DSC:140S Final Project Report

By: Trevor Arnett

Dr. Julie Butler

DSC:140S Fundamentals of Data Science

12/1/24

For this project I decided to do my analysis on a dataset of the 2018 FIFA World Cup Squads. This data set is important to me because I am a fan of soccer myself and thought it would be interesting to look at all the data and relations within the data. This data set has eleven columns and 735 rows of data. To start off the analysis I first imported the data set into my google colab notebook so that I can access the data and do my research. A computer code with text

Description automatically generated

The image above is the code for importing the dataset into my google colab folder. This code connects the google colab folder to my google drive account where the data file will be stored as a .csv file so google colab can read and interpret it. Then the library called pandas reads the file in google colab and turns it into something that I can then do my analysis on. The next step I did is I printed the data frame in google colab using the code “print(project)”. Which "project” is the name I gave to my data frame. This prints out the whole data frame so I can see all of the columns and some of the data and decide my next steps of analysis. The first statistical analysis I decided to do was find the average age of all of the players in the world cup, and the average amount of goals a player had.A black rectangle with green text

Description automatically generated

The code above is how I found the average age of a player in the world cup. The code takes my data frame called project, extracts the ‘Age’ column, then takes the average of all the ages and prints the answer below it. That answer is 27 years old.A screen shot of a computer

Description automatically generated

The code that I used to find the average amount of goals per player is the same as I did to find the age. The only difference was I took the ‘Goals’ column instead of the ‘Age’ column and took the average of it. The answer of that being 4 goals per player. After that was done I decided to make another statistical test to double check the averages I just took as well as take the average of one other column. A screenshot of a computer

Description automatically generated

This test provides the statistics for all continuous data in the data set. The column ‘Squad Number does not matter because that just tells what position on the field and on the roster the player is. So the only columns that matter are the ‘Age’, ‘Caps’, and ‘Goals’ columns. It can be seen that the mean number for age and goals are the same as the test from before so that can be verified to be the right number. The other column I wanted to see the average for is the ‘Caps’ column. A cap in soccer is the number of international games a player has played in, so I thought finding the average amount of caps throughout all of the players would be interesting since there are is a wide range of ages in the world cup. The average of this column was 34 caps. A screen shot of a computer program

Description automatically generated

This is the image for the code to graph the scatter plots for the columns of “Age vs Caps”, “Age vs Goals”, and “Caps vs Goals”. This code takes my data frame and calculates the Spearman correlation and Pearson correlation, as well as the linear regression equation and the Smirnov test statistic. All of these come together to create these graphs; A screen shot of a graph

Description automatically generated

This scatter plot shows that the older age a player is the more caps they have which should make since seeing how caps are how many international games a player has played. Meaning it makes since that the older players have more caps.A screen shot of a graph

Description automatically generated

This graph shows the amount of Caps vs Goals, this graph makes since because as the amount of caps go up so should the goals, but the points that are on zero the whole time should be the goal keepers and defenders.A screen shot of a graph

Description automatically generated

This graph shows that there is no relation between age and goals because the age doesn’t necessarily matter it’s the position of the player on the field that matters. There can be a young defender that has no goals but an older attacker that has a lot of goals, likewise there can be an old defender with no goals but a young attacker with a lot of goals, the age doesn’t matter it’s the position on the field.A computer screen shot of a program code

Description automatically generated

This Chi-squared test proves that the position on the field matters for how many goals a player has, not the age or caps.A computer screen with text

Description automatically generated

This code is called a test train split test, this test predicts a certain category that you feed it. I chose to have it try to predict the number of goals a player had based off their caps and age. The test takes a percent of the data and uses it to rain the model. Then the model takes a prediction from the data it was given and tests for its accuracy. The result of the test was that the machine learned model was 34% correct. I was lastly tasked with created two new lines of code that we have not used in class, first I chose to use a bubble chart to display goals at every players age. A graph with dots on it

Description automatically generated

This is the bubble chart that shows the amount of goals each player has a every age. The last chart I made was a strip plot to show the amount of goals every country in the world cup had. The chart looks as follows;A screenshot of a computer

Description automatically generated

This chart shows how many goals every team in the world cup had at the end of the world cup.