**Stat 133, Fall 05**
**Introduction to SQL**
**October 31, 2005**

- load the RMySQL package.

  ```
  library(RMySQL)
  ```

- load a driver for a MySQL-type database

  ```
  drv = dbDriver( "MySQL" )
  ```

- make a connection to the database management server of interest

  ```
  con = dbConnect( drv , user = "s133xx", dbname =
  "BaseballDataBank" , host = "statdocs.berkeley.edu" )
  ```

- find out what tables are available for this database.

  ```
  dbListTable(con)
  ```

**Question: What college produced the greatest number of major league baseball players?**

How can we answer this question:

(1) Which tables do we need to look at?
   - Master. Look at the table at the end of this handout to see a list of the attributes from the Master table.

(2) What are the output attributes?
   - College, COUNT(college)

(3) What are the tuples?
   - One tuple for every player who went to college.

- Let's R it up!

  ```
  query = dbGetQuery( con , "SELECT college,
       COUNT(college)
       FROM Master
       GROUP BY college;")
  ```

- Puts the colleges in order from most attended to least attended by the people in the database

  ```
  query = query[ order( query[,2], decreasing  = TRUE),]
  query[ 1:10, ]
  ```

- Note how the top two colleges are ‘  ‘ (blank), and ‘None’. This is not what we want because these are the number of players who did not attend college and went straight to the major leagues. We want to be able to return the name of a college as the top entry. Therefore we should specify not to return certain information. For instance, we do not want to count people who did not attend a college, and these people can be identified when the college information is '  ' or 'None'. We are also only concerned with baseball players and not those people in the table who are only managers, but not players. To ensure we count only individuals who were players we ask not to count those people who do not have a player ID. Let’s try a different approach using the following query.

```
colleges = dbGetQuery( con ,
  "SELECT college, COUNT(college)
  FROM Master
  WHERE playerID != '' AND college != 'None' AND college != ''
  GROUP BY college;")
```

- Now we will answer the same question using R commands. First Bring the entire Master table over into R.

```
master.table = dbGetQuery(con, “SELECT * FROM Master;” )
```

MASTER table

| | |
|---|---|
| playerID | A unique code asssigned to each player. The playerID links the data in this file with records in the other files. |
| managerID | An ID for individuals who served as managers |
| hofID | An ID for individuals who are in teh baseball Hall of Fame |
| birthYear | Year player was born |
| birthMonth | Month player was born |
| birthDay | Day player was born |
| birthCountry | Country where player was born |
| birthState | State where player was born |
| birthCity | City where player was born |
| deathYear | Year player died |
| deathMonth | Month player died |
| deathDay | Day player died |
| deathCountry | Country where player died |
| deathState | State where player died |
| deathCity | City where player died |
| nameFirst | Player's first name |
| nameLast | Player's last name |
| nameNote | Note about player's name (usually signifying that they changed their name or played under two differnt names) |
| nameGiven | Player's given name (typically first and middle) |
| nameNick | Player's nickname |
| weight | Player's weight in pounds |
| height | Player's height in inches |
| bats | Player's batting hand (left, right, or both) |
| throws | Player's throwing hand (left or right) |
| debut | Date that player made first major league appearance |
| college | College attended |
| lahman40ID | ID used in Lahman Database version 4.0 |
| lahman45ID | ID used in Lahman database version 4.5 |
| retroID | ID used by retrosheet |
| holtzID | ID used by Sean Holtz's Baseball Almanac |
| bbrefID | ID used by Baseball Reference website |