CA682 Data management and visualisation

Name	Olabode Cole
Student Number	15107591
Programme	M.Sc. in Computing
Module Code	CA682
Assignment Title	Data Visualisation
Submission date	17/12/2018
Module coordinator	Suzanne Little

I declare that this material, which I now submit for assessment, is entirely my own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my work. I understand that plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should I engage in plagiarism, collusion or copying. I have read and understood the Assignment Regulations set out in the module documentation. I have identified and included the source of all facts, ideas, opinions, and viewpoints of others in the assignment references. Direct quotations from books, journal articles, internet sources, module text, or any other source whatsoever are acknowledged and the source cited are identified in the assignment references. This assignment, or any part of it, has not been previously submitted by me or any other person for assessment on this or any other course of study.

I have read and understood the referencing guidelines found recommended in the assignment guidelines.

Name:_____Olabode Cole_____ Date: ___17/12/2018___

Data Management & Visualization

Introduction

The question being explored here is the probability of predicting which NBA basketball teams are more likely to not only reach the conference finals but win the league based on their current and over the year's statistics. Therefore, statistics such as teams performance over the years and how they have improved will be visualized, the accuracy of certain players within specific teams will also be visualized. Finally, the geolocation of each team along with the total amount of wins will also be displayed including the latitude and longitude of their location.

Datasets

The dataset utilized to implement the data visualization was gathered through google dataset and also from NBA Official sites. A total of seven datasets was utilized in creating the visualization, six of that dataset being the overall stats of each team for the past six years which includes the total amount of games played, wins, losses and total amount of 2-3 Points made per game and the seventh dataset being the over shots from each player during the All-star games.

After the end of the season, there's an event that occurs among the players which are known as the All-star games, within this event several players are selecting from each team to represent the region they are currently playing in whether it's the East or West Conference. The players selected are classified as All-Star due to their tremendous performance throughout the season and within that Event, the overall Most Valuable Player (MVP) of the season is selected.

Therefore, I decided to include the statistics of each and every player that participated within the All-star event because by analysing the efficiency of these players in terms of their productivity on the court it would be a great way to predict which teams are more likely to proceed further within the league as certain teams have more All-star players than others.

For an example, basketball starting line-ups is usually consisting of 5 players however there are certain teams Like the Golden State Warriors that has 5 All-start players for their starting position which means they would be a lot more proficient against other teams that have less All-star players but, there have been scenario where just a single All-star player within a team have managed to carry their entire team not only to the Conference finals but to win the overall league. Dirk Nowitzki is the only All-star player leading the Dallas Mavericks to their first NBA title in 2011 and (King) Lebron James doing the same with Cleveland Cavaliers in 2016.

Process

Before importing the dataset there were several alternatives that needed to consider in terms of the state of the data whether it's clean or not because a dataset consist of errors would lead to the inaccurate reading of data making visualization a little more difficult than it should be and wrong representation of data. Therefore, I decided to explore further beyond the techniques and tools we have been taught and utilizing tools such as CSV Kit.

CSV Kit is a free set of tools for validating and cleaning files like common syntax errors. The dataset utilized did not really consist of crucial errors but just to be sure I decided to run them through the CSV kit tool. As mentioned above I had to join several datasets which include the overall statistics of

the teams for the past couple of years, however, I had to manually merge those datasets together. CSV Kit join provides arguments for easily merging CSV files but because the position of each and every team varied over the years based on their stats it would have resulted in the inaccurate dataset if I was to merge the dataset using the CSV join the argument.

CSV toolkit is a useful source of data cleaning however it's functionalities is quite limited, therefore I decided to install a software called Data Cleaner. Based on the data cleaning tool provided by this tool it made cleaning and analysing the data a lot easier with just a simple click of buttons.

I created a new file by uploading the particular CSV file I was interested in and from the analytical function, I was able to select certain attributes such as duplicate detection, unique key check. After doing so a visual representation of the result based on the selected field is given. I found this particular tool very efficient as it gave me a better understanding of fields I'll need to avoid if I can't completely delete the field as it still represents some statistical data regardless of its duplications.

