Assignment Instructions- Part 1

These are the steps you need to take to complete the assignment.:

**I. Data Coding and Merging**

1. Import the “NHL\_Team.csv” data file and name the dataframe as “NHL\_Team” in Jupyter Notebook. a) Delete the following variables: “Unnamed:0”, “abbr”, “tname”, “lname”, and “sname”. b) Rename the variable “name” to “team\_name”.
2. Import the “NHL\_competition.csv” data file and name the dataframe as “NHL\_Competition” in Jupyter Notebook. a) Delete the following variables: “Unnamed: 0”, “tz”, “start”, and “end” b) Rename the variable “name” to “competition\_name”.
3. Import the “NHL\_game.csv” data file and name the dataframe as “NHL\_Game” in Jupyter Notebook. a) Delete the following variables: “X”, “period”, and “status”.
4. Merge the dataframe “NHL\_Team” into the dataframe “NHL\_Game” by “tid.” Continue to name the merged dataframe as “NHL\_Game.”
5. Merge the dataframe “NHL\_Competition” into the dataframe “NHL\_Game” by “comp\_id.” Continue to name the merged dataframe as “NHL\_Game.”
6. In the merged “NHL\_Game” dataframe, create a variable “hgd” to indicate the goal difference between home and away score (hscore – ascore) and delete observations with missing value in the variable “hgd”.
7. Drop all observations with missing values, if there is still any, from the “NHL\_Game” dataframe.
8. Convert the type of the “date” variable from “object” to “datetime.”
9. Sort the NHL games by “date” and show the first 15 observations.
10. Create two dataframes that separate the “NHL\_Game” dataframe by home and away games. Name them “NHL\_Home” and “NHL\_Away”, respectively. a) Rename variables: i) For away games, rename “ascore” to “goals\_for”; rename “hscore” to “goals\_against” ii) For home games, rename “hscore” to “goals\_for”; rename “ascore” to “goals\_against” b) Create a “win” variable that equals to 1 if the team won the game; 0 if the team lost the game; and 0.5 if it was a draw.
11. Append the “NHL\_Home” and “NHL\_Away” dataframes to be the new “NHL\_Game” dataframe.
12. Generate a team level dataframe that aggregates the total number of games won, the total number of “goals\_for” and “goals\_against” for each team in each competition (i.e. grouped by tid, competition\_name and type). Name this new dataframe “NHL\_Team\_Stats”. Make sure to convert the indexes of the new dataframe back as  variables.
13. Create a dataframe “NHL\_Game\_Count” that include the total number of games played by **each** team in **each** competition (i.e. grouped by tid, competition\_name and type). Name this new variable in the dataframe “game\_count”.
14. Merge dataframes. a) Merge the “NHL\_Game\_Count” dataframe into the “NHL\_Team\_Stats” dataframe by “tid”, “competition\_name”, and “type”. Continue to name the merged dataframe “NHL\_Team\_Stats”. b) Merge the “NHL\_Team” dataframe into the “NHL\_Team\_Stats” dataframe by “tid”. Continue to name the merged dataframe “NHL\_Team\_Stats”.
15. Import the “pp.pk.ppgf.csv” data file and name the dataframe as “NHL\_PPPK” in Jupyter Notebook. Merge the “NHL\_PPPK” dataframe into the “NHL\_Team\_Stats” dataframe by “tricode” and “competition\_name”.
16. Create new variables in the “NHL\_Team\_Stats” dataframe. a) Winning percentage (“win\_pct”)=”win”/ total number of games played b) Average goals for per game (“avg\_gf”)=total number of goals for / total number of games played      c) Average goals against per game (“avg\_ga”)=total number of goals against / total number of games played
17. In the “NHL\_Competition” dataframe, the variable “type” indicates the type of competition: type=2 – regular season. Create a dataframe that contains team statistics for games only during regular seasons. Name this dataframe “NHL\_Team\_R\_Stats”.

# Assignment Instructions- Part 2

## **II. Calculating Summary Statistics**

1. In the “NHL\_Game” dataframe, calculate summary statistics for the  “goals\_for” variable; calculate summary statistics for the “goals\_against” variable based on whether it is home or away game.
2. Create a histogram of the “goals\_against” variable by whether the game is home or away a) Make the color of the histogram green b) Set the number of bins to be 20 c) Make sure the two sub-histograms share the same ranges for the x-axis and y-axis.

# Assignment Instructions- Part 3

## **III. Correlation Analyses**

1. In the “NHL\_Team\_R\_Stats” dataframe, make a scatter plot to depict the relationship between the total number of goals for and the winning percentage. a) Plot the total number of goals for on the x-axis and winning percentage on the y-axis. b) Add a regression line to the scatter plot. c) Make the title of the graph “Relationship between Goals for and Winning Percentage” and make the font size 11. d) Label the x-axis “Total Goals for” and label the y-axis “Winning Percentage”.
2. In the “NHL\_Team\_R\_Stats” dataframe, calculate the correlation coefficient between total number of goals for and winning percentage.
3. Create a scatter plot of the total number of goals for and winning percentage similar to step 1 for regular season games. In this graph, group observations by “competition\_name”. a) Plot the total number of goals for on the x-axis and winning percentage on the y-axis. b) Add a regression line to the scatter plot. c) Make the title of the graph “Relationship between Goals for and Winning Percentage” and make the font size 11. d) Label the x-axis “Total Goals for” and label the y-axis “Winning Percentage”.
4. For the “NHL\_Team\_R\_Stats” dataframe, delete observations of 2011 and 2012 seasons. Continue to name the dataframe “NHL\_Team\_R\_Stats”.
5. In the new “NHL\_Team\_R\_Stats” dataframe, create a scatter plot of total number of goals for and winning percentage. a) Plot the total number of goals for on the x-axis and winning percentage on the y-axis. b) Add a regression line to the scatter plot. c) Make the title of the graph “Relationship between Goals for and Winning Percentage” and make the font size 11. d) Label the x-axis “Total Goals for” and label the y-axis “Winning Percentage”.
6. Calculate the correlation coefficient between total number of goals for and winning percentage in the updated “NHL\_Team\_R\_Stats” dataframe.
7. Save dataframes as csv files. a) Name the updated “NHL\_Game” dataframe as “NHL\_Game2”. b) Name the “NHL\_Team\_Stats” dataframe as “NHL\_Team\_Stats”. c) Name the “NHL\_Team\_R\_Stats” dataframe as “NHL\_Team\_R\_Stats”. d) Make sure to exclude the index as a column in the csv files.