Final Project(NBA)

Shuheng Zhao & Ziwei Zeng

5/01/2017

NBA is the National Basketball Association. It is the major men's professional basketball league in North America, and is widely considered to be the premier men's professional basketball league in the world. The NBA consists of 30 teams, 29 in the United States and 1 in Canada. We know that small ball lineups are now popular in the league. People always assume that having a small ball lineup will give the team a better socring ability than having a tradition lineup. As we know, small ball lineups are good at assists since everyone on the lineup has the ablity of passing the ball. Also, more passing usually leads to more turnovers. In addition, small ball lineups are good at attacting the basket and drawing fouls, therefore, free throws are usually going to be increased when we have a small ball lineup. Finally, we know that small ball lineups are not good defensive lineups. Having a small ball lineup will let the team have less blocks and steals.

In this project, we would like to analysis how a team's offense can be affected by the "small ball lineup" factors. We would like to find the relationships between the points socred and these factors to check whether a small ball lineup does actually give a team a better scoring ability. We will want to derive an equation.

The report of the work divided into two parts Exploring descriptive analysis and statistical modeling.

Part I EDA

Before we start perform exploring descriptive analysis, we first give overviews of the data:

```
library(readr)
library(dplyr)
library(ggplot2)
library(knitr)
data = read_csv("D:/DA101file/Final/1516NBAcleaned.csv")
data = data.frame(data)
#glimpse
glimpse(data)
```

```
## Observations: 568,333
## Variables: 47
## $ game_id        <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1,...
## $ data_set       <chr> "2015-2016 Regular Season", "2015-2016 Regula
r ...
## $ date           <date> 2015-10-27, 2015-10-27, 2015-10-27, 2015-10-
27...
## $ a1             <chr> "Marcus Morris", "Marcus Morris", "Marcus Mor
ri...
## $ a2             <chr> "Ersan Ilyasova", "Ersan Ilyasova", "Ersan Il
ya...
## $ a3             <chr> "Andre Drummond", "Andre Drummond", "Andre Dr
um...
## $ a4             <chr> "Kentavious Caldwell-Pope", "Kentavious Caldw
el...
## $ a5             <chr> "Reggie Jackson", "Reggie Jackson", "Reggie J
ac...
## $ h1             <chr> "Kent Bazemore", "Kent Bazemore", "Kent Bazem
or...
## $ h2             <chr> "Paul Millsap", "Paul Millsap", "Paul Millsap
",...
## $ h3             <chr> "Al Horford", "Al Horford", "Al Horford", "Al
 H...
## $ h4             <chr> "Kyle Korver", "Kyle Korver", "Kyle Korver",
"K...
## $ h5             <chr> "Jeff Teague", "Jeff Teague", "Jeff Teague",
"J...
## $ period         <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1,...
## $ away_score     <int> 0, 0, 0, 0, 0, 2, 2, 4, 4, 4, 4, 4, 4, 4, 5,
5,...
## $ home_score     <int> 0, 0, 0, 0, 0, 0, 2, 2, 2, 2, 2, 2, 4, 4, 4,
4,...
## $ remaining_time <int> 720, 720, 701, 699, 697, 681, 660, 644, 627,
62...
## $ elapsed        <int> 0, 0, 19, 21, 23, 39, 60, 76, 93, 95, 107, 10
9,...
```

```
## $ play_length    <chr> "00:00:00", "00:00:00", "00:00:19", "00:00:02
",...
## $ play_id        <int> 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 13, 14,
 1...
## $ team           <chr> NA, "DET", "DET", "ATL", "ATL", "DET", "ATL",
 "...
## $ OT             <chr> "1", "ATL", "ATL", "DET", "DET", "ATL", "DET",
 ...
## $ event_type     <chr> "start of period", "jump ball", "miss", "rebo
un...
## $ assist         <chr> NA, NA, NA, NA, NA, "Andre Drummond", "Kyle K
or...
## $ away           <chr> NA, "Andre Drummond", NA, NA, NA, NA, NA, NA,
 N...
## $ home           <chr> NA, "Al Horford", NA, NA, NA, NA, NA, NA, NA,
 N...
## $ block          <chr> NA, NA, "Al Horford", NA, NA, NA, NA, NA, NA,
 N...
## $ entered        <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N
A,...
## $ left           <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N
A,...
## $ num            <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N
A,...
## $ opponent       <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N
A,...
## $ outof          <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N
A,...
## $ player         <chr> NA, "Al Horford", "Andre Drummond", "Kent Baz
em...
## $ points         <int> NA, NA, 0, NA, NA, 2, 2, 2, 0, NA, 0, NA, 2,
NA...
## $ possession     <chr> NA, "Ersan Ilyasova", NA, NA, NA, NA, NA, NA,
 N...
## $ reason         <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N
A,...
## $ result         <chr> NA, NA, "missed", NA, NA, "made", "made", "ma
de...
## $ steal          <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N
A,...
## $ type           <chr> "start of period", "jump ball", "Driving Layu
p"...
## $ shot_distance  <int> NA, NA, 2, NA, NA, 13, 12, 9, 20, NA, 8, NA,
12...
## $ original_x     <int> NA, NA, -17, NA, NA, 117, 76, -68, -117, NA,
-7...
## $ original_y     <int> NA, NA, -6, NA, NA, 67, 95, 51, 164, NA, 31,
NA...
## $ converted_x    <dbl> NA, NA, 26.7, NA, NA, 13.3, 32.6, 31.8, 13.3,
 N...
```

```
## $ converted_y    <dbl> NA, NA, 4.4, NA, NA, 11.7, 79.5, 10.1, 72.6, 
NA...
## $ description    <chr> NA, "Jump Ball Horford vs. Drummond: Tip to I
ly...
## $ HT             <chr> "ATL", "ATL", "ATL", "ATL", "ATL", "ATL", "AT
L"...
## $ AT             <chr> "DET", "DET", "DET", "DET", "DET", "DET", "DE
T"...
```

