به نام خدا

آمار و احتمال مهندسی
دانشگاه صنعتی شریف - دانشکده مهندسی کامپیوتر

بهار 1401

تمرین عملی اول

طراح: سجاد سلطانیان
موعد تحویل: 3 اردیبهشت

همفکری در تمامی تمرین‌های درس توصیه می‌شود. در عین حال از شما خواسته می‌شود تا تمام پیاده‌سازی را به تنهایی و بدون مشاهده کد دیگران انجام دهید.

لطفا در فایل ارسالی تمام بلوک‌های کد اجرا شده و شامل نمودارها و خروجی‌های لازم باشند.

مقدمه

هدف از این تمرین، آشنایی با مقدمات زبان برنامه‌نویسی R در قالب کاربردی است. در هر بخش با بیان دقیق خواسته مسئله سعی شده است تا از هرگونه ابهام پیشگیری شود. با این حال، چنانچه در فهم خواسته سوال ابهامی وجود داشت در بستر کوئرا بیان کنید تا پاسخ داده شود.
توجه: ممکن است برای رسیدن به پاسخ، راه‌های متعددی وجود داشته باشد. پاسخ شما تا زمانی که خواسته مسئله را برآورده کند کاملا قابل قبول خواهد بود.

سوال اول

در این سوال قصد داریم عملیات‌های آماری مختلفی را روی یک مجموعه واقعی از داده‌ها که از طرف سایت goodreads منتشر شده است انجام دهیم.

خواندن داده از فایل (۱۰ نمره)

فایل books.csv را از پوشه resources بخوانید و آن را در متغیری به نام all-books ذخیره کنید. ۵ نمره

جهت اطمینان از بارگذاری صحیح داده‌ها، ۵ سطر ابتدایی آن را پرینت کنید. ۵ نمره

راهنمایی: از دوتابع head و tail به ترتیب جهت دریافت چند سطر ابتدایی و انتهایی جدول‌ها استفاده می‌شود.

In [1]:
```
# AmirReza Azari
# 99101087
#*********************************
all_books <- read.csv("books.csv")          # uploaded in jupyter notebook
print(head(all_books, 5))
# all_books <- read.csv('C:/Users/AmirReza81/Downloads/Compressed/practical#1/files/books.csv')
#                                                                          --> reading from pc
```

```
                                    title                series
1                         The Hunger Games   The Hunger Games #1
2 Harry Potter and the Order of the Phoenix       Harry Potter #5
3                      To Kill a Mockingbird To Kill a Mockingbird
4                        Pride and Prejudice
5                                  Twilight  The Twilight Saga #1
                                     author rating language
1                            Suzanne Collins   4.33  English
2 J.K. Rowling, Mary GrandPré (Illustrator)   4.50  English
3                                 Harper Lee   4.28  English
4 Jane Austen, Anna Quindlen (Introduction)   4.26  English
5                            Stephenie Meyer   3.60  English

genres
1                  ['Young Adult', 'Fiction', 'Dystopia', 'Fantasy', 'Science Fiction', 'Romance', 'Adventur
e', 'Teen', 'Post Apocalyptic', 'Action']
2        ['Fantasy', 'Young Adult', 'Fiction', 'Magic', 'Childrens', 'Adventure', 'Audiobook', 'Middle Grad
e', 'Classics', 'Science Fiction Fantasy']
3      ['Classics', 'Fiction', 'Historical Fiction', 'School', 'Literature', 'Young Adult', 'Historical', 'N
ovels', 'Read For School', 'High School']
4 ['Classics', 'Fiction', 'Romance', 'Historical Fiction', 'Literature', 'Historical', 'Novels', 'Historical
Romance', 'Classic Literature', 'Adult']
5           ['Young Adult', 'Fantasy', 'Romance', 'Vampires', 'Fiction', 'Paranormal', 'Paranormal Romance',
'Supernatural', 'Teen', 'Urban Fantasy']
  pages                        publisher publish.year numRatings likedPercent
1   374                 Scholastic Press         2008    6376780           96
2   870                  Scholastic Inc.         2004    2507623           98
3   324 Harper Perennial Modern Classics         2006    4501075           95
4   279                   Modern Library         2000    2998241           94
5   501        Little, Brown and Company         2006    4964519           78
  price
1  5.09
2  7.38
3
```

```
4
5   2.1
```

محبوب‌ترین کتاب‌ها از نگاه کاربران

قسمت اول: تعریف شاخص محبوبیت (۳۵ نمره)

می‌خواهیم ۱۰ کتاب برتر را در میان تمامی کتاب‌های موجود پیدا کنیم. معیار ما در اندازه‌گیری میزان محبوبیت یک کتاب، شاخص popularity خواهد بود که خود از دو شاخص rating (نظرسنجی کاربران از ۰ تا ۵)، و از likedPercent (درصد محبوبیت کتاب) بدست خواهد آمد. برای این منظور لازم است گام‌های زیر را طی کنیم:

- ابتدا کتاب‌هایی را که دارای تعداد نظرات (numRating) کمتر از ۱۰ هزار هستند را از مجموعه دادگان کنار می‌گذاریم و داده‌های باقی‌مانده را در متغیری به نام filtered_books ذخیره می‌کنیم. (در مراحل بعدی نیز از همین متغیر استفاده خواهیم کرد) ۱۰ نمره

In [2]:
```
filtered_books <- all_books[all_books$numRatings >= 10000, ]
```

- حال، پیش از محاسبه شاخص محبوبیت لازم است تا دو ستون rating و likedPercent را هم‌وزن کنیم. به این معنا که مقیاس ستون rating را از ۰-۵ به ۰ الی ۱۰۰ (مشابه likedPercent) تغییر دهیم که این کار با ۲۰ برابر کردن ستون rating قابل انجام است ۱۰ نمره

راهنمایی: برای این کار می‌توانید از تابع sapply استفاده کنید

In [3]:
```
filtered_books$rating <- filtered_books$rating * 20
# filtered_books$rating <- sapply(filtered_books$rating, function(x) 20 * x)
```

- شاخص popularity را با میانگین‌گیری وزن‌دار از دو ستون rating و likedPercent، به ترتیب با ضریب ۵ و ۲ محاسبه می‌کنیم و نتیجه را در برداری با همین نام ذخیره می‌کنیم. ۱۰ نمره

In [4]:
```
popularity <- (filtered_books$rating * 5 + filtered_books$likedPercent * 2) / 7
```

- نهایتا نیز ستون popularity را با همین نام به دیتافریم خود (filtered-books) اضافه می‌کنیم. ۵ نمره

توجه: میتوانید از کتاب‌خانه‌ها نیز استفاده کنید اما تا این مرحله، تمام کارها با استفاده از همان توابع پیشفرض موجود در زبان نیز قابل انجام بوده است.

In [5]:
```
filtered_books$popularity = popularity
```

قسمت دوم: مرتب‌سازی و یافتن بهترین‌ها (۲۵ نمره)

اکنون معیار مناسب برای قضاوت میان کتاب‌ها را داریم. حال کافیست داده‌ها را مرتب کنیم و محبوب‌ترین‌ها را استخراج کنیم. برای این منظور گام‌های زیر را طی می‌کنیم:

- کتاب‌ها را ابتدا بر اساس ستون popularity و سپس بر اساس سال انتشار و در نهایت نام نویسنده آن، به صورت نزولی مرتب کنید. ۱۰ نمره

In [6]:
```
filtered_books <- filtered_books[order(
    filtered_books$popularity,
    filtered_books$publish.year,
    filtered_books$author, decreasing = TRUE), ]
```

- تابعی بنویسید که یک دیتافریم را دریافت کند و n سطر ابتدایی آن را بازگرداند. سپس با کمک این تابع، ۱۰ کتاب برتر را در متغیر top_10 ذخیره کنید. ۱۰ نمره

In [7]:
```
return_nrows = function(df, n) head(df, n)
top_10 = return_nrows(filtered_books, 10)
```

- در نهایت، عنوان، نویسنده، سال انتشار و ناشر ۱۰ کتاب برتر را پرینت کنید. ۵ نمره

In [8]:
```
print(top_10[, c('title', 'author', 'publish.year', 'publisher')])
```

```
                                                                title
254                                       The Complete Calvin and Hobbes
5191                                                 It's a Magical World
2565                                              Mark of the Lion Trilogy
2619                                                      ESV Study Bible
6505            Harry Potter Boxed Set, Books 1-5 (Harry Potter, #1-5)
1504 The Authoritative Calvin and Hobbes: A Calvin and Hobbes Treasury
3381                                    The Indispensable Calvin and Hobbes
626                                                      Words of Radiance
9435                                    Scientific Progress Goes "Boink"
6513           Attack of the Deranged Mutant Killer Monster Snow Goons
                                                          author publish.year
254                                                Bill Watterson         2005
5191                                               Bill Watterson         1996
2565                          Francine Rivers (Goodreads Author)         1998
2619 Anonymous, Lane T. Dennis (Editor), Wayne Grudem (Editor)         2008
6505                  J.K. Rowling, Mary GrandPré (Illustrator)         2004
1504                                               Bill Watterson         1990
3381                                               Bill Watterson         1905
626                         Brandon Sanderson (Goodreads Author)         2014
9435                                               Bill Watterson         1905
6513                                               Bill Watterson         1992
                         publisher
254      Andrews McMeel Publishing
5191     Andrews McMeel Publishing
2565                 Tyndale House
2619                      Crossway
6505                   Scholastic
1504                       Sphere
3381                  Warner Books
626                      Tor Books
9435 Andrews and McMeel Publishing
6513               Turtleback Books
```

وقتی قلم اوج می‌گیرد! (۳۰ نمره)

در آخرین بخش این تمرین، میخواهیم به دو تا از مهم‌ترین عملیات‌ها در زمینه مهندسی داده بپردازیم. در دو بخش گذشته، توانستیم محبوب‌ترین کتاب‌ها را بیابیم. حال اما تصور کنید با استفاده

