Project#5

Original Dataset URL: https://www.kaggle.com/dansbecker/nba-shot-logs

Dataset Preparation

The dataset "shot_logs.csv" of the link above has 21 attributes and 128069 data points. It is the record of all shots during NBA 2014-15 season. In project#2, I have reduced the number of data points to 945. For this project, I picked 5 attributes that I found interesting during the project#3: DRIBBLES, TOUCH_TIME, SHOT_DIST, CLOSE_DEF_DIST, and FGM. The dataset using this 5 attributes is "NBA_shot_logs_used.csv" that you can find in the folder I uploaded. I don't use 10 attributes as I did in the project#3 because the visualization becomes too big to fit in a webpage.

I decided to use 4 visualizations: bar chart, PCA plot, MDS plot, and scatter plot matrix.

The attribute FGM, Field Goal Made, is the only categorical data. It is "0" or "1". "0" of FGM means that the shot was missed. "1" means that the shot was successful and got the points. I decided to represent the FGM by color on each plot. PCA, MDS, and scatter plots of the plot matrix would have points with two different colors, depending on the point's FGM value. Thus, I used 4 attributes, DRIBBLES, TOUCH_TIME, SHOT_DIST, and CLOSE_DEF_DIST, for getting PCA vectors and projections, and getting MDS points.

The python program "gen_eigval_and_eigvec_csv.py" gets the file input, "NBA_shot_logs_used.csv", and generates the two file outputs, "eig_value.csv" and "eig_vector.csv". "eig_value.csv" contains the eigenvalues of the 4 attributes: DRIBBLES, TOUCH_TIME, SHOT_DIST, and CLOSE_DEF_DIST. Accordingly, eigenvectors are in "eig_vector.csv".

The python program "gen_2d_projected_points_csv.py" gets the two file inputs, "eig_vector.csv" and "NBA_shot_logs_used.csv". By manually looking in the "eig.value.csv", I figured out the top two eigenvectors. I read and used the vectors from "eig_vectors.csv" for projecting the 4-dimensional points from "NBA_shot_logs_used.csv" to 2-dim points. The dataset actually has 5 attributes, including FGM, but as I said previously FGM is ignored when doing PCA and MDS. The program finally generates PCA projection points and write the output file "2d_projected_points.csv".

The python program "gen_mds_points.py" gets the file input, "NBA_shot_logs_used.csv", and generates the file output, "mds_points.csv". The program uses sklearn to compute MDS points.

With those PCA and MDS datasets, I consolidated all the data into one file: "NBA_shot_logs_extended.csv". The file has the 5 attributes mentioned above, and also has 4 more attributes: pca_x, pca_y, mds_x, and mds_y that came from the PCA and MDS datasets. I used this one dataset for the d3 visualization of "index.html".

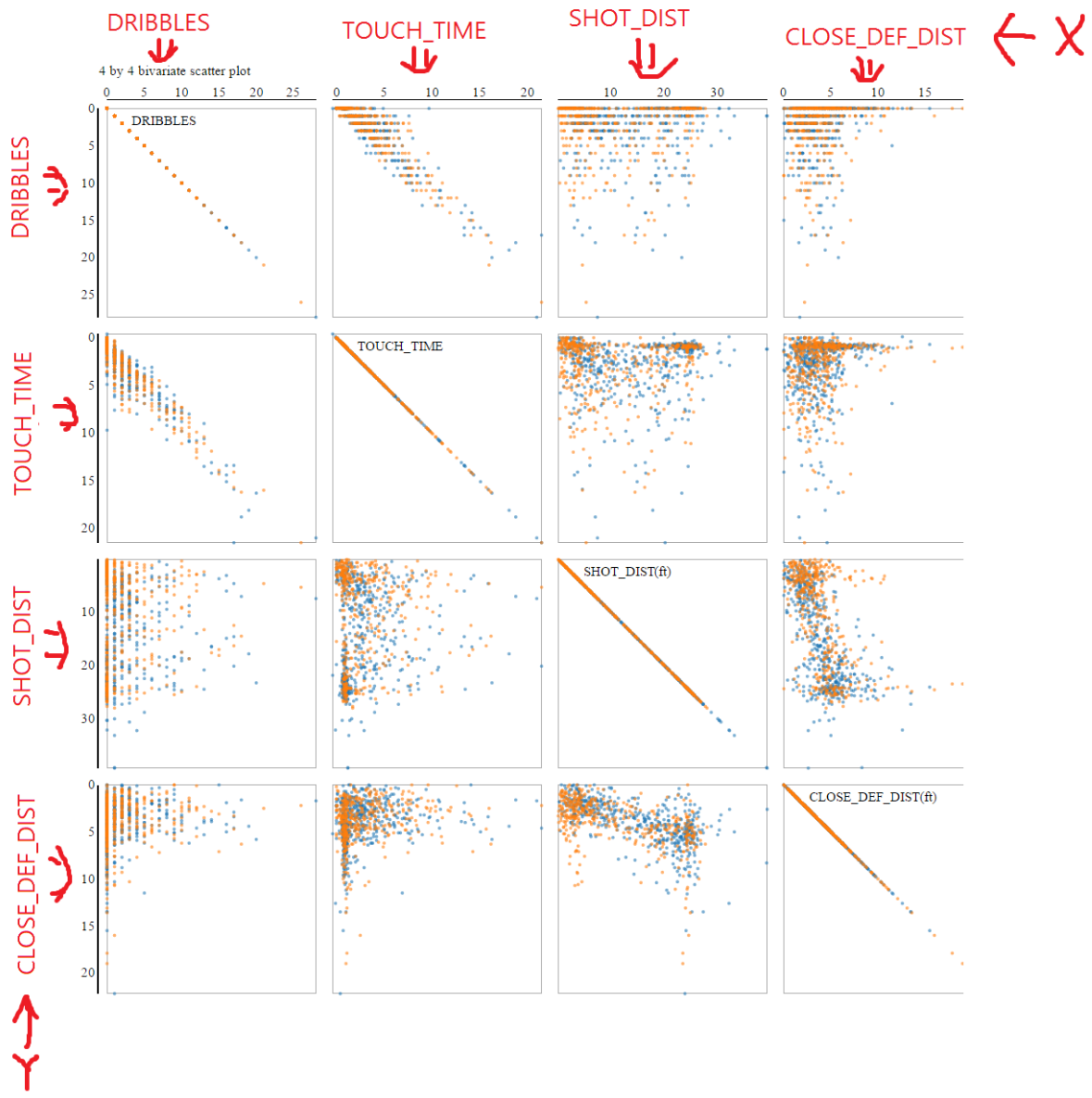
Visualization

For the scatter plots, PCA plot, and MDS plot, you can drag your mouse to brush and select data points. Only selected points are colored blue or orange. The points not selected are gray. The orange points are successful shots, and blue points are missed shots. To reset, just click anywhere on the one of the plots.

You can resize the selection box by dragging a border or an edge of the box.

The bar chart has two bars: success and missed. If you click one of the bars, e.g. success one, the all the data points of successful shots will be selected. If you check the bar chart while you are brushing, you will see the blue- or orange-colored bar is going up and down according to how many success- or missed-shot data points you selected. The number of points selected are texted above each bar.

Let me provide a scatter plot matrix axis guide below:

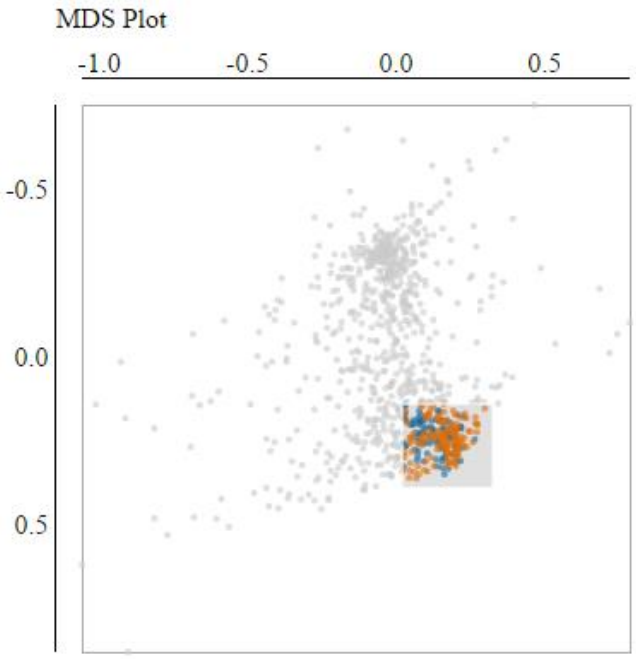


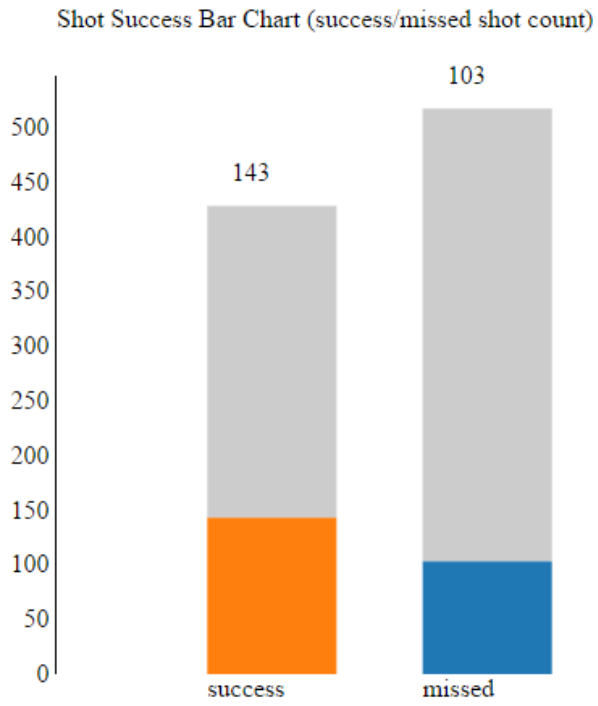
Data Stories

1. 2-point shots and 3-point shots form a cluster for each, and 2-pointers are far more likely to be successful.

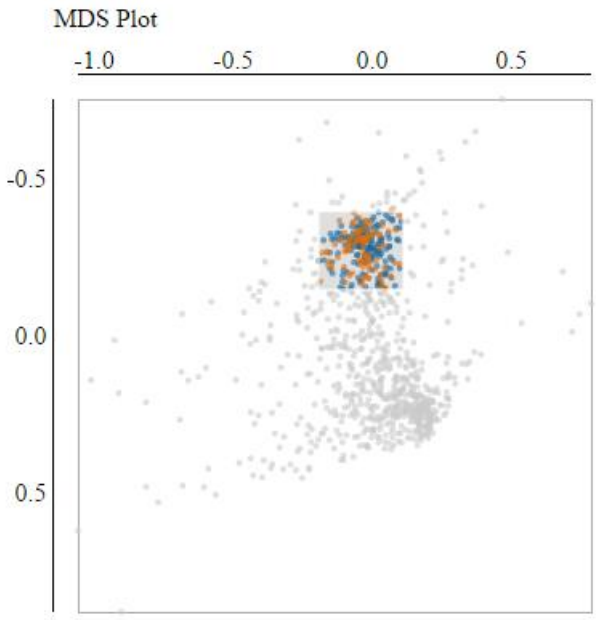
Take a look at MDS plot. It seems like data points are roughly forming two clusters.

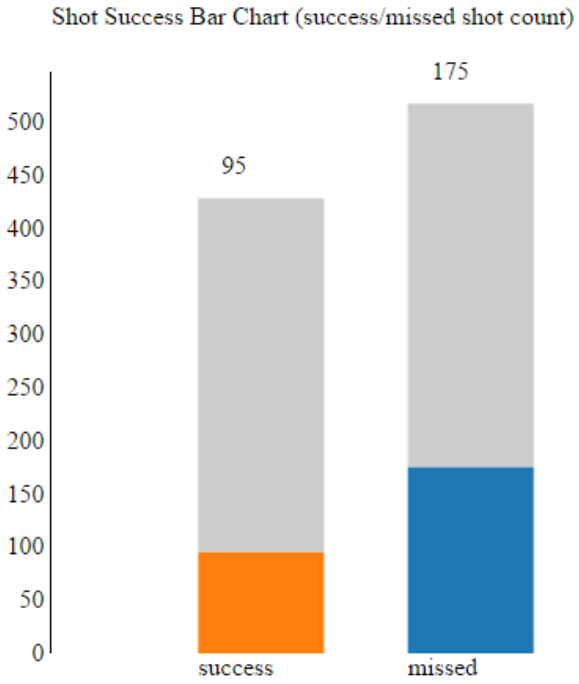
Here is the first cluster selected:





