Homework1.1-1

Group Members:
Liner Zhang, Zizhuo Shi, Wanyu Lin

Data Chosen:
https://www.kaggle.com/datasets/samikshadalvi/lungs-diseases-dataset

Individual Contribution			
CWID	**Name**	**Contribution (description)**	**Percent Contribution**
A20563408	**Liner Zhang**	**Data Description, Data Visualization Section 3.4**	**33.33%**
A20563449	**Zizhuo Shi**	**Data Visualization Section 3.1-3.3, 3.5**	**33.33%**
A20563439	**Wanyu Lin**	**Data Preparation, Results**	**33.33%**

Contents

1. Data Description

After arbitrarily downloading a dataset, we conducted a preliminary analysis of its content and found the following characteristics:

- The dataset contains a total of 8 attributes, which are Age, Gender, Smoking Status, Lung Capacity, Disease Type, Treatment Type, Hospital Visits, and Recovered.
- Specifically, the disease type refers to specific lung diseases, such as chronic obstructive pulmonary disease (COPD) or bronchitis.
- The treatment type indicates the different treatments received by patients, including therapy, medication, or surgery.
- The Age attribute is discrete, quantitative, and ratio.
- The Gender attribute is binary, qualitative, and nominal.
- The Smoking Status attribute is binary, qualitative, and nominal.
- The Disease Type attribute is discrete, qualitative, and nominal.
- The Treatment Type attribute is discrete, qualitative, and nominal.
- The Hospital Visits attribute is discrete, quantitative, and ratio.
- The Recovered attribute is binary, qualitative, and nominal.
- The dataset consists of a total of 5,200 rows.
- Many rows contain missing values, which necessitates data preprocessing.

2. Data Preparation

2.1 DataFrame

Initially, we read the CSV file and converted it into a DataFrame to facilitate data analysis.

[3]:

	Age	Gender	Smoking Status	Lung Capacity	Disease Type	Treatment Type	Hospital Visits	Recovered
0	71.0	Female	No	4.49	COPD	Therapy	14.0	Yes
1	34.0	Female	Yes	NaN	Bronchitis	Surgery	7.0	No
2	80.0	Male	Yes	1.95	COPD	NaN	4.0	Yes
3	40.0	Female	Yes	NaN	Bronchitis	Medication	1.0	No
4	43.0	Male	Yes	4.60	COPD	Surgery	NaN	Yes

Figure 1: DataFrame

2.2 Handle Missing values

Based on the basic information, we found that each attribute has 300 missing values, calculated as (5200 - 4900).

```
[4]: df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5200 entries, 0 to 5199
Data columns (total 8 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Age              4900 non-null   float64
 1   Gender           4900 non-null   object
 2   Smoking Status   4900 non-null   object
 3   Lung Capacity    4900 non-null   float64
 4   Disease Type     4900 non-null   object
 5   Treatment Type   4900 non-null   object
 6   Hospital Visits  4900 non-null   float64
 7   Recovered        4900 non-null   object
dtypes: float64(3), object(5)
memory usage: 325.1+ KB
```

Figure 2: Data Information

We chose the simplest approach: directly removing the rows that contain missing values and found that there were still 3,236 rows remaining. At the same time, we obtained the basic information of the remaining data.

```
<class 'pandas.core.frame.DataFrame'>
Index: 3236 entries, 0 to 5199
Data columns (total 8 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Age              3236 non-null   float64
 1   Gender           3236 non-null   object
 2   Smoking Status   3236 non-null   object
 3   Lung Capacity    3236 non-null   float64
 4   Disease Type     3236 non-null   object
 5   Treatment Type   3236 non-null   object
 6   Hospital Visits  3236 non-null   float64
 7   Recovered        3236 non-null   object
dtypes: float64(3), object(5)
memory usage: 227.5+ KB
None
```

Figure 3: Data Information after Handle Missing Values (1)

[7]:

	Age	Lung Capacity	Hospital Visits
count	3236.000000	3236.000000	3236.000000
mean	54.229604	3.479617	7.502163
std	20.109979	1.465408	3.983233
min	20.000000	1.000000	1.000000
25%	37.000000	2.210000	4.000000
50%	54.000000	3.440000	8.000000
75%	72.000000	4.790000	11.000000
max	89.000000	6.000000	14.000000

Figure 4: Data Information after Handle Missing Values (2)

3. Data Visualization and Explanation

3.1 Overview of the Data

To gain an overall understanding of the data, we used the *Matplotlib* library to integrate all the data into a single scatter plot. The x-axis represents the number of hospital visits, and the y-axis represents age. The shape of the markers distinguishes the type of lung disease, while the color indicates whether the patient has recovered.

