**Linear Regression**

Linear regression is an algorithm in a supervised machine learning realm. It is used to help predict the value of a dependant variable based on one or several independent variables (Madhugiri, 2023). In this case, the dependant variable is the cancer mortality rate, we will utilize US census dataset containing information and independent features of multiple counties in the US to predict the mortality rate caused by cancer.

There are a lot of literatures and studies on linear regression such as how data should be structured to best train the models and maximize the accuracy of the prediction outcomes (Brownlee, 2020). In this opportunity, we will use CRISP-DM framework to provide a more structured approach to tackle this project, properly documents every step taken and stay focused and organized on our objectives throughout the process.

Six major phases of CRISP-DM (Sridharan, 2018):

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modelling
5. Evaluation
6. Deployment

# I. Business Understanding

In 2022, there were approximately 600,000 people died from cancer in the United States alone, with lung and bronchus cancer causing the most deaths (approximately 130,000 people or 21% of total death) followed by colon and rectum cancer and pancreatic cancer (National Institute of Health, 2022).

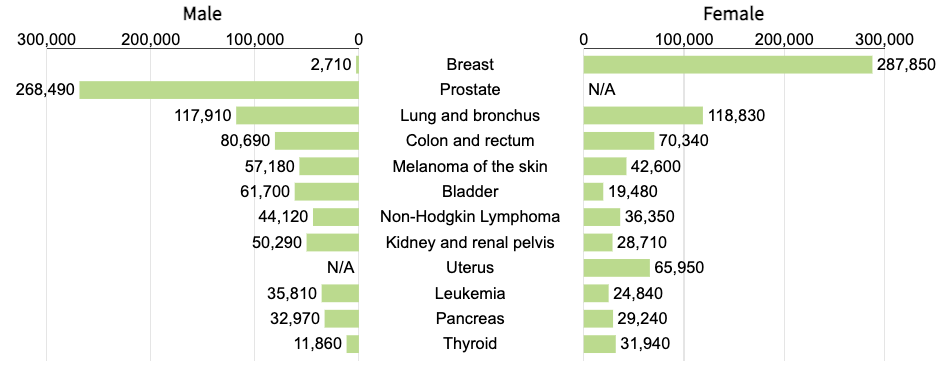
Chart, pie chart

Description automatically generated

Source: (National Institute of Health, 2022)

**Cancer cases among males and females**

If we look closer into males and females’ population, prostate cancer has the highest incidence rate among men with more than 260,000 cases. On the other hand, for women, breast cancer is the leading diagnosis, with an estimated 280,000 cases.



Source: (National Institute of Health, 2022)

**cancer mortality rates by ethnicity**

Death rates caused by cancer vary by ethnicity. As we can see from the graph below, Asian & Pacific Islander has relatively lower mortality rates in all 5 most commonly diagnose cancers.

Chart, line chart

Description automatically generated

Source: (National Institute of Health, 2022)

**Objective**

Predicting cancer mortality rates based on census data can be challenging but it has several promising advantages, including (Sekeroglu & Tuncal, 2021):

1. **Social determinants of health:** using socioeconomic status to have more comprehensive understanding of how it can impact cancer outcomes and take necessary actions.
2. **Targeted intervention:** allowing organisations or governments to better target population to improve screenings and treatments.
3. **Resources allocation:** help governments allocate resources more effectively and efficiently by directing funding to areas or population prone to cancer.
4. **Public awareness:** providing the public with valuable cancer risk information helps with prevention and can lead to cost savings to help ease government’s health care cost burden.
5. **Policy-making:** information collected can be used to assist governments in developing equitable health policies

We will explore US census dataset in depth to find correlations and utilise linear regression models to predict cancer mortality rate to raise awareness to the public and encourage actions necessary to minimize deaths caused by cancer.

