# Blog 3: Running an ANOVA

*Michael O'Donnell*

*September 28, 2020*

In Blog Two I explored the Multiple Linear Regression. Now, in this Blog I will dive into Multiple Regression with a categorical variable, an ANOVA. This will take measure a categorical variable's effect on the response variable.

To look at an ANOVA in R, let's use help

```r
help(aov)
```

```
## starting httpd help server ... done
```

Now, start by loading a dataset This dataset contains data for all NBA teams from 2014-2018

```r
nbaData <- read.csv("data/nba_data.csv")
colnames(nbaData)[1] <- "Team"

head(nbaData, 3)
```

```
##             Team Season SeasonType Win Loss MatchCount WinPercentage
## 1  Atlanta Hawks   2018        REG  28   53         81     0.3456790
## 2 Boston Celtics   2018        REG  49   33         82     0.5975610
## 3  Brooklyn Nets   2018        REG  42   40         82     0.5121951
##      Pts OppPts   Pace OffEff DefEff EFgPercentage OppEFgPercentage
## 1 112.93 119.21 103.46 108.34 114.73         0.521            0.541
## 2 112.39 107.95  98.97 112.98 108.22         0.534            0.514
## 3 112.24 112.32 100.30 110.23 110.23         0.520            0.512
##   TsPercentage OppTsPercentage RebRate EffPts OppEffPts FastBreakPts
## 1        0.555           0.580   50.07 125.25    138.43        15.26
## 2        0.567           0.550   49.25 132.42    119.59        16.24
## 3        0.556           0.548   50.18 122.98    127.00        11.62
##   OppFBPts PointsInPaint OppPointsInPaint PointsOffTO OppPointsOffTO
## 1    16.51         51.19            49.36       21.14          16.88
## 2    13.17         44.78            45.93       14.82          18.12
## 3    11.83         48.76            51.20       17.35          15.38
##   SecondChancePTS OppSecondChancePTS PersonalFoulsPG OppPersonalFoulsPG
## 1           14.11              14.51          23.519             22.124
## 2           12.48              13.52          21.500             22.037
## 3           13.82              14.40          20.354             19.537
##   ShootingFoulsPG ShootingFoulsDrawnPG LessThnEightFeedUsage
## 1          14.889               12.642                 43.55
## 2          12.268               13.415                 43.45
## 3          12.134               10.549                 36.19
##   EightToSixteenFeedUsage SixteenToTwentyFourFeetUsage
## 1                   11.46                         4.80
## 2                   11.46                         4.89
## 3                   14.82                        10.90
##   TwentyFourPlusFeetUsage AvgShotDistance OppAvgShotDistance
## 1                   39.91           13.06              13.34
## 2                   39.96           13.18              12.89
## 3                   38.00           14.00              13.49
##   AvgMadeShotDistance OppMadeAvgShotDis
## 1               10.34             10.75
## 2               10.70             10.45
## 3               11.64             10.85
```

For this analysis, we will test whether the Season (2014-2018) has any impact on Points in the Paint (PointsInPaint). Y: PointsInPaint X1: SeasonType

Run an ANOVA (first variable in Y (response))

```
model1 <- aov(PointsInPaint ~ Season, nbaData)

summary(model1)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)    
## Season        1    910     910    64.9 5.65e-14 ***
## Residuals   212   2972      14                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the above summary, Season has a statistically significant impact on Points in the Paint. The p-value is far below the significance value of 0.05 and the F value is large.

To view the descriptive statistics by Seaon, we can use the psych library:

