Kobe Bryant’s NBA career analysis

VIPUL MUNOT | SIDDHARTH JAYASANKAR | ANIRUDH K MURALIDHAR

Indiana University | Data Visualization

Contents

Abstract

The aim of this project is to analyze the career of basketball player Kobe Bryant by producing high quality visualizations. This project explores the use of innovative visualizations for the sport facilitating effective analysis of basketball players.

Introduction

Nowadays big data is being used in almost all the domains. Sports informatics and analytics is also venturing into big data to perform in-depth analysis to come up with better strategies and team composition. Through this project, we would like to explore how useful effective visualization of data regarding player's performance could help develop better strategies and help the player become better at the sport.

Motivation

Basketball is the second favorite sport in USA. It is third with regards to the money being spent. Basketball is the most played sport in America. Team USA (men) have won 15 gold medals in 19 Olympics. The NBA has 30 teams and there are 82 games per season. With so many games being played, the amount of data at hand is huge combined with the popularity and interest which this sport has, motivated us to take up this project. Added to these the team management and owners want data driven strategies and solution to enhance their performance.

This project would help,

1. Analyze the player's performance
2. Develop a player by improving his weak zones.
3. Come up with better game strategies for different opponents and different players in the opponent team.
4. In drafting new players into the team.

Baseball is a sport where advanced analytics is already being used to great extent (The movie Money Ball is a great example of how advanced analytics can be used to develop better strategies and team composition which would lead to the success of the team.). Through this project, we would like to contribute to such analytics where a whole team can be built in the game of basketball just by looking at the statistics.

Relevant Work

Baseball is the game which is primarily data driven and other sports such as NFL, soccer and basketball are catching up. Sabermetrics is the application of statistical analysis of baseball that measure in-game activity. The NBA recently hosted its first hackathon on Sep 24th, 2016. This is an indication that NBA is turning towards the data side of that game.

Interesting story of Analytics in Basketball

Daryl Morey is the General Manager of the NBA team Houston Rockets. He is predominantly a statistician. Through detailed analysis of how the game is played, Morey figured out that the average points scored from the 3-point range and close range shorts where more than the number of points scored from midrange shots.

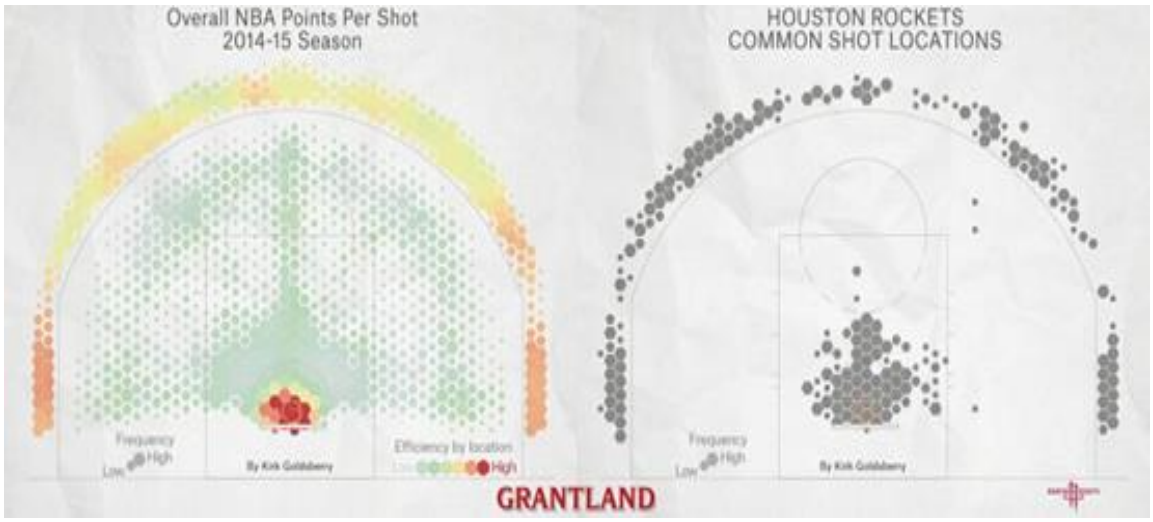

Figure 1 (a) Visualization showing efficiency of different range shots. (b) Shot map of Houston Rockets

Based on his analysis, he trained his team to shoot 3-pointers and close range shots. He ordered his team to not shoot from midrange and just concentrate on converting 3-pointers and close range shots.

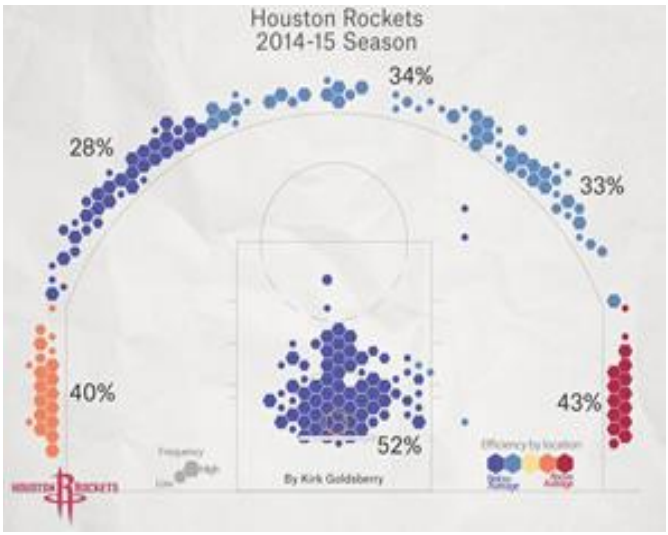

Figure 2. Efficiency of Daryl Morey tactics

Because of this tactic, Houston Rockets reached the western conference for the first time since 1997.

