Practicum II CS5200: Analyzing Data

Spring 2023

##Connect To MySQL

```
library(RMySQL)
```

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

```
db_user <- 'root'
db_password <- 'banana123'
db_name <- 'PracticumTwo'
db_host <- 'localhost'
db_port <- 3306

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

##Analytical Query I

I think a table is actually the best method of displaying the data for the first analytical query. There are only 5 journals listed, and only two values are listed for each of them. A table effeciently shows the data in a clear manner, and there does not seem to be any significant advantage to be gained by displaying the data in a different way.

I chose to show two versions of this query, one that shows the top 5 journals by articles published in the years 1975 and 1976, and one that shows the top 5 journals by articles published in all years.

A note with the way the fact table is structured is that rows with a year value of -1 represent data from all years across the given month value, and rows with a month value of -1 represent data from all months across the given year value. If both are -1, it represents the data from all years and months combined. Additionally, months 13-16 represent quarters 1-4.

The following table shows the top 5 journals by articles published in the years 1975 and 1976. The years can be easily modified.

```
SELECT JournalTitle, SUM(NumArticles) AS NumArticles FROM (SELECT * FROM FactTable WHERE (YEAR = 1975 O
```

Table 1: 5 records

JournalTitle	NumArticles
The Journal of pharmacy and pharmacology	498
Biochimica et biophysica acta	406
Annales de l'anesthesiologie francaise	307
Comparative biochemistry and physiology. A, Comparative physiology	248
The Journal of biological chemistry	245

This table shows the top 5 journals by articles published across all years.

```
SELECT JournalTitle, SUM(NumArticles) AS NumArticles FROM (SELECT * FROM FactTable WHERE YEAR = -1 AND
```

Table 2: 5 records

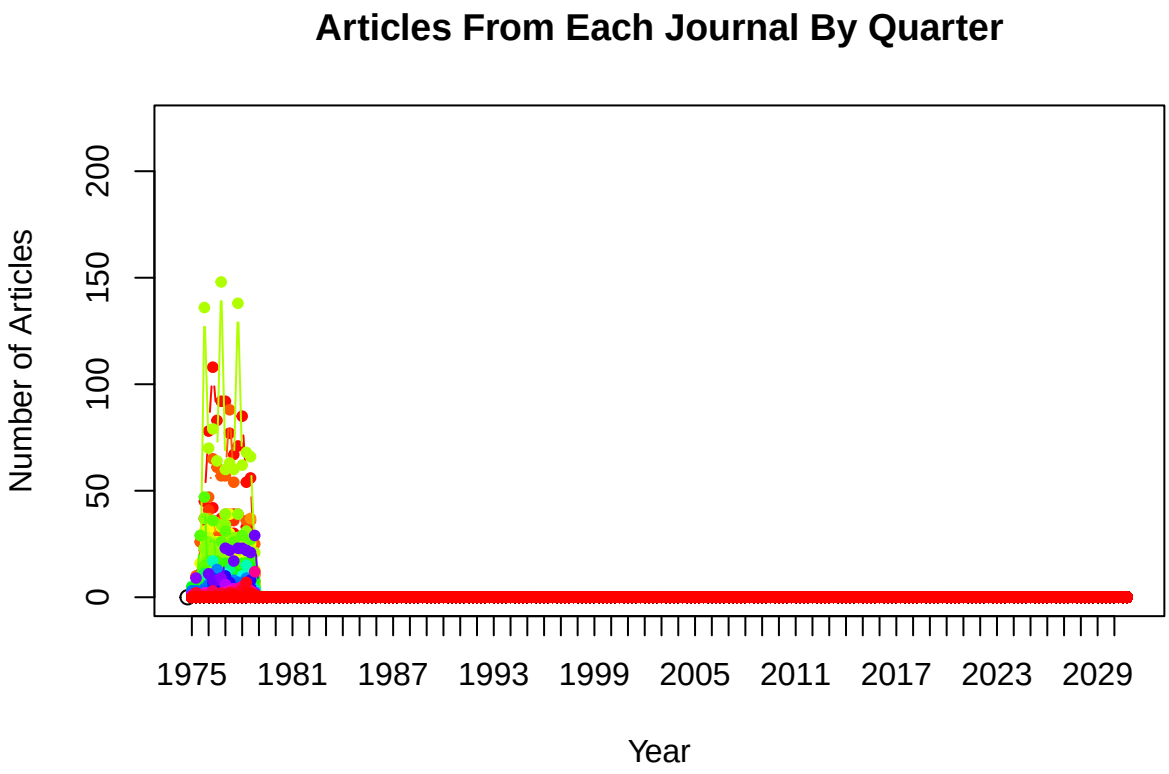
JournalTitle	NumArticles
The Journal of pharmacy and pharmacology	1036
Biochimica et biophysica acta	920
The Journal of biological chemistry	604
Annales de l'anesthesiologie francaise	542
Biochemistry	375

##Analytical Query II

```
rawData <- dbGetQuery(mydb,"SELECT JournalTitle, Year, Month, NumArticles FROM FactTable WHERE MONTH >
```

This is a plot that shows all of the data from my literal interpretation of Analytical Query II (Number of articles per journal per year broken down by quarter). It shows the amount of articles every journal published every quarter of every year. I don't see this plot as particularly useful, as there are over 1000 lines on the chart that don't have clear labels, but it gives a graph visualization of Analytical Query II. I don't see a better way to display this much data all at once. For this reason, I have an example of a more fine-tuned version of Analytical Query II in the next two code chunks after this one.

```
#Note: This chunk takes noticable time (still less than a minute for me) to run because it has to forma
uniqueJournals = rawData[!duplicated(rawData$JournalTitle),]
numJournals = nrow(uniqueJournals)
journalTitles = 1:numJournals
legendColors = 1:numJournals
maxArticles <- dbGetQuery(mydb,"SELECT NumArticles FROM FactTable WHERE MONTH > 12 AND YEAR > 1 ORDER B
plot(0,0,main="Articles From Each Journal By Quarter",
   xlab="Year", ylab="Number of Articles",xlim = c(1,224), xaxt = "n",ylim = c(0,maxArticles*1.5))
axis(side = 1, at = seq(1, 224, 4), labels = 1975:2030)
cl = rainbow(numJournals)
for (i in 1:numJournals) {
  currentTitle = uniqueJournals[i, "JournalTitle"]
  journalTitles[i] = currentTitle
  relevantData = rawData[which(rawData$JournalTitle == currentTitle),]
  months = 1:((2030-1975 + 1)*4) #equation results in 224 total quarters to get data for, kept here to
  numArticles = 1:((2030-1975 + 1)*4)
  for (j in 1:nrow(relevantData)) {
    numArticles[j] = relevantData[j, 4]
  }
  lines(1:((2030-1975 + 1)*4),numArticles,col = cl[i],type = 'b',pch=20)
}
```

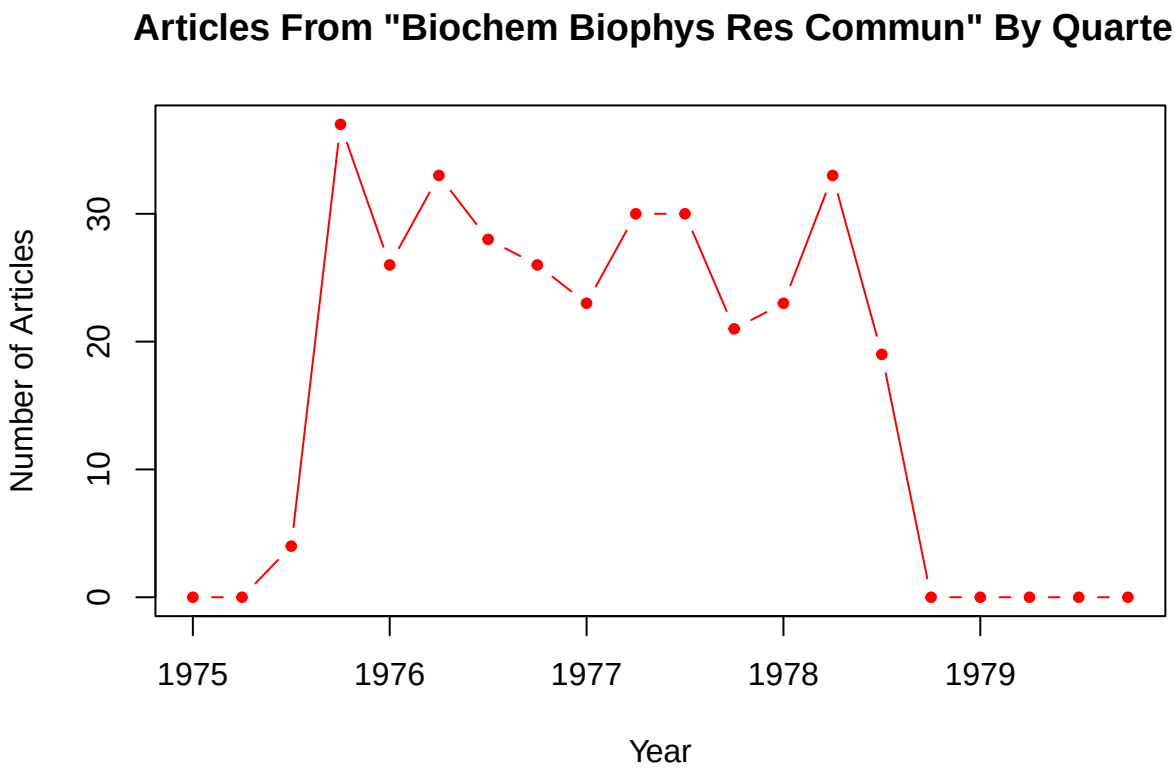



```
#There are way too many journals for a legend to be feasible, I don't see a way to reasonably show whic
#legendSize = 4/numJournals
#legend("topright", legend = journalTitles, pch=15, col = cl, cex = legendSize)
```

```
rawData <- dbGetQuery(mydb,"SELECT JournalTitle, Year, Month, NumArticles FROM FactTable WHERE MONTH > 
```

This version of Analytical Query II only shows the data from one journal, which feels much more helpful than trying to show data from all of the journals at once. It could be modified to show data from more journals if desired.

```
journalTitle = "Biochemical and biophysical research communications"
maxArticles <- dbGetQuery(mydb,"SELECT NumArticles FROM FactTable WHERE MONTH > 12 AND YEAR > 1 AND YEA
plot(0,0,main="Articles From \"Biochem Biophys Res Commun\" By Quarter",
   xlab="Year", ylab="Number of Articles",xlim = c(1,20), xaxt = "n",ylim = c(0,maxArticles))
axis(side = 1, at = seq(1, 20, 4), labels = 1975:1979)
months = 1:((1979-1975 + 1)*4) #equation results in 20 total quarters to get data for, kept here to sho
numArticles = 1:((1979-1975 + 1)*4)
for (i in 1:nrow(rawData)) {
  numArticles[i] = rawData[i, 4]
}
lines(1:((1979-1975 + 1)*4),numArticles,col = 'red',type = 'b',pch=20)
```



##Disconnect From MySQL

```
dbDisconnect(mydb)
```

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