**PAM2000 Applied Data Science 1  
Assignment 1: Visualisation**

**Introduction to the data set**

The data set used in the case of visualisation provides SPI ratings (match wise) and foresees the forecasting for every individual football league in the latest season. Following URL has been used to download it.

<https://projects.fivethirtyeight.com/soccer-api/club/spi_matches_latest.csv>

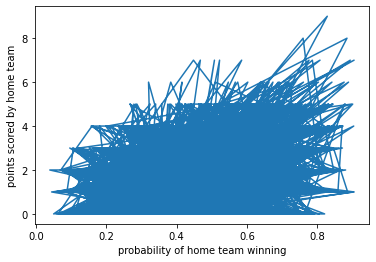
In order to predict football matches, a data set has been used. The soccer power index (SPI) is a variable that is included in a data set. A team’s strength can be measured by this. Visualization has different variables that have the ability to foresee the possibility of the joy of the home team/visiting team, scoring goals, etc.

In this investigation, the following parameters catch the attention. The labels in the company with the explanation of them are stated following:

|  |  |  |
| --- | --- | --- |
| No. 1 | Team1 | With this label home team has been identified. |
| No. 2 | Team2 | Visiting team has been recognized in this label |
| No. 3 | Prob1 | Possibility of the winning of home team |
| No. 4 | Prob2 | Possibility of the winning if the visiting team |
| No. 5 | score1 | real point scores that is done by home team |
| No. 6 | score2 | real point scores done by visiting team |
| No. 7 | xg1 | Prediction of the number of goals that “should” have been scored by home team |
| No. 8 | xg2 | Prediction of the number of goals that “should” have scored by visiting team |
| No. 9 | proj\_score1 | Projected points that has been scored by home team |
| No. 10 | proj\_score2 | Projected points that has been scored by visiting team |

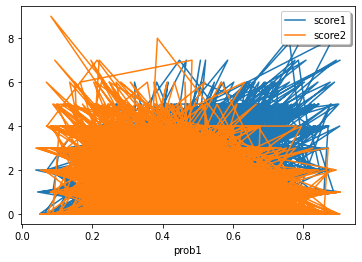
Line Plots:

The initial single line plot is basically plotted in between the prob1 as well as score 1, which is demonstrated in the below figure.

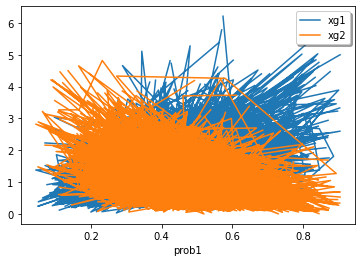


This graph is mainly zigzag format, and due to this reason, no proper information can be obtained from this figure. There are various line plots are plotted, which are shown in the following picture

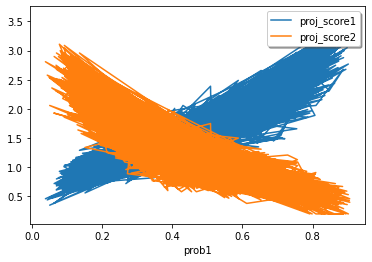
1. prob1 vs score1
2. prob1 vs score2



As these line plots are stated as a complex format, and overlap over other plots, due to this reason, they cannot provide proper meaningful interpretation. Two different multiple line graphs are represented, where the initial one plots Prob1 vs xgl, and other is prob2 vs xg2. There are also some other patterns that are observable. As probability victory for home team facilitates, the evaluated goals to shots (xg2) that are taken is reducing for team visiting, as well as is expanding for the home team (xg1)



Another plot has also been plotted between the *prob1 vs proj\_score1*, as well as *prob2 vs proj\_score2*, which is also shown in the below figure.

