Leading Cause of Cancer

The hypothetical question to explore is among cancer patients, is smoking the leading cause of cancer. The result of the analysis is that smoking was not the leading cause of cancer. Four causes of cancer were compared to the age of the cancer patient population. Among the 4 causes of cancer, the leading cause was air pollution. Air pollution made up 10% of the population of patients. Smoking made up less than 8% of the population. None of the variables seemed to have a strong statistical significance. This would apply to the specific data set that was used and may not apply to a different set of patients.

I feel that the understanding of the values of the scale used in the causes of cancer was a big miss during the analysis. This makes it seem that some variables should be combined. For example, passive smoking is still smoking, and this could have made up the low end of the smoking scale. However, the scale could have represented packs of cigarettes per week (smoking) or weekly number of cigarettes (passive smoking).

I feel that using the age variable as a base line for representing the entire population was a clever idea. I reviewed a histogram of each variable to try and determine which may be a good contender to go up against the smoking variable. However, I think I should have compared a combined variable view of something like a smoker that lives in a high air pollution environment. I could have explored how likely this would increase the risk for cancer. The remaining variables for the data set seemed to be symptoms of a cancer patient like coughing up blood and this was not as helpful in the analysis.

I would have assumed passive smoking would carry the same risk of cancer as smoking, but this was incorrect. I was correct in assuming that air pollution would cause an elevated risk of cancer. However, I was not expecting that the air pollution risk would overtake the population of smokers. I think that if the data set were different, we could have had different results. However, this data does not indicate the type of cancer the patient has. This has good and bad implications. This could mean that there is a mix of the diverse types of cancer which could validate our results. However, this could also mean that they all had the same type of cancer, and the results are invalid.

A challenge I faced was trying to have multiple variables on the regression analysis. It would have been good to know what population is affected by a combination of causes of cancer. I do not always like using data from Kaggle as I do not always fully understand the representation of variables are. For example, why was smoking on a scale of 1-8? Other than that, I feel comfortable with how the project went. Based on this one can infer that there a person needs to be exposed to multiple causes of cancer.