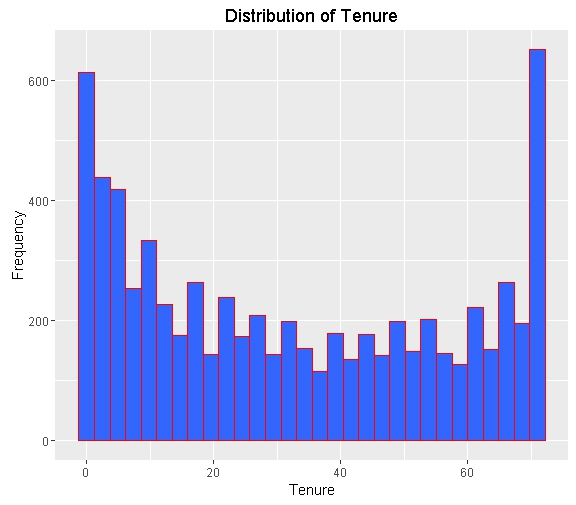
TELECOM CHURN CASE STUDY SUBMISSION

**NOTE:** This should briefly describe the important results and recommendations. The structure is suggestive; make sure to not exceed 7 pages**.**

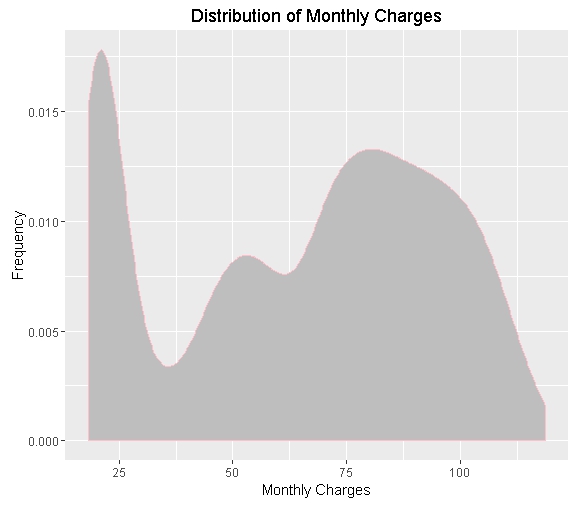
# Checkpoint-1: Data Understanding and Preparation of Master File

* The final number of rows are 7032 with 32 columns

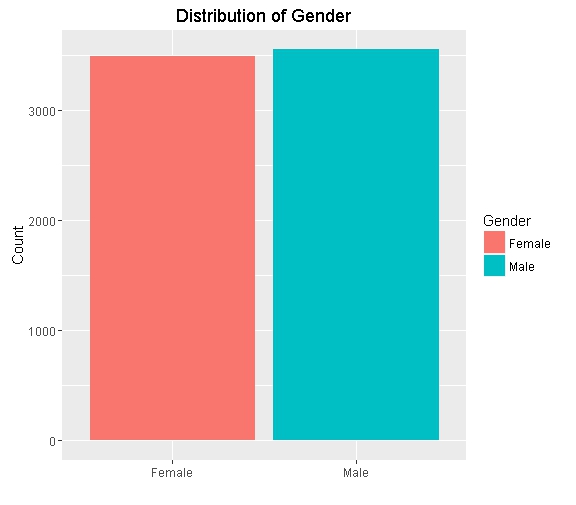
# Checkpoint 2: Exploratory Data Analysis



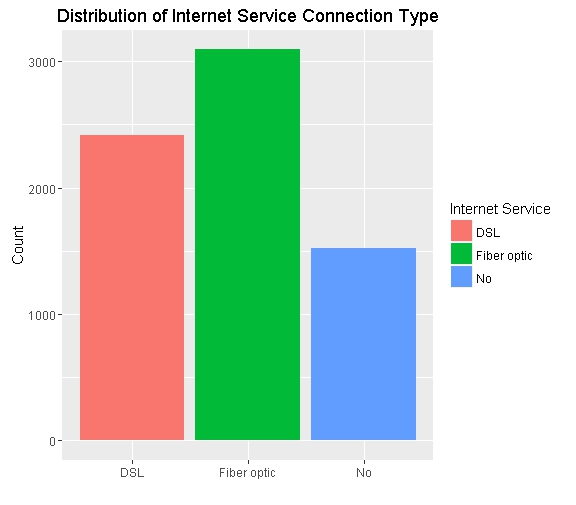
* We see that there is high number of people who are using the services for more than a year.



