WIDS Datathon 2022 – Phase 2

1. Topic –

[**US Environmental Protection Agency (EPA)**](https://www.epa.gov/): weather, air pollutant, and census data

1. Registration – Completed on 12-04-2022
2. To Complete -

[**WiDS Datathon 2022 Webinar: Data Science and Climate Change**](https://youtu.be/0rETiC0EVuQ)

<https://www.widsconference.org/blog_archive/a-beginners-tutorial-for-the-wids-datathon-2022-challenge>

[WiDS Datathon Excellence in Research Award](https://urldefense.proofpoint.com/v2/url?u=https-3A__www.kaggle.com_c_phase-2Dii-2Dwidsdatathon2022_overview&d=DwMFAw&c=jf_iaSHvJObTbx-siA1ZOg&r=chY5SL9AriF_8CgaY3bQK7xAEqBXRzpWMbegTWzu0sE&m=YByJqX3fCTJ1NowpMo6vow2e0M7EwCILVUDJnDBt7Ks&s=tdOM6QL190-6NBJYK8nnfyJTw5bQVieD3yaGDaCzpEg&e=).

Next meeting 21-02-2022

After that, the plan is to find a data scientist team member

Notes:

Sharada Kalanidhi - Blog

Data manipulation - Dplyr” is a bread-and-butter package for data manipulation

 “janitor” and fastDummies” that also help with the dataset preparation

GOSSIS data, which reflects real-world hospital data.

Comparing columns in large data frames can be tough. I found this function from the package “janitor” useful in comparing very large Training and Test data frames: compare\_df\_cols(Training, Test).

Another trick that really helped me was writing out the structure of the dataframes. For example:   
Trainstr<- capture.output(str(Train, list.len=ncol(Train)))

Creating dummies from the package fastDummies was very useful.   
Traindummies<- dummy\_cols(Train\_fctrs, remove\_first\_dummy=TRUE)

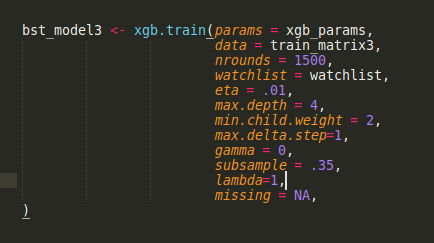
In case you decide to fit Xgboost,you will need to create a DMatrix, for which there are specific things to keep in mind.  There are many ways in which the Matrix construction can fail.  It is very important to have the Training and Test matrices exactly lined up.

In addition, there are very simple fixes that help with the construction. For example, lining up the column names of the Training and Test matrices is a simple trick that prevents DMatrix construction failure.  For example: Test<- Test[names(Train)]  
  
All of these hints and techniques are relevant to data preparation for models such as Tensorflow, SVM etc. as well.

References:

Elements of Statistical Learning:  
[**https://web.stanford.edu/~hastie/ElemStatLearn/**](https://web.stanford.edu/~hastie/ElemStatLearn/)

A question one might like to address early on is the type of patterns embedded in the data that could result in appropriate classification. For example, can the data points be separated easily with a line in the middle? Are there penalization approaches that could impact the quality of fit?   
  
Here are some resources that explore these concepts:   
[**https://statweb.stanford.edu/~owen/courses/305a/Rudyregularization.pdf**](https://statweb.stanford.edu/~owen/courses/305a/Rudyregularization.pdf)  
[**http://heather.cs.ucdavis.edu/draftregclass.pdf**](http://heather.cs.ucdavis.edu/draftregclass.pdf)  
[**http://web.engr.oregonstate.edu/~xfern/classes/cs434-18/Regularization-5.pdf**](http://web.engr.oregonstate.edu/~xfern/classes/cs434-18/Regularization-5.pdf)  
[**https://www.cs.cmu.edu/~mgormley/courses/10701-f16/slides/lecture4.pdf**](https://www.cs.cmu.edu/~mgormley/courses/10701-f16/slides/lecture4.pdf)  
[**https://www.ics.uci.edu/~xhx/courses/CS273P/04-linear-regression-273p.pdf**](https://www.ics.uci.edu/~xhx/courses/CS273P/04-linear-regression-273p.pdf)  
[**https://people.eecs.berkeley.edu/~russell/classes/cs194/f11/lectures/CS194%20Fall%202011%20Lecture%2004.pdf**](https://people.eecs.berkeley.edu/~russell/classes/cs194/f11/lectures/CS194%20Fall%202011%20Lecture%2004.pdf)  
  
Could there be nonlinear patterns in the data that if represented appropriately, would assist with the separation of the classes? If you wish to explore nonlinear separation, there are several approaches.  
  
Here are some pointers:  
[**http://cs229.stanford.edu/summer2020/cs229-notes3.pdf**](http://cs229.stanford.edu/summer2020/cs229-notes3.pdf)  
[**http://matt.colorado.edu/teaching/categories/jsw7.pdf**](http://matt.colorado.edu/teaching/categories/jsw7.pdf)  
[**https://www.ics.uci.edu/~welling/teaching/KernelsICS273B/svmintro.pdf**](https://www.ics.uci.edu/~welling/teaching/KernelsICS273B/svmintro.pdf)  (introductory)  
[**https://arxiv.org/pdf/math/0701907.pdf**](https://arxiv.org/pdf/math/0701907.pdf) (more mathematical )  
[**https://cs224d.stanford.edu/lectures/CS224d-Lecture7.pdf**](https://cs224d.stanford.edu/lectures/CS224d-Lecture7.pdf)  
[**https://www.cs.utah.edu/~zhe/teach/pdf/Tensorflow\_tutorial.pdf**](https://www.cs.utah.edu/~zhe/teach/pdf/Tensorflow_tutorial.pdf)  
  
As you build different models, you might consider trees and ensembles. These algorithms give you significant control of the fitting process in terms of parameters you can experiment with.   
[**https://arxiv.org/pdf/1603.02754&hl=th**](https://arxiv.org/pdf/1603.02754&hl=th). (Xgboost)   
[**https://xgboost.readthedocs.io/en/latest/**](https://xgboost.readthedocs.io/en/latest/)  
  
If this is your first time doing hyper parameter tuning, try an experimental approach towards model fitting. Study your fits as you tune the parameters. For example, are stumps (shorter trees) better or worse than deeper trees- for this dataset? How does eta affect the performance? Soon, you might develop an intuition for your fits- you might see steep learners that quickly fizz out or slow and steady risers that make micro improvements in performance over several thousand steps.   
  