از همان اطلاعات کتاب‌ها بخواهیم روند چاپ و نگارش کتاب را در طول تاریخ بررسی کنیم. چه باید کرد؟
این مسئله ما را به موضوع grouping در علم داده سوق می‌دهد. بدین معنا که داده‌ها را بر اساس یک معیار (در اینجا سال انتشار کتاب) دسته‌بندی می‌کنیم و در نتیجه، لیستی از سال‌ها را خواهیم داشت که به هر کدام چندین کتاب متصل شده است. در ادامه لازم است که عملیاتی روی کتاب‌های هر سال انجام دهیم. به طور مثال، میانگین قیمت آن‌ها را محاسبه کنیم تا در نهایت، یک تک مقدار برای کل آن سال داشته باشیم. به مجموعه این فرایند که به نوعی همان خلاصه‌سازی داده‌ها است، aggregation می‌گویند که با همین نام هم می‌توانید آن را در زبان R بیابید.

- حال می‌خواهیم تعداد کتاب‌های منتشر شده در هر سال را بدست بیاوریم. برای این منظور، ابتدا کتاب‌ها را بر اساس سال انتشار آن‌ها دسته‌بندی خواهیم کرد و عملیات length را روی عنوان کتاب‌ها اجرا می‌کنیم و نتیجه را در متغیر A ذخیره می‌کنیم. ۱۰ نمره

برای حل سوال بالا از تابع aggregate استفاده کنید. برای مطالعه بیشتر این لینک (/https://www.geeksforgeeks.org/how-to-use-aggregate-function-in-r) را مطالعه کنید.

In [9]:
```
A <- aggregate(filtered_books$title, list(filtered_books$publish.year),
               FUN = length)
```

- مجددا بر اساس سال انتشار دسته‌بندی انجام می‌دهیم اما این بار به جای شمارش تعداد کتاب‌ها، ماکسیسمم تعداد صفحات آن‌ها را محاسبه می‌کنیم و نتیجه را در متغیر B ذخیره می‌کنیم. ۵ نمره

In [10]:
```
B <- aggregate(filtered_books$pages, list(filtered_books$publish.year),
               FUN = max)
```

حال نگاهی به جدول A و B بیندازید. ستون سال انتشار در هر دو جدول یکسان است و تفاوتشان در ستون دوم است. فرض کنید که بخواهیم تعداد کتاب‌های هر سال و تعداد صفحات طولانی‌ترین کتاب هر سال را بررسی کنیم. لازم است مدام بین دو جدول A و B جابجا شویم! آیا می‌توانیم اطلاعات دو جدول را با هم ترکیب کنیم و همه این اطلاعات را فقط در یک جدول نگهداری کنیم؟ مفهوم join یا merge در پاسخ به نیاز فوق ایجاد شده‌اند. همانطور که از نامشان پیداست ما می‌توانیم چند جدول را بر اساس ستون‌های متناظر ترکیب کنیم و داده‌های هر دو جدول را در یک

جدول نگه‌داری کنیم. البته این موضوع کاربردهای بسیار دیگری نیز دارد که در این مقال نمی‌گنجد :)
بعدها در درس پایگاه داده نیز با چنین مفهومی مواجه خواهید شد.

- دو جدول A و B را ترکیب کرده و در متغیر C ذخیره کنید. طبیعتا تناظر بین این دو جدول نیز بر اساس همان سال انتشار خواهد بود. ۱۰ نمره
- با استفاده از تابعی که پیشتر برای دریافت top_10 نوشته‌اید، ۱۵ سطر ابتدایی جدول C را چاپ کنید. ۵ نمره

راهنمایی: در مورد تابع merge در زبان R مطالعه کنید

In [11]:
```
C <- merge(A, B, by = "Group.1")
print(return_nrows(C, 15))
```

```
   Group.1 x.x  x.y
1     1905 375 1952
2     1936   1   72
3     1939   1   56
4     1942   1   40
5     1945   1   96
6     1952   1  107
7     1953   1  289
8     1955   1   88
9     1957   2  189
10    1958   2  176
11    1959   1  282
12    1960   2  256
13    1961   1   64
14    1962   4  589
15    1963   1  325
```

حال با استفاده از اطلاعات بالا می‌توانید روند رونق چاپ کتاب را در طول سال‌های تاریخ بررسی کنید و متوجه نکات جالبی شوید.

سخن پایانی

تمرین ما به پایان رسید، اما داستان علم داده نه ...
ما در ابتدای مسیر هستیم و شاید آنچنان که باید، اهمیت اطلاعات را در دنیای امروز درک نکنیم.
برای آشنایی با جذابیت‌های این علم و آنچه که داده‌ها می‌توانند بگویند توصیه می‌کنیم خلاصه کتاب everybody lies را از اینجا

%B2%D9%88%D8%AF%20%D8%B3%D9%87%20:%20Everybody%20Lies)

بشنوید.
موفق و پیروز باشید :)