Assignment Instructions:

BusinessA has a dataset used for analysis of job markets. It is available through a REST API. They partnered with FirmB who also has market data they wish to share. BusinessA wants to do an analysis on how their data compares. The data from FirmB is stored in a remote MySQL database. All their data is in the table “businessA_records_v2”.

The scheme for the table is as follows, and the column names are one-to-one with the keys in the JSON from the BusinessA API:

Field	Type	Null	Key	Default	Additional Notes
id	int(11)	No	Primary	NULL	auto-increment
name	varchar(64)	No		NULL	
address	varchar(128)	No		NULL	
birthdate	varchar(10)	No		NULL	
sex	varchar(1)	No		NULL	
job	varchar(64)	No		NULL	
company	varchar(64)	No		NULL	
emd5	varchar(32)	No	Mull	NULL	

Answer:
> How many users are in both datasets?
> How many users are only found in each respective dataset?
> For users found in both datasets, what percent have different job titles?
> Output data for users in the intersection in a 3 column CSV with columns (emd5, BusinessA_Job/Company_JSON_list, FirmB_Job/Company_JSON_list)
> Output a CREATE TABLE statement for MySQL that could use the CSV data file

Notes:
> emd5 values are unique
> data cannot be read fully into memory from either dataset at one time. It’s too large.
> Test code
> Do it in Python.

Output should be:
Total runtime
Total records in both datasets
Total records unique to BusinessA
Total records unique to FirmB
For users in both datasets, the percent with differing job-titles
A random 10 line sample from the CSV output
The CREATE TABLE MySQL statement