**R Module 4 Assignment – Data Exploration**

Using the beta carotene data (called ‘betacar’) you created last week including the baseline average:

1. Create a box plot (whisker plot) of the baseline average by preparation. Don’t forget to label you axis. Do you think the baseline measurements were evenly distributed among the preparations – why or why not?
2. Look at the distribution of week 12 measurements overall by a histogram. Add title, labels and some color. Do you think this is normally distributed – why or why not.

Download the FEV\_missing.rda r data file from Canvas.

1. Plot the relationship between the AgeCat variable and Smoker. Describe this relationship.
2. Describe the missingness of the data including how many NAs there are for each variable and any interesting relationships you find.
3. How many missing values are there for smokers? For non-smokers? What percentage of smokers is missing FEV? Non-smokers? What does this tell you about any analysis you might do comparing smoking status to FEV?

Please submit your answers, plots and relevant R code.