For regulatory purposes, Uber is interested in tracking the number of trips that originate and end in Bay Area counties.

To accomplish this you have been asked to analyze pickup and dropoff points for a random set of drivers over the course of a week.

(1) Download a Shapefile containing all US counties by following this link: ftp://ftp2.census.gov/geo/tiger/TIGER2015/COUNTY/.

Q1. What are the measurement units of the provided projection?

(2) Create a geo-enabled PostgreSQL database by following the following steps on a Linux or Mac machine (if you've already got this set up, you can skip these instructions):

- Install homebrew by following the documentation on this site: http://brew.sh/.

- Run the following commands:

```
brew install postgres
brew install postgis
pg_ctl -D /usr/local/var/postgres -l /usr/local/var/postgres/server.log start
```

Feel free to create your own database or use the default one called "postgres". Log onto the database and enable postgis by executing the command `create extension postgis`.

Q2. What is the output of the SQL statement `select st_point(-100, 50);` inside the postgres shell?

(3) If you successfully completed step (2), you should have access to the command `shp2psql`. Use it to load the counties shapefile into your database. The table should contain unprojected (WGS84) geometry data.

Q3a. Please provide the full command you used, including any arguments, to load the table. If you had to look it up, what reference website did you use?

Q3b. What columns does your postgres table contain? What are the datatypes of each column? What SQL command did you use to extract this information?

(4) The zip file provided with this test contains the trips you're interested in analyzing. The files contain four columns:

pickup_lng, pickup_lat, dropoff_lng, dropoff_lat

They are organized into folders according to the day of the week (1=Sunday ... 7=Monday), and the name of each file identifies the driver (should be an integer between 0 and 1,000).

Q4a. Two of the files contain invalid data. Identify the files and describe the issues, then exclude these files from further analysis. Please describe the process you used to perform this step.

Q4b. Please write and attach a script (python is preferred, BASH is okay) for reading all of the attached files and loading them into your Postgres database. The end result should be a table with the following columns:

driver_id int,
day_of_week varchar,
pickup_lng float,
pickup_lat float,
dropoff_lng float,
dropoff_lat float,
pickup_geom geometry,
dropoff_geom geometry

It's okay to also include a .sql file that contains the commands you would execute either before or after you run your script (to set up the table, etc.)

Q4c. How many trips originated in each county? How many ended in each county? Which county had the highest density of trips (per square mile) that both originated and ended within its borders? Please include the SQL queries you used in your answer.

Q4d. Based on a visual inspection of the data, describe how you might find invalid/possibly fraudulent trips.

Q4e. (extra credit). Create a map showing the volume of flow between all counties (be creative in how you style it).