This story proves that effective statistical analysis will help basketball teams perform better.

Research Questions

If statistics can be used in other domains and in other sports such as baseball, why not make use of it in basketball given that huge amounts of data are available.

1. Analysis of Kobe Bryant's strong and weak zones.
2. Performance analysis across seasons.
3. Performance against various opponents.
4. Develop a predictive model, to find if Kobe can convert a shot or not.

The broad-view of this analysis is going to be how well he has performed for the team and how he did in specific conditions and how this can be extended to other players in the team as well.

Technologies

We have used Python packages such a Matplotlib, Seaborn & Plotly and D3 module in Javascript to build rich visualizations to enable effective statistical analysis.

Data

Data Collection

The data has been downloaded from kaggle.com. (Kaggle had extracted the data NBA stats website). The attributes in the data are action_type, combined_shot_type, game_event_id, game_id, lat, loc_x, loc_y, lon, minutes_remaining, period, playoffs, season, seconds_remaining, shot_distance,s hot_made_flag, shot_type, shot_zone_area, shot_ zone_basic, shot_zone_range, team_id, team_name, game_date, matchup, opponent and shot_id.

.

	season	combined_shot_type	game_id	lat	loc_x	loc_y	lon	minutes_remaining	period	playoffs	seconds_remaining	shot_distance	shot_made_flag	shot_type	shot_zone_area	shot_zone_basic	shot_zone_range
2	1996/97	Jump Shot	29600027	33.9283	-140	116	-118.4098	0	1	0	42	18	0	2PT Field Goal	Left Side Center(LC)	Mid-Range	16-24 ft.
3	1996/97	Jump Shot	29600031	33.9473	-131	97	-118.4008	10	2	0	8	16	0	2PT Field Goal	Left Side Center(LC)	Mid-Range	16-24 ft.
4	1996/97	Jump Shot	29600044	33.8633	-142	181	-118.4118	8	2	0	37	23	1	3PT Field Goal	Left Side Center(LC)	Mid-Range	16-24 ft.
5	1996/97	Jump Shot	29600044	34.0443	0	0	-118.2698	6	2	0	34	0	0	3PT Field Goal	Center(C)	Restricted Area	Less Than 8 ft.
6	1996/97	Jump Shot	29600044	33.9063	-10	138	-118.2798	5	2	0	27	13	1	2PT Field Goal	Center(C)	In The Paint (Non-RA)	8-16 ft.
7	1996/97	Jump Shot	29600057	33.8213	-64	223	-118.3338	2	2	0	16	23	1	3PT Field Goal	Center(C)	Mid-Range	16-24 ft.
8	1996/97	Jump Shot	29600057	33.8673	-79	177	-118.3488	1	3	0	53	19	0	2PT Field Goal	Left Side Center(LC)	Mid-Range	16-24 ft.
9	1996/97	Jump Shot	29600057	33.8373	-103	207	-118.3728	1	3	0	14	23	1	3PT Field Goal	Left Side Center(LC)	Mid-Range	16-24 ft.
10	1996/97	Layup	29600057	34.0443	0	0	-118.2698	0	3	0	2	0	0	2PT Field Goal	Center(C)	Restricted Area	Less Than 8 ft.
11	1996/97	Jump Shot	29600057	33.8693	-155	175	-118.4248	9	4	0	9	23	0	3PT Field Goal	Left Side Center(LC)	Mid-Range	16-24 ft.
12	1996/97	Layup	29600057	34.0443	0	0	-118.2698	8	4	0	36	0	0	2PT Field Goal	Center(C)	Restricted Area	Less Than 8 ft.

Figure 3 Sample dataset showing various features

Pre-Processing

The data consisted of some incomplete records which were excluded for the further processing. The irrelevant features such as game_event_id, game_id, matchup, team_id, etc. in the data were also removed, cutting down the attributes to nineteen.

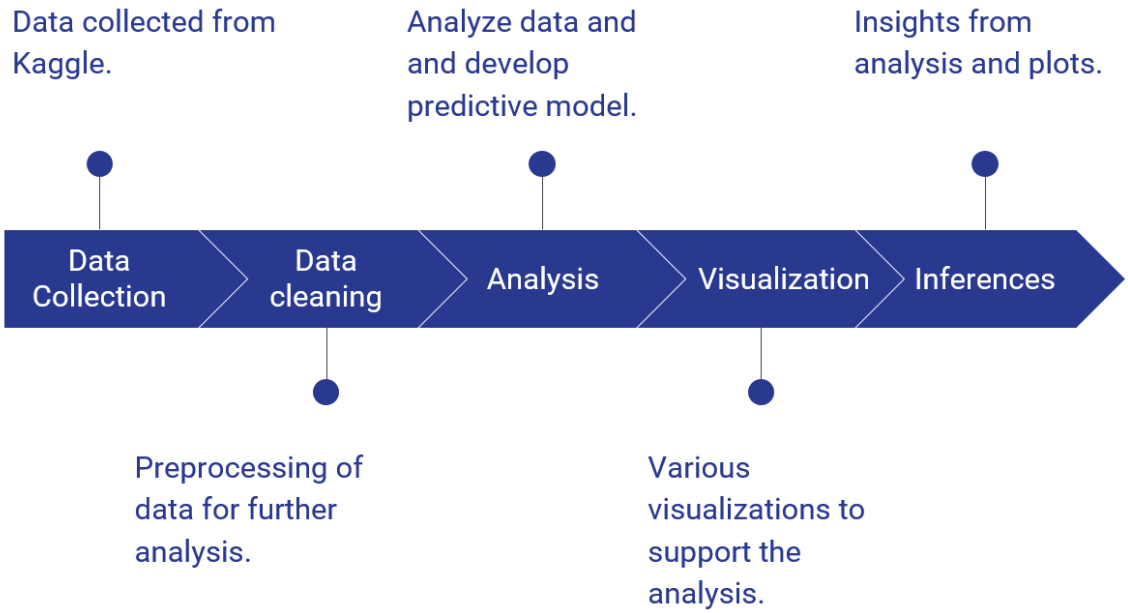

Figure 4 Image showing the overall process of the analysis

Exploratory Data Analysis

Shot Accuracy Across Seasons

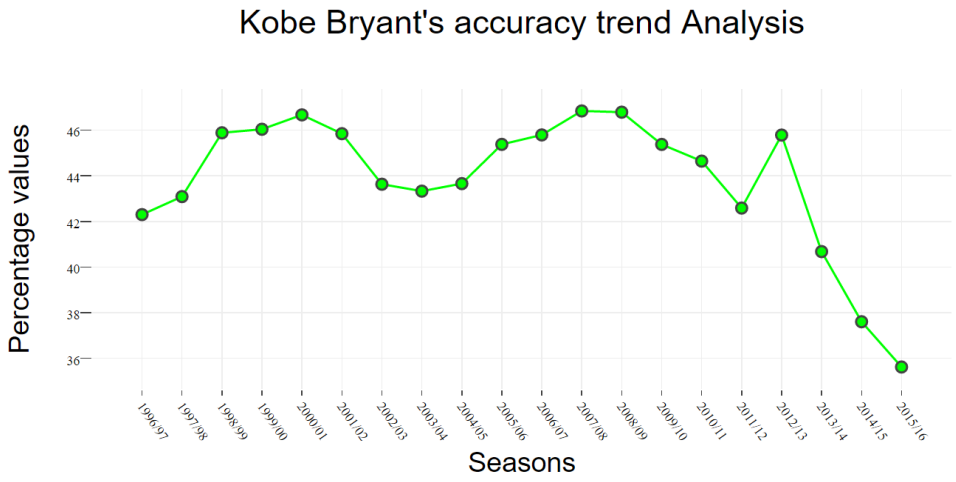

Figure 5 Trend showing Kobe's shot accuracy across seasons

The above graph shows the accuracy trend of Kobe Bryant. We could observe that he has faced ups and downs in his career with his peak during 2004-2007. Also, we could see a sharp decline from the year 2012, this might be because of the injuries that he encountered which made him play only few games.

Why line graph?

To observe this data, we had choices ranging from bar graph, scatter plot, line graph. We choose line graph from these because for plotting time-series data line graph make more sense and it was easy to visualize the trend.

Shot Accuracy Based on Shot Zone Area

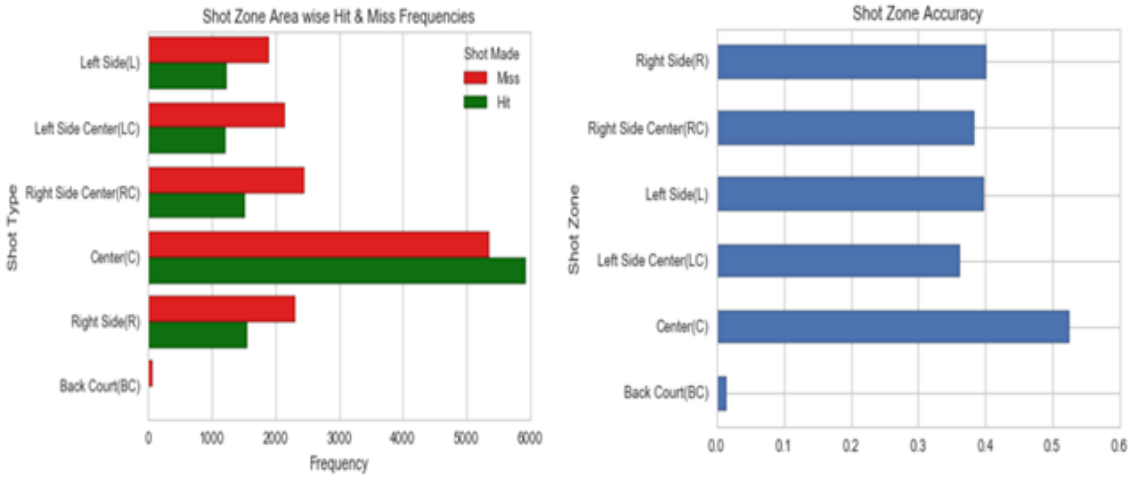

Figure 7 Plots showing Kobe's performance over various shot zones

From these graphs, we could observe the performance of Kobe Bryant based on the shot zone. He has the highest accuracy when we attempt from the center stage and his majority takes is from the center area. Also, comparing his left and right side, he prefers to shoot from the right side.

Why bar graph?

Since this data involves categorical data and frequency and accuracy, bar graph represents these data the best. A scatter plot or line graph would not be able to visualize this data with the impact a bar graph creates. Therefore, we choose bar graph over other graphs.

