1 Data

The data in this problem comes from a state's environmental health screening tool. Data were collected for each of 1,581 Census Tracts across 9 counties. (The U.S. Census Bureau has partitioned the entire country into geographical regions called census tracts that contain approximately the same number of people.) Census Tracts with missing or corrupted information have been removed, reducing the data to 1,497 Census Tracts. The raw data for those Census Tracts are provided in `Student_Data_Untransformed.csv`. The variables are listed in Table 1. They include exposure indicators (ozone, PM2.5, diesel PM, drinking water contamination, lead risk, and toxic releases), environmental effects indicators (groundwater threats, hazardous waste facilities, impaired water bodies, and solid waste facilities), and socioeconomic factor indicators (education, linguistic isolation, poverty, and unemployment).

In order to prepare input variables for effective modeling, I've already applied transformations to the predictor variables as shown in Table 2. You can reverse the transformation by applying the inverse functions listed if you need to. I recommend that you use these pre-transformed values, provided in `Student_Data.csv`, to build your models. Some of these transformations reverse the direction of effects (e.g., if $x_t = x^{-1}$, then as x increases, x_t decreases). Take this into account while discussing model effects. The response variable has not been altered, and you should still consider transformations to the response variable if they are needed.

`Asthma` is the response variable for this study. In the data file, there is no value for `Asthma` for the last 100 rows. The response values from these rows were held back by the client as a test set. You will calculate predictions and prediction intervals for those Census Tracts.

2 Task

You have been commissioned by the Air Force Exceptional Family Member Program (EFMP) office to build a predictive model that they can apply to estimate the risk of severe incidents for children with asthma. They are willing to assume that these 9 counties are representative of other areas. Due to concerns regarding fuel spills near bases affecting groundwater, leaders are particularly interested in assessing the effect (if any) of groundwater threats.

Your task is to analyze the data for the 1,397 census tracts for which you have complete data and construct *two or more* good regression models for predicting `Asthma`, which is a measure of the rate of child emergency incidents for asthma. Each model should explore different methodologies (not just stepwise regression). Compare these models' performance and select a recommended model. You may include additional explanatory variables constructed from functions of the variables on the data file if you think that they are worthwhile. *No raw Python output should be present in the report.* Summarize your analysis in a report in PDF format that includes the following sections. This simulates how a real-world analytical report would be structured.

1. A 1–2 paragraph "Executive Summary" of your major conclusions about the relationships between asthma emergency rates and the explanatory variables. You should specifically address Groundwater Threats whether it is included in the model or not. This should not contain any formulas or mathematical symbols. It should be written so that it could be easily understood by an Air Force senior leader with no formal training in statistics.
2. A description of the steps taken to identify your model(s) and select your recommended model. *Do not submit any raw Python output in this section.* Graphical analysis and summary statistics are encouraged. Simply outline the issues you considered, your decisions, and the sequence of steps you took to develop a model. Be detailed — tell me what you did, why you did it, and if it worked.
    - For the purposes of this course, consider only models with main effects, quadratic effects (X^2), and two-factor interactions. This is already a substantial design space to work with.

Variable	Description
`TotalPopulation`	Population in the community
`Ozone`	Mean of summer months (May-October) of the daily maximum 8-hour ozone concentration (ppm)
`PM25`	Annual mean concentration of PM2.5 (particulate matter with diameter of 2.5 μm or less), in $\mu g/m^3$
`DieselPM`	Spatial distribution of gridded diesel PM emissions from on-road and non-road sources (tons/year)
`DrinkingWater`	Drinking water contaminant index for selected contaminants
`Lead`	Percentage of households with likelihood of lead-based paint (LBP) hazards from the age of housing combined with the percentage of households that are both low-income (household income less than 80% of the county median family income) and have children under 6 years old
`ToxRelease`	Toxicity-weighted concentrations of modeled chemical releases to air from facility emissions and off-site incineration
`Traffic`	Sum of traffic volumes adjusted by road segment length (vehicle-kilometers per hour) divided by total road length (kilometers) within 150 meters of the census tract
`GroundwaterThreats`	Sum of weighted scores for sites within each census tract, taking into account information about type of site, its status, and its proximity to populated census blocks
`HazWaste`	Sum of weighted permitted hazardous waste facilities, hazardous waste generators, and chrome plating facilities within each census tract
`ImpWaterBodies`	Summed number of pollutants across all water bodies designated as impaired within the area
`SolidWaste`	Sum of present solid waste sites and facilities, weighted by distance from the nearest populated census blocks within a Census Tract
`Education`	Percentage of the population over age 25 with less than a high school education
`LinguisticIsolation`	Percentage of limited English-speaking households
`Poverty`	Percent of the population living below two times the federal poverty level
`Unemployment`	Percentage of the population over the age of 16 that is unemployed and eligible for the labor force.
`Asthma`	Spatially modeled, age-adjusted rate of Emergency Department visits for asthma per 10,000

Table 1: Variables used in `Student_Data_Untransformed.csv`

Variable	Transformation	Inverse
`TotalPopulation`	$x_t = \sqrt{x}$	$x = x_t^2$
`Ozone`	$x_t = x^{-1}$	$x = x_t^{-1}$
`PM25`	$x_t = x^{3.5}$	$x = x_t^{(1/3.5)}$
`DieselPM`	$x_t = \ln(x)$	$x = e^{x_t}$
`DrinkingWater`	$x_t = \ln(x)$	$x = e^{x_t}$
`Lead`	No transformation applied	
`ToxRelease`	$x_t = x^{(1/3)}$	$x = x_t^3$
`Traffic`	$x_t = \ln(x)$	$x = e^{x_t}$
`GroundwaterThreats`	$x_t = (x + 0.01)^{(1/4)}$	$x = x_t^4 - 0.01$
`HazWaste`	$x_t = (x_t + 0.01)^{(-1/4)}$	$x = x_t^{-4} - 0.01$
`ImpWaterBodies`	$x_t = (x + 0.01)^{(1/4)}$	$x = x_t^4 - 0.01$
`SolidWaste`	$x_t = (x + 0.01)^{(-1/2)}$	$x = x_t^{-2} - 0.01$
`Education`	$x_t = x^{(1/4)}$	$x = x_t^4$
`LinguisticIsolation`	$x_t = \sqrt{x + 0.01}$	$x = x_t^2 - 0.01$
`Poverty`	$x_t = \ln(x)$	$x = e^{x_t}$
`Unemployment`	$x_t = \sqrt{x}$	$x = x_t^2$

Table 2: Transformations used in `Student_Data.csv`

- Because of this model limitation, it is possible that there may be some higher order effects and/or non-linearity in the data that you cannot model. Remember this when looking at residual plots. Point out if you think there may be deficiencies caused in this way. `Asthma` can still be transformed as you see fit if you find it necessary/useful.

3. A formula for your best model(s), standard errors for coefficients, and the R^2 value. The formula should be mathematical, not Patsy language. You may rename variables to limit the length of this formula, but define your variables (e.g., `TotallPopulation` $= x_1$, `Ozone` $= x_2$, ...). Summarize Python results in tables of your own creation — do not report any values that you do not intend to discuss. Discuss and interpret any important features of your model. Pay some attention to the `GroundwaterThreats` variable as a predictor, although you may conclude that it is not important.

4. Convincing evidence that the model you selected is a good model for using some or all of the explanatory variables to predict asthma emergency rates. Discussion of residual plots and other diagnostic checks would be appropriate. Statistical tests should be formulated correctly with appropriate hypotheses and conclusions. Graphs and tables are encouraged, but raw Python output or screenshots should not be submitted.

5. (Optional) One paragraph outlining additional analyses that you would have done if you had more time or were not artificially restricted in your model parameters. You will earn points for suggestions with high potential value, but you will also lose points for suggestions with little potential value. This optional piece simulates "above and beyond" work in a sponsored study, which can add to or detract from a report.

Separate from the PDF report, submit a CSV file with your predictions for the missing `Asthma` data points (the last 100 observations). Use your final, best model to predict `Asthma` and create a 90% prediction interval for each point. Points for the "Predictive Ability" section will be based on the following

1. Root Mean Square Error (RMSE) for the test set — lower is better

2. Coverage of your confidence intervals — ideally, 90 of the 100 data points should fall within your intervals

3. Width of your confidence intervals — narrow is better as long as it still covers the data (see previous point)

3 Deliverables

1. `last_first.pdf` — PDF document with your write-up. *You should have no raw Python output in this document.* This should be a well-structured report with narrative, graphics, and tables as needed.
2. `last_first.csv` — CSV file with your predictions for the last 100 observations. This should have the following columns with the following names:
   (a) `Census Tract`
   (b) `Prediction`
   (c) `Lower Prediction CI`
   (d) `Upper Prediction CI`
3. `last_first.ipynb` — Python file with your complete analysis, including plot generation, statistical tests, and predictions. Only include relevant code and comments and/or narrative blocks to explain the code, just as it would be delivered to a client. I should be able to run it top-to-bottom without errors. It should *not* be the digital equivalent of scratch paper.

4 Grading

This is an individual assignment, and you are expected to do your own work. Do not discuss this project with anyone other than the course instructor. The primary task of this assignment is to write a report detailing how and why you came to a regression model relating asthma emergency rates to predictor variables. Write a coherent and concise report that flows well and clearly describes you analysis and conclusions. There is no absolutely best answer, and I expect to receive many different models.

This final will be worth 100 points. You will be graded on:

1. Writing (20 points): Emphasis on precision, clarity, and efficiency. You should use paragraphs, transitions, sections and incorporate any figures and tables into the flow of the document. I will count off for any raw Python or screenshots; if you cannot explain what your code is doing in words, you do not understand your code.
2. Executive Summary (20 points): Clear and concise use of language to convey your model in a limited space.
3. Model Building Process, Logic, and Conclusions (40 points): Appropriate use of tools from this course applied correctly and communicated effectively. I should be able to replicate your process based on your descriptions without any raw Python or screenshots.
4. Predictive Ability (20 points): You will provide point estimates and confidence intervals for the withheld test points. Coverage of true values in those intervals, interval width, and RMSE will determine this score.