# MAT 243 Project One Summary Report

Tyler Ellis

Tyler.Ellis1@snhu.edu

Southern New Hampshire University

## Introduction: Problem Statement

The problem we are trying to address would be “How can we, if possible, analyze and enhance the performance of the team based on historyical data?” The objective is to to study patterns the performance data and compare it to a rival team’s data, all within a specific time window to keep it accurate to eachother. The end goal is that by **pinpointing and using certain strategies and trends, it will help improve the team's overall performance and give them an overwhelming edge over the other teams.**  
  
 **We will be utilizing two dataset for this. The first data set we will be using for this is the scores and statitstics(metrics) for the great Chicaco Bulls during 1996-1998. The second data set we will be using sill be for our team, the greatest team, The Celetics during 2013 – 2015. Both of these datasets come from the NBA, if you didnt know. Some of the data included in these datasets range from Scores to game locations. No metric is to be overlooked. We have a very wide range of data that we can pull from all the way from 1947 to 2015. For this breakdown we will call the Chicago Bulls the “Assigned team” and the Celtics “Our Team”.**  
  
 **To analyze the data, we will be utilizing a combination of standard descriptive statistics, Data Visualization and Caparitive Analysis. Descriptive statistics will be used to gather the central tendencies, variations, and distributions of specific metrics pertaining to performance, meterics like Elo ratingins and points scored. If you would like to know more about Elo Ratings and how they are used in Basketball, this article has a fairly decent breakdown:** [**How We Calculate NBA Elo Ratings**](https://fivethirtyeight.com/features/how-we-calculate-nba-elo-ratings/) **To do this, we will be calculating statistics such as mean, median, standard deviation, and variance. Then we will use Data Visualization methods to build graphs and to assist in determining any sort of trends that may not be visable by staring at raw data. We can then use visualizations such as box plots, line graphs, histograms to analyze the spread of different metrics over the years and between the teams. Additonally we can use Scatter Plots to possibly determine any outlier games to which we can analyze the data to see if anything significant happened during that time which may have drastically effected the teams. The Comparative Analysis will be happening throught the course of this as we will be continously comparing the metrics and statistics between the two teams.**

## Introduction: Your Team and the Assigned Team

Table 1. Information on the Teams

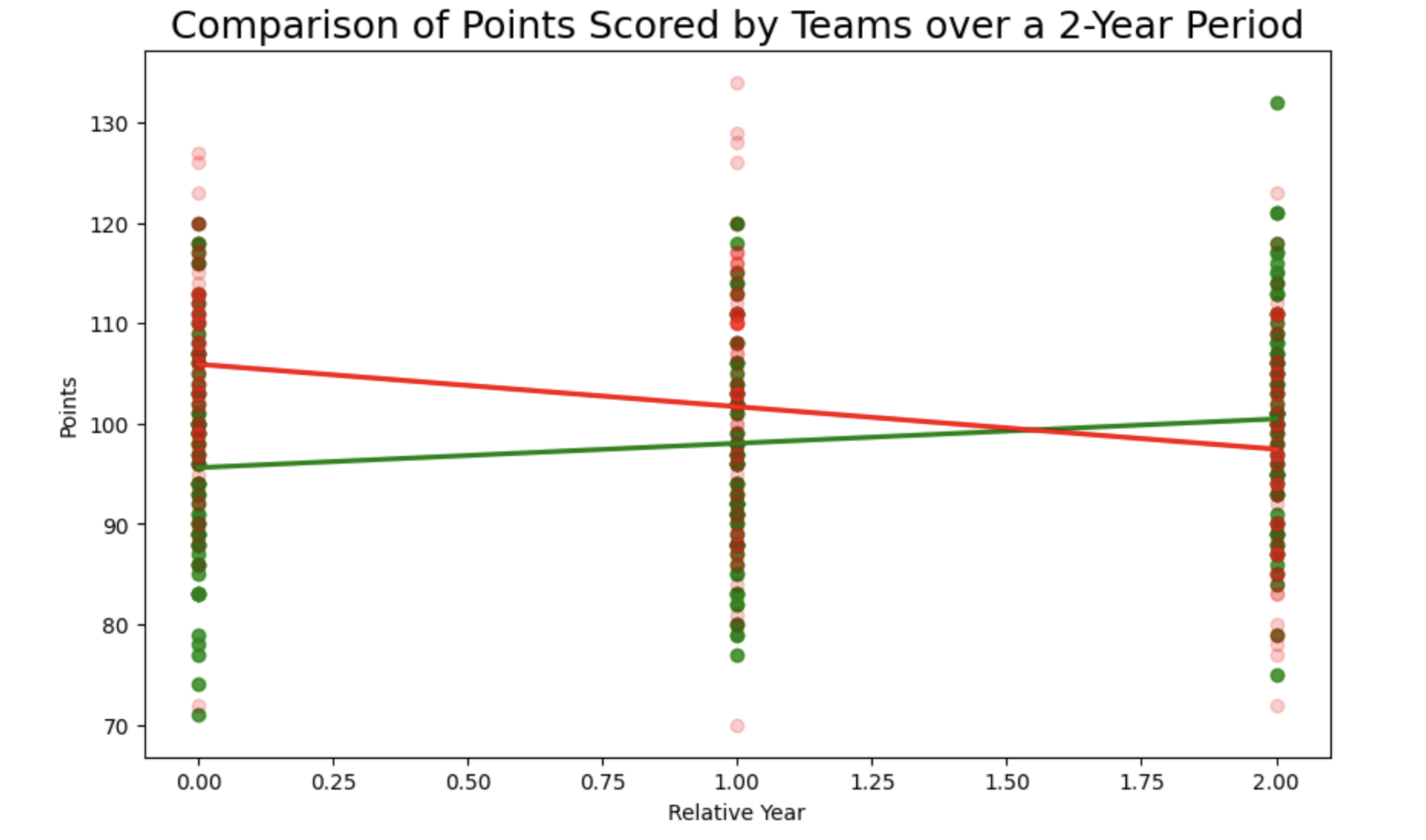
|  | **Name of Team** | **Assigned Years** |
| --- | --- | --- |
| 1. Yours | The Celtics | 2013-2015 |
| 2. Assigned | Chicaco Bulls | 1996-1998 |

## Data Visualization: Points Scored by Your Team

## Data Visualization allows for data analysts to see the data at a wider spectrum. Take scatterplots for example. Being able to see all the plots visually, If you make your x and y value for your data scaled back enough, You can Image a line that would cut through the middle of the data. This helps by allowing you to see the avertage as time progresses for any increases or decreeses in the metric stats depending on what the metric is. In turn, if there are decreases, you can check the timeframe to see what was going on or what is happening that would cause this decline, and adjust accordingly. Take Crypto Currencty for example. Looking at the average price over time for Doge coin, it was constistantly going up. But, it has a few quite drastic drops in its overall price. When reviewing the time of the drops, you can coorelate these times with times that Elon Muske has either said or done something against it causing its massive drop. For instance, April 3rd, for a short while, Elon made the logo for “X” the Doge Coin mascot, which caused the price to surge. Then Elon removed it and replaced it with its now X symbol, causing the price to drop yet again. With Data Visuallization graphs, we can easily determine trends as apposted to staring at a database of numbers. Ultimately, Data Visualizations increase the speed in which trends and outliers can be determined. For this we were to choose one of the graphs for a Data Visualization to help us in determining a trend. Reviewing both of the graphs, I chose the scatterplot. I feel this is better suited for a situation like this in respects to trying to determine areas of improvement as a scatterplot with line average through it, depending on the metrics will tell us if the metric is increasing(positive) or decreasing(potentially negative). THe Histogram is simply showing me that the team points scoring follows a bellcurve which when determining metrics to become the best, isnt always the best data for the job. This doesnt tell us which metric is lower, just that we frequently have around 100 points. Now, I chose this because it shows that over the course of 2 years, our point average(for this metrics) is increasing. Meaning that we are moving in the right direction.

## 

We can than use this type of graph with all the various metrics, in order to see what areas our metrics are decreasing. We can then use the Points graph as sort of a baseline for adjustments for the other metrics.   
Now, for this project, i was under the impression we were compareing the two teams, so I added in the Bulls data and plotted then together so we can see a visual representation of how over the course of 2 years, we compared to how they were doing in the same timeframe(2 years). From this, yes the Bulls may have had Michael Jordan, or Scotty Pippen carrying the team, but overall as a team, it does not appear they were doing so great from 1996-1998.