Here is a sample of an Xgmodel parameter set:



**3. Consider Research**  
This year, the WiDS Datathon is placing significant emphasis on research. I think this a terrific opportunity to think mathematically/ statistically and plunge into active areas of research.  The areas of pure and applied math/stat research are simply vast. Are there statistical, algebraic, geometric, number theoretic, complex analytic, topological areas that interest you?  Perhaps there are applied areas- applications to healthcare, for instance, that are of interest? Maybe you always wanted to design a new visualization technique. Well, now is the time to bring out all of those ideas and put forth a research proposal!!   
  
Above all, enjoy the data challenge. If you view it as an exercise to continue learning and growing, you will find that it opens a door to an immense and very thrilling field.

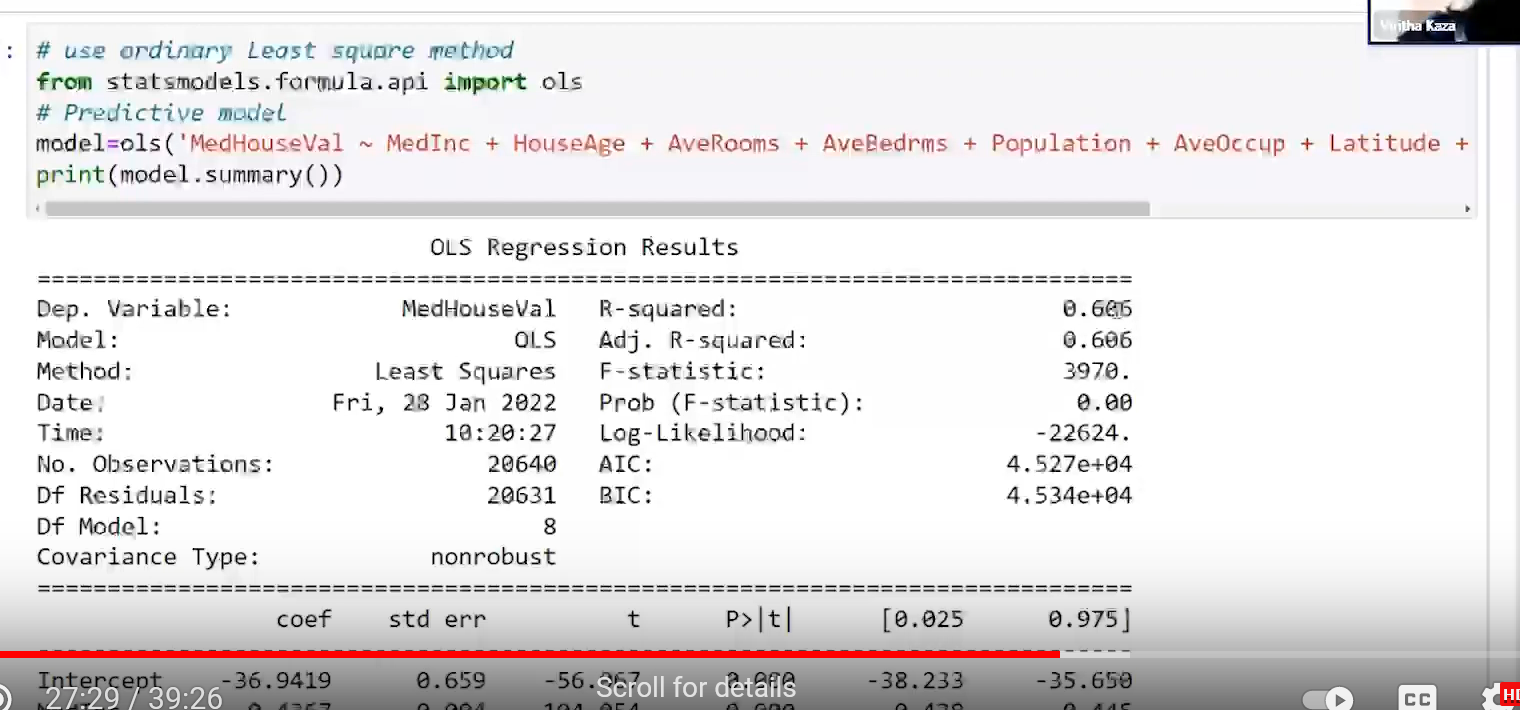
## Lesson – 2 [GM Tutorial: A Data Scientist's Perspective](https://youtu.be/79zPXUoolng)

Notes

1. Use .DESCR to describe
2. seaborn.regplot() :

This method is used to plot data and a linear regression model fit. There are a number of mutually exclusive options for estimating the regression model. For more information [click here.](https://www.geeksforgeeks.org/types-of-regression-techniques/)

1. .corr() ->heatmap
2. Ridge regression or LASSO – for multi-collinearity
3. Ols – ordinary least square method (predictive model)



