# Relevant Descriptions:

Project Timeline

• 10/26: Project proposal DUE

o Teams formed, Dataset selected, Submit based on the instructions

• 11/29: Project report & presentation slides due on Canvas (Date is FINAL)

• 11/30; 12/02: Project presentations

Project proposal

Should be 1-2 pages long and include:

• Student names

• Team name (optional)

• Project title

• Project description

• Proposed methodology/techniques/resources/datasets to be used (see details below)

o It is important to secure an adequately large dataset before committing to an idea.

General guidelines

The project should be an application that executes an in-depth data mining analysis

of a large dataset. You should integrate/combine one or more data mining techniques

studied in this course (Association analysis, Classification, Regression,

Clustering, Dimensionality reduction) into a solution for a specific problem.

Choosing a topic

In this project, the team is responsible for proposing a topic to work on. Even though

I am providing you with a list of suggestions, it is preferable (and more fun!) to pick a

problem “closer to your heart”. Just make sure that data are available that you can use

before writing the proposal (no data availability means that you will have to revise

your proposal and waste valuable time).

Therefore, it is highly recommended that you look at available data sets that look

interesting (and are large enough) and then try to formulate a problem/application

around them. Many links to datasets are posted on Canvas. However, you are free to

explore other options and/or consider writing a script that will collect data from

existing sources (again, you have to ensure that you’ll be able to do this before

committing to a project).

All the projects should include one or more (depending on the scope) data mining

algorithms. However, an analysis of some data is not sufficient. Your analysis must be

integrated in an application that has some purpose/use!

A few sample projects are (these are provided mainly to give you an idea of the scope

of the project):

• Design a program that will classify review comments (on products, blogs,

etc.) as spam/fake using sentiment analysis

• Design a predictive model for sports events or sport-related activities (e.g.

predict which team is going to win the super bowl).

• Design a program that will use/analyze data from social networks. There are

so many possible ways of doing this – e.g. analyze tweets in different US

states and find the “happiest” folks, then show results on a map.

• Use the data provided by a city (e.g. San Jose or San Francisco) and see what

interesting patterns & applications you can implement that would be useful in

a community.

• Use data collected by smart devices (e.g. activity, geo-location, etc.) and design

an application that helps the user in some way.

• Analyze video data from intersections to detect patterns in the flow of traffic,

analyze the categories of vehicles that travel on a given segment of a highway

at different times of a day, or design a traffic signal control system that uses

sensory input from video cameras to intelligently and dynamically adjust

traffic.

Topics to avoid (many others have done these before):

• Movie/Song/Book recommendation system

• Sentiment analysis/Word clouds on Twitter

• Anything using the WEKA GUI (too simple!)

Data

Datasets used for this project must have enough complexity where analysis via non-

data mining methods (e.g., using Excel) would not be possible. This can either mean

there are a lot samples and many attributes (e.g., > 200,000 samples and > 50

features, i.e., > 10M non-zeros) or that the data is heterogeneous and complex (e.g.,

mixing text profiles with extracted attributes from multiple sources).

Brainstorming:

UCI Machine learning repo :

<https://archive.ics.uci.edu/ml/datasets.php?format=&task=&att=&area=&numAtt=10to100&numIns=greater1000&type=&sort=nameUp&view=table>

List of suitable datasets:

1. <https://archive.ics.uci.edu/ml/datasets/Covertype>
2. <https://archive.ics.uci.edu/ml/datasets/PAMAP2+Physical+Activity+Monitoring>
3. Air Quality: <https://datasetsearch.research.google.com/search?query=air%20quality&docid=L2cvMTFqbnpweGdfYw%3D%3D>
4. Combine Annual Summary Data (Concentration by Monitor), from 2011 to 2020, would be well over 700k samples:

<https://aqs.epa.gov/aqsweb/airdata/download_files.html>

File Format: (<https://aqs.epa.gov/aqsweb/airdata/FileFormats.html#_format_3>)

## 

## Data Mining Techniques We Could Use:

Association analysis - If vehicles per household data is in census, could look at correlation between that and air quality levels in the area

Classification - Is a place recommended for people who have respiratory issues, can create a weighted metric based on various factors and classify as high risk, moderate risk, low risk → Bayes, (would just use the standard to assess risk)

**Regression** - Predicting future air quality levels for the subset of pollutants, and analyzing the level of changes predicted by county / state to understand air quality trends → Linear, Ridge, Lasso, Elastic Net, nearest neighbors regression, Bayesian regression (clear way to test accuracy, use some years as our training data, ex: 2001-2015, later years as our testing, ex: 2016-2020)

Clustering - Probably not relevant

Dimensionality reduction - Feature selection - We considered whether PCA would be possible to apply here, but given the nature of our dataset, and the fact that various columns were of different data types, we couldn’t come up with a satisfactory way to apply it

## Measures and Datasets We’ll Be Using:

For each of the measurements, we want to make sure that we’re preprocessing and only looking at the same parameter name, sample duration, and metric used.

