Module 4 Assignment 2: Data Cleansing

Tabitha Hagen

2022-07-29

# Module 4 - Assignment 2

## Hagen, Tabitha

### Data Cleansing

# R Setup

knitr::opts\_chunk$set(echo = TRUE)  
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(ggplot2)  
library(readxl)  
CustomerChurn <- read\_excel("CustomerChurn.xlsx", # copy original file  
 col\_types = c("text", "text", "text",   
 "text", "numeric", "text", "text",   
 "text", "numeric", "text", "text",   
 "text", "text", "text", "text", "text",   
 "text", "numeric", "numeric", "text")) # change 18 & 19th variable from char to numeric

## Warning: Expecting numeric in I2 / R2C9: got 'No'

## Warning: Expecting numeric in I3 / R3C9: got 'Yes'

## Warning: Expecting numeric in I4 / R4C9: got 'Yes'

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

## Warning: Expecting numeric in I5 / R5C9: got 'Yes'

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

## Warning: Expecting numeric in I6 / R6C9: got 'No'

## Warning: Expecting numeric in I7 / R7C9: got 'No'

## Warning: Expecting numeric in I8 / R8C9: got 'No'

## Warning: Expecting numeric in I9 / R9C9: got 'Yes'

## Warning: Expecting numeric in I10 / R10C9: got 'No'

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

## Warning: Expecting numeric in I11 / R11C9: got 'Yes'

## Warning: Expecting numeric in I12 / R12C9: got 'Yes'

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

## Warning: Expecting numeric in I13 / R13C9: got 'No internet service'

## Warning: Expecting numeric in I14 / R14C9: got 'No'

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

## Warning: Expecting numeric in I15 / R15C9: got 'No'

## Warning: Expecting numeric in I16 / R16C9: got 'Yes'

## Warning: Expecting numeric in I17 / R17C9: got 'Yes'

## Warning: Expecting numeric in I18 / R18C9: got 'No internet service'

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

## Warning: Expecting numeric in I19 / R19C9: got 'Yes'

## Warning: Coercing text to numeric in R19 / R19C18: 'NaN'

## Warning: Expecting numeric in I20 / R20C9: got 'No'

# Part 1: Dealing with missing data

#### Cleaning Missing Data

summary (CustomerChurn) # view summary of the table describing the variables/Columns

## 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   
## Min. : NA Length:19 Length:19 Length:19   
## 1st Qu.: NA Class :character Class :character Class :character   
## Median : NA Mode :character Mode :character Mode :character   
## Mean :NaN   
## 3rd Qu.: NA   
## Max. : NA   
## NA's :19   
## 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

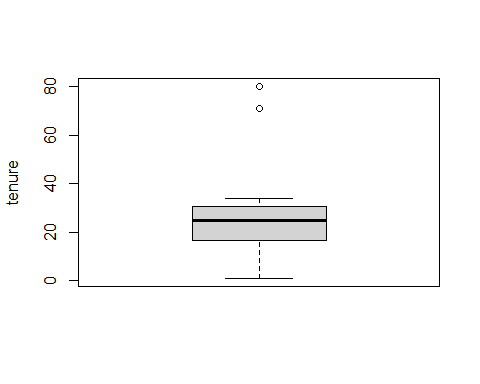
# Creating a new dataframe called CustomerChurn2 and   
# Using Mutate to replace MonthlyCharges the value of "Nan" to the Median MonthlyCharge value  
CustomerChurn2<- mutate(CustomerChurn, MonthlyCharges = replace(MonthlyCharges,   
 is.nan(MonthlyCharges), median(MonthlyCharges, na.rm = TRUE)))   
 # could have used mode or mean instead of median  
  
# Using Mutate to replace TotalCharges the value of "Na" to the Mean TotalCharge value  
CustomerChurn2<- mutate(CustomerChurn2, TotalCharges = replace(TotalCharges,   
 is.na(TotalCharges), mean(TotalCharges, na.rm = TRUE)))   
 # could have used mode or median instead of mean  
  
# Using Mutate to replace PaymentMethod the value of "NA" to "Electronic Check"  
CustomerChurn2<- mutate(CustomerChurn2, PaymentMethod = replace(PaymentMethod,   
 is.na(PaymentMethod), "Electronic Check"))  
  