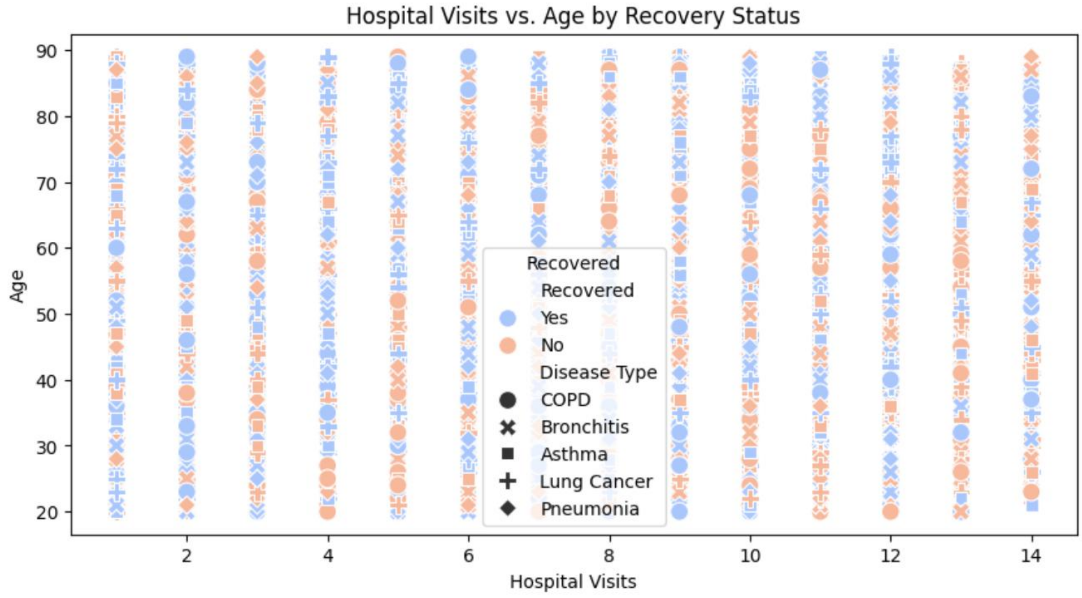

Figure 5: Line Chart of Age Distribution

Due to the large volume of data, the scatter points are densely packed, making it difficult to discern specific characteristics.

3.2 Frequency of Age Distribution

To explore the relationship between lung diseases and age, we first visualized the distribution of age.

Considering that age is a continuous variable, we used the *Matplotlib* library to create a line chart, which displays the distribution of age. The horizontal axis represents age, while the vertical axis indicates the frequency of each age.

