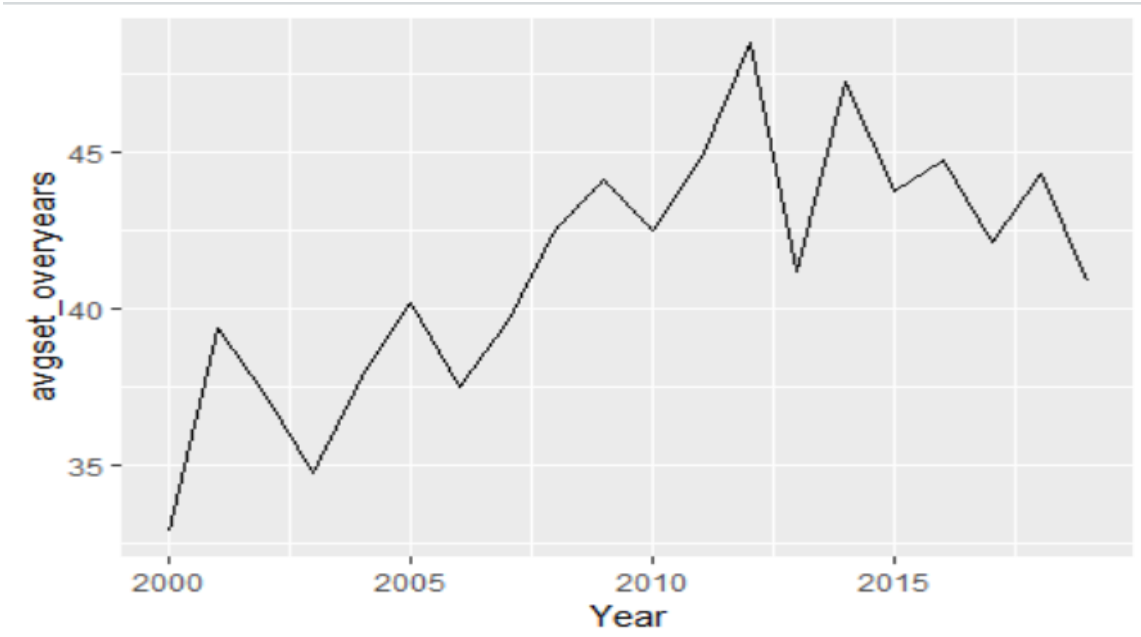
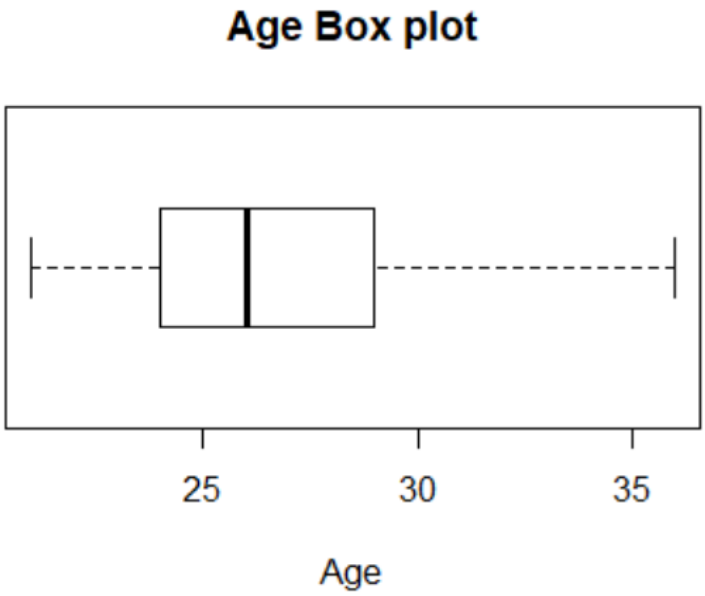
EXPLORATORY DATA ANALYSIS

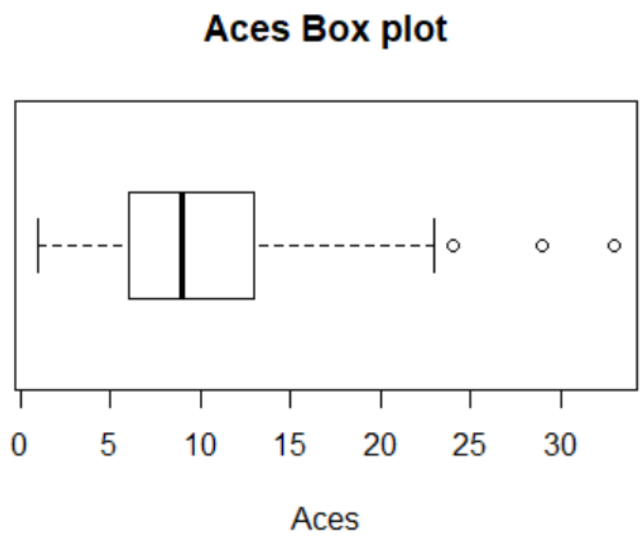
```
> summary(AustralianOpen_Finalists_allstats)
  PlayerName             Year       total_matchs   winpercentage      MatchID
 Length:277         Min.   :2000   Min.   :6.000   Min.   :0.8333   Length:27
7
 Class :character   1st Qu.:2005   1st Qu.:7.000   1st Qu.:0.8571   Class :ch
aracter
 Mode  :character   Median :2009   Median :7.000   Median :0.8571   Mode  :ch
aracter
                    Mean   :2009   Mean   :6.935   Mean   :0.9278
                    3rd Qu.:2014   3rd Qu.:7.000   3rd Qu.:1.0000
                    Max.   :2019   Max.   :7.000   Max.   :1.0000
    Round           AvgMinsPerGame  AvgSecsPerPoint AvgMinsPerSet    Tourname
nt
 Length:277         Min.   :2.930   Min.   :30.20   Min.   : 0.00   Length:27
7
 Class :character   1st Qu.:3.860   1st Qu.:37.60   1st Qu.:34.70   Class :ch
aracter
 Mode  :character   Median :4.280   Median :40.70   Median :40.60   Mode  :ch
aracter
                    Mean   :4.361   Mean   :41.25   Mean   :41.29
                    3rd Qu.:4.700   3rd Qu.:44.30   3rd Qu.:47.30
                    Max.   :9.030   Max.   :75.00   Max.   :93.30
 TotalMatchMins      Points           Age             Rank          Winner
 Min.   : 28.0   Min.   :    0   Min.   :21.0   Min.   : 1.000   Mode :logica
l
 1st Qu.:104.0   1st Qu.:    0   1st Qu.:24.0   1st Qu.: 1.000   FALSE:20
 Median :135.0   Median : 4675   Median :26.0   Median : 3.000   TRUE :257
 Mean   :144.3   Mean   : 5361   Mean   :26.8   Mean   : 9.289
 3rd Qu.:174.0   3rd Qu.: 9595   3rd Qu.:29.0   3rd Qu.: 8.000
 Max.   :353.0   Max.   :16790   Max.   :36.0   Max.   :86.000
   TotalSets        avgOdds          maxOdds         SP_Percent       RP_Perc
ent
 Min.   :0.000   Min.   :0.0000   Min.   :0.0000   Min.   :0.4000   Min.   :0
.1828
 1st Qu.:3.000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.5556   1st Qu.:0
.3644
 Median :3.000   Median :0.0000   Median :0.0000   Median :0.5984   Median :0
.4016
 Mean   :2.765   Mean   :0.6334   Mean   :0.6652   Mean   :0.5954   Mean   :0
.4046
 3rd Qu.:3.000   3rd Qu.:1.0700   3rd Qu.:1.1100   3rd Qu.:0.6356   3rd Qu.:0
.4444
 Max.   :3.000   Max.   :7.5400   Max.   :9.9500   Max.   :0.8172   Max.   :0
.6000
 BP_Win_Percentage      Aces        firstServeReturnsWon SecondServeReturnsWo
n
 Min.   :0.0000    Min.   : 1.000   Min.   : 4.00        Min.   : 3.00
 1st Qu.:0.4286    1st Qu.: 6.000   1st Qu.:17.00        1st Qu.:18.00
 Median :0.6471    Median : 9.000   Median :21.00        Median :22.00
 Mean   :0.5779    Mean   : 9.729   Mean   :22.15        Mean   :23.31
 3rd Qu.:0.8000    3rd Qu.:13.000   3rd Qu.:26.00        3rd Qu.:29.00
 Max.   :1.0000    Max.   :33.000   Max.   :47.00        Max.   :45.00
 FirstServesIn     DoubleFaults   FirstServePercentage
 Min.   : 12.00   Min.   :0.000   Min.   :0.3692
 1st Qu.: 47.00   1st Qu.:1.000   1st Qu.:0.5806
 Median : 57.00   Median :2.000   Median :0.6316
 Mean   : 62.08   Mean   :2.412   Mean   :0.6267
 3rd Qu.: 77.00   3rd Qu.:4.000   3rd Qu.:0.6754
```

