[Document title]

Jinfeng Zhu

Student ID: 47564644

Contents

[1. Overview 2](#_Toc111928567)

[2. Exploratory Data Analysis 2](#_Toc111928568)

[3. Evaluation of the Preliminary Multiple Linear Regression Model 2](#_Toc111928569)

[4. Australia Contextualization 2](#_Toc111928570)

[Appendix A – Technical Analysis 3](#_Toc111928571)

[Appendix B – Contextualisation Notes 3](#_Toc111928572)

[Appendix C – R Code 3](#_Toc111928573)

[Appendix D – Reference 3](#_Toc111928574)

# 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

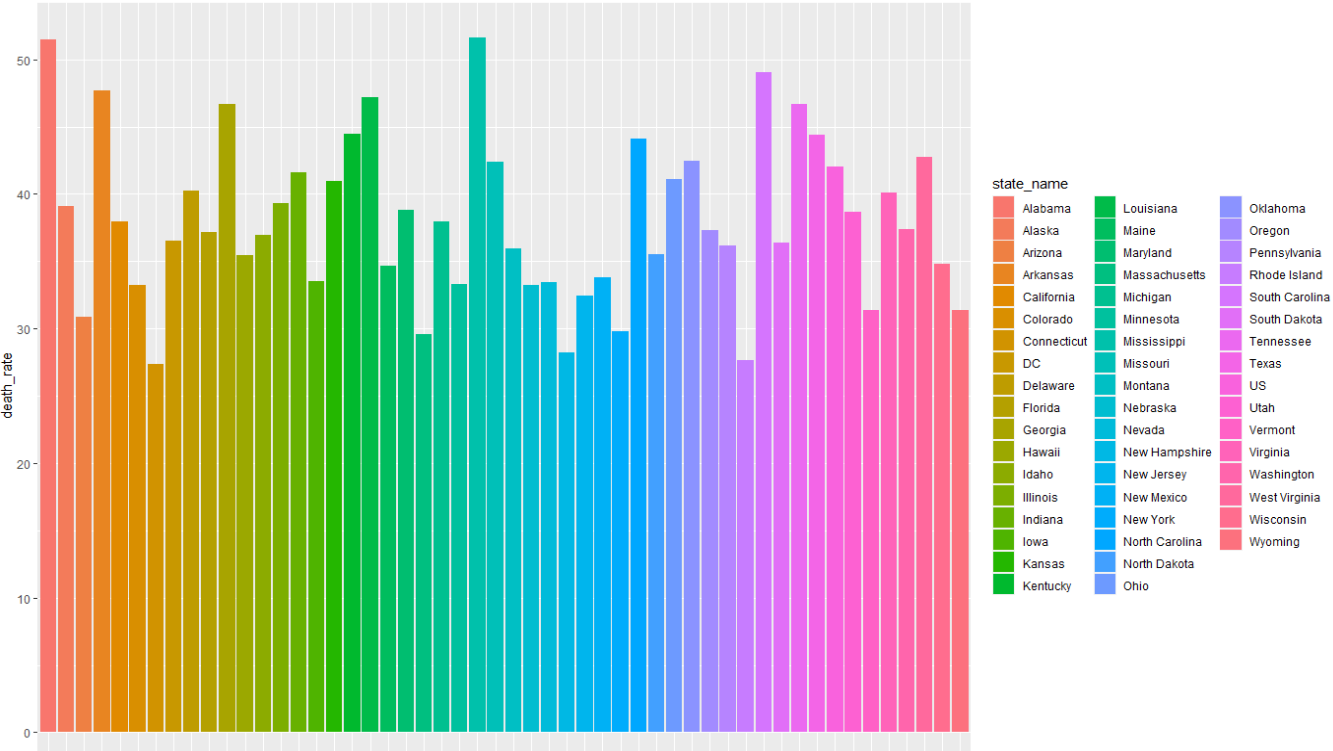
*Description of the exploratory data analysis was excellent.*

Exploratory data analysis (EDA) on the US datasets was conducted in R. The US datasets contain the information on stroke mortality and incidence rates as well as US poverty, health insurance, income and population information for the 2015 calendar year.

Initial checks were performed on duplicated values, which none of the datasets included. 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 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.

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). Besides, there is another entry for the entire country as a whole (US). Checks and corrections were performed to ensure the unique number of states is 52.



# Evaluation of the Preliminary Multiple Linear Regression Model

*Excellent evaluation of the linear model*

# Australia Contextualization

*Excellent explanations/considerations are provided*

# Appendix A – Technical Analysis

# Appendix B – Contextualisation Notes

# Appendix C – R Code

# Appendix D – Reference