# **The role of air pollution on the propagation and mortality of COVID-19**

## Theory and Research question

Since the beginning of the COVID-19 pandemic, countless studies have been performed to gain a better insight about this virus. It has been for example known that the novel coronavirus is able to spread through [droplets](https://www.frontiersin.org/articles/10.3389/fpubh.2020.00163/full?fbclid=IwAR2rhRUdRpqZ-OGRId78hKvIXT36WDTUQlpR-JRH_0C-2ElafKv62QCrrwI), however [further research](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7444490/) suggests it might also spread through air pollutants, such as in PM2.5, PM10 and NO2.

This exploratory data analysis aims to confirm whether there is a correlation between air pollution levels and COVID-19 spread and lethality.

## **Datasources**

The data for this study is automatically downloaded through the written script. The Air data is available at the official website of the United States Environmental Protection Agency (EPA), [the downloaded files](https://aqs.epa.gov/aqsweb/airdata/download_files.html#Daily) are those from the daily PM2.5 FRM/FEM Mass (88101) for the years 2020 and 2021. These specific years were chosen as they overlap with the surge and increase of COVID-19 across the USA. The number of covid cases and deaths in the USA during these two years were taken from the [New York Times Covid-19-bot’s github](https://github.com/nytimes/covid-19-data), which updates the cases on a daily basis. The downloaded data is on state-level.

## **Justification for Methods**

After importing the three different data sets these were all converted into Pandas DataFrames for better data manipulation. For organization purposes, new data frames were created with only the columns of interest. For the air data only date, location and Air Quality Index (AQI) were necessary. The selected columns for the COVID data were the date, location cases and deaths. At the moment, the air data was still split in between 2 data frames and was thus united. COVID data was filtered to select only the years 2020 and 2021.

The first visualization was for a rough visual inspection of covid cases, death was done by generating a lineplot. This seemed logical considering infections usually seem to follow a trend and do not vary erratically. Plotting the air data AQI with a line did not seem to bring good results as the data changes quite erratically. Instead, a dot plot gave a little better representation. However, due to the large amount of data points, this was still very suboptimal.

To gain a better overview of the data and to simultaneously decrease the number of data points, it was decided to look at monthly averages. This would also correct for the fact that some days did not overlap for both data sets or inconsistencies (for example, COVID numbers which were notified a day late due to technical issues).

After calculating the monthy averages for AQI, covid cases and deaths per state, both data sets were merged. To observe whether there was a visible trend, for each state a plot was generated with a time axis on x and an dual-y axis containg cases/deaths and AQI score. For many states, there seemed to be a vague linear relationship, however this was of course not certain before performing a statistical test.

Considering the independent variable was air quality and the depend variables were covid cases/deaths, a ANOVA was fitting to determine whether the groups showed a significant difference. ANOVA does require the independent variable to be categorical, thus the AQI values were split into 3, 1-50 was “good”, 51-100 was moderate and 101-150 is unhealthy for sensitive individuals (these are [pre-determined AQI values](https://www.aqi.in/)). However, this did not seem to show a significant difference. Splitting the values into more groups, considering a lower and upper threshold for each category, did in fact show a significant difference according to the performed ANOVA. By only having 3 groups, the noise possibly became to large, hiding any significant difference between groups.

By creating violin plots with jitters next to them, it became clear groups indeed seem to be different. Unhealthy lower and moderate upper, did seem t have very few data points and should perhaps not be included in the analysis. Out of the three remaining groups, no significant difference was found in between “good upper” (AQI 26-50) and “moderate lower” (AQI 51-75). The “lower good” (AQI 1-25) seemed to have far fewer covid cases or deaths, which was also confirmed by a t-test independent sample t-test.

Finally, it was chosen to perform a one way ANOVA as it is believed number of covid cases / deaths are dependent on the independent variable air quality which was split into 5 categories. Additionally, to compare each group individually, a two-sided t-test was performed to verify whether the "lower good" group was different from the "good upper" and "moderate lower".