Project\_2021

## WLoad Libraries

### Load libaries for use in current working session  
## Library "here" for workflow  
library(here)  
  
## tidyverse for data manipulation and plotting  
# Loads eight different libraries simultaneously  
library(tidyverse)  
  
## broom to extract output cleanly from many statistical models including "lm"  
library(broom)   
  
## knitr to produce pleasing tables  
library(knitr)

## Warning: package 'knitr' was built under R version 3.6.3

## plotly for 3D plotting  
library(plotly)

## Warning: package 'plotly' was built under R version 3.6.3

## help to fit of a Parametric Distribution  
library(fitdistrplus)  
  
## bootstrapping  
library(boot)

## Question 1

### Part a

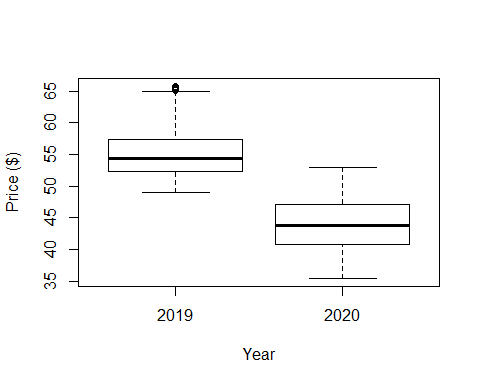
Prices <- read.table("prices.txt")

### Part b

Prices$year <- rep(2019:2020, times = c(300, 220))  
names(Prices)[1] <- "year2019"

### Part c

boxplot(year2019 ~ year, data = Prices,  
 xlab = "Year", ylab = "Price ($)")



The distribution for 2020 seems to be symmetric, whereas the distribution for 2019 is slightly skewed to the right with some outliers. In general, the prices were higher in 2019 compared to 2020. The median price in 2019 was about $54, and the median price in 2020 was about $44.

### Part d

Prices2019 <- Prices$year2019[Prices$year == 2019]

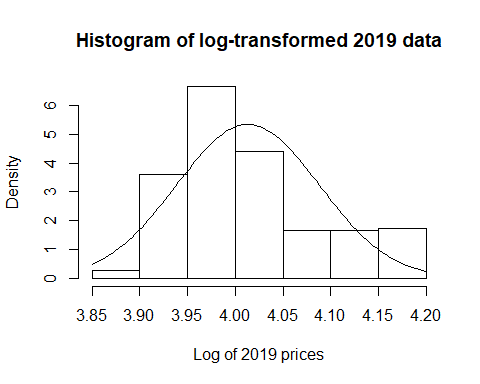
### Part e

A suitable transformation would be .

# transform 2019 data  
h <- log(Prices2019)  
# get parameter estimates  
muH <- mean(h)  
sigma <- sd(h)

### Part f

# plot desity histogram  
hist(h, probability = TRUE,  
 xlab = "Log of 2019 prices", main = "Histogram of log-transformed 2019 data")  
# superimpose the p.d.f. of the fitted normal density  
curve(dnorm(x, mean = muH, sd = sigma), add = TRUE)



The fitted normal distribution does not seem to fit the transformed data well.

### Part g

Estimated probability:

pnorm(log(70), mean = muH, sd = sigma) - pnorm(log(40), mean = muH, sd = sigma)

## [1] 0.9991807

Empirical probability:

mean(Prices2019 >= 40 & Prices2019 <= 70)

## [1] 1

## Question 2

### Part a

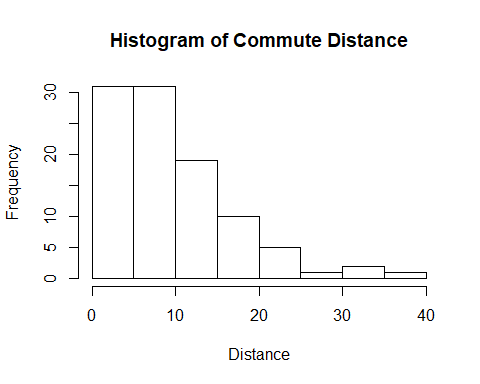
distance <- readLines("distance.txt")

## Warning in readLines("distance.txt"): incomplete final line found on  
## 'distance.txt'

# separate strings   
distance <- strsplit(distance, ",")  
# convert to numeric and assign the list into a vector  
distance <- as.numeric(unlist(distance))

### Part b

hist(distance, xlab = "Distance",  
 main = "Histogram of Commute Distance")



### Part c

distance\_fit <- fitdist(distance, distr = "gamma")  
distance\_fit$estimate

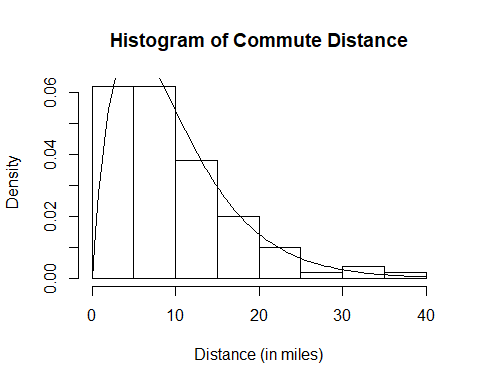
## shape rate   
## 2.0089389 0.2034194

shape <- distance\_fit$estimate[1]  
rate <- distance\_fit$estimate[2]

I fitted the Gamma distribution with shape 2.009 and rate 0.203.

### Part d

hist(distance, probability = TRUE,  
 xlab = "Distance (in miles)",  
 main = "Histogram of Commute Distance")  
# superimpose the fitted PDF on top of the histogram  
curve(dgamma(x, shape, rate), add = TRUE)



The fitted Gamma distribution seems to fit the data well.

### Part e

ks.test(distance, "pgamma", shape, rate)

## Warning in ks.test(distance, "pgamma", shape, rate): ties should not be present  
## for the Kolmogorov-Smirnov test

##   
## One-sample Kolmogorov-Smirnov test  
##   
## data: distance  
## D = 0.042473, p-value = 0.9937  
## alternative hypothesis: two-sided

Since (p-value = 0.9937) > 0.05, we can not reject the null hypothesis that distances were drawn from Gamma distribution with shape 2.009 and rate 0.203. The fitted distribution provides an adequate fit.

### Part f

Because distances are not normally distributed, We need to use bootstrapping to obtain the 95% confidence interval for the mean.

# to get mean from the data  
meanFun <- function(data, i) {  
 return(mean(data[i]))  
}  
# bootstrapping with 1000 rep  
set.seed(111)  
mean\_boot <- boot(distance, meanFun, R = 1000)  
  
# 95% confidence interval  
boot.ci(mean\_boot, type = "perc")

## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS  
## Based on 1000 bootstrap replicates  
##   
## CALL :   
## boot.ci(boot.out = mean\_boot, type = "perc")  
##   
## Intervals :   
## Level Percentile   
## 95% ( 8.431, 11.404 )   
## Calculations and Intervals on Original Scale

The 95% confidence interval for the mean distance in miles UCL staff travel each day to work is (8.458, 11.336).

## Save scripts as PDF file

#{r save\_script\_file} #knitr::stitch('SAT503\_Project\_2021\_group\_Albluwi.r')