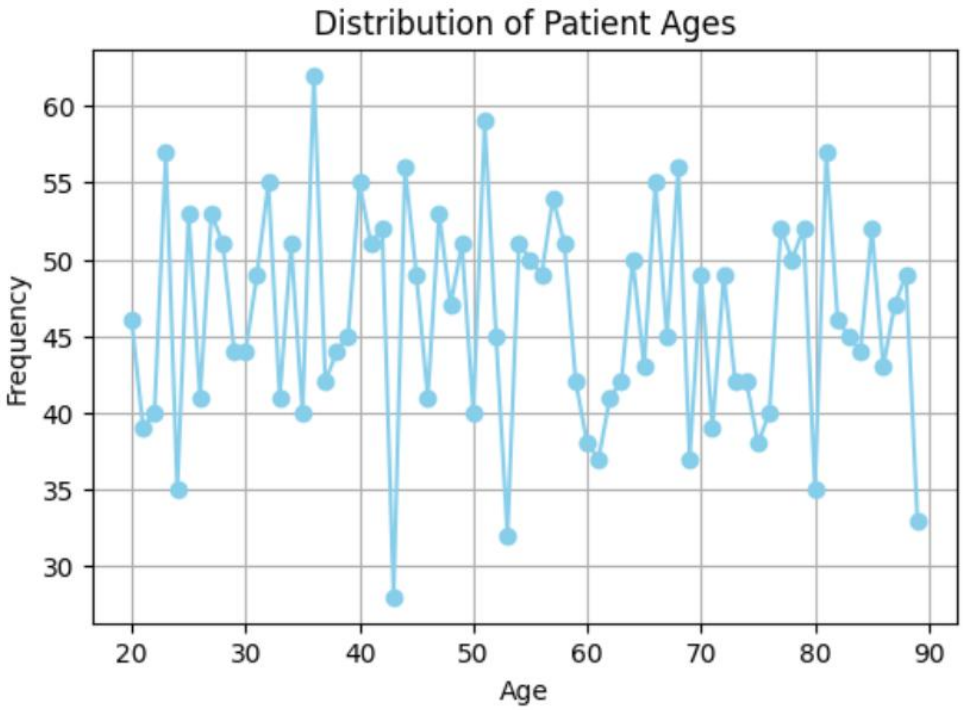

Figure 6: Line Chart of Age Distribution

From this line chart, we can observe that the overall age distribution is relatively even. However, there is a higher concentration of individuals in the 30-40 age range, with the highest frequency exceeding 60 people.

3.3 Disease Type by Gender

We explored the relationship between disease type and gender.

Using the *Seaborn* library, we created a categorical bar chart to display the distribution of different disease types across genders. The horizontal axis represents the disease type, while the vertical axis indicates the count of each disease type .

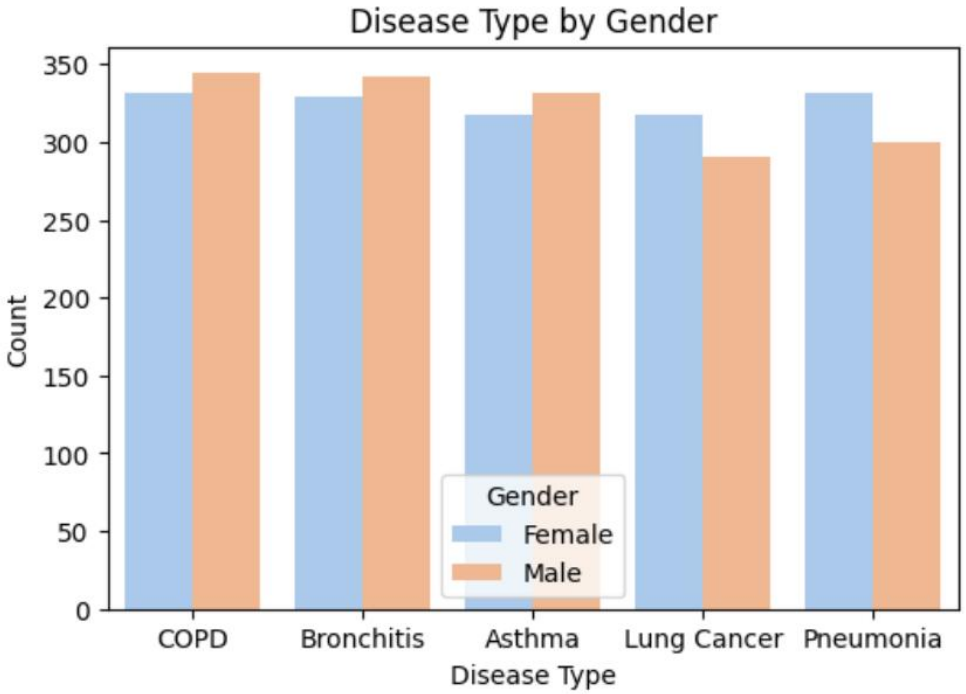

Figure 7: Relationship between Gender and Disease Type

From this chart, we can not only see that the total number of people with COPD is significantly the highest, but also observe different trends by gender. Among females, the number of people with COPD is the highest, while the number of people with Lung Cancer is the lowest. In the male group, the number of people with Pneumonia is the highest, and the number of people with Asthma is the lowest.

3.4 Hospital Visits, Treatment Type, and Recovery

To explore the relationship between the number of hospital visits and the treatment methods as well as whether the condition was cured, we used the *seaborn* library to create a box plot. The type of disease is plotted on the x-axis, the number of hospital visits on the y-axis, and the color of the box indicates whether the patient was cured.

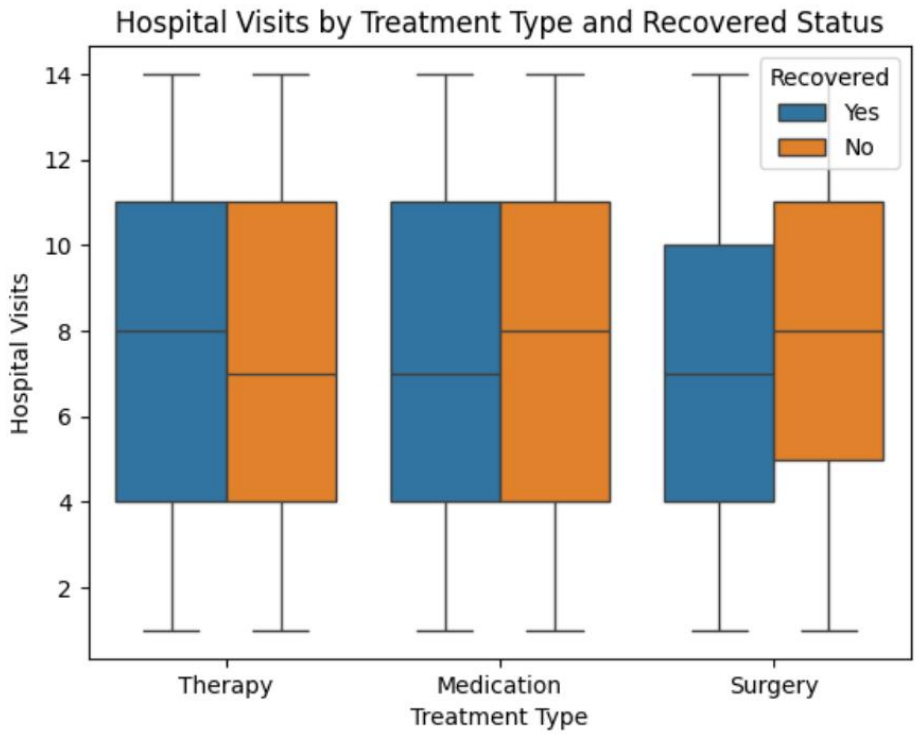

Figure 8: Box Plot of Visits, Treatment Type, and Cure Status

The figure shows that in Therapy, those who recovered had more hospital visits, while in Medication and Surgery, those who didn't recover had more hospital visits. This indicates that the number of visits is not necessarily related to whether a patient is cured， more visits do not guarantee a higher chance of recovery. Additionally, the variation in the number of visits is similar across all treatments, regardless of whether the patient was cured. Notably, the fluctuations in Therapy and Medication are identical, and all distributions are approximately normal.

3.5 Distribution of Recovery Status

To gain the most intuitive understanding of the cure rate for lung diseases, we directly created a pie chart based on whether patients recovered with *matplotlib* library.

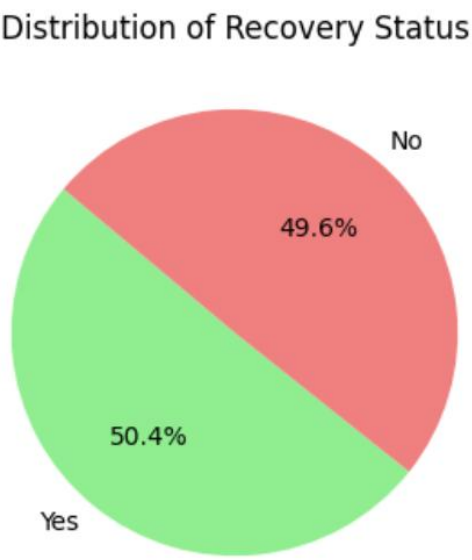

Figure 9: Pie Chart of Recovery Status

From this chart, we can see that nearly half of the patients can be cured, indicating a relatively high recovery rate for lung diseases.

4. Results

Based on the above visual analysis, we have summarized the following results：

- The age distribution of patients is relatively scattered, indicating a wide range of ages affected by lung diseases, with a significant presence across different age groups.
- The gender distribution among various lung diseases (COPD, Bronchitis, Asthma, Lung Cancer, Pneumonia) shows no significant difference in the number of male and female patients for most conditions, suggesting that these diseases affect both genders similarly.
- There is a notable relationship between patient age and the number of hospital visits, which varies by recovery status and disease type. For instance, certain diseases like COPD and Bronchitis may require more frequent hospital visits, especially among older patients.
- The overall recovery status shows that 53.7% of patients recovered, while 46.3% did not. This indicates a relatively balanced outcome, highlighting the need for

further investigation into factors influencing recovery rates.

5. Appendix

In the process of drafting this document in English, AI technology was employed to enhance the linguistic quality and ensure the accuracy and fluency of the text.