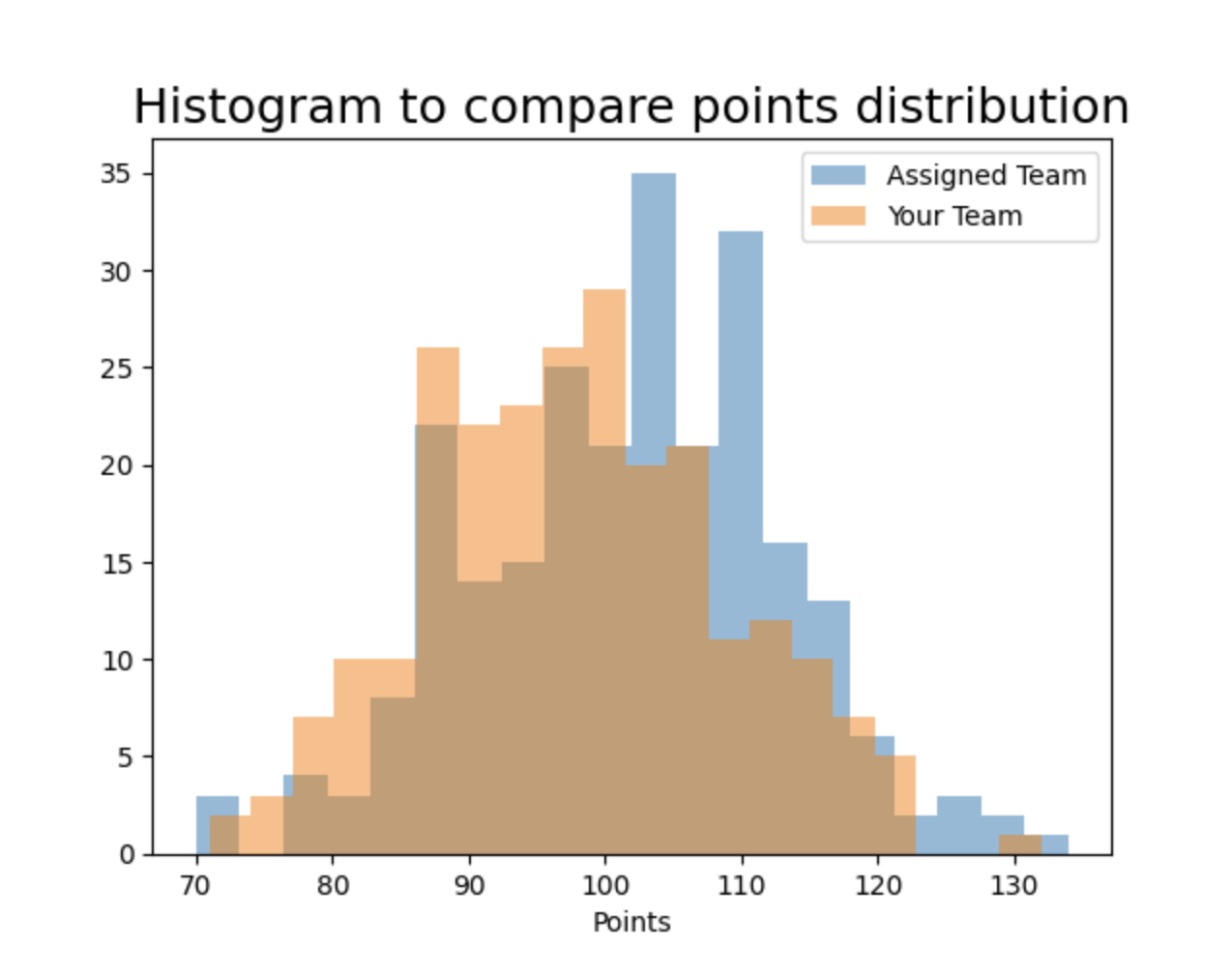


## Data Visualization: Points Scored by the Assigned Team

## For the assigned team, we’re not worried about how to make them better. Thus reviewing the scotter plot over time timeframe is not needed. This is a team that had proven its consistency. Which we can see with the histogram graph. This graph is pretty straightforward when visualizing the distribution of the points over the 2 years. In this graph, its easier for us to see which range most points fall. Because of this, we can determine that if most of the points are grouped around in specific ranges, this would signify the constistency the team has. Where as, if they are more spread, this would give us insight into performance issues they may have had. Over all, the point distribution can help determine the teams consistency. From this we can then look into what has participated in making them consistent. What worked for them, what did not work for them, what can we benefit from, where do we need to grow.

## Data Visualization: Comparing the Two Teams

When comparing two distinct data distributions, data visualization can be an important tool. Direct visual comparison is made possible, enabling the discovery of patterns, trends, and insights that might not be immediately obvious in the raw data. The presence of any outliers as well as central tendency, spread, and differences in distributions may all be seen through visual representations.



