# Module 4 - Assignment 1

# Subramanian, Subbu

# Data Cleansing

library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──  
## ✔ ggplot2 3.3.6 ✔ purrr 0.3.4  
## ✔ tibble 3.1.7 ✔ dplyr 1.0.9  
## ✔ tidyr 1.2.0 ✔ stringr 1.4.0  
## ✔ readr 2.1.2 ✔ forcats 0.5.1  
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ 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: 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)

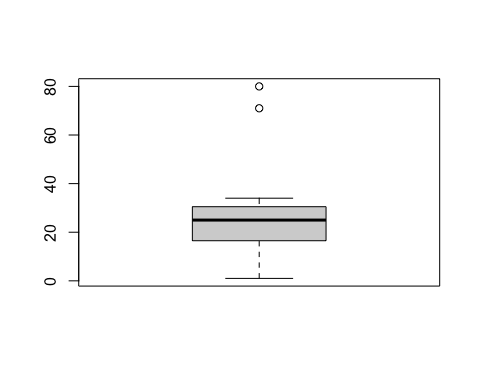
## customerID gender Partner Dependents   
## Length:19 Length:19 Length:19 Length:19   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## tenure PhoneService MultipleLines InternetService   
## 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(CustomerChurn,TotalCharges = replace(TotalCharges, is.na(TotalCharges), mean(TotalCharges, na.rm = TRUE)))  
CustomerChurn2<-mutate(CustomerChurn,PaymentMethod = replace(PaymentMethod, is.na(PaymentMethod), "ElectronicCheck"))  
  
CustomerChurn3 <- select(CustomerChurn2,MonthlyCharges,TotalCharges,PaymentMethod)  
print(CustomerChurn3)

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

## 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 × 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)

