DATA 311 - Fall 2020

Assignment #2

Name: J.Mo Yang Joanne Lee and Sejin Park

For this assignment, you will be using the ufo.csv file provided.

- To begin, import the data using pandas.
- While you have the data loaded into a DataFrame, split up the 'datetime' column into separate columns for year,month,day, and time.
- After that, use the to_sql command to turn this into a quick database
- You may receive a few warnings pertaining to the format of the data - these can be safely ignored for now

In [2]:

```
import sqlite3
import pandas as pd
!rm -f Test.db
conn = sqlite3.connect('Test.db')
curs = conn.cursor()
data = pd.read_csv('./ufo.csv')
data[['month','day','year','time']] = data['datetime'].str.replace('/', ' ').str.split(
' ', expand=True)
data = data.drop(columns='datetime')
data.to_sql('ufo', conn, index=False)
data.head(5)
```

```
/opt/conda/lib/python3.8/site-packages/IPython/core/interactiveshell.py:31
45: DtypeWarning: Columns (5,9) have mixed types.Specify dtype option on i
mport or set low_memory=False.
  has_raised = await self.run_ast_nodes(code_ast.body, cell_name,
/opt/conda/lib/python3.8/site-packages/pandas/core/generic.py:2602: UserWa
rning: The spaces in these column names will not be changed. In pandas ver
sions < 0.14, spaces were converted to underscores.
  sql.to_sql(
```

Out[2]:

	city	state	country	shape	duration (seconds)	duration (hours/min)	comments	date posted	
0	san marcos	tx	us	cylinder	2700	45 minutes	This event took place in early fall around 194...	4/27/2004	29.
1	lackland afb	tx	NaN	light	7200	1-2 hrs	1949 Lackland AFB&#44 TX. Lights racing acros...	12/16/2005	1
2	chester (uk/england)	NaN	gb	circle	20	20 seconds	Green/Orange circular disc over Chester&#44 En...	1/21/2008	
3	edna	tx	us	circle	20	1/2 hour	My older brother and twin sister were leaving ...	1/17/2004	28.
4	kaneohe	hi	us	light	900	15 minutes	AS a Marine 1st Lt. flying an FJ4B fighter/att...	1/22/2004	21.

1) In which state were the most UFO sightings reported?

Have your query return the top 5 results, sorted in descending order.

In [3]:

```
pd.read_sql("""SELECT state, count(state) as [Total_report_state]
                FROM ufo
                GROUP BY state
                ORDER BY [Total_report_state] DESC
                LIMIT 5;""", conn)
```

Out[3]:

	state	Total_report_state
0	ca	9655
1	wa	4268
2	fl	4200
3	tx	3677
4	ny	3219

2) In the state with the most UFO sightings, what is the most commonly reported "shape" of the UFO?

Have your query return the top 5 results, sorted in descending order

In [4]:

```
pd.read_sql("""SELECT shape, count(shape) as [Shape_Count]
                FROM ufo
                GROUP BY shape
                ORDER BY [Shape_Count] DESC
                LIMIT 5;""", conn)
```

Out[4]:

	shape	Shape_Count
0	light	16565
1	triangle	7865
2	circle	7608
3	fireball	6208
4	other	5649

3) In what year (at least in this data) were the most sightings reported?

Have your query return the top 5 results, sorted in descending order.

In [5]:

```
pd.read_sql("""SELECT year, count(year) as [sight_count]
                FROM ufo
                GROUP BY year
                ORDER BY [sight_count] DESC
                LIMIT 5;""", conn)
```

Out[5]:

	year	sight_count
0	2012	7357
1	2013	7037
2	2011	5107
3	2008	4820
4	2009	4541

4) For the year in which the most sightings were reported, what was the average duration of all of the sightings (in seconds) for the most commonly reported shape, in the state with the most sightings that year?

Have your query return a single record with: state, shape, year, number of sightings, and average sighting duration.

In [6]:

```
pd.read_sql("""SELECT year, state, shape, count(year) as sight_count, avg([duration (se
conds)]) AS average_duration
                FROM ufo
                WHERE year LIKE '2012'
                AND state LIKE 'ca'
                GROUP BY shape
                ORDER BY sight_count DESC
                LIMIT 1;""", conn)
```

Out[6]:

	year	state	shape	sight_count	average_duration
0	2012	ca	light	138	746.851449

5) How many sightings have been reported, by year, in Roswell, New Mexico?

Have your query return: city,state,year, and number of sightings. Sort the results in ascending order by year.

In [8]:

```
pd.read_sql("""SELECT year, city, state, count(year) as sight_count
               FROM ufo
               WHERE city LIKE 'Roswell'
               AND state LIKE 'nm'
               GROUP BY year
               ORDER BY year ASC;""", conn)
```

Out[8]:

	year	city	state	sight_count
0	1945	roswell	nm	1
1	1947	roswell	nm	1
2	1953	roswell	nm	1
3	1959	roswell	nm	1
4	1989	roswell	nm	1
5	1998	roswell	nm	2
6	2000	roswell	nm	1
7	2001	roswell	nm	2
8	2002	roswell	nm	3
9	2003	roswell	nm	1
10	2004	roswell	nm	3
11	2005	roswell	nm	1
12	2006	roswell	nm	1
13	2007	roswell	nm	1
14	2008	roswell	nm	3
15	2009	roswell	nm	2
16	2011	roswell	nm	1
17	2012	roswell	nm	1

In [ ]:

In [ ]:

In [ ]: