## Assignment 2.

### Problem 0.

I examined the first 2000 rows of data from NYC Open Data to see where creating multiple tables can save space. Because of python’s dynamic typing, using default data types does not yield maximum space savings. For example a typical value of Agency is ’DEPT OF ENVIRONMENT PROTECTION’, which on my system python stores in a string of 67 bytes. In the first 2000 rows there were 41 departments, so if we stored Agency as a separate table, its foreign key could be as small as an unsigned 1 byte integer, but, on this system, the size of a python integer is 24 bytes, so we are not realizing huge savings by using a separate table. We can further improve memory use by creating separate tables for Business Title and Civil Service Title, but these savings will not be as great as with Agency, since there are many more unique values in these fields. Salary Frequency has only 3 unique values in the first 2000 rows, so it gets a separate table. Because parsing and cleaning addresses is beyond the scope of this assignment, I chose to keep addresses in the main table.

Table 1.

Job ID: int, primary key

Agency: int, foreign key

# Of Positions: int

Business Title: string

Civil Service Title: string

Salary Range From: float

Salary Range To: float

Salary Frequency: int, foreign key

Work Location:

Division/Work Unit: string

Job Description: string

Minimum Qual Requirements: string

Preferred Skills: string

Additional Information: string

Posting date: datetime

Table 2.

Agency ID: int, primary key

Agency: string

Table 3.

Freq ID: int, primary key

Salary Frequency: string