# Illinois Tech Basketball Data Analytics

Chenjie Li,Diruba Palabiyik
Data Science
Illinois Institute of Technology

***Abstract*—In this work, we showcase our approaches and results of our data science practicum project–Illinois Tech Scarlet Hawks Men's Basketball Analytics. Starting from database desgin, we explored some of the most recognized metrics to evaluate players and lineups, such as "on-off court" data, "lineup" comparisons, "offensive and defensive tendencies". We also performed "Player Clustering" to find similar players. Based on our "scouting reports", our coaching staff made the corresponding strategies and in the end set the best record of our school history.**

## I. Introduction

"Analytics are part and parcel of virtually everything we do now." said by NBA commissioner Adam Silver.
Data Science, one of the most promising and in-demand career paths for skilled professionals, is expanding its branch to sports. The origin of sports analytics was not basketball. The education begins with the baseball analysis of Bill James [1]. With baseball leading the way, other sports began to embrace data and utilize them to help their teams. Unlike baseball which can be reasonably partitoned into a series of discrete events, basketball is much more complicated and harder to give reasonable analysis. With that being said, there are still a lot of ways to approach this exciting sport and use technology to change the game. In Stephen M. Shea's work [2], he introduces variety of ways to evaluate players and teams. Also, in Rajiv's exciting TED talk [3], he introduced the implementations of computer vision and machine learning in basketball analytics.

As a Division III team, however, the Men's basketball team of Illinois Tech hasn't benefited from this trend due to the lack of available data and technicians.

In this program practicum, we built up a database from scratch. Based on that, we explored several branches: Player evaluation, Team evaluation, Lineup Evaluation. Using Tableau software, we got straightforward yet insightful visualizations for coaching staff as references. In the end of the season, we made to the $2nd$ place in the tournament, setting the best record of our school basketball history.

## II. Data Source

You can't make bricks without straw. Getting data was one of the biggest challenges for us. Thankfully, we were lucky enough to have somethig to work on.

### A. Synergy

Synergy[4] for overall statistics from each player and each team. They provide comprehensive details for players and teams, even the distributions of offensive and defensive tendencies and the corresponding efficiencies. Based on these information, later on we created "strength and weakness", "player clustering" and so on.

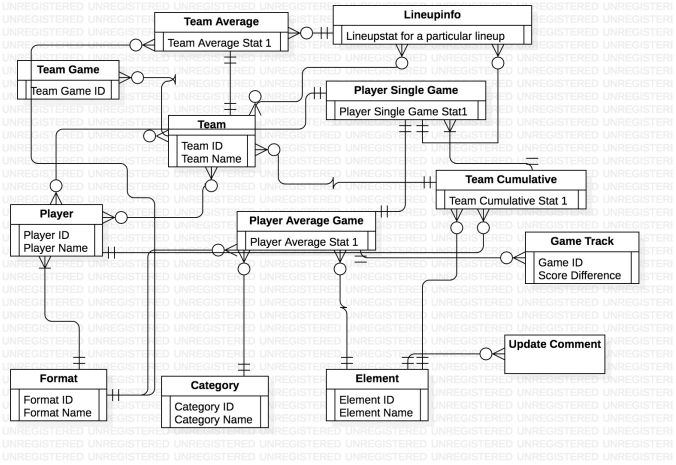


Fig. 1: ERD Diagram(simplified)

### B. NACC Official Website

Another important data source in this project was northern athletics collegiate conference website [5], in which we can get the game "Play-By-Play" files which can be used for analysis of lineups and "Player Pairs" analysis.

## III. Database Design

### A. Design

After exploring the Synergy and discussion within the team, we decided to get data from Game, Team, Player tabs.

