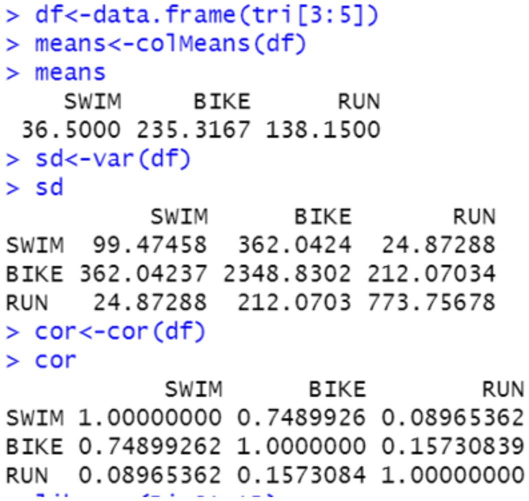
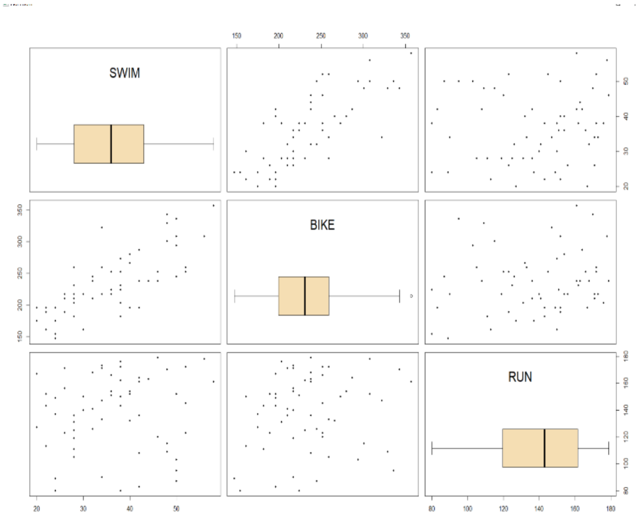
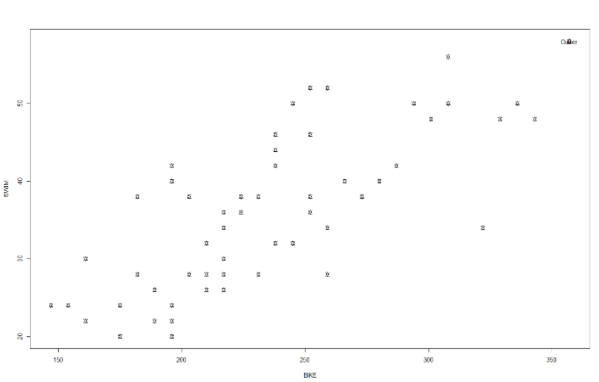
QA

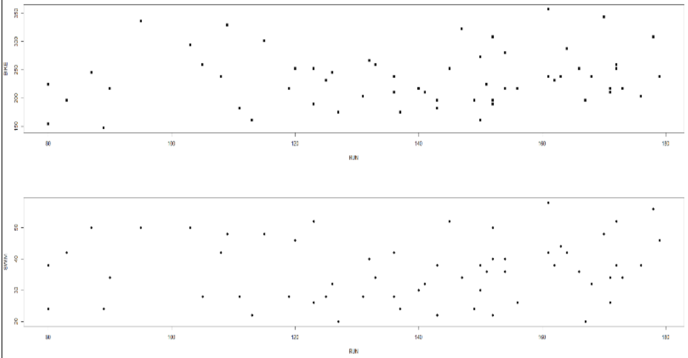


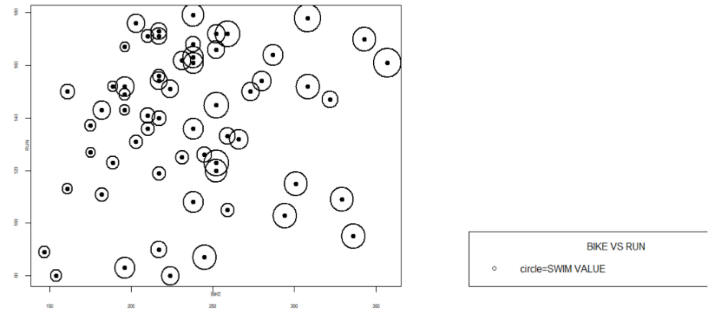
```
> df<-data.frame(tri[3:5])
> means<-colMeans(df)
> means
    SWIM     BIKE      RUN
 36.5000 235.3167 138.1500
> sd<-var(df)
> sd
          SWIM      BIKE       RUN
SWIM  99.47458  362.0424  24.87288
BIKE 362.04237 2348.8302 212.07034
RUN   24.87288  212.0703 773.75678
> cor<-cor(df)
> cor
           SWIM      BIKE        RUN
SWIM 1.00000000 0.7489926 0.08965362
BIKE 0.74899262 1.0000000 0.15730839
RUN  0.08965362 0.1573084 1.00000000
```



