# Module 4 - Assignment 2

## Tate, Levi

### Data Cleansing

library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.0 ──

## ✓ ggplot2 3.3.2 ✓ purrr 0.3.4  
## ✓ tibble 3.0.4 ✓ dplyr 1.0.2  
## ✓ tidyr 1.1.2 ✓ stringr 1.4.0  
## ✓ readr 1.4.0 ✓ 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

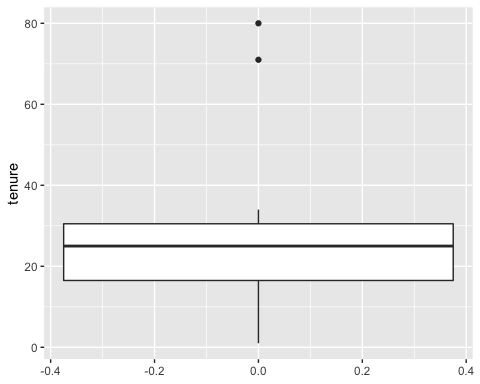
CustomerChurn2 <- mutate(CustomerChurn,MonthlyCharges = replace(MonthlyCharges, is.nan(MonthlyCharges), median(MonthlyCharges, na.rm = TRUE)))  
  
CustomerChurn2 <- mutate(CustomerChurn2 ,TotalCharges = replace(TotalCharges, is.na(TotalCharges), mean(TotalCharges, na.rm = TRUE)))  
  
CustomerChurn2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

Missing values were replaced with the *median* of each Variable in MonthlyCharges and TotalCharges. Not Applicable PaymentMethod was replaced with “ElectronicCheck.”

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

## # A tibble: 19 x 20  
## PaymentMethod MonthlyCharges TotalCharges customerID gender Partner  
## <chr> <dbl> <dbl> <chr> <chr> <chr>   
## 1 ElectronicCh… 29.8 29.8 7590-VHVEG Female Yes   
## 2 Mailed check 57.0 1890. 5575-GNVDE Male No   
## 3 Mailed check 56.2 108. 3668-QPYBK Male No   
## 4 Bank transfe… 42.3 2583. 7795-CFOCW Male No   
## 5 ElectronicCh… 70.7 152. 9237-HQITU Female No   
## 6 ElectronicCh… 99.6 820. 9305-CDSKC Female No   
## 7 Credit card … 89.1 1949. 1452-KIOVK Male No   
## 8 Mailed check 29.8 302. 6713-OKOMC Female No   
## 9 Electronic c… 56.2 3046. 7892-POOKP Female Yes   
## 10 Bank transfe… 56.2 3488. 6388-TABGU Male No   
## 11 Mailed check 50.0 2583. 9763-GRSKD Male Yes   
## 12 Credit card … 19.0 327. 7469-LKBCI Male No   
## 13 Credit card … 56.2 5681. 8091-TTVAX Male Yes   
## 14 Bank transfe… 104. 5036. 0280-XJGEX Male No   
## 15 ElectronicCh… 106. 2686. 5129-JLPIS Male No   
## 16 Credit card … 113. 7895. 3655-SNQYZ Female Yes   
## 17 Mailed check 20.6 2583. 8191-XWSZG Female No   
## 18 Bank transfe… 56.2 7382. 9959-WOFKT Male No   
## 19 Credit card … 55.2 528. 4190-MFLUW Female Yes   
## # … with 14 more variables: Dependents <chr>, tenure <dbl>, PhoneService <chr>,  
## # MultipleLines <chr>, InternetService <chr>, OnlineSecurity <chr>,  
## # OnlineBackup <chr>, DeviceProtection <chr>, TechSupport <chr>,  
## # StreamingTV <chr>, StreamingMovies <chr>, Contract <chr>,  
## # PaperlessBilling <chr>, Churn <chr>

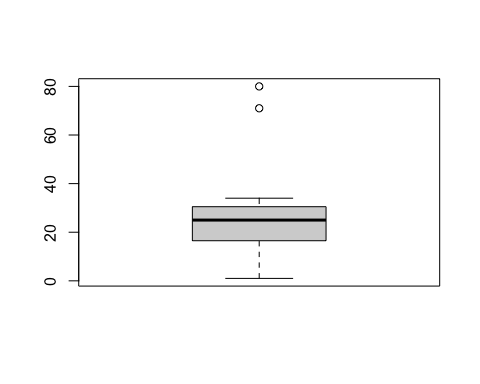
ggplot(data = CustomerChurn2, mapping = aes(y = tenure)) +  
geom\_boxplot()



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)

