I am investigating the role of Cytomegalo Virus (CMV) in breast cancer as well various personal history as well as family history variables in breast cancer or variation of behavior of these variables in breast cancer and healthy subjects. I will be simulating 250 rows and 13 columns. Here, predictor variables will be blood markers like hIL10, viral Il10, IL6, IL8, IL12, TNF alpha, age, sero-status for CMV virus, minutes of exercise per week, smoking history, history of menstruation in 2015 and family history of breast cancer. Breast cancer within 5 years will be the response variable. Blood markers will have quantitative values, age and minutes of exercise will also have quantitative values. This study is based on the on-going project in Biology department and I will be trying to maintain the distribution as well as correlation among the variables while simulating these variables. All the quantitative variables will have similar spread, statistics as well as correlation for example, with variables like age to that of original data. Sero-status, smoking history, history of menstruation in 2015, family history will have Boolean values and they will have two levels – true and false. These variables will also have similar distribution and correlation as per the original dataset. Smoking history, menstruation history will have similar correlation with age and family history will have similar correlation with history of breast cancer.

As mentioned earlier, breast cancer within 5 years will be the response variable and it would have Boolean values with two level true and false. Around 20% of the subjects in original study had history of breast cancer and hence, I would also be simulating data with similar percentage of values which will be true for breast cancer. Minutes of exercise per week had negative correlation with breast cancer which contradicts with our knowledge about the domain. I will again try to examine whether that negative correlation exists in this simulated data or not. It is seen that breast cancer is more common in elder women, menopausal women which was also discovered in the analysis of original dataset and I would try to examine this relation here as well. Smoking history and family history of breast cancer were found to have no correlation with breast cancer and I would try to examine the same here as well.

I would first observe and visualize distribution in various variables with the help of different R functions like summary, contingency tables, histograms, etc. I would also examine the correlation between different variables in the original dataset using functions like cor. I would try to preserve the spread of values as well as their correlations while generating the data. This would be an iterating process involving verification of the spread or distribution or correlation in the simulated data variables and modifying them to achieve desired values or distribution.

I would be plotting various quantitative variables in breast cancer as well as healthy subjects to see the difference in their behavior in both the groups. I would also be performing various statistical tests like test for normality of distribution, variance in two groups and two sample t test to examine the similarity or difference of these statistical concepts in healthy as well as breast cancer subjects. Lastly I would be fitting various linear and logistic regression models to predict breast cancer as a function of predictor variables and understanding the correlation between these variables and breast cancer.