Explore and Mine Data

```
require(RMySQL)
```

```
## Loading required package: RMySQL
```

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

```
# Connect to mysql database
mysqlconnection = dbConnect(RMySQL::MySQL(),
                            dbname='PracticumII',
                            host='localhost',
                            port=3306,
                            user='root',
                            password='rootroot')
```

Analytical Query I

Top five journals with the most articles published in them for the time period.

```
require(RMySQL)
journal_article_count <- function(){
  dbGetQuery(mysqlconnection, 'SELECT Journal_Id,
                                      Title,
                                      Publish_year,
                                      SUM(Article_count)
                               FROM PracticumII.Fact_Journal
                               WHERE Publish_year = 1976
                               GROUP BY 1, 2, 3
                               ORDER BY 4 DESC
                               LIMIT 5')
}
journal_article_count()
```

```
##   Journal_Id                                                                Title
## 1        444                          The Journal of pharmacy and pharmacology
## 2         11                                    Biochimica et biophysica acta
## 3        119                                The Journal of biological chemistry
## 4         48 Comparative biochemistry and physiology. A, Comparative physiology
## 5        567                                Annales de l'anesthesiologie francaise
##   Publish_year SUM(Article_count)
## 1         1976                362
## 2         1976                361
## 3         1976                230
## 4         1976                219
## 5         1976                161
```

With the fact table, if someone is interested in knowing the top journal by article count during a period of time, they can just change the filter in the where clause. This can be done by someone with limited SQL knowledge. This information is better represented by a table, since table contains more information about journal title, and the time period.

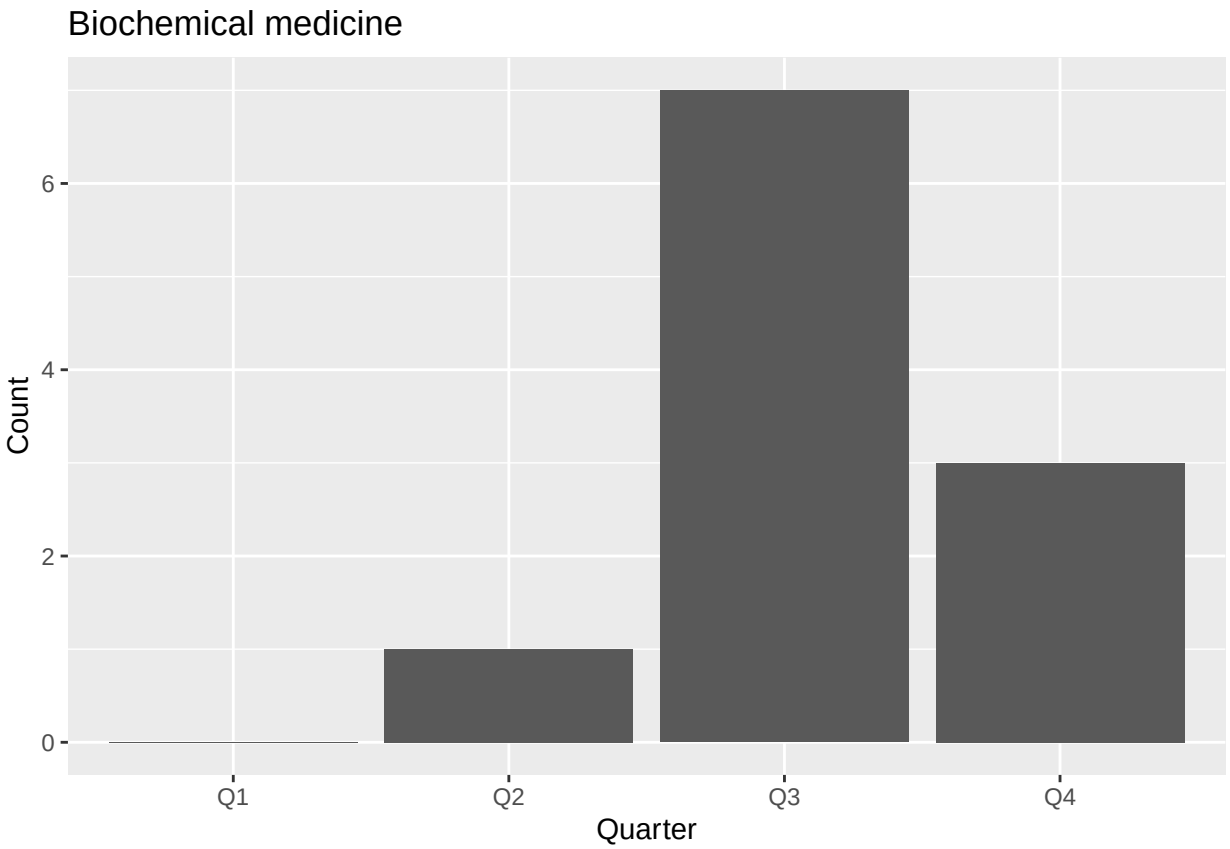
Analytical Query II

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

```
articles_per_journal <- function(){
  dbGetQuery(mysqlconnection, 'SELECT Journal_Id,
                                      Title,
                                      Publish_year,
                                      IFNULL(SUM(CASE WHEN Publish_Quarter = "Q1" THEN Article_count END
                                      IFNULL(SUM(CASE WHEN Publish_Quarter = "Q2" THEN Article_count END
                                      IFNULL(SUM(CASE WHEN Publish_Quarter = "Q3" THEN Article_count END
                                      IFNULL(SUM(CASE WHEN Publish_Quarter = "Q4" THEN Article_count END
                               FROM PracticumII.Fact_Journal
                               GROUP BY 1, 2, 3')
}
articles_per_journal <- articles_per_journal()
head(articles_per_journal, 5)
```

```
##   Journal_Id                Title Publish_year Q1 Q2 Q3 Q4
## 1          1 Biochemical medicine         1975  0  1  7  3
## 2          1 Biochemical medicine         1976  0  4  1  5
## 3          1 Biochemical medicine         1977  2  1  1  2
## 4          1 Biochemical medicine         1978  3  3  0  0
## 5         10         Biochemistry         1975  0  0  0 25
```

```
library(data.table)
library(ggplot2)
plot_articles_per_journal <- function(journal_title, year) {
  df <- articles_per_journal[articles_per_journal$Title == journal_title
                             & articles_per_journal$Publish_year == year, c("Q1", "Q2", "Q3", "Q4")]
  df_t <- transpose(df)
  df_t$Quarter <- c("Q1", "Q2", "Q3", "Q4")
  p <-ggplot(df_t, aes(x=Quarter, y=V1)) + geom_bar(stat="identity")
  p + labs(title=journal_title,
        x ="Quarter", y = "Count")
}
plot_articles_per_journal("Biochemical medicine", 1975)
```


Biochemical medicine
Count
0
2
4
6
Q1
Q2
Q3
Q4
Quarter