Assignment 1

Question 1

```
#Load the data
load("./interarrival.Rdata")
```

a)

```
mean(interarrival)
```

```
## [1] 3.285651
```

The mean inter-arrival time is 3.285651 min

b)

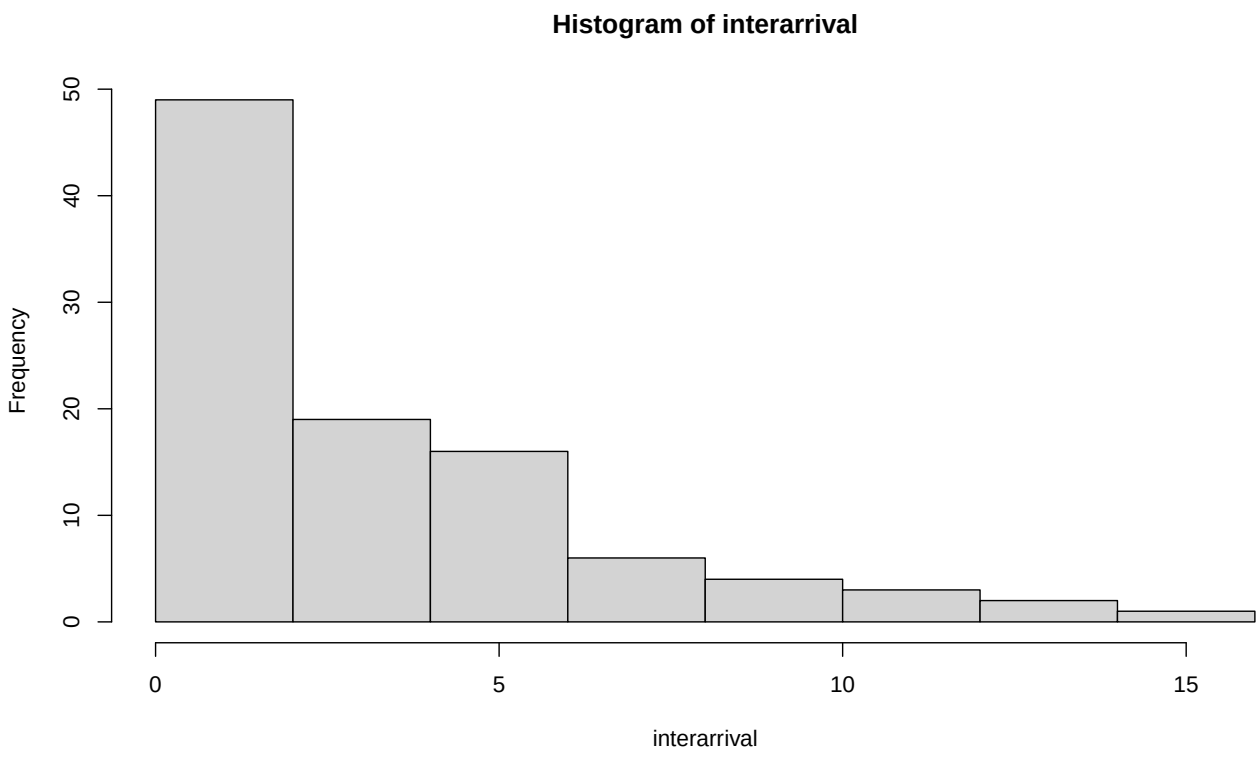
```
length(interarrival[interarrival<1])/length(interarrival)
```

```
## [1] 0.29
```

29% of the inter-arrival times are less than 1 minute.

c)

```
hist(interarrival)
```



d)

From the histogram of interarrival drawn above, it is fair to say that normal distributions is not a good fit (not symmetric, highly skewed), but the exponential distribution could be.

e)

```
fitdistr(interarrival, densfun = "exponential")
```

```
##       rate
##   0.30435371
##  (0.03043537)
```

Thus mean is $1/0.3043 = 3.2862$, variance is $1/0.3043^2 = 10.799$

The sample mean was calculated 3.2857 in (a), the variance can be calculated as:

```
var(interarrival)
```

```
## [1] 11.05406
```

The distribution estimation and the sample data show approximately the same mean and variance.

