Homework 1

- D-532 Spring 23
- Dipak Bange

Part 1 Import tables - 25pts

Step 1. Create db named income

```
import sqlite3
import pandas as pd

conn = sqlite3.connect('income.db')
cur = conn.cursor()
```

Step 2. Import `income.csv`

```
read_inc = pd.read_csv(r'income.csv')

read_inc.to_sql('income', conn, if_exists='append', index = False)
```

Step 3. Print Column names. Note - cursor with query execution has been created. Add code to extract column names.

- All_weekly = average weekly income
- M_weekly = Male employees weekly income
- F_weekly = Female employees weekly income

```
cur.execute(''' select * from income ''')
column_list = []
col_names = cur.description
for row in col_names:
    print(row[0])
    column_list.append(row[0])
```

```
Occupation
Industry
All_workers
All_weekly
M_workers
M_weekly
F_workers
F_weekly
```

Step 4. Select records where weeklly income (All_weekly) is greater than 2000

```
cur.execute('''
select * from income
where All_weekly > 2000
''')

df = pd.DataFrame(cur.fetchall(), columns=column_list)
print(df)
```

```
              Occupation    Industry  All_workers  All_weekly  M_workers  M_weekly  \
     0  Chief executives  Management         1046        2041        763      2251

        F_workers  F_weekly
     0        283      1836
```

Step 5. Select occupations and average weekly income in Business industry and where Female employees weekly income is greater than male employees weekly income. Convert results into dataframe and name columns as Job and Income.

```
cur.execute('''
select Occupation as job, AVG(All_weekly) as income  from income
where M_weekly < F_weekly and Industry = 'Business'
group by Occupation
''')

df = pd.DataFrame(cur.fetchall(), columns=['JOB', 'INCOME'])
display(df)
```

```
     Wholesale and retail buyers, except farm products
```

Step 6. You need to print the list of industries used in your database. Like you would use set() in python, you will use `distinct` with your column name to remove duplicate names. Print the results.

```
rows = cur.execute('''
select distinct Industry from income
''')
for row in rows:
  print(row[0])
```

```
     Management
     Healthcare Professional
     Legal
     Engineering
     Computational
     Business
     Arts
     Science
     Education
     Protective Service
     Social Service
     Sales
     Office
     Service
     Transportation
     Production
     Groundskeeping
     Culinary
     Healthcare Support
     Agricultural
```

Step 7. Close cursor and connection

```
cur.close()
```

Part 2. Database Design - 25pts

Q1. If you design a database schema for income data, which attribute(s) would you choose as a primary key? And why?

Answer: I would choose columns Occupation and Industry as a primary key as they are unique in the data.

Q2. You decided to redisign income database and split into several tables. What changes would you implement?

Answer:

1. All data is incorporated in one table so we will have to split the data into different tables.
2. Create new tables for weekly, monthly and all incomes i.e 3 separate tables.
3. Create a new primary key for all tables so that we can have foreign keys to other tables if required and we can uniquely idetify a row.

Note for submission: convert to PDF and submit

✓ 0s completed at 8:51 AM