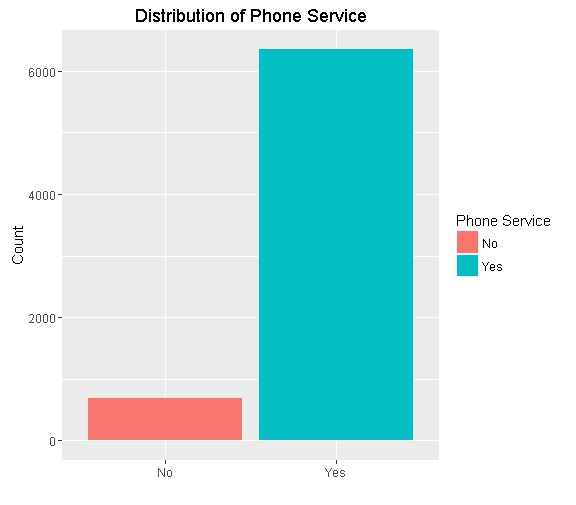
* We see that there is large number of concentration on monthly charges less than 25 and between 65 to 100.



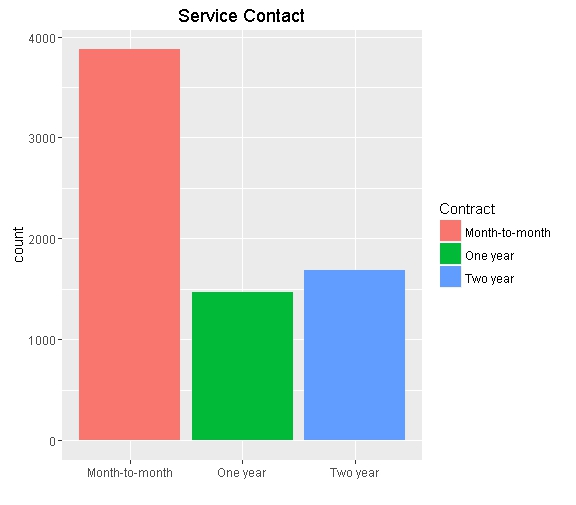
* we see that there is more of a equal distribution of male and female



