Assignment 5 – Linear Regression

Overview

- This is a continuation of Assignment 4. You must complete that assignment first.
- Generate three different views of long-term temperatures.
- Analyze how informative each view is.

Background

When we model a data set with a curve, it's important to determine how well the curve fits the data. We will use two ratios.

Coefficient of Determination (R^2)

- One measure of this is called R^2, which tells us how much of the variability in the data is captured by the model (the curve).
- R^2 ranges from 0 to 1:
    - a value of 1 means that the model is a perfect fit – i.e., it accounts for all the data's variability.
    - a value of 0 means that there is no relationship between the model and the data.

Standard Error to Slope

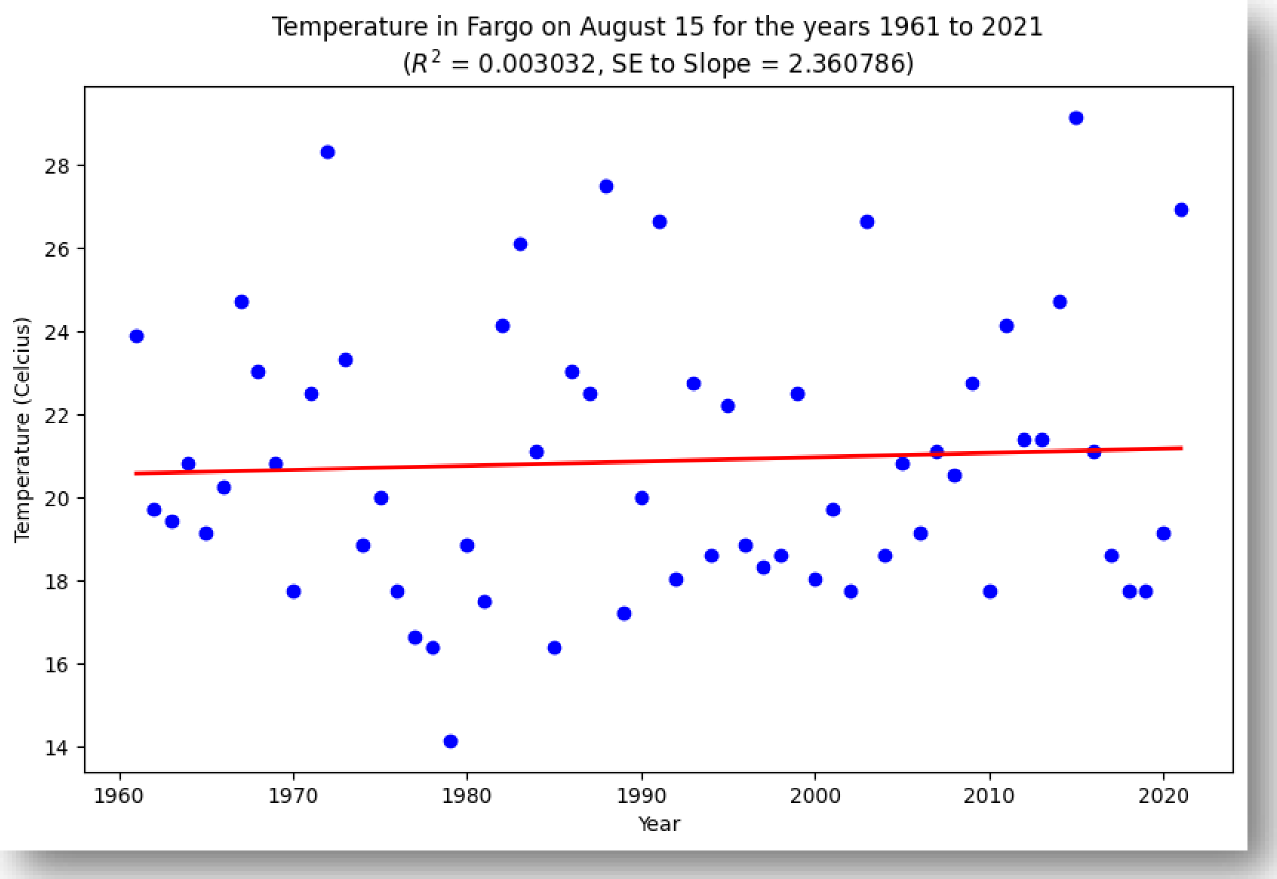
Another measure of a model's fit, in the case of a straight line, is how much of line's slope is due to "standard error" (a common statistics measure). A high ratio (say above 0.5) means that the line's slope is due to chance - or in other words, that there really isn't a linear relationship in the data. On the other hand, a very low ratio means that most of the line's slope is due to the data and not to error.

You are provided with implementations of these in the library 'LinearRegressionTools.py'.

