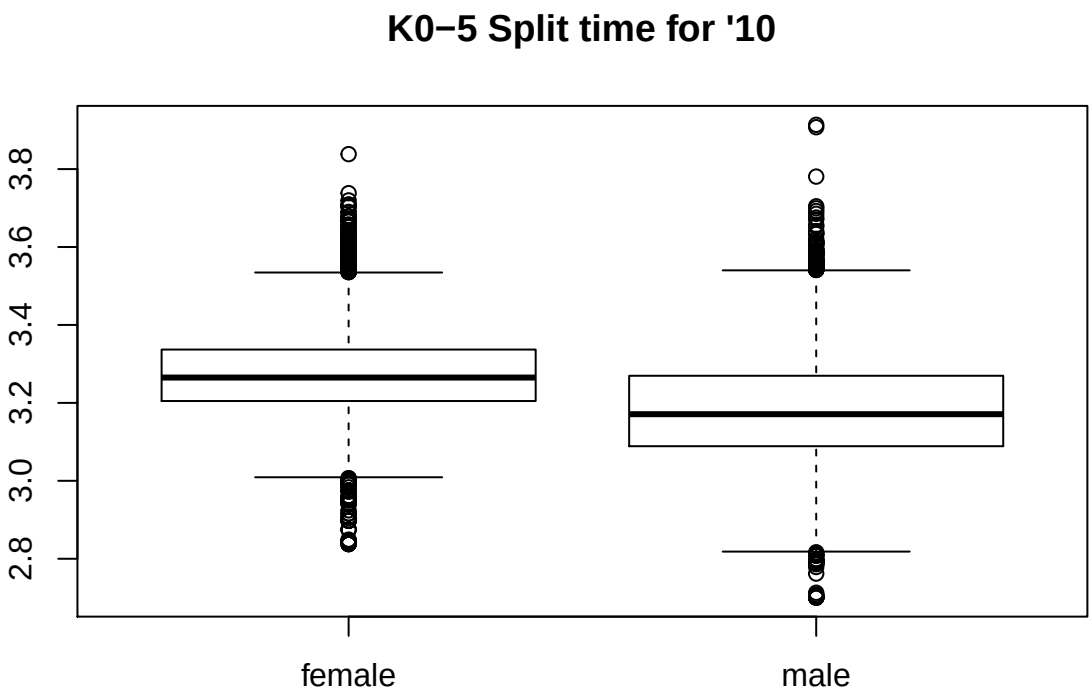
ANOVA modeling

Let's look at the ANOVA model to examine whether average time for men and women run are different for first K0-5 split time for 2010. (All the times are presented on the log scale to account for outliers and slight right skewness in the data.)

```
#1. Do men and women run differently first K0-5 - let's look at the ANOVA model
boxplot(K0.5 ~ Gender, main="K0-5 Split time for '10")
```




```
#ANOVA By Gender - first K0-5
model1 <- aov(K0.5 ~ Gender)
```

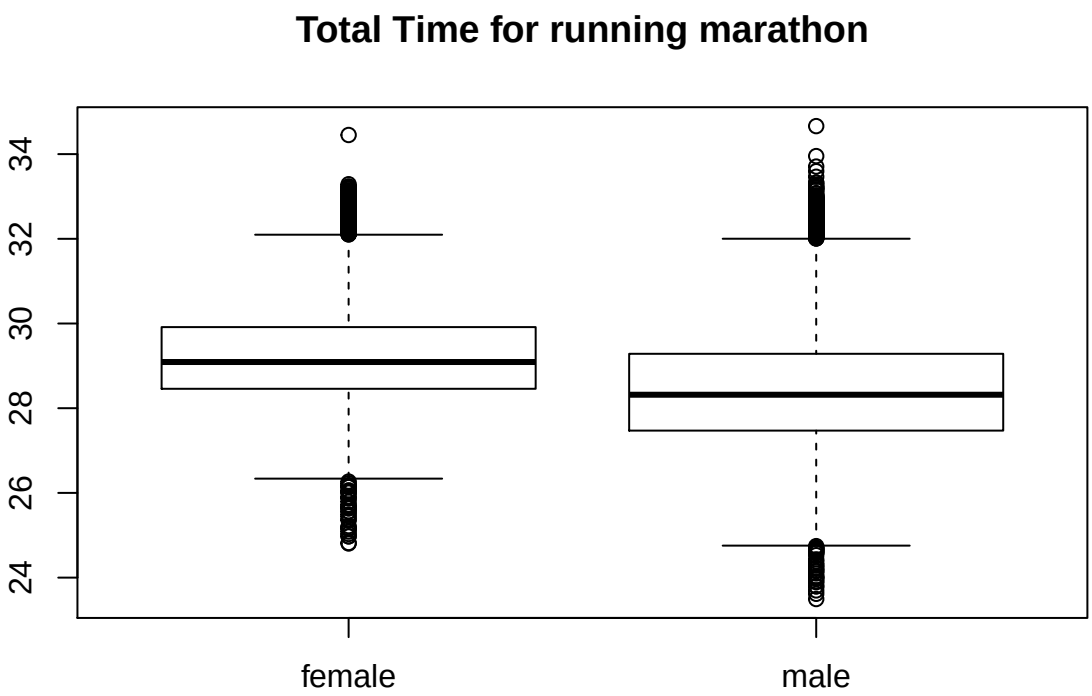
The summary of anova model is presented below which indicates significant result as p-value < 0.05 and we reject the null hypothesis and conclude that men and women have different average time for first K0-5 split time.

```
summary(model1)
```

```
##                Df Sum Sq Mean Sq F value Pr(>F)
## Gender          1   48.7   48.70    3039 <2e-16 ***
## Residuals   22668  363.2    0.02
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Similarly, we examine the total runnnig time for men and women.

