Stats Reasoning & Exploration HW 1

Shalini Mishra

10/1/2019

R Markdown

CASE 1: CONTAINER STORE
- **QUESTION 1**
- **a**

```
#95% confidence level for proportion
n <- 115
#z score for 95%
z95 <- qnorm(0.975,0,1,TRUE)

#proportion of 115 customers who answered YES to Q1
p_q1 <- 73/115
cat("proportion of 115 customers who answered YES to Q1 (%s)", round(p_q1,3))
```

```
## proportion of 115 customers who answered YES to Q1 (%s) 0.635
```

```
#upper confidence limit
ucl_1 <- p_q1+ sqrt((p_q1*(1-p_q1))/n)
#lower confidence limit
lcl_1 <- p_q1- sqrt((p_q1*(1-p_q1))/n)

cat("95 percent Confidence Interval for Q1: (%s, %s)", round(lcl_1,3), round(ucl_1,3))
```

```
## 95 percent Confidence Interval for Q1: (%s, %s) 0.59 0.68
```

```
#Interpretation
cat("We are 95 percent confident that the interval between %.1f %% and %.1f %% contains
    the true population proportion of all consumers who said Yes to Q1", lcl_1*100, ucl_1*100)
```

```
## We are 95 percent confident that the interval between %.1f %% and %.1f %% contains
##     the true population proportion of all consumers who said Yes to Q1 58.98833 67.96819
```

```
#proportion of 115 customers who answered YES to Q2
p_q2 <- 81/115
cat("proportion of 115 customers who answered YES to Q2 (%s)", round(p_q2,3))
```

```
## proportion of 115 customers who answered YES to Q2 (%s) 0.704
```

```
#upper confidence limit
ucl_2 <- p_q2+ sqrt((p_q2*(1-p_q2))/n)
#lower confidence limit
lcl_2 <- p_q2- sqrt((p_q2*(1-p_q2))/n)

cat("95 percent Confidence Interval for Q2: (%s, %s)", round(lcl_2,3),
        round(ucl_2,3))
```

```
## 95 percent Confidence Interval for Q2: (%s, %s) 0.662 0.747
```

```
#Interpretation
cat("We are 95 percent confident that the interval between %.1f %% and %.1f %% contains
       the true population proportion of all consumers who said Yes to Q2", lcl_2*100, ucl_2*100)
```

```
## We are 95 percent confident that the interval between %.1f %% and %.1f %% contains
##         the true population proportion of all consumers who said Yes to Q2 66.17943 74.69013
```

```
#proportion of 115 customers who answered YES to Q3
p_q3 <- 88/115
cat("proportion of 115 customers who answered YES to Q3 (%s)", round(p_q3,3))
```

```
## proportion of 115 customers who answered YES to Q3 (%s) 0.765
```

```
#upper confidence limit
ucl_3 <- p_q3+ sqrt((p_q3*(1-p_q3))/n)
#lower confidence limit
lcl_3 <- p_q3- sqrt((p_q3*(1-p_q3))/n)

cat("95 percent Confidence Interval for Q3: (%s, %s)", round(lcl_3,3),
       round(ucl_3,3))
```

```
## 95 percent Confidence Interval for Q3: (%s, %s) 0.726 0.805
```

```
#Interpretation
cat("We are 95 percent confident that the interval between %.1f %% and %.1f %% contains
       the true population proportion of all consumers who said Yes to Q3", lcl_3*100, ucl_3*100)
```

```
## We are 95 percent confident that the interval between %.1f %% and %.1f %% contains
##         the true population proportion of all consumers who said Yes to Q3 72.5692 80.47428
```

```
#proportion of 115 customers who answered YES to Q4
p_q4 <- 66/115
cat("proportion of 115 customers who answered YES to Q4 (%s)", round(p_q4,3))
```

```
## proportion of 115 customers who answered YES to Q4 (%s) 0.574
```

```
#upper confidence limit
ucl_4 <- p_q4+ sqrt((p_q4*(1-p_q4))/n)
#lower confidence limit
lcl_4 <- p_q4- sqrt((p_q4*(1-p_q4))/n)

cat("95 percent Confidence Interval for Q4: (%s, %s)", round(lcl_4,3),
       round(ucl_4,3))
```

```
## 95 percent Confidence Interval for Q4: (%s, %s) 0.528 0.62
```

```r
#Interpretation
cat("We are 95 percent confident that the interval between %.1f %% and
        %.1f %% contains the true population proportion of all consumers
        who said Yes to Q4", lcl_4*100, ucl_4*100)
```

```
## We are 95 percent confident that the interval between %.1f %% and
##         %.1f %% contains the true population proportion of all consumers
##         who said Yes to Q4 52.78001 62.0026
```

QUESTION 1 - b

```r
#95% confidence interval to estimate the population mean scores for each of the 6 questions
n <- 21
#population sd is not given, we are provided with sample SD
#we will have to use t-score instead of z-score
t95 <- qt(0.975,n-1,TRUE)

#q1
#Storing sample sd and mean for Q1
x_bar1 <- 42.4
s_q1 <-5.2

#Lower confidence limit
lcl1 <- x_bar1 - t95*(s_q1/sqrt(n))
#Upper confidence limit
ucl1 <- x_bar1 + t95*(s_q1/sqrt(n))

cat("95 percent Confidence Interval for Q1: (%s, %s)", round(lcl1,2), round(ucl1,2))
```

```
## 95 percent Confidence Interval for Q1: (%s, %s) 38.73 46.07
```

```r
#Interpretation
cat('We are 95 percent confident that the population mean score
    for Q1 is between %.1f and %.1f',lcl1,ucl1)
```

```
## We are 95 percent confident that the population mean score
##     for Q1 is between %.1f and %.1f 38.73495 46.06505
```

```r
#q2
#Storing sample sd and mean for Q1
x_bar2 <- 44.9
s_q2 <-3.1

#Lower confidence limit
lcl2 <- x_bar2 - t95*(s_q2/sqrt(n))
#Upper confidence limit
ucl2 <- x_bar2 + t95*(s_q2/sqrt(n))

cat("95 percent Confidence Interval for Q2: (%s, %s)",
        round(lcl2,2), round(ucl2,2))
```

```
## 95 percent Confidence Interval for Q2: (%s, %s) 42.72 47.08
```

```r
#Interpretation
cat('We are 95 percent confident that the population mean score for Q2
    is between %.1f and %.1f',lcl2,ucl2)
```

```
## We are 95 percent confident that the population mean score for Q2
##     is between %.1f and %.1f 42.71507 47.08493
```

```r
#q3
#Storing sample sd and mean for Q3
x_bar3 <- 38.7
s_q3 <-7.5

#Lower confidence limit
lcl3 <- x_bar3 - t95*(s_q3/sqrt(n))
#Upper confidence limit
ucl3 <- x_bar3 + t95*(s_q3/sqrt(n))

cat("95 percent Confidence Interval for Q3: (%s, %s)",
        round(lcl3,2), round(ucl3,2))
```

```
## 95 percent Confidence Interval for Q3: (%s, %s) 33.41 43.99
```

```r
#Interpretation
cat('We are 95 percent confident that the population
        mean score for Q3 is between %.1f and %.1f',lcl3,ucl3)
```

```
## We are 95 percent confident that the population
##         mean score for Q3 is between %.1f and %.1f 33.41387 43.98613
```

```r
#q4
#Storing sample sd and mean for Q4
x_bar4 <- 35.6
s_q4 <- 9.2

#Lower confidence limit
lcl4 <- x_bar4 - t95*(s_q4/sqrt(n))
#Upper confidence limit
ucl4 <- x_bar4 + t95*(s_q4/sqrt(n))

cat("95 percent Confidence Interval for Q4: (%s, %s)",
        round(lcl4,2), round(ucl4,2))
```

```
## 95 percent Confidence Interval for Q4: (%s, %s) 29.12 42.08
```

```r
#Interpretation
cat('We are 95 percent confident that the population mean score for
    Q4 is between %.1f and %.1f',lcl4,ucl4)
```

```
## We are 95 percent confident that the population mean score for
##     Q4 is between %.1f and %.1f 29.11568 42.08432
```

```r
#q5
#Storing sample sd and mean for Q5
x_bar5 <- 34.5
s_q5 <- 12.4

#Lower confidence limit
lcl5 <- x_bar5 - t95*(s_q5/sqrt(n))
#Upper confidence limit
ucl5 <- x_bar5 + t95*(s_q5/sqrt(n))

cat("95 percent Confidence Interval for Q5: (%s, %s)", round(lcl5,2), round(ucl5,2))
```

```
## 95 percent Confidence Interval for Q5: (%s, %s) 25.76 43.24
```

```r
#Interpretation
cat('We are 95 percent confident that the population mean score
        for Q5 is between %.1f and %.1f',lcl5,ucl5)
```

```
## We are 95 percent confident that the population mean score
##         for Q5 is between %.1f and %.1f 25.76026 43.23974
```

```r
#q6
#Storing sample sd and mean for Q6
x_bar6 <- 41.8
s_q6 <- 6.3

#Lower confidence limit
lcl6 <- x_bar6 - t95*(s_q6/sqrt(n))
#Upper confidence limit
ucl6 <- x_bar6 + t95*(s_q6/sqrt(n))

cat("95 percent Confidence Interval for Q6: (%s, %s)",
    round(lcl6,2), round(ucl6,2))
```

```
## 95 percent Confidence Interval for Q6: (%s, %s) 37.36 46.24
```

```r
#Interpretation
cat('We are 95 percent confident that the population mean score for Q6
    is between %.1f and %.1f',lcl6,ucl6)
```

```
## We are 95 percent confident that the population mean score for Q6
##     is between %.1f and %.1f 37.35965 46.24035
```

