LUNG CANCER ANALYTICS

Chatchanok Chatchaiyaroek(11) , MSc.Data Science(B)

LUNG CANCER DATASET

Lung cancer is a type of cancer that begins in the lungs, often associated with smoking but also linked to factors like air pollution, genetic predisposition, and exposure to toxins such as asbestos and radon. It is one of the leading causes of cancer-related deaths globally, as it is often diagnosed in advanced stages.

Global Perspective: In 2022, lung cancer was the leading cause of cancer death worldwide, accounting for approximately 1.8 million deaths, which represented 18.7% of all cancer-related fatalities.

LUNG CANCER DATASET

Data Shape : 300000, 30

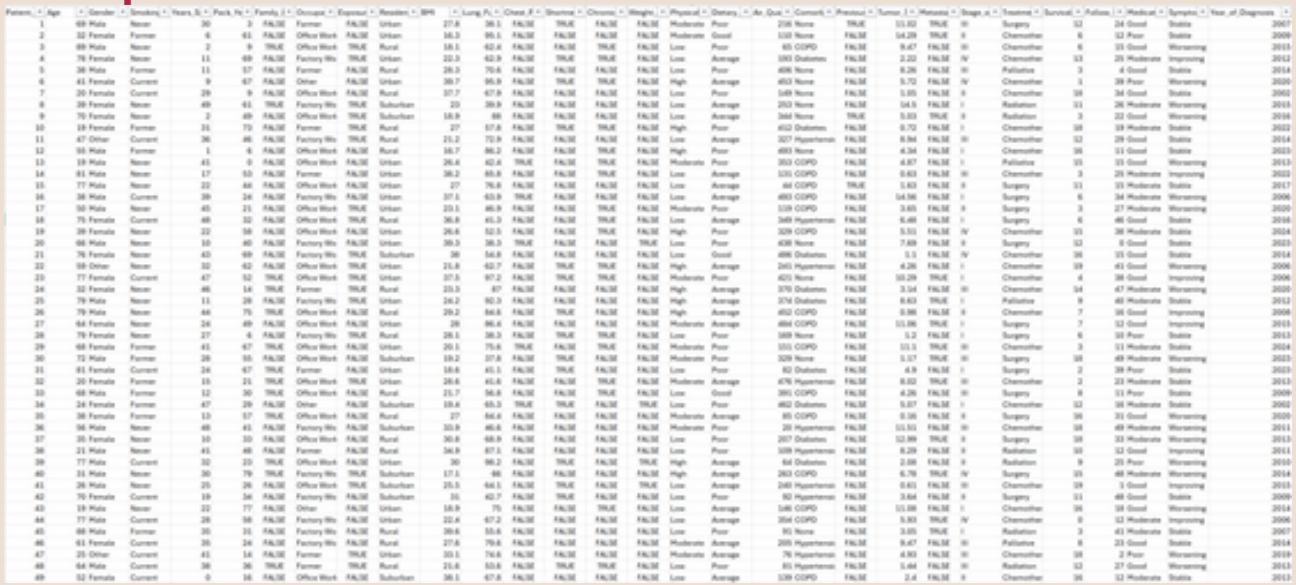
Columns

1.Patient_ID
2.Age
3.Gender
4.Smoking_History
5.Years_Smoked
6.Pack_Years
7.Family_History_Cancer
8.Occupation
9.Exposure_to_Toxins
10.Residential_Area
11.BMI
12.Lung_Function_Test_Result
13.Chest_Pain_Symptoms
14.Shortness_of_Breath
15.Chronic_Cough
16.Weight_Loss
17.Physical_Activity_Level
18.Dietary_Habits
19.Air_Quality_Index
20.Comorbidities
21.Previous_Cancer_Diagnosis
22.Tumor_Size_cm
23.Metastasis_Status
24.Stage_of_Cancer
25.Treatment_Type
26.Survival_Years
27.Follow_Up_Visits
28.Medication_Response
29.Symptom_Progression
30.Year_of_Diagnosis

LUNG CANCER DATASET

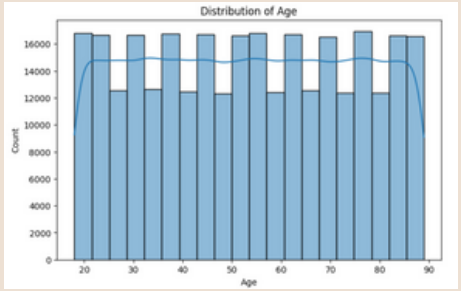
Example of dataset

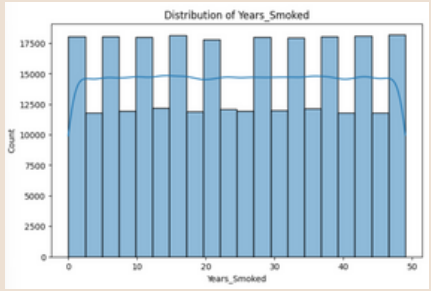
```
 #   Column                      Non-Null Count   Dtype
---  ------                      --------------   -----
 0   Patient_ID                  300000 non-null  int64
 1   Age                         300000 non-null  int64
 2   Gender                      300000 non-null  object
 3   Smoking_History             300000 non-null  object
 4   Years_Smoked                300000 non-null  int64
 5   Pack_Years                  300000 non-null  int64
 6   Family_History_Cancer       300000 non-null  bool
 7   Occupation                  300000 non-null  object
 8   Exposure_to_Toxins          300000 non-null  bool
 9   Residential_Area            300000 non-null  object
 10  BMI                         300000 non-null  float64
 11  Lung_Function_Test_Result   300000 non-null  float64
 12  Chest_Pain_Symptoms         300000 non-null  bool
 13  Shortness_of_Breath         300000 non-null  bool
 14  Chronic_Cough               300000 non-null  bool
 15  Weight_Loss                 300000 non-null  bool
 16  Physical_Activity_Level     300000 non-null  object
 17  Dietary_Habits              300000 non-null  object
 18  Air_Quality_Index           300000 non-null  int64
 19  Comorbidities               209624 non-null  object
 20  Previous_Cancer_Diagnosis   300000 non-null  bool
 21  Tumor_Size_cm               300000 non-null  float64
 22  Metastasis_Status           300000 non-null  bool
 23  Stage_of_Cancer             300000 non-null  object
 24  Treatment_Type              300000 non-null  object
 25  Survival_Years              300000 non-null  int64
 26  Follow_Up_Visits            300000 non-null  int64
 27  Medication_Response         300000 non-null  object
 28  Symptom_Progression         300000 non-null  object
 29  Year_of_Diagnosis           300000 non-null  int64
dtypes: bool(8), float64(3), int64(8), object(11)
```

