Project

2023-11-25

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com (http://rmarkdown.rstudio.com).

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
df <- read.csv('train.csv')

# Identify indices of majority and minority classes
churn_indices <- which(df$Churn == 1)
no_churn_indices <- which(df$Churn == 0)

# Randomly undersample the majority class to match the size of the minority class
set.seed(123)  # for reproducibility
no_churn_sampled_indices <- sample(no_churn_indices, length(churn_indices))
df <- df[c(churn_indices, no_churn_sampled_indices), ]


# Shuffle the rows
df <- df[sample(nrow(df)), ]

stratified_sample <- df %>%
  group_by(Churn) %>%
  sample_n(3000)

# Replace the original dataframe with the sampled data
churn_df <- data.frame(stratified_sample)
df = churn_df
```

```
df <- readr::read_csv("train.csv",show_col_types = FALSE)
#head(df)
```

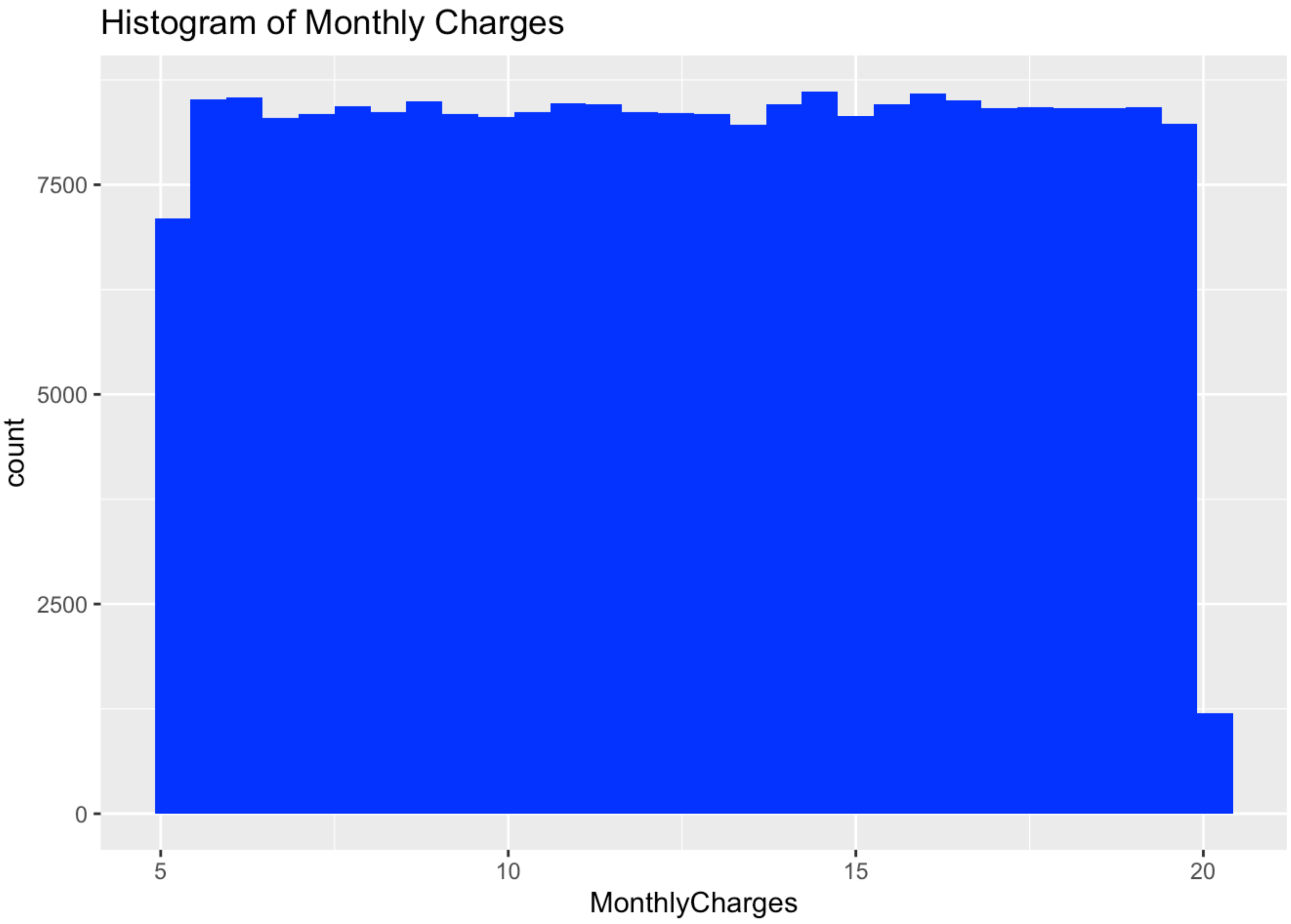
```
str(df)
```

```
## spc_tbl_ [243,787 × 21] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ AccountAge              : num [1:243787] 20 57 73 32 57 113 38 25 26 14 ...
##  $ MonthlyCharges          : num [1:243787] 11.06 5.18 12.11 7.26 16.95 ...
##  $ TotalCharges            : num [1:243787] 221 295 884 232 966 ...
##  $ SubscriptionType        : chr [1:243787] "Premium" "Basic" "Basic" "Basic" ...
##  $ PaymentMethod           : chr [1:243787] "Mailed check" "Credit card" "Mailed c
heck" "Electronic check" ...
##  $ PaperlessBilling        : chr [1:243787] "No" "Yes" "Yes" "No" ...
##  $ ContentType             : chr [1:243787] "Both" "Movies" "Movies" "TV Shows"
...
##  $ MultiDeviceAccess       : chr [1:243787] "No" "No" "No" "No" ...
##  $ DeviceRegistered        : chr [1:243787] "Mobile" "Tablet" "Computer" "Tablet"
...
##  $ ViewingHoursPerWeek     : num [1:243787] 36.8 32.5 7.4 28 20.1 ...
##  $ AverageViewingDuration  : num [1:243787] 63.5 25.7 57.4 131.5 45.4 ...
##  $ ContentDownloadsPerMonth: num [1:243787] 10 18 23 30 20 35 28 10 28 0 ...
##  $ GenrePreference         : chr [1:243787] "Sci-Fi" "Action" "Fantasy" "Drama"
...
##  $ UserRating              : num [1:243787] 2.18 3.48 4.24 4.28 3.62 ...
##  $ SupportTicketsPerMonth  : num [1:243787] 4 8 6 2 4 8 9 2 0 0 ...
##  $ Gender                  : chr [1:243787] "Male" "Male" "Male" "Male" ...
##  $ WatchlistSize           : num [1:243787] 3 23 1 24 0 2 20 22 5 18 ...
##  $ ParentalControl         : chr [1:243787] "No" "No" "Yes" "Yes" ...
##  $ SubtitlesEnabled        : chr [1:243787] "No" "Yes" "Yes" "Yes" ...
##  $ CustomerID              : chr [1:243787] "CB6SXPNVZA" "S7R2G87O09" "EASDC20BDT"
"NPF69NT69N" ...
##  $ Churn                   : num [1:243787] 0 0 0 0 0 0 0 0 1 0 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   AccountAge = col_double(),
##   ..   MonthlyCharges = col_double(),
##   ..   TotalCharges = col_double(),
##   ..   SubscriptionType = col_character(),
##   ..   PaymentMethod = col_character(),
##   ..   PaperlessBilling = col_character(),
##   ..   ContentType = col_character(),
##   ..   MultiDeviceAccess = col_character(),
##   ..   DeviceRegistered = col_character(),
##   ..   ViewingHoursPerWeek = col_double(),
##   ..   AverageViewingDuration = col_double(),
##   ..   ContentDownloadsPerMonth = col_double(),
##   ..   GenrePreference = col_character(),
##   ..   UserRating = col_double(),
##   ..   SupportTicketsPerMonth = col_double(),
##   ..   Gender = col_character(),
##   ..   WatchlistSize = col_double(),
##   ..   ParentalControl = col_character(),
##   ..   SubtitlesEnabled = col_character(),
##   ..   CustomerID = col_character(),
##   ..   Churn = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

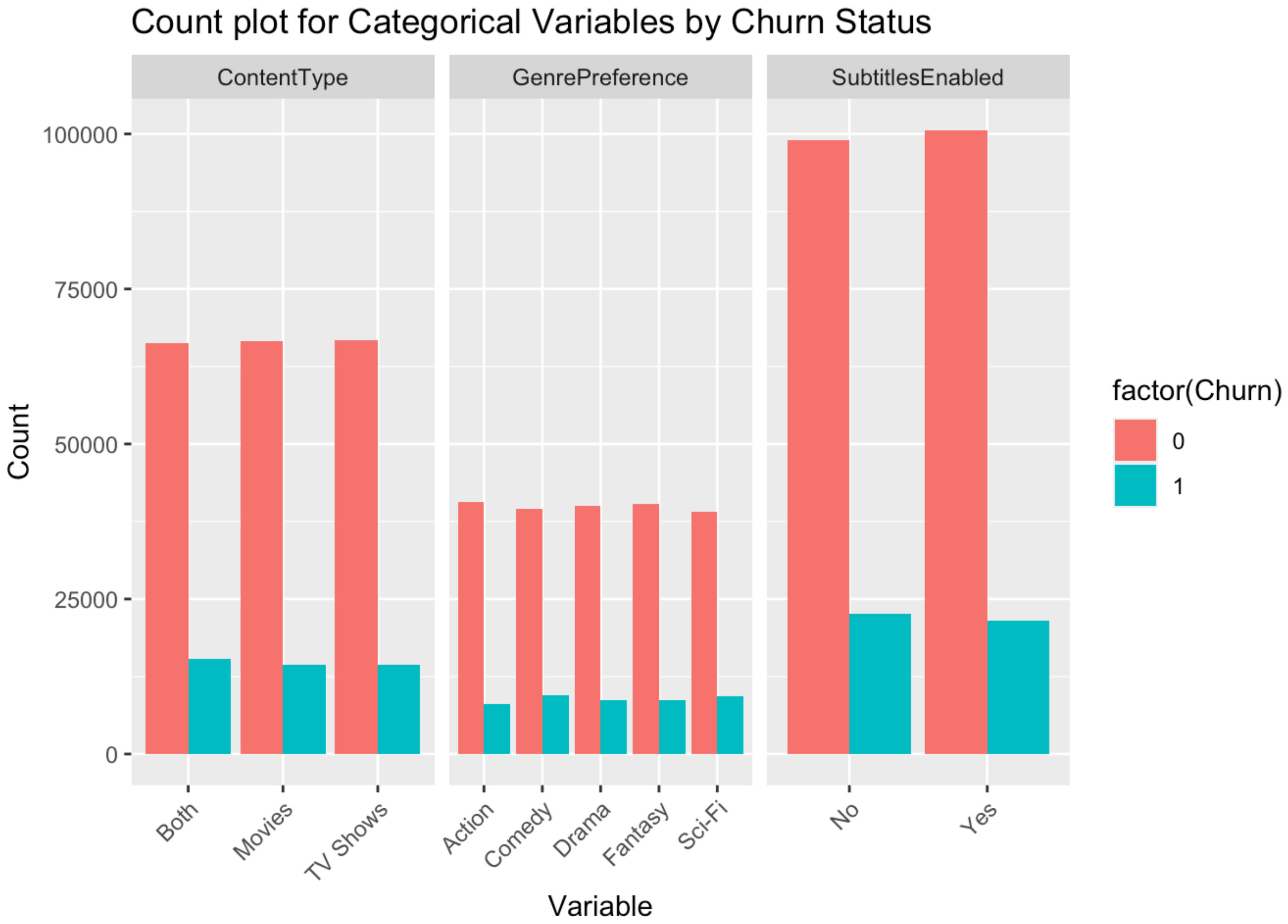
```
summary(df)
```

```
##    AccountAge     MonthlyCharges    TotalCharges      SubscriptionType
##  Min.   :  1.00   Min.   : 4.990   Min.   :   4.991   Length:243787
##  1st Qu.: 30.00   1st Qu.: 8.739   1st Qu.: 329.147   Class :character
##  Median : 60.00   Median :12.496   Median : 649.879   Mode  :character
##  Mean   : 60.08   Mean   :12.491   Mean   : 750.741
##  3rd Qu.: 90.00   3rd Qu.:16.238   3rd Qu.:1089.317
##  Max.   :119.00   Max.   :19.990   Max.   :2378.724
##  PaymentMethod      PaperlessBilling   ContentType        MultiDeviceAccess
##  Length:243787      Length:243787      Length:243787      Length:243787
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  DeviceRegistered   ViewingHoursPerWeek AverageViewingDuration
##  Length:243787      Min.   : 1.00       Min.   :  5.001
##  Class :character   1st Qu.:10.76       1st Qu.: 48.382
##  Mode  :character   Median :20.52       Median : 92.250
##                     Mean   :20.50       Mean   : 92.264
##                     3rd Qu.:30.22       3rd Qu.:135.908
##                     Max.   :40.00       Max.   :179.999
##  ContentDownloadsPerMonth GenrePreference      UserRating
##  Min.   : 0.0             Length:243787      Min.   :1.000
##  1st Qu.:12.0             Class :character   1st Qu.:2.001
##  Median :24.0             Mode  :character   Median :3.002
##  Mean   :24.5                                Mean   :3.003
##  3rd Qu.:37.0                                3rd Qu.:4.002
##  Max.   :49.0                                Max.   :5.000
##  SupportTicketsPerMonth    Gender          WatchlistSize   ParentalControl
##  Min.   :0.000          Length:243787      Min.   : 0.00   Length:243787
##  1st Qu.:2.000          Class :character   1st Qu.: 6.00   Class :character
##  Median :4.000          Mode  :character   Median :12.00   Mode  :character
##  Mean   :4.504                             Mean   :12.02
##  3rd Qu.:7.000                             3rd Qu.:18.00
##  Max.   :9.000                             Max.   :24.00
##  SubtitlesEnabled    CustomerID            Churn
##  Length:243787      Length:243787      Min.   :0.0000
##  Class :character   Class :character   1st Qu.:0.0000
##  Mode  :character   Mode  :character   Median :0.0000
##                                        Mean   :0.1812
##                                        3rd Qu.:0.0000
##                                        Max.   :1.0000
```

