Project

Jinfeng Zhu

Student ID: 47564644

Contents

[1. Introduction 2](#_Toc112968253)

[2. Exploratory Data Analysis 2](#_Toc112968254)

[2.1 Source of Data 2](#_Toc112968255)

[2.2 Data Exploration 2](#_Toc112968256)

[2.2.1 Checks on Duplicate and Data Type 2](#_Toc112968257)

[2.3 2.2 Checks on Internal Consistency 3](#_Toc112968258)

[3. Evaluation of the Preliminary Multiple Linear Regression Model 3](#_Toc112968259)

[4. Australia Contextualization 4](#_Toc112968260)

[Appendix A – Technical Analysis 5](#_Toc112968261)

[Appendix B – Contextualisation Notes 5](#_Toc112968262)

[Appendix C – R Code 5](#_Toc112968263)

[Appendix D – Reference 5](#_Toc112968264)

# 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 country for the 2015 calendar year.

## Data Exploration

### Checks on Duplicate and Data Type

Initial checks were performed on duplicated values to ensure they do not exist in any of the dataset. Investigation in missing value checks were intentionally ignored at this stage as they were known to exist due to the reasons of the insufficient data size or confidentiality. Missing value 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 the state information from the county information. For example, “Perry County, Kentucky” was separated into “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 to ensure the unique number of states is 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 checked. The number of stroke death and the stroke mortality rate was used to back solve the population for each county, which suggested the US national population in 2015 was 329 million. Therefore, 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

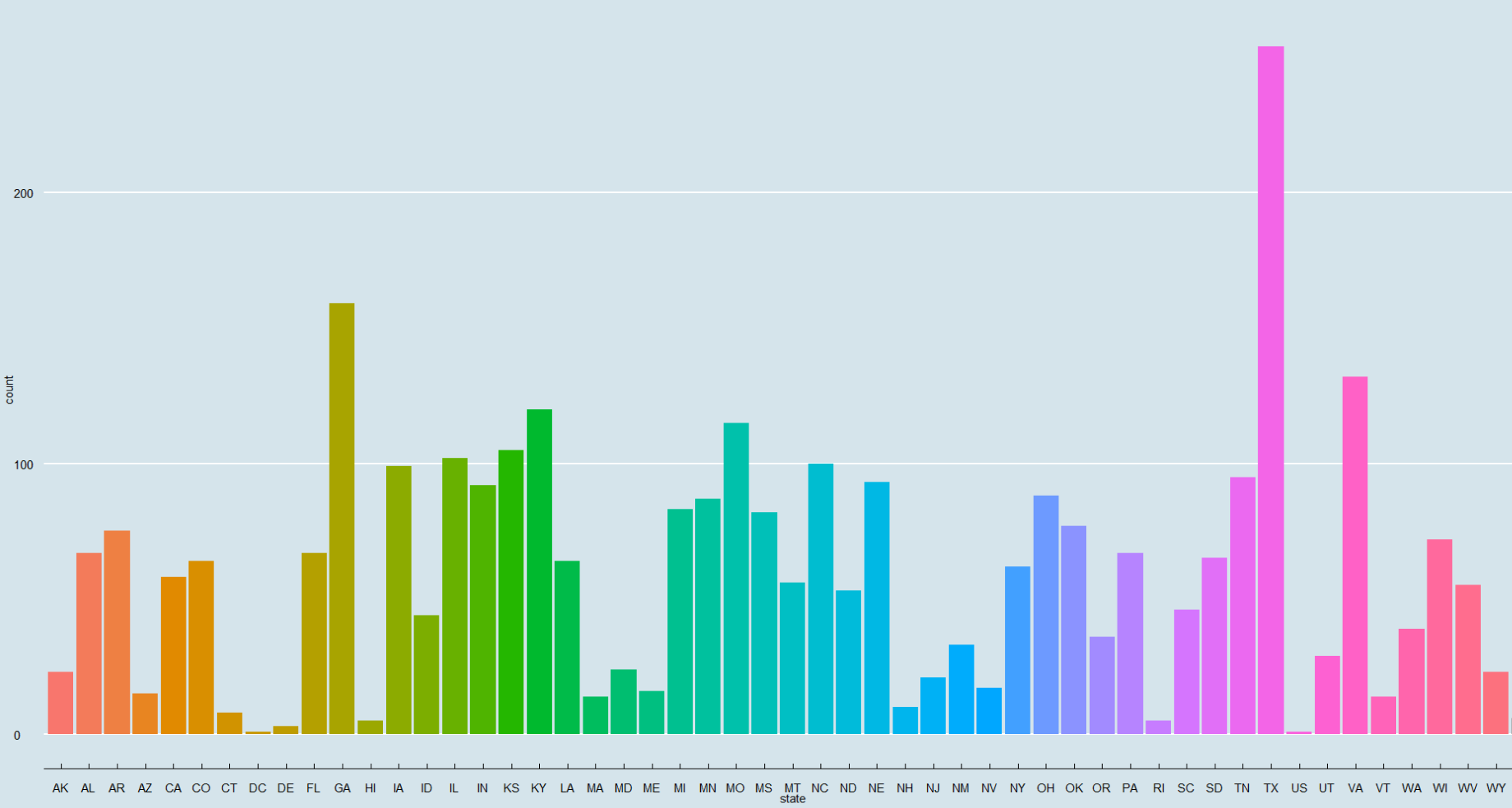
For the incidence dataset, expert opinion on the recent incidence trend was adopted where “stable” was assumed for counties whose data was suppressed. As a result, more than 90% the counties had a stable stroke incidence experience.

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

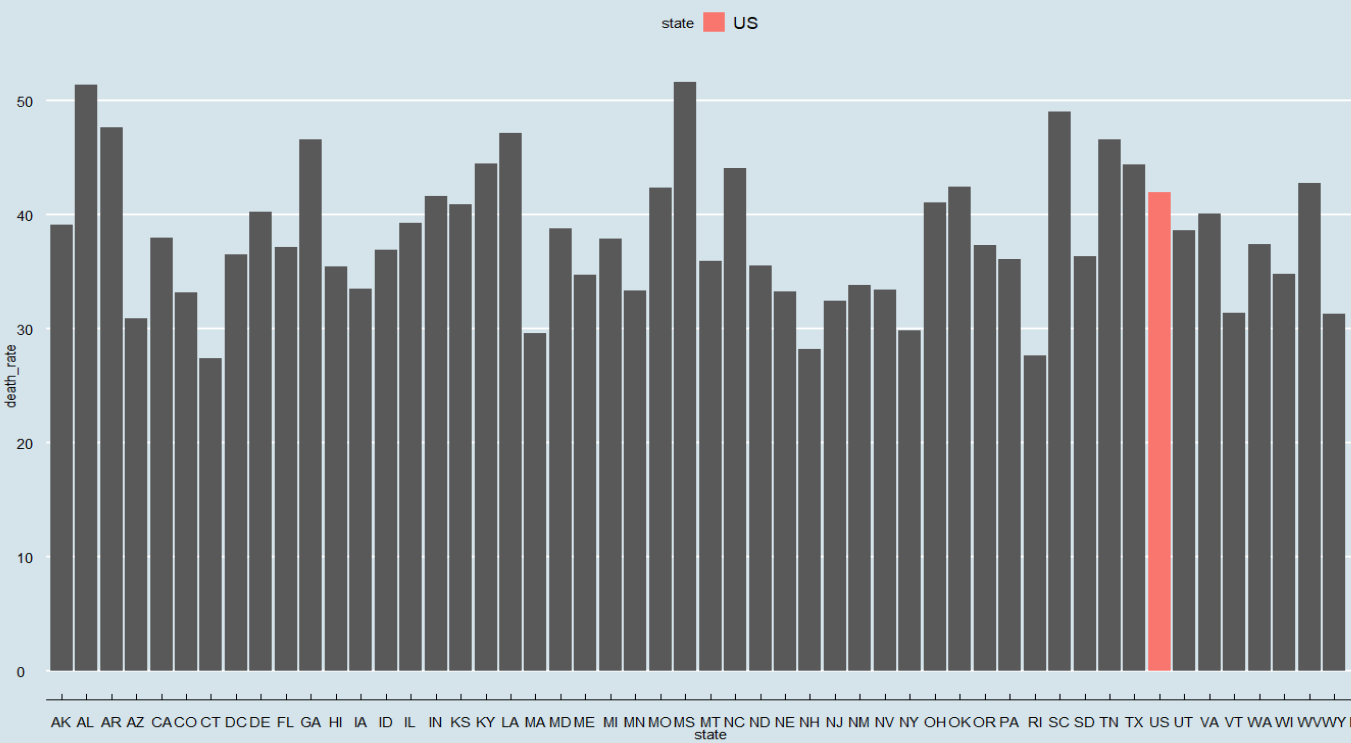
As “recent trend” is a categorical variable, dummy encoding is applied and two more numerical variables ("trend\_falling " and “trend\_rising”) taking a value of 1 or 0 were introduced.

## Visualisation

To view the completeness of the dataset, the number of counties for each state were shown in the bar plot below. All 50 states and DC were included in the datasets where majority of the states contributed around 100 counties while Texas (TX) stood out by having more than 200 valid entries.



Another bar plot compares the stroke death rates by state to the average of the nation (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. Variance



## 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 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 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 | Counties 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 dataset. |
| Incidence | Remove observations where both incidence rate and annual count were missing due to the confidentiality.  For the remaining missing values in the incidence rate, the associated counts were less than 5. Therefore, those missing values were imputed by substituting the NAs with 0. |

**Step 3 – Generate Explanatory Variables**

Some of the explanatory variables used in the preliminary multiple linear regression model were not readily available in the dataset yet.

# Evaluation of the Preliminary Multiple Linear Regression Model

*Excellent evaluation of the linear model*

Correlation



Assessment of ...

Evaluation of …

# Australia Contextualization

*Excellent explanations/considerations are provided*

3 considerations:

Heavily skewed towards one State

# Appendix A – Technical Analysis

# Appendix B – Contextualisation Notes

# Appendix C – R Code

# Appendix D – Reference