NBA Analytics - Data Science Project

Part 1:

Introduction:

During the first outbreak of COVID-19 in March, one of the world's favorite leagues, the National Basketball Association (NBA) was forced to halt all league games until further notice. Few months down the road, the NBA approved a competitive format for the comeback of the 2019-20 season with 22 teams returning to play in Orlando, Florida. Of course, with the new and improved format of the season, with an exclusion of 8 teams in the remainder of the season, some teams may or may have performed differently, which is why I chose to explore and analyze consistent and structured datasets from the regular season and the NBA Restart.

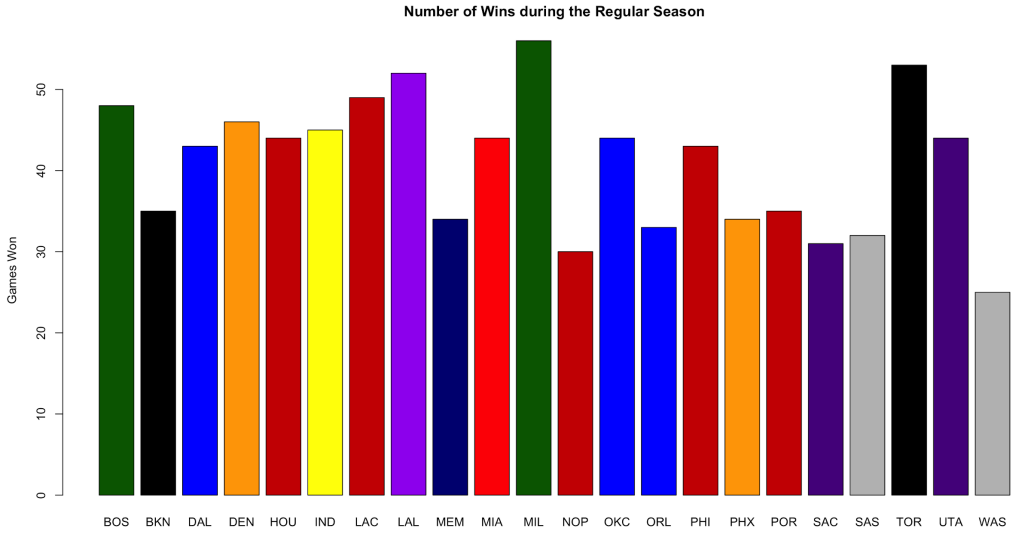
Exploration of the Data:

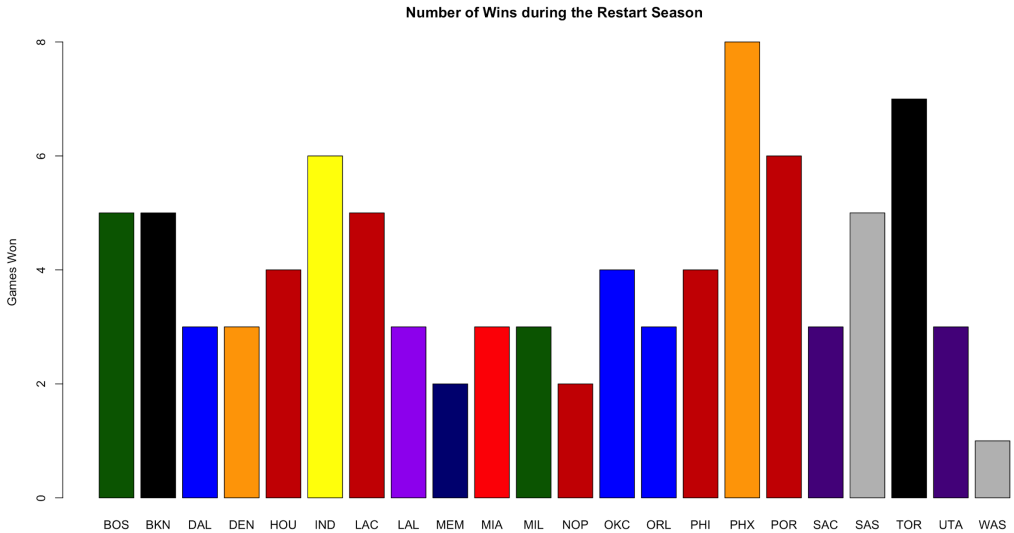
```
> head(reg)
      TEAM CONF  DIVISION GP PTS.GM aPTS.GM PTS.DIFF  PACE  OEFF  DEFF EDIFF   SOS rSOS   SAR CONS    A4F  W  L  WIN. eWIN. pWIN.    ACH STRK
1   Boston East  Atlantic 72  113.7   107.3      6.4  99.4 113.3 107.0   6.3 -0.53    0  5.77 12.0  0.020 48 24 0.667 0.701 0.711 -0.034  L 1
2 Brooklyn East  Atlantic 72  111.8   112.3     -0.5 101.4 109.0 109.5  -0.5 -0.25    0 -0.75 13.8  0.043 35 37 0.486 0.486 0.484  0.000  L 1
3   Dallas West Southwest 75  117.0   112.1      4.9  99.3 116.7 111.7   5.0 -0.10    0  4.90 14.3  0.039 43 32 0.573 0.636 0.661 -0.063  L 2
4   Denver West Northwest 73  111.3   109.2      2.1  97.1 113.1 111.0   2.1  0.01    0  2.11 12.3  0.001 46 27 0.630 0.568 0.569  0.062  L 3
5  Houston West Southwest 72  117.8   114.8      3.0 103.7 113.0 110.1   2.9 -0.35    0  2.55 14.8 -0.013 44 28 0.611 0.578 0.599  0.033  L 3
6  Indiana East   Central 73  109.4   107.5      1.9  98.9 110.0 108.0   2.0 -0.48    0  1.52 14.0  0.013 45 28 0.616 0.557 0.563  0.059  W 2
> head(restart)
      TEAM CONF  DIVISION GP PTS.GM aPTS.GM PTS.DIFF  PACE  OEFF  DEFF EDIFF   SOS rSOS   SAR CONS    A4F W L  WIN. eWIN. pWIN.    ACH STRK
1   Boston East  Atlantic  8  118.9   111.5      7.4 101.1 116.1 108.9   7.2 -0.53    0  6.67 12.0  0.080 5 3 0.625 0.726 0.744 -0.101  L 1
2 Brooklyn East  Atlantic  8  119.9   119.9      0.0 103.7 115.6 115.6   0.0 -0.25    0 -0.25 13.8  0.002 5 3 0.625 0.500 0.500  0.125  L 1
3   Dallas West Southwest  8  122.5   126.8     -4.3 101.0 116.7 120.8  -4.1 -0.10    0 -4.20 14.3 -0.035 3 5 0.375 0.388 0.358 -0.013  L 2
4   Denver West Northwest  8  118.5   123.3     -4.8  96.5 118.2 122.9  -4.7  0.01    0 -4.69 12.3 -0.028 3 5 0.375 0.351 0.342  0.024  L 3
5  Houston West Southwest  8  115.3   118.6     -3.3 106.9 106.4 109.5  -3.1 -0.35    0 -3.45 14.8 -0.087 4 4 0.500 0.417 0.391  0.083  L 3
6  Indiana East   Central  8  110.3   108.1      2.2 102.3 107.8 105.7   2.1 -0.48    0  1.62 14.0  0.011 6 2 0.750 0.559 0.572  0.191  W 2
```

