nba_player_similarity

Lucas Bishop

June 17, 2021

NBA Data Science

Load Libraries

Load all libraries that are needed for the analysis

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.1     v dplyr   1.0.6
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## Warning: package 'tibble' was built under R version 4.0.5

## Warning: package 'tidyr' was built under R version 4.0.5

## Warning: package 'dplyr' was built under R version 4.0.5

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(rvest)
```

```
## Warning: package 'rvest' was built under R version 4.0.4

##
## Attaching package: 'rvest'

## The following object is masked from 'package:readr':
##
##     guess_encoding
```

```
library(BBmisc)
```

```
## Warning: package 'BBmisc' was built under R version 4.0.5
```

```
##
## Attaching package: 'BBmisc'
```

```
## The following objects are masked from 'package:dplyr':
##
##     coalesce, collapse
```

```
## The following object is masked from 'package:base':
##
##     isFALSE
```

```
library(ggrepel)
```

```
## Warning: package 'ggrepel' was built under R version 4.0.5
```

```
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.0.5
```

```
## corrplot 0.89 loaded
```

```
library(factoextra)
```

```
## Warning: package 'factoextra' was built under R version 4.0.5
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
library(kableExtra)
```

```
## Warning: package 'kableExtra' was built under R version 4.0.5
```

```
##
## Attaching package: 'kableExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

Set up Data and Scrape (some) of the data from Internet

```
setwd("C:/Users/bisho/Documents/NBA player siMLarity/")
traditional <- read_csv("data/36minutes.csv") %>% select(-badcol)
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_double(),
##   Player = col_character(),
##   badcol = col_character(),
##   Pos = col_character(),
##   Tm = col_character()
## )
## i Use `spec()` for the full column specifications.
```

```
advanced <- read_csv("data/advancedNBA.csv") %>% select(-badcol)
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_double(),
##   Player = col_character(),
##   badcol = col_character(),
##   Pos = col_character(),
##   Tm = col_character()
## )
## i Use `spec()` for the full column specifications.
```

```
# extra pages for misc df have only qualified players, qualifications found here: https://basketball.re
extra1 <- read_html("https://basketball.realgm.com/nba/stats/2019/Misc_Stats/Qualified/dbl_dbl/All/desc/
  html_table("table#table-7945.tablesaw.compact.tablesaw-stack", header = NA, fill = TRUE)
extra2 <- read_html("https://basketball.realgm.com/nba/stats/2019/Misc_Stats/Qualified/dbl_dbl/All/desc/
  html_table("table#table-4389.tablesaw.compact.tablesaw-stack", header = NA, fill = TRUE)
extra3 <- read_html("https://basketball.realgm.com/nba/stats/2019/Misc_Stats/Qualified/dbl_dbl/All/desc/
  html_table("table#table-9644.tablesaw.compact.tablesaw-stack", header = NA, fill = TRUE)
```

Further prepare the data:

```
# need to get rid of positions with dashes in them too
# get rid of the win share stats that are redundant in misc dataset
misc <- rbind(extra1[[3]], extra2[[3]], extra3[[3]]) %>% select(-'#', -"OWS", -"DWS", -"WS")
merged_stats <- inner_join(traditional, advanced) %>% select(-Rk)
```

```
## Joining, by = c("Rk", "Player", "Pos", "Age", "Tm", "G", "MP")
```

```
## cleanup before misc merge

misc$Player=gsub(",", "", misc$Player, fixed = TRUE)

full_stats <- full_join(merged_stats, misc, by = 'Player') %>%
  unique() %>% select(-Team) %>% mutate(MPG = MP / G)
```

Originally I was going to clean up the foreign names within R but got fed up and cleaned in exel and then saved...

BUT! Now we have a full data frame with semi-cleaned data that we can add more columns to if we find more interesting stats.

