churn_data_analysis.R

sigsp

2021-11-11

```
## Author: Stephen E. Porter
## Title: Churn Data Analysis
## Course: WGU D207: Exploratory Data Analysis
## Instructor: Dr. William Sewell


################################################################################


# Libraries
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1
## Warning: package 'ggplot2' was built under R version 4.1.1
## Warning: package 'tibble' was built under R version 4.1.1
## Warning: package 'tidyr' was built under R version 4.1.1
## Warning: package 'readr' was built under R version 4.1.1
## Warning: package 'dplyr' was built under R version 4.1.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
library(cowplot)
```

```
## Warning: package 'cowplot' was built under R version 4.1.1
```

```
# Importing cleaned data file & getting basic overview

df <- read.csv(file = 'C:/WGU/D207 Exploratory Data Analysis/churn_clean.csv')
colnames(df)
```

```
##  [1] "CaseOrder"           "Customer_id"         "Interaction"
##  [4] "UID"                 "City"                "State"
##  [7] "County"              "Zip"                 "Lat"
## [10] "Lng"                 "Population"          "Area"
## [13] "TimeZone"            "Job"                 "Children"
```

```
## [16] "Age"                 "Income"              "Marital"
## [19] "Gender"              "Churn"               "Outage_sec_perweek"
## [22] "Email"               "Contacts"            "Yearly_equip_failure"
## [25] "Techie"              "Contract"            "Port_modem"
## [28] "Tablet"              "InternetService"     "Phone"
## [31] "Multiple"            "OnlineSecurity"      "OnlineBackup"
## [34] "DeviceProtection"    "TechSupport"         "StreamingTV"
## [37] "StreamingMovies"     "PaperlessBilling"    "PaymentMethod"
## [40] "Tenure"              "MonthlyCharge"       "Bandwidth_GB_Year"
## [43] "Item1"               "Item2"               "Item3"
## [46] "Item4"               "Item5"               "Item6"
## [49] "Item7"               "Item8"
```

```
# Renaming unclear columns named Item1 through Item8 for improved readability &
# confirming they have been renamed correctly

df <- df %>%
  rename(
    Response = Item1,
    Fix = Item2,
    Replacement = Item3,
    Reliability = Item4,
    Options = Item5,
    Respectful = Item6,
    Courteous = Item7,
    Listening = Item8
  )


colnames(df)
```

```
##  [1] "CaseOrder"            "Customer_id"         "Interaction"
##  [4] "UID"                  "City"                "State"
##  [7] "County"               "Zip"                 "Lat"
## [10] "Lng"                  "Population"          "Area"
## [13] "TimeZone"             "Job"                 "Children"
## [16] "Age"                  "Income"              "Marital"
## [19] "Gender"               "Churn"               "Outage_sec_perweek"
## [22] "Email"                "Contacts"            "Yearly_equip_failure"
## [25] "Techie"               "Contract"            "Port_modem"
## [28] "Tablet"               "InternetService"     "Phone"
## [31] "Multiple"             "OnlineSecurity"      "OnlineBackup"
## [34] "DeviceProtection"     "TechSupport"         "StreamingTV"
## [37] "StreamingMovies"      "PaperlessBilling"    "PaymentMethod"
## [40] "Tenure"               "MonthlyCharge"       "Bandwidth_GB_Year"
## [43] "Response"             "Fix"                 "Replacement"
## [46] "Reliability"          "Options"             "Respectful"
## [49] "Courteous"            "Listening"
```

```
# Summary statistics for each column

summary(df)
```