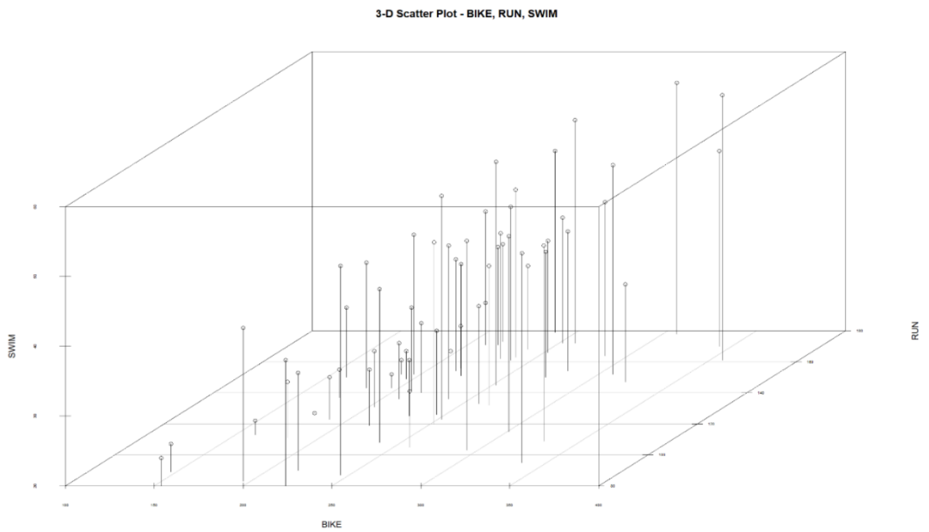
Here we can clearly see through boxplot that Bike has a univariate outlier. Regarding bivariate outlier we can clearly see through scatter plot that Bike and Swim has definite outlier and it has been labelled as such in following diagram Furthermore From the above scatter plot matrix, it can be seen that there the bivariate relationship between swim and bike is stronger as opposed to that between run and either swim or bike.



