Lab 3: Predicting Telecom Churn with tidymodels

Boutelba Houssem

1. Import Library & data

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.0     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(tidymodels)
```

```
## -- Attaching packages --------------------------------------- tidymodels 1.2.0 --
## v broom        1.0.5      v rsample      1.2.1
## v dials        1.2.1      v tune         1.2.0
## v infer        1.0.7      v workflows    1.1.4
## v modeldata    1.3.0      v workflowsets 1.1.0
## v parsnip      1.2.1      v yardstick    1.3.1
## v recipes      1.0.10
## -- Conflicts ------------------------------------------ tidymodels_conflicts() --
## x scales::discard() masks purrr::discard()
## x dplyr::filter()   masks stats::filter()
## x recipes::fixed()  masks stringr::fixed()
## x dplyr::lag()      masks stats::lag()
## x yardstick::spec() masks readr::spec()
## x recipes::step()   masks stats::step()
## * Dig deeper into tidy modeling with R at https://www.tmwr.org
```

```
library(janitor)
```

```
##
## Attachement du package : 'janitor'
##
## Les objets suivants sont masqués depuis 'package:stats':
##
##     chisq.test, fisher.test
```

```r
library(broom)
library(gridExtra)
```

```
##
## Attachement du package : 'gridExtra'
##
## L'objet suivant est masqué depuis 'package:dplyr':
##
##     combine
```

```r
library(gtExtras)
```

```
## Le chargement a nécessité le package : gt
```

```r
library(readxl)

churn_data <- read.csv("C:\\Users\\Hp\\Desktop\\Machine Learning\\Customer Churn\\Telco-Customer-Churn.
```

2. Taking a look at the data

```r
summary(churn_data)
```

```
##   customerID          gender          SeniorCitizen      Partner
##  Length:7043        Length:7043        Min.   :0.0000   Length:7043
##  Class :character   Class :character   1st Qu.:0.0000   Class :character
##  Mode  :character   Mode  :character   Median :0.0000   Mode  :character
##                                        Mean   :0.1621
##                                        3rd Qu.:0.0000
##                                        Max.   :1.0000
##
##   Dependents            tenure      PhoneService       MultipleLines
##  Length:7043        Min.   : 0.00   Length:7043        Length:7043
##  Class :character   1st Qu.: 9.00   Class :character   Class :character
##  Mode  :character   Median :29.00   Mode  :character   Mode  :character
##                     Mean   :32.37
##                     3rd Qu.:55.00
##                     Max.   :72.00
##
##  InternetService    OnlineSecurity     OnlineBackup       DeviceProtection
##  Length:7043        Length:7043        Length:7043        Length:7043
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  TechSupport        StreamingTV        StreamingMovies      Contract
##  Length:7043        Length:7043        Length:7043        Length:7043
```

```
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  PaperlessBilling   PaymentMethod      MonthlyCharges    TotalCharges
##  Length:7043        Length:7043        Min.   : 18.25   Min.   :  18.8
##  Class :character   Class :character   1st Qu.: 35.50   1st Qu.: 401.4
##  Mode  :character   Mode  :character   Median : 70.35   Median :1397.5
##                                        Mean   : 64.76   Mean   :2283.3
##                                        3rd Qu.: 89.85   3rd Qu.:3794.7
##                                        Max.   :118.75   Max.   :8684.8
##                                                         NA's   :11
##     Churn
##  Length:7043
##  Class :character
##  Mode  :character
##
##
##
##
```

```
head(churn_data) %>%
  gt() %>%
  gt_theme_excel()
```

customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetSe
7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL
5575-GNVDE	Male	0	No	No	34	Yes	No	DSL
3668-QPYBK	Male	0	No	No	2	Yes	No	DSL
7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL
9237-HQITU	Female	0	No	No	2	Yes	No	Fiber opti
9305-CDSKC	Female	0	No	No	8	Yes	Yes	Fiber opti

```
glimpse(churn_data)
```

- The churn dataset has 11 missing values.

```
## Rows: 7,043
## Columns: 21
## $ customerID       <chr> "7590-VHVEG", "5575-GNVDE", "3668-QPYBK", "7795-CFOCW~
## $ gender           <chr> "Female", "Male", "Male", "Male", "Female", "Female",~
## $ SeniorCitizen    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ Partner          <chr> "Yes", "No", "No", "No", "No", "No", "No", "No", "Yes~
## $ Dependents       <chr> "No", "No", "No", "No", "No", "No", "Yes", "No", "No"~
## $ tenure           <int> 1, 34, 2, 45, 2, 8, 22, 10, 28, 62, 13, 16, 58, 49, 2~
## $ PhoneService     <chr> "No", "Yes", "Yes", "No", "Yes", "Yes", "Yes", "No", ~
## $ MultipleLines    <chr> "No phone service", "No", "No", "No phone service", "~
```

```
## $ InternetService  <chr> "DSL", "DSL", "DSL", "DSL", "Fiber optic", "Fiber opt~
## $ OnlineSecurity   <chr> "No", "Yes", "Yes", "Yes", "No", "No", "No", "Yes", "~
## $ OnlineBackup     <chr> "Yes", "No", "Yes", "No", "No", "No", "Yes", "No", "N~
## $ DeviceProtection <chr> "No", "Yes", "No", "Yes", "No", "Yes", "No", "No", "Y~
## $ TechSupport      <chr> "No", "No", "No", "Yes", "No", "No", "No", "No", "Yes~
## $ StreamingTV      <chr> "No", "No", "No", "No", "No", "Yes", "Yes", "No", "Ye~
## $ StreamingMovies  <chr> "No", "No", "No", "No", "No", "Yes", "No", "No", "Yes~
## $ Contract         <chr> "Month-to-month", "One year", "Month-to-month", "One ~
## $ PaperlessBilling <chr> "Yes", "No", "Yes", "No", "Yes", "Yes", "Yes", "No", ~
## $ PaymentMethod    <chr> "Electronic check", "Mailed check", "Mailed check", "~
## $ MonthlyCharges   <dbl> 29.85, 56.95, 53.85, 42.30, 70.70, 99.65, 89.10, 29.7~
## $ TotalCharges     <dbl> 29.85, 1889.50, 108.15, 1840.75, 151.65, 820.50, 1949~
## $ Churn            <chr> "No", "No", "Yes", "No", "Yes", "Yes", "No", "No", "Y~
```

```
dim(churn_data)
```

```
## [1] 7043   21
```

- this dataset has 7043 rows and 21 columns

3. Cleaning Data

```
churn_data <- churn_data %>%
  select(-customerID) %>%
  mutate(SeniorCitizen = as.factor(ifelse(churn_data$SeniorCitizen==1, 'Yes', 'No'))) %>%
  clean_names()%>%
  mutate_if(is.character , as.factor) %>%
  na.omit()

glimpse(churn_data)
```

```
## Rows: 7,032
## Columns: 20
## $ gender            <fct> Female, Male, Male, Male, Female, Female, Male, Fema~
## $ senior_citizen    <fct> No, No, No, No, No, No, No, No, No, No, No, No, No, ~
## $ partner           <fct> Yes, No, No, No, No, No, No, No, Yes, No, Yes, No, Y~
## $ dependents        <fct> No, No, No, No, No, No, Yes, No, No, Yes, Yes, No, N~
## $ tenure            <int> 1, 34, 2, 45, 2, 8, 22, 10, 28, 62, 13, 16, 58, 49, ~
## $ phone_service     <fct> No, Yes, Yes, No, Yes, Yes, Yes, No, Yes, Yes, Yes, ~
## $ multiple_lines    <fct> No phone service, No, No, No phone service, No, Yes,~
## $ internet_service  <fct> DSL, DSL, DSL, DSL, Fiber optic, Fiber optic, Fiber ~
## $ online_security   <fct> No, Yes, Yes, Yes, No, No, No, Yes, No, Yes, Yes, No~
## $ online_backup     <fct> Yes, No, Yes, No, No, No, Yes, No, No, Yes, No, No i~
## $ device_protection <fct> No, Yes, No, Yes, No, Yes, No, No, Yes, No, No, No i~
## $ tech_support      <fct> No, No, No, Yes, No, No, No, No, Yes, No, No, No int~
## $ streaming_tv      <fct> No, No, No, No, No, Yes, Yes, No, Yes, No, No, No in~
## $ streaming_movies  <fct> No, No, No, No, No, Yes, No, No, Yes, No, No, No int~
## $ contract          <fct> Month-to-month, One year, Month-to-month, One year, ~
## $ paperless_billing <fct> Yes, No, Yes, No, Yes, Yes, Yes, No, Yes, No, Yes, N~
## $ payment_method    <fct> Electronic check, Mailed check, Mailed check, Bank t~
```

```
## $ monthly_charges   <dbl> 29.85, 56.95, 53.85, 42.30, 70.70, 99.65, 89.10, 29.~
## $ total_charges     <dbl> 29.85, 1889.50, 108.15, 1840.75, 151.65, 820.50, 194~
## $ churn             <fct> No, No, Yes, No, Yes, Yes, No, No, Yes, No, No, No, ~
```

