R Notebook

1. Pre-requisites

loading the tidyverse packages for data visualisation and manipulation.Using the `tidyverse` library.

```
library("tidyverse")
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

2. Loading the dataset and preview its summarized information

```
MTN_df <- read_csv("https://bit.ly/2ZlpzjF")
```

```
## Rows: 7050 Columns: 21
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (17): customerID, GENDER, PARTNER, Dependents, PhoneService, MultipleLin...
## dbl  (4): SeniorCitizen, tenure, MonthlyCharges, TotalCharges
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(MTN_df)
```

```
## # A tibble: 6 x 21
##   custom~1 GENDER Senio~2 PARTNER Depen~3 tenure Phone~4 Multi~5 Inter~6 Onlin~7
##   <chr>    <chr>    <dbl> <chr>   <chr>    <dbl> <chr>   <chr>   <chr>   <chr>
## 1 7590-VH~ Female       0 Yes     No           1 No      No pho~ DSL     No
## 2 5575-GN~ Male         0 No      No          34 Yes     No      DSL     Yes
## 3 3668-QP~ Male         0 No      No           2 Yes     No      DSL     Yes
## 4 7795-CF~ Male         0 No      No          45 No      No pho~ DSL     Yes
## 5 9237-HQ~ Female       0 No      No           2 Yes     No      Fiber ~ No
## 6 9305-CD~ Female       0 No      No           8 Yes     Yes     Fiber ~ No
## # ... with 11 more variables: OnlineBackup <chr>, DeviceProtection <chr>,
## #   TECHSUPPORT <chr>, StreamingTV <chr>, StreamingMovies <chr>,
## #   Contract <chr>, PaperlessBilling <chr>, PaymentMethod <chr>,
## #   MonthlyCharges <dbl>, TotalCharges <dbl>, Churn <chr>, and abbreviated
## #   variable names 1: customerID, 2: SeniorCitizen, 3: Dependents,
## #   4: PhoneService, 5: MultipleLines, 6: InternetService, 7: OnlineSecurity
```

Checking for collumn names.

```
colnames(MTN_df)
```

```
##  [1] "customerID"       "GENDER"           "SeniorCitizen"    "PARTNER"
##  [5] "Dependents"       "tenure"           "PhoneService"     "MultipleLines"
##  [9] "InternetService"  "OnlineSecurity"   "OnlineBackup"     "DeviceProtection"
## [13] "TECHSUPPORT"      "StreamingTV"      "StreamingMovies"  "Contract"
## [17] "PaperlessBilling" "PaymentMethod"    "MonthlyCharges"   "TotalCharges"
## [21] "Churn"
```

Checking summary of the table.

```
glimpse(MTN_df)
```

```
## Rows: 7,050
## Columns: 21
## $ customerID       <chr> "7590-VHVEG", "5575-GNVDE", "3668-QPYBK", "7795-CFOCW~
## $ GENDER           <chr> "Female", "Male", "Male", "Male", "Female", "Female",~
## $ SeniorCitizen    <dbl> 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           <dbl> 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~
```

Updating Column name ‘tenure’ to ‘tenure_in_months’ for clarity.

```
colnames(MTN_df)[6]<-"tenure_in_months"
colnames(MTN_df)[1]<-"CUSTOMER_ID"
colnames(MTN_df)[3]<-"SENIOR_CITIZEN"
colnames(MTN_df)[7]<-"PHONE_SERVICE"
colnames(MTN_df)[8]<-"MULTIPLE_LINES"
colnames(MTN_df)[9]<-"INTERNET_SERVICE"
colnames(MTN_df)[10]<-"ONLINE_SECURITY"
colnames(MTN_df)[12]<-"DEVICE_PROTECTION"
colnames(MTN_df)[11]<-"ONLINE_BACKUP"
colnames(MTN_df)[13]<-"TECH_SUPPORT"
colnames(MTN_df)[14]<-"STREAMING_TV"
colnames(MTN_df)[15]<-"STREAMING_MOVIES"
colnames(MTN_df)[17]<-"PAPERLESS_BILLING"
colnames(MTN_df)[18]<-"PAYMENT_METHOD"
colnames(MTN_df)[19]<-"MONTHLY_CHARGES"
colnames(MTN_df)[20]<-"TOTAL_CHARGES"
colnames(MTN_df) <- toupper(colnames(MTN_df))
glimpse(MTN_df)
```

```
## Rows: 7,050
## Columns: 21
## $ CUSTOMER_ID       <chr> "7590-VHVEG", "5575-GNVDE", "3668-QPYBK", "7795-CFOC~
## $ GENDER            <chr> "Female", "Male", "Male", "Male", "Female", "Female"~
## $ SENIOR_CITIZEN    <dbl> 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", "Ye~
## $ DEPENDENTS        <chr> "No", "No", "No", "No", "No", "No", "Yes", "No", "No~
## $ TENURE_IN_MONTHS  <dbl> 1, 34, 2, 45, 2, 8, 22, 10, 28, 62, 13, 16, 58, 49, ~
## $ PHONE_SERVICE     <chr> "No", "Yes", "Yes", "No", "Yes", "Yes", "Yes", "No",~
## $ MULTIPLE_LINES    <chr> "No phone service", "No", "No", "No phone service", ~
## $ INTERNET_SERVICE  <chr> "DSL", "DSL", "DSL", "DSL", "Fiber optic", "Fiber op~
## $ ONLINE_SECURITY   <chr> "No", "Yes", "Yes", "Yes", "No", "No", "No", "Yes", ~
## $ ONLINE_BACKUP     <chr> "Yes", "No", "Yes", "No", "No", "No", "Yes", "No", "~
## $ DEVICE_PROTECTION <chr> "No", "Yes", "No", "Yes", "No", "Yes", "No", "No", "~
## $ TECH_SUPPORT      <chr> "No", "No", "No", "Yes", "No", "No", "No", "No", "Ye~
## $ STREAMING_TV      <chr> "No", "No", "No", "No", "No", "Yes", "Yes", "No", "Y~
## $ STREAMING_MOVIES  <chr> "No", "No", "No", "No", "No", "Yes", "No", "No", "Ye~
## $ CONTRACT          <chr> "Month-to-month", "One year", "Month-to-month", "One~
## $ PAPERLESS_BILLING <chr> "Yes", "No", "Yes", "No", "Yes", "Yes", "Yes", "No",~
## $ PAYMENT_METHOD    <chr> "Electronic check", "Mailed check", "Mailed check", ~
## $ 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             <chr> "No", "No", "Yes", "No", "Yes", "Yes", "No", "No", "~
```

Checking for duplicates.

```
sum(duplicated(MTN_df))
```

```
## [1] 7
```

Removing Duplicates

```
distinct(MTN_df)
```

```
## # A tibble: 7,043 x 21
##    CUSTOMER_ID GENDER SENIOR_C~1 PARTNER DEPEN~2 TENUR~3 PHONE~4 MULTI~5 INTER~6
##    <chr>       <chr>       <dbl> <chr>   <chr>     <dbl> <chr>   <chr>   <chr>
##  1 7590-VHVEG  Female          0 Yes     No            1 No      No pho~ DSL
##  2 5575-GNVDE  Male            0 No      No           34 Yes     No      DSL
##  3 3668-QPYBK  Male            0 No      No            2 Yes     No      DSL
##  4 7795-CFOCW  Male            0 No      No           45 No      No pho~ DSL
##  5 9237-HQITU  Female          0 No      No            2 Yes     No      Fiber ~
##  6 9305-CDSKC  Female          0 No      No            8 Yes     Yes     Fiber ~
##  7 1452-KIOVK  Male            0 No      Yes          22 Yes     Yes     Fiber ~
##  8 6713-OKOMC  Female          0 No      No           10 No      No pho~ DSL
##  9 7892-POOKP  Female          0 Yes     No           28 Yes     Yes     Fiber ~
## 10 6388-TABGU  Male            0 No      Yes          62 Yes     No      DSL
## # ... with 7,033 more rows, 12 more variables: ONLINE_SECURITY <chr>,
## #   ONLINE_BACKUP <chr>, DEVICE_PROTECTION <chr>, TECH_SUPPORT <chr>,
## #   STREAMING_TV <chr>, STREAMING_MOVIES <chr>, CONTRACT <chr>,
## #   PAPERLESS_BILLING <chr>, PAYMENT_METHOD <chr>, MONTHLY_CHARGES <dbl>,
## #   TOTAL_CHARGES <dbl>, CHURN <chr>, and abbreviated variable names
## #   1: SENIOR_CITIZEN, 2: DEPENDENTS, 3: TENURE_IN_MONTHS, 4: PHONE_SERVICE,
## #   5: MULTIPLE_LINES, 6: INTERNET_SERVICE
```

Checking for no of missing values and deleting missing values.

```
sum(is.na(MTN_df))
```

```
## [1] 251
```

```
na.omit(MTN_df)
```

```
## # A tibble: 7,010 x 21
##    CUSTOMER_ID GENDER SENIOR_C~1 PARTNER DEPEN~2 TENUR~3 PHONE~4 MULTI~5 INTER~6
##    <chr>       <chr>       <dbl> <chr>   <chr>     <dbl> <chr>   <chr>   <chr>
##  1 7590-VHVEG  Female          0 Yes     No            1 No      No pho~ DSL
##  2 5575-GNVDE  Male            0 No      No           34 Yes     No      DSL
##  3 3668-QPYBK  Male            0 No      No            2 Yes     No      DSL
##  4 7795-CFOCW  Male            0 No      No           45 No      No pho~ DSL
##  5 9237-HQITU  Female          0 No      No            2 Yes     No      Fiber ~
##  6 9305-CDSKC  Female          0 No      No            8 Yes     Yes     Fiber ~
##  7 1452-KIOVK  Male            0 No      Yes          22 Yes     Yes     Fiber ~
##  8 6713-OKOMC  Female          0 No      No           10 No      No pho~ DSL
##  9 7892-POOKP  Female          0 Yes     No           28 Yes     Yes     Fiber ~
## 10 6388-TABGU  Male            0 No      Yes          62 Yes     No      DSL
## # ... with 7,000 more rows, 12 more variables: ONLINE_SECURITY <chr>,
## #   ONLINE_BACKUP <chr>, DEVICE_PROTECTION <chr>, TECH_SUPPORT <chr>,
## #   STREAMING_TV <chr>, STREAMING_MOVIES <chr>, CONTRACT <chr>,
## #   PAPERLESS_BILLING <chr>, PAYMENT_METHOD <chr>, MONTHLY_CHARGES <dbl>,
## #   TOTAL_CHARGES <dbl>, CHURN <chr>, and abbreviated variable names
## #   1: SENIOR_CITIZEN, 2: DEPENDENTS, 3: TENURE_IN_MONTHS, 4: PHONE_SERVICE,
## #   5: MULTIPLE_LINES, 6: INTERNET_SERVICE
```

Checking for unique values.

```
unique(MTN_df)
```

```
## # A tibble: 7,043 x 21
##    CUSTOMER_ID GENDER SENIOR_C~1 PARTNER DEPEN~2 TENUR~3 PHONE~4 MULTI~5 INTER~6
##    <chr>       <chr>       <dbl> <chr>   <chr>     <dbl> <chr>   <chr>   <chr>
##  1 7590-VHVEG  Female          0 Yes     No            1 No      No pho~ DSL
##  2 5575-GNVDE  Male            0 No      No           34 Yes     No      DSL
##  3 3668-QPYBK  Male            0 No      No            2 Yes     No      DSL
##  4 7795-CFOCW  Male            0 No      No           45 No      No pho~ DSL
##  5 9237-HQITU  Female          0 No      No            2 Yes     No      Fiber ~
##  6 9305-CDSKC  Female          0 No      No            8 Yes     Yes     Fiber ~
##  7 1452-KIOVK  Male            0 No      Yes          22 Yes     Yes     Fiber ~
##  8 6713-OKOMC  Female          0 No      No           10 No      No pho~ DSL
##  9 7892-POOKP  Female          0 Yes     No           28 Yes     Yes     Fiber ~
## 10 6388-TABGU  Male            0 No      Yes          62 Yes     No      DSL
## # ... with 7,033 more rows, 12 more variables: ONLINE_SECURITY <chr>,
## #   ONLINE_BACKUP <chr>, DEVICE_PROTECTION <chr>, TECH_SUPPORT <chr>,
## #   STREAMING_TV <chr>, STREAMING_MOVIES <chr>, CONTRACT <chr>,
## #   PAPERLESS_BILLING <chr>, PAYMENT_METHOD <chr>, MONTHLY_CHARGES <dbl>,
## #   TOTAL_CHARGES <dbl>, CHURN <chr>, and abbreviated variable names
## #   1: SENIOR_CITIZEN, 2: DEPENDENTS, 3: TENURE_IN_MONTHS, 4: PHONE_SERVICE,
## #   5: MULTIPLE_LINES, 6: INTERNET_SERVICE
```

Checking for dataframe datatypes.

```
str(MTN_df)
```

```
## spec_tbl_df [7,050 x 21] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ CUSTOMER_ID      : chr [1:7050] "7590-VHVEG" "5575-GNVDE" "3668-QPYBK" "7795-CFOCW" ...
##  $ GENDER           : chr [1:7050] "Female" "Male" "Male" "Male" ...
##  $ SENIOR_CITIZEN   : num [1:7050] 0 0 0 0 0 0 0 0 0 0 ...
##  $ PARTNER          : chr [1:7050] "Yes" "No" "No" "No" ...
##  $ DEPENDENTS       : chr [1:7050] "No" "No" "No" "No" ...
##  $ TENURE_IN_MONTHS : num [1:7050] 1 34 2 45 2 8 22 10 28 62 ...
##  $ PHONE_SERVICE    : chr [1:7050] "No" "Yes" "Yes" "No" ...
##  $ MULTIPLE_LINES   : chr [1:7050] "No phone service" "No" "No" "No phone service" ...
##  $ INTERNET_SERVICE : chr [1:7050] "DSL" "DSL" "DSL" "DSL" ...
##  $ ONLINE_SECURITY  : chr [1:7050] "No" "Yes" "Yes" "Yes" ...
##  $ ONLINE_BACKUP    : chr [1:7050] "Yes" "No" "Yes" "No" ...
##  $ DEVICE_PROTECTION: chr [1:7050] "No" "Yes" "No" "Yes" ...
##  $ TECH_SUPPORT     : chr [1:7050] "No" "No" "No" "Yes" ...
##  $ STREAMING_TV     : chr [1:7050] "No" "No" "No" "No" ...
##  $ STREAMING_MOVIES : chr [1:7050] "No" "No" "No" "No" ...
##  $ CONTRACT         : chr [1:7050] "Month-to-month" "One year" "Month-to-month" "One year" ...
##  $ PAPERLESS_BILLING: chr [1:7050] "Yes" "No" "Yes" "No" ...
##  $ PAYMENT_METHOD   : chr [1:7050] "Electronic check" "Mailed check" "Mailed check" "Bank transfer (a
##  $ MONTHLY_CHARGES  : num [1:7050] 29.9 57 53.9 42.3 70.7 ...
##  $ TOTAL_CHARGES    : num [1:7050] 29.9 1889.5 108.2 1840.8 151.7 ...
##  $ CHURN            : chr [1:7050] "No" "No" "Yes" "No" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   customerID = col_character(),
##   ..   GENDER = col_character(),
##   ..   SeniorCitizen = col_double(),
##   ..   PARTNER = col_character(),
##   ..   Dependents = col_character(),
##   ..   tenure = col_double(),
##   ..   PhoneService = col_character(),
##   ..   MultipleLines = col_character(),
##   ..   InternetService = col_character(),
##   ..   OnlineSecurity = col_character(),
##   ..   OnlineBackup = col_character(),
##   ..   DeviceProtection = col_character(),
##   ..   TECHSUPPORT = col_character(),
##   ..   StreamingTV = col_character(),
##   ..   StreamingMovies = col_character(),
##   ..   Contract = col_character(),
##   ..   PaperlessBilling = col_character(),
##   ..   PaymentMethod = col_character(),
##   ..   MonthlyCharges = col_double(),
##   ..   TotalCharges = col_double(),
##   ..   Churn = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

# Resolving the issues in the ‘senior_citizen’ variable by converting the values to either ‘YES’/‘NO’

```
MTN_df$SENIOR_CITIZEN[MTN_df$SENIOR_CITIZEN == '0'] <- 'NO'
MTN_df$SENIOR_CITIZEN[MTN_df$SENIOR_CITIZEN == '1'] <- 'YES'
sample(MTN_df)
```

```
## # A tibble: 7,050 x 21
##    DEVICE_PROTE~1 MONTH~2 TENUR~3 GENDER TECH_~4 DEPEN~5 PARTNER SENIO~6 ONLIN~7
##    <chr>            <dbl>   <dbl> <chr>  <chr>   <chr>   <chr>   <chr>   <chr>
##  1 No                29.8       1 Female No      No      Yes     NO      No
##  2 Yes               57.0      34 Male   No      No      No      NO      Yes
##  3 No                53.8       2 Male   No      No      No      NO      Yes
##  4 Yes               42.3      45 Male   Yes     No      No      NO      Yes
##  5 No                70.7       2 Female No      No      No      NO      No
##  6 Yes               99.6       8 Female No      No      No      NO      No
##  7 No                89.1      22 Male   No      Yes     No      NO      No
##  8 No                29.8      10 Female No      No      No      NO      Yes
##  9 Yes              105.       28 Female Yes     No      Yes     NO      No
## 10 No                56.2      62 Male   No      Yes     No      NO      Yes
## # ... with 7,040 more rows, 12 more variables: INTERNET_SERVICE <chr>,
## #   PHONE_SERVICE <chr>, MULTIPLE_LINES <chr>, PAYMENT_METHOD <chr>,
## #   TOTAL_CHARGES <dbl>, CHURN <chr>, CONTRACT <chr>, STREAMING_MOVIES <chr>,
## #   STREAMING_TV <chr>, ONLINE_BACKUP <chr>, CUSTOMER_ID <chr>,
## #   PAPERLESS_BILLING <chr>, and abbreviated variable names
## #   1: DEVICE_PROTECTION, 2: MONTHLY_CHARGES, 3: TENURE_IN_MONTHS,
## #   4: TECH_SUPPORT, 5: DEPENDENTS, 6: SENIOR_CITIZEN, 7: ONLINE_SECURITY
```

resolving the issues in the ‘payment_method’ variable

- Replace ‘Mailed checkkk’ with ‘Mailed check’,

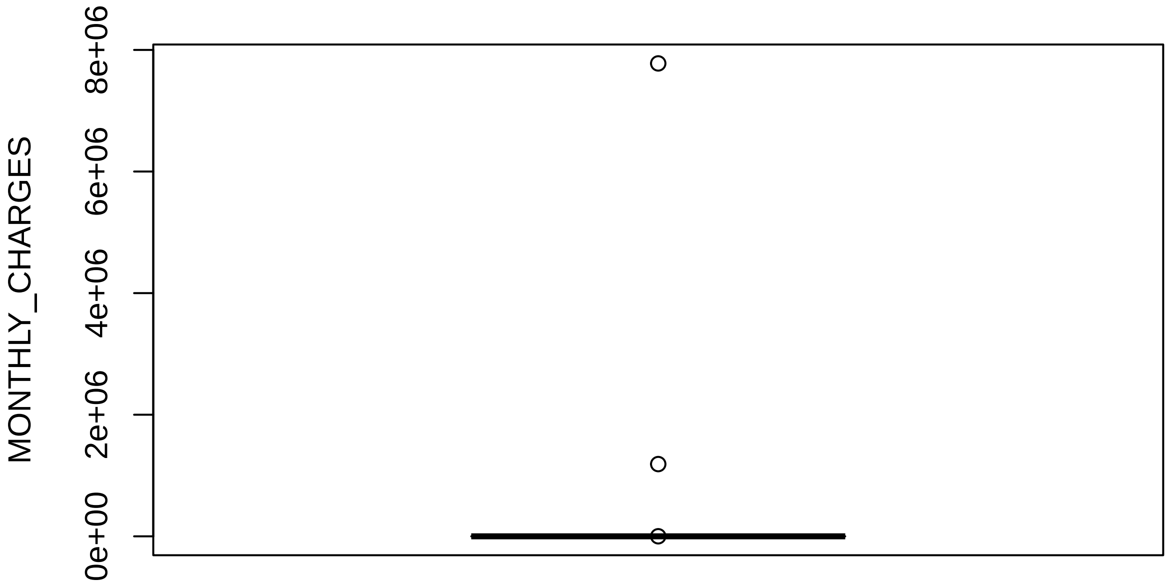
- Replace ‘Electronic checkk’, ‘Electronic check’

```
MTN_df$PAYMENT_METHOD[MTN_df$PAYMENT_METHOD == 'Mailed checkkk'] <- 'Mailed check'
MTN_df$PAYMENT_METHOD[MTN_df$PAYMENT_METHOD == 'Electronic checkk'] <- 'Electronic check'
# MTN_df$PAYMENT_METHOD
unique(MTN_df$PAYMENT_METHOD)
```

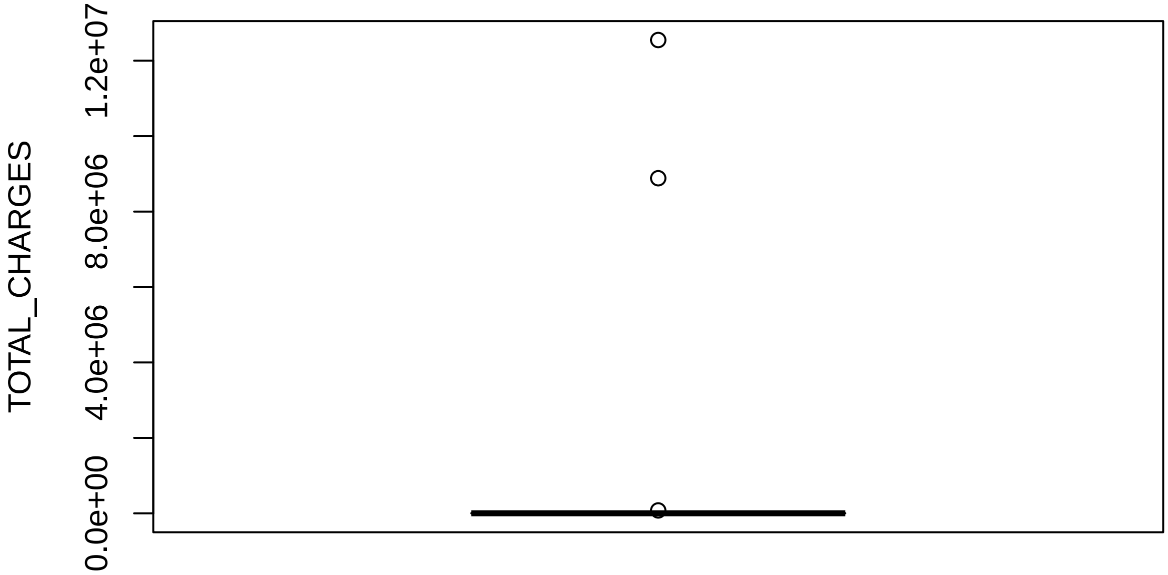
```
## [1] "Electronic check"          "Mailed check"
## [3] "Bank transfer (automatic)" "Credit card (automatic)"
## [5] NA
```

Visual distribution of the outliers using a box plot for the ‘Tenure’, ‘monthly_charges’ and ‘total_charges’

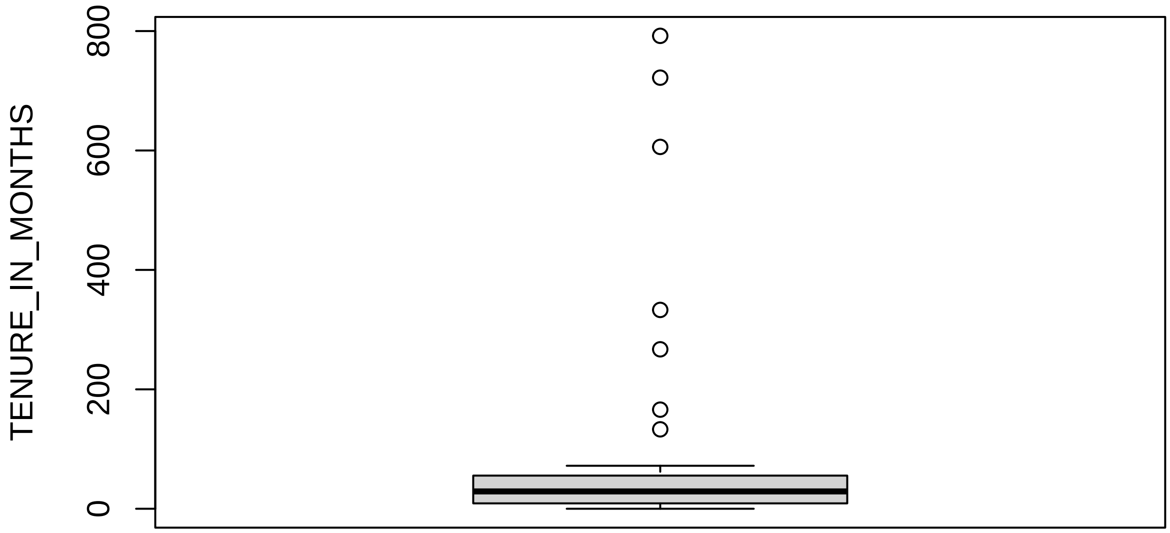
```
boxplot(MTN_df$MONTHLY_CHARGES,
  ylab = "MONTHLY_CHARGES"
)
```



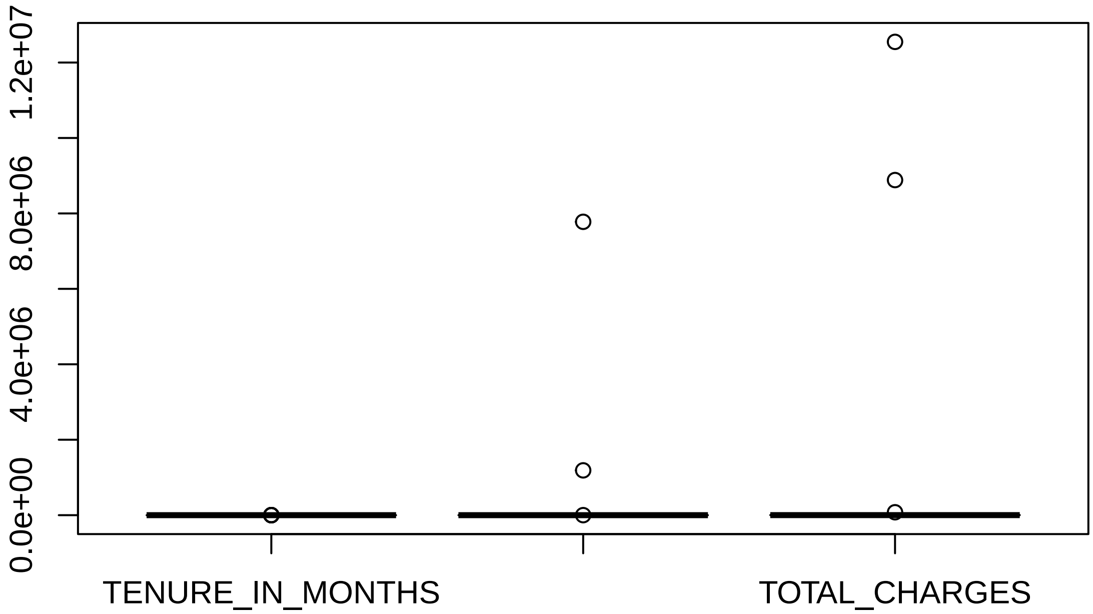
```
boxplot(MTN_df$TOTAL_CHARGES,
  ylab = "TOTAL_CHARGES"
)
```



```
boxplot(MTN_df$TENURE_IN_MONTHS,
  ylab = "TENURE_IN_MONTHS"
)
```



```
boxplot(MTN_df[,c('TENURE_IN_MONTHS', 'MONTHLY_CHARGES','TOTAL_CHARGES')])
```



# From the above visualization, it is clear that the data variables ‘TENURE_IN_MONTHS’, ‘MONTHLY_CHARGES’,’TOTAL_CHARG outliers in the data values. #OUTLIER ANALYSIS – Removal of Outliers # 1. From the boxplot, the presence of outliers are evident. That is, the data values that are present above the upper quartile can be considered as the outlier data values. # 2. Now, we will replace the outlier data values with NULL.

Replacing Outliers with NULL Values;

```
for (x in c('TENURE_IN_MONTHS', 'MONTHLY_CHARGES','TOTAL_CHARGES'))
{
  value = MTN_df[,x][MTN_df[,x] %in% boxplot.stats(MTN_df[,x])$out]
  MTN_df[,x][MTN_df[,x] %in% value] = NA
}
#Checking whether the outliers in the above defined columns are replaced by NULL or not
sum(is.na(MTN_df$TENURE_IN_MONTHS))
```

```
## [1] 11
```

```
sum(is.na(MTN_df$MONTHLY_CHARGES))
```

```
## [1] 12
```

```
sum(is.na(MTN_df$TOTAL_CHARGES))
```

```
## [1] 23
```

```
as.data.frame(colSums(is.na(MTN_df)))
```

```
##                   colSums(is.na(MTN_df))
## CUSTOMER_ID                            0
## GENDER                                 1
## SENIOR_CITIZEN                         3
## PARTNER                               12
## DEPENDENTS                            10
## TENURE_IN_MONTHS                      11
## PHONE_SERVICE                         15
## MULTIPLE_LINES                        17
## INTERNET_SERVICE                      16
## ONLINE_SECURITY                       16
## ONLINE_BACKUP                         15
## DEVICE_PROTECTION                     14
## TECH_SUPPORT                          13
## STREAMING_TV                          13
## STREAMING_MOVIES                      12
## CONTRACT                              12
## PAPERLESS_BILLING                     12
## PAYMENT_METHOD                        12
## MONTHLY_CHARGES                       12
## TOTAL_CHARGES                         23
## CHURN                                 12
```

```
colSums(is.na(MTN_df))
```

```
##       CUSTOMER_ID            GENDER    SENIOR_CITIZEN           PARTNER
##                 0                 1                 3                12
##        DEPENDENTS  TENURE_IN_MONTHS     PHONE_SERVICE    MULTIPLE_LINES
##                10                11                15                17
##  INTERNET_SERVICE   ONLINE_SECURITY     ONLINE_BACKUP DEVICE_PROTECTION
##                16                16                15                14
##      TECH_SUPPORT      STREAMING_TV  STREAMING_MOVIES          CONTRACT
##                13                13                12                12
## PAPERLESS_BILLING    PAYMENT_METHOD   MONTHLY_CHARGES     TOTAL_CHARGES
##                12                12                12                23
##             CHURN
##                12
```

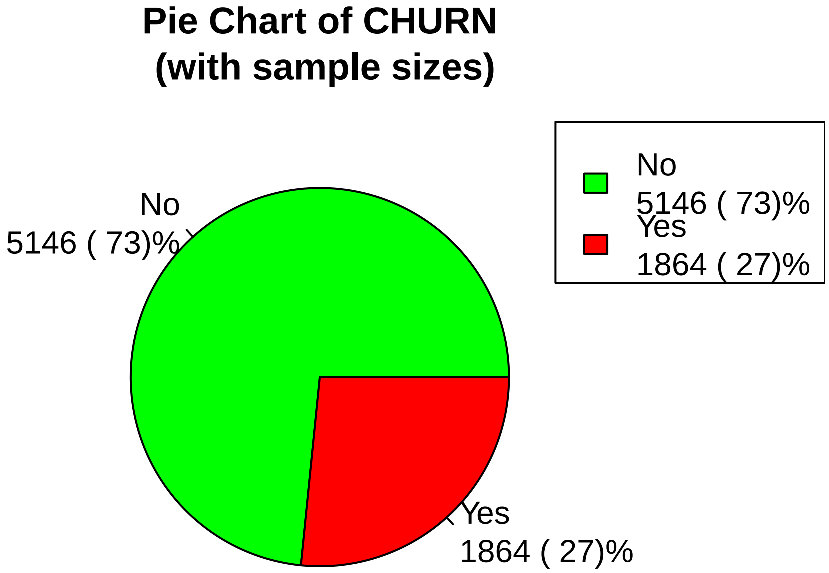
Checking for the presence of missing data i.e. whether the outlier values have been converted to missing values properly using the sum(is.na()) function.

```
MTN_df = drop_na(MTN_df)
as.data.frame(colSums(is.na(MTN_df)))
```

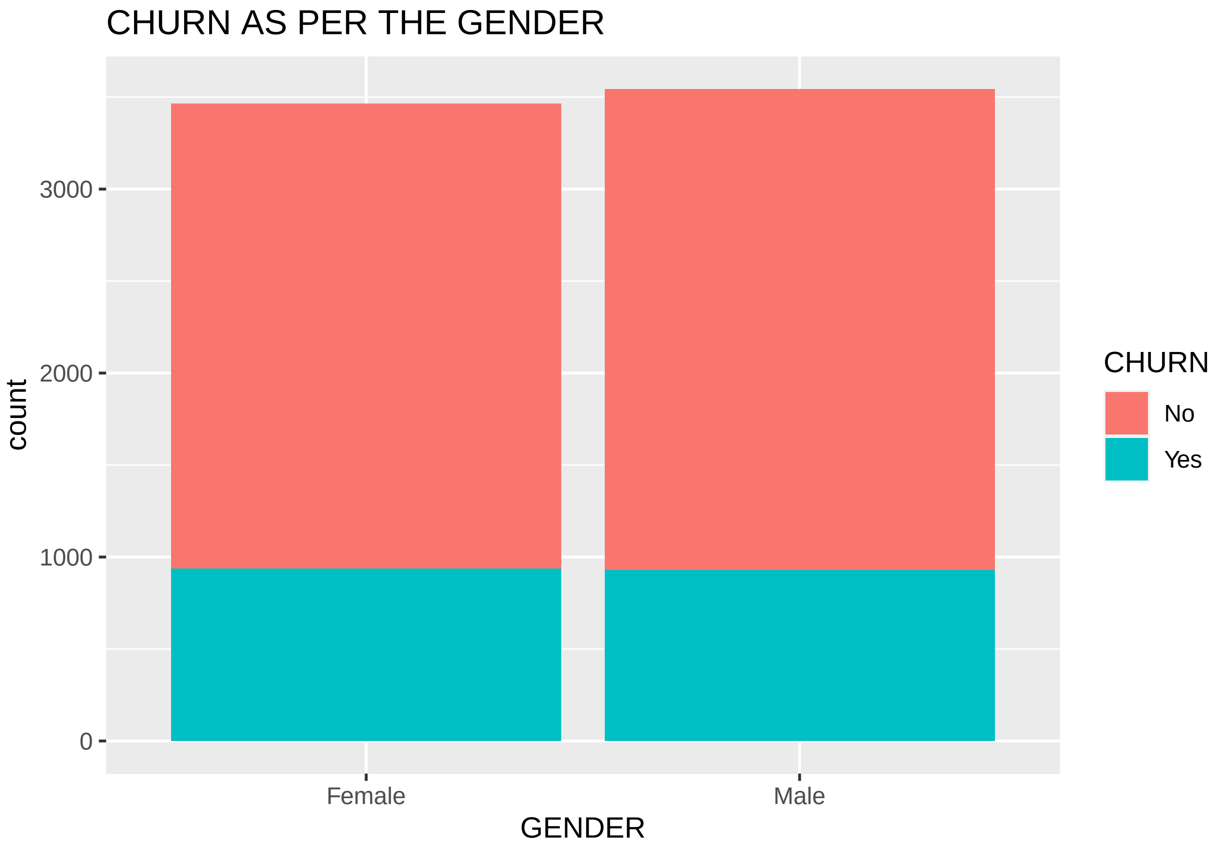
```
##                   colSums(is.na(MTN_df))
## CUSTOMER_ID                            0
## GENDER                                 0
## SENIOR_CITIZEN                         0
## PARTNER                                0
## DEPENDENTS                             0
## TENURE_IN_MONTHS                       0
## PHONE_SERVICE                          0
## MULTIPLE_LINES                         0
## INTERNET_SERVICE                       0
## ONLINE_SECURITY                        0
## ONLINE_BACKUP                          0
## DEVICE_PROTECTION                      0
## TECH_SUPPORT                           0
## STREAMING_TV                           0
## STREAMING_MOVIES                       0
## CONTRACT                               0
## PAPERLESS_BILLING                      0
## PAYMENT_METHOD                         0
## MONTHLY_CHARGES                        0
## TOTAL_CHARGES                          0
## CHURN                                  0
```

What percentage of customers from our dataset churned?

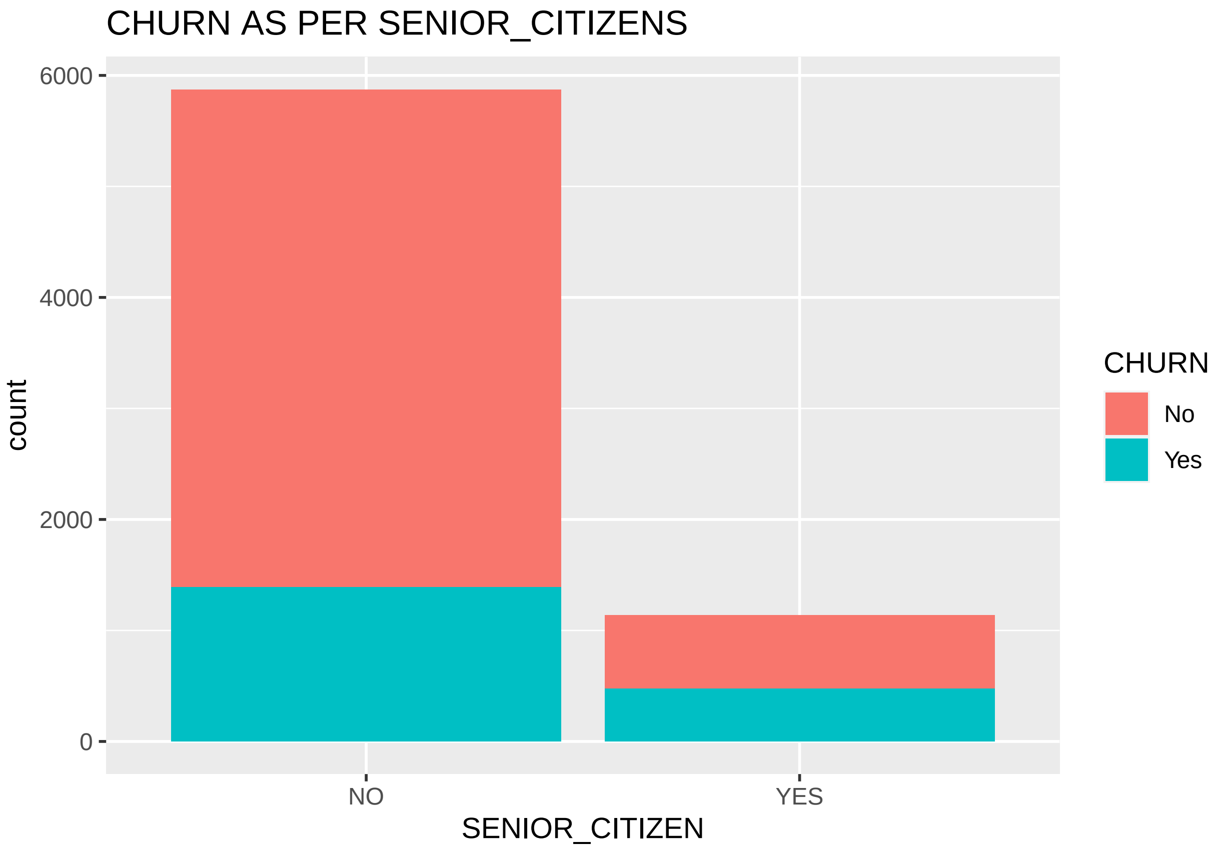
```
# Create a vector of labels
mytable <- table(MTN_df$CHURN)
lbls <- paste(names(mytable), "\n", mytable, sep="")
colors <- c("green", "red")
pct <- round(mytable/sum(mytable)*100)
lbls <- paste(lbls, "(",pct) # add percents to labels
lbls <- paste(lbls,"%",sep=")") # ad % to labels
pie(mytable, labels = lbls,col = colors,
   main="Pie Chart of CHURN\n (with sample sizes)")
legend("topright", lbls, fill = colors)
```



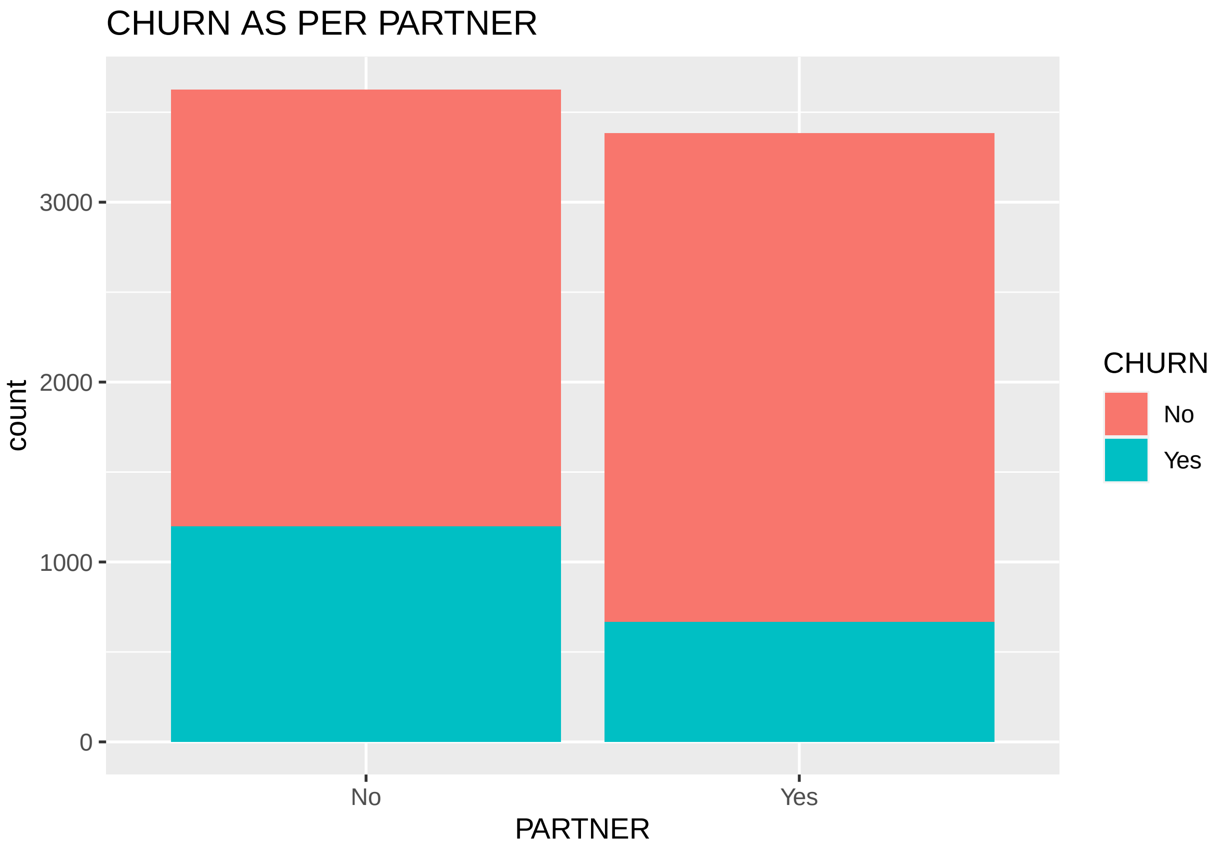
```
ggplot(MTN_df, aes(x=GENDER,fill=CHURN))+ geom_bar() + labs(title="CHURN AS PER THE GENDER", x="GENDER"
```



