Practice5S24

BilalTariq

2024-03-07

Practice5 - Due Thursday, 3/7 by midnight to Gradescope

Reminder: Practice assignments may be completed working with other individuals.

Reading

The associated reading for the week is Chapter 19 and Section 12.1.

Practicing Academic Integrity

If you worked with others or used resources outside of provided course material (anything besides our textbook, course materials in the repo, labs, R help menu) to complete this assignment, please acknowledge them below using a bulleted list.

I acknowledge the following individuals with whom I worked on this assignment:

Name(s) and corresponding problem(s)

- N/A

I used the following sources to help complete this assignment:

Source(s) and corresponding problem(s)

-

1 - MDSR 12.6 (modified)

"Baseball players are voted into the Hall of Fame by the members of the Baseball Writers of America Association. Quantitative criteria are used by the voters, but they are also allowed wide discretion. The following code identifies the position players (not pitchers) who have been elected to the Hall of Fame and tabulates a few basic statistics, include their number of career hits (`tH`), home runs (`tHR`), runs batted in (`tRBI`), and stolen bases (`tSB`)." Only players with more than 1000 total hits are included as a way to obtain the position players only (not pitchers).

```
hof <- Batting %>%
  group_by(playerID) %>%
  inner_join(HallOfFame, by = "playerID") %>%
  filter(inducted == "Y" & votedBy == "BBWAA") %>%
  summarize(tH = sum(H), tHR = sum(HR), tRBI = sum(RBI), tSB = sum(SB)) %>%
  filter(tH > 1000)
```

```
Warning in inner_join(., HallOfFame, by = "playerID"): Detected an unexpected many-to-many re
i Row 5 of `x` matches multiple rows in `y`.
i Row 72 of `y` matches multiple rows in `x`.
i If a many-to-many relationship is expected, set `relationship =
  "many-to-many"` to silence this warning.
```

```
kable(hof)
```

playerID	tH	tHR	tRBI	tSB
aaronha01	3771	755	2297	240
alomaro01	2724	210	1134	474
aparilu01	2677	83	791	506
bagweje01	2314	449	1529	202
bankser01	2583	512	1636	50
benchjo01	2048	389	1376	68
berrayo01	2150	358	1430	30
biggicr01	3060	291	1175	414
boggswa01	3010	118	1014	24
boudrlo01	1779	68	789	51
brettge01	3154	317	1596	201
brocklo01	3023	149	900	938
camparo01	1161	242	856	25
carewro01	3053	92	1015	353

playerID	tH	tHR	tRBI	tSB
cartega01	2092	324	1225	39
cobbty01	4189	117	1944	896
cochrmi01	1652	119	832	64
collied01	3315	47	1300	741
cronijo01	2285	170	1424	87
dawsoan01	2774	438	1591	314
dickebi01	1969	202	1209	37
dimagjo01	2214	361	1537	30
fiskca01	2356	376	1330	128
foxxji01	2646	534	1922	87
friscfr01	2880	105	1244	419
greenha01	1628	331	1276	58
griffke02	2781	630	1836	184
guerrvl01	2590	449	1496	181
gwynnto01	3141	135	1138	319
hartnga01	1912	236	1179	28
heilmha01	2660	183	1539	113
henderi01	3055	297	1115	1406
hornsro01	2930	301	1584	135
jacksre01	2584	563	1702	228
jeterde01	3465	260	1311	358
jonesch06	2726	468	1623	150
kalinal01	3007	399	1583	137
keelewi01	2932	33	810	495
killeha01	2086	573	1584	19
kinerra01	1451	369	1015	22
lajoina01	3243	82	1599	380
larkiba01	2340	198	960	379
mantlmi01	2415	536	1509	153
maranra01	2605	28	884	291
martied01	2247	309	1261	49
matheed01	2315	512	1453	68
mayswi01	3283	660	1903	338
mccovwi01	2211	521	1555	26
medwijo01	2471	205	1383	42
molitpa01	3319	234	1307	504
morgajo02	2517	268	1133	689
murraed02	3255	504	1917	110
musiast01	3630	475	1951	78
ortizda01	2472	541	1768	17
ottme01	2876	511	1860	89

playerID	tH	tHR	tRBI	tSB
perezto01	2732	379	1652	49
piazzmi01	2127	427	1335	17
puckeki01	2304	207	1085	134
raineti01	2605	170	980	808
riceji01	2452	382	1451	58
ripkeca01	3184	431	1695	36
robinbr01	2848	268	1357	28
robinfr02	2943	586	1812	204
robinja02	1518	137	734	197
rodriiv01	2844	311	1332	127
ruthba01	2873	714	2217	123
sandbry01	2386	282	1061	344
schmimi01	2234	548	1595	174
simmoal01	2927	307	1827	88
sislege01	2812	102	1175	375
smithoz01	2460	28	793	580
snidedu01	2116	407	1333	99
speaktr01	3514	117	1529	432
stargwi01	2232	475	1540	17
terrybi01	2193	154	1078	56
thomafr04	2468	521	1704	32
thomeji01	2328	612	1699	19
traynpi01	2416	58	1273	158
wagneho01	3420	101	1733	723
walkela01	2160	383	1311	230
wanerpa01	3152	113	1309	104
willibi01	2711	426	1475	90
willite01	2654	521	1839	24
winfida01	3110	465	1833	223
yastrca01	3419	452	1844	168
yountro01	3142	251	1406	271

- Use the `kmeans()` function to perform a cluster analysis on these players.
- Explain your choice of k, the number of clusters.
- Describe the properties that seem common to each cluster in your solution.
- Include at least one visual that helps explore the clusters found.
- Your solution should include some discussion of whether or not you chose to scale the variables and why. (You should determine whether or not you need to scale before clustering.)
- Remember that your solution must be reproducible. (Hint: this means you need to do something in your code.)