**Game** includes the details of every games stats of both teams. **Team** consists of the season overall average performance of the team. **Player** covers the individual players overall average performance of the team. The database was designed by taking the way data was represented in Synergy into consideration. We decide to design the database schema based on the way the data presented on Synergy, but since Synergy does not show how those data are connected with each other, a newer design for joining data to run different queries was needed. The detailed schema of the database design is: (Please check GitHub if you need details):

Once we obtained the database, any query can be created. For example: *"Getting the players whose average points is over 15 points?"*,*"the games that our school won/lost by within 10 points?"* In our database, the following entities were included:

- Player (Name, which team he plays for)
- Team
- Player average stats (Average Points, rebounds etc)
- Player season average stats (Season average points, rebounds, player play type distributions etc)
- Team season average stats (Average possessions, opponent average possessions, play type distributions)
- Team/Player single game stats
- Different Lineup combinations and their corresponding stats

### *B. Data Processing Workflow*

After getting our ERD diagram, we started tackling the problem of processing "unstructured" data into "clean" and "importable" data to be able to save them into our database. The workflow chart is shown in Fig 2

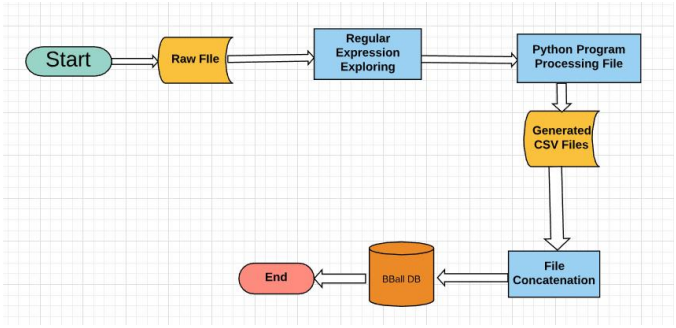


Fig. 2: Data Processing Workflow

Fig 3 shows how the data (Player) looked at the beginning.

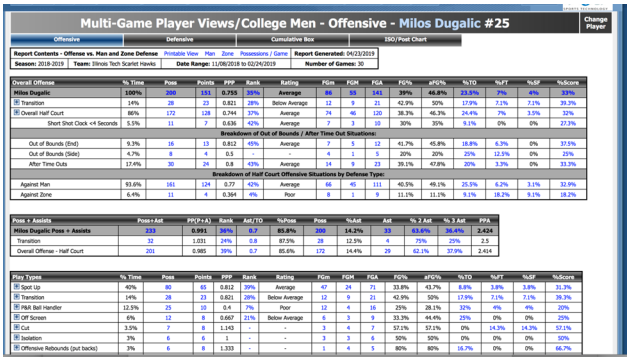


Fig. 3: Synergy Website

The steps are:

- Raw File: Copy and Paste the webpage into a .txt file.
- Exploring Regex: Using regex(regular expression), we explored the "pattern" in the text file like spaces, tabs etc.
- Write down Programs: After summarizing the "patterns" in the file, we wrote down the Python programs to process the .txt file.
- CSV contatenation: after processing each .txt file, we will have a collection of formatted .csv files(Fig 4). then we will concatenate them together as a "big" file.
- Import to Database: After getting the CSVs we want, we can easily import those files to our database.

| Player_ID | Uploaded_Date | Team_ID | GP | Min | SST | SSTexPts | Pts | Ast | TO | Ast_Turnover_Ratio | Stl | StlPos | Blk | TtlReb | OffReb | DefReb | FGA |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 12/23/2018 | 4 | 11 | 29.9 | 0.96 | 0.33 | 11.5 | 0.9 | 2 | 0.45 | 1.4 | 0.9 | 1.4 | 5.3 | 1.4 | 3.9 | 8.9 |
| 2 | 12/23/2018 | 4 | 11 | 15.6 | 0.76 | 0.19 | 6.6 | 1.1 | 0.8 | 1.33 | 0.5 | 0.3 | 0 | 1.2 | 0.1 | 1.1 | 5.2 |
| 3 | 12/23/2018 | 4 | 4 | 2.5 | 0.31 | 0.1 | 0.2 | 0.2 | 0.5 | 0.5 | 0.2 | 0.2 | 0 | 0.2 | 0 | 0.2 | 0.5 |
| 4 | 12/23/2018 | 4 | 11 | 30.2 | 0.67 | 0.18 | 10.9 | 3.3 | 3.6 | 0.9 | 1.4 | 2.1 | 0.1 | 3 | 0.4 | 2.6 | 10.5 |
| 5 | 12/23/2018 | 4 | 10 | 17.1 | 1.03 | 0.38 | 5.7 | 1.4 | 0.8 | 1.75 | 0.5 | 0.7 | 0.1 | 2.9 | 1.2 | 1.7 | 5.7 |
| 6 | 12/23/2018 | 4 | 2 | 2.5 | -0.36 | -0.2 | 0 | 0 | 0.5 | 0 | 0 | 0 | 0 | 0.5 | 0 | 0.5 | 0.5 |
| 7 | 12/23/2018 | 4 | 3 | 7.7 | 1.17 | 0.22 | 3 | 0.7 | 0.7 | 1 | 0 | 0 | 0 | 1 | 0.7 | 0.3 | 3.3 |
| 8 | 12/23/2018 | 4 | 7 | 7.1 | 0.18 | 0.1 | 1 | 0.7 | 0.4 | 1.67 | 0 | 0.4 | 0 | 0.1 | 0 | 0.1 | 1.6 |
| 9 | 12/23/2018 | 4 | 4 | 2.2 | -0.72 | -0.44 | 0 | 0 | 0.5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.8 |
| 10 | 12/23/2018 | 4 | 3 | 2.7 | 1.22 | 0.25 | 1.7 | 0 | 0 | - | 0 | 0 | 0 | 0.3 | 0.3 | 0 | 0.7 |
| 11 | 12/23/2018 | 4 | 6 | 10.5 | 0.51 | -0.05 | 4 | 0.3 | 1.5 | 0.22 | 0 | 0.3 | 0 | 1.5 | 0.3 | 1.2 | 3.5 |
| 12 | 12/23/2018 | 4 | 11 | 18.5 | 1.12 | 0.21 | 9.7 | 0.7 | 1.5 | 0.47 | 0.4 | 0.4 | 0.2 | 3.5 | 0.9 | 2.6 | 6.8 |
| 13 | 12/23/2018 | 4 | 1 | 3 | 0 | 0 | 0 | 0 | 0 | - | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 14 | 12/23/2018 | 4 | 4 | 3.8 | 0.58 | 0.33 | 1 | 0.2 | 0.2 | 1 | 0.2 | 0 | 0 | 0.5 | 0.2 | 0.2 | 1.2 |
| 15 | 12/23/2018 | 4 | 11 | 23.2 | 0.78 | 0.3 | 6.1 | 1.3 | 1.1 | 1.17 | 0.2 | 0.8 | 0 | 5 | 1.3 | 3.7 | 5.1 |
| 16 | 12/23/2018 | 4 | 4 | 2.8 | 0.65 | 0.45 | 0.5 | 0.2 | 0.2 | 1 | 0 | 0 | 0 | 1 | 0.2 | 0.8 | 0.5 |
| 17 | 12/23/2018 | 4 | 11 | 28.9 | 1.21 | 0.28 | 15.4 | 1.6 | 2.1 | 0.78 | 0.9 | 1.3 | 0.2 | 5.5 | 1.5 | 4 | 12.2 |
| 18 | 12/23/2018 | 4 | 10 | 29.7 | 1.06 | 0.46 | 10.6 | 2.3 | 1.5 | 1.53 | 0.2 | 0.8 | 0.5 | 8.1 | 2.5 | 5.6 | 8.9 |
| 19 | 12/23/2018 | 4 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | - | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Fig. 4: CSV File

## IV. Strength VS Weakness

There are many ways to evaluate players. One important aspect is "How efficient is this player?". Here, we chose to use "Points Per Possession"(ppp) as the measure to differentiate each player's strengths and weaknesses. On Synergy website, it breaks down each player(team)'s offense into different types: *Spot Up, Transition, Off Screen, Offensive Rebounds, Hand Off, Isolation, Post Up, Miscellaneous*, player(team)'s defense into different types: *Spot Up, Off Screen, P&R Ball Handler, Hand Off, Isolation, Post-Up, P&R Roll Man*. for each type, Synergy collects and computes its corresponding *Percentage of time*, which is basically the frequency of this player's all offensive(defensive) types. Also, it lists the points per possesion, which is showing how many point(s) this player can score in each offensive (or how many point(s) this player's opponent can score in each offensive possesion).

