# Module 4 - Assignment 2

## Pierce, Michaela

### Data Cleansing

library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.5  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ ggplot2 3.5.0 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.3 ✔ tidyr 1.3.1  
## ✔ purrr 1.0.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

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: Coercing text to numeric in R4 / R4C18: 'NaN'

## Warning: Expecting numeric in S5 / R5C19: got '--'

## Warning: Coercing text to numeric in R10 / R10C18: 'NaN'

## Warning: Expecting numeric in S12 / R12C19: got '--'

## Warning: Coercing text to numeric in R14 / R14C18: 'NaN'

## Warning: Expecting numeric in S18 / R18C19: got '--'

## Warning: 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

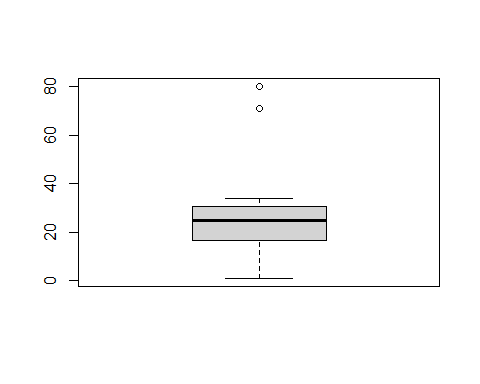
CustomerChurn2 <- mutate(CustomerChurn,MonthlyCharges = replace(MonthlyCharges,   
is.nan(MonthlyCharges), median(MonthlyCharges, na.rm = TRUE)))  
CustomerChurn2 <- mutate(CustomerChurn2, TotalCharges = replace(TotalCharges, is.na(TotalCharges), mean(TotalCharges, na.rm = TRUE)))  
CustomerChurn2 <- mutate(CustomerChurn2, PaymentMethod = replace(PaymentMethod, is.na(PaymentMethod), "ElectronicCheck"))  
  
# Create a new dataframe CustomerChurn3  
CustomerChurn3 <- CustomerChurn2 %>%  
 select(PaymentMethod, MonthlyCharges, TotalCharges)  
  
# Print the dataframe CustomerChurn3  
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.

boxplot(CustomerChurn2$tenure)$out

## [1] 80 71

outliers <- boxplot(CustomerChurn2$tenure)$out



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)