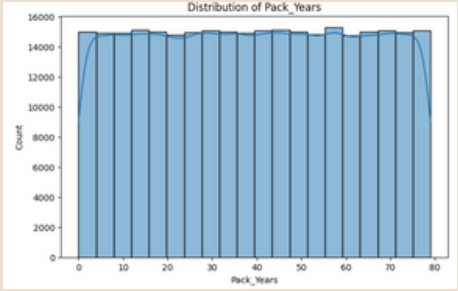


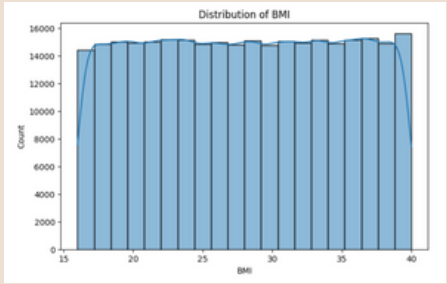
Source:https://www.kaggle.com/datasets/ashaychoudhary/comprehensive-lung-cancer-dataset/data

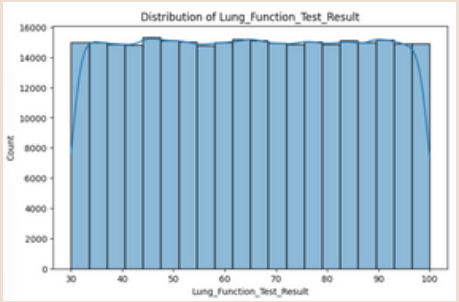
LUNG CANCER DATASET Distributions

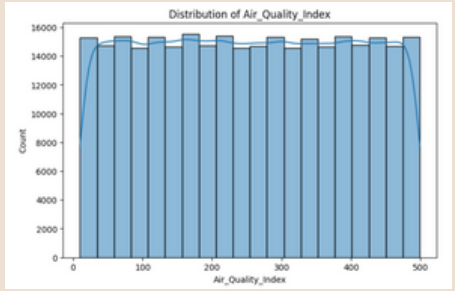


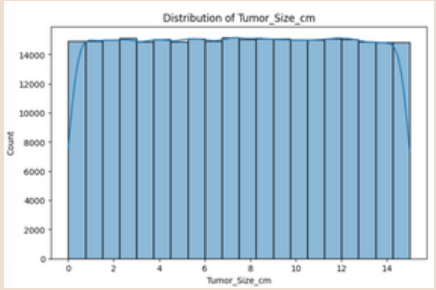


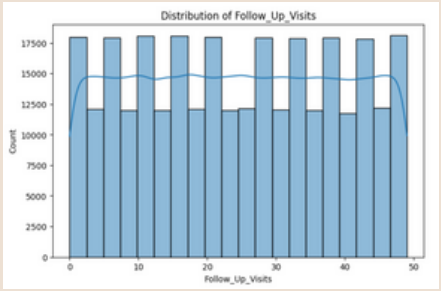


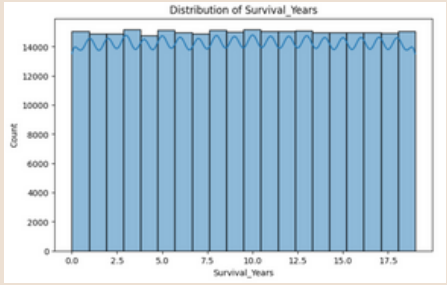


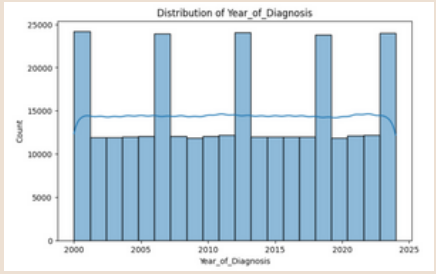




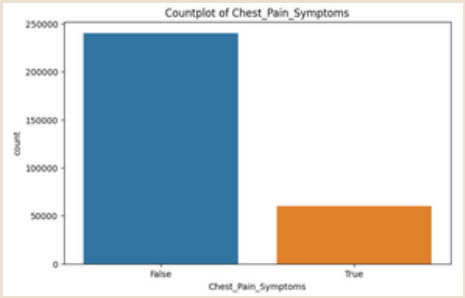