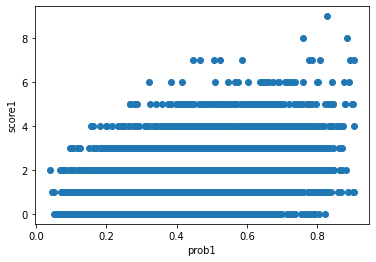


Eleven if there are some patterns that are observable here, since there are different types of lines that are still unclear about how data points are spreaded. Due to this reason, we chose scatter plots

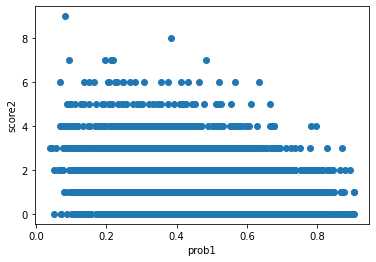
Scatter Plot:

In a scatter plot, every data point is plotted like a single dot, and there will be no line that can join them .

The initial scatter plot is mainly plotted between score 1 as well as prob1, which is figure out in the below section.

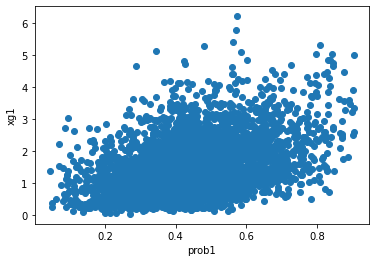


It can be said that as the possibility of the home team winning increases, the points achieved by that team also increase. Line plot cannot help to capture this information.

The second scatter plot is plotted in the middle of prob1 and score2. 

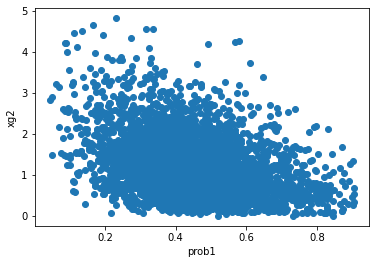
There is no doubt that, as home team victory probability increases, therefore, the visiting team’s points scored will decrease, but this data was not demonstrated by the line plot.

A scatter plot is basically plotted among xg1, and prob1, which is figured out in the below section.



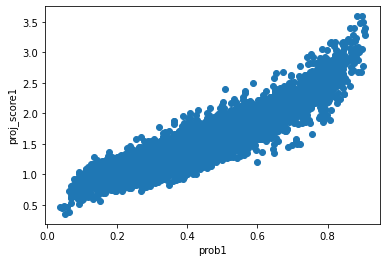
As prob1 increased as expected, therefore xg1 also expanded, and this section is not figured out clearly within the line plots, because it is understandable within the scatter plots.

Another scatter plot is represented between prob1, and xg2, which is also figured out in the below picture.



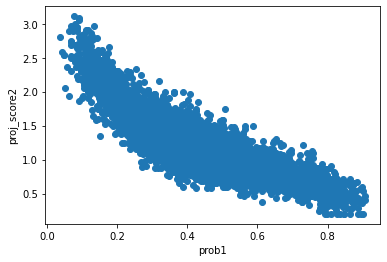
It can be clearly said that according to the possibility of the increase of home team victory, the estimated goals escorted by the respect to shots taken by the visiting team reduced. This information also cannot be captured with the help of line plots.

A scatter plot is plotted in the middle of prob1 and proj\_score1. It is shown below.



As the likelihood of winning for the home team increases, the projected points also escalate for the home team. And one can easily notice that at any given possibility, the proj\_score1 can take a range of values.

A scatter plot is plotted between prob1 and proj\_score1. It is shown below.

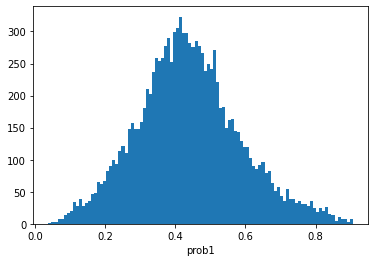


With the increase in terms of the probability of victory for the venue team, project points get reduced for the team visiting the venue. It can also be observed that at any point in time, a wide value can be taken into account by proj\_score1.

Histograms:

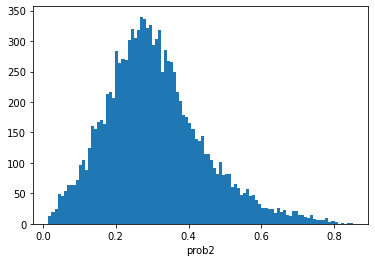
By using this specific feature of Histogram, each value and its occurrence of frequency with consumption has been shown. We intend to make some histograms and take all the relevant information from the source.