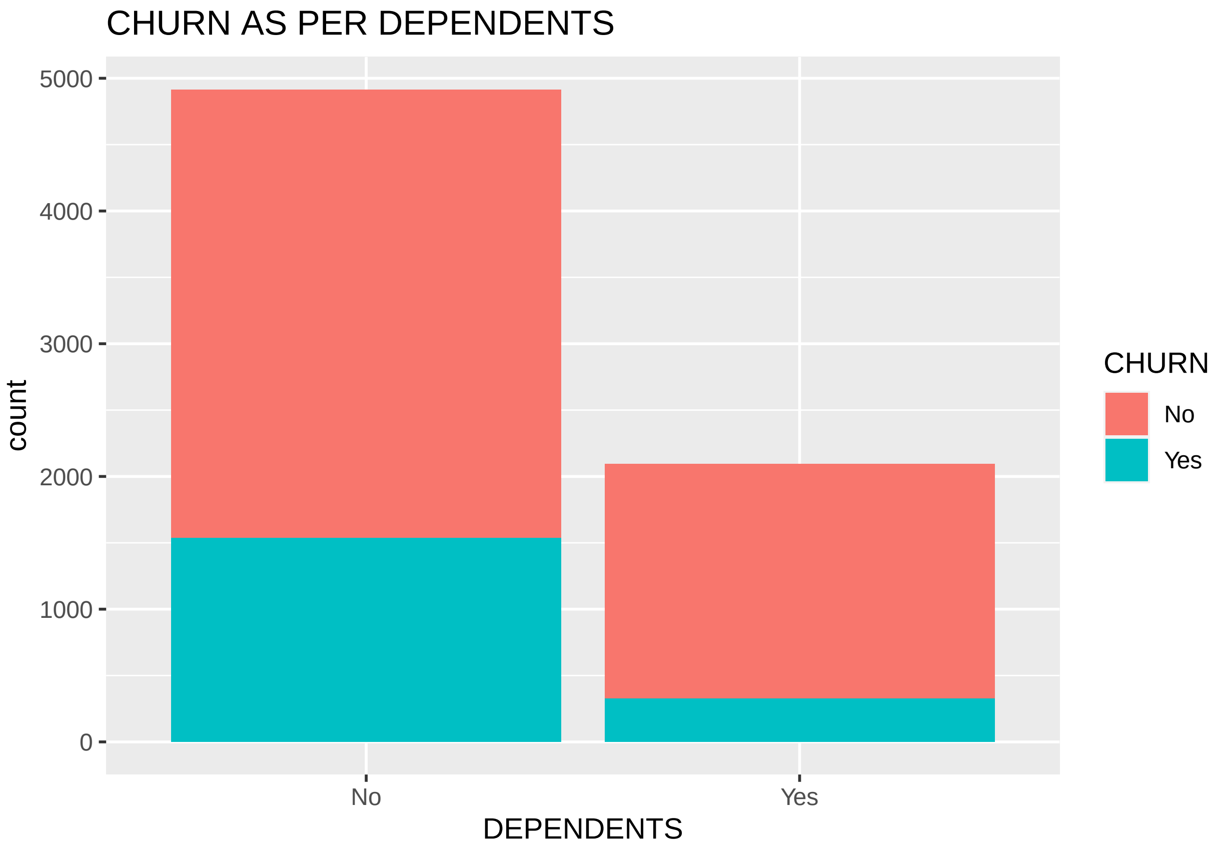
```
ggplot(MTN_df, aes(x=SENIOR_CITIZEN,fill=CHURN))+ geom_bar() + labs(title="CHURN AS PER SENIOR_CITIZENS
```



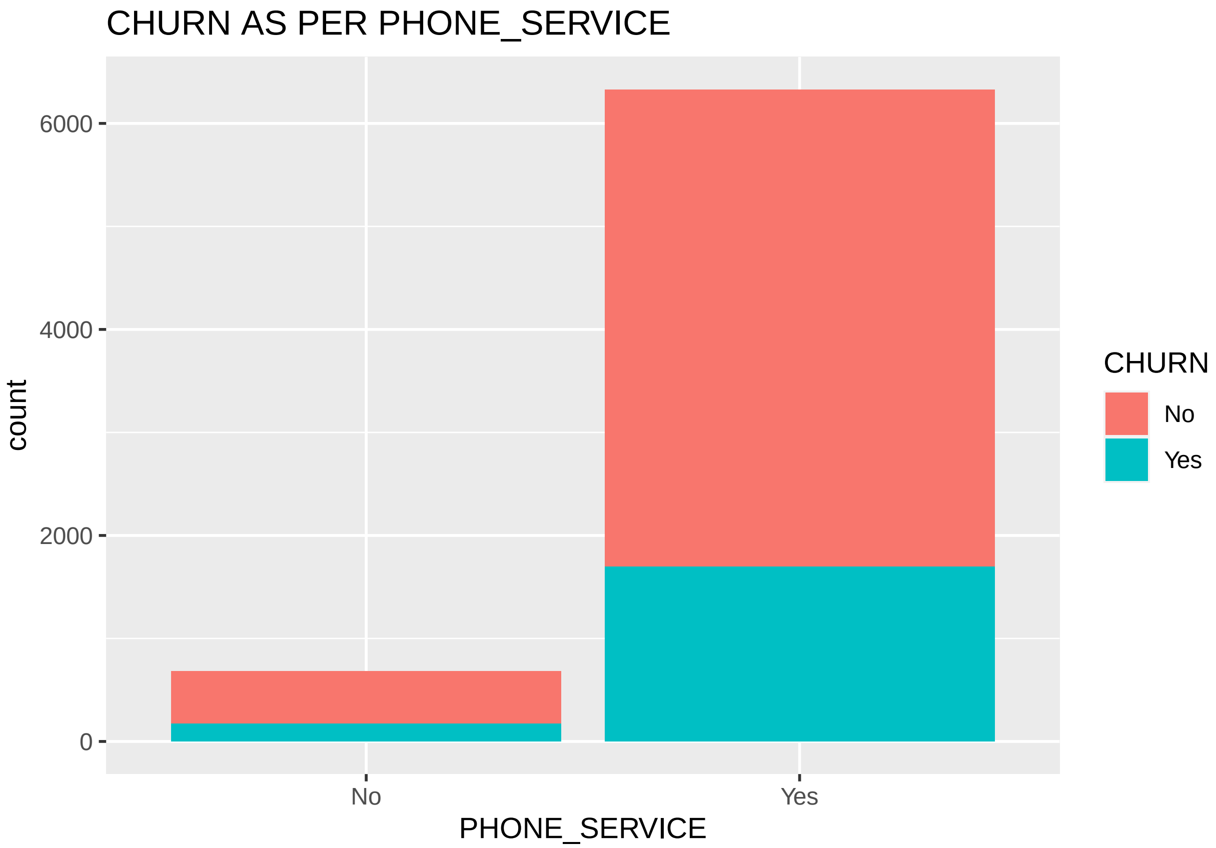
```
ggplot(MTN_df, aes(x=PARTNER,fill=CHURN))+ geom_bar() + labs(title="CHURN AS PER PARTNER", x="PARTNER",
```



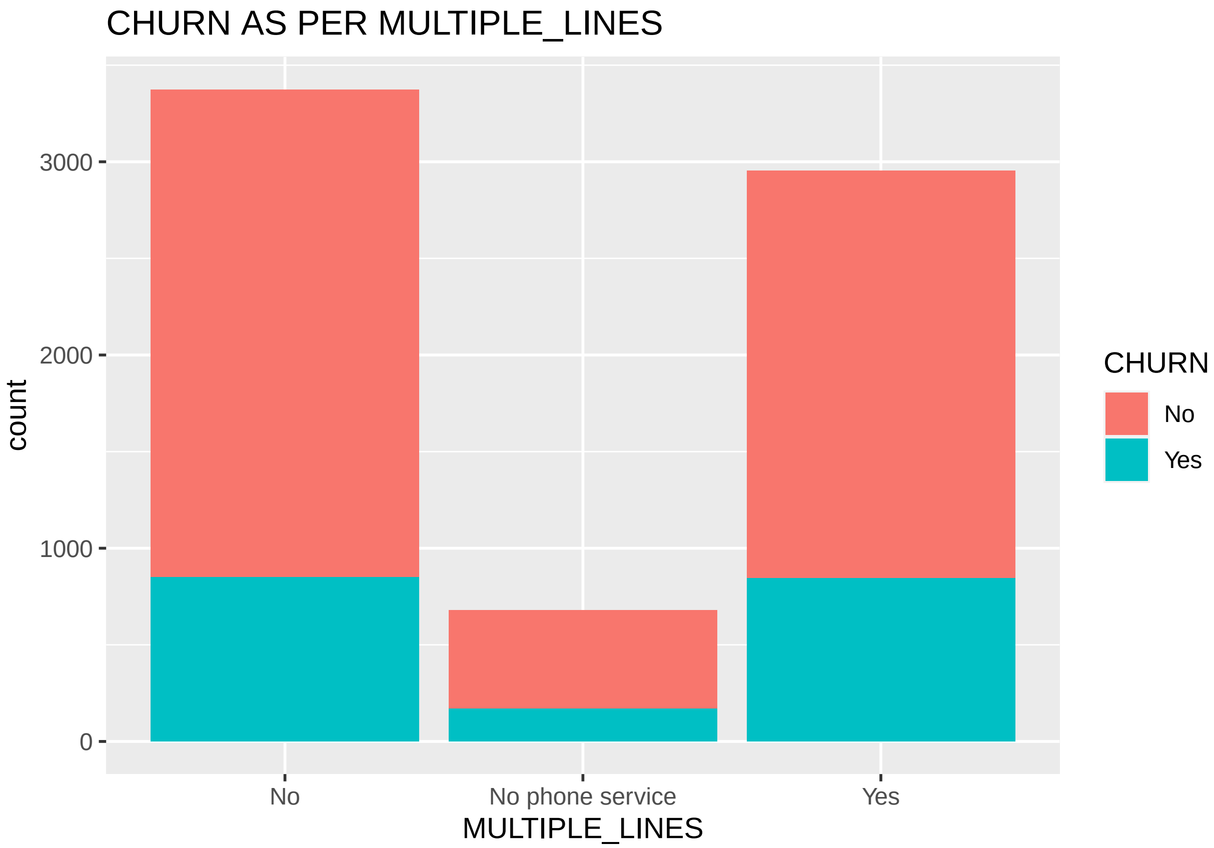
```
ggplot(MTN_df, aes(x=DEPENDENTS,fill=CHURN))+ geom_bar() + labs(title="CHURN AS PER DEPENDENTS", x="DEP
```



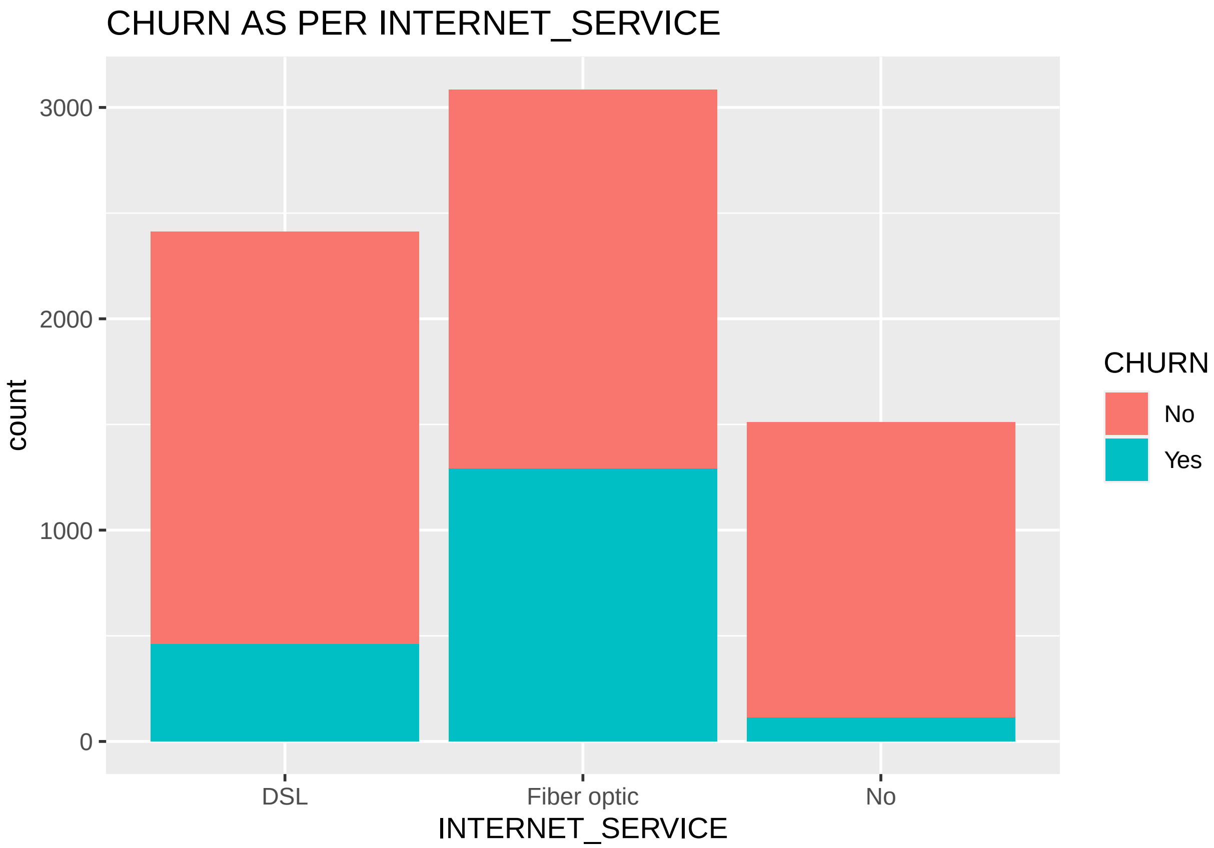
```
ggplot(MTN_df, aes(x=PHONE_SERVICE,fill=CHURN))+ geom_bar() + labs(title="CHURN AS PER PHONE_SERVICE",
```



