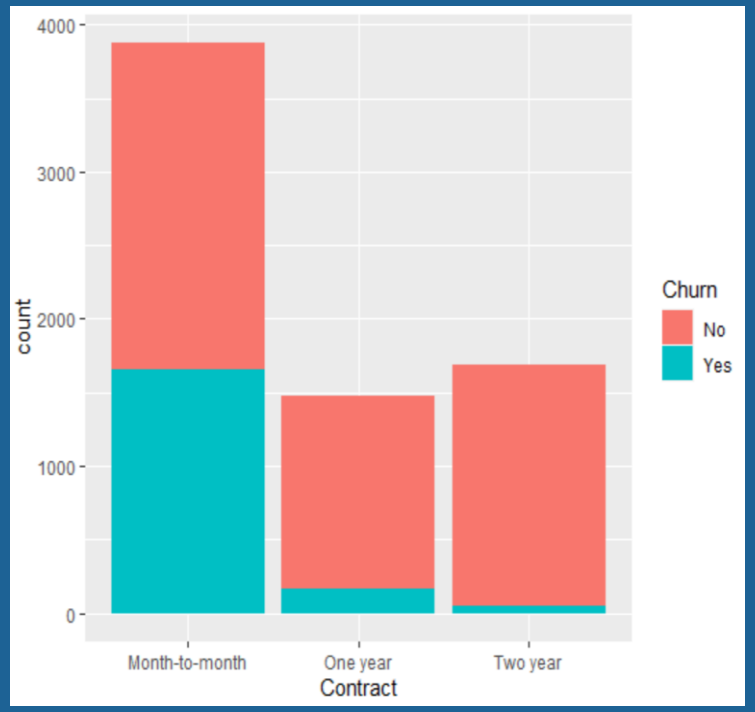
- Subscribers prefer Month-to-month contract over One or Two year contracts.
- Paperless Billing is more preferable.
- Highest number of subscribers preferred Electronic check payment method.

