Q1

It is possible that in doing this there may be a conflict with some of the baseball data and/or there is missing data, which means that the primary and/or foreign keys may prevent you from inserting some of the data. There are three ways you can resolve this problem. What are they (one sentence each, maximum)? What is the necessary SQL to implement the solution in each case.

Option 1: Disable foreign-key checks and do not modify the data.

```
drop database baseball2016;
SET FOREIGN_KEY_CHECKS=0;
source /home/z5mohamm/ece356-a1/lahman2016-tables.sql;
source /home/z5mohamm/ece356-a1/lahman2016-data.sql;
SET FOREIGN_KEY_CHECKS=1;
```

Option 2: Insert data into referenced table where a foreign key is referencing something that does not exist.

```
drop database baseball2016;
SET FOREIGN_KEY_CHECKS=0;

source /home/z5mohamm/ece356-a1/lahman2016-tables.sql;
source /home/z5mohamm/ece356-a1/lahman2016-data.sql;

INSERT INTO Schools (schoolid) (SELECT DISTINCT schoolid FROM
CollegePlaying WHERE schoolid NOT IN (SELECT schoolid FROM Schools));
INSERT INTO Master (playerid) (SELECT DISTINCT playerid FROM HallOfFame
WHERE playerid NOT IN (SELECT playerid from Master));
INSERT INTO Master (playerid) (SELECT DISTINCT playerid FROM Salaries WHERE
playerid NOT IN (SELECT playerid from Master));

SET FOREIGN_KEY_CHECKS=1;

-- Should be empty set
SELECT DISTINCT schoolid FROM CollegePlaying WHERE schoolid NOT IN (SELECT
schoolid FROM Schools);
-- Should be empty set
SELECT DISTINCT playerid FROM HallOfFame WHERE playerid NOT IN (SELECT
playerid from Master);
-- Should be empty set
SELECT DISTINCT playerid FROM Salaries WHERE playerid NOT IN (SELECT
playerid from Master);
```

Option 3. Delete anything in the referencing table that is referencing data that does not exist.

```
drop database baseball2016;
SET FOREIGN_KEY_CHECKS=0;
```

```
source /home/z5mohamm/ece356-a1/lahman2016-tables.sql;
source /home/z5mohamm/ece356-a1/lahman2016-data.sql;

DELETE FROM CollegePlaying WHERE schoolid NOT IN (SELECT schoolid FROM
Schools);
DELETE FROM HallOfFame WHERE playerid NOT IN (SELECT playerid FROM Master);
DELETE FROM Salaries WHERE playerid NOT in (select playerid FROM Master);

SET FOREIGN_KEY_CHECKS=1;
-- Should be empty set
SELECT DISTINCT schoolid FROM CollegePlaying WHERE schoolid NOT IN (SELECT
schoolid FROM Schools);
-- Should be empty set
SELECT DISTINCT playerid FROM HallOfFame WHERE playerid NOT IN (SELECT
playerid from Master);
-- Should be empty set
SELECT DISTINCT playerid FROM Salaries WHERE playerid NOT IN (SELECT
playerid from Master);
```

Q2

The SQL file has a very large number of INSERT statements in order to load the data into the database. It is typically preferred to load data directly from source files. In the case of the Baseball data, the course files are “Comma-Separated Variable” (or CSV) files. Create a LOAD statement that will load the data for the Batting CSV (Batting.csv) into its associated table. You should verify that your LOAD statement operates correctly and issues no warnings. Where is the CSV data located relative to the CLI and to the DB Server? Time how long it takes to LOAD the CSV vs. Using the equivalent INSERT statement method.

Where is the CSV data located relative to the CLI and to the DB Server?
The DB server is located at `127.0.0.1:3306`. The CSV is stored in the `secure_file_priv` as defined below.

```
+------------------+-----------------------+
| Variable_name    | Value                 |
+------------------+-----------------------+
| secure_file_priv | /var/lib/mysql-files/ |
+------------------+-----------------------+
```

Time how long it takes to LOAD the CSV vs. Using the equivalent INSERT statement method.
It takes `6s` using load and `7s` using insert.

```
    Starting INSERT test
    CURRENT_TIMESTAMP
    2020-01-19 20:28:19
    -- insert data
    CURRENT_TIMESTAMP
    2020-01-19 20:28:26
```

```
    Starting LOAD test
    CURRENT_TIMESTAMP
    2020-01-19 20:28:27
    -- load data
    CURRENT_TIMESTAMP
    2020-01-19 20:28:33
```

Q3

Create RA and SQL queries to answer each of the following questions.

a.
How many players have an unknown birthdate?
RA:

```
π COUNT(playerid) γ COUNT(playerid) ( σ birthyear ⊆ {'',0} ∨ birthmonth ⊆
{'',0} ∨ birthday ⊆ {'',0} (Master))
```

```
+--------------------------------+
| Players With Unknown Birthdays |
+--------------------------------+
|                            458 |
+--------------------------------+
```

b.
How many people are in the Hall of Fame? What fraction of each category of person are in the Hall Of Fame? Are more people in the Hall Of Fame alive or dead? Does this vary by category?
RA:

```
π COUNT(playerid) (σ inducted='Y' (HallOfFame))
```

```
+---------------------------+
| People in the Hall Of Fame |
+---------------------------+
|                       317 |
+---------------------------+
```

RA:

```
π a.category, a.CountHallOfFame, a.CountTotal, ρ
fraction(a.CountHallOfFame/a.CountTotal)
```

```
ρ a(π category, ρ CountHallOfFame(COUNT(playerid)) (γ category σ inducted =
'Y' (HallOfFame)))
```

```
+-------------------+-----------------+------------+----------+
| category          | counthalloffame | CountTotal | fraction |
+-------------------+-----------------+------------+----------+
| Manager           | 23              | 317        | 0.0726   |
| Pioneer/Executive | 34              | 317        | 0.1073   |
| Player            | 250             | 317        | 0.7886   |
| Umpire            | 10              | 317        | 0.0315   |
+-------------------+-----------------+------------+----------+
```

RA: For number of dead and alive people in the HallOfFame

```
π ρ alive(COUNT(q1.playerid)), ρ dead(COUNT(q2.playerid)) γ
COUNT(q1.playerid),COUNT(q2.playerid)
(ρ q1(σ inducted ='Y' ∧ deathyear = '' ∧ deathmonth = '' ∧ deathday = ''
(Master ⋈HallofFame)),
(ρ q2(σ inducted ='Y' ∧ deathyear != '' ∧ deathmonth != '' ∧ deathday !=
'' (Master ⋈HallofFame))))
```

There are more dead people in the HallOfFame then there are alive.

```
+-------+------+
| alive | dead |
+-------+------+
|    74 |  243 |
+-------+------+
```

The ammount of dead people does vary by category. The largest category for deaths is `player`, but proportionally the largest death fraction is present with `Pioneer/Executive`.
RA:

```
π category, status, COUNT (status) γ category, status, COUNT(status)
((π category, ρ status('dead') (σ deathyear !='' (π category, deathyear
(Master ⋈ Master.playerid = HallOfFame.playerid ∧ inducted = 'Y
HallOfFame))))
∪
(π category, ρ status('dead') (σ deathyear =='' (π category, deathyear
(Master ⋈ Master.playerid = HallOfFame.playerid ∧ inducted = 'Y
HallOfFame)))))
```

```
+-------------------+-------+------+
| category          | alive | dead |
+-------------------+-------+------+
| Player            | 65    | 185  |
| Manager           | 5     | 18   |
| Umpire            | 1     | 9    |
| Pioneer/Executive | 3     | 31   |
+-------------------+-------+------+
```

c.

What are the names and total pay (individually) of the three people with the three largest totalsalaries?

RA:

```
π namefirst, namelast, totalSalary (Master ⋈ Master.playerid = q1.playerid
τ q1 desc (ρ q1 (π playerid, ρ totalSalary( SUM (salary)) γ playerid,
SUM(salary) (Salaries))))
```

```
+-----------+-----------+-------------+
| namefirst | namelast  | TotalSalary |
+-----------+-----------+-------------+
| Alex      | Rodriguez | 398416252   |
| Derek     | Jeter     | 264618093   |
| Mark      | Teixeira  | 214275000   |
+-----------+-----------+-------------+
```

What category are these people?

RA:

```
TODO
```

```
+-----------+---------+--------+-------+
| playerid  | manager | player | other |
+-----------+---------+--------+-------+
| rodrial01 | N       | Y      | N     |
| jeterde01 | N       | Y      | N     |
| teixema01 | N       | Y      | N     |
+-----------+---------+--------+-------+
```

What are the top three in the other categories?

RA:

```
TODO
```

```
+----------+----------------+
| playerid | manager_salary |
+----------+----------------+
| rosepe01 |        1358858 |
+----------+----------------+
+-----------+--------------+
| playerid  | other_salary |
+-----------+--------------+
| belleal01 |     37417830 |
| vaughmo01 |     30333334 |
| hamptmi01 |     29003543 |
+-----------+--------------+
```

d.
What is the average number of Home Runs a player has?

```
π ρ"Average # of Homeruns" (SUM (hr) / COUNT (playerid)) (Batting)
```

```
+-----------------------+
| Average # of Homeruns |
+-----------------------+
|               15.2938 |
+-----------------------+
```

e.
If we only count players who got at least 1 Home Run, what is the average number of Home Runs a player has?
RA:

```
π ρ "Average Home Runs Excluding 0's"
(SUM(q1.total_homerun)/COUNT(q1.total_homerun))
γ SUM(q1.total_homerun), COUNT(q1.total_homerun)
ρ q1(
        π ρ total_homerun(SUM(hr))
        σ total_homerun>0
        γ playerid, SUM(hr)
        (Batting)
        )
```

```
+---------------------------------+
| Average Home Runs Excluding 0's |
+---------------------------------+
| 37.3944                         |
+---------------------------------+
```

f.
If we define a player as a good batter if they have more than the average number of Home Runs, and a player is a good Pitcher if they have more than the average number of ShutOut games, then how many players are both good batters and good pitchers?
RA:

```
π COUNT(Pitching.playerid)
(
        ρ q1(
                π Pitching.playerid, ρ
player_shut_out_total(SUM(Pitching.sho))
                σ player_shut_out_total > (π
SUM(Pitching.sho)/COUNT(Pitching.playerid)
                γ Pitching.playerid, SUM(Pitching.sho),
COUNT(Pitching.playerid)
                (Pitching)))
        ⋈ q1.playerid = q2.playerid
        ρ q2(
                π Batting.playerid, ρ
player_home_run_total(SUM(Batting.hr))
                σ player_home_run_total > (π
SUM(Batting.hr)/COUNT(Batting.playerid)
                γ Batting.playerid, SUM(Batting.hr),
COUNT(Batting.playerid)
                (Batting)))
)
```

```
+-------------------------------------------+
| Players that are good Batters and Pitchers |
+-------------------------------------------+
| 39                                        |
+-------------------------------------------+
```