Springboard Capstone Project 2: Analysis of NBA Dataset

Kaggle Dataset Repository

Final Report by: Michael Thabane

May 16, 2020

Contents

1 Introduction **3**

1.1 The Business Problem 3

1.2 Executive Summary 3

2 Approach **3**

2.1 Data Acquisition and Wrangling 3

2.2 Storytelling and Inferential Statistics 4

2.2.1 Descriptive Analysis 4

2.2.2 Hypothesis Testing 8

1 Introduction

1.1 The Business Problem

In the last 30 years the NBA has been one of the fastest growing sports in the world. With the advancement of sports analytics and the prevalence of gambling there is a high demand for player statistic predictions. The business problem we would like to solve is creating a prediction model for player stats, in particular points; using previous game data. We will also explore specific players to find potential variables that have a direct effect on the points a player scores. The plan is to model the data using time-series forecasting techniques to predict the number of points a player will score in an upcoming game.

1.2 Executive Summary

2 Approach

2.1 Data Acquisition and Wrangling

The dataset used for this project was acquired from Kaggle but was originally scraped from the NBA statistics website using the NBA API. The dataset consists of four different spreadsheets. The first spreadsheet (games.csv) contains team boxscore data which consists of total team statistics. The second spreadsheet (game_detail.csv) contains the player boxscore data which consists of individual player statistics for each game. The third spreadsheet (players.csv) contains player information for which team they play for. The fourth spreadsheet (ranking.csv) contains the nba standings for every day in the each season. In order to have all the relevant information in the same spreadsheet some data wrangling was need to get column into the game detail spreadsheet. First, we want to include the team stats for any given game to be in the player boxscore data. To achieve this we merged the team data from the games spreadsheet into the game detail spreadsheet but only included the team stats of the team the player is playing for. Then, we will manipulate data to create any variables that may have influence on the outcome of a players points. I have created the running average to include the players average over the previous three and five games.

2.2 Storytelling and Inferential Statistics

2.2.1 Descriptive Analysis

With the dataset containing information since 2003 it makes more sense to look into players that have had more of an impact scoring for their teams recently. Below is a graph which shows the top ten scorers in the last two seasons (2018 and 2019).

