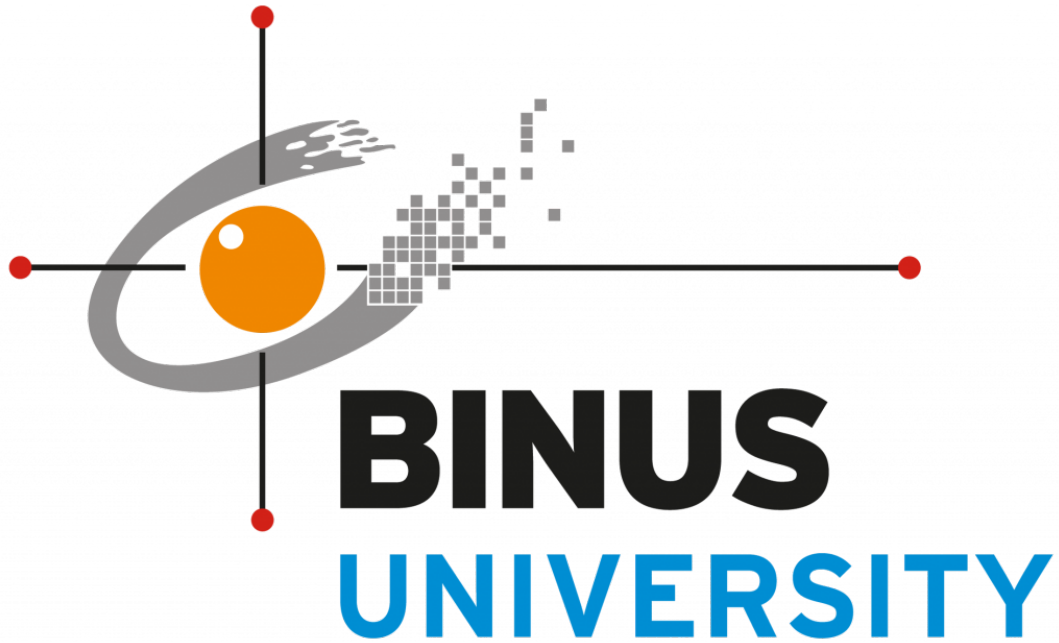
DATA MINING AND VISUALIZATION
LAB ASSIGNMENT



Created by
Eleanor Maritsa Maharani
2502014210
BA09

REPORT US COUNTIES: COVID19 + WEATHER + SOCIO/HEALTH

In US Counties: COVID19 + Weather + Socio/Health dataset, we have 3 file csv, 'US_counties_COVID19_health_weather_data.csv', 'us_county_geometry.csv', and 'us_county_sociohealth_data.csv'.

date <date>	county <chr>	state <chr>	fips <chr>	cases <dbl>	deaths <dbl>	stay_at_home_announced <chr>	stay_at_home_effective <chr>
2020-01-21	Snohomish	Washington	53061	1	0	no	no
2020-01-22	Snohomish	Washington	53061	1	0	no	no
2020-01-23	Snohomish	Washington	53061	1	0	no	no
2020-01-24	Cook	Illinois	17031	1	0	no	no
2020-01-24	Snohomish	Washington	53061	1	0	no	no
2020-01-25	Orange	California	06059	1	0	no	no

6 rows | 1-8 of 227 columns

The total rows of 'US_counties_COVID19_health_weather_data.csv' are 790331 and the total columns of 'US_counties_COVID19_health_weather_data.csv' are 227.

state <chr>	county <chr>	fips <chr>
ALABAMA	Autauga	01001
ALABAMA	Blount	01009
ALABAMA	Chambers	01017
ALABAMA	Coffee	01031
ALABAMA	Colbert	01033
ALABAMA	Covington	01039

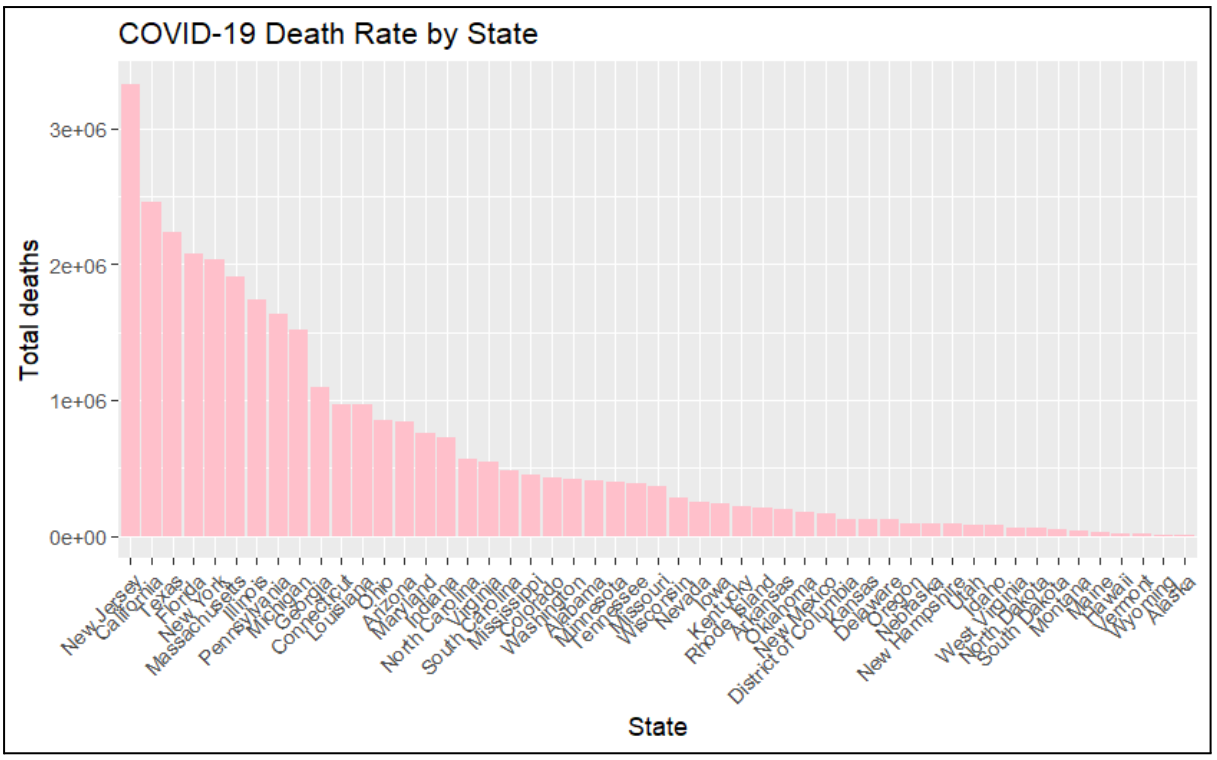
6 rows | 1-3 of 7 columns

The total rows of 'us_county_geometry.csv' are 3124 and the total columns of 'us_county_geometry.csv' are 7.

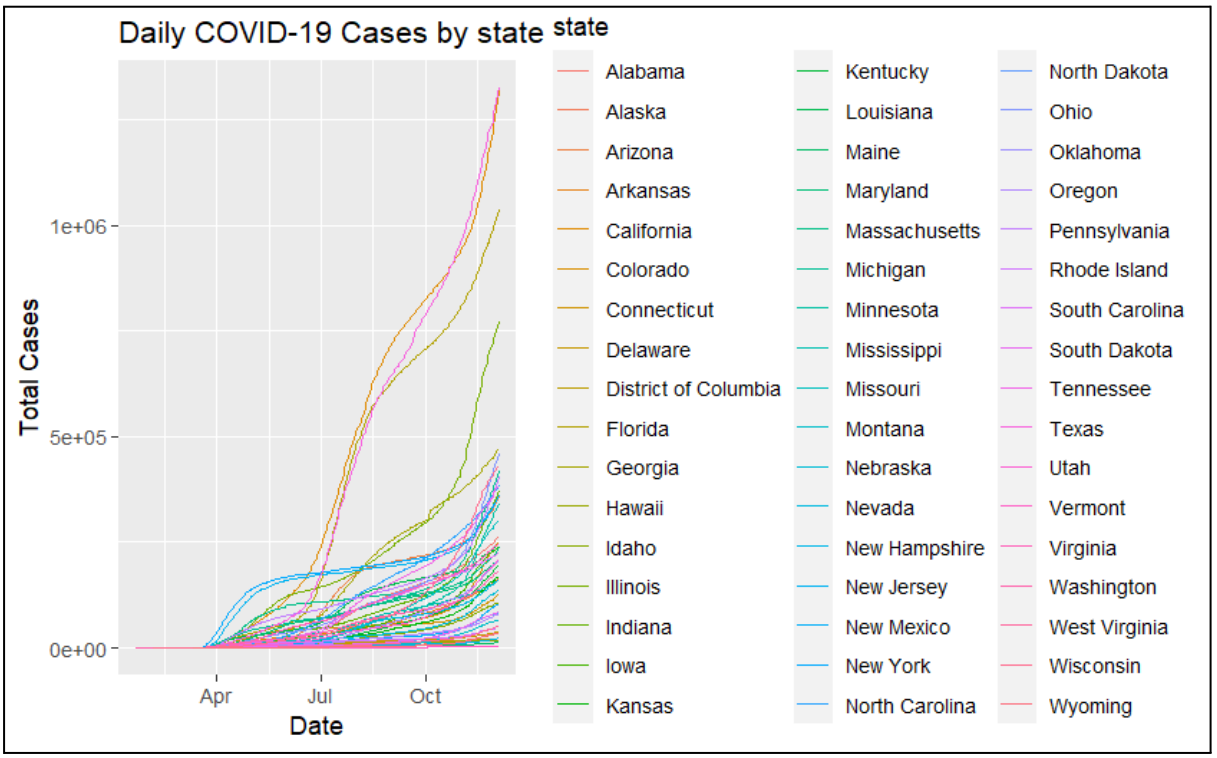
fips <chr>	state <chr>	county <chr>	lat <dbl>	lon <dbl>	total_population <dbl>	area_sqmi <dbl>	population_density_per_sqmi <dbl>
01001	Alabama	Autauga	32.53493	-86.64275	55049	594.4461	92.60553
01003	Alabama	Baldwin	30.72749	-87.72258	199510	1589.8074	125.49319
01005	Alabama	Barbour	31.86959	-85.39321	26614	884.8758	30.07654
01007	Alabama	Bibb	32.99863	-87.12648	22572	622.5824	36.25544
01009	Alabama	Blount	33.98088	-86.56738	57704	644.8065	89.49041
01011	Alabama	Bullock	32.10053	-85.71569	10552	622.8054	16.94269

6 rows | 1-8 of 181 columns

The total rows of 'us_county_sociohealth_data.csv' are 3144 and the total columns of 'us_county_sociohealth_data.csv' are 181.