# Graphs of different variables w.r.t the output variable





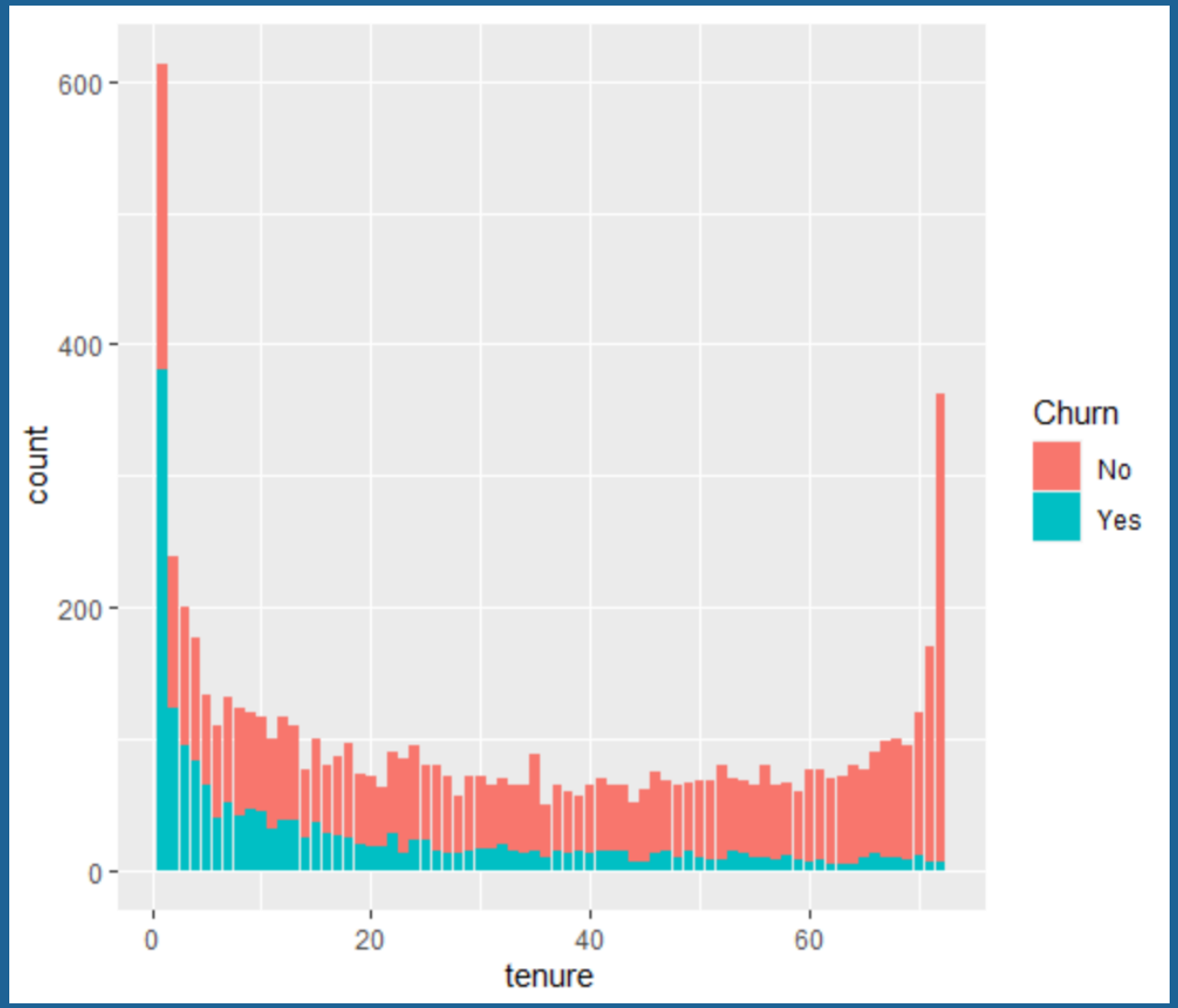



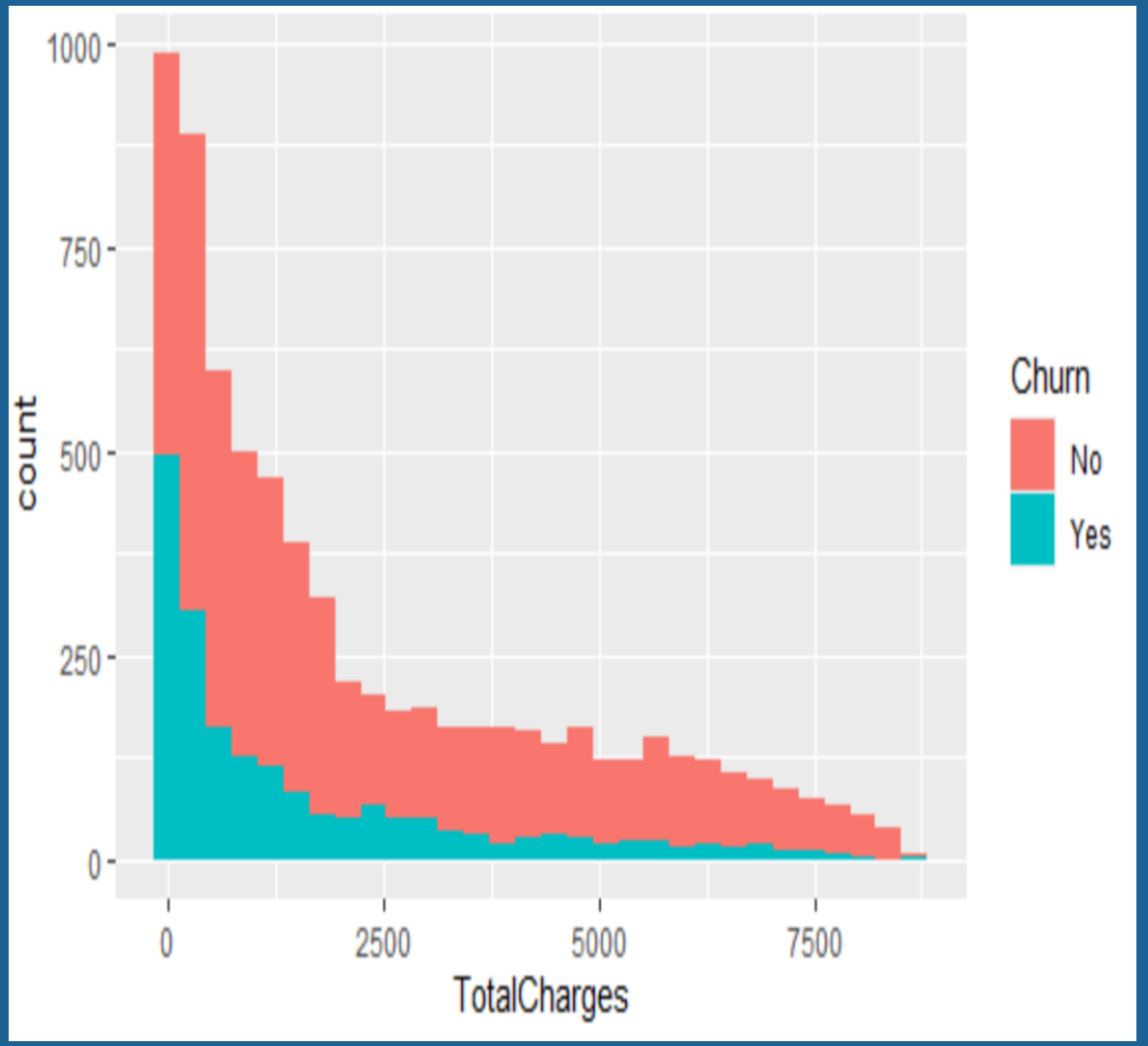




- Churning is unaffected by the Gender of the person.

