NBA User Story 2 Report and Analysis

Introduction

This report analyzes data collected on NBA players' performance statistics collected between the years of 1950 and 2021. The client has requested an analysis of players' free throws and how this variable can be used to analyze overall performance, especially the percentage of total points scored that come from free throws. The data is split into two timeframes, 1970-2021 (the years since the NBA merger) and 2000-2021. Additionally, the client asks whether total points score or minutes played affects the accuracy of the analysis.

Body

Data

This report was written using MyJupyterNotebooks. First I imported Pandas, MatPlotLib, Seaborn, and Numpy to analyze and write reports on the data. Next I imported the NBA data from a .csv file. Then I requested the number of rows and columns (26176 and 20 respectively), the data types of each column (all were float except for the players' names and city(team)), and the column names. I replaced null values with 0 so that I could perform math operations with the columns, then found the percentage of points made by free throws per year by dividing the FTM column by the PTS column and then multiplying by 100. I stored this result in a new column called FT% (for free throw percentage).

After this, I created three dataframes. The first contained the median 10 players, when sorted by average points scored (grouped by player with one data point per year recorded). For the second dataframe, I dropped all years before 1970 and then calculated the top 5 players when sorted by free throws made - this gave me my dataframe for the years since the NBA merger. To create my dataframe for the modern era, I dropped all years before 2000 and then calculated the top 5 players when sorted by free throws made.

Method

First, I used my median players dataframe to calculate the average of the minutes played, points total, and percent of points from free throws. Using these averages, I created a heatmap of the correlation. I found that there was a slight positive correlation between free throw percentage and average total points, and a strong negative correlation between free throw percentage and average minutes played. Then I created two linear regression line plots to show these two correlations.

Next, I used my NBA Merger dataframe's top 5 players to create a linear regression line plot of percentage of free throw points (from total points) by year. Karl Malone was the top player for

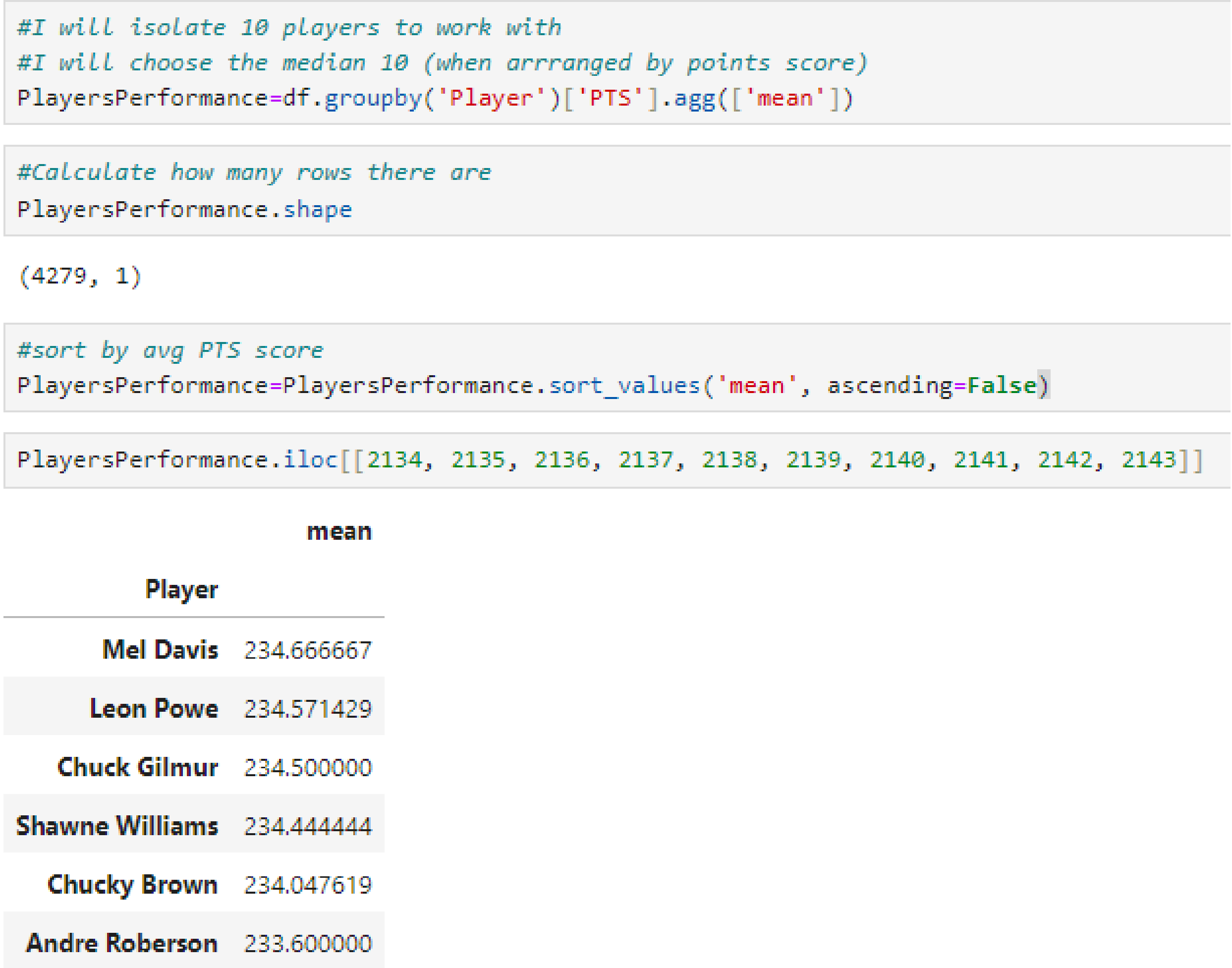
this timeframe, so I also separated his data from the others and used it to generate a line plot showing the number of his free throws and total points scored.
Finally, I used my Modern dataframe's top 5 players to create another linear regression line plot of percentage of free throw points (from total points) by year. James Harden was the top player for this timeframe, so I created a line plot for him too showing his free throws made against his total points scored.