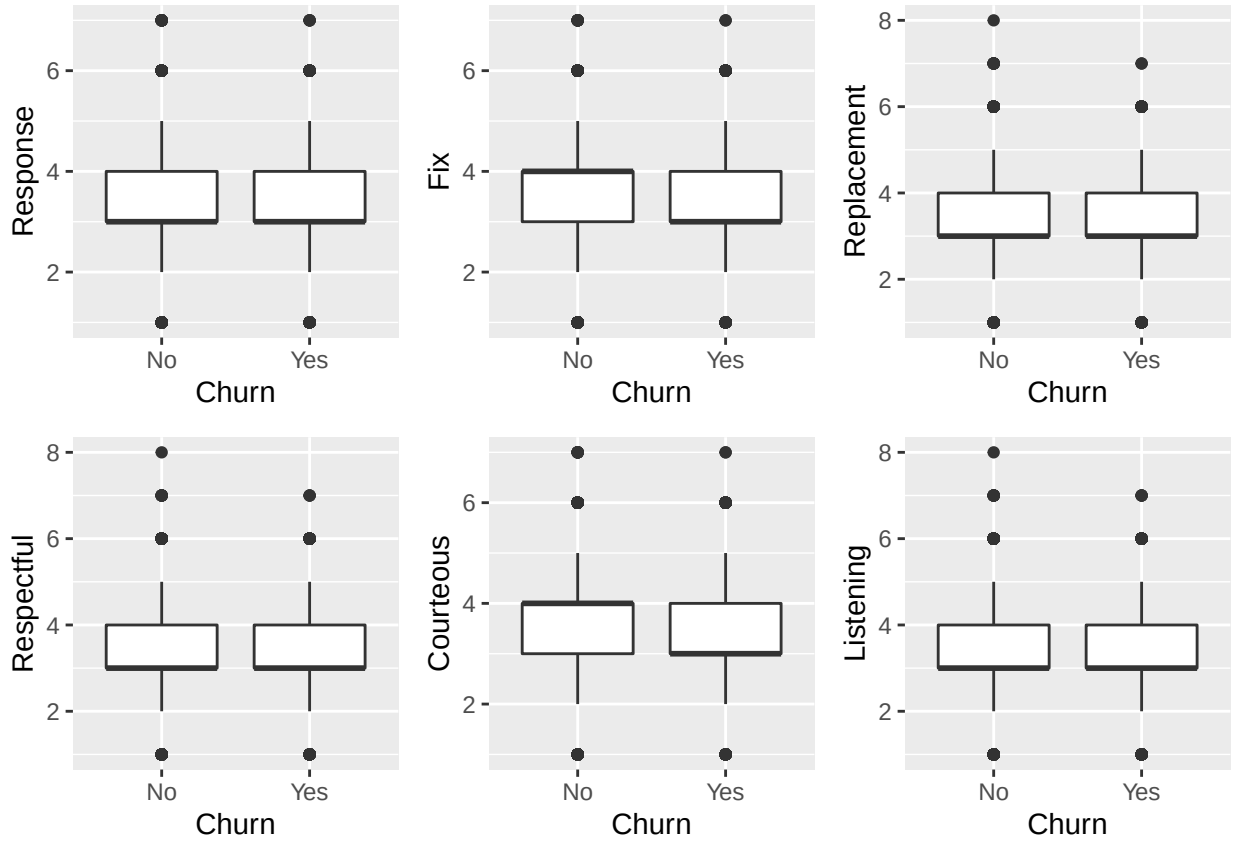
```
##    CaseOrder     Customer_id         Interaction            UID
##  Min.   :    1   Length:10000       Length:10000       Length:10000
##  1st Qu.: 2501   Class :character   Class :character   Class :character
```

```
##  Median : 5000   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 5000
##  3rd Qu.: 7500
##  Max.   :10000
##      City              State              County               Zip
##  Length:10000       Length:10000       Length:10000       Min.   :  601
##  Class :character   Class :character   Class :character   1st Qu.:26293
##  Mode  :character   Mode  :character   Mode  :character   Median :48870
##                                                           Mean   :49153
##                                                           3rd Qu.:71867
##                                                           Max.   :99929
##       Lat              Lng             Population         Area
##  Min.   :17.97   Min.   :-171.69   Min.   :     0   Length:10000
##  1st Qu.:35.34   1st Qu.: -97.08   1st Qu.:   738   Class :character
##  Median :39.40   Median : -87.92   Median :  2910   Mode  :character
##  Mean   :38.76   Mean   : -90.78   Mean   :  9757
##  3rd Qu.:42.11   3rd Qu.: -80.09   3rd Qu.: 13168
##  Max.   :70.64   Max.   : -65.67   Max.   :111850
##    TimeZone             Job               Children           Age
##  Length:10000       Length:10000       Min.   : 0.000   Min.   :18.00
##  Class :character   Class :character   1st Qu.: 0.000   1st Qu.:35.00
##  Mode  :character   Mode  :character   Median : 1.000   Median :53.00
##                                        Mean   : 2.088   Mean   :53.08
##                                        3rd Qu.: 3.000   3rd Qu.:71.00
##                                        Max.   :10.000   Max.   :89.00
##      Income            Marital             Gender             Churn
##  Min.   :   348.7   Length:10000       Length:10000       Length:10000
##  1st Qu.: 19224.7   Class :character   Class :character   Class :character
##  Median : 33170.6   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 39806.9
##  3rd Qu.: 53246.2
##  Max.   :258900.7
##  Outage_sec_perweek     Email          Contacts      Yearly_equip_failure
##  Min.   : 0.09975   Min.   : 1.00   Min.   :0.0000   Min.   :0.000
##  1st Qu.: 8.01821   1st Qu.:10.00   1st Qu.:0.0000   1st Qu.:0.000
##  Median :10.01856   Median :12.00   Median :1.0000   Median :0.000
##  Mean   :10.00185   Mean   :12.02   Mean   :0.9942   Mean   :0.398
##  3rd Qu.:11.96949   3rd Qu.:14.00   3rd Qu.:2.0000   3rd Qu.:1.000
##  Max.   :21.20723   Max.   :23.00   Max.   :7.0000   Max.   :6.000
##     Techie            Contract          Port_modem           Tablet
##  Length:10000       Length:10000       Length:10000       Length:10000
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  InternetService       Phone             Multiple         OnlineSecurity
##  Length:10000       Length:10000       Length:10000       Length:10000
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  OnlineBackup       DeviceProtection   TechSupport        StreamingTV
```

```
##  Length:10000       Length:10000       Length:10000       Length:10000
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  StreamingMovies    PaperlessBilling   PaymentMethod          Tenure
##  Length:10000       Length:10000       Length:10000       Min.   : 1.000
##  Class :character   Class :character   Class :character   1st Qu.: 7.918
##  Mode  :character   Mode  :character   Mode  :character   Median :35.431
##                                                           Mean   :34.526
##                                                           3rd Qu.:61.480
##                                                           Max.   :71.999
##  MonthlyCharge    Bandwidth_GB_Year    Response          Fix
##  Min.   : 79.98   Min.   : 155.5    Min.   :1.000   Min.   :1.000
##  1st Qu.:139.98   1st Qu.:1236.5    1st Qu.:3.000   1st Qu.:3.000
##  Median :167.48   Median :3279.5    Median :3.000   Median :4.000
##  Mean   :172.62   Mean   :3392.3    Mean   :3.491   Mean   :3.505
##  3rd Qu.:200.73   3rd Qu.:5586.1    3rd Qu.:4.000   3rd Qu.:4.000
##  Max.   :290.16   Max.   :7159.0    Max.   :7.000   Max.   :7.000
##   Replacement     Reliability       Options        Respectful      Courteous
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.00
##  1st Qu.:3.000   1st Qu.:3.000   1st Qu.:3.000   1st Qu.:3.000   1st Qu.:3.00
##  Median :3.000   Median :3.000   Median :3.000   Median :3.000   Median :4.00
##  Mean   :3.487   Mean   :3.498   Mean   :3.493   Mean   :3.497   Mean   :3.51
##  3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.00
##  Max.   :8.000   Max.   :7.000   Max.   :7.000   Max.   :8.000   Max.   :7.00
##    Listening
##  Min.   :1.000
##  1st Qu.:3.000
##  Median :3.000
##  Mean   :3.496
##  3rd Qu.:4.000
##  Max.   :8.000
```

```
################################################################################


# Analysis Question: Do any of the variables identified in the Principal
# Component Analysis have an effect on customer churn?


################################################################################


# Analysis of variables in PC1: Response, Fix, Replacement, Respectful,
# Courteous, Listening

plot_grid(

  ggplot(df, aes(x=Churn, y=Response)) +
    geom_boxplot(),

  ggplot(df, aes(x=Churn, y=Fix)) +
```

```
  geom_boxplot(),

ggplot(df, aes(x=Churn, y=Replacement)) +
  geom_boxplot(),

ggplot(df, aes(x=Churn, y=Respectful)) +
  geom_boxplot(),

ggplot(df, aes(x=Churn, y=Courteous)) +
  geom_boxplot(),

ggplot(df, aes(x=Churn, y=Listening)) +
  geom_boxplot(),

ncol = 3, nrow = 2)
```




```
response_churn <- table(df$Churn, df$Response)
summary(response_churn)
```

```
## Number of cases in table: 10000
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 4.332, df = 6, p-value = 0.6318
```

```
fix_churn <- table(df$Churn, df$Fix)
summary(fix_churn)
```

```
## Number of cases in table: 10000
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 5.272, df = 6, p-value = 0.5094
##  Chi-squared approximation may be incorrect
```

```
replacement_churn <- table(df$Churn, df$Replacement)
summary(replacement_churn)
```

```
## Number of cases in table: 10000
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 5.371, df = 7, p-value = 0.6148
##  Chi-squared approximation may be incorrect
```

```
respectful_churn <- table(df$Churn, df$Respectful)
summary(respectful_churn)
```

```
## Number of cases in table: 10000
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 4.129, df = 7, p-value = 0.7648
##  Chi-squared approximation may be incorrect
```

```
courteous_churn <- table(df$Churn, df$Courteous)
summary(courteous_churn)
```

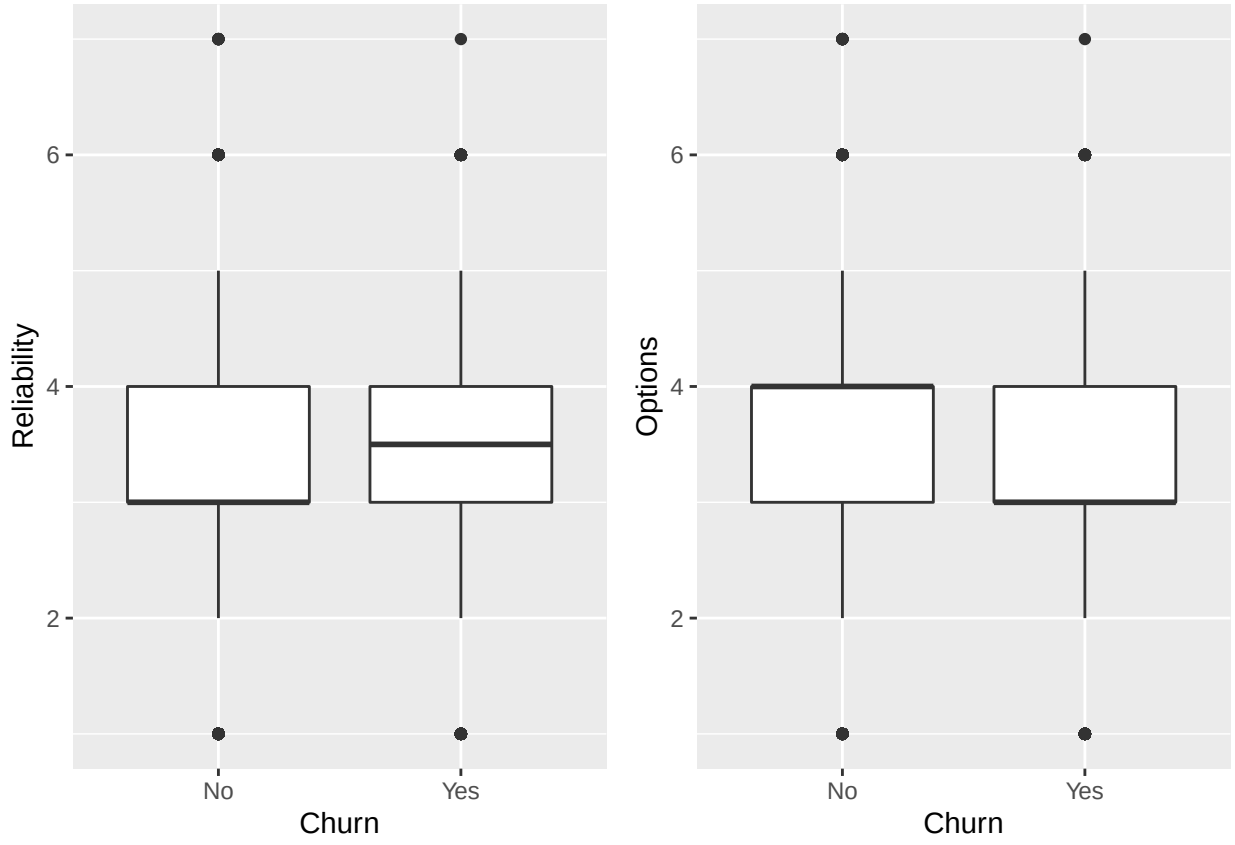
```
## Number of cases in table: 10000
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 5.638, df = 6, p-value = 0.465
##  Chi-squared approximation may be incorrect
```

```
listening_churn <- table(df$Churn, df$Listening)
summary(listening_churn)
```

```
## Number of cases in table: 10000
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 1.714, df = 7, p-value = 0.974
##  Chi-squared approximation may be incorrect
```

```
# All p-values lie outside of the standard 0.05 alpha value. We cannot reject
# the null hypothesis.

################################################################################


# Analysis of variables from PC4: Reliability, Options

plot_grid(

  ggplot(df, aes(x=Churn, y=Reliability)) +
    geom_boxplot(),

  ggplot(df, aes(x=Churn, y=Options)) +
    geom_boxplot(),
```

```
  ncol = 2, nrow = 1)
```




```
reliability_churn <- table(df$Churn, df$Reliability)
summary(reliability_churn)
```

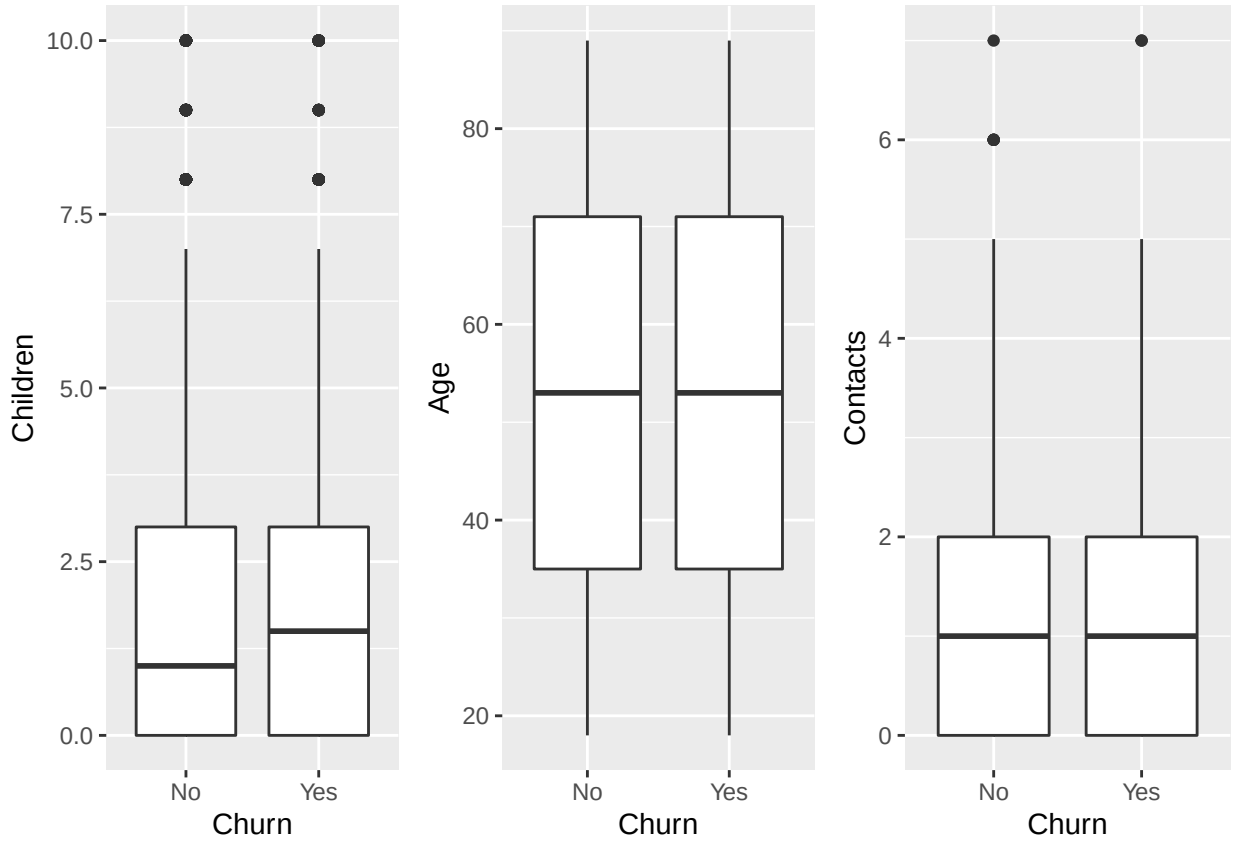
```
## Number of cases in table: 10000
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 2.9611, df = 6, p-value = 0.8137
##  Chi-squared approximation may be incorrect
```

```
options_churn <- table(df$Churn, df$Options)
summary(options_churn)
```

```
## Number of cases in table: 10000
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 5.625, df = 6, p-value = 0.4665
##  Chi-squared approximation may be incorrect
```

```
# All p-values lie outside of the standard 0.05 alpha value. We cannot reject
# the null hypothesis.


################################################################################
```

```
# Analysis of variables in PC6: Children, Age, Contacts

plot_grid(

  ggplot(df, aes(x=Churn, y=Children)) +
    geom_boxplot(),

  ggplot(df, aes(x=Churn, y=Age)) +
    geom_boxplot(),

  ggplot(df, aes(x=Churn, y=Contacts)) +
    geom_boxplot(),

  ncol = 3, nrow = 1)
```




```
children_churn <- table(df$Churn, df$Children)
summary(children_churn)
```

```
## Number of cases in table: 10000
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 6.581, df = 10, p-value = 0.7644
```

```
age_churn <- table(df$Churn, df$Age)
summary(age_churn)
```

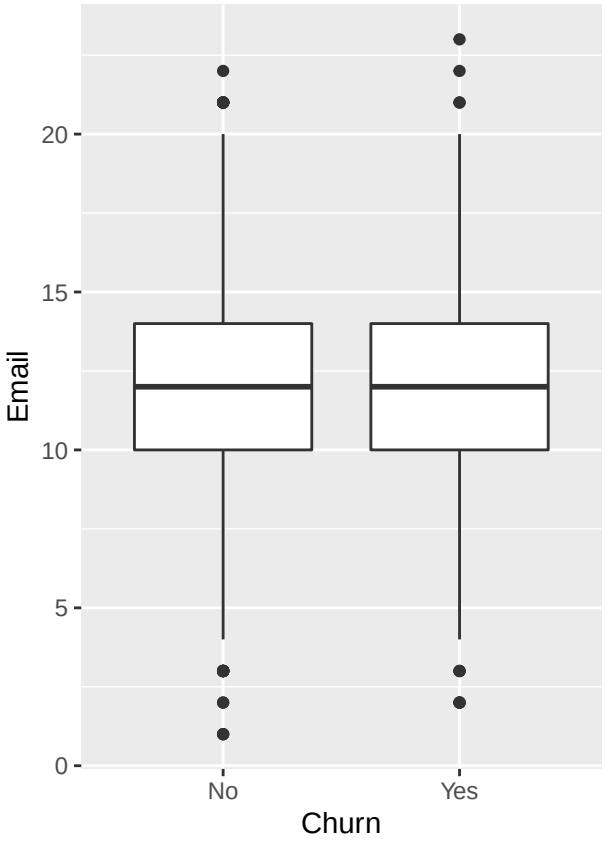
```
## Number of cases in table: 10000
```

