# Exploring the BRFSS data. Developed by Jorge Álvarez de la Fuente

## Setup

### Load packages

```
library(ggplot2)
```

```
## Warning in as.POSIXlt.POSIXct(Sys.time()): unable to identify current timezone 'H':
## please set environment variable 'TZ'
```

```
library(dplyr)
library(RColorBrewer)
library(knitr)
library(tidyverse)
```

### Load data

BRFSS data set from 2013 loaded as RData file:

```
load("brfss2013.RData")
```

The dimensions of the RData file shows there are 491775 samples and 330 variables:

```
#dimension of data
dim(brfss2013)
```

```
## [1] 491775    330
```

The data types breakdown is as follows: 330 variables containing 237 factors, 77 integers and 16 numeric data type.

```
#data types
table(sapply(brfss2013, class))
```

```
##
##  factor integer numeric
##     237      77      16
```

---

## Part 1: Data

**Introduction:** The Behavioral Risk Factor Surveillance System (BRFSS) is a collaborative project involving all of the states in the United States and other US territories and managed by Centers for Disease Control and Prevention (CDC). It was established in 1984 with 15 states and at present collects data in all 50 states in addition to District of Columbia and other U.S. territories.

The BRFSS is an ongoing surveillance system aiming to collect data on preventive health practices and risk behaviors related to chronic diseases, injuries and preventable infectious diseases concerning adult population. Participants do not get a monetary reward but they help to improve the health of U.S. residents. The number of interviews conducted in each state vary based on funding and the size of regions.

**Sampling, Population and Data Collection:** The data is collected for the non-institutionalized adult population (18 years of age and older) residing in the U.S. comprising 50 states in addition to the District of Columbia, Puerto Rico, Guam, US Virgin Islands, American Samoa, Federal States of Micronesia and Palau.

BRFSS conducts both landline telephone and cellular telephone based surveys. The states use a standardized core questionnaire, optional modules and questions added. BRFSS uses random assignment through Random Digit Dialing (RDD) techniques on telephone.

**BRFSS Generabizability:** Generalizability can be defined as the extension of research findings and conclusions from a study conducted on a sample population at large. The dataset "brfss2013" has roughly 500,000 observations and 330 likely variables. This dataset, therefore, constitutes a big sample population from 50 states and other U.S. territories. The larger the sample population, the more can be generalized the results. Given the large sample used in this study and the fact that the samples used in BRFSS surveys can be considered representative of all non-institutionalized adult population (18 years of age and older) residing in the U.S. The results, therefore, can be representative for the entire U.S. population.

**BRFSS Causality:** Causal inference is the process of drawing a conclusion about a causal connection based on the conditions of the occurrence of an effect. Given that the BRFSS is an observational study using random sampling where subjects are not assigned to groups such as treatment and control groups, BRFSS data can not be used to establish causality or causal inference. Since BRFSS data was not collected through an experiment, only associations can be made.

**Sources of Bias:** Although the BRFSS data sampling is fairly large to consider the dataset representative for the entire U.S. population, there are still some possible sources that introduce bias into the sample:

- Non-response bias: The surveys are carried out by telephone. If a low percentage (such as 30%) of the people randomly sampled for the survey actually respond, it would be unclear if the results are representative of the entire population as results can be skewed.
- Other sources of bias: The accuracy of answers may be affected by the interest in the topic shown by participants. Some people surveyed may be reluctant to answer honestly to sensitive questions related to their health condition.

---

## Part 2: Research questions

**Research question 1:**

**What is the impact of income, education levels and marital status on health?** According to Mirowsky and Ross ("Education, Social Status, and Health" page 201) the higher the level of income, the smaller the effect on health of any specific dollar amount added to it or subtracted from it. That happens partly because higher education reduces the effect of differences in income on health while also increases

the average level of household income. In addition, Robards, et al. ("Marital status, health and mortality") found that literature on health and mortality by marital status has consistently identified that unmarried individuals generally report poorer health and have a higher mortality risk than their married counterparts, with men being particularly affected in this respect. These conclusions will be explored and verified with BRFSS data sampling.

Variables used:

- genhlth: General Health.
- income2: Income Level.
- educa: Education Level.
- marital: Marital Status.

**Research question 2:**

**Which is the distribution of general health status within the sample related to Body Mass Index (BMI) of participants? How are these variables related to gender?** Ford, et al. 2012 in their research "Self-Reported Body Mass Index and Health-Related Quality of Life: Findings from the Behavioral Risk Factor Surveillance System" aimed to find the relationship between self-reported body mass index (BMI) and health-related quality of life in the general adult population in the United States. They concluded that low and increased self-reported BMI significantly impaired health-related quality of life and findings from previous studies have shown that overweight and obese persons have a worse health-related quality of life. This question aims to find a pattern between participants feedback regarding their health status with a more objective categorical variable: BMI. The same study reports a slightly more poor physical functioning as BMI in women than men. These differences between genders also reflects how men are more likely to over-report their heights than are women. These findings will be explored and verified with BRFSS data sampling.

Variables used:

- X_bmi5cat: Computed Body Mass Index Categories
- sex: Respondents Sex
- genhlth: General Health

**Research question 3:**

**Which states have the highest healthcare coverage and general health status?** According to Carmen DeNavas-Walt in its report "Income, Poverty, and Health Insurance Coverage in the United States: 2005" the South and the West had the highest uninsured rates of United States. The uninsured rate in the South increased from 18.2% to 18.6% between 2004 and 2005. The West also experienced an increase in the percentage of uninsured, from 17.4% in 2004 to 18.1% in 2005. The Midwest and the Northeast had the lowest uninsured rates in 2005, at 11.9% and 12.3% respectively. This question aims to find a relationship between healthcare coverage and general health status in the different states and other U.S. territories, as well as finding relationships of health care coverage with income levels, age groups, genders and verifying whether Carmen deNavas-Walt report findings remains nowadays using the brfss2013 dataset.

Variables used:

- hlthpln1: Have Any Health Care Coverage
- X_state: State or U.S. territory
- X_age_g: Imputed Age in Six Groups
- income2: Income Level
- sex: Respondents Sex

## Part 3: Exploratory data analysis

### Research question 1:

```
health1 <- brfss2013 %>%
  filter(!is.na(genhlth), !is.na(educa), !is.na(income2), !is.na(marital)) %>%
  select(genhlth, educa, income2, marital)
```

This excludes NA values for the target variables genhlth, educa, income2, marital and holds the result in a new variable health1.

Figure 1 shows the impact of education level on health:

```
ggplot(data = health1, mapping = aes(x = educa, fill = genhlth)) +
  geom_bar(position = "fill") + labs(x = "Education Level", y = "Participants Ratio") +
  ggtitle("Impact of Education Level on Health") +
  scale_x_discrete(label = function(x) stringr::str_trunc(x, 24)) +
  scale_fill_discrete(name = "General Health") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

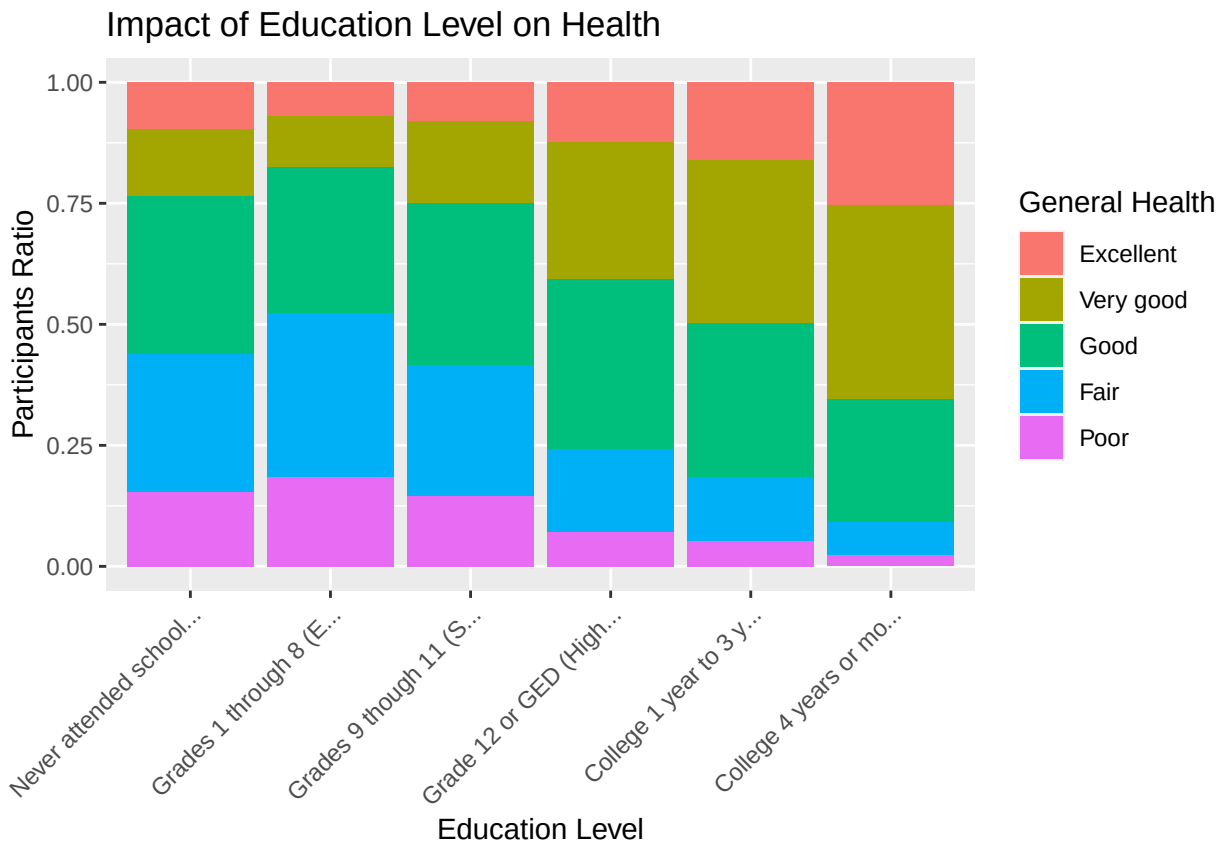


**Figure 1**. Impact of Education Level on Health.

- The proportion of reports of Excellent and Very Good health increases with higher education level, while is significantly lower with lower education level.

- Poor health reports are higher as per percentage in lower education levels compared to higher education levels.

Figure 2 shows the impact of income level on health:

```
ggplot(data = health1, mapping = aes(x = income2, fill = genhlth)) +
  geom_bar() + labs(x = "Income Level", y = "Number of Participants") +
  ggtitle("Impact of Income Level on Health") +
  scale_fill_discrete(name = "General Health") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

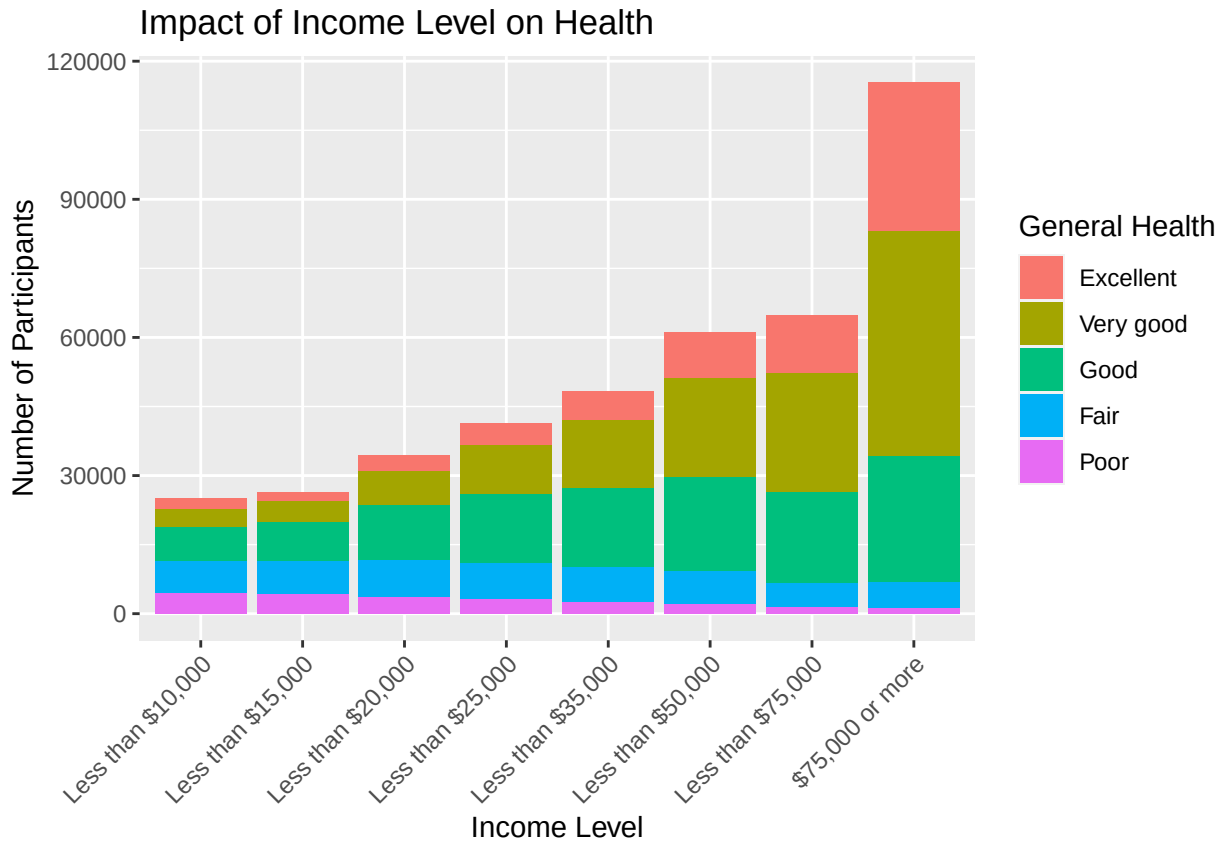


**Figure 2**. Impact of Income Level on Health.

- Likewise as for Education Level the proportion of reports of Excellent and Very Good health increases with higher income level, while is significantly lower with lower income level.
- Poor health and fair health reports are higher as per percentage in lower income levels compared to higher income levels.

Figure 3 shows the impact of marital status on health:

```
ggplot(data = health1, mapping = aes(x = marital, fill = genhlth)) +
  geom_bar () + labs(x = "Marital Status", y = "Number of Participants") +
  ggtitle("Impact of Marital Status on Health") +
  scale_fill_discrete(name = "General Health") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

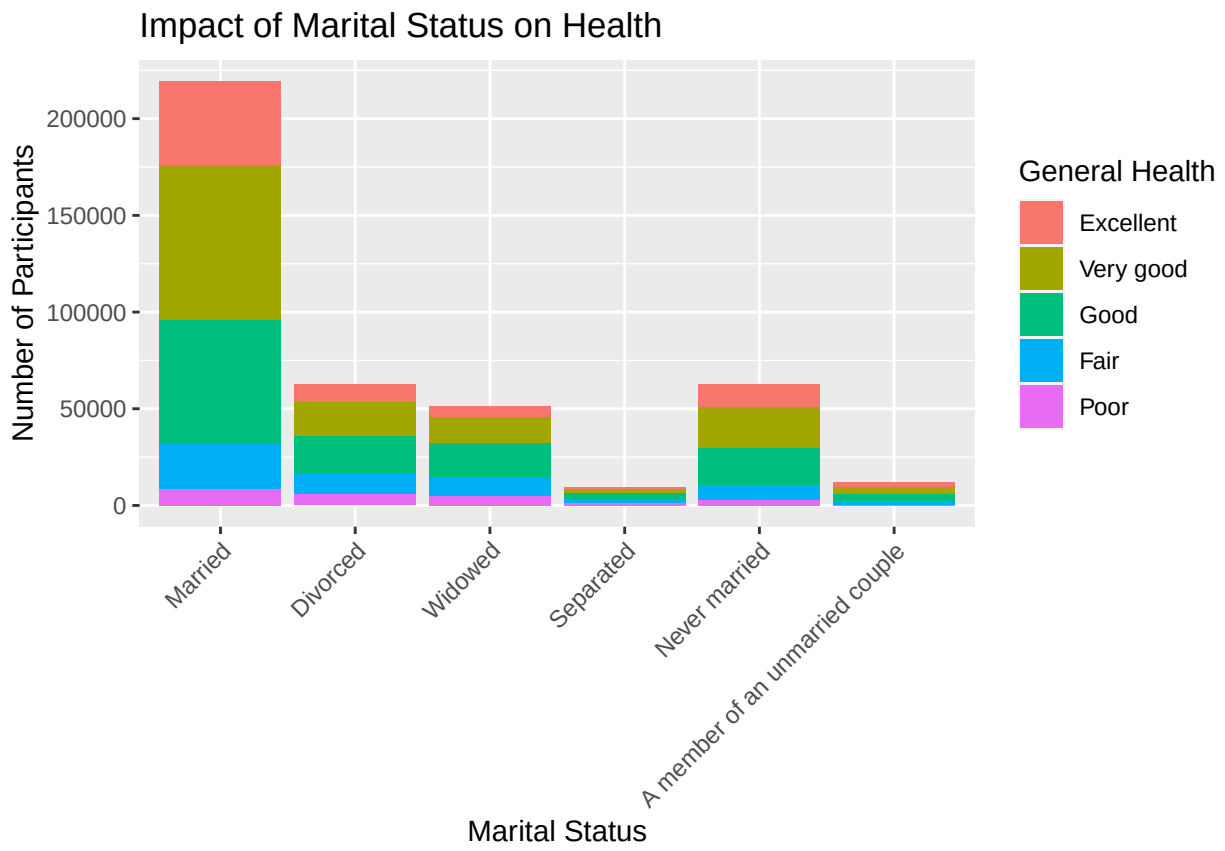


**Figure 3**. Impact of Marital Status on Health.

- The proportion of reports of Excellent and Very Good health are significantly higher for married people, in contrast to single and widowed people.

According to the results, It can be concluded and verified that higher education levels lead to a better general health condition. The same can be concluded for higher income levels and married people in contrast to single and widowed people.

**Research question 2:**

Firstly, It is interesting to know the participants health status stratified by gender. Figure 4 shows the general health status for males and females.

```
health2 <- brfss2013 %>%
  filter(!is.na(sex), !is.na(genhlth)) %>%
  group_by(genhlth, sex) %>%
  summarise(n = n(), .groups='drop')
```

This excludes NA values for the target variables sex, genhlth and holds the result in a new variable health2.

```
ggplot(health2, mapping = aes(x = genhlth, y = n, fill = sex)) +
  geom_bar(stat="identity", position = position_dodge(0.9), alpha = 0.8) +
  labs(x = "General Health Status", y = "Number of Participants") +
  ggtitle("Health Distribution Status by Gender") +
  scale_fill_brewer(name = "Gender", palette = "Set1") +
  scale_y_continuous(breaks = seq(0, 100000, by = 10000))
```

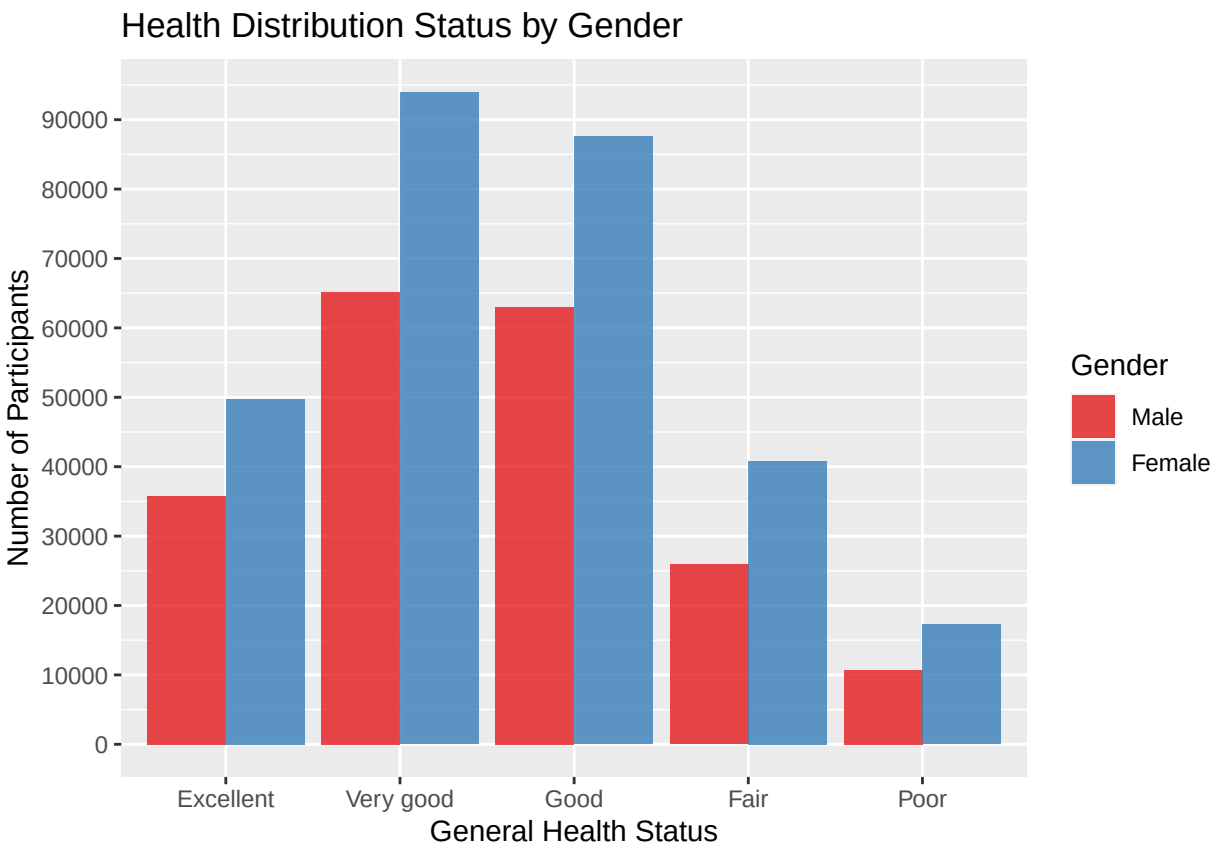


**Figure 4**. Health Distribution Status by Gender.

It is observed as the results given in Figure 4 that most participants are in good and above general health status, although there is a significant proportion of participants whose general health status falls into fair and poor categories.

Secondly, Figure 5 shows the ratio of males/females in each BMI categories related to their General Health Status:

```
BMI1 <- brfss2013 %>%
  filter(!is.na(genhlth), !is.na(sex), !is.na(X_bmi5cat)) %>%
  group_by(genhlth, sex, X_bmi5cat) %>%
  summarise(n = n(), .groups='drop') %>%
  mutate(pct_genhlth = n/sum(n), posn_pct = cumsum(pct_genhlth)-0.5*pct_genhlth)
```

This excludes NA values for the target variables genhlth, sex, X_bmi5cat and holds the result in a new variable BMI1.

```
ggplot(BMI1, aes(x = genhlth, y = pct_genhlth, fill = X_bmi5cat)) +
  geom_bar(stat = "identity", position = "fill", alpha = 0.8, col = "black") +
  facet_wrap(~sex, ncol = 2) +
  ggtitle("BMI Distribution by Gender and Health Status") +
  labs(x = "General Health Status", y = "Ratio") +
  scale_fill_manual(name="BMI", values = c("#FFCC00", "#339900", "#FF9933", "#CC0000")) +
  coord_flip()
```


**Figure 5**. BMI distribution by gender and health status.

As observed in Figure 5, for both males and females there are higher ratios of obese and overweight BMI in participants who declared having poor and fair general health status compared to participants who declared having very good and excellent general health status where the ratio of normal weight is significantly higher.

As a conclusion, the results of BMI distribution by gender and health status obtained from brfss2013 dataset verify that overweight and obese people have a worse health-related quality of life as stated by Ford, et al. 2012. This results can be generalized for the U.S. residents.

**Research question 3:**

Figure 6 shows the proportion of uninsured participants in all the states and other U.S. territories:

```
NoCoverage <- brfss2013 %>%
  filter(!is.na(X_state), !is.na(hlthpln1)) %>%
  group_by(X_state, hlthpln1) %>%
  dplyr::summarise(count=n(), .groups = 'drop') %>%
  spread(hlthpln1, count) %>%
  mutate(NoCov= No/(No+Yes))
```

This excludes NA values for the target variables X_state, hlthpln1 and holds the result in a new variable NoCoverage.

```
ggplot(NoCoverage, aes(x = reorder(X_state, NoCov), y = NoCov)) +
  geom_bar(stat="identity", fill="darkblue") +
  theme(panel.border=element_rect(colour='black', fill=NA)) +
```

```
  theme(text = element_text(size=8)) +
  labs(x="State", y="Ratio of Participants with no Health Care Coverage") +
  ggtitle("Ratio of Participants with No Health Care Coverage by state and/or U.S. territory") +
  coord_flip()
```

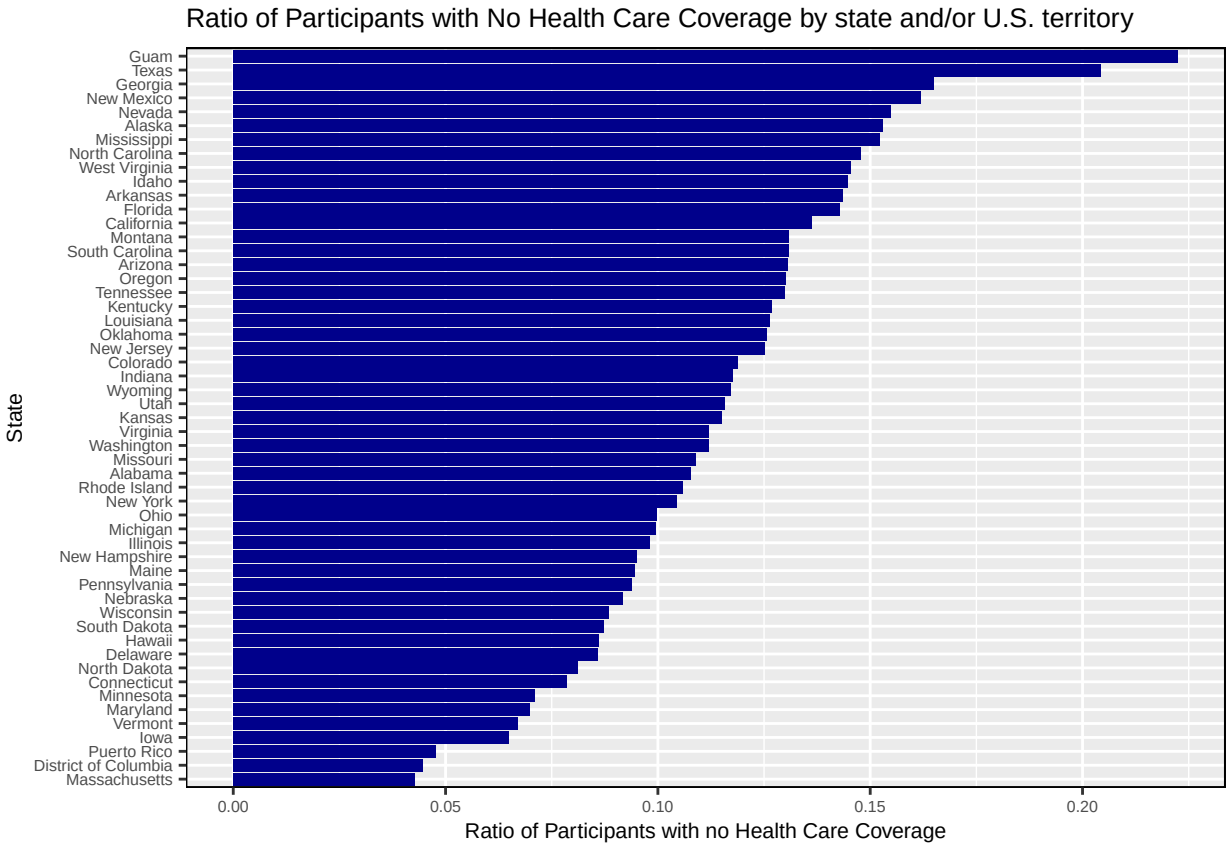


**Figure 6**. Ratio of Participants with no health care coverage by state and/or U.S. territory.

As observed in Figure 6, most of the southern and western states have the lowest health care coverage. Such are the cases of Texas, Georgia, New Mexico, Nevada and California. On the other hand, northeast and midwest states have a significant higher Health Care Coverage compared with southern and western states. Such are the cases of New York, Delaware, Michigan and Wisconsin. Thus It can be concluded that Carmen deNavas-Walt report findings remains nowadays.

Figure 7 shows the proportion of health care coverage by income level, age group and gender.

```
hmap <- brfss2013 %>%
  filter(!is.na(X_age_g), !is.na(income2), !is.na(hlthpln1)) %>%
  group_by(X_age_g, income2, sex, hlthpln1) %>%
  dplyr::summarise(count=n(), .groups = 'drop') %>%
  spread(hlthpln1, count) %>%
  mutate(Cov= Yes/(No+Yes))
```

This excludes NA values for the target variables X_age_g, income2, hlthpln1 and holds the result in a new variable hmap.

```
ggplot(hmap, aes(x = X_age_g, y = income2, fill = Cov)) +
    geom_tile() +
    scale_fill_gradient2(low = "yellow", high = "darkgreen", guide = guide_legend("Health Care Coverage
    geom_text(aes(label = round(Cov, 2)), size = 3, color = "white") + facet_grid(sex ~.) +
    theme(axis.text.x = element_text(size = 10, angle = 90, vjust = 0.8)) +
    labs( x = "Age Group" , y = "Income Level") +
    ggtitle("Health Care Coverage Ratio")
```

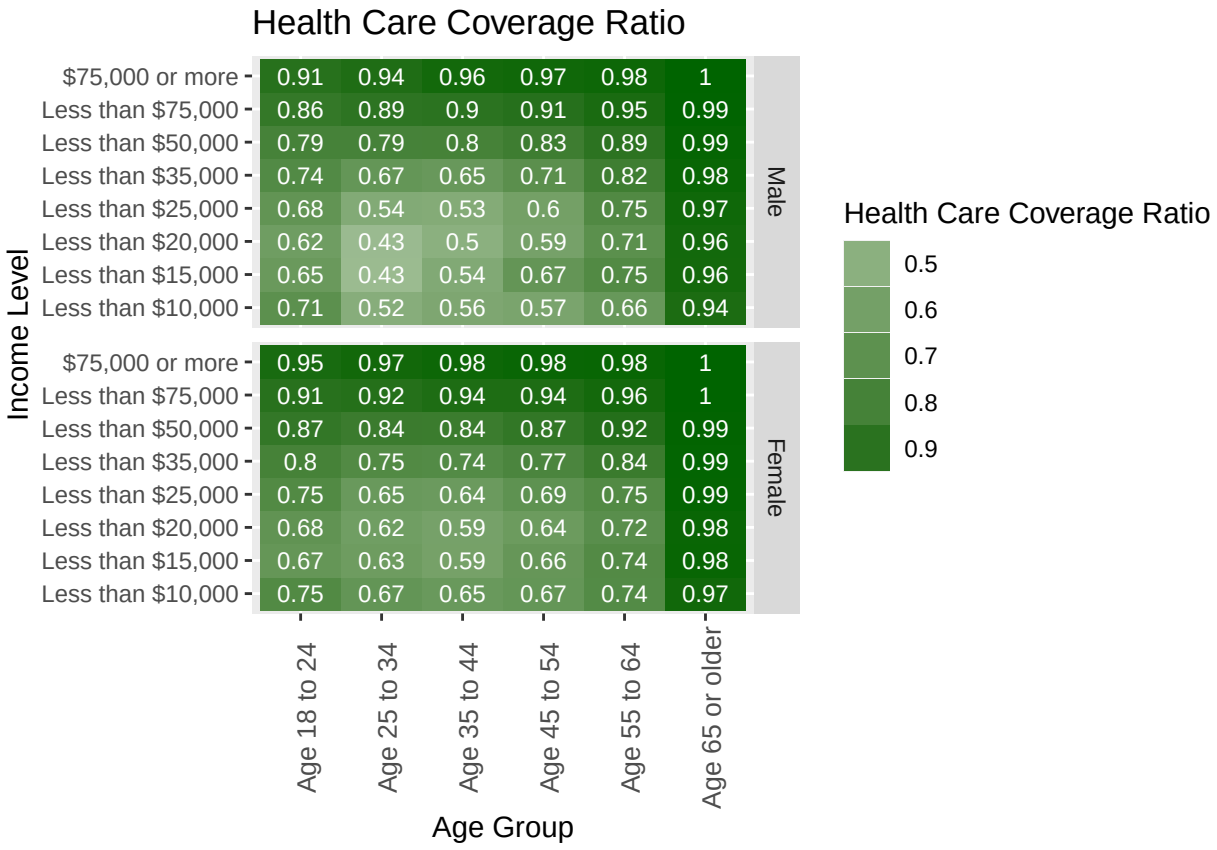


**Figure 7**. Health care coverage ratio by income level, age group and gender

The results give some interesting facts about the health care coverage in United States:

- Participants with the highest income levels ($75,000 and above) have the highest health care coverage ratio for both males and females and in a wide range of age groups (from 18 years old to 65 or older)
- Elderly participants have the highest health care coverage ratio among all age groups.
- Generally, females have higher health care coverage compared with male counterparts.

Figure 8 shows a box plot comprising the health care coverage ratio by gender:

```
ggplot(hmap, aes(x = sex, y = Cov, fill = sex)) +
  geom_boxplot() +
  scale_y_continuous(breaks = seq(0, 1, by = 0.1)) +
  scale_fill_manual(values=c("coral1", "cadetblue3"), guide = guide_legend("Gender")) +
  labs(x = "Gender", y = "Health Care Coverage Ratio")
```

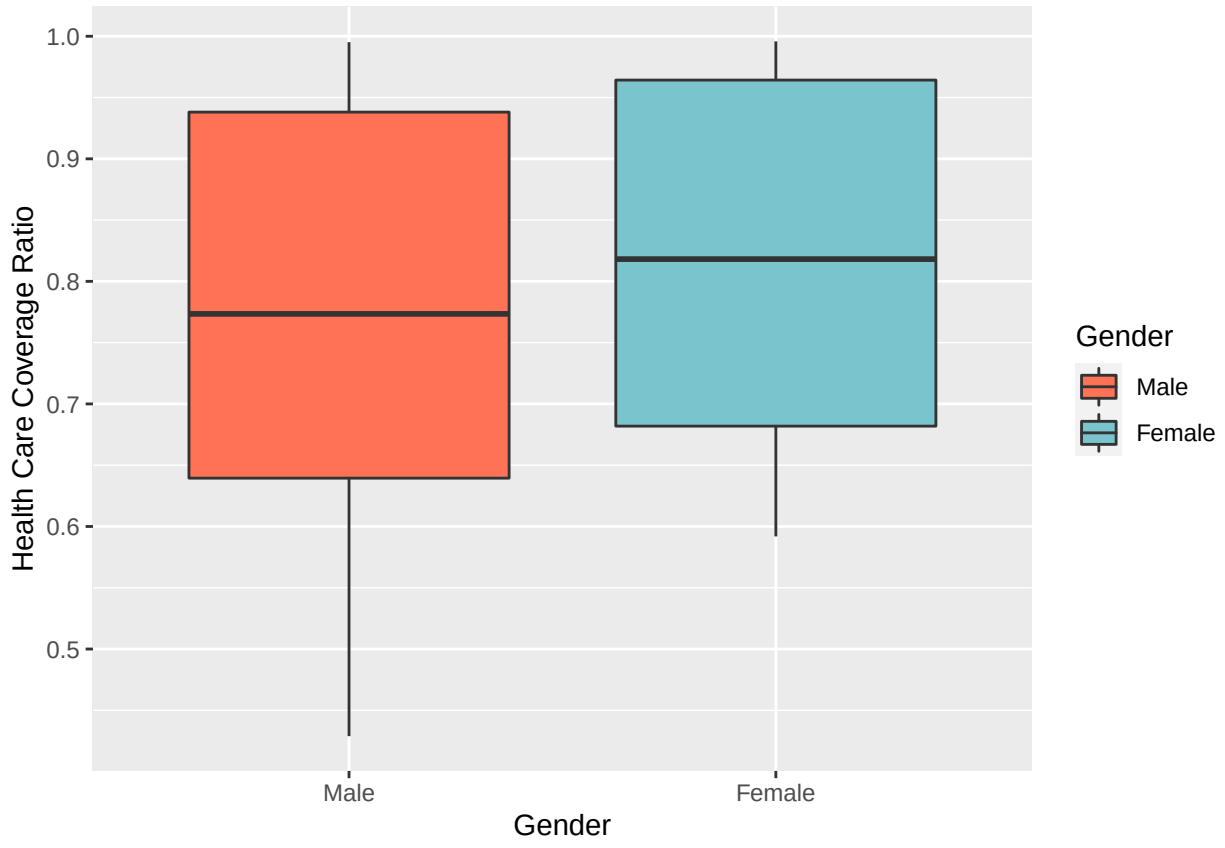


**Figure 8**. Box plot of health care coverage ratio by gender

From the box plot It can be observed that the median for the health care coverage ratio for females is higher (~0.82) than for males (~0.77). There is more variability for males than females and IQR for males is roughly 0.29 while for females is roughly 0.28. No potential outliers were found.

The box plot results confirm that generally females have a better health care coverage than males.