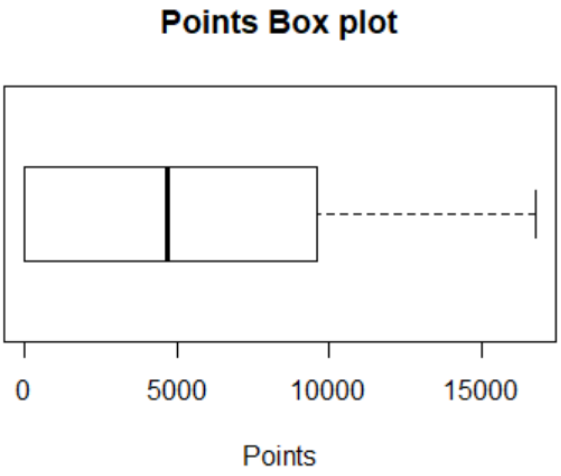
```
 Max.   :135.00    Max.   :9.000    Max.   :0.8088
```

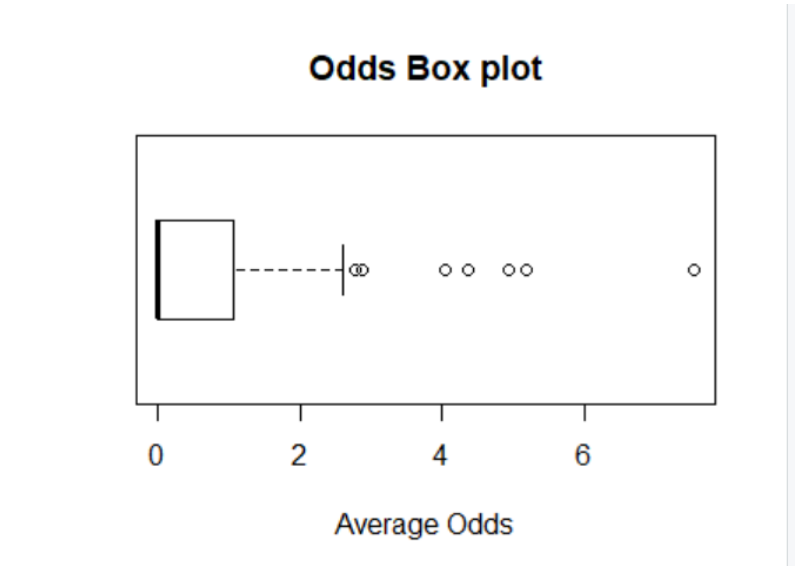
```
library(ggplot2)
> ggplot(AustralianOpen_Finalists_allstats,aes(x=Year,y=avgset_overyears))+ge
om_line()
```

