Mini Project 5 – Report

CS 6313.001 Statistical Methods for Data ScienceStatistical Methods for Data Science

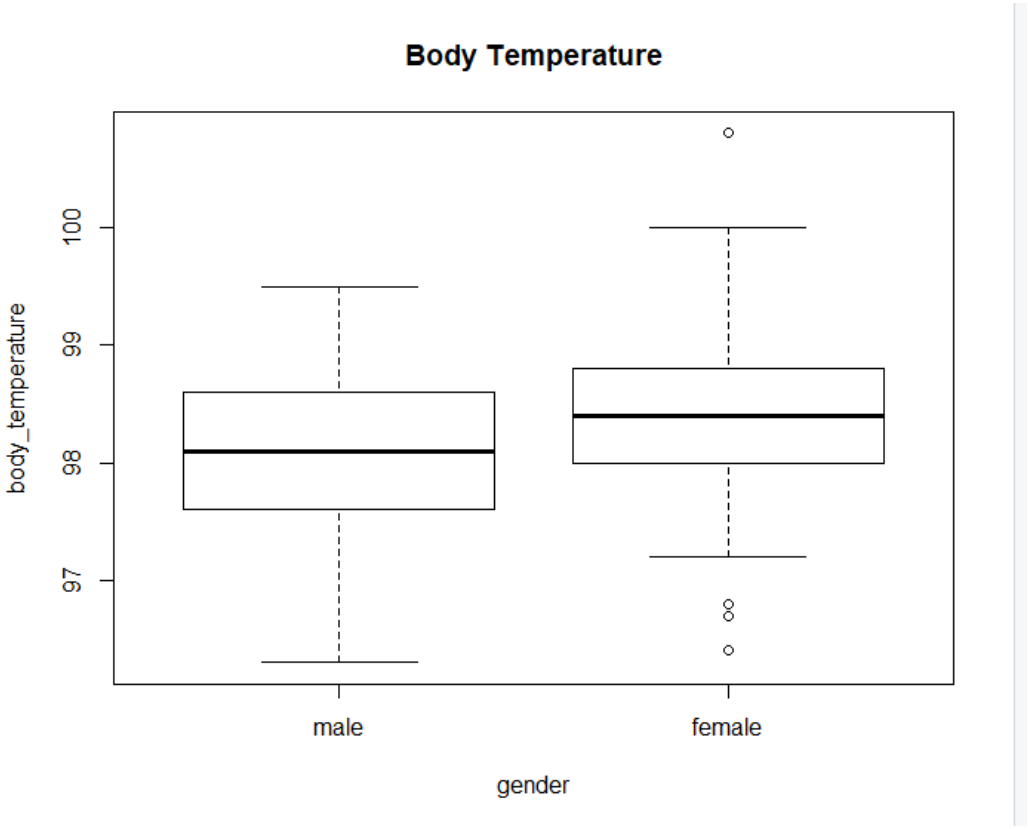
Mini Project #5

Group Members: Arpita Kumane, Harshali Dube

Contribution: Both the team members collaborated to learn R and worked on the project together. We analyzed, discussed, and efficiently worked to submit the two questions.

Question 1a.

Comparing the box plots of body temperature of male and female tells us that female boxplot have higher median, Q1 and Q3 than that of male. So, we can assume that there is a mean difference in temperature and mean of female can be higher than that of male.

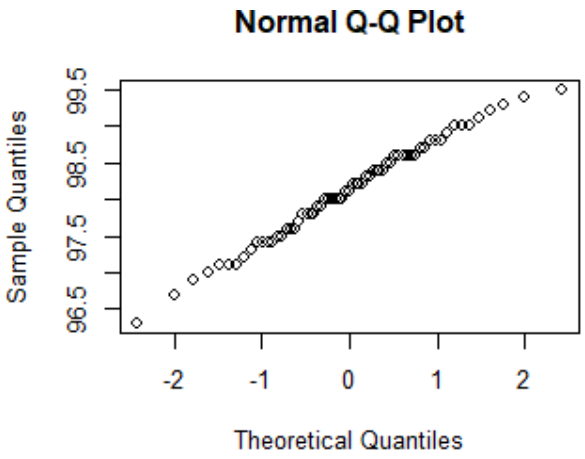


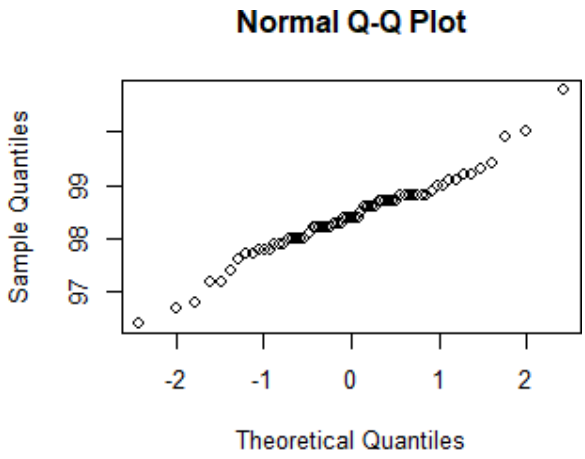
We use hypothesis test to prove this assumption.

H0: No mean difference in temperature

H1: Mean difference is not equal to zero

Sample size is 65 but we do not have population variance. So, we choose t-test for testing hypothesis. We can also verify with Q-Q plots to assume normal distribution for the sample.





After performing t-test we get the following output

```
        Welch Two Sample t-test

data:  male_temperature and female_temperature
t = -2.2854, df = 127.51, p-value = 0.02394
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.53964856 -0.03881298
sample estimates:
mean of x mean of y
 98.10462  98.39385
```

p-value is **0.02394** and we can reject the null hypothesis as the p-value is less than **0.05** and we take alpha as **0.05** here.

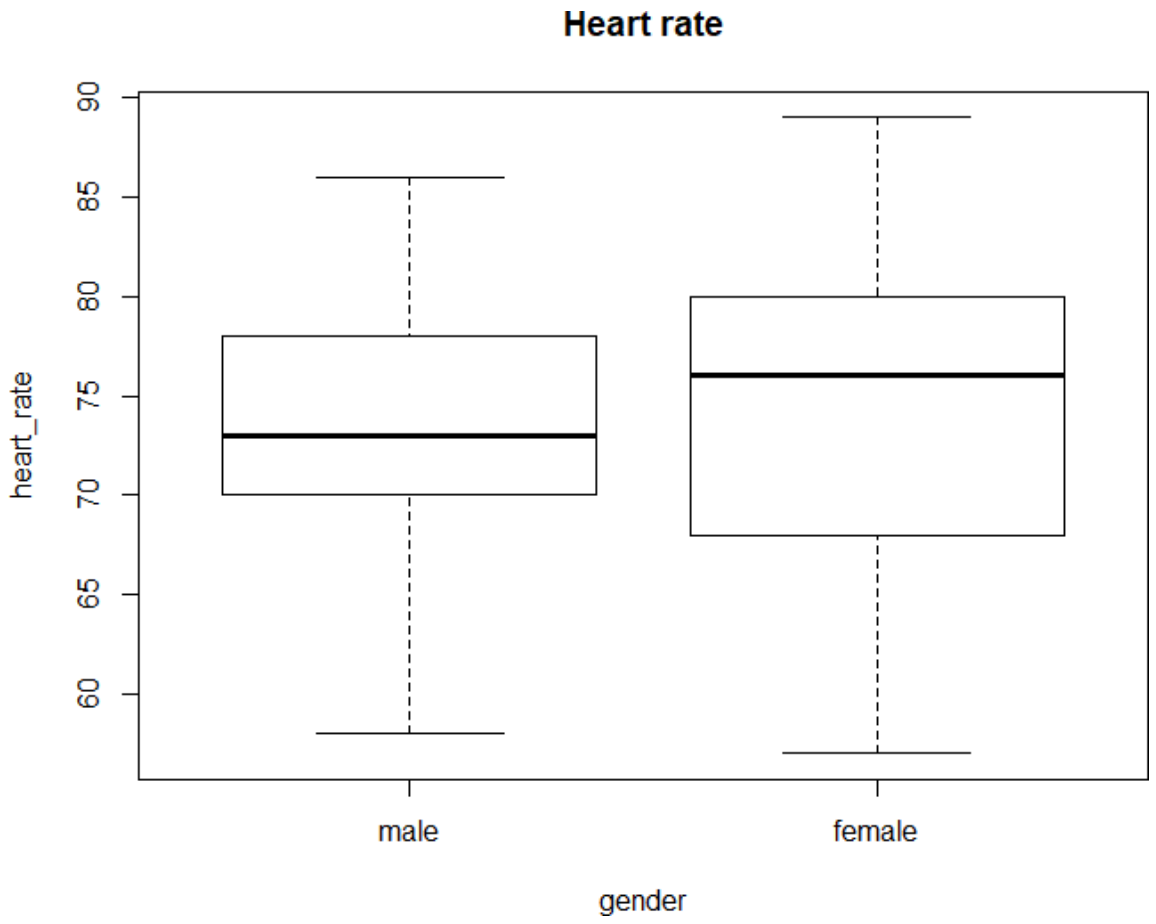
The 95% confidence interval is **(-0.53964856 -0.03881298)** and mean temperature of female is higher than that of male.

1b.

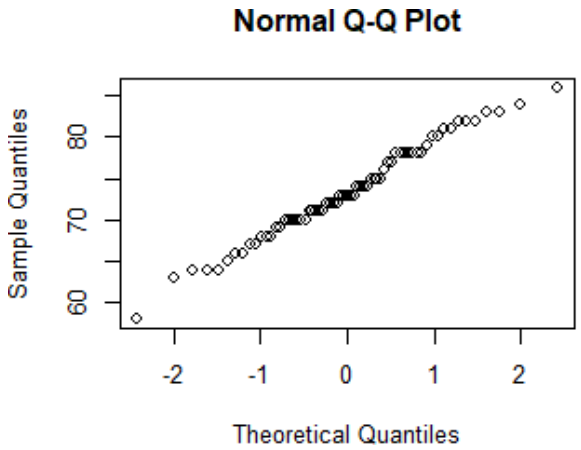
Comparing the box plots of male and female we can say about higher IQR of female than male and we cannot immediately conclude about mean. We shall do hypothesis testing to see if mean difference in heart rates is zero.

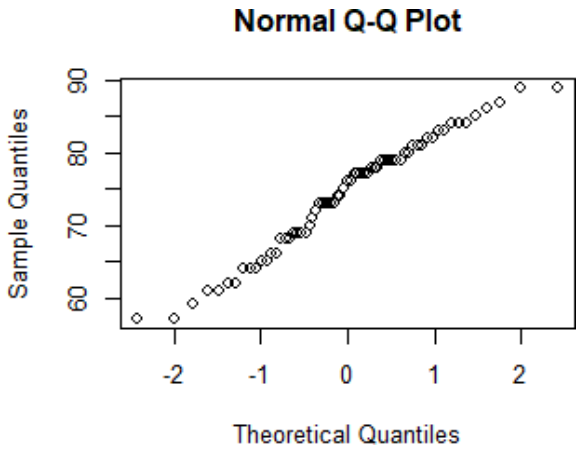
H0: No mean difference in temperature

H1: Mean difference is not equal to zero



Like 1a we shall use t-test for testing hypothesis and also observing Q-Q plots to confirm normality.





After performing t-test we get the following output

```
        Welch Two Sample t-test

data:  male_hr and female_hr
t = -0.63191, df = 116.7, p-value = 0.5287
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.243732  1.674501
sample estimates:
mean of x mean of y
 73.36923  74.15385
```