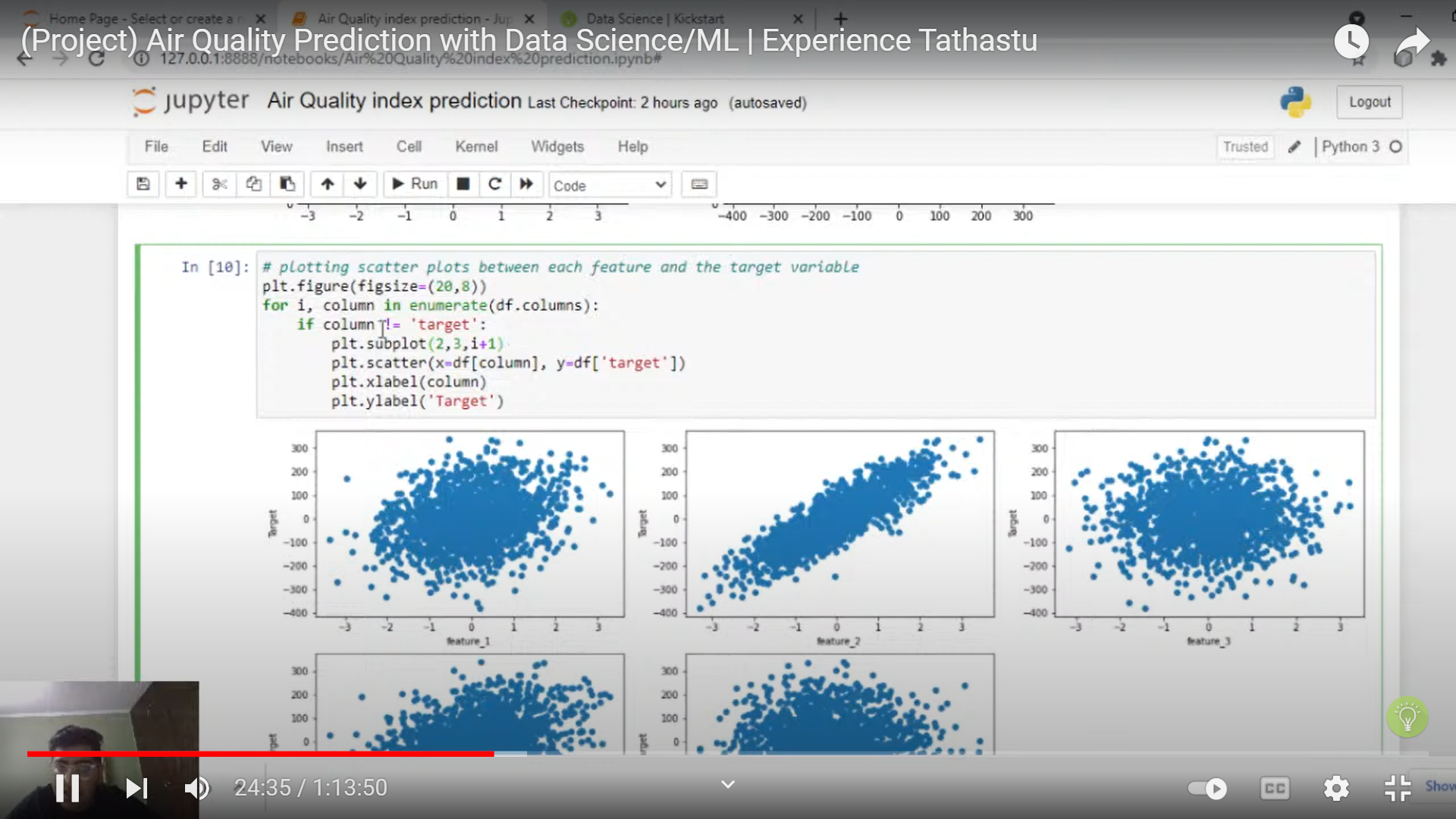
Use P value. If P value is closer to zero, they are significantly contributing

If P value is greater than 0.05, then the variable is not contributing to the dependent variable