Solution:

```r
set.seed(100)
clustering_prep_hof <- hof %>%
  select("tH","tHR") %>%
  drop_na()

glimpse(clustering_prep_hof)
```

```
Rows: 86
Columns: 2
$ tH  <int> 3771, 2724, 2677, 2314, 2583, 2048, 2150, 3060, 3010, 1779, 3154, ~
$ tHR <int> 755, 210, 83, 449, 512, 389, 358, 291, 118, 68, 317, 149, 242, 92,~
```

```r
set.seed(231)

clustering_hof <- clustering_prep_hof %>%
  select("tH", "tHR") %>%
  kmeans(centers = 2, nstart = 20)

clustering_hof$cluster
```

```
 [1] 2 2 2 1 1 1 1 2 2 1 2 2 1 2 1 2 1 2 1 2 1 1 1 2 2 1 2 1 2 1 2 2 2 1 2 2 2 2
[39] 1 1 2 1 1 1 1 1 2 1 1 2 1 2 2 1 2 2 1 1 1 1 2 2 2 1 2 2 1 1 2 2 1 1 2 1 1 1
[77] 1 1 2 1 2 2 2 2 2 2
```

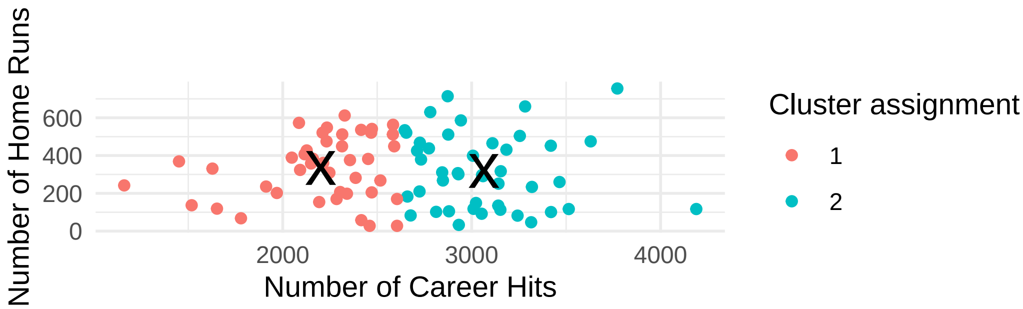
```r
clustering_hof$centers
```

```
        tH      tHR
1 2201.095 333.3333
2 3065.091 317.5455
```

```r
hof <- hof %>%
  mutate(clusters2 = factor(clustering_hof$cluster))

ggplot(data = hof, aes(x = tH, y = tHR)) +
geom_point(aes(color = clusters2)) +
coord_fixed() +
```

```
geom_point(data = data.frame(clustering_hof$centers),
aes(x = tH, y = tHR),
pch = "x", size = 8) +
labs(x = "Number of Career Hits",
y = "Number of Home Runs",
color = "Cluster assignment") + theme_minimal()
```



2 - Trump Tweets

David Robinson, Chief Data Scientist at DataCamp, wrote a blog post "Text analysis of Trump's tweets confirms he writes only the (angrier) Android half". He provides a dataset with over 1,500 tweets from the account realDonaldTrump between 12/14/2015 and 8/8/2016. We'll use this dataset to explore the tweeting behavior of @realDonaldTrump during this time period.

First, read in the file. Note that there is a `TwitteR` package which provides an interface to the Twitter web API. We'll use this R dataset David Robinson created using that package so that you don't have to set up Twitter authentication.

```
# the .rda file is also provided if this website ever breaks
load(url("http://varianceexplained.org/files/trump_tweets_df.rda"))
```

> part a - Wrangling! There are a number of variables in the dataset we won't need. First, confirm that all the observations in the dataset are from the screen-name *realDonaldTrump*. Then, create a new dataset called `tweets` that only includes the variables `text`, `created` and `statusSource`.

Solution:

> part b - Using the `statusSource` variable, compute the number of tweets from each source. How many different sources are there? How often are each used?

Hint: You could answer the questions with a nice table printed to the screen.

Solution:

> part c - We're going to compare the language used between the Android and iPhone sources, so we only want to keep tweets coming from those sources. Explain what the `extract()` function (from the **tidyverse** package) is doing below. Include in your own words what each argument is doing.

```
tweets <- tweets %>%
  extract(col = statusSource, into = "source",
          regex = "Twitter for (.*)<",
          remove = FALSE) %>%
  filter(source %in% c("Android", "iPhone"))
```

Solution:

part d - How does the language of the tweets differ by source? Create a word cloud for the top 50 words used in tweets sent from the Android. Create a second word cloud for the top 50 words used in tweets sent from the iPhone. How do these word clouds compare? (Are there some common words frequently used from both sources? Are the most common words different between the sources?)

Note: Don't forget to remove stop words before creating the word cloud. Also remove the terms "https" and "t.co".

Solution:

part e - Consider the sentiment. Compute the proportion of words among the tweets within each source classified as "angry" and the proportion of words classified as "joy" based on the NRC lexicon. How does the proportion of "angry" and "joy" words compare between the two sources? What about "positive" and "negative" words?

Solution:

part f - Lastly, based on your responses above, do you think there is evidence to support Robinson's claim that Trump only writes the Android half of the tweets from realDonaldTrump? In 2-4 sentences, please explain.

Solution: