SCS Assignment 8

Assignment 8 - Prepare a case study for statistical analysis. Take any open-source data set, apply some statistical analysis, representation technique on chosen dataset. Describe data set, write and explain code and output

Problem Statement

Analyse the batting performance of Virat Kohli in test cricket from the year 2016 to 2019. Find out the year in which he performed the best.

Dataset

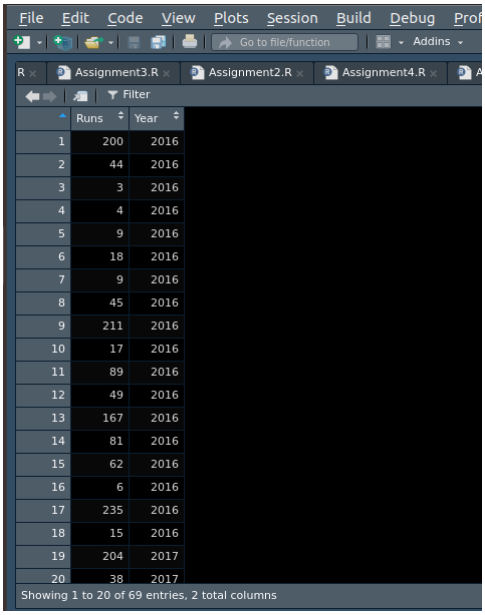
For this, the innings by innings data of the runs scored by Virat Kohli is required. This data was obtained from the website `www.espncricinfo.com`, a famous cricket website which also collects ball by ball data for every official cricket match that happens around the globe. The data obtained was stored in a csv file.

Statistical Analysis performed - code and output

Step 1 : Read the csv file

```
runs <- read.csv("~/Downloads/VKruns.csv")
```

A snapshot of the data :



Step 2 : Do the analysis for the whole 4 year period

Perform One way ANOVA to verify whether the data collected comes from the same population.

Null Hypothesis - There is no significant difference in the means of the samples

Code -

```
# perform one way anova
result <- aov(Runs~Year, data = runs)
ans <- summary(res)
print(ans)
print(result)
```

Output and interpretation -

```
             Df Sum Sq Mean Sq F value Pr(>F)
Year          1   8477    8477   3.224 0.0789 .
Residuals    48 126207    2629
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> print(result)
Call:
   aov(formula = Runs ~ Year, data = runs)

Terms:
                    Year Residuals
Sum of Squares    2658.9  322521.8
Deg. of Freedom        1        67

Residual standard error: 69.38125
Estimated effects may be unbalanced
```

At 5% significance level,
0.0789 > 0.05

Hence we can accept the null hypothesis and conclude that the data comes from the same population. Therefore, we can say that Virat Kohli performed consistently well in all these four years. Furthermore, his healthy average of 61 strengthens the claim.

Overall statistics for the 4 year period

```
> sprintf("No of innings - %d",length(runs$Runs))
[1] "No of innings - 69"
> sprintf("Total Runs - %d",sum(runs$Runs))
[1] "Total Runs - 4257"
```

```
> sprintf("Average - %f",mean(runs$Runs))
[1] "Average - 61.695652"
```

Step 3 : Apply various statistical techniques to find the best year among the four given years

Calculate the average and the standard deviation for all four years

```
runs2016 <- subset(runs,Year == 2016, select = c("Runs","Year"))
avg2016 <- mean(runs2016$Runs)
sd2016 <- sd(runs2016$Runs)

runs2017 <- subset(runs,Year == 2017, select = c("Runs","Year"))
avg2017 <- mean(runs2017$Runs)
sd2017 <- sd(runs2017$Runs)

runs2018 <- subset(runs,Year == 2018, select = c("Runs","Year"))
avg2018 <- mean(runs2018$Runs)
sd2018 <- sd(runs2018$Runs)

runs2019 <- subset(runs,Year == 2019, select = c("Runs","Year"))
avg2019 <- mean(runs2019$Runs)
sd2019 <- sd(runs2019$Runs)
```

Tabulate the number of innings, runs scored, average and standard deviation for each year

```
data <- data.frame(c("2016","2017","2018","2019"),

c(length(runs2016$Runs),length(runs2017$Runs),length(runs2018$Runs),length
(runs2019$Runs)),

c(sum(runs2016$Runs),sum(runs2017$Runs),sum(runs2018$Runs),sum(runs2019$Ru
ns)),
                  c(avg2016,avg2017,avg2018,avg2019),
                  c(sd2016,sd2017,sd2018,sd2019))

colnames(data) <- c("Year","Innings","Runs","Avg","Std Dev")

print(data)
```

Output

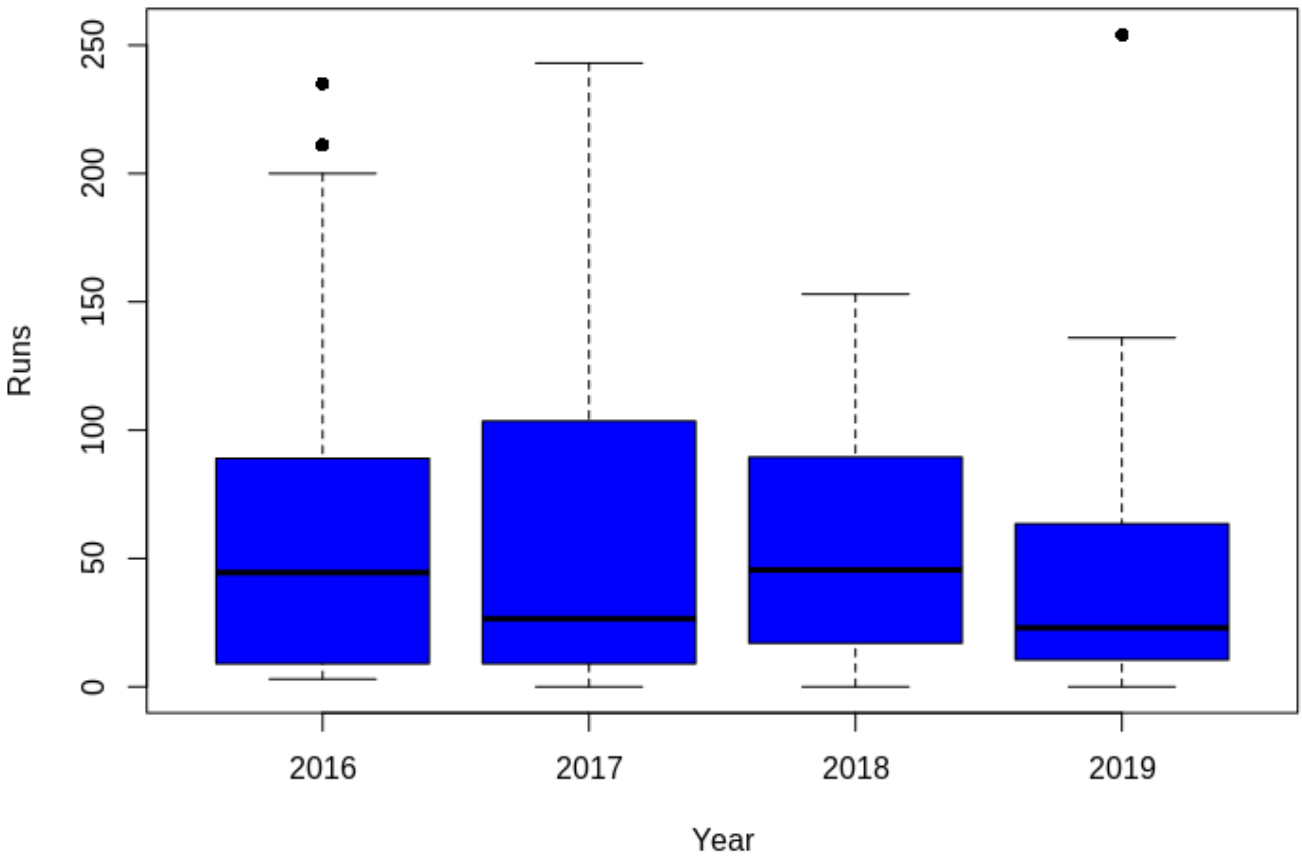
```
Year   Innings Runs    Avg   Std Dev
1 2016      18 1264 70.22222 78.44935
```

```
2 2017      16 1059 66.18750 83.19713
3 2018      24 1322 55.08333 48.54036
4 2019      11  612 55.63636 77.10677
```

It can be seen that 2018 was the best year for Virat Kohli in terms of runs scored. However, 2016 and 2017 were better in terms of his average.

Boxplot of the runs scored in each year

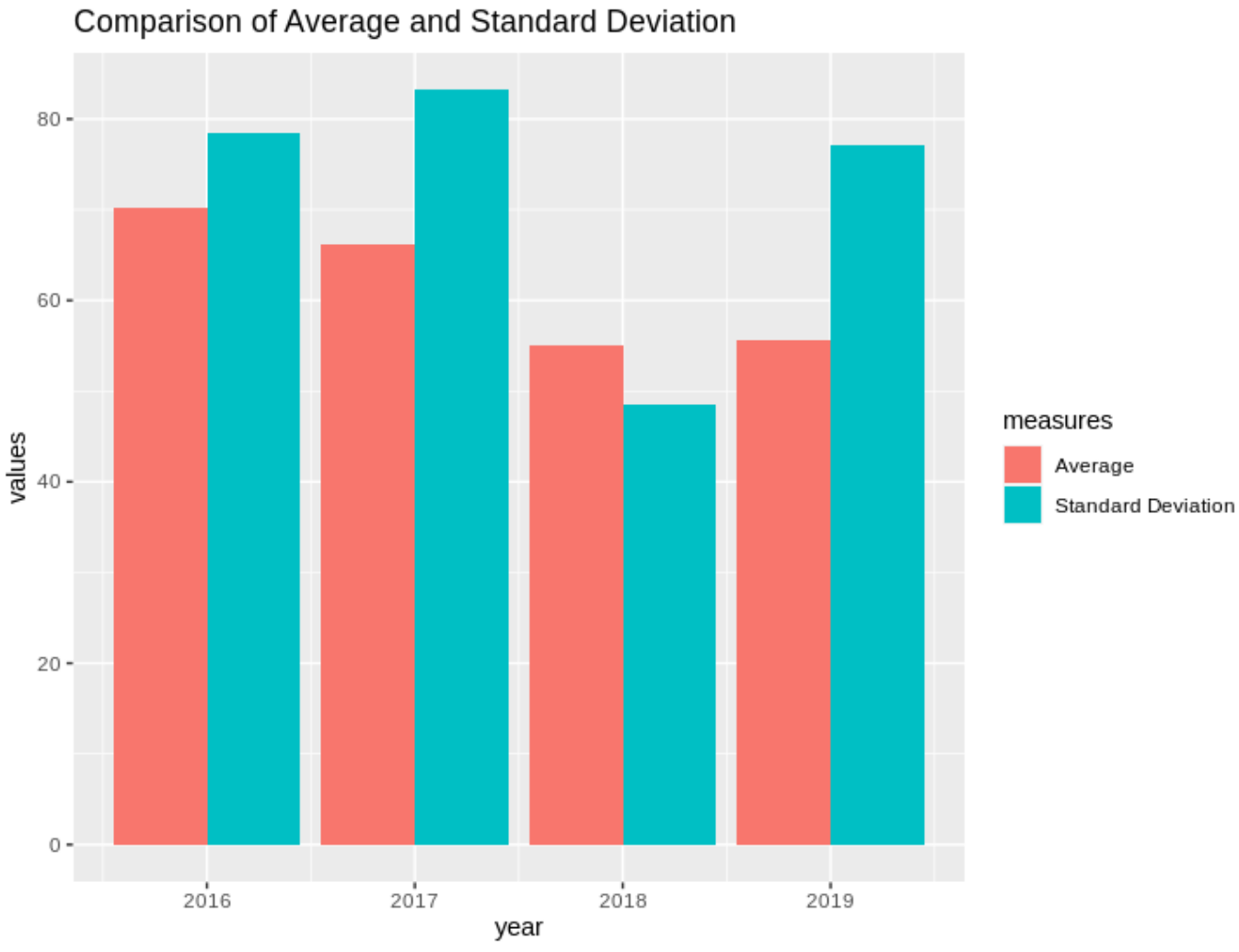
```
boxplot(runs2016$Runs,runs2017$Runs,runs2018$Runs,runs2019$Runs,col =
"Blue",xlab = "Year"
        ,ylab = "Runs",names = c("2016","2017","2018","2019"),pch = 16)
```



A boxplot is a graph that gives you a good indication of how the values in the data are spread out. From the boxplot, it can be observed that he played two innings in 2016 which are outliers. In other words, those two innings may have contributed to his high average. Similarly, there was one innings in 2019 which did the same thing. On the other hand, 2017 and 2018 have no outliers.

