Explore and Mine Data

Rohit

2022-12-20

Name: Rohit Hooda

Email: hooda.r@northeastern.edu

Course: CS5200 Fall 2022

Date: 20th December 2022

```
library(RMySQL)
```

```
## Loading required package: DBI
```

```
library(dplyr)
```

```
##
## Attaching package: ' dplyr'

## The following objects are masked from ' package:stats' :
##
##     filter, lag

## The following objects are masked from ' package:base' :
##
##     intersect, setdiff, setequal, union
```

```
db_user <- "root"
db_password <- "N0rthe@stern"
db_name <- "cs5200db2"
db_host <- "127.0.0.1"
db_port <- 3306

mySQLDB <- dbConnect(MySQL(), user = db_user, password = db_password, dbname = db_name,
                  host = db_host, port = db_port)
```

2. Analytical Query I: Top five journals with the most articles published in them for the time period.

Between the years 1975 to 1977, there were a total of 23309 articles published. The Journal of pharmacy and pharmacology published the most number of articles between 1975 and 1977 with a whopping count of 819 articles. Journal Biochimica et biophysica acta was not so far in publishing articles as well with 713 articles. While these two journals being the top publishers of article, The Journal of biological chemistry, Annales de l'anesthesiologie francaise and Biochemistry journals follow them with 483, 449, and 308 articles respectively. Looking at this data, it is certain that during the period between 1975 and 1977, research and publications were at peak for journals of Pharmacy and Pharmacology.

```
year1 <- 1975
year2 <- 1977
journal_fact_query <- paste("SELECT
  j.JournalTitle,
  COUNT(DISTINCT f.ArticleId) AS NumArticles
FROM
  JournalDim j
  JOIN JournalFact f ON j.JournalId = f.JournalId
WHERE
  j.Year BETWEEN", year1," AND ", year2,"
GROUP BY
  j.JournalTitle
ORDER BY
  NumArticles DESC limit 5", sep=" ")
top_five_journals <- dbGetQuery(mySQLDB, journal_fact_query)
top_five_journals
```

```
##                                JournalTitle NumArticles
## 1 The Journal of pharmacy and pharmacology         819
## 2            Biochimica et biophysica acta         713
## 3      The Journal of biological chemistry         483
## 4   Annales de l' anesthesiologie francaise         449
## 5                               Biochemistry         308
```

2. Analytical Query II: Number of articles per journal per year broken down by quarter

Looking at the data for number of articles published by each journal every year and every quarter of the year, we can clearly say that most of the articles are published in the first quarter of the year that can also be seen from the visualization of data below.

```
journal_fact_query <- "SELECT j.JournalTitle, j.Year, j.Quarter, COUNT(DISTINCT
f.ArticleId) AS NumArtiFROM JournalFact f
JOIN JournalDim j ON f.JournalId = j.JournalId
GROUP BY j.JournalTitle, j.Year, j.Quarter
ORDER BY j.JournalTitle, j.Year, j.Quarter"
articles_data <- dbGetQuery(mySQLDB, journal_fact_query)
nrow(articles_data)
```

```
## [1] 8197
```

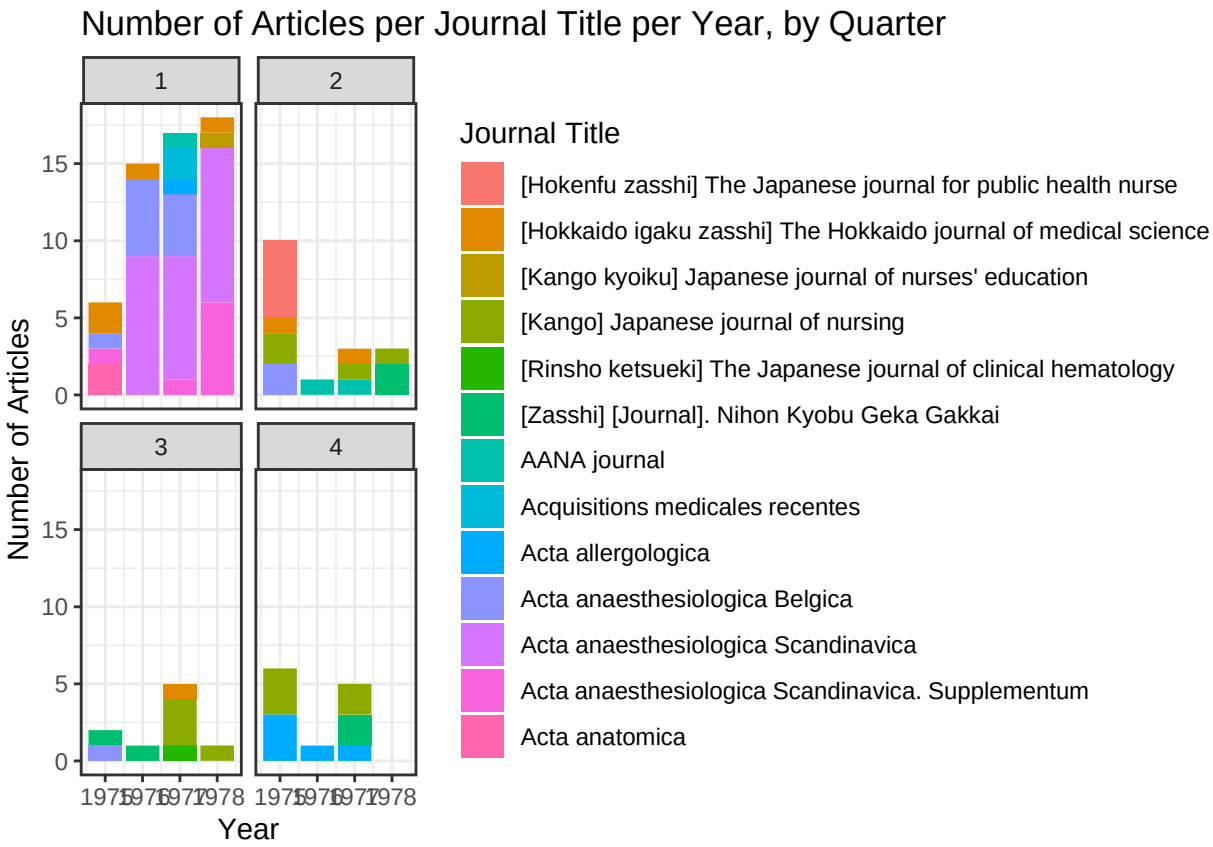
```
head(articles_data,10)
```

```
##                                                    JournalTitle  Year Quarter
## 1    [Hokenfu zasshi] The Japanese journal for public health nurse  1975       2
## 2  [Hokkaido igaku zasshi] The Hokkaido journal of medical science 1975       1
## 3  [Hokkaido igaku zasshi] The Hokkaido journal of medical science 1975       2
## 4  [Hokkaido igaku zasshi] The Hokkaido journal of medical science 1976       1
## 5  [Hokkaido igaku zasshi] The Hokkaido journal of medical science 1977       2
## 6  [Hokkaido igaku zasshi] The Hokkaido journal of medical science 1977       3
## 7  [Hokkaido igaku zasshi] The Hokkaido journal of medical science 1978       1
## 8            [Kango kyoiku] Japanese journal of nurses' education 1978       1
## 9                          [Kango] Japanese journal of nursing 1975       2
## 10                         [Kango] Japanese journal of nursing 1975       4
##    NumArticles
## 1            5
## 2            2
## 3            1
## 4            1
## 5            1
## 6            1
## 7            1
## 8            1
## 9            2
## 10           3
```

This data can be visually represented by a bar graph acoss each quarter acrross all years for each journal.

```
library(ggplot2)

# create a bar chart of the number of articles per journal title per quarter
ggplot(head(articles_data,40), aes(x = Year, y = NumArticles, fill = JournalTitle)) +
  geom_bar(stat = "identity") +
  facet_wrap(~Quarter) +
  labs(title = "Number of Articles per Journal Title per Year, by Quarter",
       x = "Year",
       y = "Number of Articles",
       fill = "Journal Title") +
  theme_bw()
```


Number of Articles per Journal Title per Year, by Quarter
1
2
3
4
Number of Articles
0
5
10
15
1975
1976
1977
1978
Year
Journal Title
[Hokenfu zasshi] The Japanese journal for public health nurse
[Hokkaido igaku zasshi] The Hokkaido journal of medical science
[Kango kyoiku] Japanese journal of nurses' education
[Kango] Japanese journal of nursing
[Rinsho ketsueki] The Japanese journal of clinical hematology
[Zasshi] [Journal]. Nihon Kyobu Geka Gakkai
AANA journal
Acquisitions medicales recentes
Acta allergologica
Acta anaesthesiologica Belgica
Acta anaesthesiologica Scandinavica
Acta anaesthesiologica Scandinavica. Supplementum
Acta anatomica