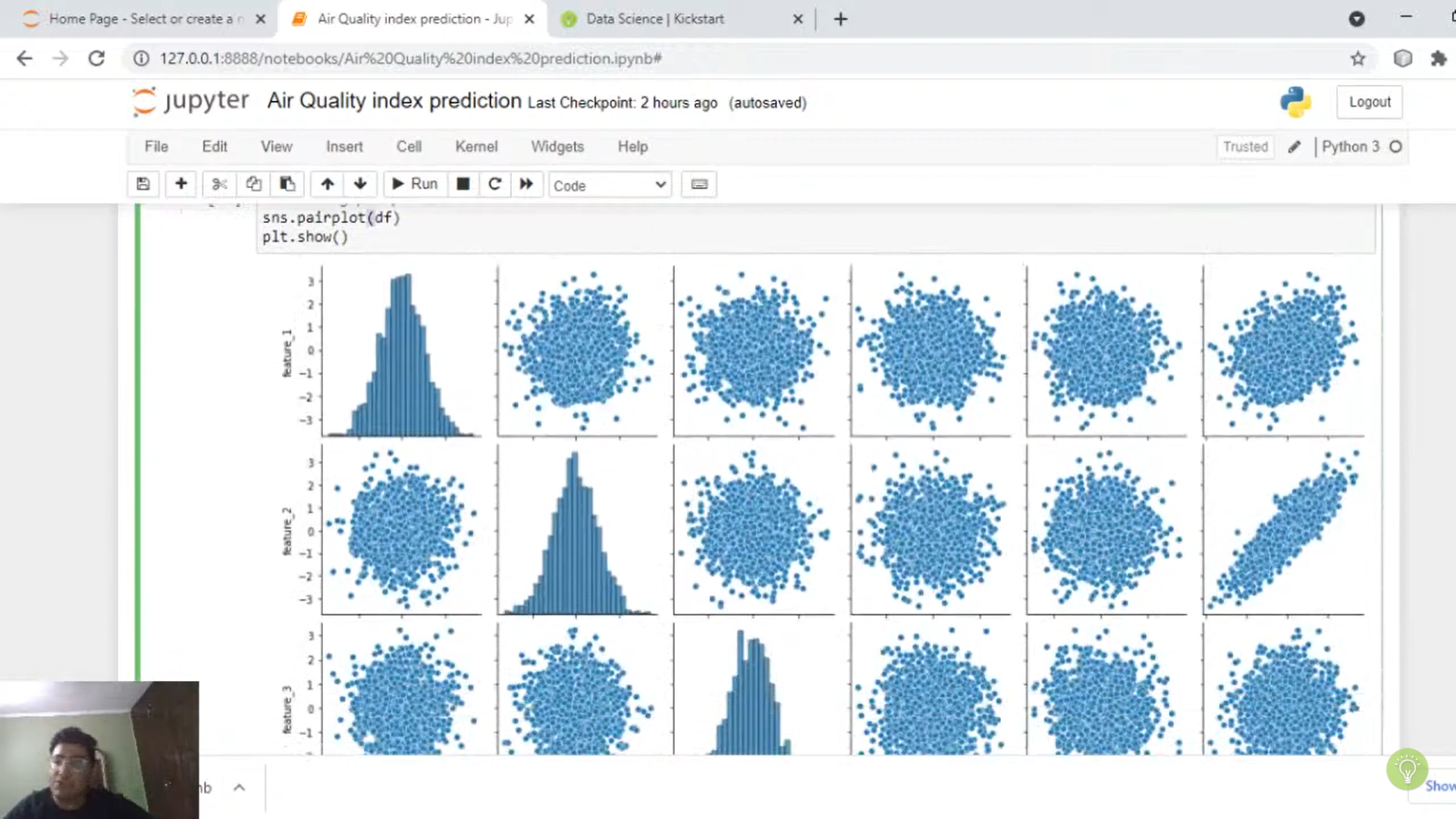
1. LocalOutlierFactor – to identifier the outliers. Have less no. of independent variables
2. Homoscedacity – to identify
3. Heteroscedacity - - use Weighted Least Square

Lessons from Youtube link

1. Cartopy – plotting location on map, what is cython?
2. When plotting locations in map, pass longitude first, then latitude
3. Try to use multiple scatter plot – one scatter plot for dependent variable with each independent variable – to show relationship
4. Use subplot to show horizontally
5. Try to use multiple histogram plots– one histogram plot for dependent variable with each independent variable – to show data distribution



