Naufal Rasyid Sutansyah

“ Telco Customer Churn Dataset ”
Report in R

1. Introduction to the dataset

This dataset that contains information about customers. Each row in the dataset represents a different customer, while each column provides details about the customer's characteristics as described in the column metadata.

The dataset includes the following information:

1. Churn: This column indicates whether a customer has recently left within the last month.
2. Services: This column lists the services that each customer has signed up for, such as phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies.
3. Customer account information: This includes details about how long each customer has been with the company, their contract type, payment method, whether they prefer paperless billing, and their monthly and total charges.
4. Demographic information: This section provides information about the customers' gender, age range, and whether they have partners and dependents.

2. Explotary Data Analysis

- **Data Set**

data	7043 obs. of 21 variables

Data Insight :
Dataset include 7043 sample dan 21 coloumn :

Feature dataset :
- customerID: A unique ID that identifies each customer.,
- gender: The customer’s gender - Male, Female,
- SeniorCitizen: Whether the customer is a senior citizen or not (1, 0),
- Partner: Whether the customer has a partner or not (Yes, No),
- Dependents: Whether the customer has dependents or not (Yes, No),
- tenure: Number of months the customer has stayed with the company,
- PhoneService: Whether the customer has a phone service or not (Yes, No),
- MultipleLines: Whether the customer has multiple lines or not (Yes, No, No phone service),
- InternetService: Customer’s internet service provider (DSL, Fiber optic, No),
- OnlineSecurity: Whether the customer has online security or not (Yes, No, No internet service),
- OnlineBackup: Whether the customer has online backup or not (Yes, No, No internet service),
- DeviceProtection: Whether the customer has device protection or not (Yes, No, No internet service),
- TechSupport: Whether the customer has tech support or not (Yes, No, No internet service),
- StreamingTV: Whether the customer has streaming TV or not (Yes, No, No internet service),

- StreamingMovies: Whether the customer has streaming movies or not (Yes, No, No internet service),
- Contract: The contract term of the customer (Month-to-month, One year, Two year),
- PaperlessBilling: Whether the customer has paperless billing or not (Yes, No),
- PaymentMethod: The customer's payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic)),
- MonthlyCharges: The amount charged to the customer monthly,
- TotalCharges: The total amount charged to the customer,
- Churn: Whether the customer churned or not (Yes or No).

- **Data Head**

customerID <chr>	gender <chr>	SeniorCitizen <dbl>	Partner <chr>	Dependents <chr>	tenure <dbl>	PhoneService <chr>
7590-VHVEG	Female	0	Yes	No	1	No
5575-GNVDE	Male	0	No	No	34	Yes
3668-QPYBK	Male	0	No	No	2	Yes
7795-CFOCW	Male	0	No	No	45	No
9237-HQITU	Female	0	No	No	2	Yes
9305-CDSKC	Female	0	No	No	8	Yes

6 rows | 1-7 of 21 columns

Data Insight :
Here is the data content looks like as we can see.

- **Checking Duplicated Data**

```
[1] 0
```

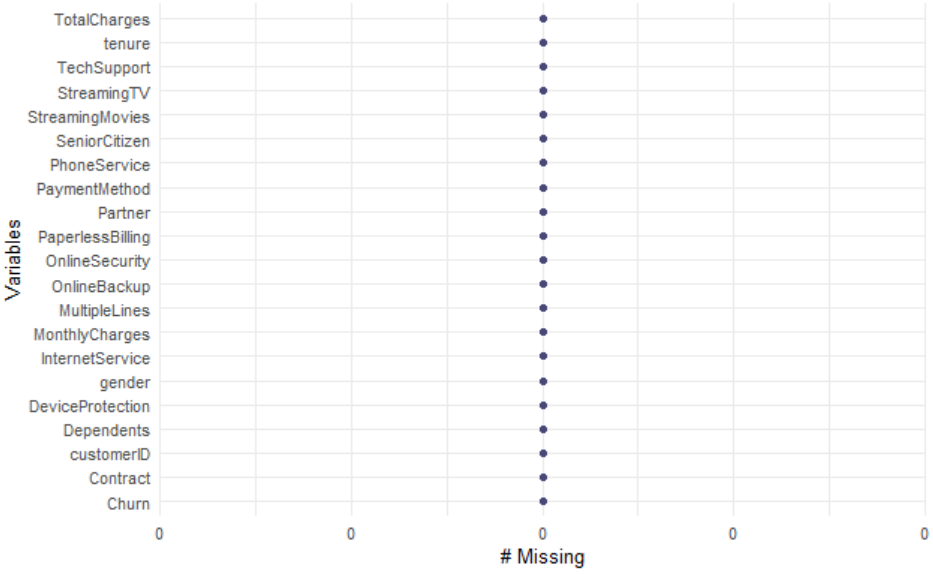
Data Insight :
There is no duplicated data in this dataset. Now let's check Missing Value.

