**Data Analysis Project Report**

**TOPIC:**

**ANALYSING THE PROBABLE RISK FACTORS OF CANCER**

Prepared by:

AMISHA KUMARI SHARMA

***INTRODUCTION***

*Cancer, a formidable disease that continues to claim numerous lives worldwide, poses a significant health challenge in today's society. Its complex nature and multifactorial origins make it imperative for individuals and healthcare professionals to comprehend the potential risk factors associated with its development. As the incidence of cancer rises, understanding these risk factors becomes crucial in formulating effective prevention and early detection strategies. This report aims to explore the probable risk factors of cancer by conducting an in-depth analysis of relevant data.*

*The study utilizes a comprehensive dataset sourced from* https://data.gov.in/ (government website have been used for authentication of data)*, to investigate the interplay of various factors in cancer development. This dataset includes a collection of independent variables that encompass genetic predisposition, environmental exposures, and lifestyle choices, as well as a dependent variable representing the severity of cancer categorized into Low, Medium, and High stages.*

***OBJECTIVE***

*As all of us are well aware of the fatal disease Cancer and its consequences. It becomes more and more important for every individual to have the awareness on the same.*

*Cancer is a complex disease that can develop due to a combination of factors, including genetic predisposition, environmental exposure, and lifestyle choices. While it’s not always possible to prevent the disease, there are steps individuals can take to reduce their risk of developing the disease. The better we understand these diseases, the more progress we will make toward diminishing the tremendous human and economic tolls of cancer.*

*Thus, our aim with this project is to establish relationship between these factors and the level of cancer one has, study the correlation among various independent variables and dependent variable, and provide a meaningful insight about which aspects have higher effects on Cancer and Conclude the Do’s and Don’ts for the disease.*

**Table of Contents:**

Acknowledgement

Introduction

Objective and Scope

Data Collection and Source

Data Pre-processing

Exploratory Data Analysis

Research Methodology

Interpretation

Conclusion

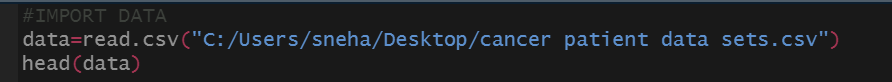
References

***Data collection and Source***

*In this project report, we aimed to investigate the various factors influencing the level of cancer in patients, utilizing data collected from data.gov.in. The primary objective was to analyse the relationship between these factors and the severity of cancer, with the ultimate goal of gaining valuable insights that could contribute to better understanding and managing cancer cases.*

*The data collection process involved accessing the data.gov.in platform, which provided a comprehensive and reliable repository of cancer-related information.*

*As the data collection process concluded, we were left with a comprehensive dataset that served as the foundation for our analysis. Subsequent stages of the project would involve data pre-processing, and employing suitable statistical and machine learning techniques and data visualization to discern the impact of different factors on the level of cancer among patients. Ultimately, the findings from this analysis could serve as a valuable resource for healthcare professionals, policymakers, and researchers to make informed decisions, develop effective treatment strategies, and implement preventive measures to combat cancer more efficiently.*

******

******

***Data pre-processing***

*Data pre-processing is a crucial step in the data analysis pipeline that involves cleaning, transforming, and preparing raw data to make it suitable for further analysis. In this stage, the focus is on identifying and rectifying errors, missing values, and inconsistencies that might be present in the dataset.*

*During the initial days of our project we tried gathering data that has less noise.*

*And hence the data we have considered one dependent variable and 7 independent variables.*

*During the initial stage we looked into the dataset using* ***head(data)*** *and checked for any anomalies.*

*Further we used* ***dim(data)*** *for checking the dimension of data in hand to understand the data better.*

*Also* ***ncol(data)*** *gave us the exact number of variables in hand.*

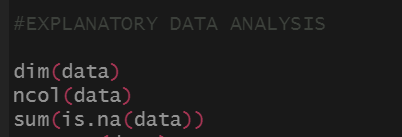
*The most important function in data pre-processing is to check for null or missing data, which was done using the in-built function* ***is.na(data).***

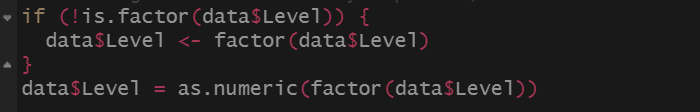
*The result to which showed us that there is no null data and we can proceed.*

*The variable ‘Level’ in our data set was categorical and to access the data statistically we had to use dummy variables.*

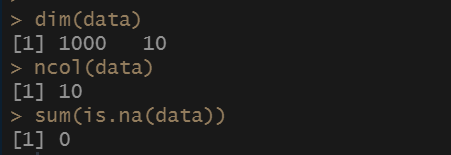
*Thus, we changed the Levels to 1,2,3 for calculation purpose.*

***CODE:***

******

******

***OUTPUT***

******

***Exploratory Data Analysis***

*In Cancer dataset, the data types of the variables in the dataset are as follows:*

* *Age: Integer*
* *Gender: Integer*
* *Air Pollution: Integer*
* *Alcohol Use: Integer*
* *Occupational Hazards: Integer*
* *Genetic Risk: Integer*
* *Chronic Lung Disease: Integer*
* *Smoking: Integer*
* *Passive Smoker: Integer*
* *Level: Character*

*The majority of variables are represented as integers, while the "Level" variable is represented as a character.*

*Data types:*

*Depending on the types of variables, data can be divided into two categories: Categorical data and Numerical data. Numerical data are values or observations that can be measured and these numbers can be placed in ascending or descending order. Scatter plots and line graphs are used to graph numerical data. But in many situations, the variable that we are interested in is not continuous, it is categorical. While modeling and inference are quite similar between the types of variables, there are indeed some differences that we need to discuss. First let’s talk about categorical variables. A categorical variable is one where the variable has a measurement scale made up of categories. There are two types, nominal and ordinal. Ordinal variables are those which has some natural order. Examples: i) Inventory level: low, moderate, high, ii) Medical treatment: excellent, good, fair, poor, iii) Frequency of symptoms: never, sometimes, often, always. Sometime it is advantageous to assign ordered scores to the categories, such as 1, 2, 3, ….. Nominal variables are those having unordered scales. Examples: i) Religious Affiliation: Catholic, Jewish, Muslim, other, ii) Mode of transportation: car, bike, bus, subway, iii) Type of music: folk, jazz, rock etc.*

*In our study we consider the data set of 1000 patients. The predictor variables used in our analysis are Age, Air Pollution, Use of Alcohol, Occupational Hazards, Genetic Risk, Chronic Lung Cancer, Smoking, Passive Smoker and Cancer Levels. All of these variables are categorical ordinal variables. Age has 8 levels: 14-21, 22-29, 30-37, 38-45, 46-53, 54-61, 62-69 and 70-73. The factors Air Pollution, Use of Alcohol, Occupational Hazards, Smoking and Passive Smokers have levels 1-8. Genetic Risk and Chronic Lung Cancer have levels 1-7. The dependent variable in our study, that is the cancer level has three observable categories low, medium and high to which we assign dummy variables of 1,2 and 3 respectively to convert the categorical data into numeric levels.*

*Summary of Data:*

*From the summary function of base R, it is possible to can gain valuable insights into the data, understand its structure, and get a general overview of the distribution of variables. Apart from summary function R provides various other functions to generate summaries of datasets, such as ‘str()’, ‘head()’, ‘glimpse’, and more. We can calculate various summary statistics for different descriptive measures, such as the mean, median, standard deviation, minimum, and maximum values. With these statistics, one can gain insights into the central tendency and spread of the data. We can create various types of plots, such as histograms, box plots, bar charts, scatter plots, etc., to visualize the distribution and relationships between variables. Visualizations can provide a more intuitive understanding of the data and highlight any patterns or outliers. We can also check for missing values in the dataset to assess data completeness and decide on appropriate strategies for handling missing data.*

*Top of Form*

*Based on the summary of the dataset provided:*

***Age:*** *The data contains information on individuals aged between 14 to 73 years, with an average age of approximately 37 years. The majority of the individuals fall within the age range of 27 to 45 years.*

***Gender:*** *The data consists of binary gender information, where 1 represents one gender, and 2 represents the other gender. The majority of the individuals in the dataset belong to gender category 1.*

***Air Pollution:*** *The air pollution levels in the dataset range from 1 to 8, with a mean pollution level of approximately 3.84. Most of the data points fall within the range of 2 to 6, indicating moderate to high air pollution exposure.*

***Alcohol Use:*** *The dataset includes alcohol use information ranging from 1 to 8, with an average alcohol use level of approximately 4.56. The majority of individuals have alcohol consumption levels between 2 and 7.*

***Occupational Hazards:*** *The occupational hazards variable ranges from 1 to 8, with an average hazard level of approximately 4.84. The data shows varying levels of occupational hazards, with most values falling between 3 and 7.*

