SPRINGBOARD DATA SCIENCE CAREER TRACK

CAPSTONE PROJECT #1 – STATISTICAL ANALYSIS

SILVIA MAIONE

The dataset used in this project comprises the entire trees population in New York City. For the purpose of statistical analyses, the variables that are considered here are the trees' diameter, health and the other categorical variables, previously encoded, which refer to specific conditions of branches, trunk, etc.

The tree diameter is the only continuous variable. Its probability density function and cumulative density function are plotted to gain insight on the type of distribution. The latter is plotted together with the one of a normal distribution having the same mean and standard deviation. The plot shows that this variable isn't normally distributed. The density functions are also plotted for the five boroughs separately, to see if any similarities can be identified. The curves look to have an exponential behavior, specifically like a gamma distribution. Moreover, permutation sampling is done for two boroughs' diameter data, to verify if they have the same distribution. From these plots, it looks like Manhattan has a different behavior. These differences seem also to confirm the finding in the exploratory data analysis.

For the categorical variables, previously encoded, the "chi-squared" test of independence (using Pearson's statistic) is run, to verify the relation with trees' health, also encoded. The purpose is to screen the independent variables to use in the model. The test is run for each encoded variable vs the health. The significance level is chosen to be 0.05. For all the variables, the p-value from the test is very small or equal to zero, so initially they can all be included in the model.