Results

There is a strong positive correlation between the top players' free throws and the total points that they score, indicating that number of free throws is a dependable variable to use when predicting total score. Additionally, there is a strong negative correlation between median players' total points *and* their free throw percentage against minutes played, indicating that the number of minutes that the player is in the game does not have an effect on their total score or the number of their free throws.

Analysis

Finding the median players:



```
#I will isolate 10 players to work with
#I will choose the median 10 (when arrranged by points score)
PlayersPerformance=df.groupby('Player')['PTS'].agg(['mean'])
```

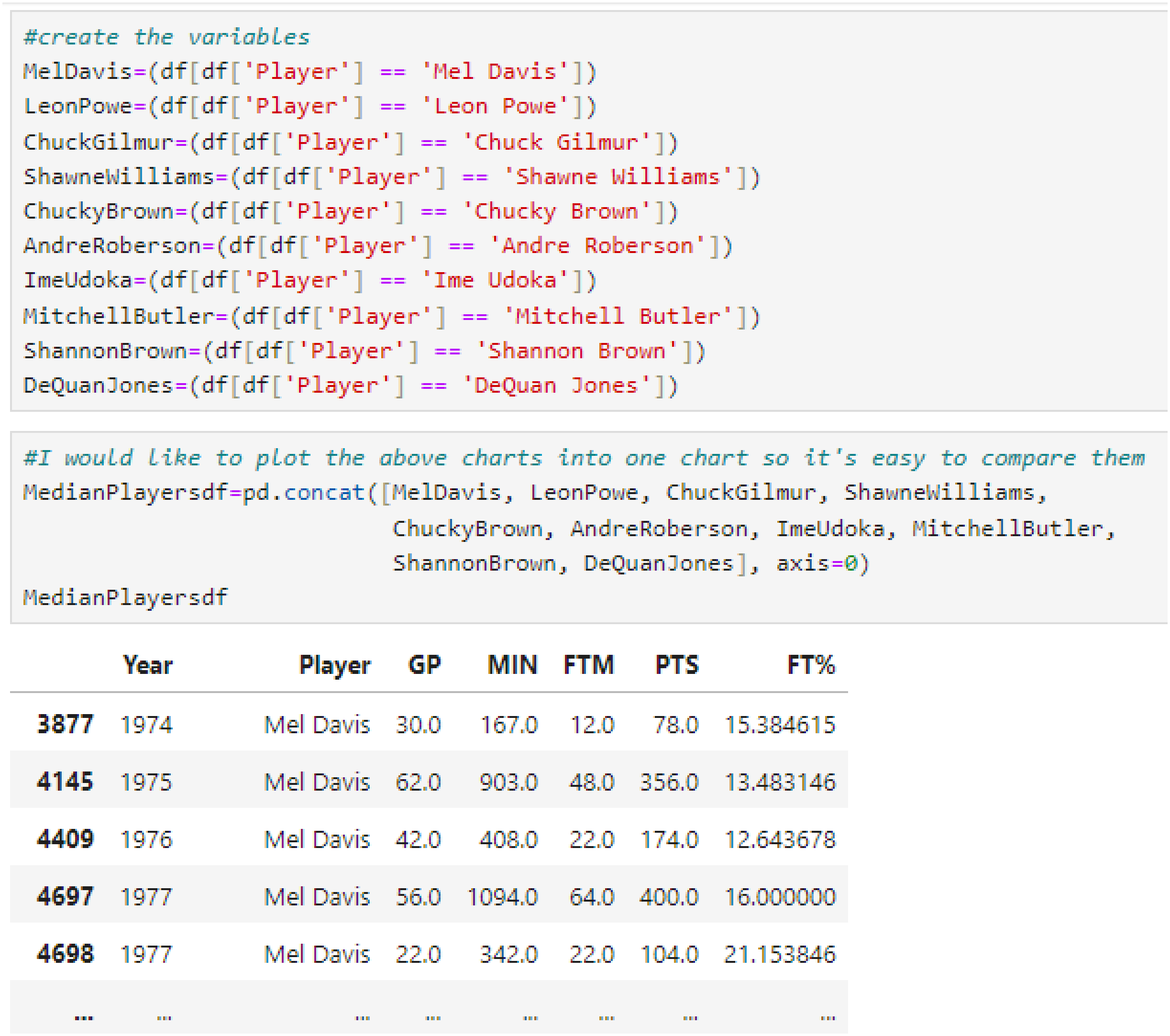
```
#Calculate how many rows there are
PlayersPerformance.shape
```

```
(4279, 1)
```

```
#sort by avg PTS score
PlayersPerformance=PlayersPerformance.sort_values('mean', ascending=False)
```

```
PlayersPerformance.iloc[[2134, 2135, 2136, 2137, 2138, 2139, 2140, 2141, 2142, 2143]]
```

	mean
Player	
Mel Davis	234.666667
Leon Powe	234.571429
Chuck Gilmur	234.500000
Shawne Williams	234.444444
Chucky Brown	234.047619
Andre Roberson	233.600000

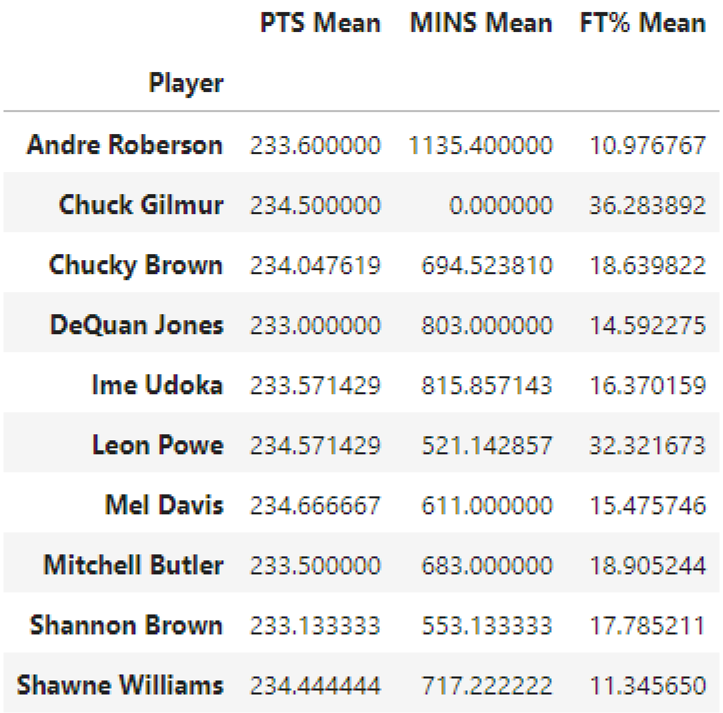


```
#create the variables
MelDavis=(df[df['Player'] == 'Mel Davis'])
LeonPowe=(df[df['Player'] == 'Leon Powe'])
ChuckGilmur=(df[df['Player'] == 'Chuck Gilmur'])
ShawneWilliams=(df[df['Player'] == 'Shawne Williams'])
ChuckyBrown=(df[df['Player'] == 'Chucky Brown'])
AndreRoberson=(df[df['Player'] == 'Andre Roberson'])
ImeUdoka=(df[df['Player'] == 'Ime Udoka'])
MitchellButler=(df[df['Player'] == 'Mitchell Butler'])
ShannonBrown=(df[df['Player'] == 'Shannon Brown'])
DeQuanJones=(df[df['Player'] == 'DeQuan Jones'])
```

```
#I would like to plot the above charts into one chart so it's easy to compare them
MedianPlayersdf=pd.concat([MelDavis, LeonPowe, ChuckGilmur, ShawneWilliams,
                           ChuckyBrown, AndreRoberson, ImeUdoka, MitchellButler,
                           ShannonBrown, DeQuanJones], axis=0)
MedianPlayersdf
```