Ex: “PM 2.5 - Local Conditions, 24 hr, Daily mean”

Important measurements: PM 2.5 for overall air quality, and specifically Nitrogen Dioxide NO2, Ozone O3, Sulfur Dioxide SO2, and Carbon Monoxide CO

## 

## Application:

Could combine with Site Description data, in particular land use and location setting, to find how these factors correlate with air quality in the area, can do data preprocessing and combine using pandas

[Only if time permits] Combining with census data, particularly income data, poverty rates, to see correlation between socioeconomic factors and air pollution

Could then do regression analysis to predict future air quality levels for the subset of pollutants, and analyse the level of changes predicted by county / state to understand air quality trends (helpfully, can cross-reference with EPA’s own trends page to some extent)

## Reference Links:

Gives Worldwide Historical Data:

<https://aqicn.org/data-platform/register/>

National Air Quality status and trends EPA page:

<https://www.epa.gov/air-trends>

Research with EPA data, looking at four major pollutants:

<https://data.world/data-society/us-air-pollution-data>

Important Pollutants and their unit/method of measurement:

<https://www.epa.gov/criteria-air-pollutants/naaqs-table>

Paper that gives some insight into amount of variability that can be explained for Air Quality Monitoring Stations PM 2.5 values:

<https://aaqr.org/articles/aaqr-20-05-oa-0217>

Past project using similar dataset (particularly

<https://towardsdatascience.com/predicting-air-pollution-with-prophet-on-gcp-42ceb1625818>

Article, that might help in feature importance

<https://towardsdatascience.com/earth-thanks-to-covid-19-i-can-breathe-easier-ebc2c11337c5>

Interesting articles on Air pollution analysis

<https://towardsdatascience.com/tagged/air-pollution>

How data science gives new insight into air pollution in the US

<https://news.mit.edu/2021/clear-data-clear-air-data-science-gives-new-insight-air-pollution-us-0602>

Link to dataset:

<https://aqs.epa.gov/aqsweb/airdata/download_files.html>

Useful(?) vehicles data (some level of granularity):

<https://www.bts.gov/ctp>

Cars per household visualization for 200 most populated cities (2015-2016):

<https://www.governing.com/archive/car-ownership-numbers-of-vehicles-by-city-map.html>

Cars per household data for cities with population >100k (2015-2016):

<https://cdn.vanderbilt.edu/vu-my/wp-content/uploads/sites/1478/2019/10/29104657/Vehicle-Ownership-in-U.S-2016.pdf>

National cars per household data (only overall listed in table, 2019):

<https://data.census.gov/cedsci/table?q=DP04&tid=ACSDP1Y2019.DP04>

County Dataset:

<https://www.kaggle.com/stansilas/us-state-county-name-codes>

Population Count by County:

<https://www.census.gov/data/tables/time-series/demo/popest/2010s-counties-total.html>

EPA’s AQI:

<https://www.airnow.gov/aqi/aqi-basics/>

## 

## Initial Data Exploration:

Create Jupyter notebook and create dataframes for a couple of years, to answer the question: can we do regression analysis and make a prediction of air quality values for 2022? Is there enough data for any of these parameters to do so?

## TO-DO: Pre-processing Steps:

* Working in Google Colab here
* Uploading all the daily summary files:
  + PM 2.5 (Lasya), Sulfur Dioxide (Lasya), Ozone (Lasya), Nitrogen Dioxide (Auni), and Carbon Monoxide (Varsha)
* Identifying which columns are relevant for each measurement:
  + PM 2.5 (arithmetic mean and AQI)
  + Ozone (Auni), Sulfur Dioxide (Lasya), Nitrogen Dioxide (Auni), and Carbon Monoxide (Varsha)
* Averaging to get MONTHLY values for relevant measurements for each unique national monitoring site (grouping by State Code, County Code, and Site Number)
  + - Grouping by address does show that some sites have >356 values or <12 values, so that does mean some of these sites did more measuring and some did way less measuring, so we may not have monthly values for all sites, just something to keep in mind
  + That way, we have 12 data points for each site in a given year, and 240+ in the duration of our project for any given monitoring site for each of our 5 air quality measurements, which should be plenty for us to do logistic regression / association analysis
* After getting monthly values per each site in any one given year, simply repeat for all 20 years, and merge the values into one big file
* 2010-2017 - Training
* 2018-2020 - Test