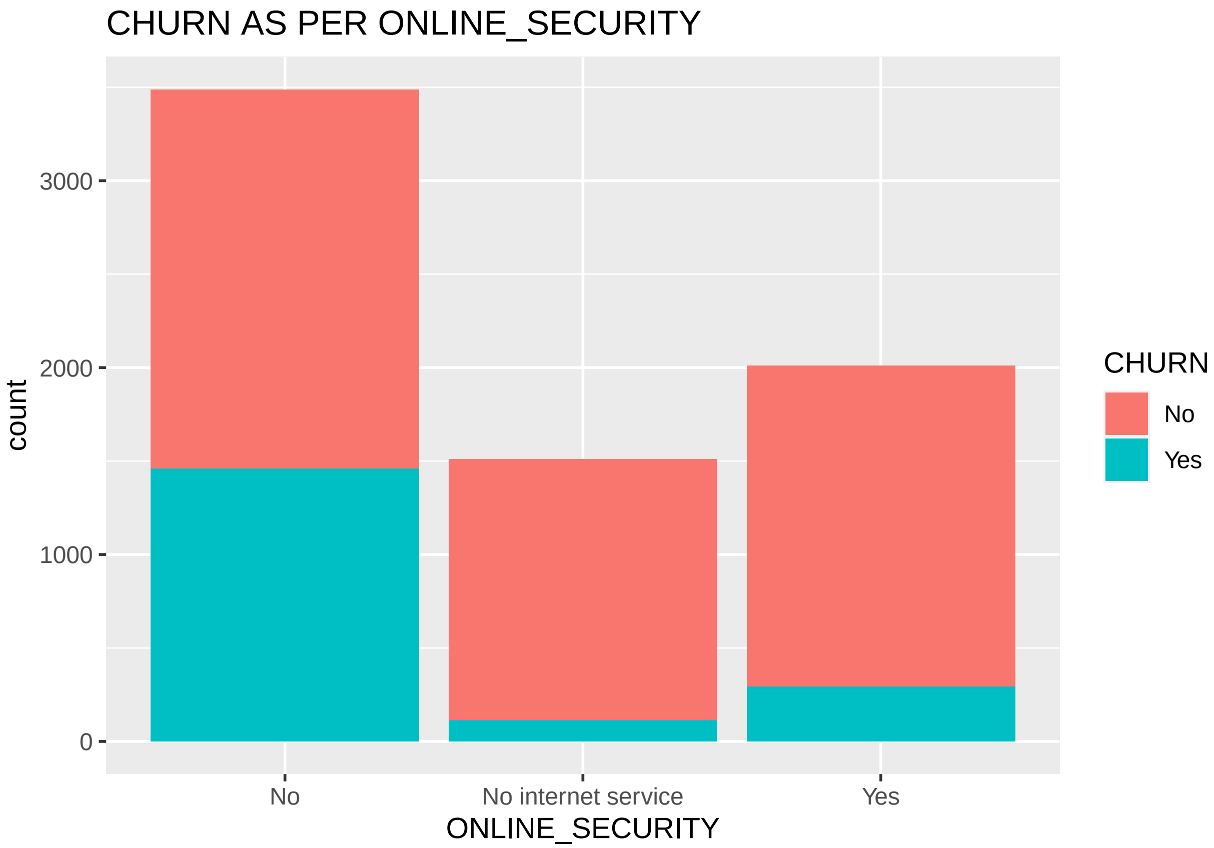
```
ggplot(MTN_df, aes(x=MULTIPLE_LINES,fill=CHURN))+ geom_bar() + labs(title="CHURN AS PER MULTIPLE_LINES"
```



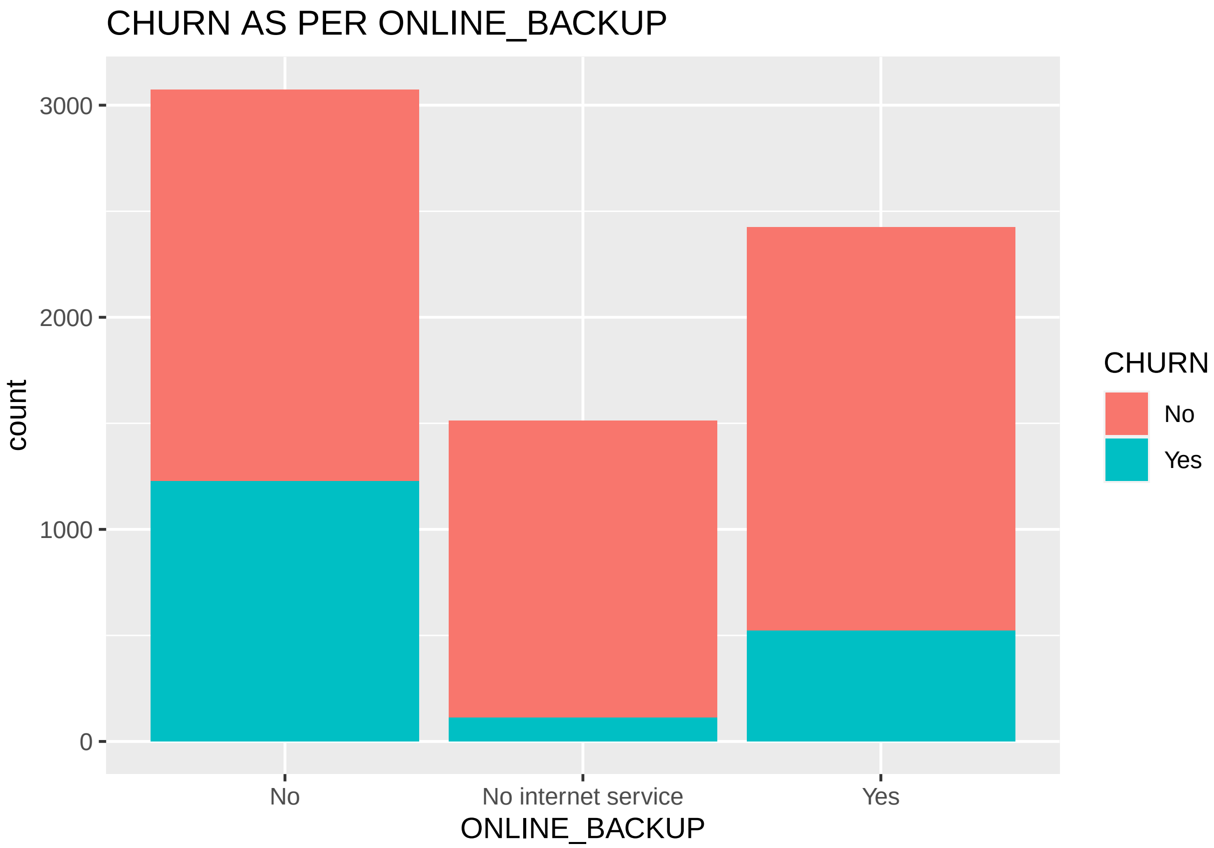
```
ggplot(MTN_df, aes(x=INTERNET_SERVICE,fill=CHURN))+ geom_bar() + labs(title="CHURN AS PER INTERNET_SERV
```



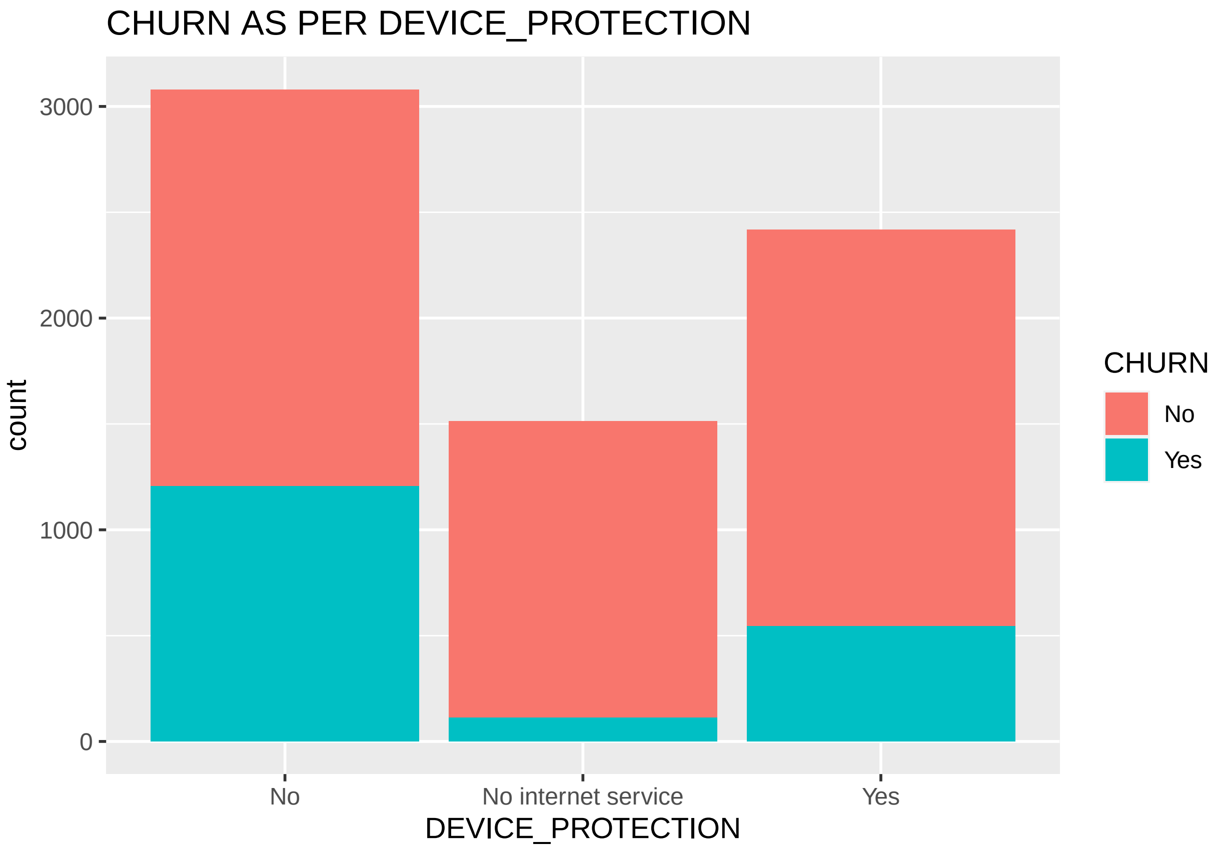
```
ggplot(MTN_df, aes(x=ONLINE_SECURITY,fill=CHURN))+ geom_bar() + labs(title="CHURN AS PER ONLINE_SECURIT
```



