
# Asthma Diagnosis Classification

**CT127-3-2-PFDA-LAB-20**

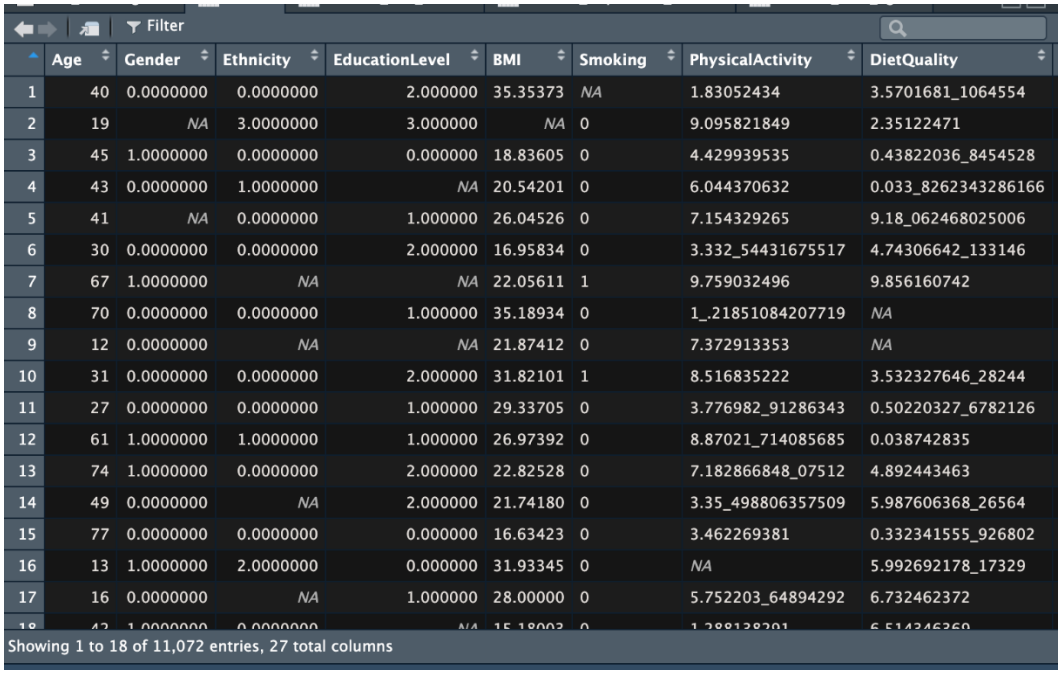
**Programming for Data Analysis**

| | |
|---|---|
| **Intake** | **APU2F2406CS(AI)** |
| **Hand Out Date** | |
| **Hand In Date** | 15th September, 2024 |
| **Lecturer** | **DR. MINNU HELEN JOSEPH** |
| **Students** | - **HTET AUNG HLAING (TP075706)**<br>- **JONATHAN NG'UA MBAI (TP075128)**<br>- **LEHWIKS RAJ A/L GUNASELAN THANGARAJ (TP061519)**<br>- **FARES ADEL OMER BA MOHRES (TP071376)** |

# Table of Contents

# 1. Introduction

## 1.1 Data Column Identification



Filter

| | Age | Gender | Ethnicity | EducationLevel | BMI | Smoking | PhysicalActivity | DietQuality |
|---|---|---|---|---|---|---|---|---|
| 1 | 40 | 0.0000000 | 0.0000000 | 2.000000 | 35.35373 | *NA* | 1.83052434 | 3.5701681_1064554 |
| 2 | 19 | *NA* | 3.0000000 | 3.000000 | *NA* | 0 | 9.095821849 | 2.35122471 |
| 3 | 45 | 1.0000000 | 0.0000000 | 0.000000 | 18.83605 | 0 | 4.429939535 | 0.43822036_8454528 |
| 4 | 43 | 0.0000000 | 1.0000000 | *NA* | 20.54201 | 0 | 6.044370632 | 0.033_8262343286166 |
| 5 | 41 | *NA* | 0.0000000 | 1.000000 | 26.04526 | 0 | 7.154329265 | 9.18_062468025006 |
| 6 | 30 | 0.0000000 | 0.0000000 | 2.000000 | 16.95834 | 0 | 3.332_54431675517 | 4.74306642_133146 |
| 7 | 67 | 1.0000000 | *NA* | *NA* | 22.05611 | 1 | 9.759032496 | 9.856160742 |
| 8 | 70 | 0.0000000 | 0.0000000 | 1.000000 | 35.18934 | 0 | 1_.21851084207719 | *NA* |
| 9 | 12 | 0.0000000 | *NA* | *NA* | 21.87412 | 0 | 7.372913353 | *NA* |
| 10 | 31 | 0.0000000 | 0.0000000 | 2.000000 | 31.82101 | 1 | 8.516835222 | 3.532327646_28244 |
| 11 | 27 | 0.0000000 | 0.0000000 | 1.000000 | 29.33705 | 0 | 3.776982_91286343 | 0.50220327_6782126 |
| 12 | 61 | 1.0000000 | 1.0000000 | 1.000000 | 26.97392 | 0 | 8.87021_714085685 | 0.038742835 |
| 13 | 74 | 1.0000000 | 0.0000000 | 2.000000 | 22.82528 | 0 | 7.182866848_07512 | 4.892443463 |
| 14 | 49 | 0.0000000 | *NA* | 2.000000 | 21.74180 | 0 | 3.35_498806357509 | 5.987606368_26564 |
| 15 | 77 | 0.0000000 | 0.0000000 | 0.000000 | 16.63423 | 0 | 3.462269381 | 0.332341555_926802 |
| 16 | 13 | 1.0000000 | 2.0000000 | 0.000000 | 31.93345 | 0 | *NA* | 5.992692178_17329 |
| 17 | 16 | 0.0000000 | *NA* | 1.000000 | 28.00000 | 0 | 5.752203_64894292 | 6.732462372 |