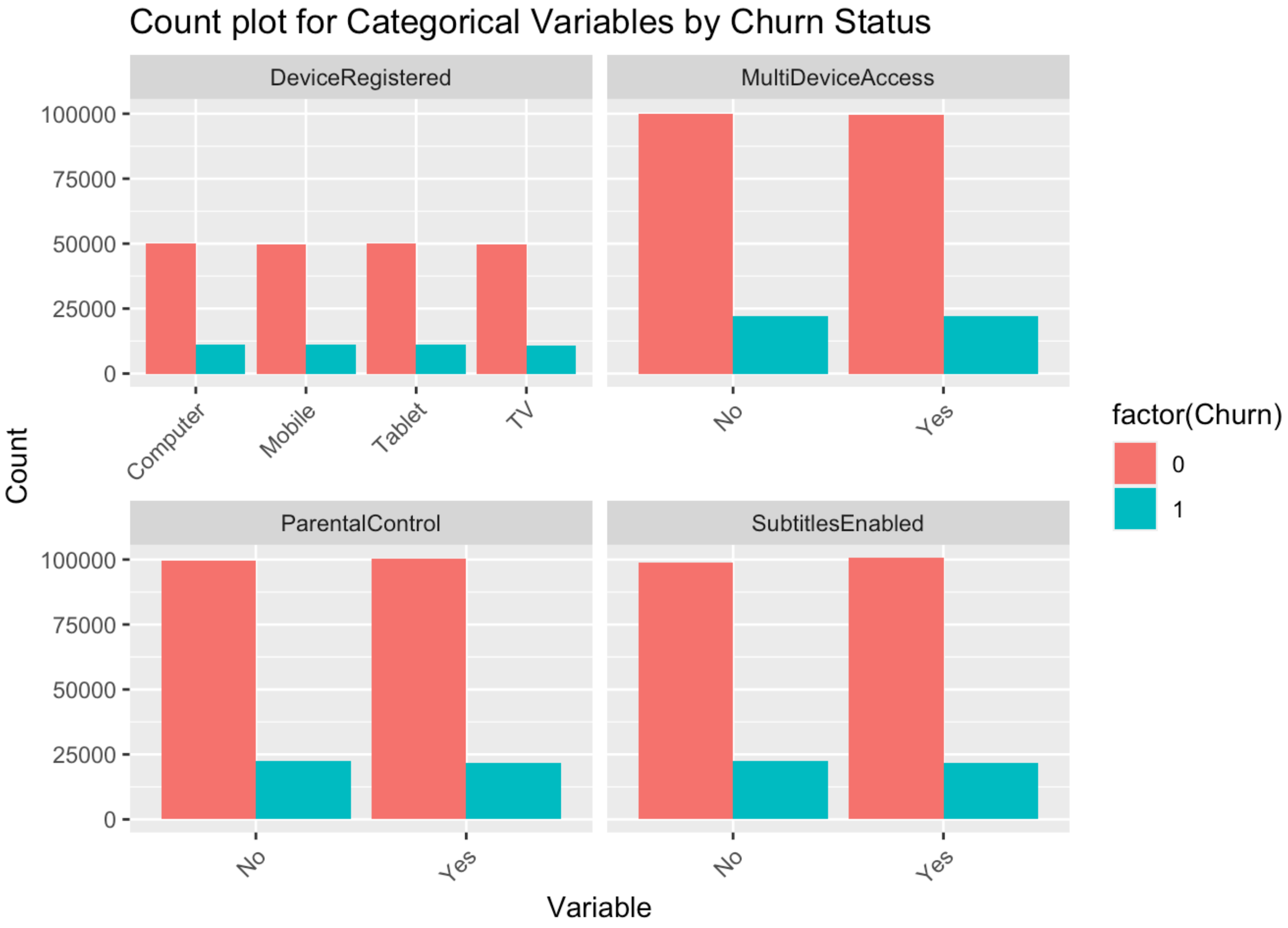
```
ggplot(df, aes(x = MonthlyCharges)) +
  geom_histogram(fill = "blue", bins = 30) +
  labs(title = "Histogram of Monthly Charges")
```



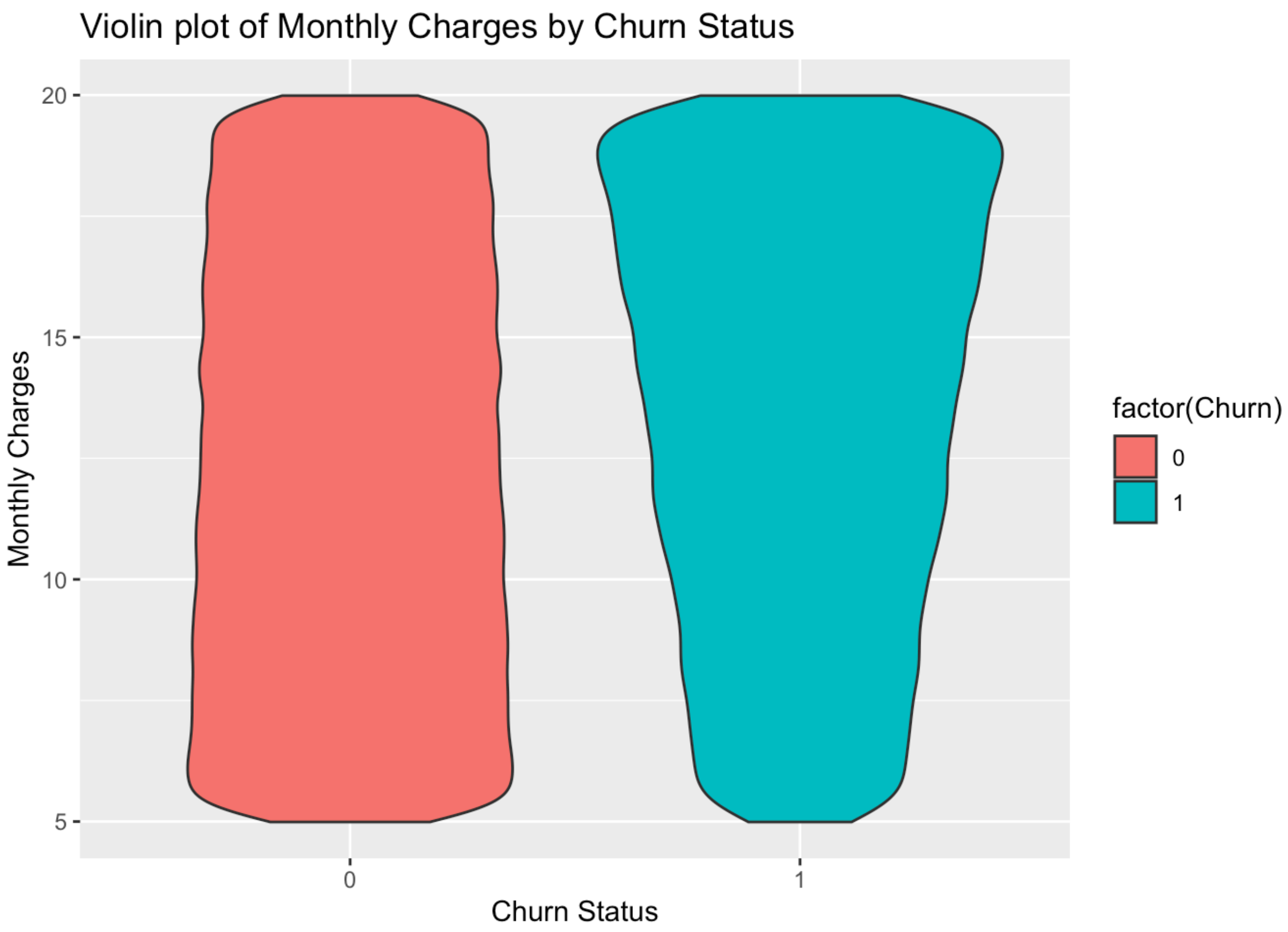
```
# Reshape data for count plots
df_long <- df %>% pivot_longer(cols = c(ContentType,GenrePreference,SubtitlesEnable
d))

# Example: Count plot for PaymentMethod
ggplot(df_long, aes(x = value, fill = factor(Churn))) +
  geom_bar(position = "dodge") +
  facet_wrap(~name, scales = "free_x") +
  labs(title = "Count plot for Categorical Variables by Churn Status", x = "Variabl
e", y = "Count") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



```
# Reshape data for count plots
df_long <- df %>% pivot_longer(cols = c(MultiDeviceAccess, DeviceRegistered,ParentalC
ontrol, SubtitlesEnabled))

# Example: Count plot for PaymentMethod
ggplot(df_long, aes(x = value, fill = factor(Churn))) +
  geom_bar(position = "dodge") +
  facet_wrap(~name, scales = "free_x") +
  labs(title = "Count plot for Categorical Variables by Churn Status", x = "Variabl
e", y = "Count") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



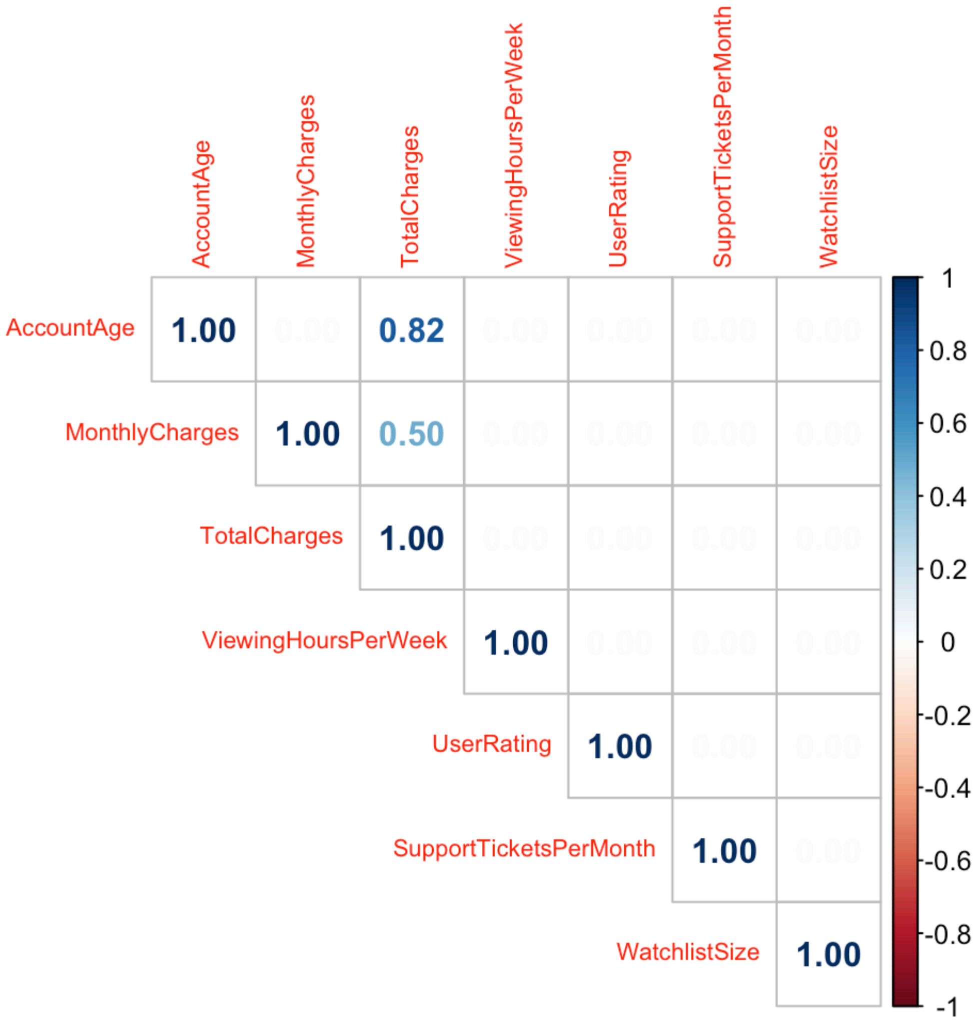
```
# Example: Violin plot for MonthlyCharges by Churn
ggplot(df, aes(x = factor(Churn), y = MonthlyCharges, fill = factor(Churn))) +
  geom_violin() +
  labs(title = "Violin plot of Monthly Charges by Churn Status", x = "Churn Status",
y = "Monthly Charges")
```



#

correlation Matrix heatmap

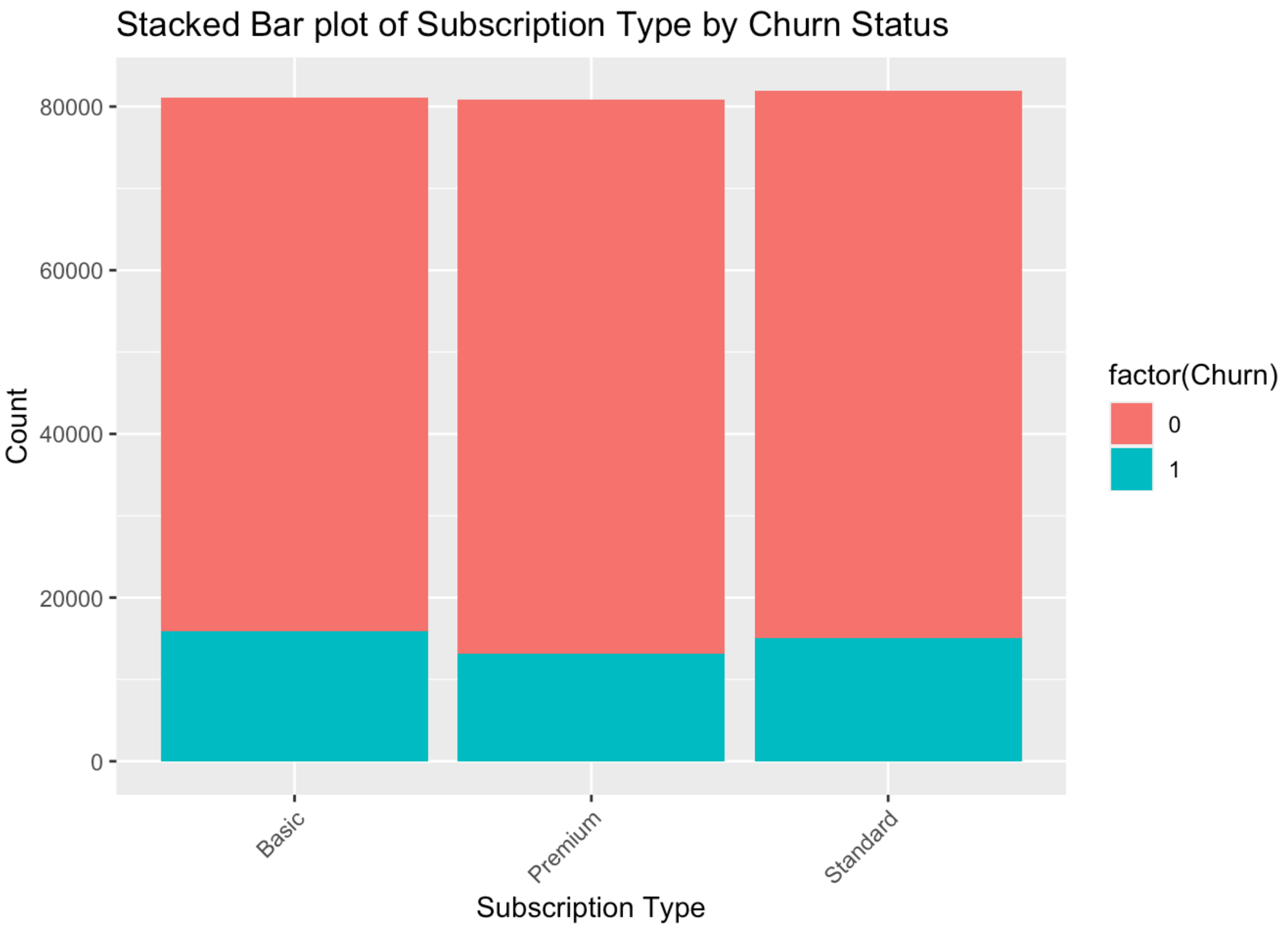
```
cor_matrix <- cor(df[, c("AccountAge", "MonthlyCharges", "TotalCharges", "ViewingHoursPerWeek", "UserRating", "SupportTicketsPerMonth", "WatchlistSize")])

corrplot(cor_matrix, method = "number", type = "upper", tl.cex = 0.7)
```