Above are the data of the first six teams that participated in both the NBA regular season and the Restart. Both datasets consist of 22 rows, each representing an NBA team. There are 23 columns each representing a statistic describing the specific team. These attributes include the following:

1. TEAM
2. CONF (Conference)
3. DIVISION
4. PTS.GM (Average points per game)
5. aPTS.GM (Average points allowed per game)
6. PTS.DIFF (Points differential)
7. PACE (Estimate of possession per 48 minutes)
8. OEFF (Offensive efficiency)
9. DEFF (Defensive efficiency)
10. EDIFF (Efficiency differential)
11. SOS (Strength of schedule)
12. rSOS (Remaining strength of schedule)
13. SAR (Schedule adjusted rating)
14. CONS (Consistency rating)
15. A4F (Adjusted four factors)
16. W (No. of wins)
17. L (No. of losses)
18. WIN% (Winning percentage)
19. eWIN% (Gaussian expected winning percentage)
20. pWIN% (Projected winning percentage)
21. ACH (Achievement level in terms of wins)
22. STRK (Winning or losing streak

The columns/variables of interest I will be using are average points per game, points differential, efficiency differential, strength of schedule, consistency, winning percentage, expected winning percentage, and achievement level. The points differential variable is the difference, on average, of the total points per game and the total points allowed per game. The efficiency differential follows the same formula; however, it is calculated among the offensive and defensive efficiency points. The strength of schedule variable is represented by the opponent efficiency differential average for the last 5 games in that team's schedule. The consistency rating is solely calculated by the variation of the game-by-game efficiency differential. Furthermore, the winning percentage is a team's net overall point differential rather than points scored and points allowed. The expected percentage shows each point differential translated to 2.7 wins over the course of the season. Finally, the datasets represent the achievement level as a metric based on the difference between expected and actual winning percentage. Positive level indicates overachievement, while negative indicates underachievement.





Above are the barplots of each team before and after the pandemic outbreak, and the number of wins they had. Of course, it is important to note that the total number of games played are different: the average number of games completed by teams in the regular season were 72.5, yet in the Restart, all teams had played a total of 8 games. The max games won during the regular season was 56, done by the Milkwaukee Bucks. However, the max games won during the Restart was 8, by the undefeated Phoenix Suns. The average number of wins during the regular season for those 22 teams was approximately 41 out of 73 games, while the average number of wins for the Restart was 4 out of 8 games.

```
> library(pastecs)
> stat.desc(reg[,c("W")])
     nbr.val     nbr.null       nbr.na          min          max        range          sum
  22.0000000    0.0000000    0.0000000   25.0000000   56.0000000   31.0000000  900.0000000
      median         mean      SE.mean CI.mean.0.95          var      std.dev     coef.var
  43.5000000   40.9090909    1.7897122    3.7219102   70.4675325    8.3944942    0.2051987
> stat.desc(restart[,c("W")])
     nbr.val     nbr.null       nbr.na          min          max        range          sum
  22.0000000    0.0000000    0.0000000    1.0000000    8.0000000    7.0000000   88.0000000
      median         mean      SE.mean CI.mean.0.95          var      std.dev     coef.var
   3.5000000    4.0000000    0.3663320    0.7618291    2.9523810    1.7182494    0.4295623
```

For this portion, I installed the *pastecs* package in order to utilize the *stat.desc()* function in R. I used this to find the quick descriptive statistics about this distribution of wins in the dataset. One thing that needs to be noted is that the standard deviation of wins during the regular season is fairly higher than the Restart, but the number of games played in both datasets are significantly different. Regardless, the regular season had more data that was collected, and the numbers of wins was pretty spread out considering each standard deviation was approximately 8.4 wins away from the mean, 40.91. Because both datasets have means that are drastically different from one another, one useful statistic we can analyze is the coefficient of variation. Because our coefficient of variations, 0.205 and 0.430, are much less than 10, we can say that the data is not that spread out while considering what the standard deviation values tell us.

End of Part 1

Part 2:

Research Question #1:

Does an NBA team's efficiency differentials differ across conferences during the regular season?

One thing special between these two datasets is that they both contain an unequal number of teams from both conferences, 9 from the East, and 13 from the West. Because of these uneven numbers of teams between the two levels, I'd like to test if there is a significant difference in their efficiency levels. The best way to answer this question is by means of an ANOVA test since there are 2 groups of conferences, each with a specific NBA team.

In order to perform this test, we need to validate that each sample is independent, that the population is normally distributed, and most importantly, see if the standard deviations are similar.

```
> sd(reg$EDIFF[reg$CONF == "East"])
[1] 4.298255
> sd(reg$EDIFF[reg$CONF == "West"])
[1] 2.856167
```

For these two categories, I will say the standard deviations are pretty similar and can assume the population is normal. Assuming that the variances are equal across groups, I will first perform the Bartlett
Test to test for equal variances in the population.

Bartlett test:

- H_0: There is homogeneity of variances between efficiency differential across the two conferences
- H_a: There is no homogeneity of variances between efficiency differential across the two conferences

```
> bartlett.test(reg$EDIFF~reg$CONF)

        Bartlett test of homogeneity of variances

data:  reg$EDIFF by reg$CONF
Bartlett's K-squared = 1.5654, df = 1, p-value = 0.2109
```

Because the p-value, 0.2109, is greater than our alpha, 0.05, we fail to reject the null hypothesis that there is homogeneity of variance in the efficiency differential of NBA teams across the two conferences. Additionally, a T-Test for means and ANOVA work well to support Bartlett's test.

Hypotheses for T-test and ANOVA:
- H_0: no difference in the mean efficiency differential across the two conferences
- H_a: difference in the mean efficiency differential across the two conferences

The null hypothesis in this case is that there is no significant difference in the efficiency differentials of NBA teams across different conferences. On the other hand, the alternative hypothesis states that there is a significant difference among these differentials. The ANOVA test will be used to compare the mean efficiency differentials of each team across the Eastern and Western conferences by using the EDIFF and CONF variables. I will also use a two-sample t-test to help support the ANOVA test.

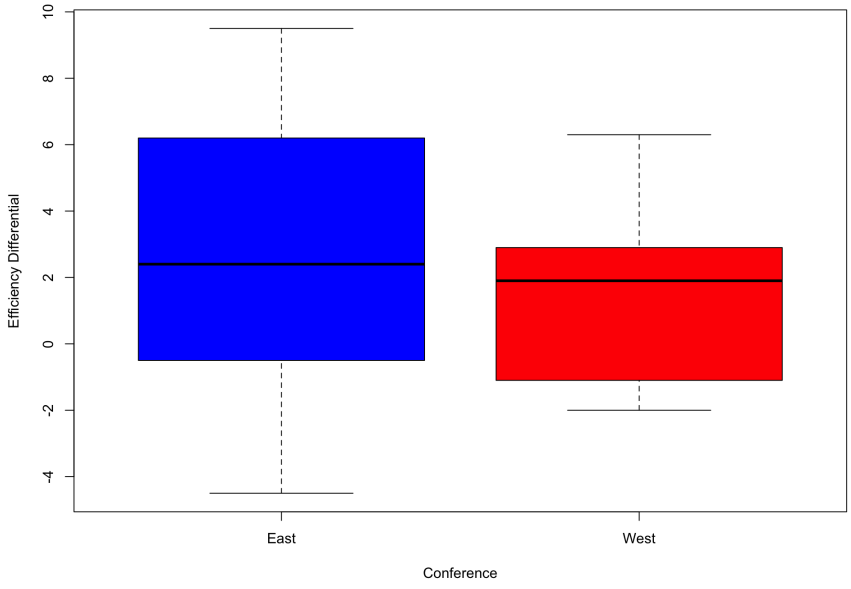
```
> summary(aov(reg$EDIFF~reg$CONF))
            Df Sum Sq Mean Sq F value Pr(>F)
reg$CONF     1   5.91   5.906   0.481  0.496
Residuals   20 245.69  12.285
> t.test(reg$EDIFF~reg$CONF)

        Welch Two Sample t-test

data:  reg$EDIFF by reg$CONF
t = 0.6437, df = 12.839, p-value = 0.5311
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.487542  4.595235
sample estimates:
mean in group East mean in group West
          2.600000           1.546154
```

For the ANOVA test, we resulted in a p-value of 0.496, and the two sample t-test gave us a p-value of 0.5311. Because we yielded with both p-values higher than our alpha, 0.05, I fail to reject the null hypothesis that there is no significant difference in the efficiency differential of NBA teams across the two conferences.

Furthermore, if we look at the boxplots comparing the East and West conference, we see that both the means and medians are pretty similar. Of course, the range and interquartile range of the two datasets, during the regular season, are pretty different; however, the center still remains relatively the same.



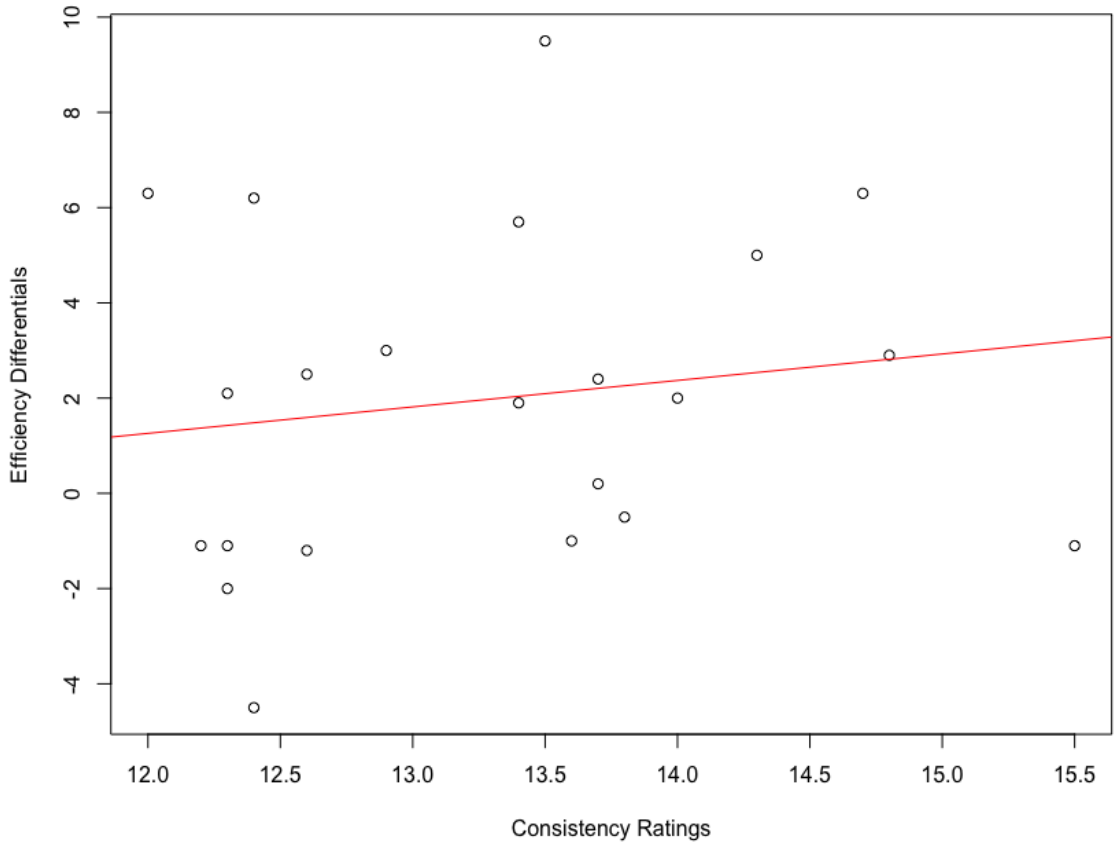

```
> summary(reg$EDIFF[reg$CONF == "East"])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   -4.5    -0.5     2.4     2.6     6.2     9.5
> summary(reg$EDIFF[reg$CONF == "West"])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -2.000  -1.100   1.900   1.546   2.900   6.300
```

Research Question #2:

Is there a correlation between consistency and efficiency differentials during both the regular season and the Restart?

This question aims to answer whether or not there is a correlation between consistency ratings and efficiency differentials. In this dataset, the consistency rating is said to have a direct relationship with the unpredictability of the team. For example, if a team has a higher consistency rating, the team is more likely to be unpredictable. This may seem contrary; however, because some teams are very consistent with their style and play of the game, it is very unpredictable when they will break the consistency. Our efficiency differential, calculated by the difference between the offensive and defensive efficiencies, can help us determine if there is a correlation between the two variables of interest.

First, I want to see what the data looks like and what trends can be found. According to the figure below (Regular Season), the data seems to have a positive association, but the strength of the linear relationship looks somewhat moderate to weak.

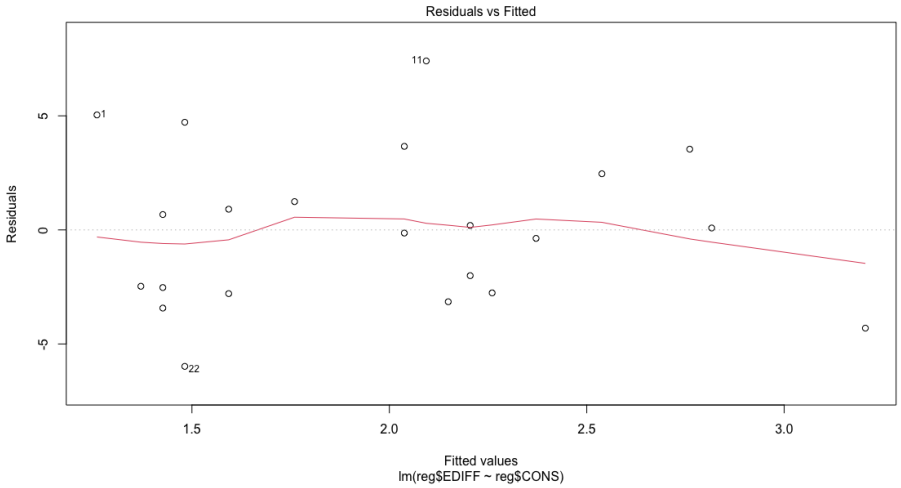


Hypothesis (Regular Season):

- H_0: There is no significant correlation between consistency ratings and efficiency differentials.
- H_1: There is a significant correlation between consistency ratings and efficiency differentials.

The null hypothesis in this case is that there is no significant correlation between consistency ratings and the efficiency differentials. On the other hand, the alternative hypothesis states that there is a significant correlation among these two variables.

When creating a linear model and interpreting the regression of the relationship of these two variables, it's important to analyze the residuals plot that R provides us to search for any unexpected trends.



Because we see a residual plot with very little to no noticeable trend, where the fitted line is somewhat straight, we may want to test more to see how well this data fits and check if the correlation is significant. In order to reach the conclusion of this test, I will perform a population correlation coefficient test.

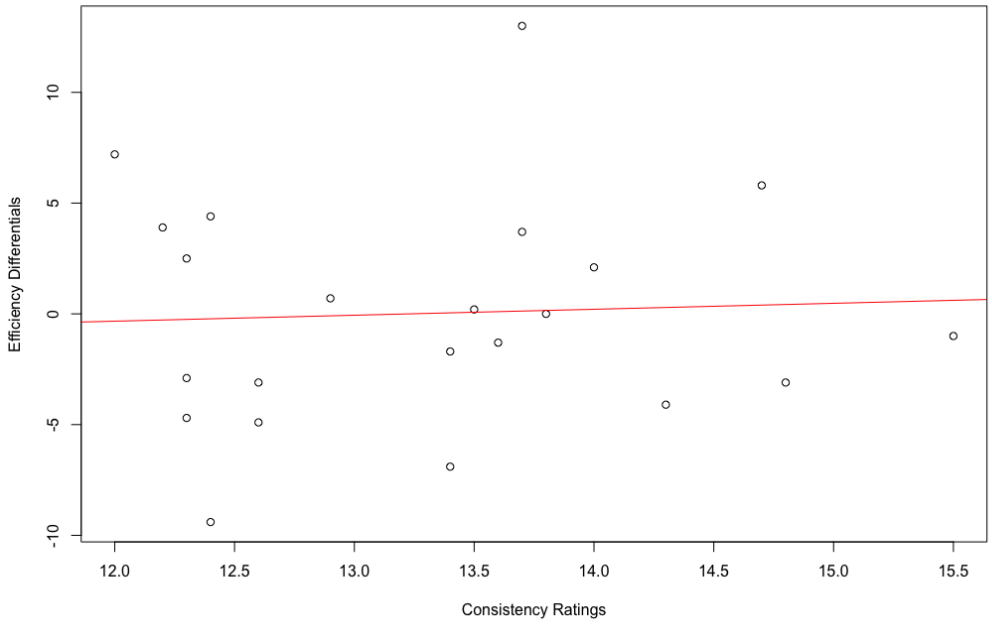
```
> cor.test(reg$EDIFF, reg$CONS)

        Pearson's product-moment correlation

data:  reg$EDIFF and reg$CONS
t = 0.7084, df = 20, p-value = 0.4869
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.2838812  0.5422906
sample estimates:
      cor
0.1564525
```

Two visualizations that can help our conclusion is the regression plot showing a weak-moderate strength and the residuals plot. Even though the residuals plot doesn't reveal much of the linear model, it helps us to dig deeper with the study. We also notice that the proportion of the variance, or R^2, is 0.0245. This means that only 2.45% of the efficiency differentials can be explained by the consistency ratings in the linear model. Nevertheless, because our p-value from the correlation test is 0.4869, which is higher than our alpha of 0.05, we fail to reject the null hypothesis that there is a significant correlation between consistency ratings and efficiency differentials.

Furthermore, I'd like to see this same analysis with the NBA Restart. Like before, I'd like to look for any noticeable trends and traits. According to the figure below, the data seems to have a positive association, but the strength of the linear relationship also looks somewhat moderate to weak just like the regular season.

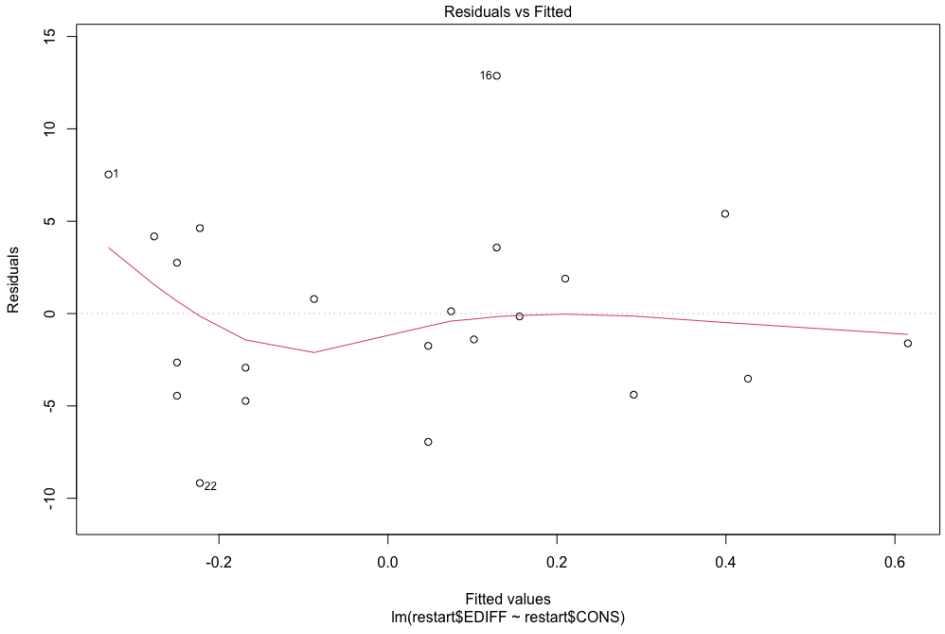


Hypothesis (Restart):

- H_0: There is no significant correlation between consistency ratings and efficiency differentials.
- H_1: There is a significant correlation between consistency ratings and efficiency differentials.

The null hypothesis in this case is that there is no significant correlation between consistency ratings and the efficiency differentials. The alternative hypothesis states that there is a significant correlation among these two variables.

Once again, I'd like to interpret this linear model by analyzing the residuals plot that R provides us to search for any unexpected trends.



Because we see a very little plot with noticeable to no where the trend, is fitted line somewhat

straight, we may want to test more to see how well this data fits and check if the correlation is significant. In order to reach the conclusion of this test, I will perform a population correlation coefficient test.

```
> cor.test(restart$EDIFF, restart$CONS)

        Pearson's product-moment correlation

data:  restart$EDIFF and restart$CONS
t = 0.23118, df = 20, p-value = 0.8195
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.3782166  0.4631517
sample estimates:
       cor
0.05162374
```

Once again, the two visualizations from the regular season, the regression plot showing a residuals plot, helped us reach a conclusion on the correlations in the restart. Although the residuals plot doesn't have sufficient insight, it helps us understand that a correlation test shall be effective. We also notice that the proportion of the variance is 0.002665. This means that only .2665% of the efficiency differentials can be explained by the consistency ratings in the linear model. Nevertheless, because our p-value from the correlation test is 0.8195, which is higher than our alpha of 0.05, we fail to reject the null hypothesis that there is a significant correlation between consistency ratings and efficiency differentials.

Because both datasets led me to a failure in rejecting the null hypothesis, I would say that there is essentially very little to no significant correlation between the two variables, consistency ratings and efficiency differentials.

Research Question #3:

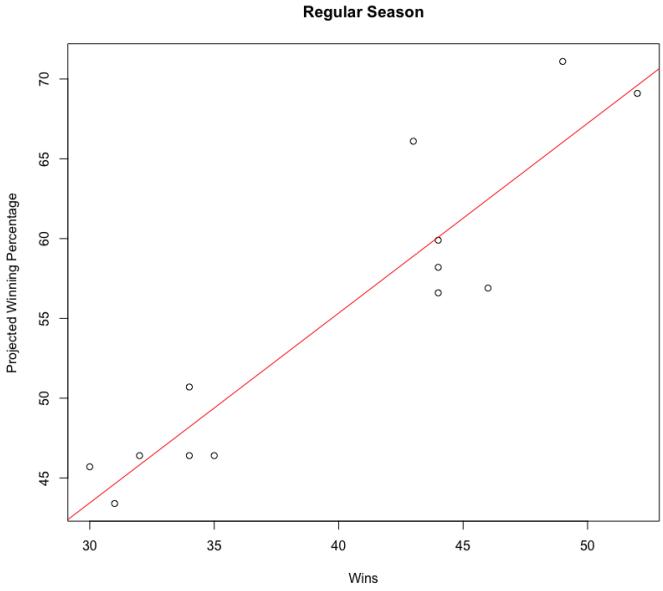
Considering there were more teams in the West during the restart, did the performances of those teams differ substantially in terms of wins?

After the NBA analyzed the teams that would be participating in the NBA Restart, the Western Conference had a more competitive bracket, so I would like to see if their statistics had a significant difference. First, I will be using a correlation test on both the regular season and the Restart to see if the correlation coefficients are significant.

I want to see what the data looks like and what trends can be found. According to the figure below, the linear model seems to have a positive association with a strong correlation.
Hypothesis (Regular Season):

- H_0: There is no significant correlation between wins and projected winning percentage.
- H_1: There is a significant correlation between wins and projected winning percentage.

The null hypothesis in this case is that there is no significant correlation between wins and projected winning percentage. On the other hand, the alternative hypothesis states is that there is a significant correlation among these two variables.




```
> lm(percentagesRegular~west_regular$W)

Call:
lm(formula = percentagesRegular ~ west_regular$W)

Coefficients:
   (Intercept)  west_regular$W
         7.762           1.189

> cor.test(percentagesRegular, west_regular$W)

        Pearson's product-moment correlation

data:  percentagesRegular and west_regular$W
t = 8.2742, df = 11, p-value = 4.734e-06
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7720944 0.9786721
sample estimates:
      cor
0.9282082
```

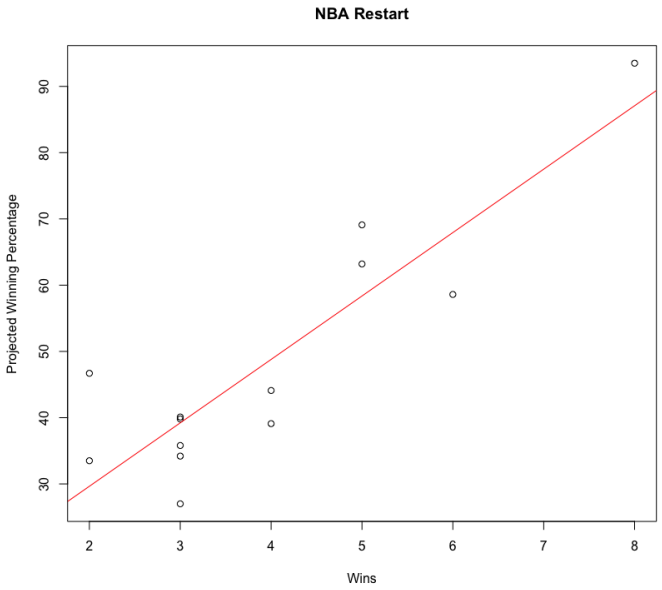
According to the model, during the regular season, the Western Conference generally had an increase of 1.89% for the projected winning percentage for every win they recorded. According to the correlation test, since the p-value, 4.37 x 10^-6, is lower than alpha, 0.05, we reject the null hypothesis that there is no significant correlation. The correlation is also 0.9282, which implies that the strength of the linear relationship is strong.

Now, I wish to repeat this process for the NBA Restart.

Hypothesis (NBA Restart):

- H0: There is no significant correlation between wins and projected winning percentage.
- H1: There is a significant correlation between wins and projected winning percentage.

The null hypothesis in this case is that there is no significant correlation between wins and projected winning percentage. On the other hand, the alternative hypothesis states is that there is a significant correlation among these two variables.




```
> lm(percentagesRestart~west_restart$W)

Call:
lm(formula = percentagesRestart ~ west_restart$W)

Coefficients:
   (Intercept)  west_restart$W
        10.506           9.571

> cor.test(percentagesRestart,west_restart$W)

        Pearson's product-moment correlation

data:  percentagesRestart and west_restart$W
t = 6.3385, df = 11, p-value = 5.533e-05
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6546370 0.9656144
sample estimates:
      cor
0.886034
```

According to this model, during the regular season, the Western Conference generally had an increase of 9.571% for the projected winning percentage for every win they recorded. According to the correlation test, since the p-value, 5.33 x 10^-5, is lower than alpha, 0.05, we reject the null hypothesis that there is no significant correlation. The correlation is also 0.8860, which implies that the strength of the linear relationship is strong.

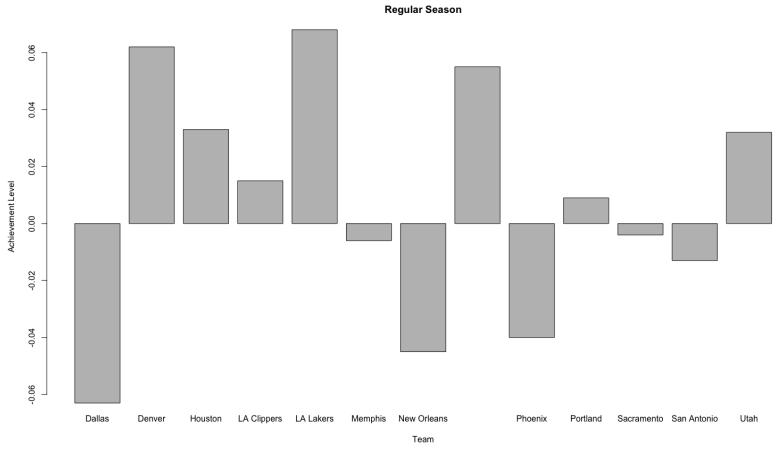
Because both tests had p-values lower than our alpha, we rejected both null hypotheses that there is no significant correlation. However, even if the two linear models had significant correlations, my main concern was to see if there is a significant difference between the two linear models from the regular season and the Restart. The 95% confidence interval was (0.7721, 0.9787) and the interval for the restart was (0.6546, 0.9656). Because there is an overlap in the two intervals, we can say that the two correlations are not significantly different.

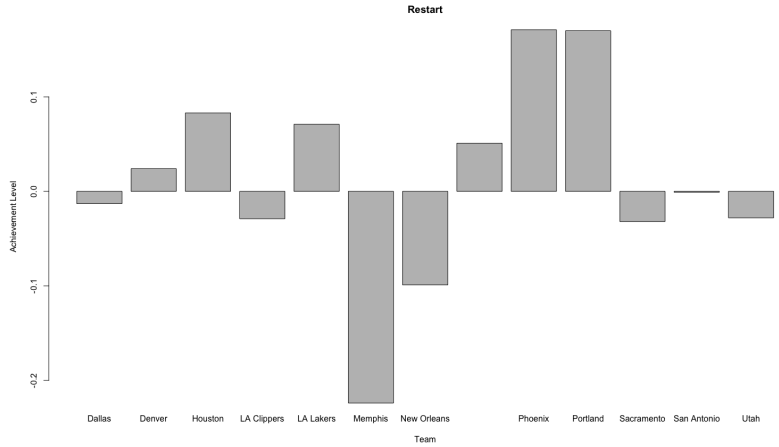
Finally, I would like to look for differences between the achievement levels from the regular season and the Restart. To begin this, we should first take a look at the teams that did the best in terms of wins. The records from the regular season are shown on the left and those of the Restart of on the right.

	TEAM	DIVISION	W	ACH
1	LA Lakers	Pacific	52	0.068
2	Denver	Northwest	46	0.062
3	Oklahoma City	Northwest	44	0.055
4	Houston	Southwest	44	0.033
5	Utah	Northwest	44	0.032
6	LA Clippers	Pacific	49	0.015
7	Portland	Northwest	35	0.009
8	Sacramento	Pacific	31	-0.004
9	Memphis	Southwest	34	-0.006
10	San Antonio	Southwest	32	-0.013
11	Phoenix	Pacific	34	-0.040
12	New Orleans	Southwest	30	-0.045
13	Dallas	Southwest	43	-0.063

	TEAM	DIVISION	W	ACH
1	Phoenix	Pacific	8	0.171
2	Portland	Northwest	6	0.170
3	Houston	Southwest	4	0.083
4	LA Lakers	Pacific	3	0.071
5	Oklahoma City	Northwest	4	0.051
6	Denver	Northwest	3	0.024
7	San Antonio	Southwest	5	-0.001
8	Dallas	Southwest	3	-0.013
9	Utah	Northwest	3	-0.028
10	LA Clippers	Pacific	5	-0.029
11	Sacramento	Pacific	3	-0.032
12	New Orleans	Southwest	2	-0.099
13	Memphis	Southwest	2	-0.224

Once again, it should be noted that the number of games played are quite different. It is important to notice that Phoenix had won 34 with a negative achievement level, but it flipped the opposite way during the restart, with an undefeated record and an achievement level of 0.171. On the other hand, some teams did a little worse compared to their regular season. For example, the Utah Jazz and LA Clippers originally were placed in 5th and 6th with a positive achievement level, but they both flipped and became negative in the Restart. To visualize all teams at once, a barplot from both seasons is provided below.





Now, I'd like to compare the variances of achievement levels between the regular season and the Restart. The two variances seem pretty close to each other, so we can see how much the distribution locates this variable in terms of the median. To do this, I will perform a Wilcoxon Test. Furthermore, the means of the achievement levels definitely did change, but now we have to see if these differences are large enough by backing performing a T-test to compare the means and ultimately support the Wilcoxon test.

```
> var(west_regular$ACH)
[1] 0.001742577
> var(west_restart$ACH)
[1] 0.01120408
```

```
> mean(west_regular$ACH)
[1] 0.007923077
> mean(west_restart$ACH)
[1] 0.01107692
```

Hypothesis for Wilcoxon Test:

- H_0: true location of median in ACH is equal to 0 across the two seasons
- H_1: true location of median in ACH is not equal to 0 across the two seasons

The null hypothesis in this case is that there is no shift in the distribution. On the other hand, the alternative hypothesis states is that there is a significant shift among these achievement levels. The Wilcoxon test will successfully compare these two location shifts.

```
> wilcox.test(west_restart$ACH, west_regular$ACH)

        Wilcoxon rank sum test with continuity correction

data:  west_restart$ACH and west_regular$ACH
W = 88.5, p-value = 0.8575
alternative hypothesis: true location shift is not equal to 0
```

Because the p-values in both seasons are higher than our significance level, 0.05, we fail to reject the null hypothesis that there is no shift in the true location of the medians. This means we can assume that there is no shift in the true location of the medians.

Hypothesis for T-Test:

- H_0: true difference in means is equal to 0 across the two seasons
- H_1: true difference in means does not equal 0 across the two seasons

The null hypothesis in this case is that there is no significant difference in the means. On the other hand, the alternative hypothesis states is that there is a significant difference in the means. The t-test will be used to compare both means and will be supported by a Wilcoxon test to look for shifts in the true location of the median.

```
> t.test(west_regular$ACH, west_restart$ACH)

        Welch Two Sample t-test

data:  west_regular$ACH and west_restart$ACH
t = -0.099939, df = 15.645, p-value = 0.9217
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.07017719  0.06386950
sample estimates:
  mean of x   mean of y
0.007923077 0.011076923
```

Because the p-value, 0.9217, is lower than our significance level, 0.05, we fail to reject the null hypothesis that there is no difference in the mean achievement levels across the two seasons. This means we can assume there is a significant difference in the means of achievement levels.

Even though we concluded that the correlations between number of wins and projected winning percentage are not that very much different from the regular season and the Restart, we noticed that the correlations itself are significant. It gives some insight that the projected winning percentages are dependent on the number of wins. Because these correlations are significant, it allows us to dive deeper and check for differences in the distributions, such as if the means and medians differ for the achievement levels. Since we were able to see the distribution shapes and centers do not differ as much as expected, we can conclude that the performances of the Western conference teams did not change. For this specific experiment, I expected the distributions of the teams achievement levels and winning records to change significantly. This was mainly due to the few months break the players had and the energy they brought to the table during the Restart. Some specific teams, like the Phoenix Suns or the Denver Nuggets, as seen from the above barplots, that changed drastically. However, when looking at the sample in its entirety, we can see that there was not much of a change overall.

Dataset source:
https://www.nbastuffer.com/2019-2020-nba-team-stats/