Data Science and R – Assignment 2

This is a graded assignment for your course. Use the template `assign2.Rmd` file provided, rename it to `<initials>-assign2.Rmd`, e.g. `HJW-assign2.Rmd` and submit it with your graph image from Q3i) as a **zip** file, e.g. `HJW-assign2.zip` through iCampus. Code should be answered within code chunks and any answer/comment that is not a runnable R code should be provided outside the code chunk where indicated in the template. Code chunks that do not run will result in loss of marks.

0) Download the file `diabetes_readmit.csv` provided with the assignment pack. Save it to your working directory.

The dataset represents 10 years (1999-2008) of clinical care at 130 US hospitals and integrated delivery networks. It includes 43 features representing patient and hospital outcomes. Information was extracted from the database for encounters that satisfied the following criteria:

(1) It is an inpatient encounter (a hospital admission).
(2) It is a diabetic encounter, that is, one during which any kind of diabetes was entered to the system as a diagnosis.
(3) The length of stay was at least 1 day and at most 14 days.
(4) Laboratory tests were performed during the encounter.
(5) Medications were administered during the encounter.

The data contains attributes such as patient number, race, gender, age, admission type, time in hospital, medical specialty of admitting physician, number of lab test performed, HbA1c test result, diagnosis, number of medication, diabetic medications, number of outpatient, inpatient, and emergency visits in the year before the hospitalisation, etc. Description of the columns are HERE.

Ensure the `ggplot2` package is loaded.

```
1.# Loading dataset into a data frame
```

a) Read in the file `diabetes_readmit.csv` using `read.csv()`. Save it into a data frame called `df`. [1]

b) Inspect `df`. Some variables have `"?"` values, e.g. the first rows of the weight column. What do you think `"?"` values mean for this dataset? [1]

Answer: _____

c) Reload `diabetes_readmit.csv` into `df` using `read.csv()` as before and add an additional parameter `na.strings ="?"` [1]

Code: _____

d) What is the effect of adding the parameter `na.strings="?"` when reading in the dataset? [2]

Answer: _____

2. # Data cleaning

a) Find columns with the most NAs (denoted by `is.na()`). Loop through all the columns to find the total number of NA values in each column. `Sort` the number of NAs in **decreasing** order. Print the column name with the number of NAs starting from the column with the most number of NAs . You can use **any** method you want to achieve this but the code must be provided. [5]

What are the two columns with the most number of NA's?

Column name with the most number of NAs: _____

Column name with the second most number of NAs: _____

Bonus: Instead of just returning the number of NAs in the column, return the number of NAs as a proportion (or percentage) of the overall data.

b) Remove the two columns with the highest number of NAs from `df`, i.e. `df` should now **NOT** contain these two columns. [4]

Code: _____

c) If you had a choice, would you **remove** any other column(s) (max 2) in order to reduce the data size without losing vital information? If so, which one(s) would you choose and why? If you choose not to remove any columns, explain why. **Note**: you do NOT have to actually remove these columns. [3]

Answer: _____

3. # EDA using visualisation

I. We want to see which age group has the highest readmissions.

a) Convert the variable `readmitted` (the number of times the patient was readmitted) to an ordered factor with levels "NO", "<30" and ">30". [3]

Code: _____

b) Plot a bar of the `age` and `fill` the bars with `readmitted`, i.e. the bars should be stacked according to the group count of readmission of all the age groups. [4]

Code: _____

c) From your graph, which age group has the highest readmission count? [1]

Answer: _____

II. We want to understand if older patients have longer stays than younger patients. To do this, we will plot some graphs to see the underlying distributions and relationships between the variables.

d) Plot a density plot of the `time_in_hospital`. Density plots are a variation of bar charts where the data is smoothed. Use `geom_density()`. [2]

Code: _____

e) Now we want to plot the `time_of_hospital` and age, but since there are many `age` groups, we'll make multiple density plots of `time_in_hospital` according to the different age groups. Add a `facet_wrap()` with the `age` as the variable to facet with. `Fill` the density plot with `age`. [4]

