Project 1

Presentation Outline

Cell 11: got JSON data

Cell 12: pprint

Cell 14: WITH EXTENSIVE SEARCHING: found out how to expand actual team stats (they weren’t available on the original information)

Then, my plan was to separate each team, then extract the stats within each team… this quickly proved to be A) inefficient B) a waste C) so much extra code D) not even practical because I needed to combine the data

*(on to Hockey Stats 2)*

Cells 92/93: added ?expand=team.stats earlier in the process

Cell 94: now have all info in json form that I’ll need

Cell 87/89: BUT thought it would be better to convert it all to a DF… wrong again

*(on to untitled)*

Tried again, useless

*(on to untitled 1)*

Cell 258: each individual team name (which worked but not what I was looking for)

Cell 258: for loop that gets all of the info I need, and can be adjusted for which stats I’ll need

Created lists, got the data, made data frame, can now create graphs

*(skip untitled 2,3 onto final: Untitled 4)*

Cell (8): assumption: more goals per game = more goals allowed per game(game opens up)

Cell (13) scatter plot for gpg & gapg: **Answer: (on to cell 14) is actually false, but the correlation is extremely weak. Kind of see a negative correlation but again, extremely weak. Answer: even tho weak correlation, actually the opposite of original assumption. Weak negative correlation**

Cell (16): more faceoffs won correlate to more goals scored

Cell (17) scatter plot is all over the place, moved on but absolutely no correlation between faceoffs won and goals per game. **Answer: faceoffs don’t matter for scoring goals. What I presume matters more (more work needed with advanced stats) is defense and attack after the faceoff.**

Cell (19): my assumption and a traditional assumption would be that more shots = more goals.

Cell (20): pattern of distribution: pvalue of 0.35, null would be accepted, not really significant in terms of the spread of shots on goal in the league.

Cell (23/24): r squared = 0.21, again it’s a weak correlation but we have to keep in mind only 31 teams. If there were a few less outliers, the r squared would greatly change.

Next assumption(66): the more you score (goals per game) the more games you win

Cells (70): r squared value is 0.7, which I think implies a pretty strong correlation. Again, 31 teams so a few outliers impact the data, but 0.7, while certainly not a perfect correlation, is pretty strong. **Answer: goals per game is a strong indicator in success. If you can determine an accurate projection of goals per game, you can make fairly accurate assumptions about overall wins.**

Cell (59) assumption: the better the goalies you have (save %) the more you’ll win… (onto cell 62) scatter plot formed, had trouble getting trend line, BUT I was able to print the r squared value, which tells me that although it appears to be a weak positive correlation, with an r squared value of 0.3, it’s a pretty weak correlation. **Answer: the pure stat of save % does not greatly impact overall wins… again, there is so much that can be done with these stats, where I’ll likely go next with this is comparing the amount of shots allowed to save %... ALSO, look for API on an advanced data website and see high danger shots etc. (the reason I didn’t start with advanced stats were that advanced stats ignore these “typical” stats like shots on goal, faceoff % and goals per game etc.**

Cell (36) do more power plays = more power play goals?

Cell (38): another fairly weak correlation. Slightly positive, it generally helps to have more power play chances, but overall a weak correlation. **Answer: amount of power plays doesn’t necessarily mean more power play goals**

Cell (40): wanted to isolate power plays and see the difference. Are some teams getting more chances?... onto cells (42) **Answer: no, teams generally get the same amount of chances, which implies that the referees aren’t playing favorites or are acting in a biased way. \*Another interesting note: by getting the data for power play chances per team, we actually can answer questions about the referees EVEN THOUGH, we don’t have any direct data ABOUT refs in this data set. This is an example of answering questions that have to do with related topics (refs) that aren’t directly incorporated in the actual data.**

Cell (44): z scores for every team, not very far from the mean

Cell (45): boxplot to get more information about power plays. Median slightly higher than the mean.

Cell (47): more work on goals per game… onto cell (49): wanted to visualize this in a bar graph, get z scores and boxplot. Z scores can be compared to earlier mean we calculated.

That’s my project. I have a bit more work to do but overall, I was able to answer the questions I posed and began to draw some good conclusions either confirming or denying some of the more “classic” or “typical” assumptions NHL fans have about certain stats.