- **Check Missing Values**


```
[1] 11
```

Data Insight :
- From this visualization we can conclude that there is a f missing values in coloum "TotalCharges" with number of missing values is 11

- **Dealing With Missing Values**
  Because there are so many missing values, we will clean them with drop rows because its only 11 row that missing, seems like small proportion from the dataset, here are the results:

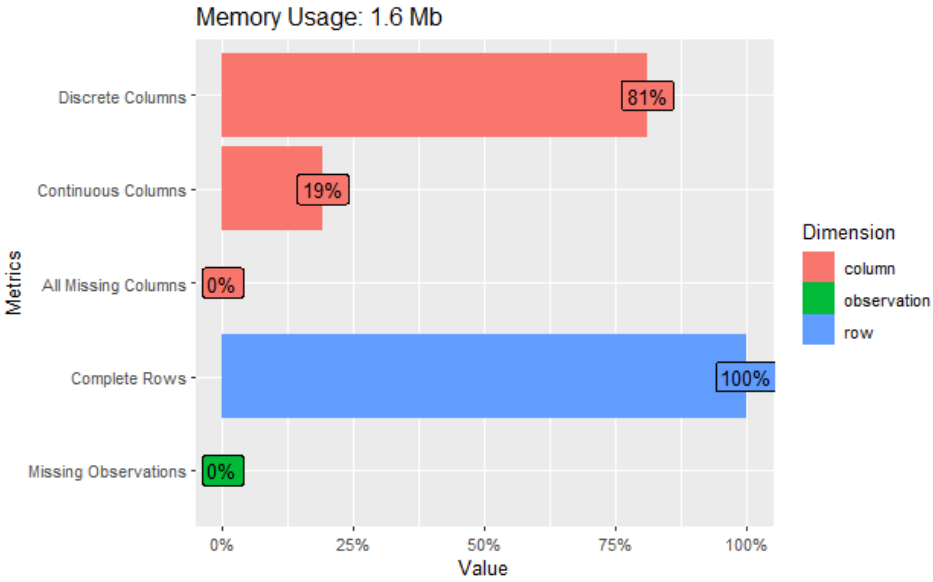



```
[1] 0
```

Data insight :

- After cleaning it can be seen from the results above that the missing values has been gone, then I will explore the dataset for futher anlysis with new dataset namely " data_clean"

