CSE 487 Assignment 1

6/3/25

Group Members: Jacob Szczudlik UBIT- jacobszc
Amadeo Vilbrun UBIT- amadeovi
Emrana Begum UBIT- emranabe
Troy Richardson UBIT - tmr32

Group contributions:

Jacob Szczudlik: Found dataset, wrote part one problem statement and contributed to EDA and graph work.

Emrana Begum: Contributed to multiple sections on EDA section, did graph work and contributed to data cleaning.

Amadeo Vilbrun: Did much of the work in the data cleaning section and contributed to EDA and graph work.

Troy Richardson: Did much of the work in finalizing the EDA and graph work.

An exploration of Lung Cancer Corollaries

1.) **Problem Statement**: Lung cancer is one of the leading causes of deaths worldwide as is the number 1 cancer related death. It's more deadly than most other forms of cancer combined. We've heard for a long time about the correlations between bad health habits and the formation of lung cancer, most notably smoking. However there are many other factors that need to be considered all related to physical health such as bmi, cholesterol levels and much more. Understanding how all of these factors contribute to lung cancer is the problem we're looking into.

a.) Cancer has been a huge point of discussion for many years now and though we understand much more about it, and treatment has improved, there is still much to learn. The main point of interest for the general public is prevention. We're not doctors, we just need to know the best course of action to give us the highest probability of cancer prevention. The objective of this data analysis is to explore some of the more likely causes of lung cancer, and see how they relate to real cancer patients to get an idea of the leading causes.

b.) The dataset we have found is such a good exploration into a huge number of people around the world who have lung cancer of varying degrees, and a look into their health and daily habits. It draws from so many different important data points, and over a huge number of people, that it gives a really good insight into the causes of lung cancer. This is obviously hugely important for producing good and accurate studies and informative literature on lung cancer prevention. It cannot be overstated how valuable data like that is for everyone.

2.) DataSources:

https://www.kaggle.com/datasets/amankumar094/lung-cancer-dataset

3.) Data Cleaning/Processing

a) Drop Duplicates

We checked for and removed duplicate rows using the function drop_duplicates. There were no changes to the dataset; it remained at (890000, 17), indicating that there were no duplicates. Although there were no duplicates, verifying data uniqueness is important before processing data.

b) Filter for Most Recent Records (2020-2025)

We filtered the dataset to only include rows from the years 2020 to 2025. We did this because we wanted to focus on recent data to keep the analysis relevant.

Health trends can change over time, so focusing on 2020 to 2025 is best for understanding what is happening right now.

c) Normalize Text Columns

We used .str.lower() to standardize the formatting of text. Sometimes the same value can be written in multiple ways (for example, “Male” versus “male”). Converting everything to lowercase keeps things consistent and makes analysis easier by avoiding confusion and potential errors.

d) Convert Yes/No to 1/0

We changed all of the “Yes” and “No” values to “1” and “0”. A lot of analysis tools and machine learning models work better with numerical values. Converting Yes/No to 1/0 makes the data easier to use and allows us to run calculations or build models without any issues.

e) Check and Handle Missing Values

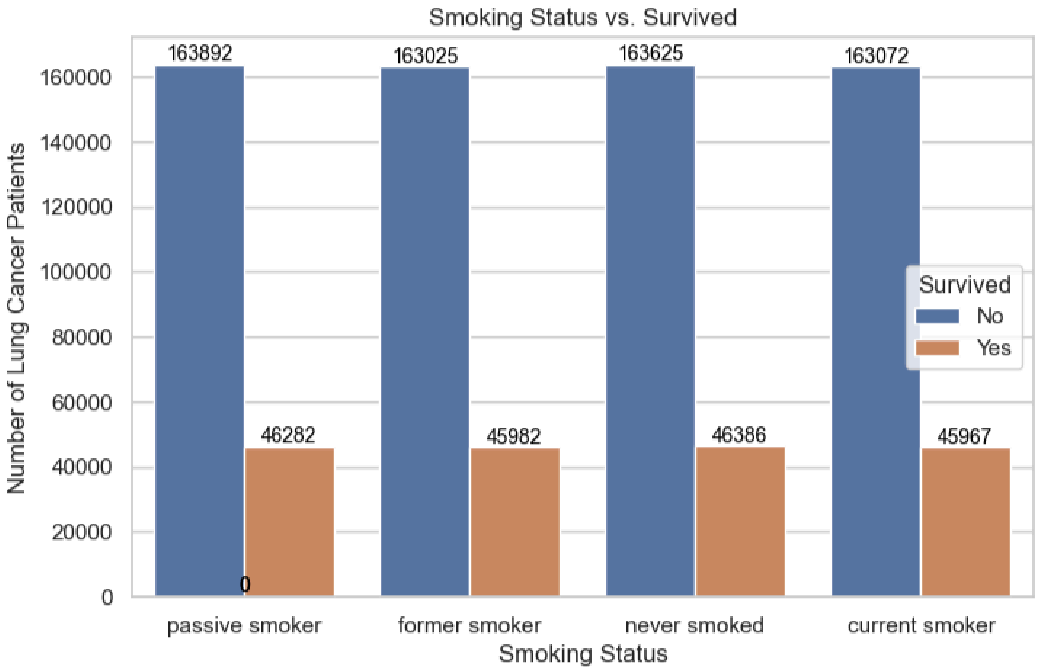
We used isnull().sum() to search for missing values in the dataset. Although we didn’t find missing values, it is still important to check for them. Missing data can throw off results or even stop certain analysis methods from working.

f) Fix Data Types (Dates)

We converted the date columns to datetime format using pd.to_datetime(). If the dates are stored as plain text, we won’t be able to do useful things like filtering or sorting by date. This also makes it easier to work with time-based data.

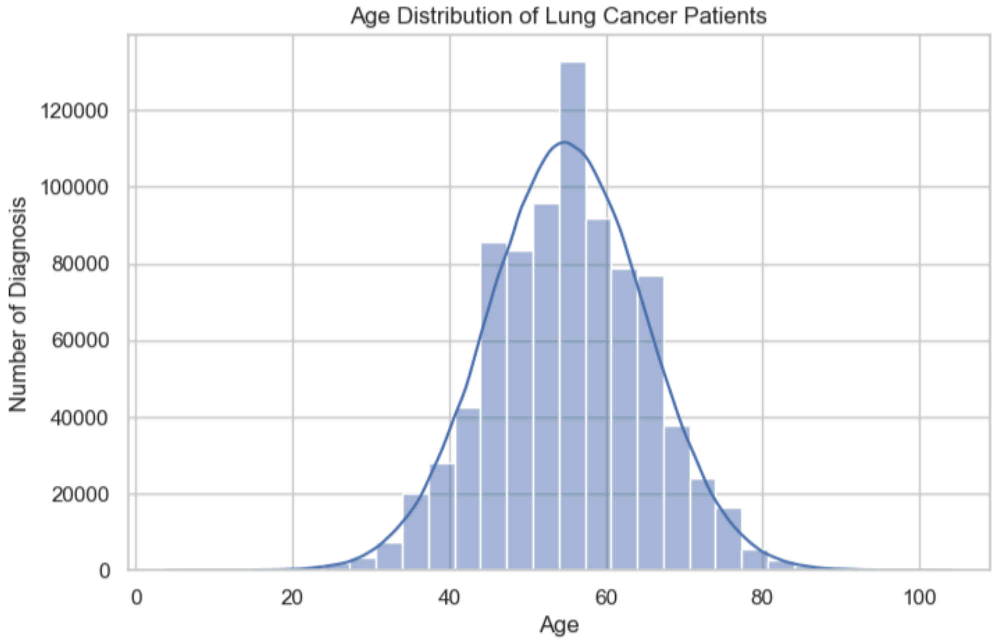
4.) Exploratory Data Analysis (EDA)

a. **Bar Graph - Smoking Status vs Survived**



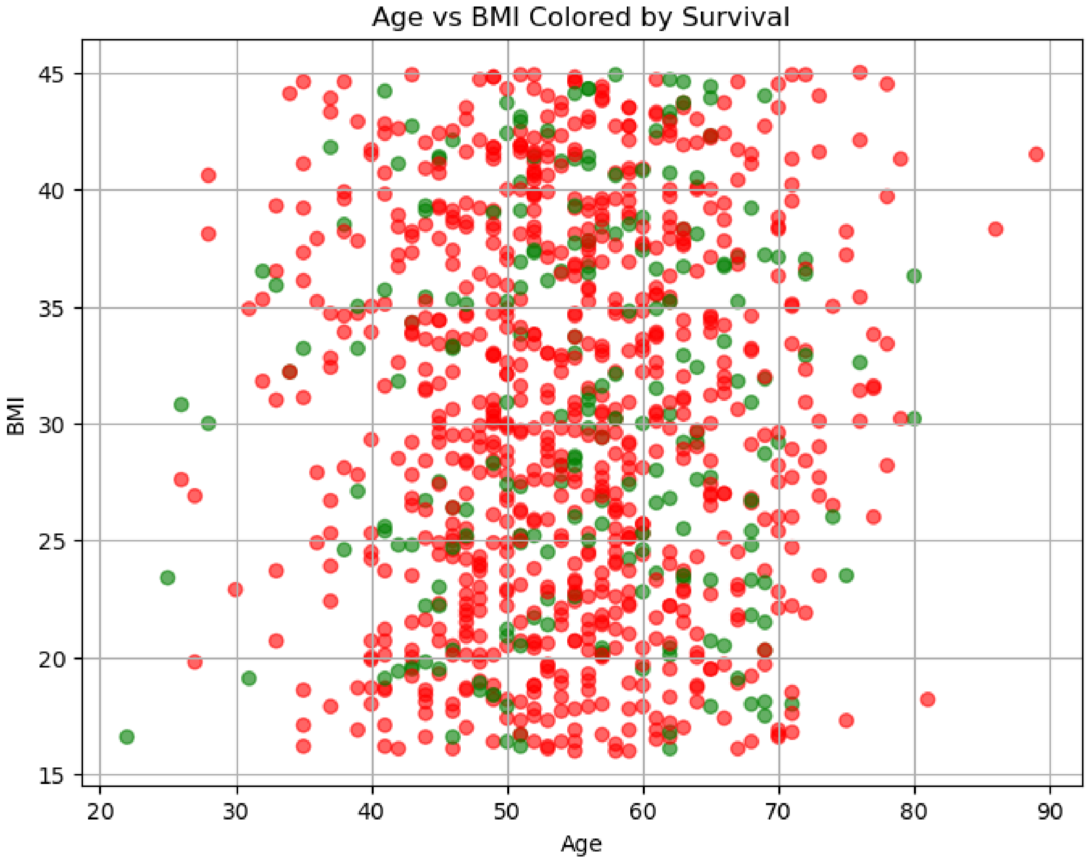
There are four smoking status categories in the dataset: passive smoker, former smoker, current smoker, and never smoked. A bar chart was used to visualize how many individuals survived versus how many did not in each category. The chart revealed that for each smoking group, roughly 163,000 did not survive and around 46,000 did survive. The proportions are very similar across all smoking statuses, resulting in a survival rate of approximately 22% regardless of whether someone smoked or not. This suggests that smoking status does not impact the survival rate of a lung cancer patient, even though it is widely known to be a risk factor in getting lung cancer.

b. **Histogram - Age Distributions**



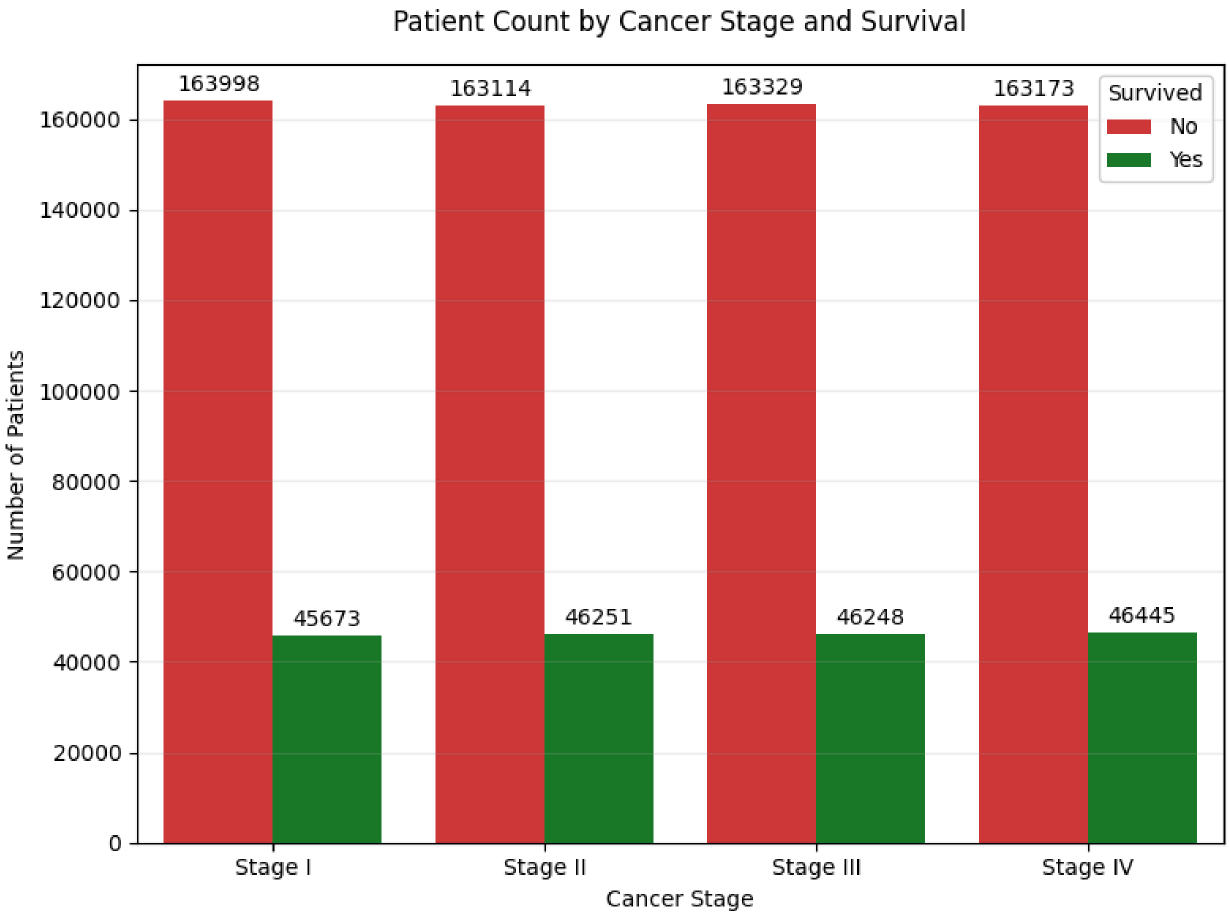
A histogram was used to display the age distribution at the time of lung cancer diagnosis. Most patients were diagnosed between the ages of 45 to 65. Cases under the age of 35 and over 85 were less common, and even more rare among younger individuals under 20.

c. **Scatter Plot - Age and BMI as they relate to survival**



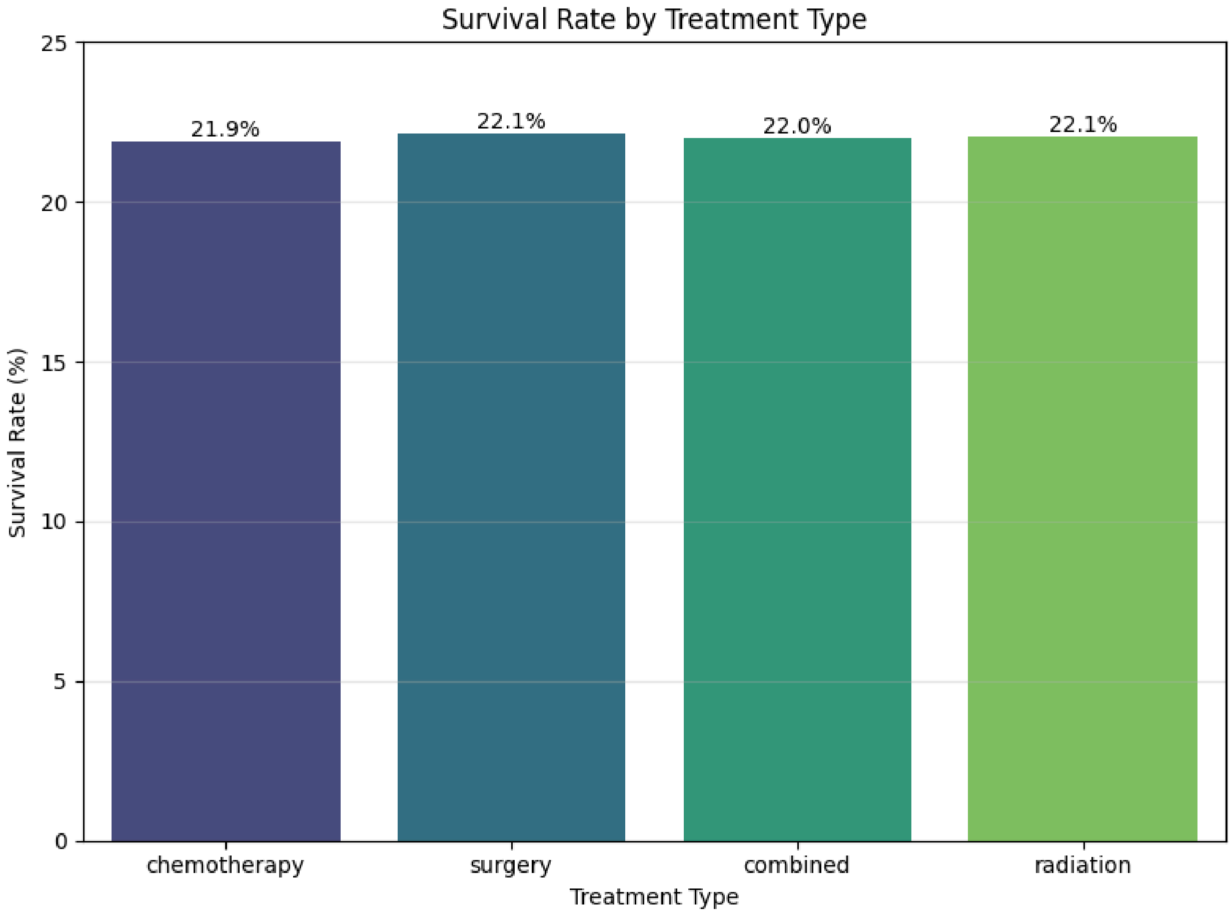
We wanted to take a look at some other factors in the dataset and how they might relate to the chance of survival in lung cancer patients. In this case we see a scatter plat showing the relationship between age and bmi and how that relates to survival. To make the visualization more digestible, I only took a small sample of the dataset so it didn't look so clumped, and also used green to signify survived and red to signify non surviving. This plot shows a pretty even distribution across all ranges. The death rates are certainly higher, but there's no particular age range or bmi that shows significant variation.

d. **Bar Graph - Cancer Stage and Survival**



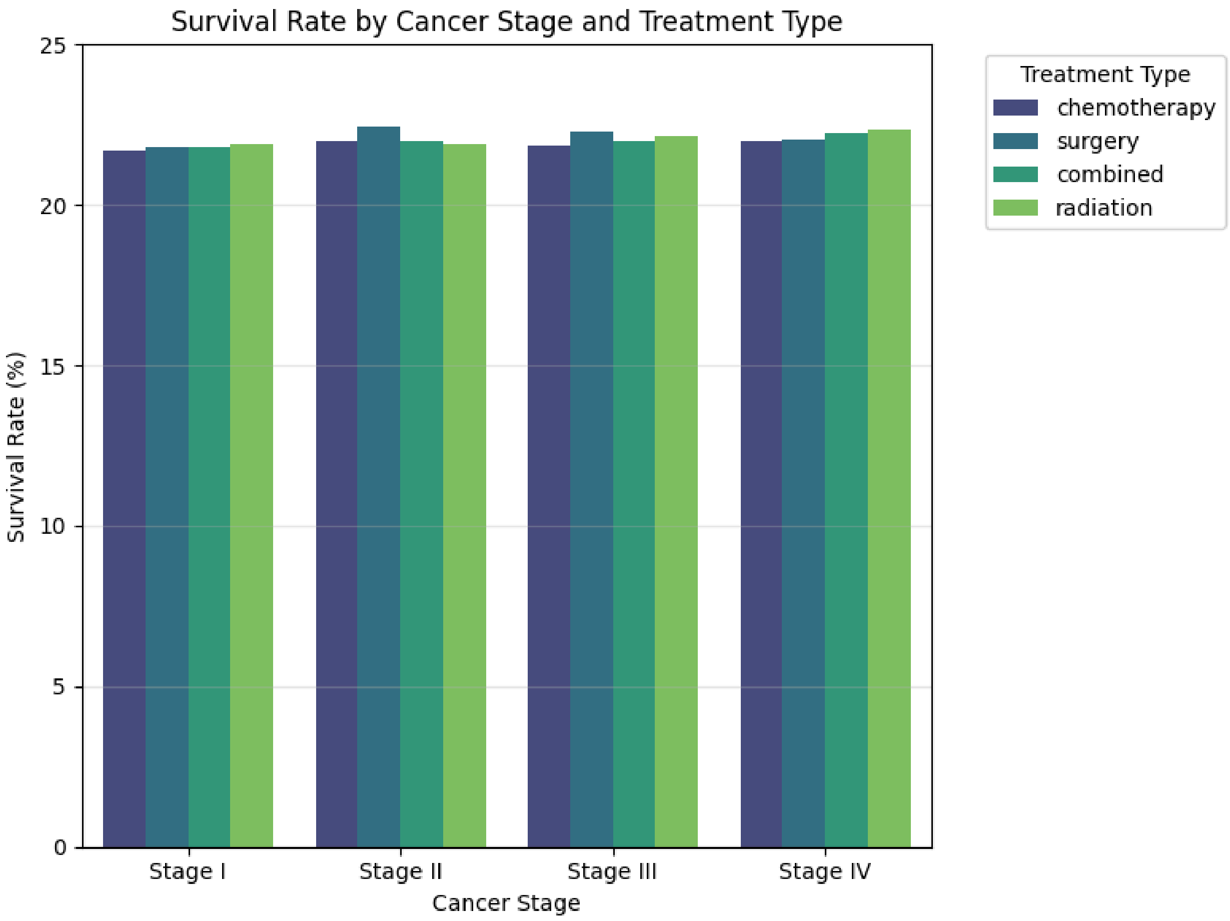
The patient count visualization shows us a pattern that seems to persist throughout the dataset, that despite the stage of cancer the survival counts and deaths are around the same. This would be unexpected for cancer, as usually the later stages would be more deadly. This leads us to believe that the collection of the data was uniformly reaching thresholds for each instead of random sampling. Along with this some downstream modeling, and analytics that would need to be taken into account would be potential misclassification, or incorrect predictions.

e. Bar Graph - Survival Rate by Treatment Type



Another column we took a look at was treatment type and whether it had a significant impact on survival rate or not. Here the treatments were broken down into chemotherapy, surgery, radiation, or a combined approach. Again seen in this data set all treatment had around the same percentage of survival with it settling around 22%. This again poses an interesting and potential dangerous downstream implication with misclassification and prediction inaccuracy.

f. Multivariate Barplot -Survival Rate by Cancer Stage and Treatment Type



Up until this point we have been looking at univariate graphs. Here because we were seeing little variation when it came to our data once we parsed it, we decided to go with a multivariate approach. Here we took the previous two graphs and data and tried to cross examine. This resulted in similar results, but some things of note such as for stage 2, and 3 of the cancer surgery seemed to perform slightly better than that of other treatment types, and for stage 4 combined and radiation seemed to slightly perform better.