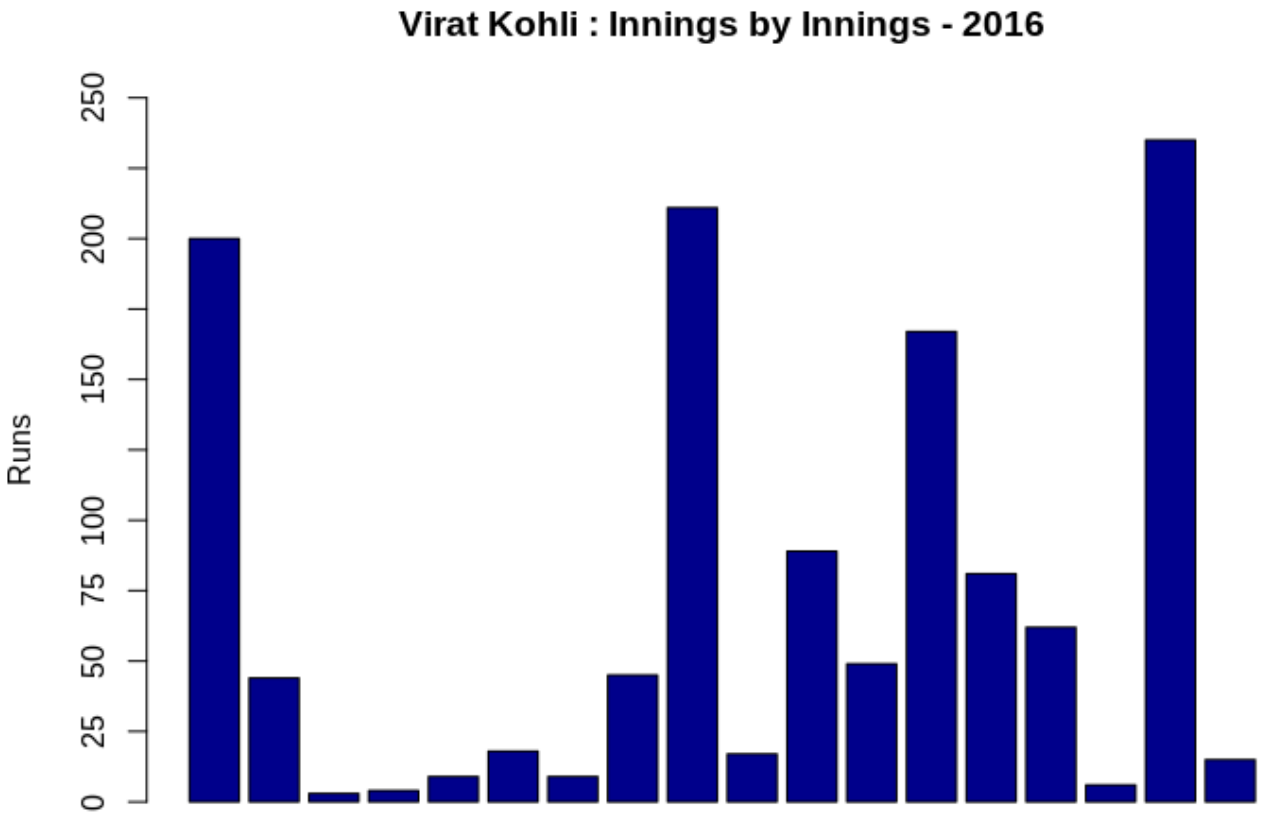
Compare the average and standard deviation for each year using a side by side bar plot

```
library(ggplot2)

avg <- c(avg2016,avg2017,avg2018,avg2019)
sd <- c(sd2016,sd2017,sd2018,sd2019)

# preparing the data
measures <- rep(c("Average","Standard Deviation"),4)
year<- c(2016,2016,2017,2017,2018,2018,2019,2019)
values <- c(avg2016,sd2016,avg2017,sd2017,avg2018,sd2018,avg2019,sd2019)
comp <- data.frame(year,measures,values)
print(comp)

# use the ggplot function to plot the bar chart
# if you explicitly say stat = "identity" in geom_bar()
# you're telling ggplot2 to skip the aggregation and that you'll provide
the y values.
ggplot(comp,aes(fill = measures,y = values,x = year)) +
geom_bar(position="dodge",stat="identity") +
ggtitle("Comparison of Average and Standard Deviation")
```