Showing 1 to 18 of 11,072 entries, 27 total columns

*Figure 1.1.1.1 Original Data*

Upon Inspection of the file, it was found that the file contains over 11,072 entries with a total of 27 columns.

By running the command:

```
colnames(data)
```

*Figure 1.1.1.2 Col Name Function on data*

A list of columns inside the csv file is returned:

```
"Age"                  "Gender"               "Ethnicity"
"EducationLevel"       "BMI"                  "Smoking"
"PhysicalActivity"     "DietQuality"          "SleepQuality"
"PollutionExposure"    "PollenExposure"       "DustExposure"
"PetAllergy"           "FamilyHistoryAsthma"  "HistoryOfAllergies"
"Eczema"               "HayFever"             "GastroesophagealReflux"
"LungFunctionFEV1"     "LungFunctionFVC"      "Wheezing"
"ShortnessOfBreath"    "ChestTightness"       "Coughing"
"NighttimeSymptoms"    "ExerciseInduced"      "Diagnosis"
```

*Figure 1.1.1.3 Columns inside CSV File*

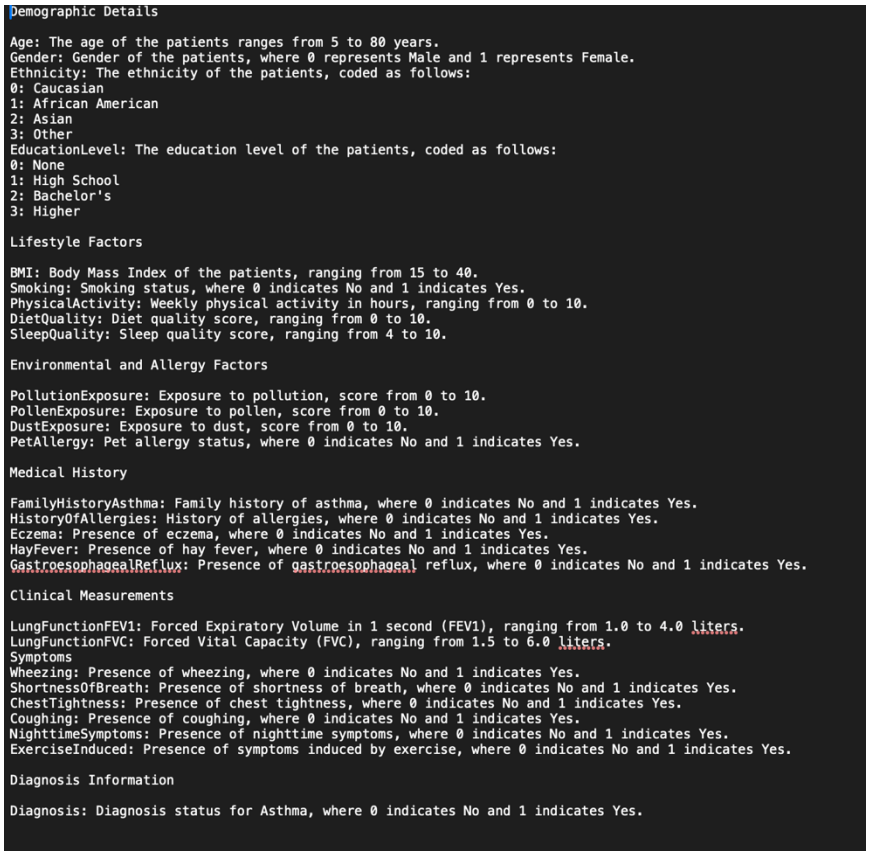



*Figure 1.1.1.4 Requirements*

As per Figure 2.1.1.4's requirements, values can be categorized into three types; Numeric, Logical and Enums.

Within numeric value types, it can further be broken down into 2 subtypes; Integer and Decimals. The only column that is of integer value is "Age" (which is greater or equal to 5 and less than or equal to 80)

The numerous decimal values inside the dataset include:

1. BMI (15-40)
2. Physical Activity (0-10)
3. Diet Quality (0-10)
4. Sleep Quality (0-10)
5. Pollution Exposure (0-10)
6. Pollen Exposure (0-10)
7. Dust Exposure (0-10)
8. LungFunctionFEVC1 (1.0-4.0)
9. LungFunctionFVC (1.5-6.0)

The columns that can be categorized as enums include:

1. Gender (0 for Male, 1 for Female)
2. Ethnicity (0 for Caucasian, 1 for African American, 2 for Asian, 3 for Other)
3. Education Level (0 for None, 1 for High School, 2 for Bachelor's, 3 for Higher)

The remaining columns can all be categorized as Logical values (true or false). This includes:

1. Smoking
2. FamilyHistoryAsthma
3. HistoryOfAllergies
4. Eczema
5. HayFever
6. GastroesophagealReflux
7. Wheezing
8. ShortnessOfBreath
9. ChestTightness
10. Coughing
11. NighttimeSymptons
12. ExerciseInduced
13. Diagnosis

# 2. Documentation

## 2.1 Data Preparation

Before the data analysis process starts, it is essential to prepare the data properly to ensure:

1. Data is clean
2. Data is properly formatted

Before the data preparation process starts, it is important to go through different stages one step after another to understand how the data should be prepared.

1. Data Validation and Reformatting
2. Missing Value Handling

### 2.1.2 Data Validation and Reformatting

First off, minimal data validation is done by replacing N.A values with mode for logical values.

```
data$Gender[is.na(data$Gender)]<-gender_mode
data$Ethnicity[is.na(data$Ethnicity)] <- ethnicity_mode
data$EducationLevel[is.na(data$EducationLevel)] <- education_mode
data$Smoking[is.na(data$Smoking)] <- smoking_mode
data$PetAllergy[is.na(data$PetAllergy)] <- petallergy_mode
data$FamilyHistoryAsthma[is.na(data$FamilyHistoryAsthma)] <- familyhistory_mode
data$HistoryOfAllergies[is.na(data$HistoryOfAllergies)] <- historyofallergies_mode
data$Eczema[is.na(data$Eczema)] <- eczema_mode
data$HayFever[is.na(data$HayFever)] <- hayfever_mode
data$GastroesophagealReflux[is.na(data$GastroesophagealReflux)] <- gastro_mode
data$Wheezing[is.na(data$Wheezing)] <- wheezing_mode
data$ShortnessOfBreath[is.na(data$ShortnessOfBreath)] <- shortnessofbreath_mode
data$ChestTightness[is.na(data$ChestTightness)] <- chesttightness_mode
data$Coughing[is.na(data$Coughing)] <- coughing_mode
data$NighttimeSymptoms[is.na(data$NighttimeSymptoms)] <- nighttimesymptons_mode
data$ExerciseInduced[is.na(data$ExerciseInduced)] <- exerciseinduced_mode
data$Diagnosis[is.na(data$Diagnosis)] <- diagnosis_mode
```

