Final Project CSCI 5040 – Domas Budrys

This report will try to answer the question: Does playing more minutes in the NBA guarantee scoring more points. Based on personal experience and passion for athletics, especially basketball, the outcome should demonstrate the results that support the idea of correlation between minutes spent on the court and points scored.

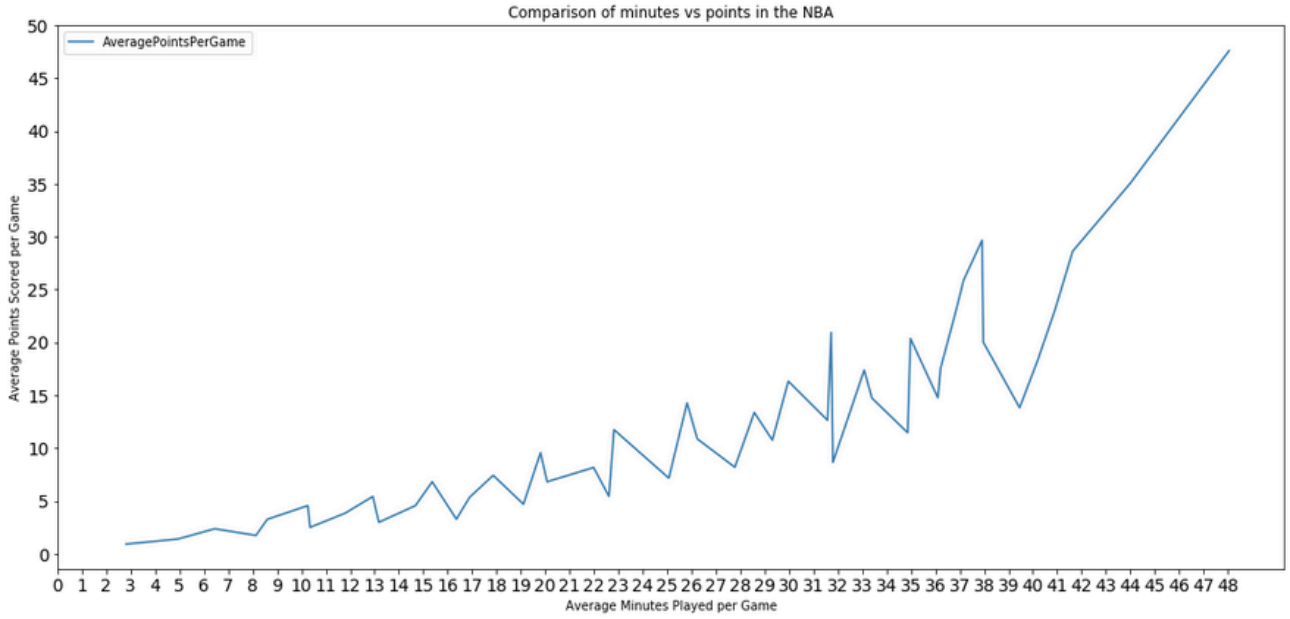
Data used for the analysis was downloaded from (1). It contains three different .csv format files named: player_data.csv, Players.csv, Seasons_Stats.csv. For the purpose of this project, only Season_Stats.csv file was used because only game statistics were takin into account to determine the benefit of spending more time on the court. Other two .csv files were mostly focused on basketball player's physical measurements.

Season_Stats.csv file contains 24691 records and its' time frame is between 1950-2017, which is an appropriate amount of information to provide answers. Before focusing on clustering, this file needed to be cleaned and well formatted. First, all of the rows containing `null` values in MP(Minutes Played) column were removed. Also, we have taken in to the account that this dataset contains entries when 24 second rule was not applied, and which led to less points scored by the players (2). Dataset was reduced to the timeframe between 1955-2017. Dataset contains records based on yearly total, such as: total games played, total points scored, total minutes played per year, and etc. Yearly averages needed to be calculated and group by players name and a year, since some of the players have played for multiple teams throughout the span of one year.

After properly formatting the dataset and creating two new attributes such as, AveragePointPerGame, and AverageMinPerGame, dataset was completed and ready to for

KMeans clustering analysis. The number of clusters was set to 48, since there are only 48 minutes played throughout the game and based on generated data there were several players that have reached that number. After fitting data on with KMeans algorithm, 48 different clusters were generated. It is important to mention that, this project was entirely focused on cluster centers rather then cluster groups itself because center values were able to proving a grouped information based on average points scored and minutes played per game.

Finally, a graph was created with Python pandas.DataFrame.plot object using cluster centers to provide visual outcome from collected results. The correlation between minutes played and points scored can be clearly seen, however there are several timeframes where players are able to provide more points wiling playing less minutes.



Web reference and source file location

(1) **Souce data:** https://www.kaggle.com/drgilermo/nba-players-stats.

(2) **24sec article**: https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=2&cad=rja&uact=8&ved=2ahUKEwjdj9Ozp4vfAhVLMt8KHU1cC1IQFjABegQIBRAF&url=http%3A%2F%2Fwww.wbur.org%2Fonlyagame%2F2015%2F04%2F22%2Fnba-shot-clock-history-basketball&usg=AOvVaw2l15teRRzmS_KzHoncCn36