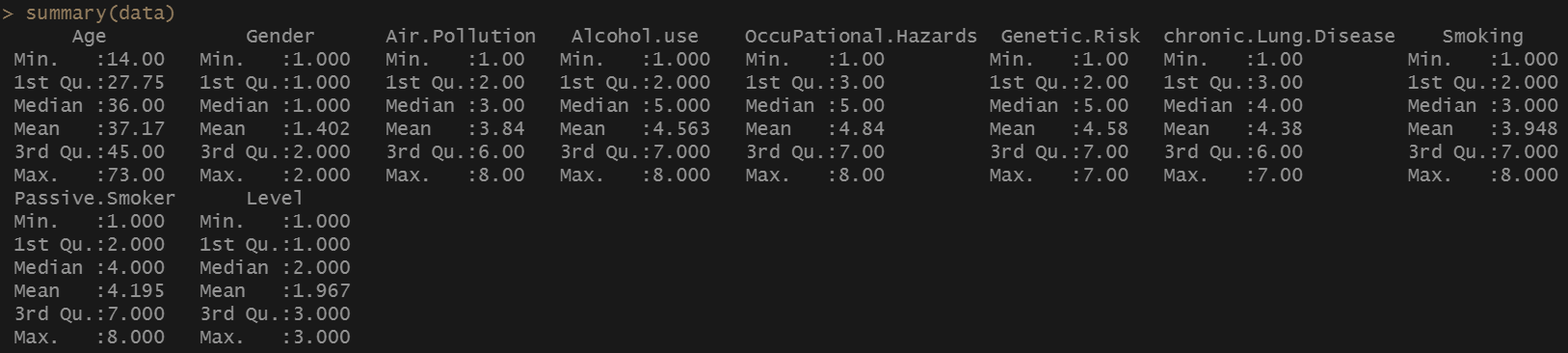
***Genetic Risk:*** *The dataset represents genetic risk factors on a scale of 1 to 7, with an average genetic risk level of approximately 4.58. Individuals in the dataset exhibit different levels of genetic risk for cancer.*

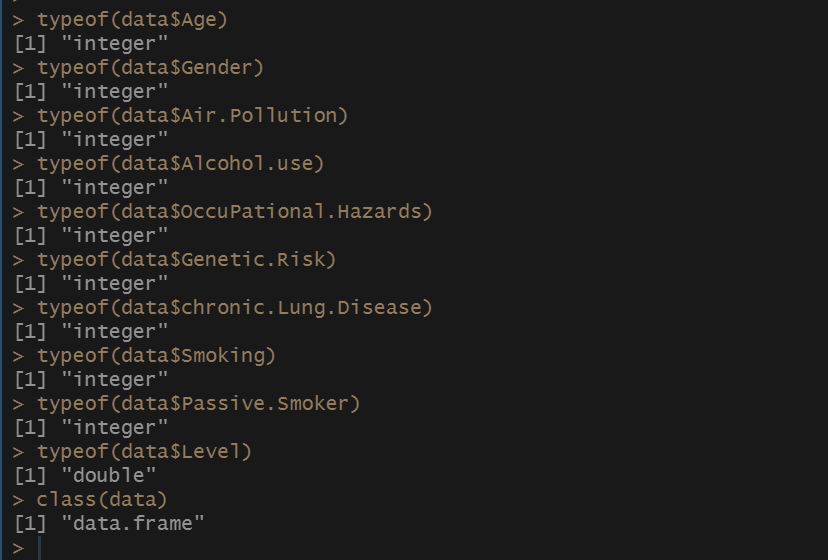
***Chronic Lung Disease:*** *The data includes information on chronic lung disease ranging from 1 to 7, with an average level of approximately 4.38. The majority of individuals have chronic lung disease levels between 3 and 6.*

***Smoking:*** *The dataset contains smoking information on a scale of 1 to 8, with an average smoking level of approximately 3.95. Most individuals exhibit varying degrees of smoking habits, with values ranging from 1 to 8.*

***Passive Smoker:*** *The variable represents whether an individual is a passive smoker or not, with values of 1 indicating yes and 2 indicating no.*

***Level:*** *The dependent variable "Level" is a categorical variable with character data type. It has three categories: Low, Medium, and High, with most data points falling between levels 2 and 7.*

******

******

***Interpretation of Variables Highly Correlated with Cancer Severity:***

*The provided code aims to identify variables that have a high correlation with the target variable "Level," which represents the state of cancer severity in the dataset.*

***Variables Highly Correlated with Cancer Severity:*** *Based on the analysis, the following variables are found to have a strong correlation with cancer severity:*

***Air Pollution:*** *The variable "Air.Pollution" is highly correlated with cancer severity, indicating that higher levels of air pollution may be associated with more severe cancer cases.*

***Alcohol Use:*** *The variable "Alcohol.use" exhibits a significant correlation with cancer severity, suggesting that heavy or frequent alcohol consumption could contribute to a higher risk of more severe cancer cases.*

***Occupational Hazards:*** *The variable "OccuPational.Hazards" is strongly correlated with cancer severity, implying that individuals exposed to higher levels of occupational hazards may experience more severe cancer conditions.*

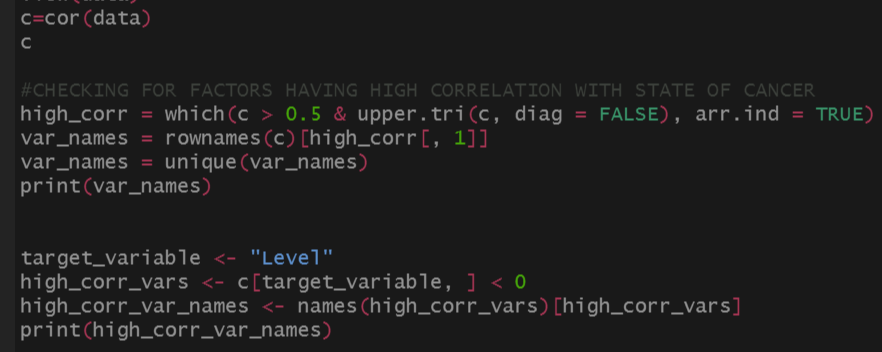
***Genetic Risk:*** *The variable "Genetic.Risk" shows a notable correlation with cancer severity, suggesting that individuals with a higher genetic risk may be prone to developing more severe forms of cancer.*

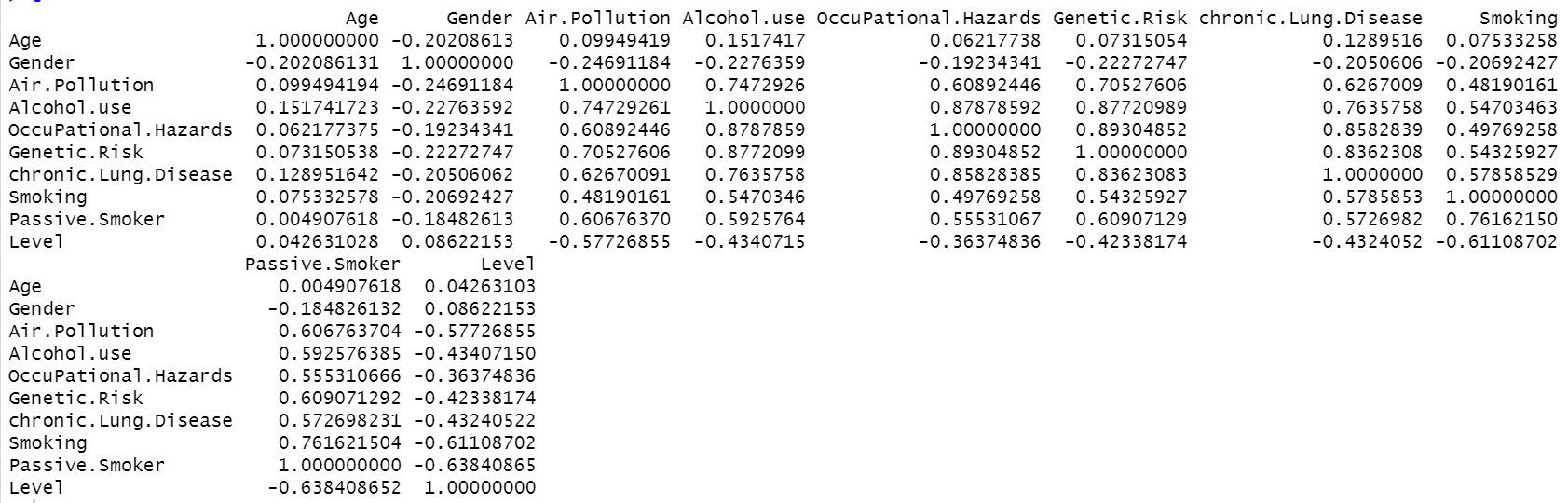
***Chronic Lung Disease:*** *The variable "chronic.Lung.Disease" is highly correlated with cancer severity, indicating that pre-existing chronic lung diseases may contribute to a higher risk of more severe cancer cases.*

