Data Cleaning and Visualization in R

Follow steps to read, clean and visualize the dataset described below after questions. There is also a link below to optional additional information on residuals of lm plots.

In your preferred R IDE (such as RStudio):

1. Download into your working directory the file *heart.csv* found at the url provided.

   http://www.cs.usfca.edu/~pfrancislyon/data/heart.csv

   Use read.csv to read the data from *heart.csv* and store in a variable named **heart**. Note that this file contains a header.

   What is the data type of **heart**?

2. In the environment, expand **heart**. Examine the variables and their data types. Correct any variable type you find that should be numeric but is not. You may find it useful to obtain documentation on read.csv by typing ***?read.csv***

3. Execute a summary of heart. Write a couple of sentences on your observations. Many researchers enter an impossible value of the correct type to indicate missing data (as 999 was used for missing gestation in the babies dataset). Correct any of these that you find in the dataset by either read.csv or some other method.

4. Fit a linear regression model named **fit_num_all** with the variable num (diagnosis of heart disease) as a function of all the other variables.

   Execute a summary of **fit_num_all**. How many observations were deleted due to missingness? How many are left in the model? Could this many points (observations) determine even a 2D (2-dimensional) line?

   Hint: Execute the following 2 lines, look up **na.omit,** to find out how the number of observations was reduced so drastically:

   summary(na.omit(heart))

   nrow(na.omit(heart))

5. Fit a new linear regression model named **fit_num_1** by removing from your model in **#4** that variable that is almost entirely NA. This will result in a 13-dimensional model.

   Execute a summary of **fit_num_1**. How many observations are included in the model? Could this many points determine a 13D hyperplane?

6. Fit a new linear regression model named **fit_num_2** by removing from your model in **#5** that variable that is mostly NA. This will result in a 12-dimensional model.

   Execute a summary of **fit_num_2**. How many observations are included in the model? Could this many points determine a 12D hyperplane (11 predictor variables plus 1 response variable)?

7. Examine the descriptions of predictor variables below to see which are categorical (can hold only a discrete number of values). List the categorical variables that are boolean (can hold only two values). List the other categorical variables (those that can hold more than two values). What type did R assign to the categorical variables? Modify the non-boolean categorical *predictor* variables (this does not include **num**) so they are represented as factor variables.

8. Fit a new linear regression model named **fit_num_3** that is the same model formula as you fit in **#6**. The only difference is that the non-boolean categorical predictor variables are now represented as factors.

   Execute a summary of **fit_num_3**. Compare it to your summary of **fit_num_2**, which has 11 predictor variables with estimates of slopes in the fitted model. How has the list of fitted predictor variables changed in **fit_num_3**? How has the fit of the significant variables changed?

9. Modify the boolean categorical variables, (there are 3 of them), so they are represented as factor variables

10. Fit a new linear regression model named **fit_num_4** that is the same model formula as you fit in **#6 and #8**. The only difference from **#8** is that the boolean categorical variables are now represented as factors.

    Execute a summary of **fit_num_4**. Compare it to your summary of **fit_num_3**. How has the list of fitted predictor variables changed in **fit_num_4**?

11. The value zero for the variable **num** indicates no heart disease, whereas the values 1, 2, 3, and 4 indicate narrowing of the arteries, interpreted as heart disease. Create a new column in **heart** named **num2**. All zeros should be propagated from **num** to **num2**. Any nonzero value of **num** should be assigned 1 in **num2**. NB: there are no missing values for **num**.

12. Fit a linear regression model named **fit_num_slope_oldpeak_sex** with the variable **num** as a function of **slope, oldpeak** and **sex.**

    Execute a summary of **fit_num_slope_oldpeak_sex**.

    Fit a linear regression model named **fit_num2_slope_oldpeak_sex** with the variable **num2** as a function of **slope, oldpeak** and **sex**.

    Execute a summary of **fit_num2_slope_oldpeak_sex**.

    What is the equation of the fitted line for **fit_num2_slope_oldpeak_sex**?

    How does the adjusted R^2 of **fit_num2_slope_oldpeak_sex** compare with the adjusted R^2 of **fit_num_slope_oldpeak_sex**? What does this indicate to you?

13. Execute the commands:

    plot(fit_num_1, 1)
    plot(fit_num_slope_oldpeak_sex, 1)
    plot(fit_num2_slope_oldpeak_sex, 1

    The plots produced by these commands indicate problems with these linear regression models. (See Assumption **(iv)** below in **Assumptions of Linear Regression.** Write a couple of sentences identifying the problems.

Assumptions of Linear Regression

There are **four principal assumptions** which justify the use of linear regression models for purposes of inference or prediction:

(i) linearity and additivity of the relationship between dependent and independent variables:

(a) The expected value of dependent variable is a straight-line function of each independent variable, holding the others fixed.

(b) The slope of that line does not depend on the values of the other variables.

(c) The effects of different independent variables on the expected value of the dependent variable are additive.

(ii) statistical independence of the errors (in particular, no correlation between consecutive errors in the case of time series data)

(iii) homoscedasticity (constant variance) of the errors

(a) versus time (in the case of time series data)

(b) versus the predictions

(c) versus any independent variable

(iv) normality of the error distribution.

http://people.duke.edu/~rnau/testing.htm

Descriptions of Variables of the **heart** dataset:

1. age - age in years
2. sex - sex (1 = male; 0 = female)
3. cp - chest pain type (1 = typical angina; 2 = atypical angina; 3 = non-anginal pain; asymptomatic)
4. trestbps - resting blood pressure (in mm Hg on admission to the hospital)
5. chol - serum cholestoral in mg/dl
6. fbs - fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
7. restecg - resting electrocardiographic results (0 = normal; 1 = having ST-T; 2 = hypertrophy)
8. thalach - maximum heart rate achieved
9. exang - exercise induced angina (1 = yes; 0 = no)
10. oldpeak - ST depression induced by exercise relative to rest
11. slope - the slope of the peak exercise ST segment (1 = upsloping; 2 = flat; 3 = downsloping)
12. ca - number of major vessels (0-3) colored by flourosopy
13. thal - 3 = normal; 6 = fixed defect; 7 = reversable defect
14. num - the predicted attribute - diagnosis of heart disease (angiographic disease status). Ranges from 0-4, representing degree of narrowing of the arteries.

Not required: If you are interested to know more about the four residual plots of a fitted linear model in R, there is an excellent explanation in response to a question raised on stackexchange:

http://stats.stackexchange.com/questions/58141/interpreting-plot-lm