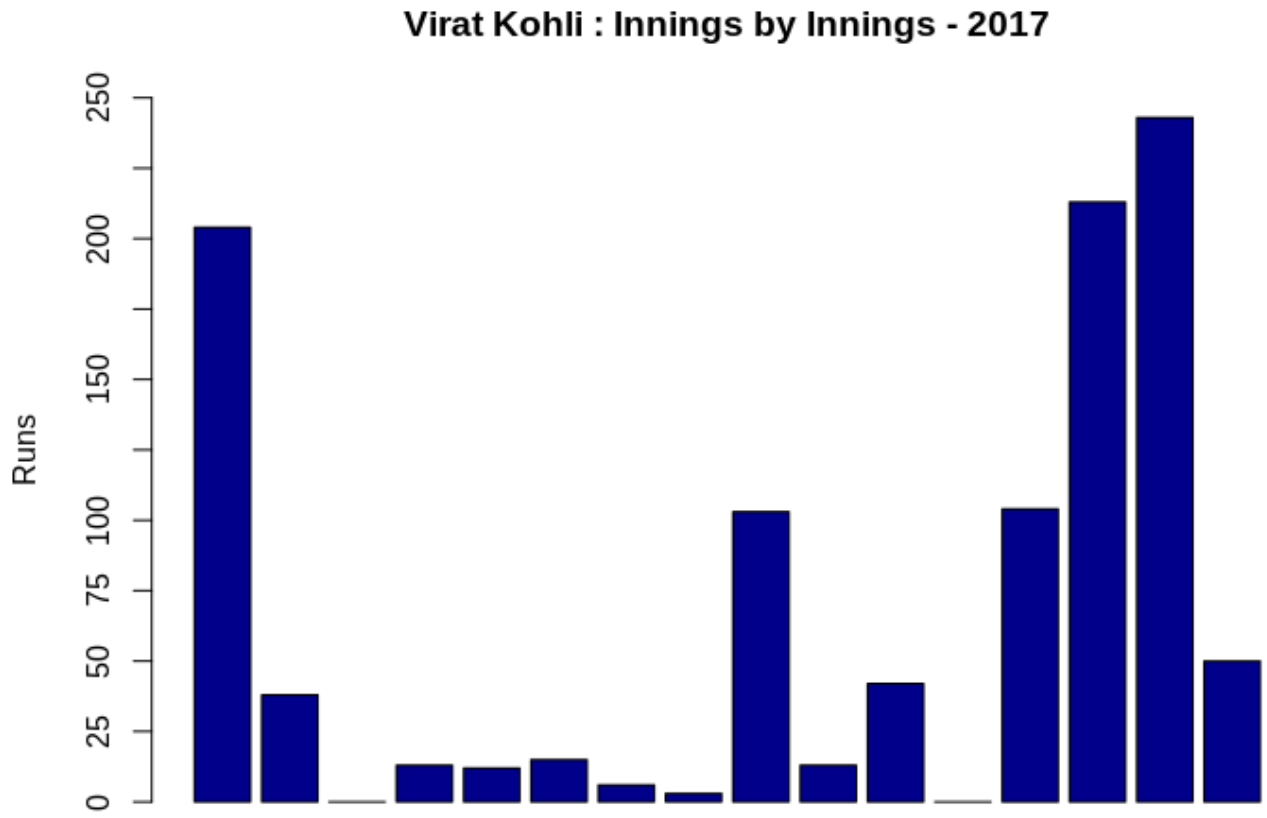
Innings by innings bar plot for the year 2016

```
barplot(runs2016$Runs,yaxp=c(0, 250, 10),ylim = c(0,250),
        main = "Virat Kohli : Innings by Innings - 2016",ylab = "Runs",col
= "Dark Blue")
```



