Session 12: Assignment 1

2.​ Problem Statement

Read the following data set:

https://archive.ics.uci.edu/ml/machine-learning-databases/adult/

Rename the columns as per the description from this file:

https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.names

Task:

Create a sql db from adult dataset and name it sqladb

* 1. Select 10 records from the adult sqladb

Code:

import random

import numpy as np

import pandas as pd

import sqlite3

from IPython.display import display

adult\_data\_df = pd.read\_csv('https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data')

display(adult\_data\_df.head(5))

adult\_data\_df.columns =["age","workclass","fnlwgt","education","education\_num","marital\_status","occupation","relationship","race","sex","capital\_gain","capital\_loss","hours\_per\_week","native\_country","income"]

display(adult\_data\_df.head(5))

import sqlalchemy

from sqlalchemy import create\_engine

engine = create\_engine('sqlite:///sqladb', echo=False)

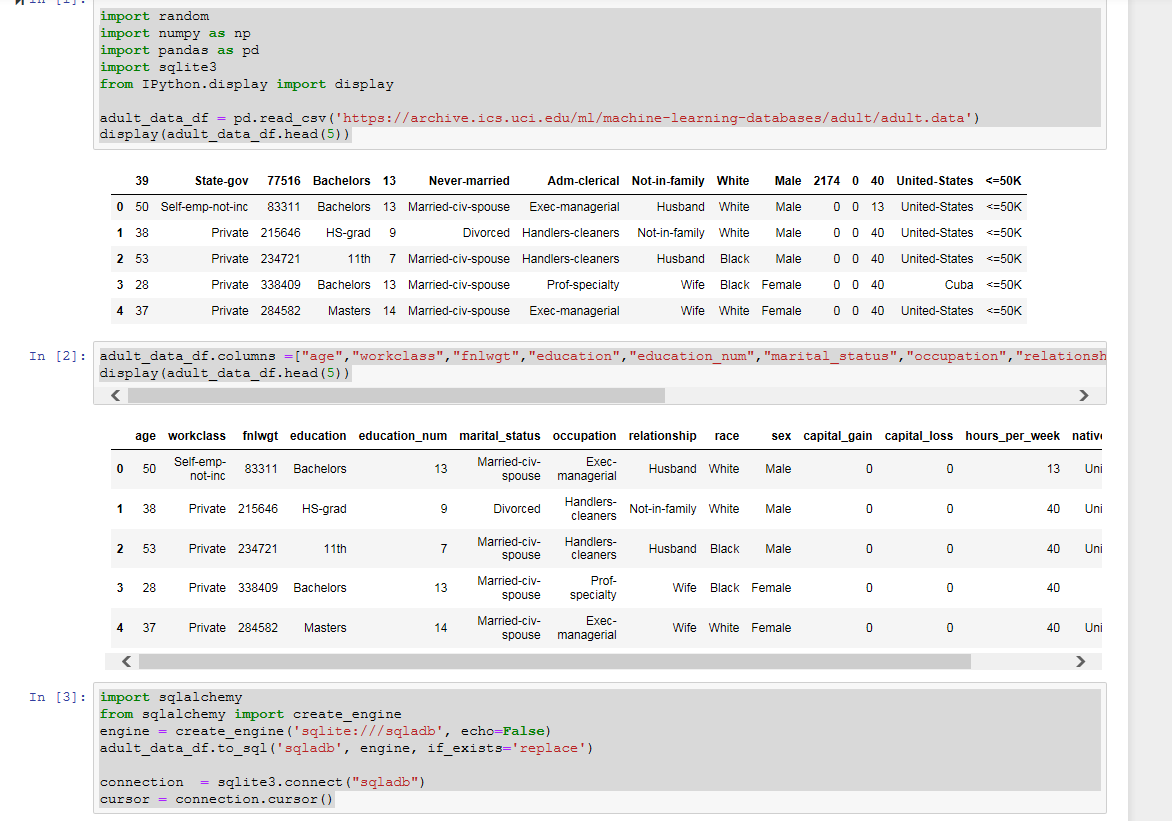
adult\_data\_df.to\_sql('sqladb', engine, if\_exists='replace')

connection = sqlite3.connect("sqladb")

cursor = connection.cursor()

print(pd.read\_sql\_query("SELECT \* FROM sqladb LIMIT 10", connection))

Output:



* 2) Show me the average hours per week of all men who are working in private sector

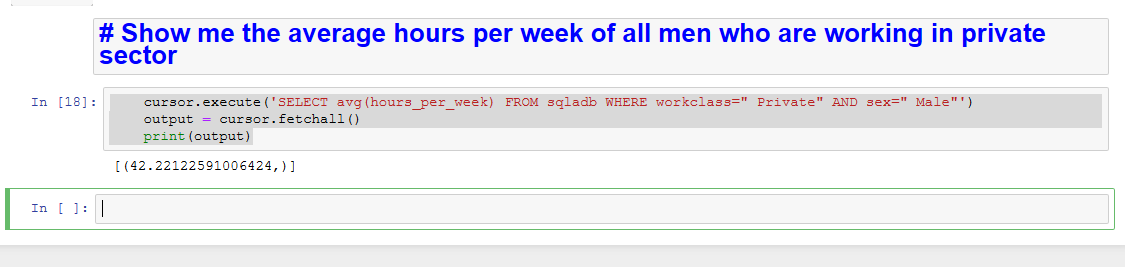
Code:

cursor.execute('SELECT avg(hours\_per\_week) FROM sqladb WHERE workclass=" Private" AND sex=" Male"')

output = cursor.fetchall()

print(output)

output:





* 3) Show me the frequency table for education, occupation and relationship, separately

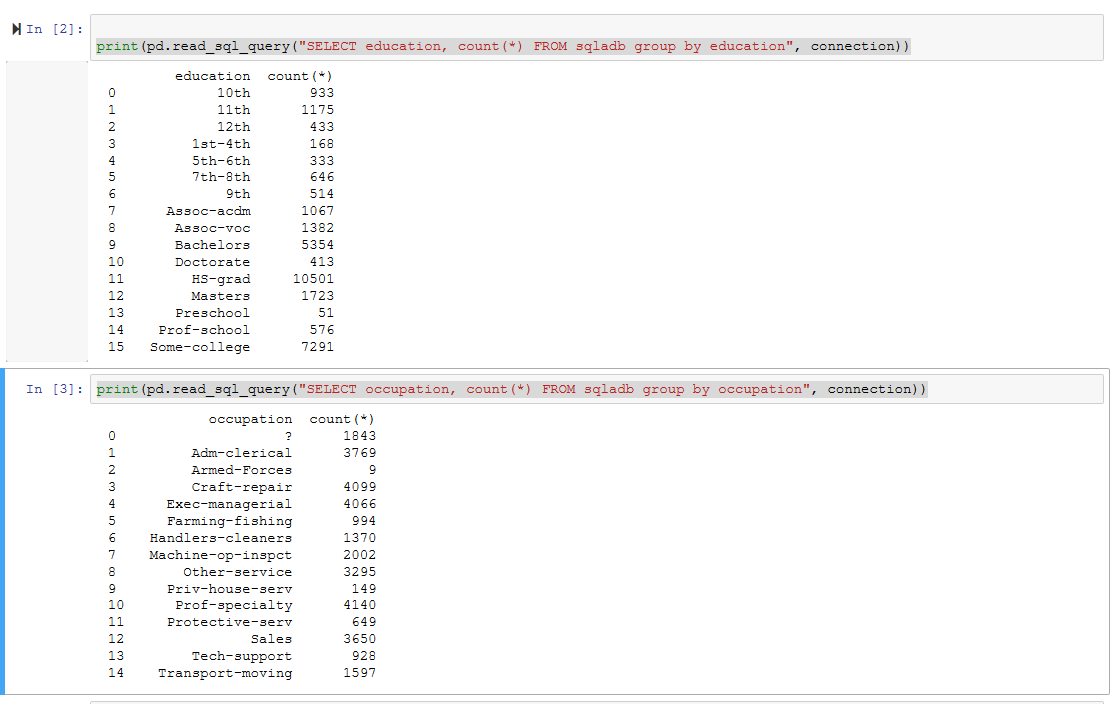
Code:

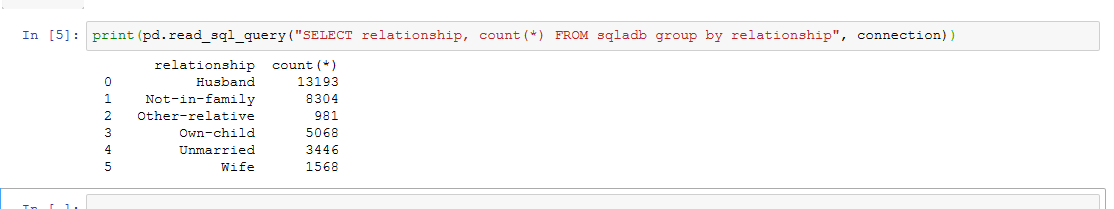
print(pd.read\_sql\_query("SELECT education, count(\*) FROM sqladb group by education", connection))

print(pd.read\_sql\_query("SELECT occupation, count(\*) FROM sqladb group by occupation", connection))

print(pd.read\_sql\_query("SELECT relationship, count(\*) FROM sqladb group by relationship", connection))

output:





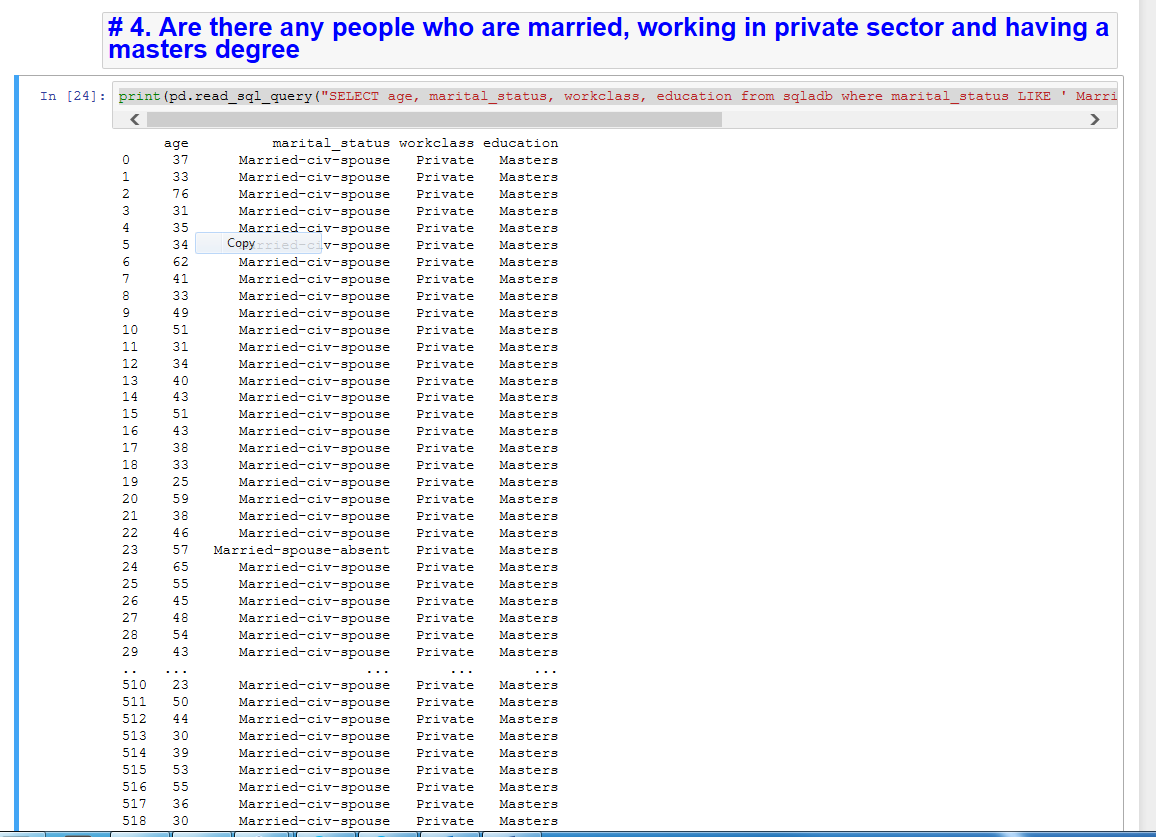
* 4. Are there any people who are married, working in private sector and having a masters

degree

Code:

print(pd.read\_sql\_query("SELECT age, marital\_status, workclass, education from sqladb where marital\_status LIKE ' Married%' and workclass = ' Private' and education = ' Masters' ", connection))

Output:

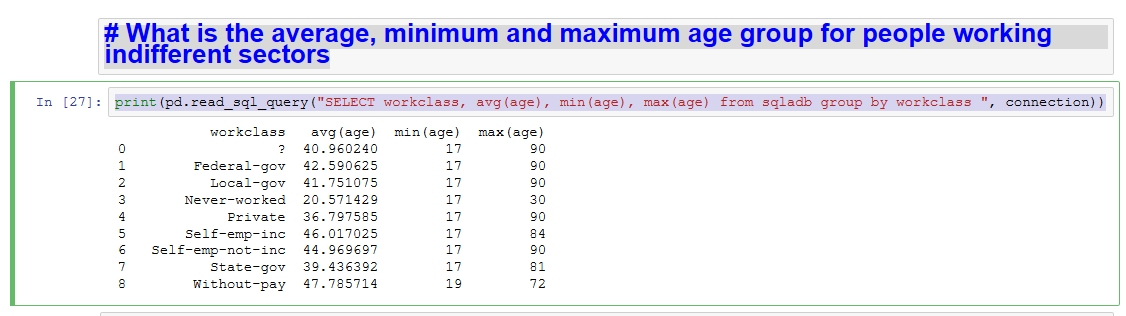


* 5 What is the average, minimum and maximum age group for people working in different sectors

