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 for both tasks 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 if 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.txt, 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_name`: Physician
3. `shiftid`: String variable denoting the date and beginning and end times of the physician's shift. If the shift spans midnight, the date corresponds to the beginning time.
4. `ed_tc`: Date and time of patient arrival to ED
5. `dcord_tc`: Date and time of patient discharge order
6. `xb_lntdc`: Measure of expected log length of stay, where length of stay is the difference between `dcord_tc` and `ed_tc`, 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 (accounting for fact that phenomena in #1 are legitimate)?
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. Describe hourly patterns of patient arrivals and the average severity of these patients. Extra credit: 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 (`shiftid`), physician (`phys_name`), and the hour of shift (`index`). `index` should be defined as follows, so that it should mostly negative values and have a maximum of 3 in the dataset: The hour ending at the same time as shift end is indexed -1, the hour beginning at shift end is indexed 0, and the hour beginning one hour after shift end is indexed 1, etc. Hint: You will need to transform the text in `shiftid` into numerical shift beginning and end times capturing both date and hour; you should ignore patient hour observations falling outside of the shift times of interest. How does the census vary with time relative to end of shift? Discuss conceptually how you construct censuses, and note (and for extra credit, address) issues with discrete time. For extra credit, you may want to 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. You may also show results graphically. Discuss how you thought about which variables you control for. What are potential threats (and any evidence of them) to your assessment (you do not need to actually address these threats)? How robust are your estimates of physician effects to various specifications? Are there any concerns you might have with a finite number of patients per physician?