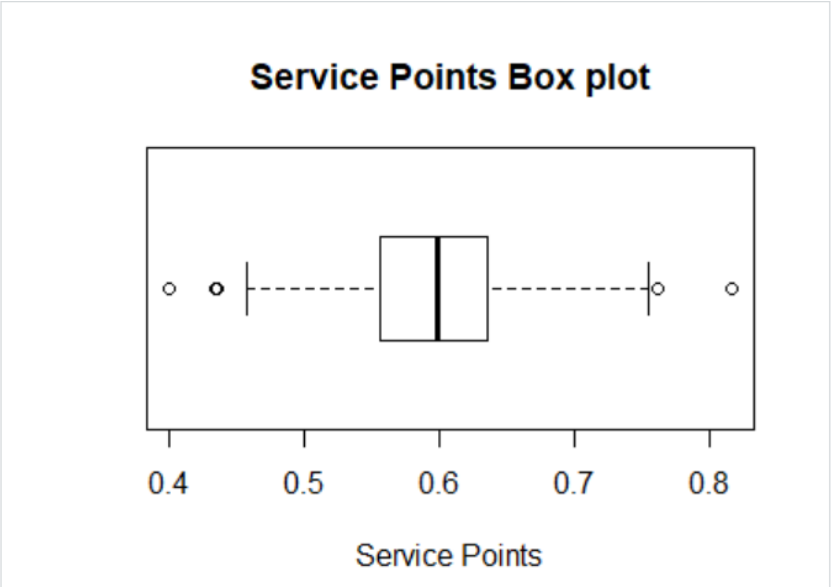


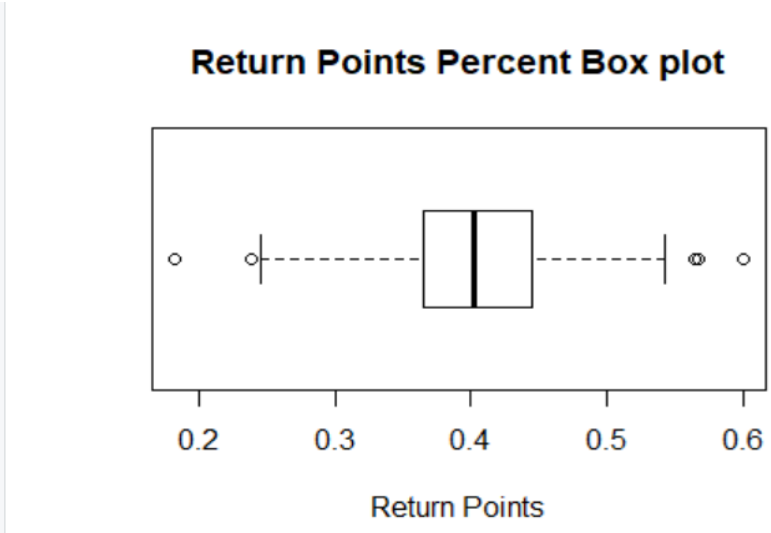



Aces Box plot
0
5
10
15
20
25
30
Aces

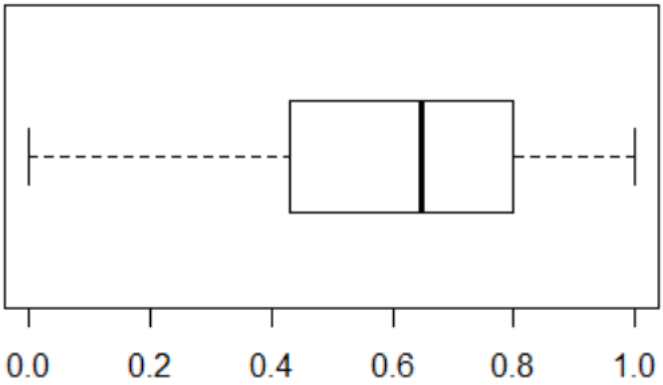

Points Box plot
0
5000
10000
15000
Points


Odds Box plot
0
2
4
6
Average Odds

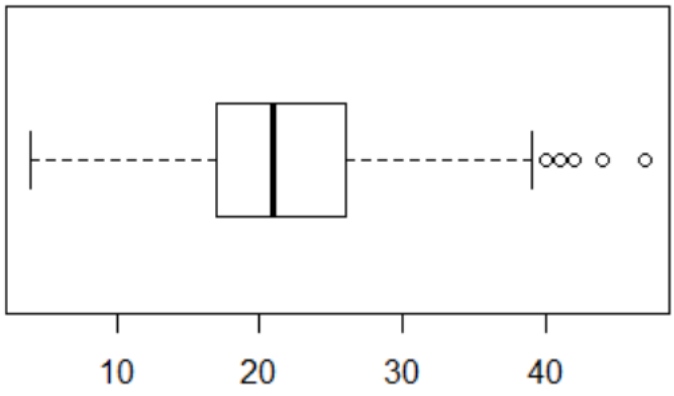

Service Points Box plot
0.4
0.5
0.6
0.7
0.8
Service Points


Return Points Percent Box plot
0.2
0.3
0.4
0.5
0.6
Return Points

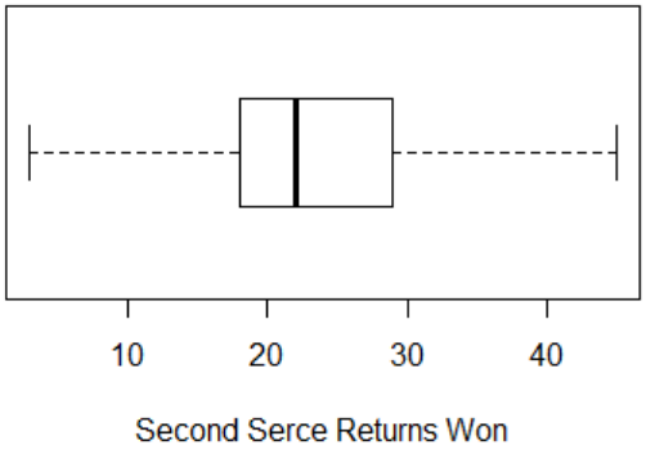
Break Points Win Percent Box plot



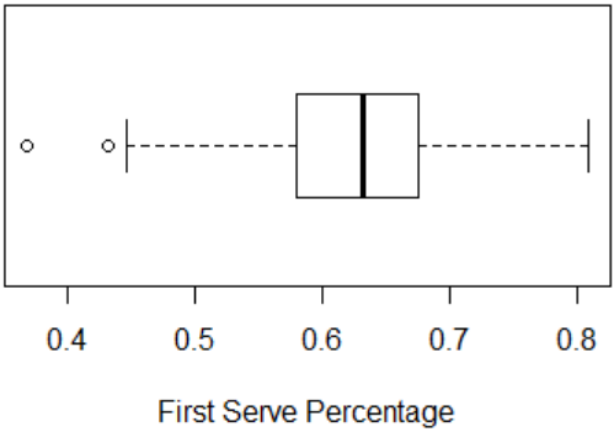
First Serve Returns Won Box plot



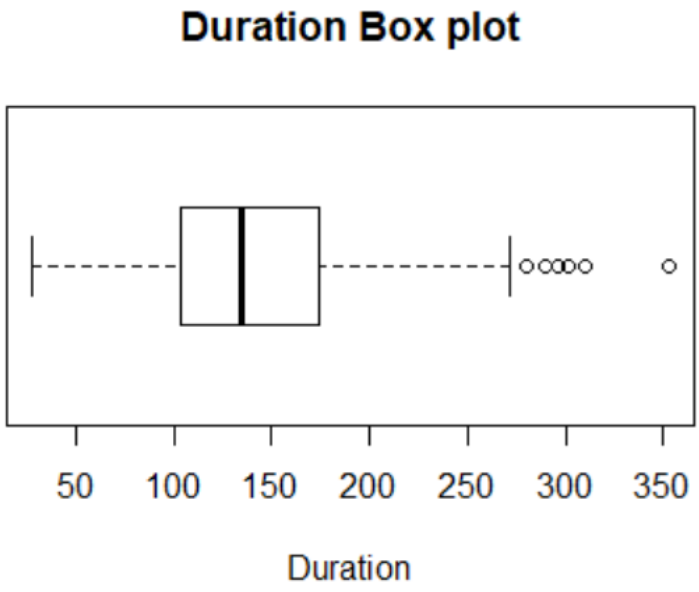
Second Serve Returns Won Box plot



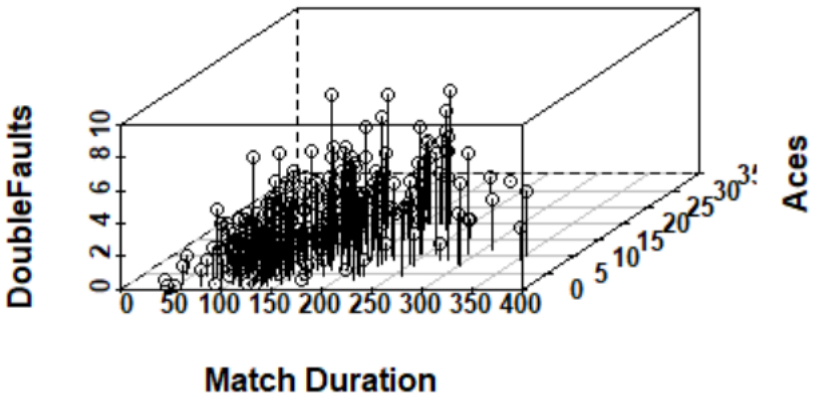
First Serve Percentage Box plot



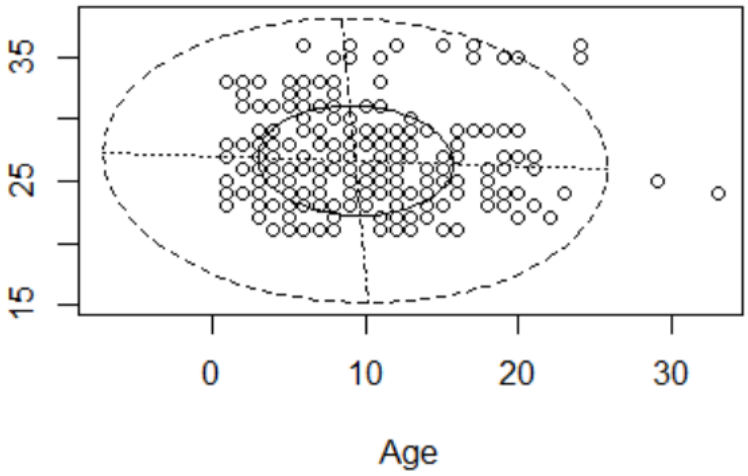

```
Sd3 <- scatterplot3d(AustralianOpen_Finalists_allstats$TotalMatchMins,Australian
Open_Finalists_allstats$Aces,AustralianOpen_Finalists_allstats$DoubleFaults,xlab="Match Duration", ylab="Aces", angle=45,zlab="DoubleFaults", lty.hide=2,type="h",y.margin.add=0.1,font.axis=2,font.lab=2)