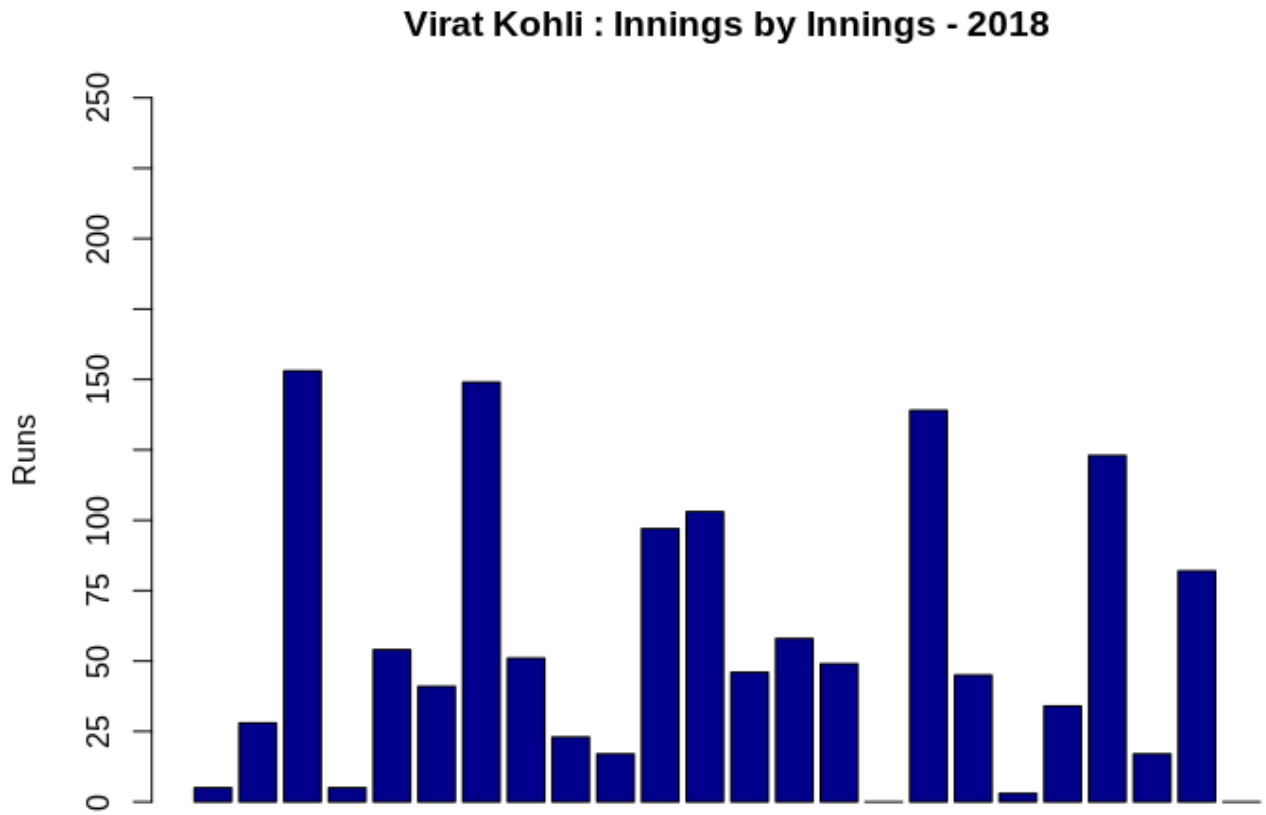
Innings by innings bar plot for the year 2017

```
barplot(runs2017$Runs,yaxp=c(0, 250, 10),ylim = c(0,250),
        main = "Virat Kohli : Innings by Innings - 2017",ylab = "Runs",col
= "Dark Blue")
```



