**Part 1: Data**

1. **BRFSS Background**

The Behavioral Risk Factor Surveillance System (BRFSS) it’s a health telephone survey conducted continuously through the year across USA by Centers for Disease Control and Prevention (CDC). Its objective is to collect health specific data on preventive health practices and risk behaviors linked to chronic diseases, injuries, and preventable infectious diseases that affect the adult population aged 18 or older across all USA at state and local level.   
BRFSS completes more than 400,000 adult interviews each year, making it the largest continuously conducted health survey system in the world. It is worth mentioning due to its success a number of countries have expressed interest in adopting the survey to their countries.

Starting in 2011 the survey is conducted using two sample methodologies:

1. By landline telephone where data is collected from a randomly selected adult in the household.
2. By cellular telephone data is collected from an adult that resides in a private residence or college housing

In 2013, additional question sets were included as optional modules to provide a measure for several childhood health and wellness indicators, including asthma prevalence for people aged 17 years or younger.

1. **Applications of BRFSS**

By gathering behavioral health risk data at the state and local level, BRFSS has become a powerful tool for targeting and building health promotion activities.

1. **Drawing Conclusions from the survey: Inference opportunity**

The survey uses *random sampling* but *no random assignment* therefore it’s an observational study.

* **Generalizability -Yes**
* **Causation -No**

The survey is conducted using Random Digit Dialing (RDD) techniques on both landlines and cell phones by generating telephone numbers at random. Once contacted interviewers asks people questions (no random assignment) out of the questionnaire. The survey it’s an observational study and analysis results can be generalized however causality cannot be inferred.

According to BRFSS site “the survey uses a weighting methodology known as iterative proportional fitting or raking in order to address some problems encountered with surveys, especially the non-response bias. This weighting methodology reduces the potential for bias and increases the representativeness of estimates. In addition, raking allows for the incorporation of a now-crucial variable—telephone ownership (landline and/or cellular telephone)—into the BRFSS weighting methodology.”

## Part 2: Research questions

Research question 1: Examine the relationship between education level and general health in male/female population

Increase in health care costs have raised concern about practical solutions to improve health in population and create tailored solutions. One possible way is to look at the association between education and general health

Keywords: education, general health, gender

Research question 2: Explore the relationship between income level and tobacco use

This study aims at looking at a possible relationship between income and tobacco use a known health risk factor. We are exploring if income level could make a positive impact in preventing tobacco use

Keywords: income, tobacco use

Research question 3: Investigate the relationship between veteran male/female population with respect of diabetes disease

This study intention is to look at the relationship between male/female veterans and diabetes a known health risk factor to better understand the frequency of diabetes among either the male or female population

Keywords: veteran, diabetes, gender

## Part 3: Exploratory data analysis

**Research question 1**: Examine the relationship between education level and general health in general population and male/female population

Keywords: education, general health, gender

**Research question 2**: Explore the relationship between income level and tobacco use

Keywords: income, tobacco use

**Research question 3**: Investigate the relationship between marital status and heavy alcohol consumption by gender

Keywords: marital status, alcohol consumption, gender

**Variables**

Q1:

**Education:** X\_educag

**General Health condition:** X\_rfhlth

**Gender:** sex

Q2:

**Tobacco Use:** X\_smoker3

**Income:** X\_incomg

Q3:

**Veteran:** veteran3

**Gender:** sex

**Diabetes:** diabete3

Q1

Total Observations in Table: 487556

| Health

Education | Good Health | Poor Health | Row Total |

-------------|-------------|-------------|-------------|

College | 262851 | 40553 | 303404 |

| 0.87 | 0.13 | 0.62 |

| 0.67 | 0.43 | |

| 0.54 | 0.08 | |

-------------|-------------|-------------|-------------|

HS | 107579 | 34788 | 142367 |

| 0.76 | 0.24 | 0.29 |

| 0.27 | 0.37 | |

| 0.22 | 0.07 | |

-------------|-------------|-------------|-------------|

No HS | 22948 | 18837 | 41785 |

| 0.55 | 0.45 | 0.09 |

| 0.06 | 0.20 | |

| 0.05 | 0.04 | |

-------------|-------------|-------------|-------------|

Column Total | 393378 | 94178 | 487556 |

| 0.81 | 0.19 | |

-------------|-------------|-------------|-------------|

The contingency table above shows frequency counts, relative frequencies (row, column, table) of health and education variables.

From the contingency table we notice there were almost 500k responders. The majority of people surveyed have college education and good health.

From the mosaic plot, we see that the percentage of people in good health steadily increases with education. For example, the percentage of people in good health went from 55% for people with no high school, to 76% for people with high school, to 87% for people with college education. This indicates health status is associated with education level.

The contingency table above shows frequency counts, relative frequencies (row, column, table) of health and education variables.

From the contingency table we notice there were almost 500k responders. The majority of people surveyed have reported non veteran status (87%) and not suffering from diabetes.

From the mosaic plot, we see there is a higher relative incidence of diabetes in veterans. For example, the percentage of veterans with diabetes is 21% but 15% for general population. However we do not know if the +7% delta for diabetes in veteran segment is due to randomness. Further analysis may be required.

From the contingency table we notice the majority of veterans are male (91%).

From the mosaic plot, we see there is a higher relative incidence of diabetes in veteran males than veteran females. For example, the percentage of veteran males with diabetes is 21% but 14% for veteran females. However we do not know if the +7% delta for diabetes in veteran male segment is due to randomness and being overrepresented in the veteran segment. Further analysis may be required.

\*\*This concludes Project 1 assignment. Thank you for your time.\*\*

Total Observations in Table: 187430

| Tobacco use

Income | Do not use tobacco | Use tobacco | Row Total |

-----------------------------|--------------------|--------------------|--------------------|

Less than $15,000 | 12378 | 14871 | 27249 |

| 0.45 | 0.55 | 0.15 |

| 0.11 | 0.21 | |

| 0.07 | 0.08 | |

-----------------------------|--------------------|--------------------|--------------------|

$15,000 to less than $25,000 | 20644 | 17460 | 38104 |

| 0.54 | 0.46 | 0.20 |

| 0.18 | 0.24 | |

| 0.11 | 0.09 | |

-----------------------------|--------------------|--------------------|--------------------|

$25,000 to less than $35,000 | 14129 | 9233 | 23362 |

| 0.60 | 0.40 | 0.12 |

| 0.12 | 0.13 | |

| 0.08 | 0.05 | |

-----------------------------|--------------------|--------------------|--------------------|

$35,000 to less than $50,000 | 18146 | 10314 | 28460 |

| 0.64 | 0.36 | 0.15 |

| 0.16 | 0.14 | |

| 0.10 | 0.06 | |

-----------------------------|--------------------|--------------------|--------------------|

$50,000 or more | 50316 | 19939 | 70255 |

| 0.72 | 0.28 | 0.37 |

| 0.44 | 0.28 | |

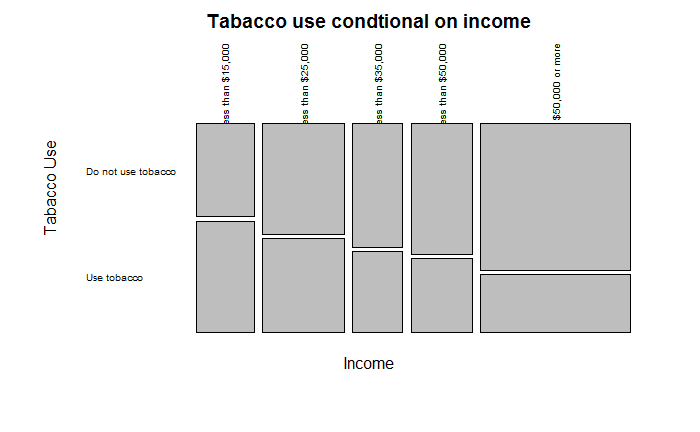
| 0.27 | 0.11 | |

-----------------------------|--------------------|--------------------|--------------------|

Column Total | 115613 | 71817 | 187430 |

| 0.62 | 0.38 | |

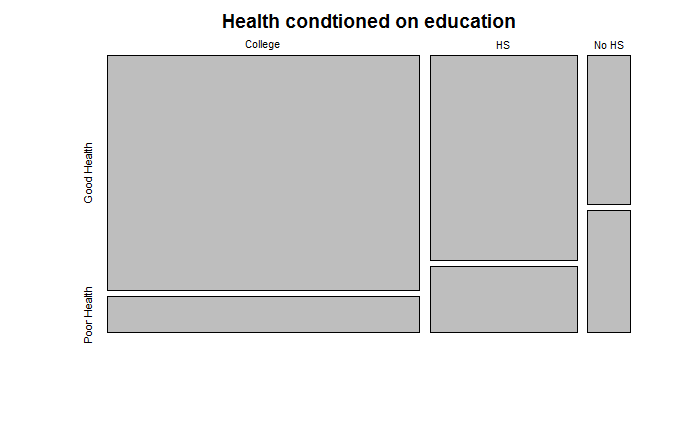
-----------------------------|--------------------|--------------------|--------------------|



The contingency table above shows frequency counts, relative frequencies (row, column, table) of health and education variables.

From the contingency table we notice there are five levels for income and almost 200k responders.

From the mosaic plot, we see that the percentage of people using tobacco steadily decreases with income. For example, the percentage of people with income of $50k or more reported that 28% use tobacco products, while at the other income end of less than $15k, 55% of people use tobacco. This suggest an association between tobacco use and income.



r