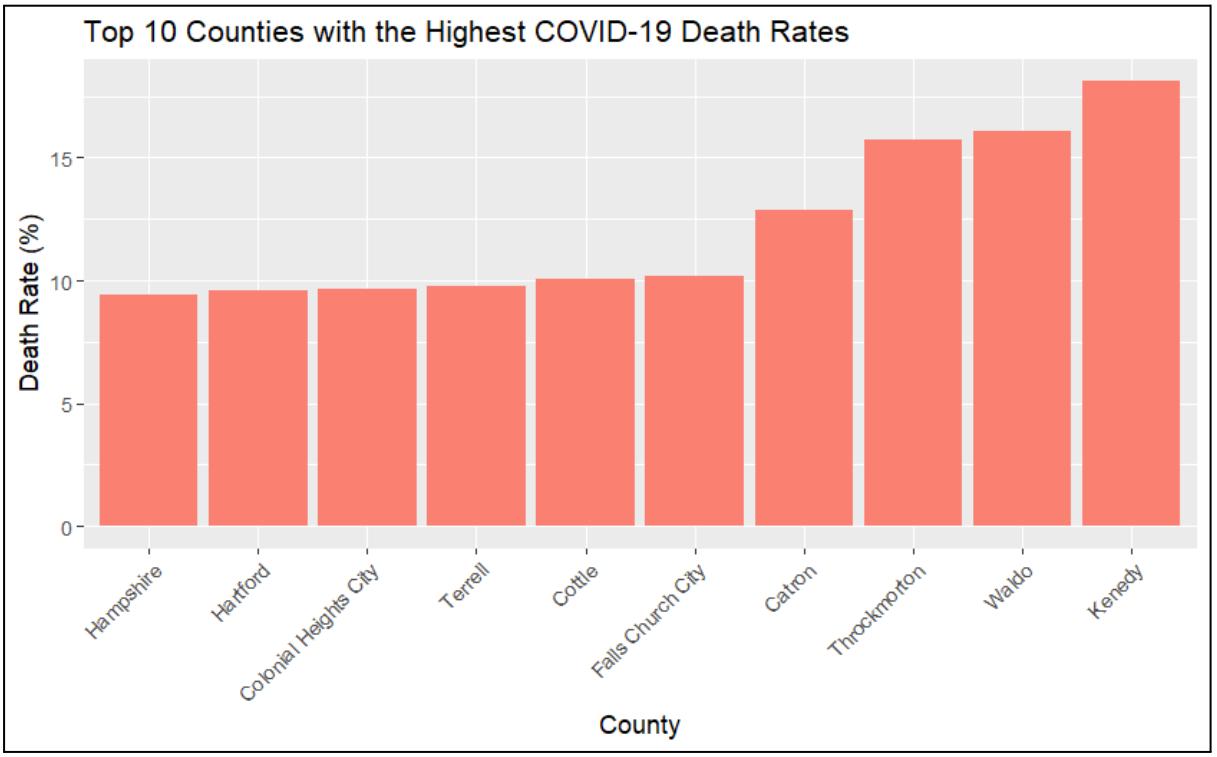


From the plot above, the state that has the highest death rate by state is New Jersey, then the second is California, and the third is Texas.

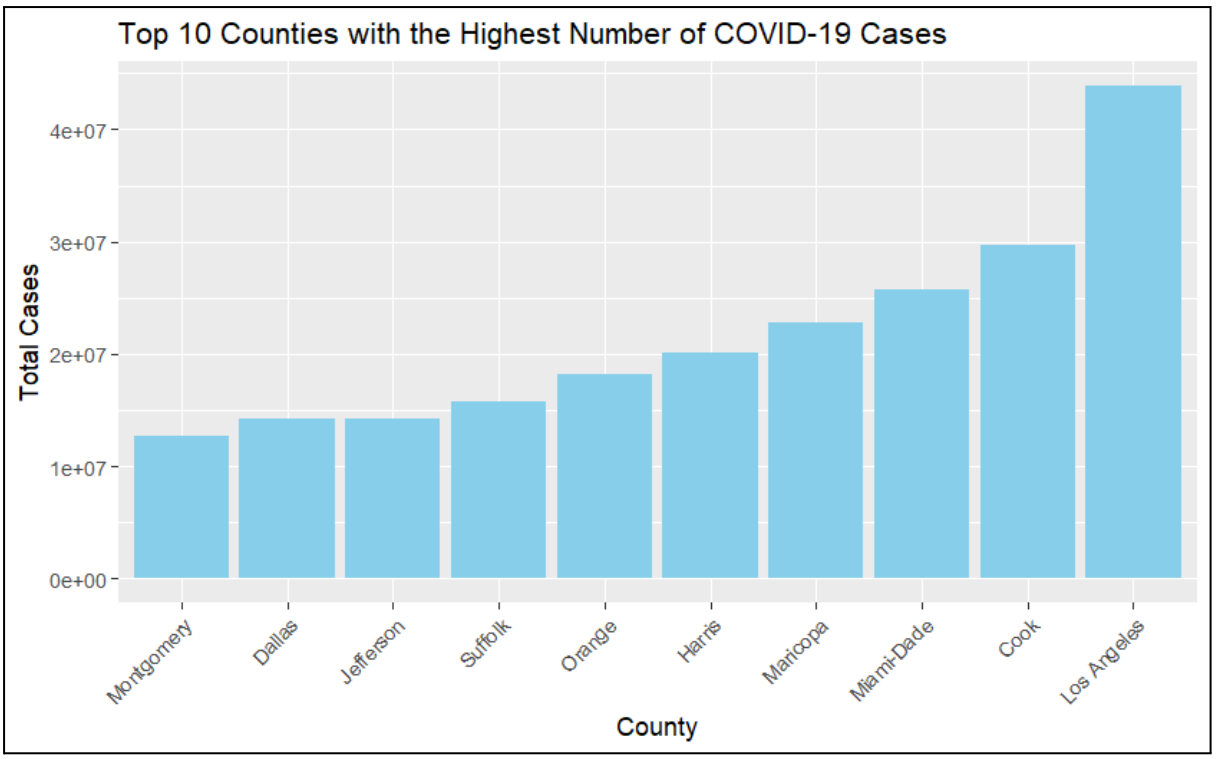


The plot showcases the daily COVID-19 cases by state over a period of time. Each state is represented by a distinct color line, allowing for easy differentiation. The x-axis displays the dates, while the y-axis represents the total number of cases reported for each respective date.

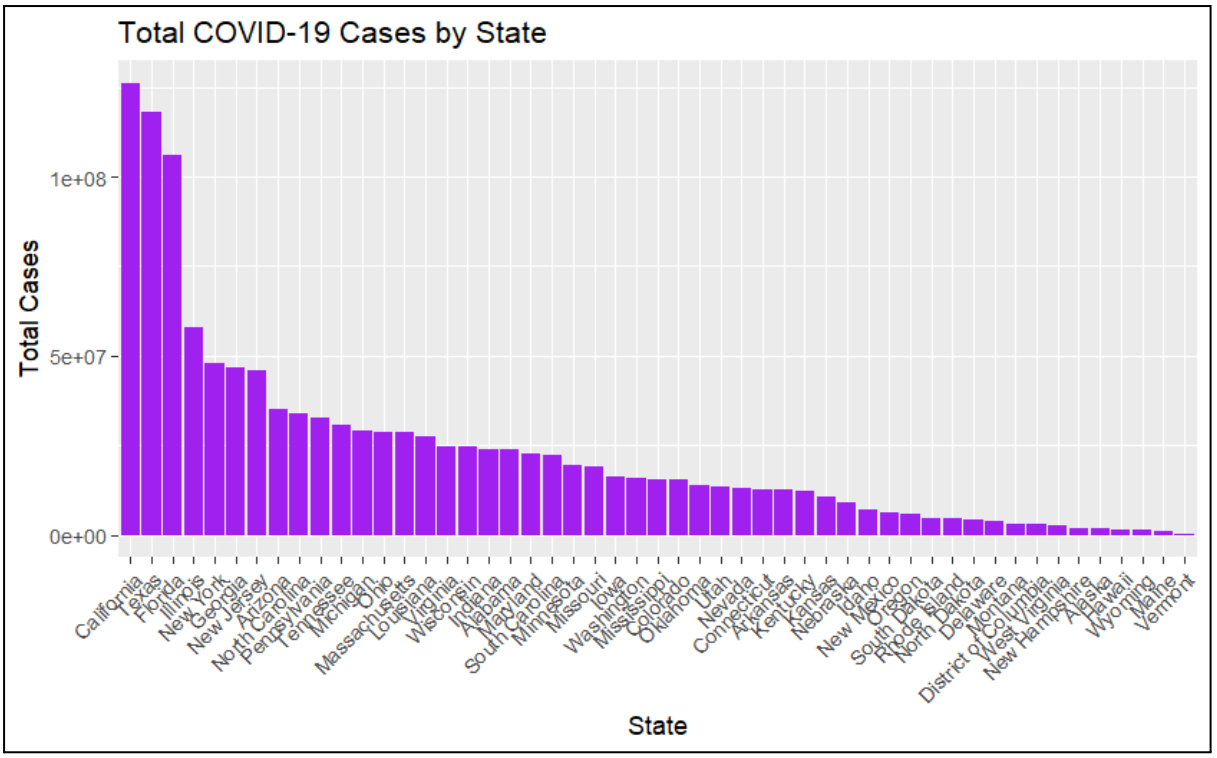
The plot is titled "Daily COVID-19 Cases by State," indicating its focus on illustrating the daily changes in cases across various states. The x-axis is labeled as "Date," signifying the progression of time, and the y-axis is labeled as "Total Cases," denoting the cumulative number of reported COVID-19 cases.