```r
nrow(churn_data)
```

```
## [1] 7032
```

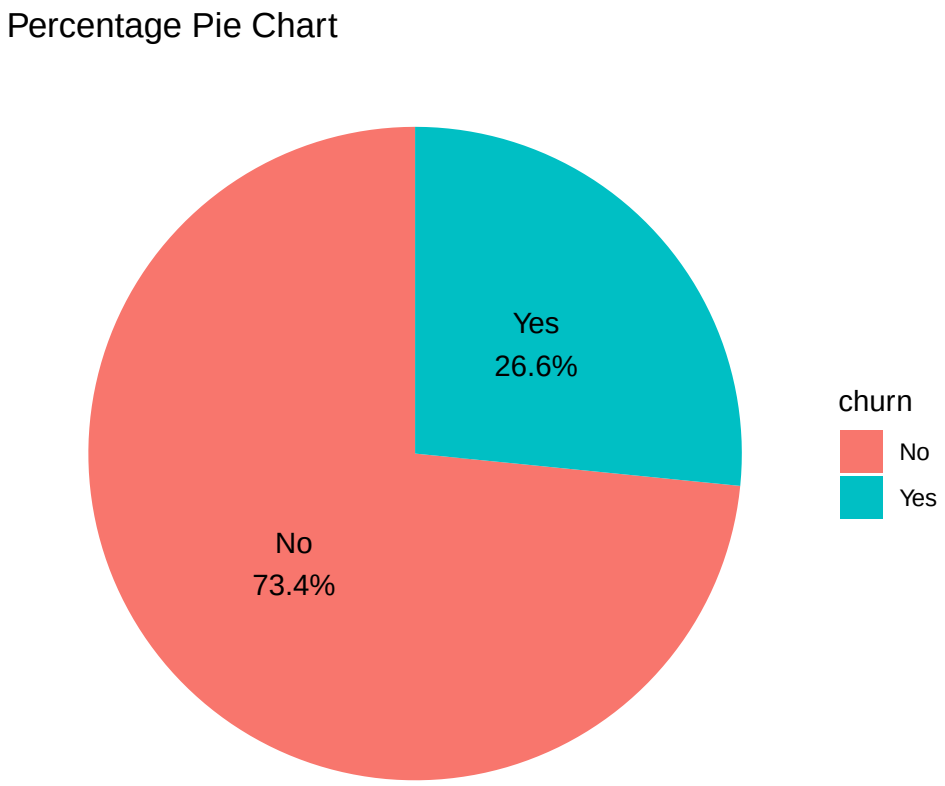
4. Explanatory Data Analysis (EDA)

a- distribution of categorical variables

```r
churn_percent <- churn_data %>%
  group_by(churn) %>%
  count()%>%
  summarise(percent = n / nrow(churn_data) * 100 )

churn_percent
```

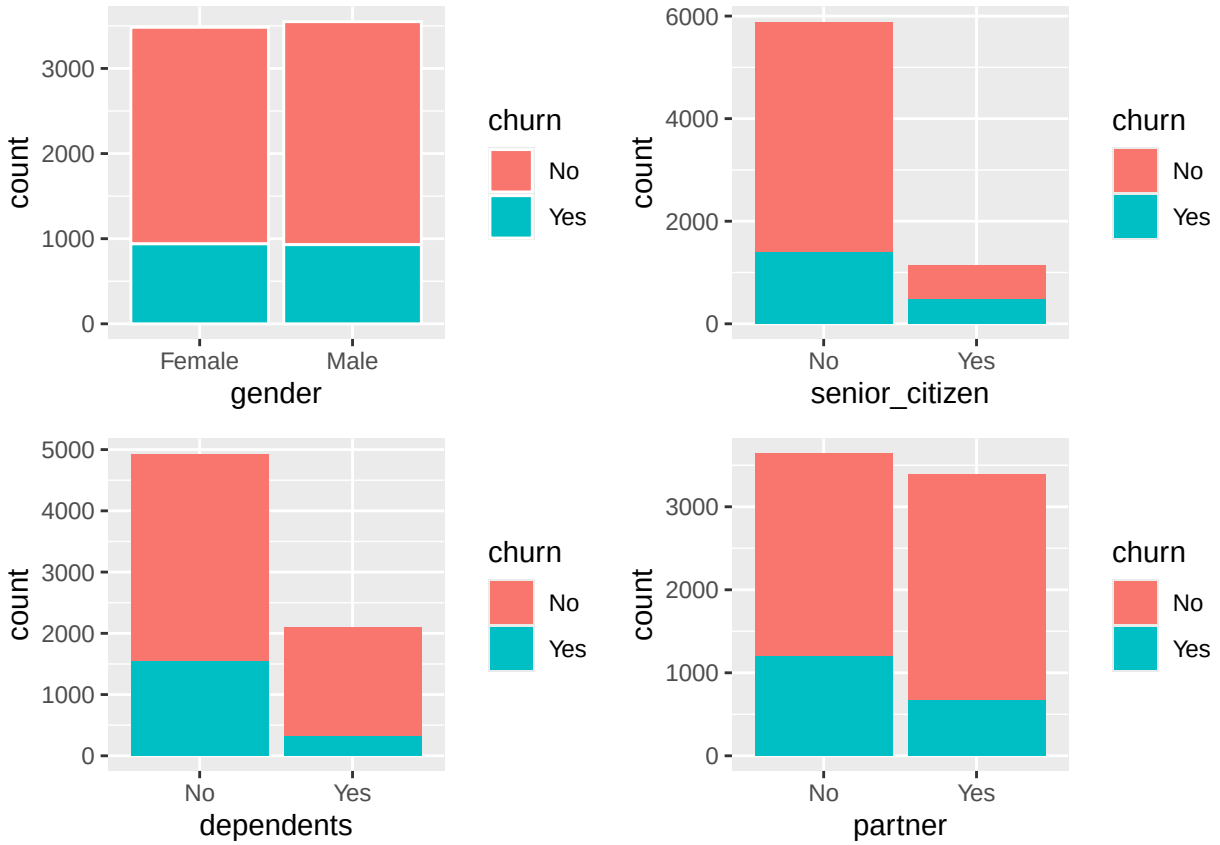
```
## # A tibble: 2 x 2
##   churn percent
##   <fct>   <dbl>
## 1 No       73.4
## 2 Yes      26.6
```

```r
churn_pie <- churn_percent %>%
  ggplot( aes(x = "", y = percent , fill = churn)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar(theta = "y") +
  theme_void() +
  labs(title = "Percentage Pie Chart") +
  geom_text(aes(label = paste0(churn, "\n", round(percent, 1), "%")),
            position = position_stack(vjust = 0.5))

churn_pie
```



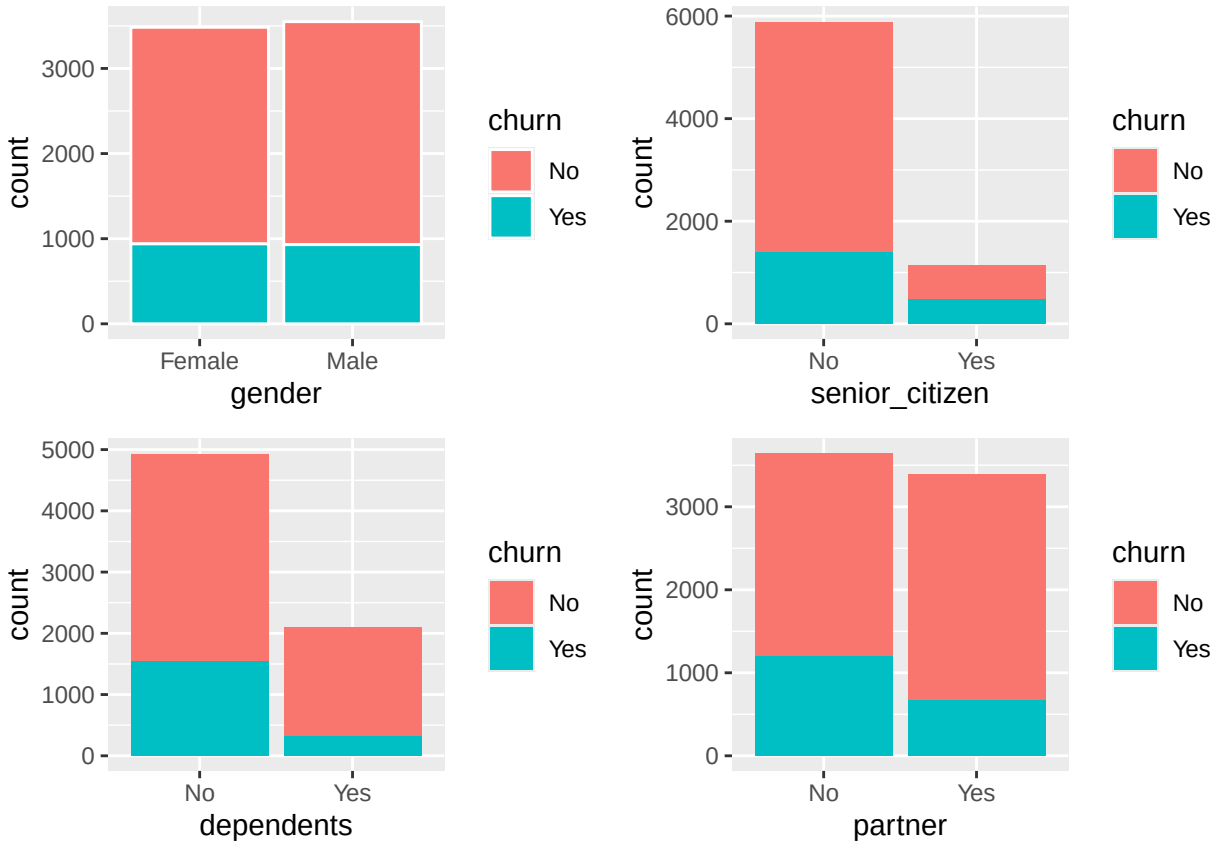
- we see that that the majority of customers didn't opt to lay off the services of the company.

