**In a .Rmd file, write the code and include written answers to the following steps:**

1a. Read the data. Tell R to treat admission\_type\_id, discharge\_disposition\_id, and admission\_source\_id as categorical.

1b. Clean the variables `diag\_1`, `diag\_2`, and `diag\_3` using the code you wrote in the Lesson 11 Homework. You should end up with 3 new variables, `diagnosis1`, `diagnosis2`, and `diagnosis3`, containing human-readable interpretations of the ICD-9 billing codes.

1c. Remove the old variables `diag\_1`, `diag\_2`, `diag\_3`. If you created any intermediate variables (such as `injury\_cause1`), remove those also.

1d. Use code to determine which columns have no variability (i.e., all rows contain the same value). Print the names of the columns, and remove those columns.

2a. Count the number of missing values that are denoted by NA. Count the number of missing values that are denoted by "?". (In a factor/character variable, R will not automatically ignore rows containing "?".) Which column has the most missing values?

2b. Replace the "?" in the `race` column with NAs. Then use mode imputation to impute the NAs in the data. (Leave the "?" in other columns alone.)

3a. Use linear regression to model the number of diagnoses a patient has as a function of `age` and `diagnosis1`.

3b. As age increases from [0-10) to [90-100), what happens to the number of diagnoses? Explain your answer based on the linear regression. Illustrate your answer with a graph using `ggformula`.

3c. Compared to people whose diagnosis1 is a blood issue (the default value), which types of diagnosis1 are significantly associated with an increased number of diagnoses? Which types are significantly associated with a decreased number of diagnoses? Explain. Illustrate your answer with a graph using `ggformula`.

4a. Under the Hospital Readmission Reduction Program, hospitals where large proportions of patients are readmitted within 30 days may be penalized in their Medicare reimbursements. Create a binary variable that tells whether `readmitted` is `<30` or not.

4b. Make a bar graph and a conditional bar graph of the relationship between `diabetesMed` and whether a patient was readmitted within 30 days. Also make a bar graph and a conditional bar graph of the relationship between `gender` and whether a patient was readmitted within 30 days.

4c. If we want to predict whether a patient will be readmitted within 30 days, which variable is more informative: `diabetesMed` or `gender`? Explain your choice, referring to specific aspects of the graphs you made.

4d. Use logistic regression to model whether a patient will be readmitted within 30 days, based on `number\_inpatient` and \*one\* of the variables `diabetesMed` or `gender` (choose the one that is more informative).

4e. Write 2-3 sentences describing the direction of association between the predictor variables and the probability that the patient will be readmitted within 30 days.

4f. Make \*one\* graph showing the predicted probability of readmission as a function of both predictor variables.

**Now, knit your .Rmd file containing all the steps above into a LastName\_FirstName.pdf file**. You will be submitting the knitted file along with the pdf file for the R portion of the project.

Your score for the R portion of the project will be based on the following:

|  |  |  |  |
| --- | --- | --- | --- |
| Criterion | Excellent | Average | Weak |
| .Rmd file  5 pts | Knitted .docx or .pdf  No huge data dumps  Easy to read, with appropriate formatting (e.g., headers)  5 pts | Knitted .docx or .pdf  No huge data dumps  3 pts | Not knitted, or contains a huge data dump (e.g., displaying the entire data set)  0 pts |
| Code completeness and efficiency  30 pts | Followed all instructions, with clean and well-organized code  30 pts | Code is disorganized or has obvious inefficiencies  23-28 pts | Did not follow all instructions  0-21 pts |
| Accuracy  15 pts | Results in problems 1d and 2a are correct. Other output accurately reflects the data source.  15 pts | Small or subtle errors. Counting NAs but not “?” in problem 2a falls into this category.  11-14 pts | Problem 1d is incomplete, or major errors in accuracy elsewhere.  0-10 pts |
| Graphs  25 pts | Selects appropriate graph types for problems 3b, 3c, 4b, 4f and implements them correctly  25 pts | Graphs are acceptable but sub-optimal  18-23 pts | Missing or inappropriate graphs  0-17 pts |
| Explanations  25 pts | Accurate, complete explanations for problems 3b, 3c, 4c, 4e  25 pts | Explanations are accurate but could be more thorough  18-23 pts | Errors in explanations  0-17 pts |

* + A knitted file (LastName\_FirstName.pdf ) from the .Rmd file in R portion.