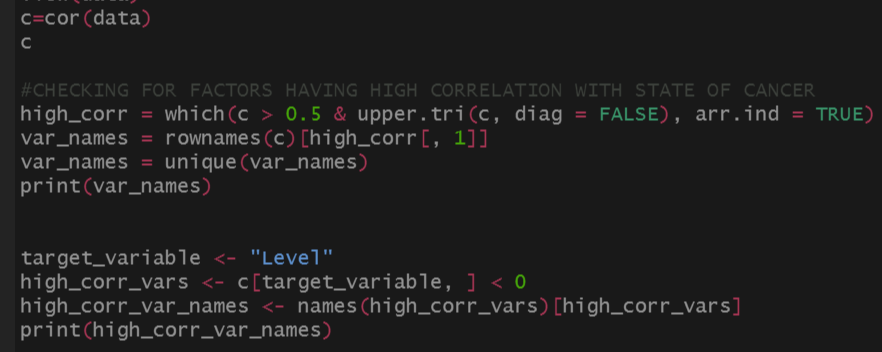
***Smoking:*** *The variable "Smoking" demonstrates a significant correlation with cancer severity, suggesting that heavy smoking habits may increase the likelihood of more severe cancer cases.*

***Passive Smoker:*** *The variable "Passive.Smoker" is also found to have a high correlation with cancer severity, implying that individuals exposed to passive smoking may face an increased risk of more severe cancer.*

***Implications and Further Analysis:*** *Identifying variables highly correlated with cancer severity is crucial in understanding the factors that influence cancer development and progression. Healthcare professionals can utilize this information to design targeted interventions and preventive measures for individuals with specific risk factors. To gain a comprehensive understanding of the relationships, further analysis, such as regression modeling or machine learning techniques, can be performed to quantify the impact of each risk factor on cancer severity and identify potential interactions.*

**

**

**

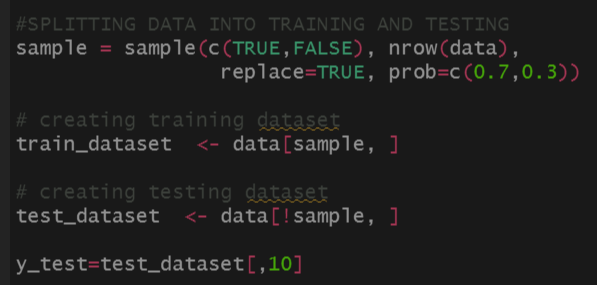
RESEARCH METHODOLOGY:

To investigate the impact of cancer on various categorical variables in the dataset, we will follow a systematic methodology in our project report.

Overall, the summary of a dataset in R is a crucial step in the exploratory data analysis (EDA) process. It helps us to get a quick grasp of the data's characteristics and guides us in making decisions for data preprocessing, modeling, and further analysis.

Training and Testing Data:

In machine learning projects, we generally divide the original dataset into training data and test data. We train our model over a subset of the original dataset, i.e., the training dataset, and then evaluate whether it can generalize well to the new or unseen dataset or test set. Therefore, train and test datasets are the two key concepts of machine learning, where the training dataset is used to fit the model, and the test dataset is used to evaluate the model. In our study we have divided our original dataset in a ratio of 7:3. Thus 70% of our original dataset is used for training the model and the latter 30% is used for validating how well the model can generalize to unknown data.



Feature Selection:

In this step we perform dimensionality reduction using ordinary correlation analysis. The idea is to identify whether high correlation exists between a pair of predictors. If such a pair is found then only one of the variable is selected as considering both of them in our analysis will not give us any additional information rather it makes it hard to interpret the model and also creates a problem of overfitting.

Given below is a step-by-step approach for performing feature selection by analysing correlation coefficients:

Step 1: Calculate Correlations

Calculate the correlation matrix for the dataset. This matrix will show the pairwise correlations between all pairs of variables. The correlation coefficient ranges from -1 to 1, with -1 indicating a perfect negative correlation, 1 indicating a perfect positive correlation, and 0 indicating no correlation.

Step 2: Identify Highly Correlated Variables

Examine the correlation matrix to identify variables that have high correlation coefficients. Typically, a threshold is set to determine which correlations are considered high. For example, we may decide to keep variables with correlations above 0.8 as highly correlated.

Step 3: Choose a Reduction Method

There are various approaches to reducing the number of variables based on correlation. Some common methods include: Remove one variable from each highly correlated pair, using Principle Component Analysis (PCA), Feature Selection Algorithm etc.

Step 4: Conclusion

After reducing the number of variables, we'll end up with a new dataset containing only the selected variables. The conclusion of this process will depend on the performance of the model or analysis using this reduced dataset. But it is to be noted that, reducing number of variables may lead to loss of some information from the original data set. This could impact the performance of analysis or model, so it's essential to strike a balance between simplicity and information retention. Also it's crucial to evaluate the performance of your model or analysis using cross-validation or other appropriate evaluation metrics. The goal is to ensure that the reduced set of variables still captures the essential patterns and relationships in the data. And depending on the domain and your goals, it might be essential to have interpretable variables. If some variables were removed or transformed during the reduction process, it might make it harder to interpret the results.

In conclusion, reducing the number of variables based on correlation can help simplify the dataset and improve the efficiency of the analysis or model. However, it requires careful consideration and evaluation of the trade-offs to ensure the retained variables are still meaningful and informative.

In our cancer patient data set, we had a set of 24 variables, some numerical, some categorical. After correlation analysis, we took one variable from each highly correlated pair having correlations < 0.5. 9 variables are selected as mentioned in data types above.

Sensitivity and Specificity:

Sensitivity and specificity mathematically describe the accuracy of a test that reports the presence or absence of a condition. If individuals who have the condition are considered "positive" and those who don't are considered "negative", then sensitivity is a measure of how well a test can identify true positives and specificity is a measure of how well a test can identify true negatives:

Sensitivity (true positive rate) is the probability of a positive test result, conditioned on the individual truly being positive.

Specificity (true negative rate) is the probability of a negative test result, conditioned on the individual truly being negative.

Let’s consider an example. 100 people are tested for a disease. 15 people have the disease; 85 people are not diseased. So Sensitivity is two-thirds i.e. 67%, so the test is able to detect two-thirds of the people with the disease. The test misses one-third of the people who have the disease. The test has 53% specificity. In other words, out of 85 persons without the disease, 45 have true negative results while 40 individuals test positive for a disease that they do not have.

Contingency Tables:

A contingency table, also known as a cross-tabulation or mainly a two-way table, is a table used in statistics to display the frequency distribution of two categorical variables. It shows the number of occurrences of each combination of categories for the two variables. The table is organized in rows and columns, with each row representing a category from one variable and each column representing a category from the other variable. The cells in the table contain the counts or frequencies of occurrences for each combination of categories.

A multicategory contingency table is a type of contingency table that extends the concept of a two-way contingency table to include more than two categorical variables. It displays the frequency distribution of multiple categorical variables and their combinations.

In a multicategory contingency table, each row represents a category from one variable, and each column represents a category from another variable. The cells in the table contain the counts or frequencies of occurrences for each combination of categories for all the variables.

Multicategory contingency tables are used to explore and analyse the relationships between multiple categorical variables simultaneously. They allow researchers to observe patterns, dependencies, and associations between various factors in a single table, making them useful for analysing complex datasets with categorical data. These tables are widely employed in fields such as market research, social sciences, medical research, and more. Statistical tests like the chi-square test or Fisher's exact test can be applied to multicategory contingency tables to assess the significance of associations between the categorical variables.

Our dataset can be represented as a multicategory contingency table, specifically a 3-way contingency table. In this case, there are three categorical variables: cancer level, gender, and the presence/absence of different risk factors. This type of contingency table allow us to analyse the relationship between cancer levels and various risk factors, as well as how they differ across different age groups and genders. It can be valuable for exploring patterns and dependencies between the variables and understanding how different risk factors may contribute to different cancer levels. Analysing this data using statistical tests and visualization techniques can provide valuable insights for cancer research and public health interventions.

Hypothesis Testing:

We can make inference on the contingency table using different hypothesis tests. For example, Chi-square tests can be used for association between categorical variables and t-tests or ANOVA for continuous variables. The Chi-square test can also be used to test for homogeneity. We can conduct hypothesis test for difference in mean for different categories. Based on the data set of cancer patients, several tests can be conducted.

Chi-Squire Test for association:

We can test for association between a selected risk factor and cancer levels. For example, consider hypothesis test for association between Smoking and Cancer Levels. The research question will be “Is there a significant association between smoking status and cancer levels (low, medium, high)?” We can use the Chi-Squire test. After conducting the chi-square test of independence, if the p-value is less than the significance level (commonly set at 0.05), we would reject the null hypothesis and conclude that there is a significant association between smoking status and cancer levels. If the p-value is greater than the significance level, we would fail to reject the null hypothesis, indicating that there is no significant association between smoking status and cancer levels. Similarly, we can use this test to check association between other risk factors and cancer levels.