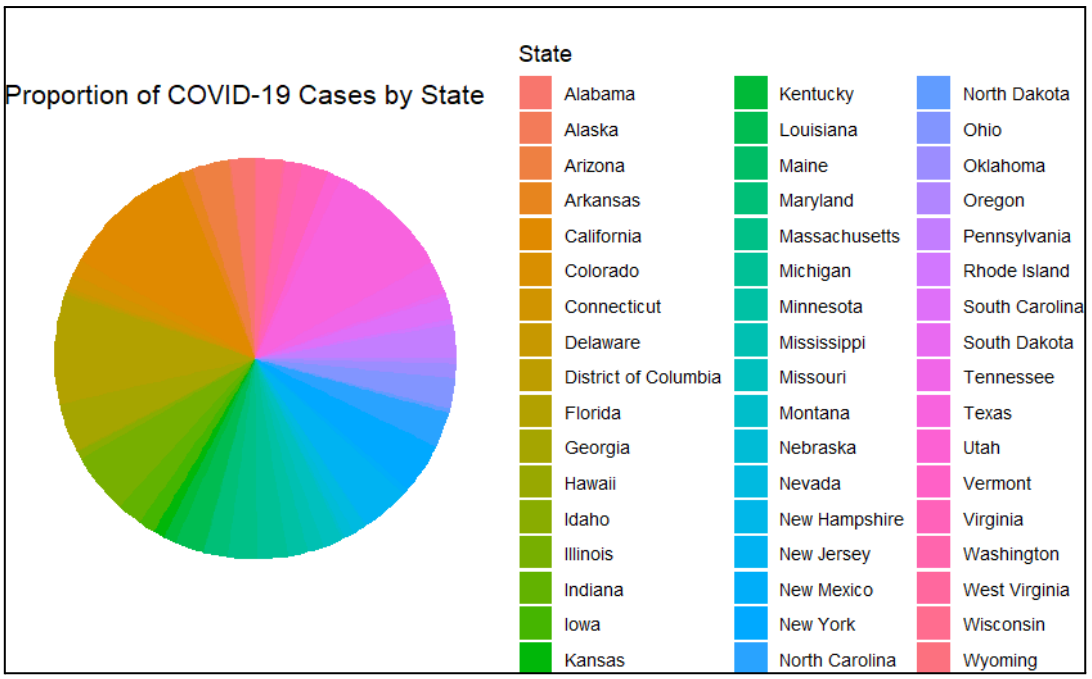
From the plot above, the state that has the highest death rate by county is Kenedy, then the second is Waldo, and the third is Throckmorton.



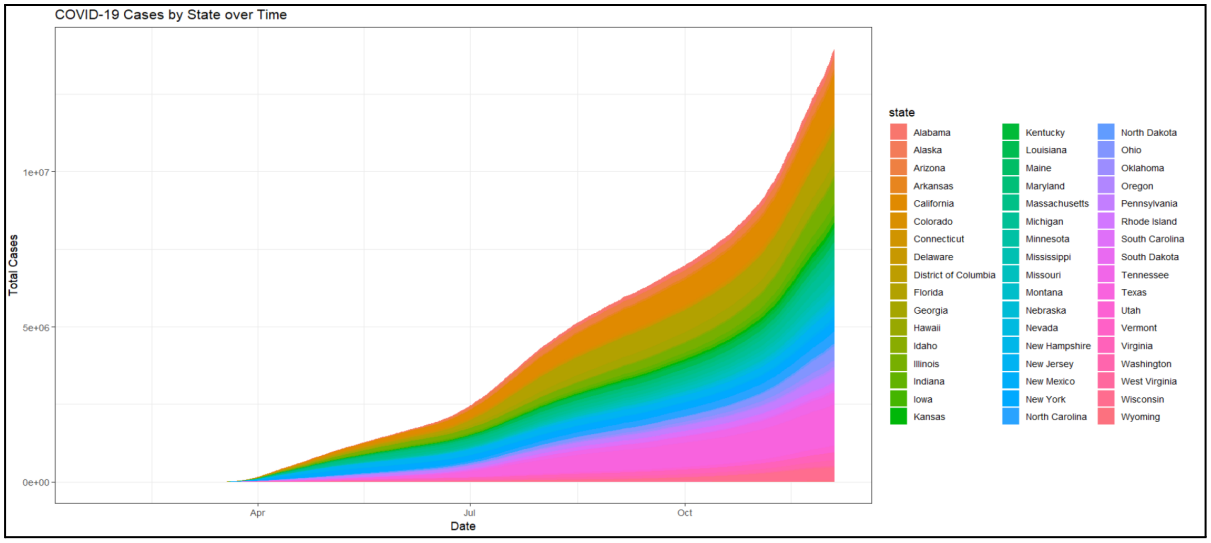
From the plot above, the county that has the highest number of COVID-19 cases is Los Angeles, then the second is Cook, and the third is Miami-Dade.



From the plot above, the state that has the highest number of COVID-19 cases by county is California, then the second is Texas, and the third is Florida.