Here is the second cluster selected (same selection box size as the first one):



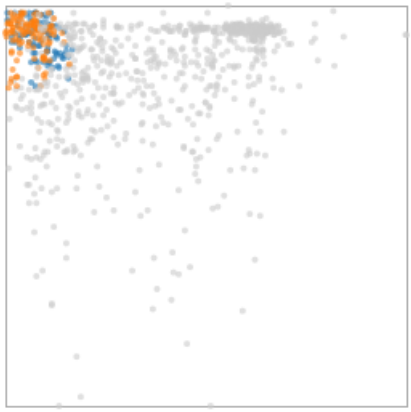


Let’s name the first cluster C1 and second cluster C2.

We can see that C1 is somewhat dominated by the success data points; about 35% of success points are gathered around in C1. C2 is opposite; about 35% of missed-shot data points are gathered in C2.

The distance of two points in MDS plots represents how different the two records are. If the two points are in the same cluster, it means the two points tend to have close values for each attribute. By the visualization above, we can assume that successful shots in NBA tend to happen in certain condition, likely to be cluster C1. For example, one successful shot in C1 has the 3 feet for the closest defender distance attribute. The other one in C1 should have similar value for the attribute to be in the same cluster. If we check the whole points in the C1, those points will have similar properties.

So what is C1 cluster? Let’s check the (x=SHOT_DIST, y= TOUCH_TIME) scatter plot in the plot matrix when C1 cluster is selected.



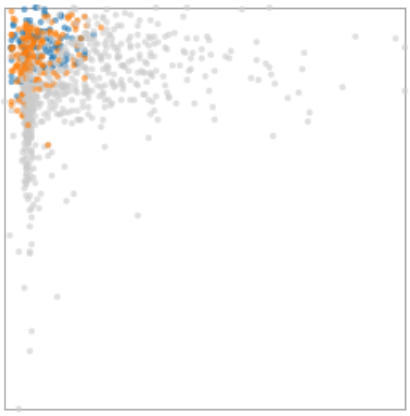
Now, here is the same plot when C2 cluster is selected:

If you check the axis ticks of the x=SHOT_DIST, you will find the C2 cluster is located around 23 feet, which is the 3-pointer line distance in NBA.

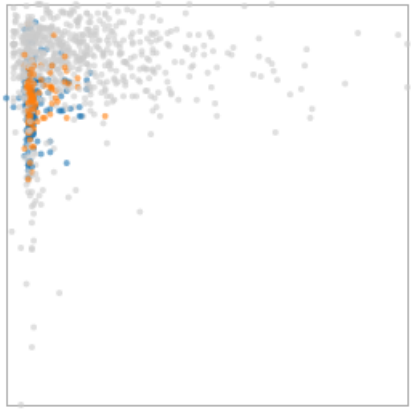
So, C1 contains all the close-distance shots, and C2 contains most of 3-pointers. And that means that 3-pointers are more likely to be missed.

2. Statistically, if a defender is closer, the shot is more likely to be successful.

Let's use the same clusters from the data story 1. I said C1 cluster is dominated by successful shots. With the cluster selected, here is the (x=TOUCH_TIME, y=CLOSE_DEF_DIST) bivariate scatter plot of the plot matrix.



Now, let's select the C2 cluster, which is dominated by missed shots, and see the same (x=TOUCH_TIME, y=CLOSE_DEF_DIST) plot.



x=TOUCH_TIME didn't change much, but there is obvious change in y= CLOSE_DEF_DIST. We can say points in C1 tend to have small closest defender distance, and points in C2 tend to have big closest defender distance. In other words, successful shots are likely to be made when the defender is close to the shooter, and the shots tend to be missed if the defender is far away. This is a strange outcome.

Here is the explanation for this strange outcome. It is because of the 3-pointers. The 3-pointers are more difficult to make in general. If you watch NBA, many players usually try their 3-pointers when it is safe to throw. While a team is passing around the ball, there comes a free-defender moment for the player outside the 3-point line. That is the condition that most players like to throw the 3-pointers. So the defender is far away, and the shot is less likely to be made because the shot distance is far.