- People who prefer Paperless Billing churn more.

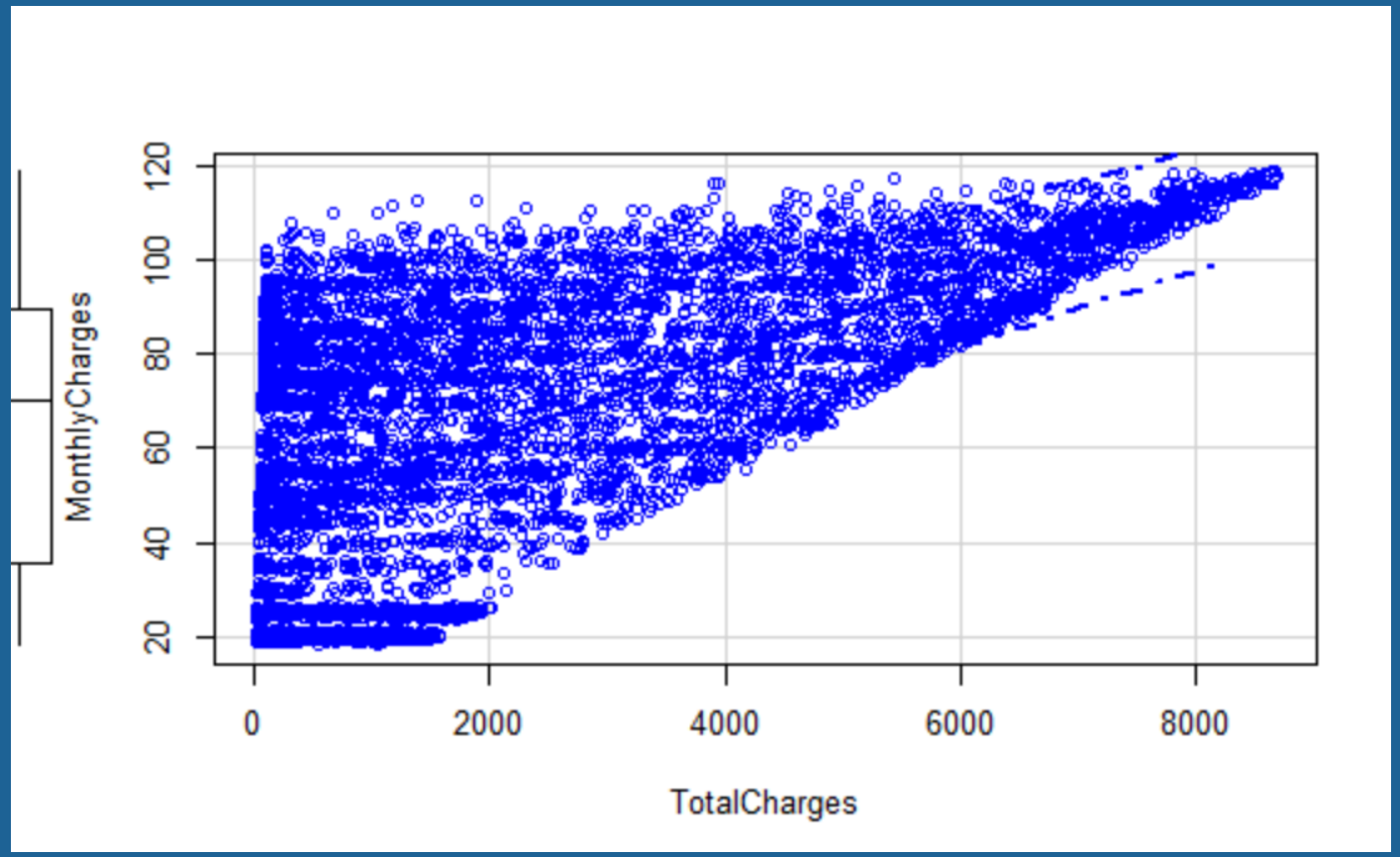
- The subscribers who prefer Month-to-month Contract have high rate of churning.




