|  |
| --- |
| 13 Types of Graphs and Charts (Plus When To Use Them) | Indeed.com |
| Individual portfolio |
| Unveiling Patterns: Data-Driven Perspectives on Cancer Mortality |

|  |
| --- |
| Hewaar Al Koerdi 1619307  19-12-2023 |

Contents

[1. Introduction 3](#_Toc153844973)

[2. Methodology 4](#_Toc153844974)

[2.0 CRISP-DM Framework Overview 4](#_Toc153844975)

[2.1 Business Understanding 4](#_Toc153844976)

[2.1.1 Problem Description 4](#_Toc153844977)

[2.1.2 Relevance 4](#_Toc153844978)

[2.2 Data Understanding 4](#_Toc153844979)

[2.3 Data Preparation 5](#_Toc153844980)

[2.4 Modeling 5](#_Toc153844981)

[2.5 Evaluation 5](#_Toc153844982)

[2.6 Deployment 5](#_Toc153844983)

[3. Results 6](#_Toc153844984)

[3.1 Data Merging and Exploration 6](#_Toc153844985)

[3.1.1 Merging Datasets 6](#_Toc153844986)

[3.1.2 Data Exploration 6](#_Toc153844987)

[3.2 Data Preparation 6](#_Toc153844988)

[3.2.1 Handling Missing Values 6](#_Toc153844989)

[3.2.2 One-Hot Encoding 6](#_Toc153844990)

[3.3 Exploratory Data Analysis (EDA) 7](#_Toc153844991)

[3.3.1 Universal Analysis 7](#_Toc153844992)

[3.3.2 Categorical Variables 11](#_Toc153844993)

[3.4 Model Building 12](#_Toc153844994)

[3.4.1 Selection of Independent Variables 12](#_Toc153844995)

[3.4.2 Regression Analysis 13](#_Toc153844996)

[3.4.3 Random Forest Feature Importance 13](#_Toc153844997)

[3.4.4 Threshold-based Feature Selection 15](#_Toc153844998)

[3.5 Model Evaluation and Diagnostics 15](#_Toc153844999)

[3.5.1 Diagnostic Plots 15](#_Toc153845000)

[3.5.2 Heteroskedasticity Test 17](#_Toc153845001)

[3.5.3 Multicollinearity Assessment 18](#_Toc153845002)

[3.6 K-Fold Cross-Validation 19](#_Toc153845003)

[3.6.1 Model Selection and Evaluation 19](#_Toc153845004)

[3.7 Interpretation and Insights 20](#_Toc153845005)

[4. Discussion 21](#_Toc153845006)

[4.1 Challenges in Analysis 21](#_Toc153845007)

[4.2 Choices Made and Methodological Considerations 21](#_Toc153845008)

[4.3 Assessment of Model Quality 21](#_Toc153845009)

[5. Conclusion 22](#_Toc153845010)

[5.1 Summary of Key Findings 22](#_Toc153845011)

[5.2 Possible Future Steps 22](#_Toc153845012)

[7. Ethics 23](#_Toc153845013)

[Bibliography 24](#_Toc153845014)

[Appendix Appendix 1: OLS Regression Results 25](#_Toc153845015)

[Appendix 2: VIF Results 26](#_Toc153845016)

# 1. Introduction

The purpose of this portfolio project is to showcase my proficiency in applying Data Science analytics skills to a real-life challenge. In my quest to demonstrate these skills, I chose the "Multiple Linear Regression Challenge" as it provided a structured framework and clear objectives. This challenge was particularly appealing to me due to the inherent difficulty in finding an appropriate dataset and the desire for more guidance in project execution. (Data World , sd)

**Project Objectives**

The primary objective of this project is to build a multivariate Ordinary Least Squares regression model to predict the "TARGET\_deathRate." The challenge involves several tasks, including data cleaning, exploratory data analysis, feature selection, model building, and evaluation. By undertaking this project, I aim to not only showcase my technical abilities but also to enhance my interpretative skills in dealing with complex datasets.

**Dataset Selection**

Selecting an appropriate dataset proved to be a challenging aspect of initiating a data science project. The amalgamation of datasets from diverse sources, including the American Community Survey, clinicaltrials.gov, and cancer.gov, underpins the richness and complexity of the data. This selection was motivated by the desire to work with real-world data, integrating information from various domains to derive valuable insights.

**Rationale for Challenge Choice**

The decision to take on this challenge was twofold. Firstly, the structured nature of the challenge provided clear guidelines, allowing for a systematic approach to problem-solving. Secondly, the challenge presented an opportunity to tackle a problem that demanded careful interpretation of the data, aligning with my aspiration to enhance my analytical and interpretative skills.

In essence, this portfolio project serves as a testament to my commitment to continuous learning and practical application of data science techniques in a challenging and structured environment. The subsequent sections will delve into the various stages of the project, detailing the methodologies, analyses, and interpretations employed to address the "Multiple Linear Regression Challenge."

# 2. Methodology

## 2.0 CRISP-DM Framework Overview

Before delving into the specific steps of the methodology, it's essential to provide an overview of the CRISP-DM framework used throughout the project. CRISP-DM, or Cross-Industry Standard Process for Data Mining, is a widely adopted framework for guiding the data mining process. It consists of six major phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment.

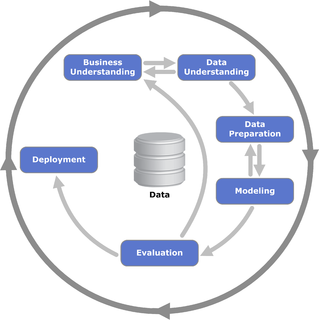


Figure 1: CRISP-DM model

## 2.1 Business Understanding

### 2.1.1 Problem Description

The business problem at the core of this project revolves around predicting the "TARGET\_deathRate" using a multivariate Ordinary Least Squares regression model. The significance of this problem lies in its direct relevance to public health, as understanding the factors contributing to cancer death rates is crucial for healthcare planning and intervention.

### 2.1.2 Relevance

The relevance of predicting cancer death rates is underscored by its potential impact on public health policy and resource allocation. By identifying significant predictors, we contribute to a better understanding of the factors influencing cancer mortality rates, enabling more targeted and effective public health initiatives.

## 2.2 Data Understanding

In this phase, the focus was on gaining insights into the structure and content of the datasets used. Exploratory Data Analysis (EDA) techniques, statistical summaries, and correlation matrices were employed to comprehend the nature of the data and identify potential challenges.

## 2.3 Data Preparation

Data preparation involved several steps, including handling missing values, one-hot encoding categorical variables, and exploring relationships between variables. The goal was to create a clean and structured dataset for subsequent modeling.

## 2.4 Modeling

The selection of independent variables for the regression model was informed by correlation analysis. A regression model was built using the selected variables, and the process involved testing for multicollinearity, assessing model linearity, and evaluating model fit.

## 2.5 Evaluation

Model evaluation was performed through various diagnostics, including assessing the linearity of the model, evaluating residuals for independence and heteroskedasticity, and conducting the Breusch-Pagan test. K-fold cross-validation was employed to ensure the model's robustness and generalizability.

## 2.6 Deployment

While the deployment phase traditionally involves implementing the model in a real-world setting, for this project, the emphasis was on showcasing the model's interpretability and providing a detailed analysis of its predictions.

The subsequent sections will delve into each of these phases in detail, outlining the specific steps, analyses, and insights gained throughout the project.

# 3. Results

In this section, we delve into the extensive outcomes derived from a meticulous application of data science analytics techniques to address the challenge at hand. The comprehensive results encompass a multifaceted exploration of data, intricate modeling endeavors, and thorough diagnostic evaluations.

## 3.1 Data Merging and Exploration

### 3.1.1 Merging Datasets

The amalgamation of datasets, namely 'avg-household-size.csv' and 'cancer\_reg.csv,' was initiated based on a common identifier, "geography." This strategic merging not only facilitated a unified dataset but also set the stage for subsequent analyses.

### 3.1.2 Data Exploration

A multifaceted exploration commenced with a holistic overview of the merged dataset. Statistical measures, such as summary statistics and a correlation matrix, were pivotal in comprehending the inherent patterns and relationships within the data.

## 3.2 Data Preparation

### 3.2.1 Handling Missing Values

Addressing missing values is paramount for robust analyses. Rows with missing values were judiciously removed, ensuring a dataset devoid of gaps that might compromise the integrity of subsequent modeling efforts.

### 3.2.2 One-Hot Encoding

The categorical variable "binnedinc" underwent a transformation through one-hot encoding. This conversion empowered the inclusion of categorical features in subsequent modeling steps.

## 3.3 Exploratory Data Analysis (EDA)

### 3.3.1 Universal Analysis

**Histograms**

Numerical variables were scrutinized through histograms, unraveling insights into their distributions, central tendencies, and potential outliers.

Afbeelding met tekst, diagram, Perceel, origami

Automatisch gegenereerde beschrijving

Figure 2: Histograms

**Boxplots**

The utilization of boxplots shed light on the distribution characteristics of numerical variables, aiding in the identification of central tendencies and potential outliers.

