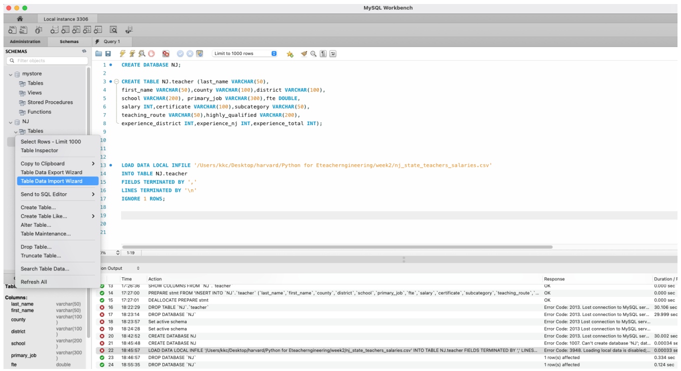
CSCI S-101 Python for Engineering

Name: Qi Chu

Data Uploading in SQL and Python

For MySQL uploading, I used mysql.connector in Jupyter notebook. It can also be done in MySQL WorkBench through Table Data Import Wizard:



```
CREATE DATABASE NJ;

CREATE TABLE NJ.teacher (last_name VARCHAR(50),
first_name VARCHAR(50),county VARCHAR(100),district VARCHAR(100),
school VARCHAR(200), primary_job VARCHAR(300),fte DOUBLE,
salary INT,certificate VARCHAR(100),subcategory VARCHAR(50),
teaching_route VARCHAR(50),highly_qualified VARCHAR(200),
experience_district INT,experience_nj INT,experience_total INT);




LOAD DATA LOCAL INFILE '/Users/kkc/Desktop/harvard/Python for Eteacherngineering/week2/nj_state_teachers_salaries.csv'
INTO TABLE NJ.teacher
FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n'
IGNORE 1 ROWS;
```

```
    Time      Action                                                                                                                                                  Response                                           Duration / F
13  17:26:36  SHOW COLUMNS FROM `NJ`.`teacher`                                                                                                                        OK                                                 0.000 sec
14  17:27:00  PREPARE stmt FROM 'INSERT INTO `NJ`.`teacher` (`last_name`,`first_name`,`county`,`district`,`school`,`primary_job`,`fte`,`salary`,`certificate`,`subcategory`,`teaching_route`,`...  OK   0.000 sec
15  17:27:01  DEALLOCATE PREPARE stmt                                                                                                                                 OK                                                 0.000 sec
16  18:22:29  DROP TABLE `NJ`.`teacher`                                                                                                                               Error Code: 2013. Lost connection to MySQL ser...  30.106 sec
17  18:23:14  DROP DATABASE `NJ`                                                                                                                                      Error Code: 2013. Lost connection to MySQL ser...  29.999 sec
18  18:23:57  Set active schema                                                                                                                                       Error Code: 2013 Lost connection to MySQL serv...
19  18:24:28  Set active schema                                                                                                                                       Error Code: 2013 Lost connection to MySQL serv...
20  18:42:52  CREATE DATABASE NJ                                                                                                                                      Error Code: 2013. Lost connection to MySQL ser...  30.002 sec
21  18:45:48  CREATE DATABASE NJ                                                                                                                                      Error Code: 1007. Can't create database 'NJ'; dat...  0.00034 se
22  18:45:57  LOAD DATA LOCAL INFILE '/Users/kkc/Desktop/harvard/Python for Eteacherngineering/week2/nj_state_teachers_salaries.csv' INTO TABLE NJ.teacher FIELDS TERMINATED BY ',' LINES...  Error Code: 3948. Loading local data is disabled;...  0.00033 se
23  18:46:57  DROP DATABASE `NJ`                                                                                                                                      1 row(s) affected                                  0.334 sec
24  18:55:35  DROP DATABASE `NJ`                                                                                                                                      1 row(s) affected                                  0.124 sec
```

Or in terminal, using the same codes.

Python uploading was just using pd.read_csv.

Data Cleaning

1. I first check the data types, and observe most of the variables are identified as objects, so will need to assign the new, appropriate data types to them.
2. Rows with any NaN value for any variable are dropped, even some data entries have only one NaN value, it is not possible to use the group average or previous/next available value since every teacher's information is independent. The NaN values do not add value to the data analysis.
3. The variables that contain text values, e.g. last_name, primary_job, are converted to strings, and are trimmed with leading and trailing spaces.
4. For the variables that are supposed to be numeric, I first check if there's any non-numeric values hiding by forcing them into float values, any errors will be replaced with NaN values.
5. Step 5 identified another 10 non-numeric / NaN values, so I drop the rows with NAs again to have the final clean data. Converting the numeric variables into float is only to be safer, but I also check if separating by float and integer is necessary. I re-convert these float variables into strings, and strip out the last character, if it's 0, that original value is an integer, otherwise it's a real float. By doing so, I figure out only 'fte' is a float while the rest variables are integers. This validates the SQL inputs too.
6. Biggest Challenge: Python identifies most variables as objects, then data reading, comparing, trimming are not possible, will need to convert to appropriate data types. There are also hidden non-numeric values that need to be cleaned.