```
graph1 <- ggplot(churn_data, aes(x=gender,fill=churn ))+
  geom_bar(color="white")
graph2 <- ggplot(churn_data, aes(x=senior_citizen,fill=churn))+
  geom_bar()
graph3 <- ggplot(churn_data, aes(x=dependents,fill=churn))+
  geom_bar()
graph4 <- ggplot(churn_data, aes(x=partner,fill=churn))+
  geom_bar()
grid.arrange(graph1,graph2,graph3,graph4,ncol=2)
```



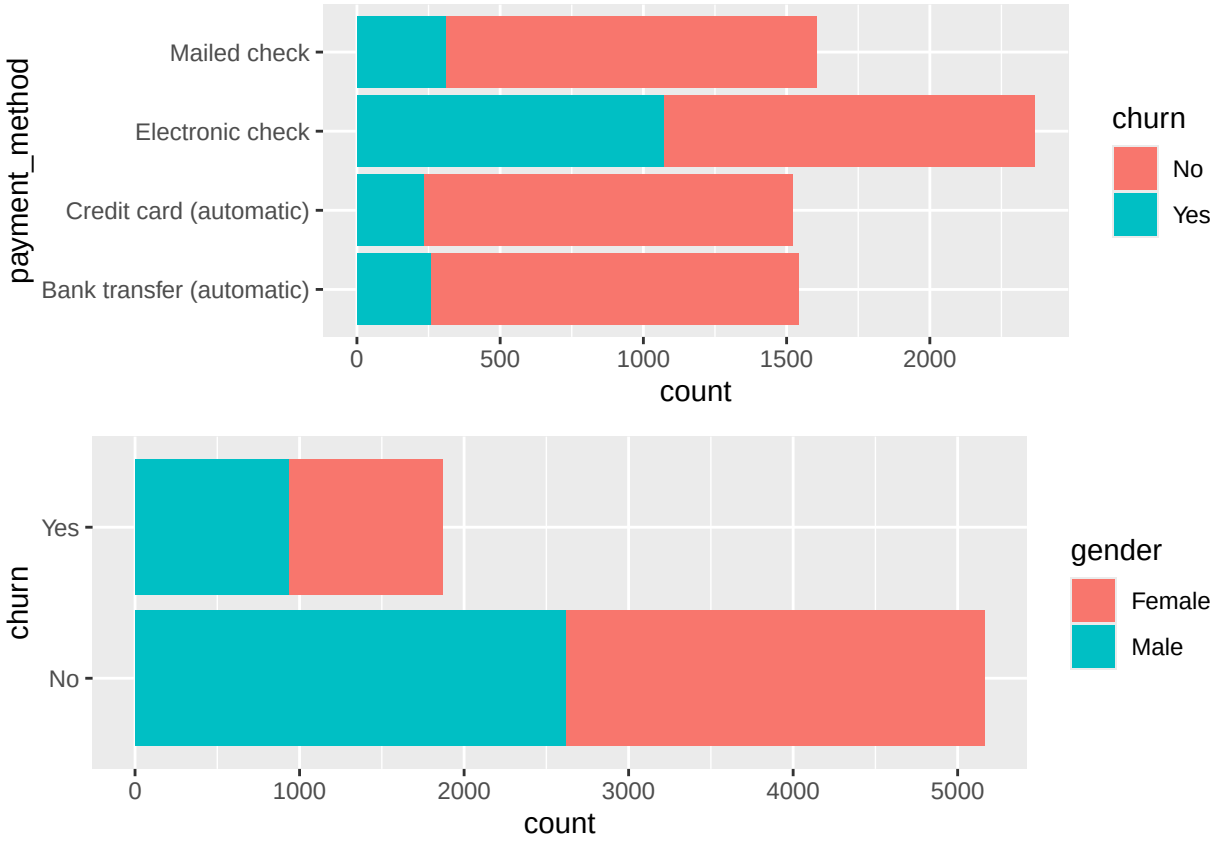
- customer churn rate is higher within senior citizens than non-senior citizens
- customer churn rate is higher within clients who don't have dependents or partners than those who don't , most likely due to the financial conditions of these clients

```
graph5 <- ggplot(churn_data, aes(x=streaming_tv,fill=churn))+
  geom_bar()
graph6 <- ggplot(churn_data, aes(x=streaming_movies,fill=churn))+
  geom_bar()
graph7 <- ggplot(churn_data, aes(x=contract,fill=churn))+
  geom_bar()
graph8 <- ggplot(churn_data, aes(x=paperless_billing,fill=churn))+
  geom_bar()
grid.arrange(graph1,graph2,graph3,graph4,ncol=2)
```



- **Streaming Services**: Customers who use streaming services (e.g., TV, movies) are less likely to churn.
- **Phone Services**: Phone service alone does not significantly impact churn.

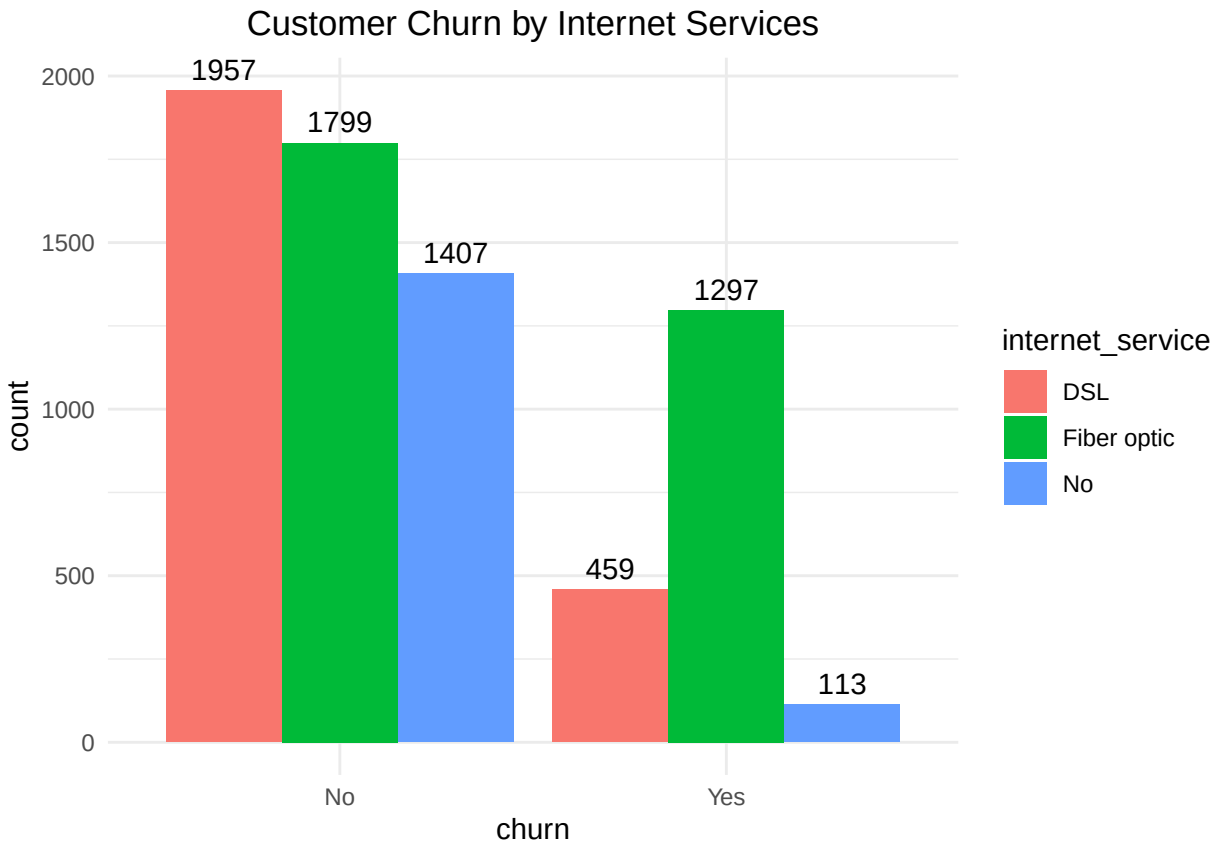
```
plot1 <- ggplot(churn_data, aes(x=payment_method,fill=churn))+
  geom_bar()+
  coord_flip()
plot4 <- ggplot(churn_data, aes(x=churn,fill=gender))+
  geom_bar()+
  coord_flip()
grid.arrange(plot1,plot4)
```



- **Electronic Check**: Customers using electronic checks have a higher churn rate.
- **Automatic Payment**: Encouraging automatic payment methods may reduce churn.

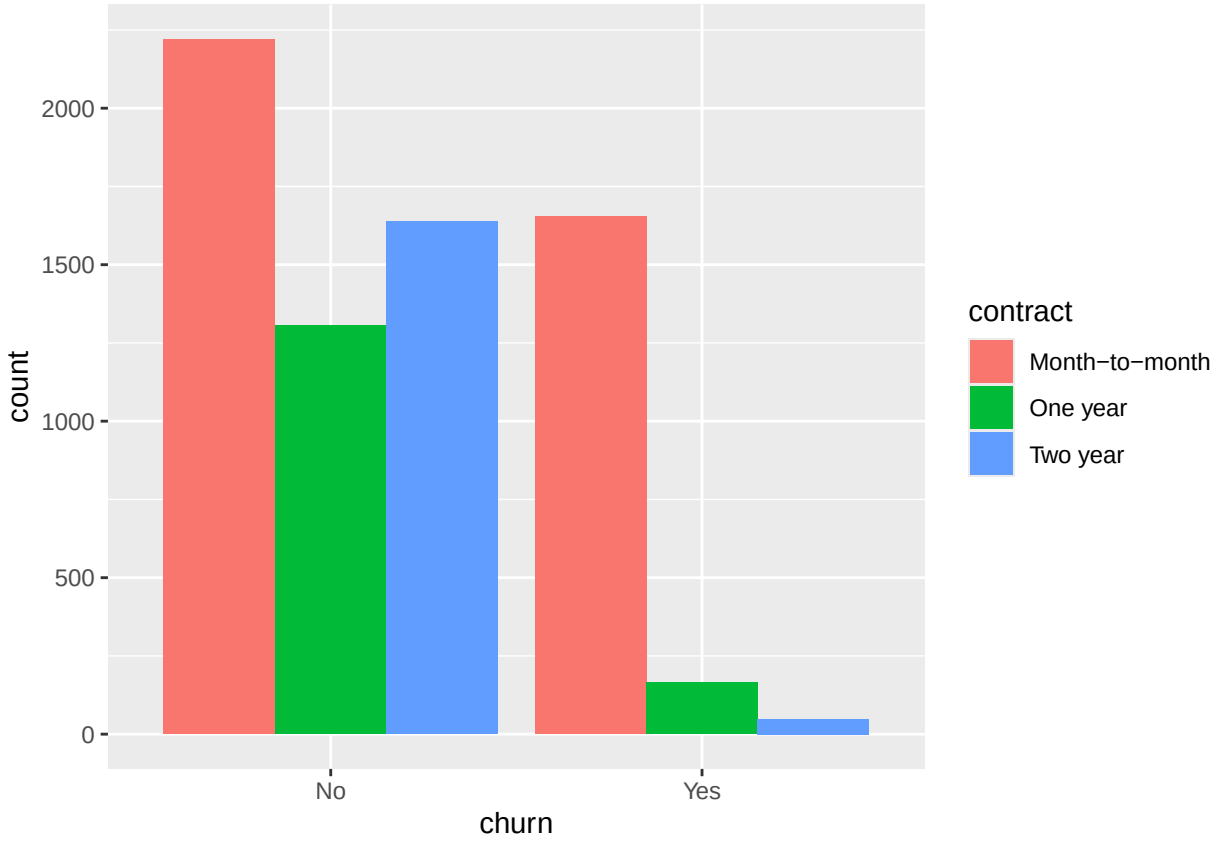
```
ggplot(data = churn_data, aes(x = churn,fill = internet_service)) +
  geom_bar(stat = "count",position = position_dodge()) +
  geom_text(stat = "count" , aes(label = paste( formatC(..count..))),vjust = -0.5 , position = position_
  ggtitle("Customer Churn by Internet Services") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```

```
## Warning: The dot-dot notation (‘..count..‘) was deprecated in ggplot2 3.4.0.
## i Please use ‘after_stat(count)‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```



- clients who opted for the fiber optic service have the highest rate of churn within those who have internet service , indicating their dissatisfaction with the services provided by the company. Although the majority of these clients use this service.