```



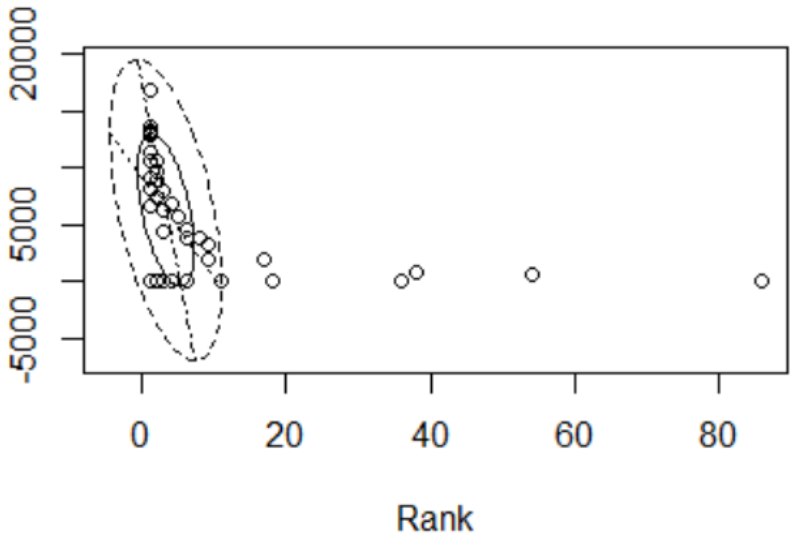

```
Sd3 <- scatterplot3d(AustralianOpen_Finalists_allstats$TotalMatchMins,Australi
anOpen_Finalists_allstats$Aces,AustralianOpen_Finalists_allstats$DoubleFault
s,xlab="Match Duration", ylab="Aces", angle=45,zlab="DoubleFaults", lty.hide=
2,type="h",y.margin.add=0.1,font.axis=2,font.lab=2)
```



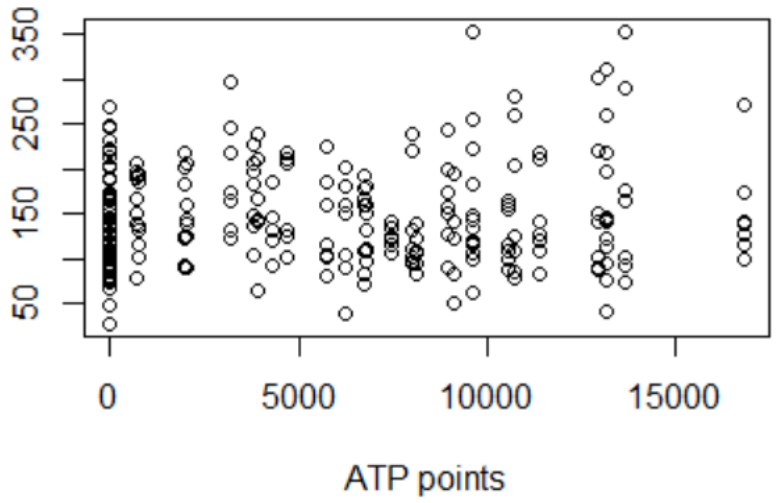

```
> mlab="Age"
> plab="Aces"
> Match_Aces_Age=data.frame(AustralianOpen_Finalists_allstats$Aces, Australia
nOpen_Finalists_allstats$Age)
> bvbox(Match_Aces_Age, mtitle = "", xlab = mlab, ylab = plab)
```




```
> mlab="Rank"
> plab="Points"
> Match_Rank_Points=data.frame(AustralianOpen_Finalists_allstats$Rank, Austra
lianOpen_Finalists_allstats$Points)
> bvbox(Match_Rank_Points, mtitle = "", xlab = mlab, ylab = plab)
```

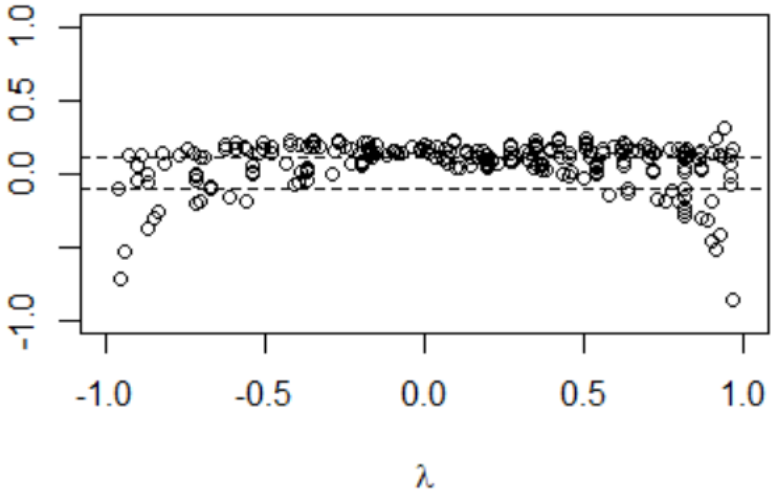



```
> plot(AustralianOpen_Finalists_allstats$Points, AustralianOpen_Finalists_all
stats$TotalMatchMins,xlab="ATP points",ylab="Match Duration")
> plot(AustralianOpen_Finalists_allstats$SP_Per
```