Then we summarize the performances of small ball lineup factors on team level, the table is given as follows:

```r
#Group_by, summarise, sub
data$team<-sub("^","15",data$team)
table1<- data%>%  group_by(team)%>%   summarise(points=sum(points,na.rm
=TRUE),OR=sum(type=="rebound offensive"))
table1<-mutate(table1,PTSPG=points/82, ORPG=OR/82)
#Remove NAs
table2<- data%>%
  group_by(team)%>%
  summarise(Steals = sum(!is.na(steal)), Blocks = sum(!is.na(block)), A
ssist = sum(!is.na(assist)), Turnovers=sum(event_type=="turnover"),TwoM
ade=length(which(points==2)),ThreeMade=length(which(points==3)),TotalFG
=sum(event_type=="shot")+sum(event_type=="miss"),EFGP=(TwoMade+1.5*Thre
eMade)/TotalFG,FreeThrow=sum(event_type=="free throw"),FTMade=length(wh
ich(result=="made"))-length(which(event_type=="shot")))
#left join
table3<-left_join(table2,table1,by=c("team"))
#Filter
table3 <- table3%>%   filter(!is.na(team))
#mutate
table3 <-mutate(table3,PTSPG=points/82)

#check first 6 and last 6 rows
tb = rbind(table3[1:6, ], table3[(nrow(table3) - 5):nrow(table3),])
kable(tb)
```

team	Steals	Blocks	Assist	Turnovers	TwoMade	ThreeMade	TotalFG	EFGP	FreeThrow	FTMade	points	OR	PTSPG	ORPG
15ATL	706	411	2100	1191	2353	815	6923	0.5164668	1638	1282	8433	679	102.84146	8.280488
15BKN	720	430	1829	1174	2605	531	6920	0.4915462	1699	1286	8089	862	98.64634	10.512195

15BOS	622	450	1981	1104	2499	717	7318	0.48845531	1929	1520	8669	950	105.71951	11.585366
15CHA	553	448	1778	973	2163	873	6922	0.50166614	1941	1534	8479	734	103.40244	8.951220
15CHI	655	465	1870	1093	2514	651	7170	0.48682001	1720	1354	8335	907	101.64634	11.060976
15CLE	590	362	1861	1056	2291	880	6888	0.52424451	1783	1333	8555	873	104.32927	10.646342
15POR	630	424	1748	1154	2303	864	7040	0.51122216	1889	1424	8622	948	105.14634	11.560976
15SAC	722	437	2009	1274	2623	660	7083	0.51009946	2089	1514	8740	868	106.58537	10.585366
15SAS	592	317	2010	1028	2719	570	6797	0.52582002	1672	1342	8490	770	103.53659	9.390244
15TOR	534	442	1536	994	2298	708	6669	0.50382237	2190	1702	8422	836	102.70732	10.195122
15UTA	652	385	1554	1162	2263	694	6593	0.50113376	1885	1402	8010	880	97.68293	10.731707
15WAS	661	356	2005	1138	2529	709	7033	0.51080862	1849	1350	8534	743	104.07317	9.060976