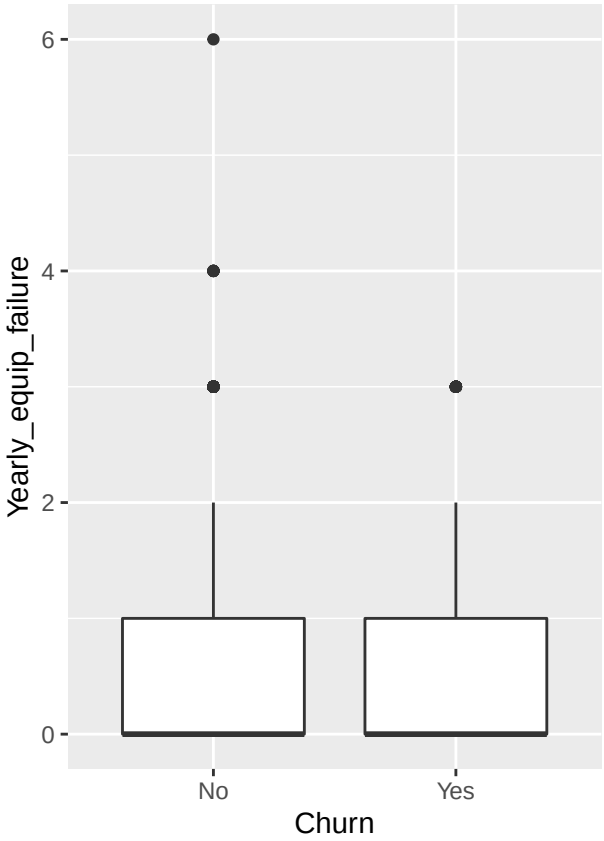
```
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 61.97, df = 71, p-value = 0.769
```

```
contacts_churn <- table(df$Churn, df$Contacts)
summary(contacts_churn)
```

```
## Number of cases in table: 10000
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 5.522, df = 7, p-value = 0.5966
##  Chi-squared approximation may be incorrect
```

```
# All p-values lie outside of the standard 0.05 alpha value. We cannot reject
# the null hypothesis.


################################################################################


# Analysis of variables in PC7: Email, Yearly_equip_failure

plot_grid(

  ggplot(df, aes(x=Churn, y=Email)) +
    geom_boxplot(),

  ggplot(df, aes(x=Churn, y=Yearly_equip_failure)) +
    geom_boxplot(),

  ncol = 2, nrow = 1)
```






```
email_churn <- table(df$Churn, df$Email)
summary(email_churn)
```

```
## Number of cases in table: 10000
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 23.111, df = 22, p-value = 0.3955
##  Chi-squared approximation may be incorrect
```

```
yef_churn <- table(df$Churn, df$Yearly_equip_failure)
summary(yef_churn)
```

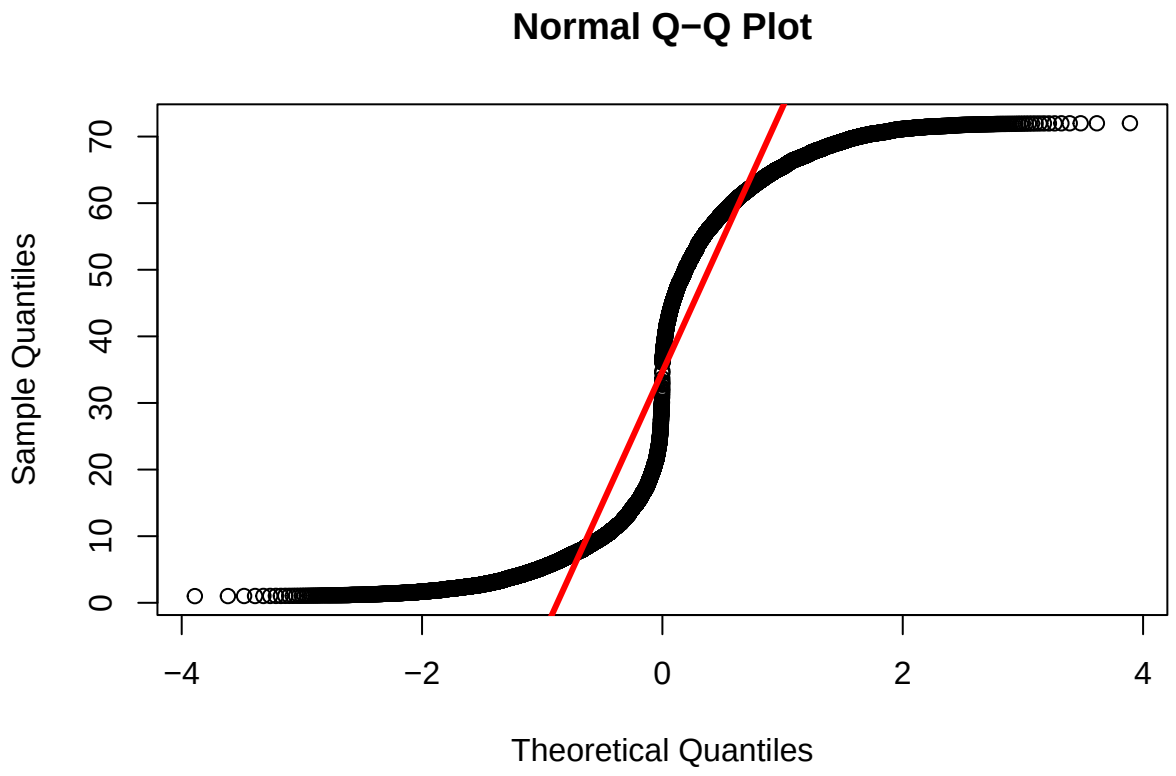
```
## Number of cases in table: 10000
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 6.925, df = 5, p-value = 0.2263
##  Chi-squared approximation may be incorrect
```

```
# All p-values lie outside of the standard 0.05 alpha value. We cannot reject
# the null hypothesis.


################################################################################


# Graphing various relationships in the data frame

# Q-Q Plots of Tenure, Bandwidth_GB_Year and MonthlyCharge
```

```
plot.new
```

```
## function ()
## {
##     for (fun in getHook("before.plot.new")) {
##         if (is.character(fun))
##             fun <- get(fun)
##         try(fun())
##     }
##     .External2(C_plot_new)
##     grDevices:::recordPalette()
##     for (fun in getHook("plot.new")) {
##         if (is.character(fun))
##             fun <- get(fun)
##         try(fun())
##     }
##     invisible()
## }
## <bytecode: 0x000000001d31cc60>
## <environment: namespace:graphics>
```

```
qqnorm(df$Tenure, pch = 1)
qqline(df$Tenure, col = "red", lwd = 3)
```

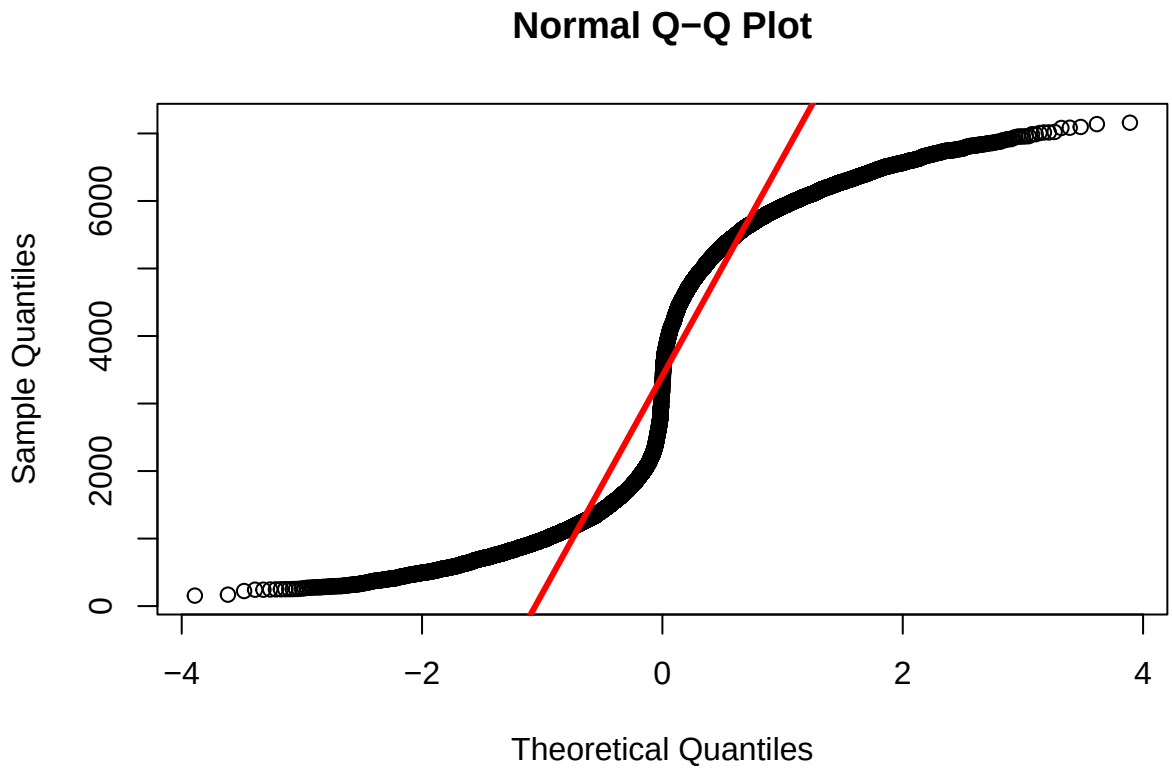



```
plot.new
```

```
## function ()
```

```
## {
##     for (fun in getHook("before.plot.new")) {
##         if (is.character(fun))
##             fun <- get(fun)
##         try(fun())
##     }
##     .External2(C_plot_new)
##     grDevices:::recordPalette()
##     for (fun in getHook("plot.new")) {
##         if (is.character(fun))
##             fun <- get(fun)
##         try(fun())
##     }
##     invisible()
## }
## <bytecode: 0x000000001d31cc60>
## <environment: namespace:graphics>
```

```
qqnorm(df$Bandwidth_GB_Year, pch = 1)
qqline(df$Bandwidth_GB_Year, col = "red", lwd = 3)
```

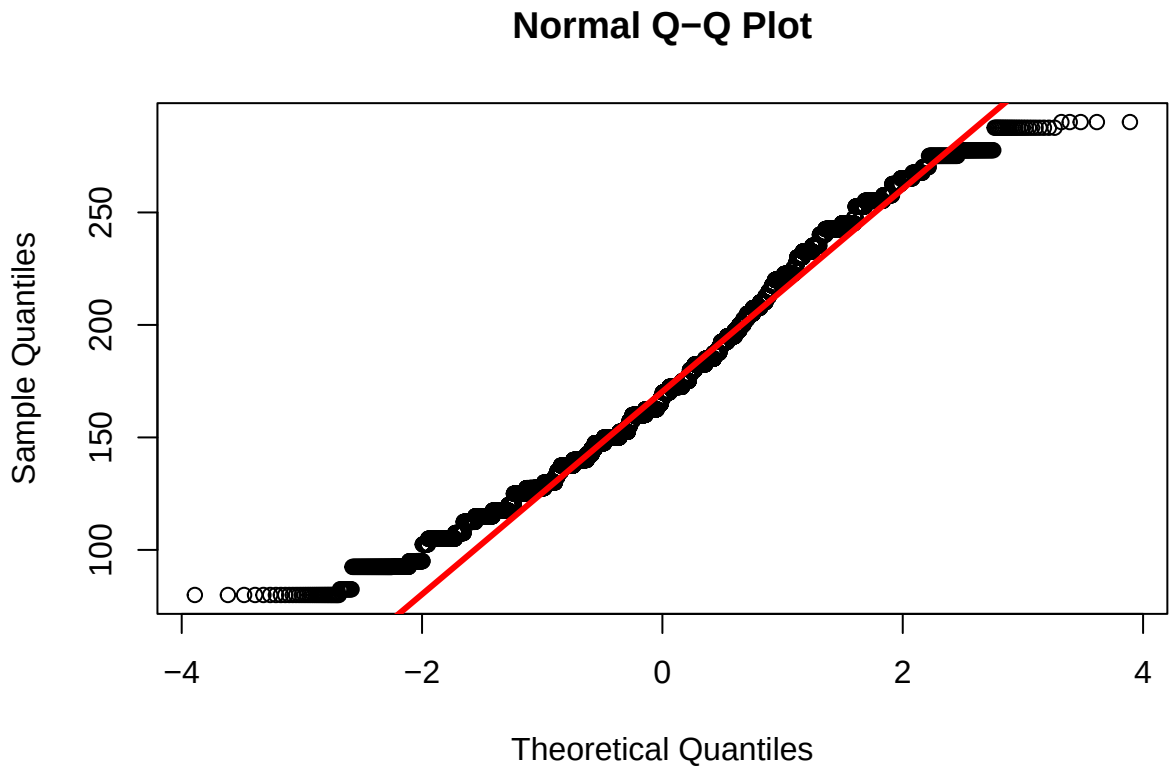



```
plot.new
```

```
## function ()
## {
##     for (fun in getHook("before.plot.new")) {
##         if (is.character(fun))
```

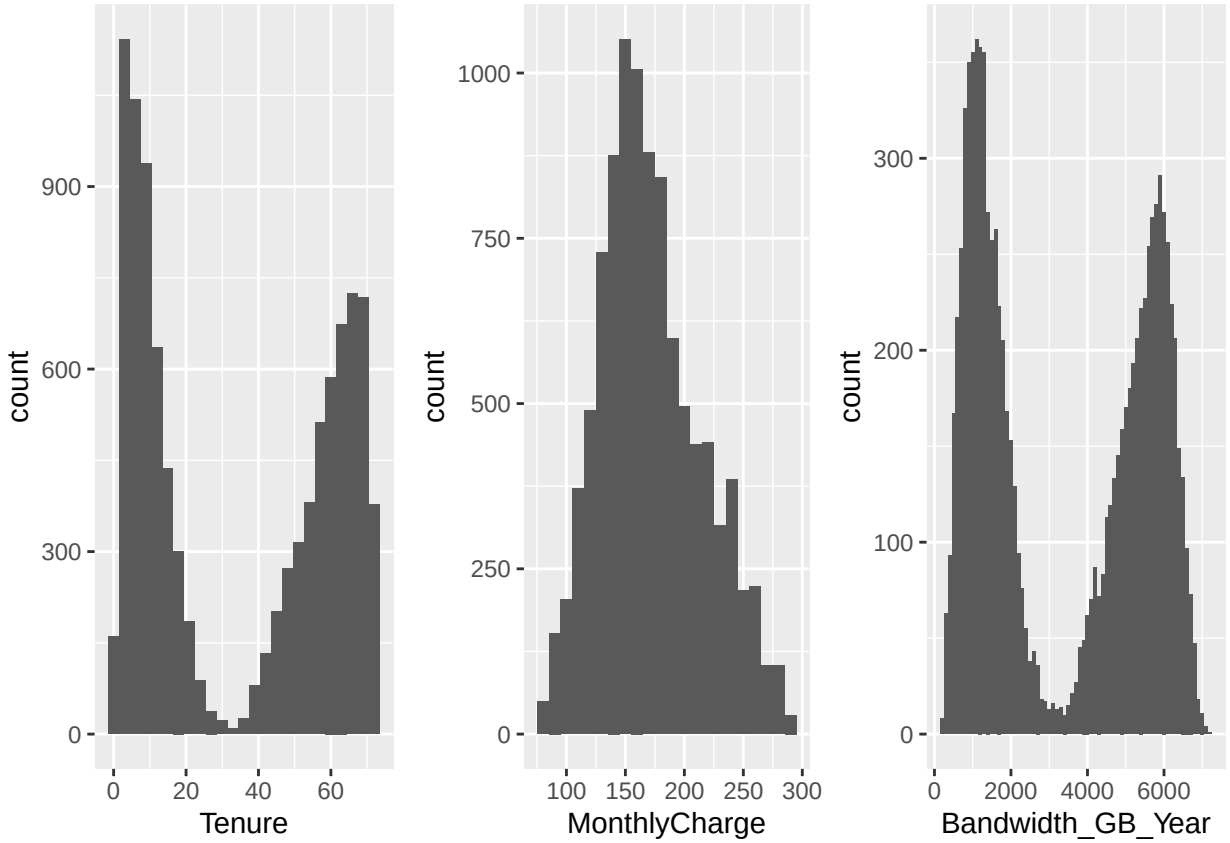
```
##             fun <- get(fun)
##         try(fun())
##     }
##     .External2(C_plot_new)
##     grDevices:::recordPalette()
##     for (fun in getHook("plot.new")) {
##         if (is.character(fun))
##             fun <- get(fun)
##         try(fun())
##     }
##     invisible()
## }
## <bytecode: 0x000000001d31cc60>
## <environment: namespace:graphics>
```

```
qqnorm(df$MonthlyCharge, pch = 1)
qqline(df$MonthlyCharge, col = "red", lwd = 3)
```