```
ggplot(data = churn_data, aes(x = churn,fill = contract)) +
  geom_bar(stat = "count",position = position_dodge())
```



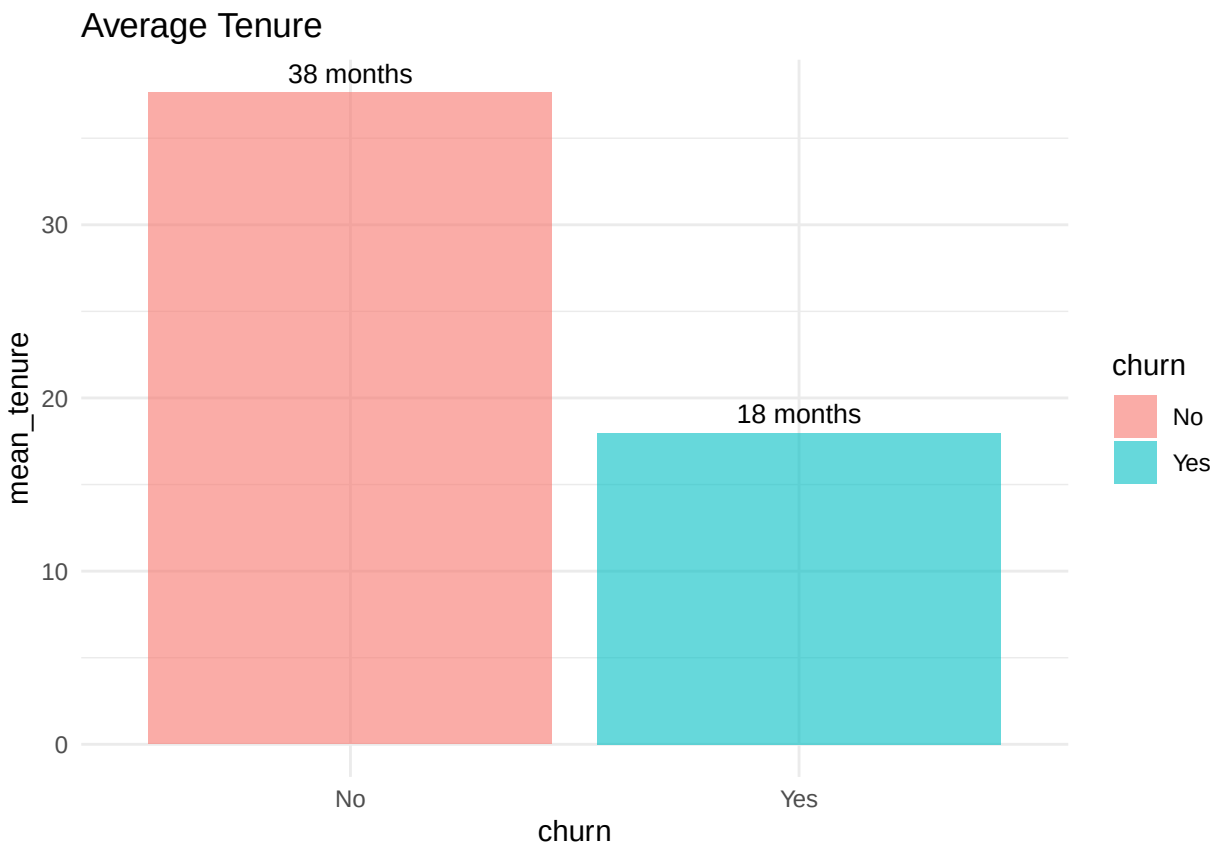
- **Contract**: Month-to-month contracts have a significantly higher churn rate compared to one-year or two-year contracts.

- **Contract Length**: Encouraging longer contract commitments may reduce churn

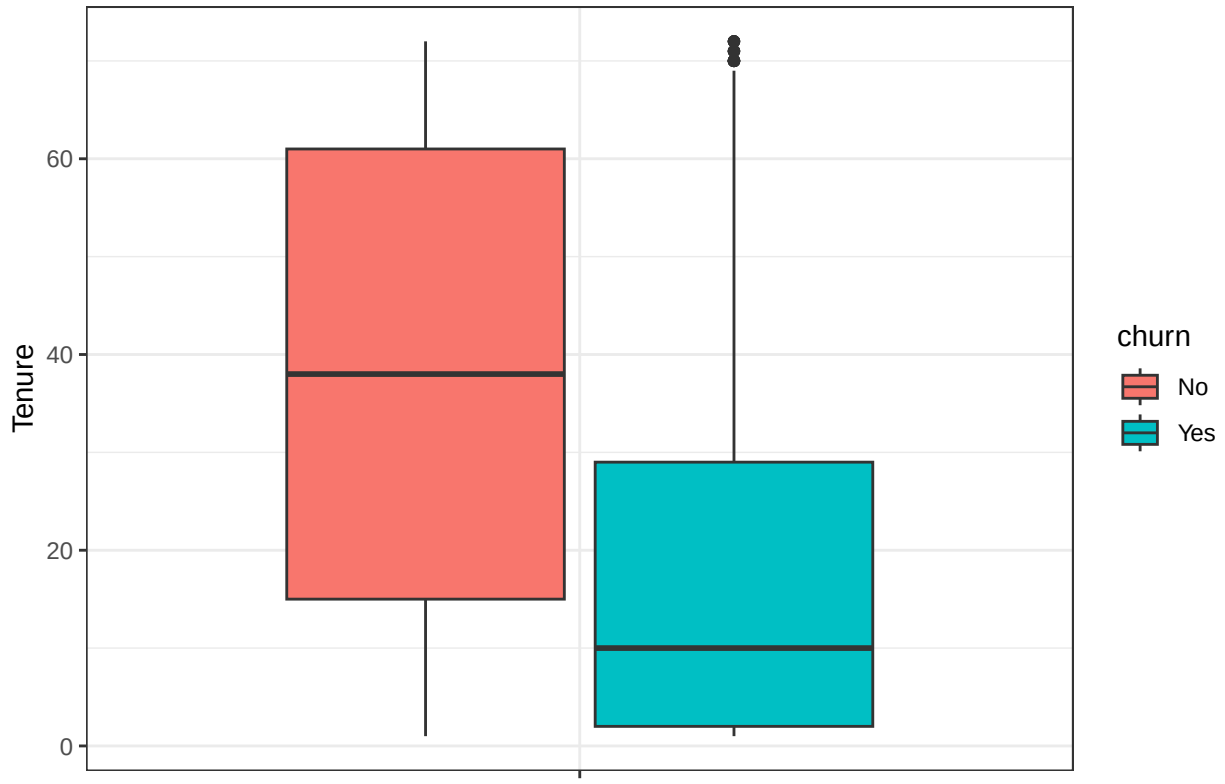
5 - Distribution of numerical variables.

```
churn_summary <- churn_data %>%
  group_by(churn) %>%
  summarize(mean_tenure = mean(tenure),
            mean_monthlyCharges = mean(monthly_charges))
```

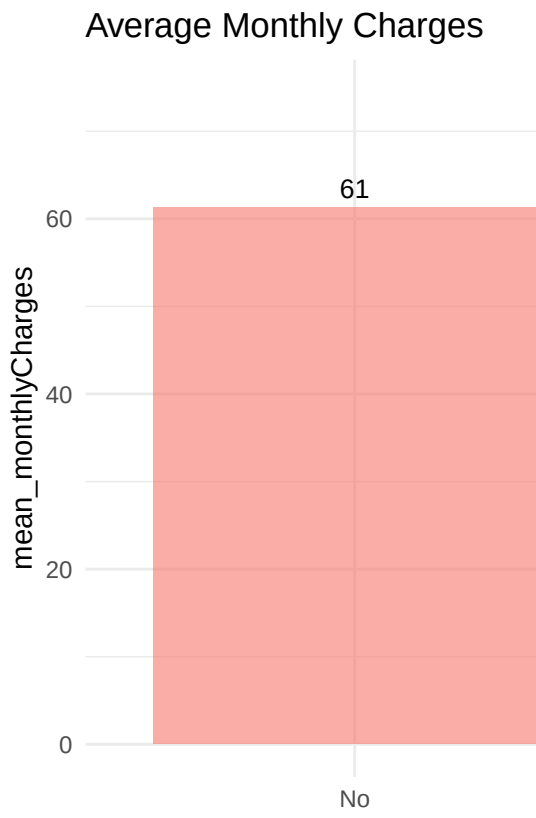
```
combined_plot <- ggplot(churn_summary, aes(x = churn)) +
  geom_bar(aes(y = mean_tenure, fill = churn), stat = "identity", alpha = 0.6) +
  geom_text(aes(y = mean_tenure, label = paste(round(mean_tenure, 0), "months")),
            size = 3.5, vjust = -0.5) +
  labs(title = "Average Tenure") +
  theme_minimal()

combined_plot
```




```
churn_data %>%
  ggplot(aes(x = "", y = tenure, fill = churn)) +
  geom_boxplot() +
  theme_bw() +
  xlab("") +
  ylab("Tenure")
```



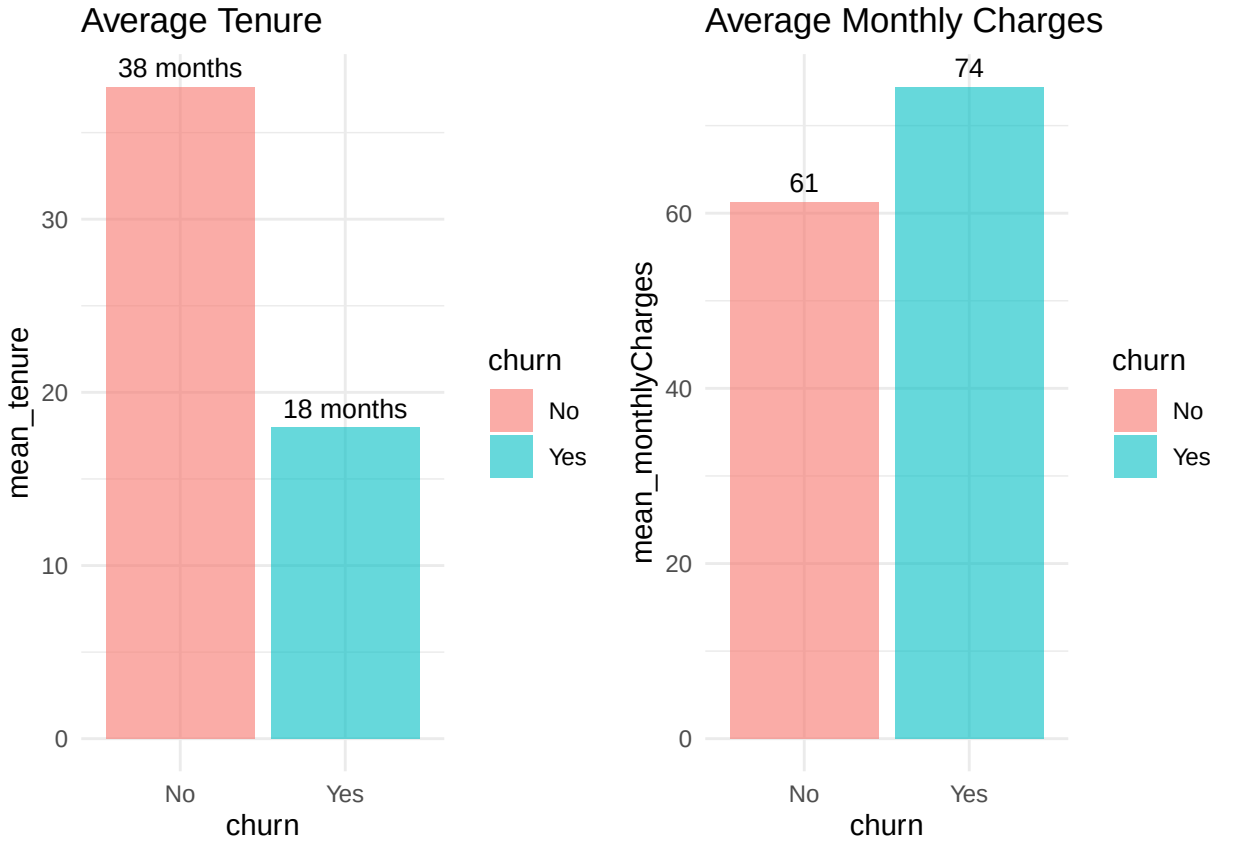

```
monthlyCharges_plot <- ggplot(churn_summary, aes(x = churn)) +
  geom_bar(aes(y = mean_monthlyCharges, fill = churn), stat = "identity", alpha = 0.6) +
  geom_text(aes(y = mean_monthlyCharges, label = paste(round(mean_monthlyCharges, 0))),
            size = 3.5, vjust = -0.5) +
  labs(title = "Average Monthly Charges") +
  theme_minimal()

monthlyCharges_plot
```



Calculate the mean tenure and monthly charges according to churn

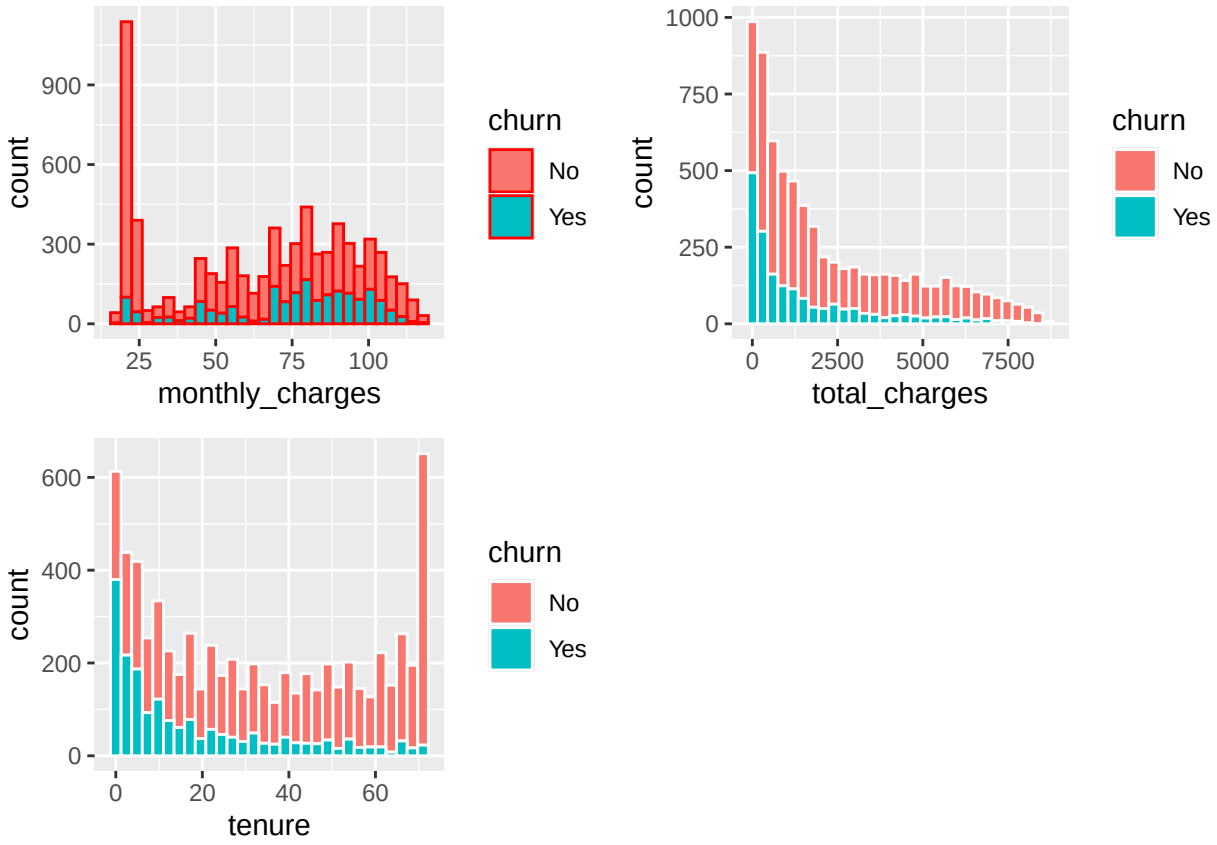
```
library(gridExtra)
combined_plots <- grid.arrange(combined_plot, monthlyCharges_plot, ncol = 2)
```

- **The average monthly charges for churn customers are more than no-churn customers.**




