Data Warehouse & Data Mining Lab

Submitted by

```
Gyanendra Shukla
CSE 1
191112040
```

Assignment Problem

Given the CSV file with field descriptions, convert the data to a structured MySQL table.

Approach Used

I've used SQL to create and import data in an MySQL table. In MySQL there is a load data method that allows us to load some data with filtering on it. First of all, I created a new table with the given schema and observing the data. For field separation, I separated them with `,` and for new entry separation with `'\n'` . I have finally displayed first 10 entries of the dataset.

Code

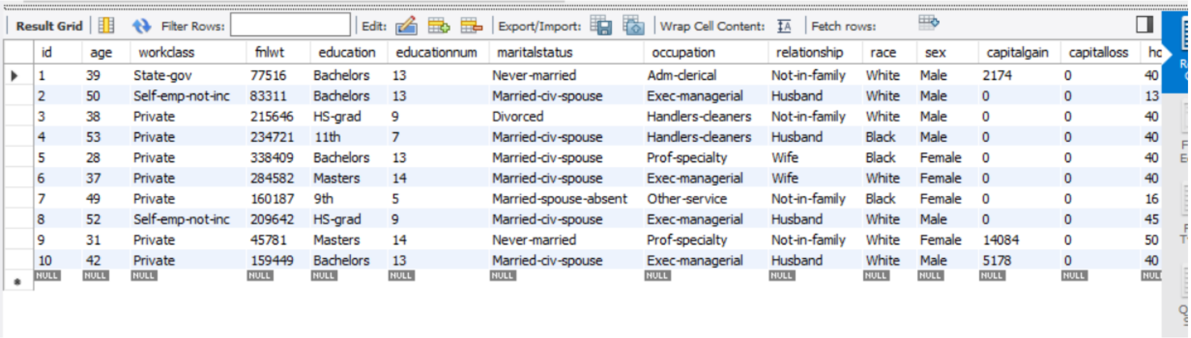
```

-- Creating a MySQL table according to the given schema.
-- The fields that were continuos were marked as float, and other
-- fields with fixed values as VARCHARS.

CREATE TABLE socialinfo (
    id INT NOT NULL auto_increment,
    age float,
    workclass VARCHAR(100),
    fnlwt float,
    education varchar(100),
    educationnum float,
    maritalstatus varchar(100),
    occupation varchar(100),
    relationship varchar(100),
    race varchar(100),
    sex varchar(100),
    capitalgain float,
    capitalloss float,
    hoursperweek float,
    nativecountry varchar(100),
    salary varchar(100),
    primary key (id)
);


```

```
-- After creating the table, we load the local infile and store it in our table
load data local infile 'C:/ProgramData/MySQL/MySQL Server 8.0/Uploads/Sample.txt'
into table socialinfo
fields terminated by ','
lines terminated by '\n'
(age, workclass, fnlwt, education, educationnum, maritalstatus, occupation, relationship,
race, sex, capitalgain, capitalloss, hoursperweek, salary);

-- showing the top 10 results
select * from socialinfo limit 10;

-- showing number of entries in the dataset
select count(*) from socialinfo;
```

Description of Code

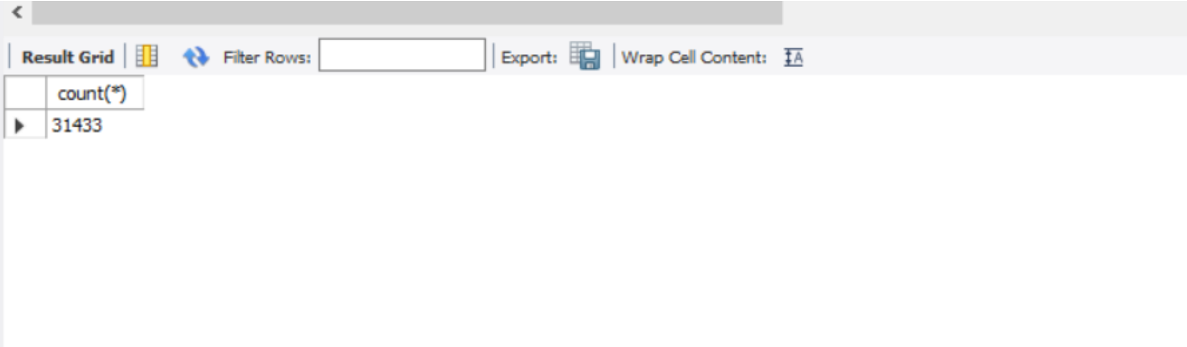
We can create a new MySQL table using CREATE TABLE method. I've created the `socialinfo` table with appropriate schema. I've also added an extra `id` field that is auto incrementing and set that as the primary key. Then, I loaded the `Sample.csv` file and inserted it in the table. For separating the data we're using `,` for fields and `\n` for entries.

Output Snapshots



	id	age	workclass	fnlwt	education	educationnum	maritalstatus	occupation	relationship	race	sex	capitalgain	capitalloss	hc
▶	1	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40
	2	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13
	3	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40
	4	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40
	5	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40
	6	37	Private	284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	40
	7	49	Private	160187	9th	5	Married-spouse-absent	Other-service	Not-in-family	Black	Female	0	0	16
	8	52	Self-emp-not-inc	209642	HS-grad	9	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	45
	9	31	Private	45781	Masters	14	Never-married	Prof-specialty	Not-in-family	White	Female	14084	0	50
	10	42	Private	159449	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	5178	0	40
*	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NUL

Fig: Top 10 entries of the dataset



	count(*)
▶	31433

Fig: Total entries in dataset

1. assignment problem, 2. approach used, 3. code with proper comments, 4. description of code, 5. output snapshots.