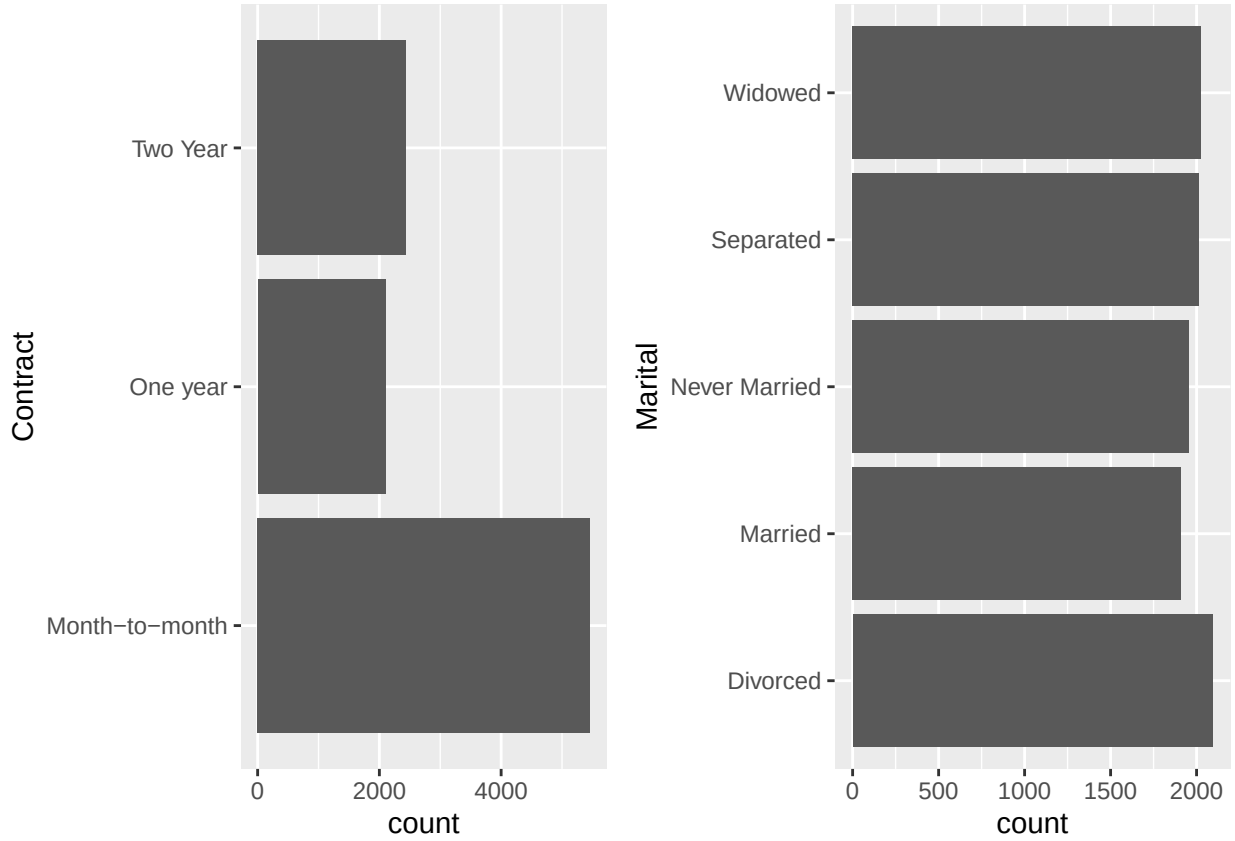



```
# Univariate graphs of continuous & categorical variables

plot_grid(
  ggplot(df, aes(x=Tenure)) +
    geom_histogram(binwidth = 3),

  ggplot(df, aes(x=MonthlyCharge)) +
    geom_histogram(binwidth = 10),
```

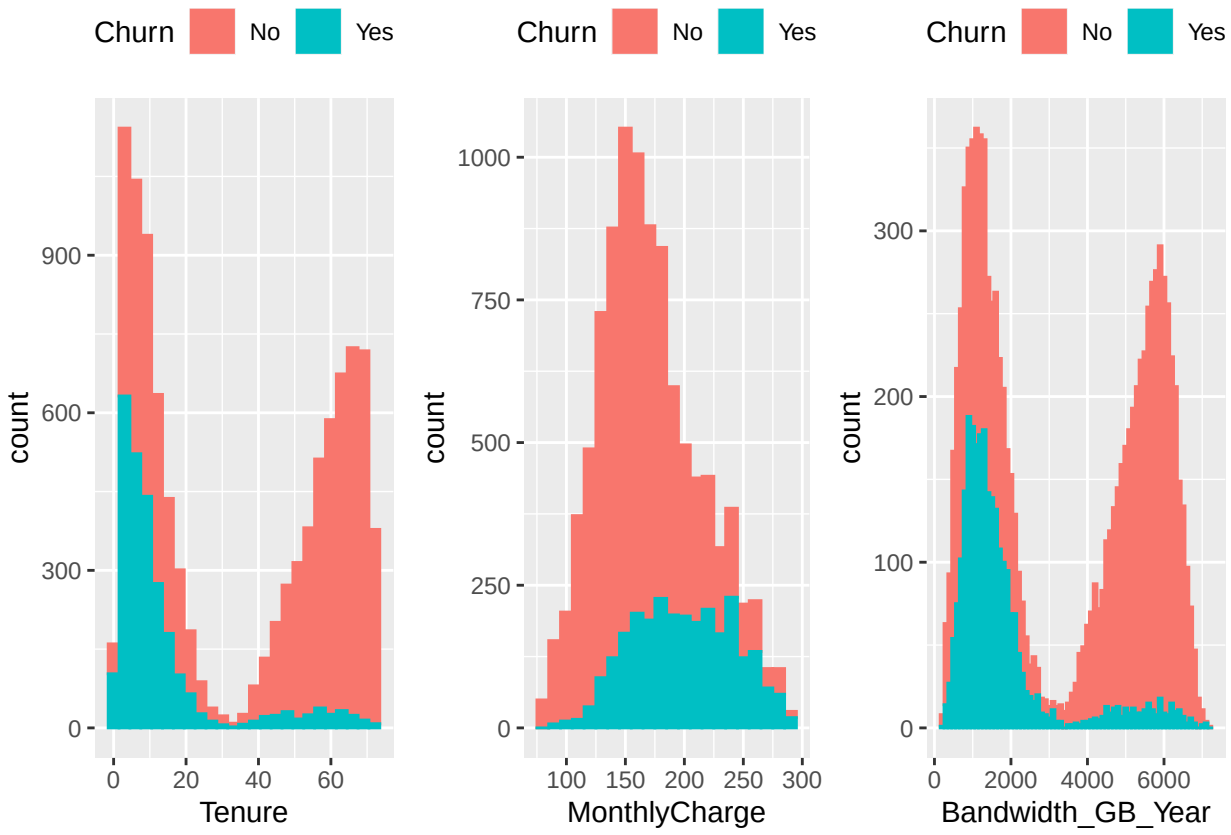
```
  ggplot(df, aes(x=Bandwidth_GB_Year)) +
    geom_histogram(binwidth = 100),

  ncol = 3, nrow = 1)
```



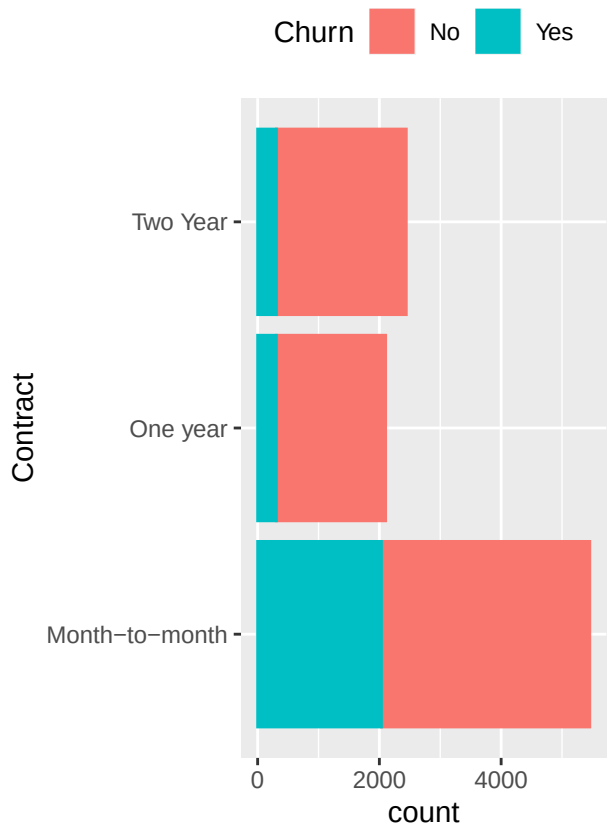

```
plot_grid(
  ggplot(df, aes(y=Contract)) +
    geom_bar(),

  ggplot(df, aes(y=Marital)) +
    geom_bar(),

  ncol = 2, nrow = 1)
```

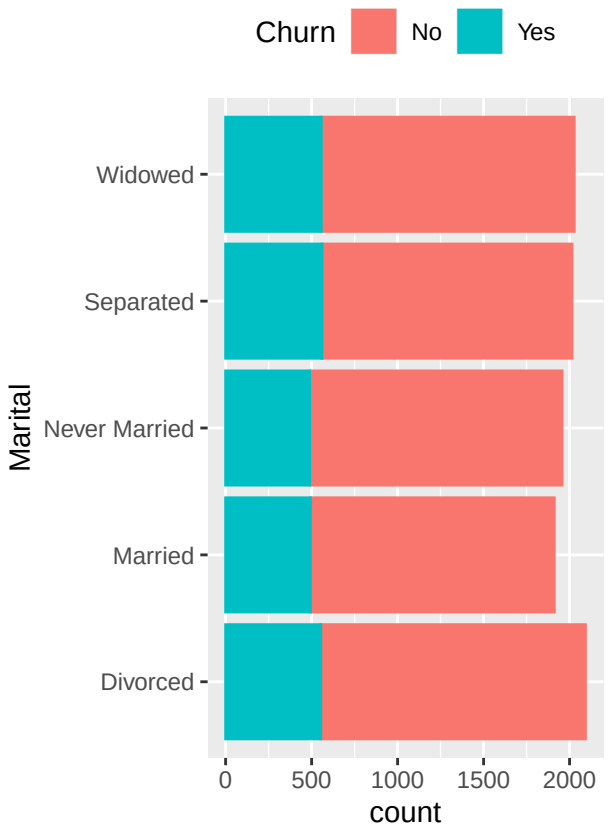



```
# Bivariate graphs of continuous & categorical variables vs Churn

plot_grid(
  ggplot(df, aes(x=Tenure, color = Churn, fill = Churn)) +
    geom_histogram(binwidth = 3) +
    theme(legend.position = 'top'),

  ggplot(df, aes(x=MonthlyCharge, color = Churn, fill = Churn)) +
    geom_histogram(binwidth = 10) +
    theme(legend.position = 'top'),

  ggplot(df, aes(x=Bandwidth_GB_Year, color = Churn, fill = Churn)) +
    geom_histogram(binwidth = 100) +
    theme(legend.position = 'top'),

  ncol = 3, nrow = 1)
```




```
plot_grid(

  ggplot(df, aes(y=Contract, color = Churn, fill = Churn)) +
    geom_bar() +
    theme(legend.position = 'top'),

  ggplot(df, aes(y=Marital,  color = Churn, fill = Churn)) +
    geom_bar() +
    theme(legend.position = 'top'),

  ncol = 2, nrow = 1)
```

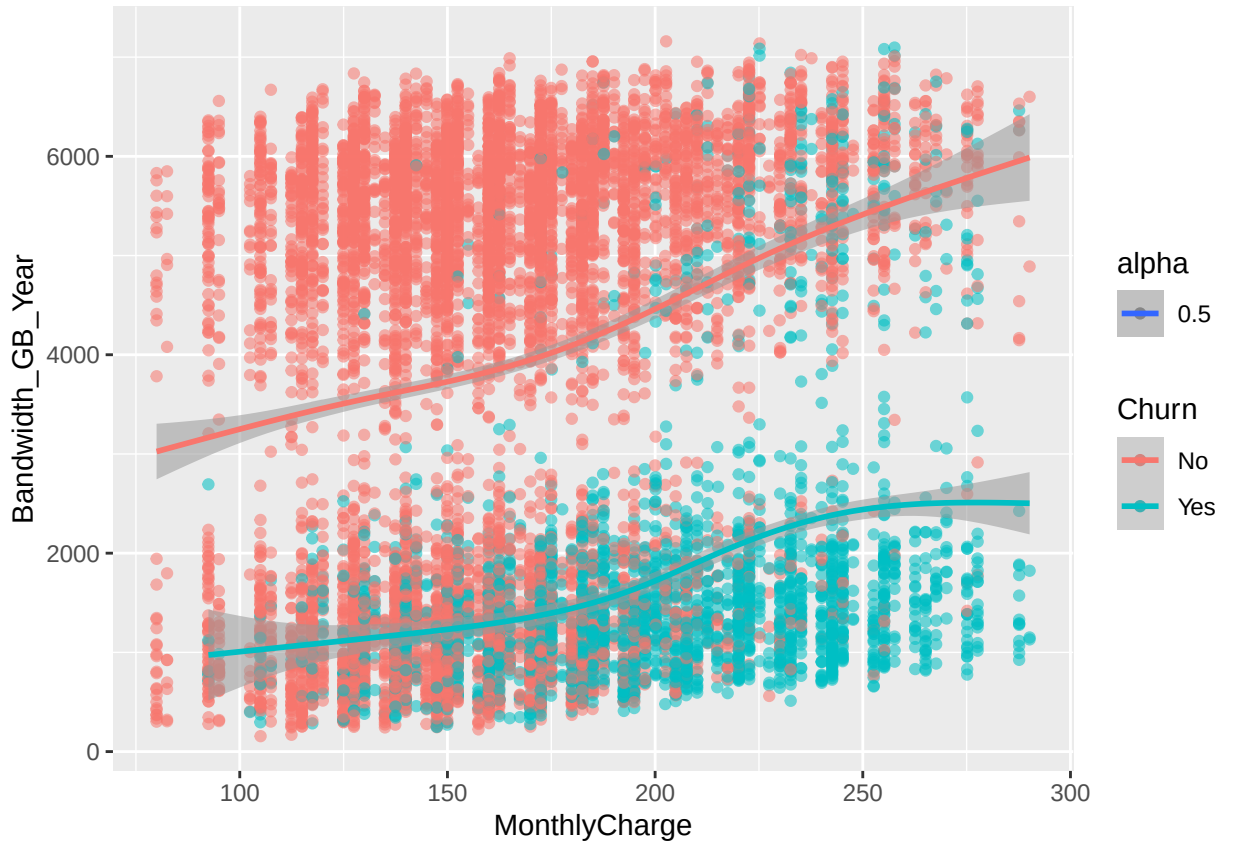





```
# Bivariate graphs of categorical variables
plot_grid(
  ggplot(df, aes(x=Contract, color = Marital, fill = Marital)) +
    geom_bar(),

  ggplot(df, aes(x=Contract, color = Area, fill = Area)) +
    geom_bar(),

  ncol = 1, nrow = 2)
```

