Nicholas Stout

ISTA 116

Professor Dylan Murphy

May 10, 2020

Final Project Report

For my final project I will look at a data set of basketball scores during the NCAA tournament. The question I am trying to answer is “Analyzing the average scores of the NCAA basketball tournament based on several factors such as year, region, seed, and round”? For my final project about the NCAA basketball scores I will get my research from the website “Data.World” the data set can be accessed by using the following URL <https://query.data.world/s/avrwqiokitxabh5skhtx33npqstsmd>

In the data set you are able to see the teams, years, regions, rounds, and what seed they were in the tournament. I created several variables some being explanatory and some being response variables. The explanatory variables that were used in my project were the year, the region, the round, and the seed. The response variables were the scores for each of the different conditions from the explanatory variables. I used the mean and median for certain years of the tournament to help answer my project question. Another association I investigated was the different years of certain teams and whether or not those teams increased their average each year that they are in the tournament. I also took the probability of the average score of each round that a team played. The combined scores for each outcome for each team helped explain the explanatory variables. I showed my results in varies plots to include scatterplot, bar plot, and boxplot. I chose to do this topic because I have always been interested in college basketball and I have been to a few of the tournament games that appear in the data. The other reason I chose this topic is because I felt like there would be enough data to answer the question I have proposed. With many different games being played each year I felt that it would be interesting to show different types of plots that show whether the games improved each year or whether they had significant variance. The calculations helped show improvement in some teams each year on the average score and other team’s scores stayed the same or even decreased.

The population of interest that I answered throughout this project was weather the scores have increased based on the variables that I have listed above. While doing these calculations and plots I was able to conclude that the population of interest for whether or not the average scores increased each year based on the factors of region, team, seed. While doing the plots you can see that the average score was able to be generalized to answer the question the project question.

While using the data to create graphs and plots you can clearly see the difference in average scores from year to year, the difference in the average scores between the three teams, and the average final score for each year that the tournament was played. The first plot was a box plot that compared three different teams.

A screenshot of a cell phone

Description automatically generated

The boxplot shows the three different teams Alabama, Arizona, and Duke. The Y axis is the average scores for those three teams for every year that they played in the tournament. I was able to conclude that Alabama has the highest median average score, with the next highest median being Duke and then Arizona. The above boxplot shows the average scores on the three teams each year they played in the tournament. The boxplot comparing these three teams depicts that every year Alabama played in the tournament averaged a combined final score of about 160. This combined average score well exceeds the two other teams. Duke had a combined average score of about 150 which means that they might have played tougher teams or that for most of the years in the tournament they did not average high scores. The last team, Arizona averaged the lowest mean about 140 which means they had a hard time scoring due to the teams they played or the region they were in.

The next plot I created was a scatterplot that compared the average final scores of all games played each year that the tournament was held. A screenshot of a cell phone

Description automatically generated

In the scatterplot the y axis is the average of the final scores of all the games for each year. The x axis is the years in increments of 5. If I tried to increase the year variable by 1 year, I can expect the average score variable to decrease by 0.3040 on average as predicted by the lm function. We can see that the scatter plot is linear because the data points on the plot follows the abline. We can see that in certain years the average final score was above 155 which mean that more teams that year scored more and less defense was being played. We can also see another outlier for the year 1985 which is below the average final score of 130 which mean teams didn’t have high scores or that there was more defense being played those years.

The final plot that I created was a boxplot that compared all the average total scores for regions in the tournament. A screenshot of a cell phone

Description automatically generated

this boxplot shows the average scores for all years based on the region. The Championship Region has a median of about 150 and has an outlier that is below 100 points. The East has one outlier that is below 100 points and has 6 outliers that have scores around 200 to 220. The East has a median of about 130. The Final Four has no outliers and has a median of about 140. Most regions have about the same median score, while the Southwest has a lower median score than the other regions. By just looking at the plot my guess is that the Southwest region has the lowest median do to the fact that the teams that played in that region did not tend to average high scores. Outliers for the Southwest region being below and above the median leads me to think that the teams that played in that region were averaging high scores and for the outlier below the mean teams scores were less then mean due to the fact that the three-point line had not yet been established.

After I had done the visualizations for the project, I then moved on to the calculations for the project. The first calculation I did was a regression model where I used the code:

model1=lm(finalscore~Year,data=Big\_Dance\_CSV)

summary(model1)

The code uses a formula of finalscore = 751.03842-0.30404\*Year. After running this code I was able to see the output and was able to conclude it as such; for a unit increase in year, I can expect the final score to go down by 0.30404 on average (as predicted from the linear regression model) assuming all the other factors remain constant. The multiple R-squared: 0.02101 value after running the regression model tells us how close the data is to the fitted regression line and in this case, we can see that it is not really close to fitting the regression line since the R-squared value is pretty small.

After I calculated the regression model, I then calculated a hypothesis testing anova. The code I used can be seen in the R file I have shared it with you. For the anova test, I was able to conclude a null hypothesis that all regions have the same average final score. The alternative hypothesis being that not all regions have the same average final score. The P value that I calculated is less than 5% , meaning that we can reject the null hypothesis and conclude the alternative hypothesis as being the preferred one. Therefore, we have sufficient evidence to conclude that at least one region has a significantly different average final score compared to the others. Southeast-East and Southeast-Southwest and West-South and Southeast-South these four pairs have a significant difference in their average final scores. By doing the anova test this is a great way to tell whether or not the P value is less than 5% for the data set that we are using and it is one of the only ways the tell if there is significant differences between the means of the variables that I am testing. Another result we can conclude from the anova test is that the Southeast-East and Southeast-Southwest and West-South and Southeast-south these four pairs have a significant difference in their average final scores. Because the one-way anova gave us a significant result i.e. we identified at least one region that had a different mean final score compared to other regions. A post hoc analysis (tukey analysis) to find out which regions were actually significantly different in their mean values of the final score.

The final calculation was a T-test. A T-test tells you whether there is a significant difference between the means of the variables that you are testing. I used a T-test for the data because it told me whether or not my guess of 140 of an average finale score is below, above, or within the mean. After running the T-test I conclude that the H0: Average final score is equal to 140 and that H1: Average final score is significantly different from 14095% confidence interval for the true mean of the final score for the college basketball games is from 141.4712 to 143.2408. Clearly this interval does not contain the hypothesized mean of 140. Therefore, I reject H0 and conclude H1is the mean final score which is significantly different from 140. This tells us that the average finale score was higher than my hypothesized number of 140.

In conclusion the question I was trying to answer of “Analyzing the average scores of the NCAA basketball tournament based on several factors such as year, region, seed, and round”, was able to be answered by the data, plots and the calculation used. There was some roadblocks encountered such as too much data to be shown in a plot or to calculate where a specific teams score increased or decreased every game they played in the tournament. Other than that, I was able to answer my question by showing every year that the tournament was played and the average score for those years in the different plots described above. By doing the calculations and interpreting the data I was able to conclude that the average score did vary each year that the tournament was played. The plots shown above identify the different average scores for the region, round, year and team that the tournament was played. I was able to conclude whether it was symmetrically about the mean or whether it was skewed. The plots also showed the normal distribution for the average scores based on the certain factors being addressed. Other additional factors to consider would be the stress level of the players, the variations of the arenas, or football stadiums, and the March Madness factor of cinderella teams playing above their typical performance.