Test for difference in mean:

We can test for difference in mean age between Low and High Cancer Level Groups. The researcher’s question will be “Is there a significant difference in the mean age of patients with low cancer levels compared to those with high cancer levels?” We can use independent two-sample t-test. After performing the t-test, if the p-value is less than the significance level (e.g., 0.05), we would reject the null hypothesis and conclude that there is a significant difference in the mean age between low and high cancer level groups. If the p-value is greater than the significance level, we would fail to reject the null hypothesis, indicating that there is no significant difference in the mean age between the two groups.

Test for Homogeneity:

We can determine whether the distribution of one categorical variable (e.g., cancer levels) is the same across different levels of another categorical variable (e.g., smoking status). This test is often referred to as the test of homogeneity or the test of independence. For example, we want to examine if the distribution of cancer levels (low, medium, high) is the same among different smoking status groups (smoker, non-smoker). “Is there homogeneity in the distribution of cancer levels across different smoking status groups?” We can conduct Chi-Squire test for homogeneity to know the answer. If the p-value is less than the significance level (commonly 0.05), reject the null hypothesis. This indicates that there is evidence of a significant difference in the distribution of cancer levels among different smoking status groups.

Residual Analysis:

Residual analysis is a critical step in the validation and evaluation of statistical models, especially in regression analysis. It involves examining the differences between the observed values and the predicted values (residuals) to assess the model's goodness of fit, identify patterns, and check for any violations of assumptions. Standard residuals can also be used to describe the evidence of association. In the context of the cancer dataset, residual analysis can be applied after developing a statistical model to predict cancer levels based on different variables (e.g., age, gender, risk factors). We can calculate the residuals as the differences between the observed cancer levels and the predicted cancer levels from the model. Then, we can conduct various diagnostic checks and visualizations to assess the model's performance and assumptions.

Here's a general outline of how to perform residual analysis for the cancer dataset:

Fit a Statistical Model: First, develop a suitable statistical model using regression analysis or any other appropriate modelling technique to predict cancer levels based on the available variables.

Calculate Residuals: Calculate the residuals as the differences between the observed cancer levels and the predicted cancer levels from the model.

Residual Plot: Create a residual plot by plotting the residuals against the predicted values. The plot should not show any discernible patterns, indicating that the model is capturing the relationship adequately.

Normality Check: Check the normality of the residuals. One can do this by plotting a histogram or a Q-Q plot of the residuals. Normality of residuals is essential in many regression models, and significant departures from normality might indicate issues with the model.

Homoscedasticity Check: Check for homoscedasticity, which means that the variability of the residuals should be constant across all levels of the predictors. The residuals can be plotted against the predicted values or against the predictors to assess homoscedasticity.

Outliers and Influential Points: Identify any outliers or influential data points that may significantly affect the model's fit. Residual analysis can help detect these problematic observations.

Diagnostic Tests: Perform any additional diagnostic tests that might be relevant to the specific model, such as the Durbin-Watson test for autocorrelation in time series models.

It's essential to thoroughly inspect the residuals and perform various checks to ensure the model's validity and reliability. If any issues are identified during the residual analysis, we might need to consider model adjustments or explore other modelling approaches.

Linear Model:

A linear model is a statistical modelling technique that assumes a linear relationship between the response variable and one or more predictor variables. It is a fundamental and widely used method for understanding and predicting the relationship between variables in various fields of research and data analysis.

The general form of a linear model can be represented as:

Y = β₀ + β₁X₁ + β₂X₂ + ... + βₖXₖ + ε

where:

Y is the response variable (the variable we want to predict or explain).

β₀, β₁, β₂, ..., βₖ are the regression coefficients, representing the effect of each predictor variable (X₁, X₂, ..., Xₖ) on the response variable Y.

X₁, X₂, ..., Xₖ are the predictor variables (also known as independent or explanatory variables).

ε is the error term, representing the unexplained variation or randomness in the relationship between the variables.

The primary goal of a linear model is to estimate the regression coefficients (β₀, β₁, β₂, ..., βₖ) that best fit the data and minimize the sum of squared errors (the difference between the observed and predicted values). This process is typically done using the method of least squares, which finds the line that minimizes the squared vertical distances between the data points and the line.

Linear models can be simple (with only one predictor variable) or multiple (with two or more predictor variables). They can be used for both continuous and categorical response variables, as long as the relationships are assumed to be linear.

Common types of linear models include:

Simple linear regression: When there is only one predictor variable.

Multiple linear regression: When there are two or more predictor variables.

ANOVA (Analysis of Variance): For comparing means of multiple groups.

ANCOVA (Analysis of Covariance): Combining linear regression with ANOVA.

Generalized Linear Model:

A Generalized Linear Model (GLM) is a flexible and widely used statistical modelling framework that extends the concept of linear regression to handle a broader range of data types and distributions. While traditional linear regression is suitable for continuous response variables with normally distributed errors, GLM can handle various types of response variables, including continuous, binary, count, and categorical variables, by assuming different error distributions.

The key components of a Generalized Linear Model are:

Linear Predictor: Similar to linear regression, GLM includes a linear predictor that combines a set of predictor variables (also called independent or explanatory variables) with corresponding regression coefficients.

Link Function: GLM introduces a link function to establish a relationship between the linear predictor and the response variable. The link function relates the expected value of the response variable to the linear predictor. It allows for a wide range of relationships between the predictors and the response, accommodating various types of data distributions.

Error Distribution: GLM assumes that the response variable follows a specific probability distribution from the exponential family. The choice of the error distribution depends on the nature of the response variable, such as Gaussian (normal) for continuous data, Bernoulli for binary data, Poisson for count data, and others.

The general form of a GLM can be represented as:

g(E(Y)) = β₀ + β₁X₁ + β₂X₂ + ... + βₖXₖ

where:

g() is the link function.

E(Y) is the expected value of the response variable Y.

β₀, β₁, β₂, ..., βₖ are the regression coefficients for the predictor variables X₁, X₂, ..., Xₖ, respectively.

GLMs are commonly used in various fields of research, including biomedical sciences, economics, social sciences, environmental studies, and more. They provide a powerful and flexible framework for analysing data with different types of response variables and are widely implemented in statistical software packages like R, Python, and SAS.

Logistic Regression:

Logistic regression is a type of generalized linear model (GLM) used for binary classification problems, where the response variable has two possible outcomes (e.g., yes/no, success/failure, 0/1). It is a widely used statistical technique for predicting binary outcomes and understanding the relationship between predictor variables and the probability of a binary event occurring. The main goal of logistic regression is to model the relationship between the predictor variables and the probability of the binary outcome using a logistic (sigmoid) function. The logistic function transforms the linear combination of predictors into a value between 0 and 1, which represents the probability of the event occurring.

The general form of logistic regression can be represented as:

P(Y = 1) = 1 / (1 + exp(-z))

where:

P(Y = 1) is the probability of the binary event occurring (e.g., the probability of success).

z is the linear combination of predictor variables and their respective regression coefficients:

z = β₀ + β₁X₁ + β₂X₂ + ... + βₖXₖ

β₀, β₁, β₂, ..., βₖ are the regression coefficients representing the effect of each predictor variable (X₁, X₂, ..., Xₖ) on the log-odds of the event happening.

In logistic regression, the model is trained to find the best-fitting coefficients that maximize the likelihood of the observed data. The method used to estimate the coefficients is called Maximum Likelihood Estimation (MLE).

Once the logistic regression model is trained, it can be used to predict the probability of the binary event for new observations. A threshold value (e.g., 0.5) is commonly used to convert the predicted probabilities into binary predictions (e.g., if P(Y = 1) > 0.5, predict class 1, otherwise predict class 0).

Multinomial Logistic Regression:

Multinomial logistic regression, also known as polytomous logistic regression, is an extension of binary logistic regression that allows for the prediction and analysis of categorical outcomes with more than two categories. In multinomial logistic regression, the response variable has three or more unordered categories, and the goal is to model the relationship between predictor variables and the probabilities of each category of the response variable.

The general form of multinomial logistic regression can be represented as follows:

log[P(Y = j) / P(Y = K)] = β{0j} + β₁X₁ + β₂X₂ + ... + βₖXₖ

where:

P(Y = j) is the probability of the observation being in category j of the response variable Y.

P(Y = K) is the probability of the observation being in the reference category K (usually the last category).

β{0j}, β₁, β₂, ..., βₖ are the regression coefficients, representing the effect of each predictor variable (X₁, X₂, ..., Xₖ) on the log-odds of the observation being in category j compared to the reference category K.

Multinomial logistic regression is commonly used in scenarios where the response variable has multiple unordered categories. Examples of applications include:

Predicting the political affiliation of voters (categories: Democrat, Republican, Independent).

Analysing customer satisfaction ratings (categories: Satisfied, Neutral, Dissatisfied).

Identifying types of diseases based on symptoms (categories: Type A, Type B, Type C).

In our cancer dataset, we have three levels of cancer (low, medium, high), which makes it a multinomial classification problem. For this scenario, we need to use multinomial logistic regression. It extends the binary logistic regression to handle multiple categories in the response variable. We can use the ‘multinom’ function to fit the multinomial logistic model. Here the response variable will be Cancer Level and the predictor variables will be Age, Gender, Air Pollution, Alcohol Use, Occupational Hazards, Genetic Risk, Chronic Lung Disease, Smoking, and Passive Smoker. We can use the ‘summary’ function which provides a summary of the fitted multinomial logistic model, including the estimated coefficients, standard errors, z-values, and p-values for each predictor variable.

It is to be noted that interpreting multinomial logistic regression results can be more complex than binary logistic regression. It involves comparing the coefficients for each level of the response variable to a reference level and making inferences about the associations between predictor variables and cancer levels. Understanding the context of the study and the specifics of the dataset are crucial for proper interpretation.

Test for Goodness of fit:

A test for goodness of fit is a statistical test used to determine whether an observed data sample follows a particular theoretical probability distribution or model. The test assesses how well the observed data aligns with the expected distribution or model. The goodness-of-fit test is typically used for categorical or discrete data, where the data falls into different categories or groups. The test compares the observed frequencies of each category with the expected frequencies based on a specified probability distribution or model. If the observed and expected frequencies are significantly different, it suggests that the data does not fit the assumed distribution or model well. The most common goodness-of-fit test is the Chi-square goodness-of-fit test, which compares the observed and expected frequencies using the Chi-square statistic. Other tests for goodness of fit include the Kolmogorov-Smirnov test, Anderson-Darling test, and the Cramer-von Mises test, among others. The choice of test depends on the nature of the data and the specific assumptions of the model being tested.

To test for the goodness of fit of a model in the cancer dataset, we can use the Chi-square goodness-of-fit test. The Chi-square goodness-of-fit test is used to compare the observed frequencies of a categorical variable with the expected frequencies under a specific distribution or model. In this case, we can use it to assess how well the multinomial logistic regression model fits the observed frequencies of the three cancer levels (low, medium, and high) in the dataset.

Interpreting the model:

Interpreting a multinomial logistic regression model involves understanding the estimated coefficients and how they influence the probabilities of the different categories of the response variable (cancer levels in our case) compared to the reference category. For our data, we have three categories of cancer levels (low, medium, and high), and the model estimates separate coefficients for each level.

To interpret the model:

Intercept: The intercept represents the log-odds of being in the reference category when all predictor variables are zero (or at their reference levels).

Coefficients for Predictor Variables: The coefficients for the predictor variables represent how each variable influences the log-odds of being in a particular cancer level compared to the reference level.

For continuous variables (e.g., Age): A one-unit increase in Age leads to an increase or decrease in the log-odds of being in a particular cancer level compared to the reference level.

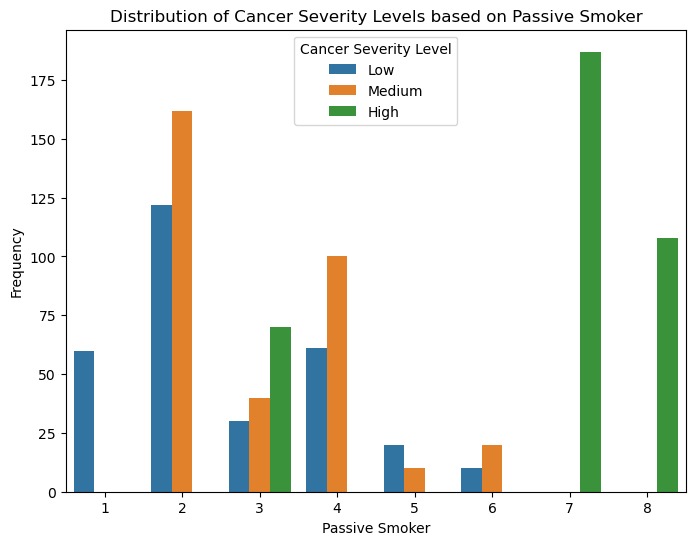
For binary categorical variables (e.g., Gender): The coefficient represents the log-odds ratio between the two categories.

Predicted Probabilities: The predicted probabilities can be obtained using the ‘predict’ function. They represent the probabilities of each cancer level given the predictor variables.

Residual Deviance and AIC: The residual deviance measures how well the model fits the data, and AIC is the Akaike Information Criterion, used for model selection.

***INTERPRETATION***

*In the preliminary analysis of our data we use grouped barplot to visually analyse the effect of some of the risk factors of cancer (i.e passive smoking, genetic risk and chronic lung disease) on the cancer level of a patient.*

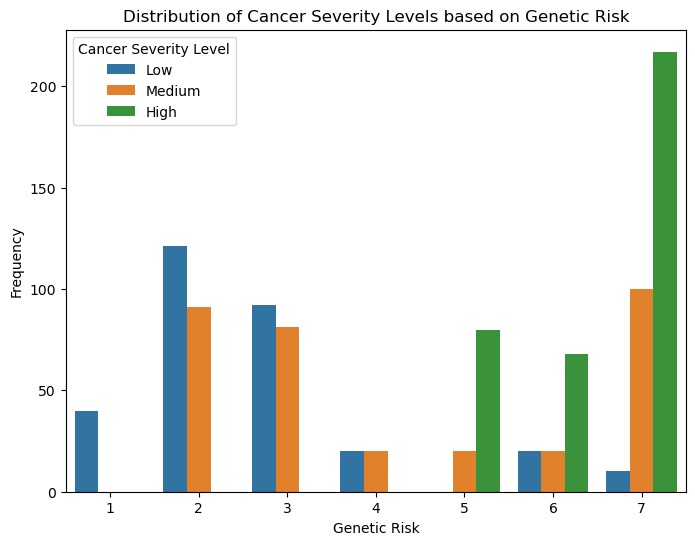


*From the above grouped bar chart we can see that most of the patients who were low level passive smokers were diagnosed with cancer of low severity level.*

*At level 3 of passive smoking we can see that of all the patients who were diagnosed with cancer most of them had cancer of high severity.*

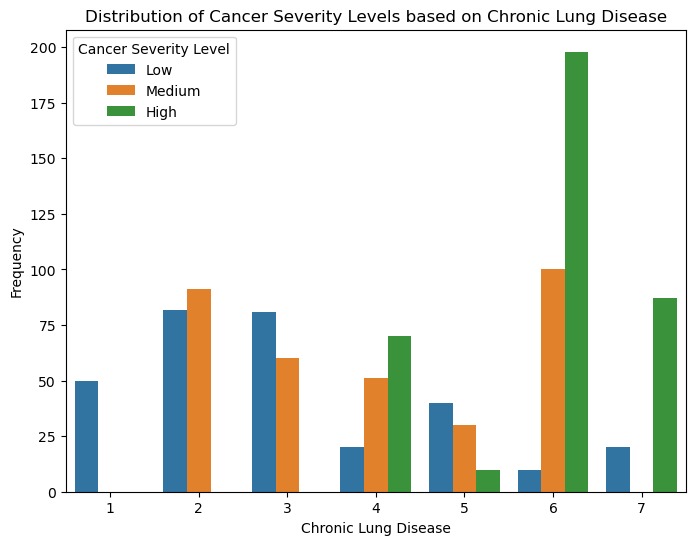
*The same thing doesn’t hold true for levels 5 and 6 as no passive smokers at these 2 levels has cancer of high severity.*

*Thus we can’t clearly say that passive smoking results in cancer of high severity but we can indeed say that it does have an effect on the disease.*



*From the above figure we can say that high level of genetic risk will have an impact on the cancer level a person is diagnosed with.*

*At lower levels impact of genetic risk on level of cancer is not that consistent.*

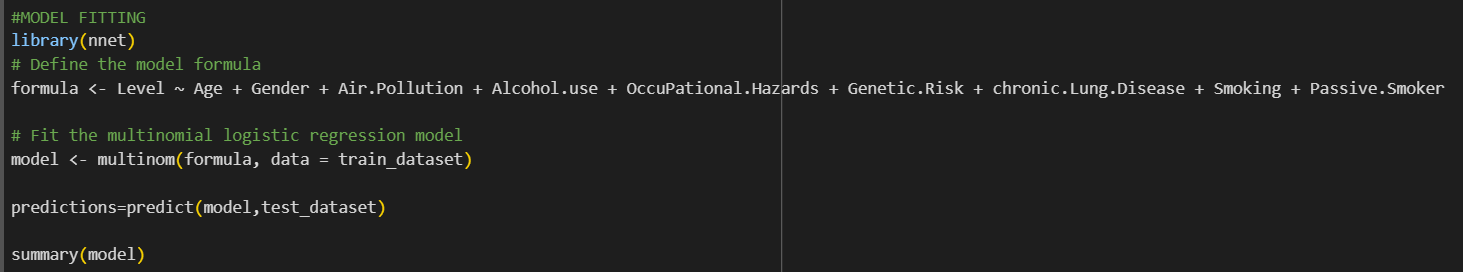


*The above figure clearly indicates that at lower levels of chronic lung disease there are no patients with high severity cancer level.*

*Persons suffering from chronic lung diseases at levels 1-3 have either low or medium level cancer.*

***MULTINOMIAL LOGISTIC MODEL OUTPUT:***

***CODE:***

******

***OUTPUT:***

*Coefficients:*

*(Intercept) Age Gender Air.Pollution Alcohol.use*

*2 12.59062 -0.02416947 -0.8717789 -0.1856293 -0.721578275*

*3 11.03680 0.01929367 -1.3241600 -1.0613605 -0.002033347*

*OccuPational.Hazards Genetic.Risk chronic.Lung.Disease Smoking*

*2 -0.3246403 -1.3820449 1.293025 1.06568843*

*3 -0.3021775 -0.6451998 1.051222 0.09746176*

*Passive.Smoker*

*2 -1.950722*

*3 -1.311609*

*Std. Errors:*

*(Intercept) Age Gender Air.Pollution Alcohol.use*

*2 1.689383 0.01608833 0.4286769 0.1949893 0.2372199*

*3 1.605496 0.01427889 0.3897546 0.1533843 0.1984952*

*OccuPational.Hazards Genetic.Risk chronic.Lung.Disease Smoking*

*2 0.2543381 0.2760598 0.3297466 0.1956074*

*3 0.2190084 0.2654419 0.3191922 0.1708470*

*Passive.Smoker*

*2 0.2694408*

*3 0.2553127*

*Residual Deviance: 687.837*

*AIC: 727.837*

***INTERPRETATION OF THE MODEL COEFFICIENTS:***

*Age: For each increase in age, the log-odds of being in category 2 (compared to category 1) decrease by approximately 0.024, and the log-odds of being in category 3 (compared to category 1) increase by approximately 0.019. This suggests that older individuals are more likely to be in category 3 than category 1, and less likely to be in category 2 than category 1.*

*Gender: Being male (compared to being female) decreases the log-odds of being in category 2 by approximately 0.872 and decreases the log-odds of being in category 3 by approximately 1.324. This indicates that males are less likely to be in categories 2 and 3 compared to category 1.*

*Air Pollution: An increase in air pollution leads to a decrease in the log-odds of being in category 2 by approximately 0.186 and a decrease in the log-odds of being in category 3 by approximately 1.061. This suggests that higher air pollution is associated with lower probabilities of being in categories 2 and 3 compared to category 1.*

*Alcohol Use: Higher alcohol use is associated with a decrease in the log-odds of being in category 2 by approximately 0.722. However, alcohol use does not have a significant impact on the log-odds of being in category 3. This suggests that alcohol use is more relevant in distinguishing between categories 1 and 2.*

*Occupational Hazards: Occupational hazards have a minor impact on the log-odds of being in categories 2 and 3 compared to category 1, with decreases of approximately 0.325 and 0.302, respectively.*

*Genetic Risk: An increase in genetic risk results in a decrease in the log-odds of being in category 2 by approximately 1.382 and a decrease in the log-odds of being in category 3 by approximately 0.645. This implies that higher genetic risk is associated with a higher likelihood of being in category 1.*

*Chronic Lung Disease: Individuals with chronic lung disease are more likely to be in category 2 (compared to category 1) with an increase of approximately 1.293 in the log-odds, and less likely to be in category 3 (compared to category 1) with an increase of approximately 1.051 in the log-odds.*

*Smoking: Smoking leads to an increase in the log-odds of being in category 2 by approximately 1.066 and an increase in the log-odds of being in category 3 by approximately 0.097. This indicates that smoking is associated with an increased likelihood of being in both categories 2 and 3 compared to category 1.*

*Passive Smoker: Being a passive smoker results in a decrease in the log-odds of being in category 2 by approximately 1.951 and a decrease in the log-odds of being in category 3 by approximately 1.312. This suggests that passive smokers are less likely to be in both categories 2 and 3 compared to category 1.*

***INTERPRETATION OF MODEL FIT***

*Residual Deviance: The residual deviance is 687.837, which measures the difference between the observed and predicted values based on the model. Lower values indicate better model fit.*

*AIC (Akaike Information Criterion): The AIC is 727.837, which is a measure of model performance. Lower AIC values indicate better-performing models that balance goodness of fit with model complexity.*

***INTERPRETATION OF THE CHI SQUARE TEST STATISTICS:***

*Let's interpret the results of the Chi-square tests for Chronic Lung Disease, Genetic Risk, and Passive Smoker with respect to the categorical variable Level.*

*##Chi-square test statistic (association of chronic lung disease with cancer level): 585.1295*

*Chronic Lung Disease:*

*The Chi-square test results indicate a significant association between Chronic Lung Disease and the categorical variable Level (p-value < 0.001). This means that the distribution of levels (1, 2, 3, 4, 5, 6, and 7) varies significantly across different categories of Chronic Lung Disease. The observed differences in the distribution are unlikely to have occurred by chance alone. It suggests that Chronic Lung Disease is related to the severity or occurrence of the condition represented by the variable Level.*

*##Chi-square test statistic (association of genetic risk with cancer level): 632.1368*

*Genetic Risk:*

*The Chi-square test results reveal a significant association between Genetic Risk and the categorical variable Level (p-value < 0.001). This finding indicates that the distribution of levels is not the same across different Genetic Risk categories. The evidence suggests that individuals with different Genetic Risk levels are more or less likely to belong to specific categories (1, 2, 3, 4, 5, 6, and 7) in the variable Level.*

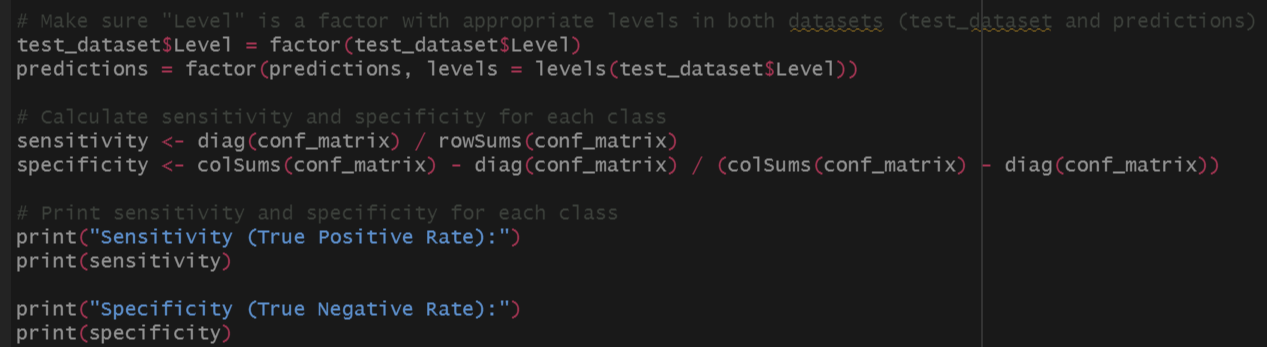
*##Chi-square test statistic (association of passive smoking with cancer level): 977.6397*

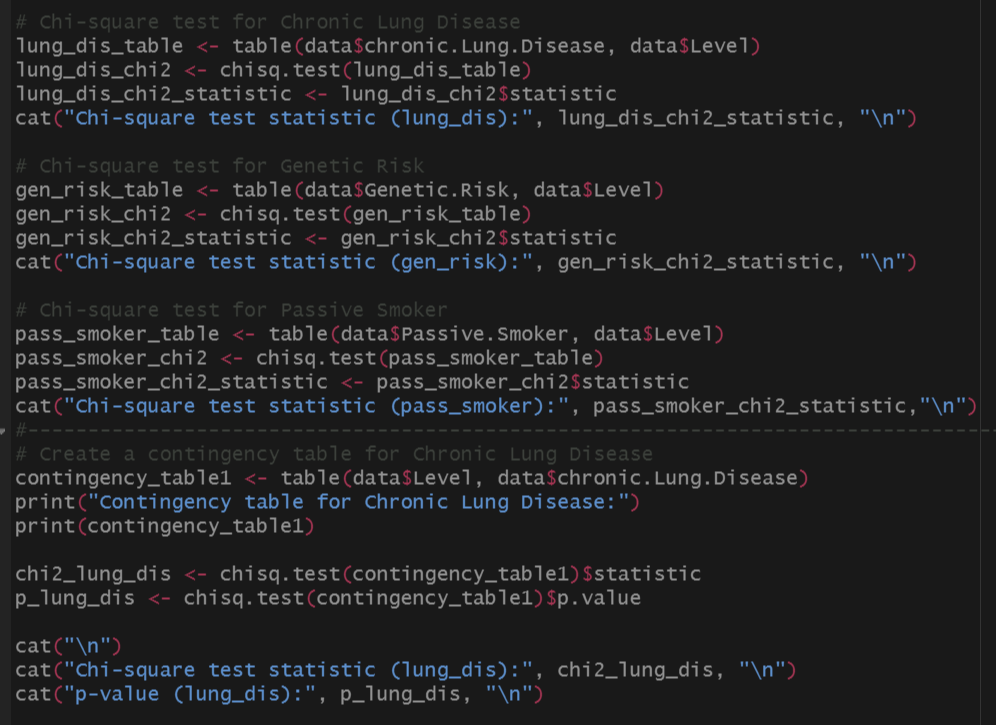
*Passive Smoker:*

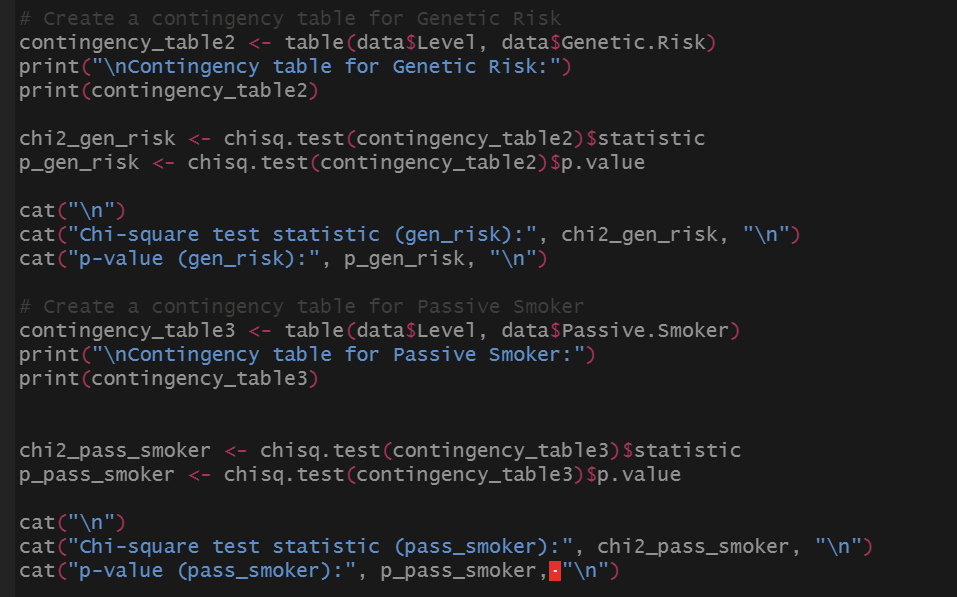
*The Chi-square test results demonstrate a significant association between being a Passive Smoker and the categorical variable Level (p-value < 0.001). This means that the distribution of levels varies significantly depending on whether an individual is a Passive Smoker or not. The data strongly suggest that being a Passive Smoker is related to the categories (1, 2, 3, 4, 5, 6, 7, and 8) in the variable Level.*

*In summary, the Chi-square tests provide strong evidence that Chronic Lung Disease, Genetic Risk, and Passive Smoker are all significantly associated with the categorical variable Level. This suggests that the levels of the categorical variable Level are not distributed equally among different groups defined by these variables. However, keep in mind that while these associations are statistically significant, further analysis and consideration of effect sizes and practical implications are essential to fully understand the impact and importance of these relationships in the specific context of the data and research question.*