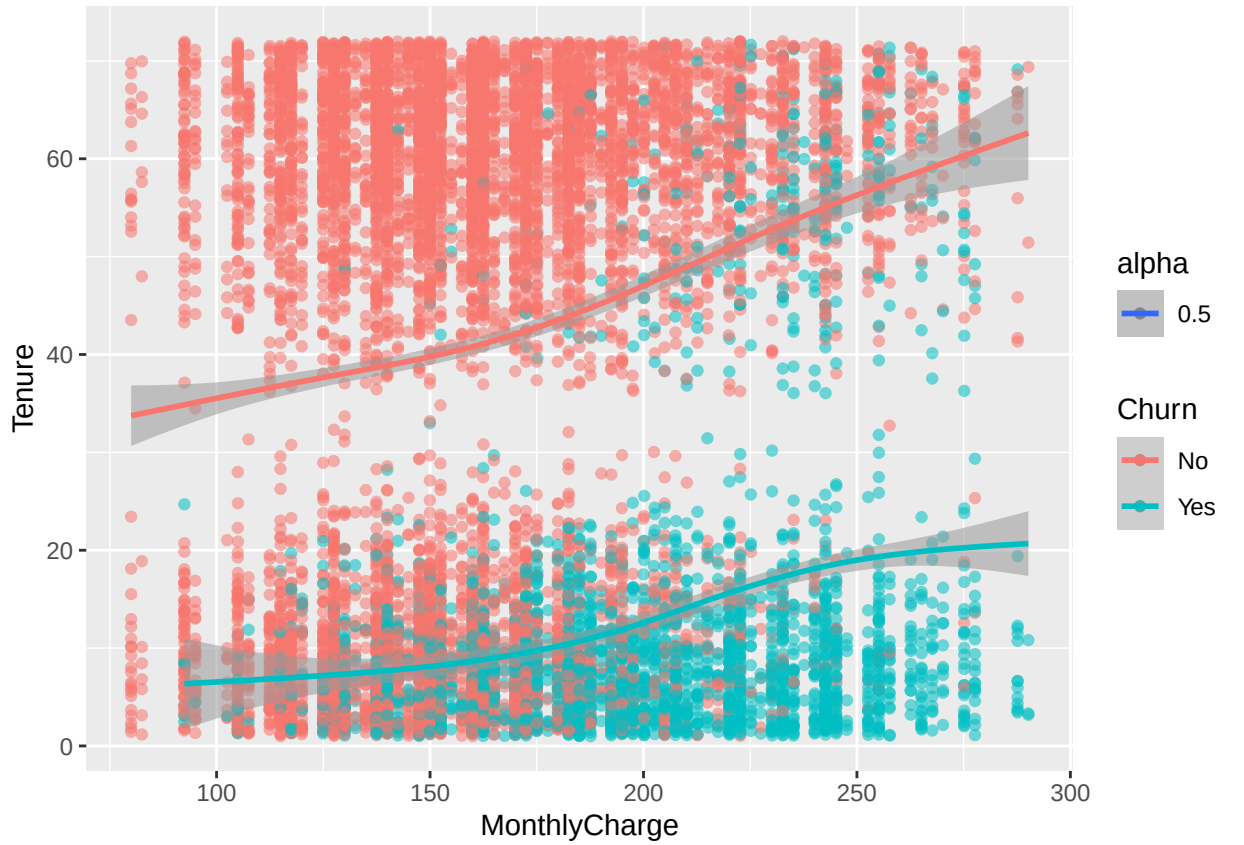

```
#Scatterplots of MonthlyCharge, Bandwidth_GB_Year, and Tenure
ggplot(df, aes(x=MonthlyCharge, y=Bandwidth_GB_Year, color = Churn, alpha = 0.5)) +
  geom_point() +
  geom_smooth(method = "auto")
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



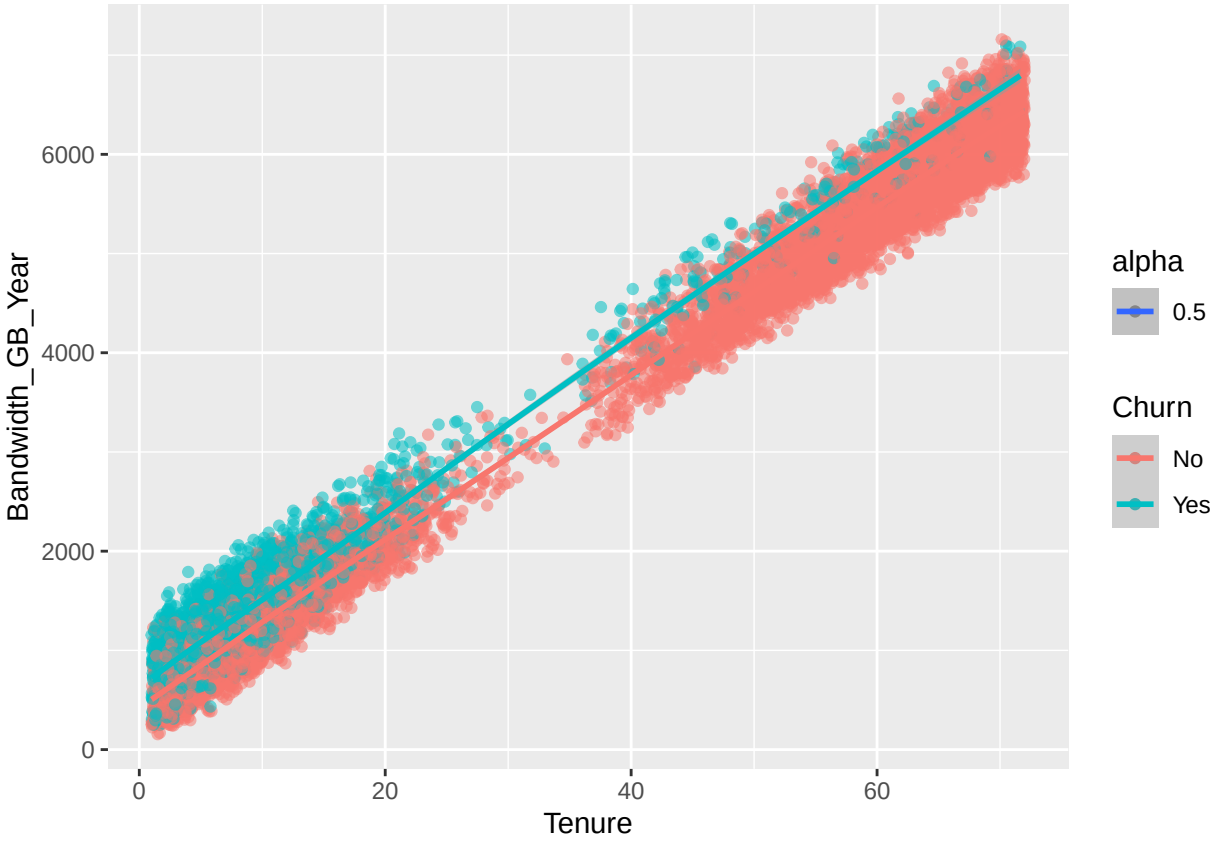

```
ggplot(df, aes(x=MonthlyCharge, y=Tenure, color = Churn, alpha = 0.5)) +
  geom_point() +
  geom_smooth(method = "auto")
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```




```
ggplot(df, aes(x=Tenure, y=Bandwidth_GB_Year, color = Churn, alpha = 0.5)) +
  geom_point() +
  geom_smooth(method = "auto")
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```


Bandwidth_GB_Year
6000
4000
2000
0
0
20
40
60
Tenure
alpha
0.5
Churn
No
Yes