1. Corr – correlations, heatmaps
2. Seaborn - Pairplot – multiple plots in one command



Find the list of suggested questions

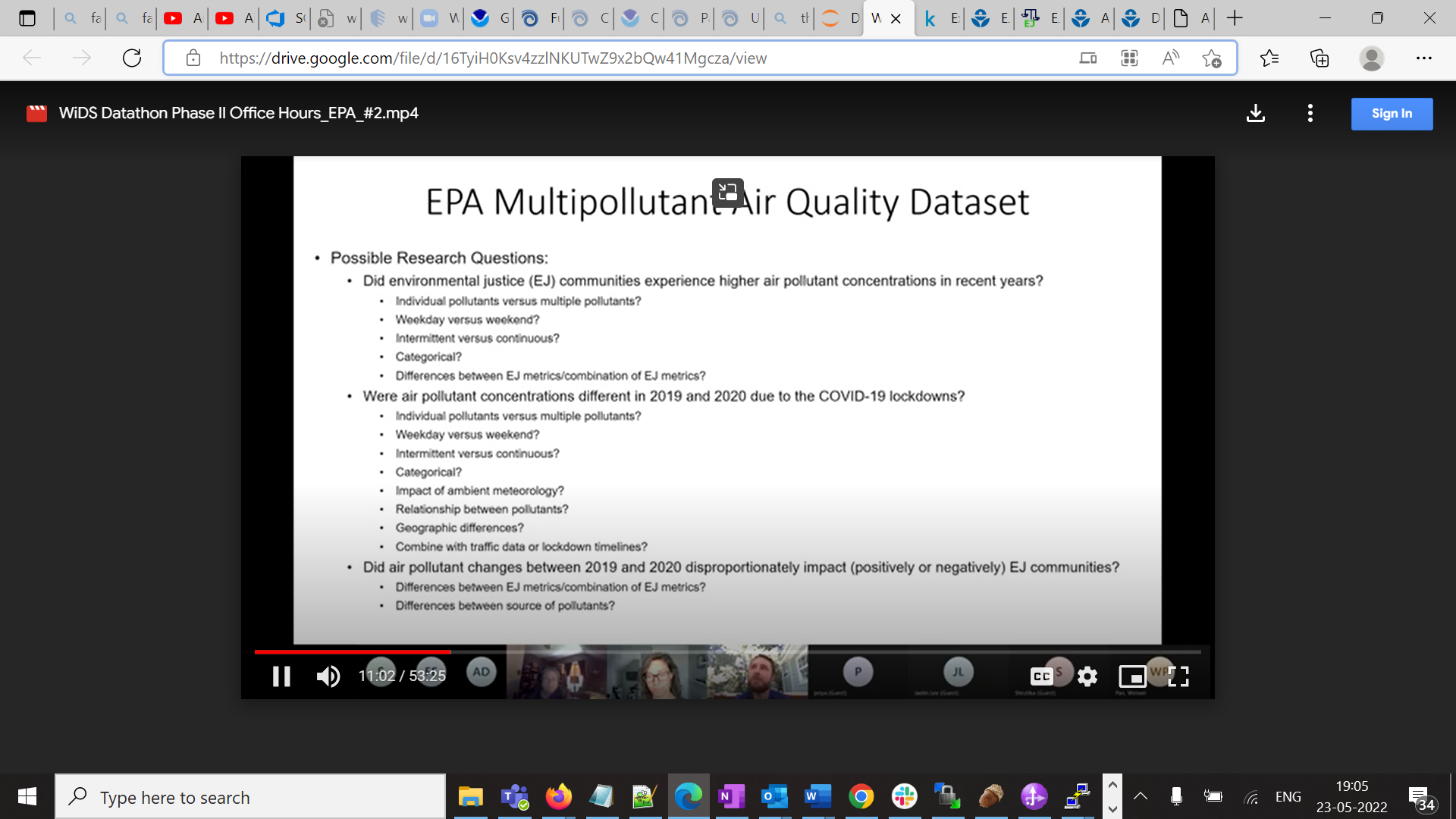
EPA

1. Is there any correlation between air pollutant concentration levels and climate change factors? (This will involve gathering additional data)

(ozone makes breathing difficult)

* + SO2, NO2, PPM, Vs Temperature, windspeed, ozone, humidity
  + Pick and choose the counties – (value\_counts() is among the top)

1. Did vulnerable and/or marginalized communities experience higher air pollutant concentrations in recent years? (2019, 2020 only)
2. Were air pollutant concentrations different in 2019 and 2020 due to the COVID-19 lockdowns?
3. Did air pollutant changes between 2019 and 2020 disproportionately impact (positively or negatively) vulnerable and/or marginalized communities?



EJ communities receive higher level of air pollutants - could be when, where , what kind of pollutants,

Regression type of analysis

Pollutant standards –, short term, long term

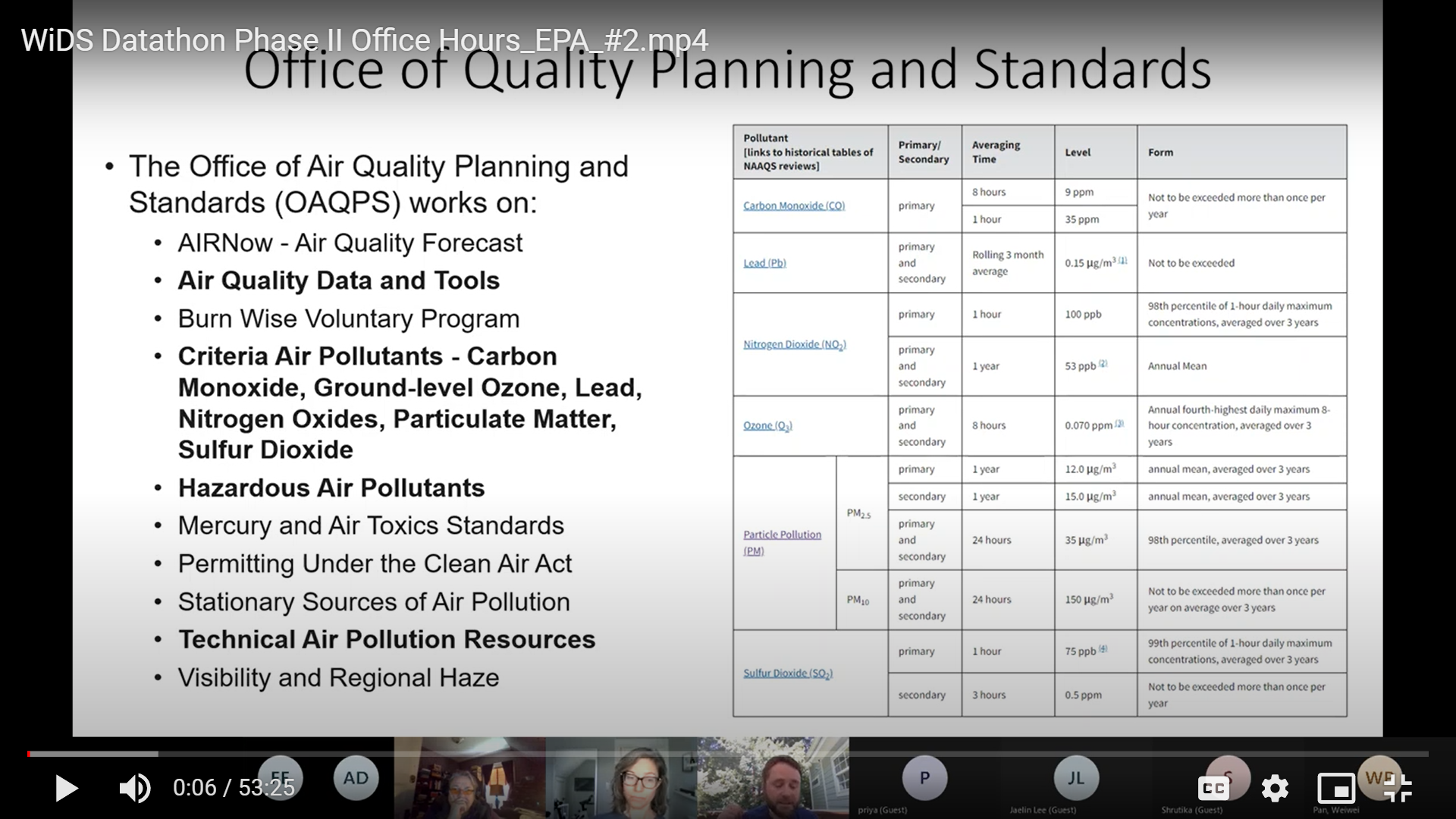
Were air pollutants in 2019 vs 2020 different due to lock down