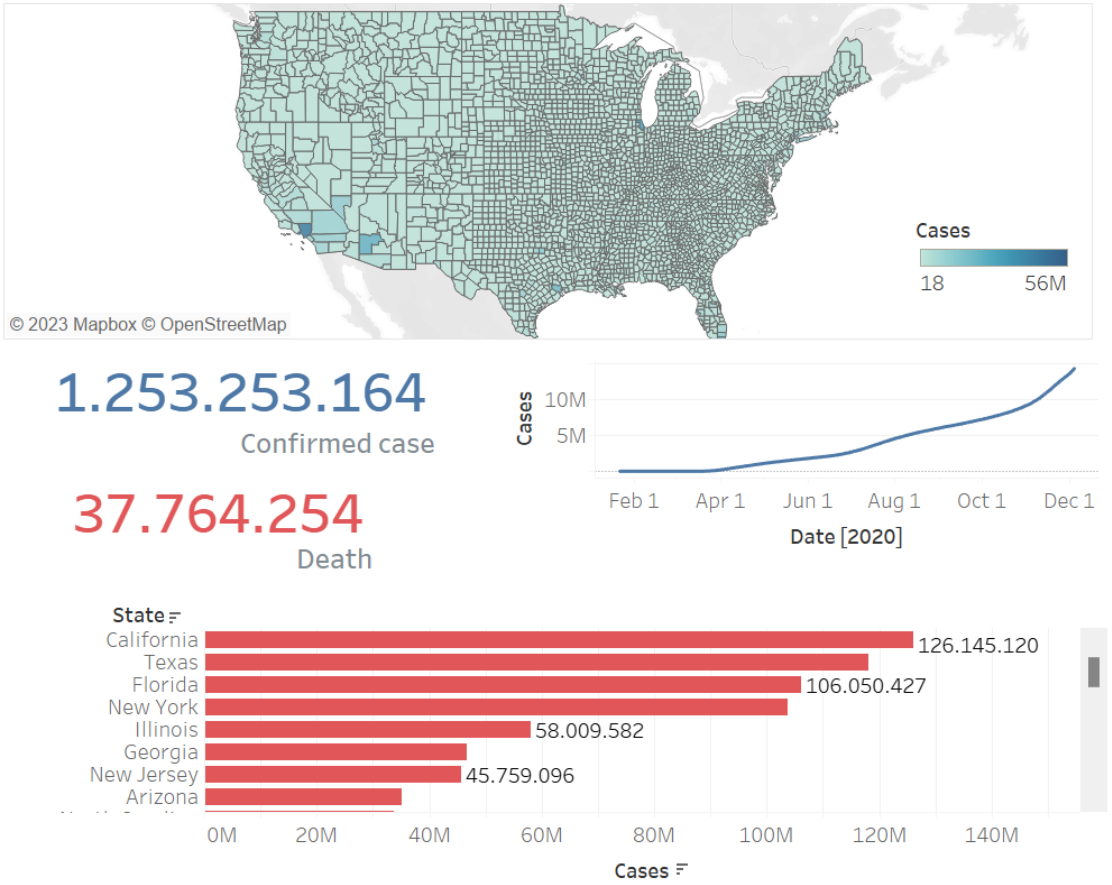
The plot illustrates the proportion of COVID-19 cases by state in a visually striking manner. Each state is represented by a segment of a circular bar, with the size of the segment indicating the proportion of total COVID-19 cases attributed to that particular state. Each state is assigned a distinct color to differentiate them visually, and the legend is positioned on the right side of the plot. The legend connects the color to the corresponding state, enabling easy identification.



The plot visualizes the COVID-19 cases by state over time using an area plot. Each state is represented by a filled area, with the height of the area representing the total number of cases reported for that state on a specific date. The plot is titled "COVID-19 Cases by State over Time," indicating its focus on illustrating the progression of cases across different states. The x-axis represents the dates, while the y-axis represents the total number of cases. The area plot provides a cumulative representation of cases over time, showcasing the growth and

cumulative totals for each state. Each state is distinguished by a unique fill color, allowing for easy identification. The legend, positioned on the right side of the plot, associates each color with its respective state.

Here is the tableau visualization:



REPORT TELCO CUSTOMER CHURN

In Telco customer dataset includes information about:

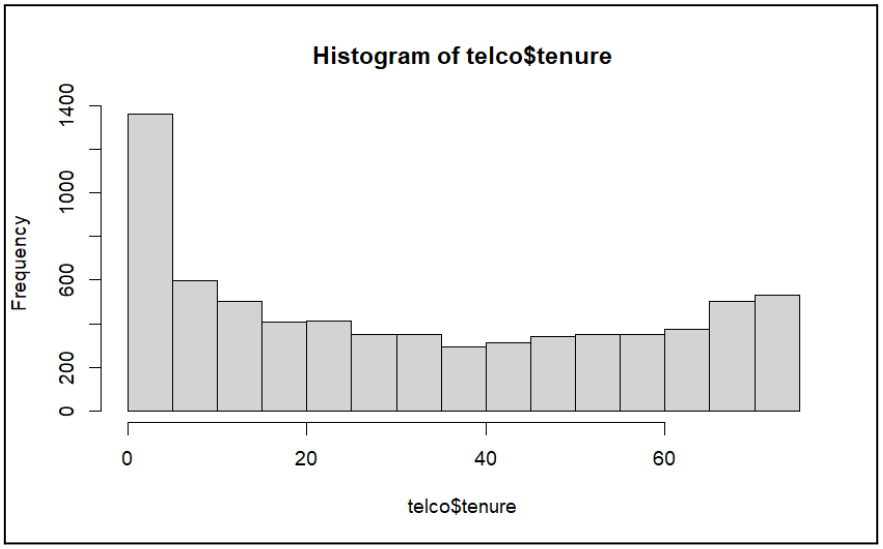
- Customers who left within the last month – the column is called Churn
- Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
- Customer account information – how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges
- Demographic info about customers – gender, age range, and if they have partners and dependents

