Lung Cancer Patients Data Analysis

Dataset Information

- Source: Kaggle
- Dataset: lung_cancer_data.csv
- Uploader: Rashad Mammadov

Reason the Problem Was Chosen

Lung cancer is one of the most common and deadly forms of cancer worldwide, with high incidence and mortality rates. Understanding the factors that influence lung cancer outcomes, such as tumor size, stage, and treatment effectiveness, is crucial for improving patient prognosis and survival rates. This problem was chosen because analyzing these variables can provide valuable insights for medical professionals, researchers, and policymakers aiming to develop better diagnostic, preventive, and treatment strategies for lung cancer.

Reason the Title Was Chosen

The title "Lung Cancer Data Analysis: Insights into Tumor Characteristics, Treatment Outcomes, and Survival" was chosen to reflect the comprehensive nature of the report. It highlights the key aspects of the analysis, including tumor characteristics, treatment outcomes, and survival rates, providing readers with a clear understanding of the report's focus and scope.

Solution

The analysis performed in this report provides a detailed examination of various factors related to lung cancer, such as tumor size across different stages, the impact of treatment types on survival months, and the relationship between smoking history and cancer stage. By identifying significant patterns and predictors, the findings can help in developing targeted interventions and personalized treatment plans. Additionally, the correlation analysis between continuous variables can guide future research in identifying biomarkers and other key indicators for lung cancer prognosis.

Relation to Sustainable Development Goal (SDG) Number 3

SDG 3 aims to ensure healthy lives and promote well-being for all at all ages. One of its targets is to reduce premature mortality from non-communicable diseases, including cancer, through prevention, treatment, and promoting mental health and well-being. This report directly contributes to SDG 3 by providing insights into lung cancer, which can lead to improved diagnostic and treatment methods, ultimately reducing mortality rates and enhancing patient outcomes. By addressing critical aspects of lung cancer management, this analysis supports the global effort to achieve better health and well-being for individuals affected by this disease.

Other Relevant Details from the Analysis

- **Comprehensive Data Exploration**: The exploratory data analysis (EDA) provided summary statistics and visualizations that offered an initial understanding of the dataset. This step is essential for identifying key trends and potential areas of interest for further analysis.
- **Detailed Statistical Analysis**: Multiple statistical tests, including ANOVA, Kruskal-Wallis, chi-square tests, and logistic regression, were performed to examine the relationships between variables. These tests helped identify significant differences and associations, providing a deeper understanding of the factors influencing lung cancer outcomes.
- **Interactive Visualizations**: The report includes interactive visualizations, such as histograms and boxplots, which allow for a more engaging and insightful presentation of the data. These visualizations help in effectively communicating complex statistical findings to a broader audience.
- **Implications for Future Research**: The findings from this analysis suggest several avenues for future research, including exploring genetic markers, environmental exposures, and longitudinal studies to understand the disease's progression over time. This can further enhance our knowledge of lung cancer and contribute to the development of more effective prevention and treatment strategies.

Objective

- Identify key factors influencing lung cancer survival.
- Assess the statistical significance of these factors.
- Provide visualizations and interpretative insights.

The primary objectives of this project are to uncover the main factors affecting lung cancer survival, evaluate their statistical significance, and present the findings through visualizations and interpretative insights to assist researchers and healthcare professionals.

Methods to Analyze the Dataset

1. Exploratory Data Analysis (EDA): Understanding the distribution and relationships between variables.
2. KaplanMeier Survival Curves: Estimating the survival function over time.
3. Cox Proportional Hazards Model: Assessing the impact of various factors on the risk of death.

We utilized several methods to analyze the lung cancer data:

- EDA helps in understanding the basic structure and distribution of the data.
- KaplanMeier Survival Curves are used to estimate the survival probabilities over time.
- Cox Proportional Hazards Model evaluates the effect of multiple covariates on survival.

Descriptive Stats

- The average age of patients is 65 years.
- 60% of patients are male.
- 70% of patients had a history of smoking.
- Median tumor size was 25 mm.
- The majority of patients were diagnosed at stage III.

- Median survival time is 18 months.

From the dataset, we derive key descriptive statistics:

- Patients' average age is 65 years.
- There is a higher prevalence of males (60%).
- A significant portion of the patients (70%) have a smoking history.
- The median size of the tumors is 25 mm.
- Most patients were diagnosed at stage III, indicating advanced cancer.
- The median survival time for patients is 18 months.

Key Findings

1. Cancer Stages: Advanced stages of cancer tend to have larger tumor sizes.

   The analysis shows that tumors tend to be larger in patients diagnosed at more advanced stages of cancer. This is visually represented through the "Cancer Stages" section in the poster.

2. Tumor Size: Larger tumor size slightly reduces the risk of death, which is unusual and may need further investigation.

   Interestingly, the Cox model suggests that larger tumor sizes are associated with a slightly reduced risk of death (HR: 0.99, CI: 0.97 1, PValue: 0.03). This counterintuitive finding highlights the need for further investigation to ensure the accuracy of the model and coding.

3. Ethnicity: Some ethnic groups have a slightly lower risk of death compared to the reference group, indicating that ethnicity may affect lung cancer survival.

The analysis also reveals that certain ethnic groups have a statistically significant lower risk of death compared to the reference group (HR: 0.98, CI: 0.97 1, PValue: 0.017). This suggests that ethnicity can be an important factor in lung cancer survival.

4. Treatments: There are significant differences in survival probabilities among different treatment groups. Patients receiving certain treatments show better survival rates compared to others.

   Different treatment modalities exhibit varying impacts on survival probabilities, indicating the effectiveness of certain treatments over others. This information is crucial for tailoring patientspecific treatment plans to improve outcomes.

Conclusion

This comprehensive report provides valuable insights into lung cancer, addressing various factors that influence its prognosis and outcomes. By aligning with SDG 3, the analysis supports global efforts to reduce the burden of non-communicable diseases and promote health and well-being. The findings from this report can inform future research, clinical practices, and policy decisions, ultimately contributing to better health outcomes for lung cancer patients.

References

GeeksforGeeks.org Exploratory Data Analysis in R Programming

DataCamp.com KaplanMeier Method

DataCamp.com Cox Proportional Hazards Models

Investopedia.com Statistical Significance: Definition, Types, and How It’s Calculated

StatisticsByJim.com Hazard Ratio: Interpretation & Definition

Our Team

260274595 Rayyan

2602028946 Azrie Muhammad Nurlan

2602028246 Fiya Geneta Amrezi

260207483 Faiz Abisha Santoso