```
#2. Do men and women total time are different - let's look at the ANOVA model
boxplot(TotalTime ~ Gender, main="Total Time for running marathon")
```



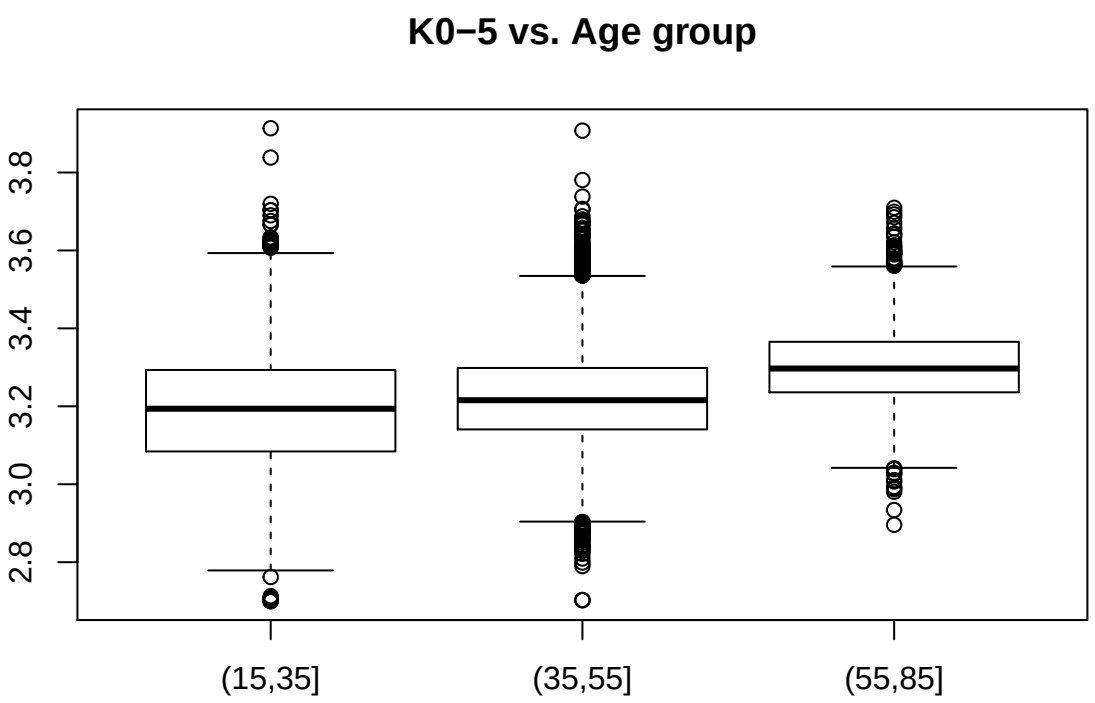
From the p-value for the ANOVA test we reject the null hypothesis and accept the alternate hypothesis that men and women have different average total time.

```
#ANOVA TotalTime By Gender
model2 <- aov(TotalTime ~ Gender)
summary(model2)
```

```
##                Df Sum Sq Mean Sq F value Pr(>F)
## Gender          1   3315    3315    1831 <2e-16 ***
## Residuals   22668  41047       2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Next, we examine whether men and women run differently by age group by comparing the total time in three different age groups, namely, (15,35], (35,55] and (55,85]. From the anova summay we conclude that average time is different among the age groups for first K0-5 split time.

```r
#3. Do men and women run differently by age group - let's look at the ANOVA model
boxplot(K0.5 ~ AgeGroup, main="K0-5 vs. Age group")
```