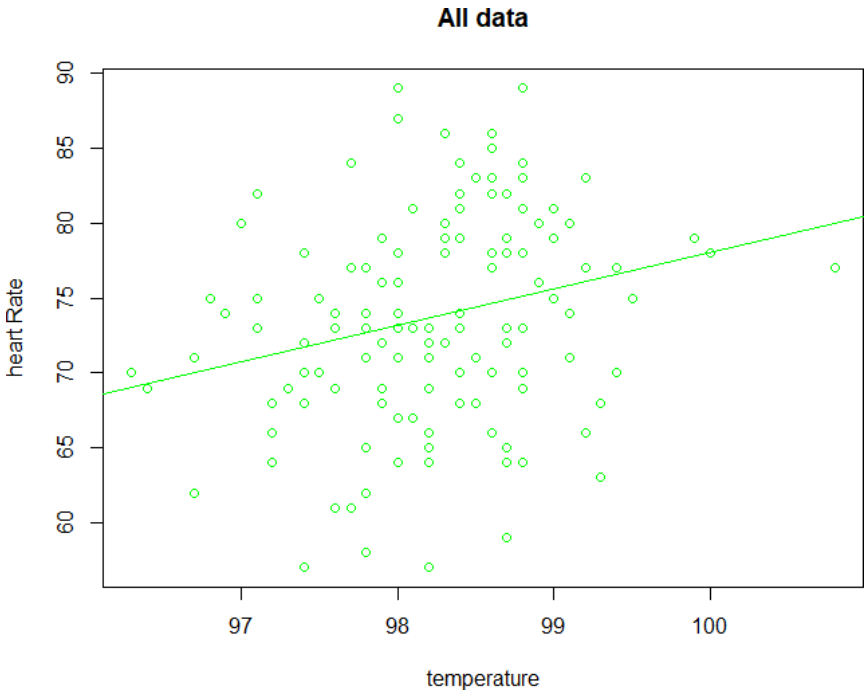
p-value is **0.5287** and we can accept the null hypothesis as the p-value is greater than **0.05** and we take alpha as **0.05** here.

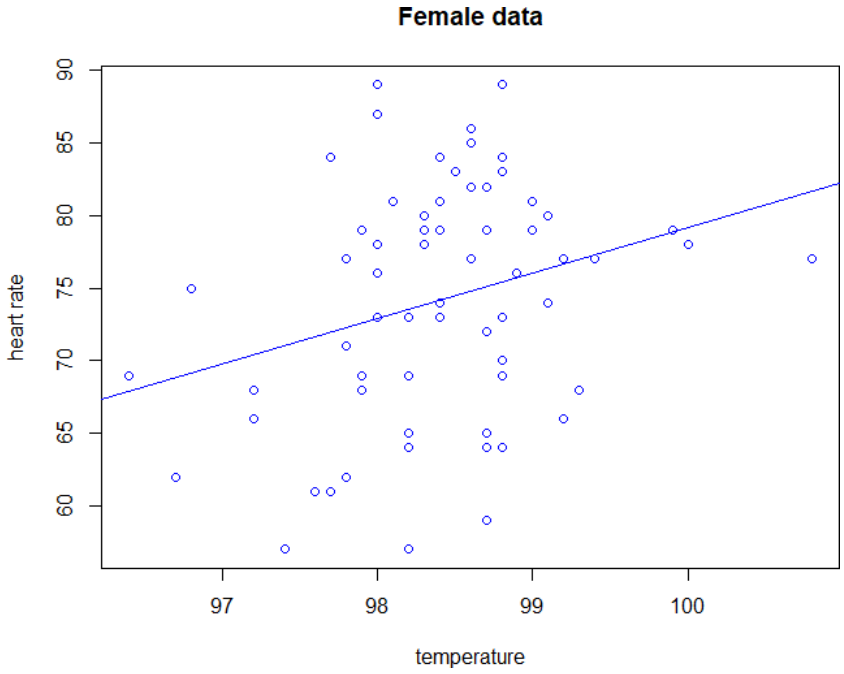
The 95% confidence interval is **(-3.243732, 1.674501)** and zero falls in this range.

Conclusion: There is no difference in means of heart rates between males and females.

1c.

Following figures shows the linear relationship between heart rate and temperature for all the data, male data, and female data.





From the figures, we can observe that there is some linear relationship in heart rate and temperature but does not seem to be strong. Female data seems to have stronger relationship than that of male data. Let's observe the correlation coefficients.

```
> cor(dataset$body_temperature, dataset$heart_rate)
[1] 0.2536564
> cor(male_data$body_temperature, male_data$heart_rate)
[1] 0.1955894
> cor(female_data$body_temperature, female_data$heart_rate)
[1] 0.2869312
```

From this, we can confirm that there is some linear relationship but not strong enough from the given data.

Question 2a and 2b.

Different values of sample size and lambda are taken and calculated confidence intervals using large sample z-interval and bootstrap percentile method. This experiment is repeated 5000 times and following are their accuracies for various combinations.

	Size	Lambda	LargeSample	BootstrapMethod
1	5	0.01	0.8080	0.8938
2	5	0.10	0.8094	0.9024
3	5	1.00	0.8136	0.8996
4	5	10.00	0.8100	0.9010
5	10	0.01	0.8664	0.9182
6	10	0.10	0.8694	0.9180
7	10	1.00	0.8692	0.9178
8	10	10.00	0.8650	0.9182
9	30	0.01	0.9204	0.9438
10	30	0.10	0.9138	0.9366
11	30	1.00	0.9152	0.9372
12	30	10.00	0.9180	0.9384
13	100	0.01	0.9406	0.9528
14	100	0.10	0.9352	0.9456
15	100	1.00	0.9372	0.9438
16	100	10.00	0.9344	0.9424

R-code:

```
#choose alpha value
alpha = 0.05
#consider different sizes and lambdas
sizes = c(5, 10, 30, 100)
lambdas = c(0.01, 0.1, 1, 10)
#use dataframe to store each combination of size and lambda
dataframe = data.frame()


for(size in sizes){
  for(lambda in lambdas){
    #initialize coverage
    coverage_z = 0
    coverage_b = 0
    rate = 1/lambda

    #repeat experiment for 5000times
    for (i in 1:5000) {
      #interval1 using large sample z interval
      large_sample = rexp(size, lambda)
      sample_mean = mean(large_sample)

      z_ci = sample_mean + c(-1,1) * qnorm(1-(alpha/2)) *(sd(large_sample)/sqrt(size))
      if(z_ci[1] < rate && z_ci[2] > rate)
        coverage_z = coverage_z + 1

      #interval2 using parametric bootstrap percentile method
      bootstrap = rexp(size*1000, 1/sample_mean)
      bootstrap = matrix(bootstrap, size, 1000)
      #sort the data and take 25th and 975th data points to get 95%ci
      bootstrap = sort(colMeans(bootstrap))
      bootstrap_ci = c(bootstrap[25], bootstrap[975])

      if(bootstrap_ci[1] < rate && bootstrap_ci[2] > rate)
        coverage_b = coverage_b + 1

    }

    #find accuracy of estimates
    coverage_mean = coverage_z/5000
    coverage_b_mean = coverage_b/5000
    dataframe = rbind(dataframe, list(size, lambda, coverage_mean, coverage_b_mean))
  }
}
names(dataframe) = list("Size", "Lambda", "LargeSample", "BootstrapMethod")
```

2c.

We can observe that as the size of the sample increases, the accuracy increased. In case of large sample interval n needed is 100 or more for the interval to be accurate and we got accuracy near to 95%. For the bootstrap interval, n needed can be 30 because there is only slight variation in accuracy when n is 30 or 100. So, 30 or more gives good accuracy for bootstrap interval. These values do not depend on lambda as there is only slight differences in accuracies when lambda is changes for same sample size.

We cannot always say that one interval is better than the other interval as depending on the sample size accuracies are varied. We would recommend using bootstrap method when the sample size is small as bootstrap have better accuracies for small n. When we have large sample size then large sample interval can be used as bootstrap method can be computationally heavy for bigger samples.

2d. Our conclusions do not depend on specific values of lambda that were fixed in advance. Here we have chosen only small subset of lambda values. For same sample size, there is only slight difference in accuracies. Moreover, for some samples increase in lambda increased the accuracy but for some samples increase in lambda decreased the accuracy. We can also observe that with increase in lambda accuracy initially increased then it decreased. With all these observations, we do not consider specific lambdas when drawing conclusions.