```
g1 <- churn_data %>%
  ggplot(aes(x=monthly_charges,fill=churn ))+
  geom_histogram(color="red")
g2 <- churn_data %>%
  ggplot(aes(x=total_charges,fill=churn ))+
  geom_histogram(color="white")
g3 <- churn_data %>%
  ggplot(aes(x=tenure,fill=churn ))+
  geom_histogram(color="white")
grid.arrange(g1,g2,g3,ncol=2)
```

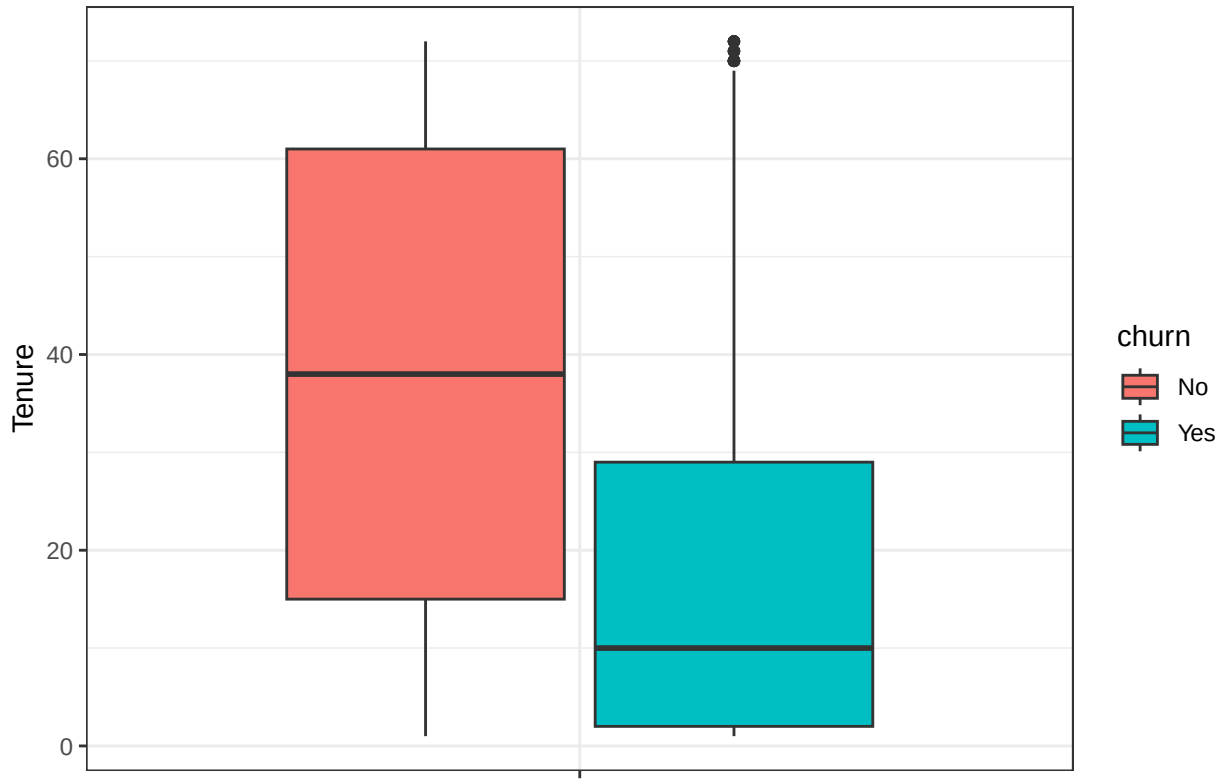
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



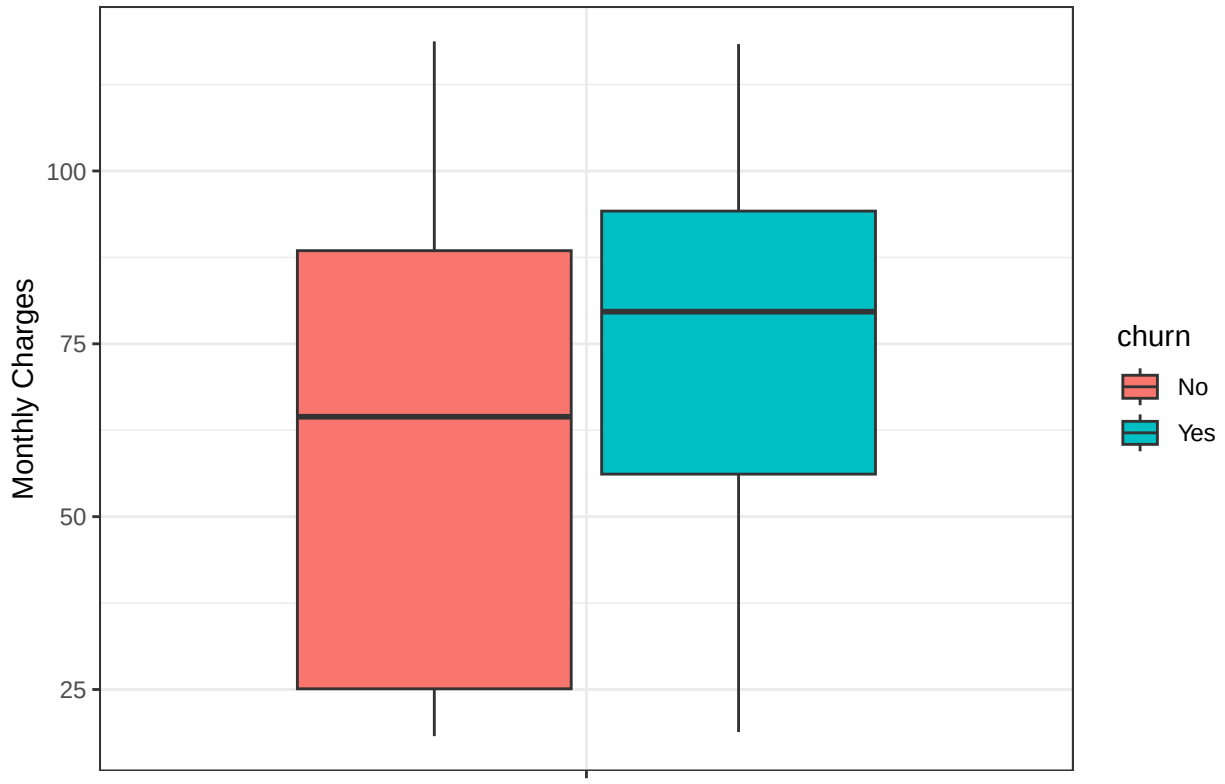
- from the graphs we notice that :
  - Shorter tenure correlates with higher churn rates. New customers are more likely to leave.
  - Customers with longer tenure (e.g., more than 60 months) are loyal and less likely to churn.
  - Higher monthly charges are associated with higher churn rates.

6. the relationship between Churn and the numerical variables.

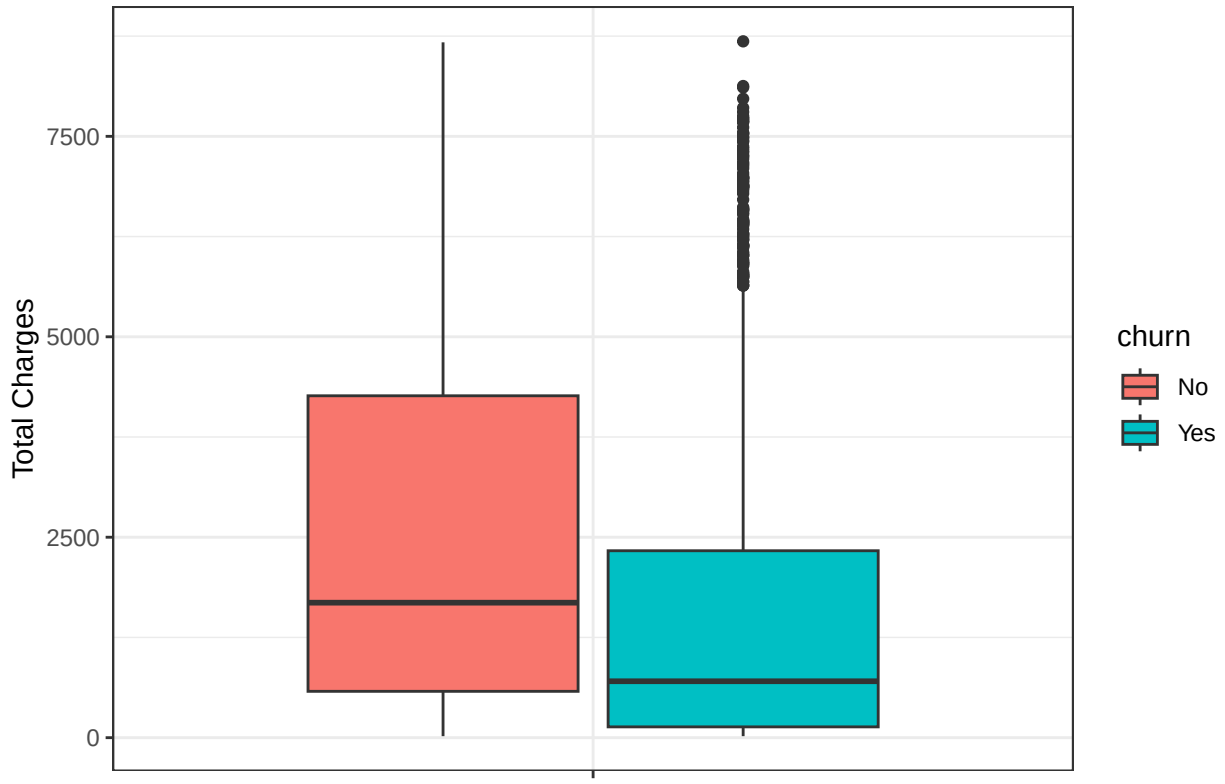
```
churn_data %>%
  ggplot(aes(x = "", y = tenure, fill = churn)) +
  geom_boxplot() +
  theme_bw() +
  xlab("") +
  ylab("Tenure")
```