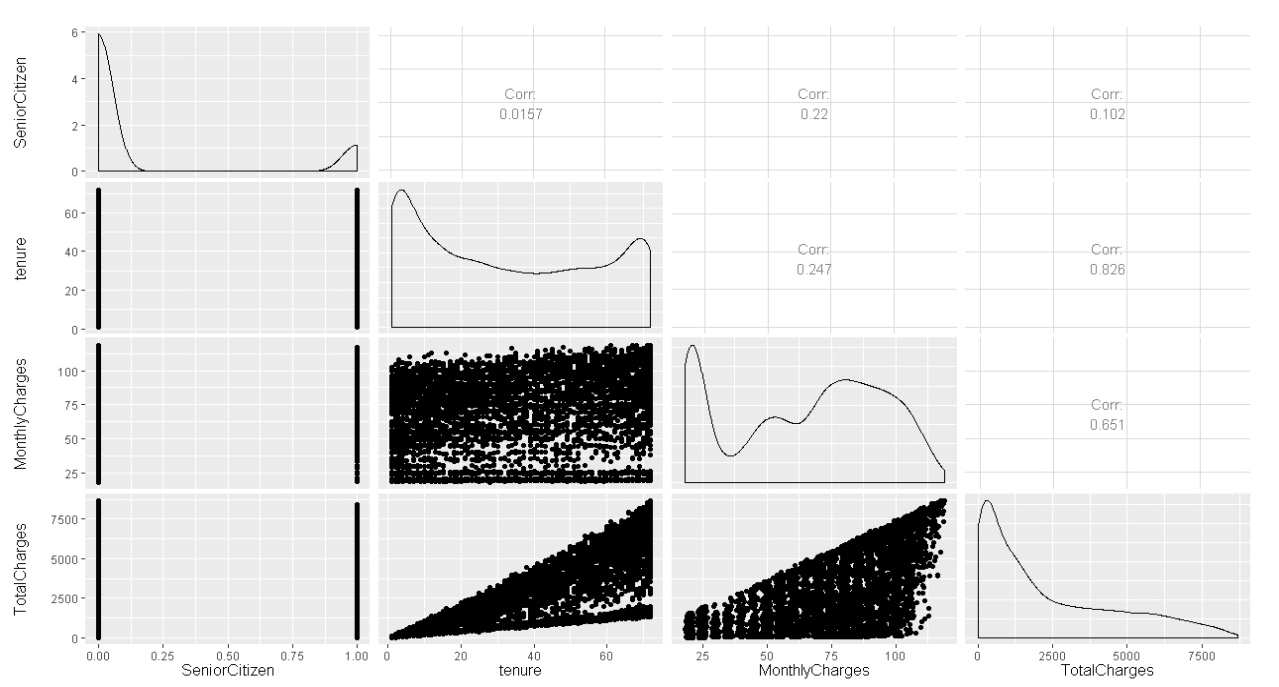
* we see that about 1500 customers don't use internet service



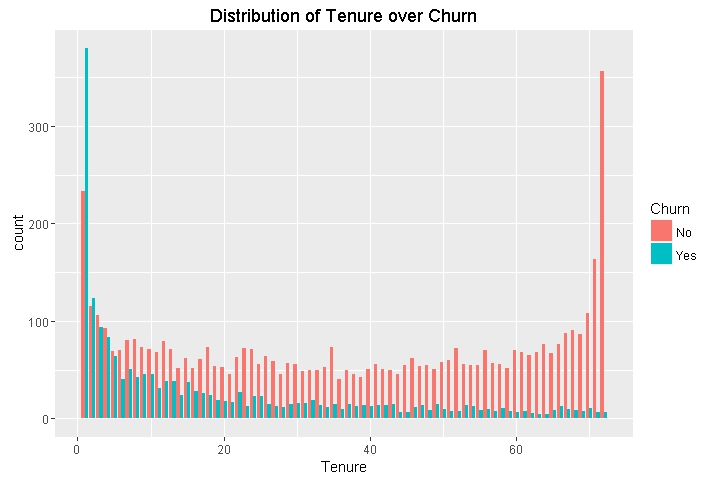
* we see that about 700 customers don't use phone service.



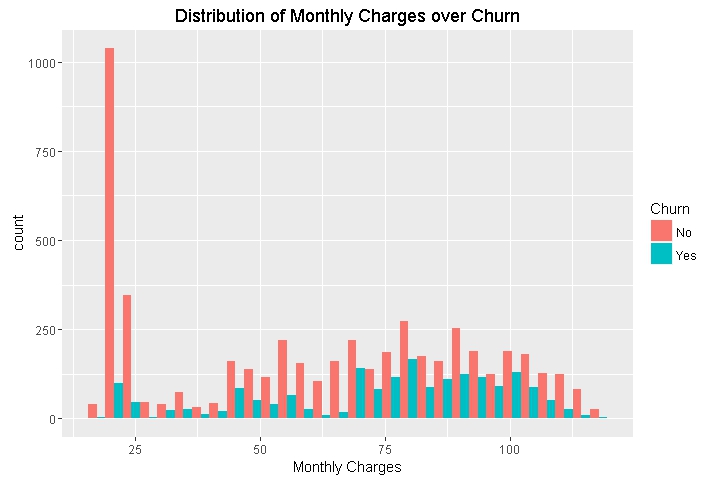
* we see that most customer are on month to month contract



