In this set of tasks, you will be using made-up data of the flow of patients in an emergency department (ED). You should think how best to structure the data, analyze it, and communicate your approach and findings. Please submit your code and indicate how long it took for you to answer each question. Please also prepare a report with your answers. You will be evaluated both on your program (organization, simplicity, commenting) and your discussion.

Physicians work in shifts, in which they begin work at a set time and stay until they discharge their patients (usually past the end of shift). Patients arrive and are immediately assigned to a physician unless the physician has not started his or her shift yet. In the latter case, the patient is assigned to the physician at the beginning of the shift. In the dataset test_data.csv, you will see comma-separated data in which each row represents a patient visit. The variables are as follows:

1. visit_num: Row identifier for the patient visit
2. phys_id: Physician
3. shift_date: Denotes the date on which the physician's shift started
4. shift_start: Denotes the time at which the physician's shift began
5. shift_end: Denotes the time at which the physician's shift ended
6. arrive: Date and time of patient arrival to ED
7. leave: Date and time of patient discharge order
8. pred_lnlos: Measure of expected log length of stay, where length of stay is the difference between leave and arrive, based on patient demographics and medical conditions (you can think of this as "patient severity")

Using a statistical program, perform the following tasks:

0. Summarize the data. Do some observations appear to be data entry errors?

1. Some patients may arrive before their physician's shift starts and therefore would have to wait. Other patients may be discharged after their physician's shift ends (and the physician would have to stay past the end of shift). What percentages of visits fall in these categories?

2. Show hourly patterns of patient arrivals and the average severity of these patients. How might one formally test whether patient severity is or is not predicted by hour of the day?

3. Create and include with your solutions a dataset recording the "census," or number of patients under a physician's care (patients who have arrived and have not yet been discharged), during each hour of a physician's shift from beginning to 4 hours past the end of shift. The observations in this dataset should correspond to the shift (defined by the shift date, shift start time, and shift end time), physician (phys_id), and the hour of shift (hour, where hour = 1 in the first hour of the shift, hour = 2 in the second hour of the shift, etc). How does the census vary with time relative to end of shift? Discuss conceptually how you construct censuses and address issues with discrete time. Produce a "lower bound" census, "upper bound" census, and "exact" census for each observation.

4. Which physician appears to be the fastest at discharging patients? You should answer this with a regression of log length of stay and present your results graphically. What are potential threats to the validity of your assessment? Show the robustness of your estimates of physician effects to various specifications.