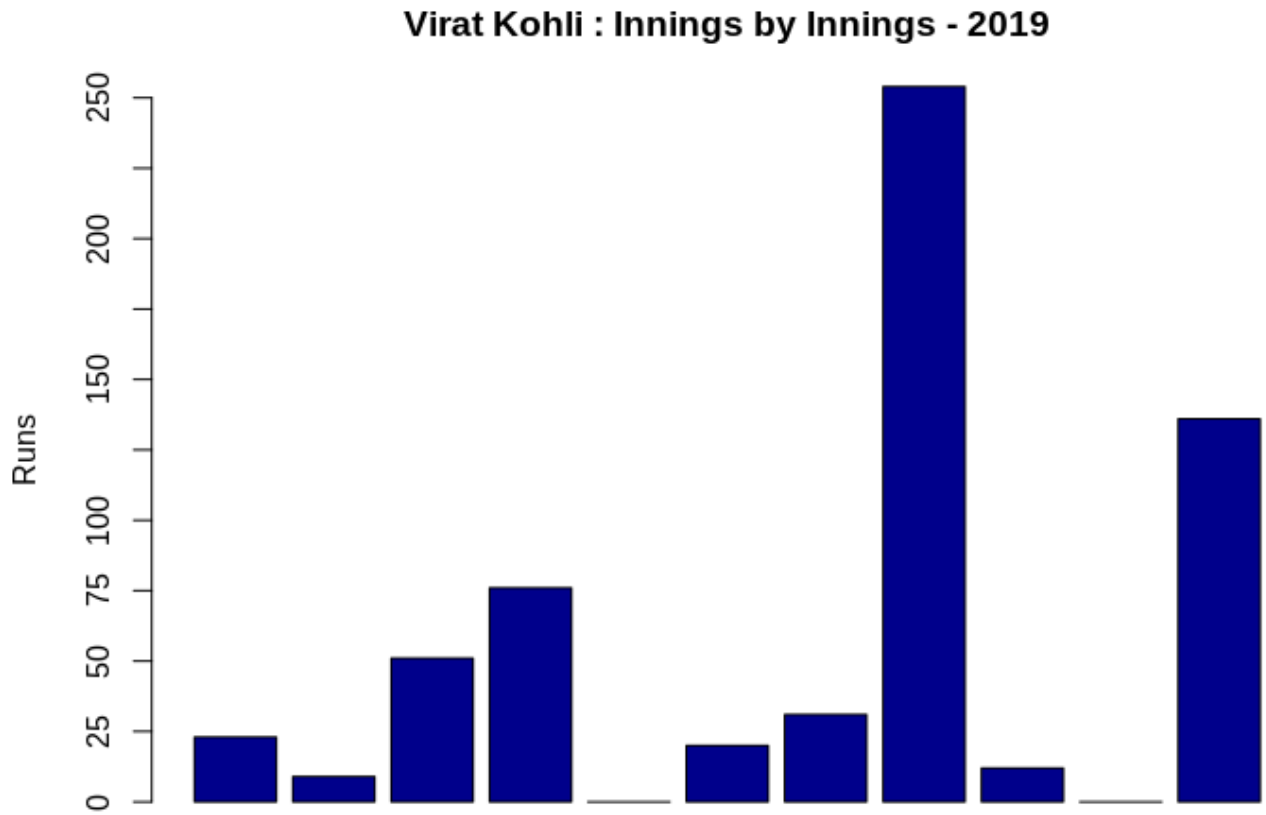
Innings by innings bar plot for the year 2018

```
barplot(runs2018$Runs,yaxp=c(0, 250, 10),ylim = c(0,250),
        main = "Virat Kohli : Innings by Innings - 2018",ylab = "Runs",col
= "Dark Blue")
```



Innings by innings bar plot for the year 2019

```
barplot(runs2019$Runs,yaxp=c(0, 250, 10),ylim = c(0,250),
        main = "Virat Kohli : Innings by Innings - 2019",ylab = "Runs",col
= "Dark Blue")
```



Conclusion

Using the various statistical methods applied above, it can be concluded that 2018 was Virat Kohli's best year. This is because he had a high average of 50+ in that year and the standard deviation for that year was the lowest among the four. This means that he was most consistent in the year 2018. From the innings by innings bar charts plotted for every year, it is observed that there were a lot of scores between 50-100, which is the hallmark of a consistent batsman in cricket.