Part A - Single Day, One City

1. In A4, you wrote the function, *singleDayPerCity*, which creates one chart per city, where each chart shows the temperatures on a single day for multiple years. Write a new function, *singleDayPerCityWithTrendline*, that adds trend lines to the charts. This function should:
    a. Use *polyfit* to and *polyval* to find the best fitting line for this data.
    b. Calculate R^2 and the Standard Error to Slope Ratio.
    c. Produce a graph showing all this information.
2. Choose four cities and four days. Produce sixteen charts with trend lines, one for each city/day combination. Be sure to include the R^2 and SE to Slope Ratio in the title.

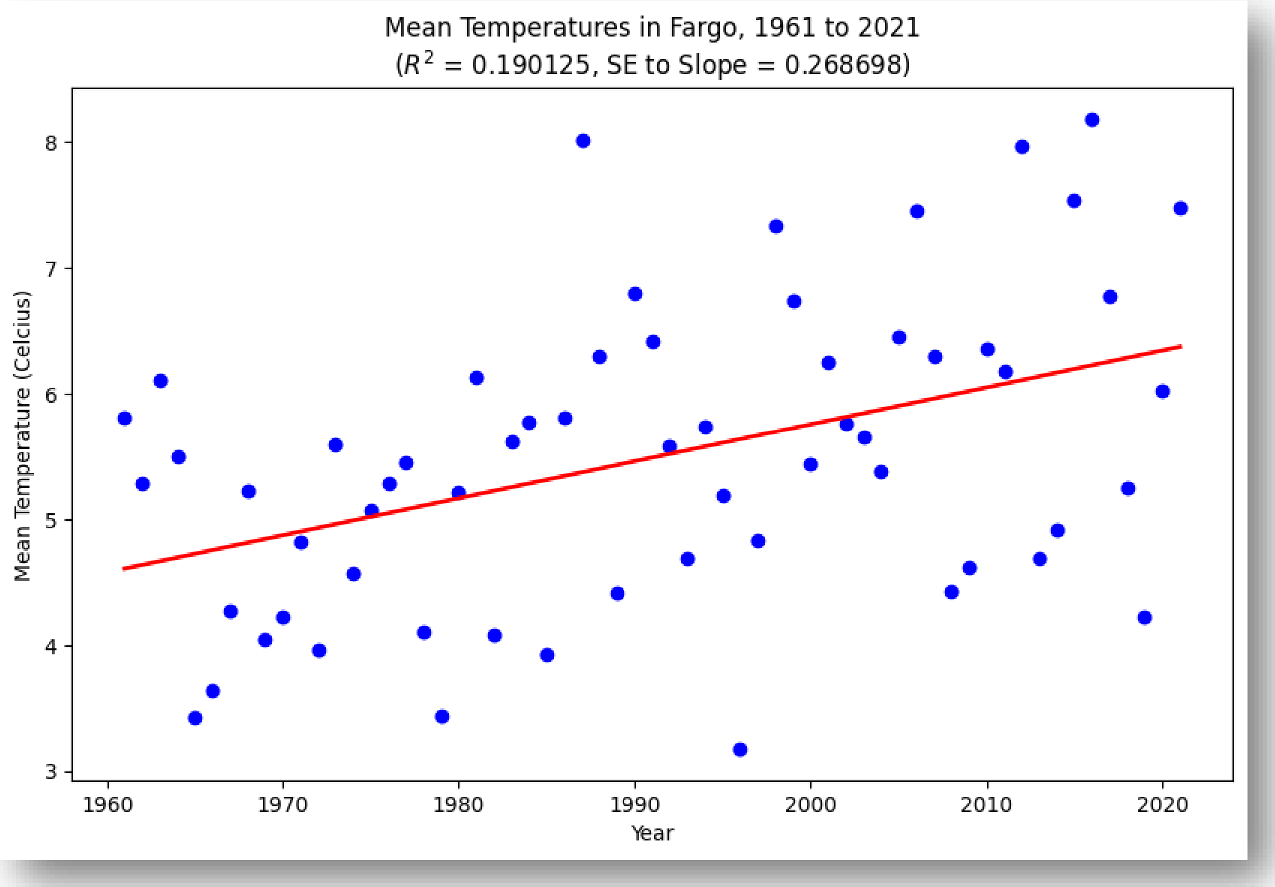
Sample Result



Part B - Annual Means, One City

1. In A4, you wrote the function, *annualMeansPerCity*, which creates one chart per city, where each chart shows the mean annual temperatures over for multiple years. Write a new function, *annualMeansPerCityWithTrendline*, that adds trend lines to the charts. This function should:
    a. Use *polyfit* to and *polyval* to find the best fitting line for this data.
    b. Calculate R^2 and the Standard Error to Slope Ratio.
    c. Produce a graph showing all this information.
2. Produce four charts for the same four cities you used in Part A.

