ECO239 Group R Assignment 1

Due date: November 6th (Monday) at 16:00.
No late homework will be accepted.

Things to submit:
1. Your report (code, output including figures, your findings/comments).
2. Contribution Sheet (signed by all the group members)
3. Honor Code (signed by all the group members)
*Just stable together. No plastic covers, please.

This is a group assignment. Create a group of maximum 4 members. Carefully read Honor Code before starting your assignment.

Data set: cdc.csv
Download either from our course web page or at *http://www.openintro.org/stat/data/cdc.R*

Basic Tasks

1. Select one numerical-discrete variable and one numerical-continuous variable and numerically summarize data. Report your finding.

2. Select one numerical variable and generate (a) Frequency Distribution Table and (b) Histogram. For Histogram, comment the shape of the distribution. Report your finding.

3. Select one categorical variable and generate (a) Frequency Distribution Table, (b) bar plot and (c) pie chart. Report your finding.

4. Select one categorical variable and one related numerical variable to generate a meaningful box-plot. Report your finding.

5. Select two related numerical variables and generate a scatter plot. Report your finding.

6. Define the Body Mass Index (BMI) in R. BMI is a weight to height ratio and can be calculated as

$$BMI = \frac{weight(lb)}{(height(in))^2} * 703$$

703 is the approximate conversion factor to change units from metric (meters and kilograms) to imperial (inches and pounds).

```
bmi <- (cdc$weight/cdc$height^2) * 703
```

Generate a boxplot and report your finding.

7. Find the mean and standard deviation of weight and determine what proportion of the weights are within one standard deviation of the mean.

Research Task

For this section, you have to analyze the relevant variables using numerical and graphical summaries to prove your findings.

a. Examine the relationship between smoking habit and gender.

b. Examine the relationships among BMI, smoking habit, general health, and exercise habit.

c. Set up an original research question of your own, conduct necessary data analysis and report your findings.

Variable Descriptions

- **genhlth**: A categorical vector indicating general health, with categories excellent, very good, good, fair, and poor.
- **exerany**: A categorical vector, 1 if the respondent exercised in the past month and 0 otherwise.
- **hlthplan**: A categorical vector, 1 if the respondent has some form of health coverage and 0 otherwise.
- **smoke100**: A categorical vector, 1 if the respondent has smoked at least 100 cigarettes in their entire life and 0 otherwise.
- **height**: A numerical vector, respondent's height in inches.
- **weight**: A numerical vector, respondent's weight in pounds.
- **wtdesire**: A numerical vector, respondent's desired weight in pounds.
- **age**: A numerical vector, respondent's age in years.
- **gender**: A numerical vector, respondent's gender

Data Information

This a data set is collected by the Centers for Disease Control and Prevention (CDC). The Behavioral Risk Factor Surveillance System (BRFSS) is an annual telephone survey of 350,000 people in the United States. As its name implies, the BRFSS is designed to identify risk factors in the adult population and report emerging health trends. For example, respondents are asked about their diet and weekly physical activity, their HIV/AIDS status, possible tobacco use, and even their level of health care coverage. The BRFSS Web site (http://www.cdc.gov/brfss) contains a complete description of the survey, including the research questions that motivate the study and many interesting results derived from the data.

We will focus on a random sample of 20,000 people from the BRFSS survey conducted in 2000. While there are over 200 variables in this data set, we will work with a small subset.