**Homework 2**

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

20000 cases and 9 variables.

|  |
| --- |
| **Question 2**: **What type of variable is genhlth?** |

Categorical – Ordinal – Qualitative

|  |
| --- |
| **Question 3: What type of variable is weight?** |

Numerical – Quantitative – Continuous

|  |
| --- |
| **Question 4**: **What type of variable is smoke100?** |

Categorical – Nominal(Binary) – Qualitative

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

9569

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

0.23285

|  |
| --- |
| **Question 7: What does the mosaic plot reveal about smoking habits and gender?** |

Males – High number of males have smoked at least 100 cigarettes.

Females – Low number of females have smoked at least 100 cigarettes.

**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\_smoke = 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?**

620

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

(a) The median BMI is roughly 25 for all general health categories, and there is a slight increase in median BMI as general health status declines (from excellent to poor).

(b) The IQR increases slightly as general health status declines (from excellent to poor).

(c) Among people with excellent health, there are some with unusually low BMIs compared to the rest of the group.

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

**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.**

Created boxplot for bmi vs gender. Gender may affect bmi values as both male and female bodies are structured differently. For females median bmi is slightly lower than the males and also there are certain males who have bmi less than 15 also.

**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?**

(a) moderately weak negative linear

(b) moderately weak positive linear - True

(c) moderately strong positive linear

(d) moderately weak negative linear