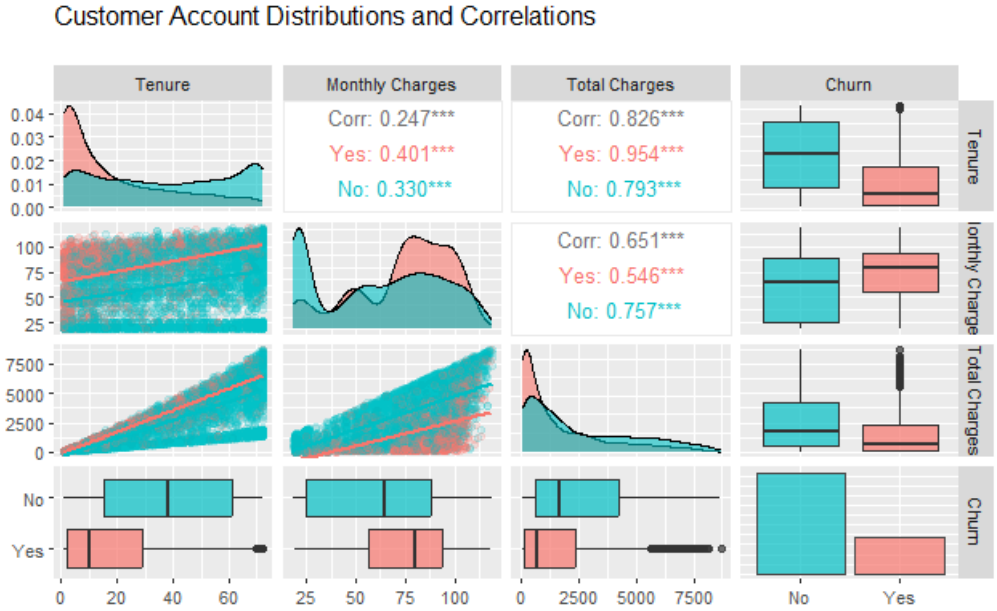
- **Outlier Check**



Data Insight:

- As u can see complete rows says its 100% so there is no outliers in this data set.

- Distribution and Corelations



Data Insight :

Based on our analysis of the numerical variables, we found a strong correlation between TotalCharges and customer tenure, particularly among customers who have left our service (Churn = Yes). This correlation coefficient is higher than 0.95, indicating a significant relationship. Additionally, we observed a small positive relationship between MonthlyCharges and tenure, with a correlation coefficient of 0.25, which is statistically significant.

The histogram plot of MonthlyCharges revealed a distinctive shape, suggesting the presence of multiple peaks or modes in the distribution. On the other hand, the distribution of customer tenure appeared relatively uniform among current customers, but it showed a right-skewed pattern among those who have churned or left our service. This implies that there is a difference in tenure distribution between current and churned customers.

- **Prop table**

```
                          No        Yes
No                 0.58221333 0.41778667
No internet service 0.92565789 0.07434211
Yes                0.85359801 0.14640199
```

Data Insight :

- The proportion of customers without dependents who did not churn (No/No) is approximately 0.687 or 68.7%.
- The proportion of customers without dependents who churned (No/Yes) is approximately 0.313 or 31.3%.
- The proportion of customers with dependents who did not churn (Yes/No) is approximately 0.845 or 84.5%.
- The proportion of customers with dependents who churned (Yes/Yes) is approximately 0.155 or 15.5%.
- Based on this analysis, it appears that customers without dependents have a higher churn rate (31.3%) compared to customers with dependents (15.5%). This information

can be useful for understanding the relationship between the Dependents variable and customer churn.

	No	Yes
No	0.58221333	0.41778667
No internet service	0.92565789	0.07434211
Yes	0.85359801	0.14640199

Data Insight:

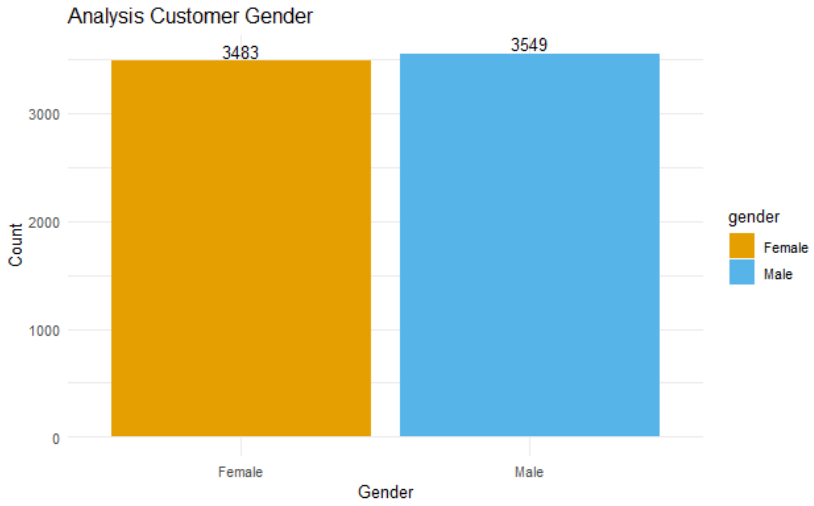
- The proportion of customers without online security who did not churn (No/No) is approximately 0.582 or 58.2%.
- The proportion of customers without online security who churned (No/Yes) is approximately 0.418 or 41.8%.
- The proportion of customers with no internet service who did not churn (No internet service/No) is approximately 0.926 or 92.6%.
- The proportion of customers with no internet service who churned (No internet service/Yes) is approximately 0.074 or 7.4%.
- The proportion of customers with online security who did not churn (Yes/No) is approximately 0.854 or 85.4%.
- The proportion of customers with online security who churned (Yes/Yes) is approximately 0.146 or 14.6%.
- Based on this analysis, it appears that customers without online security or with no internet service have higher churn rates compared to customers with online security. This information can be valuable for understanding the relationship between the OnlineSecurity variable and customer churn.