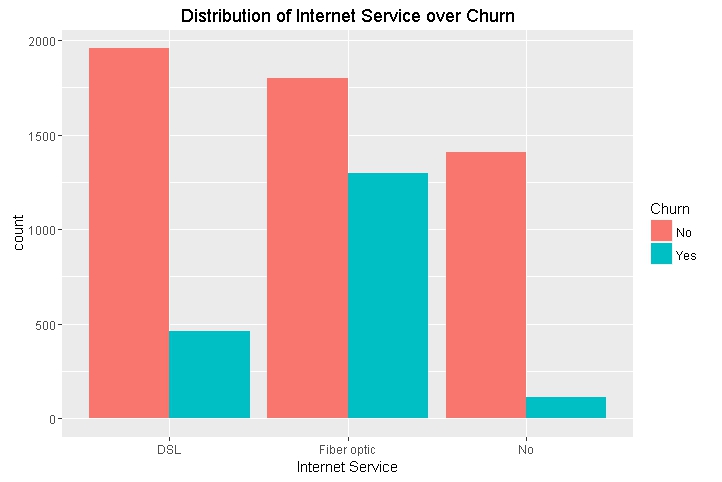
* we see that tenure and monthly charges are correlated with total charges, which means if tenure increases total charges also increases and when monthly charges increases total charges also increases.



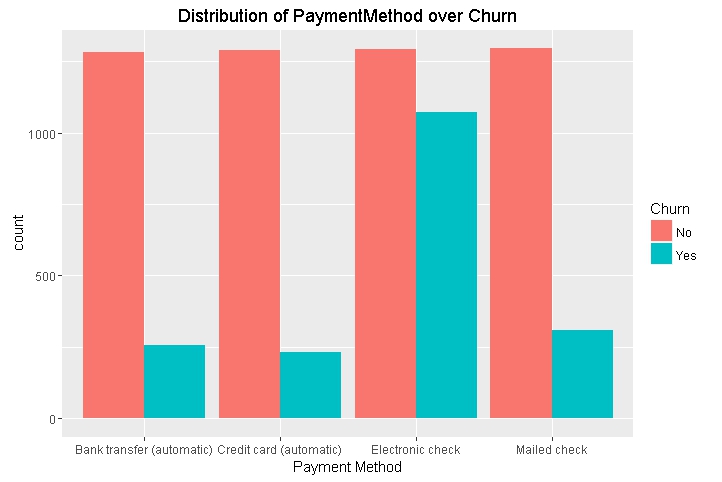
* we see that if the tenure is increases the churn is reduces, so we can conclude that tenure is negatively correlated with churn



* we see that highest frequency of customers who churn are having monthly charges between 65 to 105



* we see that most customers who churn are on fibre optic.



* we see that most customers who churn are using payment method electronic check

# Checkpoint 3: Data Preparation

* There were no duplicate found in the data
* Explain the methodology of Missing value treatment and additionally fill the below table:

|  |  |
| --- | --- |
| **Questions** | **Results(Numeric)** |
| Total number of observations in the dataset | 7043 |
| Total number of variables in the dataset | 21 |
| Total missing values in the dataset | 11 |

* Explain the methodology of Outlier treatment and fill the below table: **boxplot**
* Bring the data in the correct format. Report the number of variables for which the format was changed.