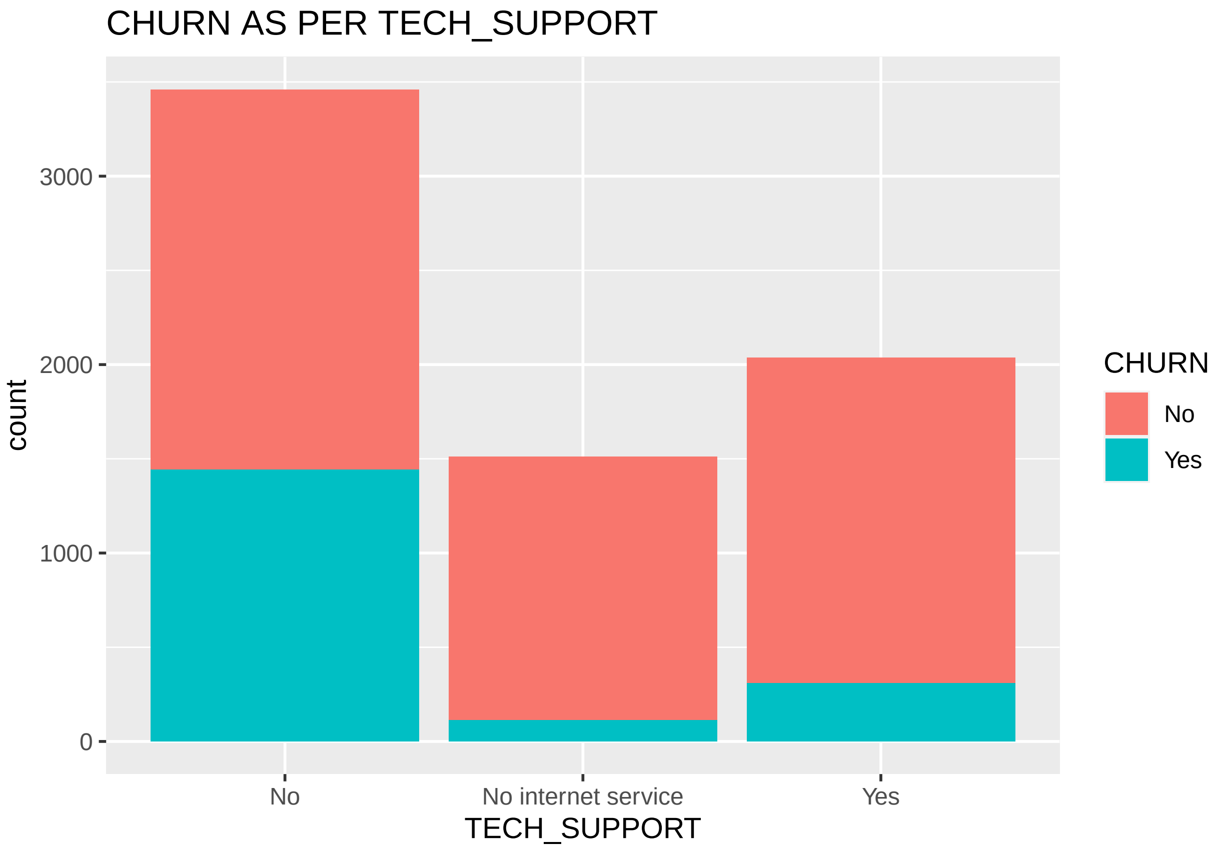
```
ggplot(MTN_df, aes(x=ONLINE_BACKUP,fill=CHURN))+ geom_bar() + labs(title="CHURN AS PER ONLINE_BACKUP",
```



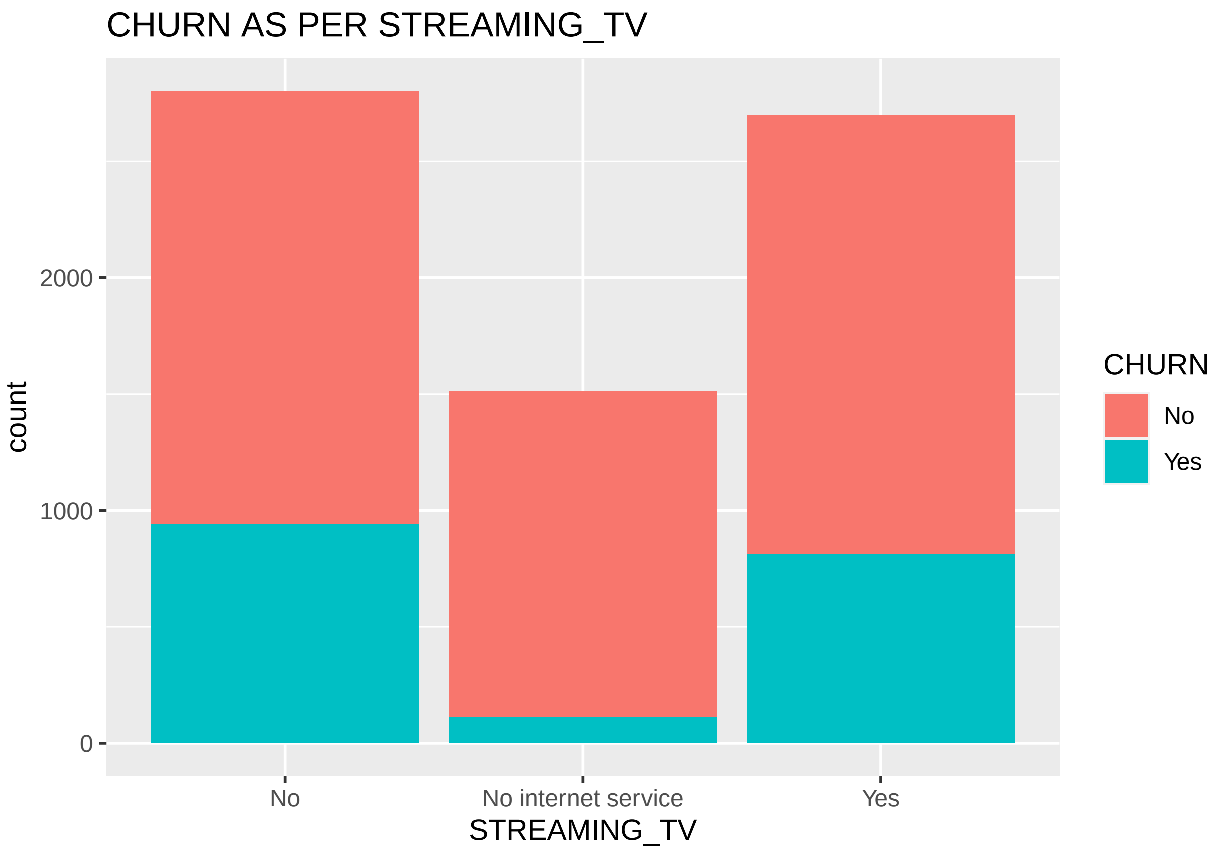
```
ggplot(MTN_df, aes(x=DEVICE_PROTECTION,fill=CHURN))+ geom_bar() + labs(title="CHURN AS PER DEVICE_PROTE
```



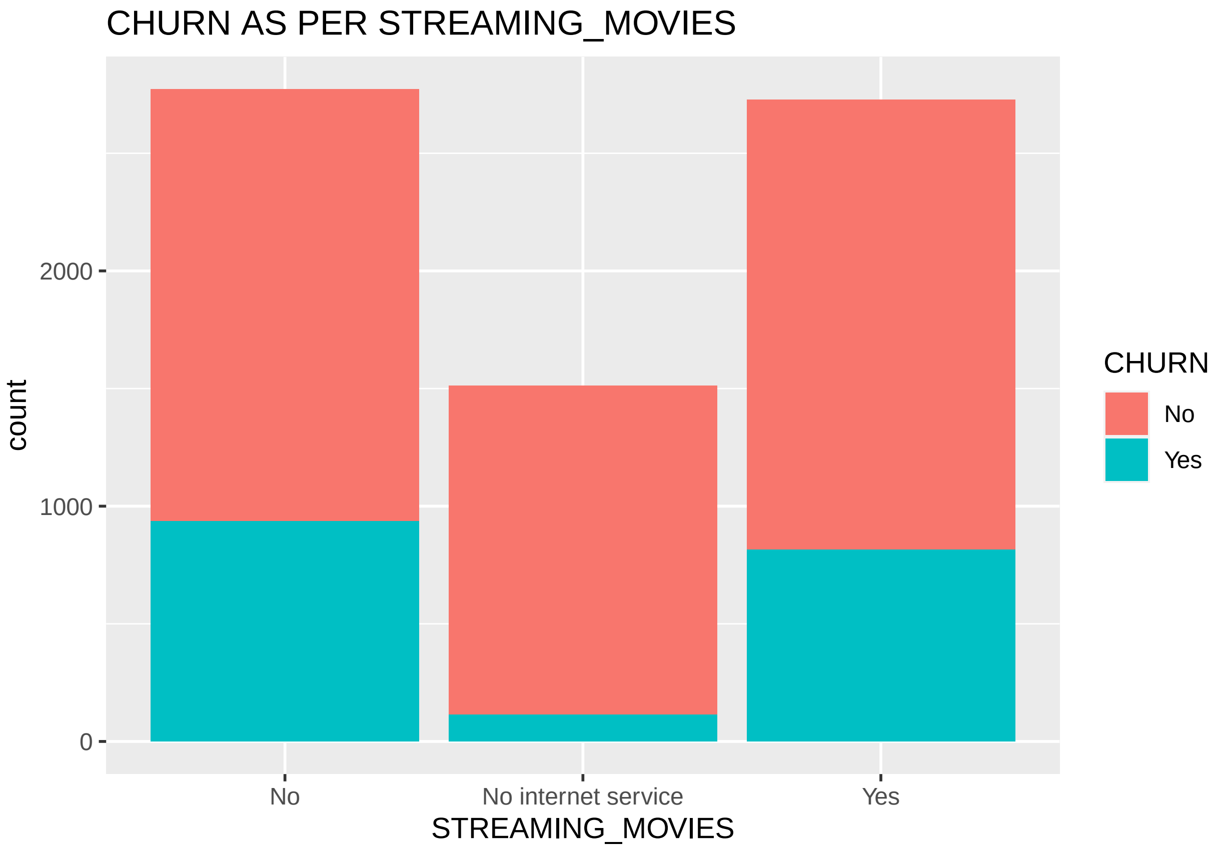
```
ggplot(MTN_df, aes(x=TECH_SUPPORT,fill=CHURN))+ geom_bar() + labs(title="CHURN AS PER TECH_SUPPORT", x=
```



