**Appendix**

Jessica Wheeler, Earl Shaw, Vishal Ahir

Code used:

USE mydb;

#drop table batting;

CREATE TABLE batting (

#record\_id varchar(25),

player\_id varchar(25),

year INTEGER,

stint INTEGER,

team\_id varchar(10),

league\_id varchar(10),

g INTEGER,

ab INTEGER,

r INTEGER,

h INTEGER,

dble INTEGER,

triple INTEGER,

hr INTEGER,

rbi INTEGER,

sb INTEGER,

cs INTEGER,

bb INTEGER,

so INTEGER,

ibb INTEGER,

hbp INTEGER,

sh INTEGER,

sf INTEGER,

g\_idp INTEGER

#primary key (player\_id)

);

#IMPORT BATTING FILE

load data local infile 'C:\\Users\\Earl\\Documents\\MSDS 7330\\Projects\\batting\_2.csv'

into table batting fields terminated by ',' ENCLOSED BY '"'

ignore 1 lines

select count(\*) from batting

select \*

from batting

limit 5

#drop table pitching;

CREATE TABLE pitching (

#record\_id varchar(25),

player\_id varchar(25),

year INTEGER ,

stint INTEGER,

team\_id varchar(10),

league\_id varchar(10),

w INTEGER,

l INTEGER,

g INTEGER,

gs INTEGER,

cg INTEGER,

sho INTEGER,

sv INTEGER,

ipouts NUMERIC,

h INTEGER,

er INTEGER,

hr INTEGER,

bb INTEGER,

so INTEGER,

baopp NUMERIC(5,2),

era NUMERIC (5,2),

ibb NUMERIC (5,2),

wp NUMERIC (5,2),

hbp NUMERIC (5,2),

bk INTEGER ,

bfp NUMERIC (5,2),

gf NUMERIC (5,2),

r INTEGER,

sh NUMERIC (5,2),

sf NUMERIC (5,2),

g\_idp NUMERIC (5,2));

#primary key (record\_id));

#IMPORT PITCHING FILE

load data local infile 'C:\\Users\\Earl\\Documents\\MSDS 7330\\Projects\\pitching\_2.csv'

into table pitching fields terminated by ',' ENCLOSED BY '"'

ignore 1 lines

select count(\*) from pitching

select \*

from pitching

limit 5

#drop table salary;

CREATE TABLE salary (

Year INTEGER,

Team\_id VARCHAR (5),

League\_id VARCHAR (5),

Player\_id VARCHAR (40),

Salary INTEGER);

#IMPORT SALARY FILE

load data local infile 'C:\\Users\\Earl\\Documents\\MSDS 7330\\Projects\\salary\_2.csv'

into table salary fields terminated by ',' ENCLOSED BY '"'

ignore 1 lines

select count(\*) from salary

where year in (2013,2014,2015)

select \*

from salary

limit 5

#drop table Player;

CREATE TABLE player (

player\_id VARCHAR(40),

birth\_year INTEGER,

birth\_month INTEGER,

birth\_day INTEGER,

birth\_country VARCHAR(40),

birth\_state VARCHAR(40),

birth\_city VARCHAR(40),

death\_year INTEGER,

death\_month INTEGER,

death\_day INTEGER,

death\_country VARCHAR(40),

death\_state VARCHAR(40),

death\_city VARCHAR(40),

name\_first VARCHAR(40),

name\_last VARCHAR(40),

name\_given VARCHAR(40),

weight INTEGER,

height INTEGER,

bats VARCHAR(40),

throws VARCHAR(40),

debut VARCHAR(40),

final\_game VARCHAR(40),

retro\_id VARCHAR(40),

bbref\_id VARCHAR(40));

/\* IMPORT PLAYER FILE\*/

load data local infile'C:\\Users\\Earl\\Documents\\MSDS 7330\\Projects\\player.csv'

into table player fields terminated by ',' ENCLOSED BY '"'

ignore 1 lines

#-------------------------------------------------------------------------------------------

#-------------------------------------------------------------------------------------------

#Getting year and salary information from "salary.csv" file for players from 2013-2015

drop table player\_2;

create table player\_2 as

select distinct a.\*,

b.year as salary\_year,

case when a.birth\_country='USA' then 1

else 0

end as american\_born,

case when a.birth\_country <> 'USA' then 1

else 0

end as foreign\_born

from player a

inner join salary b on a.player\_id=b.player\_id #"inner join" b/c we only want 2013-2015 information, and "salary.csv" file has 2013-2015 information.

order by a.player\_id;

#-------------------------------------------------------------------------------------------

#-------------------------------------------------------------------------------------------

#Count of American-born and Foreign-born players by birth month

select birth\_month,

salary\_year,

sum(american\_born) as american\_born\_count,

sum(foreign\_born) as foreign\_born\_count

from player\_2

group by birth\_month,salary\_year

order by salary\_year,birth\_month;

#-------------------------------------------------------------------------------------------

#-------------------------------------------------------------------------------------------

#Creating table with batting and salary information for player from 2013-2015

drop table batting\_2;

create table batting\_2 as

select distinct concat(a.name\_first,' ',a.name\_last) as full\_name,

a.birth\_month,

a.birth\_country,

c.salary,

b.\*,

case when birth\_month=8 then 'august'

else 'other'

end as birth\_group

from player a

inner join batting b on a.player\_id=b.player\_id #"inner join" b/c we only want 2013-2015 information, and "batting.csv" file has 2013-2015 information.

left join salary c on b.player\_id=c.player\_id

and b.year=c.year

and b.team\_id=c.team\_id;

#-------------------------------------------------------------------------------------------

#-------------------------------------------------------------------------------------------

#Summary Statistics by Birth Group

select birth\_group,

avg(ab) as avg\_at\_bats,

avg(r) as avg\_runs,

avg(h) as avg\_hits,

avg(dble) as avg\_doubles,

avg(hr) as avg\_home\_runs,

case when (sum(ab)>=20) then sum(h)/sum(ab)

else 0

end as batting\_avg

from batting\_2

where birth\_country='USA'

group by birth\_group

order by birth\_group;

#-------------------------------------------------------------------------------------------

#-------------------------------------------------------------------------------------------

#Top 5 Batting Averages by Year

select year,

full\_name,

birth\_month,

sum(h)/sum(ab) as batting\_avg

from batting\_2

where birth\_country='USA' and year='2015' /\*'2013' '2014'\*/ and ab>=20 #Batter must have at least 20 at bats

group by year,full\_name,birth\_month

order by year,batting\_avg desc

limit 5;

#-------------------------------------------------------------------------------------------

#-------------------------------------------------------------------------------------------

#Creating table with home run bins

drop table hr\_bin;

create table hr\_bin as

select distinct player\_id,

year,

case when sum(hr)>=0 and sum(hr)<=10 then '0-10'

when sum(hr)>10 and sum(hr)<=20 then '11-20'

when sum(hr)>20 and sum(hr)<=30 then '21-30'

when sum(hr)>30 and sum(hr)<=40 then '31-40'

when sum(hr)>40 then '40+'

else ''

end as hr\_bin

from batting\_2

group by player\_id,year;

#-------------------------------------------------------------------------------------------

#-------------------------------------------------------------------------------------------

#Joining table with home run bins to main table

drop table batting\_3;

create table batting\_3 as

select distinct a.\*,

b.hr\_bin

from batting\_2 a

left join hr\_bin b on a.player\_id=b.player\_id

and a.year=b.year

where a.stint=1 #Keeping only 1 record for each player

order by a.year,a.player\_id;

#-------------------------------------------------------------------------------------------

#-------------------------------------------------------------------------------------------

#Average Salary by Home Run Bin and Year

select hr\_bin,

year,

avg(salary)

from batting\_3

where birth\_country='USA'

group by hr\_bin,year

order by year,hr\_bin;

#-------------------------------------------------------------------------------------------

#-------------------------------------------------------------------------------------------

#Frequency of Birth Months in the '21-30' HR bin

select year,

birth\_month,

count(birth\_month) as count\_birth\_month

from batting\_3

where hr\_bin='21-30' and birth\_country='USA'

group by year,birth\_month,hr\_bin

order by year,birth\_month;

#-------------------------------------------------------------------------------------------

**SALARY**

/\*SALARY -- List of highest salaries by year\*/

SELECT \*

FROM player natural join salary

WHERE year = 2013 /\*Also 2014 and 2015\*/

ORDER BY salary desc;

/\*Adding Months to birth\_month\*/

CREATE TABLE players\_2 as

SELECT \*,

case when birth\_month=1 then 'January'

when birth\_month=2 then 'February'

when birth\_month=3 then 'March'

when birth\_month=4 then 'April'

when birth\_month=5 then 'May'

when birth\_month=6 then 'June'

when birth\_month=7 then 'July'

when birth\_month=8 then 'August'

when birth\_month=9 then 'September'

when birth\_month=10 then 'October'

when birth\_month=11 then 'November'

when birth\_month=12 then 'December'

else 'Missing'

end as birth\_month\_text,

case when birth\_country='USA' then 1

else 0

end as american,

case when birth\_country <> 'USA' then 1

else 0

end as non\_american

FROM player;

/\*View of table\*/

select \*

from players\_2;

select \*

from salary natural join player;

/\*Count by Birthplace\*/

select birth\_month,

birth\_month\_text,

sum(american) as AMERICAN\_COUNT,

sum(non\_american) as FOREIGN\_COUNT

from players\_2

group by birth\_month, birth\_month\_text

order by birth\_month;

/\* Count by birthplace 2013 - replace for each year\*/

select birth\_month,

birth\_month\_text,

sum(american) as AMERICAN\_COUNT,

sum(non\_american) as FOREIGN\_COUNT

from players\_2 natural join salary

WHERE year=2005

group by birth\_month, birth\_month\_text

order by birth\_month;

**PITCHING**

select year, stint, max(r)

from batting

group by 1,2

order by year, stint, max(r) desc

select year, player\_id, league\_id, r,@curRank := @curRank + 1 AS rank

from

batting,(

SELECT @curRank := 0

) q

-- where year = 2015

order by r desc

select year, player\_id, league\_id, r, w,@curRank := @curRank + 1 AS rank

from

pitching,(

SELECT @curRank := 0

) q

-- where year = 2015

order by w desc

2281

select count(\*)

from

pitching p1,

player p2

where p1.player\_id = p2.player\_id

2281

-- filtering only 2015, 2014 and 2013 records for player file

-- as for pitching and batting has only 3 years of data

delete from player

where player\_id not in

( select player\_id from pitching);

commit

select

from

pitching p1,

player p2

where p1.player\_id = p2.player\_id

-- Vishal

-- Queries for identifying top 5 pitchers in terms of Wins, Runs and StrikeOuts

-- Query to get birth years for top 5 players based on wins, strike outs and runs

-- select a.player\_id, rank, p.birth\_year, p.birth\_month, p.birth\_day

select p.birth\_city, count(\*)

from

(select year, player\_id, league\_id, w,@curRank := @curRank + 1 AS rank

from

pitching,(

SELECT @curRank := 0

) q

-- where year = 2015

order by w desc) a,

player p

where rank <= 50

and p.player\_id = a.player\_id

group by 1

-- order by rank;

-- query to identify top 5 rows, yearwise for runs, wins and strikeout.

-- column names are as follows:

-- runs = r

-- wins = w

-- strikeout = so

select year, concat(name\_last, ", ", name\_first) , league\_id, w, birth\_year,

case birth\_month

when 1 then 'January'

when 2 then 'February'

when 3 then 'March'

when 4 then 'April'

when 5 then 'May'

when 6 then 'June'

when 7 then 'July'

when 8 then 'August'

when 9 then 'September'

when 10 then 'October'

when 11 then 'November'

when 12 then 'December'

else 'unknown'

end as birth\_mon

from

(

select year, p1.player\_id, league\_id, w,@curRank := @curRank + 1 AS rank

from

pitching p1,(

SELECT @curRank := 0

) q

where year = 2015

order by w desc

) a,

player p

where a.player\_id = p.player\_id

and rank <= 5

order by rank