datascience_assignment_13.1

July 14, 2018

0.1 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

0.2 Steps

- Import numpy, pandas, sqllite3, pandasql
- Read the csv using the URL provided using pandas method read_csv nad store into pandas dataframe sqladb which will act as the sql db
- Define the column for sqladb as provided in URL for description of csv file
- Define panda sqldf name it as pysqldf

```
In [2]: import numpy as np
        import pandas as pd
        from pandas import DataFrame
        import sqlite3 as db
        from pandasql import sqldf
```

```
In [3]: # Read the csv using the URL provided using pandas method read_csv nad store into pand
        # which will act as the sql db
        sqladb = pd.read_csv('https://archive.ics.uci.edu/ml/machine-learning-databases/adult/a

        # Define the column for sqladb as provided in URL for description of csv file
        sqladb.columns =['age', 'workclass', 'fnlwgt', 'education', 'education_num', 'marital_s
```

```
In [4]: # Show first few records of sqladb using head method
        sqladb.head()
```

```
Out[4]:    age          workclass  fnlwgt   education  education_num  \
        0   50   Self-emp-not-inc   83311   Bachelors             13
        1   38            Private  215646     HS-grad              9
        2   53            Private  234721        11th              7
        3   28            Private  338409   Bachelors             13
        4   37            Private  284582     Masters             14

                marital_status          occupation    relationship    race      sex  \
```

```
0   Married-civ-spouse     Exec-managerial         Husband   White     Male
1             Divorced   Handlers-cleaners   Not-in-family   White     Male
2   Married-civ-spouse   Handlers-cleaners         Husband   Black     Male
3   Married-civ-spouse      Prof-specialty            Wife   Black   Female
4   Married-civ-spouse     Exec-managerial            Wife   White   Female

   capital_gain  capital_loss  hours_per_week  native_country  income
0             0             0              13   United-States   <=50K
1             0             0              40   United-States   <=50K
2             0             0              40   United-States   <=50K
3             0             0              40            Cuba   <=50K
4             0             0              40   United-States   <=50K
```

```
In [5]: # Define panda sqldf name it as pysqldf
        pysqldf = lambda q: sqldf(q, globals())
```

0.3 Select 10 records from the adult sqladb

0.4 Steps:

- Select record from sqladb and limit to 10

```
In [6]: pysqldf("SELECT * FROM sqladb LIMIT 10;")
```

```
Out[6]:    age          workclass  fnlwgt      education  education_num  \
        0   50   Self-emp-not-inc   83311      Bachelors             13
        1   38            Private  215646        HS-grad              9
        2   53            Private  234721           11th              7
        3   28            Private  338409      Bachelors             13
        4   37            Private  284582        Masters             14
        5   49            Private  160187            9th              5
        6   52   Self-emp-not-inc  209642        HS-grad              9
        7   31            Private   45781        Masters             14
        8   42            Private  159449      Bachelors             13
        9   37            Private  280464   Some-college             10

                marital_status          occupation    relationship    race  \
        0      Married-civ-spouse     Exec-managerial         Husband   White
        1                Divorced   Handlers-cleaners   Not-in-family   White
        2      Married-civ-spouse   Handlers-cleaners         Husband   Black
        3      Married-civ-spouse      Prof-specialty            Wife   Black
        4      Married-civ-spouse     Exec-managerial            Wife   White
        5   Married-spouse-absent       Other-service   Not-in-family   Black
        6      Married-civ-spouse     Exec-managerial         Husband   White
        7           Never-married      Prof-specialty   Not-in-family   White
        8      Married-civ-spouse     Exec-managerial         Husband   White
        9      Married-civ-spouse     Exec-managerial         Husband   Black

              sex  capital_gain  capital_loss  hours_per_week  native_country  income
```

0	Male	0	0	13	United-States	<=50K
1	Male	0	0	40	United-States	<=50K
2	Male	0	0	40	United-States	<=50K
3	Female	0	0	40	Cuba	<=50K
4	Female	0	0	40	United-States	<=50K
5	Female	0	0	16	Jamaica	<=50K
6	Male	0	0	45	United-States	>50K
7	Female	14084	0	50	United-States	>50K
8	Male	5178	0	40	United-States	>50K
9	Male	0	0	80	United-States	>50K

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

- Filter sqladb and get those records for which workclass column is 'Private' and sex column is 'Male'
- Use avg aggregate function on hours_per_week to get average of hours worked

```
In [7]: #Filter sqladb and get those records for which workclass column is 'Private' and sex c
        #Use avg aggregate function on hours_per_week to get average of hours worked
        pysqldf("SELECT avg(hours_per_week) AS 'Average Hours Per Week' FROM sqladb  WHERE TRI
```

Out[7]:

	Average Hours Per Week
0	42.221226

0.6 3. Show me the frequency table for education, occupation and relationship, separately

0.7 Steps

- Select count of records from sqlsdb as frequency group by education column
- Select count of records from sqlsdb as fequency group by occupation column
- Select count of records from sqlsdb as frequency group by relationship column

```
In [8]: # Select count of records as frequency from sqlsdb group by education column
        pysqldf("SELECT education, count(*) AS frequency FROM sqladb GROUP BY education")
```

Out[8]:

	education	frequency
0	10th	933
1	11th	1175
2	12th	433
3	1st-4th	168
4	5th-6th	333
5	7th-8th	646
6	9th	514
7	Assoc-acdm	1067
8	Assoc-voc	1382
9	Bachelors	5354
10	Doctorate	413

```
11        HS-grad      10501
12        Masters       1723
13      Preschool         51
14    Prof-school        576
15   Some-college       7291
```

```
In [9]: # Select count of records from sqlsdb as fequency group by occupation column
        pysqldf("SELECT occupation, count(*) AS frequency FROM sqladb GROUP BY occupation")
```

```
Out[9]:             occupation  frequency
0                            ?       1843
1          Adm-clerical              3769
2          Armed-Forces                 9
3          Craft-repair              4099
4       Exec-managerial              4066
5       Farming-fishing               994
6     Handlers-cleaners              1370
7     Machine-op-inspct              2002
8         Other-service              3295
9       Priv-house-serv               149
10       Prof-specialty              4140
11      Protective-serv               649
12                Sales              3650
13         Tech-support               928
14    Transport-moving              1597
```

```
In [10]: # Select count of records from sqlsdb as frequency group by relationship column
         pysqldf("SELECT relationship, count(*) AS frequency FROM sqladb GROUP BY relationship"
```

```
Out[10]:       relationship  frequency
0           Husband      13193
1     Not-in-family       8304
2    Other-relative        981
3         Own-child       5068
4         Unmarried       3446
5              Wife       1568
```

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

0.9 Steps:

- Select count of records as fequency from sqladb where marital_status column starts with 'Married' and workclass column is 'Private' and education column is 'Masters'

```
In [11]: #  Select count of records as fequency from sqladb where marital_status column starts
         # and workclass column is 'Private' and education column is 'Masters'
         pysqldf("SELECT count(*) AS frequency FROM sqladb WHERE marital_status LIKE 'Married%
```

```
Out[11]:    frequency
0           0
```

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

0.11 Steps:

- Select occupation, AVG on age column to get average age, MAX on age column to get Maximum Age, MIN on age column to get Minimum Age from sqladb group by occupation

```
In [12]: #Select occupation, AVG on age column to get average age, MAX on age column to get Ma
         # MIN on age column to get Minimum Age from sqladb group by occupation
         pysqldf("SELECT occupation, AVG(age) 'Average Age', MAX(age) AS 'Maximum Age', MIN(ag
```

Out[12]:

	occupation	Average Age	Maximum Age	Minimum Age
0	?	40.882800	90	17
1	Adm-clerical	36.963916	90	17
2	Armed-Forces	30.222222	46	23
3	Craft-repair	39.031471	90	17
4	Exec-managerial	42.169208	90	17
5	Farming-fishing	41.211268	90	17
6	Handlers-cleaners	32.165693	90	17
7	Machine-op-inspct	37.715285	90	17
8	Other-service	34.949621	90	17
9	Priv-house-serv	41.724832	81	17
10	Prof-specialty	40.517633	90	17
11	Protective-serv	38.953775	90	17
12	Sales	37.353973	90	17
13	Tech-support	37.022629	73	17
14	Transport-moving	40.197871	90	17

0.12 6. Calculate age distribution by country

0.13 Steps:

- Select native_country, age, count from sqladb group by native_country, age

```
In [13]: # Select native_country, age, count from sqladb group by native_country, age
         pysqldf("SELECT native_country,age, count(*) AS frequency FROM sqladb GROUP BY native
```

Out[13]:

	native_country	age	frequency
0	?	17	2
1	?	18	8
2	?	19	5
3	?	20	10
4	?	21	11
5	?	22	12
6	?	23	6
7	?	24	14
8	?	25	11
9	?	26	18
10	?	27	15

11	?	28	19
12	?	29	12
13	?	30	19
14	?	31	18
15	?	32	17
16	?	33	13
17	?	34	24
18	?	35	18
19	?	36	23
20	?	37	22
21	?	38	20
22	?	39	19
23	?	40	12
24	?	41	22
25	?	42	24
26	?	43	14
27	?	44	10
28	?	45	17
29	?	46	15
...	...	...	...
1251	Vietnam	37	2
1252	Vietnam	38	1
1253	Vietnam	40	1
1254	Vietnam	41	1
1255	Vietnam	43	2
1256	Vietnam	44	3
1257	Vietnam	45	3
1258	Vietnam	46	1
1259	Vietnam	48	1
1260	Vietnam	50	1
1261	Vietnam	51	1
1262	Vietnam	52	1
1263	Vietnam	53	1
1264	Vietnam	54	1
1265	Vietnam	63	1
1266	Vietnam	70	1
1267	Vietnam	73	2
1268	Yugoslavia	20	1
1269	Yugoslavia	22	1
1270	Yugoslavia	25	1
1271	Yugoslavia	29	1
1272	Yugoslavia	31	1
1273	Yugoslavia	35	2
1274	Yugoslavia	36	1
1275	Yugoslavia	40	1
1276	Yugoslavia	41	2
1277	Yugoslavia	43	1
1278	Yugoslavia	45	1

```
1279     Yugoslavia   56          2
1280     Yugoslavia   66          1

[1281 rows x 3 columns]
```

0.14 7. Compute a new column as 'Net-Capital-Gain' from the two columns 'capital-gain' and 'capital-loss'

0.15 Steps:

- Select a new column Net-Capital-Gain by subtracting capital_loss from capital_gain

```
In [14]: # Select a new column Net-Capital-Gain by subtracting capital_loss from capital_gain
         pysqldf("SELECT (capital_gain - capital_loss)  AS 'Net-Capital-Gain',* FROM sqladb")
```

Out[14]:

	Net-Capital-Gain	age	workclass	fnlwgt	education
0	0	50	Self-emp-not-inc	83311	Bachelors
1	0	38	Private	215646	HS-grad
2	0	53	Private	234721	11th
3	0	28	Private	338409	Bachelors
4	0	37	Private	284582	Masters
5	0	49	Private	160187	9th
6	0	52	Self-emp-not-inc	209642	HS-grad
7	14084	31	Private	45781	Masters
8	5178	42	Private	159449	Bachelors
9	0	37	Private	280464	Some-college
10	0	30	State-gov	141297	Bachelors
11	0	23	Private	122272	Bachelors
12	0	32	Private	205019	Assoc-acdm
13	0	40	Private	121772	Assoc-voc
14	0	34	Private	245487	7th-8th
15	0	25	Self-emp-not-inc	176756	HS-grad
16	0	32	Private	186824	HS-grad
17	0	38	Private	28887	11th
18	0	43	Self-emp-not-inc	292175	Masters
19	0	40	Private	193524	Doctorate
20	0	54	Private	302146	HS-grad
21	0	35	Federal-gov	76845	9th
22	-2042	43	Private	117037	11th
23	0	59	Private	109015	HS-grad
24	0	56	Local-gov	216851	Bachelors
25	0	19	Private	168294	HS-grad
26	0	54	?	180211	Some-college
27	0	39	Private	367260	HS-grad
28	0	49	Private	193366	HS-grad
29	0	23	Local-gov	190709	Assoc-acdm
...	...	...	...	...	...
32530	0	30	?	33811	Bachelors
32531	0	34	Private	204461	Doctorate

32532	0	54	Private	337992	Bachelors
32533	0	37	Private	179137	Some-college
32534	0	22	Private	325033	12th
32535	0	34	Private	160216	Bachelors
32536	0	30	Private	345898	HS-grad
32537	15020	38	Private	139180	Bachelors
32538	0	71	?	287372	Doctorate
32539	0	45	State-gov	252208	HS-grad
32540	0	41	?	202822	HS-grad
32541	0	72	?	129912	HS-grad
32542	0	45	Local-gov	119199	Assoc-acdm
32543	0	31	Private	199655	Masters
32544	0	39	Local-gov	111499	Assoc-acdm
32545	0	37	Private	198216	Assoc-acdm
32546	0	43	Private	260761	HS-grad
32547	1086	65	Self-emp-not-inc	99359	Prof-school
32548	0	43	State-gov	255835	Some-college
32549	0	43	Self-emp-not-inc	27242	Some-college
32550	0	32	Private	34066	10th
32551	0	43	Private	84661	Assoc-voc
32552	0	32	Private	116138	Masters
32553	0	53	Private	321865	Masters
32554	0	22	Private	310152	Some-college
32555	0	27	Private	257302	Assoc-acdm
32556	0	40	Private	154374	HS-grad
32557	0	58	Private	151910	HS-grad
32558	0	22	Private	201490	HS-grad
32559	15024	52	Self-emp-inc	287927	HS-grad

	education_num	marital_status	occupation	\
0	13	Married-civ-spouse	Exec-managerial	
1	9	Divorced	Handlers-cleaners	
2	7	Married-civ-spouse	Handlers-cleaners	
3	13	Married-civ-spouse	Prof-specialty	
4	14	Married-civ-spouse	Exec-managerial	
5	5	Married-spouse-absent	Other-service	
6	9	Married-civ-spouse	Exec-managerial	
7	14	Never-married	Prof-specialty	
8	13	Married-civ-spouse	Exec-managerial	
9	10	Married-civ-spouse	Exec-managerial	
10	13	Married-civ-spouse	Prof-specialty	
11	13	Never-married	Adm-clerical	
12	12	Never-married	Sales	
13	11	Married-civ-spouse	Craft-repair	
14	4	Married-civ-spouse	Transport-moving	
15	9	Never-married	Farming-fishing	
16	9	Never-married	Machine-op-inspct	
17	7	Married-civ-spouse	Sales	

```
18                14                Divorced     Exec-managerial
19                16      Married-civ-spouse      Prof-specialty
20                 9           Separated       Other-service
21                 5      Married-civ-spouse     Farming-fishing
22                 7      Married-civ-spouse    Transport-moving
23                 9                Divorced        Tech-support
24                13      Married-civ-spouse        Tech-support
25                 9           Never-married        Craft-repair
26                10      Married-civ-spouse                   ?
27                 9                Divorced     Exec-managerial
28                 9      Married-civ-spouse        Craft-repair
29                12           Never-married     Protective-serv
...              ...                     ...                 ...
32530             13           Never-married                   ?
32531             16      Married-civ-spouse      Prof-specialty
32532             13      Married-civ-spouse     Exec-managerial
32533             10                Divorced        Adm-clerical
32534              8           Never-married     Protective-serv
32535             13           Never-married     Exec-managerial
32536              9           Never-married        Craft-repair
32537             13                Divorced      Prof-specialty
32538             16      Married-civ-spouse                   ?
32539              9           Separated        Adm-clerical
32540              9           Separated                   ?
32541              9      Married-civ-spouse                   ?
32542             12                Divorced      Prof-specialty
32543             14                Divorced       Other-service
32544             12      Married-civ-spouse        Adm-clerical
32545             12                Divorced        Tech-support
32546              9      Married-civ-spouse   Machine-op-inspct
32547             15           Never-married      Prof-specialty
32548             10                Divorced        Adm-clerical
32549             10      Married-civ-spouse        Craft-repair
32550              6      Married-civ-spouse   Handlers-cleaners
32551             11      Married-civ-spouse               Sales
32552             14           Never-married        Tech-support
32553             14      Married-civ-spouse     Exec-managerial
32554             10           Never-married     Protective-serv
32555             12      Married-civ-spouse        Tech-support
32556              9      Married-civ-spouse   Machine-op-inspct
32557              9                 Widowed        Adm-clerical
32558              9           Never-married        Adm-clerical
32559              9      Married-civ-spouse     Exec-managerial

          relationship                race      sex  capital_gain  \
0              Husband               White     Male             0
1        Not-in-family               White     Male             0
2              Husband               Black     Male             0
```