CASE 2: PAYMENT TIME

QUESTION 2

- **a**

```r
#As per the comsulting firm, the billing system would be effective
#if population mean of billing payment time < 19.5 days

payment_case <- read.csv('PaymentTimeCase.csv')
```

```r
#population standard deviation is given
pop_sd <- 4.2
#sample size = 65 invoices
n <- nrow(payment_case)
#sample mean of payment time
x_bar <- mean(payment_case$PayTime)

#let's calculate 95% confidence interval for the given sample
#compute z score for 95%
z95 <- qnorm(0.975,0,1,TRUE)
#lower confidence limit
lcl_95 <- x_bar - (z95*pop_sd/(sqrt(n)))
#upper confidence limit
ucl_95 <- x_bar + (z95*pop_sd/(sqrt(n)))

cat("95 percent Confidence Interval for mean payment time: (%s, %s)",
        round(lcl_95,2), round(ucl_95,2))
```

```
## 95 percent Confidence Interval for mean payment time: (%s, %s) 17.09 19.13
```

#Interpretation we are 95% confident that the population mean payment time lies between 17.09 and 19.13 i.e. we are 95% confident that the system is effective as both the limits are less than 19.5

- **b**
  #Answer Yes The upper limit of the confidence interval(worst case scenario) is less than 19.5 as claimed by the consultancy firm for effectiveness both upper limit(worst case scenario) and lower limits(best case scenario) <19.50 the value can be as low as 17.09 and as high as 19.13 but not exceeding 19.13 we are 95% confident that the population mean payment time lies between 17.09 and 19.13 i.e. we are 95% confident that the system is effective
- **c**

```r
#let's calculate 99% confidence interval for the given sample
#compute z score for 99%
z99 <- qnorm(0.995,0,1,TRUE)
#lower confidence limit
lcl_99 <- x_bar - (z99*pop_sd/(sqrt(n)))
#upper confidence limit
ucl_99 <- x_bar + (z99*pop_sd/(sqrt(n)))

cat("99 percent Confidence Interval for mean payment time: (%s, %s)",
        round(lcl_99,2), round(ucl_99,2))
```

```
## 99 percent Confidence Interval for mean payment time: (%s, %s) 16.77 19.45
```

#Interpretation Yes we are 99% confident that the population mean payment time lies between 17.09 and 19.13 i.e. we are 99% confident that the system is effective as the upper limits, 19.45(the worst case scenario) $<$ 19.50 In the very worst case as per our estimations from the given sample, mean =19.45 which is less tham 19.5

- **d**

```r
#given in the question
#population mean
pop_mean <- 19.5
#population sd
pop_sd <- 4.2
#probability of sample mean < 18.1077 days
pnorm(18.1077,pop_mean,pop_sd)
```

```
## [1] 0.3701334
```

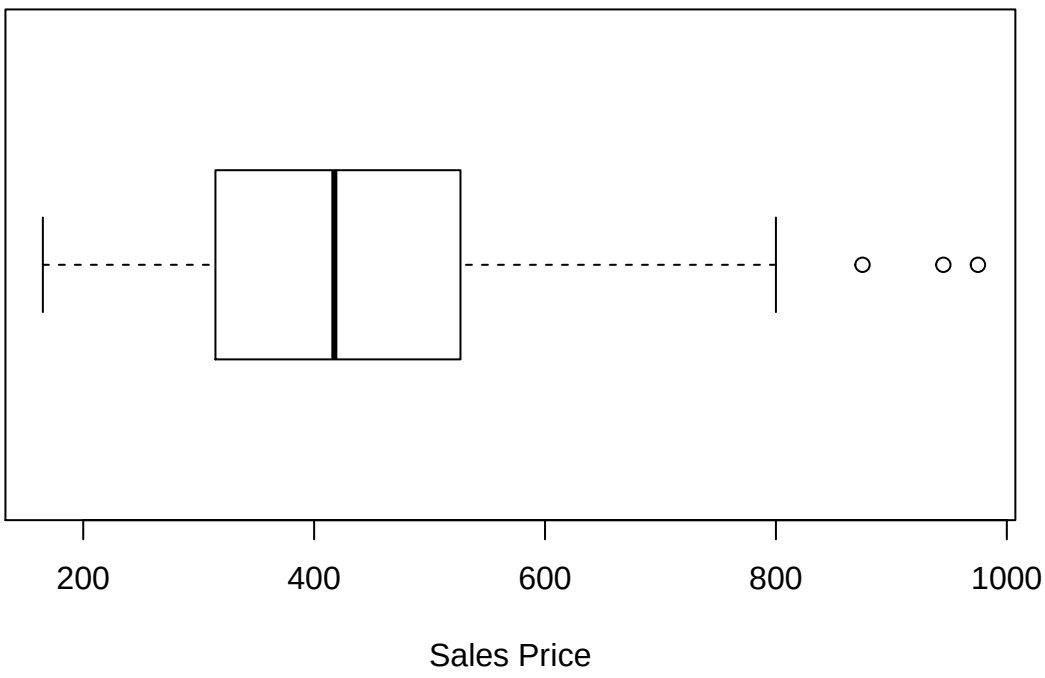
#Premium Real Estate Case
QUESTION 3
- a

```r
realestate_data <- read.csv('PremiumRealEstate.csv',stringsAsFactors = FALSE)
#beach-facing data
estate_data <- realestate_data[2:41,1:3]
#park-facing data
estate_data_2 <- realestate_data[2:19,4:6]

colnames(estate_data) <- c('Listing','Sales Price','Days to sell')
colnames(estate_data_2) <- c('Listing','Sales Price','Days to sell')
```

#Variable-Sales Price #Beach Facing Data

```r
#Boxplot
boxplot(as.double(estate_data$`Sales Price`),horizontal = TRUE,
        main='Beach facing',xlab='Sales Price')
```

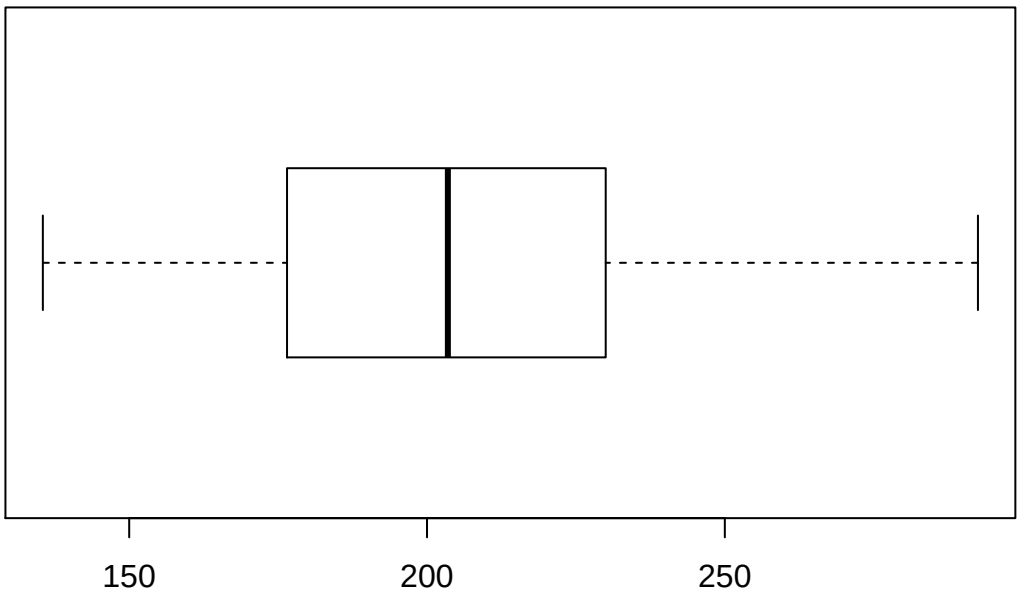



```
#Five number summary
summary(as.double(estate_data$`Sales Price`))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   165.0   314.8   417.5   454.2   522.9   975.0
```

#Park Facing Data

```
#Boxplot
boxplot(as.double(estate_data_2$`Sales Price`),horizontal = TRUE,
        main='Park facing',xlab='Sales Price')
```

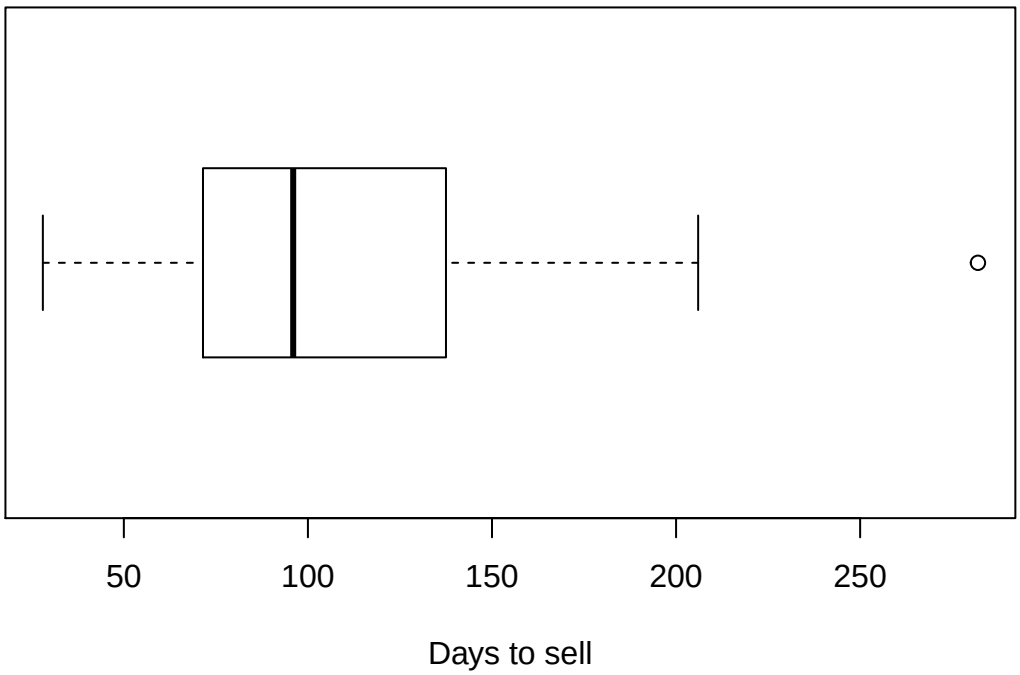



```
#Five number summary
summary(as.double(estate_data_2$`Sales Price`))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   135.5   177.1   203.5   203.2   229.2   292.5
```

#Variable-Days to sell #Beach facing

```
#Boxplot
boxplot(as.numeric(estate_data$`Days to sell`),horizontal = TRUE,
        main='Beach facing',xlab='Days to sell')
```

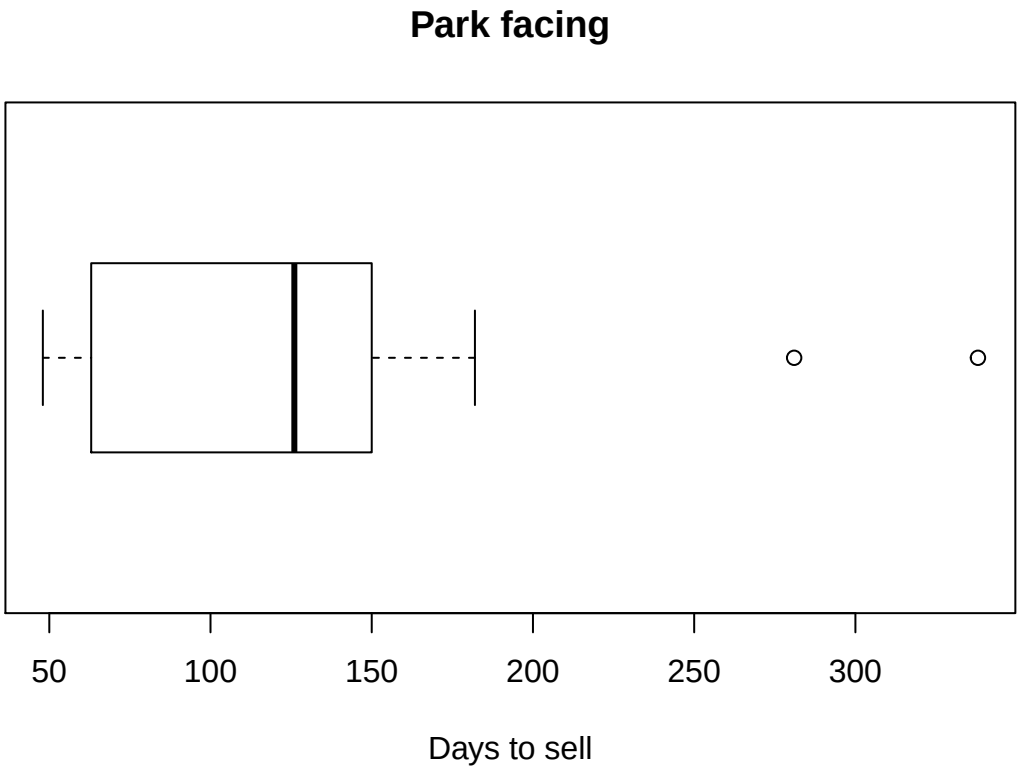



```
#Five number summary
summary(as.numeric(estate_data$`Days to sell`))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   28.00   71.75   96.00  106.00  136.25  282.00
```

#Park facing Data

```
#Boxplot
boxplot(as.numeric(estate_data_2$`Days to sell`),horizontal = TRUE,
        main='Park facing',xlab='Days to sell')
```




```
#Five number summary
summary(as.numeric(estate_data_2$`Days to sell`))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   48.00   74.75  126.00  135.00  149.75  338.00
```

#Interpretation Park-facing properties are way cheaper than beach-facing properties Mean Selling price of Beach facing properties is 454 K dollars compared to 203 K dollars for park facingproperties Park facing properties tend to sell sooner by few days than beach facing properties in general Beach-facing properties are more profitable than park facing The 3rd quantile for Park Facing and beach facing differs only by 10 days which isn't much when compared to revenue one beach property can bring in expense of 10 days

- **b**

```
#95% confidence interval for Beach facing properties
##mean sales price

#sample mean
x_bar <- mean(as.double(estate_data$`Sales Price`))
n <- nrow(estate_data)
#sample standard deviation
s_b <- sd(as.double(estate_data$`Sales Price`))
#as population standard deviation is not known, computing t-score
t95 <- qt(0.975,n-1,TRUE)
```

```r
#upper confidence interval limit for mean price
ucl_b <- x_bar + (t95*s_b/sqrt(n))
#lower confidence interval limit for mean price
lcl_b <- x_bar - (t95*s_b/sqrt(n))

cat('95 percent confidence limit for the mean sales price of beach facing properties:',
    round(lcl_b,2),round(ucl_b,2))
```

```
## 95 percent confidence limit for the mean sales price of beach facing properties: 360.05 548.39
```

```r
##mean days to sell
x_bar <- mean(as.numeric(estate_data_2$`Days to sell`))
#sample standard deviation
s_b <- sd(as.numeric(estate_data$`Days to sell`))
#as population standard deviation is not known, computing t-score
t95 <- qt(0.975,n-1,TRUE)
#upper confidence interval limit for mean days to sell
ucl_b <- x_bar + (t95*s_b/sqrt(n))
#lower confidence interval limit for mean days to sell
lcl_b <- x_bar - (t95*s_b/sqrt(n))

cat('\n95 percent confidence limit for the mean days to sell for beach facing properties:',
    round(lcl_b,0),round(ucl_b,0))
```

```
##
## 95 percent confidence limit for the mean days to sell for beach facing properties: 109 161
```

We are 95% confident that the population mean sales price of beach facing properties lie between the interval 360 thousand dollars and 548.4 thousand dollars We are 95% confident that the population average number of days to sell beach facing properties lie between 109 days and 161 days

- **c**

```r
#95% confidence interval for Park facing properties
##mean sales price

#sample mean
x_bar <- mean(as.double(estate_data_2$`Sales Price`))
n <- nrow(estate_data_2)
#sample standard deviation
s_b <- sd(as.double(estate_data_2$`Sales Price`))
#as population standard deviation is not known, computing t-score
t95 <- qt(0.975,n-1,TRUE)
#upper confidence interval limit for mean price
ucl_b <- x_bar + (t95*s_b/sqrt(n))
#lower confidence interval limit for mean price
lcl_b <- x_bar - (t95*s_b/sqrt(n))

cat('95 percent confidence limit for the mean sales price of park facing properties:',
    round(lcl_b,2),round(ucl_b,2))
```

```
## 95 percent confidence limit for the mean sales price of park facing properties: 169.24 237.14
```

```r
##mean days to sell
x_bar <- mean(as.numeric(estate_data_2$`Days to sell`))
#sample standard deviation
s_b <- sd(as.numeric(estate_data_2$`Days to sell`))
#as population standard deviation is not known, computing t-score
t95 <- qt(0.975,n-1,TRUE)
#upper confidence interval limit for mean days to sell
ucl_b <- x_bar + (t95*s_b/sqrt(n))
#lower confidence interval limit for mean days to sell
lcl_b <- x_bar - (t95*s_b/sqrt(n))

cat('\n95 percent confidence limit for the mean days to sell
    for park facing properties:',round(lcl_b,0),round(ucl_b,0))
```

```
##
## 95 percent confidence limit for the mean days to sell
##     for park facing properties: 76 194
```

We are 95% confident that the population mean sales price of park facing properties lie between the interval 169.24 thousand dollars and 237.14 thousand dollars We are 95% confident that the population average number of days to sell park facing properties lie between the interval 76 days and 194 days

- **d**

```r
#BEACH FACING
#Margin of Error for beach facing=40k $
E <- 40
n <- nrow(estate_data) #count of rows in our existing sample
s_b <- sd(as.double(estate_data$`Sales Price`))
#t-score for 95% confidence interval
t95 <- qt(0.975,n-1,TRUE)

#Estimated Sample size for the desired margin of error
n_est <- ceiling(((t95*s_b)/E)^2)
n_est
```

```
## [1] 222
```

At least a sample of 222 beach facing properties should be considered

```r
#PARK FACING
#Margin of Error for beach facing=15k $
E <- 15
n <- nrow(estate_data_2) #count of rows in our existing sample
s_b <- sd(as.double(estate_data_2$`Sales Price`))
#t-score for 95% confidence interval
t95 <- qt(0.975,n-1,TRUE)

#Estimated Sample size for the desired margin of error
n_est <- ceiling(((t95*s_b)/E)^2)
n_est
```

```
## [1] 93
```

At least a sample of 93 Park facing properties should be considered

- **e** beach-facing townhome with a list price of $589,000 and a park-facing townhome with a list price of $285,000

```
estate_data <- rbind(estate_data,c(589,NA,NA))
knn_5_imp <- VIM::kNN(estate_data, variable = c('Sales Price'), k = 5)
knn_5_imp <- VIM::kNN(knn_5_imp, variable = c('Days to sell'), k = 5)
```

Using KNN imputation for the sample provided, the beach facing property with listing price 589 K, can get sold at an estimated price of 534.5$ in approx. 71 days

```
estate_data_2 <- rbind(estate_data_2,c(285,NA,NA))
knn_5_imp <- VIM::kNN(estate_data_2, variable = c('Sales Price'), k = 5)
knn_5_imp <- VIM::kNN(knn_5_imp, variable = c('Days to sell'), k = 5)
```

Using KNN imputation for the sample provided, the park facing property with listing price 285 K, can get sold at an estimated price of 135.5$ in approx. 338 days

QUESTION 4

```
#95% confidence interval for population standard deviation

#Loading Waiting time data
bank_data <- read.csv('Question 4.csv')
n <- nrow(bank_data)
#BANK A (Single line) standard deviation
### For 95% confidence level, find the chi-sq (lower) and chi-sq (upper)
l_sq <- qchisq(0.025, df = n - 1, lower.tail = TRUE)
u_sq <- qchisq(0.975, df = n - 1, lower.tail = TRUE)

#sample standard deviation for Bank A
s <- sd(bank_data$BankA)

# Find the 95% CI Limits
l95 <- round(sqrt((n - 1)*(s^2)/u_sq),2)
r95 <- round(sqrt((n - 1)*(s^2)/l_sq),2)

cat("95 percent Confidence Interval for variability in waiting time in Bank A:
        (%s, %s)", l95, r95)
```

```
## 95 percent Confidence Interval for variability in waiting time in Bank A:
##         (%s, %s) 0.33 0.87
```

```
#BANK B (Multiple lines) standard deviation
### For 95% confidence level, find the chi-sq (lower) and chi-sq (upper)
l_sq <- qchisq(0.025, df = n - 1, lower.tail = TRUE)
u_sq <- qchisq(0.975, df = n - 1, lower.tail = TRUE)

#sample standard deviation for Bank A
```

```r
s <- sd(bank_data$BankB)

# Find the 95% CI Limits
l95 <- round(sqrt((n - 1)*(s^2)/u_sq),2)
r95 <- round(sqrt((n - 1)*(s^2)/l_sq),2)

cat("95 percent Confidence Interval for variability in waiting time in Bank B: (%s, %s)", l95, r95)
```

```
## 95 percent Confidence Interval for variability in waiting time in Bank B: (%s, %s) 1.25 3.33
```

We are 95% confident that the variability in waiting time for population in Bank A lies between the interval 0.33 and 0.87 minutes Whereas in Bank B, we are 95% confident that the variability in population waiting time lies between 1.25 and 3.33 minutes The variation in waiting period in Bank B seems higher than in bank B as the best case variability in B > worst case variability in Bank A(1.25>0.87)

```r
##Calculating 95% confidence interval for waiting time for both to get more visibility
#As population sd is unknown, we will use t score instead of z score
#Bank A (Single Line) mean
###95% confidence interval for mean waiting time
x_a <- mean(bank_data$BankA)
t95 <- qt(0.975,n-1,TRUE)
s_a <- sd(bank_data$BankA)
ucl_a <- x_a + (t95*s_a/sqrt(n))
lcl_a <- x_a - (t95*s_a/sqrt(n))

#Bank B (Multiple Lines) mean
###95% confidence interval for mean waiting time
x_b <- mean(bank_data$BankB)
t95 <- qt(0.975,n-1,TRUE)
s_b <- sd(bank_data$BankB)
ucl_b <- x_b + (t95*s_b/sqrt(n))
lcl_b <- x_b - (t95*s_b/sqrt(n))

cat("95 percent Confidence Interval for mean waiting time in Bank A: (%s, %s)", round(lcl_a,2) ,round(u
```

```
## 95 percent Confidence Interval for mean waiting time in Bank A: (%s, %s) 6.6 7.7
```

```r
cat("95 percent Confidence Interval for mean waiting time in Bank B: (%s, %s)", round(lcl_b,2), round(u
```

```
## 95 percent Confidence Interval for mean waiting time in Bank B: (%s, %s) 5.07 9.23
```

We can say a system is better than the other only when the mean and the variability in waiting time is less than the other. To get more visibility, I tried to compute 95% Confidence intervals for mean waiting time. But it is inconclusive as the UCL of Bank A>LCLof Bank B and LCL of Bank A> LCL of Bank B and with variability we can't come to a conclusion for sure But as far as variability in waiting time is considered, Bank A better than Bank B

QUESTION 5

- **a**

```r
#Case-sample proprtion not known
# we are going to take a approximation of p-hat as 0.5
p_hat <- 0.5
#accurate within 4% (+/- 4%)
margin_of_error <- 0.04
#ATQ, 95% confidence interval for proportion, computing z score
z95 <- qnorm(0.975,0,1,TRUE)

#estimate the value of sample size, n
n_est <- ceiling(p_hat*(1-p_hat)*((z95/margin_of_error)^2))
n_est
```

```
## [1] 601
```

#Answer We need atleast 601 US Adults as a sample

- **b**

```r
#Case- we know sample proprtion
#Using a prior study that found that 48% of U.S. adults
#think the president can do a lot about the price of gasoline
p_hat <- 0.48
#accurate within 4% (+/- 4%)
margin_of_error <- 0.04
#ATQ, 95% confidence interval for proportion, computing z score
z95 <- qnorm(0.975,0,1,TRUE)

#estimate the value of sample size, n
n_est <- ceiling(p_hat*(1-p_hat)*((z95/margin_of_error)^2))
n_est
```

```
## [1] 600
```

#Answer We need atleast 600 US Adults as a sample

- **c** #Answer We need one adult more when we approximate our sample proportion as 0.5 than when we take 0.48 to get desired results The estimated sample size also depends on sample proportion.