Explore and Mine Data (Omer Koc and Etki Acilan)

Connect to the database.

```
# 1. Library (must be installed prior to loading)
# Load the required packages
library(RMySQL)
```

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

```
# 2. My SQL Settings
db_user <- 'acilankocdbp2'
db_password <- 'SQL12345'
db_name <- 'acilankocdbp2'
db_host <- 'db4free.net'
db_port <- 3306 # always this port unless you change it during installation

# 3. Connect to DB (mydb)
mydb = dbConnect(MySQL(), user = db_user, password = db_password,
                 dbname = db_name, host = db_host, port = db_port)
```

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

We suppose our manager aims to determine the five most published journals that have the highest number of articles within a certain time frame. This data will be helpful in assessing the journals' performance and help make more accurate decisions about future publishing strategies. In order to simplify the presentation of the data, a bar chart or table can be constructed that displays the journal titles in decreasing order based on the number of articles published. In this case, we preferred to visualize our data in a simple table format as it is easily readable.

```
query_result <- dbSendQuery(mydb, "SELECT Title, SUM(NumArticlesYear) AS NumArticlesYear
                            FROM JournalF WHERE Year BETWEEN 1977 AND 1980 GROUP BY Title
                            ORDER BY SUM(NumArticlesYear) DESC LIMIT 5")
```

```
## Warning in .local(conn, statement, ...): Decimal MySQL column 1 imported as
## numeric
```

```
# Fetch the results
query_data <- dbFetch(query_result)

# Table
query_data
```

```
##                                     Title NumArticlesYear
## 1 The Journal of pharmacy and pharmacology            2148
## 2            Biochimica et biophysica acta            2056
## 3      The Journal of biological chemistry            1436
## 4                             Biochemistry             832
## 5                    Aktuelle Gerontologie             720
```

Our manager could conclude from this table that 'The Journal of pharmacy and pharmacology' has the highest performance in publishing articles in given year interval. It is seen that there is a rivalry between 'The Journal of pharmacy and pharmacology' and 'Biochimica et biophysica acta', and other journals can not seem to be catching them.

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

In this case, per the report requested by our manager, the article distribution of a journal called "American Journal of Hospital Pharmacy" is requested by years. In addition, our manager also asks the quarterly performance of the magazine in order to assess the performance of the magazine in depth.

This report will provide information on the journal's past performance and will be an important tool for future publication strategies and goals.

The collection, analysis and visualization of data is an important step in understanding the journal's performance and should be presented in a manner consistent with the manager's demands.

```
# All data
query_result <- dbSendQuery(mydb, "SELECT Title, Year, Quarter, SUM(NumArticlesQuarter)
                                   as NumArticlePerYearQuarter
                                   FROM JournalF
                                   GROUP BY Title, Year, Quarter LIMIT 5")
```

```
## Warning in .local(conn, statement, ...): Decimal MySQL column 3 imported as
## numeric
```

```
# Fetch the results
query_data <- dbFetch(query_result)
query_data
```

```
##                                    Title Year Quarter NumArticlePerYearQuarter
## 1                          AANA journal 1976       2                        1
## 2                          AANA journal 1977       1                        1
## 3                          AANA journal 1977       2                        1
## 4 AJR. American journal of roentgenology 1976       1                        1
## 5 AJR. American journal of roentgenology 1976       2                        1
```

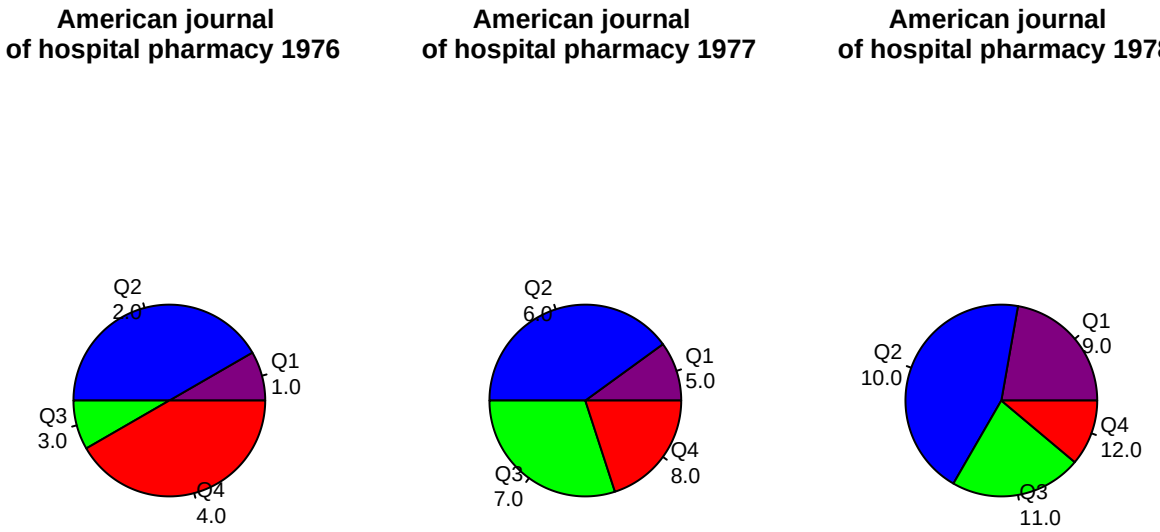
Here, we showed that our query and the fact table support the required analytical query.

```
query_result2 <- dbSendQuery(mydb, "SELECT Title, Year, Quarter, SUM(NumArticlesQuarter)
                                as NumArticlePerYearQuarter FROM JournalF
                               WHERE Title = 'American journal of hospital pharmacy'
                               GROUP BY Title, Year, Quarter")
```

```
## Warning in .local(conn, statement, ...): Decimal MySQL column 3 imported as
## numeric
```

```r
# Fetch the results
query_data2   <- dbFetch(query_result2)
years_unique  <- unique(query_data2$Year)
quarter       <- unique(query_data2$Quarter)
labels        <- c("Q1", "Q2", "Q3", "Q4")
colors <- c("#800080", "#0000FF", "#00FF00", "#FF0000")


par(mfrow=c(1,length(years_unique)))
for (i in years_unique) {
  values <- which(query_data2$Year == i)
  labels_with_values <- sprintf("%s\n%.1f", labels, values)
  pie(query_data2$NumArticlePerYearQuarter[values], labels = labels_with_values,
      col = colors, main = paste("American journal \n of hospital pharmacy", i))
}
```



