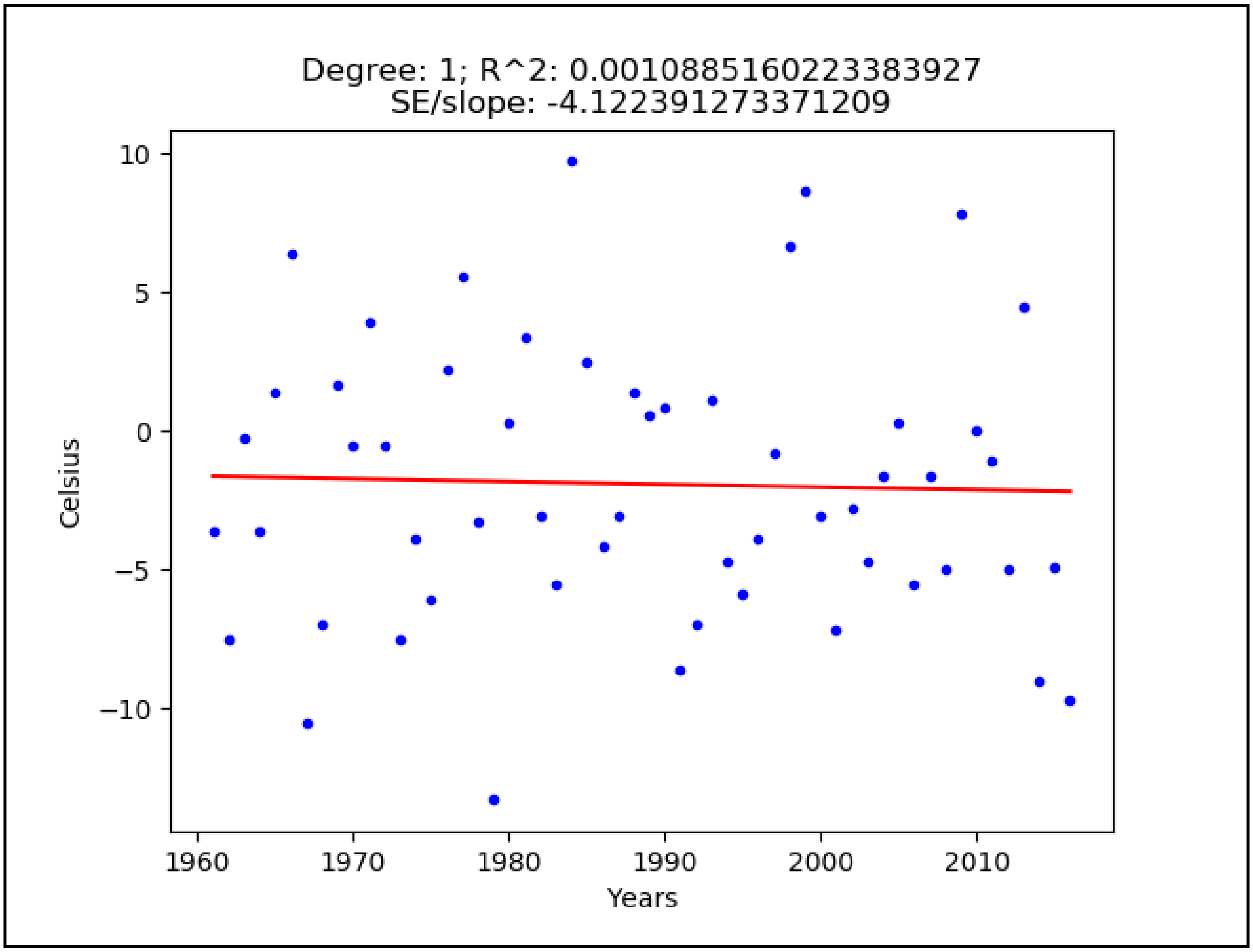
**Name: Alvin Zhu**
**Kerberos: alvinzhu**

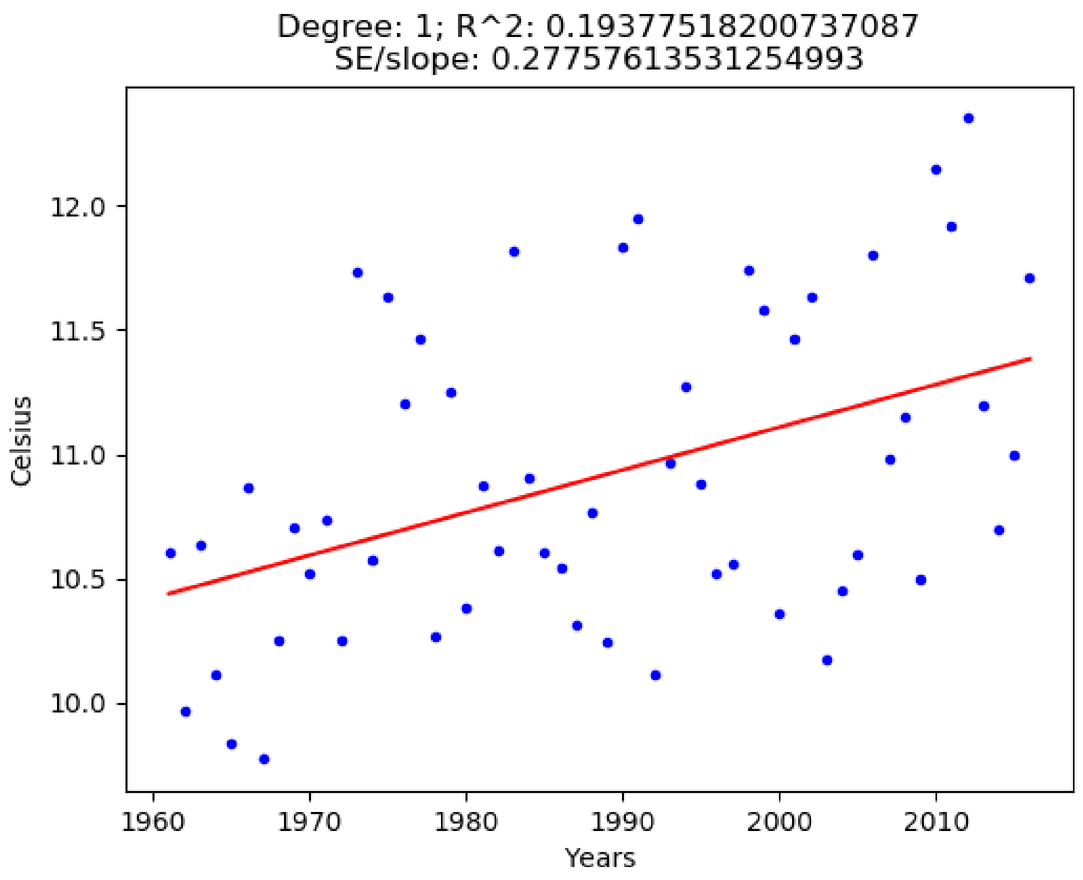
# Problem Set 5: Modeling Temperature Change

**Problem 4**

**Plot 4A:** *Average Daily Temp for Boston on 2/12 (1961-2016)*



**Plot 4B:** *Average Yearly Temp for Boston (1961-2016)*



**4.1** What difference does choosing a specific day to plot the data versus calculating the yearly average have on the goodness of fit of the model? Interpret the results.

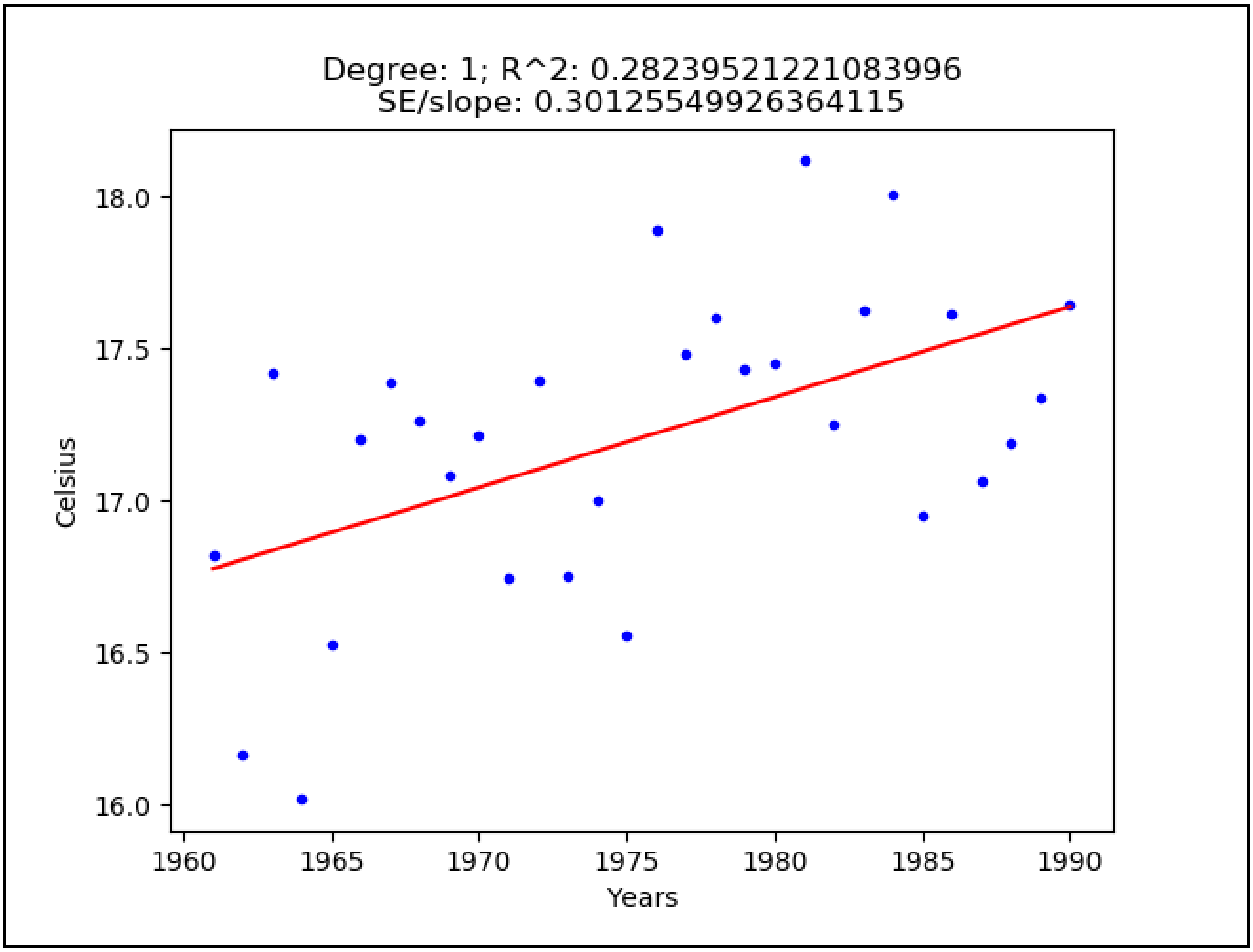
Choosing the yearly average produces a better fit for the plot because there is less variation between the average temperature of a given year than a given day. This is analogous to how DNA variation between individuals is greater than that between populations. Furthermore, due to the seasons of Boston, choosing a specific day especially during winter or summer will not produce a good model.

**4.2** Why do you think these graphs are so noisy?

These graphs are so noisy due to external and more-or-less random factors that affect the temperature of a given day.

**Problem 5**

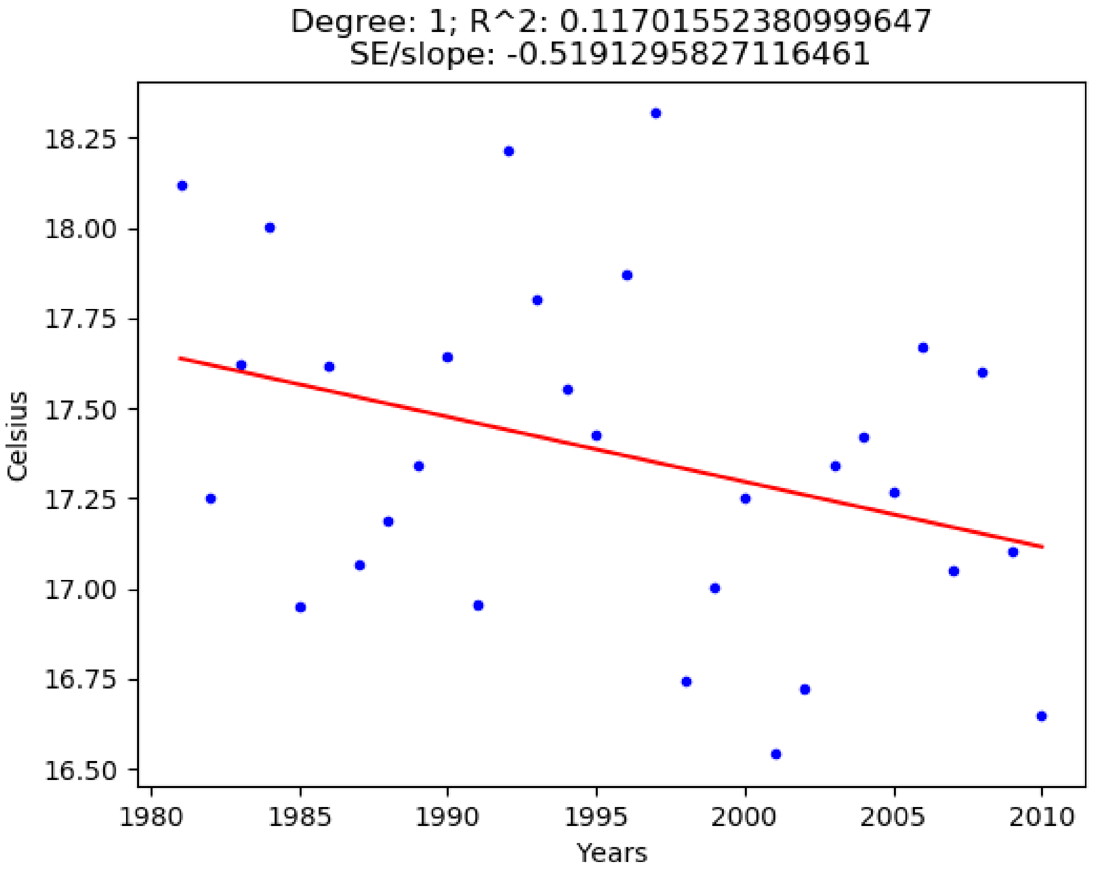
**Plot 5.A** *Increasing Interval (Los Angeles, length=30)*



**5.1** What was the start and end year for your window? What was the slope?

Start year: 1961
End year (inclusive): 1990
Slope = 2.971*10$^{-2}$

**Plot 5.B** *Decreasing Interval (Los Angeles, length=30)*

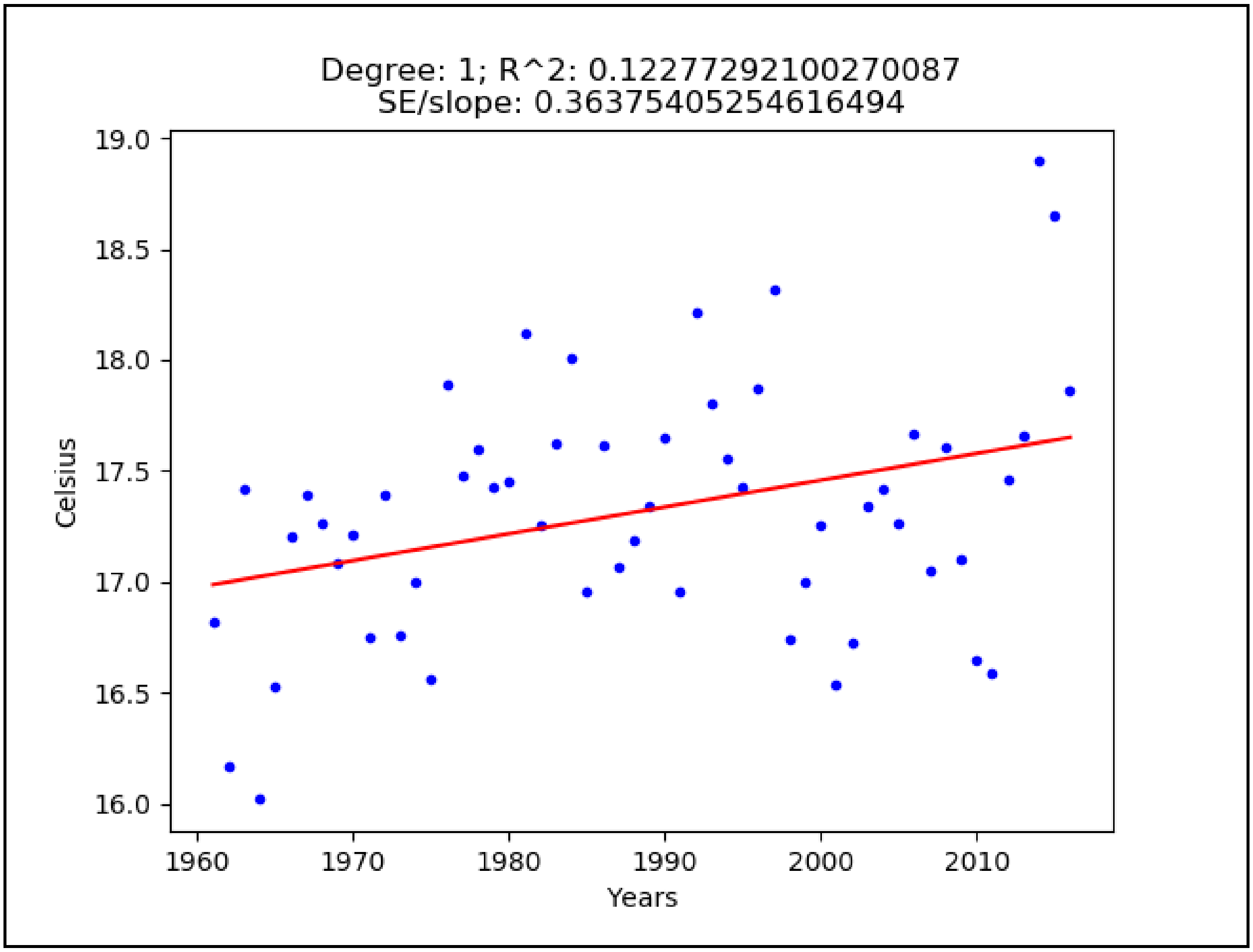


**5.2** What was the start and end year for your window? What was the slope?

Start year: 1981
End year (inclusive): 2010
Slope = $-1.801*10^{-2}$

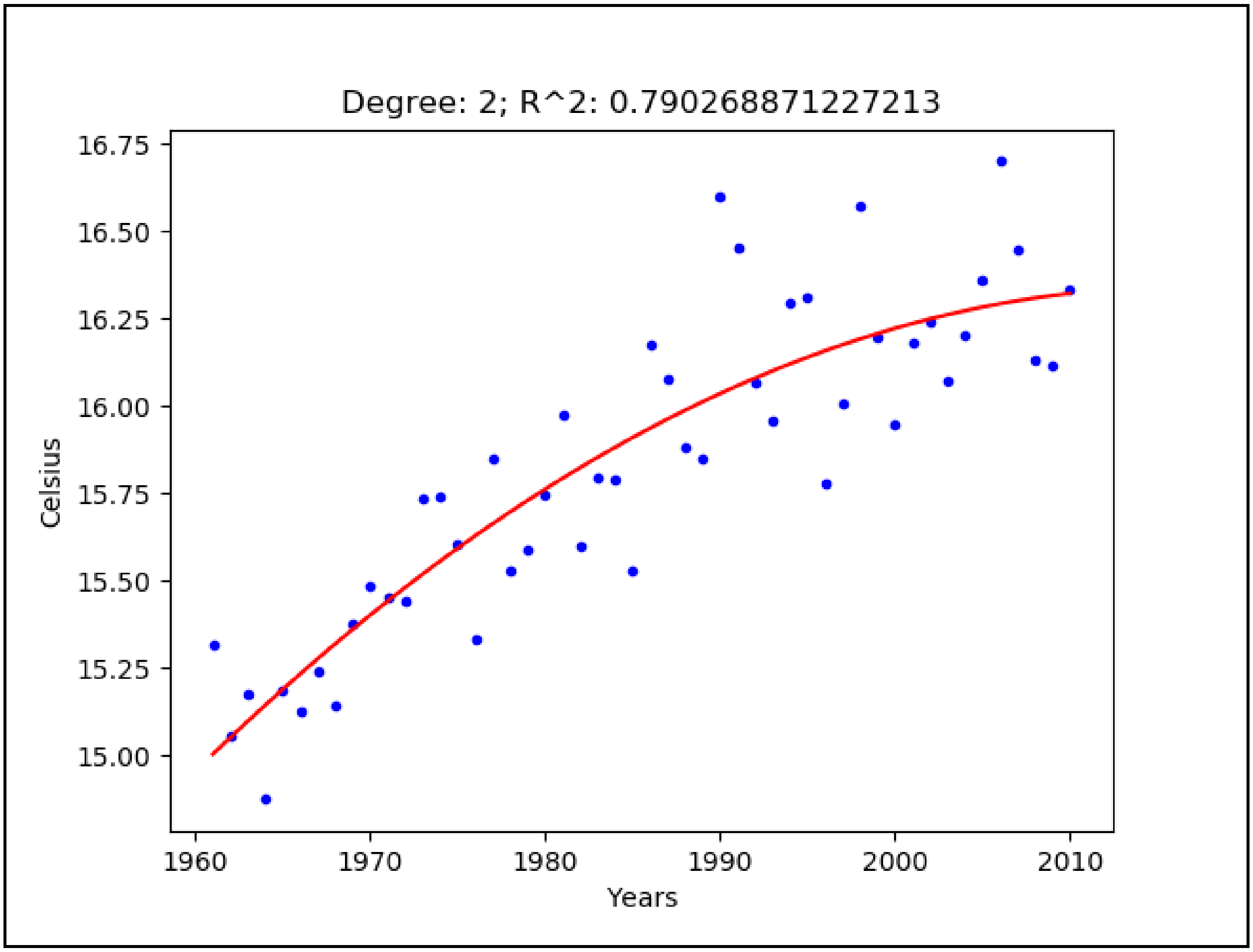
**5.3** Considering *both* plots, what conclusions might you make with respect to how temperature is changing over time?