It was learned that probability-based plot for histogram of the win for the venue team has been demonstrated in this graph.



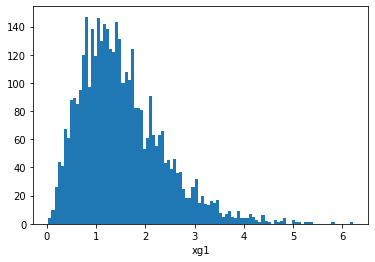
This articulates that most of the teams from venue win approximately 40-45 % in their matches played at their home venue.

Victory for home visiting for the “ histogram plot” has been shown in the above graph.



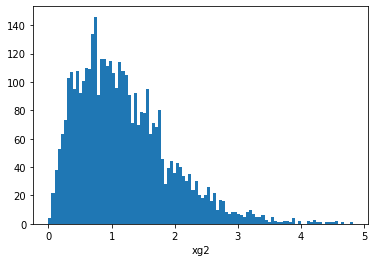
This graph succinctly demonstrates that teams who visit the venue win approximately 25-35 % of each of their matches.

The graph demonstrates that the estimated goals of the histogram plot could have been scored considering the number of shots (Xg1) undertaken for the venue team.



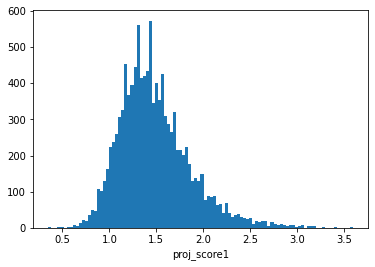
The above figure categorically demonstrates that teams from home venue are likely to make score of 1-2 goals in those league matches to be played at home.

The figure depicts that histogram plots must have scored considering the shots that were taken for the team from other region.



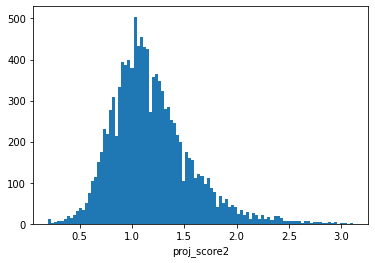
The above graph outlines that teams visiting from other region are likely to make score of 0.5-1.5 goals in all those matches to be played at home.

The above pasted figure demonstrated project points as scored by the team from hone in Histogram plot.



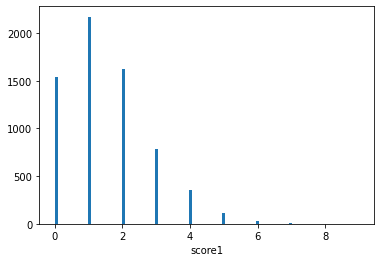
It depicts that teams from home region are capable of scoring 1.5 points in the matches to be played at home region.

Project points and its histogram plot as has been scored by visiting team has been demonstrated in the above graph.



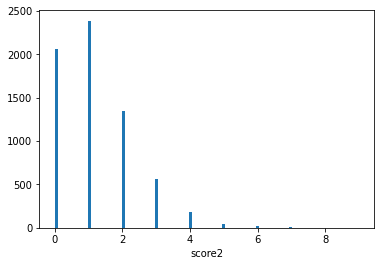
It can be seen that most team visiting from other regions are expected to make good score if 1 point in their league away matches.

Actual points and its histogram plots as have been scored by the team at home has been demonstrated in this following graph.



This critically demonstrates that most of the teams from home are likely to get the score of 1-2 goals in the matches they would play in home.

In the above figure, the plot of histogram in line with actual points that was scored by visiting teams has been shown.



It has been shown that all those visiting means are likely to make score of 1-2 gaols in the matches to be played. Notwithstanding, it can be stated that the specific number for 2 is considerably less compared to the number in score 2. It is apparent that visiting team lost more matches than the home team.