Anissa Khan
EECS 731 - Project 1 Report

By the end of the Jupyter Notebook, I have created 2 useful datasets:

1. dfGlyphCombined: contains pesticide data for glyphosate merged with cancer data
2. dfCombined: contains all pesticide data merged with cancer data

My reasoning behind selecting a single pesticide to create a dataset is to make it easier to determine a single pesticide's effect on cancer rates. One could create a dataframe for any pesticide of one's choosing in a similar manner.

A strong caveat made by selecting these initial datasets:

1. The cancer dataset contains data aggregated from the years 2012-2016. It contains both medians and averages over that time period. The pesticide dataset contains counts from 2014 and 2015 separately. By using this cancer dataset to determine any relationship between cancer and pesticide use, we are assuming pesticide use prior to 2014 is similar to its usage in 2014 and 2015. Ideally, we would find a cancer dataset that contains incidence rates from 2014 and 2015 separately so it matches up to the pesticide dataset.