Consider using the following data set of 1000 patients to create a model to predict lung cancer. There are 4 columns of interest:

1. **Cancer:** Lung cancer status (0 = no, 1 = yes)
2. **Age:** Age of patient
3. **BMI:** Body Mass Index of patient
4. **Smoking:** Smoking status (0 = non-smoker, 1 = smoker)

|  |  |
| --- | --- |
| **Question 1**  What is the response variable for this model? | |
| A | Age |
| B | BMI |
| C | Cancer |
| D | Smoking |

|  |  |
| --- | --- |
| **Question 2**  What is an appropriate method to explore the association between Cancer and BMI | |
| A | Scatter Plot |
| B | Box Plot |
| C | Histogram |
| D | Contingency Table |

|  |  |
| --- | --- |
| **Question 3**  Consider the following box plot of Age and Cancer. Select the most appropriate response.  A diagram of a cancer  Description automatically generated with medium confidence | |
| A | There is no association, Age should not be included in the model. |
| B | There is an association, Age should be included in the model. |
| C | There is an association, Age should not be included in the model. |
| D | There is no association, Age should be included in the model. |

|  |  |
| --- | --- |
| **Question 4**  Select all the appropriate methods to explore the association between Cancer and Smoking. | |
| A | Odds Ratio |
| B | Box Plot |
| C | Contingency Table |
| D | Correlation coefficient |

|  |  |
| --- | --- |
| **Question 5**  Given this photo, which of the following lines of code correctly calculates the odds ratio for the odds of developing lung cancer in smokers compared to non-smokers?  A close-up of a computer screen  Description automatically generated | |
| A | (tab[1] \* tab[2]) / (tab[3] \* tab[4]) |
| B | (tab[3] \* tab[4]) / (tab[1] \* tab[2]) |
| C | (tab[2] \* tab[3]) / (tab[1] \* tab[4]) |
| D | (tab[1] \* tab[4]) / (tab[2] \* tab[3]) |