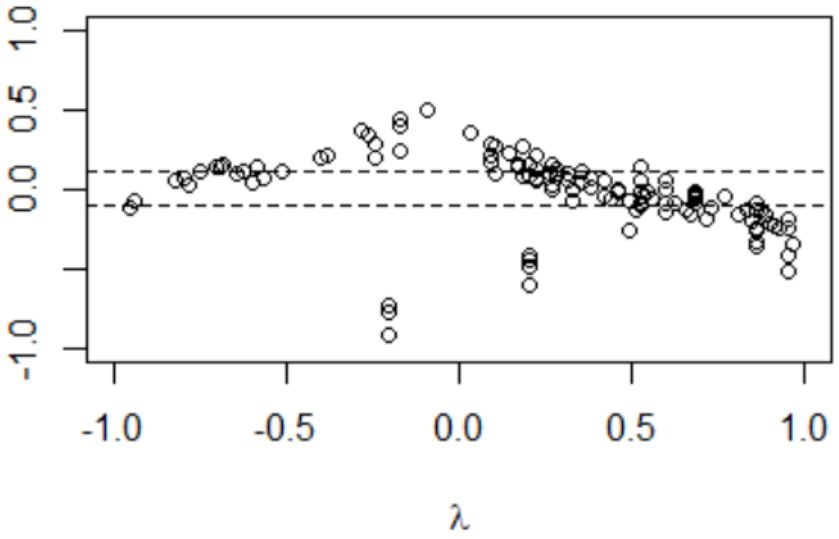


CHI plots

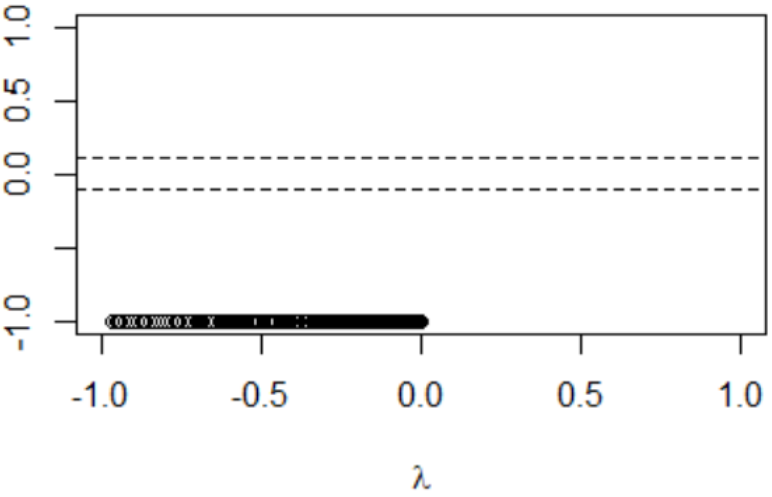
```
> plab = "Match Duration"
> with(AustralianOpen_Finalists_allstats, plot(Aces, TotalMatchMins, xlab = m
lab , ylab = plab, cex.lab = 0.9))
> with(AustralianOpen_Finalists_allstats, chiplot(Aces, TotalMatchMins))
```



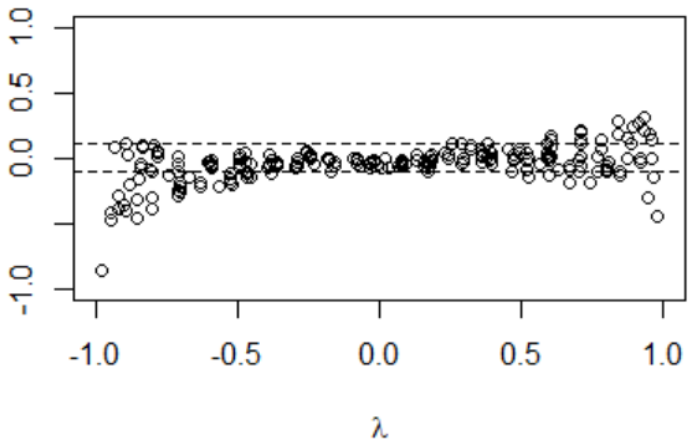

```
> mlab = "Average Odds"
> plab = "Points"
> with(AustralianOpen_Finalists_allstats, plot(avgOdds, Points, xlab = mlab ,
ylab = plab, cex.lab = 0.9))
> with(AustralianOpen_Finalists_allstats, chiplot(avgOdds, Points))
```



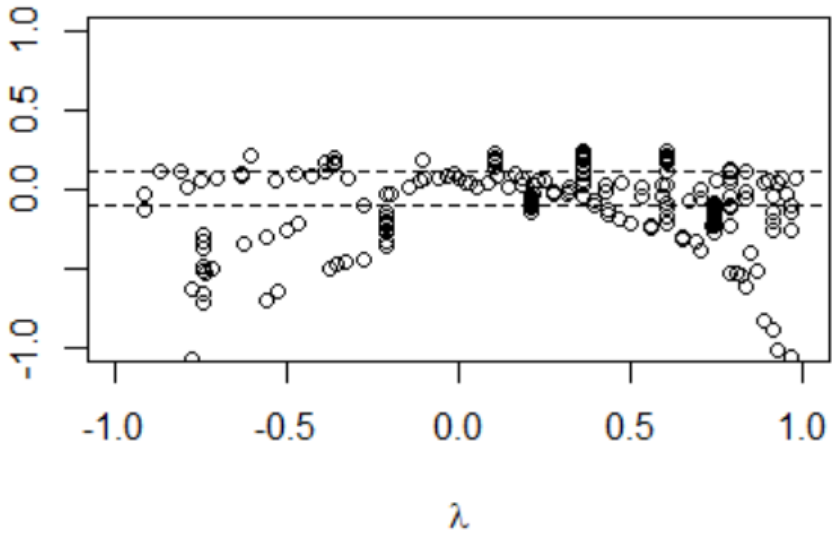

```
> mlab = "Service Points"
> plab = "Return Points"
> with(AustralianOpen_Finalists_allstats, plot(SP_Percent, RP_Percent, xlab =
mlab , ylab = plab, cex.lab = 0.9))
> with(AustralianOpen_Finalists_allstats, chiplot(SP_Percent, RP_Percent))
```




```
> mlab = "First Serve Returns Won"
> plab = "Second Serve Returns Won"
> with(AustralianOpen_Finalists_allstats, plot(firstServeReturnsWon, SecondSe
rveReturnsWon, xlab = mlab , ylab = plab, cex.lab = 0.9))
> with(AustralianOpen_Finalists_allstats, chiplot(firstServeReturnsWon, Secon
dServeReturnsWon))
```

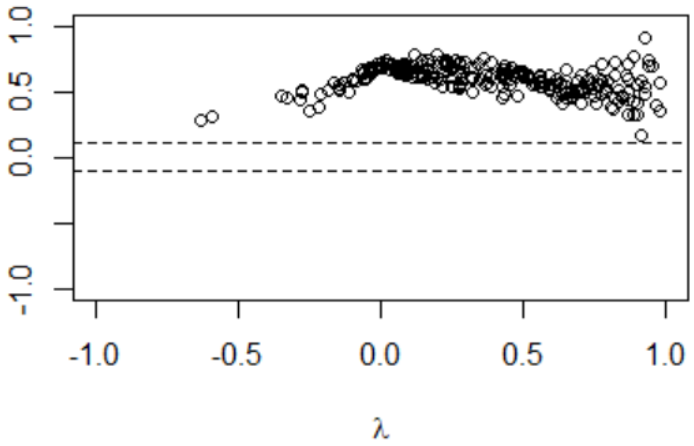



```
> mlab = "First Serve In"
> plab = "Double Faults"
> with(AustralianOpen_Finalists_allstats, plot(FirstServesIn, DoubleFaults, x
lab = mlab , ylab = plab, cex.lab = 0.9))
> with(AustralianOpen_Finalists_allstats, chiplot(FirstServesIn, DoubleFaults
))
```