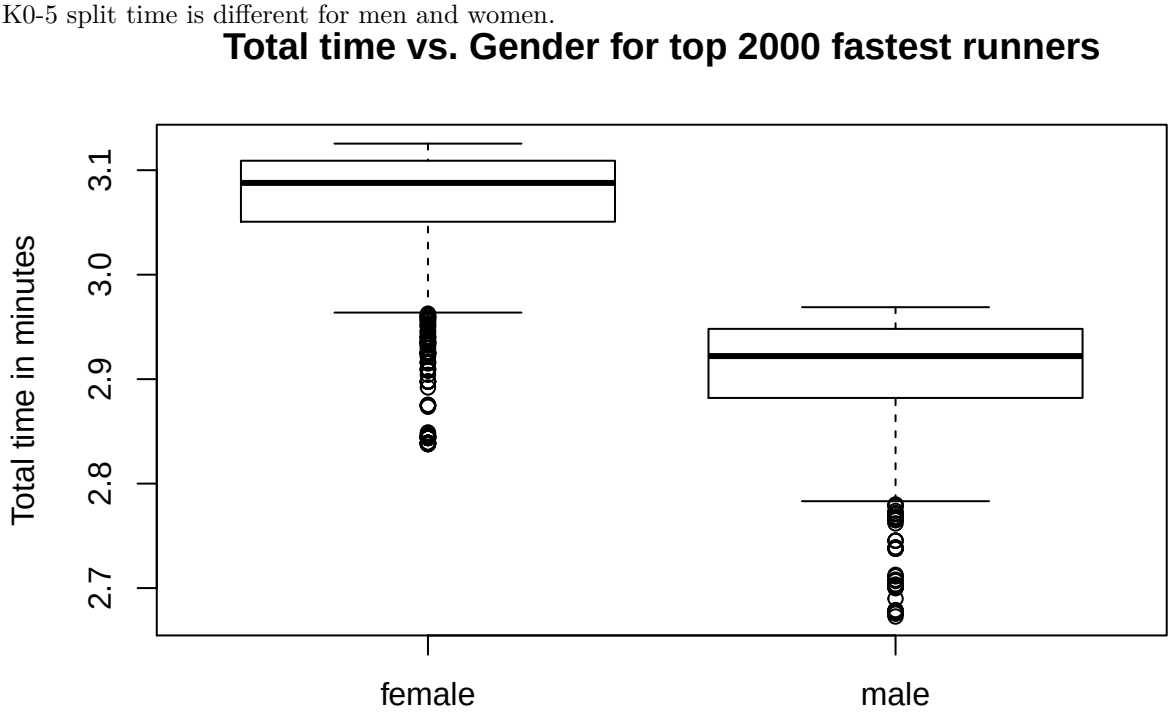

```r
#ANOVA By Age group
model3 <- aov(K0.5 ~ AgeGroup)
summary(model3)
```

```
##                Df Sum Sq Mean Sq F value Pr(>F)
## AgeGroup        2   19.4   9.685   559.2 <2e-16 ***
## Residuals   22667  392.6   0.017
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now, lets examine the interaction effect of age and gender together. We also run the TukeyHSD model and find that for first K0-5 split time, female in age group (35,55] and male in age group (55,85] do not have signigficant result. However this is not the same when we run the Tukey test for total time as seen below. Highlighted the main TukeyHSD output below,
TukeyHSD(aov(K0.5 ~ AgeGroup * Gender))
Fit: aov(formula = K0.5 ~ AgeGroup * Gender)
$AgeGroup:Gender
p adj
(55,85]:male-(35,55]:female 0.761009

```
##                    Df Sum Sq Mean Sq F value Pr(>F)
## AgeGroup            2   19.4    9.69  670.31 <2e-16 ***
## Gender              1   63.9   63.88 4421.01 <2e-16 ***
## AgeGroup:Gender     2    1.2    0.61   42.52 <2e-16 ***
## Residuals       22664  327.5    0.01
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We also compared fastest 2000 runners for male and female group and found that the average time for first K0-5 split time is different for men and women.




```
model6 <- aov(K0.5 ~ Gender)
summary(model6)
```

```
##                Df Sum Sq Mean Sq F value Pr(>F)
## Gender          1   48.7   48.70    3039 <2e-16 ***
## Residuals   22668  363.2    0.02
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The various ANOVA models for TotalTime were analyzed and summarized below with the decreasing R^2 values.

```
results
```

```
##      model                                    R^2
## [1,] "aov(TotalTime~SecondHalf)"              "0.959405885490947"
## [2,] "aov(TotalTime~FirstHalf+Age+Gender)" "0.923333773353598"
## [3,] "aov(TotalTime~FirstHalf)"               "0.920393858126625"
## [4,] "aov(TotalTime~K0.5)"                    "0.841914367377936"
```

We still have to analyze ANOVA in for interaction effect of both Age and Gender on the Split time...