#

Example: Stacked bar plot for SubscriptionType by Churn

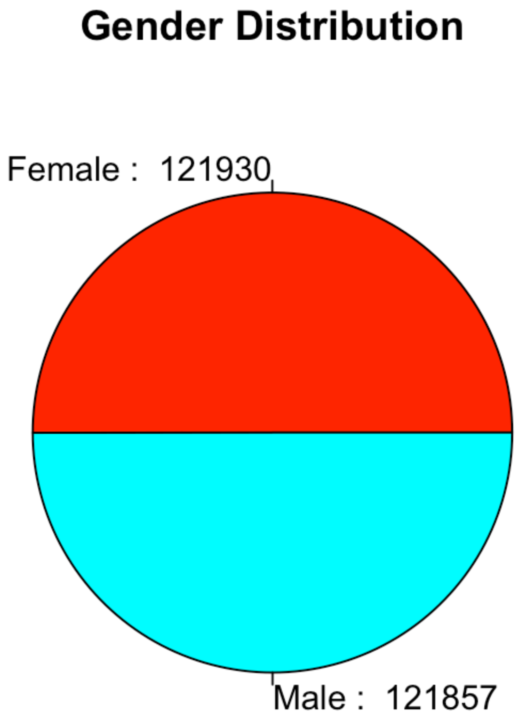
```
ggplot(df, aes(x = SubscriptionType, fill = factor(Churn))) +
  geom_bar(position = "stack") +
  labs(title = "Stacked Bar plot of Subscription Type by Churn Status", x = "Subscrip
tion Type", y = "Count") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



#

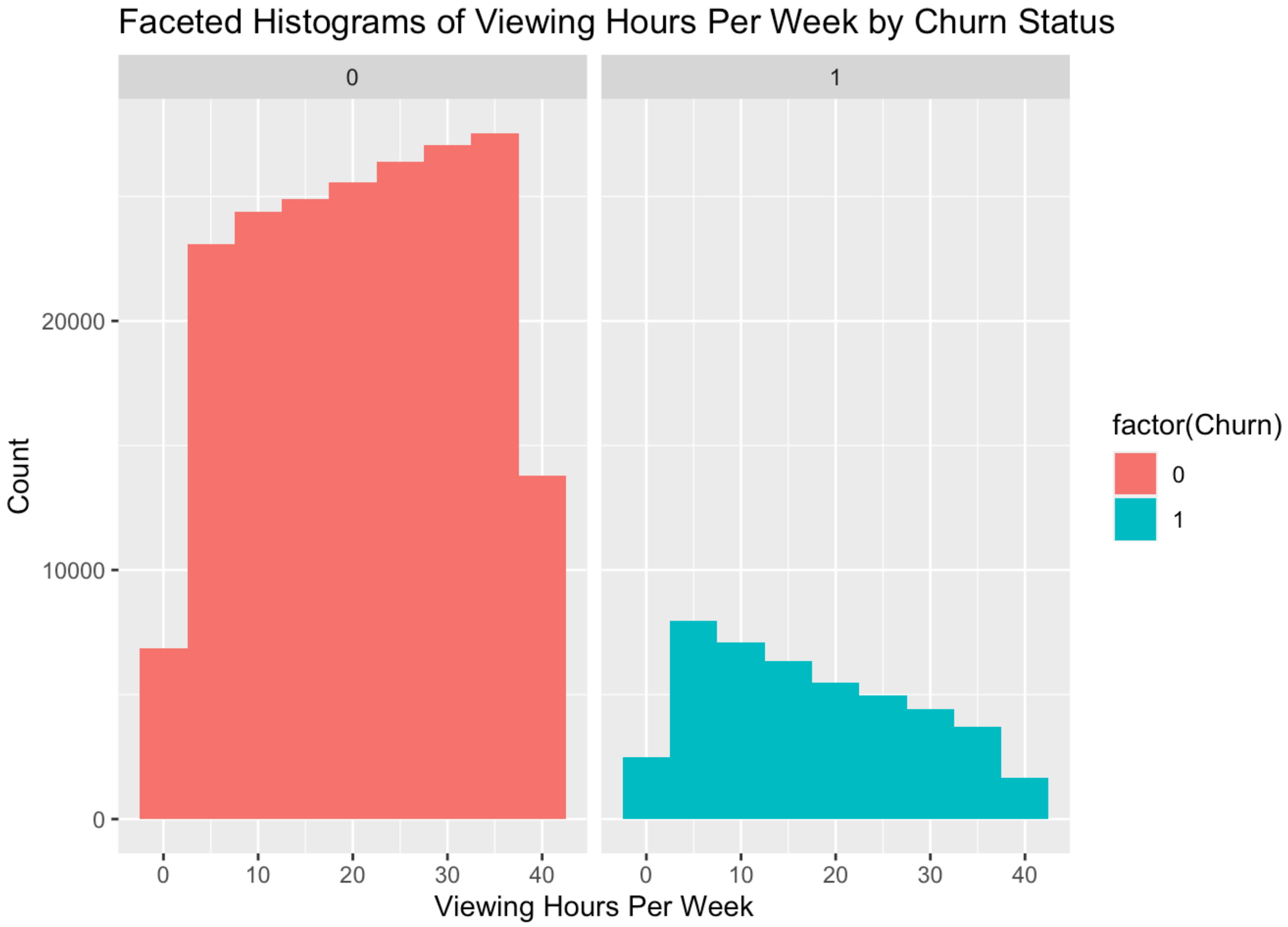
Pie chart for Gender distribution

```
gender_counts <- table(df$Gender)
pie(gender_counts, labels = paste(names(gender_counts), ": ", gender_counts), main =
"Gender Distribution", col = rainbow(length(gender_counts)))
```



#histograms for Viewing Hours Per Week by Churn

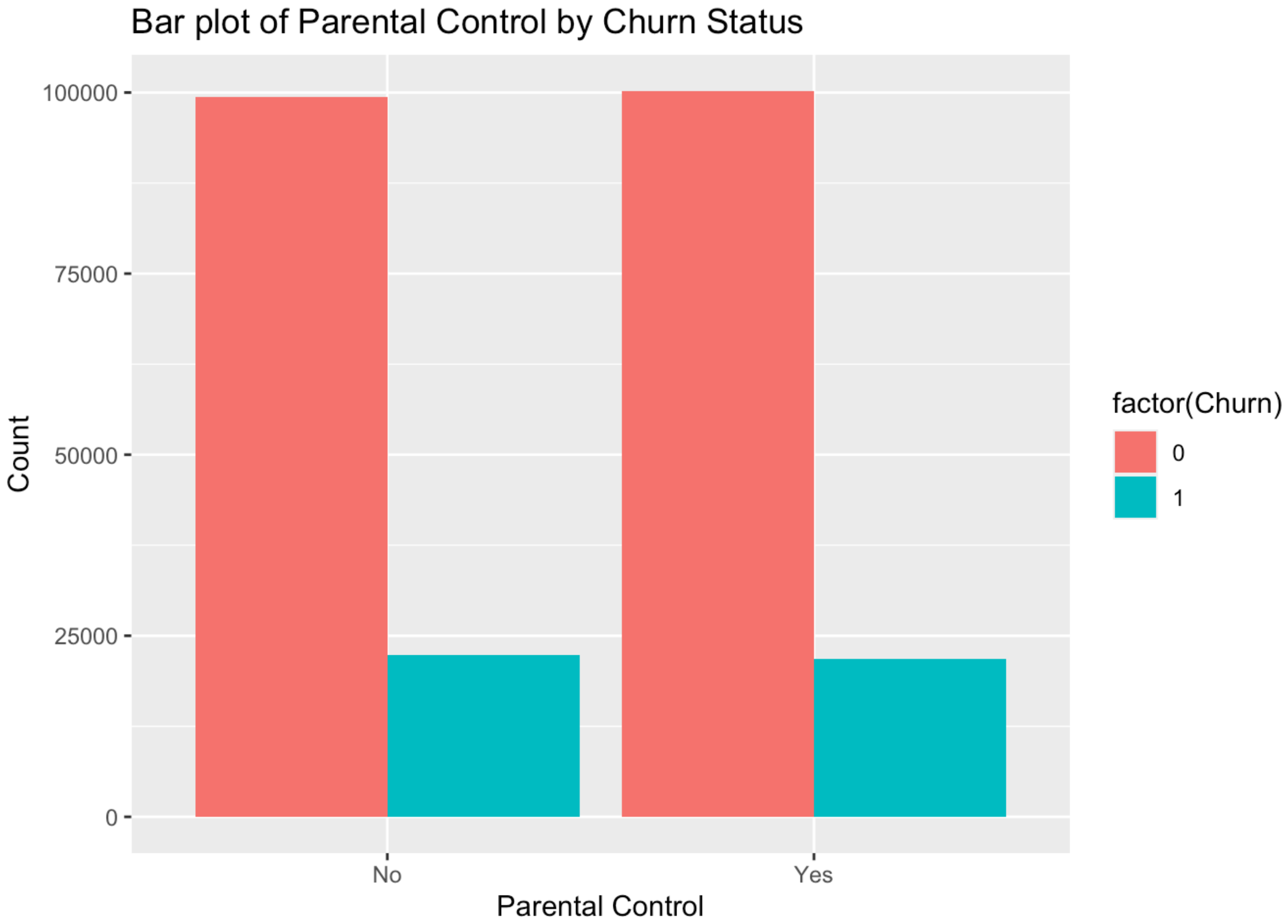
```
ggplot(df, aes(x = ViewingHoursPerWeek, fill = factor(Churn))) +
  geom_histogram(binwidth = 5, position = "dodge") +
  facet_wrap(~Churn) +
  labs(title = "Faceted Histograms of Viewing Hours Per Week by Churn Status", x = "V
iewing Hours Per Week", y = "Count")
```



#

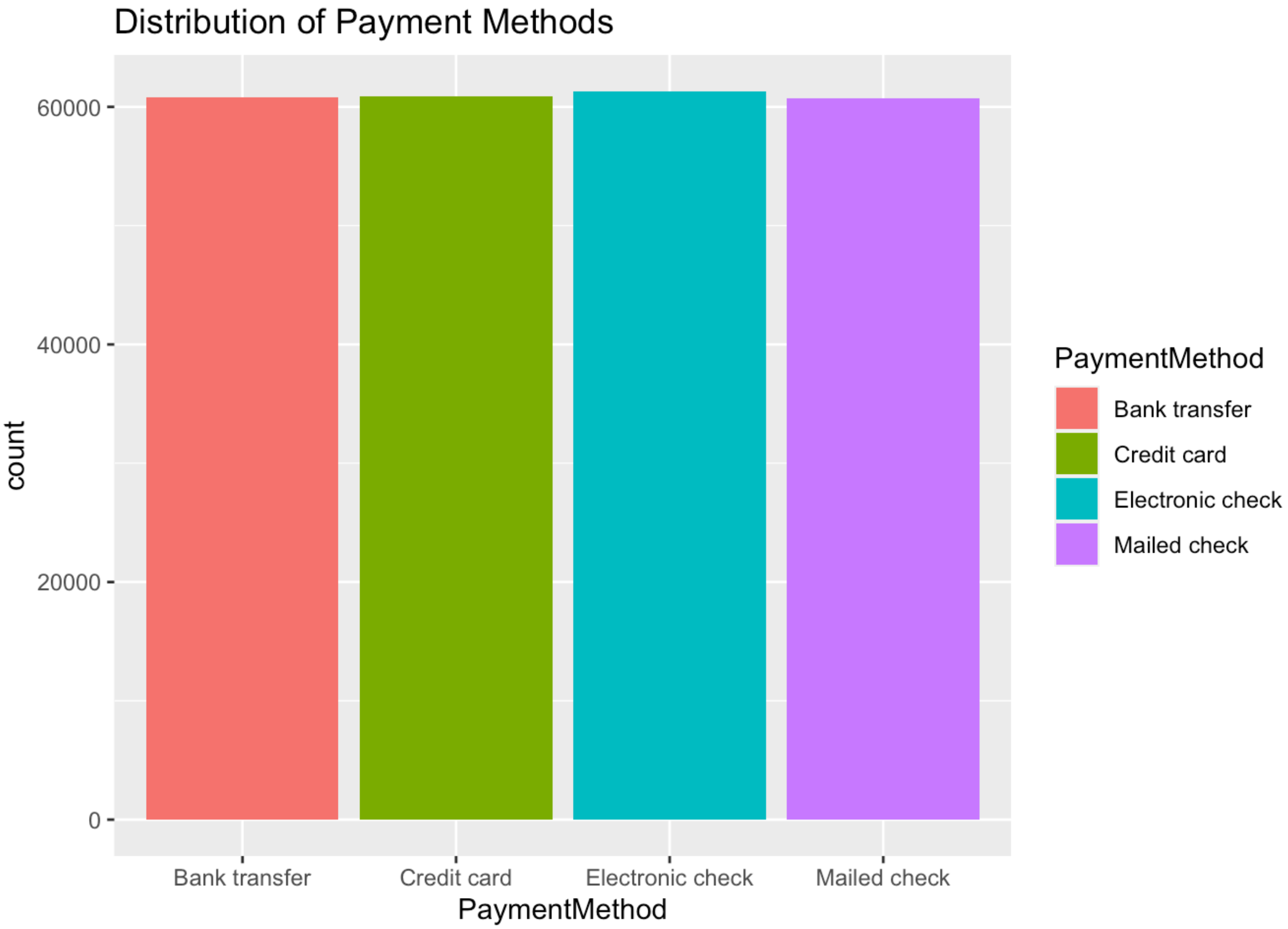
Bar plot for Parental Control