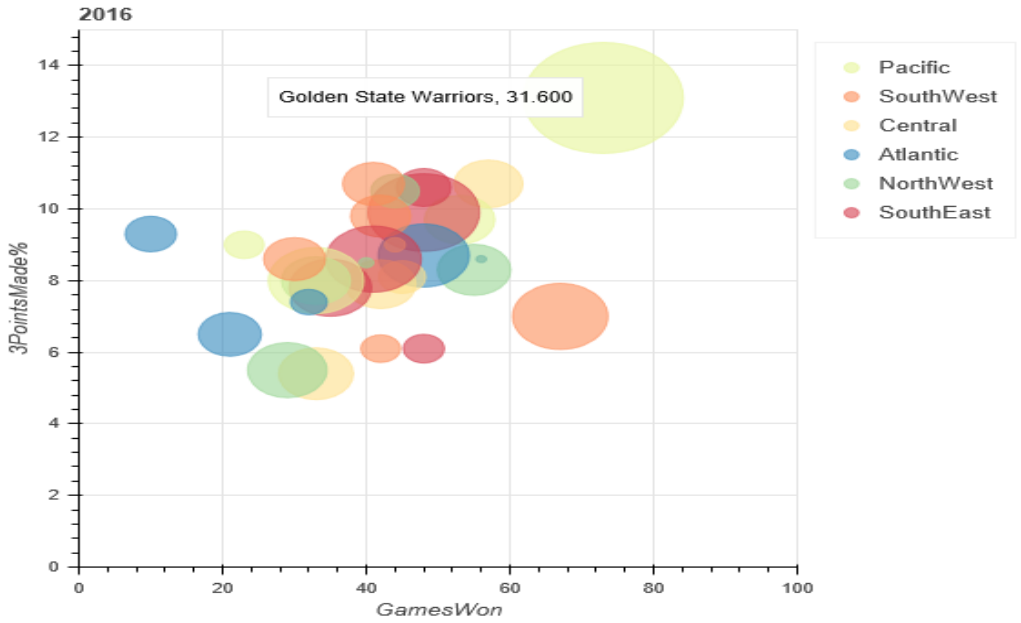
Tools

For representing the data visually, I utilized python but two different Python Libraries which are Bokeh, and Plotly. I used Bokeh to create an interactive scattered bubble chart consisting of all the teams and how they stats varied over the years. To visually represent the total amount of shot taken by the players during the All-Star games I had to create a shot chart that shows where the shots were taken from and whether it was made or not. The last visualization chart is an interactive map that displays were each and every team in the NBA are located within United Stated while displaying their names and statistics such as the total amount of games they have won since they have been in the League. These charts will be discussed in further details broken down into several sections later on in the report.

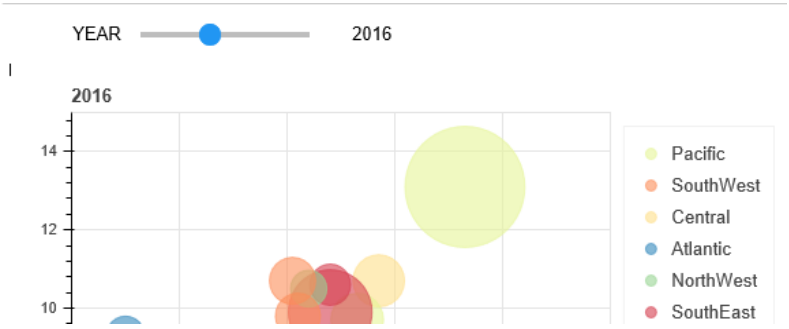
Results

Here is a visual representation of the scattered bubble chart that consists of all the team in the NBA for the past six years. The X-axis represents the total amounts of games won within a given year while the Y-axis represents the average percentages of the 3 points made per year. The reason why I decided to visualize these particular statistics is that by analysing the total amounts of games won you can easily tell whether a team is improving or their performance is decreasing over the years and by also analysing their 3 points average ratio. This is because it is quite challenging to get into the Paint, also known as the free throw range or rim in order to score an easy 2 points layup or try to draw a foul 2 free 1 point shots. Therefore, teams often try to shoot 3's from downtown or outside the semi-circle as it results in more point, less hassle compared to getting into the paint but requires more technical skills.

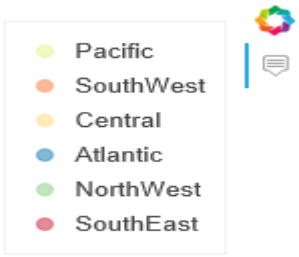
The over tool within this chart acts as an identifier so if the user were to hover over a specific bubble it will display the name of the team and the percentage of the 3Point Attempted per year. Note that this is different from the average amount of 3 pointers made per year.



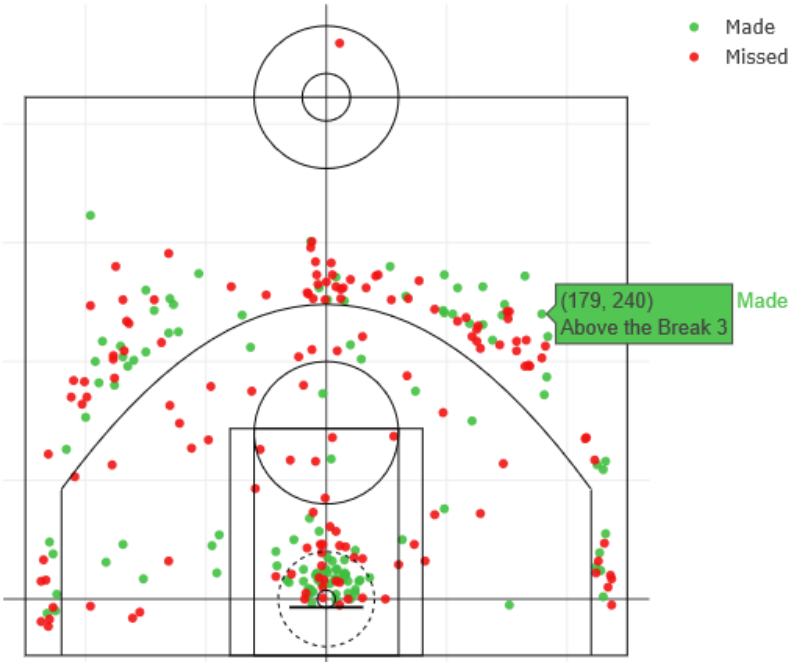
The interactions within this charts are interactive slider and cursor hover tool. Since the dataset used to design this chart consist of 6 different years, I included a slider that allows the users to select a particular year they are interested in, however, the bubble scattered chart moves along to the selected year.



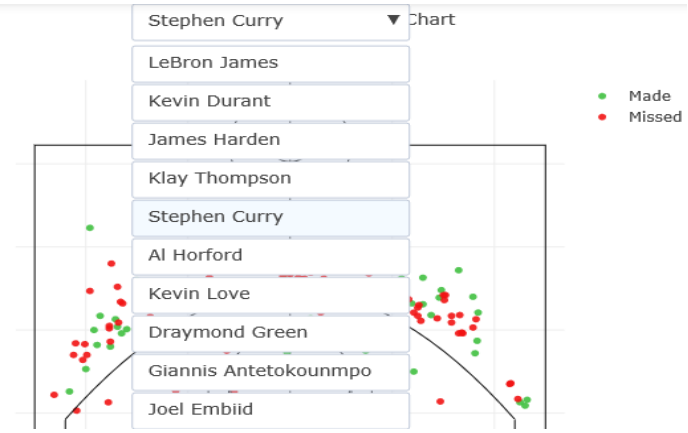
There are 2 different conferences within NBA which are East & West as mentioned above, however, within both conferences, there are 6 regions known as South-west, North-west, Pacific, Central, South-east and Atlantic. These regions are colour coded with each region having different colour cues. If you were to analyse the bubble scattered chart the bubbles are also coloured, but however the colours the bubbles are displayed in corresponding to the colour of the region the team is from whether its North-west or South-east.



The chart displayed here is the representation of the number of shots taken by a select player during the All-Star games at the end of the season. This chart pretty much speaks for its self as its very easy to understand what's going on due to its simplicity but yet efficient data visualization.



The main colour cues within this chart are red and green. Red indicating that the shot was missed by the player and Green indicated that the shot was Made. Once again I applied the hover tool, that also displayed whether the shot was made or not and the location the specific shot came from. Above the break indicating a shot from outside the 3-point mark, In the Paint indicate shots taken from the free through range etc.



Things I'll do differently

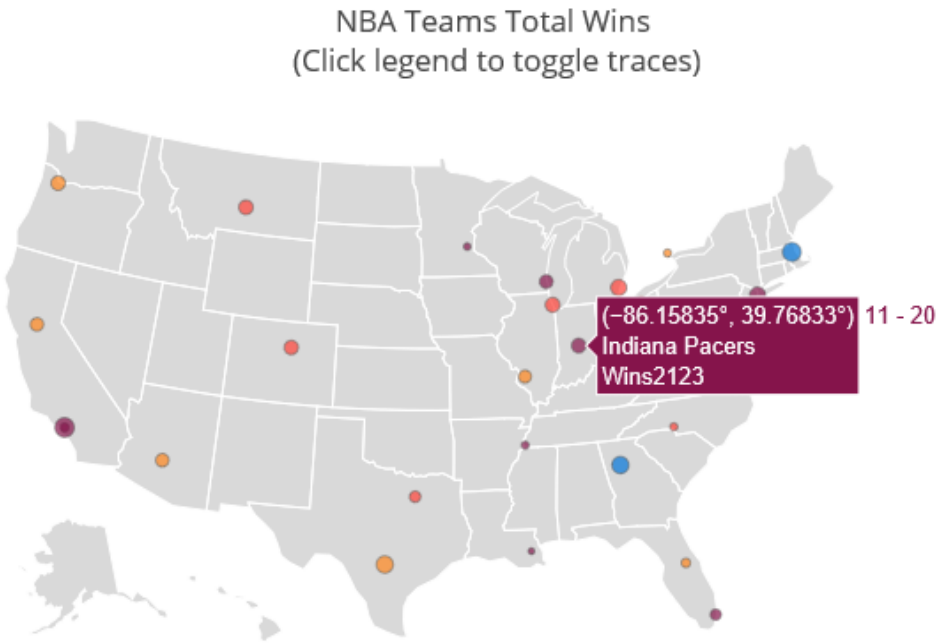
If we didn't have such tight deadlines and if I was to continue working on the chart. I would like to implement a slider that would allow the users to either display only the specific shots that they are interested in (Made, Miss). I know this specific statistic isn't displayed within the chart but I would also like to allow the users to be able to select certain shots taken by the player based on certain quarters as they are 4 within a basketball game but the shots being displayed the total shot within the game.

Technical Difficulties

The functionalities I was having issues with implementing was adding the picture of the selected player for better visual representations because not a lot of people are good with names compared to how they are with faces. So I believe a visual representation of the player's images would have been really interesting.

Third Chart

Here is the 3rd and final chart within my project and it's a choropleth map visualization of all the teams within the NBA. The map displays were each team are from based on their region within the united states. The interactions within this chart is that it allows the users to zoom in and out of the maps and if the users hover over certain teams it displays the details such as the longitude and latitude of the team locations, the name of the team and the total amount of wins accumulated by the teams since they have been promoted to the league.



If you look closely within the dataset utilized to implement this chart there is a column know as row, which means each and every team are in a specific row so, therefore, I created a set of colours that would group the teams within a specific range (0,2), (3,10), (11,20), (21,50) and (50-3000). So teams within the range 3-10 would have the specific colours initialized for that range. Since there are only 32 teams in NBA only 4 colour cues were displayed and the last range (50, 3000) was discarded.