Now we want to filter this merged data for only entries that we are interested in:

```
full_stats_qual <- full_stats %>% drop_na()
#This leaves us with a data frame of players that have data entries in all 65 variables, or 'relevant' p
# now to subset this data for only looking at players that play significant minutes. I chose 25
cutoff <- 25.5
full_stats_qual <- subset(full_stats_qual, MPG >= cutoff)
```

Lets see what that looks like:

```
full_stats_qual %>% kable()
```

Player	Pos	Age	Tm	G	GS	MP	FG	FGA	FG%	3P	3PA	3P%	2P	2
Steven Adams	C	25	OKC	80	80	2669	6.5	10.9	0.595	0.0	0.0	0.000	6.5	1
LaMarcus Aldridge	C	33	SAS	81	81	2687	9.2	17.7	0.519	0.1	0.6	0.238	9.0	1
Jarrett Allen	C	20	BRK	80	80	2096	5.8	9.8	0.590	0.1	0.8	0.133	5.7	
Al-Farouq Aminu	PF	28	POR	81	81	2292	4.0	9.3	0.433	1.5	4.4	0.343	2.5	
Kyle Anderson	SF	25	MEM	43	40	1281	4.2	7.8	0.543	0.3	1.0	0.265	4.0	
Giannis Antetokounmpo	PF	24	MIL	72	72	2358	11.0	19.0	0.578	0.8	3.1	0.256	10.2	1
Trevor Ariza	SF	33	TOT	69	69	2349	4.5	11.3	0.399	2.2	6.7	0.334	2.3	
D.J. Augustin	PG	31	ORL	81	81	2269	5.0	10.7	0.470	2.1	4.9	0.421	3.0	
Deandre Ayton	C	20	PHO	71	70	2183	8.4	14.3	0.585	0.0	0.1	0.000	8.4	1
Lonzo Ball	PG	21	LAL	47	45	1423	4.7	11.5	0.406	1.9	5.8	0.329	2.8	
Harrison Barnes	PF-SF	26	TOT	77	77	2533	6.1	14.6	0.420	2.5	6.3	0.395	3.7	
Will Barton	SF	28	DEN	43	38	1189	5.6	13.9	0.402	2.0	5.9	0.342	3.6	
Bradley Beal	SG	25	WAS	82	82	3028	9.1	19.1	0.475	2.5	7.1	0.351	6.6	1
Patrick Beverley	PG	30	LAC	78	49	2137	3.3	8.0	0.407	1.9	4.8	0.397	1.4	
Eric Bledsoe	PG	29	MIL	78	78	2272	7.4	15.4	0.484	2.0	6.0	0.329	5.5	
Devin Booker	SG	22	PHO	64	64	2242	9.4	20.2	0.467	2.2	6.6	0.326	7.2	1
Jaylen Brown	SG	22	BOS	74	25	1913	6.9	14.9	0.465	1.8	5.2	0.344	5.1	
Jimmy Butler	SF-SG	29	TOT	65	65	2185	6.9	14.9	0.462	1.1	3.2	0.347	5.8	1
Willie Cauley-Stein	C	25	SAC	81	81	2213	6.7	12.1	0.556	0.0	0.0	0.500	6.7	1
Jordan Clarkson	SG	26	CLE	81	0	2214	8.6	19.2	0.448	2.3	7.2	0.324	6.3	1
John Collins	PF	21	ATL	61	59	1829	9.2	16.4	0.560	1.1	3.1	0.348	8.1	1
Darren Collison	PG	31	IND	76	76	2143	5.2	11.1	0.467	1.3	3.3	0.407	3.8	
Mike Conley	PG	31	MEM	70	70	2342	7.5	17.2	0.438	2.4	6.5	0.364	5.1	1
DeMarcus Cousins	C	28	GSW	30	30	771	8.3	17.3	0.480	1.2	4.4	0.274	7.1	1
Robert Covington	SF	28	TOT	35	35	1203	4.7	10.8	0.431	2.5	6.7	0.378	2.1	
Allen Crabbe	SG	26	BRK	43	20	1133	4.4	11.9	0.367	3.1	8.2	0.378	1.2	
Jae Crowder	SF	28	UTA	80	11	2166	5.3	13.2	0.399	2.9	8.7	0.331	2.4	
Stephen Curry	PG	30	GSW	69	69	2331	9.8	20.7	0.472	5.5	12.5	0.437	4.3	
Anthony Davis	C	25	NOP	56	56	1850	10.3	20.0	0.517	0.9	2.8	0.331	9.4	1
DeMar DeRozan	SG	29	SAS	77	77	2688	8.5	17.6	0.481	0.1	0.6	0.156	8.4	1
Spencer Dinwiddie	PG	25	BRK	68	4	1914	6.9	15.6	0.442	2.3	7.0	0.335	4.6	
Luka Doncic	SG	19	DAL	72	72	2318	7.9	18.4	0.427	2.6	8.0	0.327	5.2	1
Damyean Dotson	SG	24	NYK	73	40	2004	5.2	12.5	0.415	2.3	6.1	0.368	2.9	
Andre Drummond	C	25	DET	79	79	2647	7.6	14.3	0.533	0.1	0.5	0.132	7.6	1
Kris Dunn	PG	24	CHI	46	44	1389	5.6	13.1	0.425	0.9	2.5	0.354	4.7	1
Kevin Durant	SF	30	GSW	78	78	2702	9.6	18.4	0.521	1.8	5.2	0.353	7.8	1
Joel Embiid	C	24	PHI	64	64	2154	9.7	20.0	0.484	1.3	4.4	0.300	8.4	1
Bryn Forbes	SG	25	SAS	82	81	2293	5.7	12.4	0.456	2.8	6.5	0.426	2.9	
Evan Fournier	SG	26	ORL	81	81	2553	6.6	15.1	0.438	2.2	6.3	0.340	4.4	
De'Aaron Fox	PG	21	SAC	81	81	2546	7.1	15.6	0.458	1.2	3.3	0.371	5.9	1
Danilo Gallinari	SF	30	LAC	68	68	2059	7.2	15.5	0.463	2.8	6.5	0.433	4.3	
Marc Gasol	C	34	TOT	79	72	2436	5.8	12.9	0.448	1.5	4.0	0.363	4.3	
Rudy Gay	PF	32	SAS	69	51	1842	7.3	14.6	0.504	1.4	3.6	0.402	5.9	1
Paul George	SF	28	OKC	77	77	2841	9.0	20.5	0.438	3.7	9.6	0.386	5.3	1
Shai Gilgeous-Alexander	PG	20	LAC	82	73	2174	5.6	11.9	0.476	0.8	2.3	0.367	4.8	
Aaron Gordon	PF	23	ORL	78	78	2633	6.4	14.3	0.449	1.7	4.7	0.349	4.8	
Jerami Grant	PF	24	OKC	80	77	2612	5.6	11.3	0.497	1.6	4.0	0.392	4.1	
Danny Green	SG	31	TOR	80	80	2216	4.8	10.2	0.465	3.2	7.1	0.455	1.5	
Draymond Green	PF	28	GSW	66	66	2065	3.3	7.4	0.445	0.8	2.9	0.285	2.5	
Jeff Green	PF	32	WAS	77	44	2097	5.6	11.8	0.475	1.9	5.5	0.347	3.7	
Blake Griffin	PF	29	DET	75	75	2622	8.5	18.4	0.462	2.6	7.2	0.362	5.9	1
Tim Hardaway Jr.	SG	26	TOT	65	63	2057	6.8	17.4	0.393	2.8	8.3	0.340	4.0	
James Harden	PG	29	HOU	78	78	2867	10.6	24.0	0.442	4.7	12.9	0.368	5.8	1
Montrezl Harrell	C	25	LAC	82	5	2158	9.1	14.8	0.615	0.1	0.3	0.176	9.1	1
Tobias Harris	PF	26	TOT	82	82	2847	7.7	15.9	0.487	2.0	5.0	0.397	5.8	1
Josh Hart	SG	23	LAL	67	22	1715	4.0	9.7	0.407	1.9	5.8	0.336	2.0	
Gordon Hayward	PF	28	BOS	72	18	1863	5.7	12.3	0.466	1.5	4.5	0.333	4.2	