```
ggplot(df, aes(x = ParentalControl, fill = factor(Churn))) +
  geom_bar(position = "dodge") +
  labs(title = "Bar plot of Parental Control by Churn Status", x = "Parental Contro
l", y = "Count")
```

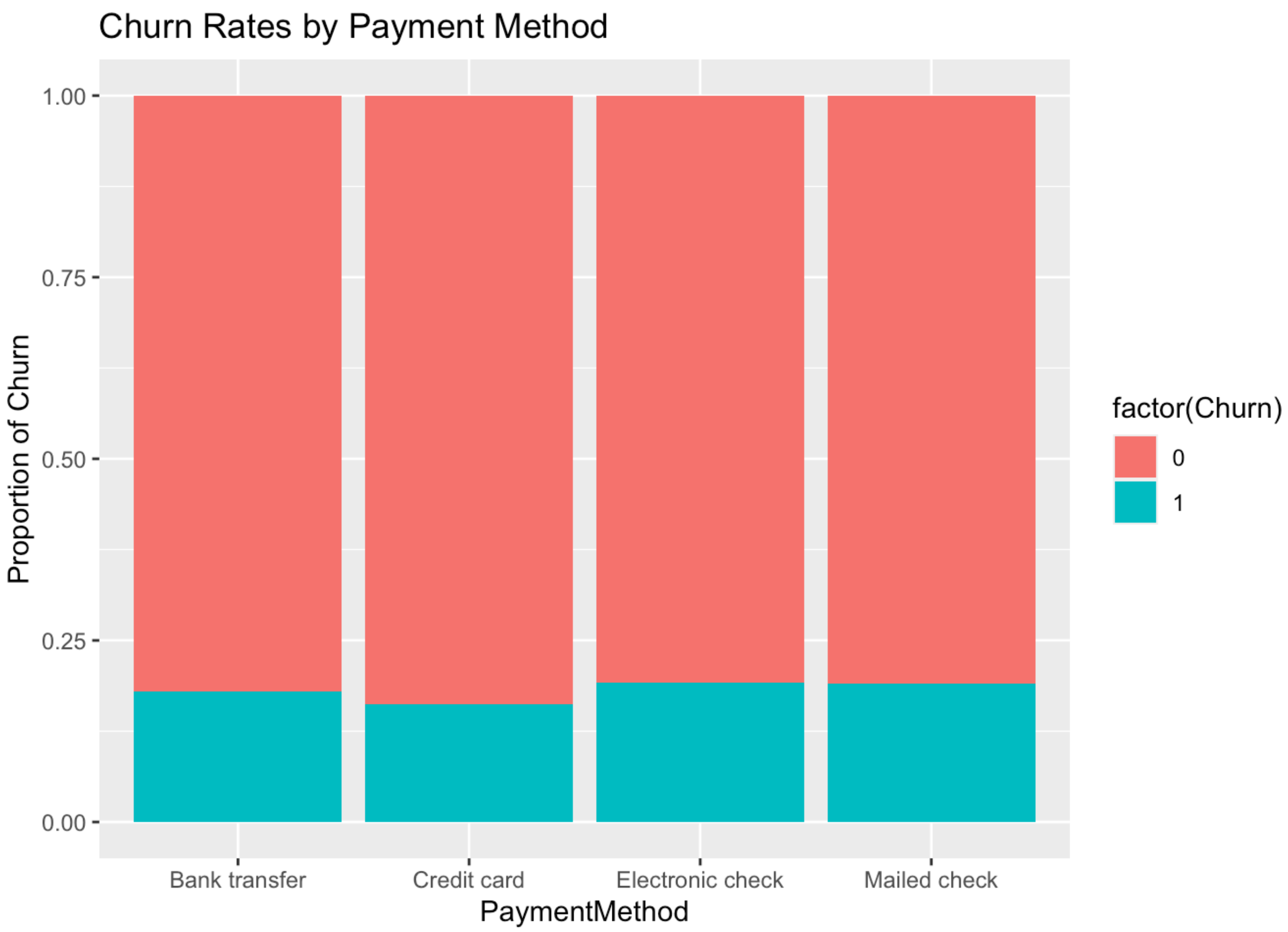


#Churn Rates and box plot for Payment Method

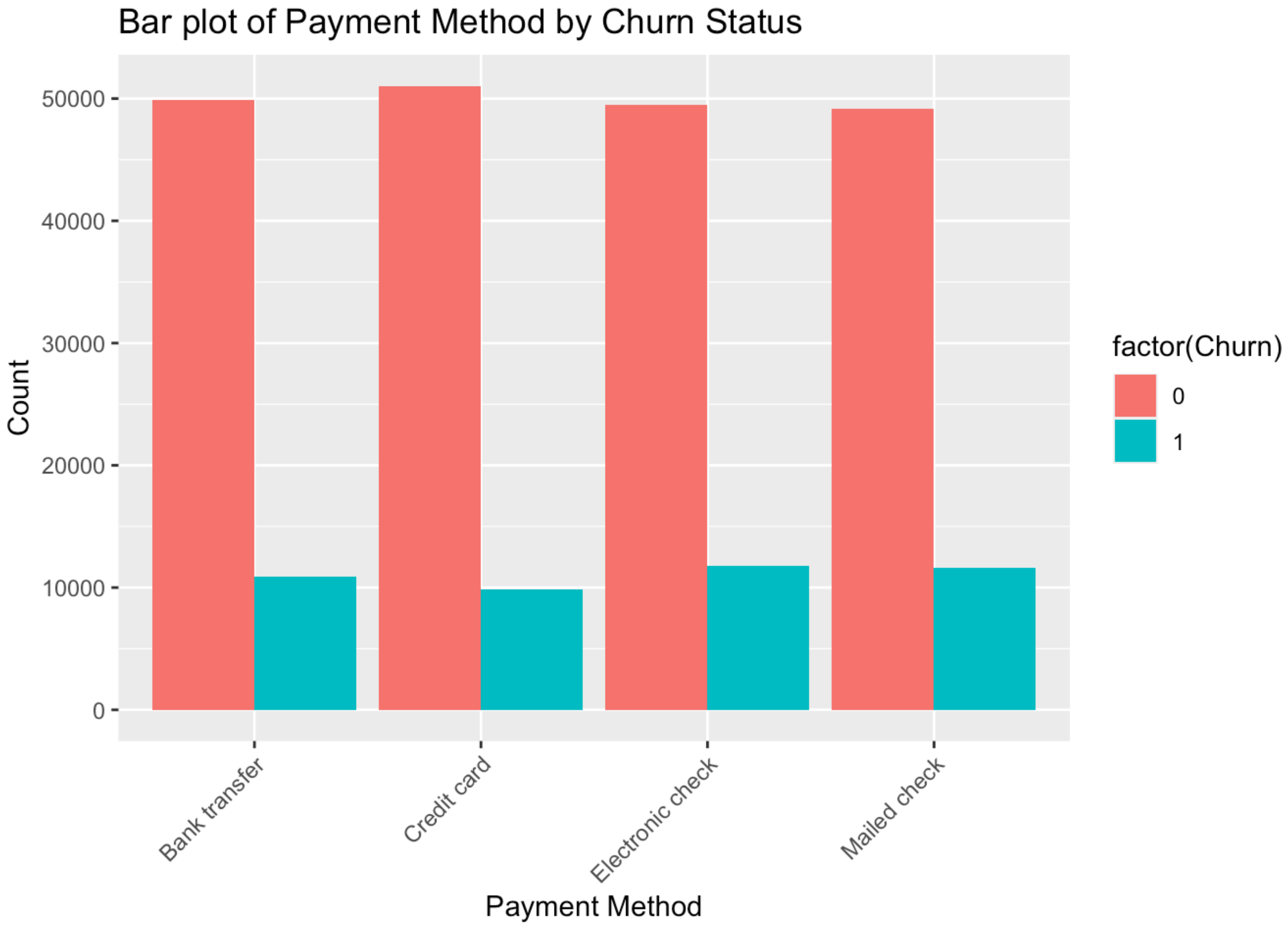
```
ggplot(df, aes(x = PaymentMethod, fill = PaymentMethod)) +
  geom_bar() +
  labs(title = "Distribution of Payment Methods")
```



```
ggplot(df, aes(x = PaymentMethod, fill = factor(Churn))) +
  geom_bar(position = "fill") +
  labs(title = "Churn Rates by Payment Method", y = "Proportion of Churn")
```



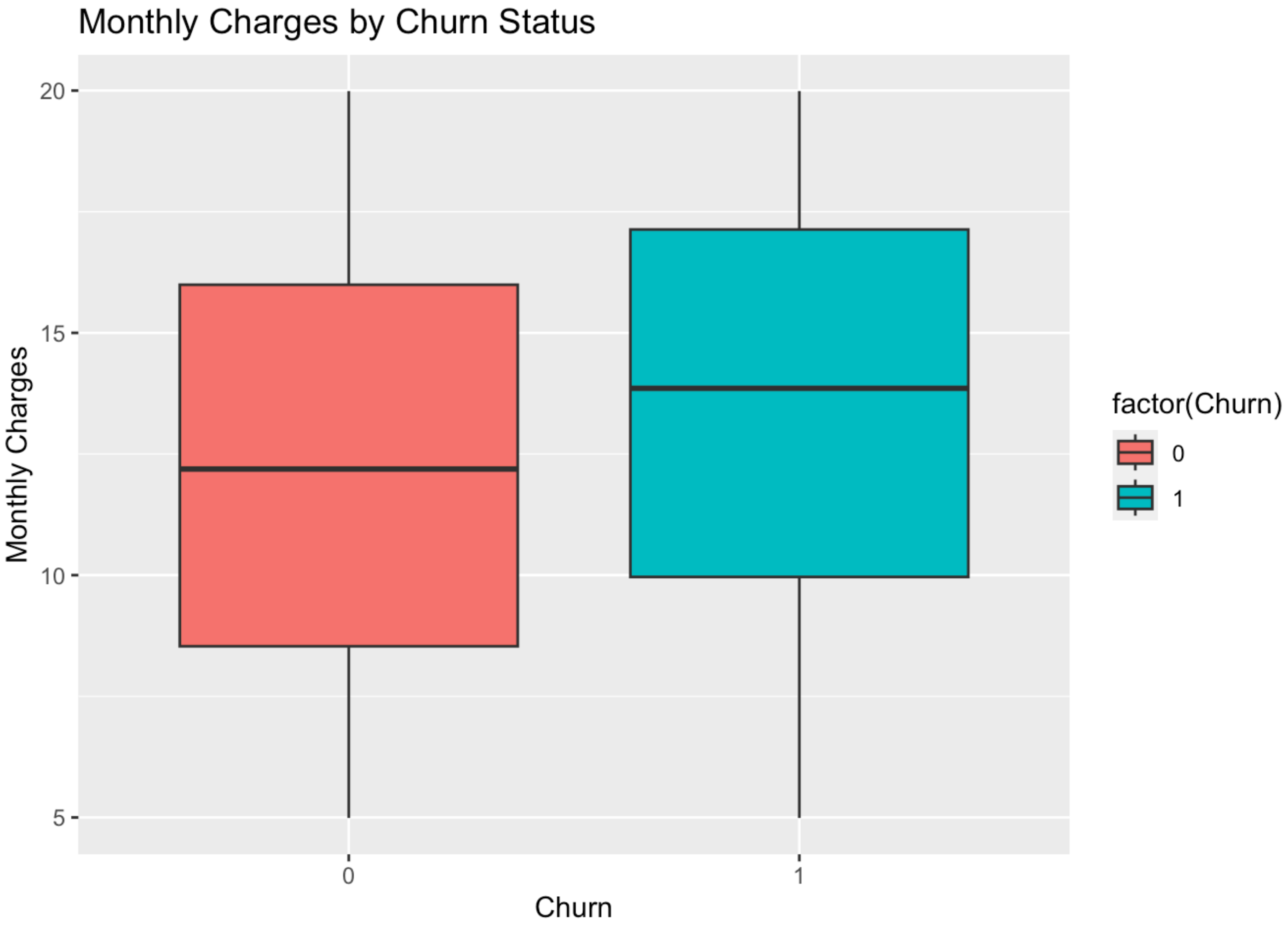
```
ggplot(df, aes(x = PaymentMethod, fill = factor(Churn))) +
  geom_bar(position = "dodge") +
  labs(title = "Bar plot of Payment Method by Churn Status", x = "Payment Method", y
= "Count") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



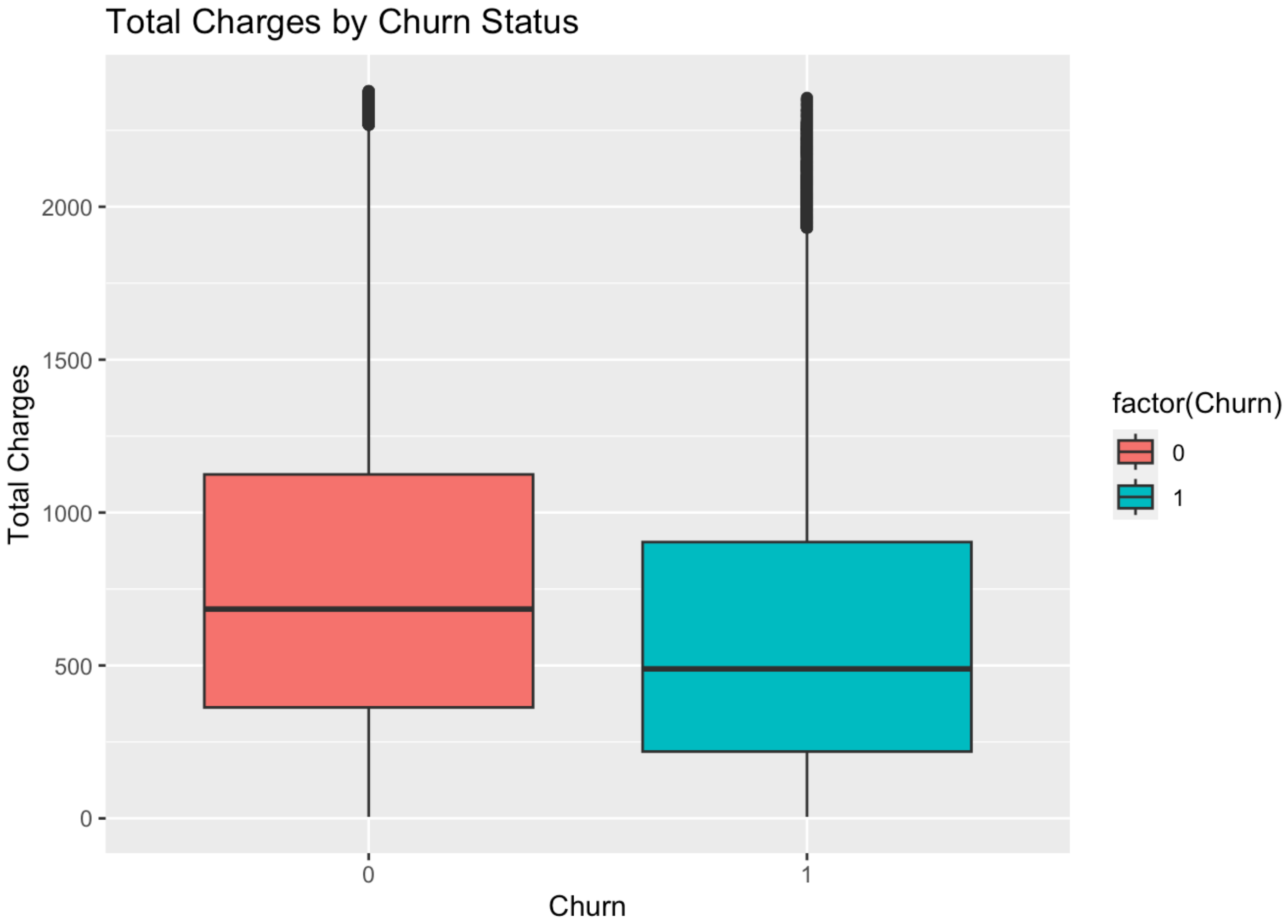
#

Monthly Charges and total charges by Churn - Boxplot

```
ggplot(df, aes(x = factor(Churn), y = MonthlyCharges, fill = factor(Churn))) +
  geom_boxplot() +
  labs(title = "Monthly Charges by Churn Status", x = "Churn", y = "Monthly Charges")
```