* For this, i chose the overlaping histogram charts. I chose this because i felt it was to best to view the frequency of scores within the certain ranges. I feel that this offeres a more detailed view of how often the two teams score within various point brackets. From this, we can then see what the differences are when scoring in this area. What worked that allowed them to score higher than we did. What worked for us that didnt work for them? How do the two graphs compare to each other? Overall, when reviewing the two, i can see that the bulls bell curve is skewed slightly further right than the Celtics bell curve meaning they have slightly more high scoring games than the Celtics do. Adiitionally, i see higher points in the Bulls which means they have more games with these higher points. The majority of our games i would say are mid range, or just slightly under 100 points.

## Descriptive Statistics: Points Scored By Your Team in Home Games

| **-** | **Value** |
| --- | --- |
| Mean  Median  Variation  Standard Deviation | 98.82  99.5  107.67 10.38 |

* The mean of the DataSet gives the average of the data being calculated. This average can be one of the numbers or between two of the numbers. This average can be skewed drastically if you have multiple ourliers. The median of the dataset gives the middle value. If you have outliers, it may be beneficial to review this number as apposed to the median as this will theoretically be where the majority of the data points cluster up thus telling you if the majority of the points per game are higher, or lower. The variance measures the spread of the points from the mean. A high variance indicates that the points are spread out over a large range, while lower variance indicates that the points are closer to the mean. The Standard Deviation, is a measure of the amount of spread of points. A lower standard deviation would signify that the points fall close to the mean, whereas a higher standard deviation signify that the points fall farther from the mean, or spread out more. Because we have a lower Standard Deviation, this would inidcate that most of the points fall closer to the mean of 98.82 which correlates with the median as the median would typically fall with the most clustered of the points which would be around the 99 mark. To determine if this is a skewed left or skewed right graph we can run the skew() function in scipy.stats. For this, this gave us a skew of –0.031 which would typically indicate that this is a skewed left graph but, this is so close to 0 that i believe this to be a bell shaped graph. This is backed by the fact that not everyone or every team is equal in sports and there will enevitably be less games with higher scores and less games with lower scores, thus creating a bell shape.

## Descriptive Statistics: Points Scored By Your Team in Away Games

| **Statistic Name** | **Value** |
| --- | --- |
| Mean Median  Variance Standard Deviation | 97.28  96.0  121.43  11.02 |

* The mean is roughly about the same as the Home games, which is about 97 points per game. Again, the median is trailing closely with around 96 points telling us that the the majority of the games are around this 96-97 area. The Variation is 121 which is fairly high. This tells us that the team has had very high scoring games as well have relatively lower scoring games when in Away matches, which tells us that the points per game is inconsistent and can be improved. The Standard Deviation in the instance is indicating that the majority of the points fall within the mean, but ultimately because of the variance, this will probably be ignored and the coaches should focus on lowering the variance.

Ultimately what we can determine from this data is that our team performs better when in Home games. Our games seem to be more consistant. Our focus at the time should be to lower the varience in the Away games but todo that we would need to review all the day from the Away games.

## Confidence Intervals for the Average Relative Skill of All Teams in Your Team’s Years

| **Confidence Level (%)** | **Confidence Interval** |
| --- | --- |
| 95% | (1502.02 , 1507.18) |

* Confidence intervals are an important technique in statistics for estimating the range of an actual population measurements, such as the mean. They are produced using sample data and contain a margin of error which allows for variation in the sample and is determined by the confidence level chosen (for example, 95% or 99%). This interval displays population parameters more accurately by recognizing the natural variability in estimates formed by data samples. They aid in evaluating the significance of results and directing educated choices in a wide range of areas by allowing users to evaluate the accuracy of their projections and compare different groups or solutions with a higher degree of probability.
* For the average relative competence of teams from 2013 to 2015, the confidence interval of (1502.02,1507.18) at a 95% confidence level suggests that if we took several random samples from the same population within that time frame and calculated the confidence interval for each sample, we would find that around 95% of those intervals contained the average relative skill of all teams in that time frame. In other words, we are 95% certain that the average ability level of teams from 2013 to 2015 is between 1502 and 1507.18. Rather than a single point estimate, it presents a range of values that include each team’s skill level. This accounts for and recognizes the fluctuation that takes place when estimating a population parameter from a sample. It means that, while the average for our sample falls within this range, there remains some uncertainty, and actual average for the entire population of teams could fall somewhere within this range.
* The overall width of the confidence interval changes as the confidence level changes. A higher confidence level, like 99%, results in a wider interval which reflects more variability ensuring greater confidence, but with less precision. A lower confidence level, such as 90%, yields a shorter interval, allowing for more precision but at a cost of being less certain that it actually contains the base population value. There is an imbalance between confidence and precision.

## Confidence Intervals for the Average Relative Skill of All Teams in the Assigned Team’s Years

In the Python script, you calculated a 95% confidence interval for the average relative skill of all teams in the league during the years of the assigned team. Additionally, you calculated the probability that a given team in the league has a relative skill level less than that of the assigned team.

Confidence Interval for Average Relative Skill of Teams in Assigned Team’s Years

| **Confidence Level (%)** | **Confidence Interval** |
| --- | --- |
| 95% (for example, 95%) | (1487.66 , 1493.65)  \*Round off to 2 decimal places. |

* The computed 95% confidence interval for average relative skill (ELO) of teams in 1996-1998 is (x,y), where x is the "lower bound" and y is the "upper bound".  
    
  The 95% confidence interval for team average relative skill (ELO) in 1996-1998 is (x,y). This means that we are 95% positive that the genuine mean relative skill of all teams during these years falls within this range. If we sampled teams' relative talents numerous times, the average relative skill would fall between x and y 9.5 out of 10 times. Based on the data we have, this interval indicates a range in which the genuine average relative skill of teams for those years is likely to be observed. It assists in determining the precision and dependability of our projection, and the smaller the interval, the more precise our estimate.
* Like the previous interval, when the confidence level of an interval is changed, the width of that interval changes accordingly. A higher level of confidence like 99%, results in a wider interval, suggesting greater "confidence" that the population parameter is within this range. But, this comes at the cost of precision. A lower confidence level like 90%, yields a smaller interval, yielding a more precise estimate but with less certainty. Basically, there is an imbalance between the confidence level chosen and the precision of the interval.

When comparing two interval ranges, the range and overlapping must be examined. If the intervals overlap, it suggests the average relative skills of the teams between the two timeframes are not significantly different. If the intervals do not overlap or show minimal overlay, it suggests a significant variance in average relative skill between the two time periods. If the interval from the assigned team's time frame is larger, it suggests the average relative skill of teams during that time frame was likely higher compared to the rest of the time frame provided, and vice versa. If one interval is wider than the other, it suggests the relative skill for that time frame is not clear.

## Conclusion

**We wanted to understand the average relative skill level of basketball teams across two time frames using statistical analysis and the 'elo\_n' metric. We presented a range in which we are 95% convinced the genuine mean relative skill level of teams falls by calculating confidence intervals across both timeframes. This research' practical value stems from its potential to present data on how team skills have changed or remained consistent through time. By definition, confidence intervals offer a range of probable values rather than a single point estimate, allowing for an improved understanding of the data.**  
 If the interval from the assigned team's time frame is larger, it suggests the average relative skill of teams during that time frame was likely higher compared to the rest of the time frame provided, and vice versa. If one interval is wider than the other, it suggests the relative skill for that time frame is not clear.

In the context of our scenario, examining the confidence intervals between the two timeframes will give insight into whether team performance standards are increasing or decreasing. If one interval is substantially greater than the other, it suggests a variation in average skill level between the time periods. Overlapping intervals, on the other hand, can indicate consistency in performance over time. This data is essential to clients such as team managers or coaches. It may guide decisions regarding training requirements, strategy for upcoming matches against formidable teams, and/or even recruiting efforts. In the end, these results offer a historical framework for the development of skills in the league, outlining the basis for future efforts.

## Citations

zyBooks. (2023). Module 1: escriptive Statistics and Probability. MAT-243. <https://learn.zybooks.com/zybook/MAT-243-J1254-OL-TRAD-UG.23EW1/chapter/1/>zyBooks. (2023). Module 2: Probability Distributions. MAT-243. [https://learn.zybooks.com/zybook/MAT-243-J1254-OL-TRAD-UG.23EW1/chapter/2/](https://learn.zybooks.com/zybook/MAT-243-J1254-OL-TRAD-UG.23EW1/chapter/1/)zyBooks. (2023). Module 3: Confidence Intervals and Hypothesis Testing. MAT-243. [https://learn.zybooks.com/zybook/MAT-243-J1254-OL-TRAD-UG.23EW1/chapter/3/](https://learn.zybooks.com/zybook/MAT-243-J1254-OL-TRAD-UG.23EW1/chapter/1/)