*Definition 1 (Strengh and Weakness):* Given a player's offensive average $ppp_{off\_overall}$, the strength for the player is $ppp_{off\_strength} > ppp_{off\_overall}$, the weakness for the player is $ppp_{off\_weakness} < ppp_{off\_overall}$; Given a player's defensive average $ppp_{def\_overall}$, the strength for the player is $ppp_{def\_strength} < ppp_{def\_overall}$, the weakness for the player is $ppp_{def\_weakness} < ppp_{def\_overall}$.

The reason for choosing " $ppp_{def\_strength} < ppp_{def\_overall}$" is that it shows how good this player is in making his opponent inefficient.

For this part, we store those information in *"player(team) average"* table

We chose 2 different ways of showing the visualizations in Tableau. The first one, Fig 5 is showing the distributions of each player's play types. It has different filters on this dataset. For example, users can filter based on team,offensive/defensive, player etc. Fig 6 is showing in a more straightforward way. The red reference line is player's $ppp_{overall}$.

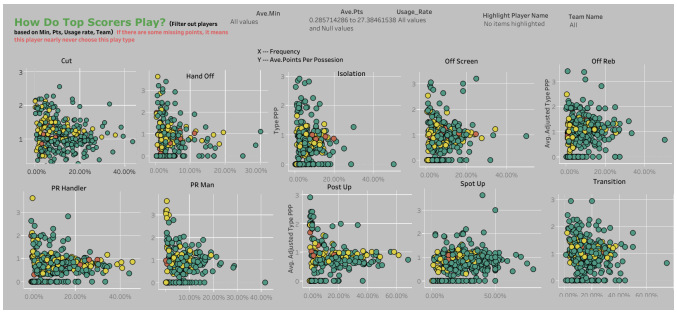


Fig. 5: Strength and Weakness 1

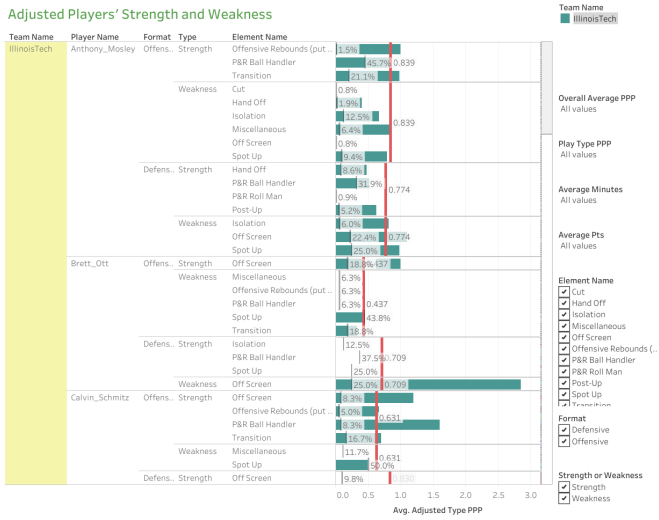


Fig. 6: Strength and Weakness 2

## V. Lineup Anaysis

In this section, we essentially did 2 things, "individual player's contribution" to team and "pairs' performance". The reason why we consider both of these 2 aspects are: A player might be normal or average when evaluated individually, but his "pair" with some other teammates may give us an surprising result(pairwise); At the same time, we do need individual player's performance as a reference to be able to give ourselves a sense of what kind of a player is. For instance, consider "Plus Minus" as the measuring metric, player A's personal plus minus is +100 in the season, whereas his average plus minus with other teammates together(pairs) is only +10. this could give us 2 insights: player A is very good; he might need to work on involving his teammates. We do admit that this is not necessarily true. For example, A is just in another level compared with his teammates, we can't jump to the conclusion like this solely based on one metric.

With the idea above, there's actually a lot more we should do: we didn't have the stats in that detail. For example, to get the "pairwise performance", we need to get the stats only when those pairs were on the court. But we didn't have those at hand.