Question 2

```r
# Setting
A = matrix(c(1,3,-1,9,5,-2,-1,-1,2,2,-3,1),nrow = 4)
b = matrix(c(2,1,-1),nrow = 3)
```

a)

```r
A %*% b
```

```
##      [,1]
## [1,]    5
## [2,]    2
## [3,]    0
## [4,]   16
```

b)

```r
t(b) %*% b
```

```
##      [,1]
## [1,]    6
```

c)

```r
b %*% t(b)
```

```
##      [,1] [,2] [,3]
## [1,]    4    2   -2
## [2,]    2    1   -1
## [3,]   -2   -1    1
```

d)

```r
t(A) %*% A
```

```
##      [,1] [,2] [,3]
## [1,]   92   -9   20
## [2,]   -9   31    8
## [3,]   20    8   18
```

e)

```r
# Since the result is square
sum(diag(t(A) %*% A))
```

```
## [1] 141
```

f)

```r
solve(t(A) %*% A)
```

```
##             [,1]        [,2]        [,3]
## [1,]  0.01720655  0.01121560 -0.02410310
## [2,]  0.01121560  0.04374782 -0.03190526
## [3,] -0.02410310 -0.03190526  0.09651689
```

Question 3

Section1

```r
library(dplyr)
library(tidyr)
library(ggplot2)
library(stringr)
data = read.csv("./NFLdraft.csv")
# Remove the space on the two side of any strings
data = data %>% mutate_if(is.character,str_trim)
# LS => C
data = data %>% mutate(ifelse(Pos=="LS","C",Pos))
# Create a new column based on Pos
data["Positions"] = 0
data = data %>% mutate(Positions=ifelse(Pos %in% c("C","OG","OT","TE","DT","DE"),"Linemen",
                                        ifelse(Pos %in% c("CB","WR","FS"),"Small Backs","Big Back")))
data$Positions = as.factor(data$Positions)
# Modify the Ht column
data = data %>% separate(Ht, c("Foot","Inch"),"-")
data$Foot = as.numeric(data$Foot)
data$Inch = as.numeric(data$Inch)
data$Ht_in_Inch = data$Foot * 12 + data$Inch
# Divide the Drafted column into 4 columns by '/'
data = data %>% separate(Drafted, c("Team","Round","Pick","Year")," / ")
```

Section2

a)

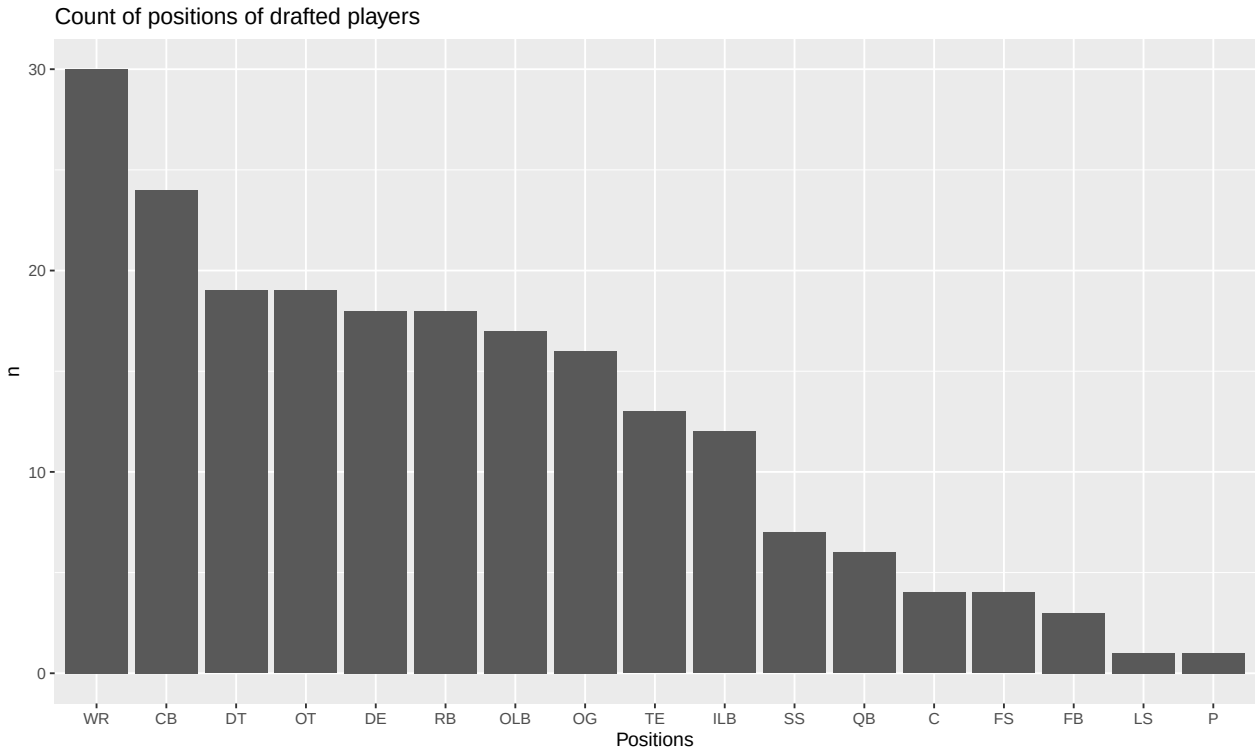
```r
data %>% count(Team) %>% dplyr::filter(n==max(n))
```

```
##               Team  n
## 1 Cleveland Browns 11
```