3	Wife	Black	Female	0
4	Wife	White	Female	0
5	Not-in-family	Black	Female	0
6	Husband	White	Male	0
7	Not-in-family	White	Female	14084
8	Husband	White	Male	5178
9	Husband	Black	Male	0
10	Husband	Asian-Pac-Islander	Male	0
11	Own-child	White	Female	0
12	Not-in-family	Black	Male	0
13	Husband	Asian-Pac-Islander	Male	0
14	Husband	Amer-Indian-Eskimo	Male	0
15	Own-child	White	Male	0
16	Unmarried	White	Male	0
17	Husband	White	Male	0
18	Unmarried	White	Female	0
19	Husband	White	Male	0
20	Unmarried	Black	Female	0
21	Husband	Black	Male	0
22	Husband	White	Male	0
23	Unmarried	White	Female	0
24	Husband	White	Male	0
25	Own-child	White	Male	0
26	Husband	Asian-Pac-Islander	Male	0
27	Not-in-family	White	Male	0
28	Husband	White	Male	0
29	Not-in-family	White	Male	0
...	...	...	...	...
32530	Not-in-family	Asian-Pac-Islander	Female	0
32531	Husband	White	Male	0
32532	Husband	Asian-Pac-Islander	Male	0
32533	Unmarried	White	Female	0
32534	Own-child	Black	Male	0
32535	Not-in-family	White	Female	0
32536	Not-in-family	Black	Male	0
32537	Unmarried	Black	Female	15020
32538	Husband	White	Male	0
32539	Own-child	White	Female	0
32540	Not-in-family	Black	Female	0
32541	Husband	White	Male	0
32542	Unmarried	White	Female	0
32543	Not-in-family	Other	Female	0
32544	Wife	White	Female	0
32545	Not-in-family	White	Female	0
32546	Husband	White	Male	0
32547	Not-in-family	White	Male	1086
32548	Other-relative	White	Female	0
32549	Husband	White	Male	0

32550	Husband	Amer-Indian-Eskimo	Male	0	
32551	Husband	White	Male	0	
32552	Not-in-family	Asian-Pac-Islander	Male	0	
32553	Husband	White	Male	0	
32554	Not-in-family	White	Male	0	
32555	Wife	White	Female	0	
32556	Husband	White	Male	0	
32557	Unmarried	White	Female	0	
32558	Own-child	White	Male	0	
32559	Wife	White	Female	15024	

	capital_loss	hours_per_week	native_country	income
0	0	13	United-States	<=50K
1	0	40	United-States	<=50K
2	0	40	United-States	<=50K
3	0	40	Cuba	<=50K
4	0	40	United-States	<=50K
5	0	16	Jamaica	<=50K
6	0	45	United-States	>50K
7	0	50	United-States	>50K
8	0	40	United-States	>50K
9	0	80	United-States	>50K
10	0	40	India	>50K
11	0	30	United-States	<=50K
12	0	50	United-States	<=50K
13	0	40	?	>50K
14	0	45	Mexico	<=50K
15	0	35	United-States	<=50K
16	0	40	United-States	<=50K
17	0	50	United-States	<=50K
18	0	45	United-States	>50K
19	0	60	United-States	>50K
20	0	20	United-States	<=50K
21	0	40	United-States	<=50K
22	2042	40	United-States	<=50K
23	0	40	United-States	<=50K
24	0	40	United-States	>50K
25	0	40	United-States	<=50K
26	0	60	South	>50K
27	0	80	United-States	<=50K
28	0	40	United-States	<=50K
29	0	52	United-States	<=50K
...	...	...	...	...
32530	0	99	United-States	<=50K
32531	0	60	United-States	>50K
32532	0	50	Japan	>50K
32533	0	39	United-States	<=50K
32534	0	35	United-States	<=50K

```
32535             0             55   United-States    >50K
32536             0             46   United-States   <=50K
32537             0             45   United-States    >50K
32538             0             10   United-States    >50K
32539             0             40   United-States   <=50K
32540             0             32   United-States   <=50K
32541             0             25   United-States   <=50K
32542             0             48   United-States   <=50K
32543             0             30   United-States   <=50K
32544             0             20   United-States    >50K
32545             0             40   United-States   <=50K
32546             0             40          Mexico   <=50K
32547             0             60   United-States   <=50K
32548             0             40   United-States   <=50K
32549             0             50   United-States   <=50K
32550             0             40   United-States   <=50K
32551             0             45   United-States   <=50K
32552             0             11          Taiwan   <=50K
32553             0             40   United-States    >50K
32554             0             40   United-States   <=50K
32555             0             38   United-States   <=50K
32556             0             40   United-States    >50K
32557             0             40   United-States   <=50K
32558             0             20   United-States   <=50K
32559             0             40   United-States    >50K

[32560 rows x 16 columns]
```