Code:

print(pd.read\_sql\_query("SELECT workclass, avg(age), min(age), max(age) from sqladb group by workclass ", connection))

Output:

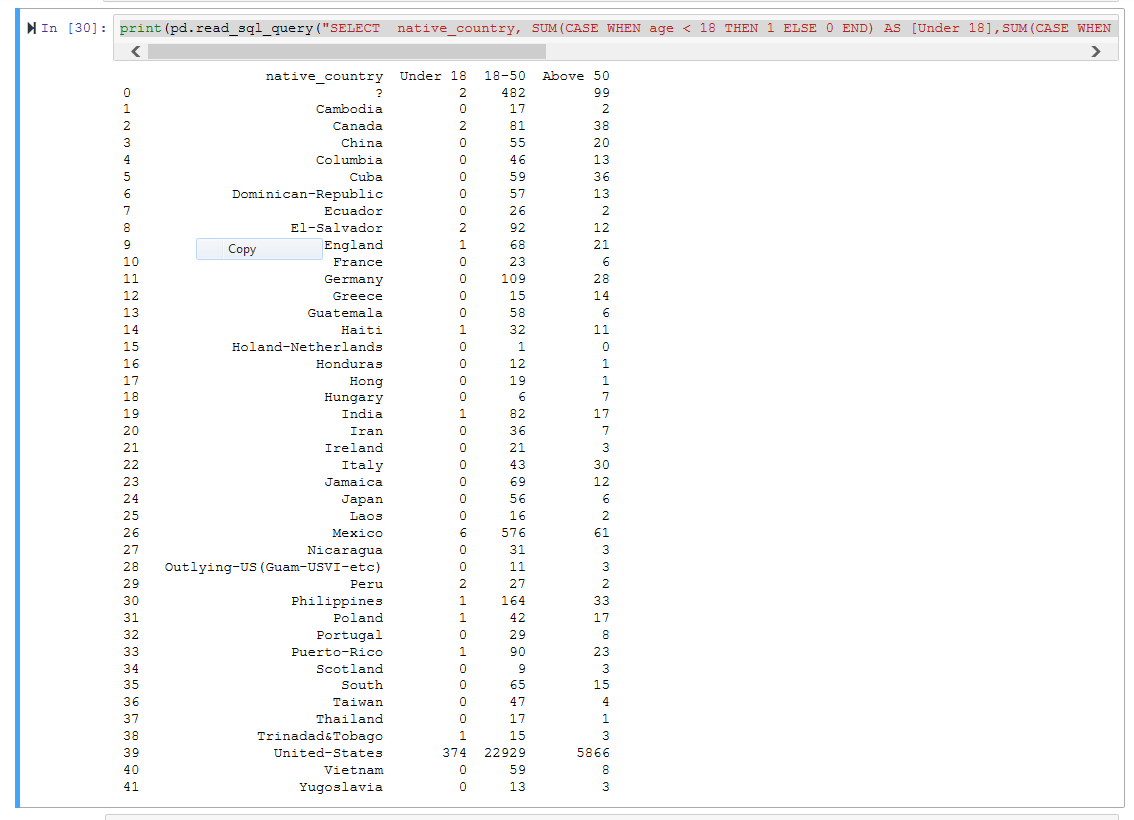


* 6) Calculate age distribution by country

Code:

print(pd.read\_sql\_query("SELECT native\_country, SUM(CASE WHEN age < 18 THEN 1 ELSE 0 END) AS [Under 18],SUM(CASE WHEN age BETWEEN 18 AND 50 THEN 1 ELSE 0 END) AS [18-50], SUM(CASE WHEN age > 50 THEN 1 ELSE 0 END) AS [Above 50] from sqladb group by native\_country ", connection))

Output:



* 7) Compute a new column as 'Net-Capital-Gain' from the two columns 'capital-gain' and

'capital-loss'

print(pd.read\_sql\_query("SELECT native\_country, SUM(CASE WHEN age < 18 THEN 1 ELSE 0 END) AS [Under 18],SUM(CASE WHEN age BETWEEN 18 AND 50 THEN 1 ELSE 0 END) AS [18-50], SUM(CASE WHEN age > 50 THEN 1 ELSE 0 END) AS [Above 50] from sqladb group by native\_country ", connection))

Output:

