# Module 4 - Assignment 2

## Kiser, Stephen

### Data Cleansing

library(tidyverse)

## -- Attaching packages ------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.0 v purrr 0.3.3  
## v tibble 2.1.3 v dplyr 0.8.5  
## v tidyr 1.0.2 v stringr 1.4.0  
## v readr 1.3.1 v forcats 0.5.0

## -- Conflicts ---------------------------------- tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

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'

#### Cleaning Missing Data

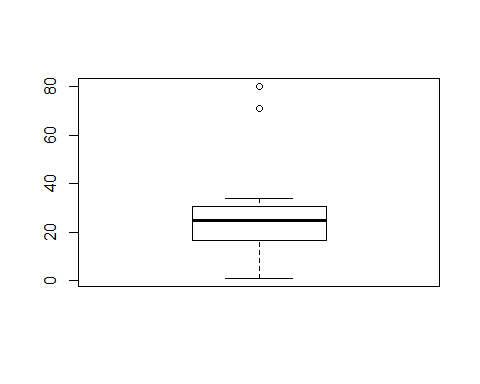
summary(CustomerChurn)

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

#### Part 2 - Outliers

boxplot(CustomerChurn2$tenure)  
boxplot(CustomerChurn2$tenure)$out



## [1] 80 71

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

## # A tibble: 2 x 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   
## # ... with 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)