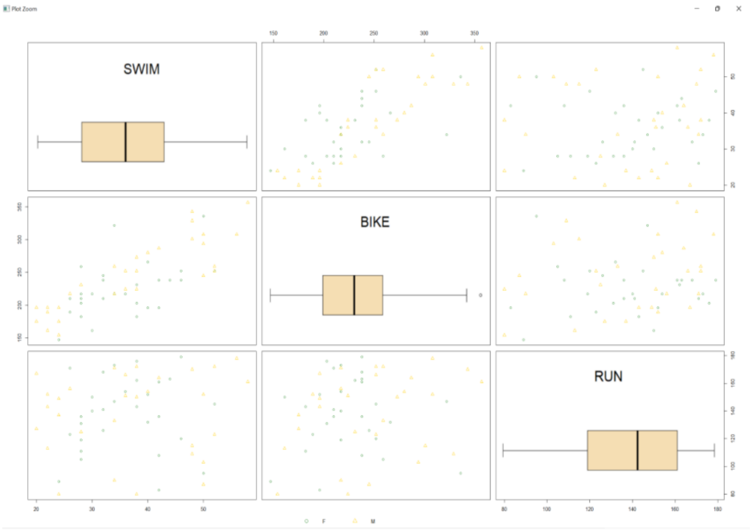


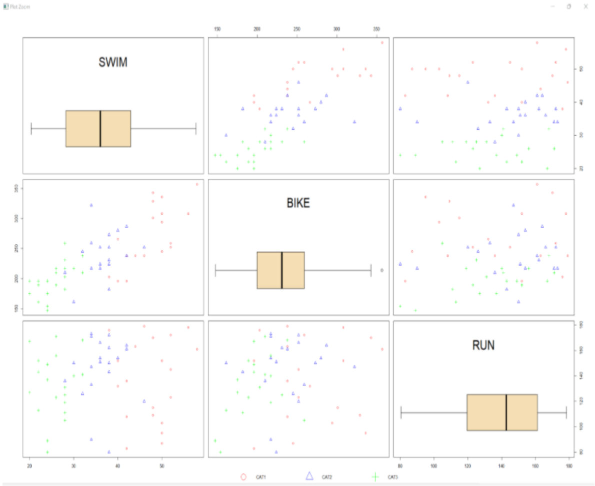


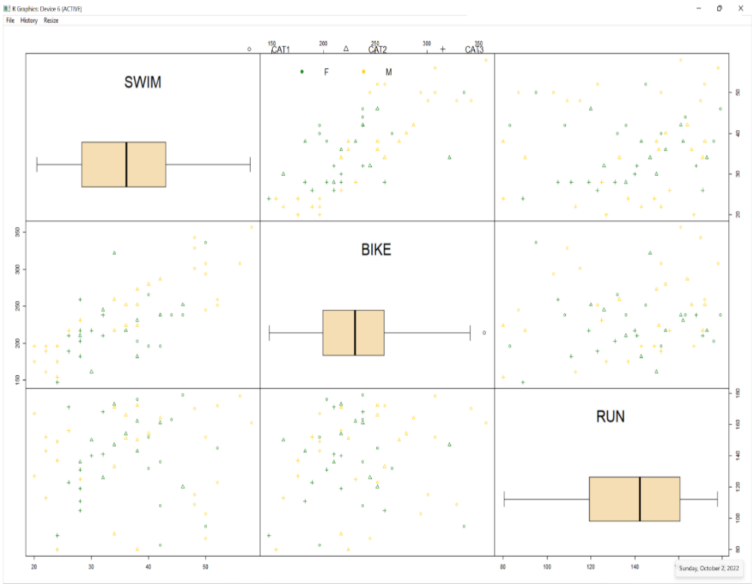
QC) One pattern that can be discerned from the above scatterplot is that lower swim race times are associated with the lower bike and run race times as most of the smaller circles appear in the lower left portion of the plot. Moreover, the highest swim race times are associated with the bike and run race times that fall in the average range. Hence if you are good at bike and run , you are bound to be good at swimming too.



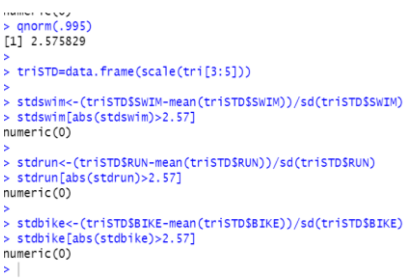
QD) From the above 3d scatterplot, it can be seen that there is no definite pattern in the relationship between bike, run and swim race times, except perhaps that bike race times seem more dispersed as compared to the run and swim race times. This points toward the fact that the participants showed divergent bike riding capabilities while being more at parity as far as running and swimming were concerned.



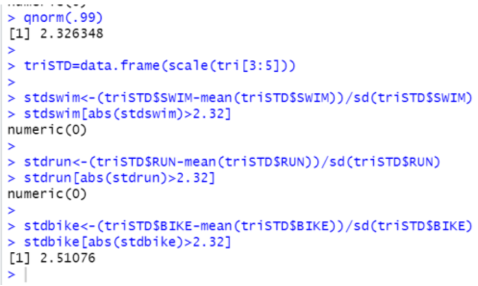




QD) Here, both symbols and colors have been used at the same time for scatterplot matrix . top left one contains gender as factor, top right contains age category as factor while bottom one contains both age and gender combined as factor. From the above three scatterplot matrices, it can be discerned that the relationship between male and females triathlete was stronger in bike-swim race times while it was weaker for triathletes belonging to the different age categories. On the other hand, the run race times against bike and swim race times were far more dispersed for all the age and gender criteria. Hence we can safely assume that gender has more to do with performance regarding overall triathlon than age criteria.



```
> qnorm(.995)
[1] 2.575829
>
> triSTD=data.frame(scale(tri[3:5]))
>
> stdswim<-(triSTD$SWIM-mean(triSTD$SWIM))/sd(triSTD$SWIM)
> stdswim[abs(stdswim)>2.57]
numeric(0)
>
> stdrun<-(triSTD$RUN-mean(triSTD$RUN))/sd(triSTD$RUN)
> stdrun[abs(stdrun)>2.57]
numeric(0)
>
> stdbike<-(triSTD$BIKE-mean(triSTD$BIKE))/sd(triSTD$BIKE)
> stdbike[abs(stdbike)>2.57]
numeric(0)
> |
```



```
> qnorm(.99)
[1] 2.326348
>
> triSTD=data.frame(scale(tri[3:5]))
>
> stdswim<-(triSTD$SWIM-mean(triSTD$SWIM))/sd(triSTD$SWIM)
> stdswim[abs(stdswim)>2.32]
numeric(0)
>
> stdrun<-(triSTD$RUN-mean(triSTD$RUN))/sd(triSTD$RUN)
> stdrun[abs(stdrun)>2.32]
numeric(0)
>
> stdbike<-(triSTD$BIKE-mean(triSTD$BIKE))/sd(triSTD$BIKE)
> stdbike[abs(stdbike)>2.32]
[1] 2.51076
> |
```

QE) After looking at boxplot in Ques. D we thought that perhaps bike has outlier. However, when we got marginal outlier using qnorm(.995) it didn't show any outlier for any data but when we used lower thresholds which is qnorm(.99) it showed one outlier for bike data. Hence we can assume that perhaps R has lower threshold for an outlier than what we are used to for marginal outlier calculation. Hence depending on your calculation we only have one marginal outlier for bike data.