**1. Gender Distribution:**

* **Goal**: Visualize and calculate the proportion of male vs female patients.
* **Metric**: Percentage of male and female patients.
* **Tools**:
  + SQL Query:

*SELECT gender, COUNT(\*) \* 100.0 / (SELECT COUNT(\*) FROM patients) AS percentage*

*FROM patients*

*GROUP BY gender;*

* + **Chart**: Use a **pie chart** or **bar chart** in Excel/SQL visualization tools like **Tableau**.

**2. Age Distribution:**

* **Goal**: Understand the distribution of patient ages and identify age groups with higher occurrences of cancer.
* **Metric**: Mean, median, and age group breakdown.
* **SQL Query**:

*SELECT AVG (age) AS mean\_age, MEDIAN(age) AS median\_age FROM patients;*

Group ages:

*SELECT CASE*

*WHEN age BETWEEN 0 AND 30 THEN '0-30'*

*WHEN age BETWEEN 31 AND 50 THEN '31-50'*

*WHEN age > 50 THEN '51+'*

*END AS age\_group, COUNT(\*) AS frequency*

*FROM patients*

*GROUP BY age\_group;*

* **Chart**: Use **histogram** or **boxplot** to display distribution.

**3. Treatment Distribution Analysis:**

* **Goal**: Analyze which treatments for lung cancer (e.g., chemotherapy, radiation, surgery) are most used.
* **Key Metric**: Frequency of each treatment type.
* **SQL Query**:

*SELECT treatment\_type, COUNT(\*) AS frequency*

*FROM patients*

*GROUP BY treatment\_type*

*ORDER BY frequency DESC;*

* **Visualization**: **Bar chart** or **table** showing the most frequent treatments.

**4. Survival Status Analysis:**

* **Goal**: Analyze the survival rates based on gender, age, and treatment type.
* **Key Metric**: Percentage of patients surviving (or high survival rate) across different groups.
* **SQL Query**:

*SELECT gender, treatment\_type,*

*SUM(CASE WHEN survival\_rate >= 0.5 THEN 1 ELSE 0 END) \* 100.0 / COUNT(\*) AS survival\_rate*

*FROM patients*

*GROUP BY gender, treatment\_type;*

* **Visualization**: **Stacked bar chart** showing differences in survival rate based on gender and treatment type.

**5. Treatment and Outcome Correlation:**

* **Goal**: Investigate correlations between treatment types (e.g., chemotherapy, radiation, surgery) and survival outcomes.
* **Key Metric**: Survival rate per treatment.
* **SQL Query**:

*SELECT treatment\_type, AVG(survival\_rate) AS avg\_survival\_rate*

*FROM patients*

*GROUP BY treatment\_type;*

* **Visualization**: **Boxplots** to visualize survival rate variation across treatment types.

**6. Smoking and Alcohol Consumption:**

* **Goal**: Explore the impact of smoking and alcohol consumption on lung cancer survival rates.
* **Key Metric**: Survival rates and lung cancer occurrences by smoking or alcohol consumption status.
* **SQL Query**:

*SELECT smoking\_status, alcohol\_consumption, AVG(survival\_rate) AS avg\_survival\_rate*

*FROM patients*

*GROUP BY smoking\_status, alcohol\_consumption;*

* **Visualization**: **Contingency tables** or **grouped bar charts** to compare survival between smokers vs non-smokers, and alcohol users vs non-users.

**7. Age vs. Survival Rate:**

* **Goal**: Evaluate how survival rates vary with age.
* **Key Metric**: Mean survival rate by age group.
* **SQL Query**:

*SELECT age, AVG(survival\_rate) AS avg\_survival\_rate*

*FROM patients*

*GROUP BY age;*

* **Visualization**: **Scatterplot** or **line chart** to display survival rate trends across different ages.

**8. BMI and Survival:**

* **Goal**: Analyze the relationship between BMI and lung cancer survival rates.
* **Key Metric**: Mean BMI of survivors vs non-survivors.
* **SQL Query**:

*SELECT AVG(BMI) AS avg\_bmi,*

*SUM(CASE WHEN survival\_rate >= 0.5 THEN 1 ELSE 0 END) / COUNT(\*) AS survival\_rate*

*FROM patients*

*GROUP BY survival\_rate;*

* **Visualization**: **Scatterplot** showing the relationship between BMI and survival rate.

**ML analysis:**

**9. Mortality Risk Factors:**

* **Goal**: Use multivariate analysis to identify key risk factors (e.g., smoking, BMI, age) associated with mortality.
* **Key Metric**: Odds ratios or risk metrics for various factors like age, gender, smoking, BMI.
* **Method**: Use **logistic regression** to predict survival outcomes based on these factors.
* **Tool**: Perform logistic regression using **Python**, **R**, or a similar tool. For example, in Python, you can use the statsmodels package to run logistic regression.

**Statistical analysis:**

**10. Statistical Significance Testing:**

* **Goal**: Conduct statistical tests (e.g., chi-square, t-tests) to see if survival differences between different groups are significant (e.g., gender, treatment type, smoking).
* **Key Metric**: P-values for group comparisons.
* **SQL Query** (for setting up a chi-square test for survival and smoking status):

*SELECT smoking\_status, survival\_rate, COUNT(\*)*

*FROM patients*

*GROUP BY smoking\_status, survival\_rate;*

* **Tool**: Use **Python** or **R** for chi-square or t-tests to determine statistical significance.

**11. Survival Rate by Treatment Type and Age Group**

* **Goal**: Examine the survival rate for each treatment type, broken down by age groups.
* **SQL Query**:

*SELECT treatment\_type,*

*CASE*

*WHEN age BETWEEN 0 AND 30 THEN '0-30'*

*WHEN age BETWEEN 31 AND 50 THEN '31-50'*

*WHEN age > 50 THEN '51+'*

*END AS age\_group,*

*AVG(survival\_rate) AS avg\_survival\_rate*

*FROM patients*

*GROUP BY treatment\_type, age\_group;*

* **Insight**: See how survival rates for different treatments vary by age group.

**12. Alcohol Consumption and Survival Rate**

* **Goal**: Analyze whether alcohol consumption affects survival rates.
* **SQL Query**:

*SELECT alcohol\_consumption, AVG(survival\_rate) AS avg\_survival\_rate*

*FROM patients*

*GROUP BY alcohol\_consumption;*

* **Insight**: Explore the potential impact of alcohol consumption on survival outcomes.