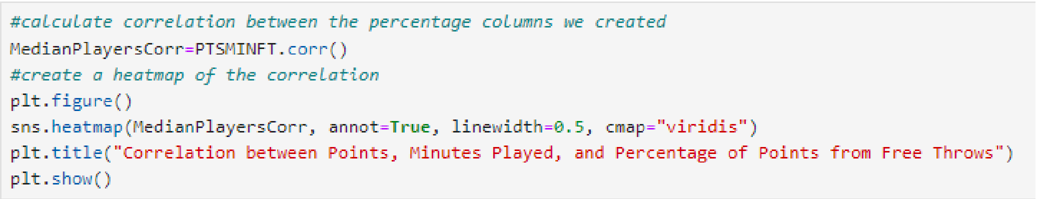
	Year	Player	GP	MIN	FTM	PTS	FT%
3877	1974	Mel Davis	30.0	167.0	12.0	78.0	15.384615
4145	1975	Mel Davis	62.0	903.0	48.0	356.0	13.483146
4409	1976	Mel Davis	42.0	408.0	22.0	174.0	12.643678
4697	1977	Mel Davis	56.0	1094.0	64.0	400.0	16.000000
4698	1977	Mel Davis	22.0	342.0	22.0	104.0	21.153846
...	...	...	...	...	...	...	...

Calculating percentage score of the metrics:

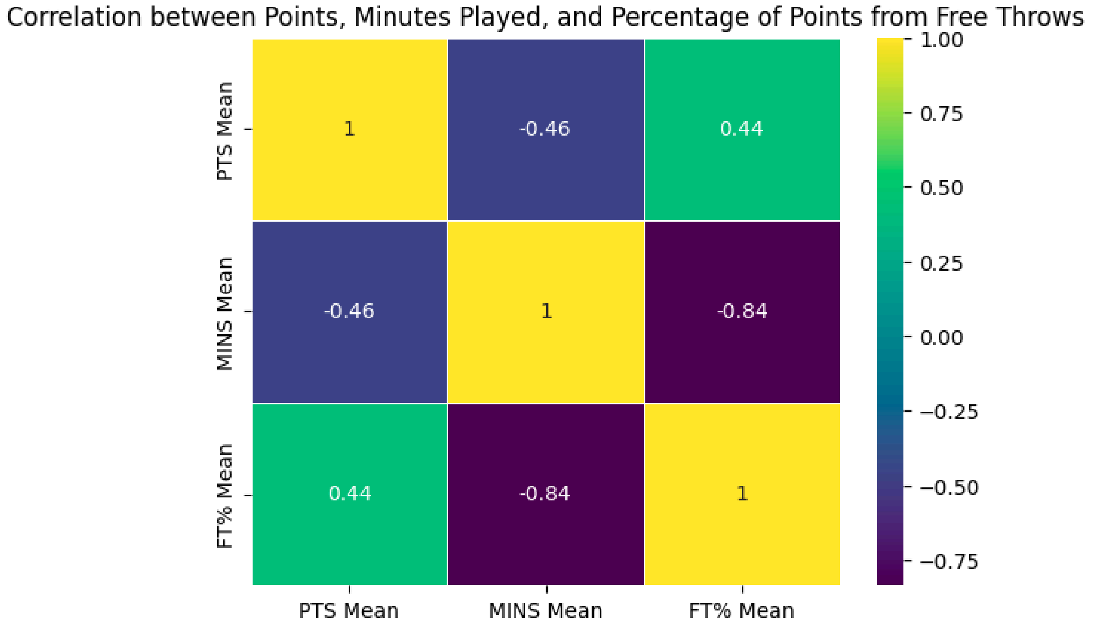


Player	PTS Mean	MINS Mean	FT% Mean
Andre Roberson	233.600000	1135.400000	10.976767
Chuck Gilmur	234.500000	0.000000	36.283892
Chucky Brown	234.047619	694.523810	18.639822
DeQuan Jones	233.000000	803.000000	14.592275
Ime Udoka	233.571429	815.857143	16.370159
Leon Powe	234.571429	521.142857	32.321673
Mel Davis	234.666667	611.000000	15.475746
Mitchell Butler	233.500000	683.000000	18.905244
Shannon Brown	233.133333	553.133333	17.785211
Shawne Williams	234.444444	717.222222	11.345650