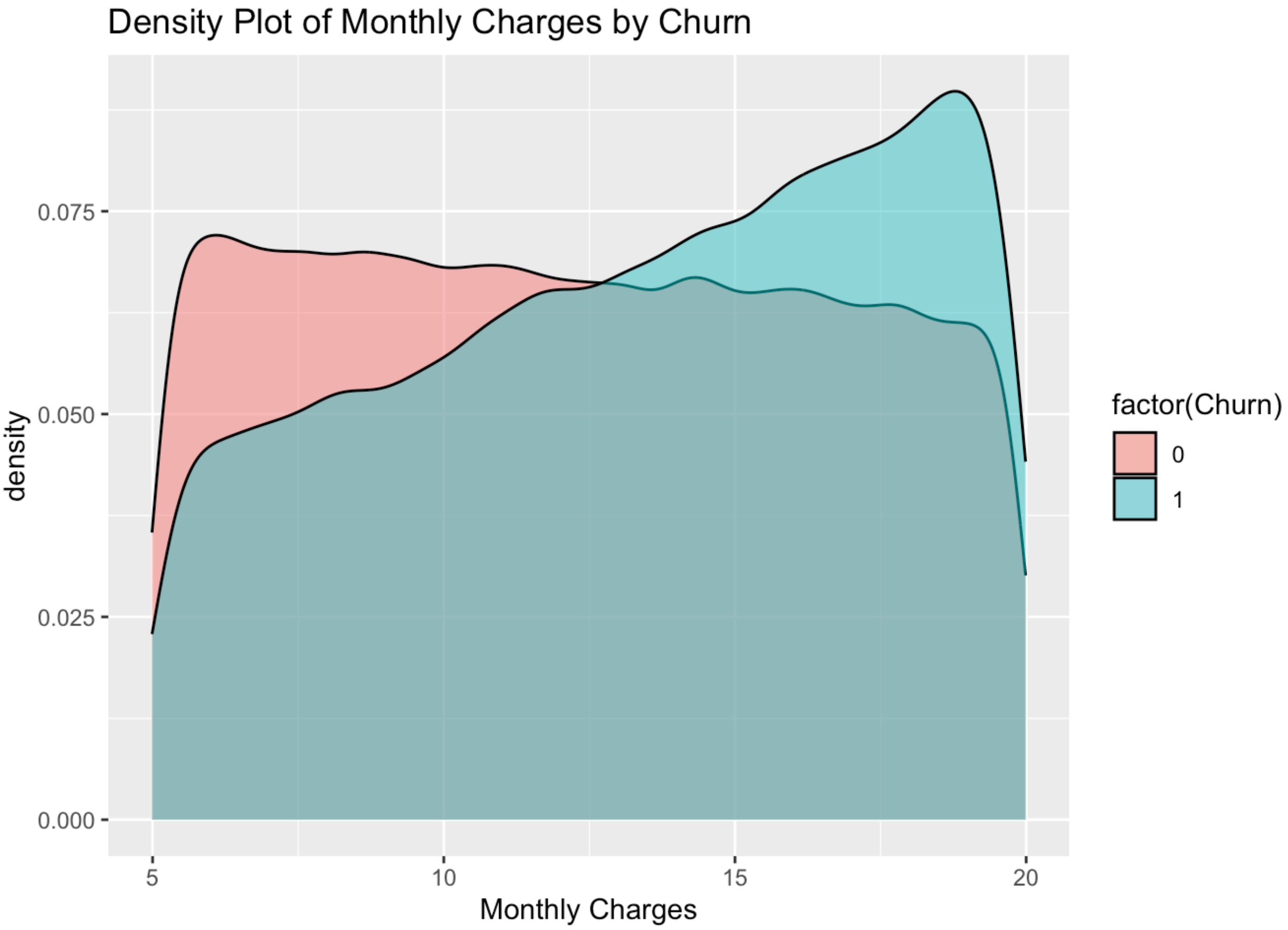
```
ggplot(df, aes(x = factor(Churn), y = TotalCharges, fill = factor(Churn))) +
  geom_boxplot() +
  labs(title = "Total Charges by Churn Status", x = "Churn", y = "Total Charges")
```



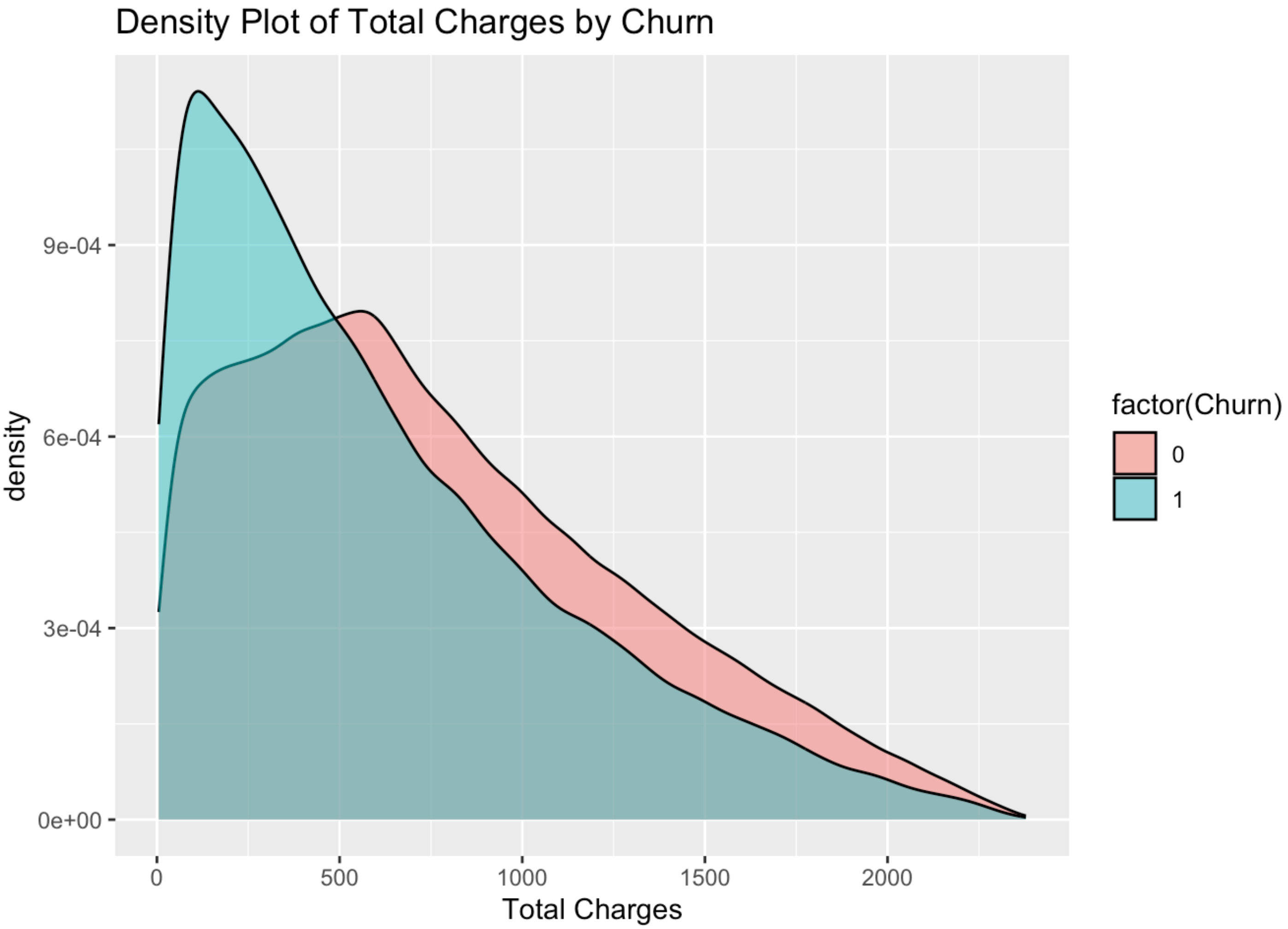
#

Monthly charges distribution and Total charges distribution by churn density plot

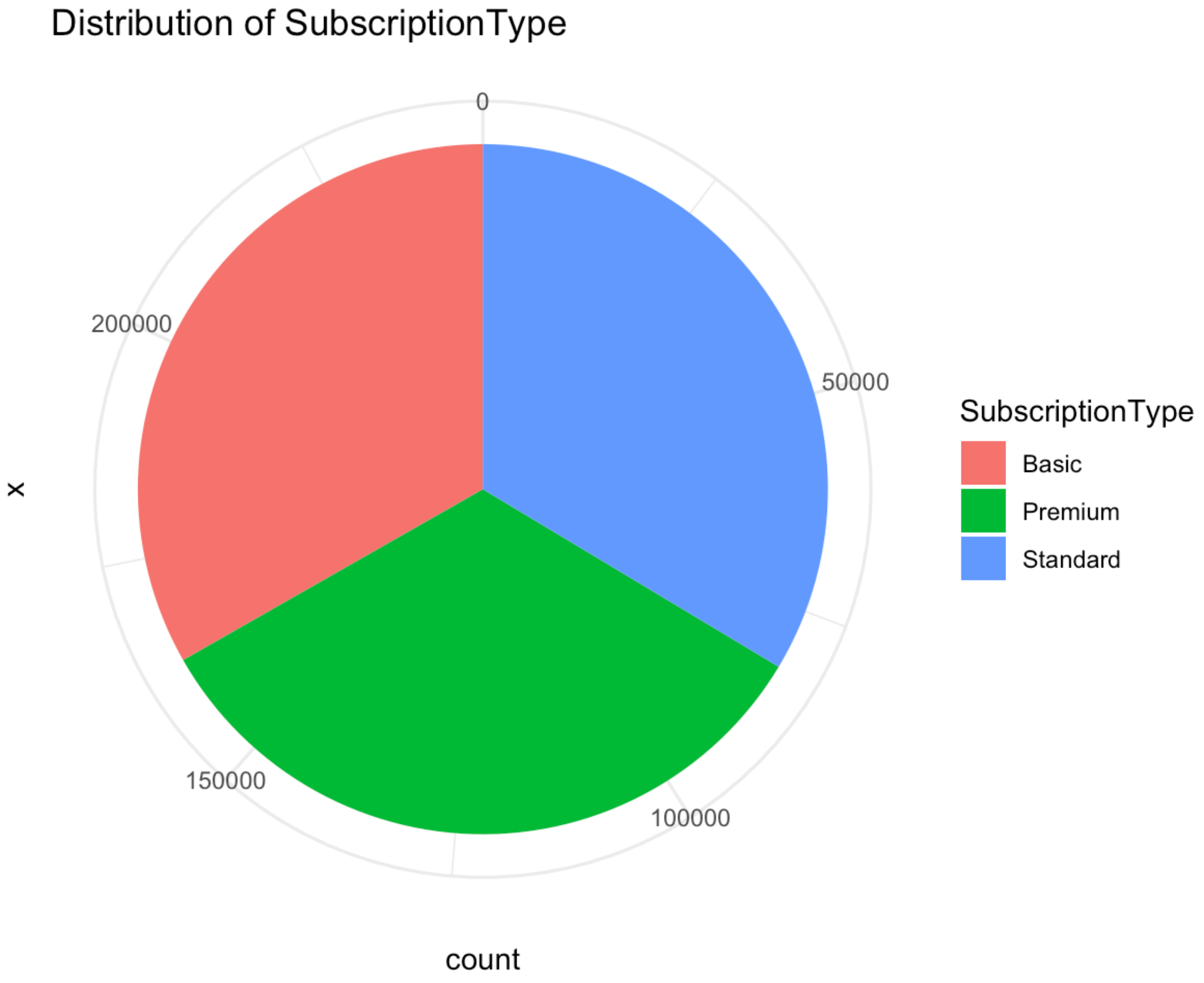
```
ggplot(df, aes(x = MonthlyCharges, fill = factor(Churn))) +
  geom_density(alpha = 0.5) +
  labs(title = "Density Plot of Monthly Charges by Churn", x = "Monthly Charges")
```



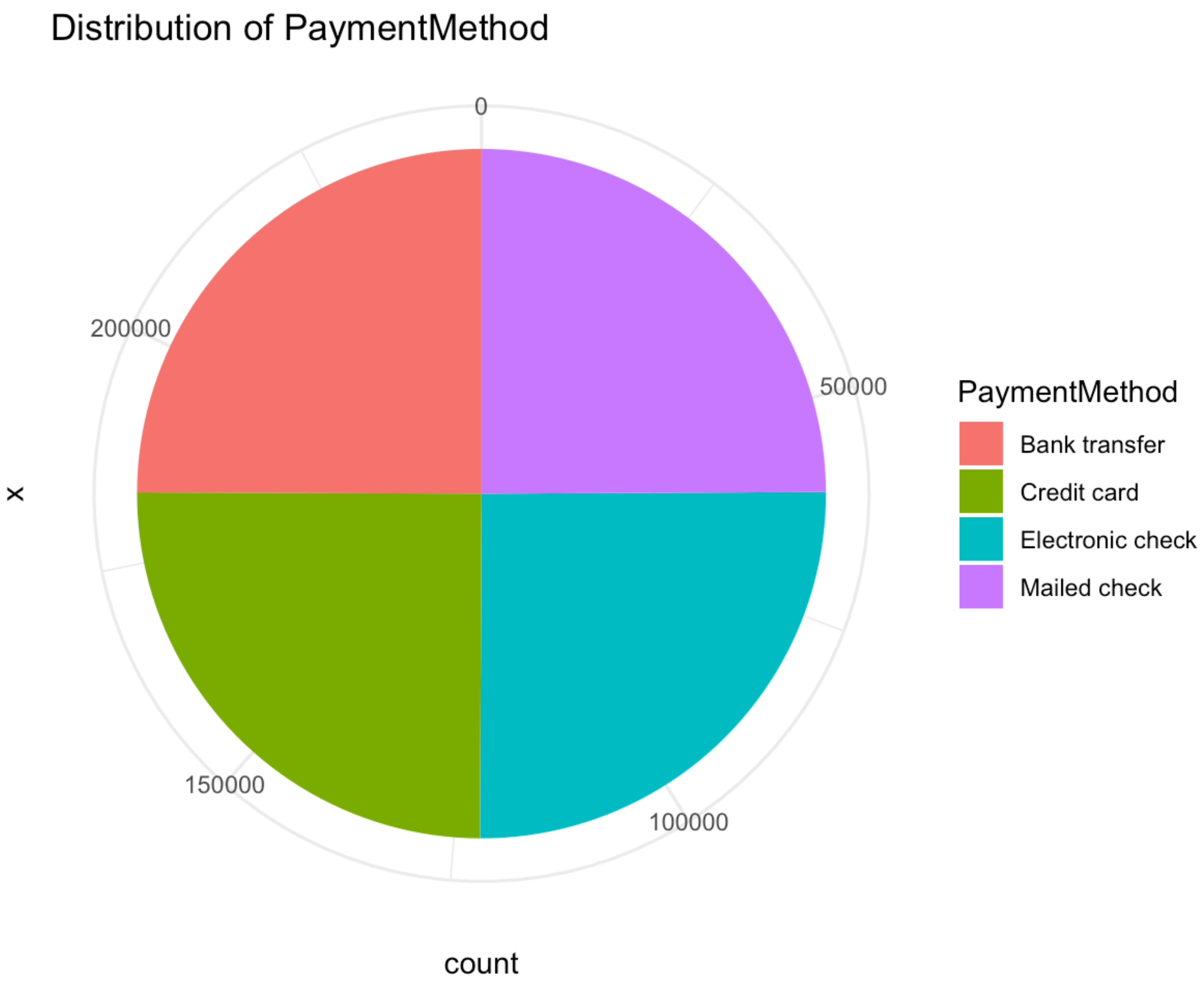
```
ggplot(df, aes(x = TotalCharges, fill = factor(Churn))) +
  geom_density(alpha = 0.5) +
  labs(title = "Density Plot of Total Charges by Churn", x = "Total Charges")
```