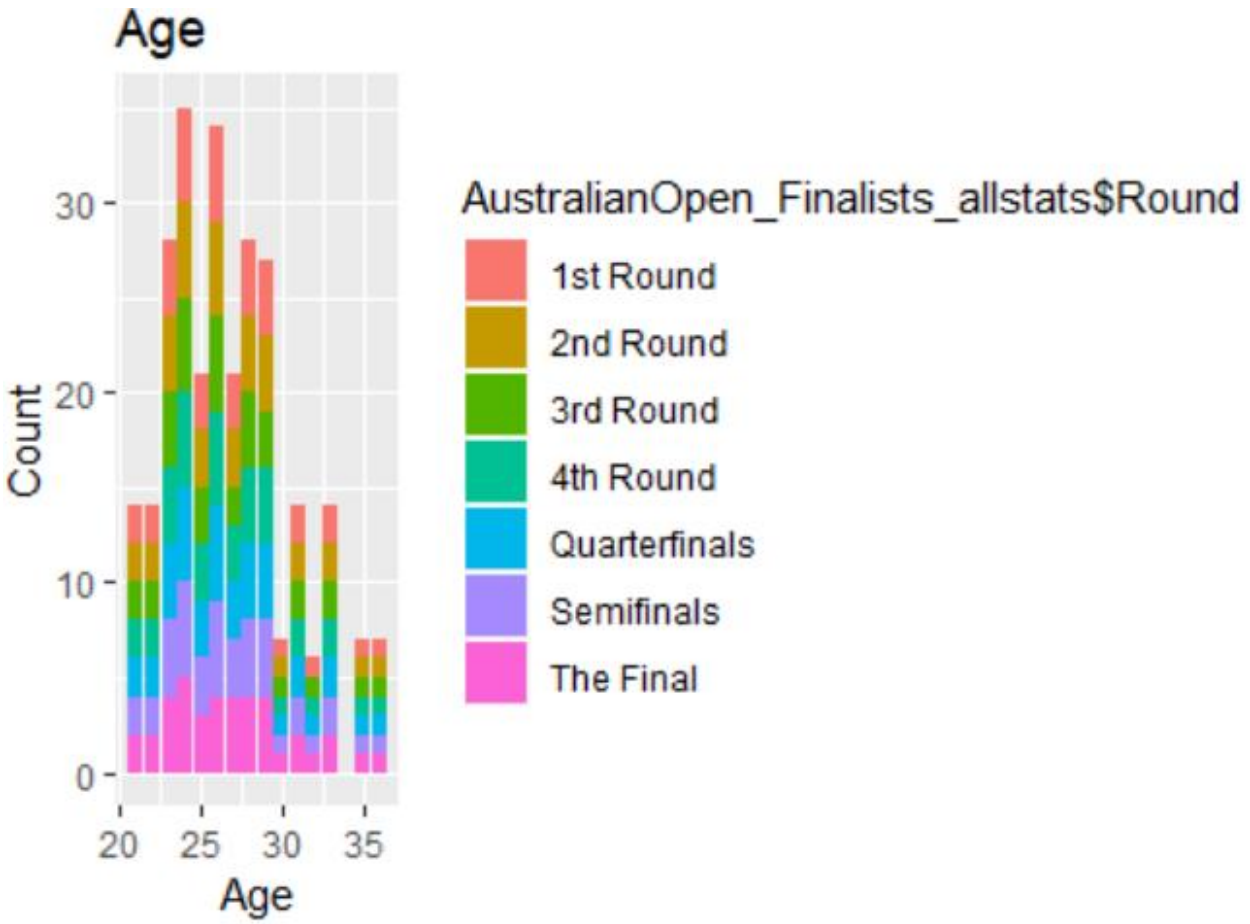


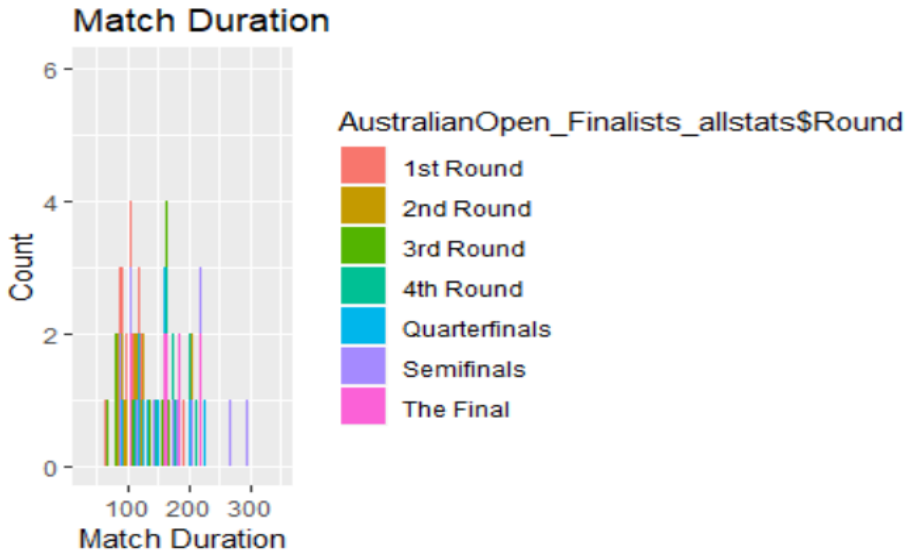

```
> mlab = "First Serve In"
> plab = "Match Duration"
```

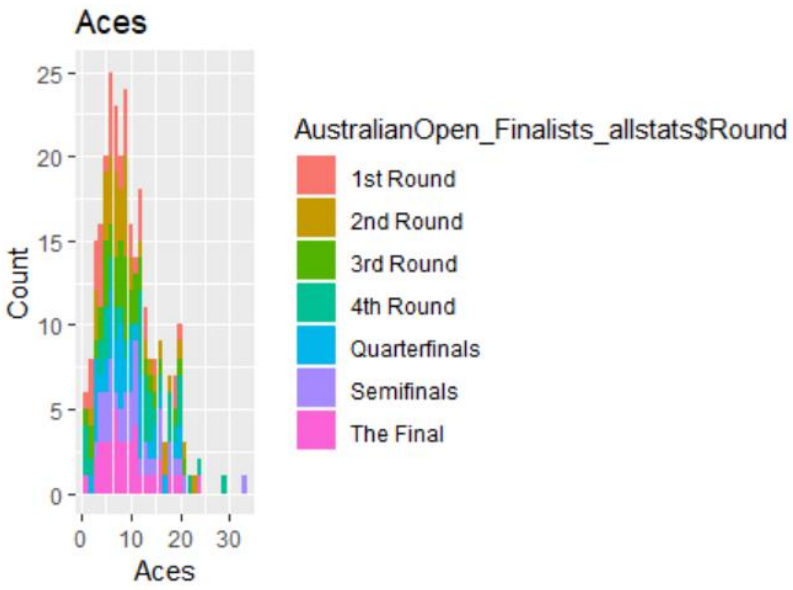
```
> with(AustralianOpen_Finalists_allstats, plot(FirstServesIn, TotalMatchMins,
xlab = mlab , ylab = plab, cex.lab = 0.9))
> with(AustralianOpen_Finalists_allstats, chiplot(FirstServesIn, TotalMatchMi
ns))
```



