1. **How many data samples are included in the dataset?**  
   The dataset contains **3,047 samples**.
2. **Which problem will this dataset try to address?**  
   This dataset aims to explore the relationship between **cancer mortality rates** (TARGET\_deathRate) and various characteristics of the population that may affect it, such as **% employment**, **age**, **% race**, **% coverage**, etc.
3. **What is the minimum value and the maximum value in the dataset?**
   * **Maximum TARGET\_deathRate**: 362.8
   * **Minimum TARGET\_deathRate**: 59.7
4. **How many features in each data sample?**  
   Each data sample contains **34 features**.
5. **Does the dataset have any missing information? E.g., missing features.**  
   Yes, the dataset has missing values in the following features:
   * **'PctSomeCol18\_24'**
   * **'PctEmployed16\_Over'**
   * **'PctPrivateCoverageAlone'**
6. **What is the label of this dataset?**  
   The label we are trying to predict is **TARGET\_deathRate**.
7. **How many percent of data will you use for training, validation, and testing?**  
   The dataset is split into:
   * **80% training set**
   * **20% testing set**
   * We have enough data to spare 20% for testing and this is also conventionally common
8. **What kind of data pre-processing will you use for your training dataset?**  
   The following pre-processing steps were applied:
   * **Handling Missing Values**:
     + **'PctSomeCol18\_24'** was dropped because more than **50%** of its entries were missing.
     + The other two missing features were imputed by replacing missing values with the **mean** of their respective columns.
   * **Geography**: The geography feature was converted to **state names** and **one-hot encoded**.
   * **Income**: Binned income values were replaced with the **average** of the corresponding bin ranges.
   * **Encoding**: One-hot encoding was used for categorical features to prevent assigning an ordinal relationship that numerical encoding might imply. Which would make no sense
   * **Log Transformation**: The entire dataset, excluding the target variable and state values, was **log-transformed** to reduce skewness.
   * **Normalization**: The data was then **normalized** using **z-score standard scaling** to standardize the features across a common scale.

|  |  |
| --- | --- |
| **Model** | LR: 0.0001 (R²) |
| **Linear Regression** | 0.7656 |
| **DNN-16** | 0.7761 |
| **DNN-30-8** | 0.8342 |
| **DNN-30-16-8** | 0.8267 |
| **DNN-30-16-8-4** | 0.7936 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **LR: 0.1 (R²)** | **LR: 0.01 (R²)** | **LR: 0.001 (R²)** | **LR: 0.0001 (R²)** |
| **Linear Regression** | -4222358585038829780992 | 0.7932 | 0.7881 | 0.7656 |
| **DNN-16** | nan | nan | 0.8219 | 0.7761 |
| **DNN-30-8** | -0.0029 | -0.0004 | -0.0001 | 0.8342 |
| **DNN-30-16-8** | nan | nan | -0.0001 | 0.8267 |
| **DNN-30-16-8-4** | nan | -0.0004 | -0.0001 | 0.7936 |