Additionally, fill the below table:

|  |  |
| --- | --- |
| **Operations performed** | **Variable Name** |
| Outlier treatment | No Outlier Found |
| Dummy creation |  |
| Binning of variables | tenure |

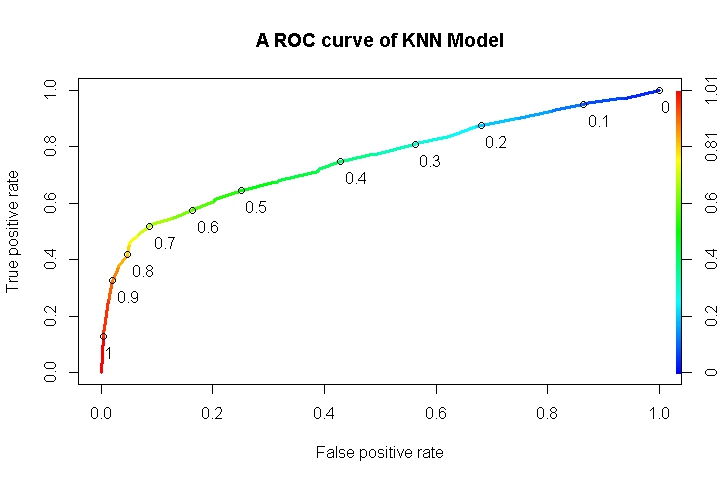
# Checkpoint 4: Modelling

* **Model – K-NN**
  + Explain the Data Preparation step for K-NN modelling.
    - Loading all the three the data into R
    - Checking for any duplicate in customer ID since it is the unique key
    - Merging the three data sets by using customer id as the unique key
    - Removing 11 na values, since its consists of only 0.15% of the entire data we have removed the 11 rows
    - Binning the variable tenure into short, medium and long
    - Converting all the character variable into factor class
    - Scaling the continuous variable
    - creating dummy variable for all the factor variable using n-1 levels except for churn variable
    - creating final data by binding the continuous data with the dummy variables
    - creating split index with a split ration of 0.70 on churn variable
    - use the split index to create training and testing data
  + Explain the methodology of building the model with optimal value of K?
    - Started with a value of k=1 and obtain a accuracy of 72%, Sensitivity of 80% and Specificity of 50%
    - Then with a value of k=3 and obtain a accuracy of 74%, Sensitivity of 83% and Specificity of 50%
    - Then with a value of k=7 and obtain a accuracy of 77%, Sensitivity of 85% and Specificity of 55%
    - After continuous iteration we found that at k=47 we are getting a accuracy of 80, Sensitivity of 87% and Specificity of 61%, which is the highest in each metrics, if we are increasing any more iteration we are either getting lower accuracy, or lower sensitivity or lower specificity. So we conclude that 47 will be the optimal value for k.
    - We can also see the optimal value of K in the ROC curve which is shown below.

Additionally, fill the below table:

|  |  |
| --- | --- |
| **Threshold value** | **Values (Numeric)** |
| Overall Accuracy | 80 |
| Sensitivity | 87 |
| Specificity | 61 |
| AUC | 75 |

* + Display the AUC curve.



* **Model – Naïve Bayes**
  + Explain the Data Preparation step for Naïve Bayes modelling.
    - Loading all the three the data into R
    - Checking for any duplicate in customer ID since it is the unique key
    - Merging the three data sets by using customer id as the unique key
    - Removing 11 na values, since its consists of only 0.15% of the entire data we have removed the 11 rows
    - Binning the variable tenure into short, medium and long
    - Converting all the character variable into factor class
    - Scaling the continuous variable
    - creating dummy variable for all the factor variable using n-1 levels except for churn variable
    - creating final data by binding the continuous data with the dummy variables
    - creating split index with a split ration of 0.70 on churn variable
    - use the split index to create training and testing data
  + Explain the methodology of building the model.

Additionally, fill the below table:

|  |  |
| --- | --- |
| **Threshold value** | **Values (Numeric)** |
| Overall Accuracy | 70 |
| Sensitivity | 65 |
| Specificity | 83 |
| AUC |  |

* + Display he AUC curve.
* **Model – Logistic Regression**
* Explain the methodology of building the model? In the final model, interpret what the coefficients of the variable imply. Check if the coefficients make business sense

Ans:-We have converted the dependant variable with 0 and 1, Since Logistic regression takes numeric variables as inputs.