Thus Cleveland Browns has the most pick in 2015, with 11 picks.

b)

```r
pos_data = data %>% count(Pos) %>% arrange(desc(n))
ggplot(pos_data,aes(x=reorder(Pos,-n),y=n))+geom_bar(stat = "identity")+
  labs(title="Count of positions of drafted players")+
  xlab("Positions")
```



c)

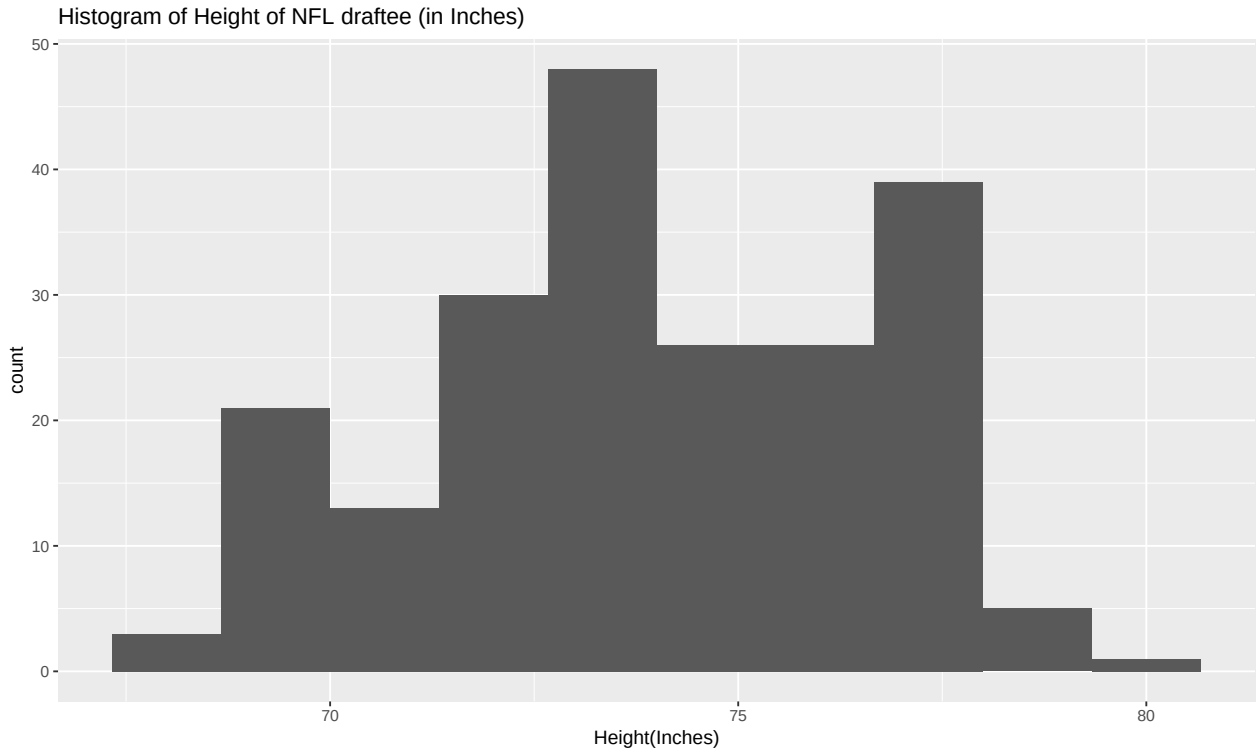
```
summary(data$Ht_in_Inch)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   68.00   72.00   74.00   74.04   76.00   80.00
```

