Problem Statement

The "Big" IDEA:
To find the correlation between a **chronic disease- Household income - Presence of industrial establishment** for a specific location among over 500 cities in the United States.

Why should others care about it?

a) Household income and its effect on Health has a very strong correlation. Along with this, the Industrial establishments makes it worse.
b) City health authorities can understand the burden in their jurisdictions and help assist in planning better public health interventions.

How did you choose this problem?

- Even in the presence of good health infrastructure and schemes in the US, its affordability plays a vital role.
- Households having low income plus living in a toxic environment has a drastic effect on physical health.

This is a pressing issue and we wanted to analyse and explore more about it.

HYPOTHESIS SO FAR:
The main idea is to find the correlation between the household income, its incurred lifestyle and choice of living near an industrial establishment on the physical well being of a person.

Data and its Sources

Data Source Agency: The Centers for Disease Control and Prevention (CDC), US Health and Human Services

Data hosting Website: https://www.data.gov/

Description of Dataset and available formats: https://tinyurl.com/yxzpkrtq

Downloadable link for JSON file: https://tinyurl.com/y6gt5cxd

We are using "500 Cities Local Data for Better Health" which was obtained from US Government Open (https://www.data.gov/) Data Sets.The data is available in CSV and JSON format which is updated regularly.

Dataset specifics:
Size: About 810K rows and 24 columns
Types of Attributes (A GIST);
Nominal: CityName; GeographicLevel; Category; UniqueID; Measure; Short_Question_Text; CategoryID
Ordinal: StateAbbr
Quantitative: Year; GeoLocation; PopulationCount
Primary Classification: Prevention, Unhealthy behaviors and Healthy outcomes

Solution Strategy

How do you plan to approach the problem?

- Exploratory data analysis using bar-graphs, histograms and various other visualization technique available.
- Validating the hypothesis with the results of visualizations.
- Then extracting only the relevant data pertaining to a specific city and implementing machine learning algorithms for further insights
- Using predictive modelling techniques and proposing a possible event/outcome

What is the proposed scope of your project and the next steps?

- Utilize and demonstrate the learnings in all the stages of a data science project development life cycle.
- Our ability to apply key data science concepts in this project.

What do you envision the end result to be?

- A detailed report, rather an interactive dashboard with clear visualizations describing the correlation and predicting the future outcomes if things go in the same manner.

What techniques do you think you will use to analyze the data?

- The referenced source has used multilevel regression modeling with poststratification framework for their analysis.
- A Similar approach with other techniques available - Ensemble learning (Random Forest), Logistic Regression etc.

Do you envision your system to be interactive or static?

- Interactive Dashboard where a user can filter results according to necessity.

What do you hope to have achieved for the Progress report?

A working prototype which would show the proposed correlation along with the planned changes till final submission.