- The subscribers who preferred shorter tenures churn more than the one's opting for longer tenures.

- The number of customers churning decreases as the Total charges increase.




- The Monthly Charges and Total Charges has high positive correlation, which can be seen in the scatter plot as well.
- As these two variables are correlated, only one is required for the analysis and its safe to drop the other.

# Logistic Regression Model

- We chose to split the data in 3:1 ratio and use 75% of the data to build the model and rest 25% for prediction and checking the accuracy.
- First created the model with all the independent variables.
- On analysing the summary of the first model, we built different models with all those variables which had high **Significance**.
- The best fit model with the lowest AIC and Residual deviance was the one built using following variables:

tenure+InternetService+PhoneService+MultipleLines+OnlineSecurity+OnlineBackup+PaperlessBilling+TechSupport+Contract+PaymentMethod+TotalCharges

```
    Null deviance: 5998.3  on 5181  degrees of freedom
Residual deviance: 4319.1  on 5159  degrees of freedom
AIC: 4365.1
```

# Performance of Logistic Regression Model

| | y_pred | |
|---|---|---|
| | 0 | 1 |
| No | 1206 | 151 |
| Yes | 208 | 285 |

Confusion matrix

| | |
|---|---|
| Accuracy | 80.59% |
| Error percent | 19.41% |
| Accuracy of predicting “YES” | 57.81% |
| Accuracy of predicting “NO” | 88.87% |

- **Overall accuracy of the model is fairly good.**
- **Accuracy of predicting customers churned is low compared to the accuracy of predicting not churning customers**

# Random Forest model

- After splitting the data into training and test sets, we created an Initial Random forest model which had an Out-of-bag error rate of 20.19%.

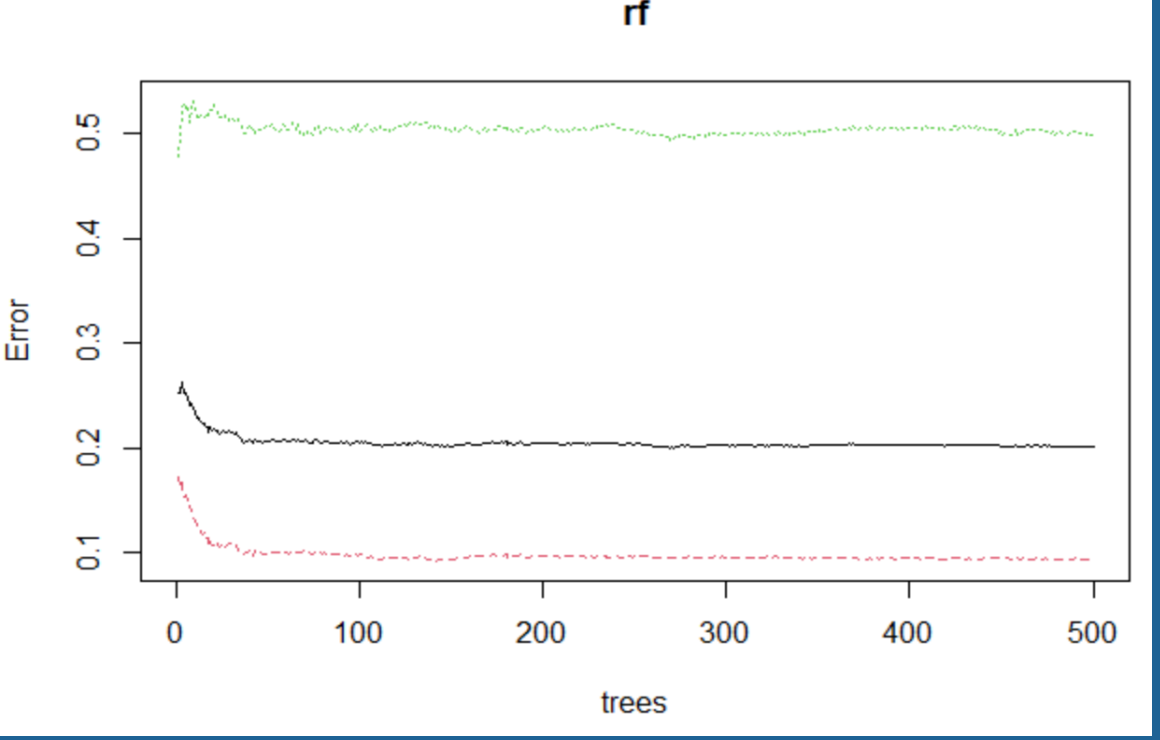
```
        OOB estimate of  error rate: 20.19%
Confusion matrix:
      No Yes class.error
No  3453 361  0.09465128
Yes  685 682  0.50109729
```