Shot Accuracy Based on Shot Range

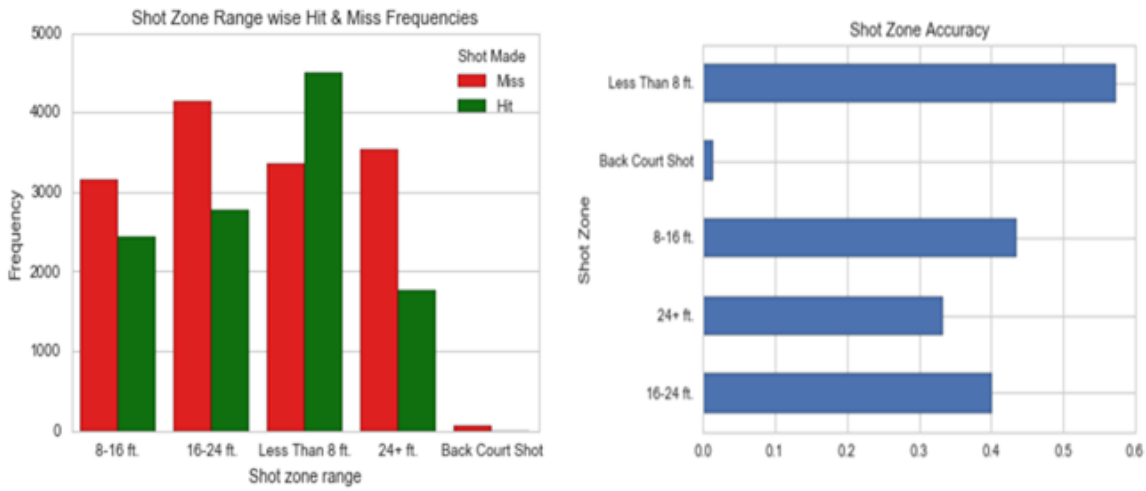

Figure 8 Plots showing Kobe's performance over various shot ranges

This plot shows the effectiveness of Kobe Bryant across shot ranges. We could see that he is successful when nearing the basket court and his success rate decreases as the distance increases.

Why bar graph?

Since this data involves categorical data and frequency and accuracy, bar graph represents these data the best.

A scatter plot or line graph would not be able to visualize this data with the impact a bar graph creates. Therefore, we choose bar graph over other graphs.

Visualizations

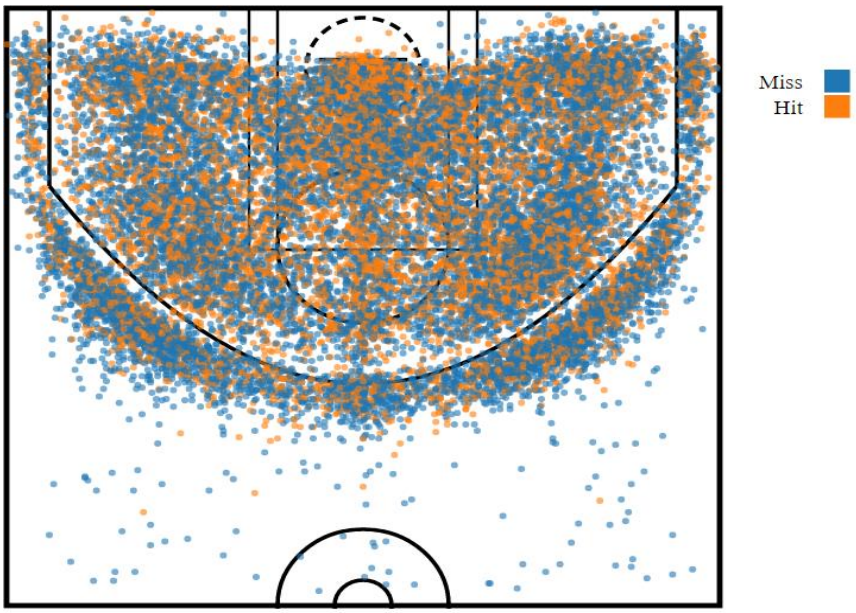

Fig 9: Hit and misses of Kobe Bryant on court

The above scatter plot shows the hit and miss location of Kobe Bryant on basketball court, this helps us to visualize his positive and negative areas on court.

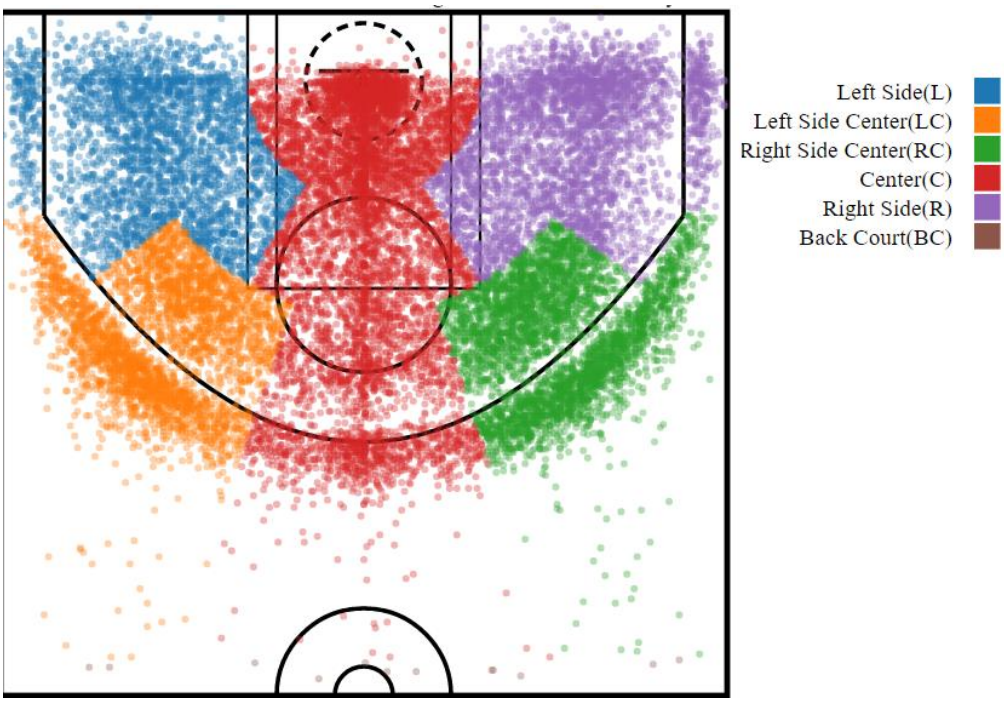

Fig 10: Kobe Bryant attempts from which area of court

This graph shows the areas from which Kobe Bryant tries, this gives us a clear idea of his preference and how he has done in each area on court.