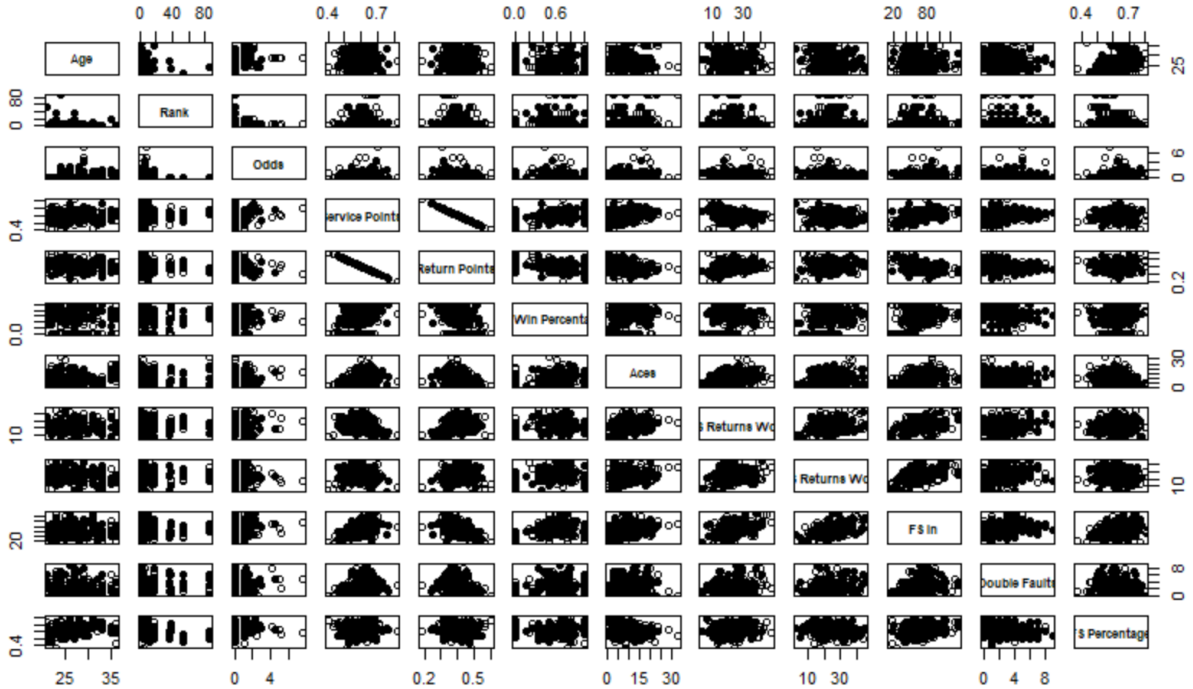

```
> ggplot(AustralianOpen_Finalists_allstats,aes(x=AustralianOpen_Finalists_all
stats$Age,fill=AustralianOpen_Finalists_allstats$Round)) + geom_bar() +
+   labs(y= "Count", x="Age", title = "Age")
```


Age
Count
30
20
10
0
20
25
30
35
Age
AustralianOpen_Finalists_allstats$Round
1st Round
2nd Round
3rd Round
4th Round
Quarterfinals
Semifinals
The Final


Match Duration
Count
6
4
2
0
100
200
300
Match Duration
AustralianOpen_Finalists_allstats$Round
1st Round
2nd Round
3rd Round
4th Round
Quarterfinals
Semifinals
The Final



Correlation plot



t.tests

```
> t.test(AustralianOpen_Finalists_allstats$Age[AustralianOpen_Finalists_allstats$Winner=="TRUE"],AustralianOpen_Finalists_allstats$Age[AustralianOpen_Finalists_allstats$Winner=='FALSE'],var.equal=TRUE)
```

```
        Two Sample t-test

data:  AustralianOpen_Finalists_allstats$Age[AustralianOpen_Finalists_allstat
s$Winner ==  and AustralianOpen_Finalists_allstats$Age[AustralianOpen_Finalis
ts_allstats$Winner ==     "TRUE"] and     "FALSE"]
t = 0.93807, df = 275, p-value = 0.349
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.8897711  2.5096155
sample estimates:
mean of x mean of y
 26.85992  26.05000
```

Not Significant

```
> t.test(AustralianOpen_Finalists_allstats$Rank[AustralianOpen_Finalists_alls
tats$Winner=='FALSE'],AustralianOpen_Finalists_allstats$Rank[AustralianOpen_F
inalists_allstats$Winner=="TRUE"],var.equal=TRUE)

        Two Sample t-test

data:  AustralianOpen_Finalists_allstats$Rank[AustralianOpen_Finalists_allsta
ts$Winner ==  and AustralianOpen_Finalists_allstats$Rank[AustralianOpen_Final
ists_allstats$Winner ==     "FALSE"] and     "TRUE"]
t = 1.58, df = 275, p-value = 0.1153
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.501002 13.704503
sample estimates:
mean of x mean of y
14.950000  8.848249
```

Not Significant

```
> t.test(AustralianOpen_Finalists_allstats$avgOdds[AustralianOpen_Finalists_a
llstats$Winner=='TRUE'],AustralianOpen_Finalists_allstats$avgOdds[AustralianO
pen_Finalists_allstats$Winner=='FALSE'],var.equal=TRUE)

        Two Sample t-test

data:  AustralianOpen_Finalists_allstats$avgOdds[AustralianOpen_Finalists_all
stats$Winner ==  and AustralianOpen_Finalists_allstats$avgOdds[AustralianOpen
_Finalists_allstats$Winner ==     "TRUE"] and     "FALSE"]
t = -2.9655, df = 275, p-value = 0.003287
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.0493503 -0.2120077
sample estimates:
mean of x mean of y
 0.587821  1.218500
```

Significant

```
> t.test(AustralianOpen_Finalists_allstats$SP_Percent[AustralianOpen_Finalist
s_allstats$Winner=='TRUE'],AustralianOpen_Finalists_allstats$SP_Percent[Austr
alianOpen_Finalists_allstats$Winner=='FALSE'],var.equal=TRUE)
```

```
        Two Sample t-test

data:  AustralianOpen_Finalists_allstats$SP_Percent[AustralianOpen_Finalists_
allstats$Winner ==  and AustralianOpen_Finalists_allstats$SP_Percent[Australi
anOpen_Finalists_allstats$Winner ==     "TRUE"] and     "FALSE"]
t = -5.4811, df = 275, p-value = 9.561e-08
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.10123025 -0.04772923
sample estimates:
mean of x mean of y
0.5899766 0.6644563
```