Thankfully, we finally found something to get started. Nowadays, most basketball games' broadcasting service will feature a section called "play-by-play", which is a document-like file that will keep what happened on the court in it. So we decided to write a program that can process it and generate the stats we want. A sample of "play-by-play" file: ($www.illinoistechathletics.com/sports/mbkb/2018\text{-}19/boxscores/20190223\_a3e3.xml?view=plays$)

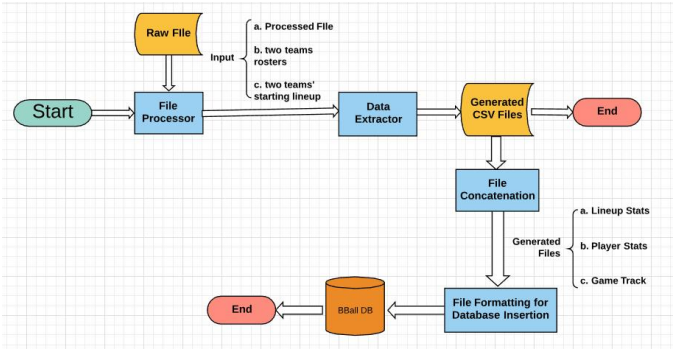


Fig. 7: Program Pipeline

Fig 7 shows the flow of our designed program. **Raw File** is just "Play-By-Play" file which we can easily get just using traditional "Copy and Paste" action(Ctrl+A in the website and Ctrl+C, Ctril+V to a .txt file). **File Processor** is a program that can handle some special structural problems in raw file, like cleaning up a miscellaneous collection of redundant data and strange time sequence (some incorrect sequence of the game timeline).**Data Extractor** is the main part of the program. when using this program, user needs to specify the starting lineups of both teams, and we made the assumption that for both first and second half, both teams were using the same lineups. In reality, of course, this is a bad assumption, but user can manually change the input in the program if there's any changes in the starting lineups in the second half. After running the program, it will generate 3 files: *Lineup Stats, Player Stats, Game Track*. *Lineup Stats* contains the stats for the whole lineup (5 players combined). It also has the corresponding stats for the opponent stats.*Lineup Stats* were saved in $Lineup$ Table of our database.*Player Stats* contains each individual player's stats in each different lineup, i.e. $PlayerA$'s stats in Lineup $a$ is a row of data, whereas $PlayerA$'s stats in Lineup $b$ is another row of data.*Player Stats* were saved in $Player\_Game$ Table of our database. *Game Track* is a file that records the change of the score difference between two teams as time changes.*Game Track* data were saved in $Game\_Track$ Table of our database.In **File Concatenation**, in the case of generating files for multiple games at once, we might want to concatenate them together before formatting them in the next step. In **File Formatting For Database Insertion**, we map the field names into the *keys* in the database tables we created.

After all these steps above, we finally got the stats we need to perform the analysis we want in the first place.

For the rest of the section, we will discuss the algorithm we used for generating **Top-K** pairs for each player in terms of the given metric.

*Definition 2 (Top-K Pairs):* Given the direction of the metric $\theta \in \{high, low\}$, the measuring metric $\sigma$, player $P$, $P^k_{\sigma,\theta}$ is a set of players that are best $K$ pairs with this player $P$.

**Algorithm 1** Top-K pairs

$R \leftarrow$ *the set of all the players in the team (Rosters)*
$p_0 \leftarrow$ *target player*
**procedure** TOP-K GENERATION($R$, $\theta$, $\sigma$,$p_0$)
  **for** each p in R **do**
    **if** $p == p_0$ **then**
      *Continue*
    **else**
      $Candidate\_Pair \leftarrow SORT\{p, p_0\}$
      **if** $\theta == low$ **then**
        $heap \leftarrow MinHeap$
      **else**
        $heap \leftarrow MaxHeap$
      $Q$ =
      *SELECT p as target_player, Candidate_Pair as pair,σ as metric FROM Lineup*
      *WHERE lineup_players like %Candidate_Pair%*
      $result \leftarrow R(Q)$
      **if** *heap.size* $<$ *K* **then**
        *heap.insert(result)*
      **else if** $heap.peek(\leq\ if\ minheap\ or\ \geq\ if\ maxheap)\ result$ **then**
        *heap.insert(result)*

In the Algorithm 1, we ignored the number of players in each pair, i.e. we treat "2-men" pair and "3-men" pairs the same way. Combine those work together, we generated the visualization for pairs and individual "On and Off court" using "Plus Minus" as our evaluating metric as shown in Fig 8