And the average hight of NFL draftee is 74.04 feet.

d)

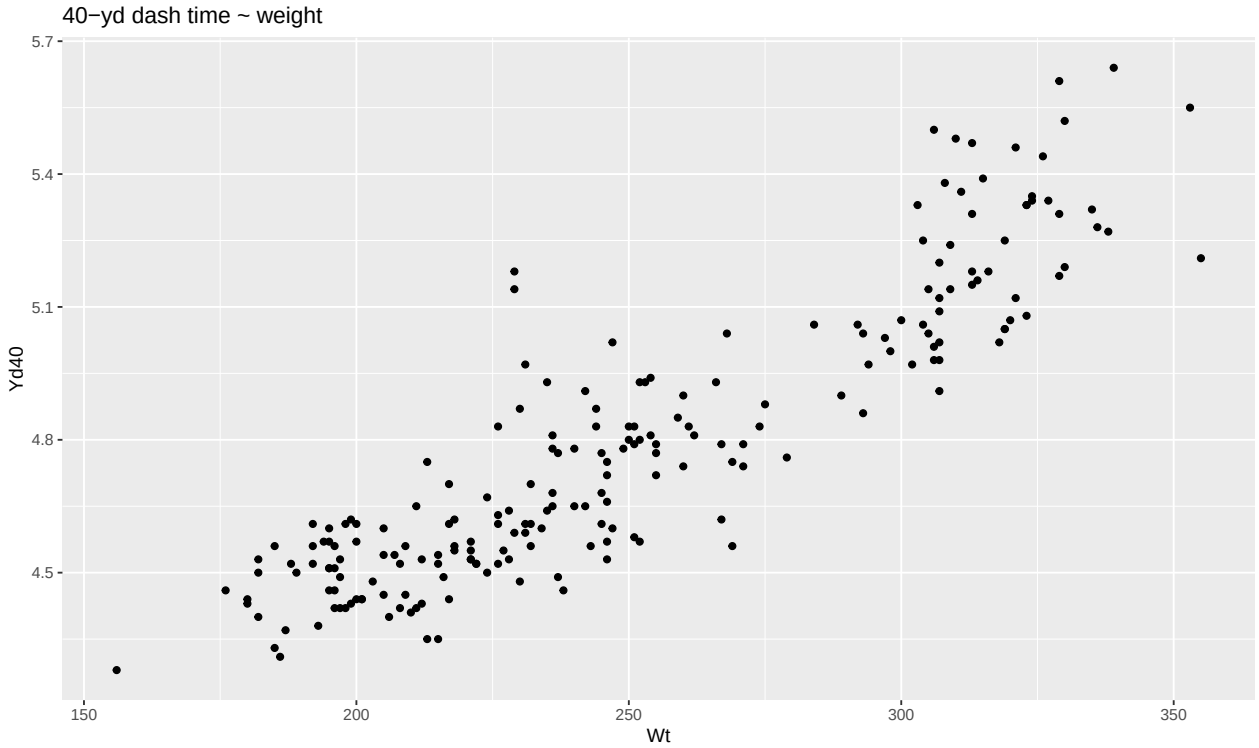
```
ggplot(data,aes(x=Ht_in_Inch))+geom_histogram(bins=10)+labs(title="Histogram of Height of NFL draftee (
```