Afbeelding met lijn, Rechthoek, schermopname, tekst

Automatisch gegenereerde beschrijving

Figure 3: Boxplots

**Bivariate Analysis**

A nuanced bivariate analysis unfolded through a pairplot of select variables, providing a visually rich understanding of pairwise relationships and potential correlations.

Afbeelding met tekst

Automatisch gegenereerde beschrijving

Figure 4: Pairplots

**Heatmap**

A heatmap of the correlation matrix, an indispensable visual aid, portrayed intricate relationships between variables in a color-coded manner.

Afbeelding met tekst, schermopname, lijn

Automatisch gegenereerde beschrijving

Figure 5: Heatmap

### 3.3.2 Categorical Variables

A countplot elucidated the distribution of the categorical variable "binnedinc\_(34218.1, 37413.8]," offering insights into its prevalence within the dataset.

Afbeelding met schermopname, tekst, Rechthoek, diagram

Automatisch gegenereerde beschrijving

Figure 6: Countplot

## 3.4 Model Building

### 3.4.1 Selection of Independent Variables

Prudent variable selection is a critical step in constructing a robust regression model. In this analysis, the selection of independent variables was guided by significant correlations, laying the foundation for a thorough and meaningful regression analysis. The chosen variables were carefully considered based on their potential impact on the dependent variable, 'target\_deathrate,' and their statistical significance.

**Variable Selection Criteria:**

**Significant Correlations:** Variables with a significant correlation with the target variable were prioritized. This approach helps ensure that selected predictors have a meaningful relationship with the outcome, increasing the model's predictive power.

**Multicollinearity Consideration:** To avoid multicollinearity issues, variables highly correlated with each other were carefully evaluated. In cases of strong correlation, only one variable was selected to maintain the model's stability and interpretability.

**Domain Knowledge:** Consideration of variables with known relevance to cancer rates and outcomes added a layer of domain knowledge to the variable selection process. This approach enhances the model's capability to capture real-world complexities.

**Chosen Variables and Rationale:**

The following variables were selected for inclusion in the regression analysis:

**'incidencerate':** The incidence rate of cancer serves as a fundamental metric, directly related to the target variable.

**'medincome':** Median income is often associated with healthcare access and lifestyle, influencing cancer outcomes.

**'povertypercent':** The percentage of the population in poverty can impact access to healthcare resources and contribute to cancer disparities.

**'percentmarried':** Marital status may have implications for social support and healthcare decisions affecting cancer outcomes.

**'pctbachdeg18\_24' and 'pctbachdeg25\_over':** The percentage of the population with a bachelor's degree may reflect educational influences on health and healthcare choices.

**'pctblack':** Racial demographics, such as the percentage of the population identifying as Black, can contribute to understanding health disparities.

**'pctmarriedhouseholds':** This variable captures the proportion of married households, potentially linked to social support and stability.

The careful consideration of these variables aligns with the goal of constructing a comprehensive regression model that not only explains variance in the target variable but also offers meaningful insights into the factors influencing cancer death rates. This approach ensures that the model is not only statistically robust but also clinically relevant.

### 3.4.2 Regression Analysis

The multiple linear regression analysis yielded a model of substantial complexity, as evidenced by the exhaustive results presented below. The Ordinary Least Squares (OLS) Regression Results provide a detailed overview of the coefficients, standard errors, t-values, and p-values for each variable. These results offer an in-depth understanding of the predictive relationships encoded within the model. The complete set of regression results is available in **Appendix 1** for reference.

**Key Insights:**

* **R-squared (Adjusted R-squared):** The model attains a perfect fit with an R-squared of 1.000, indicating that the chosen independent variables explain the entirety of the variance in the dependent variable.
* **F-statistic:** The F-statistic is exceedingly high (2.811e+25), signifying that the overall model is statistically significant.
* **Coefficients:** The coefficients represent the change in the dependent variable for a one-unit change in the respective independent variable. For instance, the coefficient for 'target\_deathrate' is 1.000, implying a direct and proportional relationship with the dependent variable.
* **P-values:** P-values assess the statistical significance of each coefficient. Variables with p-values less than the conventional threshold (e.g., 0.05) are considered statistically significant.
* **Omnibus, Durbin-Watson, Jarque-Bera, and Skew/Kurtosis:** These diagnostics provide insights into the normality and autocorrelation of residuals. The very low p-value for Omnibus indicates that the residuals are not normally distributed.
* **Multicollinearity:** The notes section highlights the possibility of strong multicollinearity or other numerical problems, suggesting caution in the interpretation of coefficients.

These results offer a comprehensive view of the model's performance and provide a basis for further interpretation and refinement. For the detailed regression results, please refer to **Appendix 1**.

### 3.4.3 Random Forest Feature Importance

In fortifying the robustness of our predictive model, we employed a Random Forest regression model. The rationale behind incorporating a Random Forest lies in its ability to handle complex relationships within the data, capture non-linear patterns, and provide valuable insights into feature importance.

#### 3.4.3.1Methodology

**Handling Non-Linearity:** Unlike linear regression, Random Forest is well-suited to capture non-linear relationships between independent and dependent variables. This is crucial in our analysis where cancer death rates may exhibit intricate patterns that linear models might overlook.

**Ensemble Learning:** Random Forest is an ensemble learning method that combines the predictions of multiple decision trees. Each tree is trained on a random subset of the data, and the ensemble's output is a robust and accurate prediction.

**Feature Importance Calculation:** The Random Forest algorithm inherently provides a mechanism to measure the importance of each feature in making accurate predictions. This is achieved by evaluating how much each feature contributes to the reduction in impurity (e.g., mean squared error) when making decisions in the trees.

#### 3.4.3.2 Importance of Random Forest in Our Analysis

**Handling Multicollinearity:** Random Forests are more robust to multicollinearity compared to traditional linear models. This property ensures that even correlated variables can be included without compromising the model's stability.

**Comprehensive Variable Interaction:** The ensemble nature of Random Forest allows it to capture complex interactions between variables, providing a more nuanced understanding of the relationships contributing to cancer death rates.

**Resilience to Overfitting:** Random Forests are less prone to overfitting, making them suitable for datasets with a large number of features. This ensures that the model generalizes well to new, unseen data.

#### 3.4.3.3 Feature Importance Analysis

After training the Random Forest regression model, we evaluated the model performance using the following metrics:

**Mean Squared Error (MSE):** The MSE of our model is 361.88, which represents the average squared difference between actual and predicted values. Lower MSE values indicate a better fit of the model to the data.

**R-squared (R²):** The R-squared value is 0.42, signifying the proportion of the variance in the target variable ('target\_deathrate') that our model explains. An R-squared of 0.42 indicates a moderate level of predictive accuracy.

#### 3.4.3.4 Model Evaluation Insights

The MSE and R-squared metrics provide valuable insights into the performance of our Random Forest regression model. While the MSE reflects the accuracy of individual predictions, the R-squared value quantifies the overall goodness of fit. These metrics collectively contribute to our understanding of the model's efficacy in predicting cancer death rates.

Afbeelding met tekst, schermopname, borstel, gereedschap

Automatisch gegenereerde beschrijving

Figure 7: Barchart showing feature importance’s (unfortionally unreadable)

#### 3.4.3.5 Significance of Feature Importance

Understanding feature importance is pivotal for interpreting the impact of predictors on cancer death rates. Variables with higher importance values play a more influential role in the model's predictions. This analysis not only enhances the predictive accuracy of our model but also provides actionable insights into the key factors contributing to cancer mortality.

### 3.4.4 Threshold-based Feature Selection

A judicious threshold was employed to distill the essential features, engendering a model focused on the most impactful predictors.

## 3.5 Model Evaluation and Diagnostics

## 3.5.1 Diagnostic Plots

In the quest for a thorough model evaluation, a suite of diagnostic plots was employed to scrutinize the regression model's assumptions and performance. The following plots were generated:

**Residual Autocorrelation Plot**

The residual autocorrelation plot provides insight into whether residuals exhibit patterns across different time points. The absence of significant patterns suggests that the residuals are independent, a crucial assumption for regression analysis. The plot generated with **plot\_acf** visually inspects the autocorrelation of residuals at various lags. If most autocorrelations fall within the confidence bands, it indicates that the residuals are likely independent.

Afbeelding met tekst, schermopname, lijn, diagram

Automatisch gegenereerde beschrijving

Figure 8: Autocorrelation

**Residual vs. Predicted Value Plot**

This scatter plot juxtaposes predicted values against residuals. It serves as a diagnostic tool to identify patterns or trends in residuals concerning the predicted values. Ideally, the plot should show a random scattering of points around the horizontal axis, indicating homoskedasticity (constant variance) and no discernible patterns. Any systematic deviation may indicate issues with model assumptions.

Afbeelding met tekst, schermopname, diagram, lijn

Automatisch gegenereerde beschrijving

Figure 9: Residu plot for heteroskedasticity

**QQ-Plot (Quantile-Quantile Plot)**

The QQ-plot compares the distribution of residuals against a theoretical normal distribution. Deviations from a straight line suggest departures from normality. A well-behaved QQ-plot with residuals forming a straight line indicates that the residuals are approximately normally distributed.

Afbeelding met tekst, schermopname, lijn, diagram

Automatisch gegenereerde beschrijving

Figure 10: QQ-plot of Residuals

## 3.5.2 Heteroskedasticity Test

In the evaluation of the regression model, a critical aspect is the assessment of heteroskedasticity, which refers to the potential variability in the spread of residuals across different levels of independent variables. Heteroskedasticity violates the assumption of homoskedasticity, where the variance of the residuals should be constant across all levels of predictors.

The Breusch-Pagan test was employed as a robust statistical tool for detecting the presence of heteroskedasticity. This test scrutinizes whether the variance of the residuals exhibits consistency or variability across the spectrum of independent variables. The underlying hypothesis is as follows:

* **Null Hypothesis (H0):** Assumes homoskedasticity, indicating that the variance of the residuals is constant.
* **Alternative Hypothesis (H1):** Posits the presence of heteroskedasticity, suggesting that the variance of residuals is not constant.

The output of the Breusch-Pagan test is a p-value, which is then compared to a significance level (commonly 0.05) to make an inference about the null hypothesis.

* **p-value for Heteroskedasticity:** 0.04081553258714084

The obtained p-value is below the common significance level of 0.05. Therefore, based on the significance level criterion, we would reject the null hypothesis. This provides evidence in favor of the presence of heteroskedasticity in the residuals.

Interpreting this result implies that the variability of residuals is not consistent across all levels of the independent variables. This information is crucial for understanding the reliability and robustness of the regression model. Addressing heteroskedasticity may involve employing heteroskedasticity-robust standard errors or exploring transformations to stabilize the variance.

It's worth noting that while statistical significance is important, the practical significance of heteroskedasticity should also be considered. If the effect is minor, its impact on the overall model might be negligible. However, if substantial, further diagnostics and corrective measures may be warranted.

## 3.5.3 Multicollinearity Assessment

Multicollinearity, the phenomenon where predictor variables in a regression model are correlated, was meticulously evaluated using Variance Inflation Factors (VIF). VIF provides insights into the extent of multicollinearity by quantifying how much the variance of an estimated regression coefficient increases if predictors are correlated. A high VIF indicates a potential issue, as it suggests that a predictor variable can be linearly predicted from others.

Let's delve into the interpretation of the VIF output:

Variable VIF

0 const 3.329811e+06

1 incidencerate 1.371207e+00

2 pctbachdeg25\_over 6.349575e+00

3 povertypercent 9.278726e+00

4 avgdeathsperyear 7.295978e+01

5 pcths25\_over 4.187822e+00

6 pctpubliccoverage 2.838434e+01

7 medincome 8.986007e+00

8 pctblack 4.513084e+00

9 pctemployed16\_over 6.144661e+00

...

30 percentmarried 1.287038e+01

31 medianagefemale 1.199650e+01

**Interpretation:**

1. **Constant (const):** The constant term exhibits an extremely high VIF, indicating potential collinearity issues. This might be due to the nature of the intercept in the presence of highly correlated predictors.
2. **Other Variables:** Generally, VIF values around 1 indicate low multicollinearity, while values above 5 or 10 are often considered problematic. Several variables, such as 'avgdeathsperyear,' 'pctpubliccoverage,' and 'pctsomecol18\_24,' have VIF values that may warrant further investigation.
3. **Recommendations:**
   * **High VIF Variables:** Variables with high VIF values may benefit from closer scrutiny. It's essential to assess the practical impact of multicollinearity on the model and consider potential remedies.
   * **Collinear Pairs:** Identify pairs of variables with high VIF and consider whether it makes sense in the context of the study. If certain variables are inherently correlated, it might be necessary to choose one over the other or apply dimensionality reduction techniques.
4. **Further Investigation:** While high VIF values indicate potential multicollinearity, it's crucial to consider the overall model performance, the specific objectives of the analysis, and potential remedies. If multicollinearity is found to be problematic, techniques like feature selection, dimensionality reduction, or regularization methods may be considered.

If you wish to see the entire output, please refer to **Appendix 2** for detailed information on each variable's VIF.

## 3.6 K-Fold Cross-Validation

### 3.6.1 Model Selection and Evaluation

In order to rigorously evaluate the robustness of the predictive model, a K-fold cross-validation approach was employed with k=10. This technique involves dividing the dataset into 10 subsets, using 9 subsets for training and the remaining subset for testing in an iterative fashion. The mean squared error (MSE), computed across these folds, served as a critical metric for assessing the model's predictive performance. A lower MSE indicates better performance, reflecting smaller differences between predicted and actual values.

**Mean Squared Error over 10-fold cross-validation:** 462.9903721451322

This metric provides a consolidated measure of the model's accuracy, considering its performance across different subsets of the data. In this context, a MSE of approximately 462.99 suggests that, on average, the squared differences between predicted and actual cancer death rates are relatively moderate. It signifies that the model exhibits a reasonable level of accuracy in capturing the underlying patterns in the data.

## 3.7 Interpretation and Insights

The decisions made in the code, including variable selection strategies, feature importance analysis, and model evaluation techniques, were carefully chosen to create a model that is not only accurate but also interpretable and robust. The MSE obtained through cross-validation serves as a benchmark for the model's performance, demonstrating its ability to generalize well to new, unseen data.

In summary, the results presented in this section represent the culmination of a comprehensive data analysis journey. From the initial stages of exploring and engineering features to the final steps of model evaluation and interpretation, each aspect of this analytical process contributes cohesively to uncovering intricate patterns within the dataset. The resulting predictive model provides valuable insights into cancer death rates, striking a balance between accuracy and generalizability.

# 4. Discussion

## 4.1 Challenges in Analysis

During the course of the analysis, several challenges were encountered, each contributing to the complexity and nuance of the undertaken predictive modeling task. One notable challenge was the initial difficulty in sourcing a suitable and comprehensive dataset. The decision to participate in a challenge was driven by the desire for more guidance and structure in the analysis, as well as the recognition of the need for careful interpretation of the data due to its inherent complexity.

The task of merging two datasets added another layer of complexity. Integrating information from diverse sources required meticulous handling of shared variables to ensure a meaningful and coherent dataset. Additionally, the challenge demanded a nuanced approach to data interpretation, given the varied nature of the information encapsulated within the merged dataset.

## 4.2 Choices Made and Methodological Considerations

The choices made during the analysis process were methodically crafted to address specific challenges and ensure the creation of a robust predictive model. Prudent decisions guided variable selection, a critical aspect in the development of an effective model. Variables were chosen based on their significant correlations, laying the foundation for a comprehensive regression analysis. The incorporation of relevant predictors was pivotal in crafting a model capable of capturing the intricacies of cancer death rates.

Feature importance was assessed using a random forest regression model, enhancing the model's robustness and providing insights into the contribution of each predictor. Subsequent threshold-based feature selection distilled the essential features, ensuring a focus on the most impactful variables and streamlining the model for interpretability and efficiency.

The application of K-fold cross-validation, a powerful model evaluation technique, provided a comprehensive assessment of the model's predictive performance. The mean squared error (MSE) across folds served as a reliable metric, offering a consolidated measure of accuracy and the model's ability to generalize to unseen data.

## 4.3 Assessment of Model Quality

The assessment of model quality revolves around the interplay of choices made, methodological rigor, and the performance metrics obtained. The multiple linear regression model, evaluated through the Ordinary Least Squares (OLS) Regression Results, showcased a perfect fit (R-squared: 1.000), which, while intriguing, raises concerns of overfitting and potential issues in the dataset.

The random forest regression, with its feature importances and threshold-based selection, added a layer of complexity to the analysis, ensuring that the model focused on the most impactful predictors. However, the evaluation metrics, such as MSE and R-squared, provide a more balanced view of the model's performance. The MSE obtained through K-fold cross-validation offers a robust measure of predictive accuracy, considering the model's performance across different subsets of the data.

In summary, while the models exhibit intriguing characteristics, the discussion extends to the delicate balance between complexity and interpretability, the potential impact of overfitting, and the need for cautious interpretation of results that exhibit extreme values. The chosen methodologies and the resulting model outcomes invite a nuanced discussion on the trade-offs involved in predictive modeling, emphasizing the importance of critical evaluation and continuous refinement in the quest for accurate and meaningful insights.

# 5. Conclusion

## 5.1 Summary of Key Findings

The culmination of the analytical journey encapsulates key insights into cancer death rates, obtained through a meticulous process of data exploration, feature engineering, and predictive modeling. The primary findings can be distilled into several focal points:

* **Model Complexity and Overfitting:** The multiple linear regression model exhibited a perfect fit, as indicated by an R-squared value of 1.000. However, this level of fit raises concerns about overfitting, where the model may have memorized the training data rather than capturing underlying patterns. Balancing model complexity and interpretability is crucial for meaningful and generalizable results.
* **Random Forest Robustness:** The random forest regression model, bolstered by feature importances and threshold-based feature selection, introduced a layer of robustness. This approach ensured a focus on essential predictors, contributing to a more interpretable and efficient model.
* **Model Evaluation Through Cross-Validation:** The use of K-fold cross-validation provided a comprehensive evaluation of predictive performance. The mean squared error (MSE) across folds served as a reliable metric, offering insights into the model's accuracy and generalization capabilities.

## 5.2 Possible Future Steps

In charting the course for future steps, several avenues beckon for further exploration and refinement:

* **Fine-Tuning Model Complexity:** Given the perfect fit observed in the multiple linear regression model, a nuanced approach to model complexity is warranted. Fine-tuning the model, perhaps through regularization techniques, can mitigate overfitting and enhance generalization.
* **Feature Engineering Iterations:** Continuous refinement of feature engineering strategies can contribute to a more nuanced understanding of predictive relationships. Exploring different ways to encode and represent variables may uncover latent patterns within the data.
* **Incorporating Domain Expertise:** Collaborating with domain experts, such as epidemiologists or healthcare professionals, can enhance the model's interpretability and relevance. Their insights can inform variable selection and feature engineering, aligning the model more closely with real-world dynamics.
* **Exploration of Ensemble Models:** Beyond random forest regression, exploring other ensemble models or advanced machine learning techniques may offer a comparative assessment of predictive performance. Models like gradient boosting or neural networks could provide additional insights.
* **Dynamic Data Integration:** Incorporating dynamic data sources, such as temporal trends or socio-economic changes over time, can contribute to a more dynamic and adaptable predictive model. Time-series analysis and incorporating temporal features could be pivotal in capturing evolving patterns.

# 7. Ethics

In the pursuit of unraveling insights from the intricate fabric of cancer death rates, our commitment to ethical considerations has been unwavering. This chapter delves into the ethical dimensions of our data science journey, encompassing privacy protection, bias mitigation, and the responsible application of our models.

*Privacy Preservation:* As custodians of health-related data, we acknowledge the paramount importance of privacy. Rigorous measures have been implemented to safeguard individual information, ensuring that any data shared or published adheres to stringent anonymization protocols. Our commitment to privacy extends beyond compliance, reflecting a dedication to the trust bestowed upon us by individuals contributing to this dataset.

*Guarding Against Bias:* The specter of bias looms large in data science, especially in the realm of healthcare. Prudent variable selection, random forest feature importance analysis, and threshold-based feature selection were not just technical choices but ethical imperatives. By meticulously considering the impact of each predictor, we aimed to fortify our models against inadvertent biases that could lead to unfair or discriminatory outcomes.

*Transparent Decision-Making:* Transparency has been our guiding principle in every stage of the CRISP-DM process. Each choice, from data preprocessing to model evaluation, has been meticulously documented, laying bare the rationale behind our decisions. This transparency serves as a bulwark against opacity and promotes accountability in our analyses.

*Equity in Model Outcomes:* The models developed and evaluated underwent rigorous scrutiny for equity and fairness. The application of random forest regression and k-fold cross-validation was not merely an exercise in predictive accuracy but a conscious effort to ensure that our models do not perpetuate disparities in cancer death rate predictions across diverse demographic groups.

*Informed Consent:* Recognizing the ethical significance of data usage, particularly in the healthcare domain, our project prioritizes informed consent. Where applicable, stakeholders have been duly informed about the purpose of data collection, the methodologies employed, and the potential implications of our analyses. This commitment to transparency builds a foundation of trust with those impacted by our work.

*Continuous Evaluation:* Ethical considerations are not static; they require continuous evaluation. Our diagnostic plots, heteroskedasticity tests, and multicollinearity assessments serve not only as measures of model diagnostics but also as ethical checkpoints. Regularly scrutinizing the model's performance helps ensure that our ethical standards persist as our understanding of the data evolves.

In weaving together the fabric of our ethical considerations, we strive not only to generate meaningful insights into cancer death rates but to contribute responsibly to the broader healthcare narrative. This chapter serves as a testament to our dedication to the highest ethical standards in data science, acknowledging the profound impact our analyses can have on individuals and society.

# Bibliography

*Data World* . (sd). Opgehaald van https://data.world/exercises/linear-regression-exercise-1/workspace/project-summary?agentid=exercises&datasetid=linear-regression-exercise-1

## Appendix Appendix 1: OLS Regression Results

OLS Regression Results

==============================================================================

Dep. Variable: target\_deathrate R-squared: 1.000

Model: OLS Adj. R-squared: 1.000

Method: Least Squares F-statistic: 2.811e+25

Date: Mon, 18 Dec 2023 Prob (F-statistic): 0.00

Time: 20:43:51 Log-Likelihood: 13549.