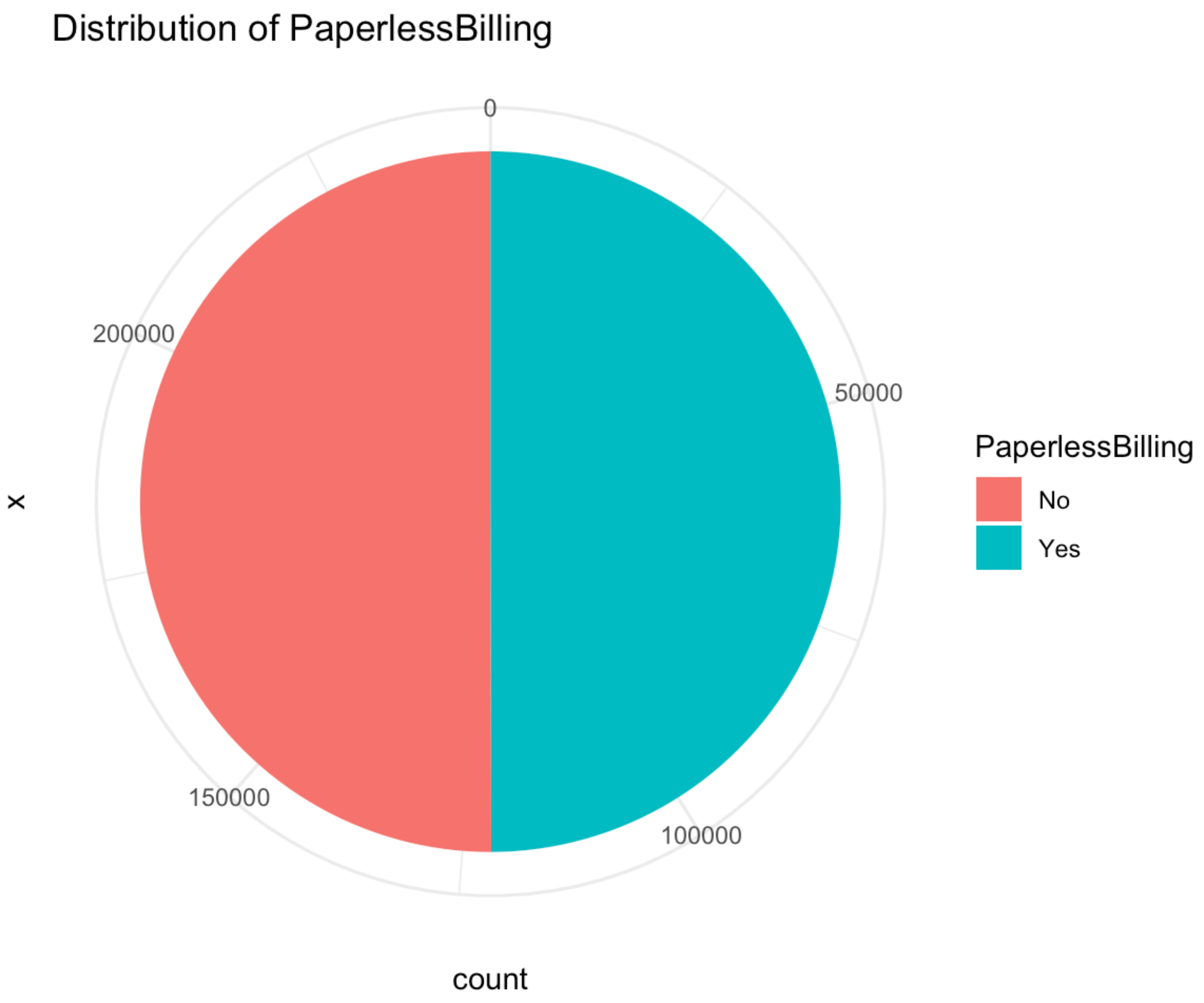
```
pie_chart <- function(data, variable) {
  ggplot(data, aes(x = "", fill = !!as.symbol(variable))) +
    geom_bar(width = 1, stat = "count") +
    coord_polar("y") +
    labs(title = paste("Distribution of", variable))
}

# Example pie charts for some variables
pie_chart(df, "SubscriptionType") +
  theme_minimal()
```



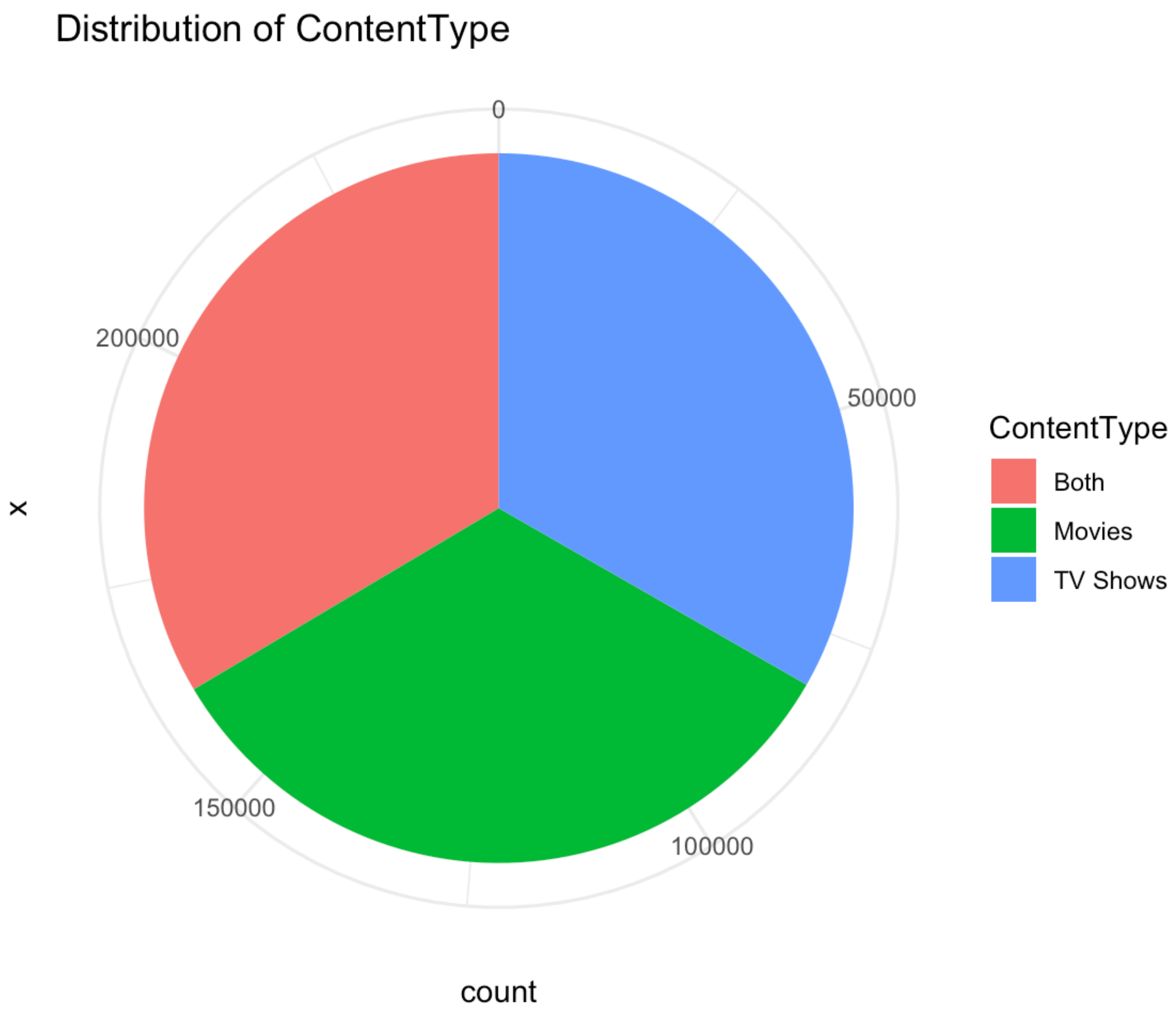
```
pie_chart(df, "PaymentMethod") +
  theme_minimal()
```



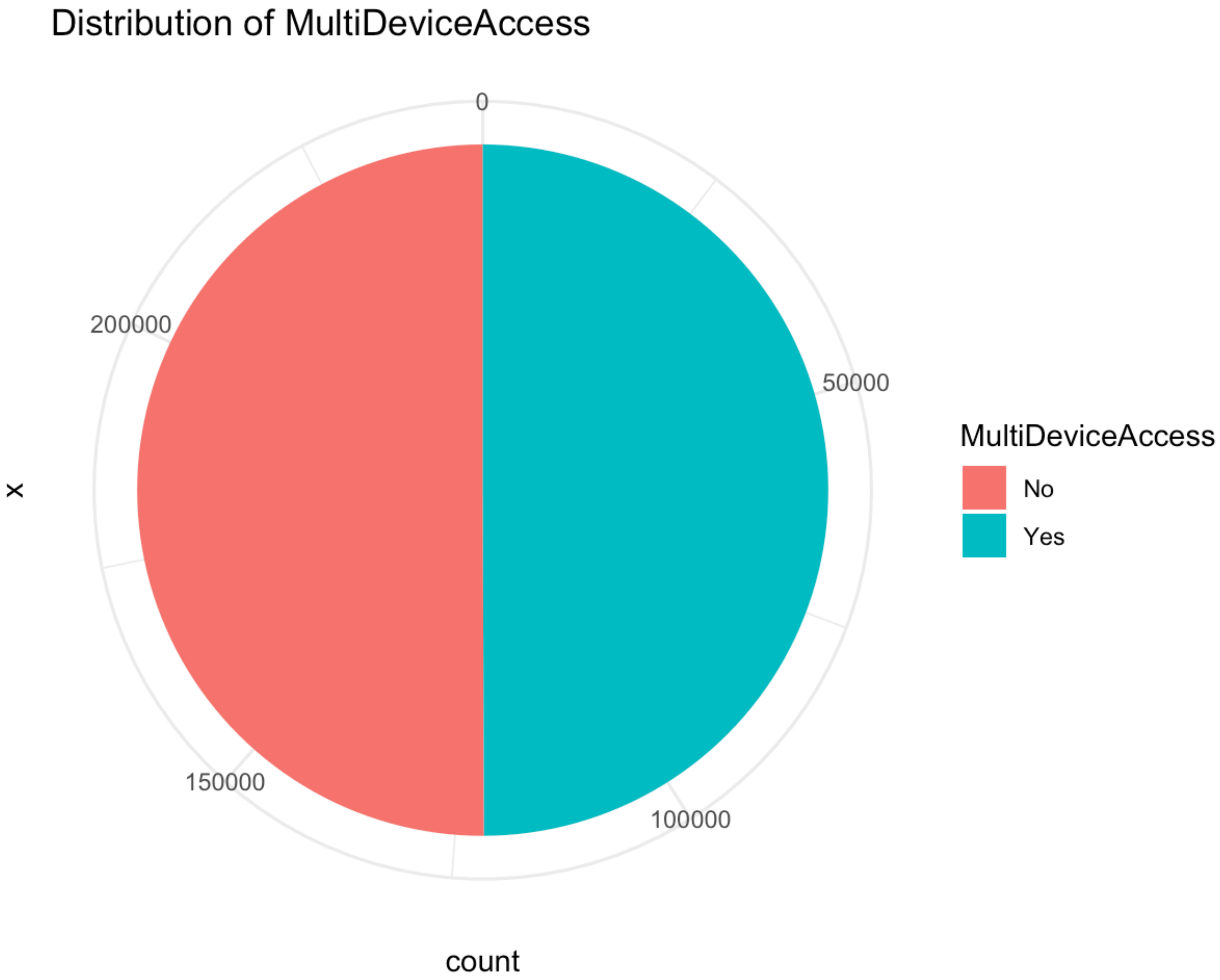
```
pie_chart(df, "PaperlessBilling") +
  theme_minimal()
```



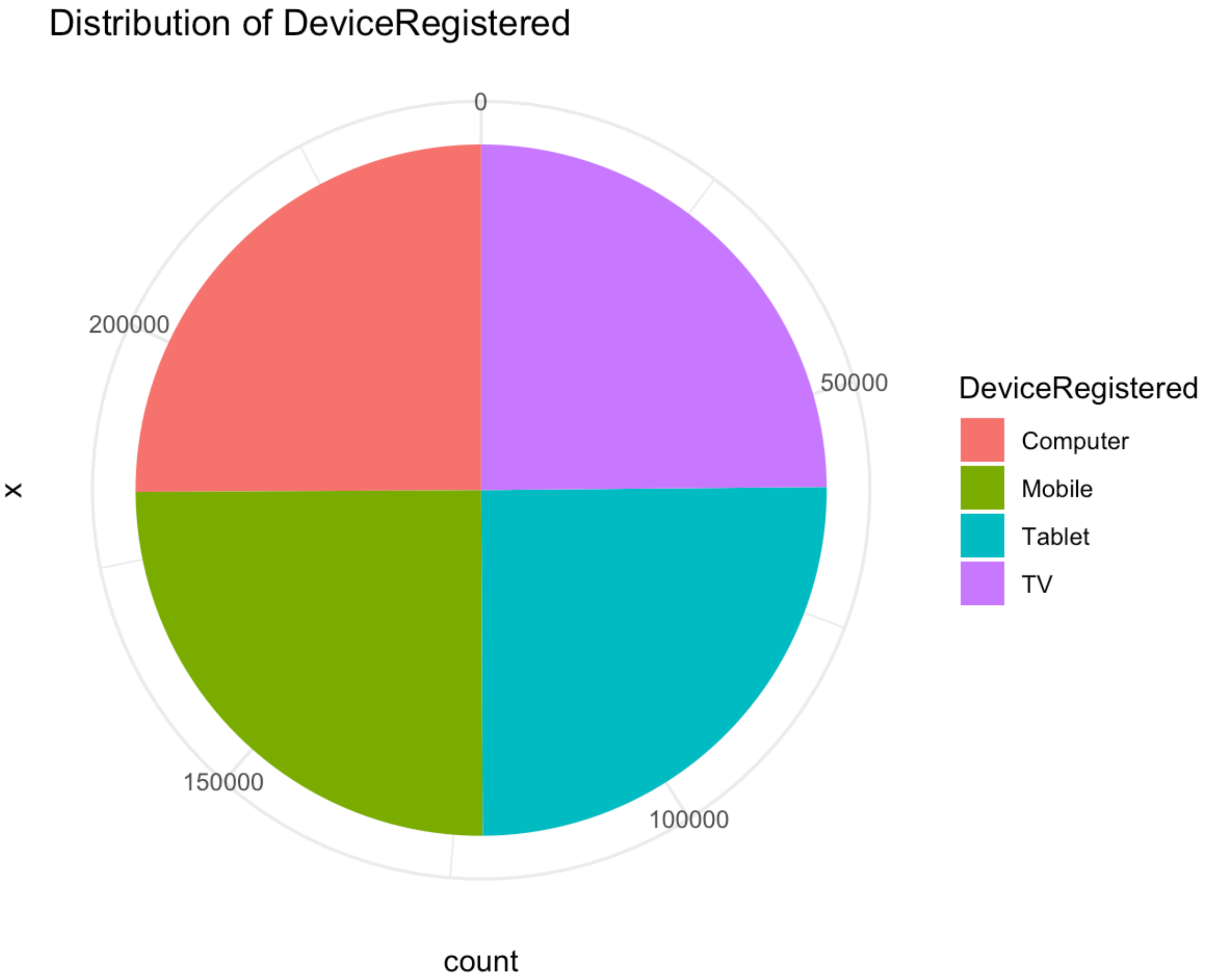
```
pie_chart(df, "ContentType") +
  theme_minimal()
```



