Levi Nickerson
CSCI 4830
109340569

Project Initial Design

Project Type:

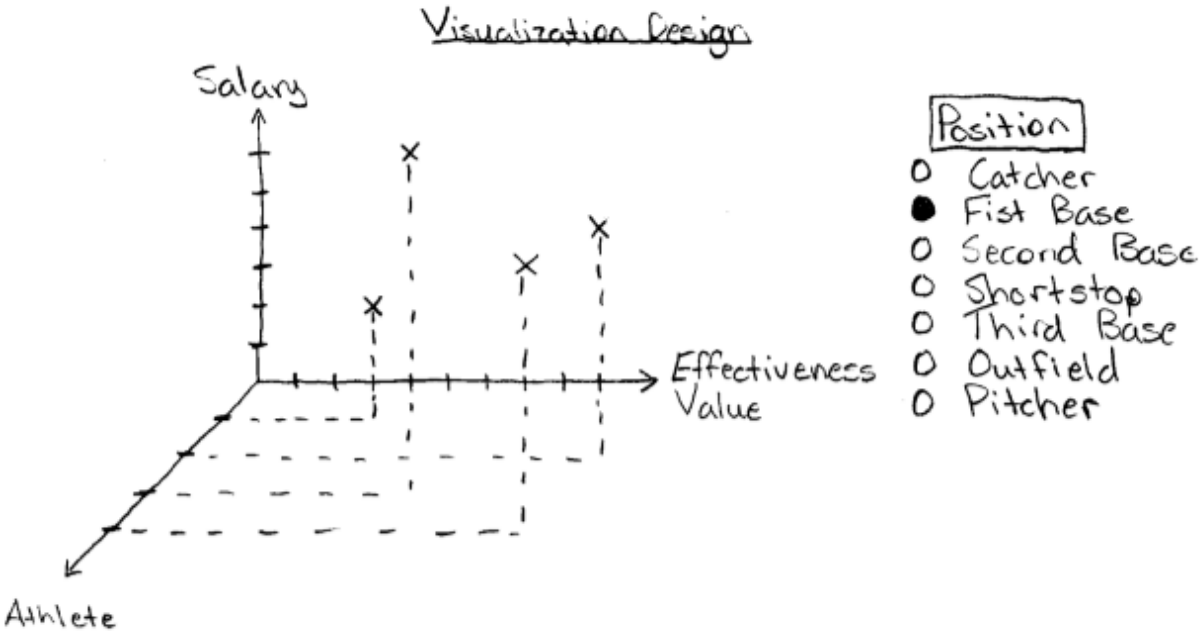
I am working on a programming project. The overall goal of this project is to develop a program that aids the user is selecting a lineup for daily fantasy baseball. The program will produce enough information for the user to make an educated selection.

Design:

The overall task of the project is to develop a program that aids in the selection of a contending lineup for daily fantasy. In terms of what we have learned so far the primary task I am concerned with is analysis. Within the analyze action I am venturing into the subsets of consume and produce. I am producing a derived attribute and then consuming the information that the attribute displays. I am deriving my own attribute from a dataset for analysis. Once this attribute has been derived I am presenting it in a way that allows the user to look for outliers. Identifying outliers is at the core of what this program does. It is intended to highlight athletes who have a low salary in comparison to their effectiveness value. This effectiveness value is the derived attribute that is dependent on many different factors. The secondary task that is most important to me is to enjoy the visualization. I am an avid sports fan and have always been fascinated by the numbers behind sports. This project is an enjoyable way for me to dive deeper into baseball and see how well I can predict outcomes. I also believe that there are many sports fans that are interested in a project of this realm.

To accomplish these tasks I am focused on two different visualizations. The first design is a 3D visualization comparing the salary and effectiveness value for each athlete. The salary value for each athlete is pulled from a data set produced from the daily fantasy company. The effectiveness value is what the program will compute. This value is dependent on many different statistics and can be manipulated in the second visualization design.

