# APPENDIX

Table of Contents

[APPENDIX 1](#_Toc131415428)

[1.Business Understanding: 2](#_Toc131415429)

[2.Data Understanding 2](#_Toc131415430)

[3.EXPLORATORY DATA ANALYSIS: 4](#_Toc131415431)

[a.DATA CLEANING AND PRE-PROCESSING 5](#_Toc131415432)

[b.Imputation of missing values, duplicate values 5](#_Toc131415433)

[c.Handling categorical variable 7](#_Toc131415434)

[4.Correlation Matrix: 8](#_Toc131415435)

[a.Univariate analysis 9](#_Toc131415436)

[b.Multivariate Analysis 11](#_Toc131415437)

[c.Multivariate Linear Regression using feature engineering: 12](#_Toc131415438)

[5.Evaluation: 13](#_Toc131415439)

[6.Deployment: 13](#_Toc131415440)

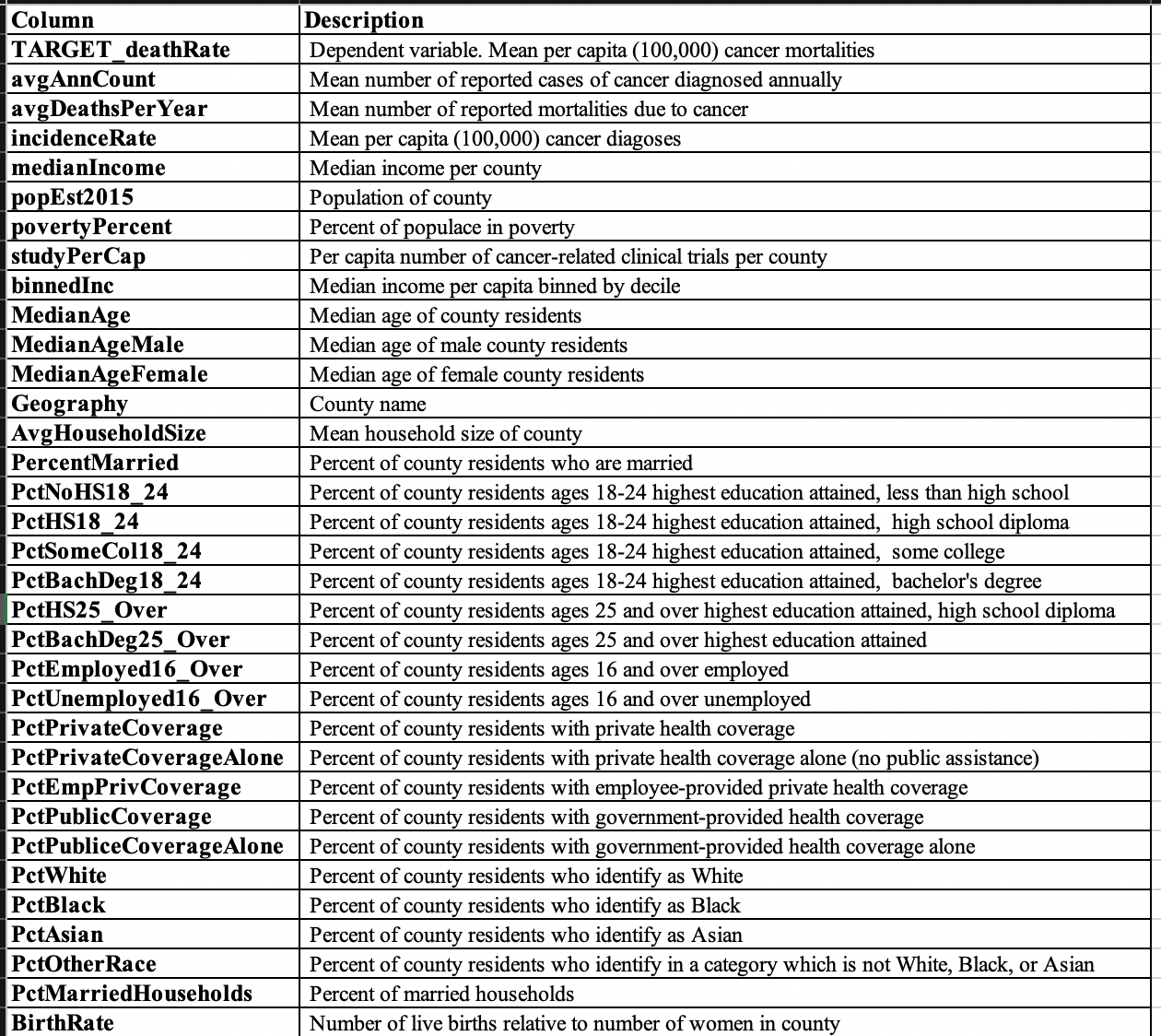
[7.Conclusion: 14](#_Toc131415441)

# 1.Business Understanding:

This project aims to accurately predict cancer mortality rates in US counties based on census data. Accurate predictions can help healthcare organizations allocate resources effectively and identify areas needing targeted interventions. The potential impact of accurate results includes better resource allocation, informed policy-making, and improved public health outcomes. Inaccurate results could lead to misallocating resources and missed opportunities to reduce cancer mortality rates.

# 2.Data Understanding

The dataset used for this project is ‘Cancer Mortality Rates in the US for 2010–2016 at the county level. This includes the data for all 50 states and the District of Columbia. Based on this data, we want to build a model that estimates the cancer death rates in the US and tries to predict the risk that a person from that country might have cancer. The data dictionary is as shown below:



# 3.EXPLORATORY DATA ANALYSIS:

Exploratory analysis is often the initial step of data analysis. Here we get familiar with data, visualize the data in a number of forms, analyze relationships between the variables, look for outliers, patterns, and trends in the data.

The very first step in EDA is to check the dimension of the input dataset and the type of variables.

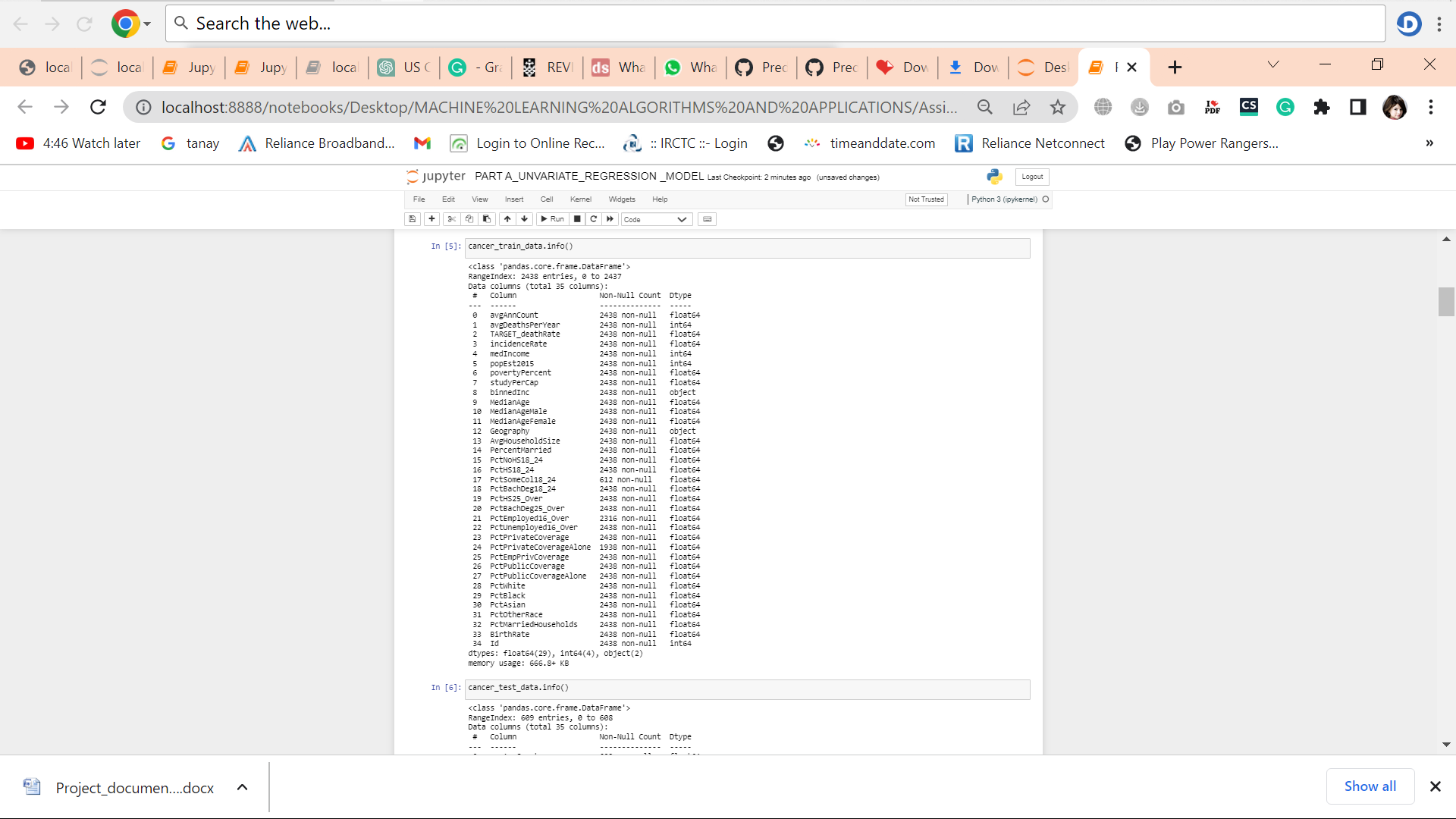


Figure 1

From Figure 1 we can sight that the dataset contains 2438 rows and 35 columns. Majority of the variables are numeric with two categorical variables that are Geography and binned Income.

## a.DATA CLEANING AND PRE-PROCESSING

## b.Imputation of missing values, duplicate values

Our second step in the EDA is to perform some pre-processing and check if the given input data has any missing values, before diving deep into the analysis.

