Pset 5

\*\*\*Existing Issues:

1: Title formatting is strange.

2: Title can’t line break (/n ?)

Q4:

Part A1: Temperature on Jan 10 in New York

![Chart, scatter chart

Description automatically generated]()

Part A2: Annual Average Temperature in New York

![Chart, scatter chart

Description automatically generated]()

Part B: Annual Average Temperature of Various Aggregated Cities in the US

![Chart, scatter chart

Description automatically generated]()

Part C: Five Year Moving Average

![Chart, scatter chart

Description automatically generated]()

Writeups:

* How does this graph compare to the graphs from part A and B (​i.e., in terms of the R2​ ​values, the fit of the resulting curves, and whether the

graph supports/contradicts our claim about global warming)? Interpret the results.

The R value is much higher using the five year moving average compared to Parts A and B, and the Measured Points appear to be much smoother. Does this support our claim about global warming? I believe that for cities in particular, it could be used as an argument to support global warming over time.

* Why do you think this is the case?

Using a moving average does help increase the effective sample size of each measured point, taking into account more data per point. However, there are additional variables that must be considered (i.e. cities in other countries, temperature measurements in areas outside of cities)

4.D.2:

Testing Estimates using Training Interval, Aggregated cities annual average temperature estimate

![Chart, scatter chart

Description automatically generated]()

4.D.2.I

![Chart, scatter chart

Description automatically generated]()

* How do these models compare to each other?
  + R values are higher for the higher degree polynomials, and the estimates fit the measured points better
* Which one has the best R2 ? Why?
  + Degree 20 has the highest R2 because it has more degrees of freedom to fit the curve
* Which model best fits the data? Why?
  + Degree 20 best fits the existing data because it has more degrees of freedom to fit the curve.
  + However, my best estimate of the best predictor of the test data is the degree 2 polynomial fit, because the general trend \*appears\* to be a slightly decreasing rate of temp increase vs year

4.D.2.II

![Chart, line chart

Description automatically generated]()

* How did the different models perform? How did their RMSEs compare?
  + Deg.1 and 2 performed somewhat similarly, Deg 20 was way off.
  + RMSE of Deg 1 performed the best, RMSE of Deg 20 was way off
* Which model performed the best? Which model performed the worst? Are they the same as those in part D.2.I? Why?
  + Best performer based on RMSE is Deg 1. Deg 20 performed the worst.
  + The results are the opposite of the training data R^2, due to overfitting of the curve to the training data.

4.D.2.II. New York no MA

![Chart

Description automatically generated]()

* If we had generated the models using the A.4.II data (i.e. average annual temperature of New York City) instead of the 5-year moving average over 22 cities, how would the prediction results 2010-2015 have changed?
  + We end up underestimating the Measured Points, mainly because New York is a colder overall climate compared to the rest of the country.

4.E: Standard deviation of daily temperatures across cities over the year

![Chart, scatter chart

Description automatically generated]()

* Does the result match our claim (i.e., temperature variation is getting larger over these years)?
  + No, it appears that temperature variations are very slightly decreasing over the time period (<5% change in std dev over 50 years)
* Can you think of ways to improve our analysis?
  + Some ideas:
    - Gathering data across more areas
    - Instead of taking the standard deviation of the average temp of each day in the year, maybe take the standard deviation of all temperatures in the year? That way we’re taking more granular data into account for a given year