**Feedback from Dr. Geist for Preliminary Project Proposal**

Close

**Attempt 2 Feedback:**

[See comments details](https://merrimack.instructure.com/courses/15730/assignments/262694)

**Mar 26 at 10:03pm**

Wonderful job, love the professionalism you've put into this already!

**Some suggestions RE: variables:**

1. Given that you want to look at pollutant concentration, I think you will also want to grab population density. Given that average concentration is likely to be higher in denser communities, this will help to account for this.

2. Given distance to treatment centers is going to be affected by the sheer size of the tract (and the density of a census tract, but now I have you grabbing pop density) - make sure to include land area too. You will want to control for this in your model as well. Presumably larger tracts (in terms of land area) will have larger distances to facilities. You could also consider making new variables instead, if you'd rather.

3. What about the percentage of renter-occupied units? This can also tell you something about the economy of the tract, as well as the health: (1) more impoverished folks *tend* to rent (albeit not exclusively) and (2) rental homes, depending on location, may be a source for increased prevalence of disease risk and could indirectly impact hospitalization rates.

4. When you mentioned poverty, were you going to explore all poverty variables (I think you could consider this) or just one?

5. Another environmental variable (although they list it under physical infrastructure) that you could consider would be ACS\_PCT\_HU\_COAL.

I know I've given you a lot more to consider - but don't hesitate to be comprehensive here, since we will take an unsupervised method (PCA, clustering, etc.) to try to reduce dimensionality. I also allow penalized regression here if you're using it only as an unsupervised analysis. Point is - you can include more than you might have thought!

I would **strongly** suggest, since you're planning to include Medicare especially, the percentage of 65+ persons as well as the percentage of 65+ persons who live alone.

While I like that you have proposed feature engineering, I have concerns that you may not want to do that out of the gate. Here's why: some of these variables are about **access to care** (distance, insurance, poverty) while others are about **intrinsic and extrinsic causes of illness** (disability, pollution,  and possibly poverty to some extent). I would wait to see what your EDA and unsupervised analyses show, and then if you do feature engineering, it could be based on principal components or by these broader classifications I provided above. Feature engineering is great if we have a ton of complexity and we're needing to lump like things together, but in this case (and knowing what bits I do), I suspect it would mess up your model if you do it from the beginning like this. Think about this a bit, and feel free to follow up or fight me on it if you disagree!

**Some additional suggestions:**

Don't worry too much about t-tests/ANOVA here. While you're 100% right that if we wanted to publish in a public health journal we'd want them, I will suggest you may skip over them OR plan to put in your Appendix (totally fine with that if that's something your feel strongly about doing!). My suggestion is based on moving us toward our predictive models, which we can still use for hypothesis testing, especially if you undertake a multiple regression or penalized regression first.

**I wouldn't bother doing both regression and classification unless the granularity of regression didn't provide a robust model**. I think, for the sake of your sanity and the time limitation, I would drop the classification unless you're desperate to do both.

Great thoughts on penalized regression - move that sooner. I would strongly recommend that OR a PCA / cluster for your unsupervised model. If you do PCA, you can save Lasso for your first model.

I would bring your Stakeholder Information up sooner. I am even going to encourage you to start **addressing your stakeholder directly, since they are your audience!**