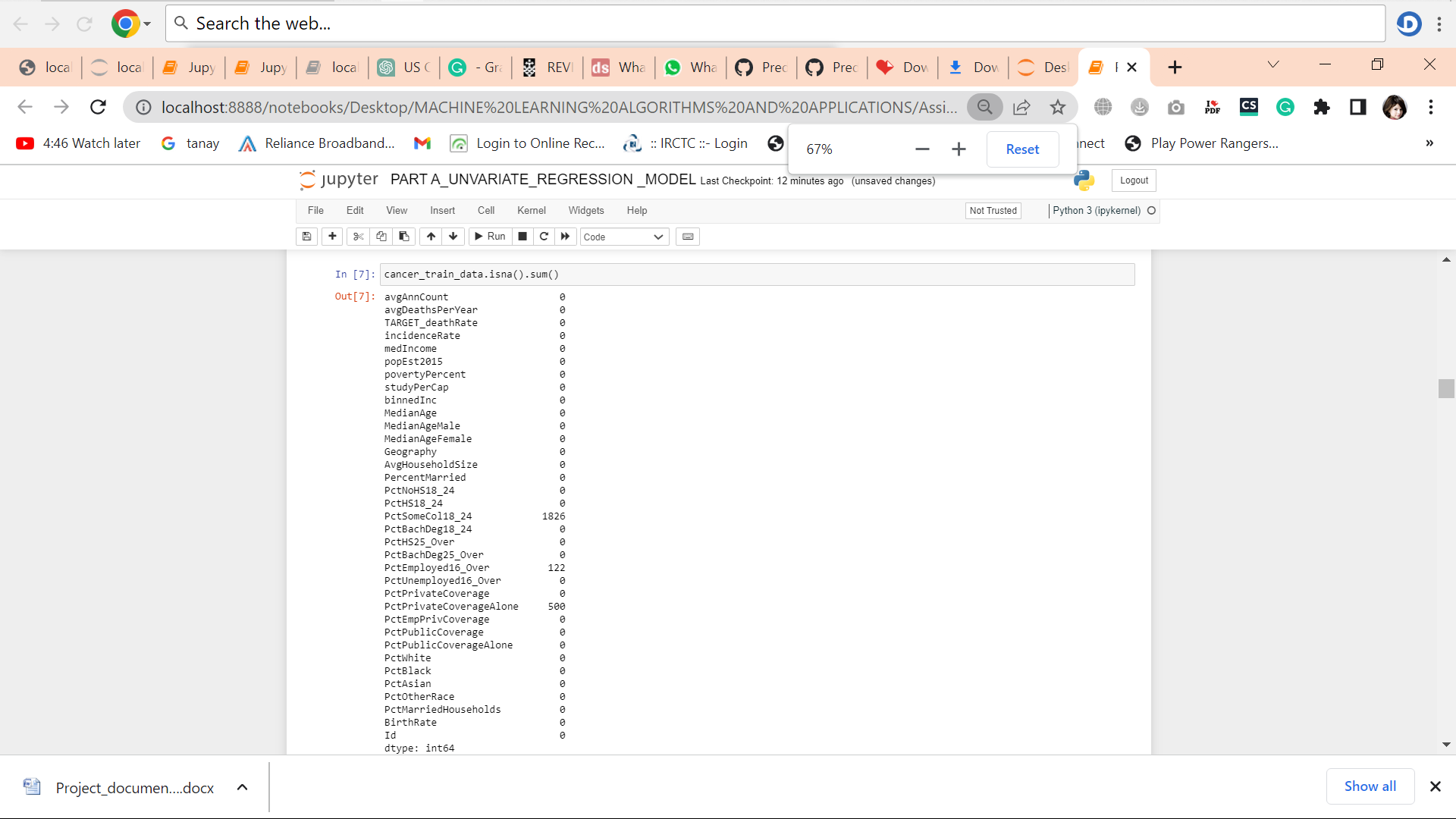


Figure missing values

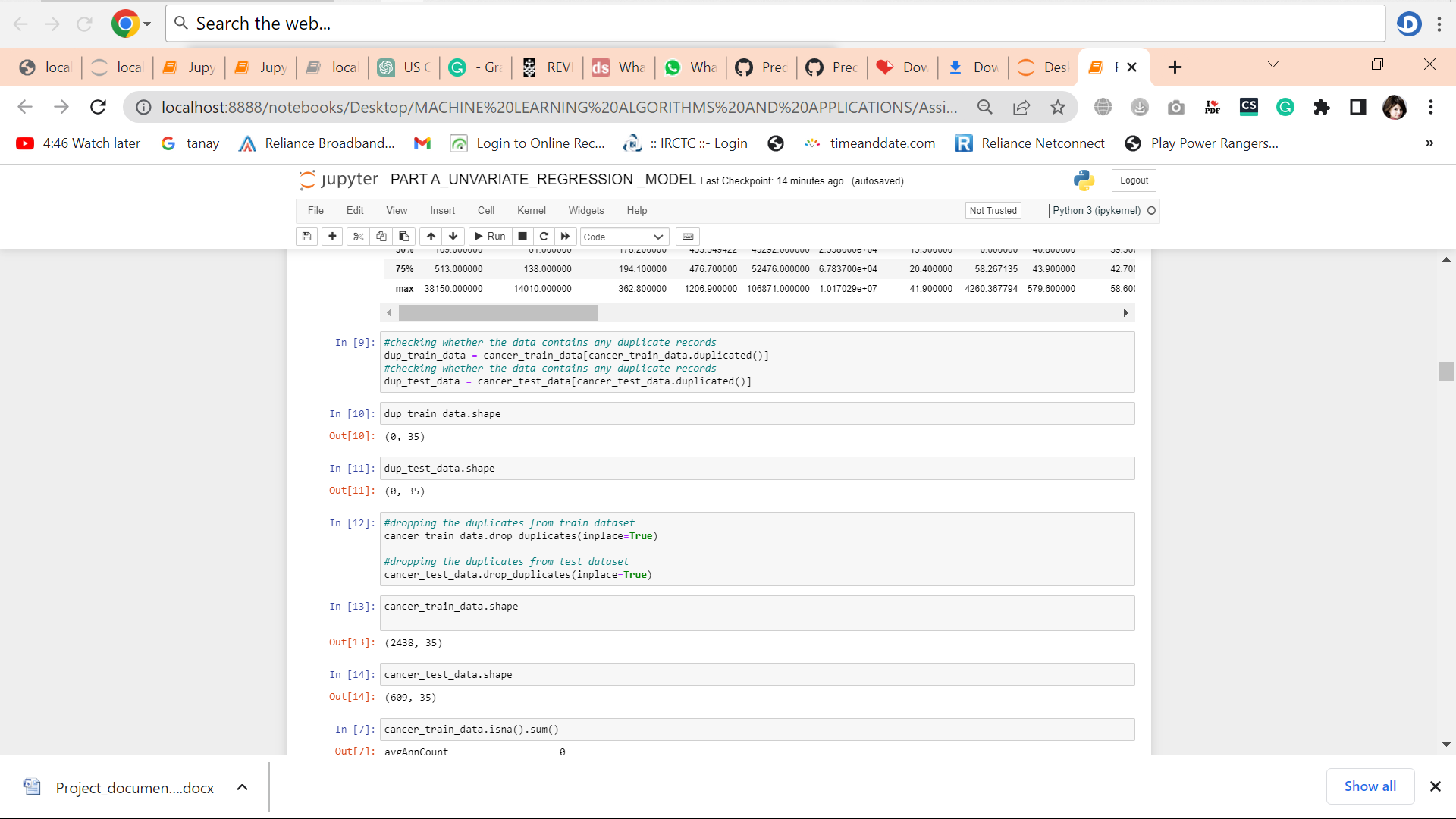


Fig Duplicate values

## c.Handling categorical variable

The geography column contains two information in it. Therefore, split the geography column into two columns as county and state. This can be viewed in Figure 3.

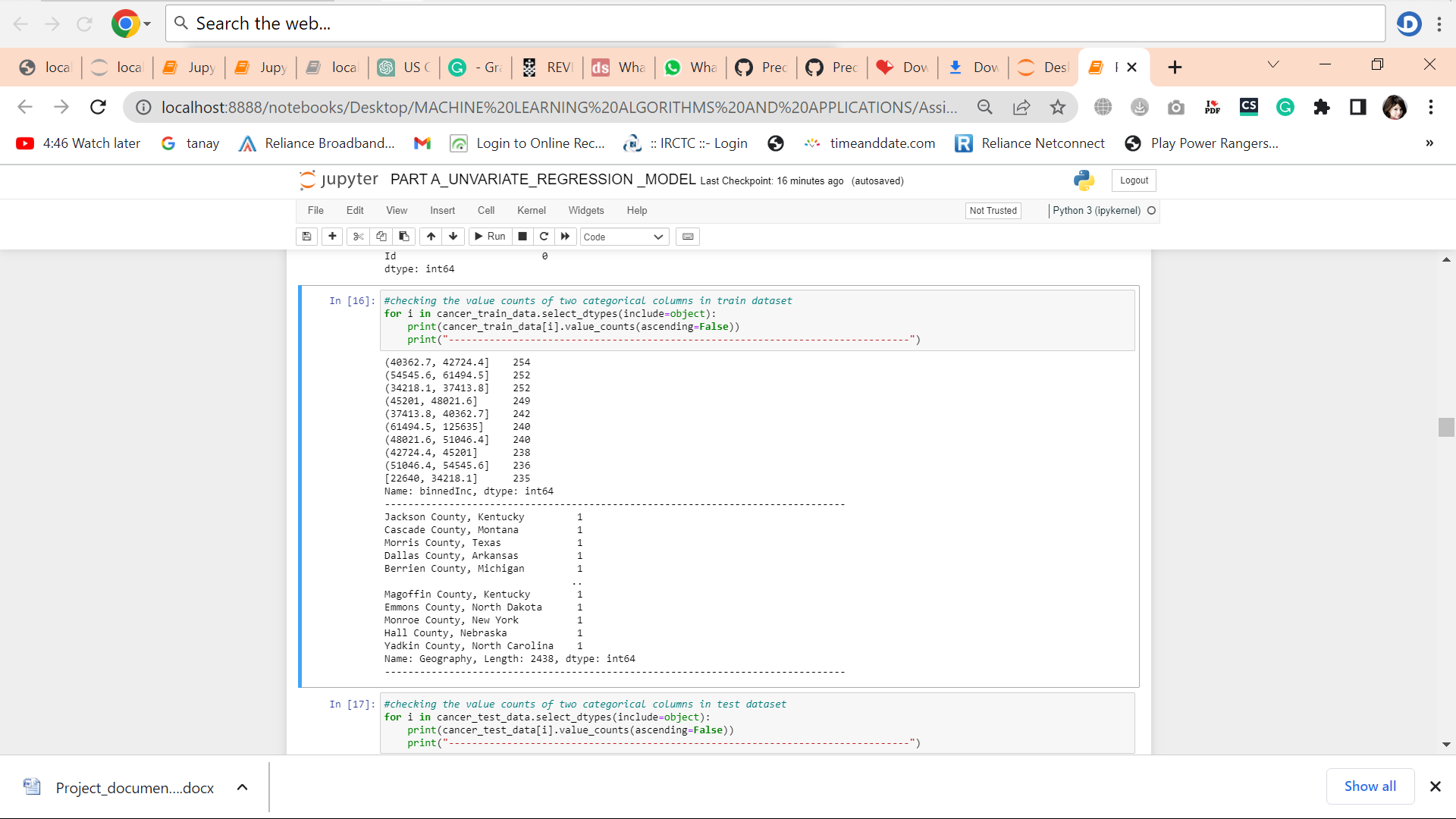


Figure Handling Categorical Data

There are three categorical columns in this data set - binnedInc, county and state.

Every observation in this data set corresponds to one county since the data is collected by county. Hence there are 2438 unique county values. The state variable contains 51 states of the US country. The binnedInc has 10 levels and this variable has already been divided into brackets for better interpretation of the median income per capita binned by decile.

