# Module 4 - Assignment 2

## Angelo, Elizabeth

### Data Cleansing

library(readxl)  
CustomerChurn <- read\_excel("CustomerChurn.xlsx",  
col\_types = c("text", "text", "text",  
 "text", "numeric", "text", "text",  
 "text", "text", "text", "text", "text",  
 "text", "text", "text", "text", "text",  
 "numeric", "numeric", "text"))

## Warning in read\_fun(path = enc2native(normalizePath(path)), sheet\_i = sheet, :  
## Coercing text to numeric in R4 / R4C18: 'NaN'

## Warning in read\_fun(path = enc2native(normalizePath(path)), sheet\_i = sheet, :  
## Expecting numeric in S5 / R5C19: got '--'

## Warning in read\_fun(path = enc2native(normalizePath(path)), sheet\_i = sheet, :  
## Coercing text to numeric in R10 / R10C18: 'NaN'

## Warning in read\_fun(path = enc2native(normalizePath(path)), sheet\_i = sheet, :  
## Expecting numeric in S12 / R12C19: got '--'

## Warning in read\_fun(path = enc2native(normalizePath(path)), sheet\_i = sheet, :  
## Coercing text to numeric in R14 / R14C18: 'NaN'

## Warning in read\_fun(path = enc2native(normalizePath(path)), sheet\_i = sheet, :  
## Expecting numeric in S18 / R18C19: got '--'

## Warning in read\_fun(path = enc2native(normalizePath(path)), sheet\_i = sheet, :  
## Coercing text to numeric in R19 / R19C18: 'NaN'

#### Clean Missing Data

library(dplyr)

## Warning: package 'dplyr' was built under R version 4.1.2

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

summary(CustomerChurn)

## customerID gender Partner Dependents   
## Length:19 Length:19 Length:19 Length:19   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## tenure PhoneService MultipleLines InternetService   
## Min. : 1.00 Length:19 Length:19 Length:19   
## 1st Qu.:16.50 Class :character Class :character Class :character   
## Median :25.00 Mode :character Mode :character Mode :character   
## Mean :26.42   
## 3rd Qu.:30.50   
## Max. :80.00   
##   
## OnlineSecurity OnlineBackup DeviceProtection TechSupport   
## Length:19 Length:19 Length:19 Length:19   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## StreamingTV StreamingMovies Contract PaperlessBilling   
## Length:19 Length:19 Length:19 Length:19   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## PaymentMethod MonthlyCharges TotalCharges Churn   
## Length:19 Min. : 18.95 Min. : 29.85 Length:19   
## Class :character 1st Qu.: 36.08 1st Qu.: 320.57 Class :character   
## Mode :character Median : 56.15 Median :1919.45 Mode :character   
## Mean : 62.78 Mean :2582.56   
## 3rd Qu.: 94.38 3rd Qu.:3875.04   
## Max. :113.25 Max. :7895.15   
## NA's :4 NA's :3

CustomerChurn2 <- mutate(CustomerChurn, MonthlyCharges = replace(MonthlyCharges, is.nan(MonthlyCharges), median(MonthlyCharges, na.rm = TRUE)))   
CustomerChurn2 <- mutate(CustomerChurn2, TotalCharges = replace(TotalCharges, is.na(TotalCharges), mean(TotalCharges, na.rm = TRUE)))  
print(CustomerChurn2)

## # A tibble: 19 × 20  
## customerID gender Partner Dependents tenure PhoneService MultipleLines   
## <chr> <chr> <chr> <chr> <dbl> <chr> <chr>   
## 1 7590-VHVEG Female Yes No 20 No No phone service  
## 2 5575-GNVDE Male No No 34 Yes No   
## 3 3668-QPYBK Male No No 1 Yes No   
## 4 7795-CFOCW Male No No 30 No No phone service  
## 5 9237-HQITU Female No No 15 Yes No   
## 6 9305-CDSKC Female No No 25 Yes Yes   
## 7 1452-KIOVK Male No Yes 22 Yes Yes   
## 8 6713-OKOMC Female No No 31 No No phone service  
## 9 7892-POOKP Female Yes No 28 Yes Yes   
## 10 6388-TABGU Male No Yes 80 Yes No   
## 11 9763-GRSKD Male Yes Yes 17 Yes No   
## 12 7469-LKBCI Male No No 16 Yes No   
## 13 8091-TTVAX Male Yes No 1 Yes Yes   
## 14 0280-XJGEX Male No No 25 Yes Yes   
## 15 5129-JLPIS Male No No 25 Yes No   
## 16 3655-SNQYZ Female Yes Yes 27 Yes Yes   
## 17 8191-XWSZG Female No No 33 Yes No   
## 18 9959-WOFKT Male No Yes 71 Yes Yes   
## 19 4190-MFLUW Female Yes Yes 1 Yes No   
## # ℹ 13 more variables: InternetService <chr>, OnlineSecurity <chr>,  
## # OnlineBackup <chr>, DeviceProtection <chr>, TechSupport <chr>,  
## # StreamingTV <chr>, StreamingMovies <chr>, Contract <chr>,  
## # PaperlessBilling <chr>, PaymentMethod <chr>, MonthlyCharges <dbl>,  
## # TotalCharges <dbl>, Churn <chr>

CustomerChurn2 <- mutate(CustomerChurn2, PaymentMethod = replace(PaymentMethod, is.na(PaymentMethod), "ElectronicCheck"))  
print(CustomerChurn2)

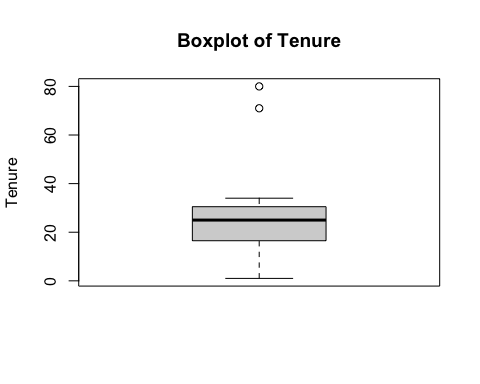
## # A tibble: 19 × 20  
## customerID gender Partner Dependents tenure PhoneService MultipleLines   
## <chr> <chr> <chr> <chr> <dbl> <chr> <chr>   
## 1 7590-VHVEG Female Yes No 20 No No phone service  
## 2 5575-GNVDE Male No No 34 Yes No   
## 3 3668-QPYBK Male No No 1 Yes No   
## 4 7795-CFOCW Male No No 30 No No phone service  
## 5 9237-HQITU Female No No 15 Yes No   
## 6 9305-CDSKC Female No No 25 Yes Yes   
## 7 1452-KIOVK Male No Yes 22 Yes Yes   
## 8 6713-OKOMC Female No No 31 No No phone service  
## 9 7892-POOKP Female Yes No 28 Yes Yes   
## 10 6388-TABGU Male No Yes 80 Yes No   
## 11 9763-GRSKD Male Yes Yes 17 Yes No   
## 12 7469-LKBCI Male No No 16 Yes No   
## 13 8091-TTVAX Male Yes No 1 Yes Yes   
## 14 0280-XJGEX Male No No 25 Yes Yes   
## 15 5129-JLPIS Male No No 25 Yes No   
## 16 3655-SNQYZ Female Yes Yes 27 Yes Yes   
## 17 8191-XWSZG Female No No 33 Yes No   
## 18 9959-WOFKT Male No Yes 71 Yes Yes   
## 19 4190-MFLUW Female Yes Yes 1 Yes No   
## # ℹ 13 more variables: InternetService <chr>, OnlineSecurity <chr>,  
## # OnlineBackup <chr>, DeviceProtection <chr>, TechSupport <chr>,  
## # StreamingTV <chr>, StreamingMovies <chr>, Contract <chr>,  
## # PaperlessBilling <chr>, PaymentMethod <chr>, MonthlyCharges <dbl>,  
## # TotalCharges <dbl>, Churn <chr>

CustomerChurn3 <- select(CustomerChurn2, PaymentMethod, MonthlyCharges, TotalCharges)  
print(CustomerChurn3)

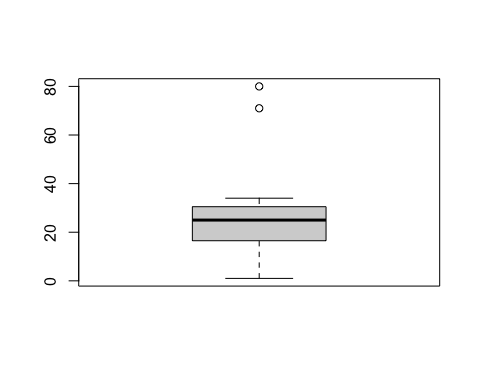
## # A tibble: 19 × 3  
## PaymentMethod MonthlyCharges TotalCharges  
## <chr> <dbl> <dbl>  
## 1 ElectronicCheck 29.8 29.8  
## 2 Mailed check 57.0 1890.   
## 3 Mailed check 56.2 108.   
## 4 Bank transfer (automatic) 42.3 2583.   
## 5 ElectronicCheck 70.7 152.   
## 6 ElectronicCheck 99.6 820.   
## 7 Credit card (automatic) 89.1 1949.   
## 8 Mailed check 29.8 302.   
## 9 Electronic check 56.2 3046.   
## 10 Bank transfer (automatic) 56.2 3488.   
## 11 Mailed check 50.0 2583.   
## 12 Credit card (automatic) 19.0 327.   
## 13 Credit card (automatic) 56.2 5681.   
## 14 Bank transfer (automatic) 104. 5036.   
## 15 ElectronicCheck 106. 2686.   
## 16 Credit card (automatic) 113. 7895.   
## 17 Mailed check 20.6 2583.   
## 18 Bank transfer (automatic) 56.2 7382.   
## 19 Credit card (automatic) 55.2 528.

The missing values for ‘MonthlyCharges’ were replaced with the median value of ‘MonthlyCharges’. The missing values for ‘TotalCharges’ were replaced with the mean value of ‘TotalCharges’. The missing values for ‘PaymentMethod’ were replaced with the string “ElectronicCheck”.

boxplot(CustomerChurn$tenure, main = "Boxplot of Tenure", ylab = "Tenure")



outliers <- boxplot(CustomerChurn2$tenure)$out



print(outliers)

## [1] 80 71

CustomerChurn2[which(CustomerChurn2$tenure %in% outliers),]

## # A tibble: 2 × 20  
## customerID gender Partner Dependents tenure PhoneService MultipleLines  
## <chr> <chr> <chr> <chr> <dbl> <chr> <chr>   
## 1 6388-TABGU Male No Yes 80 Yes No   
## 2 9959-WOFKT Male No Yes 71 Yes Yes   
## # ℹ 13 more variables: InternetService <chr>, OnlineSecurity <chr>,  
## # OnlineBackup <chr>, DeviceProtection <chr>, TechSupport <chr>,  
## # StreamingTV <chr>, StreamingMovies <chr>, Contract <chr>,  
## # PaperlessBilling <chr>, PaymentMethod <chr>, MonthlyCharges <dbl>,  
## # TotalCharges <dbl>, Churn <chr>

CustomerChurn3 <- CustomerChurn2[-which(CustomerChurn2$tenure %in% outliers),]  
boxplot(CustomerChurn3$tenure)