Relationship between pollutants – sources – which sources were most, effect of weather

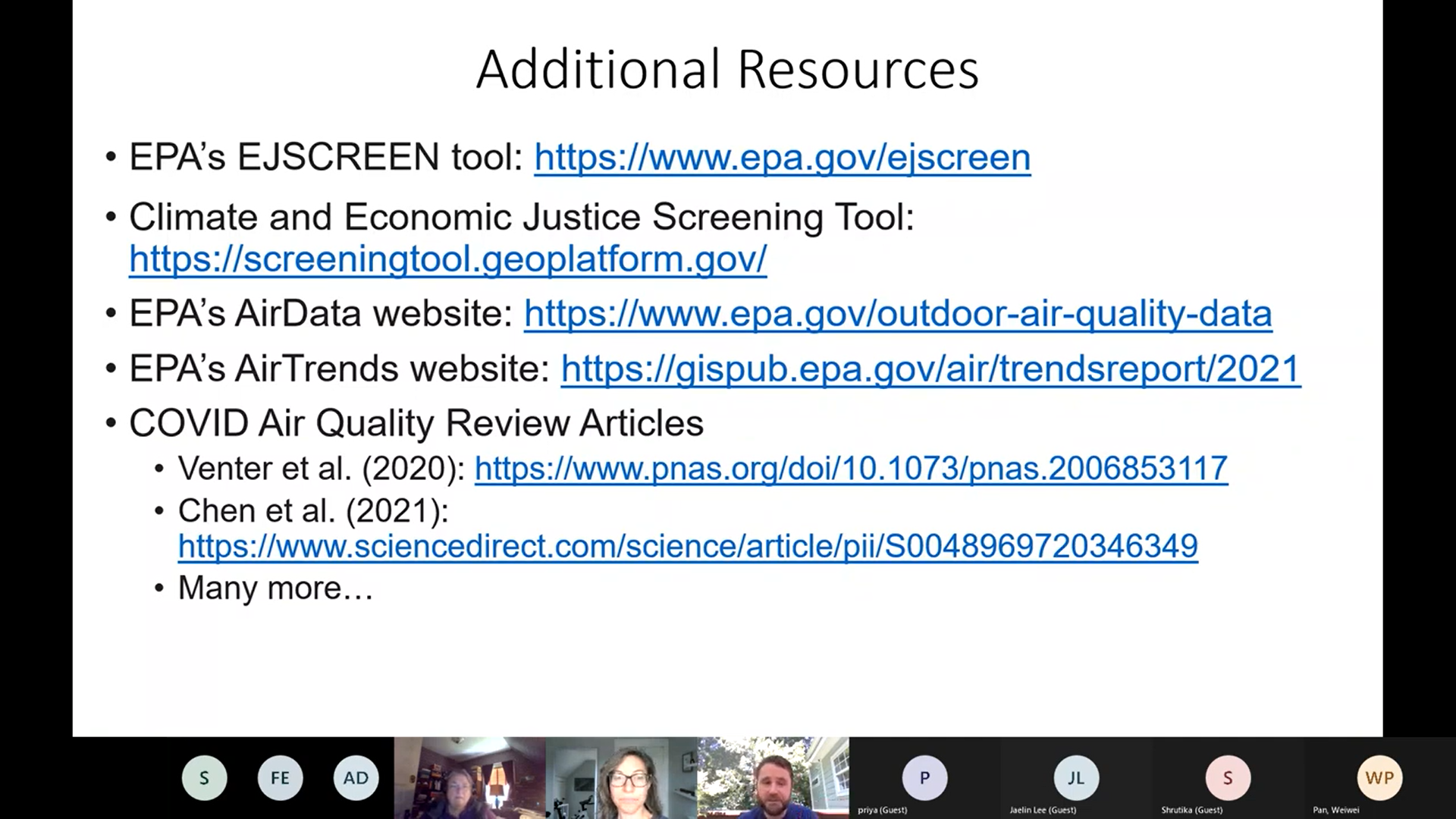
Climate Change AI – France

1. Generalization questions: for example, how well can we predict energy use across France?
2. Explainability/pattern recognition questions: for example, what are the main observed drivers of energy consumption?
3. Scenario questions: for example, how could energy use change if…?

EPA recording notes:



Data Sources:



AQI -

Action item:

1. Complete listening to the 2 office hours

**Target**:

1. What is the problem statement that we are going to address?
2. Identify the specific data set
3. What is objective of the project –
   1. Predict –
   2. Suggest –

# Parameters affecting Air Pollution

1. Dew
2. NO2
3. SO2
4. CO2
5. PM10
6. PM25
7. OZONE (O3)
8. Temperature
9. Weather?

To Do:

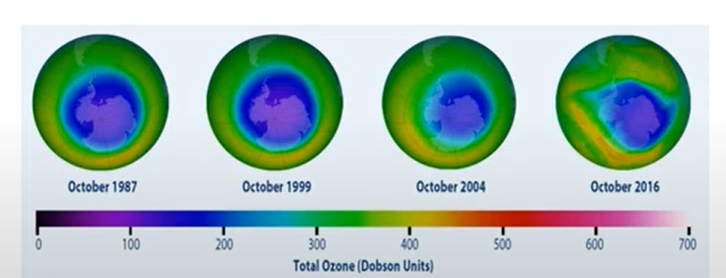
1. REST API – or direct data - use latitude and longitude information> CBSA>State
2. Check if CBSA has too many missing, then see if it can be ignored…
3. Identify the CBSAs that have valid data max and min
4. Try to identify the correlation between parameters, for CBSAs that have maximum number of data (very minimal missing data)