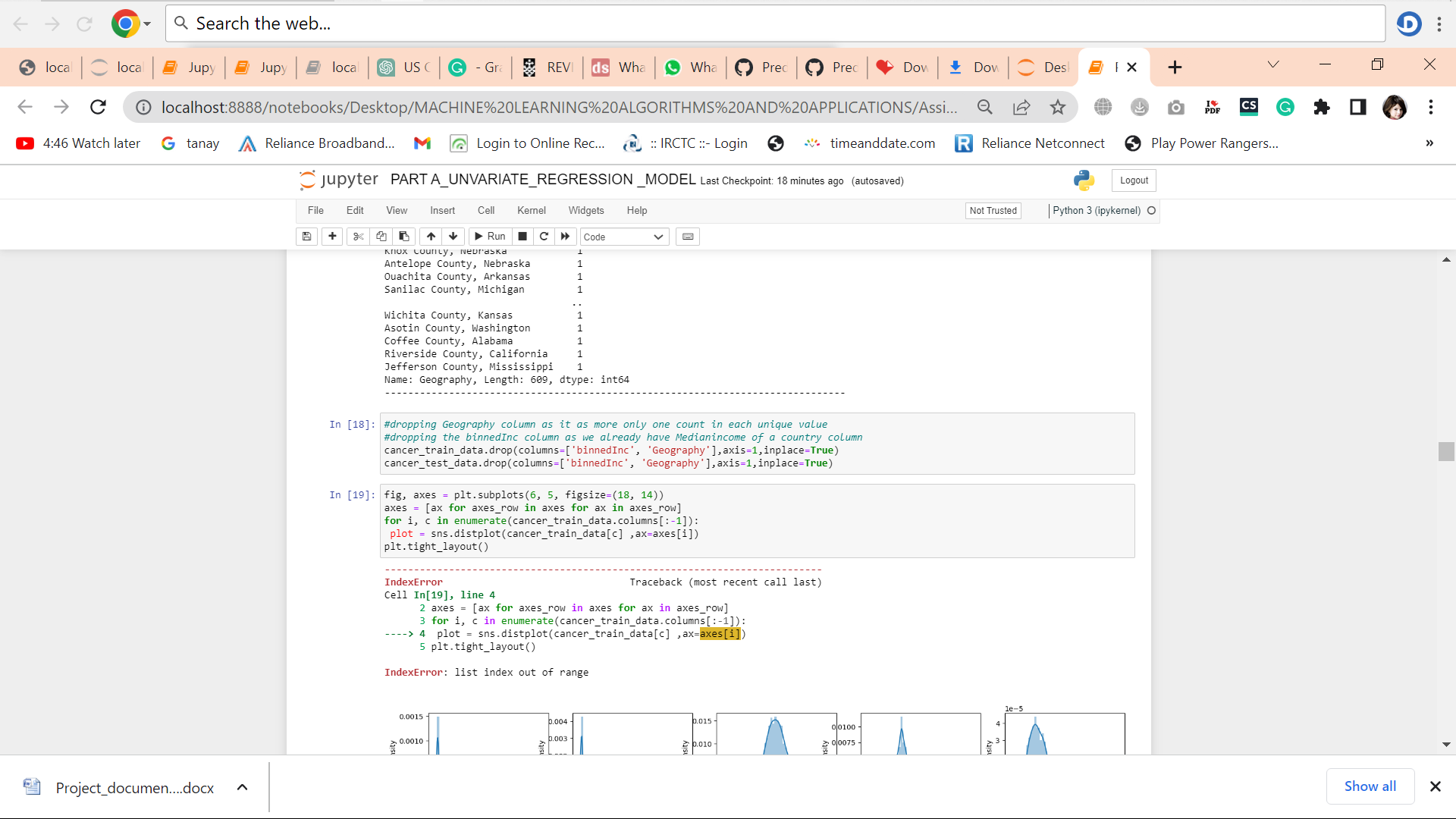


Fig Dropping Some Columns

In this experiment I have dropped the geography and binnedInc columns as geography colum has more count in each unique value and as we have medianIncome of county we don’t need binnedInc column.

## 4.Correlation Matrix:

## download.png

Fig Correlation Matrix

We see moderate to strong correlations for TARGET\_deathRate with incidenceRate, povertyPercent, PctPublicCoverageAlone and strong correlation for PctPublicCoverage with PctPublicCoverageAlone and PctEmpPrivCoverage, PercentMarried with PctMarriedHouseholds, avgDeathsPerYear with avgAnnCount and popEst2015.

These relationships can also be observed visually with the scatterplot matrix as shown in Figure 10. It consists of a collection of scatterplots for each variable-combination of cancer dataset. Each of the scatterplot in the matrix pictures the relationship between a pair