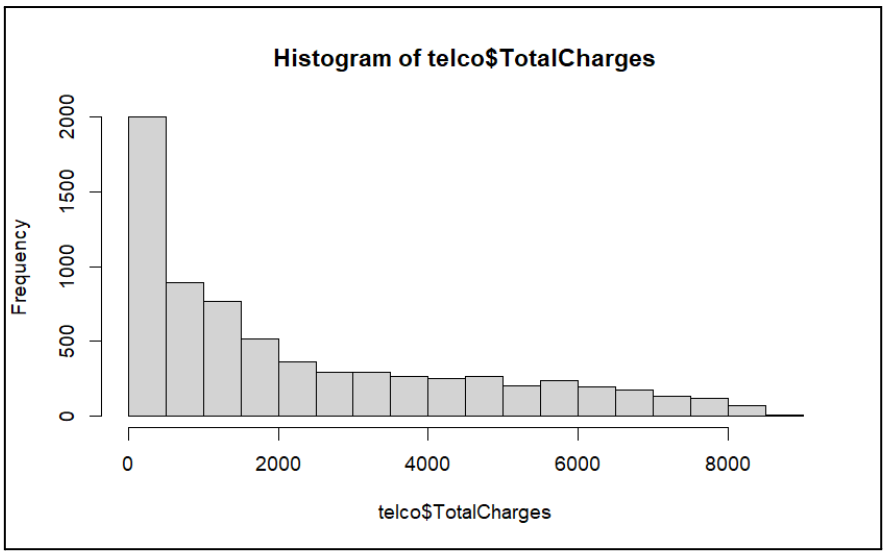
customerID <chr>	gender <chr>	SeniorCitizen <dbl>	Partner <chr>	Dependents <chr>	tenure <dbl>	PhoneService <chr>
7590-VHVEG	Female	0	Yes	No	1	No
5575-GNVDE	Male	0	No	No	34	Yes
3668-QPYBK	Male	0	No	No	2	Yes
7795-CFOCW	Male	0	No	No	45	No
9237-HQITU	Female	0	No	No	2	Yes
9305-CDSKC	Female	0	No	No	8	Yes

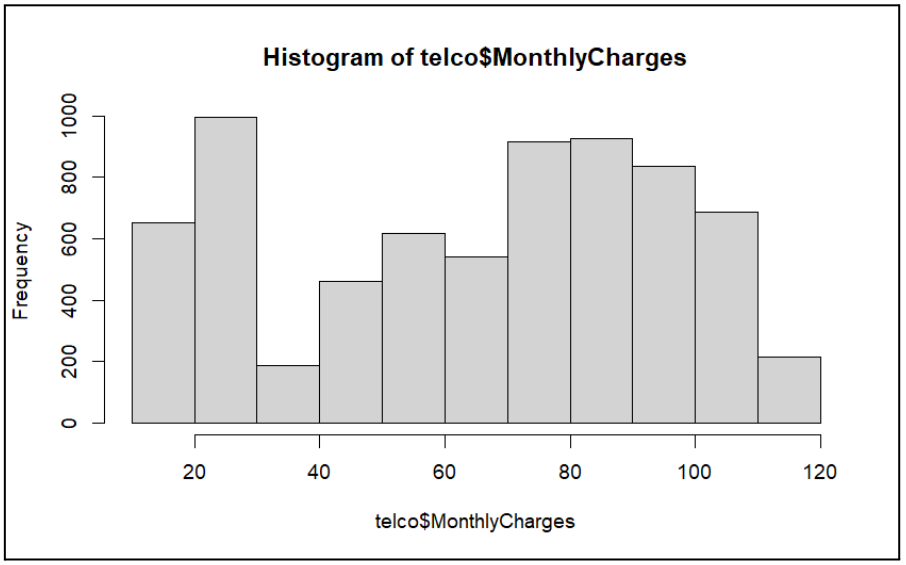
6 rows | 1-7 of 21 columns

The total rows of this dataset are 7043 and the total columns of this dataset are 21, but the telco dataset has 11 missing values, so I'm using the na.omit() function to remove missing values in telco. The telco data set after removing missing values has 7032 rows.

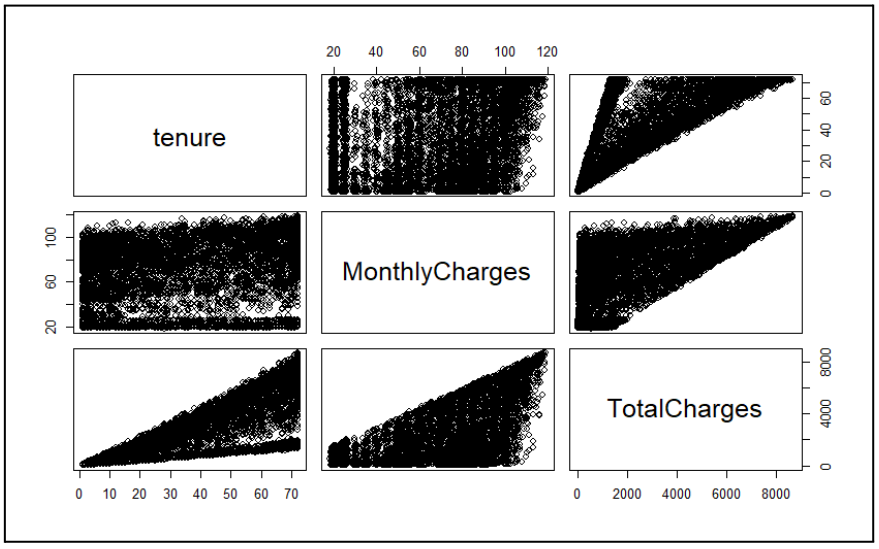
I would utilize the histogram plot for the three numerical variables (tenure, total_charges, and monthly_charges) to look for outliers. No outliers are seen in the 3 plots below.