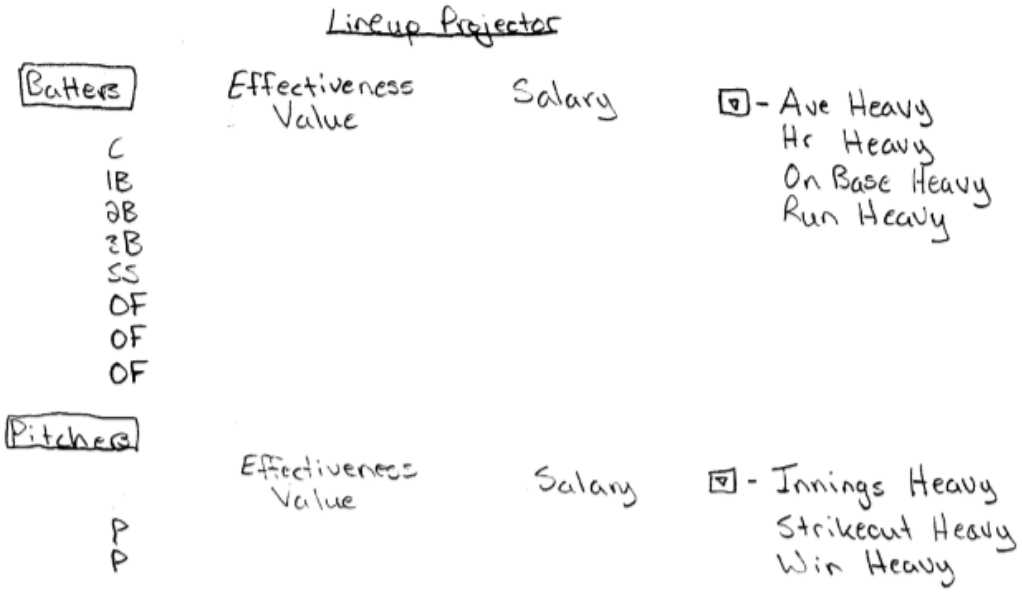
Figure 1 – Visualization 1



The program will initially ask the user for an input. This input should be the position of interest. The input options are the positions in baseball: catcher, first base, second base, shortstop, third base, outfielders, and pitchers. Once a position has been selected the program will determine the athletes at that position that are playing on the current day. The visualization will then be generated and the user will be able to highlight athletes that have a high effectiveness value and a low salary. In terms of tasks, this visualization is concerned with producing and consuming. It is also the visualization that is used to identify outliers. Finding outliers is the core of the program and this visualization provides enough information to correctly determine these outliers.

The second visualization will be determined by user inputs. This visualization is intended to put the decision making into the user's hands and let them determine the experience. It will produce lineups that are projected to be successful.

Figure 2 - Visualization 2



The user will input what kind of lineup they wish to see. Currently the inputs for batters are: Average Heavy, Home Run Heavy, On Base Heavy, and Run Heavy. These options are meant to cater to different experiences. For example, a user might select a "Home Run Heavy" lineup if they are planning on watching games with friends and want to have a lineup fueled by excitement. A "Home Run Heavy" lineup has a greater potential of hitting home runs, something that is very exciting to watch. There are also selections for pitchers: Innings Heavy, Strikeout Heavy, and Win Heavy. Do you want to use a pitcher who usually has long outings (pitching 6+ innings) or are you more concerned with pitchers that get the win? This portion of the program is intended to be enjoyable. Predicting the outcome of sporting events can never be done with 100 percent accuracy. Knowing that, this program focuses on informing the user while also let them enjoy the process of selecting a lineup.

For future possible designs I have an idea to include another data set. Baseball is a very interesting sport in that consistency is most desired, but most difficult to achieve. Batters can be hot for a week and suddenly fall off and struggle for weeks. Similarly, pitchers can have an incredible outing (such as a no hitter) and immediately follow it by getting run out of the stadium in 2 innings. The data set I would consider adding is the current hot and cold streaks for each athlete. The current data set focuses on the athlete's consistency for the entire season up to date. The new data set would look at the local consistency for each athlete. This would allow for better predictions by further highlighting low salary athletes that have potential for successful outings.

Scalability is not necessarily an issue that I am concerned with. There are only 30 teams in the MLB. This means that at most there are 15 games going on for a given day (this is an uncommon occurrence). Each team has one starter for each position which means that for a given position there is at most 30 athletes to consider (except for the outfield which is grouped and results in around 90 potential athletes). Having 30 data points in a 3D visualization is not overwhelming. Especially considering that the user is primarily considered with a specific region within the visualization. The user is looking for athletes that have a higher effectiveness value for a lower salary. However, if the visualization does become overloaded with clutter an easy solution is attainable. I would allow the user to set bounds on the output. The user could input an effectiveness value that would limit what athletes are displayed.

The second visualization is not very dependent on scalability. The only issue that arises would be run time. However, as I mentioned prior the maximum number of athletes to be analyzed is relatively small. In essence scalability is not a primary issue of the project because of the nature of the dataset. The MLB is a professional sports league that does not change often. Teams are rarely added to the league and the number of games played has stayed relatively constant. If more teams were added and the number of games played in a season was increased scalability would start to become a concern.

Infrastructure:

I am currently building the code in Python. From previous experiences I would rather be working in Matlab (a coding environment that I have substantial experience in). However, Matlab is not a free programing environment and I do not need the intense mathematical possibilities. Python is more than capable of handling the mathematics behind the program it will just take me time to learn the certain things that I will need. I need to figure out how to properly produce a 3D visualization. 3D visualization is a challenging aspect that will take research to determine an acceptable module that can properly plot the visualization I am picturing.