Master's students extra project is worth 20 points. Problems are worth 3 points each except problem one (2 points). It is due March 12th at 11:59 PM. Since SQL will not let you do regressions, python or R are going to be necessary. You can query the database with python/R and then do your analysis and plots. Any table joins should be done with SQL, only the final plots and calculations should be done in python/R. When the question asks you to plot the relationship between X and Y, make sure you add a regression line to the scatterplot. If it asks for the mathematical relationship, get the slope and intercept of the regression.

Here are some links to look at t-tests and regressions:

linear regression code from: https://towardsdatascience.com/the-complete-guide-to-linear-regression-in-python-3d3f8f06bf8

t-test code from: https://www.pythonfordatascience.org/independent-samples-t-test-python/

1) Plot a histogram of the players age as of January 1st, 2014. When the question asks for players, be sure and exclude the managers.
2) Do a regression on players heights and weights to get the mathematical relationship between the two. (get the slope and intercept of the regression).
3) Plot the relationship between players age (as of Jan. 1, 2014) and batting average. Batting average is determined by dividing a player's hits by his total at-bats for a number between zero (shown as .000) and one (1.000). Do players get better or worse with age? What is the best age for a player?
4) Do heavier players hit more home runs per at bat than lighter players? Plot the relationship between weight vs home runs per at bat. Compare home runs per at bat for above average weight players to below average weight players? Is the difference statistically significant? Remember when calculating the average, don't average the averages. Sum up the home runs, sum up the at bats, then divide them.
5) Which team has the oldest average age for players? Is there a relationship between average age of players and winning percentage?
6) Calculate the percentage of times a player is caught stealing (from batting table: caught_stealing/(stolen_bases + caught_stealing) * 100. Is there a statistically significant difference between the best manager and the worst manager? Best manager being defined as the manager with the lowest percentage of times their players get caught stealing.
7) For each team, calculate the batting average (hits/at_bats) for the team. Do all teams with better than average batting also have better than 50% winning percentage? Print a table with all the teams that have above average batting and less than 50% winning percentage.