# II. Data Understanding

**Summary**

The dataset we will be using is derived from US county census data comprising 35 different columns. Python and pandas are used to understand the dataset. 33 of the columns are numerical which are features to be used for this analysis:

Table

Description automatically generated with medium confidence

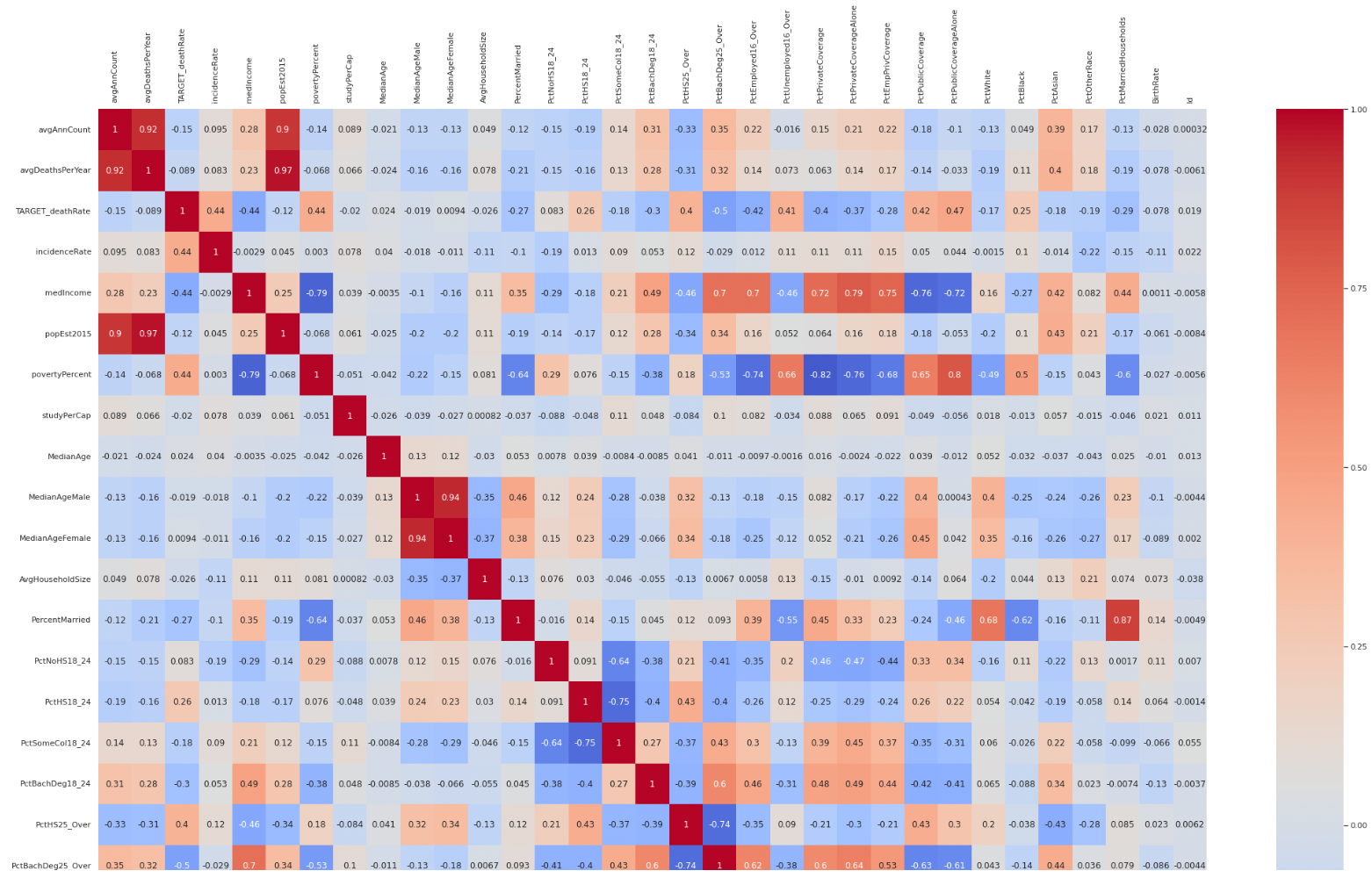
There are over 3,000 US counties in the dataset which covers most if not all of US counties and should provide comprehensive view of each county in the US. The entire dataset consists of 3,048 rows and 35 columns and we will split this dataset into 3 smaller sets for training, validation and testing. Below is the screenshot of the data:

Graphical user interface

Description automatically generated with medium confidence

**Relationships**

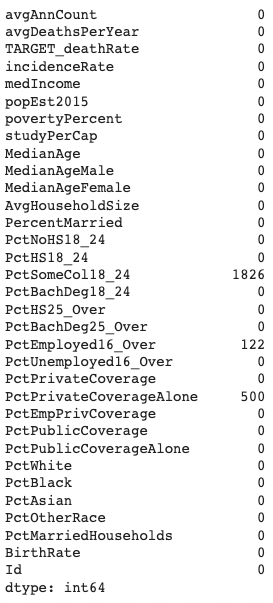
Correlations between variables are pictured in the heatmap below. We can see there are 17 features that have relatively strong correlation with target variable (0.25 and above), these features tend to be the socioeconomic condition of the population such as the county’s poverty level, unemployment rate, median income level and education level.



