# SAS project 2019-2020: Exploratory analysis of the US pollution data

The data file **pollution\_us\_2000\_2016.csv** was downloaded from

<https://www.kaggle.com/ksaulakh/r-analysis-pollution-data>

Information on reported values are given on <https://www.epa.gov/outdoor-air-quality-data/about-air-data-reports>. The focus will be on daily mean values of sulphur dioxide (SO2) or ozone (O3) and their air quality index (AQI) values.

The climate data **avg\_max\_temp.csv** was downloaded from <https://www.ncdc.noaa.gov/ghcn/comparative-climatic-data>

The data file **us-state-ansi-fips.csv** provides state names, state name FIPS codes (st) and postal abbreviations (stusps).

More information on air quality index thresholds are given on <https://airnow.gov/index.cfm?action=aqibasics.aqi>

Answer following questions:

1. Import the 3 data sets.
2. Using the climate data, calculate the mean temperature value in each state and month and output the results.
3. Starting from the output data from question 2, use an array statement to obtain the average monthly temperature data from each state in a long format.
4. Merge the data table obtained in Question 3 with the FIPS code data set using **stsups** as by variable and outputting only matching rows.
5. Merge the data table obtained in Question 4 with the pollution data. *Hint: you’ll have to create a month variable in the pollution data set first and then perform a merge by state and month.* Output only matching rows.
6. Use an appropiate procedure for comparing the number of levels of **county\_code** and **county**. What do you observe? Write down your findings. Remedy the problem by creating a unique county code by concatenating state\_code and county\_code separated by an underscore. Search for a SAS function that can do this.
7. Using the data set obtained from question 6, create a permanent data set keeping all observations from the state **PA** and the variables
   * County
   * County\_Code
   * Stname
   * Stusps
   * St
   * Date\_Local
   * Month variable from question 5
   * **O3\_units**
   * **O3\_mean**
   * **O3\_AQI**
   * Temperature variable from question 3
8. Calculate average values of **O3\_mean** and **O3\_AQI** for each county and month and include the table in your report. Use the tabulate procedure. Make use of the **Date\_local** variable in combination with an appropiate format. Report 3 decimals. Discuss.
9. Plot the average concentration of **O3** in different counties, using barplots, in descending order. Use appropiate axis titles. In case you are trying the unicodes for subscripts, use **labelattrs=(family='Times New Roman')** since the unicode will not work for the standard font. Include the graph in your report and discuss.
10. Calculate the yearly average of the **ozone** concentration for each county and output these values to a SAS data set. *Hint: Use an appropiate format statement to achieve this*. Using this outputted data set, use the HEATMAPPARM statement in PROC SGPLOT to create a continuous heat map with **Date\_local** in the x-axis, **county** in the Y-axis and the average **ozone** concentration as response. *You might want to have a look at the SAS blog from Rick Wicklin* [*https://blogs.sas.com/content/iml/2019/07/15/create-discrete-heat-map-sgplot.html*](https://blogs.sas.com/content/iml/2019/07/15/create-discrete-heat-map-sgplot.html)*.* Include the heatmap in your report and discuss.
11. Using again the data set created in question 7, draw a lineplot that displays **O3\_AQI** over the years for the 3 counties with highest concentration as concluded from question 9. When plotting, take monthly mean values as the representative values by using an appropiate format for **Date\_local**. The Y-axis limits should be 0 and 220. Draw a reference line at 50, 100, 150 and 200 labeled *good*, *moderate*, *unhealthy for some groups* and *unhealthy*. The following SAS paper covers this topic: <http://support.sas.com/resources/papers/proceedings09/158-2009.pdf> Include the graph in your report and discuss.
12. Subset the data: select the observations from the year 2015 and verify whether there is an association between temperature and **O3\_AQI.** Report and discuss your findings.
13. Write a pdf report with the ODS system of max 5 pages containing the following:

* Names of the group members with a mentioning of who did what.
* A short introduction on the topic.
* Answers to the questions with requested plots and tables inserted.
* The name of the report should be as follows: **SASproject\_groupX.pdf**. The code should be sent as a separate file named **SASproject\_groupX.sas.** This code is expected to be between 300 and 400 lines.