`Code:` _____

f) Adding gender in. We also want to investigate the stay period of different genders. Keeping age as the variable to facet the density plot of the `time_in_hospital`, change the `fill` to `gender`. [3]

`Code:` _____

g) Do you see some erroneous values in the gender category? We are going to redo the plot by removing these values by subsetting `df` to contain only valid gender values. WITHOUT modifying `df` or creating a new data frame, replot the graph using valid gender categories. [2]

`Code:` _____

h) To see a slightly smoother visualisation, set `alpha` and `adjust` arguments to appropriate values within `geom_density()` Your plot should look similar to the one shown below with the appropriate labels (i.e. title with YOUR name, x, y and legend labels). The legend title can be changed by adding the layer `scale_fill_discrete()`. [6]

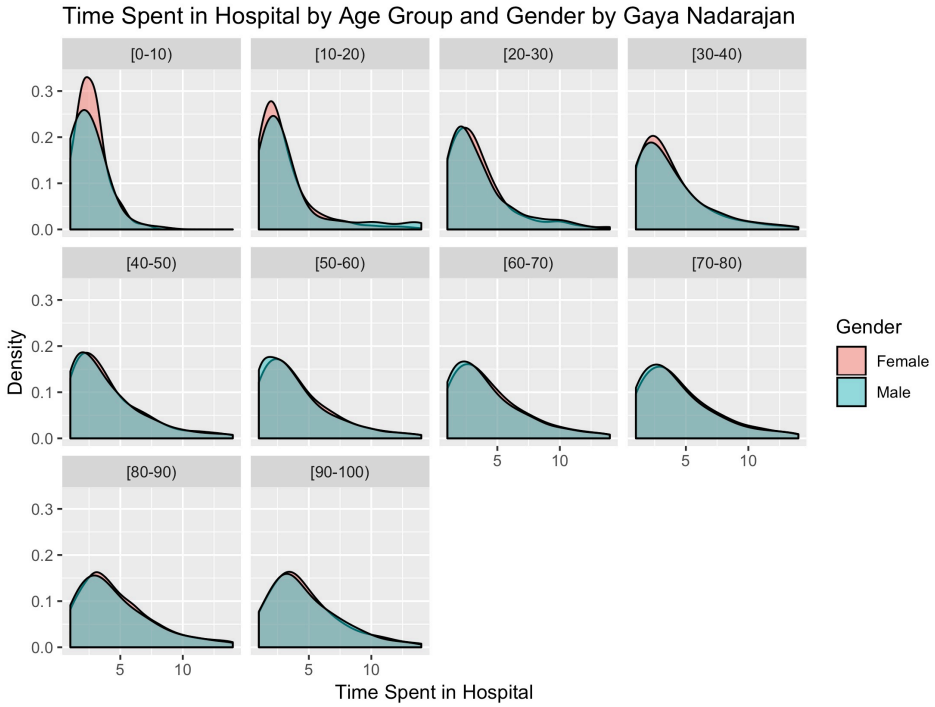
`Code:` _____

i) Save your plot as `"<initials>-diabetes.jpg"` (e.g. `HJW-diabetes.jpg`) and attach it with your submission. Provide the code for saving the file. [1]

`Code:` _____

j) How would you interpret this graph? [2]

`Answer:` _____



Notes:

- Marks will be deducted for code that don't not run, so fix all syntax errors! When loading your data file (Q1a), leave the path to the file as it worked for you, we can work with our own data file when marking.
- For questions 3h)-i), full marks will not be awarded if you provide correct working code but fail to submit your graph image. No marks will be awarded for a graph image without evidence of the code that generated it.
- For efficiency reasons, do not create duplicates or large subsets of the data frame when completing the assignment, i.e. work with only one copy of the data frame loaded to memory. This is good practice in general, especially when working with large datasets.