Working with Health Data: Data Cleaning and EDA (Exploratory Data Analysis)

Information about the data set: https://archive.ics.uci.edu/ml/datasets.html
Abstract: From National Institute of Diabetes and Digestive and Kidney Diseases; Includes cost data (donated by Peter Turney)

Attribute Information:
1. Number of times pregnant
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. 2-Hour serum insulin (mu U/ml)
6. Body mass index (weight in kg/(height in m)^2)
7. Diabetes pedigree function
8. Age (years)
9. Class variable (0 or 1)

For this assignment you will perform some cleaning and EDA of the diabetes dataset from the UC Irvine Machine Learning Repository, described at the bottom of this specification.

1. Read the file ***diabetes2.csv*** into a dataframe named **diab**.

2. EDA: Calculate summary statistics on **diab** and identify which variables seem to have invalid data in need of cleaning.

3. EDA: Plot the frequency distributions of all variables in **diab** and identify which variables seem to have invalid data in need of cleaning. Reconcile differences in #2 and #3, if there are any.

4. For the variable **insu** produce a new variable of logical that has the value TRUE at any position in the vector **diab$insu.**

5a. For the variable **pres** (ie: blood pressure) produce a new variable of logical that has the value ZOMBIE at any position in the vector **diab$pres** where the blood pressure is that of a zombie rather than that of a human. ***Hint***: zombies do not have a heartbeat – what BP results from no heartbeat?

5b. For the variable **pres** use a logical expression from #4 to take a slice of the vector **diab$pres**. This slice will contain all the elements of **diab$pres** that need to be replaced with NA.

6. For the variable **pres** use your slice of **diab$pres** from #5 to replace the values that should be NA with the value NA.
Hint: one way is to use something like this, filling in with your logical expression from #4 and #5:
diab$pres[<logical expression>] <- NA

7. For the variable **skin** (ie: measure of skin fold thickness at triceps) use a logical expression to take a slice of the vector **diab$skin**. This slice will contain all the elements of **diab$skin** that need to be replaced with NA.

Hint: the slice will contain all values of **diab$skin** where the skin fold thickness is that of a skeleton rather than that of a human. (Skeletons have zero skin thickness).

8. For the variable **skin** use your slice of **diab$skin** from #7 to replace the values that should be NA with the value NA.

9. Confirm that variable ***insu*** needs cleaning, clean them by replacing values that should be NA, with values within normal range (16 -180 mu U/ml).

10. For all the variables other than **pres** and **skin** ***that need cleaning***, clean them by replacing values that should be NA, with values between the 1st interquatile and mean of the variable.

11. Execute the following commands, each one on a different column of diab:
summary, range, IQR, quantile, median, mean
Very briefly describe what the commands reveal

12. Create the grouped histogram for

- **class** variable of **diab**. Include a title, x and y labels in your plot.
- **class variable insu**

13. As in #12, plot the **class** variable of **diab**, but first typecast (coerce) the **class** variable to type factor. Include title and axis labels as before.

14. Create a plot of your own choosing to show the distribution of a single variable OR the relation between 2 variables of the **diab** dataset. Include a title and axis labels in your plot.

15. Note: this question is meant to be strange; figure out what is going on:
Execute the following 2 lines of code:
diab$col1 <-**diab$pres**/(**diab$pres**-***median***(**diab$pres**, na.rm=T))
diab$col2 <- ***sqrt***(**diab$pres**/(**diab$pres**-***median***(**diab$pres**, na.rm=T)))

Then execute the two **which** commands below:

which is introduced in Book of R ch 4, the other functions in ch 6
which(***is.infinite***(**diab$col1**))
which(***is.nan***(**diab$col2**))

After executing these, explain what happened in the creation of **col1** and **col2** to result in the output of each of the two calls to **which**