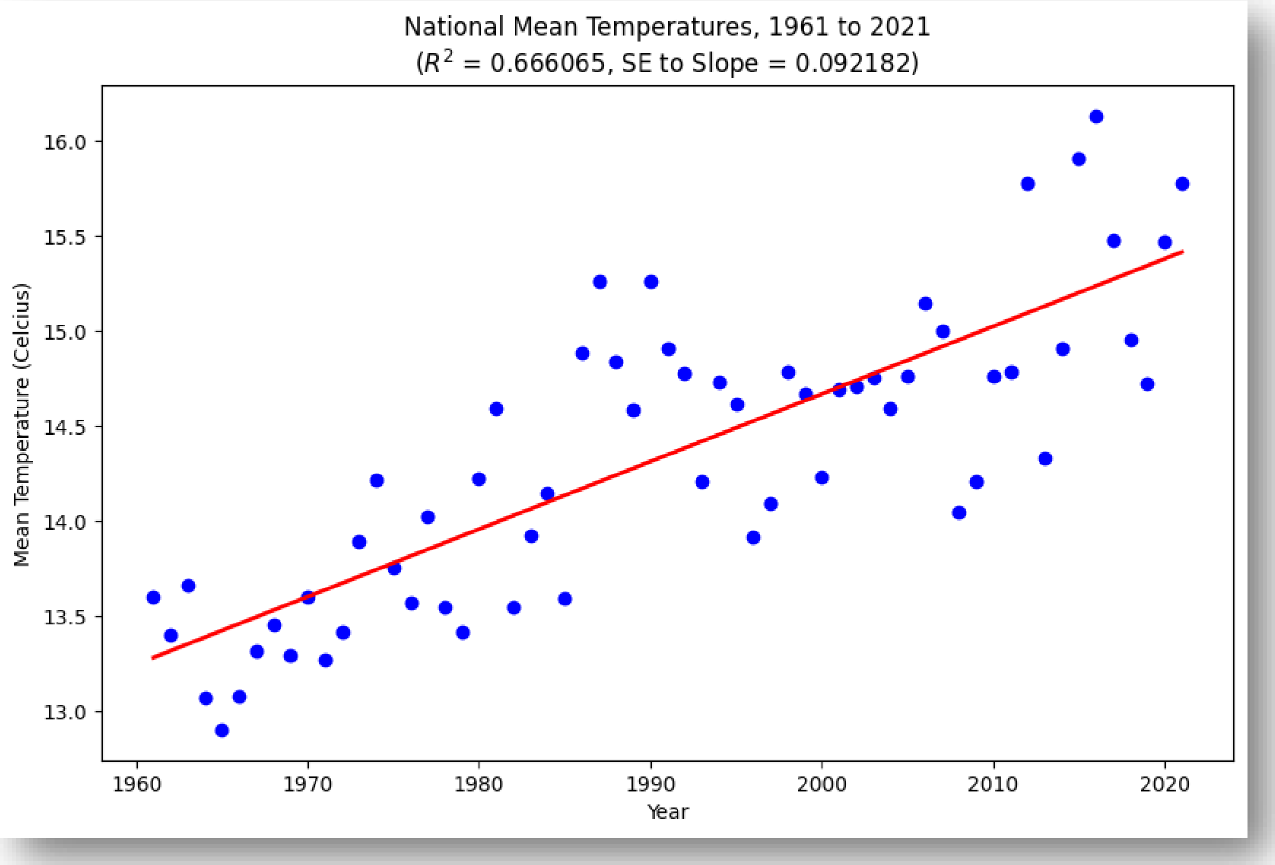
Sample Result



Part F - National Annual Mean Temperatures

1. Write the function *calcNationalAnnualMeans*(cities, years). This function should calculate the national annual mean temperature for each year specified.
   - The national annual mean is the average of the annual means of each specified city. So, first find the annual means for each city, and then find the average of those values.
2. Write the function *nationalAnnualMeans(cities, years)*. This function should create a chart showing the national annual mean temperatures for the years specified, as well as the trend line through the data. (Call *calcNationalAnnualMeans* to get the data you need.)
   a. Use *polyfit* to and *polyval* to find the best fitting line for this data.
   b. Calculate R^2 and the Standard Error to Slope ratio.
   c. Produce a graph showing all this information.
3. Call *nationalAnnualMeans* using the same cities and years that you used in Parts A-B.

Sample Result



What to Hand In

- The source code for all functions.
- Twenty-one charts produced (sixteen in Part A, four in Part B, one in Part C).