Since I don’t know where to put the EDA part, I’m writing in another file.

Presentation: Each project is to be presented using PowerPoint in a modified Pecha Kucha style, and the PPT file will have to be submitted to be published on course web site.

Report: The report will be written in the format of a formal paper and include the following 6 sections: introduction, literature review, data description, analysis & results, conclusion, and future work. The final report should be approximately 8-10 pages long, single spaced. The literature review for the final report consists of reading and summarizing about 6-7 published papers on the review topic. The final report should describe in detail the problem being addressed, the actual analysis and results obtained, and then some conclusion providing a broader interpretation of those results as they apply to solving the problem. Finally, students should brainstorm about potential future work, or other data that could be obtained, to help better address the problem beyond the scope of the work they performed during the project.

if we had data related to frequency of smoking or drink,

[Data Description]

1. Data Source: Describe the source of the data, such as the dataset name, organization, or database from which it was obtained. If the data is collected specifically for the study, mention the data collection methods used.
2. Data Collection Period: Specify the time frame during which the data was collected. This is particularly important if the data represents a specific period or if there are temporal aspects relevant to the study.
3. Data Size and Structure: Provide information about the size of the dataset, including the number of observations or samples and the number of variables or features. Mention whether the data is structured (e.g., in tabular format) or unstructured (e.g., text or image data).
4. Data Variables: List and briefly describe the variables included in the dataset. Mention their names, types (e.g., numerical, categorical), and units of measurement if applicable. If there are any specific or derived variables used in the study, explain how they were calculated or derived.
5. Data Preprocessing: Describe any preprocessing steps performed on the data before analysis. This may include data cleaning, handling missing values, outlier treatment, normalization, or feature engineering. Explain the rationale behind these steps and how they may have affected the dataset.
6. Ethical Considerations: If applicable, discuss any ethical considerations related to data collection or usage, such as data privacy, consent, or adherence to ethical guidelines.

Originally, it had 10000 rows and 34 columns.

[Description of all columns]

|  | Description | Type | Value | Null |
| --- | --- | --- | --- | --- |
| No | Number of index | int64 | max: 9999 / min: 0 | 10000 non-null |
| Age | age at diagnosis | int64 | max: 88 / min: 31 | 10000 non-null |
| Adenocarcinoma | Histological Diagnostic Name | int64 | 0: false / 1: true | 10000 non-null |
| Large cell carcinoma | Histological Diagnostic Name | int64 | 0: false / 1: true | 10000 non-null |
| Squamous cell carcinoma | Histological Diagnostic Name | int64 | 0: false / 1: true | 10000 non-null |
| TX, T0, T1, T1a, T1b, T1c, T2, T2a, T2b, T3, T4, N1, N2, N3, M1a, M1b, M1c | Stage | int64 | 0: false / 1: true | 10000 non-null |
| Type of Drink | Type of Drinking | int64 | 1: beer / 2: soju / 3: liquor / 99: etc | 10000 non-null |
| Smoke | Smoking Status | int64 | 0: not / 1: current / 2: past | 10000 non-null |
| Height | Height | float64 | max: 188.3 / min: 142.0 | 10000 non-null |
| Weight | Weight | float64 | max: 105.1 / min: 34.3 | 10000 non-null |
| FEV1\_FVC\_P | FEV test result | int64 | max: 99 / min: 31 | 10000 non-null |
| DLCO\_VA\_P | DLCO test result | int64 | max: 161 / min: 27 | 10000 non-null |
| EGFR mutation Detection | Whether EGFR mutation is found | int64 | max: 99 / min: 0 | 10000 non-null |
| Operation | Surgery or not | int64 | 0: false / 1: true | 10000 non-null |
| Chemotherapy | Chemotherapy | int64 | 0: false / 1: true | 10000 non-null |
| Radiation Therapy | Radiation Therapy | int64 | 0: false / 1: true | 10000 non-null |
| Death | Death | int64 | 0: false / 1: true | 10000 non-null |
| Survival period | Survival period | int64 | max: 730 / min: 16 | 10000 non-null |

Before proceeding with EDA, unnecessary columns (‘No’, 'TX', 'T0', 'N1', 'N2', 'N3', 'M1a', 'M1b', 'M1c') were removed in the preprocessing step. There was no null value and no value to be judged as an outlier in the data, so any operations were not performed other than checking the information (data type, max, mean, etc.). However, several columns that were deemed necessary in the EDA stage were added.

[Description of the added column]

|  | Description | Type | Value | Null |
| --- | --- | --- | --- | --- |
| Cancer | Cancer type  0 (false) when 0 for all diagnostic criteria\*, 1 (true) when any one is 1 | int64 | 0: false(non-cancer) / 1: true(cancer) | 10000 non-null |
| Drink\_mapped | Specify what kind of alcohol it is by mapping it into an object type | object | Beer, Soju, Liquors, Others | 10000 non-null |
| Smoke\_mapped | Specify which smoking type corresponds to by mapping it to the object type. | object | Not, Current, Past | 10000 non-null |
| Smoke\_re\_mapped | Not divided into not/current/past, but only whether it is smoked or not. | object | Not, Smoke | 10000 non-null |
| BMI | Calculate BMI using Height and Weight | float64 | max: 51.32 / min: 10.70 | 10000 non-null |
| BMI\_mapped | Categorized according to the calculation method\*\* | int64 | 0: Underweight /  1 : Normal /  2: Overweight + Obese | 10000 non-null |
| Smoke\_re\_int\_mapped | Replace the 'Smoke\_re\_mapped' column with int form and mark 1 (true) for smoke and 0 (false) for Not | int64 | 0: false(Not) /  1: true(Smoke) | 10000 non-null |
| FEV\_filter | 0 (false) and 1 (true) indicate whether the value of 'FEV1\_FVC\_P' is less than 70 | int64 | 0: false /  1: true(below 70) | 10000 non-null |
| DLCO\_filter | 0 (false) and 1 (true) indicate whether the value of DLCO\_VA\_P is less than 80 | int64 | 0: false /  1: true(below 80) | 10000 non-null |

\* 'Adenocarcinoma', 'Large cell carcinoma', 'Squamous cell carcinoma', 'T1', 'T1a', 'T1b', 'T1c', 'T2', 'T2a', 'T2b', 'T3', 'T4'

\*\* provided by the Korea Centers for Disease Control and Prevention

Also, for the second EDA, one column that was not included for the first EDA was added. To be precise, the 'Cancer' column changed because the cancer classification was different. Previously, when all diagnostic criteria were 0, it was determined as 0 (false, non-cancer). However, this time, through "Adenocarcinoma," "Large cell carcinoma," and "Squamous cell carcinoma," cancer was also divided into cancers of small cells and cases that were not. The three diagnoses above are not small-sized cancers. Therefore, if one of the three is 1, it is classified as 2 (non-small-sized cancel), if all three are 0, and if there is even one of the remaining diagnostic criteria\*, it is classified as 1 (small-sized cancel), and if it is 0 in all cases, it is classified as non-cancer. And other columns were the same for the first EDA.

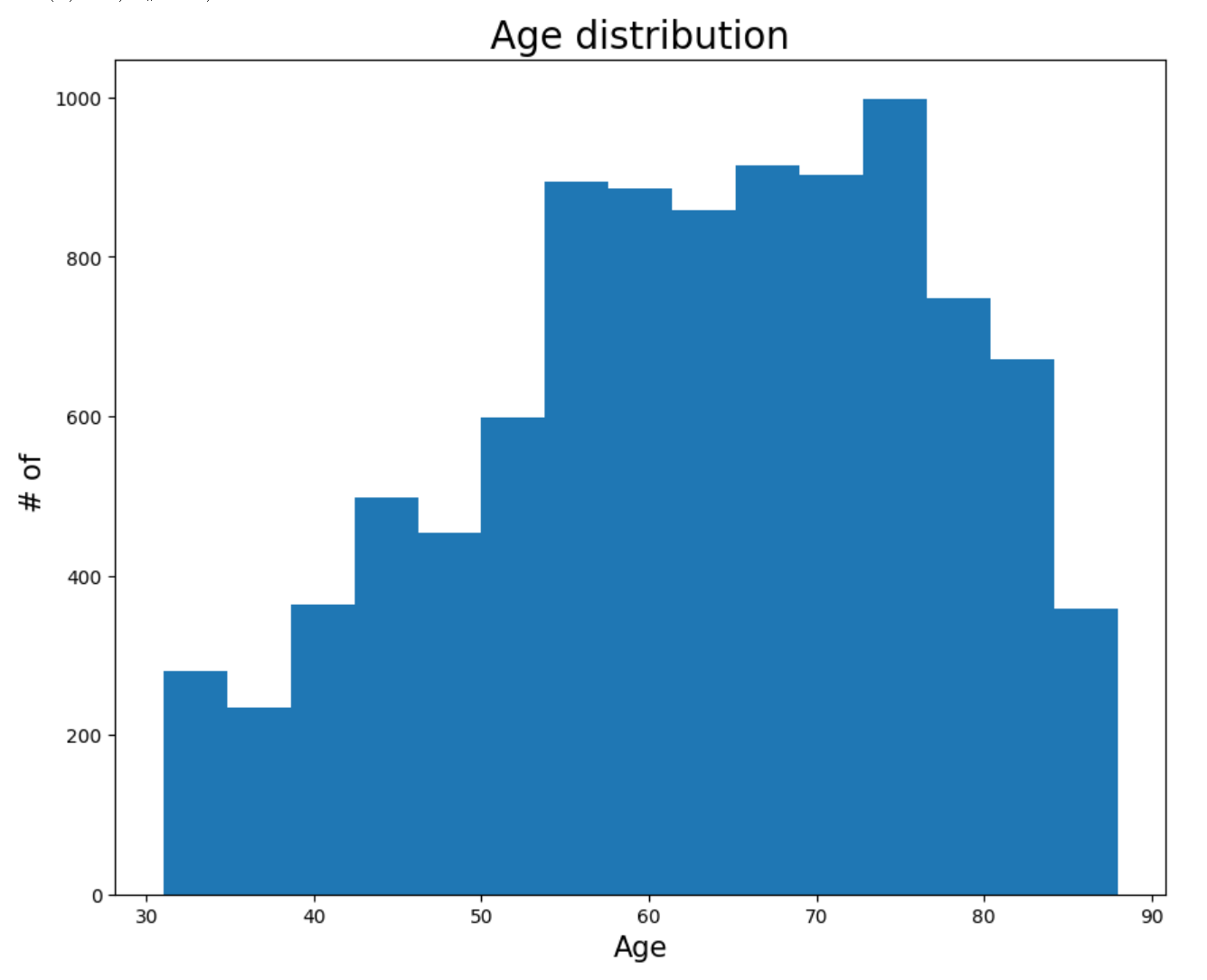
|  | Description | Type | Value | Null |
| --- | --- | --- | --- | --- |
| Cancer | Cancer type | int64 | 0: non-cancer / 1: small-sized cancer / 2: non-small sized cancer | 10000 non-null |

\* 'T1', 'T1a', 'T1b', 'T1c', 'T2', 'T2a', 'T2b', 'T3', 'T4'

[Analysis]