Temperature is rising in general since the slope of the greatest temperature increase interval is greater than that of the slope of the greatest temperature decrease (increase outpaces decrease). Also see below for the average temperatures of Los Angeles in general:

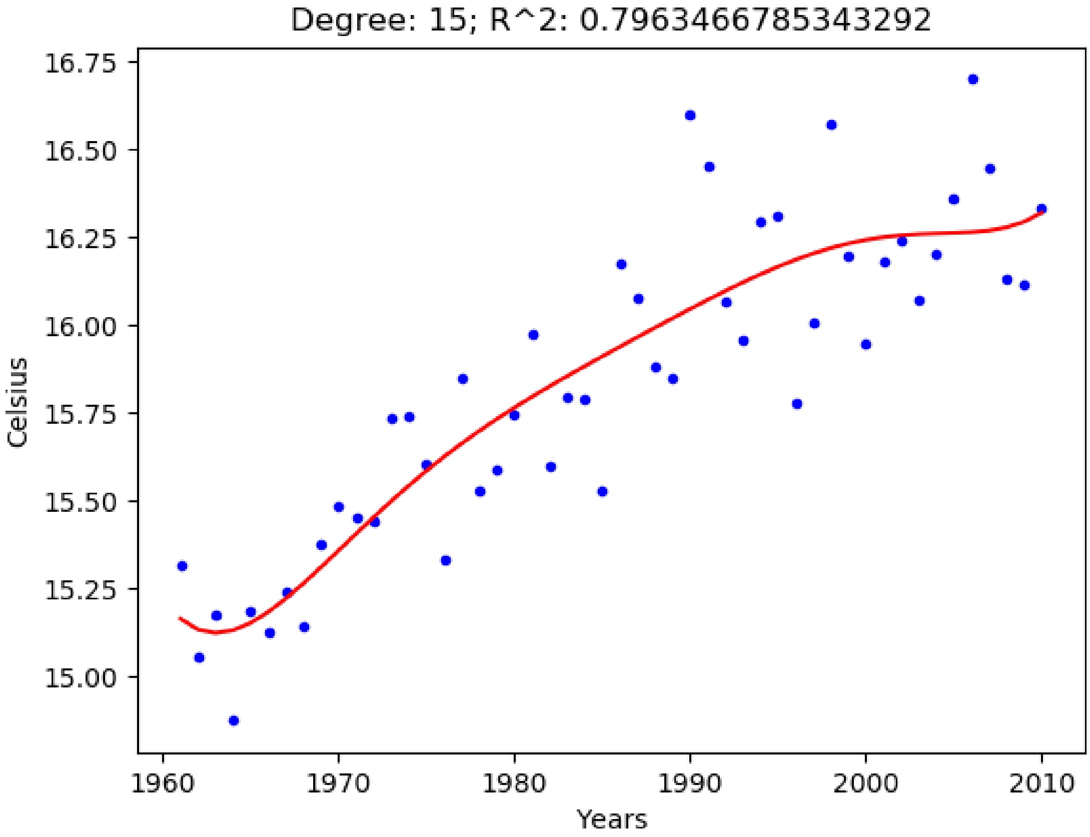


## Problem 6

**Plot 6.A** *Training Data, Degree 2*



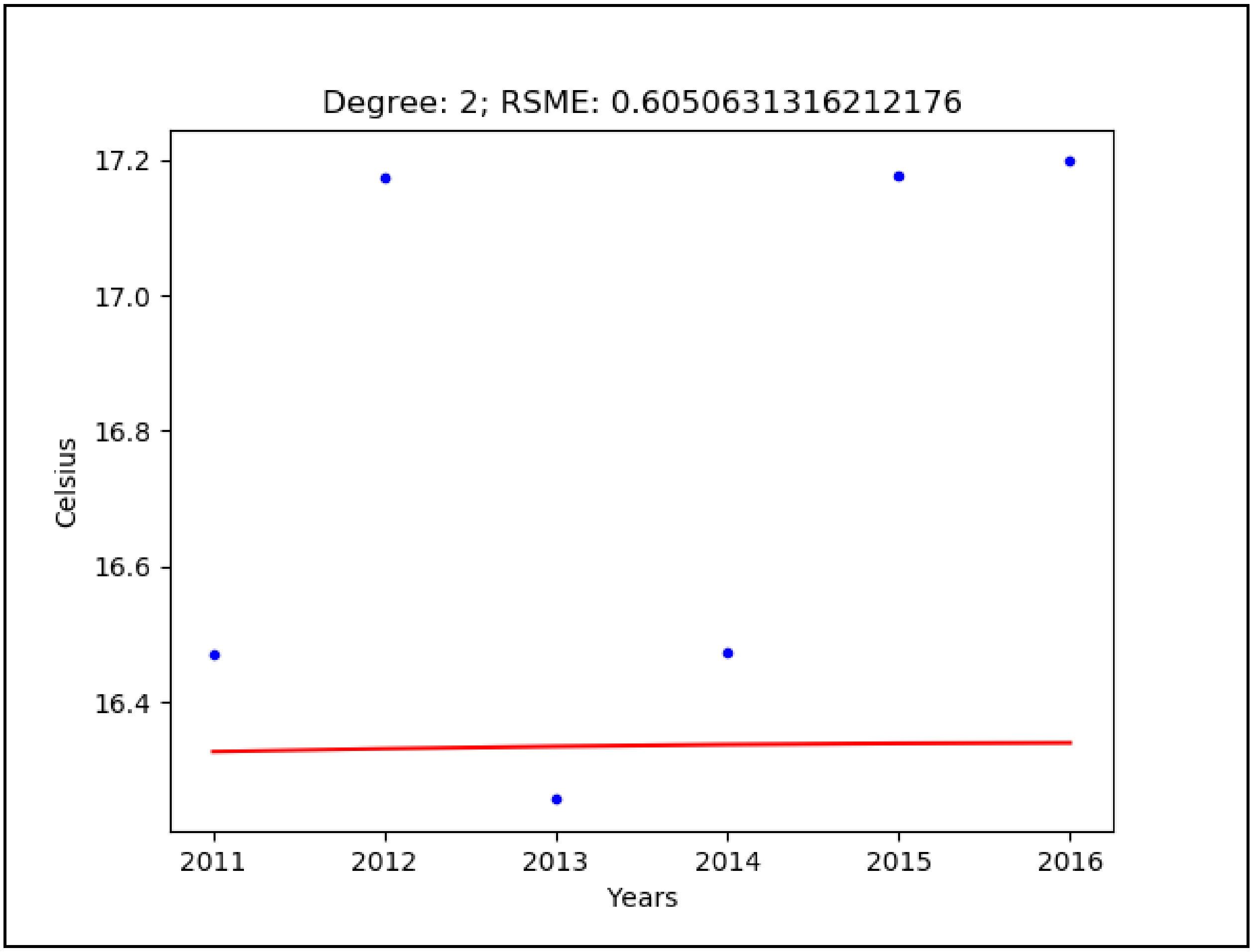
**Plot 6.B** *Training Data, Degree 15*



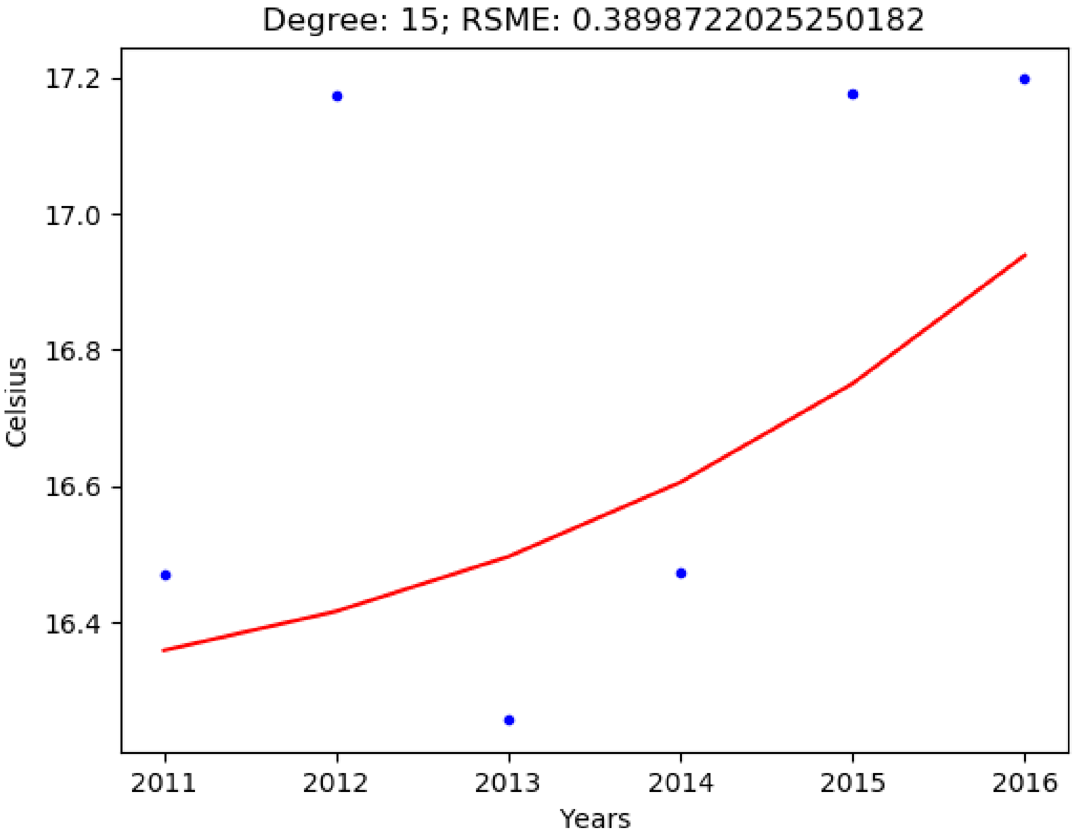
**6.1** How do these models compare to each other in terms of R^2 and fitting the data?

The model for the Degree 2 plot has a lower $R^2$ value than that of the Degree 15 plot. Therefore, the model for Degree 15 better fits the training data. With this being said, however, because the $R^2$ scores are very close, there is also the possibility that the Degree 15 may be outfitted (dip in the 1960s seem to suggest a drop in temperature but that may not actually be true). Thus although it matches my data well, it may not be so applicable for a wider use.

**Plot 6.C** *Test Data, Degree 2*



**Plot 6.D** *Test Data, Degree 15*



**6.4** Which model performed the best? Which model performed the worst? Is this different from the training performance in the previous section? Why?

The model for Degree 15 performed better than the model for Degree 2. This is not different from the training performance in the previous section because logically if the R$^2$ term (how closely a model matches the training data) is closer to 1, the deviation between a model's estimated and true values should be lower (closer to 0). Furthermore, graphically, the model for Degree 15 fits the testing data much better than Degree 2 (whereas before, they were quite similar), thus justifying how Degree 15 is a better fit than Degree 2.

**6.5** If we had generated the models using the data from Problem 4B (i.e. the average annual temperature of Boston) instead of the national annual average over the 22 cities, how would the prediction results on the national data have changed?

The prediction results would have been much lower than expected (as in the temperature would be lower) because Boston is a colder city compared to many other US cities.