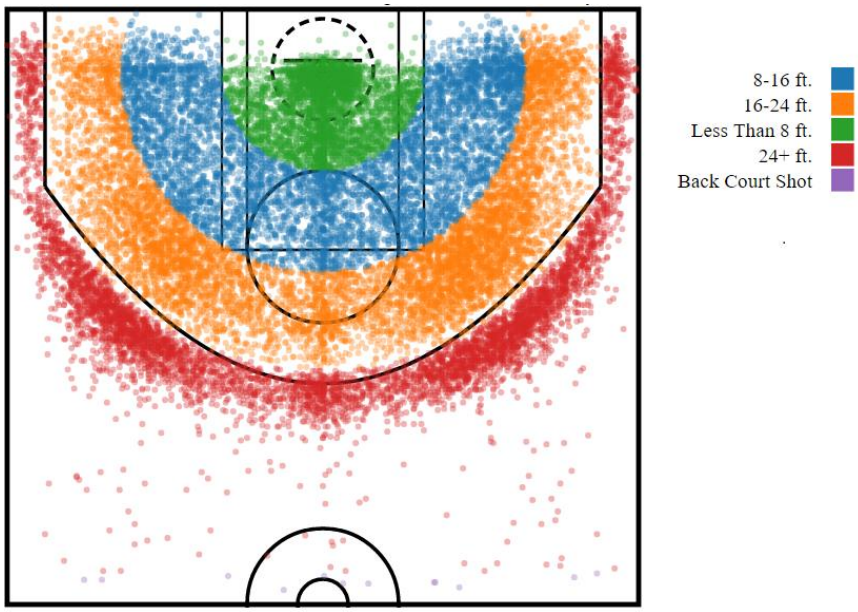

Fig 11: Kobe Bryant attempts across various shot zones

This graph shows the attempts made from Kobe Bryant based on various shot distances. This graph can give us an idea the range that he prefers against various teams and across seasons.

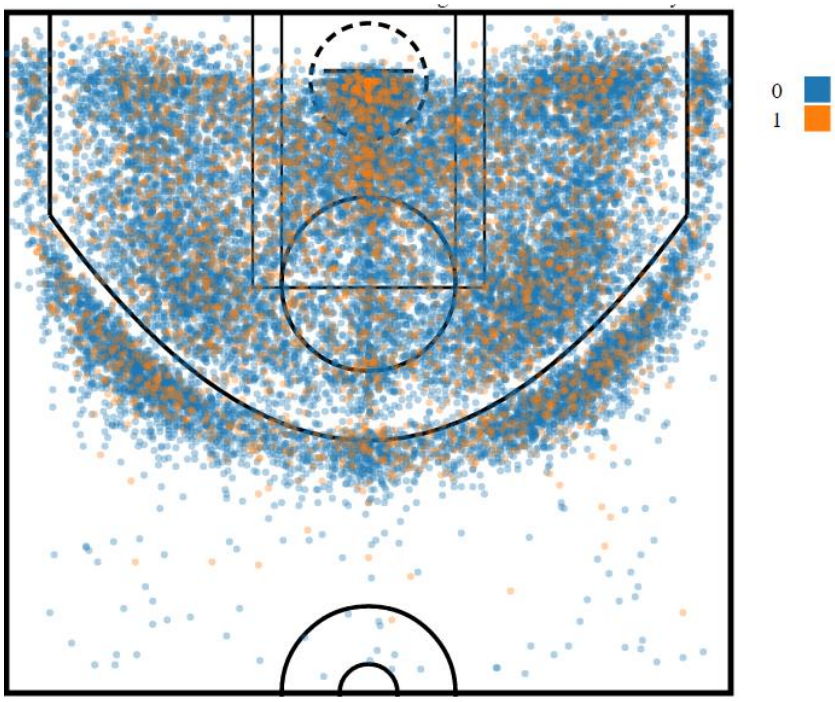

Fig 12: Kobe Bryant shot attempts based on playoffs

This graph shows the attempts made by Kobe Bryant based on if the match was a playoff or not. We could see that during playoff matches Kobe prefers to shoot from a close range.

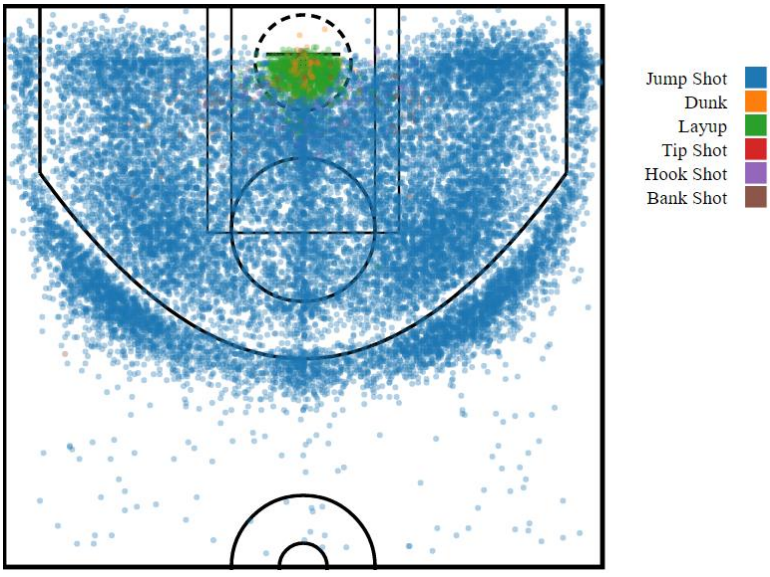

Fig 14: Kobe Bryant shot attempts based on types

We could see that Kobe's stock shot is the 'jump shot'. He does dunk a lot during his initial phase of his career, but as time goes the dunk and close in shots fades away.