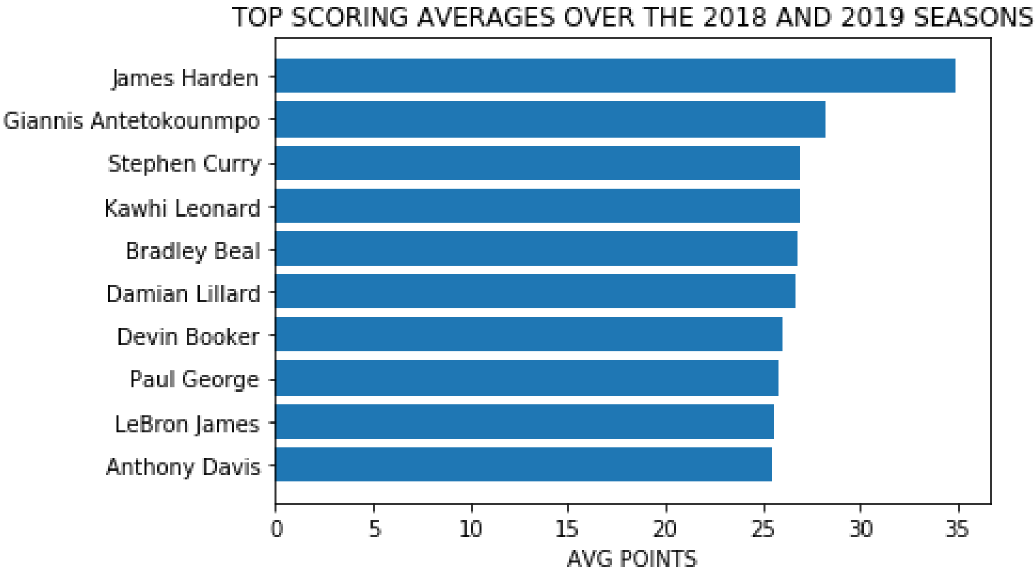

Figure 1: Top 10 Scorers Over 2018 and 2019 Seasons

From above you will notice that there are no centers in the top 10 list of scorers. Based on the recent NBA rule changes the game has strayed from making centers the focal point of an offense. Now the center position has been watered down with a lot less dominant big man in the league. Then new style of play has more spread out offenses and less focus on dominant interior play or in and out basketball. To examine this we can at look distribution of points scored by starters from the positions forward, guard and center and look to see if we can find this trend.

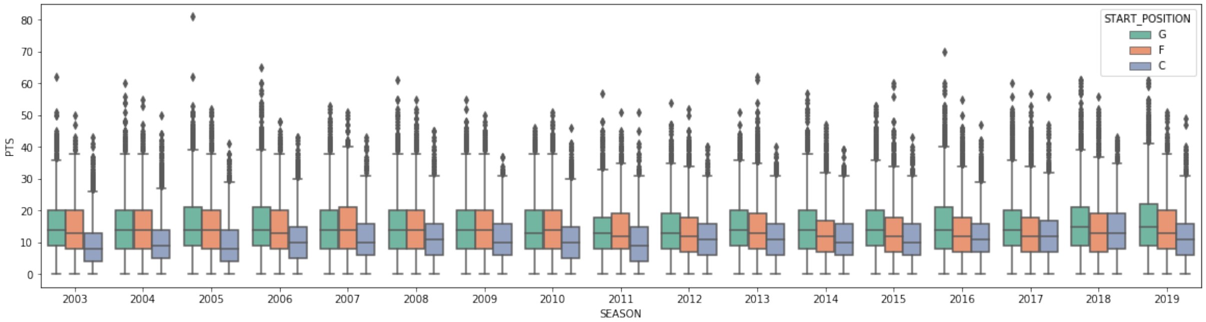

Figure 2: Boxplot of Starters Points

From the plot above it does not seem to show this trend as centers are scoring at a similar rate over the last 20 years. However if we plot the number of three-point field goals attempted it might show that interior players have gone from never shooting threes to spreading the floor.

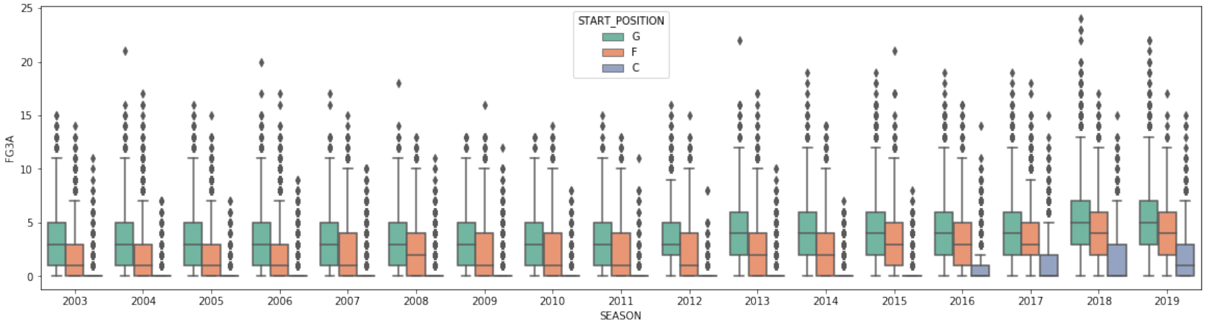

Figure 3: Boxplot of Starters 3PT Field Goals Attempted

From the plot above we can see that in the last four years starting centers have begun to spread the floor by shooting more threes. If we had the data of points scored in the paint that should show a trend of less points being scored in the paint by centers as the game has evolved.

From Figure 1 we can see that James Harden has had the highest scoring average. However, this has not always been the case throughout his career. His role for the OKC Thunder which is the team that drafted him was to provide scoring off the bench. Now his role with his current team is to be the focal point of the offense and provide scoring and leadership. To further investigate this we will use a violin plot to examine his point distribution throughout his career.

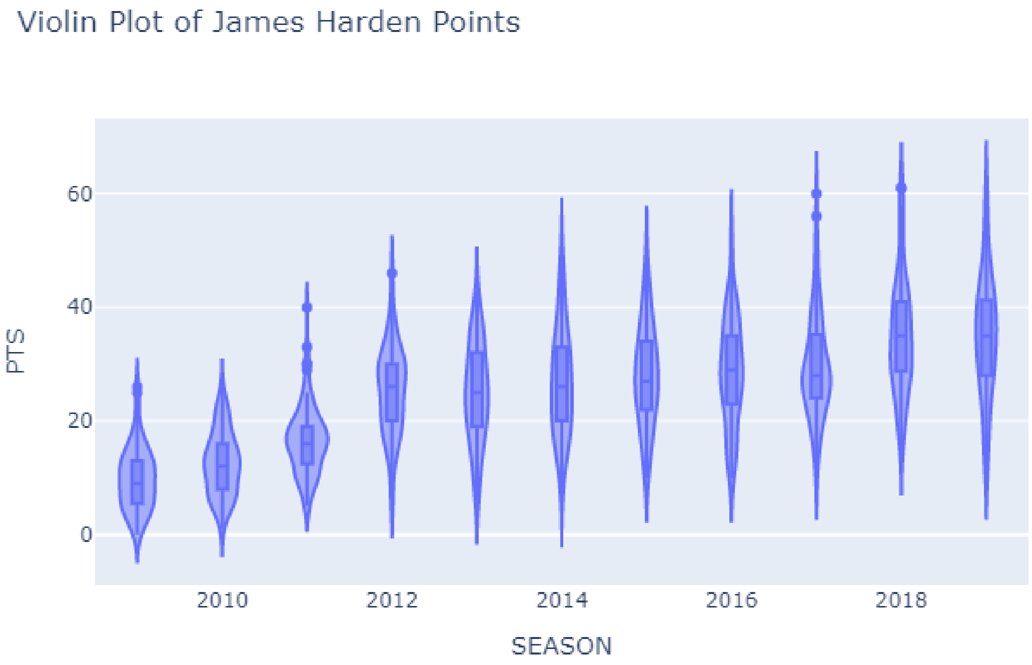

Figure 4: Violin Plot of James Harden

The plot above shows that when James Harden was traded from OKC to Houston at the start of the 2012 season his point average and distribution made a significant increase. This shows how player movement whether it is through free-agency or trades can make a significant impact on the volume of scoring a player will provide for his team. However, it is possible for a player's role to change with the team over time without a trade occurring. This can be attributed to both player development as well as team needs change from season to season. This can be perfectly illustrated by Pascal Siakam who has both developed as a player and has had an increased role due to the departure of the teams star player. A violin plot of Siakam's points help illustrate this.

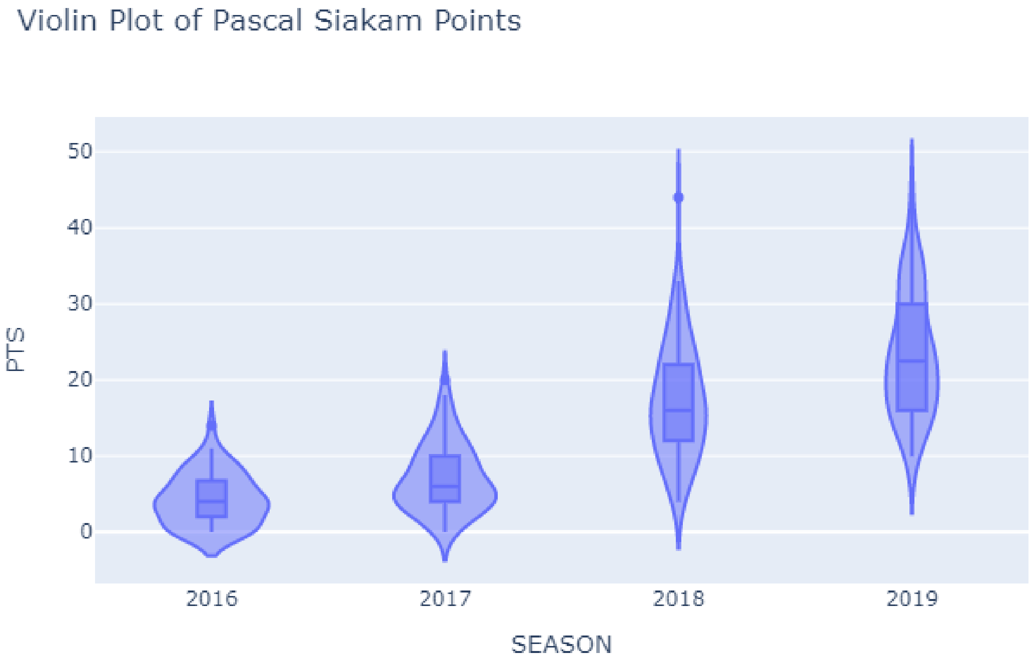

Figure 5: Violin Plot of Pascal Siakam Points

The above plot shows a big jump in point production from the 2017 to 2018 and slight increase from 2018 to 2019. However, the violin plot alone will not give you an idea of what has caused the increase. I will use a violin plot of Siakam's minutes to illustrate how his starting role change in the 2018 season impacted his points compared to how Kawhi's departure in the 2019 season has affected his point distribution.

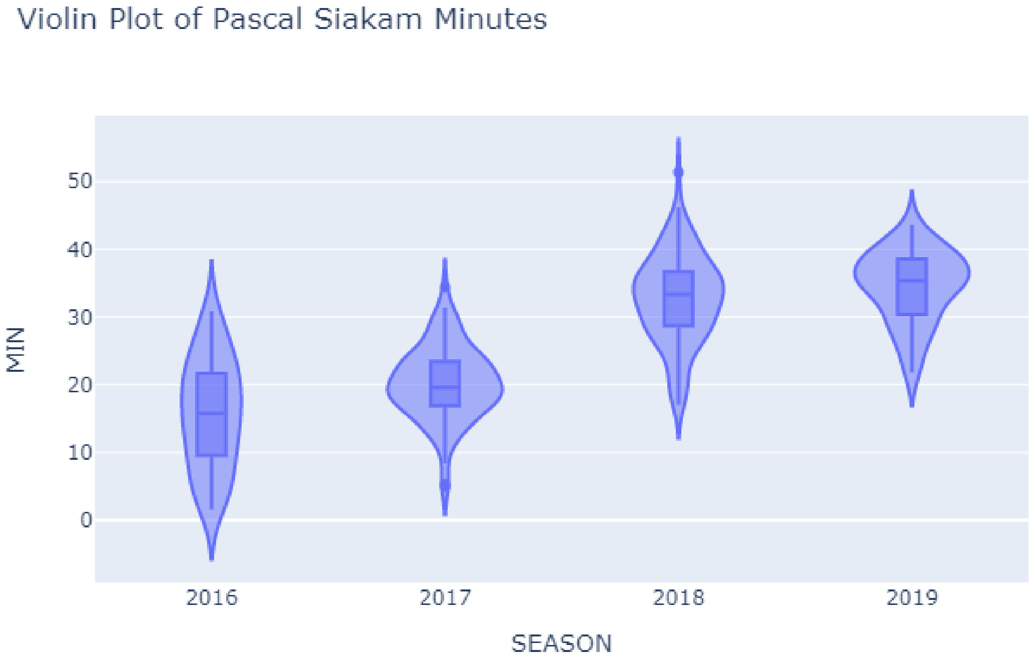

Figure 6: Violin Plot of Pascal Siakam Mins

2.2.2 Hypothesis Testing