# III. Data Preparation

**Missing Values**

We will be checking for missing values in the dataset and take necessary actions to remedy the issue. We found out there are 3 columns with missing values.

****

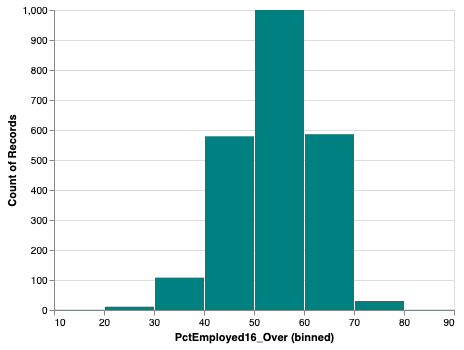
Since there are significant number of missing values, dropping the rows will reduce the size of the data significantly, so we opt to replace the missing values with their mean values instead. Prior to replace the values, we conducted checks on values of each column to see if the distribution is normal:

-Checking for unusual value (negative or more 100%) in 'PctSomeCol18\_24’

Chart, histogram

Description automatically generated

-Checking for unusual value (negative or more 100%) in 'PctEmployed16\_Over'

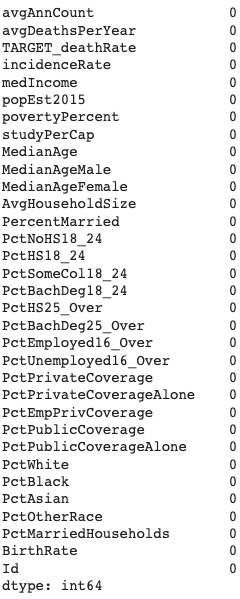


-Checking for unusual value (negative or more 100%) in 'PctPrivateCoverageAlone'

Chart, histogram

Description automatically generated

The value distributions looks normal (above 0% and under 100%), so we will go ahead and replace the missing values with mean value, result is shown below:

****

**Duplicate Values**

We are checking for duplicate values in ‘Geography’ and ‘Id’ columns, since these are county identifier so there should be no duplicate counties in the dataset. The rest of the features possibly have duplicates, but these are numerical values such as poverty rate and median income so any duplicate values are normal.

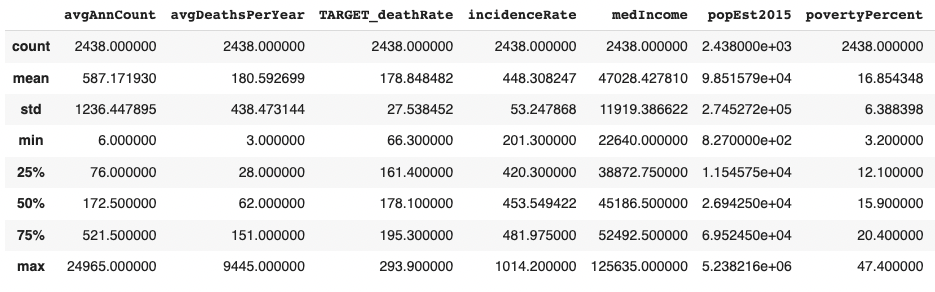
There are no duplicate values in ‘Geography’ and ‘Id’:

**Graphical user interface, text, application, email

Description automatically generated**

**Outliers**

We used describe() to look at the statistical calculation of every features and found and removed a number of outliers in “MedianAge” column where median age is more than 100. 24 rows were removed from the dataset in total.



We also performed checks on several other features, for example:

-Outliers in median income make sense since some counties have much higher income per capita compared to others

Chart

Description automatically generated

-Outliers in study per capita is also logical since the number of population between counties vary from hundreds to millions of people

Chart, scatter chart

Description automatically generated

**Dropping Non-Numerical Features**

Since linear regression can only work with numerical values, we removed non-numerical columns (‘binnedInc’ and ‘Geography’) from the dataset before feeding it to the regression model.

# IV. Modelling

Linear regression will be used in this analysis and will be running several experiments to see which models fit better:

**Models:**

**Model 1: Univariate linear regression**

We developed univariate linear regression using 2 features with higher correlation with the target variable (cancer mortality rate). The purpose is to see if we can accurately predict mortality rate using single variable.

**Model 2: Multivariate linear regression**

We developed multivariate linear regression using all the available features in the dataset. The objective is to see how well the model performs if we all the features available.

**Model 3: Multivariate linear regression**

In this experiment, we used regularization and feature engineering to try to achieve better accuracy compared to previous experiments.

**Dataset Splitting**

Initially, the dataset contains 3,047 rows and 35 columns, 20% of the dataset will be reserved for final testing and remaining 80% will be used for training and validation, breakdown:

1. Training dataset (80%):

-Actual training (80%)

-Validation (20%)

2. Testing dataset (20%)

**Assessment method:**

Root Mean Squared Error (RMSE): we used RMSE to evaluate the models’ accuracy, compare their performance with each other, detect overfitting or underfitting and identify sources of error (outliers, missing data etc).

# V. Evaluation

It’s not surprising to see that model 3 performed better than the other 2 models, this is due to feature engineering and regularization. The breakdown of each model’s performance is as follow:

**Model 1:**

1. Baseline performance (average): 27.2903

Using 2 different features:

Feature #1 (povertyPercent):

-Training dataset: 24.644

-Test dataset: 26.3111

Feature #2 (medIncome):

-Training dataset: 24.6753

-Test dataset: 26.2344

Chart, line chart

Description automatically generated

**Model 2:**

1. Baseline performance (average): 27.5093

2. Using all numeric features:

-Validation dataset: 19.9466

-Test dataset: 20.6126

Chart, line chart

Description automatically generated

**Model 3:**

1. Baseline performance (average): 27.5093

2. After feature engineering:

-Validation dataset: 19.8352

-Test dataset: 20.5568

Chart, line chart

Description automatically generated

3. Using regularization techniques:

-Lasso: 19.5773

Chart, line chart

Description automatically generated

-Ridge: 19.4689

Chart, line chart

Description automatically generated

-ElasticNet: 19.5613

Chart, line chart

Description automatically generated

In model 3, we dropped 2 features (‘Id’ and ‘MedianAgeFemale’) since ID is basically an identifier, therefore does not have significant meaning and both features have low correlation with the target variable. We can see that model 3 performed better than model 1 and 2, but further fine tuning can still be done to achieve better accuracy if given more time. For the time being, we can trial model 3 before deploying it to production to ensure the model is accurate and perform as expected in real-world setting. The feed gathered from trial can also be used adjust and fine tune the model.

# VI. Deployment

It could be a good idea to trial the model in the real world and in a closely monitored setting, gather feedbacks from users to make adjustment and fine tune the model before deploying it to a larger audience. Thorough deployment, monitoring, maintenance and review planning should be developed to capture issues (technical, ethical etc), implement improvements/tuning and assess the impacts of the model to maximize the model’s potential.

# References

1. Madhugiri, D. (2023). Linear Regression in Machine Learning: A Comprehensive Guide.

<https://www.knowledgehut.com/blog/data-science/linear-regression-for-machine-learning>

1. Brownlee, J. (2020). Linear Regression for Machine Learning.

<https://machinelearningmastery.com/linear-regression-for-machine-learning/>

1. Sridharan, M. (2018). CRISP-DM – A Framework for Data Mining and Analysis.

<https://thinkinsights.net/data/crisp-dm/>

1. National Institute of Health. (2022). Cancer Stat Facts: Common Cancer Sites

<https://seer.cancer.gov/statfacts/html/common.html#:~:text=In%202022%2C%20an%20estimated%20609%2C360,common%20cause%20of%20cancer%20death>

1. Sekeroglu, B., Tuncal, K. (2021) Prediction of cancer incidence rates for the European continent using machine learning models. Health Informatics Journal. 2021;27(1). doi:10.1177/1460458220983878

https://journals.sagepub.com/doi/10.1177/1460458220983878