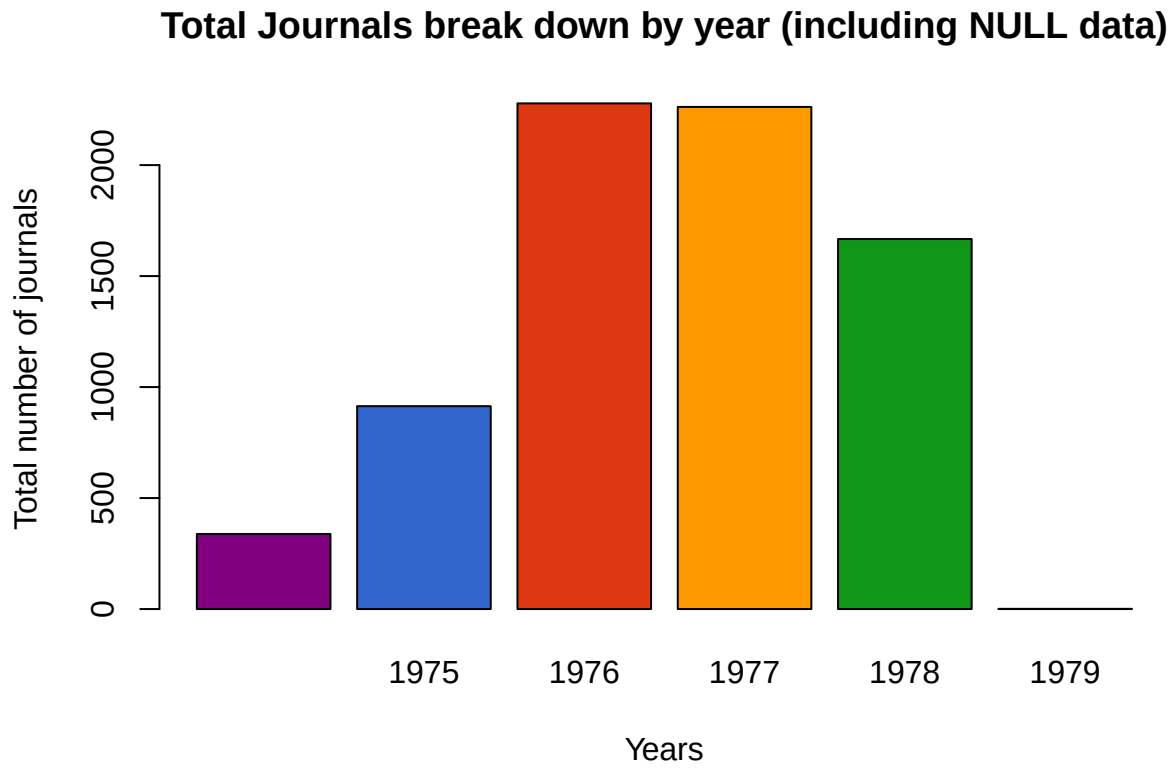
From the pie chart, our manager could conclude that generally there is high performance in Q2. Also, there is an increase in number of articles from years 1976 to 1978. Moreover, it is obvious from the pie chart that there is no balance between the quarters in publication number.

Analytical Query III: A Custom Query designed by us.

A company executive wants to use data to make decisions that will help the company grow. They want to look at all the journals in our database and see how they did each year. This will help them make better plans for marketing, sales, and growth. And also this will show the best times in the past and help plan for the future.

```
custom_query <- dbSendQuery(mydb, "
SELECT Year, COUNT(*) as TotalJournals
FROM JournalF
WHERE JournalID IS NOT NULL
GROUP BY Year
ORDER BY Year")
# Fetch the results
custom_query_data <- dbFetch(custom_query)
color = c("#800080","#3366CC", "#DC3912", "#FF9900", "#109618", "#990099")


barplot(custom_query_data$TotalJournals, names.arg=custom_query_data$Year,
        xlab="Years", ylab="Total number of journals",
        main="Total Journals break down by year (including NULL data)",
        col=color)
```



In this bar chart, there is a significant increase in number of published articles after year 1975 and it continues until 1978. Our data lack information after 1978. Our manager can conclude that the same strategies should be applied in order to increase the total number of published articles. (Please note that the purple bar in the chart corresponds to NULL years.)