CSE 516 Project Proposal (Marques and Ramiro)

Question to explore – How does the performance of Azure w/ Spark compare to a regular Azure SQL database?

-**Our Dataset: Shots taken by NBA players during the regular season**

- NBA Shot Dataset and scrapping code
- Data Sample (1 player/season combination)
- List of shots taken by NBA players during the regular season (data starts in 1996)
- The shot data consists of one CSV made by vertically joining several smaller files, all with the same schema, that have all of the shot data for each player/season combination
- Format: CSV, to be imported using any native Azure tools

-**Queries to compare across regular Azure SQL databases vs Azure with Spark:**

1. Whether popular "regions" of the court to attempt shots for have shifted over time
2. What is the average field goal percentage for home teams vs away teams? How has this changed over time?
3. What players attempt the most 3 point shots? What percentage of the time does each player prefer Layup Shots, Jump Shots, Slam Dunk Shots, etc.?
4. What section of the court (as determined by SHOT_ZONE_BASIC and SHOT_ZONE_AREA) does each player shoot the most from? Which section does each player have the highest field goal percentage?
5. What is the average shot distance per period?
6. What team has the highest field goal percentage for any team in any year? Which has the lowest?
7. What is the average points per attempt for each player with a minimum of 300 attempts? Who had the highest return value each season in the league (of players who had a minimum of 300 attempts)? Who has the highest average points per attempt during the last 5 minutes of a game (minimum 200 attempts)?

-**What we hope to report**:

-Three box plots comparing the runtimes of non-optimal SQL queries, optimal SQL queries, and queries with Spark

-Two Polynomial Regression lines (one for spark, one for no spark) from a random sample of data, comparing the size of the dataset vs the runtime, using a randomly selected dataset size and a fixed query

-Three separate two sample t-tests comparing the runtimes of a Spark query vs No Spark, using a fixed dataset sample. The three queries will vary in complexity from simple, average, and complex, and we want to compare there are meaningful differences in runtime for each query.

-Average runtime for Spark queries with different numbers of joins

Agreement for splitting work:

-Ramiro does queries 1, 2, 5 and 6, and will work on system logistics (getting the data prepped, and up on the DBs)

-Marques will generate queries for question 3, 4, and 7, and also work on the statistical reports