Note: All these scatter plots on court are designed in JavaScript, so filtering based on opponents and seasons can be done for all these graphs. We choose a scatter plot because it is best way to represent data on court. There can't be any other alternative graph for this representation.

Shots attempted by Kobe every 24 seconds

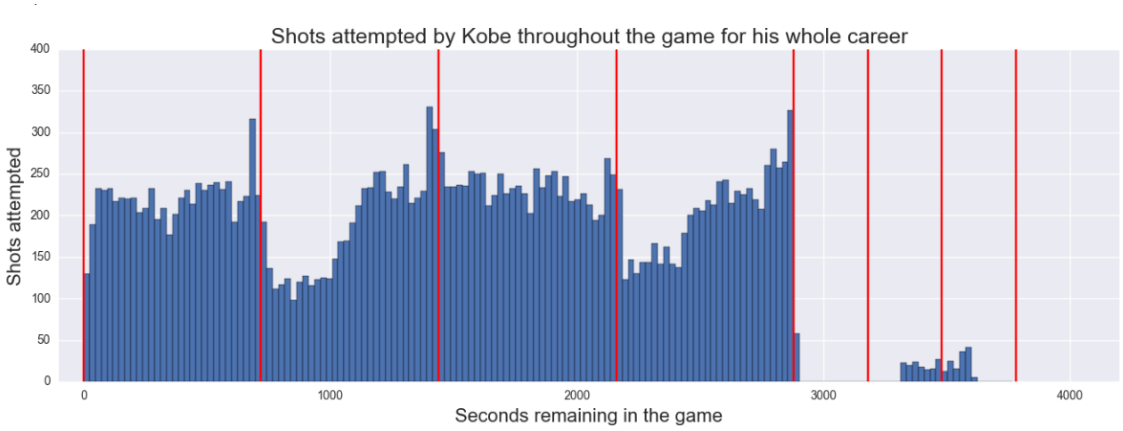

Fig 15. Shot attempted by Kobe across each 24 seconds

This graph shows the number of shots attempted by Kobe Bryant for each 24 seconds from the start of first quarter till the game ends for all games that he has played. Also, the red bars correspond to the end of a quarter.

We could observe that Kobe shot attempts increases during the end of all quarters, this might be because he is being treated as the go to man for the LA Lakers. Also, the number of shots attempted during the initial period of second and fourth quarter is less compared to first and third. This might be because he has been rested during this period so that he can excel during the quarter end.

Why histogram?

Since we want to find the frequency with a gap of 24 seconds, we prefer a histogram since it can bin the data and give us the count needed. Hence, histogram is the best choice in this case.

Average shot distance across seasons

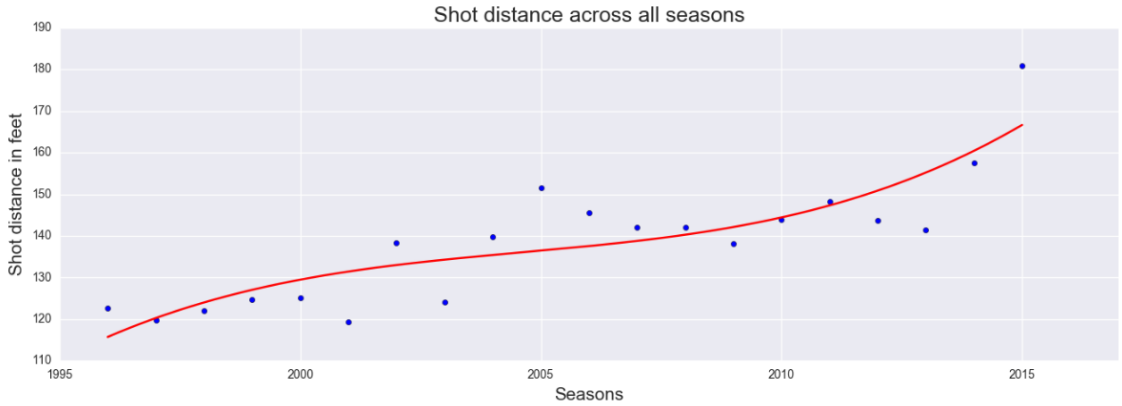

Fig 16. Average shot distance across seasons

This graph shows the average distance from which Kobe Bryant shoots for all the seasons that he has played. We could observe that as season passes his average distance increases. This might be because of factors such as his change in game play, age, change in team strategy and others. Also, we fitted a cubic line to this data to observe the trend.

Why scatter plot?

To visualize this data, we have options ranging from bar graph, line graph, scatter plot. We choose scatter plot as that can be clean in this case. Bar graph can be good choice only if we have categorical data on the x-axis. We did not go with a line graph because we can't have one smooth like the red line which shows the trend. So, we decided to go with a scatter plot with the red smooth line showing the trend.