```
churn_data %>%
  ggplot(aes(x = "", y = monthly_charges, fill = churn)) +
  geom_boxplot() +
  theme_bw() +
  xlab("") +
  ylab("Monthly Charges")
```




```
churn_data %>%
  ggplot(aes(x = "", y = total_charges, fill = churn)) +
  geom_boxplot() +
  theme_bw() +
  xlab("") +
  ylab("Total Charges")
```



Data Preprocessing

treating the target variables

```
churn_data <- churn_data %>% mutate(churn= as.factor(ifelse(churn=="Yes",1,0)))

head(churn_data)%>%
  gt() %>%
  gt_theme_excel()
```

gender	senior_citizen	partner	dependents	tenure	phone_service	multiple_lines	internet_service	online_
Female	No	Yes	No	1	No	No phone service	DSL	N
Male	No	No	No	34	Yes	No	DSL	Y
Male	No	No	No	2	Yes	No	DSL	Y
Male	No	No	No	45	No	No phone service	DSL	Y
Female	No	No	No	2	Yes	No	Fiber optic	N
Female	No	No	No	8	Yes	Yes	Fiber optic	N

Spliting the data

```
set.seed(123)
churn_split <-initial_split(churn_data,
                            prop = 0.8,
                            strata ="churn" )
churn_train <- training(churn_split)
churn_test <- testing(churn_split)
churn_split
```

```
## <Training/Testing/Total>
## <5625/1407/7032>
```

recipe

```
rec_churn <- recipe(churn~.,churn_train)
```

```
churn_rec <- rec_churn %>%
  step_normalize(all_numeric_predictors()) %>%
  step_dummy(all_nominal_predictors()) %>%
  prep()
  churn_rec
```

```
##

## -- Recipe ----------------------------------------------------------------

##

## -- Inputs

## Number of variables by role

## outcome:    1
## predictor: 19

##

## -- Training information

## Training data contained 5625 data points and no incomplete rows.

##

## -- Operations

## * Centering and scaling for: tenure and monthly_charges, ... | Trained

## * Dummy variables from: gender, senior_citizen, partner, ... | Trained
```

- baking the data in the recipe

```
churn_train_process <-bake(churn_rec,churn_train)
head(churn_train_process)%>%
  gt() %>%
  gt_theme_excel()
```

tenure	monthly_charges	total_charges	churn	gender_Male	senior_citizen_Yes	partner_Yes	dependent
-1.27844131	-1.1692918	-0.9943489	0	0	0	1	
0.06459603	-0.2662516	-0.1744290	0	1	0	0	
0.51227514	-0.7544265	-0.1959229	0	1	0	0	
-0.91215840	-1.1726240	-0.8744020	0	0	0	0	
-0.79006410	-0.4995092	-0.7485030	0	1	0	1	
1.04135046	1.1799455	1.4972878	0	1	0	1	

- setting the engine to :

##1- logistic regression

```
logic_specification <- logistic_reg() %>%
 set_engine("glm") %>%
  set_mode("classification")
```

- model training

```
logit_fit <- logic_specification %>%
  fit(churn  ~. , churn_train_process)
```

- baking the testing data

```
test_train_process <-bake(churn_rec,churn_test)
head(test_train_process)%>%
  gt() %>%
  gt_theme_excel()
```

tenure	monthly_charges	total_charges	churn	gender_Male	senior_citizen_Yes	partner_Yes	dependents
-0.4237812	0.8050672	-0.1480191	0	1	0	0	
-0.1795926	1.3282307	0.3354940	1	0	0	1	
1.2041429	-0.2929096	0.5303277	0	1	0	0	
-0.6679698	-1.5325072	-0.8634236	0	1	0	0	
0.6750675	1.2915759	1.2129954	1	1	0	0	
-0.8307622	-1.5041831	-0.9183377	0	1	0	1	

- prediction

```
churn_pred <- predict(logit_fit, test_train_process)
churn_pred
```

```
## # A tibble: 1,407 x 1
##    .pred_class
##    <fct>
##  1 1
##  2 1
##  3 0
##  4 0
##  5 0
##  6 0
##  7 0
##  8 0
##  9 1
## 10 1
## # i 1,397 more rows
```

```
churn_test_proc_results <- test_train_process %>%
    dplyr::bind_cols(churn_pred)
churn_test_proc_results
```

```
## # A tibble: 1,407 x 32
##     tenure monthly_charges total_charges churn gender_Male senior_citizen_Yes
##      <dbl>           <dbl>         <dbl> <fct>       <dbl>               <dbl>
##  1 -0.424            0.805        -0.148 0               1                   0
##  2 -0.180            1.33          0.335 1               0                   0
##  3  1.20            -0.293         0.530 0               1                   0
##  4 -0.668           -1.53         -0.863 0               1                   0
##  5  0.675            1.29          1.21  1               1                   0
##  6 -0.831           -1.50         -0.918 0               1                   0
##  7 -1.28            -1.49         -0.999 0               1                   0
##  8 -1.28            -0.656        -0.988 0               1                   1
##  9  0.0646           1.38          0.557 1               1                   0
## 10 -0.871            1.10         -0.520 1               0                   0
## # i 1,397 more rows
## # i 26 more variables: partner_Yes <dbl>, dependents_Yes <dbl>,
## #   phone_service_Yes <dbl>, multiple_lines_No.phone.service <dbl>,
## #   multiple_lines_Yes <dbl>, internet_service_Fiber.optic <dbl>,
## #   internet_service_No <dbl>, online_security_No.internet.service <dbl>,
## #   online_security_Yes <dbl>, online_backup_No.internet.service <dbl>,
## #   online_backup_Yes <dbl>, device_protection_No.internet.service <dbl>, ...
```

- evaluation of the model

```
yardstick::accuracy(churn_test_proc_results,churn,.pred_class)
```

```
## # A tibble: 1 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary         0.807
```

##2- KNN(K-nearest neighbors)

- set the specification :

```
knn_spec <- nearest_neighbor() %>%
 set_engine("kknn") %>%
  set_mode("classification")
```

- set the model :

```
knn_fit <- knn_spec %>%
  fit(churn ~. , churn_train_process)
```

- prediction :

```
knn_churn_pred <- predict(knn_fit, test_train_process)
knn_churn_pred
```

```
## # A tibble: 1,407 x 1
##    .pred_class
##    <fct>
##  1 1
##  2 1
##  3 0
##  4 0
##  5 1
##  6 0
##  7 1
##  8 1
##  9 1
## 10 1
## # i 1,397 more rows
```

```
knn_churn_test_results <- test_train_process %>%
     dplyr::bind_cols(knn_churn_pred)
knn_churn_test_results
```

```
## # A tibble: 1,407 x 32
##     tenure monthly_charges total_charges churn gender_Male senior_citizen_Yes
##      <dbl>           <dbl>         <dbl> <fct>       <dbl>               <dbl>
##  1 -0.424            0.805        -0.148 0               1                   0
##  2 -0.180            1.33          0.335 1               0                   0
##  3  1.20            -0.293         0.530 0               1                   0
##  4 -0.668           -1.53         -0.863 0               1                   0
##  5  0.675            1.29          1.21  1               1                   0
##  6 -0.831           -1.50         -0.918 0               1                   0
##  7 -1.28            -1.49         -0.999 0               1                   0
##  8 -1.28            -0.656        -0.988 0               1                   1
##  9  0.0646           1.38          0.557 1               1                   0
## 10 -0.871            1.10         -0.520 1               0                   0
## # i 1,397 more rows
```

```
## # i 26 more variables: partner_Yes <dbl>, dependents_Yes <dbl>,
## #   phone_service_Yes <dbl>, multiple_lines_No.phone.service <dbl>,
## #   multiple_lines_Yes <dbl>, internet_service_Fiber.optic <dbl>,
## #   internet_service_No <dbl>, online_security_No.internet.service <dbl>,
## #   online_security_Yes <dbl>, online_backup_No.internet.service <dbl>,
## #   online_backup_Yes <dbl>, device_protection_No.internet.service <dbl>, ...
```

- evaluation of the model

```
yardstick::accuracy(knn_churn_test_results,churn,.pred_class)
```

```
## # A tibble: 1 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary         0.684
```

3. Decision Tree

```
decision_spec <- decision_tree() %>%
 set_engine("rpart") %>%
  set_mode("classification")
```

set the model :

training the model

```
decision_fit <- decision_spec %>%
  fit(churn  ~. , churn_train_process)
```

prediction

```
decision_churn_pred <- predict(decision_fit, test_train_process)
head(decision_churn_pred)%>%
  gt() %>%
  gt_theme_excel()
```

.pred_class
0
0
0
0
0
0

```
decision_churn_test_results <- test_train_process %>%
     dplyr::bind_cols(decision_churn_pred)
head(decision_churn_test_results)%>%
  gt() %>%
  gt_theme_excel()
```

tenure	monthly_charges	total_charges	churn	gender_Male	senior_citizen_Yes	partner_Yes	dependents
-0.4237812	0.8050672	-0.1480191	0	1	0	0	
-0.1795926	1.3282307	0.3354940	1	0	0	1	
1.2041429	-0.2929096	0.5303277	0	1	0	0	
-0.6679698	-1.5325072	-0.8634236	0	1	0	0	
0.6750675	1.2915759	1.2129954	1	1	0	0	
-0.8307622	-1.5041831	-0.9183377	0	1	0	1	

accuracy

```
yardstick::accuracy(decision_churn_test_results,churn,.pred_class)
```

```
## # A tibble: 1 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary         0.781
```

4. Random Forest

setting the model

```
rand_forest_spec <- rand_forest() %>%
 set_engine("ranger") %>%
  set_mode("classification")
```

training the model

```
random_fit <- rand_forest_spec %>%
  fit(churn  ~. , churn_train_process)
```

prediction

```
random_churn_pred <- predict(random_fit, test_train_process)
random_churn_pred
```

```
## # A tibble: 1,407 x 1
##    .pred_class
##    <fct>
##  1 0
##  2 1
##  3 0
##  4 0
##  5 0
##  6 0
##  7 0
##  8 1
##  9 0
## 10 1
## # i 1,397 more rows
```

the most accurate model is that of logical regresion with 0.8066809

6-new_customers_data

```
new_churn_data <- read_xlsx("C:\\Users\\Hp\\Desktop\\Machine Learning\\Customer Churn\\new_customers_da
new_churn_data
```

```
## # A tibble: 50 x 20
##    customerID gender SeniorCitizen Partner Dependents tenure PhoneService
##    <chr>      <chr>          <dbl> <chr>   <chr>       <dbl> <chr>
##  1 25795      Male               1 No      No             51 Yes
##  2 10860      Male               1 Yes     Yes            61 No
##  3 86820      Male               0 Yes     No             57 Yes
##  4 64886      Female             1 Yes     Yes            51 Yes
##  5 16265      Male               1 Yes     Yes            11 No
##  6 92386      Female             1 Yes     No             38 No
##  7 47194      Male               0 Yes     No              1 Yes
##  8 97498      Female             0 No      No              2 Yes
##  9 54131      Male               0 Yes     No             55 No
## 10 70263      Female             0 Yes     No             58 No
## # i 40 more rows
## # i 13 more variables: MultipleLines <chr>, InternetService <chr>,
## #   OnlineSecurity <chr>, OnlineBackup <chr>, DeviceProtection <chr>,
## #   TechSupport <chr>, StreamingTV <chr>, StreamingMovies <chr>,
## #   Contract <chr>, PaperlessBilling <chr>, PaymentMethod <chr>,
## #   MonthlyCharges <dbl>, TotalCharges <dbl>
```

taking a look at the data

```
glimpse(new_churn_data)
```

```
## Rows: 50
## Columns: 20
## $ customerID       <chr> "25795", "10860", "86820", "64886", "16265", "92386",~
## $ gender           <chr> "Male", "Male", "Male", "Female", "Male", "Female", "~
## $ SeniorCitizen    <dbl> 1, 1, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1,~
## $ Partner          <chr> "No", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "No",~
## $ Dependents       <chr> "No", "Yes", "No", "Yes", "Yes", "No", "No", "No", "N~
## $ tenure           <dbl> 51, 61, 57, 51, 11, 38, 1, 2, 55, 58, 1, 1, 53, 0, 18~
## $ PhoneService     <chr> "Yes", "No", "Yes", "Yes", "No", "No", "Yes", "Yes", ~
## $ MultipleLines    <chr> "No", "Yes", "No phone service", "Yes", "Yes", "No", ~
## $ InternetService  <chr> "DSL", "Fiber optic", "No", "DSL", "DSL", "DSL", "DSL~
## $ OnlineSecurity   <chr> "No", "No internet service", "Yes", "Yes", "Yes", "No~
## $ OnlineBackup     <chr> "No", "No internet service", "No", "Yes", "No", "Yes"~
## $ DeviceProtection <chr> "No", "No", "Yes", "No internet service", "No", "No",~
## $ TechSupport      <chr> "No", "No", "No internet service", "Yes", "No interne~
## $ StreamingTV      <chr> "No internet service", "No internet service", "No", "~
## $ StreamingMovies  <chr> "Yes", "Yes", "No", "Yes", "No internet service", "No~
## $ Contract         <chr> "Two year", "Month-to-month", "Two year", "Two year",~
## $ PaperlessBilling <chr> "No", "No", "No", "Yes", "Yes", "Yes", "Yes", "No", "~
## $ PaymentMethod    <chr> "Electronic check", "Electronic check", "Mailed check~
## $ MonthlyCharges   <dbl> 31.96, 19.71, 53.48, 77.54, 57.67, 62.22, 109.12, 53.~
## $ TotalCharges     <dbl> 5459.98, 726.82, 7589.16, 7999.04, 547.59, 2417.82, 7~
```

```
summary(new_churn_data)
```

```
##   customerID          gender          SeniorCitizen    Partner
##  Length:50          Length:50          Min.   :0.00   Length:50
##  Class :character   Class :character   1st Qu.:0.00   Class :character
##  Mode  :character   Mode  :character   Median :0.00   Mode  :character
##                                        Mean   :0.42
##                                        3rd Qu.:1.00
##                                        Max.   :1.00
##   Dependents            tenure      PhoneService       MultipleLines
##  Length:50          Min.   : 0.00   Length:50          Length:50
##  Class :character   1st Qu.:16.50   Class :character   Class :character
##  Mode  :character   Median :38.00   Mode  :character   Mode  :character
##                     Mean   :36.58
##                     3rd Qu.:56.75
##                     Max.   :72.00
##  InternetService    OnlineSecurity     OnlineBackup       DeviceProtection
##  Length:50          Length:50          Length:50          Length:50
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  TechSupport        StreamingTV        StreamingMovies      Contract
##  Length:50          Length:50          Length:50          Length:50
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
```

```
##
##
##
##  PaperlessBilling   PaymentMethod      MonthlyCharges    TotalCharges
##  Length:50          Length:50          Min.   : 19.71   Min.   : 192.2
##  Class :character   Class :character   1st Qu.: 35.31   1st Qu.:2242.4
##  Mode  :character   Mode  :character   Median : 54.17   Median :3961.3
##                                        Mean   : 59.24   Mean   :3993.4
##                                        3rd Qu.: 79.96   3rd Qu.:5577.1
##                                        Max.   :116.74   Max.   :8567.9
```

cleaning the data

```
new_churn_data <- new_churn_data %>%
  select(-customerID) %>%
  mutate(SeniorCitizen = as.factor(ifelse(new_churn_data$SeniorCitizen==1, 'Yes', 'No'))) %>%
  clean_names()%>%
  mutate_if(is.character , as.factor) %>%
  na.omit()
head(new_churn_data)%>%
  gt() %>%
  gt_theme_excel()
```

gender	senior_citizen	partner	dependents	tenure	phone_service	multiple_lines	internet_service	online
Male	Yes	No	No	51	Yes	No	DSL	
Male	Yes	Yes	Yes	61	No	Yes	Fiber optic	No inte
Male	No	Yes	No	57	Yes	No phone service	No	
Female	Yes	Yes	Yes	51	Yes	Yes	DSL	
Male	Yes	Yes	Yes	11	No	Yes	DSL	
Female	Yes	Yes	No	38	No	No	DSL	

spliting the data

```
new_churn_split <-initial_split(new_churn_data,
                                prop = 0.8 )
new_churn_train <- training(new_churn_split)
new_churn_test <- testing(new_churn_split)
new_churn_split
```

```
## <Training/Testing/Total>
## <40/10/50>
```

baking the data

```
new_churn_train_process <-bake(churn_rec,new_churn_train)
head(new_churn_train_process) %>%
  gt() %>%
  gt_theme_excel()
```

tenure	monthly_charges	total_charges	gender_Male	senior_citizen_Yes	partner_Yes	dependents_Yes	p
0.4308789	-0.6911137	-0.03487974	1	1	1	0	
-0.9121584	0.5904702	0.23122992	1	0	0	1	
0.2680865	-0.3358957	0.38790377	1	1	1	0	
-0.5458755	-0.9786870	1.59544096	1	0	1	1	
1.0006524	-0.3818808	2.33855162	1	0	1	0	
1.4483315	-1.0469982	0.44216976	1	0	1	0	

predictions

```
new_churn_pred <- predict(logit_fit, new_churn_train_process)
```

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
## prediction from rank-deficient fit; attr(*, "non-estim") has doubtful cases
```

```
head(new_churn_pred)%>%
  gt() %>%
  gt_theme_excel()
```

.pred_class
0
0
1
1
0
0

results

```
new_churn_results <- new_churn_train_process %>%
     dplyr::bind_cols(new_churn_pred)
```