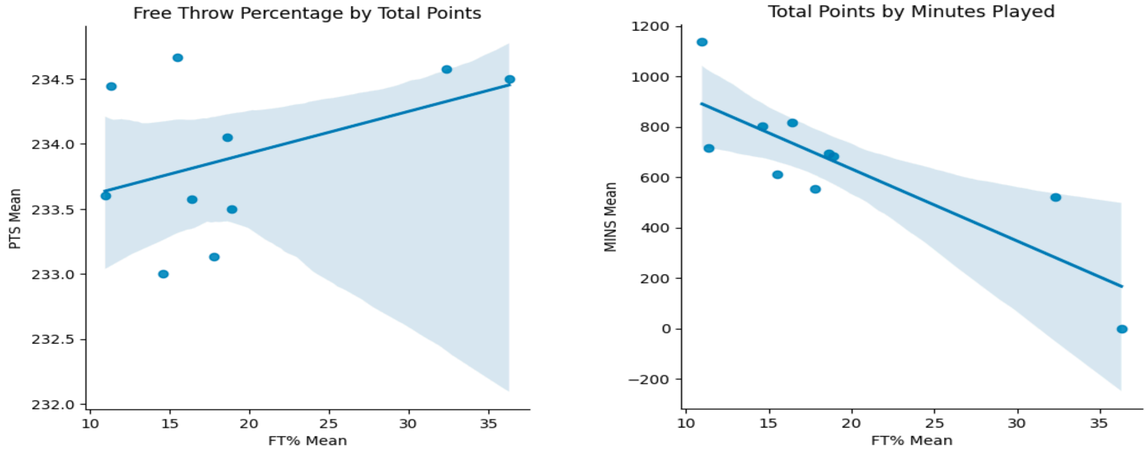
Heatmap showing correlation between free throw percentage, total points, and minutes played:



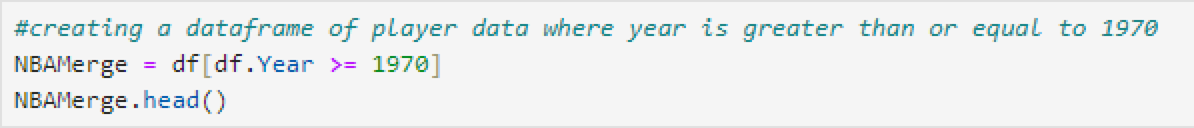
```
#calculate correlation between the percentage columns we created
MedianPlayersCorr=PTSMINFT.corr()
#create a heatmap of the correlation
plt.figure()
sns.heatmap(MedianPlayersCorr, annot=True, linewidth=0.5, cmap="viridis")
plt.title("Correlation between Points, Minutes Played, and Percentage of Points from Free Throws")
plt.show()
```


Correlation between Points, Minutes Played, and Percentage of Points from Free Throws

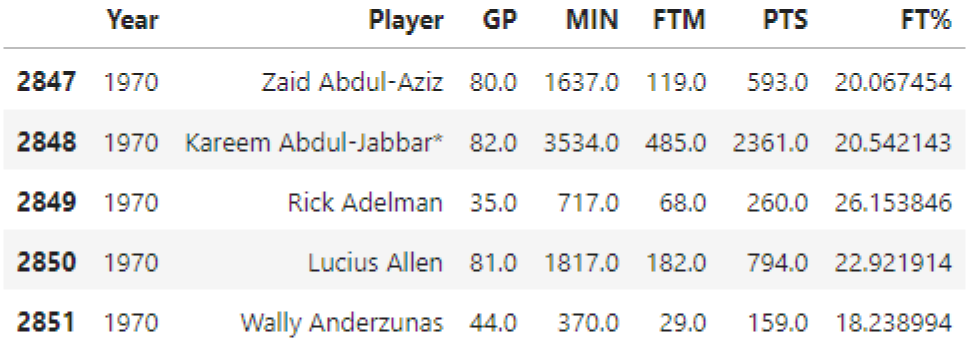
Linear regression lines of the correlation between free throws, total points, and minutes played:



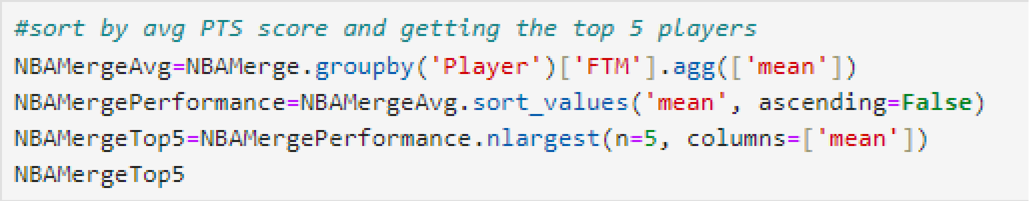
Finding the top players since the NBA merge:



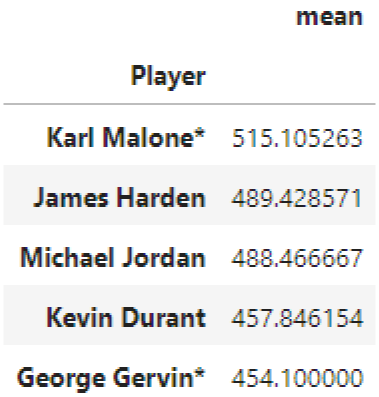
```
#creating a dataframe of player data where year is greater than or equal to 1970
NBAMerge = df[df.Year >= 1970]
NBAMerge.head()
```



	Year	Player	GP	MIN	FTM	PTS	FT%
2847	1970	Zaid Abdul-Aziz	80.0	1637.0	119.0	593.0	20.067454
2848	1970	Kareem Abdul-Jabbar*	82.0	3534.0	485.0	2361.0	20.542143
2849	1970	Rick Adelman	35.0	717.0	68.0	260.0	26.153846
2850	1970	Lucius Allen	81.0	1817.0	182.0	794.0	22.921914
2851	1970	Wally Anderzunas	44.0	370.0	29.0	159.0	18.238994

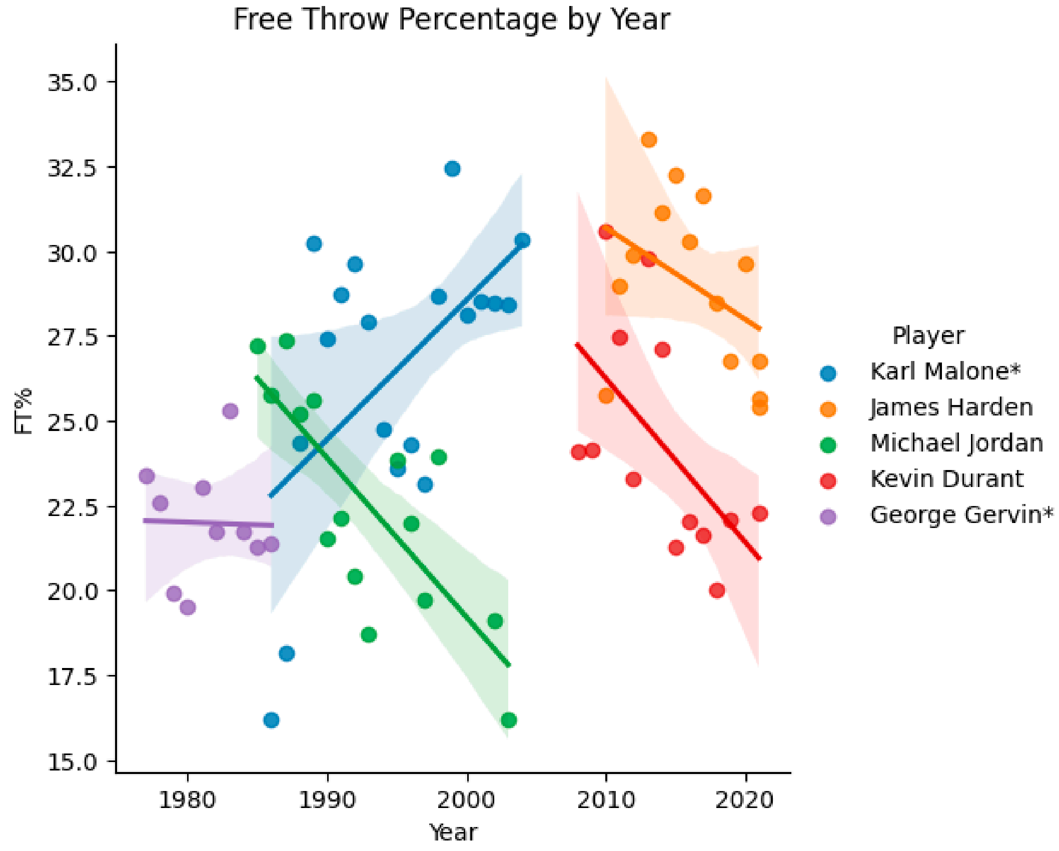


```
#sort by avg PTS score and getting the top 5 players
NBAMergeAvg=NBAMerge.groupby('Player')['FTM'].agg(['mean'])
NBAMergePerformance=NBAMergeAvg.sort_values('mean', ascending=False)
NBAMergeTop5=NBAMergePerformance.nlargest(n=5, columns=['mean'])
NBAMergeTop5
```

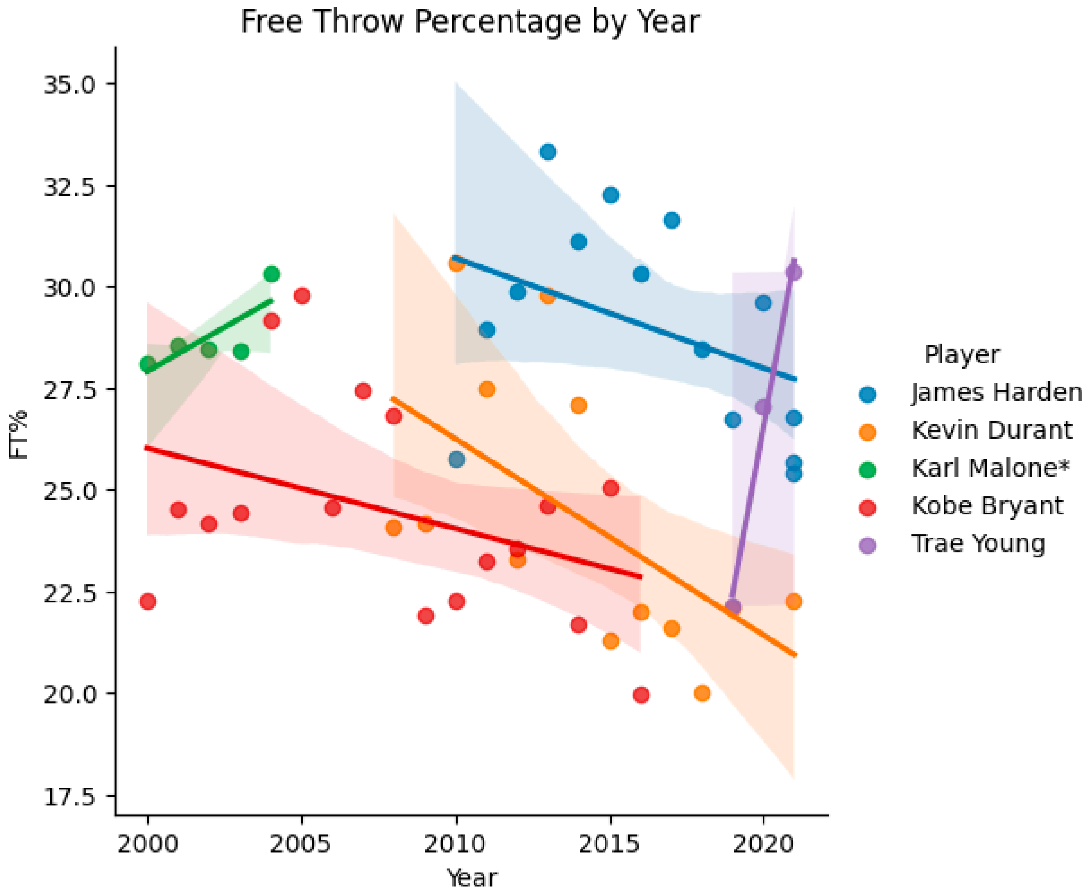


	mean
Player	
Karl Malone*	515.105263
James Harden	489.428571
Michael Jordan	488.466667
Kevin Durant	457.846154
George Gervin*	454.100000

Using the same method as in the median players, I created a dataframe of the above 5 players from the data since the NBA merge. Then I used that to create a linear regression plot:

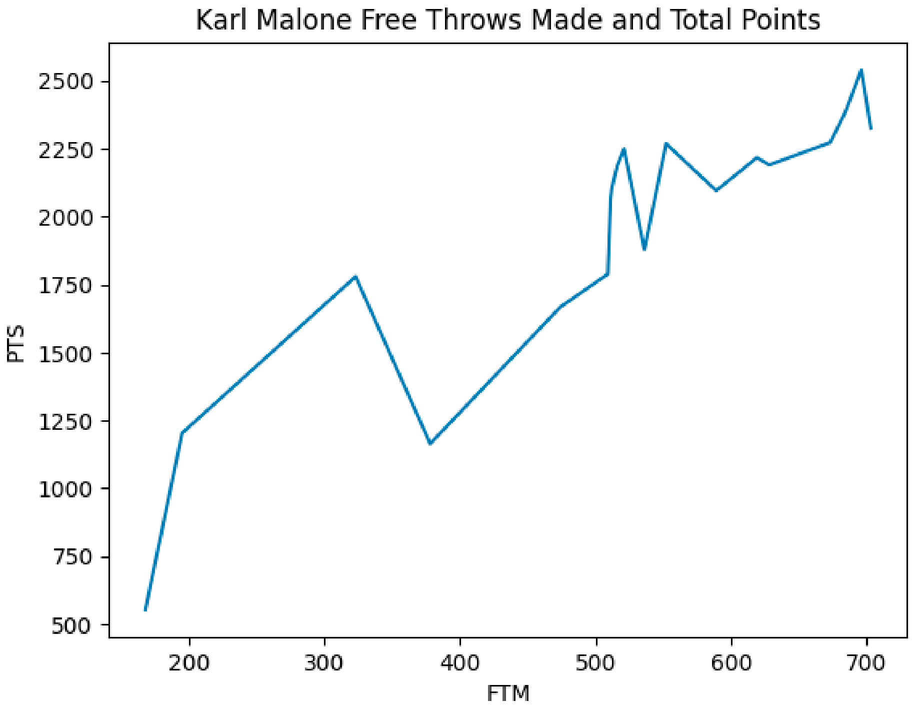


Using the same method, I created a linear regression plot showing the top 5 players from the year 2000 on:

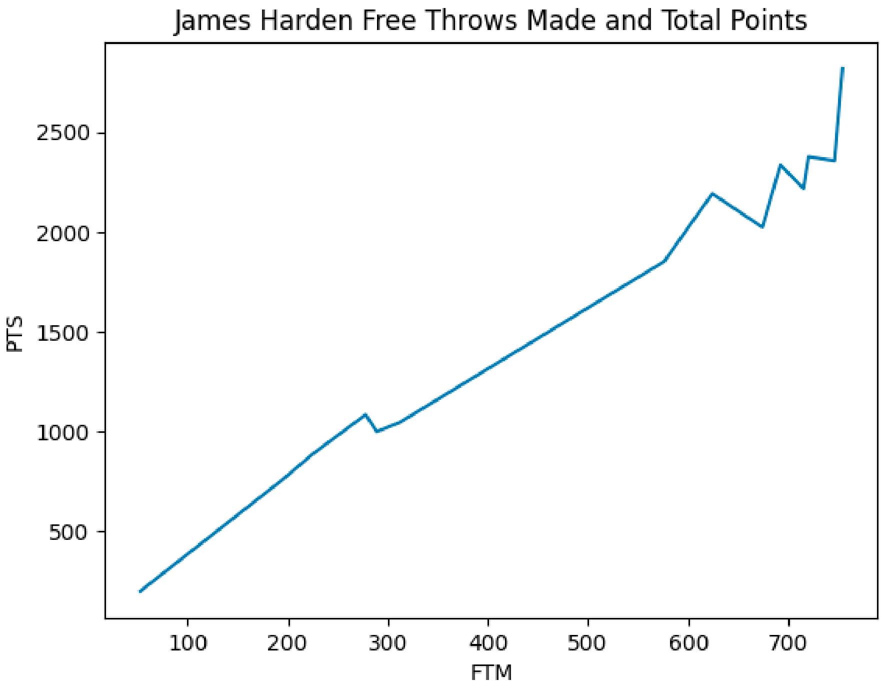


Conclusion

Since Karl Malone is the top player since the NBA merge, I created a line plot showing his free throws compared to total points:



And James Harden is the top player from the year 2000-2021:



Using this data, it is possible to draw the conclusion that more free throws correlates to a higher points score, and that free throw percentage could be an accurate predictor for points score.

Although not the hypothesis of this analysis, a potential further analysis is needed to investigate a possible inverse link between minutes played and points score, indicating that higher average minutes played per year could have a negative effect on points score and number of free throws.