DATA.TABLE ASSIGNMENT: HEALTHCARE DATA

INTRODUCTION

For this assignment, we will be using some simulated electronic health records (these are not real data!). This is a common sort of dataset for health care systems to use when tracking all the patients and the outpatient activity. You should take a few minutes to review the datasets using Excel, read the descriptions, and understand how they fit together. We will only use a few datasets in this exercise, but you could explore all of them on your own.

- OutpatientVisit.csv
  - This is a table with a row for every outpatient visit. This file can be linked to the patient table using Patient ID, the clinic table using ClinicID, the ICD table using the ICD_1-3 columns, and the staff table using StaffID.
- Patient.csv
  - This file has one row for every patient, and includes demographic information about each patient. This file can be linked to the outpatient visit file using Patient ID.
- Clinic.csv
  - This is a table of all the clinic IDs, and if the clinic is primary care, specialty care, or emergency department. There should be one row per clinic ID. You can link this file to the OutpatientVisit file by clinic code to see if a visit occurred in a primary care clinic, a specialty clinic, or the emergency department.
- Staff.csv
  - This is a table with information about each staff member. This can be linked to the visit file using staffID
- Mortality.csv
  - This is a table of date of death for each patient (if they have died). Patients who have not died are not listed in this file. You can link this to the other files using PatientID
- ICDCodes.csv
  - This is a table with the description for each ICD code. ICD stands for 'international classification of disease'. Each outpatient visit is associated with 1-3 of these codes, indicating what diseases the patient has that were addressed during the visit.

- DiseaseMap.csv
    - This is a table that maps different ICD codes to clinically meaningful categories. For example, there are many different diagnoses for diabetes, but we might wish to generally call all of them 'diabetes'.

QUESTIONS

You will use these datasets to answer some questions listed below. You must be careful to think about what the appropriate denominator is for each question. Use Python and Pandas to solve these questions. Create a Jupyter notebook, then save it as an html file or PDF (using LaTeX)

1) Are men more likely to die than women in this group of patients? Assume people without a date of death in the mortality table are still alive.
2) I am interested to know if there are patterns in the disease groups across gender. For every patient with at least one outpatient visit, identify if they have been diagnosed with any of the 22 conditions listed in the diseaseMap table at any time point. You will need to consider all three ICD columns in the outpatientVisit file (not just one). Create a table with the rate of disease for each condition for men, women, and all. It should look like this, where the XX% is the percent with the condition:

.	Men	Women	All
Alcohol	XX%	XX%	XX%
Cancer	XX%	XX%	XX%
…	XX%	XX%	XX%
Stroke	XX%	XX%	XX%

3) Calculate the mortality rate for every year between 2005 and 2018. Is it generally increasing, or decreasing? Assume patients are only at risk of death as of their first visit (in the outpatient Visit file). Once they have died, they are no longer at risk in subsequent years…
    a. This is a harder question to answer than at first glance. What should the denominator of patients be for every year? How will you calculate it?

