Part 3 - Explore and Mine Data

Phwe Thant Chay

2023-04-19

1. Connect to Database

```
# 1. Library
library(RMySQL)
```

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

```
library(ggplot2)

# Settings
db_user <- 'root'
db_password <- '12345678'
db_name <- 'pubmed_snowflake'
db_host <- 'localhost'
db_port <- 3306

# Read data from db
mydbcon <-  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. (year - 1977, quarter - 2)

```
SELECT journal_id, year, quarter, SUM(articles_published)
as total_articles_published
FROM journal_facts
WHERE year = 1977 AND quarter = 2
GROUP BY journal_id
ORDER BY total_articles_published DESC
LIMIT 5;
```

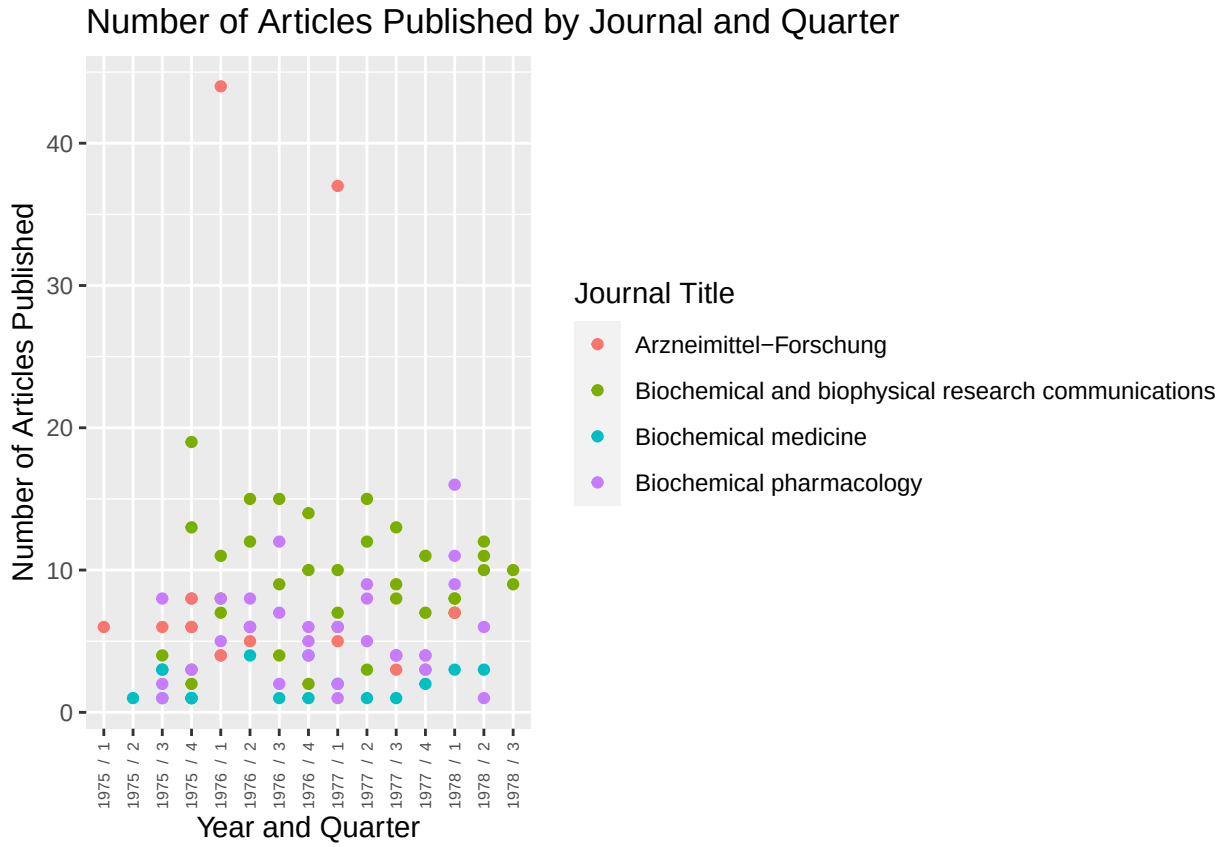
Table 1: 5 records

journal_id	year	quarter	total_articles_published
381	1977	2	88
59	1977	2	77
1463	1977	2	63
955	1977	2	39
373	1977	2	39

3.Number of articles per journal per year broken down by quarter

```
results <- dbGetQuery(mydbcon, "SELECT
                                    journal_id,
                                    title,
                                    year,
                                    quarter,
                                    articles_published
                                FROM
                                    journal_facts
                                WHERE year = 1977 OR year = 1978 OR year = 1976
                                OR year = 1975
                                ORDER BY
                                    journal_id,
                                    year,
                                    quarter
                                LIMIT 100")

ggplot(results, aes(x = paste(year, " / ", quarter),
                    y = articles_published, color = title)) +
  geom_point() +
  labs(title = "Number of Articles Published by Journal and Quarter",
       x = "Year and Quarter",
       y = "Number of Articles Published",
       color = "Journal Title") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5,
                                   size = 6))
```



4. Disconnect database

```
dbDisconnect(mydbcon)
```

```
## [1] TRUE
```