SEIS 631

Dan Ward

Assignment 2

**Question 1: How many cases (i.e. observations) and how many variables (i.e. attributes/features) are there in this data set?**

There are 20,000 observations and 9 attributes

**Question 2: What type of variable is genhlth (represented in the dataset)?** Categorical Ordinal

**Question 3: What type of variable is weight?** Numerical Continuous

**Question 4: What type of variable is smoke100?** Categorical Nominal

**Question 5: Create a numerical summary for gender. How many males are in the sample?** There are 9,569 males in the sample.

**Question 6: Compute the relative frequency distribution of genhlth. What proportion of the sample reports being in excellent health?** 23.285% report being in excellent health.

**Question 7: What does the mosaic plot reveal about smoking habits and gender?** The mosaic plot reveals that the percentage of men that have smoked at least 100 cigarettes is more than the men who have not. The percentage of men that have smoked at least 100 cigarettes is higher than the percentage of women that have smoked at least 100 cigarettes. More women have not smoked at least 100 cigarettes than have smoked at least 100 cigarettes. There were more women than men in the sample group.

**Question 8: Create a new object called under23 and smoke that contains all observations of respondents under the age of 23 that have smoked at least 100 cigarettes in their lifetime. Write the command you used to create the new object?**

under23\_and\_smoked100 = subset(cdc,cdc$age<23 & cdc$smoke100==1)

**Question 9: How many observations are in the subset under23 and smoke that you created in the previous exercise, i.e. how many people in the sample are under the age of 23 and have smoked at least 100 cigarettes in their lifetime?** There are 620 people that are under the age of 23 and have smoked at least 100 cigarettes in their lifetime in this sample group.

**Question 10: Which of the following is false based on the box plot of BMI vs. general health?**

(d) The distributions of BMIs within each health status group is left skewed.

**Question 11: Pick another categorical variable from the data set and see how it relates to BMI. List the variable you chose, why you might think it would have a relationship to BMI, and indicate what the figure seems to suggest.** I chose exercise, which I assumed that those that exercised would have a lower BMI. The figure indicates that the average BMI is lower for those that exercise and the interquartile range is smaller, and the whiskers have less gap between the quartiles.

**Question 12: In the last assignment, when exploring how percentages of boys and girls born vary over time (two numerical variables) we used a scatterplot. Using the same tools (the plot function) make a scatterplot of weight versus desired weight. Based on the plot you made, which of the following is true about the relationship between weight and desired weight?**

(c) moderately strong positive linear