**DESCRIPTION**

Build a model to predict cancer mortality rates in United states of America. The data dictionary for the factors that can help in cancer mortality rates across different states and counties is provided along with the problem statement in a separate excel sheet.

### ****Background****

These data were aggregated from a number of sources including the American Community Survey ([census.gov](http://census.gov/)), [clinicaltrials.gov](http://clinicaltrials.gov/), and [cancer.gov](http://cancer.gov/). Credits to *Noah Rippner* for consolidating this data.

### ****Your Task: Build a multivariate Regression model to predict****"TARGET\_deathRate"

### Some requirements while modelling the solution –

### This is not a hackathon. While the RMSE/R-squared error of the solution is important, each step towards the final solution is even more important. Hence it is required to indicate the logical flow through documentation Ex – What you are doing in the step ? Why you are doing the current step ? Statistical / Logical Inference from each step …

### Copied solutions will be Disqualified / Rejected.

### You can refer python library API’s/code snippets from internet during the course of solving the problem.

### The expected deliverables are indication of what is essential to showcase the solution but not limited to the below points. You could go beyond these steps to put your solution in perspective.

##### ****Deliverables :****

a. Your final model equation. Significance of the Beta values or co-efficient.

b. The statistical software output including **(adjusted) R-squared and Root Mean Squared Error (RMSE)**

c. Your python Notebook (.ipynb). Ensure that the code and solution is visible in the notebook before submissions. Note - Comments is highly encouraged in the python code.

d. Before modelling split the Train / Test with 80 / 20 using Random seed = 42. You can choose to do validation using the Train data set. Solutions not following this will not be considered for evaluation or rejected.

d. Model diagnostics including statistics and visualizations:

- Assess linearity of model (parameters)

- Assess multicollinearity

e. Model selection (Optional step):

- Implement k-fold (k=10) cross validation

f. Your interpretation of the model

g. Other important factors to consider:

- Are there any outliers?

- Are there missing values?

- How will you handle categorical variables ? Are there any categorical variables in this problem ?

**Exploratory Data Analysis(EDA) :**

* Univariate/Bivariate analysis
* Important Visualizations and interpretation of the same..