Project

Jinfeng Zhu

Student ID: 47564644

Contents

[1. Introduction 2](#_Toc113370887)

[2. Exploratory Data Analysis 2](#_Toc113370888)

[2.1 Source of Data 2](#_Toc113370889)

[2.2 Data Exploration 2](#_Toc113370890)

[2.2.1 Checks on Duplicate and Data Type 2](#_Toc113370891)

[2.2.2 Tidy Form Conversion 3](#_Toc113370892)

[2.2.3 Checks on Internal Consistency 3](#_Toc113370893)

[2.2.4 Sense Checking 3](#_Toc113370894)

[2.3 Data Manipulation 3](#_Toc113370895)

[2.4 Visualisation 4](#_Toc113370896)

[2.5 Final Dataset 5](#_Toc113370897)

[3. Evaluation of the Preliminary Multiple Linear Regression Model 9](#_Toc113370898)

[3.1 Correlation 9](#_Toc113370899)

[3.2 Fitting Regression Model 10](#_Toc113370900)

[3.2.1 The Suggested Preliminary Model 10](#_Toc113370901)

[3.2.2 Model Improvement 11](#_Toc113370902)

[4. Australia Contextualization 13](#_Toc113370903)

[Appendix A – Technical Analysis 15](#_Toc113370904)

[Appendix B – Contextualisation Notes 15](#_Toc113370905)

[Appendix C – R Code 18](#_Toc113370906)

[Appendix D – Reference 18](#_Toc113370907)

# Introduction

Stroke has become the most fatal disease in Australia. Therefore, Stroke Foundation (SF) Australia has approached AA consulting firm with the aim of deepening their understanding of risk factors associate with stroke mortality rates.

This technical report, produced by AA consulting firm, provides a preliminary exploratory data analysis on the US national datasets and evaluates a preliminary regression model with suggestions on potential improvement.

Besides, this report also explores the feasibility of the same type of analysis to Australian context along with contextualisation notes on business understanding and stakeholder analysis.

# Exploratory Data Analysis

Exploratory data analysis (EDA) on the US datasets was conducted in R.

## Source of Data

The US datasets were published by various US entities including the US Centers for Disease Control and Prevention (cdc.gov) for stroke mortality and incidence data and Census (census.gov) for poverty, income, health insurance and population data. All these datasets contain the information for over 3,100 counties across the United States for the year of 2015.

## Data Exploration

[uncertainty: incidence rate over 75?]

### Checks on Duplicate and Data Type

Initial checks were performed on duplicated values to ensure their non-existence in any of the dataset. Investigation in missing values was intentionally ignored at this stage as they were known to exist due to the insufficient data size or the reason of confidentiality. Missing values will be explored in the later stage of EDA.

Checks on data format were also performed for each dataset and types of some variables were converted. For example, variable “Age-Adjusted Death Rate” and “Average Deaths per Year” in the stroke mortality dataset were stored as character, which have been converted to numeric variables for further analysis.

### Tidy Form Conversion

The stroke mortality dataset was reshaped by separating out the state information from the county information. For example, “Perry County, Kentucky” was separated into 2 columns as “Perry County” and “Kentucky”.

Based on the domain knowledge, the United States is made up of a total of 50 states, plus the District of Columbia (DC). Checks and corrections were performed for some observations to ensure the unique number of states was 52, with an additional entry for US as a whole.

### Checks on Internal Consistency

To gain a better understanding of the datasets, internal consistency checks on a few variables were performed, especially for those being used for the preliminary regression model. For instance, the relationship among certain variables was examined. Below lists several examples:

* non-institutionalized population was proved to be equal to the sum of male non-institutionalized population and female non-institutionalized population within the health insurance dataset.
* Population for whom poverty status is determined was proved to be equal to the sum of below poverty level population and above poverty level population.

### Sense Checking

Sense checking was performed on the population dataset, where the total US national population at the mid-year 2015 was 642 million. However, based on the domain knowledge, the population of the United States should be in the range of 300 to 350 million. The number in the population dataset appeared to be doubled in some way. To resolve the uncertainty, death dataset was cross referenced. The number of stroke death and the stroke mortality rate was used to estimate the population for each county, which suggested the US national population in 2015 was 329 million. Therefore, a decision was made that all the county level data in the population dataset was halved and the total population was reduced to 321 million, which is more consistent with the real world.

## Data Manipulation

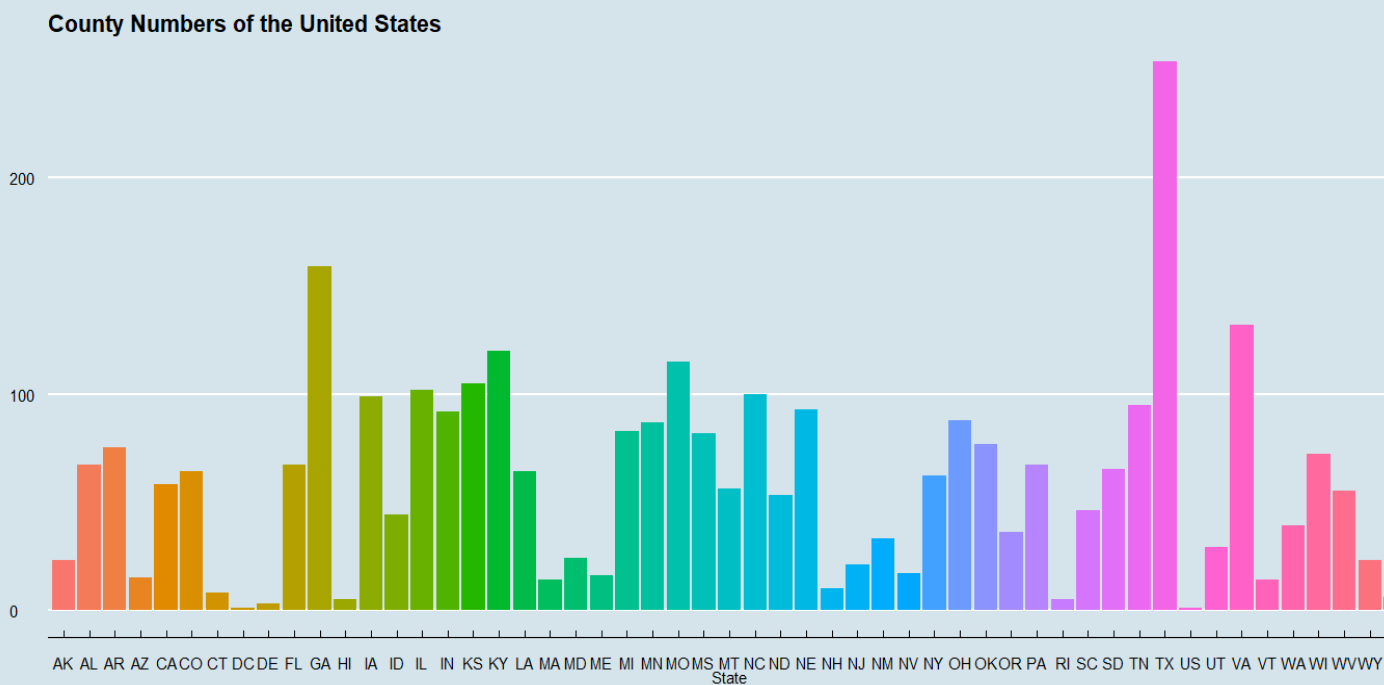
For the incidence dataset, expert opinion on the recent incidence trend was adopted where “stable” was assumed for counties whose data was suppressed. As a result, more than 90% of the counties had a stable stroke incidence experience.

|  |  |  |  |
| --- | --- | --- | --- |
| Falling | Rising | Stable | Total |
| 200 | 43 | 2,898 | 3,141 |
| 6.4% | 1.4% | 92.3% | 100% |

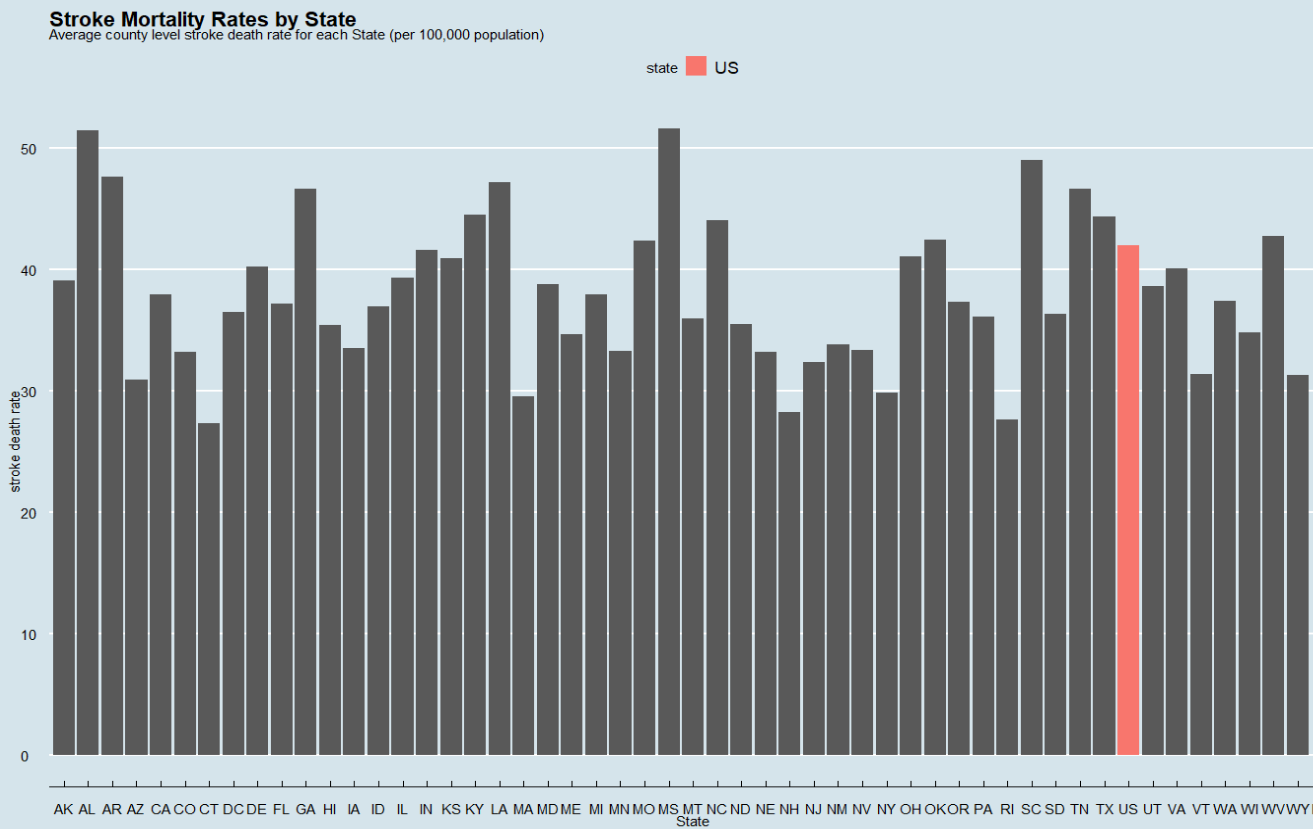
As “recent trend” is a categorical variable, dummy encoding was applied and two more numerical variables ("trend\_falling " and “trend\_rising”) taking a value of 1 or 0 were introduced.

## Visualisation

To view the completeness of the dataset, the number of counties for each state was shown in the bar plot below. All 50 states and DC were included in the datasets. Majority of the states contributed around 100 counties or less, while Texas (TX) stood out by having more than 200 valid entries, which seemed to be an outlier. However, Texas covers the most (254) counties among all the US states[[1]](#footnote-1), followed by Georgia which covers 159 counties, which is consistent with the observation from our dataset.



Another bar plot compares the stroke death rate by state against the national average (US). It appears that the average US stroke mortality rate was around 41 per 100,000 population while the stroke mortality rates for most states ranged from 30 to 50 per 100,000 population in the year of 2015. Overall, the death rate seemingly had a stable distribution without enormous variances.



## Final Dataset

After the initial understanding and checking on each dataset, several steps were taken to reach the final dataset for the modelling purpose.

**Step 1 – Create Unique Key**

To join the 6 datasets into a final dataset for modelling, a 5-digit FIPS was derived from the State FIPS and the County FIPS in each dataset and was treated as the unique key. 6 datasets were then combined by applying mutating join in sequence.

**Step 2 – Check Missing Values**

Checks on missing values were performed at this point on the combined dataset. In order to facilitate the detection of missing values, all the suppressed values "\*" were converted to NA. As expected, missing values were observed in various features from each dataset due to different reasons.

|  |  |
| --- | --- |
| **Source of Missing Values** | **Action** |
| Response variable – stroke death rate | Observations without a response variable were completely removed from the dataset. |
| The observation for US | The entire US observation was removed because none of the social determinant datasets contained information for "US". |
| Income | Since only the median income for the whole population (income\_001) was chosen for the preliminary model, missing values in other income variables were ignored. |
| Nevada Data | All Nevada observations were removed as they were not available in death or incidence datasets. |
| Incidence Dataset | Remove observations where both incidence rate and annual count were missing due to the confidentiality. As a result, Kansas and Minnesota observation were removed.  For the remaining missing values in the incidence rate, the associated counts were all less than 5. Therefore, those missing values were imputed by substituting the NAs with 0. |

After dealing with all the missing values, observations from 48 States remained in the dataset. Observations from Nevada, Kansas and Minnesota were removed during the course of handling missing values.

**Step 3 – Generate Explanatory Variables**

Two explanatory variables used in the preliminary multiple linear regression model were not readily available in the existing dataset. Hence another step was needed to create those required variables in the final dataset.

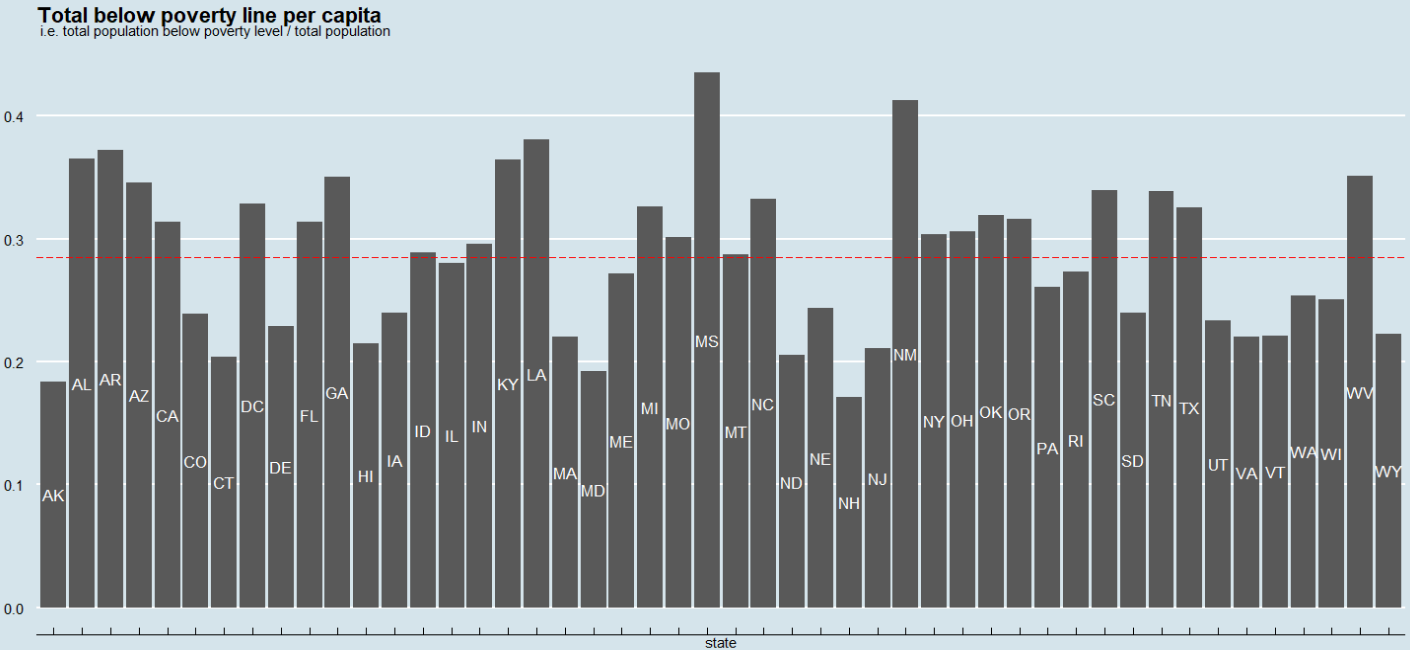
1. ***Total below poverty line per capita*** *= total population below poverty level / total population*
2. ***Total without health insurance per capita*** *= sum of the population without health insurance coverage for each age group / total population*

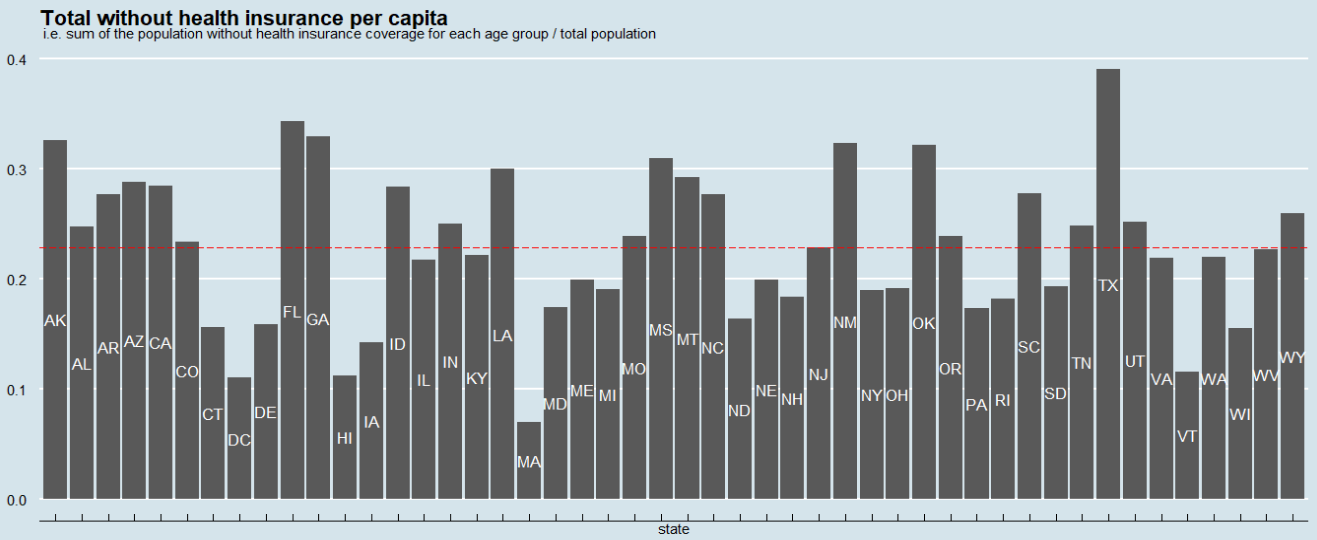
**Step 4 – Reasonableness Checking**

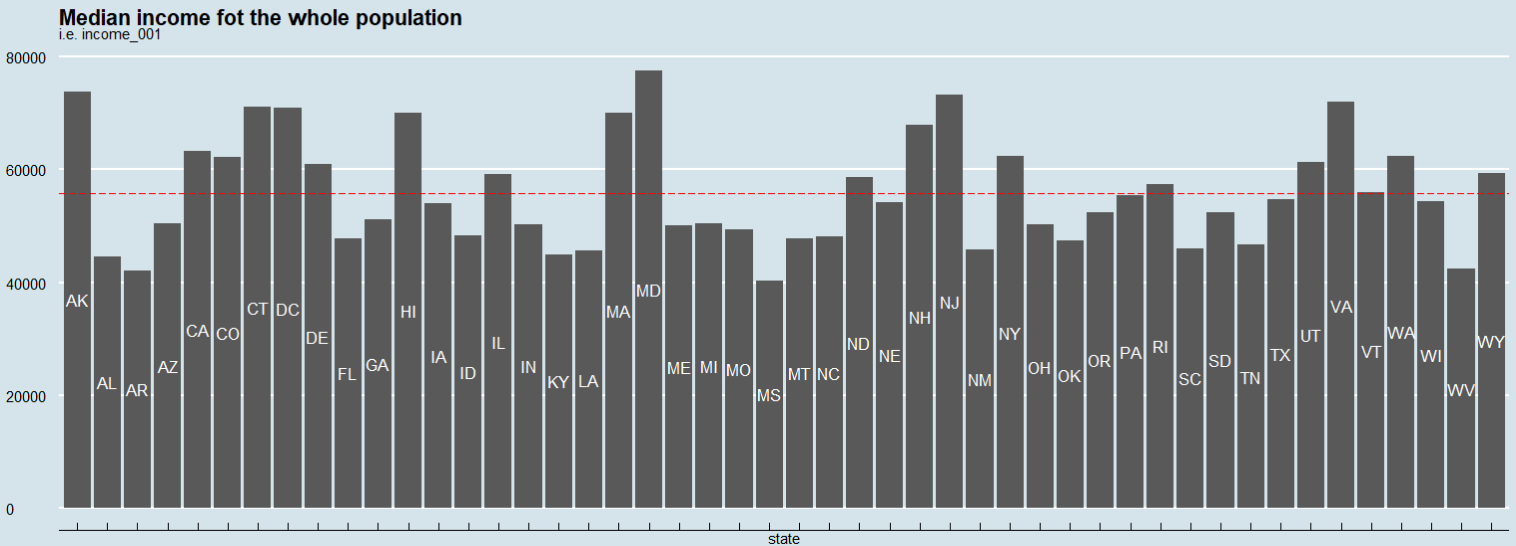
Before reaching the final dataset, some reasonableness checking was performed on the selected explanatory variables at the State level. This was achieved by applying the data visualisation to all those explanatory variables to identify any anomalies such as outliers or unusual trends.

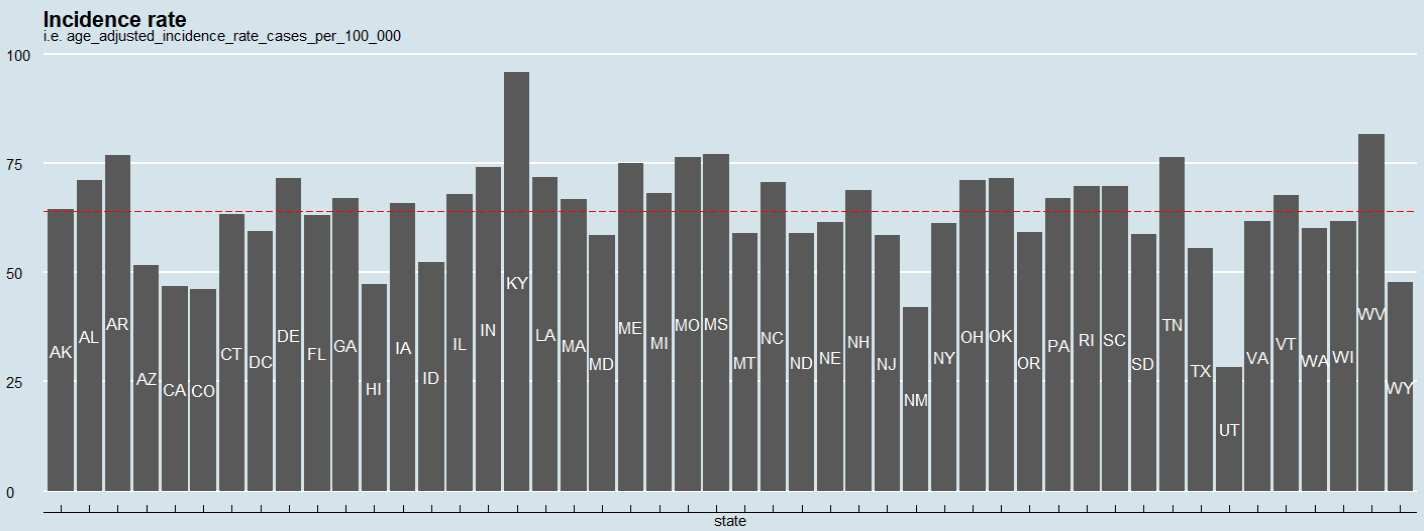
From the below 4 plots, no anomalies were noticed and the national average was as follows:

|  |  |
| --- | --- |
| **Explanatory Variable** | **National Average** |
| *Total below poverty line per capita* | *29%* |
| *Total without health insurance per capita* | *23%* |
| *Median income on average* | *USD$56,000* |
| *Incidence rate* | *65 per 100,000 population* |

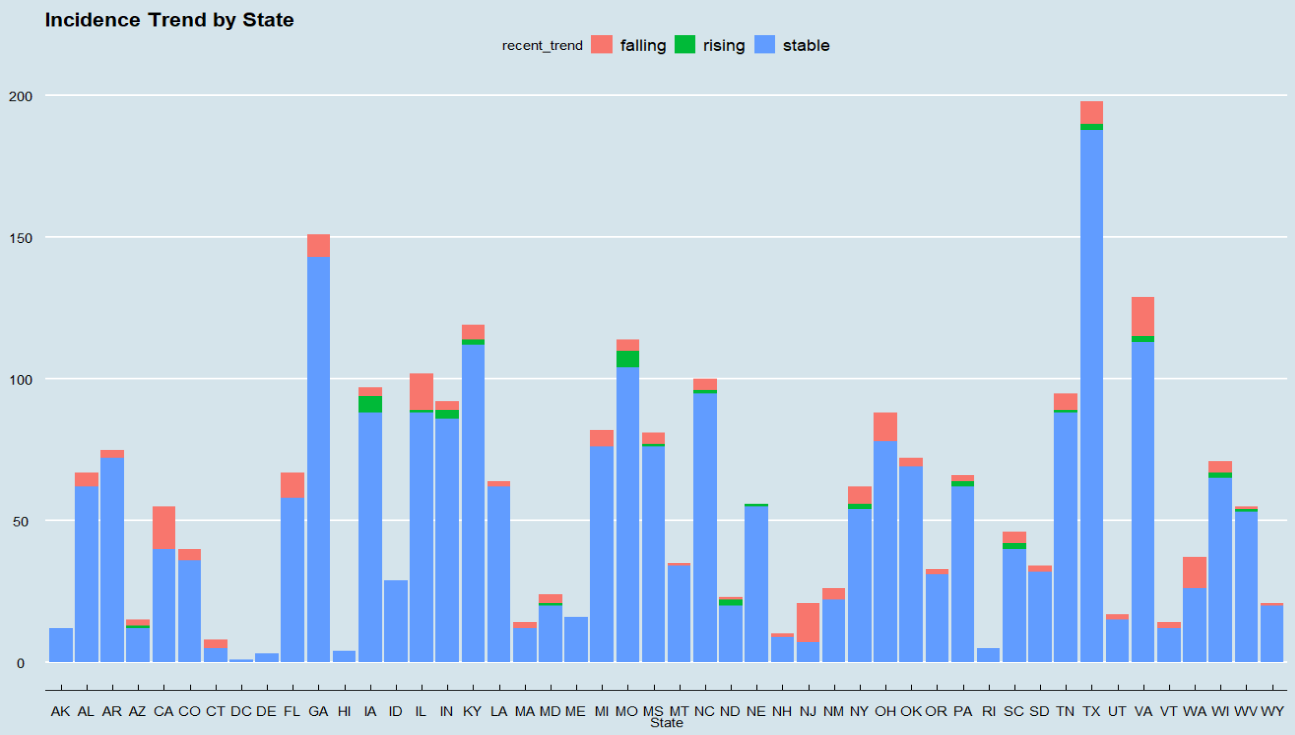


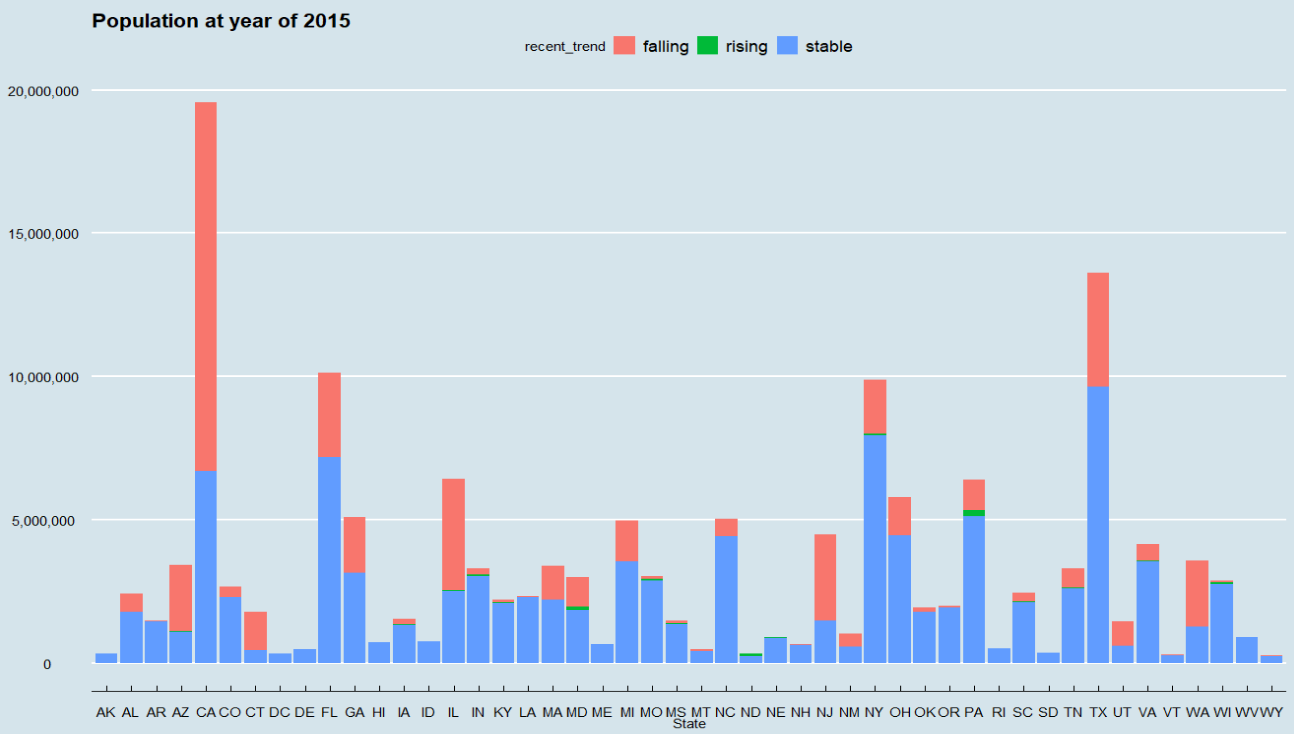






When exploring the incidence trend and population at the same time, it can be seen that falling trend existed in a small number of counties with large population. For example, less than 10% of the counties in California experienced a falling trend while more than 50% of the population in California resided in those counties. The reason may be that people who live in large counties can easily access better medical treatment when compared to those who live in small counties.





**Step 5 – Reaching the Final Dataset**

The final dataset was generated by selecting a subset of the joined dataset with the following variables:

* State
* Stroke mortality rate (the response variable)
* Total below poverty line per capita
* Total without health insurance per capita
* Medium income for the whole population
* Incidence rate
* Falling trend for incidence rate
* Rising trend for incidence rate
* Population

# Evaluation of the Preliminary Multiple Linear Regression Model

This section focuses on the evaluation of the preliminary multiple linear regression model suggested by Sam. It also explores the potential improvements that can be made to the preliminary model.

## Correlation

Before fitting the linear regression model, an initial understanding of the model was established by examining the correlation among all the variables in the final dataset, which is shown in the following graph.



Conceptually, the death rate of a certain disease is associated with patients’ wealth level, which can be depicted from different aspects such as the level of income and the status of insurance coverage. In addition, population is indicative of the size of the county and it may imply the level of urbanization and the quality of health care system to some extent in that county, which then could possibly affect the stroke death rate.

As expected, the graph shows that the response variable “stroke mortality” was positively correlated with covariates “below poverty line per capita”, “no health insurance per capita” and “incidence rate”, while it was negatively correlated with covariate “median income”. However, response variable’s relationship with incidence trend and population was not significant based on the dataset.

It is worth noting that “median income” and “below poverty line per capita” appeared to be significantly correlated at the county level. This may suggest the removal of one of them in the regression model.

## Fitting Regression Model

### The Suggested Preliminary Model

First of all, a multiple linear regression model was fitted with all the covariates in the final dataset as suggested by Sam:

Where,

* Y is the response variable stroke death rate;
* is the coefficient for explanatory variables;
* is the residual

The fitting summary is shown as below.

|  |  |  |
| --- | --- | --- |
| **Fit 1** | **Coefficient** | **p-value** |
| (Intercept) | 29.65 | 0.0000 |
| below\_poverty | 13.03 | 0.0000 |
| no\_health\_ins | 13.03 | 0.0000 |
| median\_incom | 0.00 | 0.0000 |
| incidence | 0.12 | 0.0000 |
| incidence\_falling | -1.06 | 0.0530 |
| incidence\_rising | -1.57 | 0.1661 |
| population | 0.00 | 0.0553 |
| R-squared | 0.3052 |  |
| Adjusted R-squared | 0.3033 |  |

As per the statistical diagnosis (*p-value*) result above, the preliminary model contained insignificant covariates (*incidence\_falling, incidence\_rising, and population*). Addtionally, R2 and adjusted R2 indicated that around 30% of the variability in the response variable was explained by the preliminary model with those selected explanatory variables.

Other statistical diagnostic[[2]](#footnote-2) was also performed to examine the validity of the linear regression model, which indicated that the choice of multiple linear regression model was reasonable for this project.

Overall, the preliminary linear model incorporated redundant variables and it was not a good fit for predicting the stroke death rate.

### Model Improvement

To make further improvement on the suggested model, insignificant covariates need to be removed in sequence.

**Step 1 – Remove “incidence\_rising”**

“Incidence\_rising” was removed first as it had the highest p-value in the previous fitting, which suggested a low significance. R2 was hardly affected by the removal of “Incidence\_rising”.

|  |  |  |
| --- | --- | --- |
| **Fit 2** | **Coefficient** | **p-value** |
| (Intercept) | 29.60 | 0.0000 |
| below\_poverty | 12.96 | 0.0000 |
| no\_health\_ins | 13.14 | 0.0000 |
| median\_income | 0.00 | 0.0000 |
| incidence | 0.12 | 0.0000 |
| incidence\_falling | -1.04 | 0.0579 |
| population | 0.00 | 0.0581 |
| R-squared | 0.3047 |  |
| Adjusted R-squared | 0.3031 |  |

**Step 2 – Remove “incidence\_falling”**

“Incidence\_falling” was removed to exclude all the incidence trend related impact from the fitting. Notably, “population” was significant at 5% level but not at 1% level. Besides, its coefficient was nearly zero. Therefore, “population” is considered to be removed from the model as the next step.

|  |  |  |
| --- | --- | --- |
| **Fit 3** | **Coefficient** | **p-value** |
| (Intercept) | 29.73 | 0.0000 |
| below\_poverty | 12.81 | 0.0000 |
| no\_health\_ins | 13.16 | 0.0000 |
| median\_income | 0.00 | 0.0000 |
| incidence | 0.12 | 0.0000 |
| population | 0.00 | 0.0136 |
| R-squared | 0.3037 |  |
| Adjusted R-squared | 0.3024 |  |

**Step 3 – Remove “population”**

“Population” was removed in this step.

|  |  |  |
| --- | --- | --- |
| **Fit 4** | **Coefficient** | **p-value** |
| (Intercept) | 30.46 | 0.0000 |
| below\_poverty | 11.94 | 0.0000 |
| no\_health\_ins | 13.07 | 0.0000 |
| median\_income | 0.00 | 0.0000 |
| incidence | 0.12 | 0.0000 |
| R-squared | 0.3021 |  |
| Adjusted R-squared | 0.3011 |  |

**Step 4 – Remove “median\_income”**

Although “median\_income” was extremely significant in the previous fitting, its coefficient was essentially zero, which means it did not contribute to the variability of stroke mortality. This may also be explained by the collinearity between “median\_income” and “below\_poverty”. From the correlation plot in section 3.1, these 2 features were strongly negatively correlated with each other, which means they likely represent the same information in a different way. As a result, “median\_income” has been removed while R2 slightly decreased by 1% when compared to the preliminary model.

|  |  |  |
| --- | --- | --- |
| **Fit 5** | **Coefficient** | **p-value** |
| (Intercept) | 21.66 | 0.0000 |
| below\_poverty | 19.98 | 0.0000 |
| no\_health\_ins | 13.77 | 0.0000 |
| incidence | 0.13 | 0.0000 |
| R-squared | 0.2921 |  |
| Adjusted R-squared | 0.2913 |  |

**Step 5 – Model assessment**

At this stage, all the variables were significant according to the statistical diagnosis. Besides, they were relatively independent features representing the level of wealth and health in a certain county, although a mild correlation has been detected among those 3 covariates in the correlation plot.

In conclusion, the suggested linear model can be improved by reducing the unnecessary model complexity and only incorporating three intuitively reasonable explanatory variables.

# Australia Contextualization

*Excellent explanations/considerations are provided*

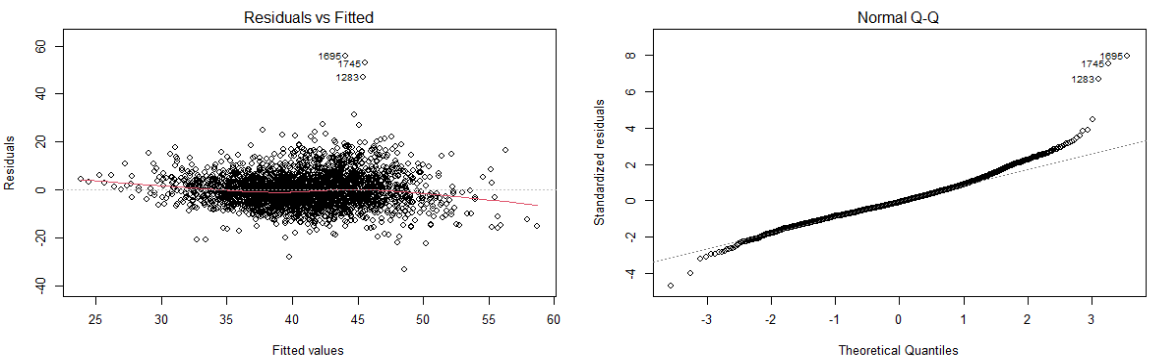
3 considerations:

Heavily skewed towards one State

# Appendix A – Technical Analysis

**Examination on Linearity**

In addition to the fitting summary, the residual plot is used to examine the suitability of the linear assumption for the regression model.



Based on the residuals vs. fitted plot, although the spread of the residuals seemed to be increasing towards the right on the horizontal axis, the residuals roughly scattered around 0 without any patterns.

Overall, linearity assumption appears to be reasonable for the regression model to predict the stroke death rate.

# Appendix B – Contextualisation Notes

**Business Understanding**

1. **Determine the Business Objectives**
   1. Background

Stroke has become the most fatal disease in Australia and hence raised the attention of Stroke Foundation (SF) Australia, who wishes to understand the associated risk factors to facilitate their work on stroke prevention, detection and support for all Australians.

* 1. Environment

The key concern relating to the death from stroke arises in Australia while the analysis is based on the US stroke mortality, incidence and other social determinants data in the year of 2015.

* 1. Reason of the Analysis

SF wishes to gain some initial insights into the risk factors of stroke death under the US environment, with the hope of facilitating its further exploration and investigation in Australia.

1. **Assess the situation**
   1. Available Resources

* US publicly available datasets on stroke mortality and incidence as well as US poverty, health insurance, income and population information on a national level for the 2015 calendar year. All the datasets are provided by Sam, who works at the American Stroke Association (ASA).
* Corresponding data dictionaries.
* The structure of a preliminary multiple linear regression model from Sam.
* Taylor, as a health economist working at SF, may be able to provide more domain knowledge in this field.
  1. Constraints
* Data is not available for some States (e.g. Nevada)
* Some datasets may be incomplete due to confidentiality reasons or scanty data in some counties.
* The feasibility of a similar analysis in Australia on a national level is unclear.
  1. Assumptions
* Expert opinion on the recent incidence trend is adopted where “stable” is assumed for counties with suppressed information.
  1. Uncertainties and Risks of the Project
* Whether the insights gained from the US analysis is relevant and instructive to Australian context is uncertain.
* The analysis is based on the data in year of 2015. Its relevance to the current circumstances may be weak.

1. **Determine Project Goals**

If meaningful learnings can be gained from this analysis, SF aims to showcase end-to-end analysis in stroke research to the public, enhancing the transparency of the work they do, which may in turn bring greater awareness of risk factors associated to deaths caused by strokes in Australia.

1. **Project Plan**

No specific project plan for this analysis based the information given.

**Stakeholder Analysis**

1. **Client**

Stroke Foundation Australia.

More specifically, Taylor, the health economist working at SF.

1. **Other Stakeholders and their Vested Interests**

* Sam, who works at the American Stroke Association (ASA) and wishes to compare this analysis to ASA’s analysis and see if the preliminary model can be improved.
* Manager / the Board at AA consulting firm, who have the interest of ensuring the overall quality of the analysis to maintain the reputation of the firm.
* Australian government, who may wish to understand the actions that could be taken to tackle the stroke problem across the nation.

# Appendix C – R Code

# Appendix D – Reference

1. https://www.wikiwand.com/en/County\_statistics\_of\_the\_United\_States [↑](#footnote-ref-1)
2. See appendix A. [↑](#footnote-ref-2)