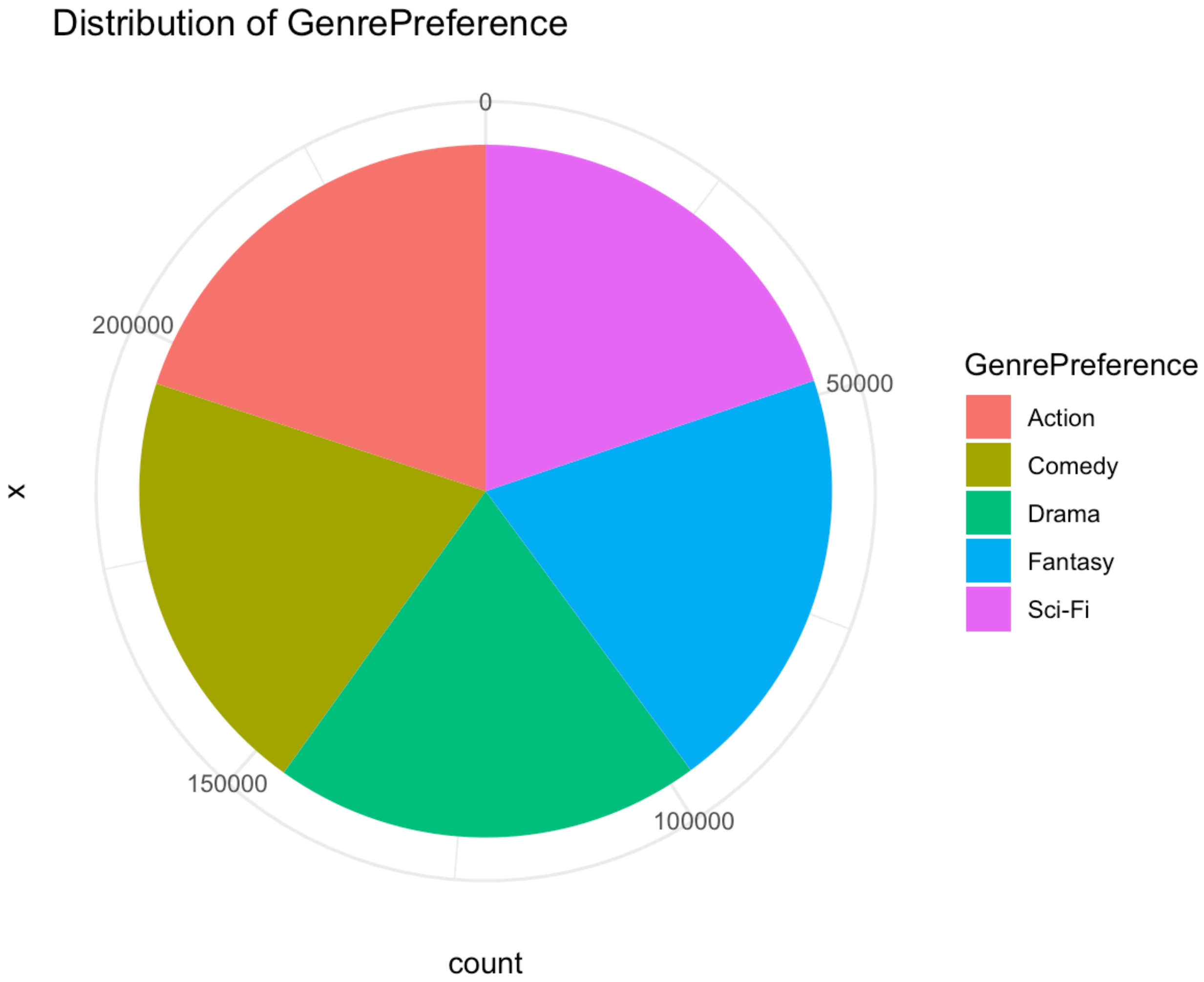
```
pie_chart(df, "MultiDeviceAccess") +
  theme_minimal()
```



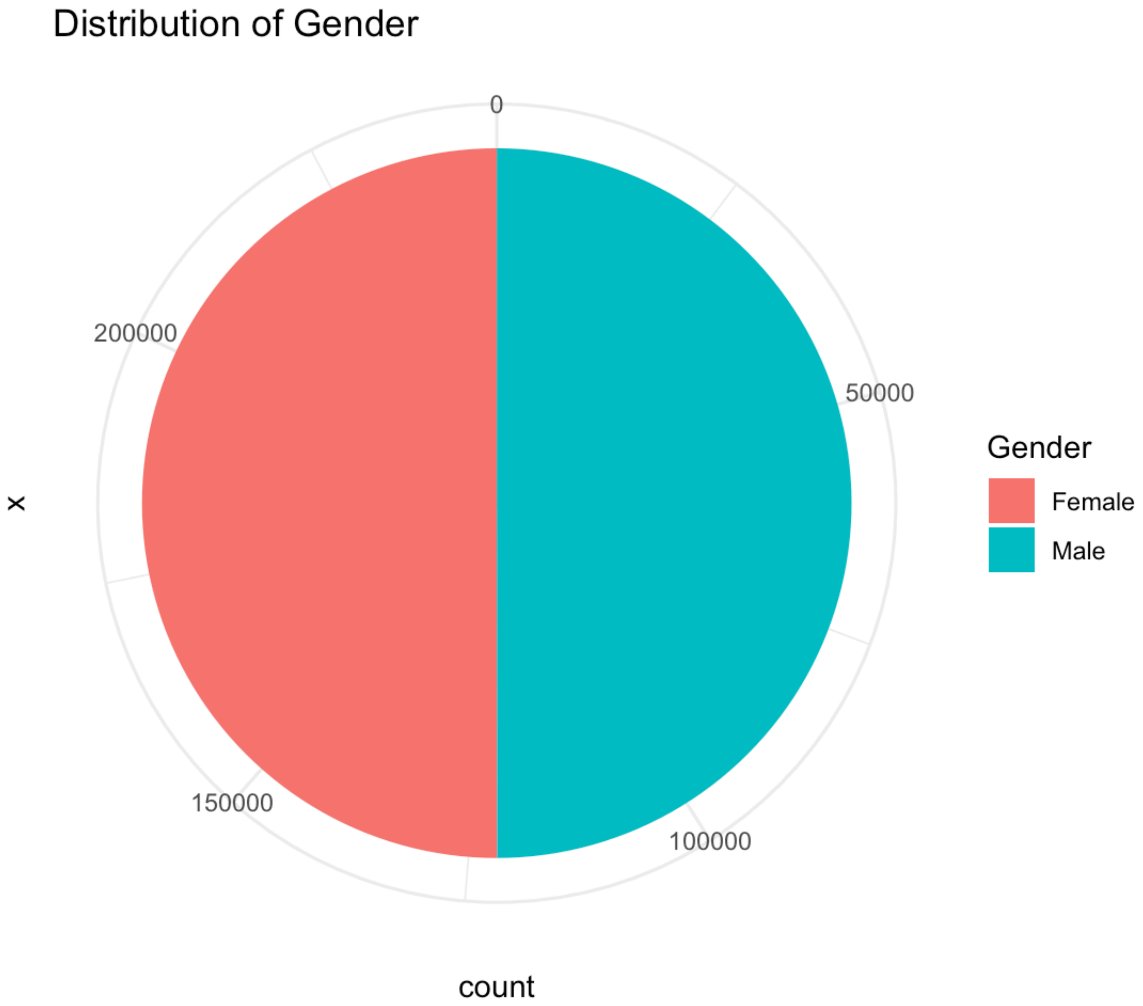
```
pie_chart(df, "DeviceRegistered") +
  theme_minimal()
```



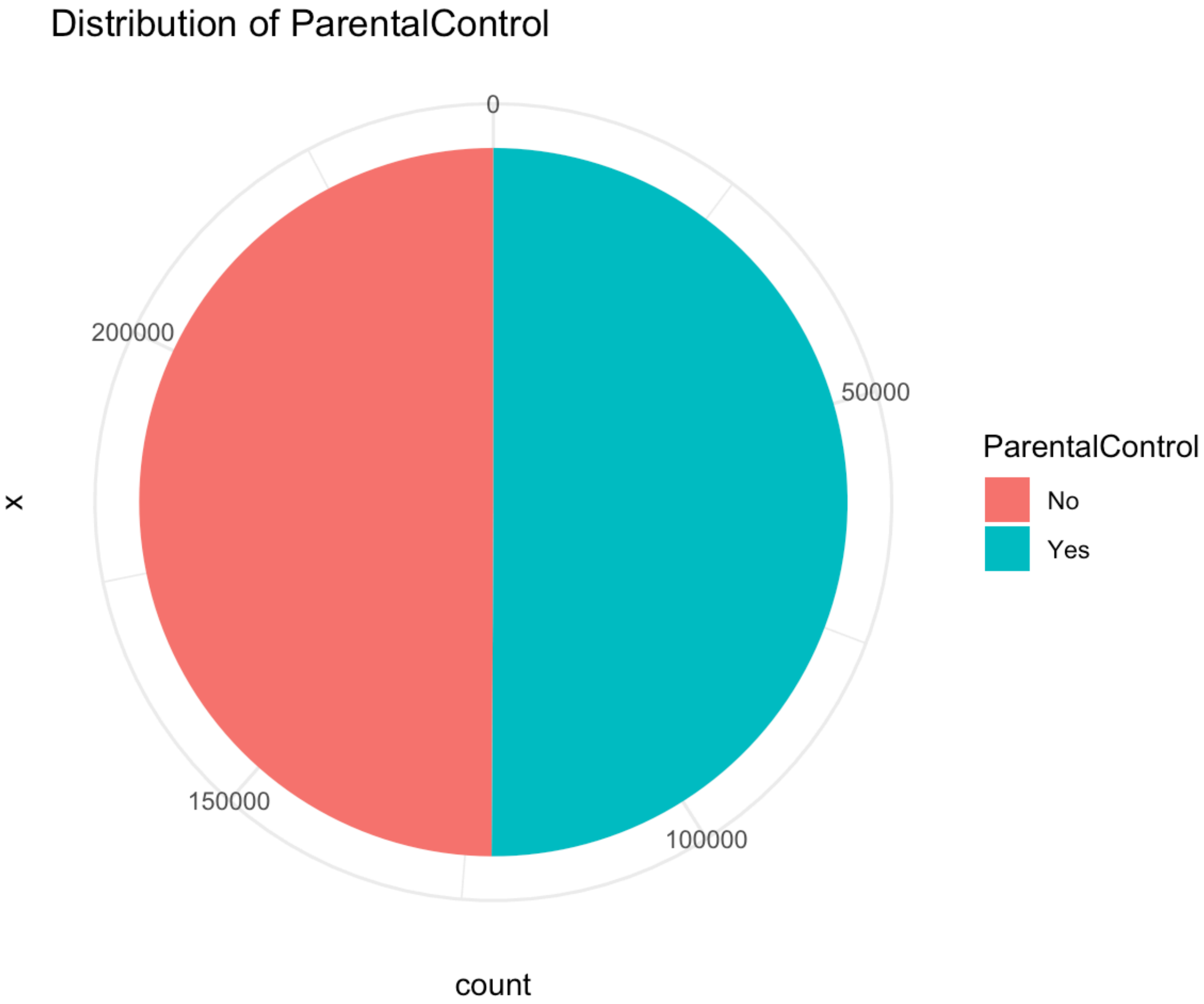
```
pie_chart(df, "GenrePreference") +
  theme_minimal()
```



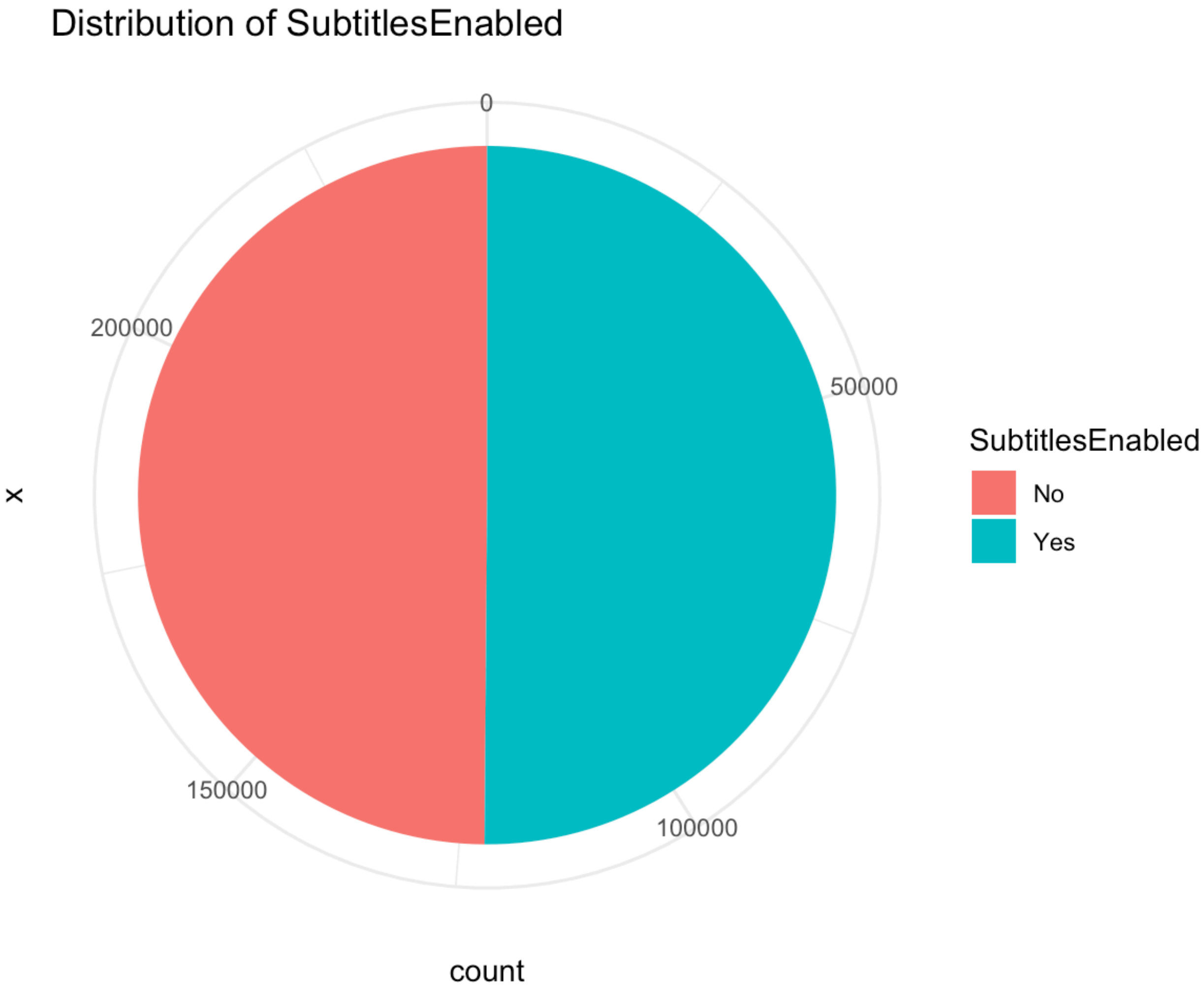
```
pie_chart(df, "Gender") +
  theme_minimal()
```