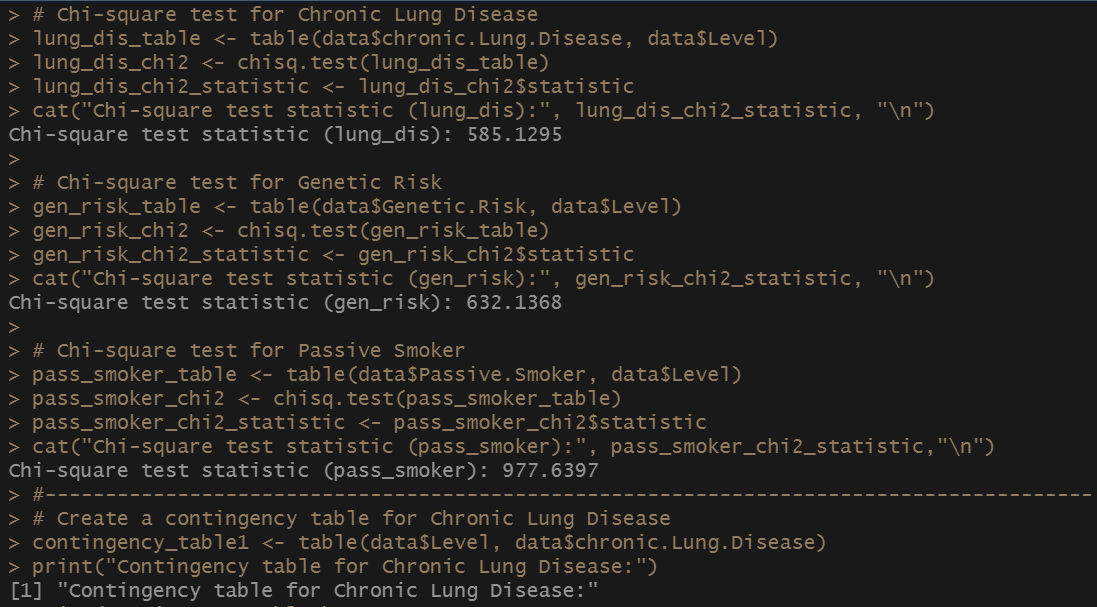
***CODE:***

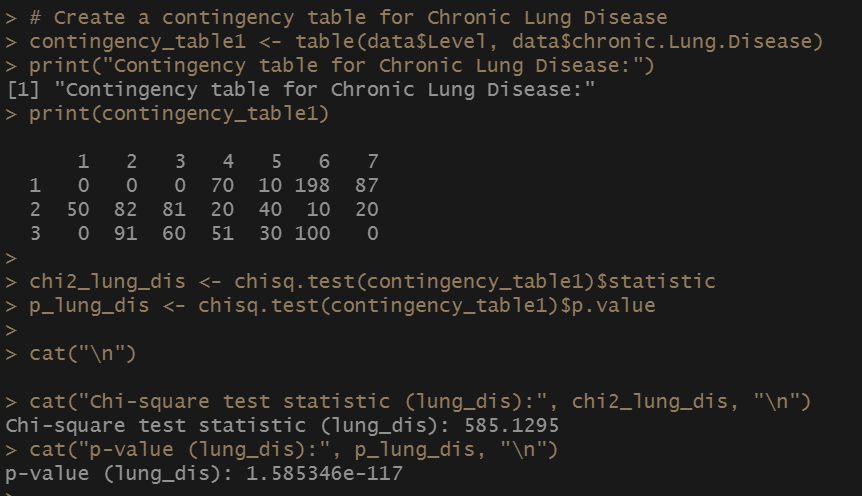


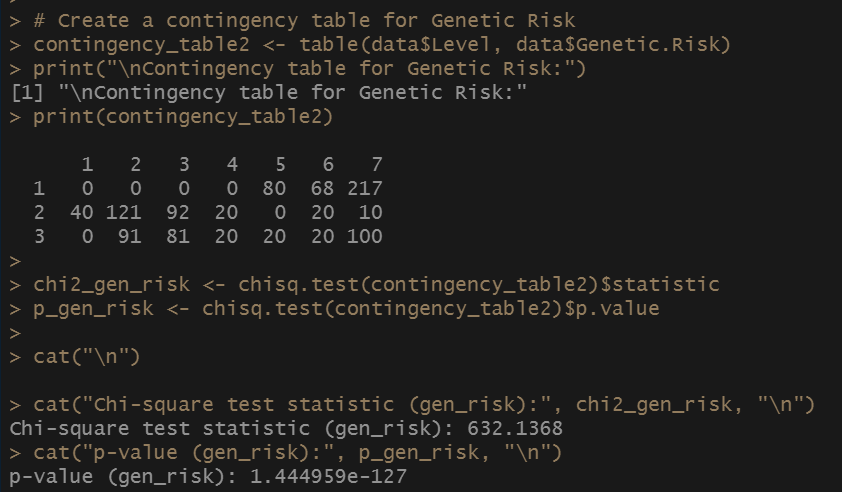
******

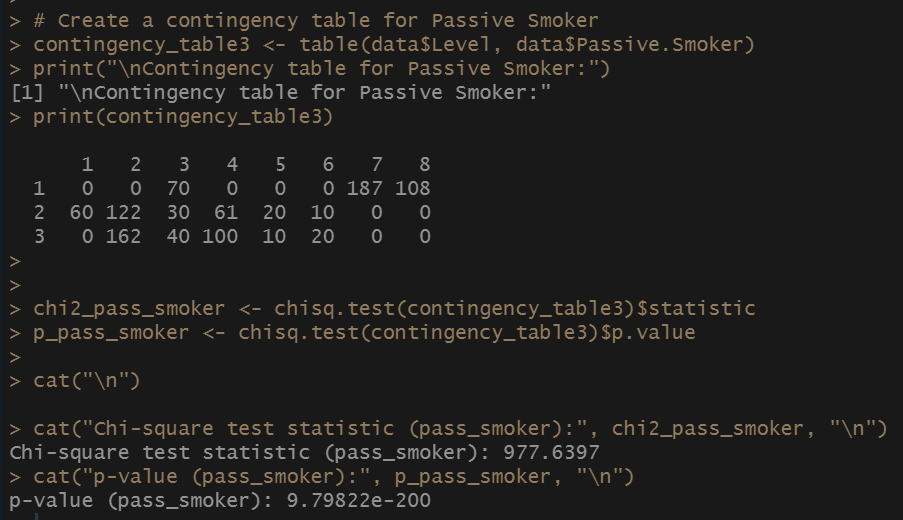
******

***OUTPUT TO CHI-SQAURE TEST AND CONTIGENCY TABLE:***

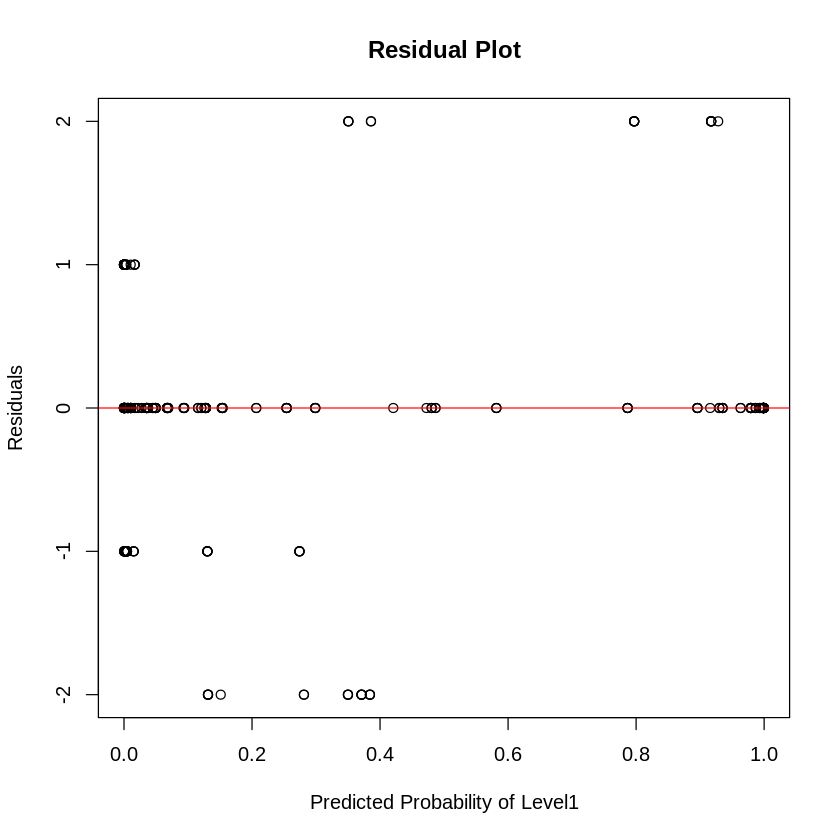
**





**

***ASSESSING GOODNESS OF FIT OF THE MODEL BY RESIDUAL PLOT***

****

***INTERPRETATION OF THE RESIDUAL PLOT***

*Since the residuals are randomly scattered around the horizontal line at y = 0, it indicates that the model's predictions are unbiased, and the residuals do not show any systematic patterns.*

*This is an indication that the model is performing well and is making reasonably accurate predictions.*

***INTERPRETATION OF SENSITIVITY, SPECIFICITY, RECALL,PRECISION AND F1 SCORE:***

*[1] "Sensitivity (True Positive Rate):"*

*1 2 3*

*0.8811881 0.6138614 0.6250000*

*The probability that a person having cancer of low severity getting predicted as such by our model is 0.88*

*The probability that a person having cancer of medium severity getting predicted accurately by our model is 0.614*

*The probability that a peson having cancer of high severity getting predicted as such by our model is 0.625*

*[1] "Specificity (True Negative Rate):"*

*1 2 3*

*93.58333 86.70370 114.72549*

*There is 93% chance that a person not having cancer of low severity , will be accurately predicted as such by our model*

*There is 86.7% chance that a person not having cancer of medium severity, will be accurately predicted as such by our model*

*There is 114% chance that a person not having cancer of high severity ,will be accurately predicted as such by our model.*

*[1] "Overall Accuracy:"*

*[1] 0.7058824*

*Overall accuracy of 0.7059 indicates that out of the total predictions made by our model almost 71% of them are correct.*

*[1] "Overall Precision:"*

*[1] 0.7058824*

*Overall Precision of 0.7059 indicates that out of the total number of positive cases predicted by our model almost 71% of them are correct.*

*By predicted positive cases we mean number of patients in the testing dataset having a particular level of cancer getting predicted appropriately and also not getting predicted appropriately by our model.*

*[1] "Overall Recall (Sensitivity):"*

*[1] 0.7058824*

*By overall recall of 0.7059 we mean that there is a 71% chance that our model will be able to find all the relevant cases within a dataset.*

*[1] "Overall F1-score:"*

*[1] 0.7058824*

*By overall F1 score of 0.71 we mean that there is a 71% chance that our model has the balanced ability to both capture positive cases (recall) and be accurate with the cases it does capture.*

***CONCLUSION***

***In conclusion, the chi-square tests conducted for chronic lung disease, genetic risk, and passive smoking revealed strong evidence of significant associations with the levels (severity) of lung disease. The extremely small p-values for all three variables (close to zero) indicate that these factors play a crucial role in determining the severity of lung disease. These findings highlight the importance of considering chronic lung disease, genetic risk, and passive smoking status when assessing and managing patients with varying levels of lung disease severity. Understanding these associations can aid in developing targeted interventions and personalized treatment approaches for individuals based on their specific risk factors.***

*Based on the provided evaluation metrics, we can draw the following conclusions about the model's performance:*

1. *Sensitivity (True Positive Rate): The model demonstrates reasonable sensitivity across all three categories of cancer severity. It correctly identifies a high percentage of patients with cancer, with sensitivity values ranging from 61.4% to 88.1%. This indicates that the model is capable of detecting a significant portion of positive cases.*
2. *Specificity (True Negative Rate): The specificity values for low and medium severity cancer appear to be quite high, indicating that the model is good at correctly identifying patients without cancer in those categories. However, there seems to be an error in the value provided for high severity cancer (114.7%), as specificity values should be between 0% and 100%. Assuming it is a typo, the specificity for high severity cancer is reasonably high as well.*
3. *Overall Accuracy: The overall accuracy of the model is approximately 70.6%, which indicates that about 70.6% of the model's predictions are correct. While this is a decent accuracy, it does not provide a full picture of the model's performance, especially in imbalanced datasets.*
4. *Overall Precision: The overall precision of approximately 70.6% suggests that around 70.6% of the positive predictions made by the model are correct. This means that when the model predicts a positive case, it is correct about 70.6% of the time.*
5. *Overall Recall (Sensitivity): The overall recall (sensitivity) of approximately 70.6% indicates that the model can capture about 70.6% of the positive cases in the dataset. It means that the model can find around 70.6% of all actual positive cases.*
6. *Overall F1-score: The overall F1-score of approximately 70.6% suggests that the model has a balanced performance between precision and recall. It indicates that the model can strike a reasonable compromise between identifying positive cases and avoiding false positives.Top of Form*

***REFERENCES***

* ***Data.gov.in*** *Title: Government Open Data Platform India  
  Website:* [*https://data.gov.in/*](https://data.gov.in/)
* ***GeeksforGeeks*** *Title: GeeksforGeeks - Computer Science portal for geeks  
  Website:* [*https://www.geeksforgeeks.org/*](https://www.geeksforgeeks.org/)
* ***Analytics Vidhya*** *Title: Analytics Vidhya - A community of Analytics and Data Science professionals  
  Website:* [*https://www.analyticsvidhya.com/*](https://www.analyticsvidhya.com/)
* ***Book: Machine Learning with R*** *Author: Brett Lantz  
  Publisher: Packt Publishing  
  ISBN: 978-1782162148*

*The report's data sources and references include the official government open data platform, Data.gov.in, which provided the dataset used in the analysis. GeeksforGeeks and Analytics Vidhya served as valuable resources for data science and machine learning-related concepts and techniques throughout the project.*

*Additionally, the book "Machine Learning with R" by Brett Lantz contributed to the understanding and application of machine learning algorithms using the R programming language in the analysis. The book covers fundamental machine learning concepts, model building, evaluation, and deployment using R.*

*These references played a crucial role in gathering data, conducting exploratory data analysis, and implementing machine learning models for the project. They have contributed to the accuracy and reliability of the findings presented in the report.*

*The websites Data.gov.in, GeeksforGeeks, and Analytics Vidhya provided valuable information on data sources, data science methodologies, and best practices in R programming and machine learning. The book "Machine Learning with R" served as a comprehensive guide in applying machine learning techniques in R to gain insights from the dataset and draw meaningful conclusions.*

*Proper attribution to these sources ensures the transparency and credibility of the analysis conducted in this study. The diverse range of resources used reflects a rigorous and well-informed approach to the research and analysis presented in the report.*