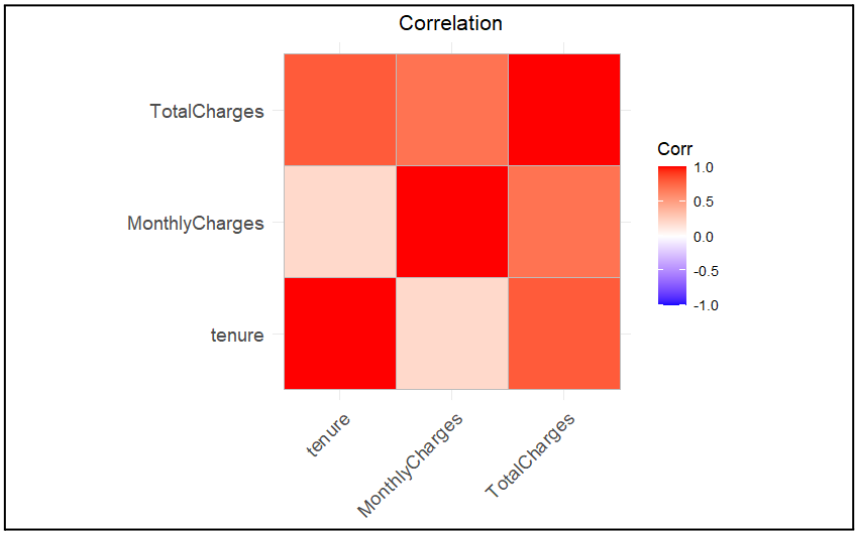




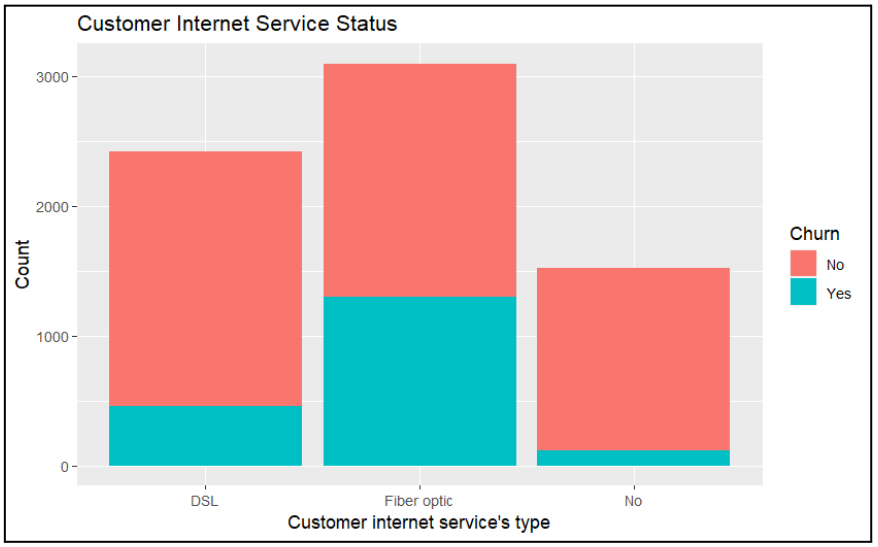


Next, i want to see correlation between tenure, monthlyCharges, and Total Charges

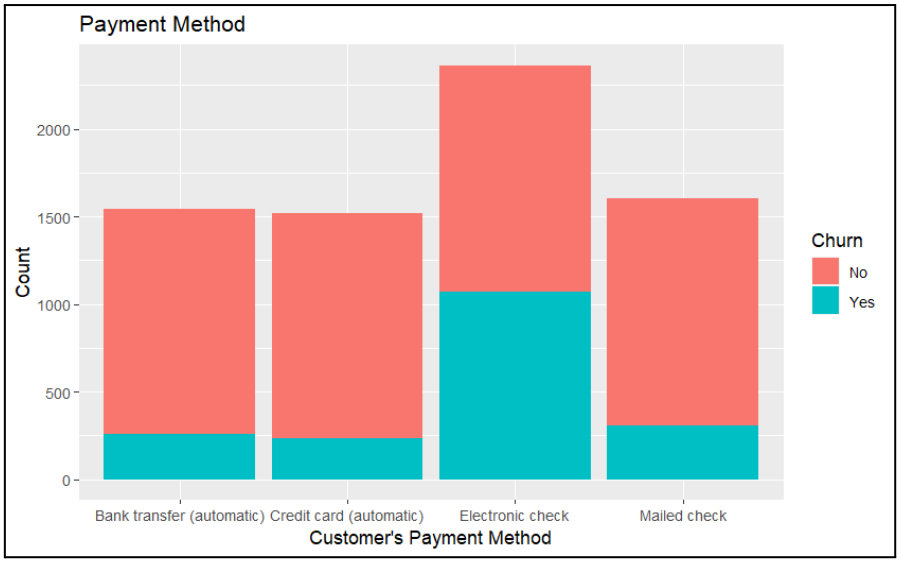




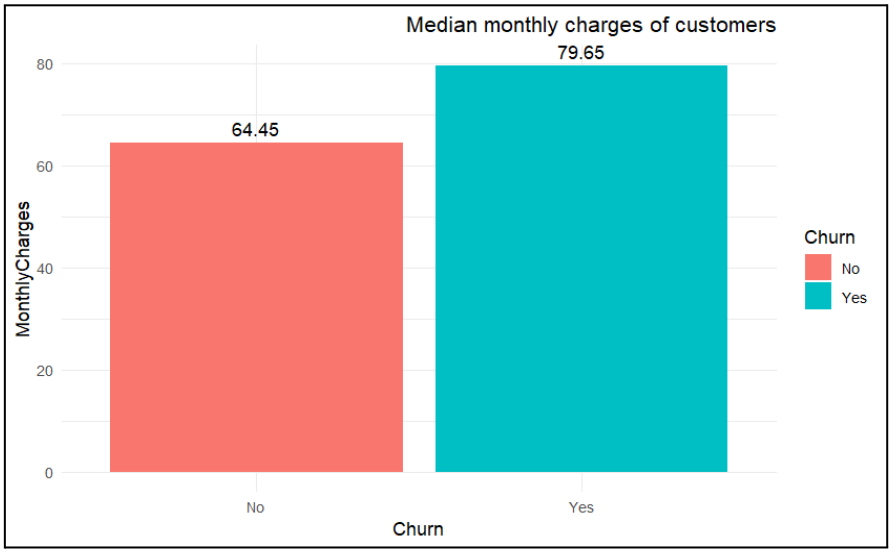
From the graph above the Total Charge has a positive correlation with the Monthly Charge.



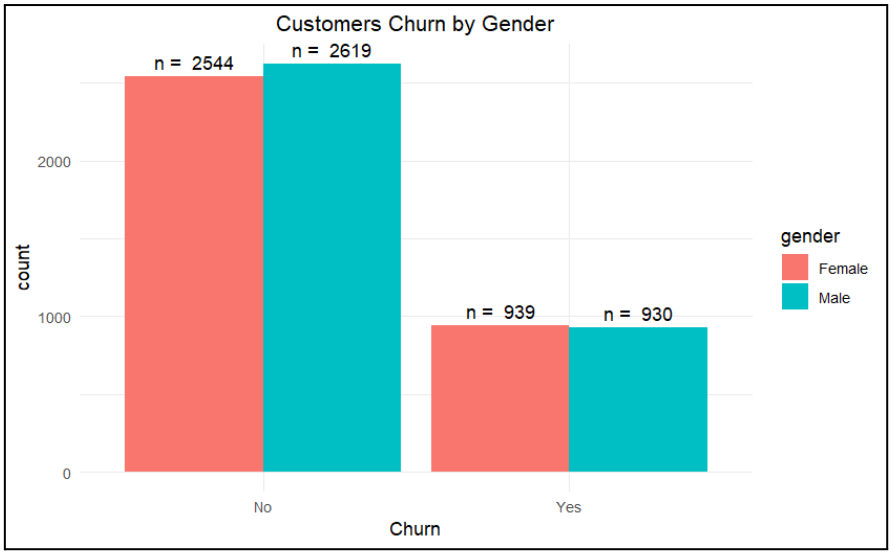
It's fascinating to note that out of all the customers, those that have fiber-optic internet access at home are much more likely to churn.



From the graph above we can see that customers mostly use electronic check as their payment method, electronic check in this graph is 1071, while the least used payment method is credit card which is 232.



The average monthly cost for churning customers is higher than the cost for non-churning customers by about 15 USD/Month.



Both for male and female customers, an approximation of the number of turnover is displayed. Customer females with churn are 939 and customer male with churn are 939.

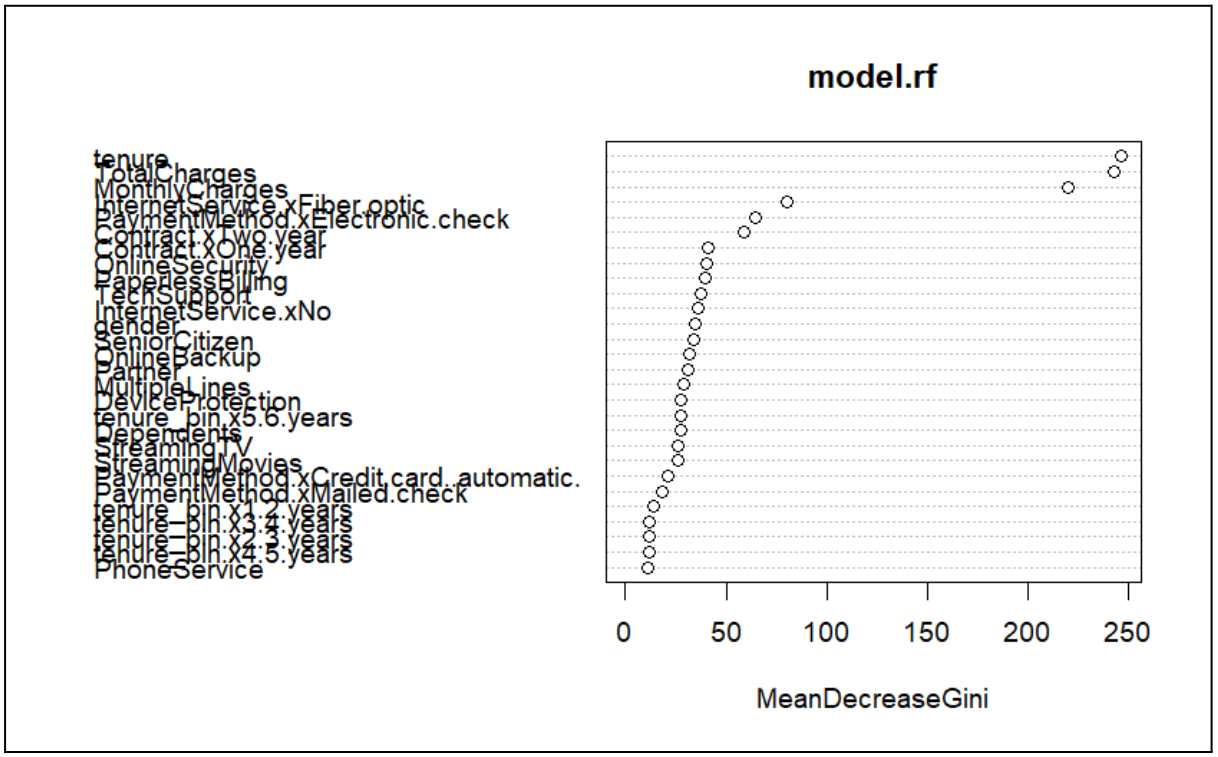
The next step is that I will create a predictive model, here I use the random forest model. The Random Forest selects random samples from a dataset to train each tree in the forest using the bootstrapped aggregation technique. An accumulation of individual tree predictions makes up the final forecast in a RandomForest. One of RandomForest's benefits is that it provides out-of-bag (OOB) error estimates, or the mean prediction error on a training sample, using the trees whose bootstrap samples do not include that training sample. It might function as a cross-validation mistake, obviating the necessity for test/validation data and boosting training data.

```
Call:
 randomForest(formula = Churn ~ ., data = train, proximity = FALSE,      importance = FALSE, ntree = 500, mtry = 4, do.trace =
FALSE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 4

        OOB estimate of  error rate: 19.59%
Confusion matrix:
     0   1 class.error
0 3258 356  0.09850581
1  608 700  0.46483180
```

The OOB error estimate comes to around 19.59%, so the model has around 80.41% out of sample accuracy for the training set. Let's check the prediction and accuracy on our validation data.

```
testPred    0    1
       0 1392  283
       1  157  278
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1392  157
         1  283  278

               Accuracy : 0.7915
                 95% CI : (0.7735, 0.8086)
    No Information Rate : 0.7938
    P-Value [Acc > NIR] : 0.6182

                  Kappa : 0.4246

 Mcnemar's Test P-Value : 2.536e-09

            Sensitivity : 0.8310
            Specificity : 0.6391
         Pos Pred Value : 0.8986
         Neg Pred Value : 0.4955
             Prevalence : 0.7938
         Detection Rate : 0.6597
   Detection Prevalence : 0.7341
      Balanced Accuracy : 0.7351

       'Positive' Class : 0
```

The basic RandomForest model gives an accuracy of 79.15%(almost close enough to the OOB estimate), Sensitivity 83.10% and Specificity 63.91%.



The plot above shows the variable importance plot. The importance values are typically calculated based on the decrease in Gini index or other impurity measures as each variable is included in the model. Higher values indicate greater importance, suggesting that the variable has a stronger impact on the model's predictions. In this case, the "Tenure" variable appears to have the highest importance. So we can conclude that the random forest model is one of the predictive models that is good enough to be used in this dataset because it achieved an accuracy value of 79.15%.