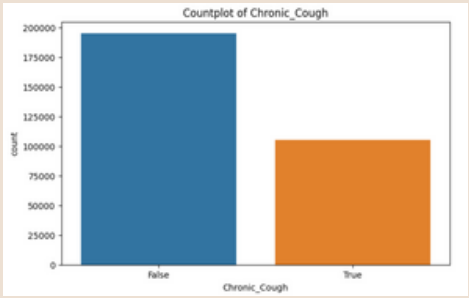


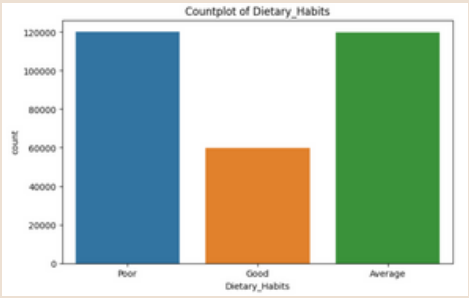


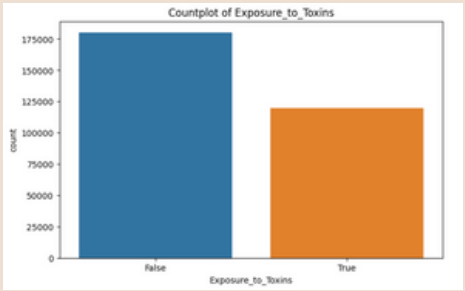


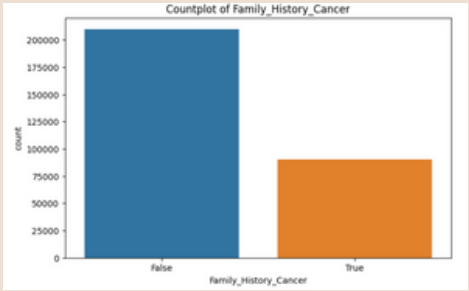
LUNG CANCER DATASET Distributions

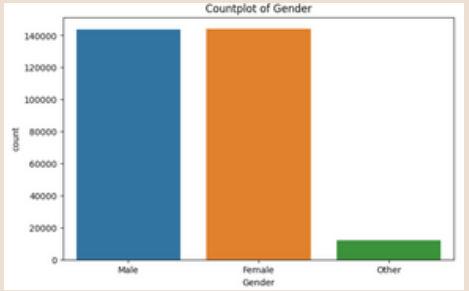


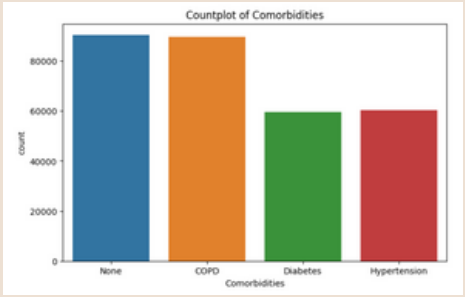


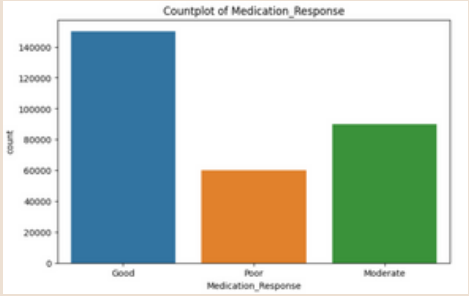


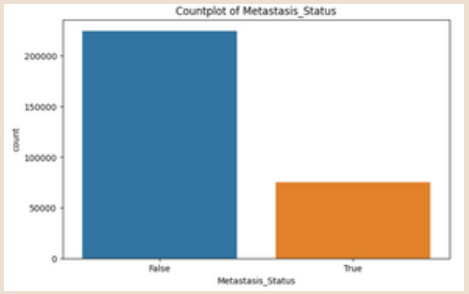




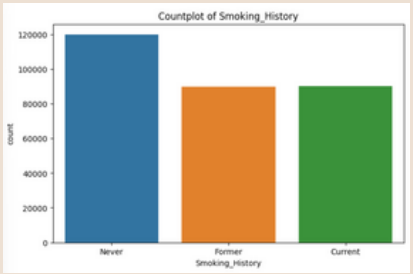


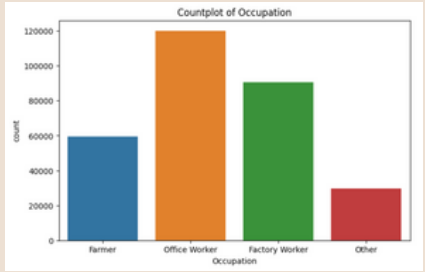


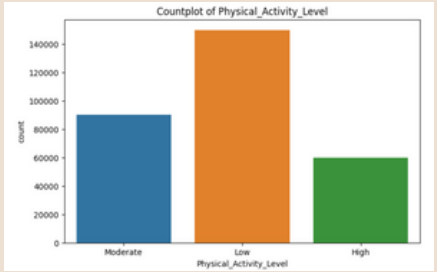


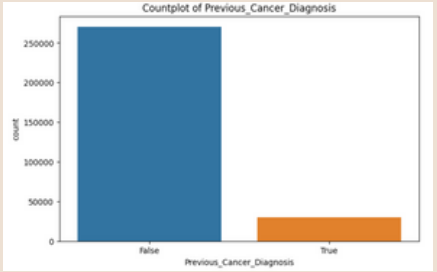


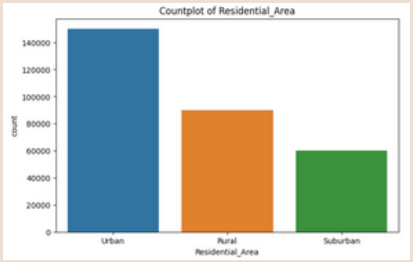
LUNG CANCER DATASET Distributions

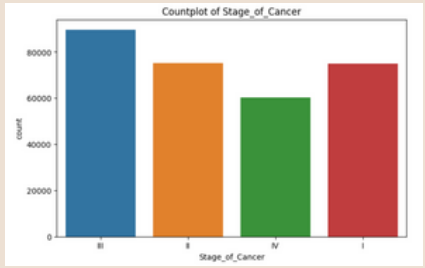


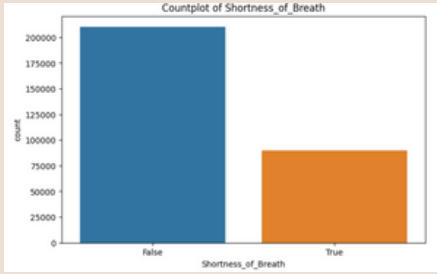


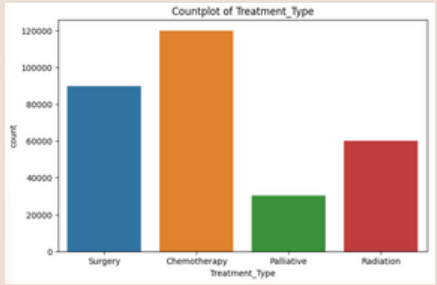


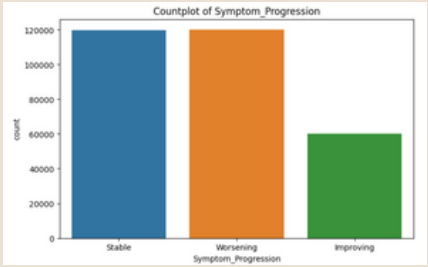




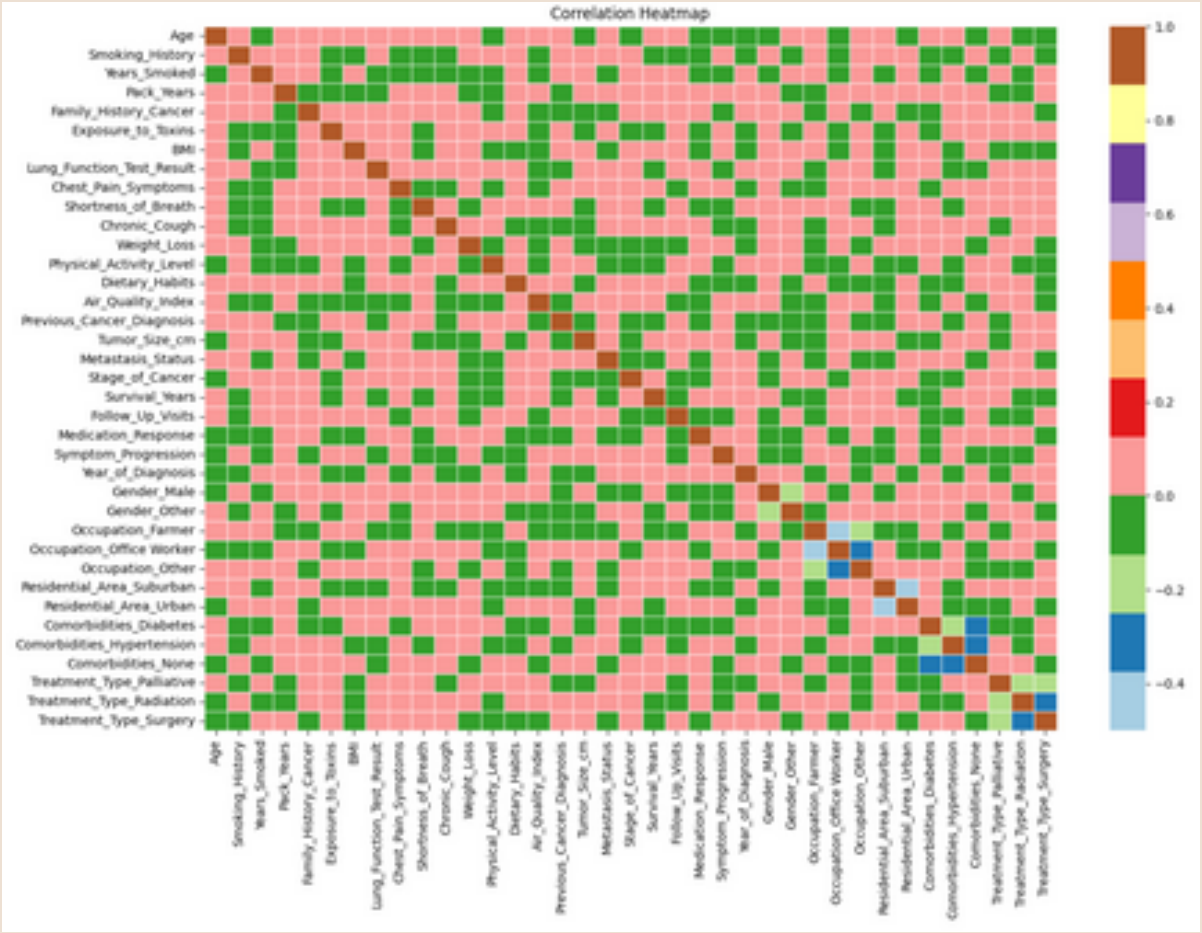




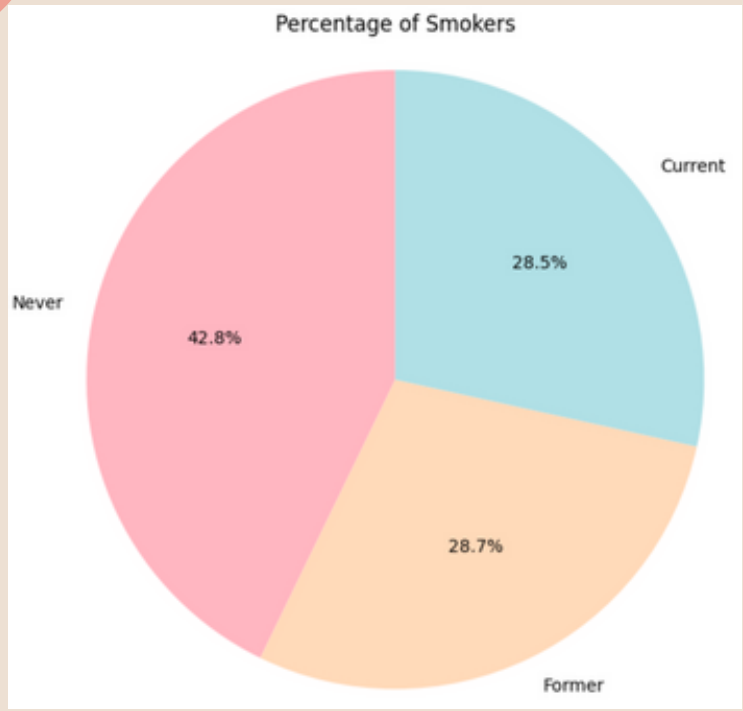




LUNG CANCER DATASET Correlations

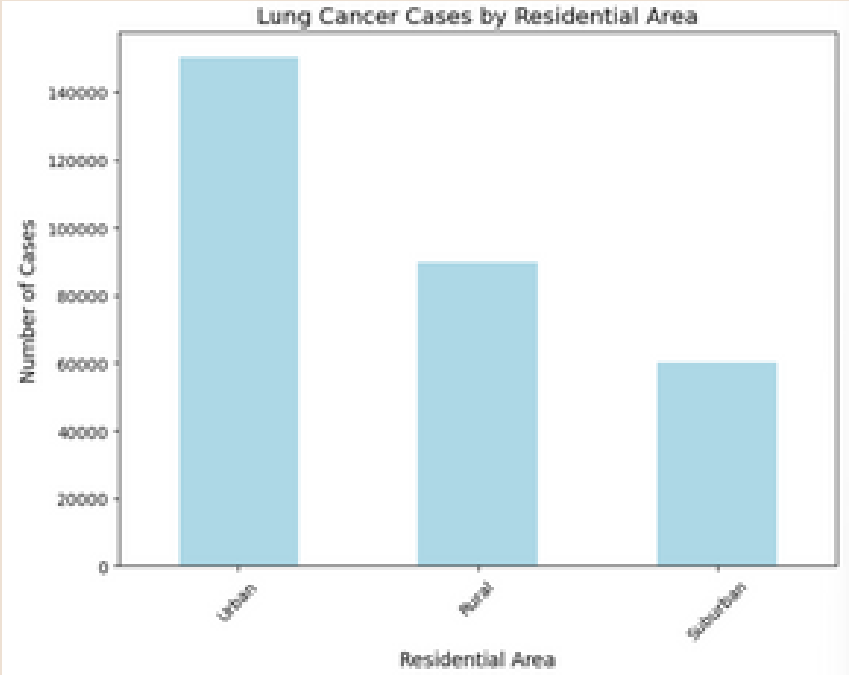


QUESTION 1 :WHAT PERCENTAGE OF LUNG CANCER PATIENTS ARE SMOKERS ?



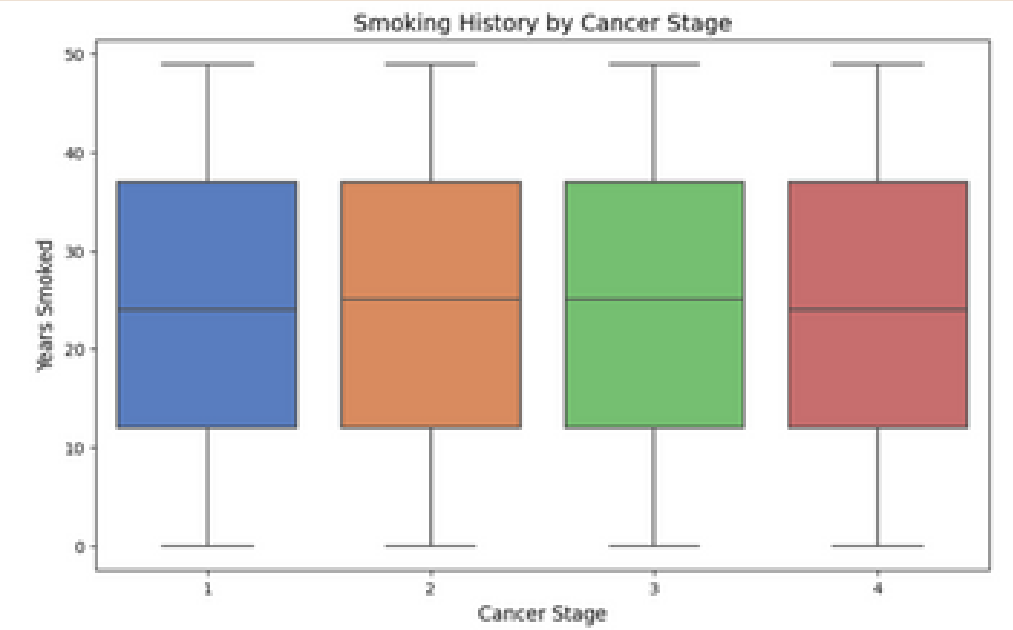
- Never Smoked (42.8%): A significant portion of lung cancer patients have never smoked, indicating that factors other than smoking could also contribute to lung cancer.
- Former Smokers (28.7%): About one-third of the patients are former smokers, suggesting a potential long-term impact of smoking on lung cancer risk even after quitting.
- Current Smokers (28.5%): Another one-third are current smokers, reinforcing the well-documented connection between smoking and lung cancer development.

QUESTION2 :HOW DOES LUNG CANCER PREVALENCE VARY ACROSS DIFFERENT RESIDENTIAL AREAS?



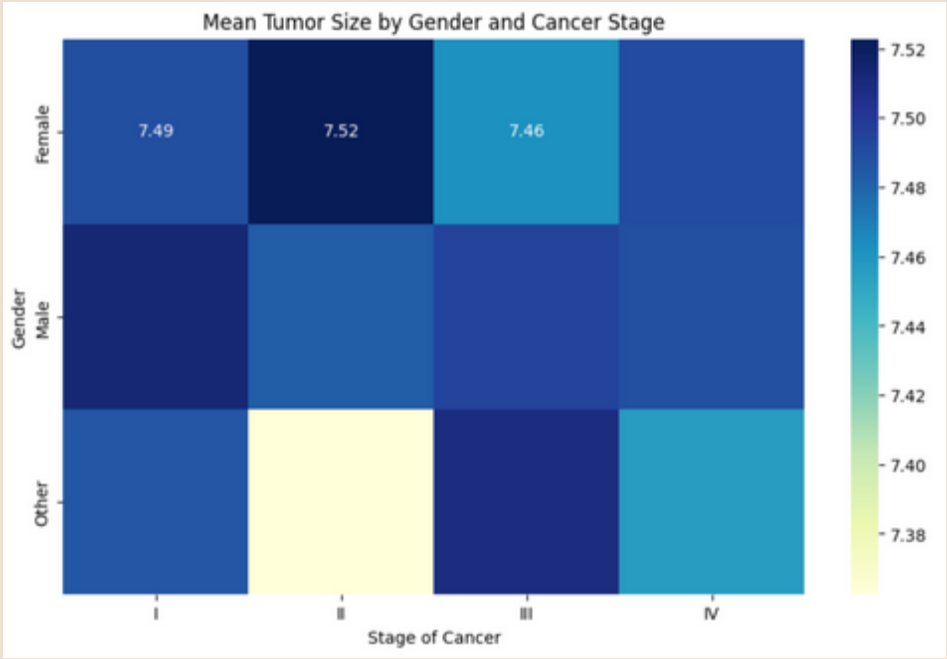
Lung cancer appears to be most prevalent in urban areas, with approximately 150,000 cases. Rural areas have the second highest prevalence, with around 90,000 cases. Suburban areas have the lowest prevalence, with roughly 60,000 cases.

QUESTION3 : DOES SMOKING HISTORY CORRELATE WITH THE STAGE OF LUNG CANCER AT DIAGNOSIS?



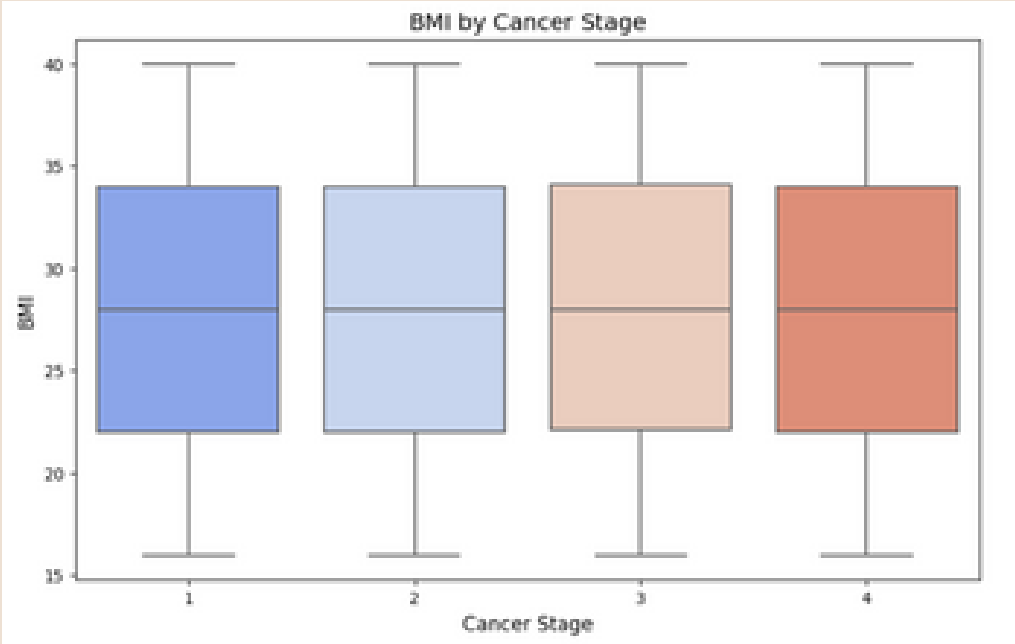
The median years smoked remains relatively consistent across all four cancer stages. This suggests that the typical smoking duration is similar for patients regardless of their stage at diagnosis.
There is no clear correlation between smoking history and lung cancer stage. The number of years smoked appears similar across all stages.

QUESTION4 : "WHAT ARE THE DIFFERENCES IN MEAN TUMOR SIZE BETWEEN GENDERS ACROSS VARIOUS CANCER STAGES?"



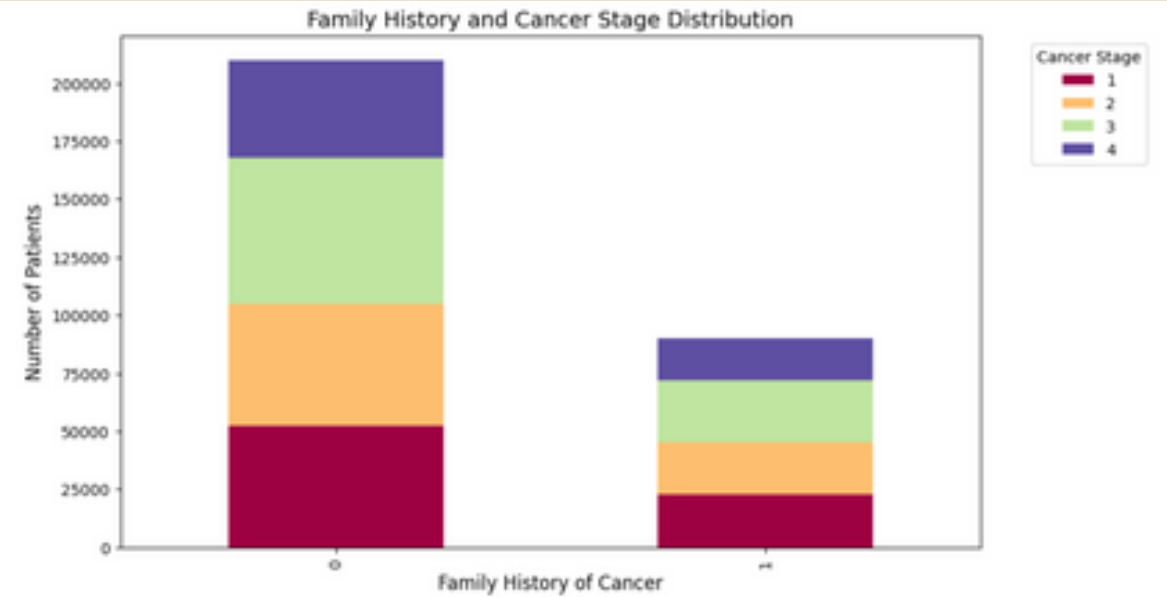
- Males: Tend to have larger tumors in the later stages of cancer (Stage II and IV).
- Females: Generally have slightly larger tumors in the earlier stages (Stage I and III).
- Other: Individuals in this category consistently have the smallest average tumor sizes across all stages.

QUESTION5 : HOW DOES BMI (BODY MASS INDEX) RELATE TO LUNG CANCER STAGE OR PROGRESSION?



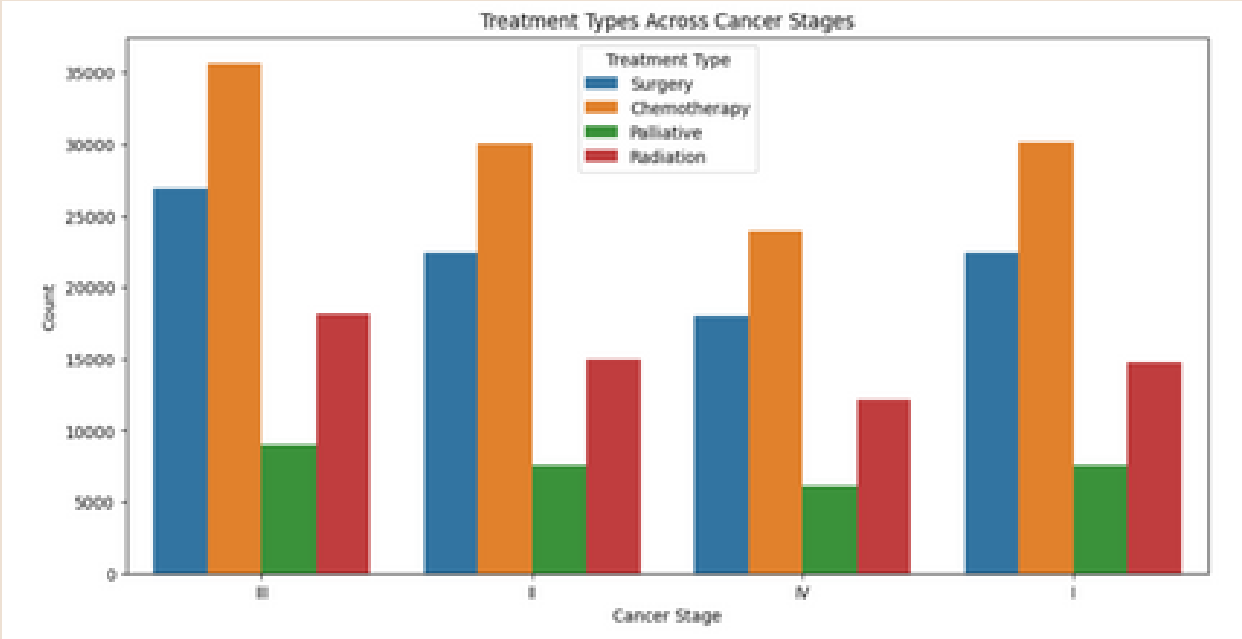
The overall distribution of BMI values seems to be relatively similar across all stages.
The boxes overlap considerably, indicating that the spread of BMI values is similar for patients at different stages.

QUESTION6 :DOES FAMILY HISTORY OF CANCER INCREASE THE RISK OF ADVANCED-STAGE LUNG CANCER?



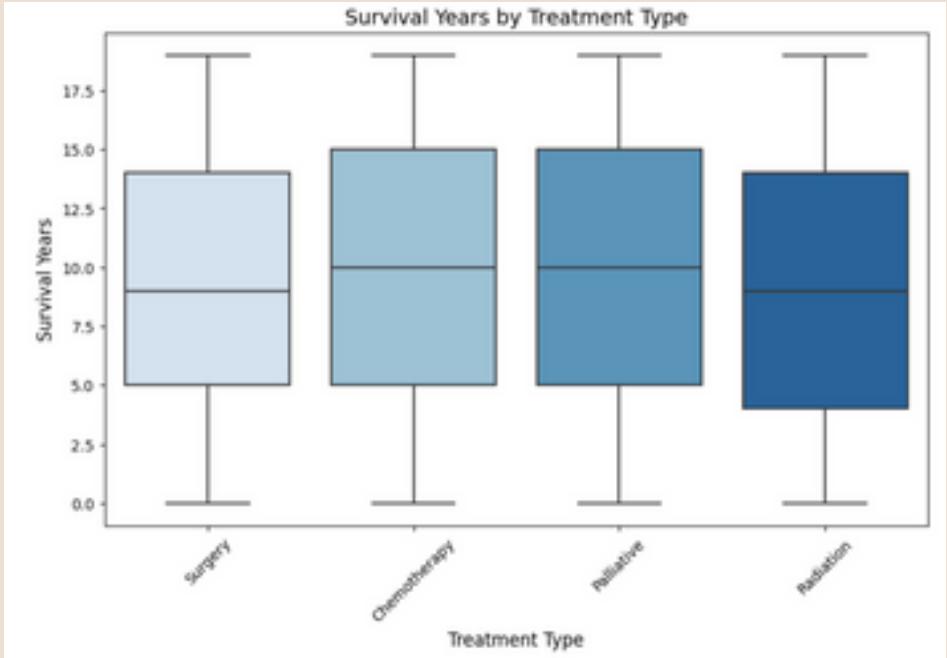
Patients with a family history of cancer seem to have a higher likelihood of being diagnosed with advanced-stage lung cancer (Stage 3 or 4) compared to those without a family history. This may suggest a genetic predisposition or delayed diagnosis due to other factors.

QUESTION7 : HOW DO TREATMENT TYPES VARY ACROSS DIFFERENT STAGES OF CANCER?



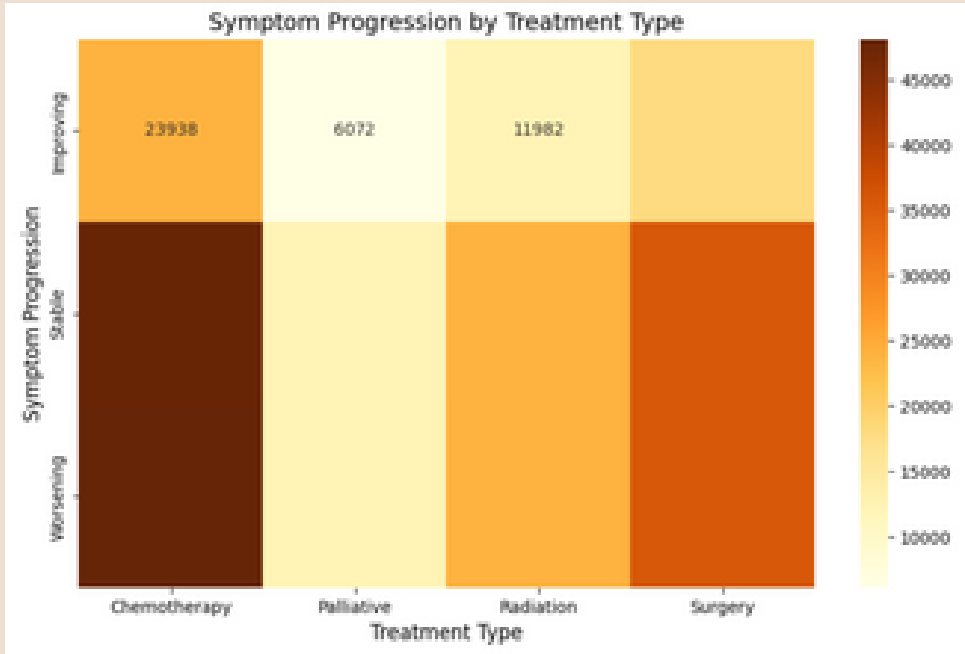
Chemotherapy is the most frequently used treatment across all cancer stages, followed by Surgery. Radiation and Palliative care are typically used less frequently."

QUESTION8 : HOW DOES THE TREATMENT TYPE IMPACT SURVIVAL YEARS?



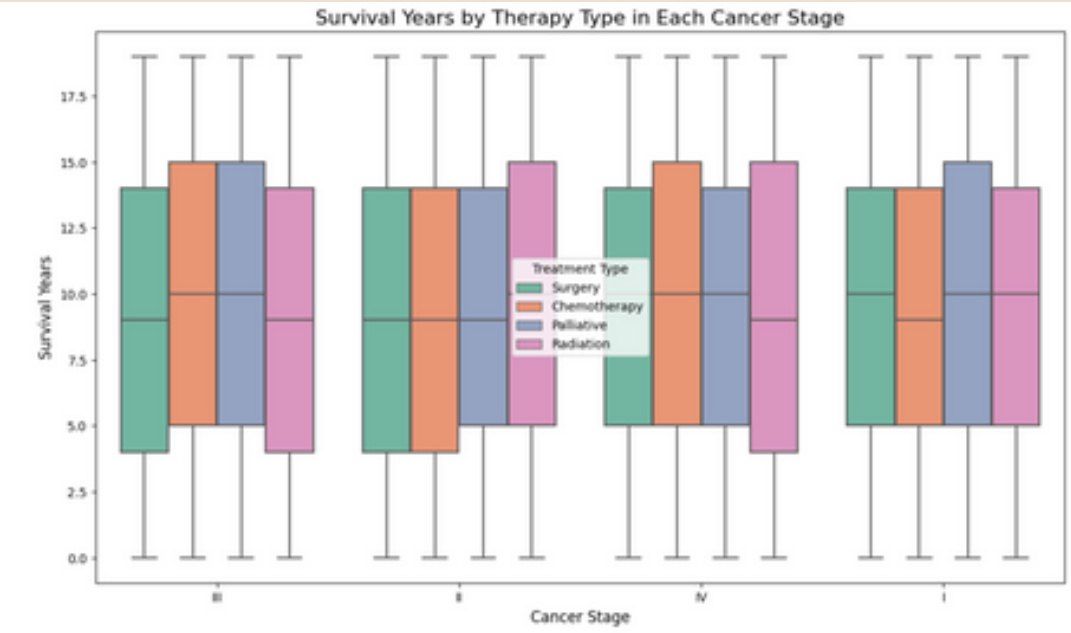
Even though , Chemotheraphy is the most popular treatment from the previous slide and Palliative is the least , the survival years of patients who got these treatments seem to be similar and a little bit higher than the others.

QUESTION9 : DOES SYMPTOM PROGRESSION VARY BY TREATMENT TYPE?



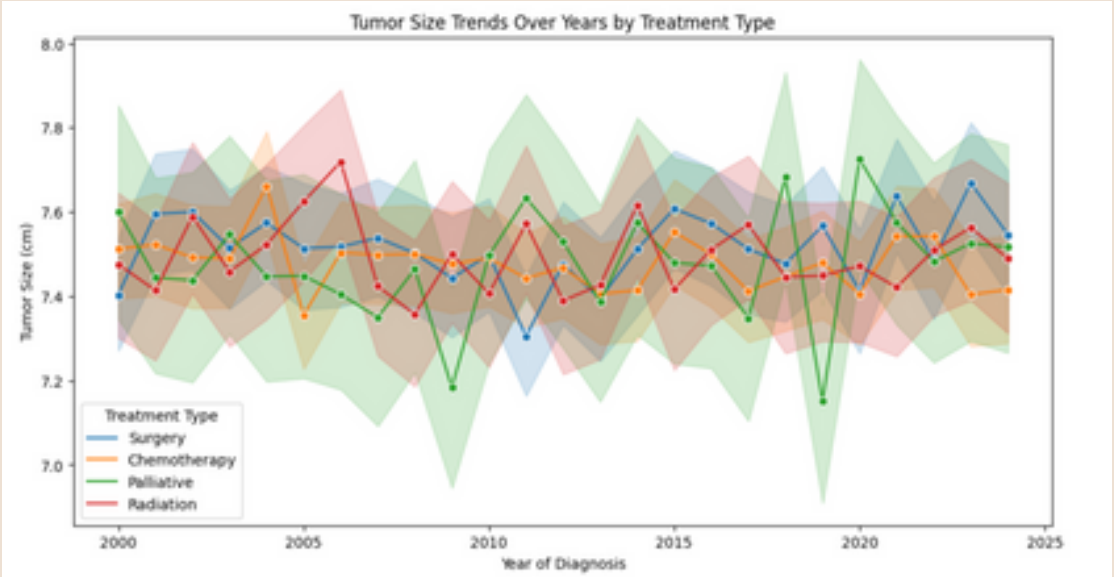
The treatment that gives the best improving symptom progression is ‘Chemotheraphy’ , then followed by Surgery , Radiation and Palliative.

QUESTION10 : "HOW DO SURVIVAL YEARS VARY ACROSS DIFFERENT MEDICAL RESPONSES FOR EACH TREATMENT TYPE?"



- Surgery appears to result in longer survival years across most stages compared to other treatments.
- Chemotherapy and Radiation have overlapping distributions, with slightly lower medians.
- Palliative care shows the shortest survival years overall, reflecting its use in advanced or end-stage care.

QUESTION11 : ARE THERE PATTERNS IN TUMOR SIZE BASED ON YEAR OF DIAGNOSIS AND TREATMENT TYPE?



- No clear overall trend: Tumor sizes fluctuate over the years for all treatment types without a consistent upward or downward direction.
- Treatment-specific variations: There are slight variations in tumor size trends for each treatment type, with some showing initial increases followed by decreases and vice versa.
- Year-to-year differences: Tumor sizes vary across different years of diagnosis, with some years showing higher sizes for specific treatment types.

CONCLUSION

The causes of lung cancer, the symptoms that manifest, and the body's response to treatment vary significantly between individuals and are not universally dependent on any single factor.

Causes

The correlations between lung cancer and its potential causes, such as smoking, pollution, and genetics, are relatively low, indicating a complex interplay of factors. While smoking remains a significant risk factor, non-smokers are also susceptible due to other influences such as poor air quality, occupational hazards, or genetic predisposition. Interestingly, people living in urban areas or working in office environments appear to have a higher prevalence of lung cancer, possibly linked to environmental pollution or other urban lifestyle factors.

Symptoms

The symptoms of lung cancer vary greatly among individuals. Common signs include chest pain, persistent coughing, and shortness of breath, but these are not universal. Some patients develop lung cancer despite being otherwise healthy, with no underlying conditions or comorbidities, emphasizing the unpredictable nature of the disease.

Treatment

Chemotherapy is one of the most widely used treatments and shows the most significant overall improvement in disease progression. However, its effectiveness varies based on the stage of cancer at diagnosis. Notably, patients receiving palliative treatments often demonstrate comparable survival years to those undergoing more aggressive therapies, highlighting the individualized nature of treatment outcomes.