The coefficients of our model are:- `Contract : Month-to-month` + `Dependents : No` + `DeviceProtection : No internet service` + `InternetService : DSL` + `MultipleLines : No` + `OnlineSecurity : No` + `PaperlessBilling : No` + PaymentMethod : Electronic check` + `StreamingTV : No` + `TechSupport : No` + TotalCharges. We observe that `DeviceProtection : No internet service` has -1.67553 as coefficients, which means if the customer have no internet service his probability of churning will be very less, which makes sense, also we see that Contract : Month-to-month has 1.06600, which means that if the customer is on a month to month contract his changes of churning will be very high, which completely makes sense.

Additionally, fill the below table:

|  |  |
| --- | --- |
| **Significant variables in final model (add more rows if requires)** | **Coefficients value (Numeeric)** |
| Contract : Month-to-month` | 1.066 |
| `Dependents : No` | 0.27851 |
| `DeviceProtection : No internet service` | -1.67553 |
| `InternetService : DSL` | -0.99918 |
| `MultipleLines : No` | -0.30452 |
| `OnlineSecurity : No` | 0.44351 |
| `PaperlessBilling : No` | -0.27778 |
| `PaymentMethod : Electronic check` | 0.3707 |
| `StreamingTV : No` | -0.51884 |
| `TechSupport : No` | 0.42661 |
| TotalCharges | -0.70983 |

|  |  |
| --- | --- |
| **Final model metrics** | **Values (Numeric)** |
| AIC value | 4222.2 |
| Null deviance | 5699.5 |
| Residual Deviance | 4198.2 |

* Calculate c-statistic and KS-statistic. What can you tell about the model based on their values?
  + We see that the c-statistic for test data is 8.36, which is close to 1, we can safely say that our model has good discriminative power.
  + Our ks-statistic is at 0.40 which is good, in 4th decile in test data set.

Additionally, fill the below tables:

**Note**: Write the numeric value of c-statistic and KS-statistic after applying your final model to the train dataset and test dataset.

|  |  |  |  |
| --- | --- | --- | --- |
| **Train Dataset** | | **Test Dataset** | |
| C-statistic | 8.37 | C-statistic | 8.36 |
| KS-statistic | 0.51 | KS-statistic | 0.54 |
| Model Evaluation (write Accept or Reject) | | Accept | |

|  |  |
| --- | --- |
| **Threshold value** | **Values (Numeric)** |
| Overall Accuracy(.30) | 0.76 |
| Sensitivity | 0.7702 |
| Specificity | 0.7629 |
| AUC | 0.83 |

* **Model – SVM**
  + Explain the Data Preparation step for SVM modelling.
    - Loading all the three the data into R
    - Checking for any duplicate in customer ID since it is the unique key
    - Merging the three data sets by using customer id as the unique key
    - Removing 11 na values, since its consists of only 0.15% of the entire data we have removed the 11 rows
    - Binning the variable tenure into short, medium and long
    - Converting all the character variable into factor class
    - Scaling the continuous variable
    - creating dummy variable for all the factor variable using n-1 levels except for churn variable
    - creating final data by binding the continuous data with the dummy variables
    - creating split index with a split ration of 0.70 on churn variable
    - use the split index to create training and testing data

* + Explain the methodology of building the model.

Additionally, fill the below table:

|  |  |
| --- | --- |
| **Threshold value** | **Values (Numeric)** |
| Overall Accuracy | 0.50 |
| Sensitivity | 0.90 |
| Specificity | 0.51 |
| AUC |  |

* Report the best model and its performance metrics.

|  |  |
| --- | --- |
| **Threshold value** | **Values (Numeric)** |
| Overall Accuracy | 70.24 |
| Sensitivity | 65.33 |
| Specificity | 83.78 |
| AUC |  |

# Checkpoint 6: Threshold value

* Select an appropriate threshold value and calculate the confusion matrix and overall accuracy, sensitivity and specificity

Additionally, fill the below table:

|  |  |
| --- | --- |
| **Threshold value** | **Values (Numeric)** |
| Overall Accuracy |  |
| Sensitivity |  |
| Specificity |  |