Significant

```
> t.test(AustralianOpen_Finalists_allstats$RP_Percent [AustralianOpen_Finalis
ts_allstats$Winner=='TRUE'],AustralianOpen_Finalists_allstats$RP_Percent [Aus
tralianOpen_Finalists_allstats$Winner=='FALSE'],var.equal=TRUE)

        Two Sample t-test

data:  AustralianOpen_Finalists_allstats$RP_Percent[AustralianOpen_Finalists_
allstats$Winner ==  and AustralianOpen_Finalists_allstats$RP_Percent[Australi
anOpen_Finalists_allstats$Winner ==     "TRUE"] and     "FALSE"]
t = 5.4811, df = 275, p-value = 9.561e-08
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.04772923 0.10123025
sample estimates:
mean of x mean of y
0.4100234 0.3355437
```

Significant

```
> t.test(AustralianOpen_Finalists_allstats$RP_Percent [AustralianOpen_Finalis
ts_allstats$Winner=='TRUE'],AustralianOpen_Finalists_allstats$RP_Percent [Aus
tralianOpen_Finalists_allstats$Winner=='FALSE'],var.equal=TRUE)

        Two Sample t-test

data:  AustralianOpen_Finalists_allstats$RP_Percent[AustralianOpen_Finalists_
allstats$Winner ==  and AustralianOpen_Finalists_allstats$RP_Percent[Australi
anOpen_Finalists_allstats$Winner ==     "TRUE"] and     "FALSE"]
t = 5.4811, df = 275, p-value = 9.561e-08
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.04772923 0.10123025
sample estimates:
mean of x mean of y
0.4100234 0.3355437

> t.test(AustralianOpen_Finalists_allstats$BP_Win_Percentage[AustralianOpen_F
inalists_allstats$Winner=='TRUE'],AustralianOpen_Finalists_allstats$BP_Win_Pe
rcentage[AustralianOpen_Finalists_allstats$Winner=='FALSE'],var.equal=TRUE)

        Two Sample t-test
```

```
data:  AustralianOpen_Finalists_allstats$BP_Win_Percentage[AustralianOpen_Fin
alists_allstats$Winner ==  and AustralianOpen_Finalists_allstats$BP_Win_Perce
ntage[AustralianOpen_Finalists_allstats$Winner ==     "TRUE"] and     "FALSE"
]
t = -0.26861, df = 275, p-value = 0.7884
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.1659194  0.1260774
sample estimates:
mean of x mean of y
 0.576470  0.596391
```

Significant

```
> t.test(AustralianOpen_Finalists_allstats$Aces[AustralianOpen_Finalists_alls
tats$Winner=='TRUE'],AustralianOpen_Finalists_allstats$Aces[AustralianOpen_Fi
nalists_allstats$Winner=='FALSE'],var.equal=TRUE)

        Two Sample t-test

data:  AustralianOpen_Finalists_allstats$Aces[AustralianOpen_Finalists_allsta
ts$Winner ==  and AustralianOpen_Finalists_allstats$Aces[AustralianOpen_Final
ists_allstats$Winner ==     "TRUE"] and     "FALSE"]
t = 1.3264, df = 275, p-value = 0.1858
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.824079  4.228359
sample estimates:
mean of x mean of y
  9.85214   8.15000
```

Not Significant

```
> t.test(AustralianOpen_Finalists_allstats$firstServeReturnsWon[AustralianOpe
n_Finalists_allstats$Winner=='TRUE'],AustralianOpen_Finalists_allstats$firstS
erveReturnsWon[AustralianOpen_Finalists_allstats$Winner=='FALSE'],var.equal=T
RUE)

        Two Sample t-test

data:  AustralianOpen_Finalists_allstats$firstServeReturnsWon[AustralianOpen_
Finalists_allstats$Winner ==  and AustralianOpen_Finalists_allstats$firstServ
eReturnsWon[AustralianOpen_Finalists_allstats$Winner ==     "TRUE"] and     "
FALSE"]
t = 2.4803, df = 275, p-value = 0.01373
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.8667385 7.5359853
sample estimates:
mean of x mean of y
 22.45136  18.25000
```

Significant

```
> t.test(AustralianOpen_Finalists_allstats$SecondServeReturnsWon[AustralianOp
en_Finalists_allstats$Winner=='TRUE'],AustralianOpen_Finalists_allstats$Secon
dServeReturnsWon[AustralianOpen_Finalists_allstats$Winner=='FALSE'],var.equal
=TRUE)
```

```
        Two Sample t-test

data:  AustralianOpen_Finalists_allstats$SecondServeReturnsWon[AustralianOpen
_Finalists_allstats$Winner ==  and AustralianOpen_Finalists_allstats$SecondSe
rveReturnsWon[AustralianOpen_Finalists_allstats$Winner ==     "TRUE"] and
"FALSE"]
t = 2.8927, df = 275, p-value = 0.004125
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.621884 8.532201
sample estimates:
mean of x mean of y
 23.67704  18.60000
```

Significant

```
> t.test(AustralianOpen_Finalists_allstats$FirstServesIn[AustralianOpen_Final
ists_allstats$Winner=='TRUE'],AustralianOpen_Finalists_allstats$FirstServesIn
[AustralianOpen_Finalists_allstats$Winner=='FALSE'],var.equal=TRUE)

        Two Sample t-test

data:  AustralianOpen_Finalists_allstats$FirstServesIn[AustralianOpen_Finalis
ts_allstats$Winner ==  and AustralianOpen_Finalists_allstats$FirstServesIn[Au
stralianOpen_Finalists_allstats$Winner ==     "TRUE"] and     "FALSE"]
t = -2.5272, df = 275, p-value = 0.01206
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -22.952773  -2.851507
sample estimates:
mean of x mean of y
 61.14786  74.05000
```

Significant

```
> t.test(AustralianOpen_Finalists_allstats$DoubleFaults[AustralianOpen_Finali
sts_allstats$Winner=='TRUE'],AustralianOpen_Finalists_allstats$DoubleFaults[A
ustralianOpen_Finalists_allstats$Winner=='FALSE'],var.equal=TRUE)

        Two Sample t-test

data:  AustralianOpen_Finalists_allstats$DoubleFaults[AustralianOpen_Finalist
s_allstats$Winner ==  and AustralianOpen_Finalists_allstats$DoubleFaults[Aust
ralianOpen_Finalists_allstats$Winner ==     "TRUE"] and     "FALSE"]
t = -3.9623, df = 275, p-value = 9.464e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.5626752 -0.8614493
sample estimates:
mean of x mean of y
 2.287938  4.000000
```

Significant

```
> t.test(AustralianOpen_Finalists_allstats$FirstServePercentage[AustralianOpe
n_Finalists_allstats$Winner=='FALSE'],AustralianOpen_Finalists_allstats$First
ServePercentage[AustralianOpen_Finalists_allstats$Winner=="TRUE"],var.equal=T
RUE)
```

```
        Two Sample t-test

data:  AustralianOpen_Finalists_allstats$FirstServePercentage[AustralianOpen_
Finalists_allstats$Winner ==  and AustralianOpen_Finalists_allstats$FirstServ
ePercentage[AustralianOpen_Finalists_allstats$Winner ==     "FALSE"] and
"TRUE"]
t = -1.8677, df = 275, p-value = 0.06287
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.065260012  0.001717735
sample estimates:
mean of x mean of y
0.5972012 0.6289723
```

Not Significant