Comments to Aaron

1. This project aims to use the PurpleAir data to predict the level of pollution at times where the readings from EPA monitors are missing. I think this topic is generally very interesting, and the missingness itself is of crucial importance: it would be very cool to broadly document which data are more likely to be missing during what time at which stations (in which counties). I image that monitoring stations with consistently above median readings are more likely to be missing.
2. This seems to be a statistics / data science paper, but if you want to make it more ‘econ’, it would be great to speak to the literature of misreporting and principal-agent problems if you detect any selection bias. I’m sure there are many papers about the U.S. environmental policies, and there is an interesting paper about misreporting in China (Ghanem and Zhang, 2014). Like what they did, it may be useful to also test if the distribution of EPA readings has any discontinuities close to the cutoff of non-attainment (similar to the test proposed by McCrary (2008)). Assuming there is no *direct* manipulation of data (i.e. change readings), any discontinuities/ bunching below the cutoff could suggest some reporting bias.
3. The above could be a motivating fact (if there are any irregularities). It’s great that this project proposes a way to estimate the ‘missingness bias,’ i.e. the difference between average predicted EPA readings for the missing times and the average EPA readings for the non-missing times. This difference seems to include two components: the difference between actual missing EPA readings (had the readings been non-missing) and the actual non-missing EPA readings, and the prediction error for the missing EPA readings. What you want to measure here is actually the first component, so ideally you may want to bound the prediction error as well as the ‘missingness bias.’ I’m slightly worried that the EPA readings were missing during very bad, highly polluting days, and when extrapolating for these missing readings, the prediction error could be large and the prediction error for the non-missing EPA readings may not be comparable.
4. I imagine there are some systematic differences between EPA readings and PurpleAir readings because the latter are more likely to reflect indoor pollution level. Since you are using a correction factor published by EPA, I assume that it has already corrected for this difference in placement.
5. If I’m understanding correctly that you use the non-missing EPA readings to train a predictive model using PurpleAir data, why not use a more flexible model instead of a linear one? I’m not sure which one may have better out-of-sample performance though.