Seth Pennebaker

1) Selecting Data
    a. My dataset is a NHL stats with over 11,000 games in the dataset with games dating back to 2011.
2) Defining Tasks
    a. Section 1
        i. Took game_team_stats and got rid of the header so I can handle the data
        ii. Mapped the faceoffprecentage and the team_id by splitting the lines.
        iii. I then countedbyKey with this new data
        iv. Summed the values of all the faceoffs that each team had.
        v. I then averaged out the value of the faceoffs with avgRatings02
        vi. I then took the team_info.csv and extracted the team info which I then mapped the team names to the face off percentages of avgRatings02
        vii. I then augmented the data by joining avgRatings02 and teams
    b. Section 2
        i. I first took the game_skater_stats.csv data and extracted the header away from it.
        ii. I then mapped out the data from this set by splitting the lines for the players_id and if they got a goal during that game.
        iii. CountedByKey on this data set.
        iv. I then summed the value of all the goals that each player scored
        v. I then took player_info and extracted the header out so I could deal with the data.
        vi. I then mapped that dataset using rplit to the goals and the players and then being able to order them
    c. Section 3
        i. Created an sql table with the game_skater_stats where I extracted player_id and penalty minutes.
        ii. Created another sql table with player_info.csv where I extracted the first name and last name
        iii. Connected the first/last names with the lasyer id and was able to display the players with the most penalty minutes
    d. Section 4
        i. Created an sql table with game_team_stats where I extracted the head_coach and the faceoffpercentages
        ii. Displayed the data where I ordered the games with the highest percentage and the current head coach that was during that game
    e. Section 5

        i. Created an sql table with game_skater_stats where I took the player id and connected it with the first/last name players and displayed how may hits that player had.

3) Values of data analytical outcomes
    a. Section 1 – Team with the highest FaceOffPercentage throughout the whole dataset
        i. The value of this dataset is that we can use it for future analytics like if were to take the individuals games and find out the percentage of games that are won when the faceOffPercentage is above 50% for the game.
    b. Section 2 – Most Goals
        i. The value of this dataset is to show the top goal scores of this dataset with Alex Ovechkin being the top of the list. This dataset can be used for future analytics by comparing two teams and seeing which one has more top scorers and trying to predict which team would win.
    c. Section 3 – SQL, Top 20 in Penalty Minutes
        i. The value of this data set is to show the top 20 goons in the league are. These players are typically the ones the end up getting in fights the most out of each other. This can be used in the future as a way of prediciting who has the best chance of getting a penalty throughout a game that may be upcoming.
    d. Section 4 – SQL, Individual games with the highest faceoffpercentage with the current head coach of that team.
        i. This shows off the top 20 games with the highest faceoffpercentae for each game and what the current head coach of the game was. This can be used for future analytics
    e. Section 5 – Players with most Hits
        i. This shows the top 20 players with the most hits throughout all the games. We can even use this information for future analytics like for which players in this top 20 list also have the most Penalty minutes which we can probably guess that these players in this new list would probably be the dirtiest players out of everyone.