	No	Yes
DSL	0.8104089	0.1895911
Fiber optic	0.5810724	0.4189276
No	0.9259502	0.0740498

Data Insight :

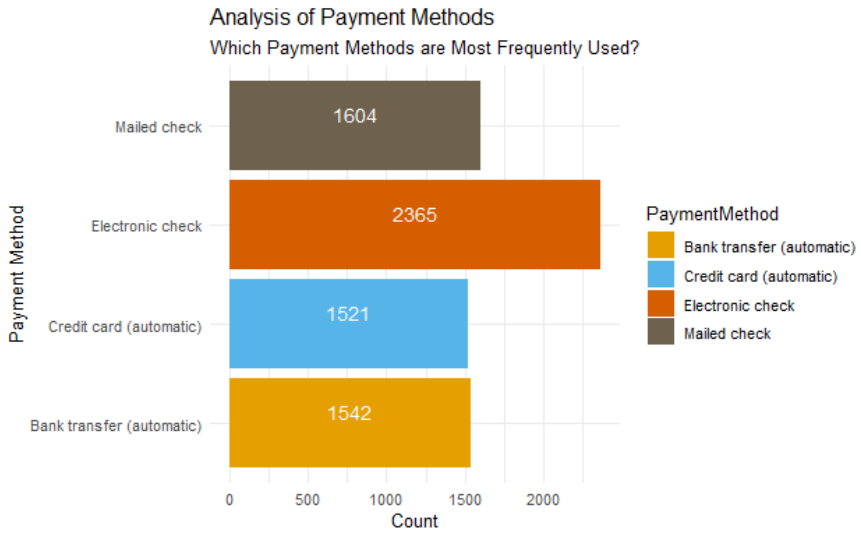
- The proportion of customers with DSL internet service who did not churn (DSL/No) is approximately 0.810 or 81.0%.
- The proportion of customers with DSL internet service who churned (DSL/Yes) is approximately 0.190 or 19.0%.
- The proportion of customers with fiber optic internet service who did not churn (Fiber optic/No) is approximately 0.581 or 58.1%.
- The proportion of customers with fiber optic internet service who churned (Fiber optic/Yes) is approximately 0.419 or 41.9%.
- The proportion of customers with no internet service who did not churn (No/No) is approximately 0.926 or 92.6%.
- The proportion of customers with no internet service who churned (No/Yes) is approximately 0.074 or 7.4%.
- Based on this analysis, it appears that customers with fiber optic internet service have a higher churn rate (41.9%) compared to customers with DSL internet service (19.0%) or no internet service (7.4%). This information can be insightful for understanding the relationship between the InternetService variable and customer churn.

- **Visualization**



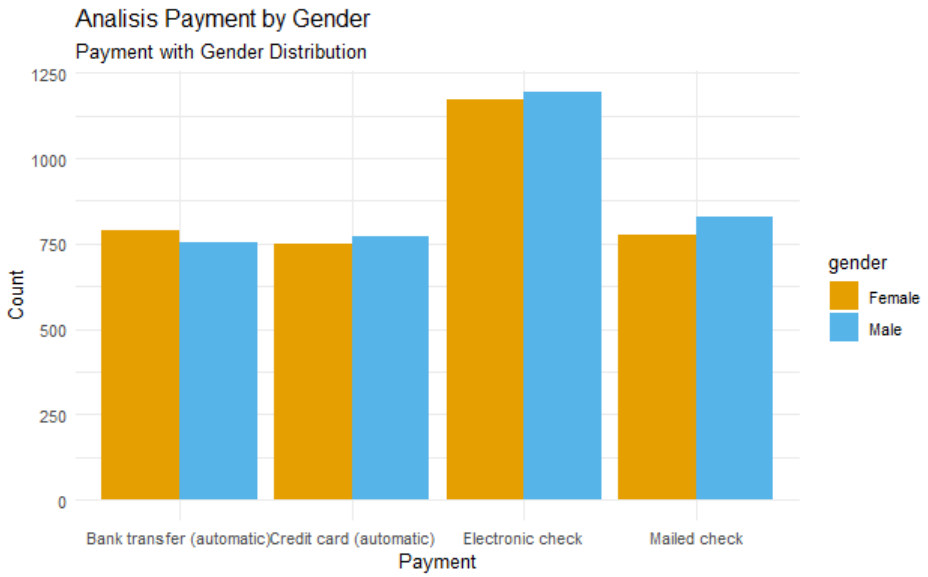
Data Insight :

As we can most customer gender is male with values 3555



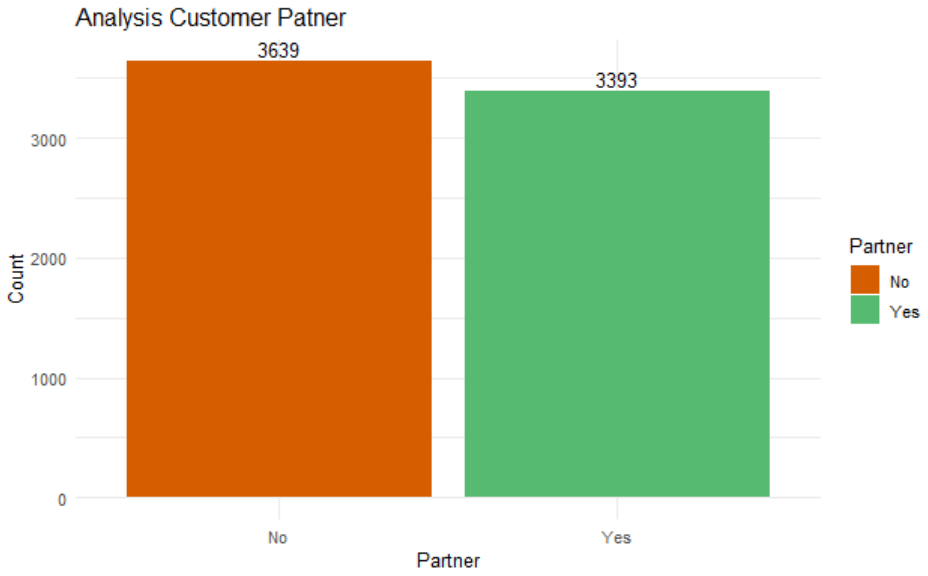
Data Insight :

As we can see most frequently payment method is electronic check then mailed check

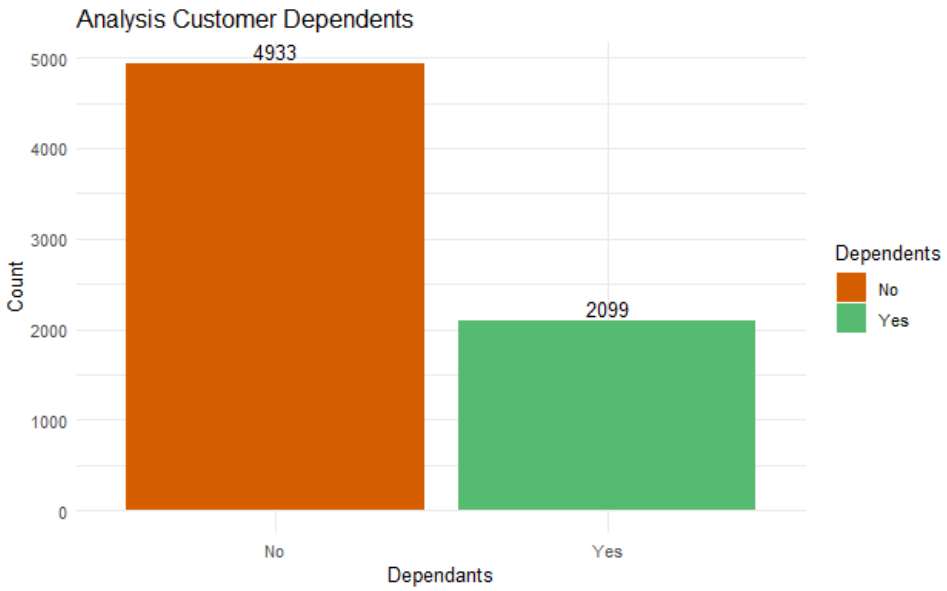


	Bank transfer (automatic)	Credit card (automatic)	Electronic check	Mailed check
Female	787	751	1170	775
Male	755	770	1195	829

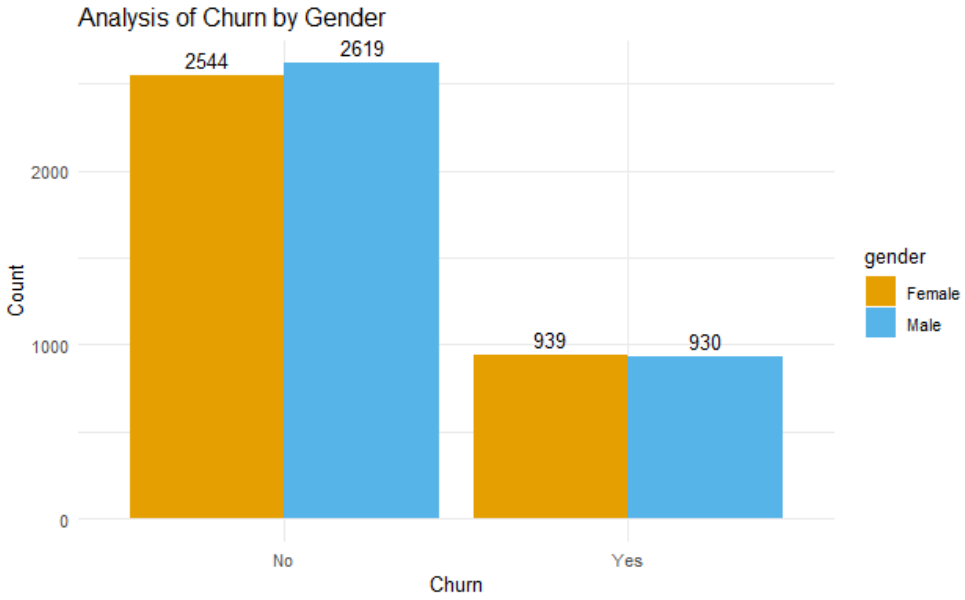
Data Insight :
As we can see most frequently each payment method by gender



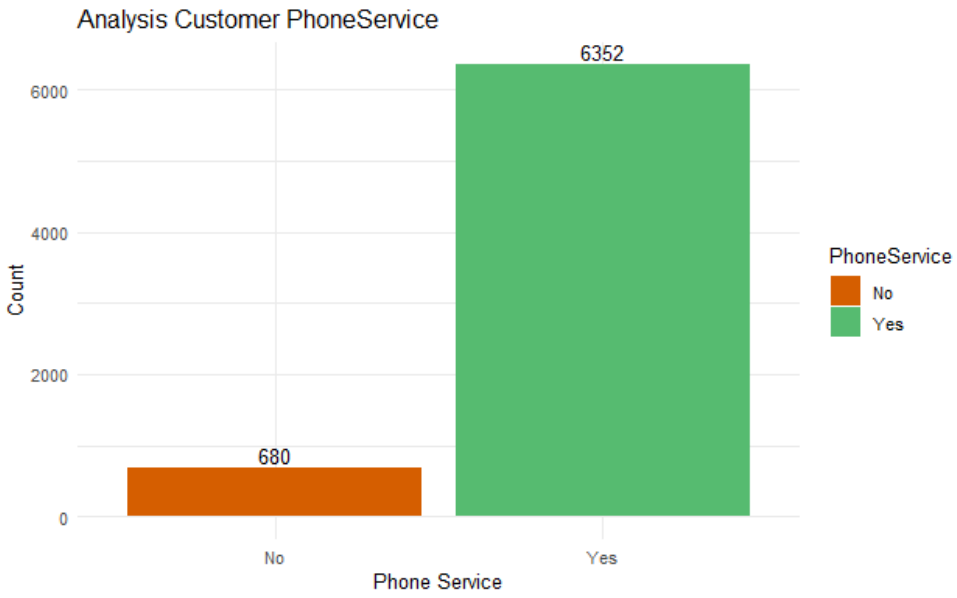
Data Insight :
It can be seen from the results that customers who do not have a partner tend to be more numerous but only have a slight difference with those who have a partner.



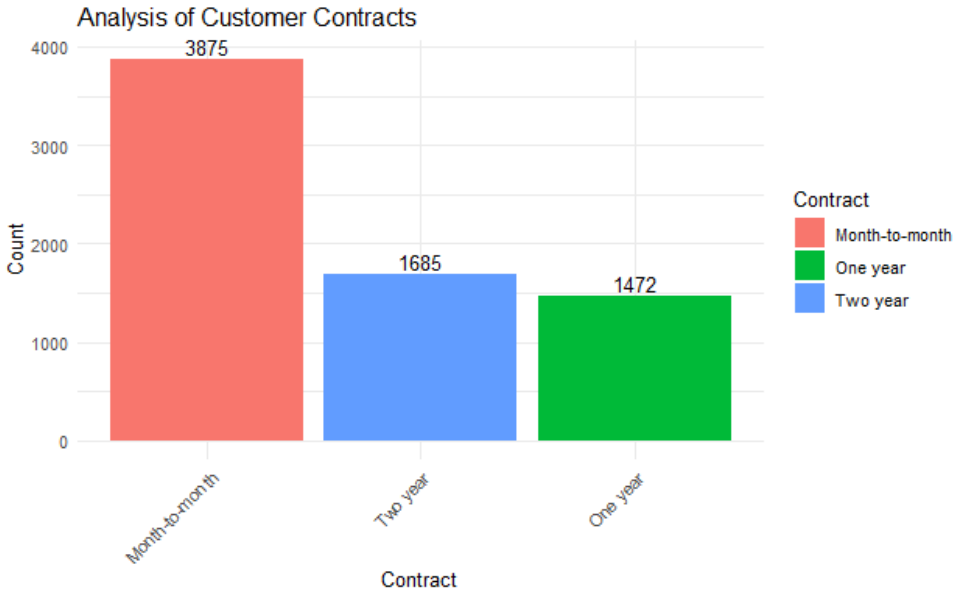
Data Insight :
Most customer don’t have a dependent only a half of them have a dependents.



Data Insight :
Most customer doesn't have a churn only few that have churn



Data Insight :
Most customer have a phone service before



Data Insight :
Most customer have contract month to moth, the second most customer contract month is two year, lastly is one year contract

3. Predictive Modeling for Customer Churn

- **Random Forest Model**
  I used random forest to model the following results :

```
Train Accuracy: 0.9648
Train Error: 0.0352
Test Accuracy: 0.7974414
Test Error: 0.2025586
Precision: 0.6394984
Recall: 0.5454545
F1 Score: 0.5887446
```

Insight :
Based on the output of the random forest modeling with the evaluation metrics, we can draw the following conclusions:

- **Train Accuracy: 0.9648**
  The model achieved a high accuracy of 96.48% on the training dataset, indicating that it is able to predict the churn behavior of customers with a high level of accuracy.

- **Train Error: 0.0352**
  The train error rate of 3.52% indicates that the model has a low rate of misclassification on the training dataset.

- **Test Accuracy: 0.7974414**
  The model achieved an accuracy of 79.74% on the test dataset. This suggests that the model performs reasonably well on unseen data, although the accuracy is slightly lower compared to the training accuracy.

- **Test Error: 0.2025586**
  The test error rate of 20.26% indicates that the model misclassifies approximately 20.26% of the churn cases in the test dataset.

- **Precision: 0.6394984**
  Precision represents the proportion of correctly predicted positive cases (churn) out of all predicted positive cases. In this case, the precision of 63.95% indicates that when the model predicts a customer will churn, it is correct about 63.95% of the time.

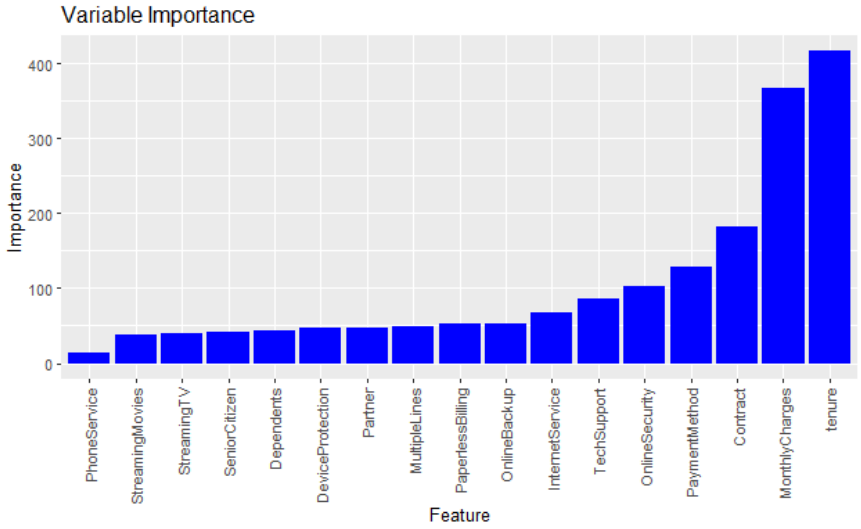
- **Recall: 0.5454545**
  Recall (also known as sensitivity or true positive rate) represents the proportion of correctly predicted positive cases (churn) out of all actual positive cases. A recall of 54.55% suggests that the model is able to identify approximately 54.55% of the customers who actually churn.

- **F1 Score: 0.5887446**
  The F1 score is the harmonic mean of precision and recall. It provides a single measure that combines both precision and recall. The F1 score of 0.5887446 indicates a moderate balance between precision and recall.

Conclusion :

Overall, the random forest model shows good performance on the training dataset, but there is some drop in accuracy and performance on the test dataset. This could suggest some level of overfitting. Improvements can be made by fine-tuning the model parameters or considering other modeling techniques to further enhance the predictive accuracy and minimize errors.

- **Random Forest Features**



From the feature importance results obtained from the random forest modeling, we can analyze and draw the following conclusions:

1. Tenure
The "tenure" feature has the highest importance score of 417.26872. This suggests that the duration of the customer's tenure with the company plays a significant role in predicting customer churn. Customers with longer tenure may have a lower probability of churning compared to those with shorter tenure.

2. MonthlyCharges
The "MonthlyCharges" feature has a high importance score of 366.42337. It indicates that the monthly charges incurred by customers are an influential factor in predicting churn. Higher monthly charges may potentially lead to a higher likelihood of churn.

3. Contract
The "Contract" feature has an importance score of 182.46051. This indicates that the type of contract a customer has (e.g., month-to-month, one-year, or two-year) is an important predictor of churn. Customers with shorter-term contracts, particularly month-to-month contracts, might have a higher probability of churning.

4. PaymentMethod
The "PaymentMethod" feature has an importance score of 128.05098. It suggests that the payment method chosen by customers could be a relevant factor in predicting churn. Different payment methods might have varying effects on customer loyalty and churn behavior.

5. OnlineSecurity, TechSupport, InternetService, OnlineBackup, PaperlessBilling, MultipleLines and more:
These features have moderate importance scores ranging from 66.87870 to 48.69807. They represent various aspects of the company's services, such as online security, technical support, internet service type, online backup, paperless billing, and multiple lines. These features

contribute to the overall prediction of churn, albeit to a slightly lesser extent than the previously mentioned features.

Conclusion :

The random forest model highlights several important features that strongly influence the prediction of customer churn. Tenure, monthly charges, contract type, payment method, and certain service-related features are key factors to consider when analyzing and addressing customer churn. Companies can focus their efforts on improving customer retention strategies by paying attention to these influential features and taking appropriate actions based on their impact on churn prediction.