e)

```
ggplot(data,aes(y=Yd40,x=Wt))+geom_point()+labs(title="40-yd dash time ~ weight")
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```

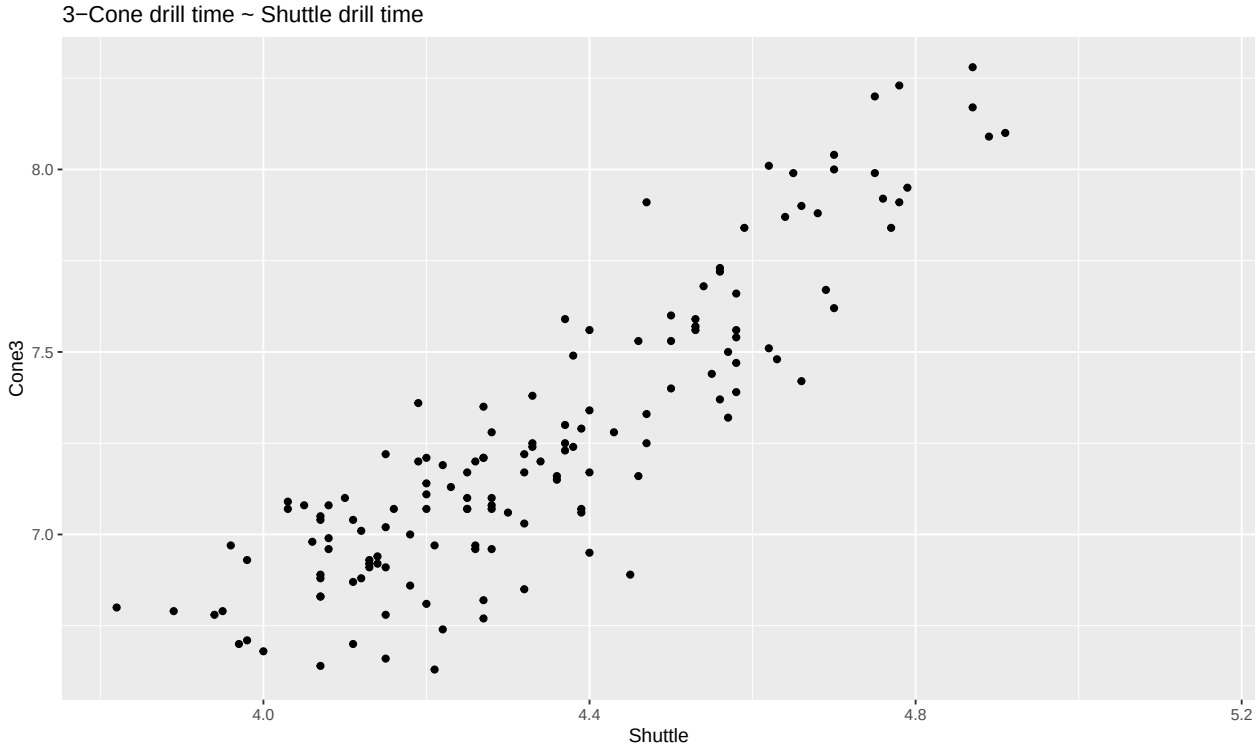


The relationship looks linear. There are outliers that have weight of over 350.

f)

```
ggplot(data,aes(y=Cone3,x=Shuttle))+geom_point()+labs(title="3-Cone drill time ~ Shuttle drill time")
```

```
## Warning: Removed 68 rows containing missing values (geom_point).
```

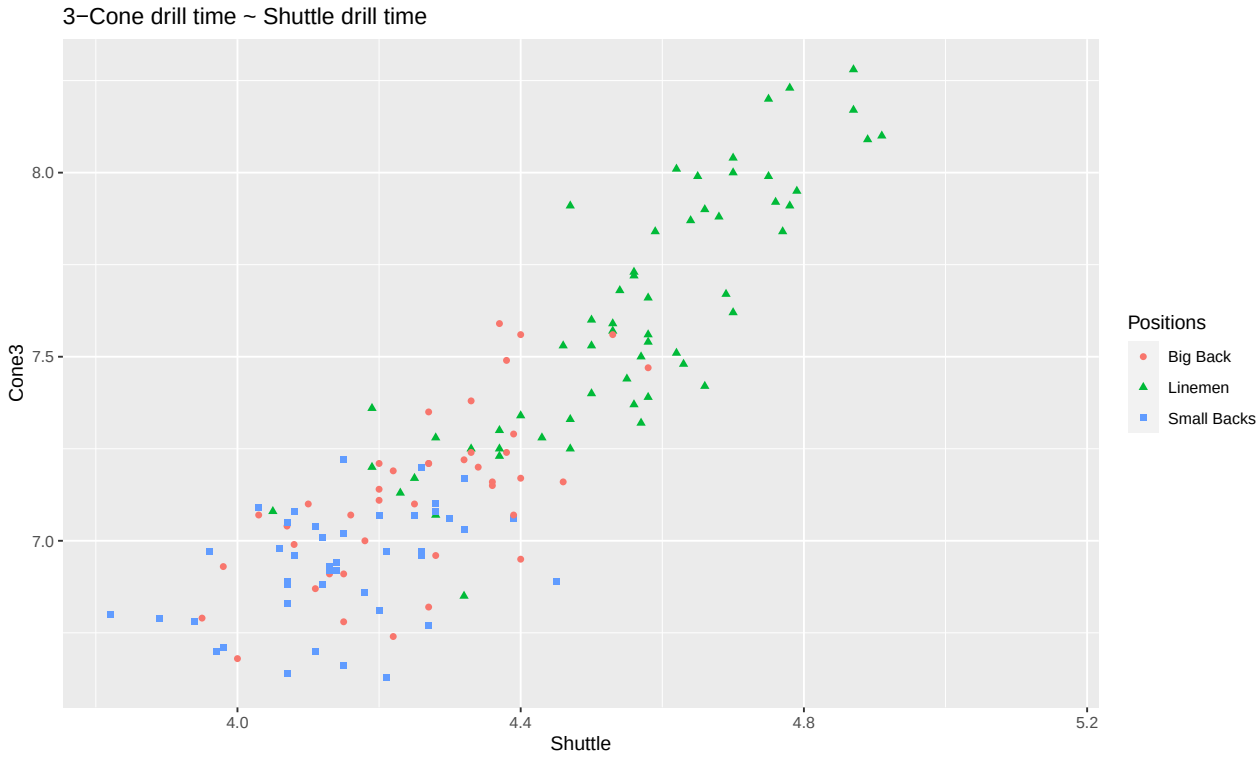


There seems to be a linear relationship.

g)

```
ggplot(data,aes(y=Cone3,x=Shuttle,color=Positions,shape=Positions))+geom_point()+labs(title="3-Cone dri
```

```
## Warning: Removed 68 rows containing missing values (geom_point).
```



Section3

a)

```
fit = lm(Cone3~Shuttle,data = data)
summary(fit)
```

```
##
## Call:
## lm(formula = Cone3 ~ Shuttle, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.52696 -0.10194 -0.00166  0.13035  0.46398
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.95098    0.27431   3.467 0.000697 ***
## Shuttle      1.45303    0.06308  23.034  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1823 on 142 degrees of freedom
##   (68 observations deleted due to missingness)
## Multiple R-squared:  0.7889, Adjusted R-squared:  0.7874
## F-statistic: 530.6 on 1 and 142 DF,  p-value: < 2.2e-16
```

The estimated regression equation is

$$3 - Cone\ drill\ time = 0.95098 + 1.45303 \times Shuttle\ drill\ time$$

b)

The intercept(0.95098) indicates the 3-cone drill time of a draftee when his shuttle drill time is 0.

The slope(1.45303) indicates that for draftees, the average increase in the time of 3-cone drill will increase by 1.45303s when one's shuttle drill time increases 1s.