# 

# Proposal:

Should be 1-2 pages long and include:

• Student names

• Team name (optional)

• Project title

• Project description

• Proposed methodology/techniques/resources/datasets to be used (see details below)

o It is important to secure an adequately large dataset before committing to an idea.

**CMPE 255: Group 9: Project Proposal**

**Team Members:**

Varsha Suresh (015245347)

Lasya Bheemendra Nalini (014618695)

Auni Bagchi (011256895)

**Project Title: Predicting Air Quality and Pollutant Levels in the US**

**Project Description:**

Through this project, we aim to predict the level of air pollution in each US county based on past trends. We will also be performing an association analysis between the pollutant levels and each of land use, population in the area, and location setting to determine how the levels vary in relation to these factors.

These analyses and predictions will help government bodies as well as residents in mitigating and addressing current and future local pollution levels. It will also give people a better understanding of overall air quality, which is necessary in order to take the steps towards ensuring a cleaner environment and healthier population.

The results from our analyses and predictions will be presented in a US map plot highlighting the pollutant levels across counties. We will be using EPA’s AQI as a reference to standardize our findings and display them.

The air quality, land use, and location setting data are sourced from the EPA, and will be collated for a period of 20 years for the purposes of our analysis. The population data and estimates will be sourced from the US Census Bureau, looking at a period of 20 years for the purpose of this project. A description of land use and location setting in the context of our project can be found in the following table:

| Land Use | A category describing the predominant land use within a 1/4 mile radius of the site (ex: residential, commercial, agricultural, forest, desert, etc.) |
| --- | --- |
| Location Setting | A description of the setting within which the monitoring site is located (ex: rural, urban and city center, suburban, etc.) |

**Methodology:**

* Pre-processing:
  + Collate daily summary pollutant data for all years
  + Feature subset selection from daily summary pollutant data to retain relevant data
  + Combine pollutant data with site description data based on FIPS code
  + Combine pollutant data with census data based on FIPS code
* Data Mining:
  + Regression analysis to predict pollutant levels by county and state level
  + Association analysis between air quality and population, air quality and location setting, air quality and land use
* Visualization:
  + Plot past and future levels of pollutant levels on a US Map, displaying gradients based on EPA’s Air Quality Index

**Datasets:**

* Combined 2000 - 2020 Daily Summary Data for Critical Gases, including Ozone (O3), Sulfur Dioxide (SO2), Nitrogen Dioxide (NO2), and Carbon Monoxide (CO)
  + Sample for one year, we will be combining 20, 2020 Ozone: <https://aqs.epa.gov/aqsweb/airdata/daily_44201_2020.zip>
* Combined 2000 - 2020 Daily Summary Data for Particulates, PM 2.5 FRM/FEM Mass
  + Sample for one year, we will be combining 20, 2020 PM 2.5 FRM/FEM Mass: <https://aqs.epa.gov/aqsweb/airdata/daily_88101_2020.zip>
* Site Description Data, in particular Land Use and Location Setting
  + <https://aqs.epa.gov/aqsweb/airdata/aqs_sites.zip>
* Combined 2000 - 2020 Annual Summary Data Concentration by Monitor
  + Sample for one year, we will be combining 20, 2020: <https://aqs.epa.gov/aqsweb/airdata/annual_conc_by_monitor_2020.zip>
* Census County Population Totals, 2000 - 2019
  + 2010 - 2019: <https://www2.census.gov/programs-surveys/popest/tables/2010-2019/counties/totals/co-est2019-annres.xlsx>
  + 2000 - 2010: <https://www2.census.gov/programs-surveys/popest/datasets/2000-2010/intercensal/county/co-est00int-tot.csv>

**Sources:**

* Pollutant data and Air Quality Index details: United States Environmental Protection Agency (EPA)
  + <https://aqs.epa.gov/aqsweb/airdata/download_files.html>
* Census data: The United States Census Bureau
  + <https://www.census.gov/data/tables/time-series/demo/popest/2010s-counties-total.html>