Practical Data Science
(Explorative Data Analysis)

Write R Scripts or use R to perform any mathematical operations while solving the following problems.

Problem1: Another attempt at Predictive Analytics Problems
Go through the following problems of kaggle:
https://www.kaggle.com/c/sf-crime/data

Do the following tasks:

a. Explore the train data set for crime incidents numerically and visually. Find out the patterns and hardcode those patterns and submit the solution to kaggle.
b. Find out the features that might be useful for machine discovery?
c. Apply tree learning approach and findout the parameter values for the model, which provides the best performance?(Use resamping techniques)
d. Predict the outcome for test data with the model and submit the solution to kaggle. Is the accuracy given by kaggle matches with your local validation set based accuracy?

Problem 2: Explore Car dataset
Download the car.data from datasets branch of algorithmica github repository. Here are the descriptions of the attributes of the car dataset:

buying: vhigh, high, med, low.
maint: vhigh, high, med, low.
doors: 2, 3, 4, 5more.
persons: 2, 4, more.
lug_boot: small, med, big.
safety: low, med, high.

The output class attribute can take one of the following values:

unacc, acc, good, vgood

Do the following tasks:

a. Load the dataset into frame and convert all the attributes to factor type.
b. Explore all the attributes individually using univariate numerics and graphics.
c. What kind of preprocessing do you suggest after the univariate explorations.
d. Explore all the bivariate relationships numerically and graphically.
e. What features do you recommend for predicting class category and why?
f. What kind of patterns have you discovered with the above explorations?

Practical Data Science
(Explorative Data Analysis)

Problem 3: Exploring Kidney data

Download the chronic_kidney_data.txt from datasets branch of algorithmica github repository. The description of the dataset can be found at following link:

http://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease

Do the following tasks:

a. Load the dataset into frame and convert all the attributes to factor type.
b. Explore all the attributes individually using univariate numerics and graphics.
c. What kind of preprocessing do you suggest after the univariate exploration.
d. What kind of missing value handling mechanism do u suggest and why?
e. Explore all the bivariate relationships numerically and graphically.
f. What features do you recommend for predicting the disease is chronic or not and why?
g. What kind of patterns have you discovered with the above explorations?