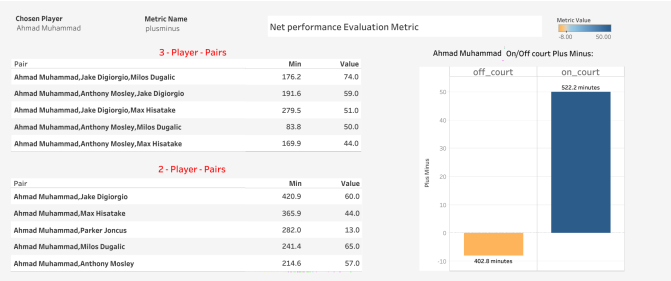


Fig. 8: Pairwise and Individual Plus Minus

By implementing the data from *Game_Track* Table, we generated the "Game Flow" chart as shown in Fig Different colors means the different lineups and the barchart above are the score differences between 2 teams.

Besides, after our team made to the tournament, we came up with a "scouting report" by combining the strength and weakness in section IV with the play-by-play data to give our coach a more efficient and more convenient way to study our opponents. One example of the "Scouting Report" is shown in Fig 10

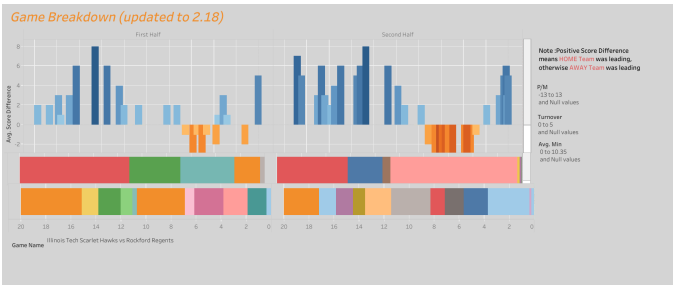


Fig. 9: Game Flow Visualization

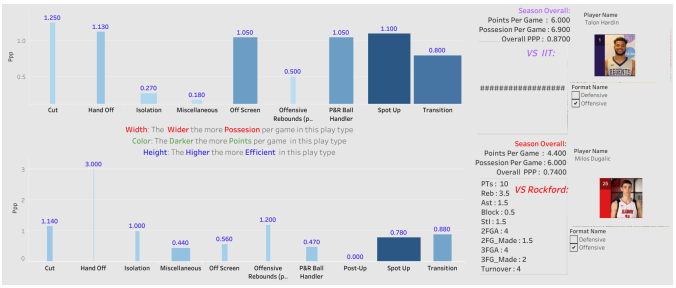


Fig. 10: Sample Scouting Report

## VI. CLUSTERING

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters) The purpose of the clustering is to categorize the players to help coach to make comparison between players that uses different play types, which players are similar etc. The attributes used in Clustering:

- Effective field goal percentage
- Spot up time percentage
- PR Ball Handler Time time percentage ( each player used) Field Goal Attempted
- Two Field Goal Attempted
- Three Field Goal Attempted
- Free Throw Attempted
- Cut Off Percentage
- PR Roll Man Percentage
- Hand Off Percentage
- Isolation Percentage
- Off screen percentage
- Post up percentage
- Pr ball handler percentage
- Spot up percentage
- Transition percentage
- Assist
- Total Rebounds

To find the Time Percentages(frequecny) for different play types: Queries are used to create for each different play type;

Views are created and then used to “left-join” with each other. To get the *offensive cut off* percentage, the query we wrote for that is:

**SELECT** p.player_name, pa.percentage_time as cutoff_percentage, f.format_name c.category_name, e.element_name
**FROM** player p, player_Average pa,category c, format f, element e
**WHERE** p.player_id = pa.player_id
**AND** c.category_id = pa.category_id
**AND** e.elemen_id = pa.element_id
**AND** pa.uploaded_date = ’2019-02-12’
**AND** f.format_name = ’Offensive’
**AND** c.category_name = ’Play Types’ **AND** f.format_id = pa.formatid
**AND** e.element_name = ’Cut’;

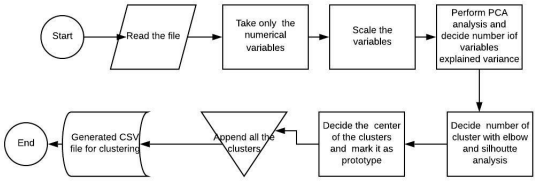


Fig. 11: Clustering Workflow

We followed the steps shown in Fig 11 to perform the clustering process:

- Only numeric variables are extracted from the program
- PCA analysis performed to explain how many of the Variables explain which percentage of the overall variance.
- Deciding on the number of components
- Deciding on the right number of clusters
- Elbow and silhouette score techniques are used Since at number 10 , an elbow shape formed and the at 10 silhouette score the line stabilized.
- According to results of the two tests we, selected 10 as the optimal number of clusters based on the 16 attributes mentioned above.
- The center of the clusters are taken as any of the IIT players player in the selected cluster and named as the prototype
- Based on the different feature averages in each cluster the clusters are named. For example 3 pointers includes the player with a higher 3 FGA rating compared to other clusters, assist ballhandler cluster had the players with the better rating of assist, ball handler features compared to other features.

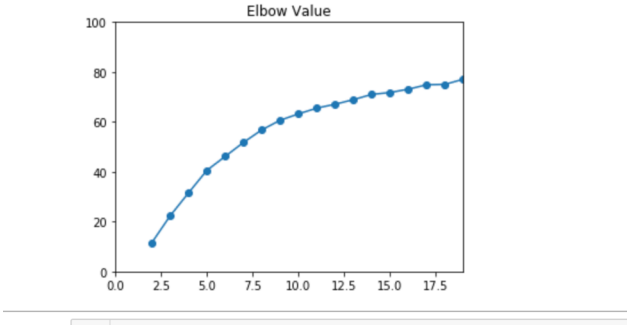


Fig. 12: Elbow Analysis

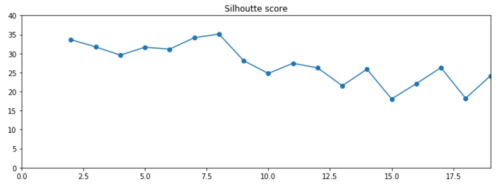


Fig. 13: Silhouette Score Analysis

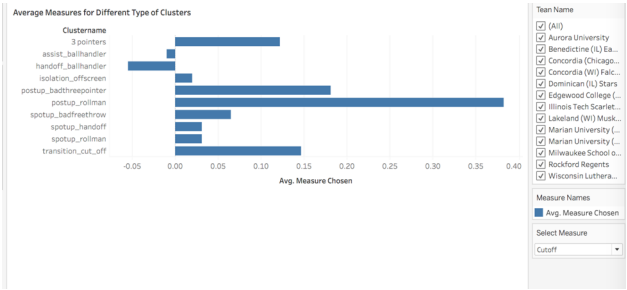


Fig. 14: Brief Clustering Summary for Teams


Fig. 15: Brief Clustering Summary for Players

As can be seen from the figures above with the help of the clustering, strengths and weaknesses of players/teams in terms of attributes mentioned above can be found on both team and player level.

## VII. Conclusion and Future work

In this project, we started from the foundation of the previous work, built up a lot more interesting and helpful visualizations and analysis that helped coaching staff gain the insights and story behind the data.

What's still lacking in our project, as we initially planned but couldn't finish, is "Automating the process of copy pasting data from Website", in other words, **Scraping Data** from website. Once this step achieved, we will actually build a complete pipeline for our team for getting the data and do any analysis on it!

We thank **Prof. Shlomo Argamon** for his brilliant advice and ideas to help us achieve those goals. We thank **coach Todd Kelly** for his feedbacks to help us modify our results to really help our basketball team.

## VIII. Appendix

This work's source codes are available in Git Repository:
*https://github.com/JayLi2018/ScarletHawksAnalysis*

The visualizations are available in authors' Tableau Public Profile:
*https://public.tableau.com/profile/chenjie.li#!/*
*https://public.tableau.com/profile/dilruba6464#!/*

## References

[1] B. James. *The New Bill James Historical Baseball Abstract*. Free Press, 2010.
[2] Stephen M. Shea and Christopher E. Baker. *Basketball Analytics: Objective and Efficient Strategies for Understanding How Teams Win*. CreateSpace Independent Publishing Platform, 2013.
[3] Rajiv Maheswaran. The math behind basketball's wildest moves.
[4] Synergy sports technology.
[5] Nacc men's basketball.