Project

Jinfeng Zhu

Student ID: 47564644

1. 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.

2. Exploratory Data Analysis (EDA)

2.1 Source of Data

The US datasets, including stroke mortality and incidence, as well as poverty, income, health insurance and population data, were published by various US entities[1]. All these datasets contain the information for over 3,100 counties across the United States for the year of 2015.

2.2 Data Exploration

2.2.1 Checks on Duplicate and Data Type

Initial checks were performed on duplicated values to ensure their non-existence in any of the dataset.

Checks on data format were also performed for each dataset and types of some variables were converted. For example, “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.

Investigation in missing values was intentionally ignored at this stage as they were known to exist due to the scanty data or the reason of confidentiality. Missing values will be explored for the aggregated dataset in the later stage of EDA.

[1] US Centers for Disease Control and Prevention (cdc.gov) and Census (census.gov)

2.2.2 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 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.

2.2.3 Internal Checks

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

- 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.

2.2.4 Sense Checking

Sense checking was performed on the population dataset, where 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.

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. After further investigation, the population dataset seemed to include the subtotal population at state level. Those observations were removed so that only county level data remained and total population was 321 million.

2.3 Data Manipulation

For the incidence dataset, expert opinion on the recent incidence trend[2] 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

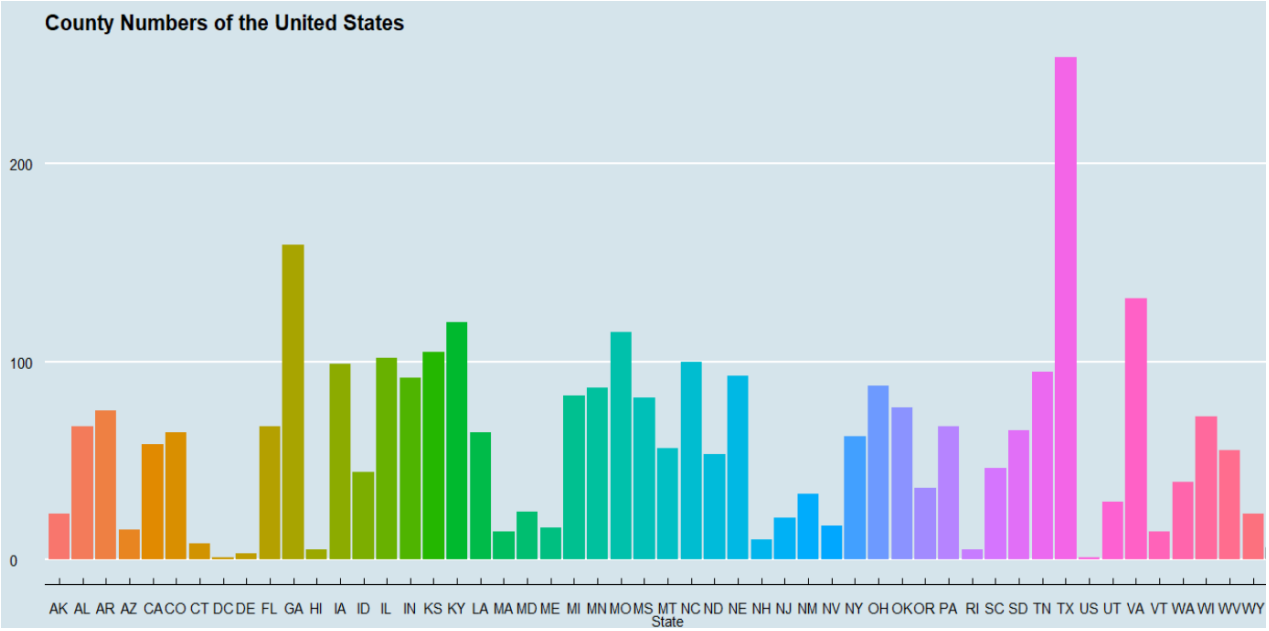
[2] See appendix A for data transformation on incidence trend.

6.4%	1.4%	92.3%	100%

Furthermore, incidence rate was defined as new hospitalizations from stroke for those aged 75 and above. However, no action was taken as the uncertainty herein was assessed not to have a material impact on this project, assuming the majority of the new stroke hospitalizations were above age 75.

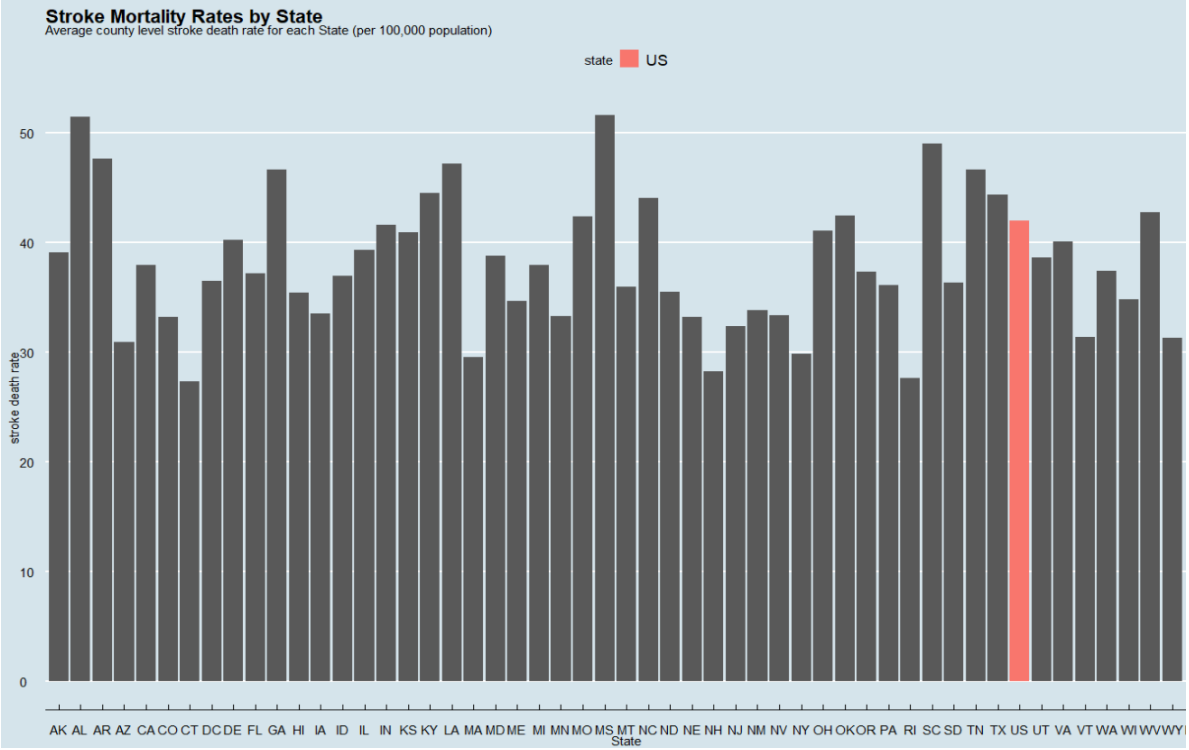
2.4 Visualisation

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[3], 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.

[3] https://www.wikiwand.com/en/County_statistics_of_the_United_States



2.5 Final Dataset

After initial exploration 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 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 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.
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 median value.

As a result, observations from 48 States remained in the dataset. 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 regression model were not readily available in the existing dataset. Hence another step was needed to generate those required variables in the final dataset.

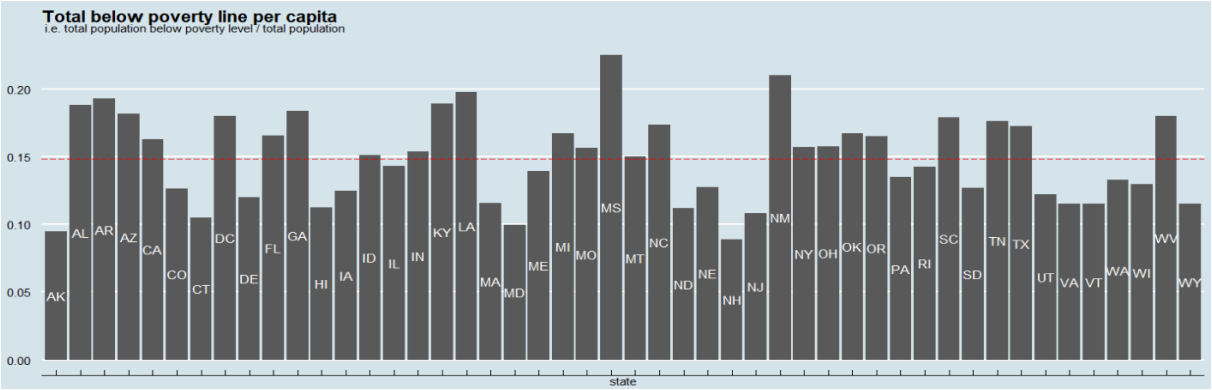
1) ***Total below poverty line per capita*** *= total population below poverty level / total population whom poverty status Is determined*
2) ***Total without health insurance per capita*** *= sum of the population without health insurance coverage for each age group / total population*

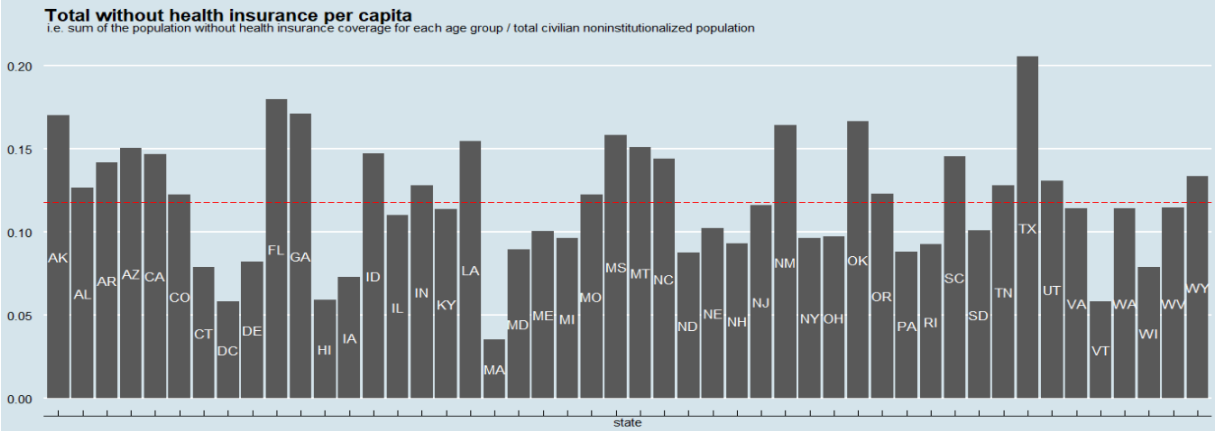
Step 4 – Reasonableness Checking

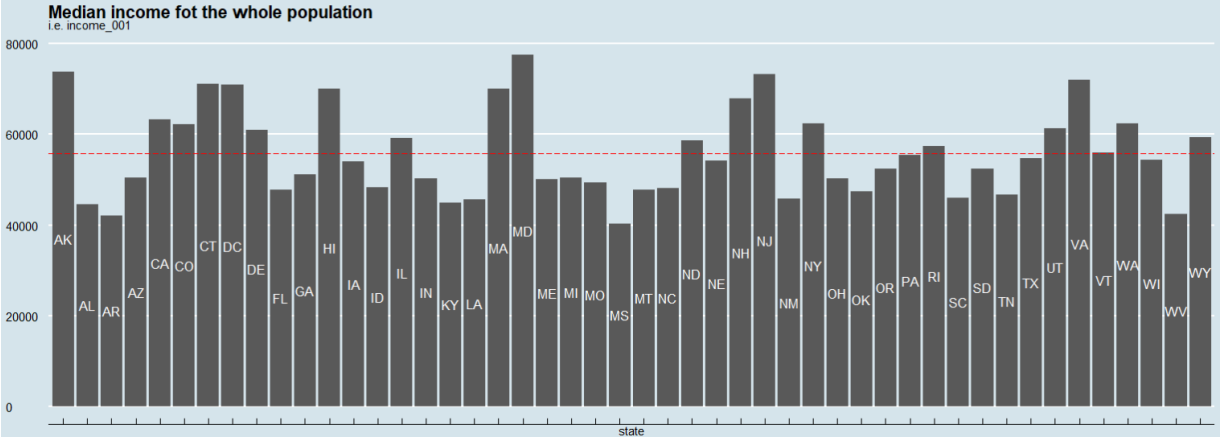
Before reaching the final dataset, some reasonableness checking was performed on the selected explanatory variables at State level. This was achieved by applying the data visualisation to those explanatory variables to identify any anomalies such as outliers or unusual trends.

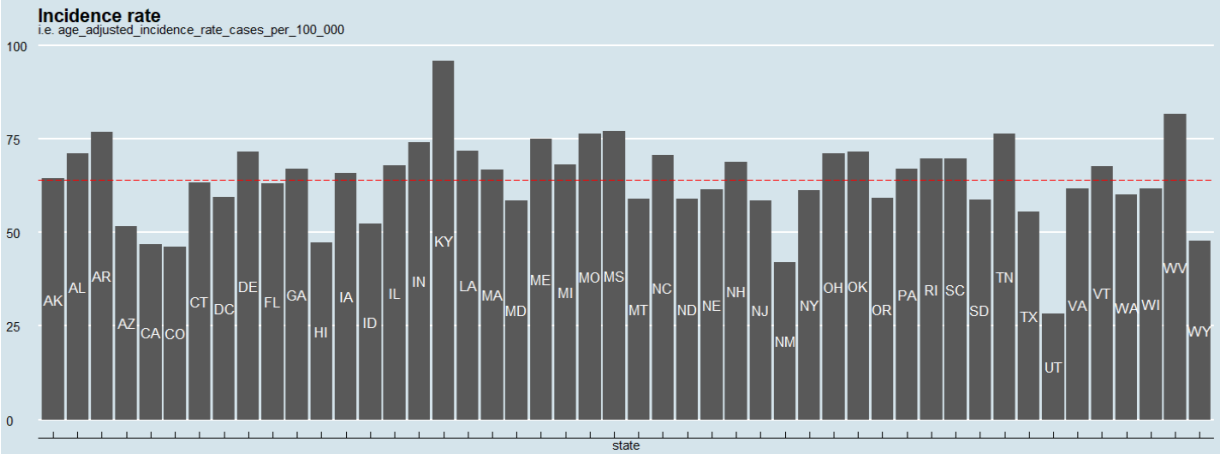
From the 4 plots below, no anomalies were noticed and the national average was as follows:

Explanatory Variable	National Average
Total below poverty line per capita	*15%*
Total without health insurance per capita	*12%*
Median income	*USD$56,000*
Incidence rate	*64 per 100,000 population*

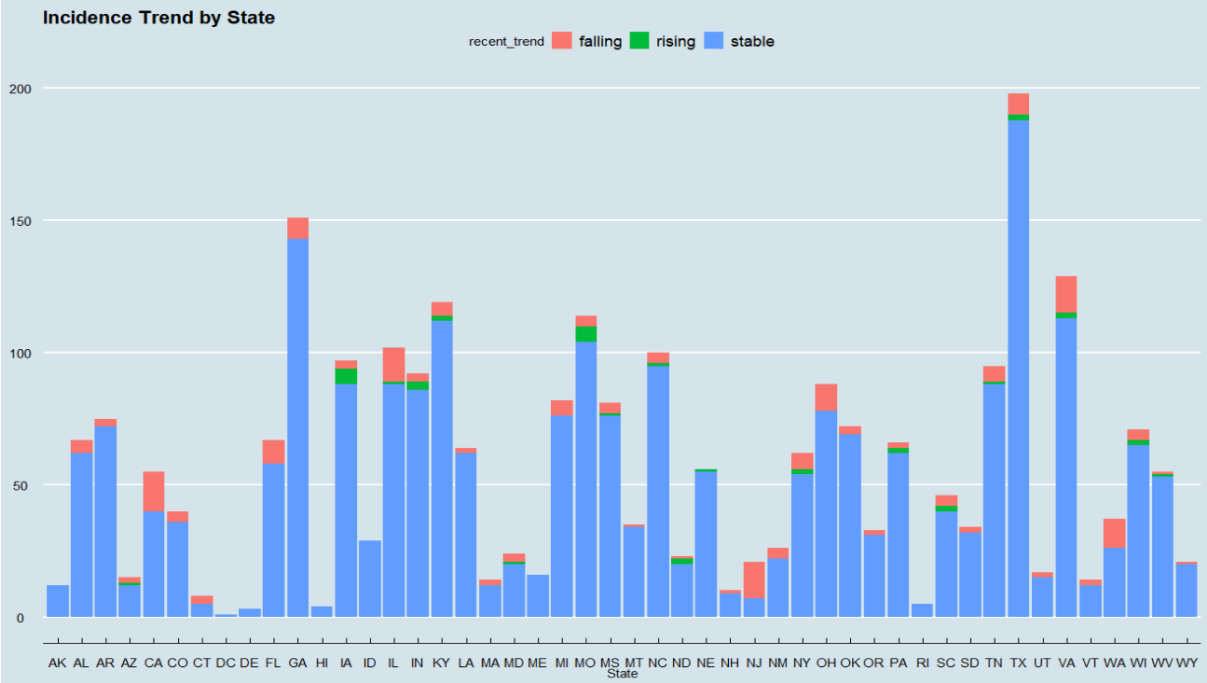

Total below poverty line per capita
i.e. total population below poverty level / total population
0.20
0.15
0.10
0.05
0.00
AK AL AR AZ CA CO CT DC DE FL GA HI IA ID IL IN KY LA MA MD ME MI MO MS MT NC ND NE NH NJ NM NY OH OK OR PA RI SC SD TN TX UT VA VT WA WI WV WY
state

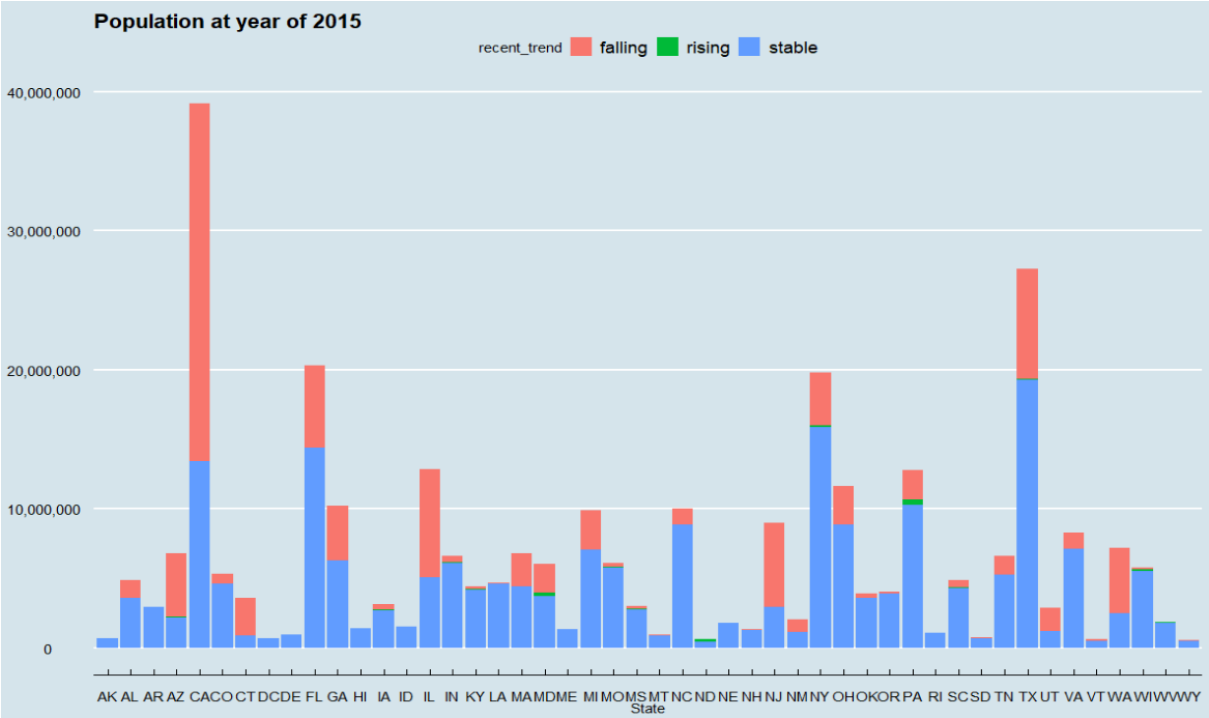

Total without health insurance per capita
i.e. sum of the population without health insurance coverage for each age group / total civilian noninstitutionalized population
0.20
0.15
0.10
0.05
0.00
AK AL AR AZ CA CO CT DC DE FL GA HI IA ID IL IN KY LA MA MD ME MI MO MS MT NC ND NE NH NJ NM NY OH OK OR PA RI SC SD TN TX UT VA VT WA WI WV WY
state


Median income fot the whole population
i.e. income_001
80000
60000
40000
20000
0
AK AL AR AZ CA CO CT DC DE FL GA HI IA ID IL IN KY LA MA MD ME MI MO MS MT NC ND NE NH NJ NM NY OH OK OR PA RI SC SD TN TX UT VA VT WA WI WV WY
state



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 people who live in large counties can easily access better medical treatment when compared to people from 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

3. Evaluation of 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.

3.1 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 level of income and insurance coverage. In addition, population is indicative of the county size and it may imply the level of urbanization and the quality of health care system to some extent, 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 "median income". However, response variable's relationship with incidence trend and population did not appear to be significant.

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

3.2 Preliminary Model Evaluation

3.2.1 Model Evaluation

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

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_d x_d + \epsilon$$

Where,

- Y is the response variable stroke death rate;
- β_i is the coefficient for explanatory variables;
- ϵ is the residual

The fitting summary is shown as below.

Fit 1	Coefficient	p-value
(Intercept)	29.55	0.0000
below_poverty	24.58	0.0000
no_health_ins	25.65	0.0000
median_incom	0.00	0.0000
incidence	0.12	0.0000
incidence_falling	-1.01	0.0652
incidence_rising	-1.42	0.2117
population	0.00	0.0616
R-squared	0.3049	
Adjusted R-squared	0.3031	

As per the statistical diagnosis (*p-value*) result above, the preliminary model contained insignificant covariates (*incidence_falling, incidence_rising, and population*). Additionally, R^2 and adjusted R^2 indicated around 30% of the variability in the response variable was explained by the preliminary model.

Other statistical diagnostic[4] 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. Nevertheless, variable transformation or other non-linear model forms could also be explored to improve the percent of variance explained.

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

3.2.2 Model Improvement

Apart from considering other non-linear model forms, one possible improvement that can be made on the preliminary model is to remove insignificant covariates in sequence.

Step 1 – Remove "incidence_rising"[5]

[4] See appendix A.
[5] Please refer to appendix for fitting result summary for each step of the model improvement.

"Incidence_rising" was removed first as it had the highest p-value in the previous fitting, which suggested a low significance.

Step 2 – Remove "incidence_falling"

"Incidence_falling" was removed to exclude all the incidence trend related impact from the fitting. After that, "population" was significant at 5% level but not at 1% level with a nearly zero coefficient. Therefore, "population" is considered to be removed as the next step.

Step 3 – Remove "population"

"Population" was removed. All the remaining covariates appear to be significant.

Step 4 – Remove "median_income"

Although "median_income" was extremely significant in the previous fitting, its coefficient was almost zero, which means it did not contribute much to the variability of stroke mortality. This may be explained by the collinearity between "median_income" and "below_poverty". From the correlation plot above, these 2 features were strongly negatively correlated with each other, which suggests they likely represent the same information. As a result, "median_income" has been removed while R^2 slightly decreased by 1% when compared to the preliminary model.

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 for each county, although a mild correlation has been detected among those 3 covariates based on the correlation plot.

In conclusion, the preliminary model can be improved by reducing the unnecessary model complexity and only incorporating significant explanatory variables.

4. Australian Contextualization

In Australia, a visualization dashboard for stroke hospitalization and death is publicly available[6] on the website of Australian Institute of Health and Welfare (AIHW). Those statistics can be viewed in segregated age or sex groups, as well as population group including indigenous/non-indigenous, remoteness area and socioeconomic group.

[6] https://www.aihw.gov.au/reports/heart-stroke-vascular-diseases/hsvd-facts/contents/data-visualisations

However, to gain insights into the stroke death rate, postcode level observations (similar to the US county level) would be ideal for the regression analysis. Remoteness only classifies geographical locations into 5 groups, which is significantly less granular than the US data where geographical area is classified into 3,000 counties. Moreover, Australian Bureau of Statistics (ABS) does not provide more granular stroke data either.

In conclusion, this insufficiency in the publicly available data on stroke death events makes both of the regression analysis and "social determinants of health maps" infeasible in Australia. Other resources may need to be explored.

Three Considerations in Australian Context

1) Population Distribution

The population distribution in Australia may be quite different from the US. Majority of the population reside in New South Wales, Victoria and Queensland. Any potential variations and patterns for smaller geographical states may not be identified due to the scarce observations.

This concern may also apply to aboriginal and Torres Strait Islander people, since limited national information on the occurrence of stroke is available for the Indigenous population.

2) Feature Selection

Features representing similar information should not be considered in the analysis. For example, to prevent the collinearity, socioeconomic group can be selected alone to indicate socioeconomic aspect without including other similar features such as income.

Besides, important features under Australian context may be different from what is included in the US analysis. Domain knowledge and statistical diagnosis should be relied on to ensure the soundness of feature selection.

3) Health Insurance

In Australian context, health insurance coverage may not be a good socioeconomic variable as all Australians are covered by the Medicare health system, therefore people are less incentive to purchase additional commercial health insurance.

In this regard, Medicare system can be a robust medical data source. SF could explore cooperation with Medicare to obtain more stroke related information.

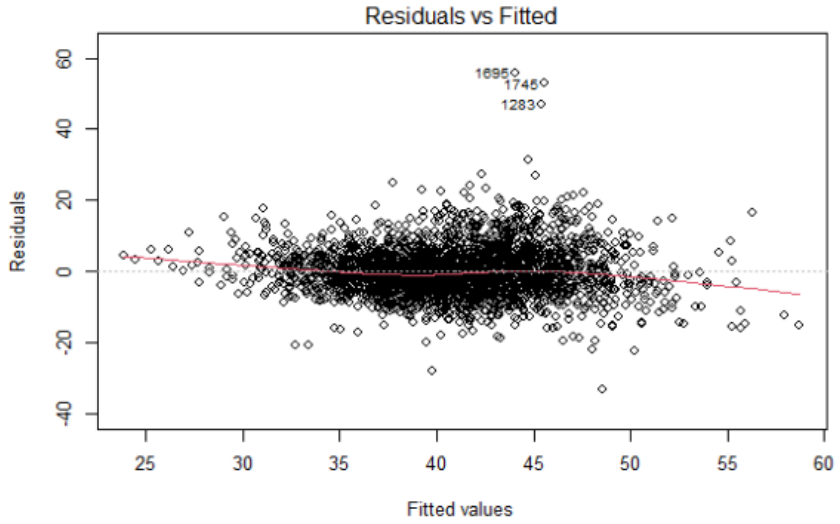
Appendix A – Technical Analysis

1. Data Transformation on Categorical Variable

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.

2. 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, most of the residuals in the middle roughly scattered around 0 without any patterns.

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

3. Fitting Result Summary for Model Improvement

Step 1 – Remove “incidence_rising”

Fit 2 (Step 1)	Coefficient	p-value
(Intercept)	29.49	0.0000
below_poverty	24.49	0.0000
no_health_ins	25.85	0.0000
median_income	0.00	0.0000
incidence	0.12	0.0000
incidence_falling	-0.99	0.0704
population	0.00	0.0639

R-squared	0.3045
Adjusted R-squared	0.3029

Step 2 – Remove “incidence_falling”

Fit 3 (Step 2)	Coefficient	p-value
(Intercept)	29.60	0.0000
below_poverty	24.26	0.0000
no_health_ins	25.88	0.0000
median_income	0.00	0.0000
incidence	0.12	0.0000
population	0.00	0.0163
R-squared	0.3037	
Adjusted R-squared	0.3023	

Step 3 – Remove “population”

Fit 4 (Step 3)	Coefficient	p-value
(Intercept)	30.31	0.0000
below_poverty	22.63	0.0000
no_health_ins	25.66	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”

Fit 5 (Step 4)	Coefficient	p-value
(Intercept)	21.41	0.0000
below_poverty	38.05	0.0000
no_health_ins	27.25	0.0000
incidence	0.13	0.0000
R-squared	0.292	
Adjusted R-squared	0.2912	

Appendix B – Contextualisation Notes

Business Understanding

1. Determine the Business Objectives

1.1. 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.

1.2. 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.

1.3. 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.

2. Assess the situation

2.1. 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.

2.2. 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.

2.3. Assumptions

- Expert opinion on the recent incidence trend is adopted where “stable” is assumed for counties with suppressed information.

2.4. 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.
- Stroke incidence rate was only based on the age 75 and above.

3. **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.

4. **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.

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

   - Sam, who works at the American Stroke Association (ASA) and wishes to compare this analysis to ASA's one 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

````
---
title: "DAP_Project"
author: "Jinfeng Zhu"
date: "2022-08-15"
output: html_document
editor_options:
---
```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
# Installing libraries
# install.packages("simputation")
library(tidyverse)
library(readr)
library(dplyr)
library(tidyr)
library(ggplot2)
library(ggthemes)
library("corrplot")
library(janitor) # to clean column names
library(psych)   # for the describe function
library("stringr") # to add leading zeros
library(naniar) # to check missing values
library(simputation) # to replaces NAs
```
# 1. Exploratory Data Analysis
## 1.1 Import Data
```{r import data, include=FALSE}
data_death      <- read_csv("Data/death.csv")
data_health_ins <- read_csv("Data/healthinsurance.csv")
data_incidence  <- read_csv("Data/incidence.csv")
data_income     <- read_csv("Data/income.csv")
data_population <- read_csv("Data/population.csv")
data_poverty    <- read_csv("Data/poverty.csv")
```
## 1.2 Clean variable names
```{r naming convention of header}
data_death_clean      <- clean_names(data_death)
data_health_ins_clean <- clean_names(data_health_ins)
data_incidence_clean  <- clean_names(data_incidence)
data_income_clean     <- clean_names(data_income)
data_population_clean <- clean_names(data_population)
data_poverty_clean    <- clean_names(data_poverty)

names(data_death_clean)
names(data_health_ins_clean)
names(data_incidence_clean)
names(data_income_clean)
names(data_population_clean)
````
````

```
names(data_poverty_clean)
```

1.3 Data Exploration

```{r view data}
# Overall checks
head(data_death_clean) # "age_adjusted_death_rate" and “average_deaths_per_year” are character
head(data_health_ins_clean)
head(data_incidence_clean) # "age_adjusted_incidence_rate_cases_per_100_000" and “average_annual_count” are character
head(data_income_clean)
head(data_population_clean)
head(data_poverty_clean)

summary(data_death_clean)
summary(data_health_ins_clean)
summary(data_incidence_clean)
summary(data_income_clean)
summary(data_population_clean)
summary(data_poverty_clean)
```

1.3.1 Check duplicates

```{r check duplicates}
# Check duplicates
data_death_clean %>% mutate(dup = duplicated(data_death_clean)) %>% filter(dup == TRUE)
data_health_ins_clean %>% mutate(dup = duplicated(data_health_ins_clean)) %>% filter(dup == TRUE)
data_incidence_clean %>% mutate(dup = duplicated(data_incidence_clean)) %>% filter(dup == TRUE)
data_income_clean %>% mutate(dup = duplicated(data_income_clean)) %>% filter(dup == TRUE)
data_population_clean %>% mutate(dup = duplicated(data_population_clean)) %>% filter(dup == TRUE)
data_poverty_clean %>% mutate(dup = duplicated(data_poverty_clean)) %>% filter(dup == TRUE)
  # No duplicates for all datasets
```

1.3.2 Convert Data Type

```{r data type conversion}
# [Death] Convert the data type from character to numeric
data_death_clean2 <- data_death_clean %>%
  mutate(age_adjusted_death_rate=as.double(age_adjusted_death_rate)) %>%
  mutate(average_deaths_per_year=as.double(average_deaths_per_year))

miss_var_summary(data_death_clean)
miss_var_summary(data_death_clean2)
  # 331 suppressed values "*" are automatically converted to NA by coercion
```

```
# [Incidence] Convert the data type from character to numeric
data_incidence_clean2 <- data_incidence_clean %>%

mutate(age_adjusted_incidence_rate_cases_per_100_000=as.double(age_adjusted_inc
idence_rate_cases_per_100_000)) %>%
  mutate(average_annual_count=as.double(average_annual_count))

miss_var_summary(data_incidence_clean)
miss_var_summary(data_incidence_clean2)
  # 442 suppressed values "*" are automatically converted to NA by coercion

# [Income] Convert the data type from character to numeric
summary(data_income_clean)

data_income_clean$income_b_001 <- as.double(data_income_clean$income_b_001)
data_income_clean$income_c_001 <- as.double(data_income_clean$income_c_001)
data_income_clean$income_d_001 <- as.double(data_income_clean$income_d_001)
data_income_clean$income_e_001 <- as.double(data_income_clean$income_e_001)
data_income_clean$income_f_001 <- as.double(data_income_clean$income_f_001)
data_income_clean$income_g_001 <- as.double(data_income_clean$income_g_001)
data_income_clean$income_h_001 <- as.double(data_income_clean$income_h_001)
data_income_clean$income_i_001 <- as.double(data_income_clean$income_i_001)

data_income_clean
```

1.3.3 Reshaping Data into Tidy Form

```
```{r reshaping}
# Separate state names from county
data_death_state <- data_death_clean2 %>% separate(county,
c("county_name","state_name"), ", ")
  # Missing pieces are autimatically filled with `NA` in 2 rows, needs to check

data_death_state %>% filter(is.na(county_name)) # no "NA"
data_death_state %>% filter(is.na(state_name))  # one is United States, the other is
District of Columbia (State) (DC)

data_death_state1 <- data_death_state %>% mutate(state_name  = ifelse(county_name
== "District of Columbia (State)", "DC", state_name))

data_death_state2 <- data_death_state1 %>% mutate(state_name  =
ifelse(county_name == "United States", "US", state_name))

data_death_state2 %>% filter(is.na(state_name)) # no more missing values in
state_name

# Check the number of the distinct states
n_distinct(data_death_state2$state_name) # Unique number of states is 54. Based on
the domain knowledge, the United States is made up of a total of 50 states, plus the
District of Columbia.
```
```

```
# Check the list of states
unique(data_death_state2$state_name) # "Arizona<sup>3</sup>" and
"Alaska<sup>3</sup>" appear to be errors

data_death_state3 <- data_death_state2 %>% mutate(state_name = ifelse(state_name
== "Alaska<sup>3</sup>", "Alaska", state_name))

data_death_state4 <- data_death_state3 %>% mutate(state_name = ifelse(state_name
== "Arizona<sup>3</sup>", "Arizona", state_name))

# Check the number of the distinct states again
n_distinct(data_death_state4$state_name) # 52 = 50 states + 1 federal district + US as
a whole
unique(data_death_state4$state_name) # looks alright
```

1.3.4 Internal Checks

```{r Nevada}
data_death_state4 %>% filter(state_name == "Nevada")
  # There are still 12 valid Nevada entries.
  # [Uncertainty] According to Sam, Nevada data is not available. But some Nevada data
are actually valid in the death datasets. Include them for further analysis at the moment,
but will ask Sam for clarification.

data_incidence_clean2 %>% filter(grepl('Nevada', county)) # Nevada data is not
available
```

```{r health insurance}
# Check if Non-institutionalized Population (hi_001) = Non-institutionalized
Population_male (hi_002) + Non-institutionalized Population female (hi_030)
data_health_ins_clean %>%
  mutate(check = hi_001 - hi_002 - hi_030) %>%
  group_by(state) %>%
  summarise(sum(check))
  # all zero, passed the check
```

```{r poverty}
# Check if:
# 1) below poverty level population (poverty_002) = below poverty level male
(poverty_003) + below poverty level female (poverty_017)
data_poverty_clean %>%
  mutate(check = poverty_002 - poverty_003 - poverty_017) %>%
  group_by(state) %>%
  summarise(sum(check))
  # all zero, passed the check

# 2) above poverty level population (poverty_031) = above poverty level male
(poverty_032) + above poverty level female (poverty_046)
data_poverty_clean %>%
  mutate(check = poverty_031 - poverty_032 - poverty_046) %>%
  group_by(state) %>%
```

```
  summarise(sum(check))
  # all zero, passed the check

# 3) Population For Whom Poverty Status Is Determined (poverty_001) = below poverty level population (poverty_002) + above poverty level population (poverty_031)
data_poverty_clean %>%
  mutate(check = poverty_001 - poverty_002 - poverty_031) %>%
  group_by(state) %>%
  summarise(sum(check))
  # all zero, passed the check
```

```{r population}
# Check the total population at mid-year 2015 on a national level.
sum(data_population_clean$popestimate2015)/10**6 # 642 million

# [Uncertainty] Based on the domain knowledge, the population of the United States should be around 320 million.
# This can be verified by the death dataset
data_death_state4 %>%
  mutate(population = average_deaths_per_year /
(age_adjusted_death_rate/100000)) %>%
  filter(state_name == "US") %>%
  pull(population)/10**6
  # 329 million

# Look into datasets: this was caused by the subtotal amount of each group
# Remove these subtotals
data_population_clean <- data_population_clean %>% filter(county != "000")
sum(data_population_clean$popestimate2015)/10**6 # 321 million
```

1.4 Manipulate and Cleanse the Data

1.4.1 Clean FIPS

```{r FIPS}
# Make FIPS a 5-digit code
# For death and incidence data, most of the FIPS codes are 5-digits but if they are 4-digits, you will need to add a 0 in front.
data_death_clean_fips <- data_death_state4 %>%
  mutate(fips_clean = str_pad(data_death_state4$fips, width = 5, pad = "0"))

data_incidence_clean_fips <- data_incidence_clean2 %>%
  mutate(fips_clean = str_pad(data_incidence_clean2$fips, width = 5, pad = "0"))

# For poverty, health insurance, income and population data, you will need to combine State FIPS and County FIPS to get the 5-digit code.
data_health_ins_clean_fips <- data_health_ins_clean %>%
  mutate(state_fips_clean  = str_pad(data_health_ins_clean$state_fips, width = 2, pad = "0")) %>%
  mutate(county_fips_clean = str_pad(data_health_ins_clean$county_fips, width = 3, pad = "0")) %>%
  mutate(fips_clean = paste(state_fips_clean, county_fips_clean, sep=""))
```

```
data_income_clean_fips <- data_income_clean %>%
  mutate(state_fips_clean  = str_pad(data_income_clean$state_fips, width = 2, pad =
"0")) %>%
  mutate(county_fips_clean = str_pad(data_income_clean$county_fips, width = 3, pad =
"0")) %>%
  mutate(fips_clean = paste(state_fips_clean, county_fips_clean, sep=""))

data_population_clean_fips <- data_population_clean_half %>%
  mutate(state_fips_clean  = str_pad(data_population_clean_half$state, width = 2, pad =
"0")) %>%
  mutate(county_fips_clean = str_pad(data_population_clean_half$county, width = 3,
pad = "0")) %>%
  mutate(fips_clean = paste(state_fips_clean, county_fips_clean, sep=""))

data_poverty_clean_fips <- data_poverty_clean %>%
  mutate(state_fips_clean  = str_pad(data_poverty_clean$state_fips, width = 2, pad =
"0")) %>%
  mutate(county_fips_clean = str_pad(data_poverty_clean$county_fips, width = 3, pad =
"0")) %>%
  mutate(fips_clean = paste(state_fips_clean, county_fips_clean, sep=""))
```

1.4.2 Check Duplicate FIPS

```{r check FIPS}
dim(data_death_clean_fips)[1] == length(unique(data_death_clean_fips$fips_clean)) #
No duplicates
dim(data_incidence_clean_fips)[1] ==
length(unique(data_incidence_clean_fips$fips_clean)) # No duplicates
dim(data_health_ins_clean_fips)[1] ==
length(unique(data_health_ins_clean_fips$fips_clean)) # No duplicates
dim(data_income_clean_fips)[1] == length(unique(data_income_clean_fips$fips_clean))
# No duplicates
dim(data_population_clean_fips)[1] ==
length(unique(data_population_clean_fips$fips_clean)) # No duplicates
dim(data_poverty_clean_fips)[1] == length(unique(data_poverty_clean_fips$fips_clean))
# No duplicates
```

1.4.3 Expert Opinion

```{r expert opinion on incidence trend}
# For Incidence dataset, expert opinion suggests suppressed cells of * for Recent Trend
is likely to be stable.
table(data_incidence_clean_fips$recent_trend) # observe 422 *

data_incidence_clean_fips_expert <- data_incidence_clean_fips %>%
mutate(recent_trend, recent_trend = ifelse(recent_trend == "*", "stable", recent_trend))

table(data_incidence_clean_fips_expert$recent_trend) # 422 more "stable" entries

# Apply Dummy Encoding for categorical variable "recent_trend"
data_incidence_clean_fips_expert_encoding <- data_incidence_clean_fips_expert %>%
```

```
  mutate(trend_falling = ifelse(recent_trend == "falling", 1, 0)) %>%
  mutate(trend_rising  = ifelse(recent_trend == "rising", 1, 0))

table(data_incidence_clean_fips_expert_encoding$trend_falling)
table(data_incidence_clean_fips_expert_encoding$trend_rising)
```

1.5 Visualise and Analyse Patterns in Data

1.5.1 Create State Abbreviations Column in Death Dataset

```{r add state abbr.}
data_death_clean_fips_with_state <- left_join(data_death_clean_fips,
data_health_ins_clean_fips %>% select(state, fips_clean), by="fips_clean")

# Assign "US" to for the national level
data_death_clean_fips_with_state <- data_death_clean_fips_with_state %>%
mutate(state = ifelse(state_name == "US", "US", state))

table(data_death_clean_fips_with_state$state_name)
n_distinct(data_death_clean_fips_with_state$state_name)

table(data_death_clean_fips_with_state$state)
n_distinct(data_death_clean_fips_with_state$state, na.rm = T)
```

1.5.2 Visualisation

```{r Visualisation: number of state}
# Check number of entries for each state
data_death_clean_fips_with_state %>% ggplot(aes(state, fill = state)) +
  geom_bar(show.legend = F) +
  theme_economist() +
  labs(x = "State", title = "County Numbers of the United States") +
  theme(axis.title.y=element_blank())
```

```{r visualisation: stroke mortality rates}
# Check the average stroke mortality rates by state and see where the average of the nation (US) sits among other states
summary_death_data_all_state <- data_death_clean_fips_with_state %>%
group_by(state) %>% summarize(death_rate = mean(age_adjusted_death_rate,
na.rm=TRUE))

summary_death_data_us <- data_death_clean_fips_with_state %>% filter(state ==
"US")

ggplot() +
  geom_col(data = summary_death_data_all_state, aes(state, death_rate)) +
  geom_col(data = summary_death_data_us, aes(state, age_adjusted_death_rate, fill =
state)) +
  theme(axis.title.x=element_blank(),
        # axis.text.x=element_blank(),
        axis.ticks.x=element_blank()) +
  theme_economist() +
  labs(x = "State", y = "stroke death rate", title = "Stroke Mortality Rates by State",
```

```
      subtitle = "Average county level stroke death rate for each State (per 100,000
population)")
```

1.6 Create Final Dataset

1.6.1 Join Datasets

```{r join data step 1}
# Step 1: death + incidence
dim(data_death_clean_fips_with_state)
dim(data_incidence_clean_fips_expert_encoding)
names(data_incidence_clean_fips_expert_encoding)

# Remove columns that are already contained in the left-hand data
data_incidence_join <- data_incidence_clean_fips_expert_encoding %>%
  select(c(3:8))

# Left Join - death / incidence
joined_data1 <- left_join(data_death_clean_fips_with_state, data_incidence_join,
by="fips_clean")
joined_data1
names(joined_data1)
```

```{r join data step 2}
# Step 2: + health
dim(data_health_ins_clean_fips)
names(data_health_ins_clean_fips)

# Remove columns that are already contained in the left-hand data
data_health_ins_join <- data_health_ins_clean_fips %>%
  select(c(5:61,64))

# Left Join - death / incidence / health
joined_data2 <- left_join(joined_data1, data_health_ins_join, by="fips_clean")
joined_data2
names(joined_data2)
```

```{r join data step 3}
# Step 3: + income
dim(data_income_clean_fips)
names(data_income_clean_fips)

# Remove columns that are already contained in the left-hand data
data_income_join <- data_income_clean_fips %>%
  select(c(5:14,17))

# Left Join - death / incidence / health / income
joined_data3 <- left_join(joined_data2, data_income_join, by="fips_clean")
joined_data3
names(joined_data3)
```

```{r join data step 4}
```

```
# Step 4: + population
dim(data_population_clean_fips)
names(data_population_clean_fips)

# Remove columns that are already contained in the left-hand data
data_population_join <- data_population_clean_fips %>%
  select(c(5,8))

# Left Join - death / incidence / health / income / population
joined_data4 <- left_join(joined_data3, data_population_join, by="fips_clean")
joined_data4
names(joined_data4)
```

```{r join data step 5}
# Step 5: + poverty
dim(data_poverty_clean_fips)
names(data_poverty_clean_fips)

# Remove columns that are already contained in the left-hand data
data_poverty_join <- data_poverty_clean_fips %>%
  select(c(5:63,66))

# Left Join - death / incidence / health / income / population / poverty
joined_data_all <- left_join(joined_data4, data_poverty_join, by="fips_clean")

glimpse(joined_data_all)
```

1.6.2 Check Missing Values

```{r missing values}
# To ensure all the suppressed values "*" are shown as "NA"
joined_data_all_with_NA <- joined_data_all %>% mutate(across(everything(),
function(x){ifelse(x == "*" | x == ".", NA, x)}))

# Overview on missing values
miss_var_summary(joined_data_all_with_NA)

# Response variable age_adjusted_death_rate: 331 rows have missing values.
# Remove those rows.
joined_data_all_y <- joined_data_all_with_NA %>%
filter(!is.na(age_adjusted_death_rate))
miss_var_summary(joined_data_all_y)

# Remove "US" entry as none of the social determinants datasets contain input for "US"
joined_data_all_y_no_us <- joined_data_all_y %>% filter(state != "US")
miss_var_summary(joined_data_all_y_no_us)

# Since only income_001 is one of the explanatory variables, missing values in other
income variables can be ignored.
# Hence those variables are entirely removed.
drop_cols <- c(names(data_income_clean)[6:14])
```

```
joined_data_all_y_no_us_income <- joined_data_all_y_no_us %>% select(-
one_of(drop_cols))
miss_var_summary(joined_data_all_y_no_us_income)

# Remove Nevada Data
joined_data_all_y_no_us_income_no_nevada <-
joined_data_all_y_no_us_income %>% filter(state_name != "Nevada")
joined_data_all_y_no_us_income %>% filter(state_name == "Nevada")

miss_var_summary(joined_data_all_y_no_us_income)

# Check on incidence missing values
joined_data_all_y_no_us_income_no_nevada %>%
filter(is.na(age_adjusted_incidence_rate_cases_per_100_000))
  # for these rows, "average_annual_count" is also missing or less than 5

# Check on incidence missing values where both "age_adjusted_incidence_rate" and
"average_annual_count" are missing
joined_data_all_y_no_us_income_no_nevada %>% group_by(state) %>%
  filter(is.na(age_adjusted_incidence_rate_cases_per_100_000) &
is.na(average_annual_count)) %>%
  summarise(sum(age_adjusted_incidence_rate_cases_per_100_000)) # Zero for KS
and MN, needs to check

joined_data_all_y_no_us_income_no_nevada %>% filter(state %in% c("KS", "MN")) #
Incidence data is missing for KS and MN

# Remove observations where both "age_adjusted_incidence_rate" and
"average_annual_count" are missing (i.e. KS and MN observations)
joined_data_all_y_no_us_income_no_nevada_incidence <-
joined_data_all_y_no_us_income_no_nevada %>% filter(!is.na(average_annual_count))

miss_var_summary(joined_data_all_y_no_us_income_no_nevada_incidence)

# Replace remaining 5 NAs in incidence rate with median value
index <-
which(is.na(joined_data_all_y_no_us_income_no_nevada_incidence$age_adjusted_inci
dence_rate_cases_per_100_000))
joined_data_all_y_no_us_income_no_nevada_incidence$age_adjusted_incidence_rate
_cases_per_100_000[index]

joined_data_all_final <- joined_data_all_y_no_us_income_no_nevada_incidence %>%
impute_proxy(age_adjusted_incidence_rate_cases_per_100_000~median(age_adjusted
_incidence_rate_cases_per_100_000, na.rm=TRUE))

joined_data_all_final$age_adjusted_incidence_rate_cases_per_100_000[index]
  # The median was 70.1 per 100,000 population
```

```
miss_var_summary(joined_data_all_final) # no more NAs

n_distinct(joined_data_all_final$state_name) # only 48 States remain in the dataset.
# 51 - Nevada - Kansas - Minnesota = 48
```

1.6.3 Generate Explanatory Variables for the Preliminary Multiple Linear Regression Model

```{r total_below_poverty_level_per_capita}
# Total below poverty level per capita
joined_data_all_final <- joined_data_all_final %>%
mutate(total_below_poverty_level_per_capita = poverty_002 / poverty_001)

# Check min and max value
describe(joined_data_all_final$total_below_poverty_level_per_capita)

summary_below_poverty_level_per_capita_by_state <- joined_data_all_final %>%
  group_by(state) %>%
  summarise(below_poverty_level_per_capita = sum(poverty_002) / sum(poverty_001) )

ggplot_poverty_level <- summary_below_poverty_level_per_capita_by_state %>%
  ggplot(aes(state, below_poverty_level_per_capita)) +
  geom_col(show.legend = F) +
  theme_economist() +
  theme(axis.text.x=element_blank(),
        axis.title.y=element_blank()) +
  geom_text(aes(label = state, y = below_poverty_level_per_capita/2), size = 4.5, color =
"white") +
  geom_hline(aes(yintercept=mean(below_poverty_level_per_capita)), linetype = 5,
col="red")

ggplot_poverty_level + labs(title = "Total below poverty line per capita",
                            subtitle = " i.e. total population below poverty level / total population
whom poverty status Is determined")
```

```{r total_without_health_ins_per_capita}
# Total without health insurance
joined_data_all_final <- joined_data_all_final %>%
  mutate(total_without_health_ins =
           hi_005 + hi_008 + hi_011 + hi_014 + hi_017 + hi_020 + hi_023 + hi_026 + hi_029
+
           hi_033 + hi_036 + hi_039 + hi_042 + hi_045 + hi_048 + hi_051 + hi_054 +
hi_057)

# Total without health insurance per capita
joined_data_all_final <- joined_data_all_final %>%
mutate(total_without_health_ins_per_capita = total_without_health_ins / hi_001)

# Check min, max
describe(joined_data_all_final$total_without_health_ins_per_capita)
```

```
summary_without_health_ins_per_capita_by_state <- joined_data_all_final %>%
  group_by(state) %>%
  summarise(without_health_ins_per_capita = sum(total_without_health_ins) /
sum(hi_001) )

ggplot_without_health_ins <- summary_without_health_ins_per_capita_by_state %>%
  ggplot(aes(state, without_health_ins_per_capita)) +
  geom_col(show.legend = F) +
  theme_economist() +
  theme(axis.text.x=element_blank(),
        axis.title.y=element_blank()) +
  geom_text(aes(label = state, y = without_health_ins_per_capita/2), size = 4.5, color =
"white") +
  geom_hline(aes(yintercept=mean(without_health_ins_per_capita)), linetype = 5,
col="red")

ggplot_without_health_ins + labs(title = "Total without health insurance per capita ",
                         subtitle = " i.e. sum of the population without health insurance
coverage for each age group / total civilian noninstitutionalized population")
```
```{r median_income}
# Median income - income_001
# Check median income by State (assume the average income is similar to the median
income)
summary_median_income_by_state <- joined_data_all_final %>%
  group_by(state) %>%
  summarise(median_income_by_state = sum(income_001*popestimate2015) /
sum(popestimate2015) )

ggplot_income <- summary_median_income_by_state %>%
  ggplot(aes(state, median_income_by_state)) +
  geom_col(show.legend = F) +
  theme_economist() +
  theme(axis.text.x=element_blank(),
        axis.title.y=element_blank()) +
  geom_text(aes(label = state, y = median_income_by_state/2), size = 4.5, color =
"white") +
  geom_hline(aes(yintercept=mean(median_income_by_state)), linetype = 5, col="red")

ggplot_income + labs(title = "Median income fot the whole population",
                subtitle = "i.e. income_001")
```
```{r incidence_rate}
# Incidence rate (Age-Adjusted Incidence Rate - cases per 100,000)
# [Uncertainty: only for aged 75 and above?]
# Check incidence rate by state
summary_incidence_by_state <- joined_data_all_final %>%
  group_by(state) %>%
```

```
  summarise(incidence_rate_per_100000 =
sum(age_adjusted_incidence_rate_cases_per_100_000 * popestimate2015) /
sum(popestimate2015) )

ggplot_incidence <- summary_incidence_by_state %>%
  ggplot(aes(state, incidence_rate_per_100000)) +
  geom_col(show.legend = F) +
  theme_economist() +
  theme(axis.text.x=element_blank(),
        axis.title.y=element_blank()) +
  geom_text(aes(label = state, y = incidence_rate_per_100000/2), size = 4.5, color =
"white") +
  geom_hline(aes(yintercept=mean(incidence_rate_per_100000)), linetype = 5,
col="red")

ggplot_incidence + labs(title = "Incidence rate",
                        subtitle = "i.e. age_adjusted_incidence_rate_cases_per_100_000")
```
```{r incidence_trend & population}
ggplot_trend <- joined_data_all_final %>%
  ggplot(aes(state, fill = recent_trend)) +
  geom_bar() +
  theme_economist() +
  scale_y_continuous(labels = function(x) format(x, big.mark = ",", scientific = FALSE))

ggplot_trend + labs(x = "State", title = "Incidence Trend by State") +
  theme(axis.title.y=element_blank())

ggplot_pop <- joined_data_all_final %>%
  ggplot(aes(state, popestimate2015, fill = recent_trend)) +
  geom_col() +
  theme_economist() +
  scale_y_continuous(labels = function(x) format(x, big.mark = ",", scientific = FALSE))

ggplot_pop + labs(x = "State", y = "Population", title = "Population at year of 2015") +
  theme(axis.title.y=element_blank()) ```
```{r final dataset}
# The final dataset for modelling purpose
joined_data_all_final_LRM <- joined_data_all_final %>% select(county_name,
                                        state_name,
                                        state,
                                        fips_clean,
                                        age_adjusted_death_rate,
                                        total_below_poverty_level_per_capita,
                                        total_without_health_ins_per_capita,
                                        income_001,
                                        age_adjusted_incidence_rate_cases_per_100_000,
                                        trend_falling,
                                        trend_rising,
                                        popestimate2015) %>% rename(
```
```

```
                                        stroke_mort = "age_adjusted_death_rate",
                                        below_poverty =
"total_below_poverty_level_per_capita",
                                        no_health_ins =
"total_without_health_ins_per_capita",
                                        median_income = "income_001",
                                        incidence =
"age_adjusted_incidence_rate_cases_per_100_000",
                                        incidence_falling = "trend_falling",
                                        incidence_rising = "trend_rising",
                                        population = "popestimate2015")
```

2. Model Evaluation

2.1 Check on Correlations

```{r correlations}
corrplot(cor(joined_data_all_final_LRM[, 5:12], use="pairwise.complete.obs"))

plot(joined_data_all_final_LRM[, 5:12])
```

2.2 Fit Preliminary Linear Regression

2.2.1 With All Variables

```{r linear regression all}
fit1 <-lm(stroke_mort ~. , joined_data_all_final_LRM[, 5:12]) # regression wrt all
covariates in dataframe
summary(fit1)
```

2.2.2 Remove incidence_rising

```{r linear regression all - incidence_rising  }
fit2 <-lm(stroke_mort ~. - incidence_rising, joined_data_all_final_LRM[, 5:12])
summary(fit2)
```

2.2.3 Remove incidence_falling

```{r linear regression all - incidence_rising - incidence_falling}
fit3 <-lm(stroke_mort ~. - incidence_falling - incidence_rising,
joined_data_all_final_LRM[, 5:12])
summary(fit3)
```

2.2.4 Remove population

```{r linear regression all - incidence_falling - incidence_rising - population}
fit4 <-lm(stroke_mort ~. - incidence_falling - incidence_rising - population,
joined_data_all_final_LRM[, 5:12])

summary(fit4)
```

2.2.5 Remove income

```{r remove income}
fit5 = update(fit4,~.- incidence_falling - incidence_rising - population - median_income )
summary(fit5)
```

2.2.6 Final Improved Preliminary Regression Model

```{r final improved regression model}
fit_final <-  fit5

par(mfrow=c(2,2))
plot(fit_final) # 4 plots of various things for linear fit
```