**A Summary of DSC 530 Final Project: NCAA March Madness**

Ramsey King

Bellevue University

DSC 530: Introduction to Data Science

Prof. Matthew Metzger

June 4, 2021

For this summary, I will be reviewing the following questions: the Outcome of my EDA, What I feel was missed during the analysis, variables that I felt could have helped in the analysis, incorrect assumptions, and challenges faced.

The hypothesis of my EDA was to determine if the best seeded teams made the NCAA Men’s basketball tournament more often than the other seeds. Going through the statistical analysis, it was shown that this is the case. There were many ways of looking at the data (PMF, CDF, correlations tests, and regression analysis. Each of these analysis led to the conclusion that the #1 seeds reached the Final Four at a higher rate than the other seeds.

What I felt that was missed during the analysis was a deeper dive into the data. There were many variables that were available to me, such as which region (the tournament is broken up into 4 16-team regions) the team that made the Final Four came from. There could have been some information gain that could have shown that a particular region produces more Final Four teams than others.

Variables that I felt that could have helped during the analysis was a couple of columns that had significantly incomplete data. These columns were the OWP (opponent’s winning percentage), and OOWP (opponent’s opponent’s winning percentage). These two metrics show how strong of an opponent the team in question has faced during the season prior to the tournament. There may have been more information gain if these two columns contained complete data.

The assumptions made that I felt that were incorrect was that the data was going to be easy to gather. I know this is not specific to the project, however, there is something to be said I feel in terms of being able to obtain the information needed. There was a lot of pre-processing of the data that took place in Excel. It was a really good experience for me to see how gathering the data can be cumbersome. It has also provided me with the motivation to get better at working with data.

Challenges faced during this project was making the decision to use the Thinkstats classes and functions or to use the more traditional libraries and functions found in the data science industry. I decided on using the more traditional libraries, functions, and methods found in the data science industry because I feel that these are going to be the libraries that I will use most often. Going this route actually helped me learn the Thinkstats library a little better. It gave me the opportunity to undertand what was going on behind the code, because I had to figure out how to do it without using their libraries.

This class was wonderful and I look forward to seeing what else the curriculum will help me learn!