# Business Understanding.

### Business Objective:

The primary objective of the business is to reduce the rate of cancer related deaths across various counties. Taking a look at the columns and their description, at a high level one can assume that the organisation wishes to affect the percentage distributions by implementing policies based on our study, in turn lowering the rate of cancer related deaths.

### Situation Assessment:

The resources allocated to us the data analyst include the dataset which contains information for the various counties and the relevant parameters of measurement for each field within the dataset.

The project requires us to study the dataset and leverage the data to train a model that can help predict the trends between the target variable (TARGET\_deathRate) and other independent variables.

If successful in our analysis, we can recommend areas of focus for various policies that could in turn result in lower rates of cancer related deaths and an overall increase in living quality for these people.

### Technical Assessment:

As far as the technical goals are concerned, we shall primarily focus on evaluating the interactions between the independent and target variables, narrowing down the relevant fields. The secondary goals will involve studying correlation between the trends for these variables and the target variable, and lastly training a model that can visually and statistically quantify this correlation.

### Project Plan:

Split into 3 stages, this project will first study the various interactions amongst the variables and shortlist variables of interest. The second stage will involve building a model that utilises these variables and predicts a trend that we can use to suggest policy changes. The last and final stage will be used to engineer new features and possibly scale our dataset and errors down to better understand the effects of dimensionality.

# Data Understanding.

### Initial Data Collection:

As the dataset was provided by the business itself, there was no collection phase required. At most, the only step required was to download the zipped folder and extract it to reveal the training and testing dataset.

### Data Description:

The dataset is split into two, the training and testing dataset. As per the norm, we shall not perform any work on the testing dataset until the time of evaluation, so our descriptive work will cover only the training set.

The raw dataset contains 35 columns and 2414 corresponding entries. The columns cover the various fields of income diversity, age, healthcare coverage, education level, marital status and the target variable.

### Data Exploration:

As the data is mostly numerical, we can drop the two categorical fields containing nominal data and plot scatter graphs for each variable against the target variable. Here we notice the various trends in the data and note down the columns we see as valuable for training and modelling our project.

Namely, the following fields hold the most correlation and maintain a somewhat linear relationship with the target variable:

[ "medIncome", "povertyPercent", "PctUnemployed16\_Over", "PctEmployed16\_Over", "PctHS25\_Over", "PctBachDeg25\_Over", "PctPrivateCoverage", "PctPrivateCoverageAlone", "PctEmpPrivCoverage", "PctPublicCoverage", "PctPublicCoverageAlone", "PctMarriedHouseholds", "PercentMarried" ]

### Data Quality Verification:

Performing a perfunctory visual inspection of the dataset, we notice a few outliers in the columns for median age, noting it down for later data cleaning and preparation. We also perform datafield descriptive statistics to narrow down the margin of error in terms of outliers and perform a few transformations to ensure data validity.

The dataset appears to be mostly clean and there is very little in terms of actual cleaning that would need to be done to use this dataset, however, this is taken care of in the next stage.

# Data Preparation:

In this stage of the project, we refer back to the fields we noticed as having outliers or inaccurate data. With the median age in mind specifically, we apply a filter mask and prune the entries that do not align with logical norms of ages over 100 years.

The next stage is to drop the entries with rows that contain Na values, however, to reduce the number of rows gotten rid of, we first reduce our dataset to the columns we wish to actually train and observe.

Once this is done, we can verify the fields once again to ensure no null values or any other data quality issues and proceed with the next stages of this project. The final stage of the data preparation differs within the separate experiments.

In Part A, we first extract two columns from the dataset, those of “PctBachDeg25\_Over” and “PctPublicCoverageAlone” in order to narrow down the training of the univariate linear regression model that we will implement.

In Part B, we extract an array of columns; []. These columns form the sub-dataset that we will use to train the multivariate linear regression model, as well as the different variants of linear regression with regularisation implemented.

Finally, in Part C we perform feature engineering and feature scaling to provide more detailed statistics for the overall rate of education in the given counties so as to better facilitate the policy decisions after providing our results.

These individual stages had unique results that were used to reapply logic to previous and future stages and helped refine the overall experiment.

# 

# Modelling:

As part of the modelling process, the experiment constraints performed somewhat limited the manner in which we approached the issue at each stage. Nonetheless, we managed to implement every variation that we needed to, albeit with a bit of trial and error at each independent stage, going back and forth between experiments to apply our learnings.

Part A of our experiment relied solely on discovering the trends in the data and shortlisting the variables we needed for univariate linear regression. As this stage called for no advanced training or modelling, we simply focussed on understanding the dataset and existing correlations.

Part B of the experiment opened up the opportunity to experiment with regularisation within linear regression models, and we successfully implemented Lasso, Ridge and ElasticNet regularisation.

Finally, Part C was the chance we needed to work with feature engineering and normalisation of the dataset. Here, we used the same models as before, but allowed ourselves to create new data points and features to hopefully reduce our error. However, while one may notice that the initial stages seem to increase the error rate, when normalised, the dataset in fact shows a decrease in the overarching error rates from the previous experiment.

# 

# 

# Evaluation:

The primary metric for evaluating model performance was Mean Square Error rate. Using the scikit-learn library, we were able to seamlessly import and calculate the performance at each stage of the experiment. Additionally, we drew line and scatter plots to visualise where we were missing the most predictions, and understand the reason for such errors.

The univariate linear regression models showed great ability to follow trends in the data, highlighting the variables that had the most significant relationship and correlations with the target variable. The model trained on “PctPublicCoverageAlone” scored a MSE of 695.6542515936903, while the model trained on “PctBachDeg25\_Over” scored a MSE of 659.5891932136362. While these seem to be high error rates, a quick look at the scatter plot for the test dataset shows that the outliers on either end of the graph seem to contribute largely to this. Normalising the dataset later shows a change in this trend.

The multivariate regression trained on :

[ "medIncome", "povertyPercent", "PctUnemployed16\_Over", "PctEmployed16\_Over", "PctHS25\_Over", "PctBachDeg25\_Over", "PctPrivateCoverage", "PctPrivateCoverageAlone", "PctEmpPrivCoverage", "PctPublicCoverage", "PctPublicCoverageAlone", "PctMarriedHouseholds", "PercentMarried" ]

scored an MSE of 547.7813962647862. With regularisation, the error rate was reduced slightly, to 546.3250861443997. The reason for high error seems to be the same as in the previous experiment, and Part C would go on to change the degree of the error.

Finally, with feature engineering and normalisation, the multivariate linear regression models had an error rate that scored at 0.7463373849051153. This is a rather low error rate, however, this would need to be independently verified before applying it in the real world.

# Conclusion:

Depth of discussion of ethics/privacy issues, value, benefits and recommendation for business

After much exploration and experimentation with the dataset, the above documents the process behind our results. However, applying these results in another problem altogether.

The primary issues with the dataset’s infringement on ethics and privacy are inconsequentially small. Since there are no obvious names or identifiers in the census based dataset, it should be no issue to apply the learning of this project in policies and suggestions.

The value to be gleaned from our experiments is of paramount importance. From our study, if we simply promote a system that results in higher education rates and perhaps cheaper or better quality in the public healthcare system, we could effectively lower the rate of cancer related deaths. However, there are some variables that we cannot openly suggest to policy changes such as increasing the percentage of married people, and instead these must be approached with appropriate caution and regard.

The benefits of applying our results to governmental policies to promote specific sectors of counties to grow and expand would greatly benefit not the business, but the people living in said counties as the overall quality of life should drastically improve.

It is the recommendation of us, the data analyst to move forward with the results listed and shown above to draft ideas and policies that positively influence the end goals of the business, that is to reduce the overall rate of cancer deaths and in turn improve the population’s lifestyle and nation’s growth.