It can be found that each team has different performances in the various ascepts. And we check the types frequency and show the top 10 most types:

```r
#table
tb = table(data$type)
tb10 = sort(tb, decreasing = T)[1:10]
tb10
```

```
##
##           Jump Shot rebound defensive               sub rebound offens
ive
##               98820             82015             55030              25
621
## Free Throw 1 of 2 Free Throw 2 of 2            s.foul       team rebo
und
##               24179             24162             22873              20
319
##               Layup           unknown
##               18143             16825
```

Now we check the relationships between ORPG and PTSPG use Bivariate Regression:

```r
#summary
reg<-lm(PTSPG~ORPG,table3)
summary(reg)
```

```
##
## Call:
## lm(formula = PTSPG ~ ORPG, data = table3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.4510 -2.3515  0.0182  1.8618 12.4105
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  98.3637     6.8274  14.407 1.78e-14 ***
## ORPG          0.4136     0.6521   0.634    0.531
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.828 on 28 degrees of freedom
## Multiple R-squared:  0.01417,    Adjusted R-squared:  -0.02104
## F-statistic: 0.4023 on 1 and 28 DF,  p-value: 0.531
```

The 95% confindence interval for the estimated slope is:

```r
#Confidence intervals
confint(reg)
```

```
##                  2.5 %     97.5 %
## (Intercept) 84.3783932 112.348984
## ORPG        -0.9221246   1.749362
```

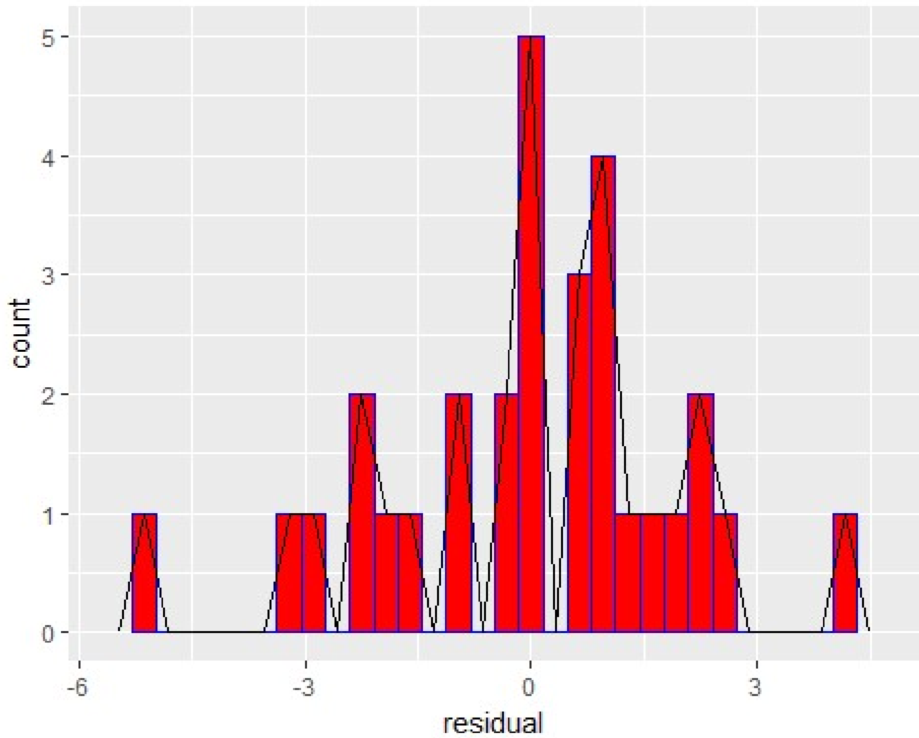
It can be found the predictor ORPG is significant with p value less than 0.05, and then we check the relationships between ORPG, EFGP and PTSPG use Multivariate regression:

```
reg2<-lm(PTSPG~ORPG+EFGP,table3)
summary(reg2)
```

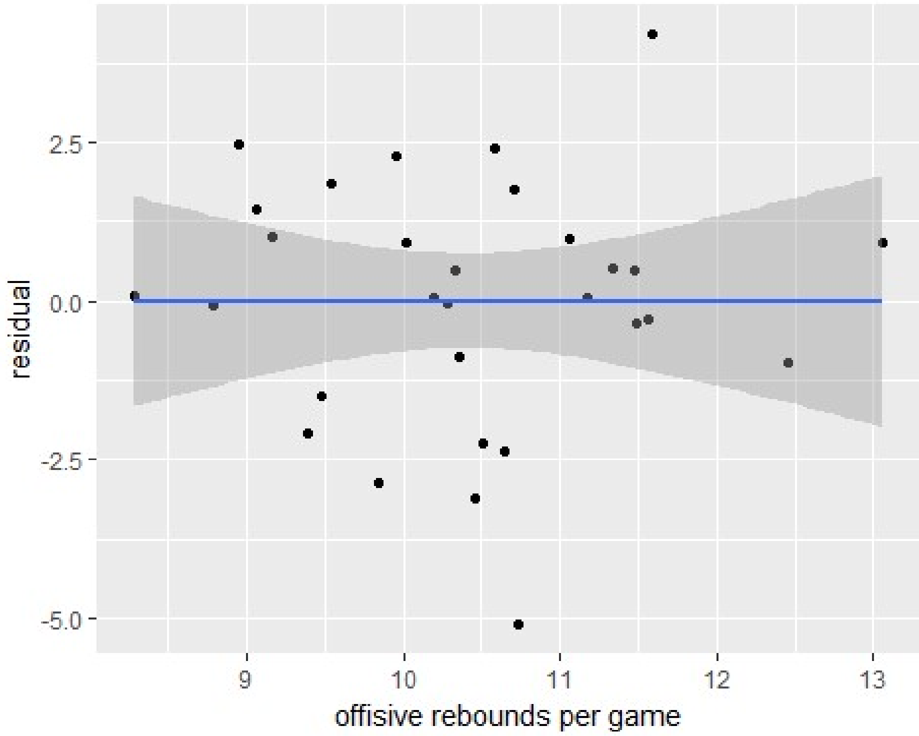
```
##
## Call:
## lm(formula = PTSPG ~ ORPG + EFGP, data = table3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1004 -0.9496  0.0620  0.9979  4.2079
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.907     11.627   0.336   0.7394
## ORPG           1.094      0.354   3.090   0.0046 **
## EFGP         173.876     20.344   8.547 3.68e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.025 on 27 degrees of freedom
## Multiple R-squared:  0.734,  Adjusted R-squared:  0.7143
## F-statistic: 37.24 on 2 and 27 DF,  p-value: 1.725e-08
```

The output also show the both of the predictors are significant, then we check the residuals of the model:

```
res = table3$PTSPG  - predict(reg2)
table3<-mutate(table3,resid2=res)
table3 <-rename(table3, residual = resid2)
#Histogram, freqpoly
table3%>%
  ggplot(aes(residual))+
  geom_histogram(colour="blue",fill="red",bins=30) +  geom_freqpoly()
```




```
#scatter
table3%>%
  ggplot(aes(ORPG,residual))+
  geom_point()+geom_smooth(method="lm")+
  xlab("offisive rebounds per game")
```



Shapiro-wilk test to test the noramlity of the residuals:

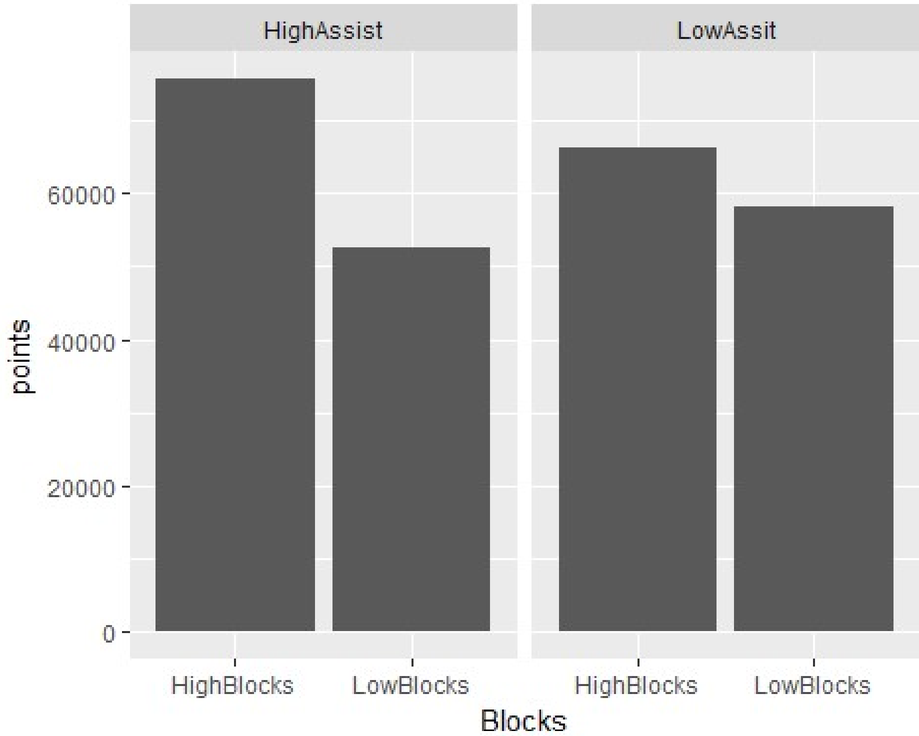
```
shapiro.test(res)

##
##  Shapiro-Wilk normality test
##
## data:  res
## W = 0.97588, p-value = 0.7086
```

The p value is 0.7086 > 0.05 which means the normality is true for the regression.

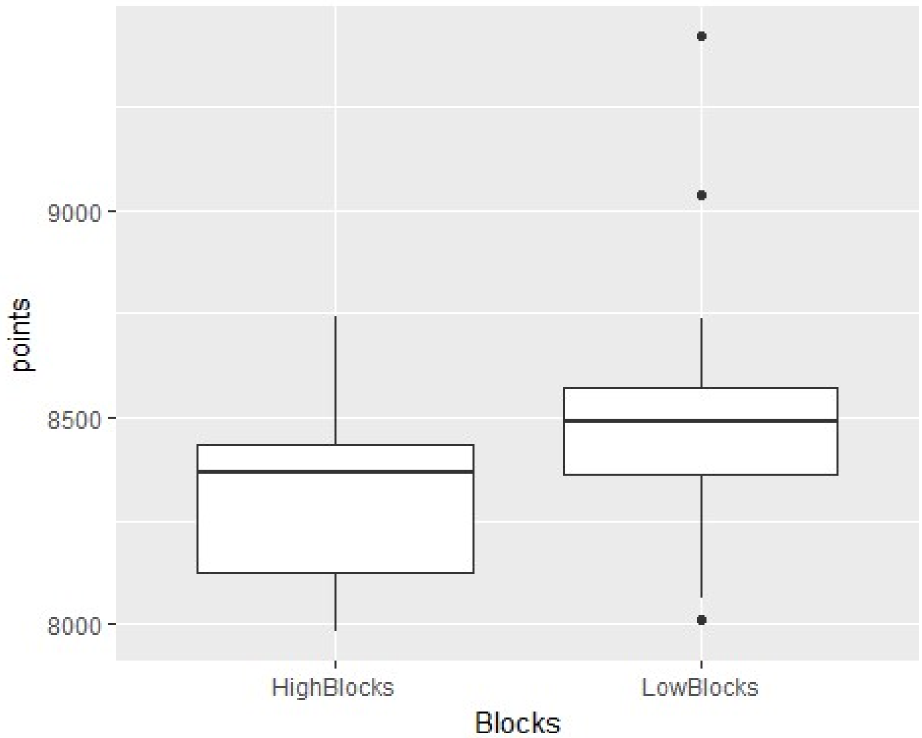
Next. We seperate the teams by blocks and assists. We would like to find the points socred by teams that have blocks above and below the average, and assists above and below the average.

```
#Bar
table3 = data.frame(table3)
table3 <- mutate(table3, HighAssist = if_else(Assist > mean(Assist), "H
ighAssist", "LowAssit"), Highblocks = if_else(Blocks > mean(Blocks), "H
ighBlocks", "LowBlocks"))
ggplot(table3, aes(Highblocks,points)) + geom_bar(stat = "identity") +
facet_wrap(~HighAssist) + xlab("Blocks")
```



The bar plots show there are difference between the high block teams and high assist teams in points. And we check the average points between the two groups of high blocks and low blocks teams:

```
ggplot(table3, aes(Highblocks,points)) + geom_boxplot()+xlab("Blocks")
```



So low block teams have an average higher points which means they are more focus on getting points.

Finally, we perform some formal statistical tests including Independent samples T-test, One-sample T-test and Correlation test:

First, we check if there is correlation between Two Made shot and Three MAde ones, the result is:

```
#select
table4 = select(table3, c(TwoMade, ThreeMade))
with(table4, cor.test(TwoMade, ThreeMade))

## 
##  Pearson's product-moment correlation
## 
## data:  TwoMade and ThreeMade
## t = -4.8764, df = 28, p-value = 3.888e-05
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.8342646 -0.4199347
## sample estimates:
##        cor 
## -0.6776783
```

The p value is much lower than 0.05, so it means we should reject the independent of the two variables.

For One-sample T-test, we want to t test if the average shot distance is less than 12.

```
with(data, t.test(shot_distance, mu = 12, alternative = "greater"))
```

```
##
##  One Sample t-test
##
## data:  shot_distance
## t = 43.071, df = 208050, p-value < 2.2e-16
## alternative hypothesis: true mean is greater than 12
## 95 percent confidence interval:
##  12.91064      Inf
## sample estimates:
## mean of x
##   12.9468
```

So the p value is less than 0.05, we should reject H0 and conclude the average shot distance is greater than 12.

For independent samples T-test, we want to test if the away score and home score are equal.

```
with(data, t.test(away_score, home_score))
```

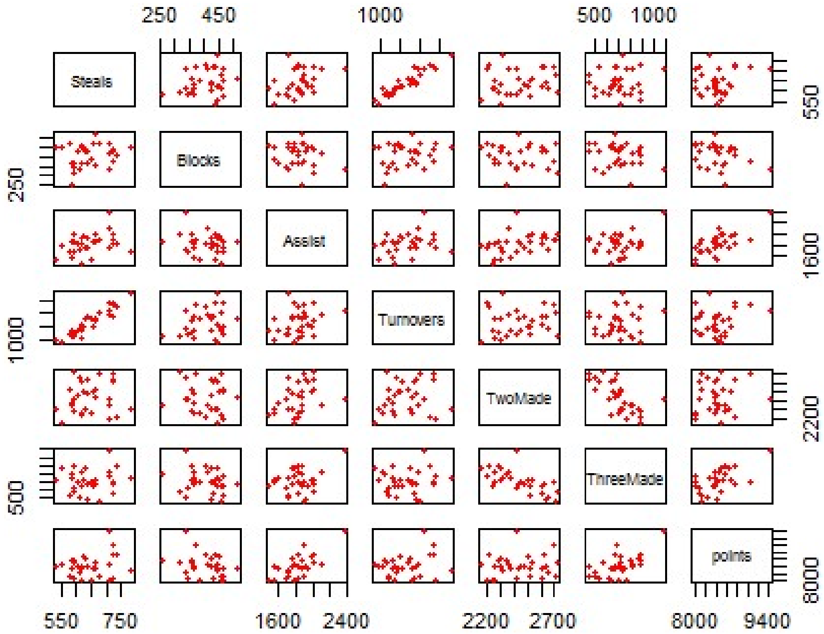
```
##
##  Welch Two Sample t-test
##
## data:  away_score and home_score
## t = -25.903, df = 1136000, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.614309 -1.387195
## sample estimates:
## mean of x mean of y
##  52.04732  53.54807
```

So the p value is less than 0.05, we should reject H0 and conclude the two scores are not equal.

Part II Modeling

In this section, based on the exploring results from Part I, we are aim to find relationships among the variables via linear regression model, in our case, we want to find the relationships among the points with steals, blocks, assist , Turnovers and FreeThrow, we would like to regress points on the 4 factors and based on the model, we can not only interpret the predictors but also can predict the response variable in the future when given the predictors.

First, we show an overall of relationships among the variables using scatter plot matrix among the response points and the 5 predictors:



This plot shows all the relationship among the variables. By using this plot, we don't have to plot multiple times to observe the relationships. From this plot, we can't observe that there are any direct relationships between any specific variables besides steals and turnovers.

Therefore, we need to build a model which decribe the relationshiop between points socred and all the factors above, the model is:

$$Points = \beta_0 + \beta_1 Steals + \beta_2 Blocks + \beta_3 Assist + \beta_4 Turnovers + \beta_5 FreeThrow$$

And the result of the model is:

```
## Residuals:
##      Min       1Q   Median       3Q      Max
```

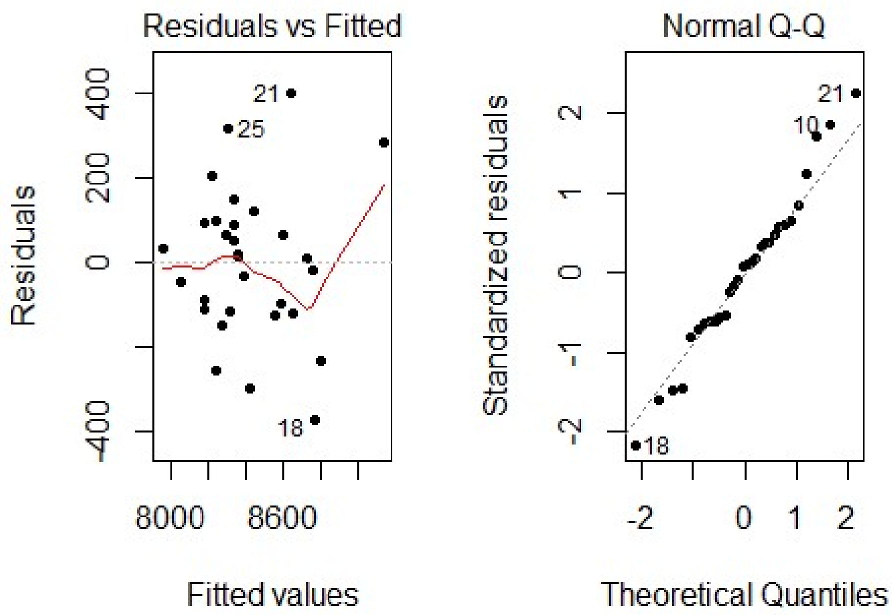
```
## -368.86 -112.60   14.76   95.15  401.39
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5012.4582   730.6464   6.860 4.28e-07 ***
## Steals        -1.2889     1.8456  -0.698  0.49166
## Blocks        -1.0504     0.6798  -1.545  0.13537
## Assist         1.3791     0.2358   5.850 4.94e-06 ***
## Turnovers      0.7083     1.1413   0.621  0.54073
## FreeThrow      0.7001     0.2190   3.197  0.00387 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 193.8 on 24 degrees of freedom
## Multiple R-squared:  0.6778, Adjusted R-squared:  0.6107
## F-statistic:  10.1 on 5 and 24 DF,  p-value: 2.674e-05
```

The p-value is small (2.674e-05) so the null hypothesis that the true slope is 0 is accepted. Thus, the variables combination has strong relationship with points made, and so these variables are significant. From the analysis, we can get the coefficients of the linear model. Therefore, we can have an equation shows the relationship between the factors and the points socred. We also find that the R squred value is 0.6, which is pretty large. Therefore, this linear model fits pretty well.

So the model estimated is:

$$Points = 5012.4582 - 1.2889Steals - 1.0504Blocks + 1.3791Assist + 0.7083Turnovers + 0.7001FreeThrow$$

Now we perform model diagnostic plots to check the assumptions of the above model, the plots are shown as below:



"

These two plots test the normality of the linear model. The left residuals plot shows that the spread of points do not change across the x-axis which means the constant variance is satisfied and there is no special curve which means linearity is true and the right normal qq plot shows the points fit the straight line quite well which means the normality assumption is true, so our model is valid.

Therefore, we can conclude that the relationship between the points socred and the "small ball lineup" factors is the following equation:

$$Points = 5012.4582 - 1.2889 Steals - 1.0504 Blocks + 1.3791 Assist + 0.7083 Turnovers + 0.7001 FreeThrow$$

From this equation, we can see that the coefficient before steals and blocks are negative numbers. Therefore, having more steals and blocks leads to scoring less points. Since having a traditional lineup will give the team a better defense, in other words, more steals and blocks. We can interpret this result as having a tradition lineup leads to scoring less points. Then, the coefficients before Assists, Turnovers and Freethrows are all positive. Therefore, having more assists, turnovers and freethrows leads to scoring more points. And since small ball lineups bring more of those stats than the traditional lineups, having a small ball lineup will let the team to score more points.

In conclusion, having a small ball lineup does give a team a better scoring ability than having a traditional lineup.