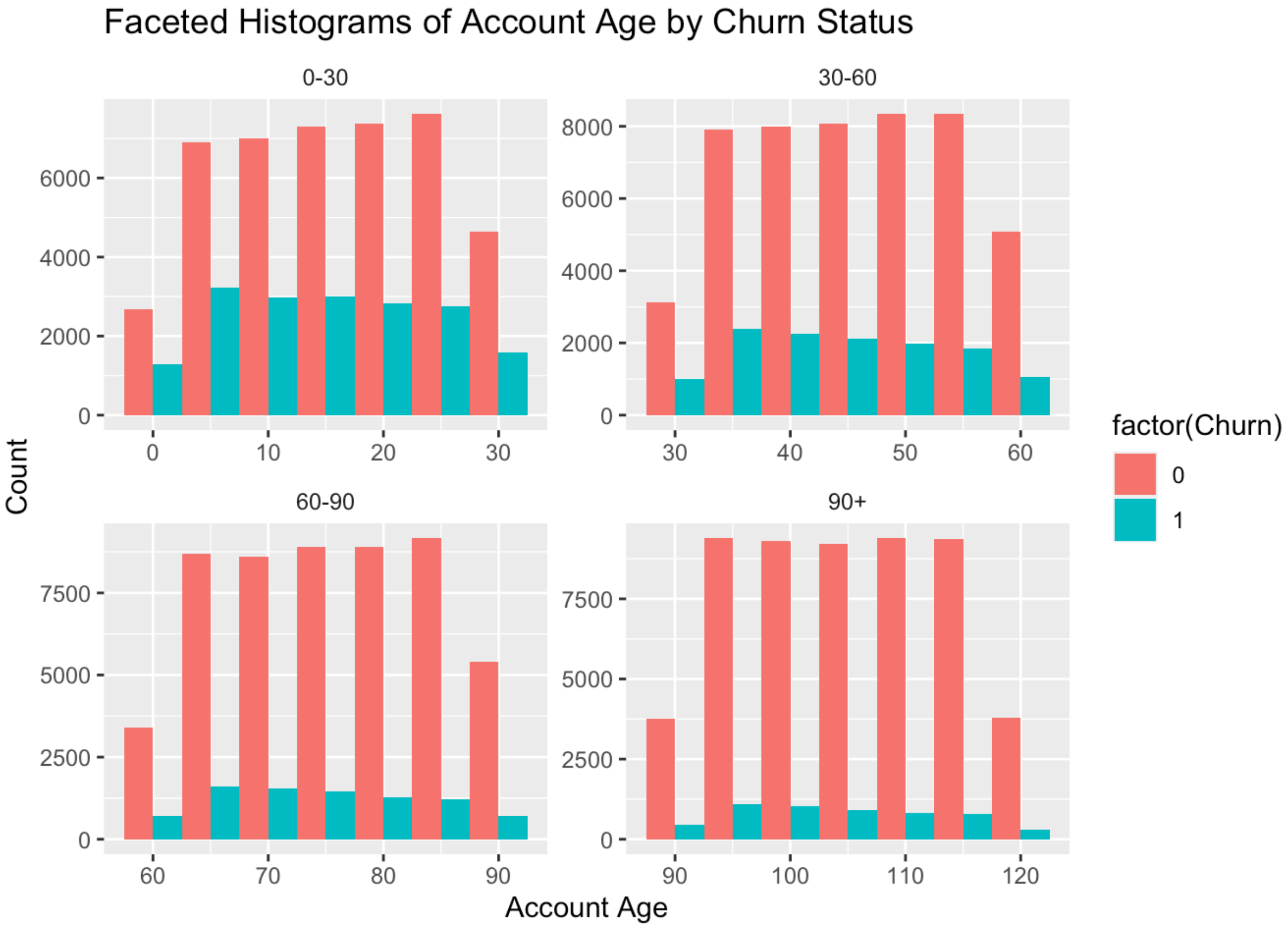
```
pie_chart(df, "ParentalControl") +
  theme_minimal()
```



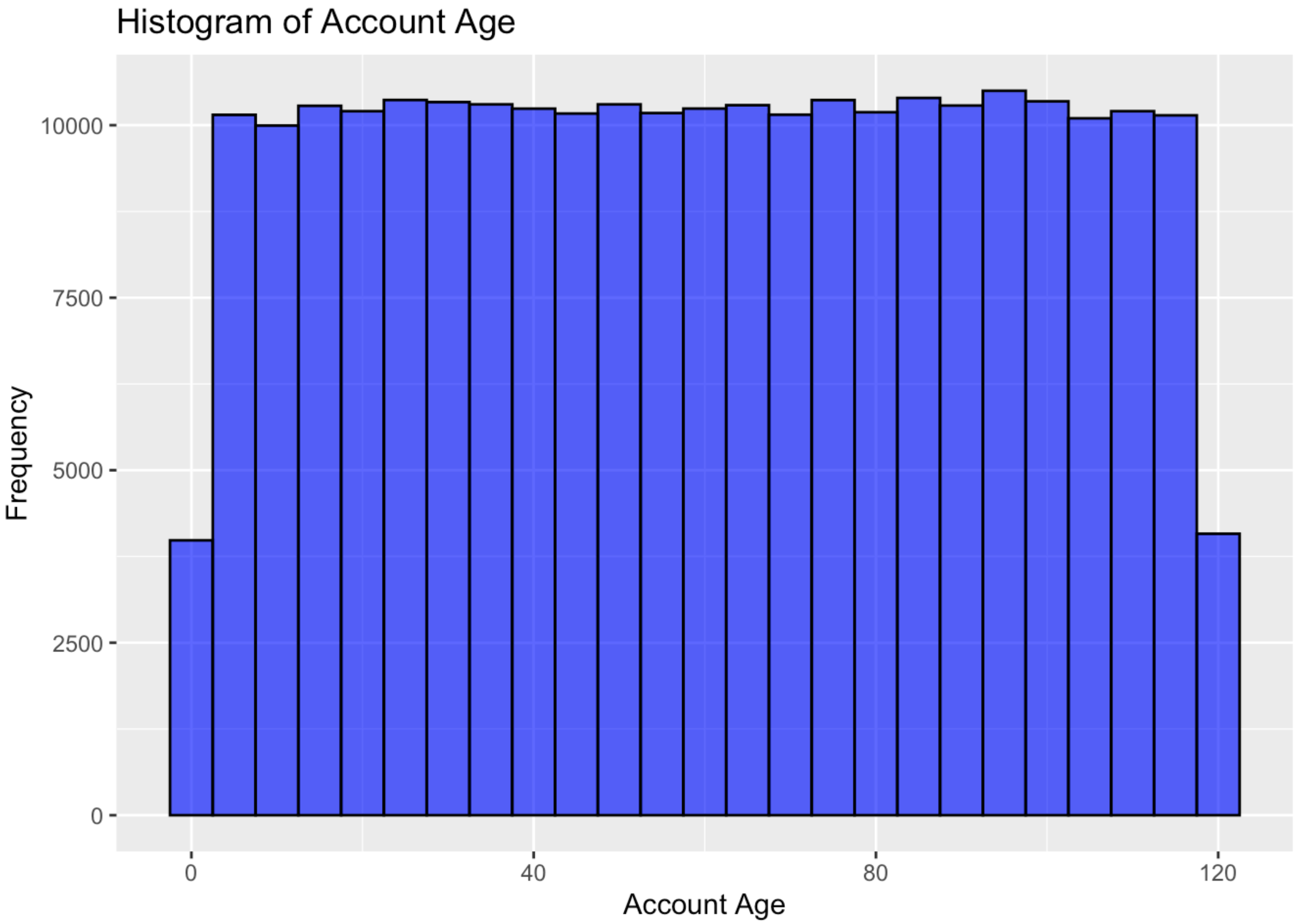
```
pie_chart(df, "SubtitlesEnabled") +
  theme_minimal()
```



```
library(ggplot2)

# Create bands for AccountAge
df$AgeBand <- cut(df$AccountAge, breaks = c(0, 30, 60, 90, Inf), labels = c("0-30", "
30-60", "60-90", "90+"))

# Plot histogram with gaps
ggplot(df, aes(x = AccountAge, fill = factor(Churn))) +
  geom_histogram(binwidth = 5, position = "dodge") +
  facet_wrap(~ AgeBand, scales = "free") +
  labs(title = "Faceted Histograms of Account Age by Churn Status", x = "Account Ag
e", y = "Count") +
  theme(strip.placement = "outside", strip.background = element_blank())
```



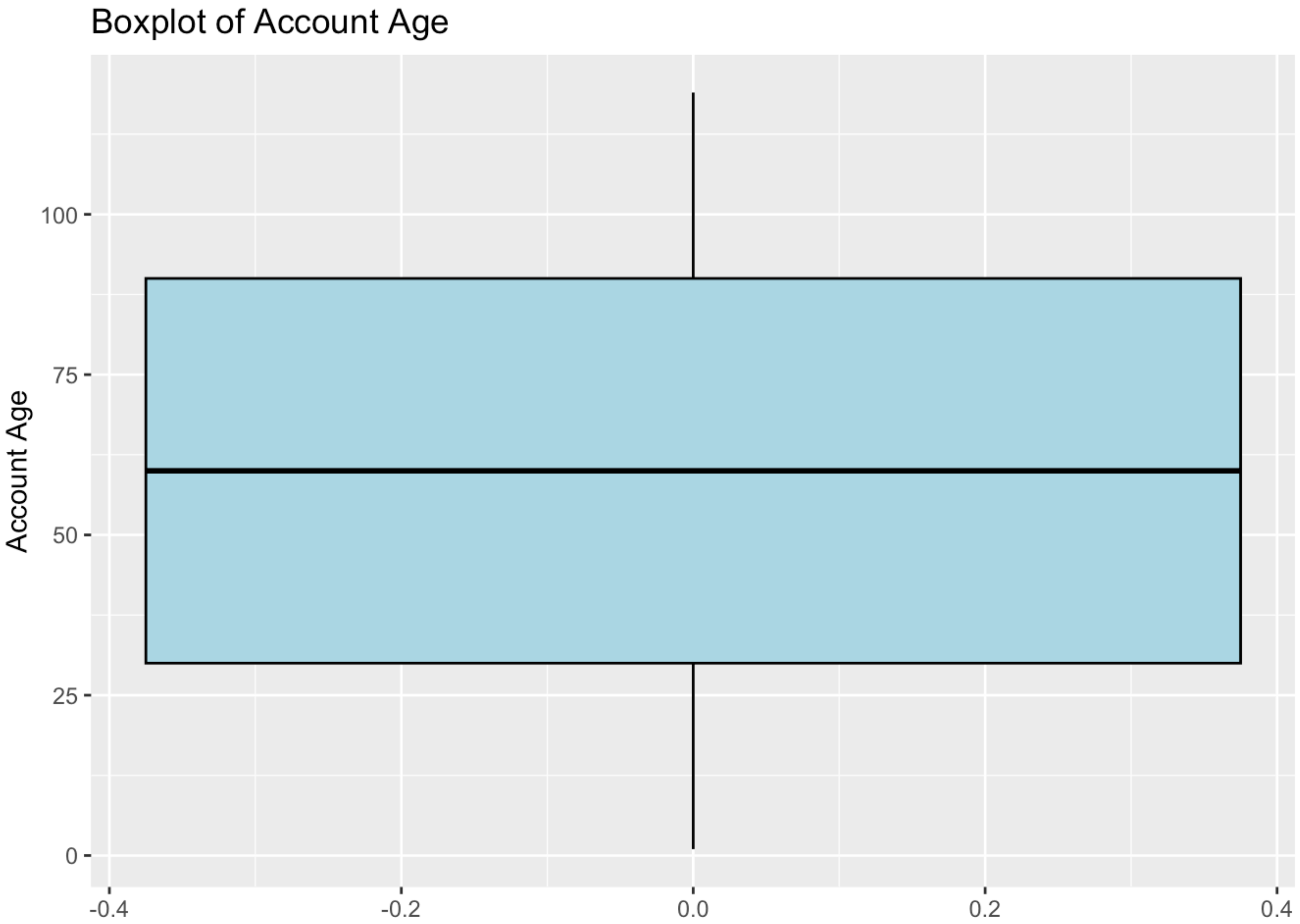
```
# Histogram and summary statistics
ggplot(df, aes(x = AccountAge)) +
  geom_histogram(binwidth = 5, fill = "blue", color = "black", alpha = 0.7) +
  labs(title = "Histogram of Account Age", x = "Account Age", y = "Frequency")
```



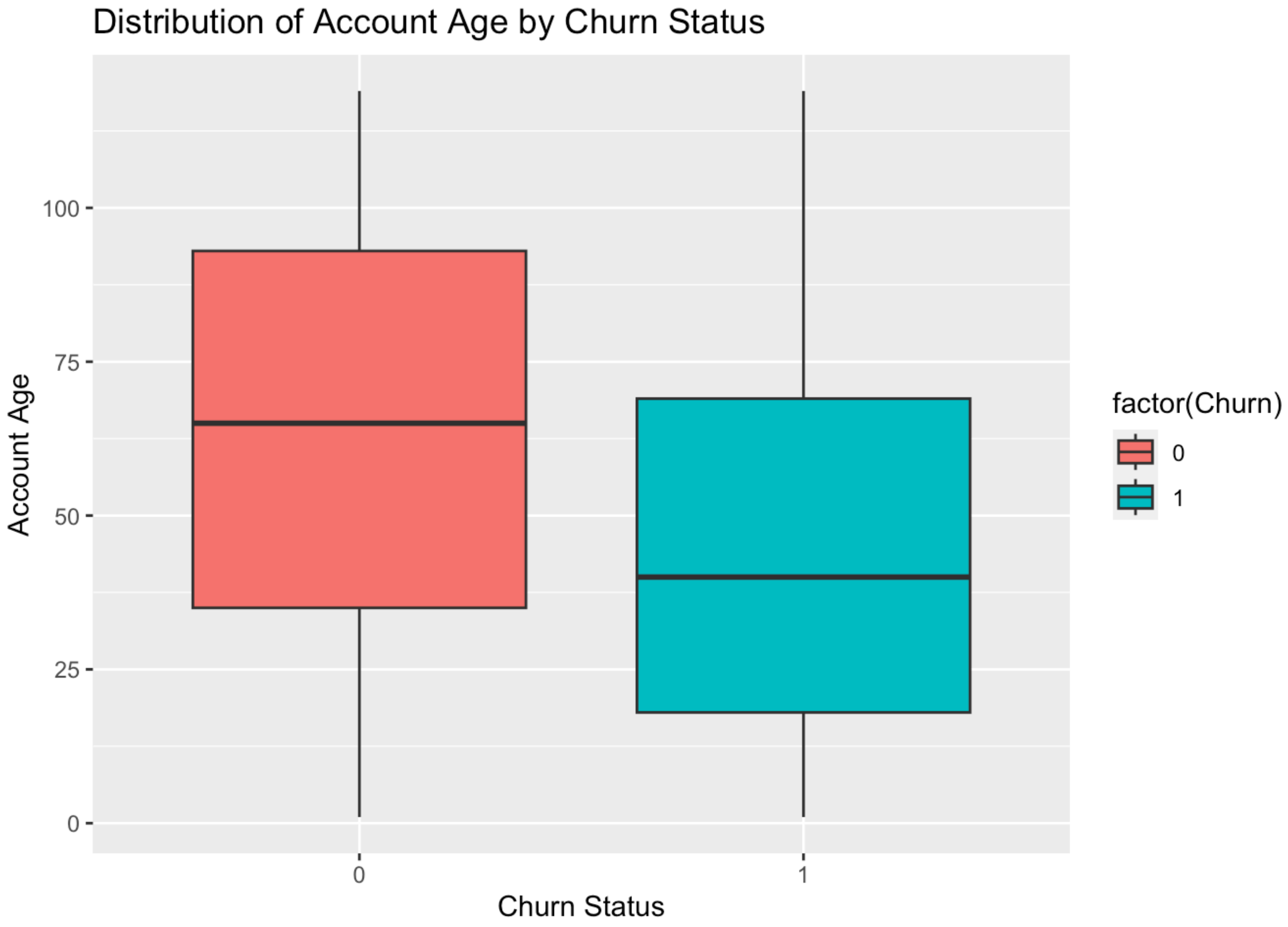
```
summary(df$AccountAge)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    1.00   30.00   60.00   60.08   90.00  119.00 
```

```
ggplot(df, aes(y = AccountAge)) +
  geom_boxplot(fill = "lightblue", color = "black") +
  labs(title = "Boxplot of Account Age", y = "Account Age")
```



```
# Distribution of AccountAge by Churn status
ggplot(df, aes(x = factor(Churn), y = AccountAge, fill = factor(Churn))) +
  geom_boxplot() +
  labs(title = "Distribution of Account Age by Churn Status", x = "Churn Status", y =
"Account Age")
```



```
# Correlation analysis
cor(df$AccountAge, df$Churn)
```

```
## [1] -0.1977356
```