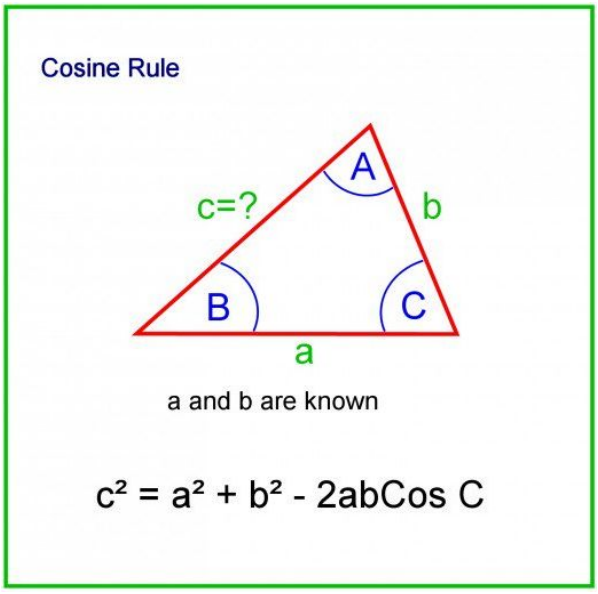
Building an expected goals model from shot-event data

Data wrangling report

The dataset selected for this project is the Statsbomb open dataset; specifically the women's soccer dataset.

Cleaning steps:
The data wrangling, cleaning and engineering steps were as follows:

1. Downloaded the data to a local hard drive
    a. The data consisted of separate folders named 'events','lineups' and 'matches'.
    b. There was also a separate json file name 'competitions' which listed all the competitions and their corresponding IDs.
2. From competitions.json, extract the competition IDs for the women's soccer leagues and tournaments - namely, the NWSL (USA), FA WSL (UK) and the 2019 women's world cup.
3. Using the competitions IDs obtained from competitions.json, the match IDs were extracted from the corresponding matches.json file and appended to a match ID list.
4. Iterating through each entry in the match ID list, the events data was loaded for each match as a dataframe from the corresponding events json file.
5. The events data for each match was comprehensive and documented each event in a particular match. Since the project only requires shot and shot-related events, all the other events are dropped from the dataframe.
6. The data-structure for shot and shot-related events contain the following information:
    a. Location
    b. Time
    c. Player information (shooter and non-shooters)
    d. Type of play
    e. Technique used in shot
    f. Body part used for shot
    g. Outcome of shot
    h. Type of key-pass preceding the shot
    i. Preceding event information
    j. xG predicted by Statsbomb
7. Each of these specific factors are collated into a separate list and once the shot-related data from all the matches are collated, they are joined into a dataframe.
8. Some of the features needed to be engineered. The following features were engineered:
    a. The shot location in x-y coordinates were translated into the Euclidean distance from the center of the goal.
    b. The angle of the shot from the goal face is calculated using the cosine rule as given below



c. The number of players in the triangle formed by the shot location and the edges of the goal was designated as the packing density. This was calculated by the barycentric algorithm given here.

The total number of shots in the dataframe assimilated was 5929.

Missing values:

1. After assimilating all the features and samples into a dataframe, the following features were found to have missing data
    a. Shot angle
        i. One instance of the shot angle was observed to be NaN. Upon closer inspection it was found that the shot location was on the same vertical coordinates as the edges of the goal. Therefore, for this shot, the missing angle is imputed as 0 degrees.
    b. Preceding pass
        i. Of the 5929 shots in the dataframe, 1845 shots don't have preceding pass information. This could be because some shots are not made in open play, while some shots did not result from a pass but from a dribble. The missing data will be imputed according to the context of the shot

Outliers:

Outliers will be handled in the EDA phase.