of variables which allows many relationships to be explored in just one diagram. If the data points make a straight line going from the left corner out to high x- and y-values, then the variables are said to have a positive correlation**.** If the data points go from a high-value on the y-axis down to a high-value on the x-axis, the variables have a negative correlation.

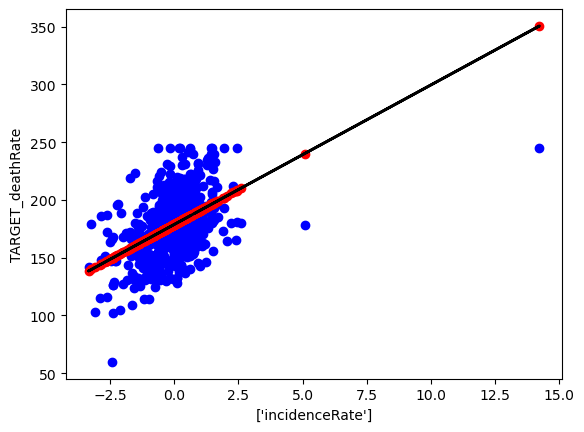
## a.Univariate analysis

Based on the above correlation seen, I selected two features that are highly correlated with target variable.

The features that are highly correlated with the target variable are as follows:

1. incidenceRate
2. povertyPercent

The line graphs for linear regression between incidenceRate and TARGET\_deathRate are as follows:

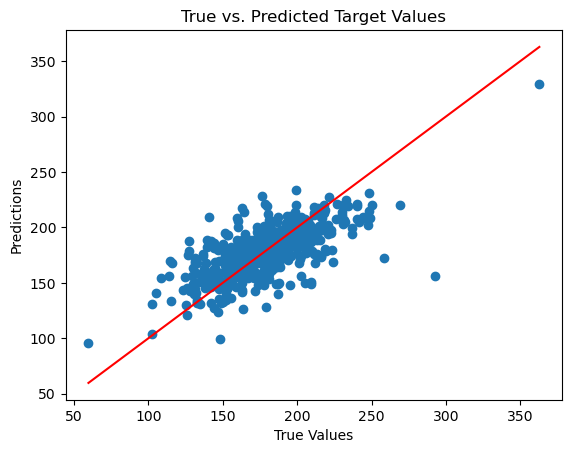


The line graphs for linear regression between povertyPercent and TARGET\_deathRate are as follows:

## download (2).png

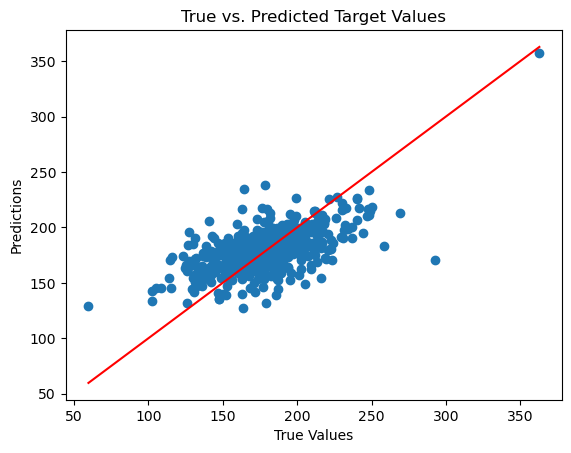
## b.Multivariate Analysis

Firstly, I selected all numeric variables with target variable death rate and the graph for predicted and true values are as follows:



I wanted to check whether the highly correlated values can predict the model. I selected the top three correlated values i.e, incidenceRate, povertyPercent and PctPublicCoverageAlone.

The graph for this model is as follows:



## c.Multivariate Linear Regression using feature engineering:

We trained a multivariate linear regression model with feature engineering for this experiment. The reasons for choosing this model are as follows:

1. Multivariate linear regression can handle multiple features, which allows us to capture complex relationships between the features and the target variable.
2. Linear regression models are simple, interpretable, and can perform well when the relationships between features and the target variable are linear or close to linear.
3. Feature engineering can help improve model performance by transforming the raw data into a more suitable format, such as scaling or applying log transformation to the features.

There are no hyperparameters to tune for a linear regression model. However, we applied feature engineering techniques, including log transformation and standardization, to improve the model's performance. The rationale for using these techniques is as follows:

1. Log transformation: This can help reduce the influence of outliers and transform skewed distributions into more symmetric ones. This can lead to better performance for linear regression models.
2. Standardization: Scaling the features with a mean of 0 and a standard deviation of 1 can help the model converge faster and improve performance.

We decided not to train the following models:

1. Decision Trees, Random Forests, and Gradient Boosting: These models are more complex than linear regression models and may not be necessary for this problem. However, if the linear regression models do not provide satisfactory results, these models can be considered for future experiments.
2. Support Vector Machines: SVMs can be computationally expensive, and linear regression models perform better with fewer computational resources. If linear regression models are insufficient, SVMs with different kernels could be considered for future experiments.

For future experiments, it might be worthwhile to investigate more complex models, such as ensemble methods or deep learning models, to improve performance further. More advanced feature engineering techniques and feature selection methods could also be explored.

## 5.Evaluation:

The model's performance was evaluated using the mean squared error (MSE) metric. Additionally, R-squared values were calculated to assess the proportion of variance explained by the models. The best-performing model was selected based on these evaluation metrics. The insights gained from the models were then analyzed in the context of the business objective.

## 6.Deployment:

If the model achieves the required outcome for the business, the following steps can be taken to deploy the solution into production:

· Retrain the model on the entire dataset and perform a final validation.

· Integrate the model into production systems, including databases, APIs, and user interfaces.

· Continuously monitor the model's performance and update it as needed.

· Document the model and train stakeholders on its usage.

· Establish a feedback loop to collect feedback from end-users and stakeholders for continuous improvement.

## 7.Conclusion:

Based on census data, the project successfully developed regression models to predict cancer mortality rates in US counties. The models provided valuable insights into the relationship between various features and the target variable. Although the current approach showed promise, there is still room for improvement. Further experimentation with different models, feature engineering techniques, and data preparation methods can lead to better performance and more accurate predictions, ultimately benefiting healthcare organizations and public health outcomes.

Top of Form

|  |
| --- |
|  |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
|  |
|  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |