Name: Tanubrata Dey (td2201)
UID- N10633072
BIOL GA 1009 HW2

Part A:

1. Which columns in the table are redundant? Once you simplify the number of columns, how would you reduce the redundancy?

   In partA.csv file containing the table, the columns Alias2, Experiment1_type & Experiment2_type are redundant because in Alias2 there are null values and in Experiment1_type and Experiment2_type, all values are repeating. It would be good to reduce the redundancy inside the table by joining Alias & Alias2 together and create a separate table as Alias, and for Experiment1_type and Experiment2_type create a different table named experimenttype so there is no repetition of data. Also, I created experiment table by joining the two columns of experimental values for each gene and connected it with experimenttype table and the main genes table to reduce redundancy the redundancy.

2. Given an example of the following types of relationships:

   a) One to one:

      Each gene has only one annotation
      OR
      Each gene has only one start point or one end point.

   b) One to many:

      Each gene may have two alias (i.e. one gene have many alias)

   c) Many to many

      Each gene is used in two different experiment types or have two different experimental values and alternatively one experiment is performed with many genes. So together many genes are used for many different experiment types or have many experimental values.

3. Create your new normalized database. Simplest way to do this is to create separate files that represent the different tables in your new normalized database and then import them into a new database. Call your new database “HW2<netid>.db” where <netid> represents your netid.

   HW2td2201.db

4. Provide the create statements for all tables in your final normalized database.
   a) Make sure to create primary and foreign keys where appropriate. It should be clear from your create statements which fields are your primary keys and which fields are your foreign keys?

Name: Tanubrata Dey (td2201)
UID- N10633072
BIOL GA 1009 HW2

Genes Table:

```
CREATE TABLE genes (
        gene_id         INTEGER       PRIMARY KEY,
        Gene            VARCHAR (9),
        Chromosome INTEGER,
        Start           INTEGER,
        Stop            INTEGER,
        Strand          CHAR,
        Annotation      TEXT
);
```

Alias Table:

```
CREATE TABLE alias (
        alias_id        INTEGER     PRIMARY KEY,
        gene_id         INTEGER     REFERENCES       genes (gene_id),
        alias           TEXT
);
```

Experiment Table:

```
CREATE TABLE experiment (
        exp_id          INTEGER     PRIMARY KEY,
        gene_id         INTEGER     REFERENCES       genes (gene_id),
        code            INTEGER     REFERENCES       experimenttype (code),
        value           DOUBLE
);
```

Experimenttype Table:

```
CREATE TABLE experimenttype (
        code   INTEGER     PRIMARY KEY,
        name  TEXT,
        type   TEXT
);
```

Part B:

5. Write a select statement, using JOIN, to return average experiment values for each gene.

```
SELECT a.gene_id AS Gene_ID,
a.Gene AS Gene_Name,
avg(b.value) AS Average
FROM genes a
INNER JOIN
experiment b ON a.gene_id = b.gene_id
GROUP BY (a.gene);
```