Univariate analysis? – analyse one independent variable

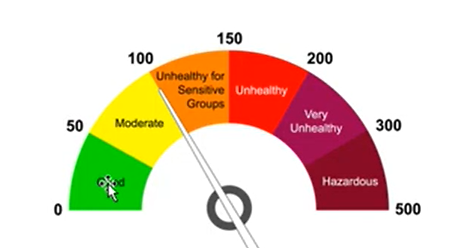
Bi/Multi-variate analysis – correlation between more than one

**------------------------RESTART------------------**

**Total Ozone Depletions**



**Air Quality Index:**



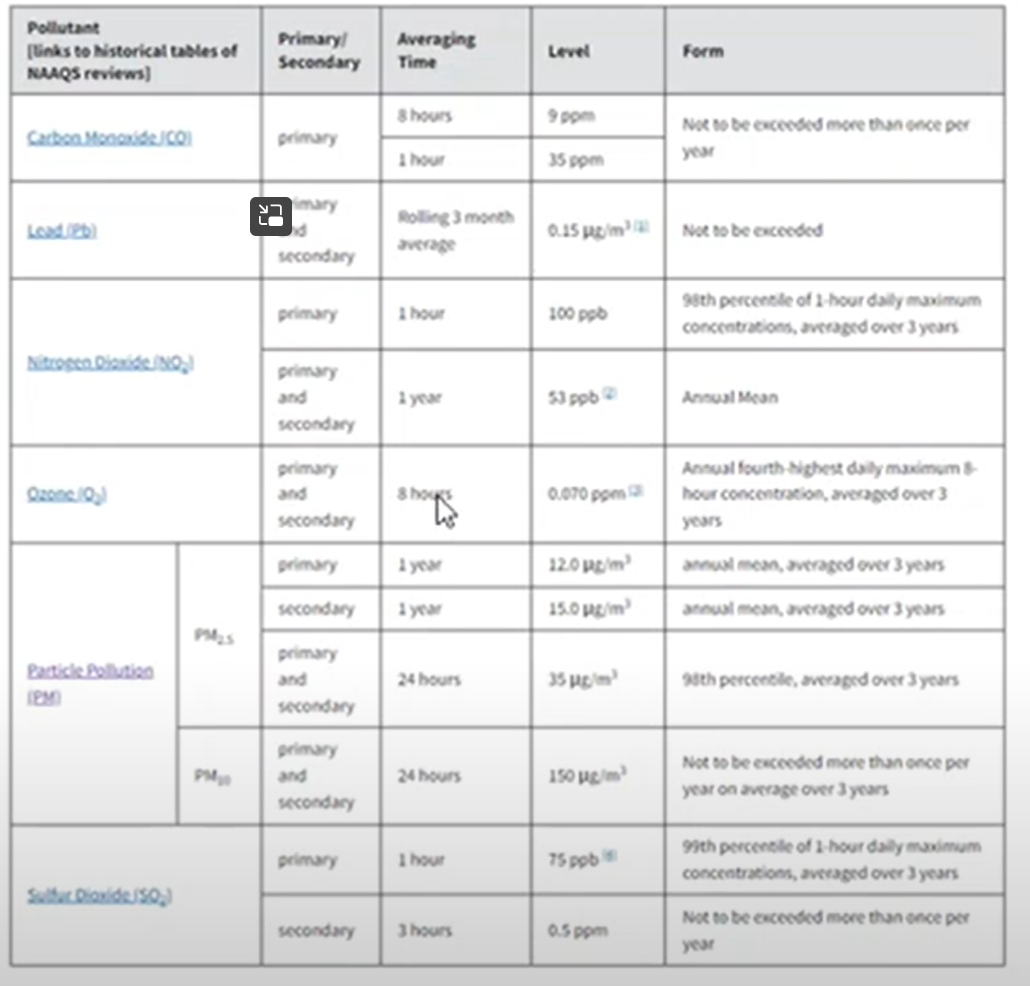
**Air Quality Planning and Standards:**

**FINAL important points:**

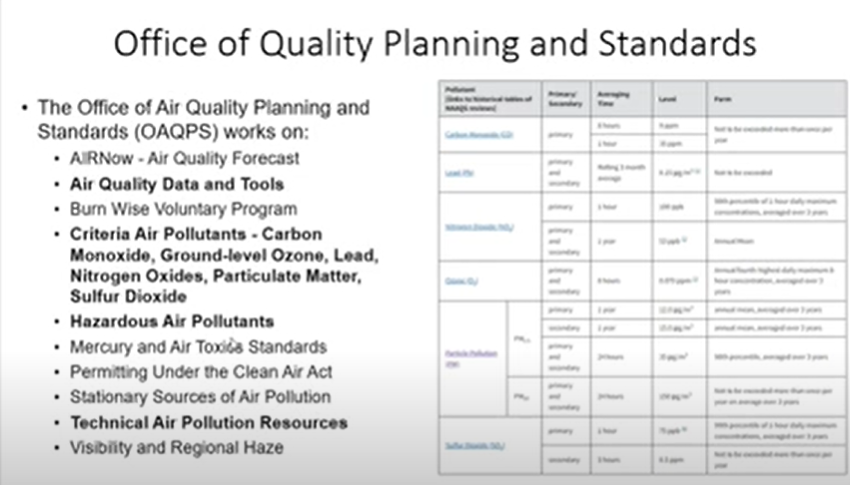
1. Data Set shared is EPA Multipollutant Air Quality DataSet
2. Classfication of data:

* Demographic information -
* Meteorological information – temperature, wind speed, wind direction, humidity
* Major Air pollutants – Ozone, PPM2.5
* Other Major Air pollutants – NO2, SO2, CO2, Pb (Lead), Benzene – come from Vehicles

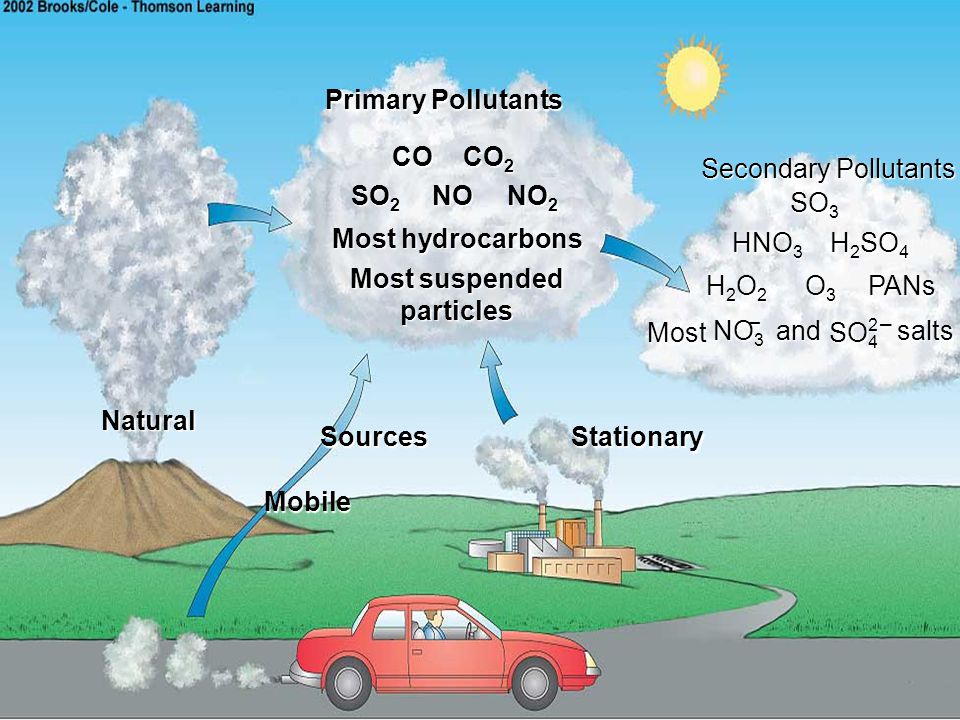
1. The permissible level of pollutants has been provided
2. Particularly particulate matter and ozone

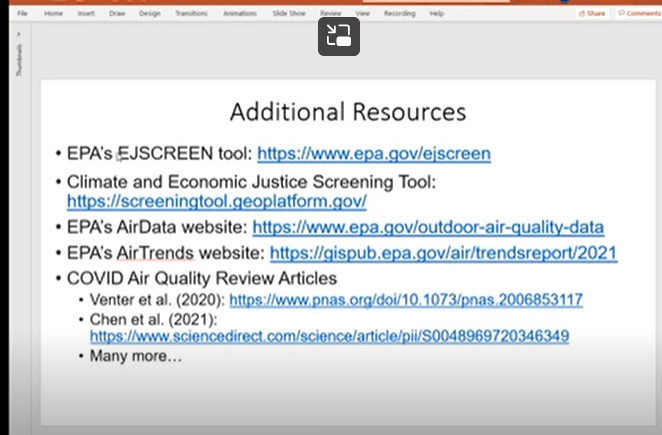


1. Criteria for Air Pollutants –



1. Domain Information:
   1. Hazardous pollutants – benzene, toluene, arsenic, formaldehyde
   2. Volatile Organic Compounds - benzene
   3. Visibility is also an important factor when considering Air pollution
   4. Fine particles –
   5. Pollutants fall in 2 categories
      1. Pollution is released –directly emitted into the atmosphere
         * These are radiated from the source itself
         * In the atmosphere, these are found in the original state.
         * e.g., Wildfire, chimney, cars,
      2. Secondary formations - Secondary pollutants are defined as the pollutants which form in the atmosphere. These secondary pollutants do not come directly from a source (like vehicles or power plants). It forms as a result of the pollutants the sources emit and reacting with the molecules in the atmosphere.
         * e.g., ozone – is never emitted, it is formed in the atmosphere Formation of Ozone. The reaction of UV radiation with O 2 or dioxygen molecules results in the formation of O 3. The UV radiation splits the oxygen molecule into the free oxygen or O atoms. These O atoms combine with molecular form of oxygen to form ozone (O 3 ).
   6. NH3- ammonia
   7. CO – carbon monoxide – incomplete combustion
   8. Certain sources like heavy vehicles – are major highways, near low income/minority neighbourhoods? – how do they impact these neighbourhoods



1. 
2. Climate Justice tool – screening tool
3. Benzene is the least frequent data in the data set
4. Missing data means – no measurement – not measured

Wild fires contribute to CO

**FINAL Preparation**

Expectations from the Office hours discussions

1. Most Important Criteria – Ozone and PPM2.5. Analysis should be directly related to this
2. The analysis should reflect the data given with the sources of pollution

To do (from our side)?

1. How to calculate Air Quality Index. (if AQI is identified for the given areas/data, can it be analysed). Can we compare Air Quality index between 2019 and 2020?
2. Look at possibilities of extrapolation of the data using data science models/tools. But not so high priority – should be really relevant – and should be comparable

**Good to have aspects:**

1. Visibility, haze