Shot accuracy based on distance

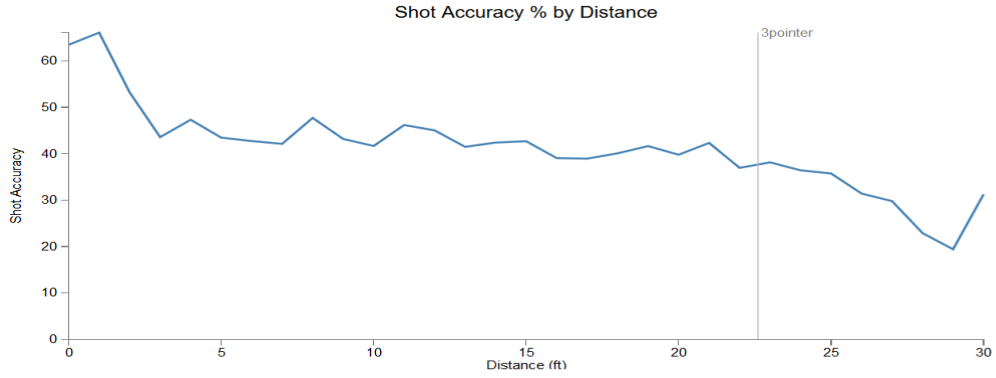

Fig 17. Shot accuracy based on distance

From this graph, we could observe that the shot accuracy decreases based on the distance. Kobe is exceptionally well near the basket court and as distance increases his accuracy decreases which is obvious.

Why line graph?

With line graph this trend is much easier to show compared to other graphs. It gives a clear idea how the shot accuracy behaves as a function of distance. This is the reason why we choose line graph.

Shot type used across seasons

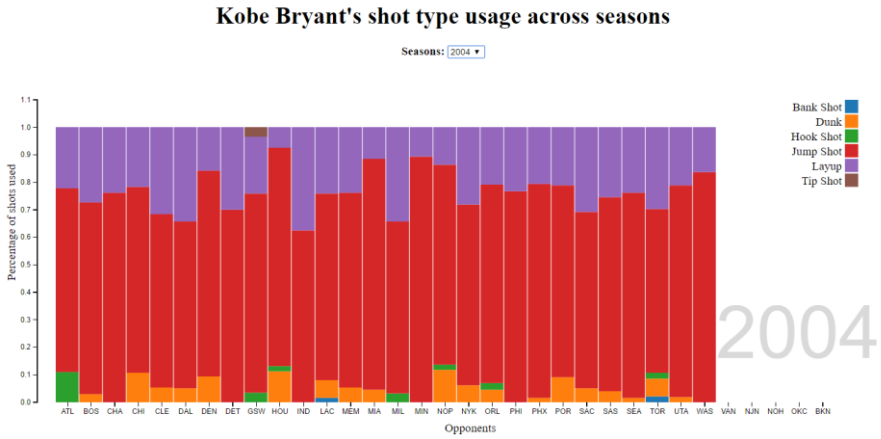

Fig 18. Shot type against opponents across various seasons

From this graph, we could see the percentage of various shots that Kobe Bryant prefers against various opponents for all the seasons. We could observe that during his initial career and during his peak he uses dunk shots which fades during his career end. The stock shot for Kobe remains as jump shot.

Why bar graph?

We prefer bar graph here because of the stacked nature of the data, we could have used multiple line graphs to represent the data but it would have been bit difficult for comparison, that is the reason why a stacked bar graph is chosen.

Kobe's accuracy over each minute through various periods

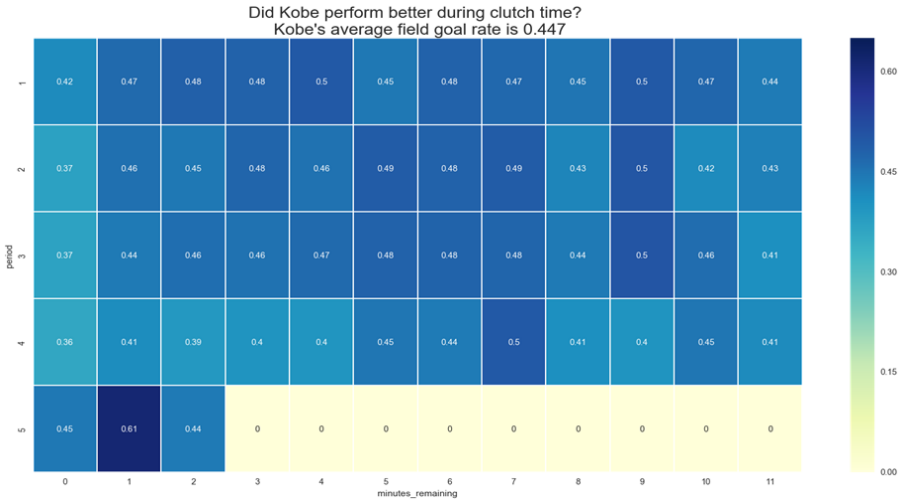

Fig 19. Kobe's shot accuracy across minutes and periods

This heatmap shows the accuracy of Kobe Bryant across periods and in each minute. We could see that he does well during the end minutes and as well as during the middle of quarters.

Why this heat map?

We choose this map because this can easily represent the data that we want to represent, since there are parameters such as period, minutes, accuracy. We can't easily fit a bar graph or a line graph with this data.

Inferences

From the various visualizations and graphs that we generated we can assert the following statements.

1. Kobe Bryant has been the go to man for LA Lakers, proves why he has been a star.
2. As season passes he focuses more on the three-pointer compared to two pointer.
3. Kobe has a slight inclination towards the right side of the court compared to the left and center side.
4. He performs takes a sharp dip during his end, this might be because of injuries and less number of games that he played.
5. There is a decrease in the usage of dunk shots as his career progresses.
6. He has missed more shots than he converted.

Predictive Model

Model Development Overview

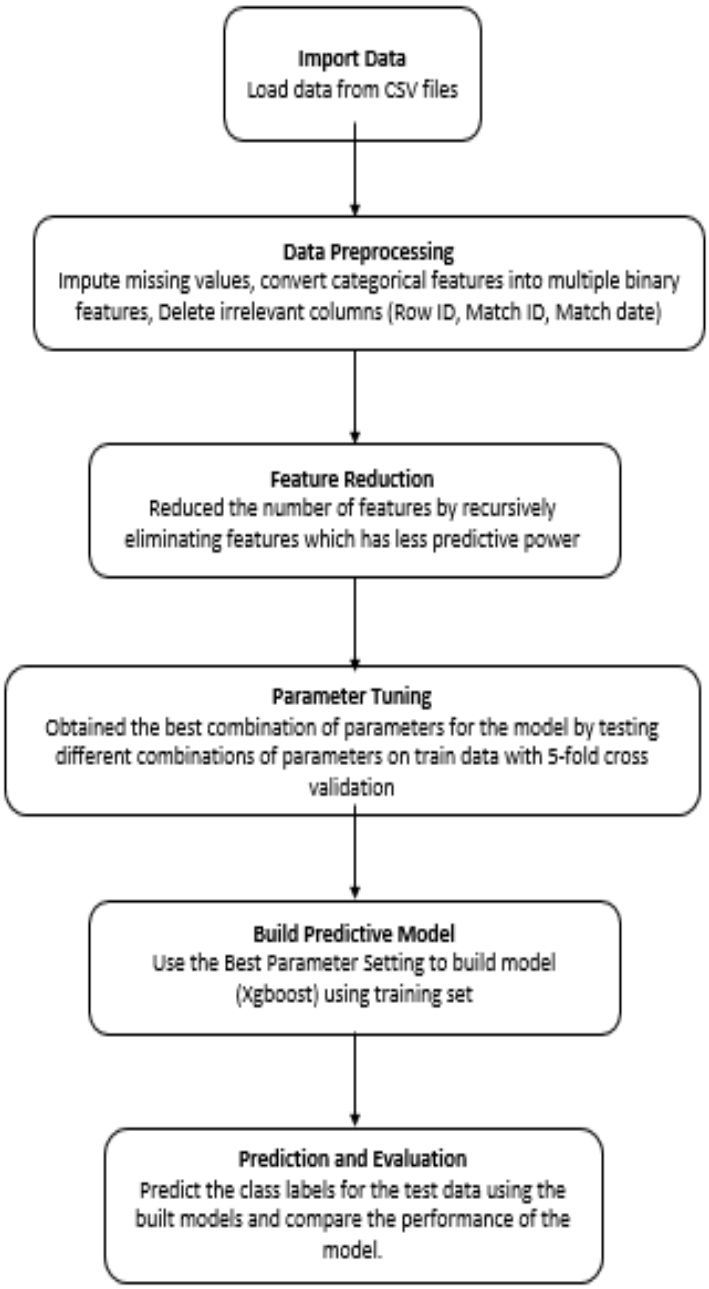

Fig 20: Predictive Model Process

Data Preprocessing

Missing data were imputed with mean for numerical features and mode for categorical features. The categorical features were converted to series of binary features.

Feature Selection

The features were ranked using the method of recursive feature elimination. The process works by eliminating the least important feature in each iteration. The features were ranked based on their importance.

Parameter Tuning

The parameters values to consider were varied. Then Xgboost was built using all possible combinations of parameters and the model was evaluated using 5-fold cross validation of the train set. Then the best set of parameters were chosen based on the result of the 5-fold cross validation.

Best Model Parameters - max_depth=4, n_estimators=600, learning_rate=0.01

Model Building

The Xgboost model was built using the best combination of parameters obtained from the previous step.

Evaluation

Metrics such as AUC, Accuracy, Precision, Recall and F1-Score were calculated.

Result

Accuracy	68.28%
Precision	69%
Recall	68%

Table 1. Performance metrics of the predictive model

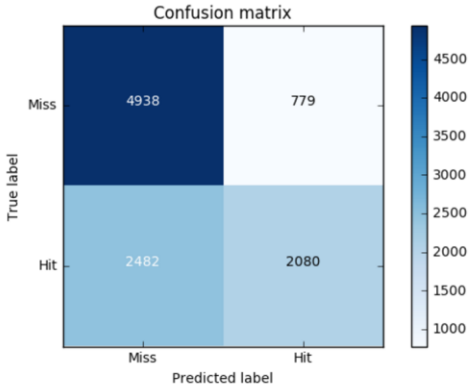

Fig 21. Confusion matrix of the predictive model

Conclusion

From our analysis, we could make the following conclusions

1. These types of visualization and analysis can really assist the sports informatics and can elevate the player and team's performance.
2. It can also serve in comparing players' performance and extended to teams performance as well.
3. Machine Learning algorithms can be utilized to predict a player's performance to a certain extent.

Future work

1. Several players' data could be collected and the same analysis can be done for comparison. This can be extended to team as well.
2. These types of analysis can be extended to other sports like cricket, tennis etc where statistics are not used extensively.

References

1. Kaggle site: https://www.kaggle.com/c/kobe-bryant-shot-selection
2. NBA stats website: http://stats.nba.com/
3. Kobe Bryant: https://en.wikipedia.org/wiki/Kobe_Bryant
4. Python plotly: https://plot.ly/
5. http://www.houstonchronicle.com/news/article/The-cult-of-analytics-with-Rockets-general-4342412.php
6. Our Code: https://goo.gl/cLorgU
7. Live Demo: http://www.vipulmunot.com/DataViz/