# Create a new CustomerChurn3 Dataframe using MonthlyCharges, TotalCharges, & Payment Method   
CustomerChurn3 <- select(CustomerChurn2, MonthlyCharges, TotalCharges, PaymentMethod)  
print (CustomerChurn3) # print the subset dataframe CustomerChurn3

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

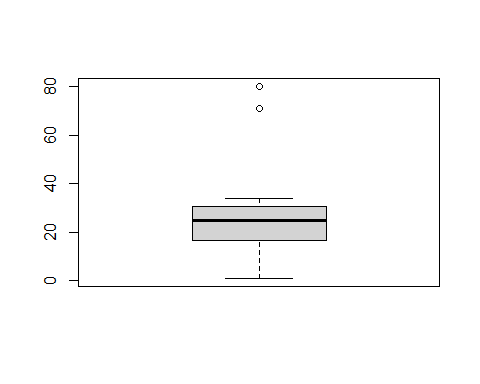
# Part 2: Outliers

#### 2A Outliers

boxplot(CustomerChurn2$tenure, ylab = "tenure") # Draw a boxplot to draw the Histogram to detect the 2 Outliers

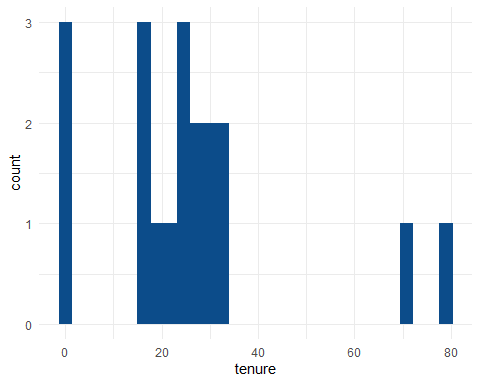


boxplot(CustomerChurn2$tenure)$out # shows 2 outliers

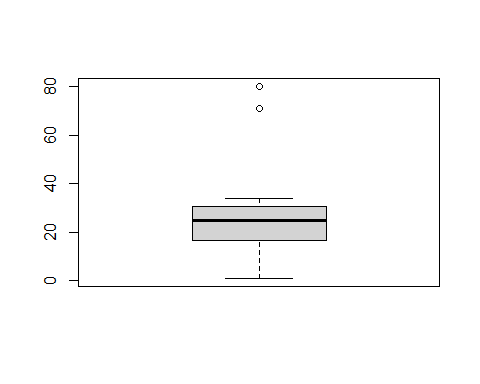


## [1] 80 71

ggplot(CustomerChurn2) + # Additionally, using ggplot2 to draw the Histogram   
 aes(x = tenure) + # to show how many times these outliers occurred  
 geom\_histogram(bins = 30L, fill = "#0c4c8a") +  
 theme\_minimal()



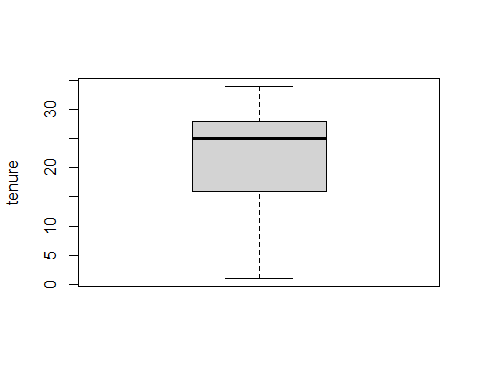
# assign the outliers to a new vector  
outliers <- boxplot(CustomerChurn2$tenure)$out



# use the outlier values to find the outliers in CustomerChurn 2  
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 <dbl>,  
## # OnlineBackup <chr>, DeviceProtection <chr>, TechSupport <chr>,  
## # StreamingTV <chr>, StreamingMovies <chr>, Contract <chr>,  
## # PaperlessBilling <chr>, PaymentMethod <chr>, MonthlyCharges <dbl>,  
## # TotalCharges <dbl>, Churn <chr>

#remove the two outliers from the file in a new dataframe  
CustomerChurn3 <- CustomerChurn2[-which(CustomerChurn2$tenure %in% outliers),]   
  
# determine if the outliers are indeed removed  
boxplot(CustomerChurn3$tenure, ylab = "tenure")



# another way of looking at the data in a histogram  
ggplot(CustomerChurn3) + # Additionally, using ggplot2 to draw the Histogram   
 aes(x = tenure) + # to show how many times these outliers occurred  
 geom\_histogram(bins = 30L, fill = "green") +  
 theme\_minimal()