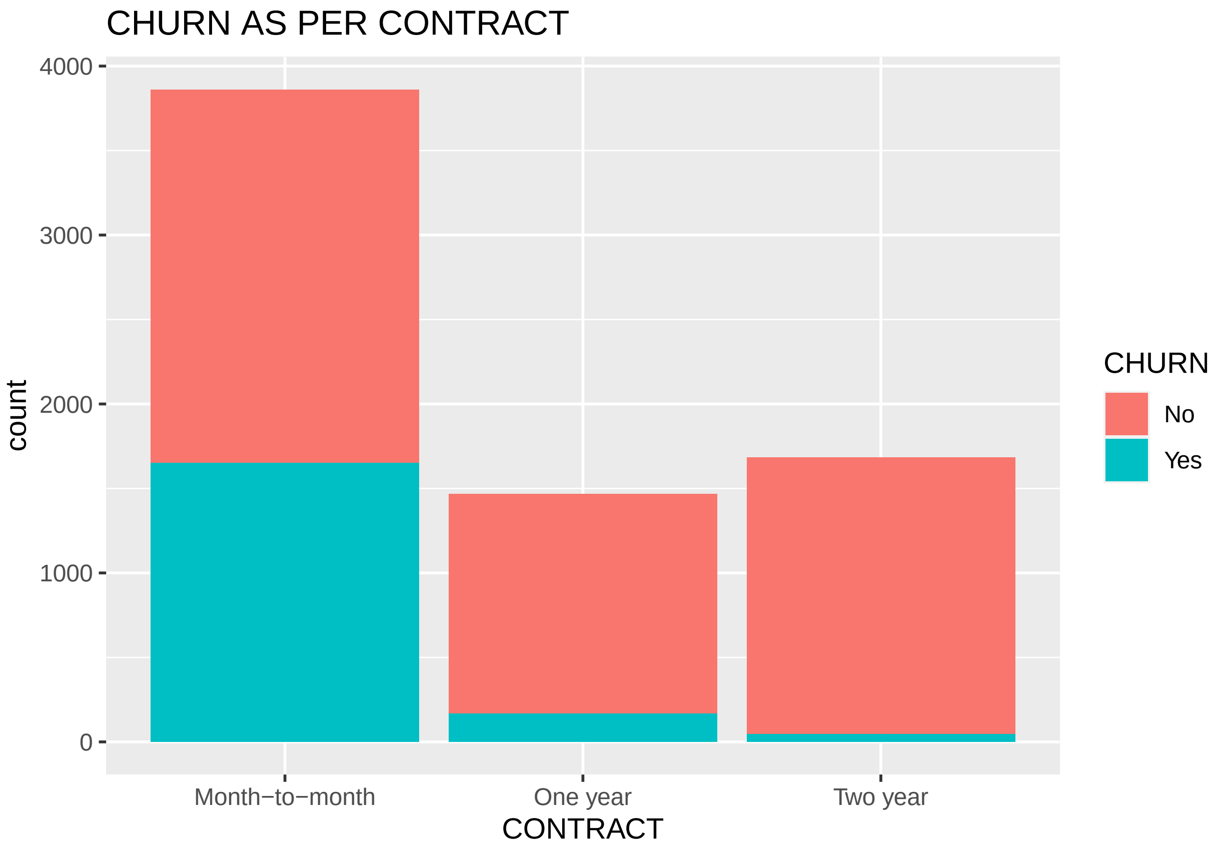
```
ggplot(MTN_df, aes(x=STREAMING_TV,fill=CHURN))+ geom_bar() + labs(title="CHURN AS PER STREAMING_TV", x=
```



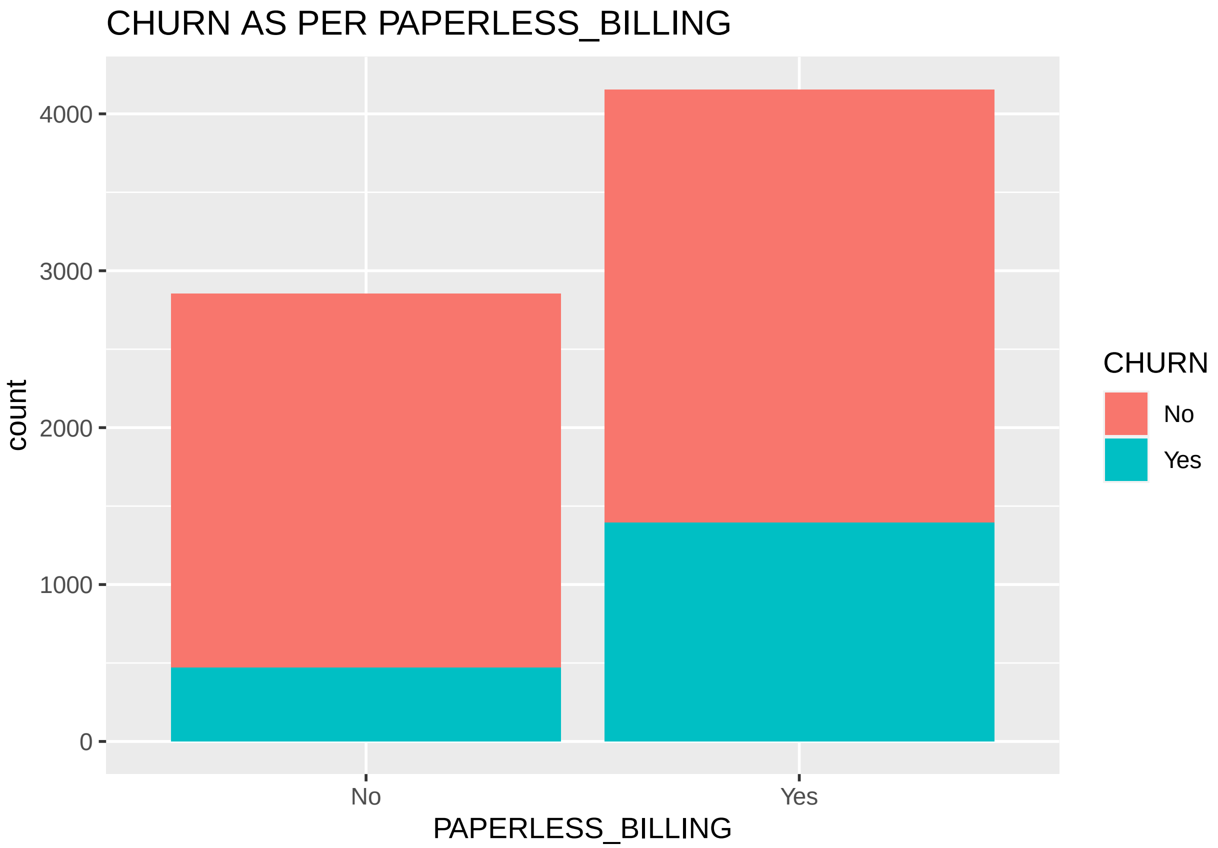
```
ggplot(MTN_df, aes(x=STREAMING_MOVIES,fill=CHURN))+ geom_bar() + labs(title="CHURN AS PER STREAMING_MOVI
```



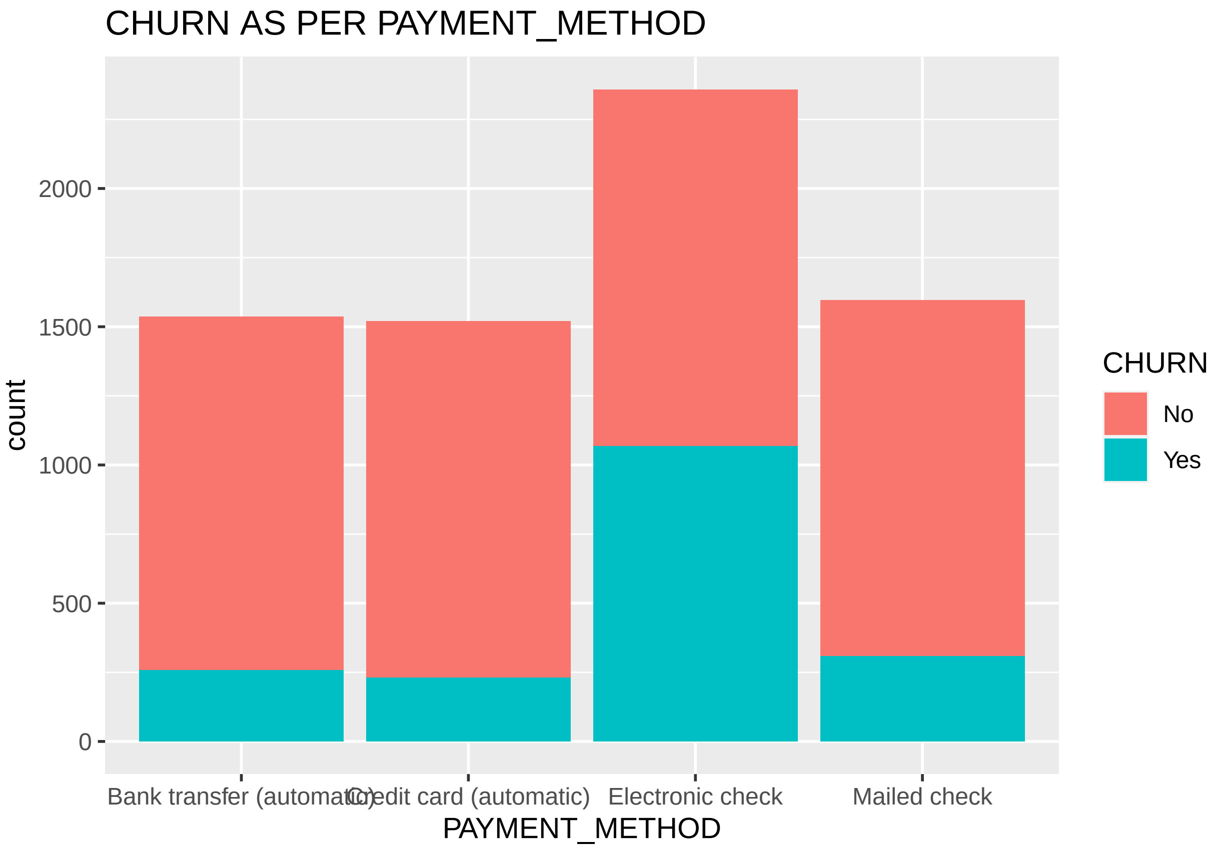
```
ggplot(MTN_df, aes(x=CONTRACT,fill=CHURN))+ geom_bar() + labs(title="CHURN AS PER CONTRACT", x="CONTRAC
```



```
ggplot(MTN_df, aes(x=PAPERLESS_BILLING,fill=CHURN))+ geom_bar() + labs(title="CHURN AS PER PAPERLESS_BIL
```



```
ggplot(MTN_df, aes(x=PAYMENT_METHOD,fill=CHURN))+ geom_bar() + labs(title="CHURN AS PER PAYMENT_METHOD"
```



## Observation:

1.The distribution of gender(male and female) is fifty, and the churn rate is almost the same. 2.According to the senior citizen chart, we can see that most of the customers in the data set are younger people. 3.Almost 50% of customers have a partner, and the churn rate is lower than customers who don't have a partner. 4.Online security, online backup, device protection, tech support, streaming tv, and streaming movies are services used by customers with internet service. The churn rate for customers who use the add-ons service is lower than for those who don't use the service. For example, customers who have used tech support's churn rate is much lower. 5.Customers with the monthly plan have the highest churn rate.

Recommendations:

–In order to create an effective customer retention program, management should take the following measures:

1.Focus more on meeting the needs of non-senior citizens. 2.Focus more on having customers that have partners and/or dependents since these people are less likely to churn. Alternatively, management can come up with services specifically designed for customers without parters and/or dependents. This would require additional research. 3.Focus more on getting customers to long term contract e.g. two year contracts which had low churn rates.