**Is Punxsutawney Phil an Accurate Predictor of Spring for Pennsylvania or the US?**

For the past several years, the Midwest seems to have experienced relatively mild winters, and early springs. This year, it seems that winter has lasted much longer, with snowfall events occurring into May. This year on Groundhog’s Day, the prediction was for an early spring. These events made me wonder how accurate the groundhog prediction is for Pennsylvania, as well as the United States in general. My hypothetical question was: is there a positive correlation between the groundhog’s prediction and the onset of spring? My null hypothesis is that there is no correlation whatsoever between the prediction and the recorded temperatures. Based on the data set, higher temperatures in February and March would indicate an early spring, therefore, there should be a correlation between the groundhog not seeing his shadow and higher temperatures. A full or partial shadow should be correlated with the lower temperatures. To test this, I created an integer variable for the prediction. No shadow was assigned to 0, and full shadow or partial shadow was assigned to 1.

Using EDA, I did not find any outliers in the temperatures, they all seemed to be reasonably possible outcomes. I did remove any lines of data that did not have a recorded prediction or temperature. I only kept the data for years that were complete. The data indicated that temperatures follow a normal distribution. The covariance and correlation numbers indicate that there is a very small negative correlation between the prediction and the average temperatures in various regions of the US, as well as the overall average temperature. I interpreted this as there is no correlation between the prediction and the temperatures. The hypothesis testing indicated that the p-values were large. This means that the findings are not significant, and the results can be explained by chance or coincidence, rather than a correlation between the predictions and the temperatures. The regression model was based on the temperature being a function of the prediction. This also indicated there is no correlation, and the null hypotheses is true. The R-squared values accounted for less than half a percent of the explanation of the temperature values. The F-statistic and t values also indicate that the probability of the results occurring without the prediction are well over half. The confidence intervals also include 0, so there is very little reason to believe that the prediction has any impact on the temperature outcome.

One thing I feel I missed during the analysis was binning the temperatures on the histograms to make those results more meaningful. This was a relatively straight forward analysis of temperatures, so the model was not very robust. For a more extensive research project, the method used to determine the groundhog’s prediction could be examined. There could be other potential variables that arise from this. For example, is there any consultation with local or national meteorologists for weather predictions? Do they use the Farmer’s Almanac or any other type of predicting mechanism to influence the prediction? Some of these methods could have an impact on the outcome. If included in the regression model, they could reveal any possible correlations. Based solely on the variables in the data set, my assumption was that the prediction had no impact on the temperature results, and I think this was correct. I also assumed that average temperatures are a normal distribution, and I feel that was correct as well.

My biggest challenge was interpreting the results of the regression model. While I feel that they supported what I thought the outcome would be, I do not yet fully understand all of the t values and F-statistics as completely as I would like.