(a) Provide a summary of your definition and your approach.

I defined clutch as good performance under pressure. I investigated time as the primary factor that influenced pressure. I limited myself to only investigating 2013 since that was the most recent year in the retrosheet database. Instead of starting off with the players, I thought it would be fun to look at team performance. After learning a little bit about overall team performance I also looked at team performance late in the game (assuming there were nine innings).

(b) Describe the data you used to support your work.

The data I worked with was numerical. I used Sean Lahman's baseball database for the season averages of team or player performance and I also used retrosheet to pull numbers based off of innings/events that occurred.

(c) Describe your methods for evaluating the data. Include your queries with the .sql file you submit for questions 1 and 2.

The methods that I used for evaluating data was simple math to calculate statistics such as averages, win loss ratios and batting averages. Based off of the relationship between the figures I calculated and the attributes the datasets provided I identified correlations and compared the values I calculated to a baseline average to determine if the trend/player I was investigating was significant.

(d) Describe your results and conclusions that you draw from the results.

One of the first things that I looked at was team performance. Luckily, the lahman dataset has an entire table dedicated to team based statistics. I calculated the batting average and the win/loss ratio for each team in 2013. Unsurprisingly, I found that there was a strong positive

correlation between the two variables. Teams with a higher batting average usually win more games. I also saw a few outliers where even with a poor batting average the team would still have a decent win/loss ratio or vice versa. This was pretty inconclusive. Although this told me that better batters means more wins I also wanted to check out the individual players and their batting averages as well. I discovered that the same correlation exists, there is a lot more noise since you have more situational outliers. For example, Zach Duke batted 1000 for both the Washington Nationals and the Cincinnati Reds in 2013 even though he was primarily a pitcher.

I decided to turn to retrosheet for more in-depth data. I modified one of the queries we used in class and looked at how many teams were losing after the 6th inning but still managed to win the game. The Minnesota Twins were the comeback team of 2013 with a total of 16 games won this way. However, when I went back to the lahman dataset to check their win/loss ratio it was only 0.6875. To ensure that this was an outlier I checked the rest of the teams that won games after the 6th inning and most of those teams won more games than they lost.

I also adjusted the query to pull comebacks in the 8th inning which only removed the Mariners from the table which was surprising. It seems that as far as comebacks go, there are teams that do better than others in performing and pulling the game back.

(e) Based on what you learned, what is the next question you would ask in this area and how would you evaluate that question?

Since I have identified teams that have a propensity for making a comeback, I would then investigate why they do so. I would like to get into the individual player statistics and see if a few players contribute to the win expectancy significantly more than others under these high pressure situations. Primarily, I would use the retrosheet data to extract data from the 2013

season of the Twins and possibly use pandas and matplotlib to help visualize the results. I found that staring at the result of queries can only get me so far in an analysis.