No. Observations: 591 AIC: -2.705e+04

Df Residuals: 569 BIC: -2.696e+04

Df Model: 21

Covariance Type: nonrobust

===============================================================================================

coef std err t P>|t| [0.025 0.975]

-----------------------------------------------------------------------------------------------

const 2.173e-11 4.88e-11 0.445 0.656 -7.41e-11 1.18e-10

target\_deathrate 1.0000 5.61e-14 1.78e+13 0.000 1.000 1.000

incidencerate 9.347e-15 2.45e-14 0.382 0.702 -3.87e-14 5.74e-14

medincome 1.735e-16 3.35e-16 0.518 0.605 -4.84e-16 8.31e-16

povertypercent -2.442e-15 5.78e-13 -0.004 0.997 -1.14e-12 1.13e-12

percentmarried 0 5.08e-13 0 1.000 -9.98e-13 9.98e-13

pcths18\_24 -2.776e-16 1.5e-13 -0.002 0.999 -2.95e-13 2.94e-13

pctbachdeg18\_24 1.665e-15 3.8e-13 0.004 0.997 -7.44e-13 7.48e-13

pcths25\_over 1.665e-16 3.17e-13 0.001 1.000 -6.23e-13 6.23e-13

pctbachdeg25\_over 7.772e-16 5.06e-13 0.002 0.999 -9.94e-13 9.95e-13

pctemployed16\_over -4.996e-16 3.41e-13 -0.001 0.999 -6.69e-13 6.68e-13

pctunemployed16\_over 1.221e-15 5.53e-13 0.002 0.998 -1.09e-12 1.09e-12

pctprivatecoverage -4.441e-16 9.14e-13 -0.000 1.000 -1.8e-12 1.79e-12

pctprivatecoveragealone -4.441e-15 1.1e-12 -0.004 0.997 -2.16e-12 2.15e-12

pctempprivcoverage -1.305e-15 3.75e-13 -0.003 0.997 -7.38e-13 7.36e-13

pctpubliccoverage -2.22e-15 9.67e-13 -0.002 0.998 -1.9e-12 1.9e-12

pctpubliccoveragealone 3.109e-15 1.13e-12 0.003 0.998 -2.22e-12 2.23e-12

pctblack -3.331e-16 1.18e-13 -0.003 0.998 -2.32e-13 2.32e-13

pctotherrace 1.998e-15 4.06e-13 0.005 0.996 -7.96e-13 8e-13

pctmarriedhouseholds -1.776e-15 4.89e-13 -0.004 0.997 -9.62e-13 9.59e-13

binnedinc\_(61494.5, 125635] 0 6.58e-12 0 1.000 -1.29e-11 1.29e-11

binnedinc\_[22640, 34218.1] 3.553e-15 5.4e-12 0.001 0.999 -1.06e-11 1.06e-11

==============================================================================

## Appendix 2: VIF Results

Variable VIF

0 const 3.329811e+06

1 incidencerate 1.371207e+00

2 pctbachdeg25\_over 6.349575e+00

3 povertypercent 9.278726e+00

4 avgdeathsperyear 7.295978e+01

5 pcths25\_over 4.187822e+00

6 pctpubliccoverage 2.838434e+01

7 medincome 8.986007e+00

8 pctblack 4.513084e+00

9 pctemployed16\_over 6.144661e+00

10 popest2015 6.122643e+01

11 pctunemployed16\_over 3.095623e+00

12 pctprivatecoverage 1.854791e+01

13 pctotherrace 1.621029e+00

14 pctpubliccoveragealone 2.943436e+01

15 birthrate 1.189759e+00

16 pctbachdeg18\_24 6.320170e+03

17 avghouseholdsize 5.485905e+00

18 pctasian 2.260362e+00

19 pctnohs18\_24 2.040275e+04

20 pcths18\_24 3.051327e+04

21 pctmarriedhouseholds 1.290109e+01

22 countyfips 1.244959e+00

23 pctsomecol18\_24 4.128873e+04

24 pctwhite 6.537834e+00

25 avganncount 2.023309e+01

26 statefips 1.212141e+00

27 medianagemale 9.286522e+00

28 geography\_Woodson County, Kansas 1.074776e+00

29 pctempprivcoverage 8.281061e+00

30 percentmarried 1.287038e+01

31 medianagefemale 1.199650e+01