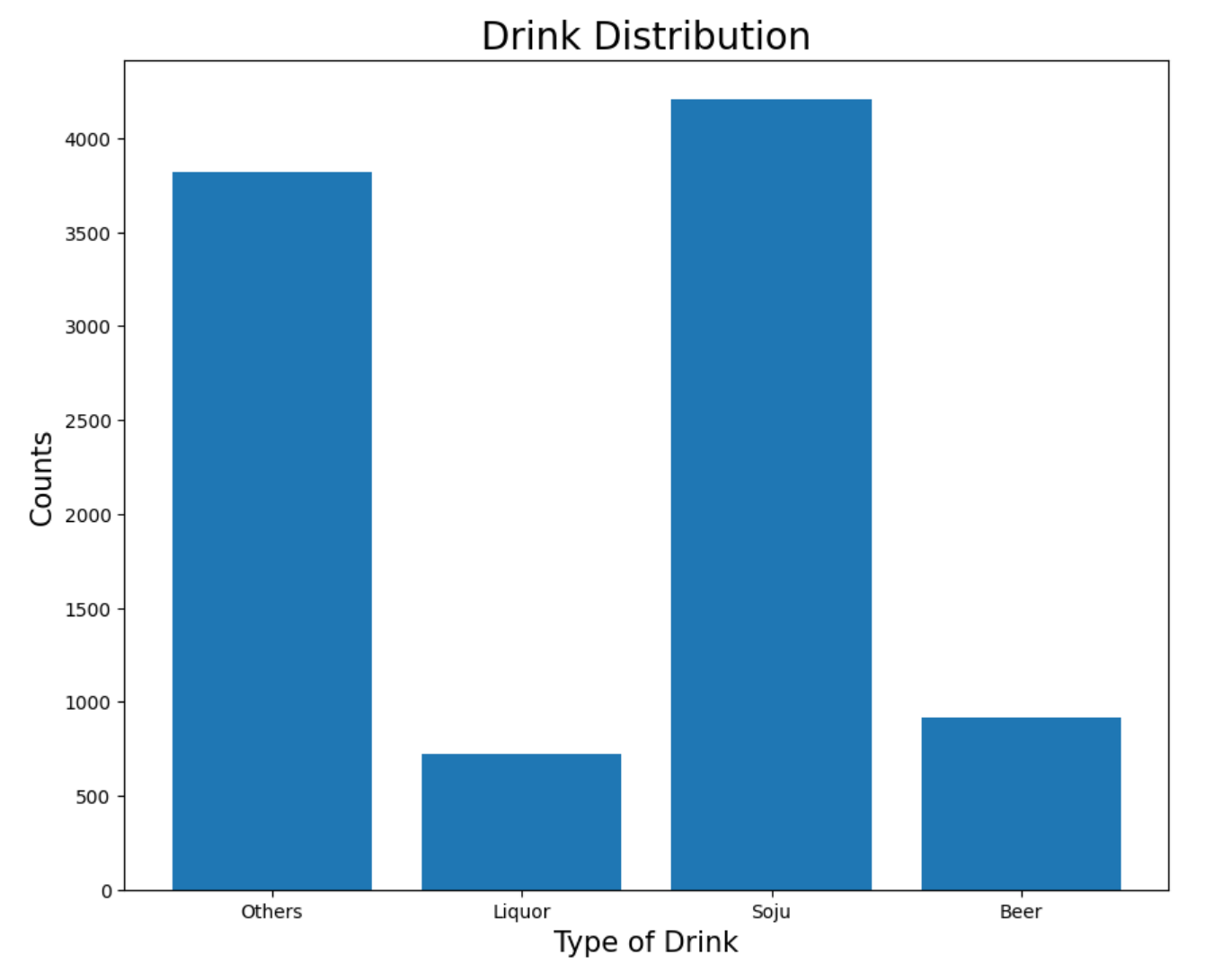
The first EDA focused on cancer and non-cancerous cases and created EDAs. The hypotheses are 1) the older the person, the higher the probability of lung cancer, 2) the more smoking, the higher the probability of lung cancer, 3) the higher the BMI, the higher the probability of lung cancer, and 4) treatment will increase the survival rate.

As a result of drawing the age distribution of cancer patients as a histogram, the proportion of the older group was high(figure 1). Therefore, for hypothesis 1, it can be judged that the higher the age, the higher the incidence of cancer. However, further verification and analysis are needed to confirm the hypothesis.



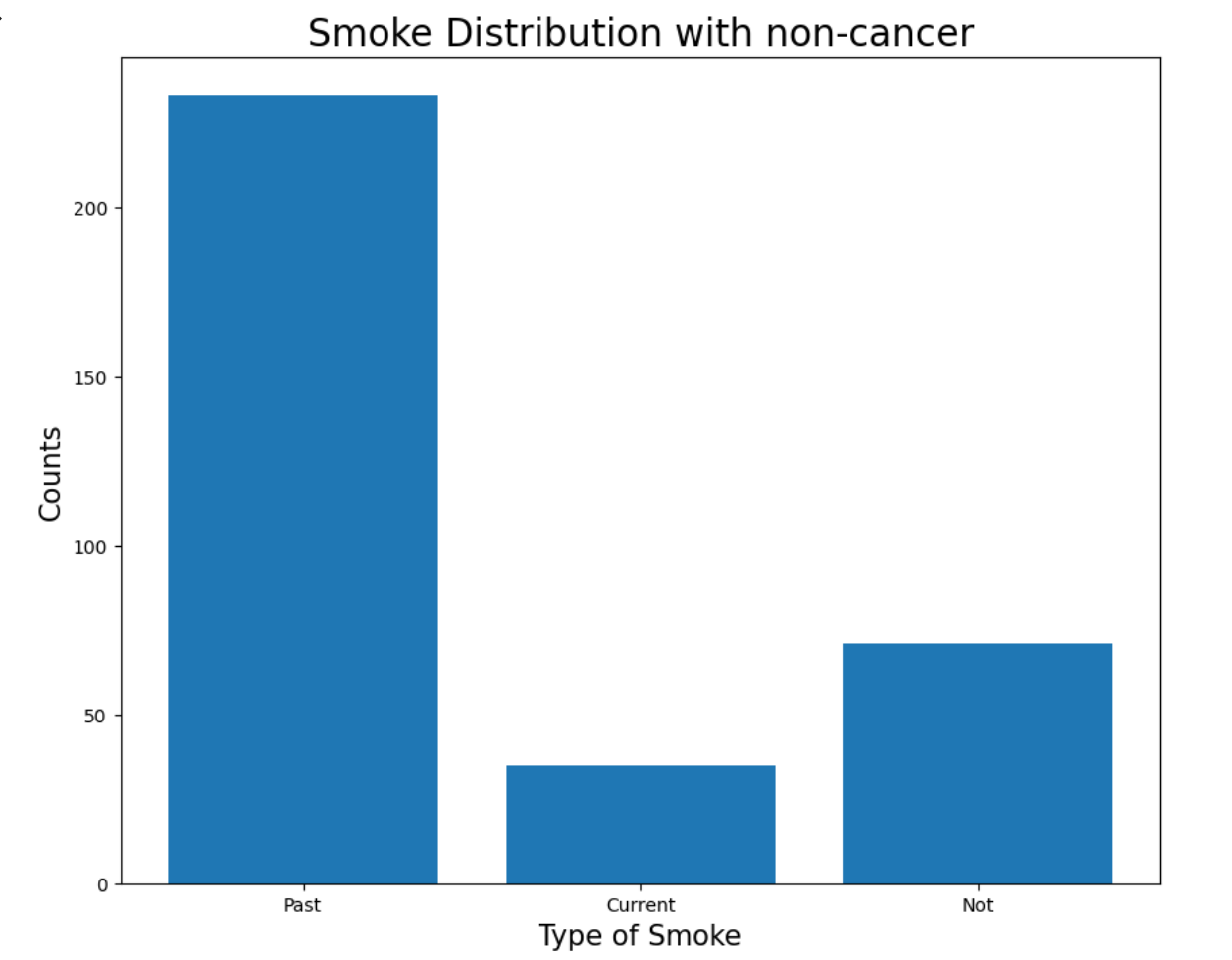
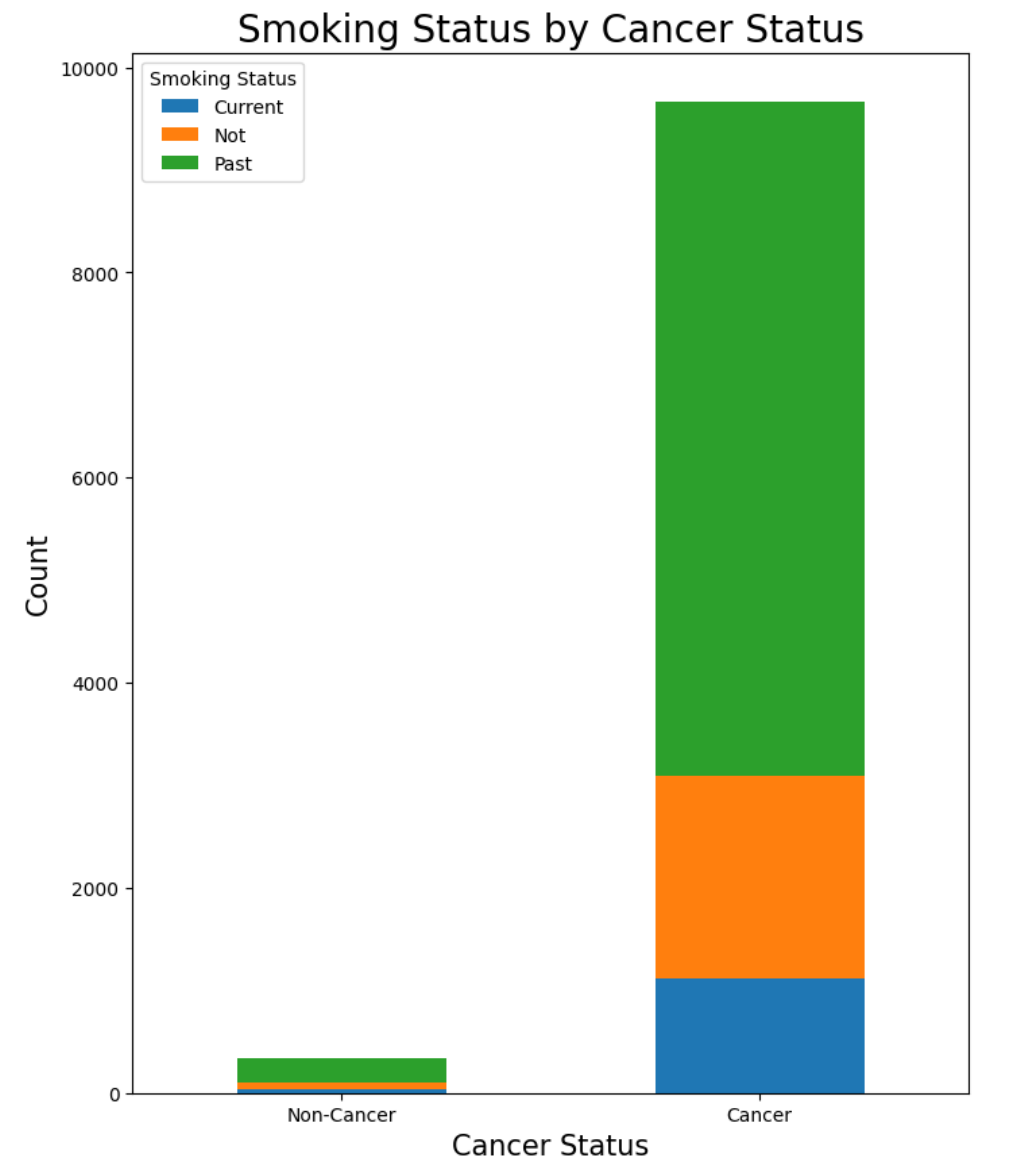
[figure 1]

The correlation between drink and cancer was tried to be examined, but it could not be confirmed because the drink data was represented by the type of drink, not by the presence or intake of the drink. So, only what type of alcohol people with cancer enjoyed a lot could be found out, and as a result, soju and others were high (figure 2).

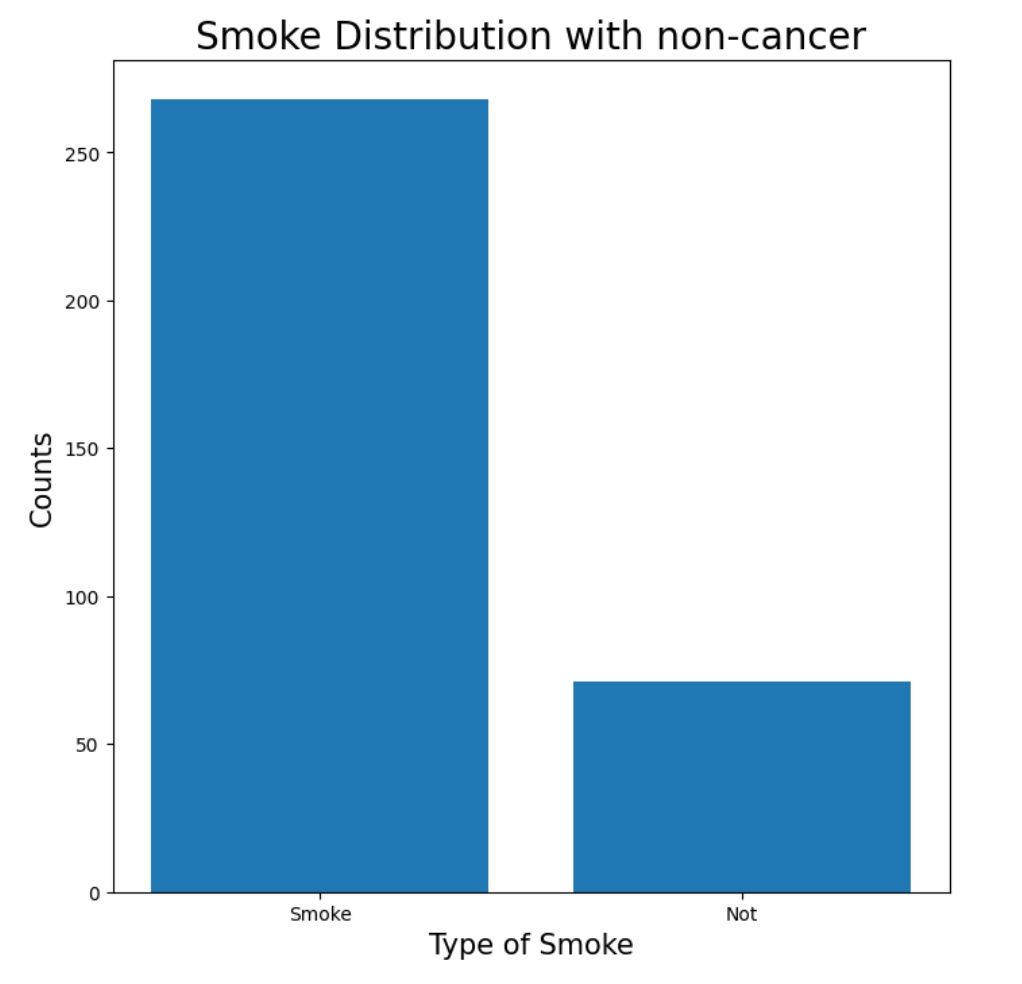
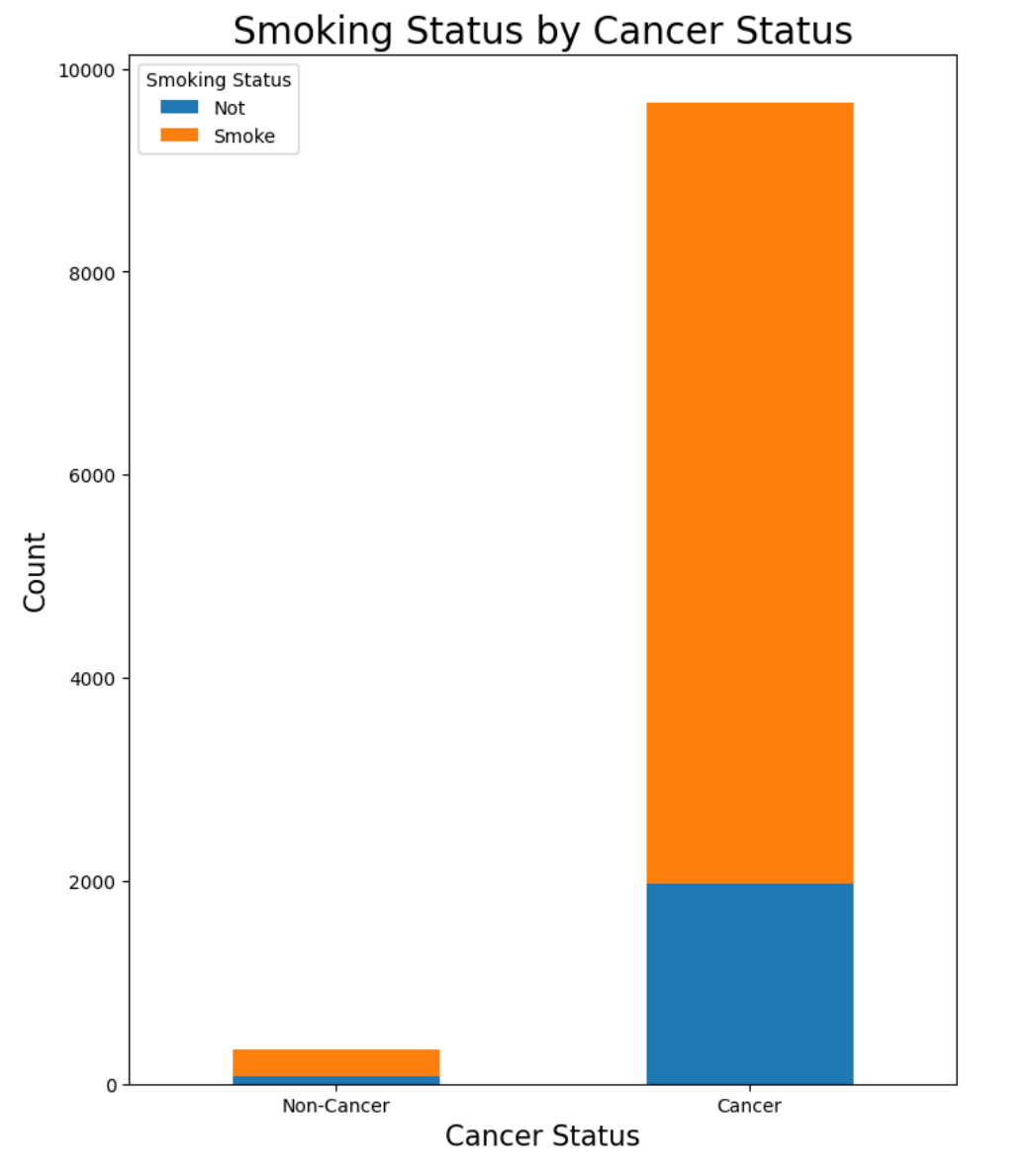


[figure 2]

To find out the relationship between smoking and the onset of cancer, a bar plot of accumulated values was drawn (figure 3). People with cancer were found to be past smoking > non-smoking > current smoking. The difference in data size between those with cancer and those without cancer made it difficult to determine the results of those without cancer, so more bar plots were created with only data from those without cancer (figure 4). As a result, like people with cancer, past smoking > non-smoking > current smoking resulted in results. What this shows is that smoking and the onset of cancer are not related. Even if someone quit smoking six months ago, it can be recorded as smoking in the past. Also, according to an Oxford academic journal article, smoking in the past can also affect his or her lungs, it is classified as not/current/past, but not/smoke. So even if someone has smoked in the past, it is included as smoking. Even when the classification was changed, a cumulative form of bar plot was drawn (figure 5), and another bar plot was created to determine the distribution of people who did not have cancer (figure 6). As a result, the proportion of smoke was high in both people with and without cancer. In conclusion, in Hypothesis 2, it can be said that smoking is not related to the occurrence of cancer.

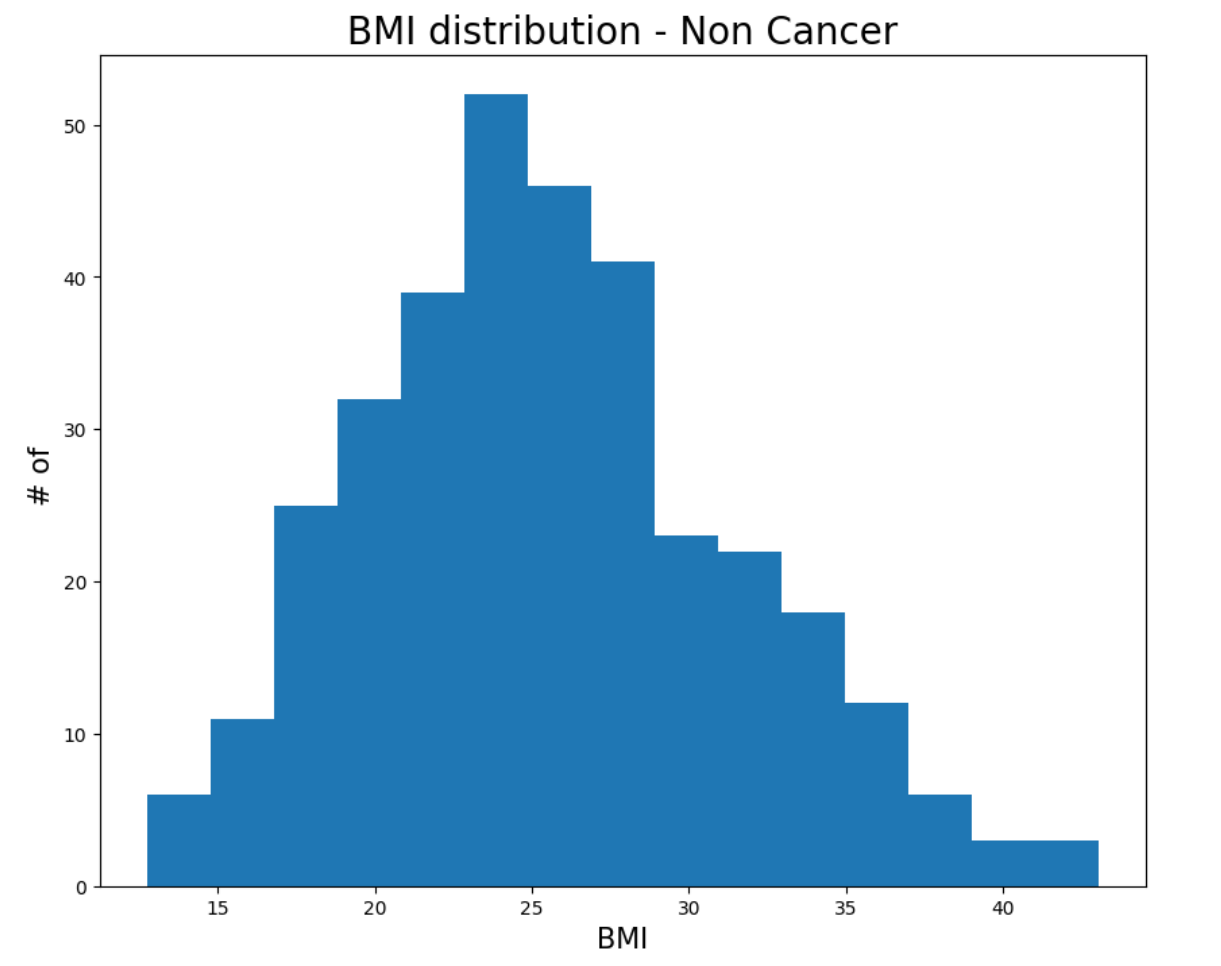
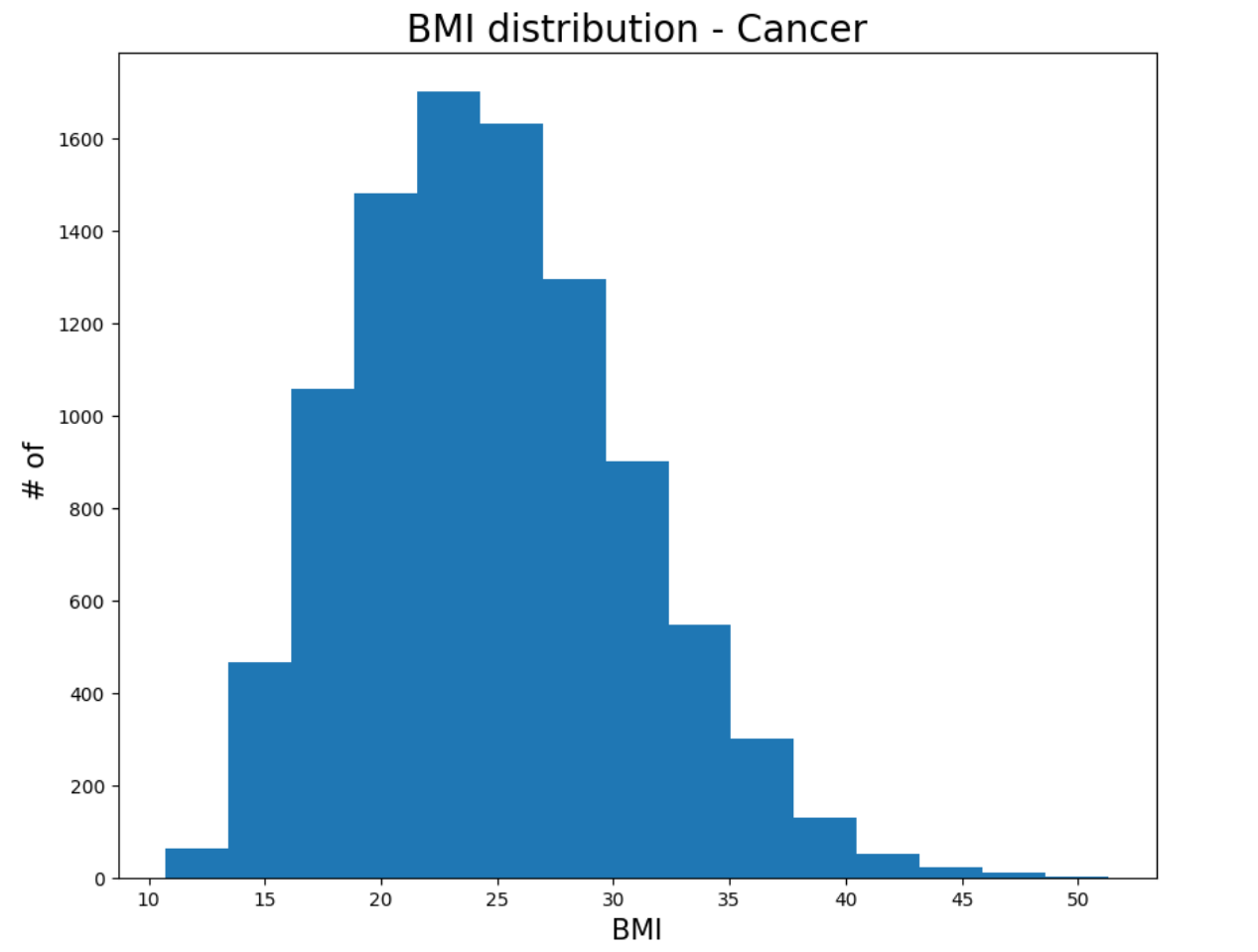


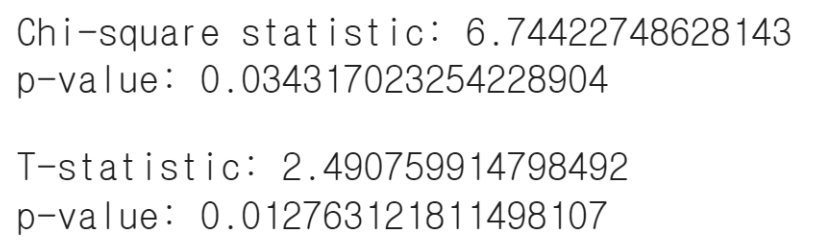
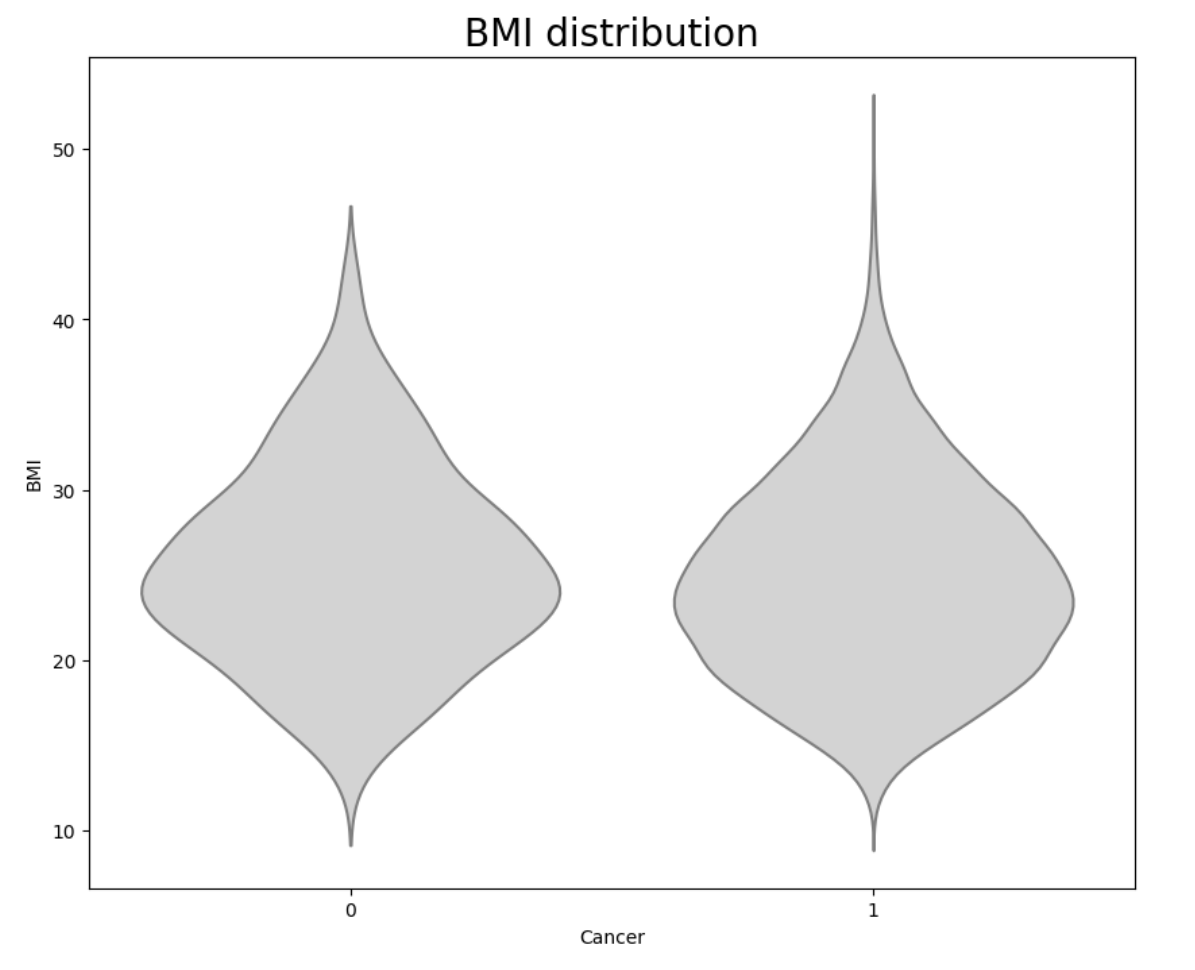
[figure 3] [figure 4]



[figure 5] [figure 6]

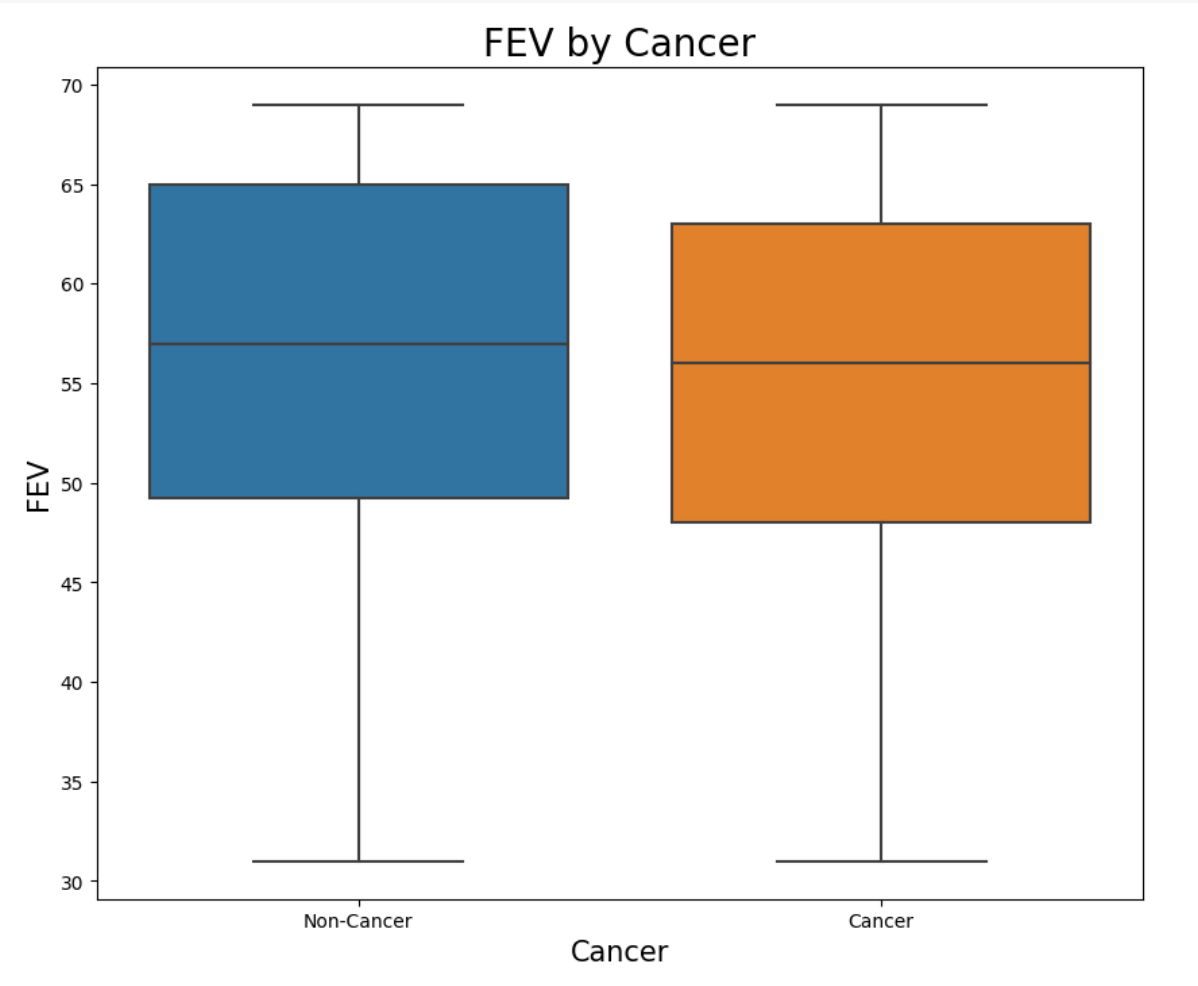
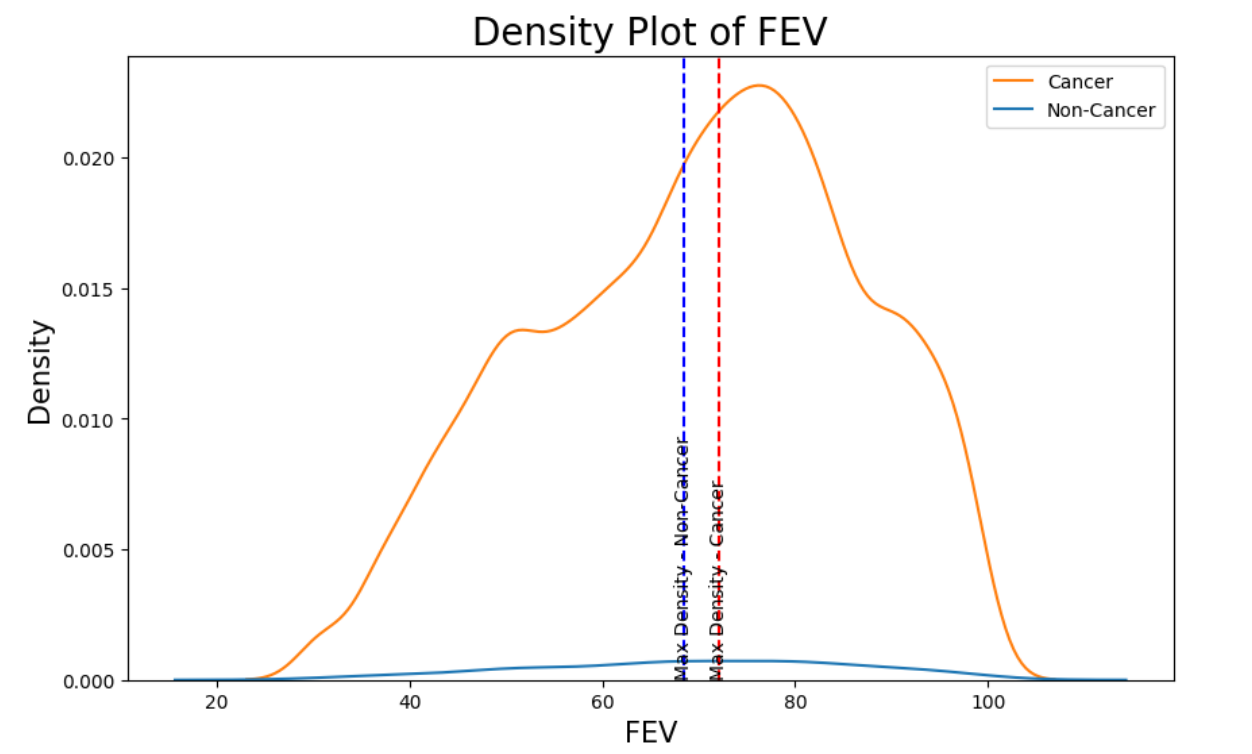
To find out the relationship between BMI(weight(kg)/height(m)\*\*2) and the onset of cancer, BMI histograms and boxplots were drawn according to the presence or absence of cancer. However, since it was difficult to confirm whether it was related or not only with histograms and boxplots, a qui-square test was conducted(figure 7,8,9). The qui-square test is a method of testing relevance (independence) through the difference between expected and observed values. In this test, the null hypothesis is 'no link between BMI and the onset of cancer'. As a result of the test, since p is < .05, the null hypothesis can be rejected and it can be judged that there is a relationship between BMI and lung cancer(figure 10). In addition, since the result of the independent t-test is p < .05, it can be said that there is a significant difference between the BMI of the cancer group and the BMI of the non-cancer group(figure 11). Therefore, there is a significant relationship between BMI and lung cancer in Hypothesis 3, but further analysis is needed to confirm that the hypothesis is true.



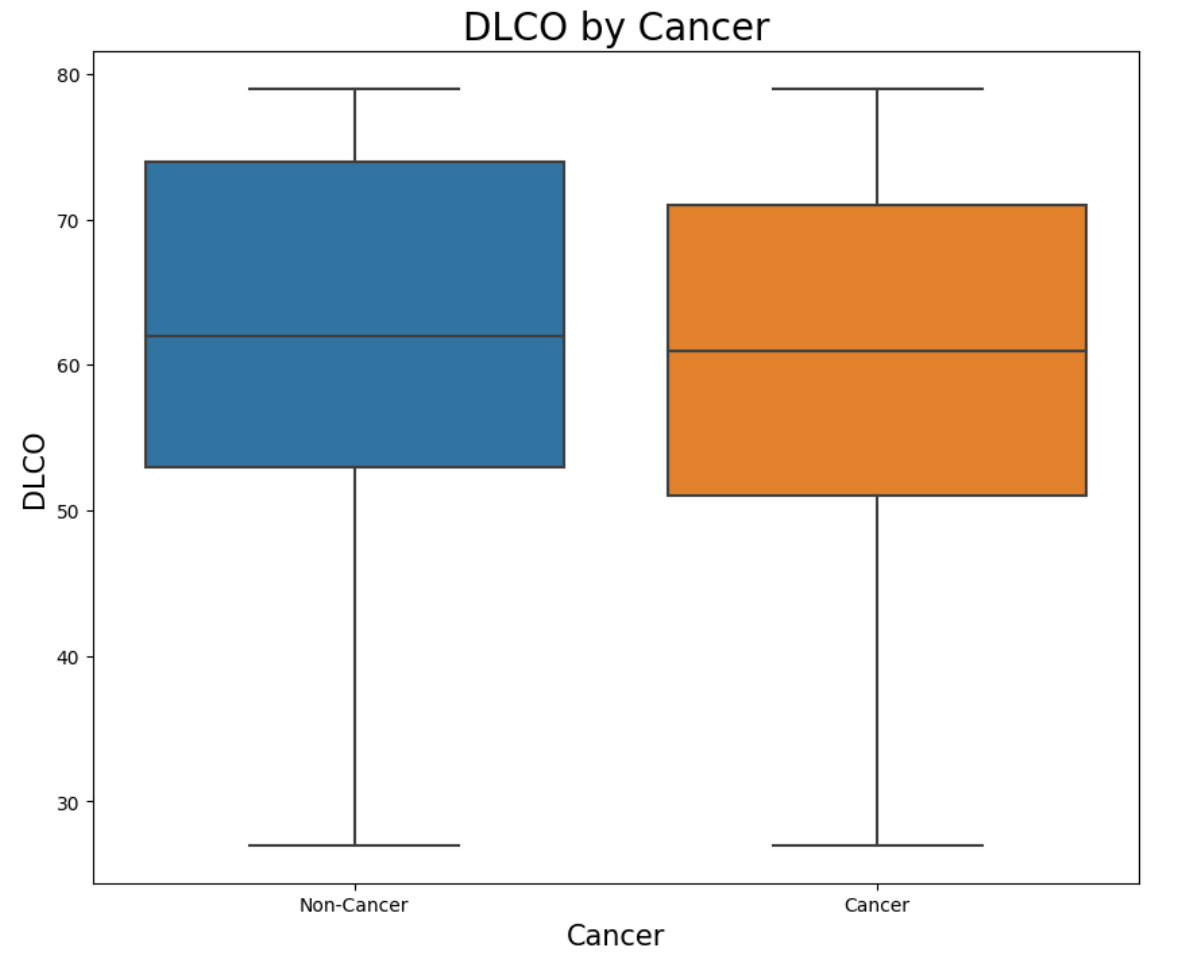
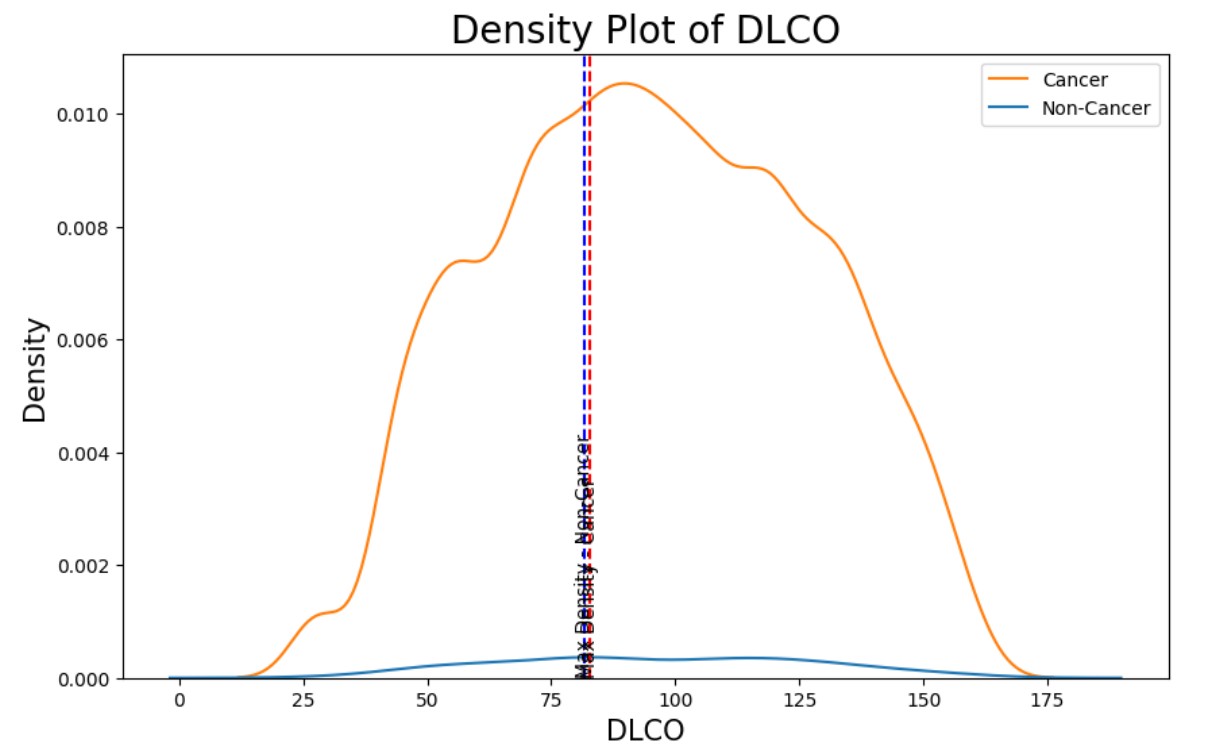
[figure 7] [figure 8] 

[figure 9] [upper: figure 10, lower: figure 11]

The purpose of this analysis was to investigate the relationship between lung function tests and cancer incidence. The first is the FEV test. In this data, the FEV test has a value representing the FEV1/FVC ratio, and the FEV1/FVC is an indicator of bronchial obstruction and is in the normal range of 70% to 80%. If it is less than 70%, it is judged that there is a problem with breathing. When the box plot and KDE plot were drawn using the entire data, rather, the Q2 (median) and the densest point in non-cancer patients were rather lower than in cancer patients(figure 12). The figure is over 70%, but there is confusion in determining the relevance. Therefore, it was decided to check separately only for less than 70%. In this case, the test results of people with cancer were distributed lower in the box plot. Q1 (lower quartile), Q2 (median), and Q3 (upper quartile) all had low values in cancer patients(figure 13). In conclusion, if the FEV is less than 70, the lower the value from 70, the higher the likelihood of cancer patients, but it is difficult to determine whether there is a relationship between cancer and FEV testing. The following is the DLCO test. DLCO is an indicator of how efficiently gas exchange is performed in the lungs, and 80 to 120 is the normal range. If the DLCO is less than 80, lung function is judged to be impaired. When drawing box plots and KDE plots using the entire data, Q2 (medium) and the highest density showed similar values in the two groups(figure 14). In addition, only the case of less than 80, such as FEV, was extracted separately and a box plot was drawn(figure 15). In this case, data from people with cancer were distributed low. In conclusion, if it is less than 80 in DLCO, the lower the value, the more likely it is a cancer patient, but it is difficult to determine whether there is a relationship between DLCO and the onset of cancer.

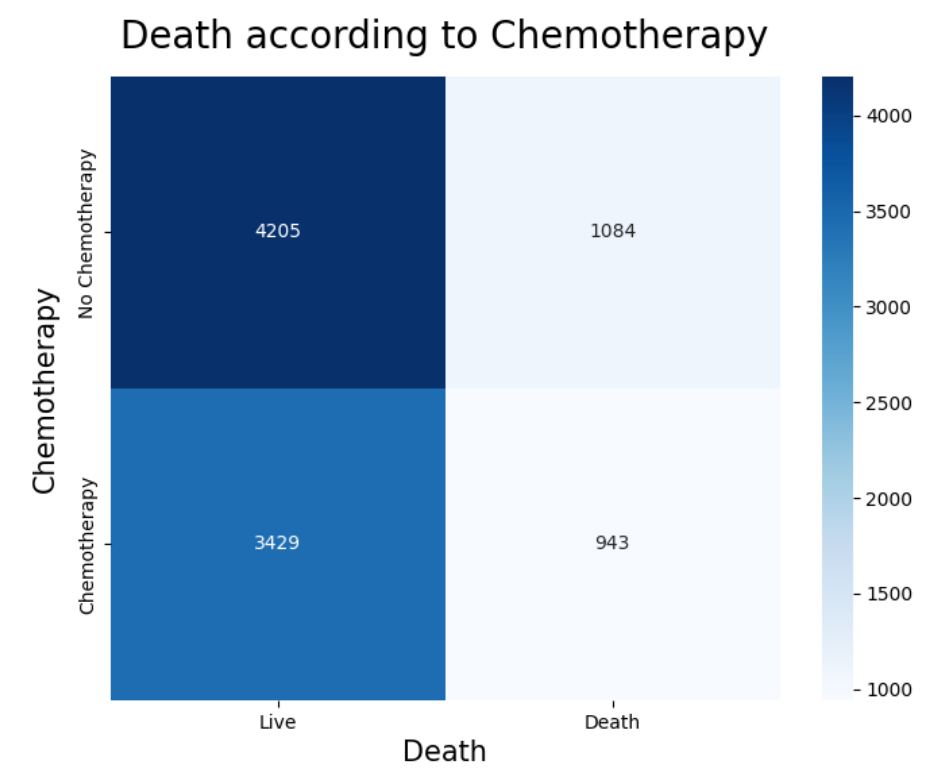
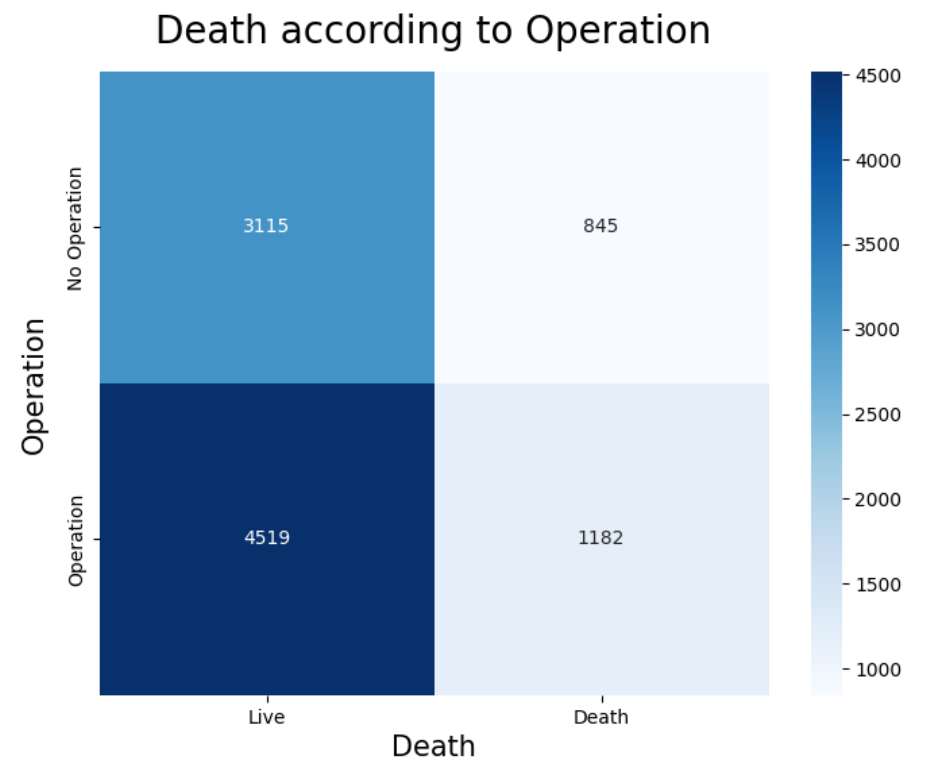


[figure 12] [figure 13]

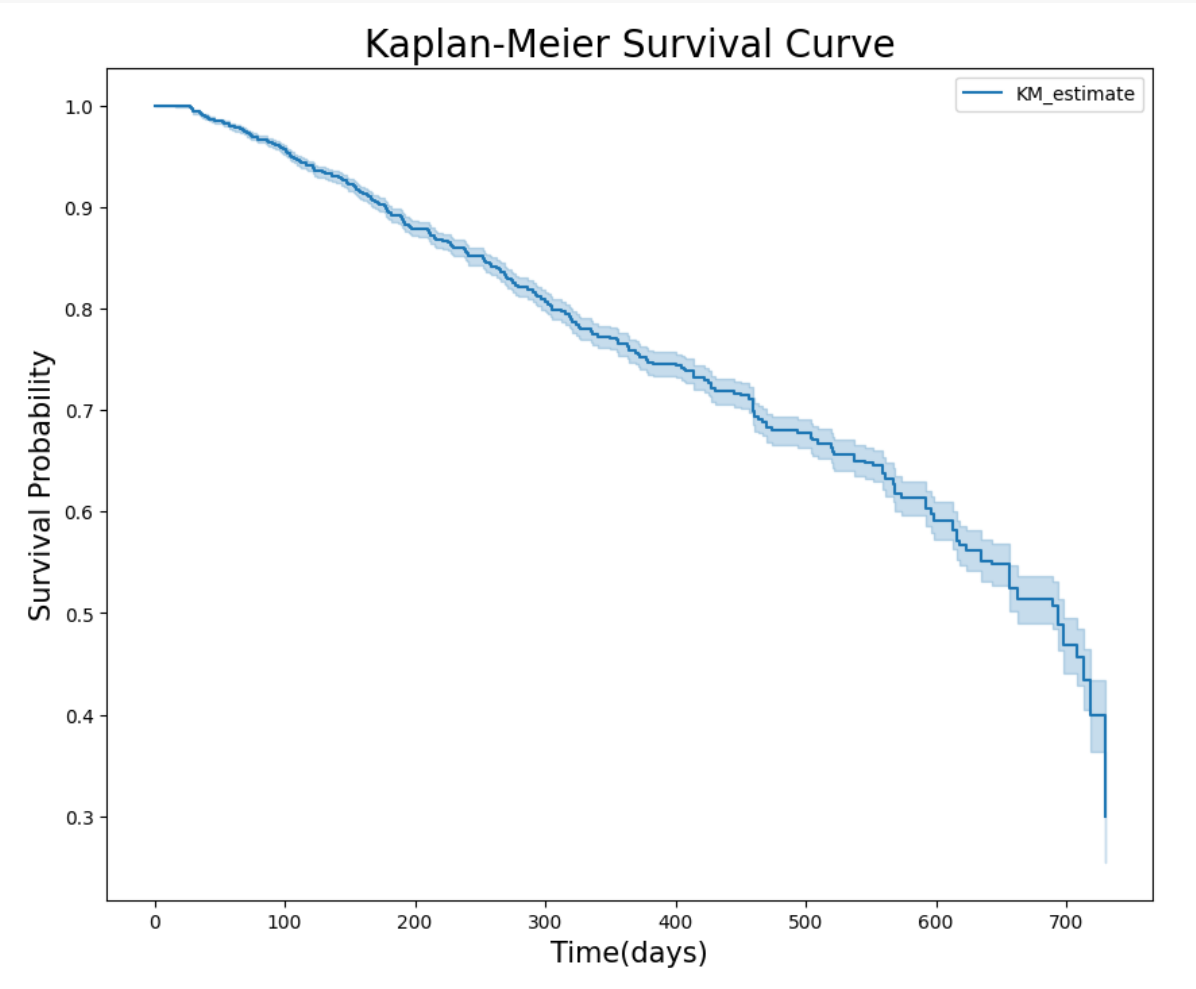
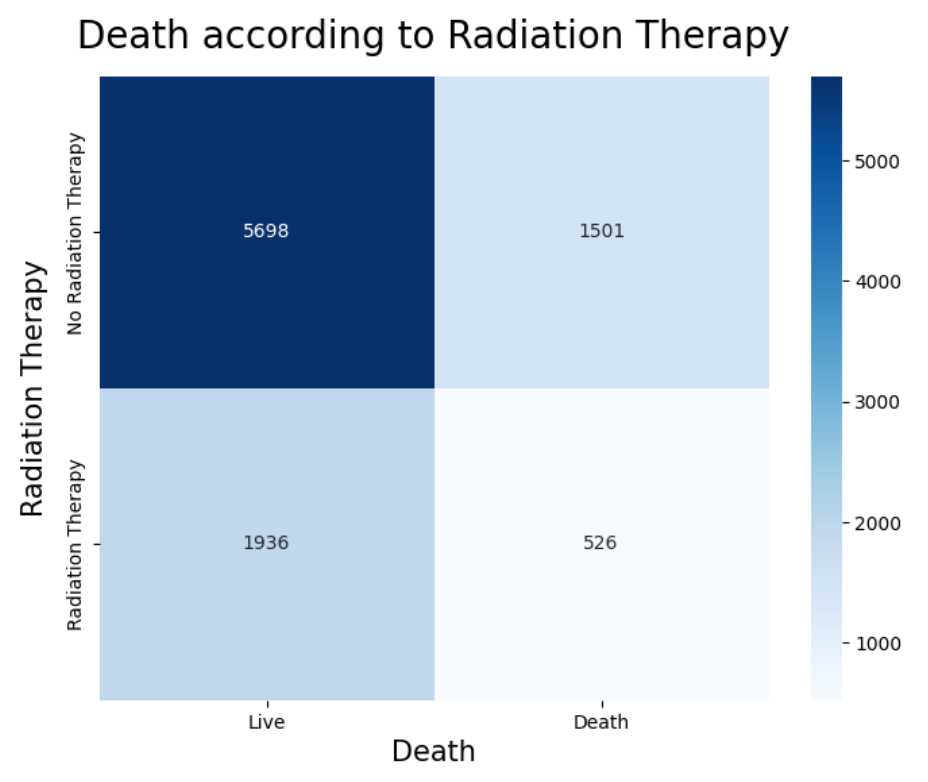


[figure 14] [figure 15]

The association between death and each treatment was investigated. There are three treatments: Operation, Chemotherapy, and Radiation Therapy. Since the treatment and Death data are represented by 0 (false) or 1 (true), a heat map was drawn. And then conditional probabilities were used to determine whether each treatment increased the survival rate. First is the case of operation(figure 16). The probability of survival after operation was 79.2%, and the probability of survival without operation was 78.6%, which was slightly higher in the case of surgery. Next is the case of chemotherapy. The probability of survival after receiving chemotherapy was 78.5%, and the probability of surviving without chemotherapy was 79.5%, which was slightly higher when not receiving chemotherapy. Finally, it is the case of radiation therapy(figure 18). The probability of survival after receiving radiation therapy was 88.1%, and the probability of survival without radiation therapy was 79.1%, and the survival rate was higher when receiving radiation therapy. In addition, the survival probability curve was drawn using the survival time of the deceased(figure 19). Due to the narrow range of opaque areas, uncertainty is low, and over time, it is possible to see what the survival rate is.



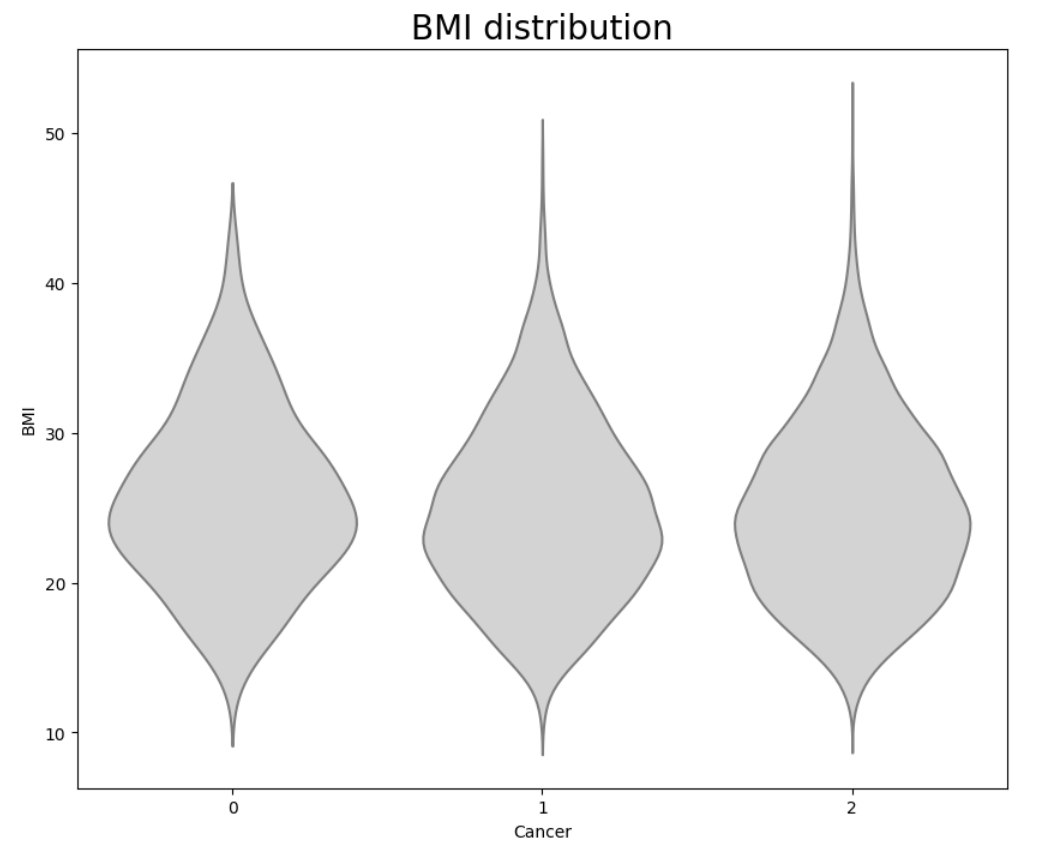
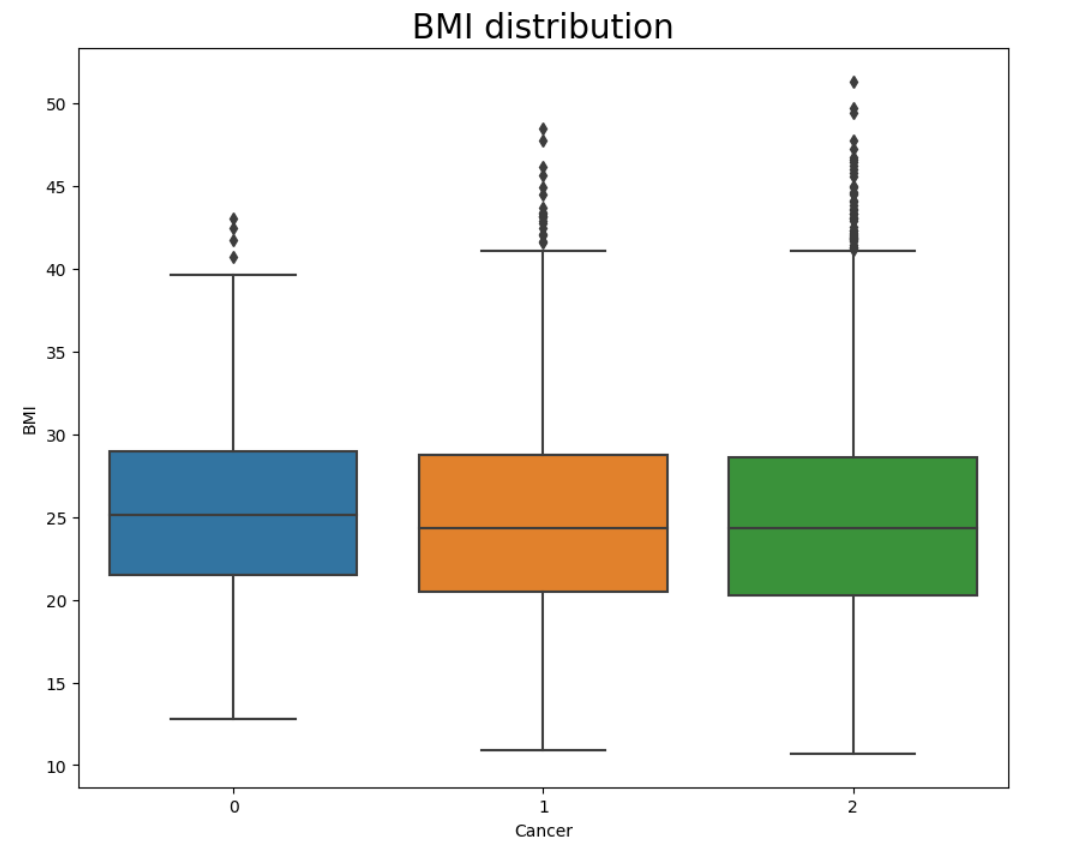
[figure 16] [figure 17]



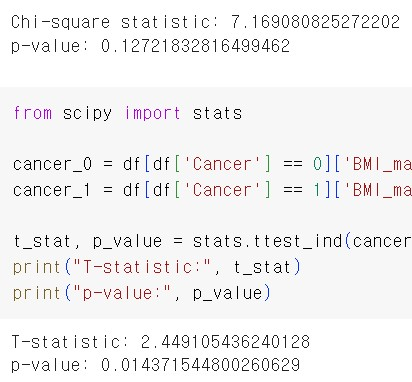
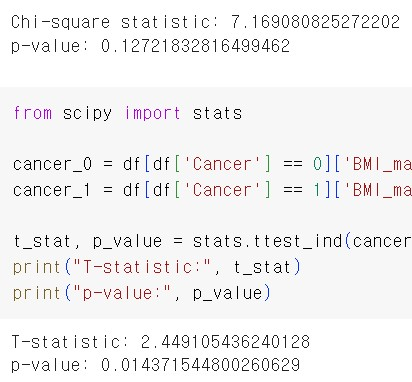
[figure 18] [figure 19]

The second EDA distinguished cancer patients in more detail, indicating whether they have small-sized cancer or not. First of all, I will summarize the features that have similar conclusions as the first EDA. First, age and cancer may be related because both non-small-sized and small-sized cancers had more people with cancer as they were older. Second, all three groups (non-cancer, small-sized cancer, and non-small-sized cancer) were high in the order of past smoking > non-smoking > current smoking. In addition, as in the first EDA, both past and current smoking were classified as smoking and divided into smoking or non-smoking. All three groups had a high rate of smoking, making it difficult to confirm the significance of smoking and cancer. Third, both the FEV test and the DLCO test were difficult to confirm the test results and the significance of the cancer when considering the entire data. However, considering the data below 70 and 80, which are the criteria for lung function disorders, respectively, the distribution of cancer cases was lower than that of non-cancer cases. However, in both tests, the distribution when it was a non-small-sized cancer was not lower than when it was a small-sized cancer.

It explains the part that appears differently from the first EDA. First, it's BMI. The first EDA confirmed the significance between BMI and cancer incidence. Also in this EDA, it was difficult to confirm the relationship only with the box plot and the violin plot, so chi-square test and independent t-test were conducted (figure 20, 21). Since the result of the chi-square test is p >.05, it cannot be ignored that there is no relationship between cancer types(non-cancer, small-sized cancer, non-small-sized cancer) and BMI (figure 22). On the other hand, the independent t-test shows that there is a statistically significant difference since p is lower than .05 (figure 23). In conclusion, it means that there are differences between groups, but not related. Therefore, when combined with the first EDA, it can be seen that BMI is associated with cancer and non-cancer classification, but not with cancer size.

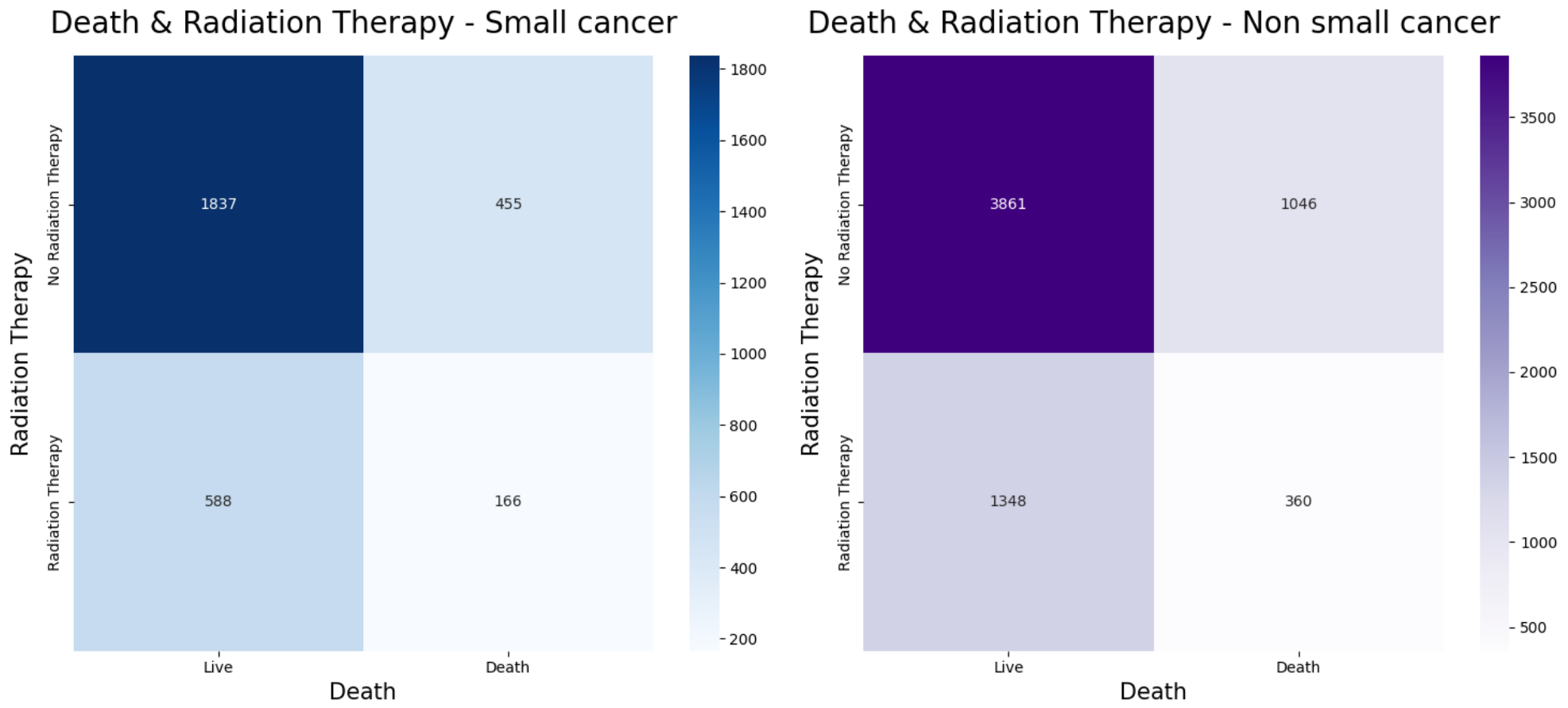


[figure 20] [figure 21]

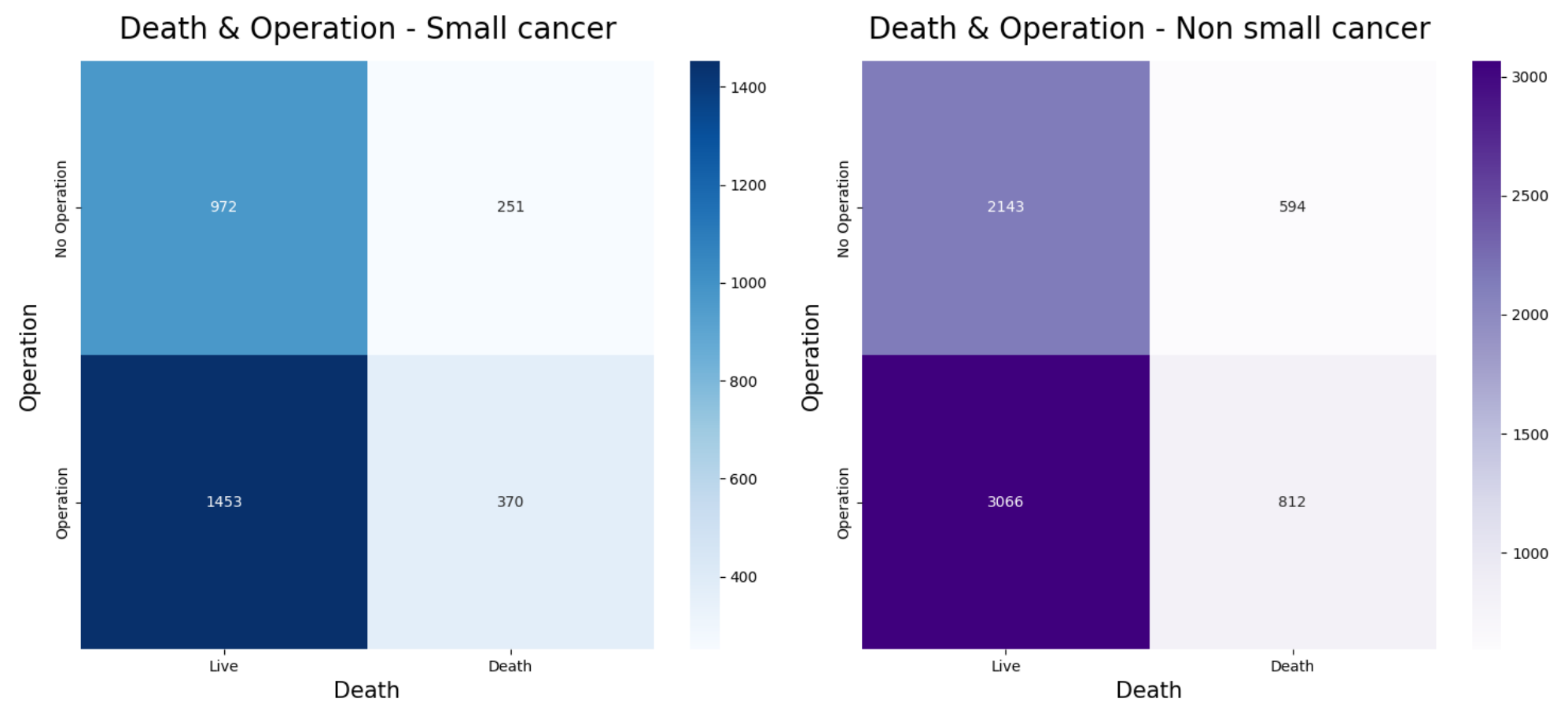


[figure 22] [figure 22]

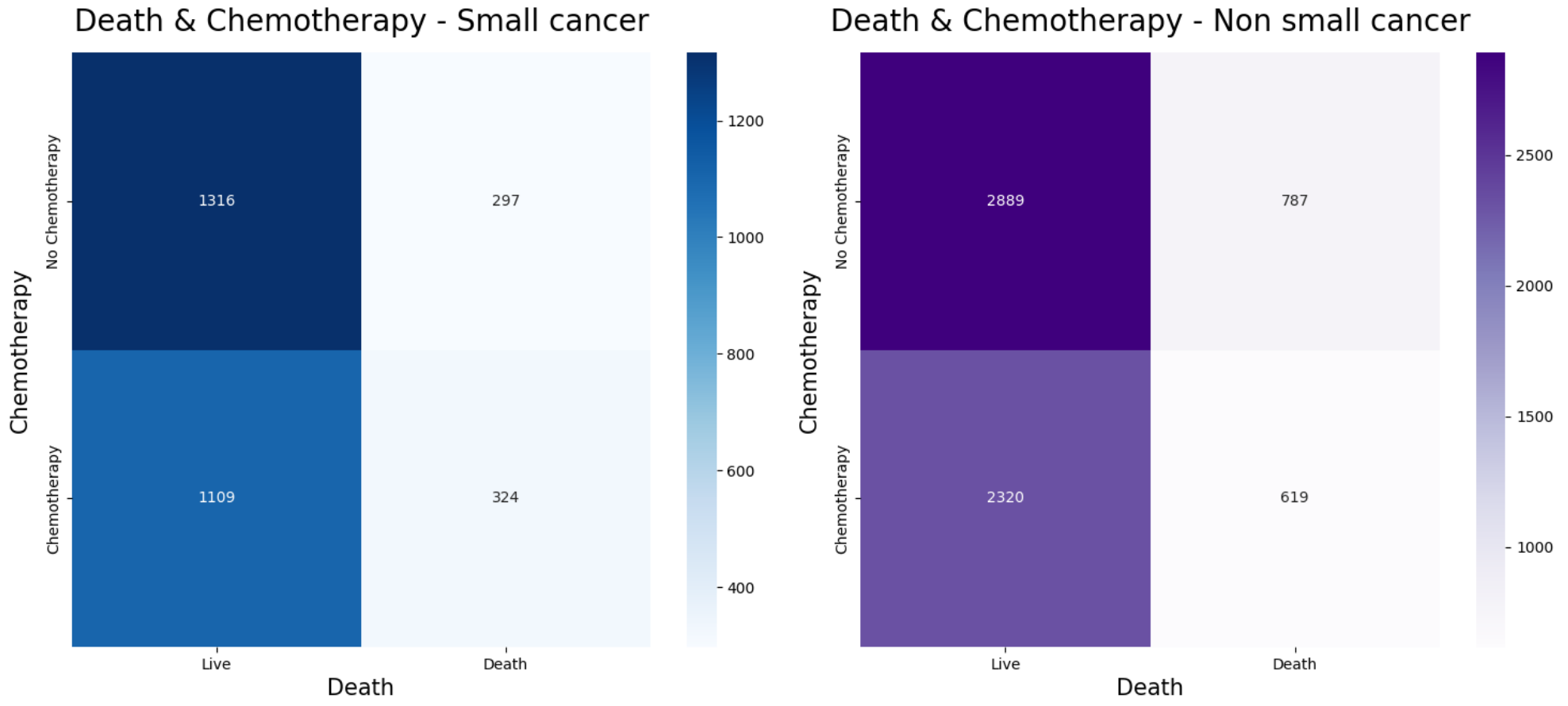
It was also examined whether the treatment method increased the survival rate by dividing it into a small-sized cancer and a non-small-sized cancer. After drawing the heatmap, the survival rate when treatment was received and when treatment was not received was calculated for each case using the conditional probability. In the case of small-sized cancers, the probability of survival when undergoing operation was 79.7%, and the probability of survival without operation was 79.4% (figure 23). When chemotherapy was received, the probability of survival was 77.3%, and the probability of survival without chemotherapy was 81.5% (figure 24). In addition, when receiving radiation therapy, the probability of survival was 77.9%, and the probability of survival without radiation therapy was 80.1% (figure 25). Therefore, even though there was only a small difference in the small-sized cancer, it cannot be said that it helped increase the survival rate other than surgery. In the case of non-small-sized cancers, the probability of survival when undergoing operation was 79.0%, and the probability of survival without operation was 78.2% (figure 26). When chemotherapy was received, the probability of survival was 78.9%, and the probability of survival without chemotherapy was 78.6%. (figure 27). In addition, when receiving radiation therapy, the probability of survival was 78.9%, and the probability of survival without radiation therapy was 78.6% (figure 28). For non-small-sized cancer, all treatment methods were effective in increasing the survival rate even a little.



[figure 23] [figure 26]

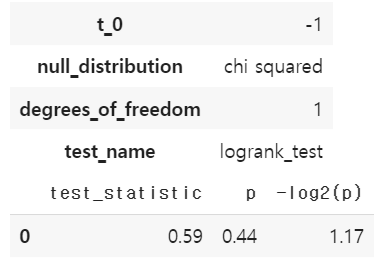
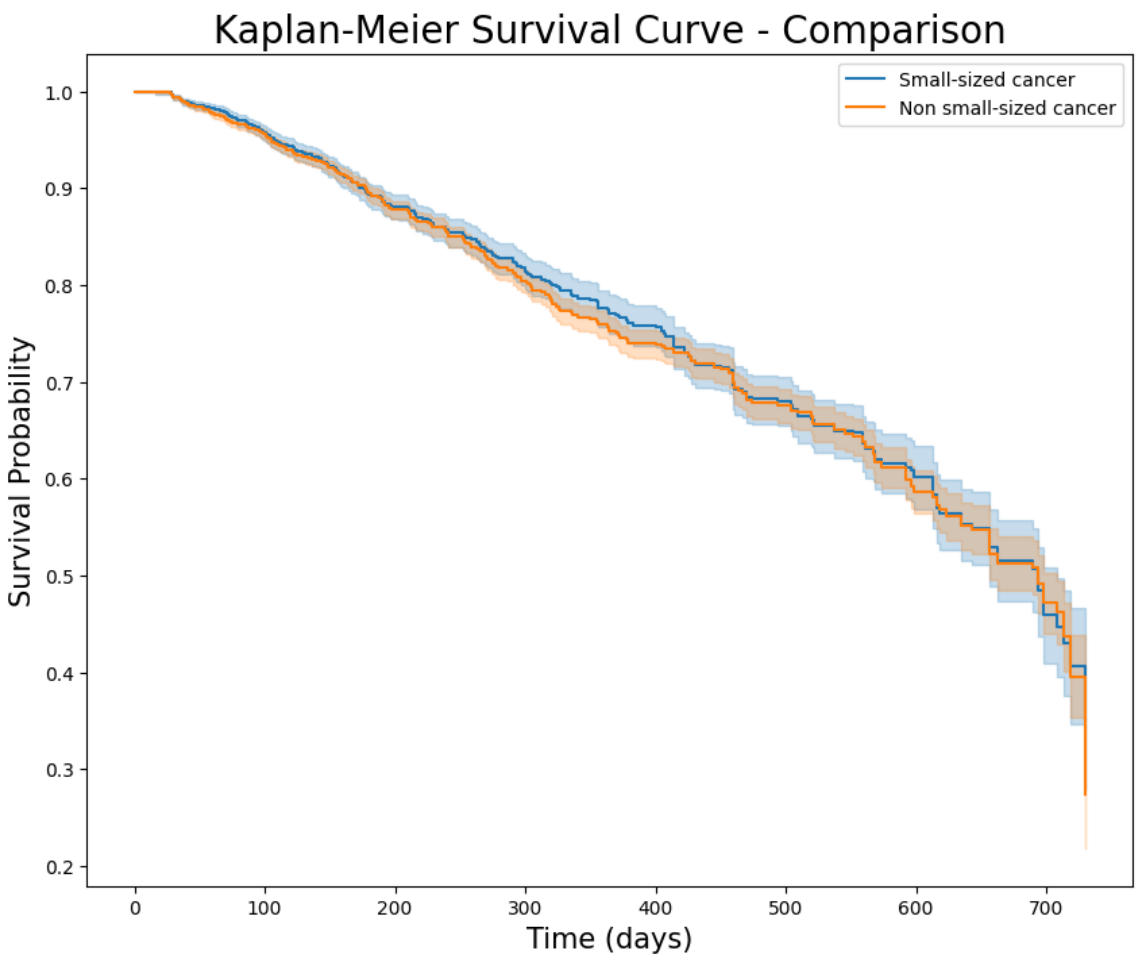


[figure 24] [figure 27]



[figure 25] [figure 28]

Finally, the survival curves of the non-small-sized canceler and the small-sized canceler were drawn. Previously, it was hypothesized that the probability of survival would be higher in the case of a small-sized cancer than in the case of a non-small-sized cancer. However, in Kaplan-Meier survival analysis, the two curves almost overlapped (figure 29). Since this means that the survival rate is similar without a significant difference, the difference in survival rate was statistically evaluated through the log-rank test. The log-rank test is a statistical test method to see if there is a difference in survival rates between groups by comparing survival time data. Since the test result was p > 0.05, the null hypothesis that there is no significant difference in survival rates between the two groups cannot be rejected (figure 30). Therefore, it can be concluded that the size of the cancer is not related to the survival rate.



[figure 29] [figure 30]

[Reference]