Capstone Project 1 - Statistical Data Analysis

The section of statistical analysis is based on the previous EDA section and uses techniques for analyzing the CO2 trend and CO2 yearly differences.

I noticed a slight anomaly in the time series data, between the years 1990…1995. I used several techniques for a better understanding of the anomaly, like:

- sns.regplot for showing graphically the confidence interval for CO2 data for 1, 2 and 3 sigma;
- sns.regplot for the yearly difference
- sns.jointgrid for contour plots
- sns.residplot for plotting the residuals against the fitted values
- sns.qqplot for plotting the sample quantiles against the theoretical quantiles
- sns.boxplot for analyzing the 1990…1995 anomaly against the data before 1990 and after 1995
- CDF for investigating of the CO2 differences have a normal distribution
- Polynomial interpolation with sns.lmplot and np.polyfit, with the advantage of polyfit that it returns the coefficients for interpolation and other elements
- np.polyfit interpolation for the data until 1990 and extrapolation until present
- bootstrap replicates for getting statistical parameters even if there is only one set of data available
- Hypothesis testing using permutation replicates
- Test statistics and p-value

The results show that the 1990…1995 anomaly effects were diminished in time. The increase in CO2 concentration is currently even higher than before 1990 and the yearly standard deviation is relatively unchanged. This confirms the H0 hypothesis that the CO2 increase before 1990 and after 1995 are similar. The p-value is 1.0 for test regarding different standard deviations before 1990 and after 1995.

However, the anomaly created a delay in CO2 increase. As for now (January 2020), the current CO2 level should have been reached several years ago.