```
library(psych)
describeBy(nbaData$PointsInPaint, nbaData$Season)
```

```
## 
##  Descriptive statistics by group 
## group: 2014
##    vars  n  mean   sd median trimmed  mad   min   max range  skew kurtosis
## X1    1 46 42.03 3.51  42.25   42.14 3.98 33.37 49.06 15.69 -0.33    -0.37
##      se
## X1 0.52
## -------------------------------------------------------- 
## group: 2015
##    vars  n mean   sd median trimmed mad min   max range skew kurtosis   se
## X1    1 46 41.8 3.57  41.87   41.77 3.2  34 50.34 16.34 0.09    -0.24 0.53
## -------------------------------------------------------- 
## group: 2016
##    vars  n  mean   sd median trimmed  mad   min   max range  skew kurtosis
## X1    1 46 43.16 3.33  43.09    43.2 3.42 32.78 49.88  17.1 -0.36     0.68
##      se
## X1 0.49
## -------------------------------------------------------- 
## group: 2017
##    vars  n  mean sd median trimmed  mad  min   max range skew kurtosis
## X1    1 46 45.01  4  44.28   44.88 3.95 37.6 54.89 17.29 0.38    -0.41
##      se
## X1 0.59
## --------------------------------------------------------
## group: 2018
##    vars  n  mean   sd median trimmed  mad  min   max range skew kurtosis
## X1    1 30 48.58 3.87  49.11    48.4 3.97 42.1 58.35 16.25  0.3    -0.38
##      se
## X1 0.71
```

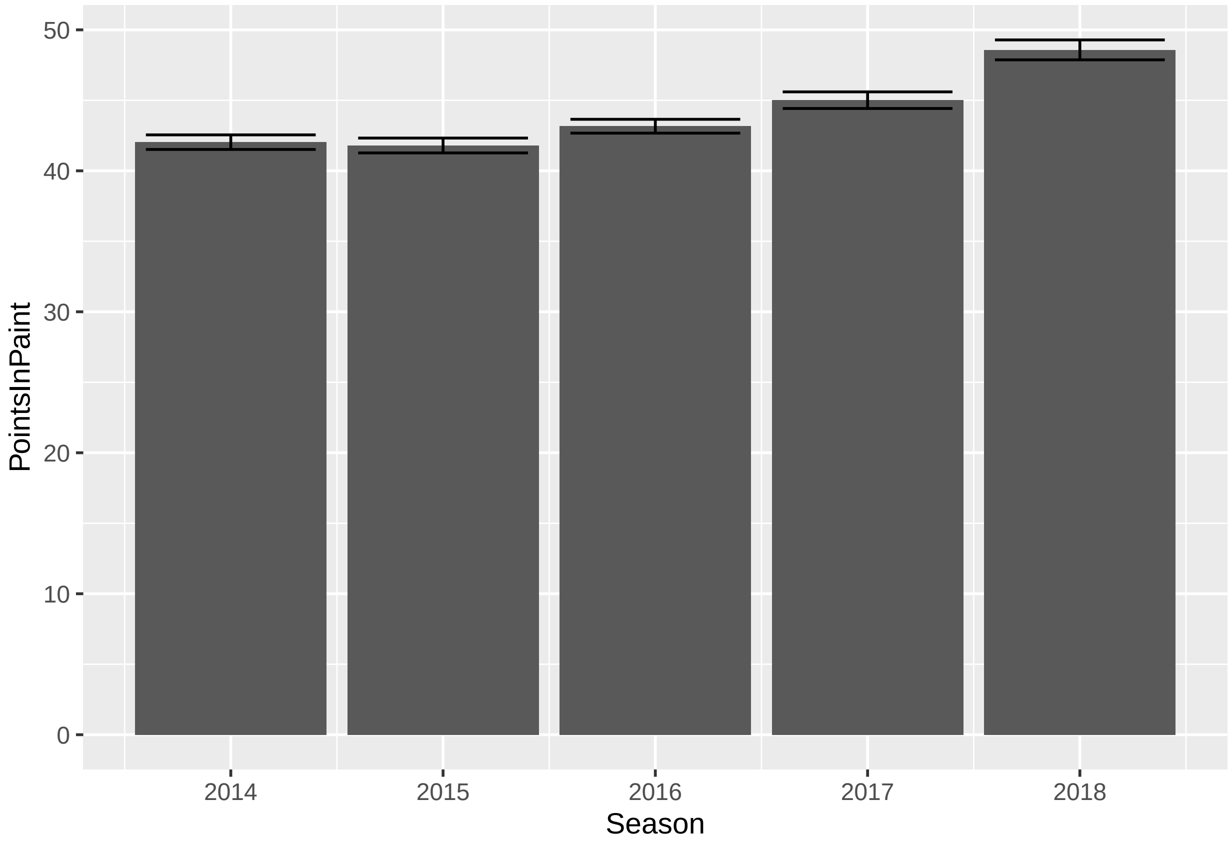
To visualize the data above, we can use ggplot to graph the Points in the Paint by Season

```
library(ggplot2)
```

```
##
## Attaching package: 'ggplot2'
```

```
## The following objects are masked from 'package:psych':
##
##     %+%, alpha
```

```
ggplot(nbaData,aes(y=PointsInPaint, x=Season))+
  stat_summary(fun="mean", geom="bar",position="dodge")+
  stat_summary(fun.data = mean_se, geom = "errorbar", position="dodge",width=.8)
```


PointsInPaint
Season