**Accuracy of the initial random forest model**

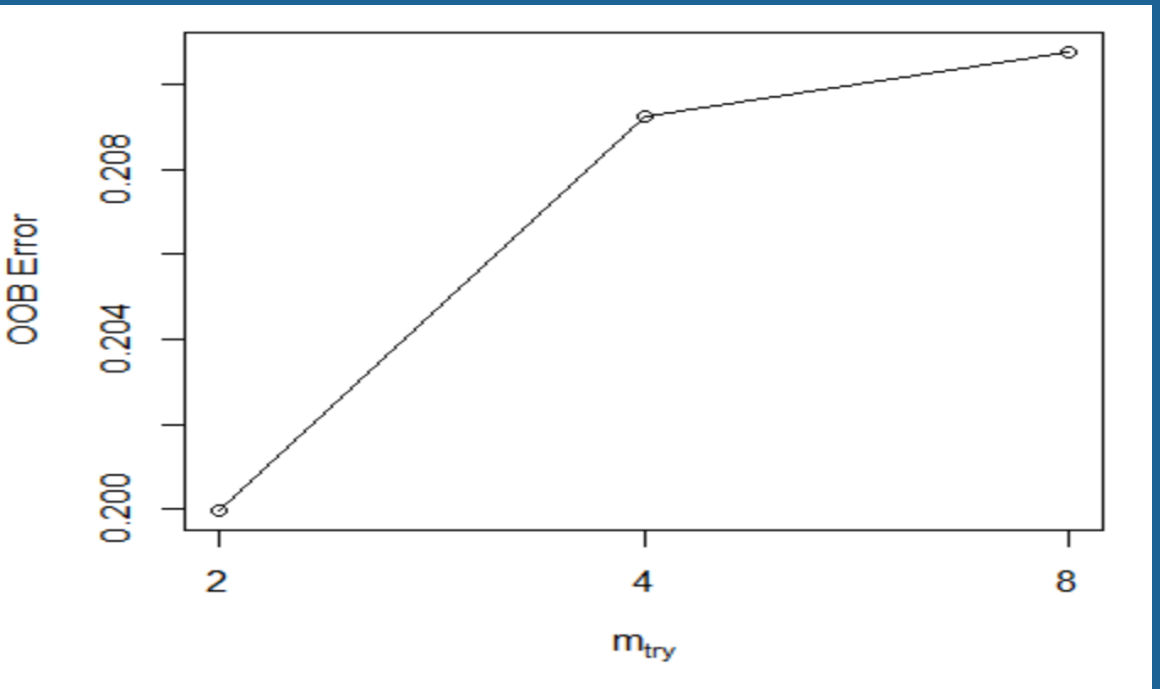
```
    pred
      No  Yes
No  1243  106
Yes  272  230
```

Confusion matrix

| | |
|---|---|
| Accuracy | 79.58% |
| Error percent | 20.42% |
| Accuracy of predicting “YES” | 45.82% |
| Accuracy of predicting “NO” | 92.14% |




- The visual representation of the initial model suggests that the OOB error rate decreases only till 200 trees after that it becomes constant and does not decrease further.




- After tuning the model, graph depicts that the OOB error is the least when $m_{try} = 2$.

# Performance of Random Forest Model

- Using all the previous findings, we built a better random forest model.
- On taking number of trees = 200 and mtry = 2, the OOB error decreases from 20.19% to 20.03%

```
        OOB estimate of  error rate: 20.03%
Confusion matrix:
      No Yes class.error
No  3516 298  0.07813319
Yes  740 627  0.54133138
```

## Accuracy Calculations

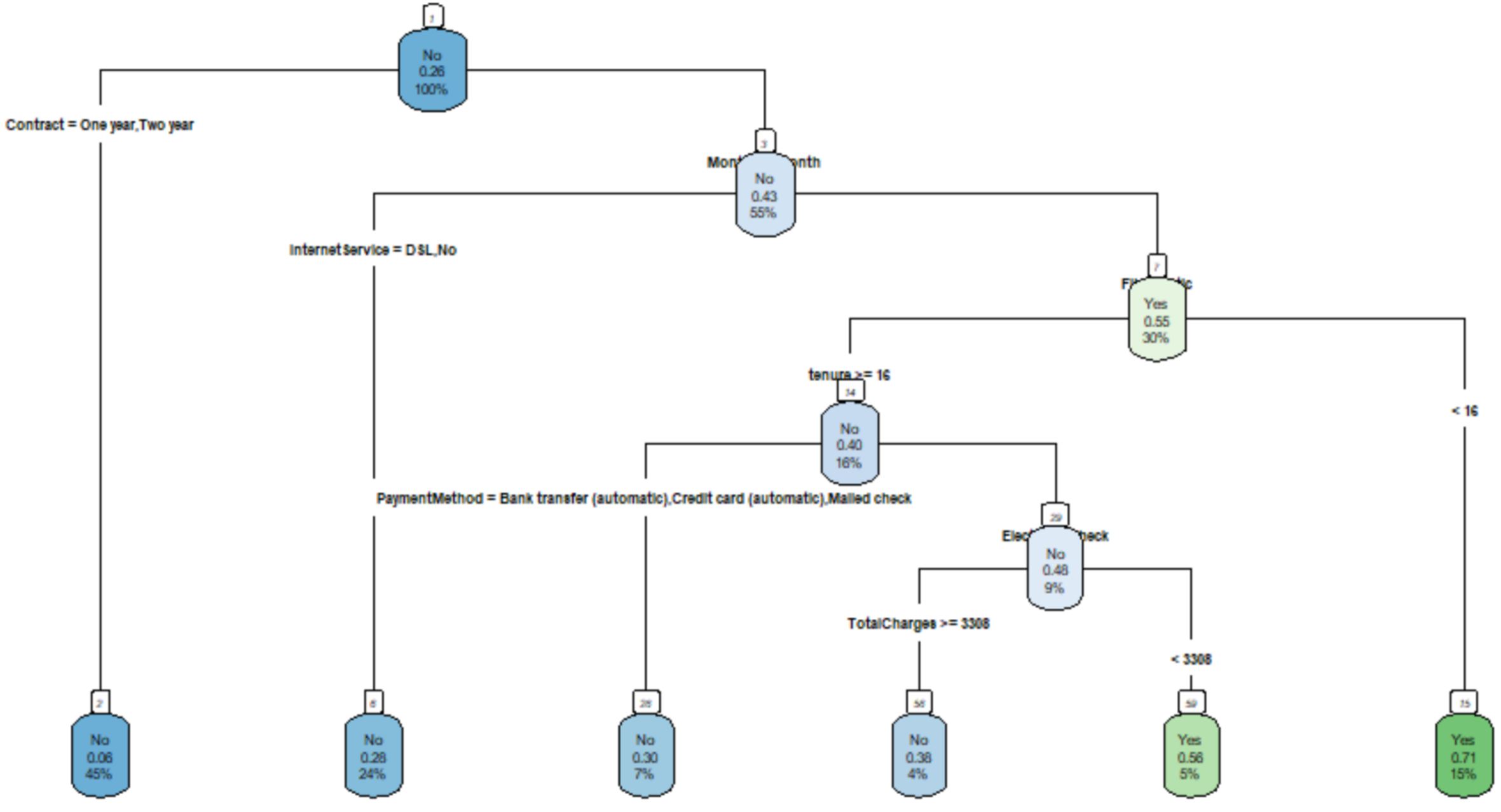
| | pred | |
|---|---|---|
| | No | Yes |
| No | 1216 | 133 |
| Yes | 244 | 258 |

Confusion matrix

| | |
|---|---|
| Accuracy | 79.63% |
| Error percent | 20.37% |
| Accuracy of predicting “YES” | 51.40% |
| Accuracy of predicting “NO” | 90.14% |

➢ **Overall accuracy is lower than Logistic regression model, though accuracy for predicting the non Churning customers is high.**

# Decision Tree



# Observations

- The tree illustrates that the variable Contract is the most crucial in predicting Customer churn.
- If a customer has a one year or a two year contract, it is less likely that they will churn (with 7% probability of loosing)
- If a customer is in a Month-to-month contract and has opted for fibre optics as internet service, there is a probability of 55% that he(she) will churn, also if such a customer is in a tenure less than 16 months then the probability of churning is even high (71% probability)
- Further, the churning of a customer (with fibre optics) who is in a contract for more than 16 months depends on the payment method and total charges, if the payment method is Electronic check and the Total charge > 3308 then there is probability of 56% that such a customer will churn.

# Performance of Decision Tree algorithm

| | predictChurn_test | |
|---|---|---|
| | 1 | 2 |
| No | 1385 | 150 |
| Yes | 322 | 253 |

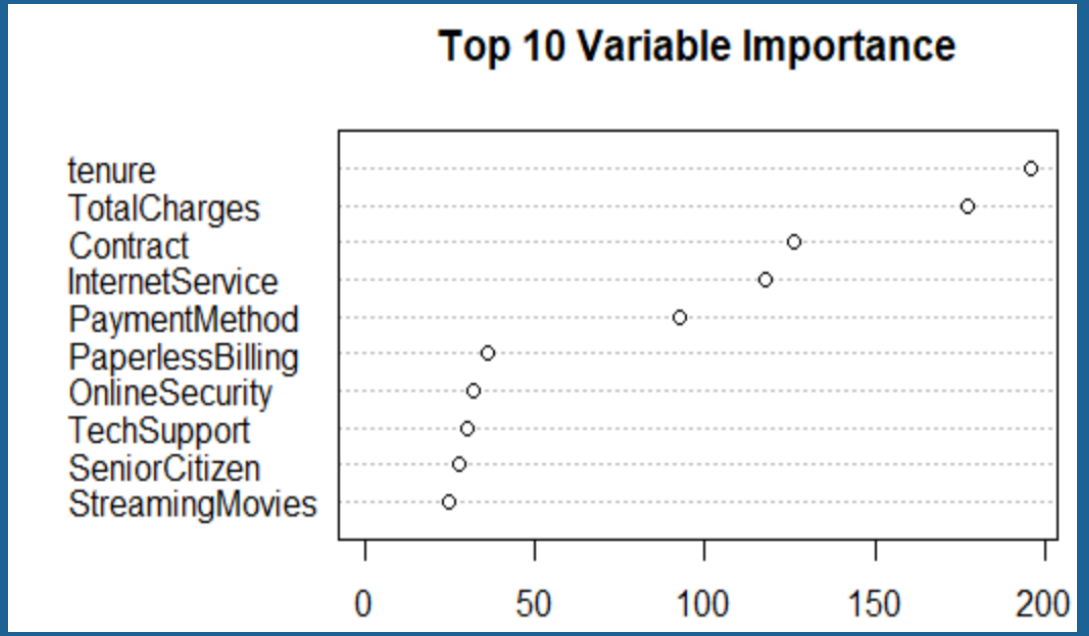
Confusion matrix

| | |
|---|---|
| Accuracy | 77.63% |
| Error percent | 22.37% |
| Accuracy of predicting “YES” | 56.00% |
| Accuracy of predicting “NO” | 90.02% |

- **The model is efficient in predicting the customers who did not churn. The overall accuracy of the model is lower than the previous two models we built.**

# Summary

- All the three models suggest that the variables such as Tenure, Total charges, Contract, Internet Service, Paperless Billing and Payment Method have high importance and play a significant role in customer churn.
- Even though all the three models have fairly good accuracy level, but we are more interested in the customers who churn so we will go with the model which predicts that accurately enough.
- Logistic regression model had the highest Accuracy in predicting “Yes” (i.e. 57.81%) so we will use this for future predictions.



# Conclusion

- Churn is dependent on contract & tenure of association of customer with telecom company so the company should target customers who are interested in long duration contract either 1 year or more.
- Churn rate is also dependent on Monthly charges so telecom company should revise the monthly rate.
- Maximum churning occurs if the tenure is less & customer opt for month to month contract.