*Figure 2.1.2.1 R Snippet of replacing N.A values with mode for Logical value*

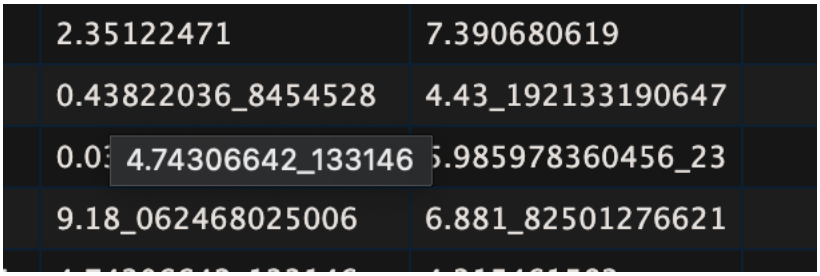



*Figure 2.1.2.2 malformed decimal values in the dataset*

For numeric values, it was essential to convert malform values as shown in Figure 2.1.2.2 to proper format by removing the '_' from the value, and converting them into proper numerical type. Right after the operation, the cell with empty rows are replaced with the mean value of the column as per Figure 2.1.2.3.

```r
data$PhysicalActivity <- gsub("_", "", data$PhysicalActivity)
data$PhysicalActivity <- as.numeric(data$PhysicalActivity)
data$PhysicalActivity[is.na(data$PhysicalActivity)]<-mean(data$PhysicalActivity,na.rm=TRUE)

data$DietQuality <- gsub("_", "", data$DietQuality)
data$DietQuality <- as.numeric(data$DietQuality)
data$DietQuality[is.na(data$DietQuality)]<-mean(data$DietQuality,na.rm=TRUE)

data$SleepQuality <- gsub("_", "", data$SleepQuality)
data$SleepQuality <- as.numeric(data$SleepQuality)
data$SleepQuality[is.na(data$SleepQuality)]<-mean(data$SleepQuality,na.rm=TRUE)

data$PollenExposure <- gsub("_", "", data$PollenExposure)
data$PollenExposure <- as.numeric(data$PollenExposure)
data$PollenExposure[is.na(data$PollenExposure)]<-mean(data$PollenExposure,na.rm=TRUE)

data$DustExposure <- gsub("_", "", data$DustExposure)
data$DustExposure <- as.numeric(data$DustExposure)
data$DustExposure[is.na(data$DustExposure)]<-mean(data$DustExposure,na.rm=TRUE)

data$LungFunctionFEV1 <- gsub("_", "", data$LungFunctionFEV1)
data$LungFunctionFEV1 <- as.numeric(data$LungFunctionFEV1)
data$LungFunctionFEV1[is.na(data$LungFunctionFEV1)]<-mean(data$LungFunctionFEV1,na.rm=TRUE)

data$LungFunctionFVC <- gsub("_", "", data$LungFunctionFVC)
data$LungFunctionFVC <- as.numeric(data$LungFunctionFVC)
data$LungFunctionFVC[is.na(data$LungFunctionFVC)]<-mean(data$LungFunctionFVC,na.rm=TRUE)

data$PollutionExposure[is.na(data$PollutionExposure)]<-mean(data$PollutionExposure,na.rm=TRUE)
data$BMI[is.na(data$BMI)]<-mean(data$BMI,na.rm=TRUE)
```

*Figure 2.1.2.3 R Snippet of replacing N.A values with mean for Numeric value*

After that, enum columns are converted to their respective representative values with the snippet shown in figure 2.1.2.4.

```r
ethnicity_labels <- c("Caucasian", "African American", "Asian", "Other")
data$Ethnicity <- factor(data$Ethnicity, levels = 0:3, labels = ethnicity_labels)

education_labels <- c("None", "High School", "Bachelors's", "Higher")
data$EducationLevel <- factor(data$EducationLevel, levels = 0:3, labels = education_labels)

data$Diagnosis <- as.logical(data$Diagnosis)

gender_labels <- c("Male", "Female")
data$Gender <- factor(data$Gender, levels = 0:1, labels = gender_labels)
```

*Figure 2.1.2.4 R Snippet of replacing numerical representation to actual representation*

Then, the logical value are converted from numeric value to actual TRUE or FALSE value with the snippet below.

```
data$PetAllergy <- as.logical(data$PetAllergy)
data$HistoryOfAllergies <- as.logical(data$HistoryOfAllergies)
data$Eczema <- as.numeric(as.character(data$Eczema))
data$Eczema <- as.logical(data$Eczema)
data$HayFever <- as.numeric(as.character(data$HayFever))
data$HayFever <- as.logical(data$HayFever)
data$GastroesophagealReflux <- as.numeric(as.character(data$GastroesophagealReflux))
data$GastroesophagealReflux <- as.logical(data$GastroesophagealReflux)
data$FamilyHistoryAsthma <- as.numeric(as.character(data$FamilyHistoryAsthma))
data$FamilyHistoryAsthma <- as.logical(data$FamilyHistoryAsthma)
data$Wheezing <- as.numeric(as.character(data$Wheezing))
data$Wheezing <- as.logical(data$Wheezing)
data$ShortnessOfBreath <- as.numeric(as.character(data$ShortnessOfBreath))
data$ShortnessOfBreath <- as.logical(data$ShortnessOfBreath)
data$ChestTightness <- as.numeric(as.character(data$ChestTightness))
data$ChestTightness <- as.logical(data$ChestTightness)
data$Coughing <- as.numeric(as.character(data$Coughing))
data$Coughing <- as.logical(data$Coughing)
data$NighttimeSymptoms <- as.numeric(as.character(data$NighttimeSymptoms))
data$NighttimeSymptoms <- as.logical(data$NighttimeSymptoms)
data$ExerciseInduced <- as.numeric(as.character(data$ExerciseInduced))
data$ExerciseInduced <- as.logical(data$ExerciseInduced)
```

*Figure 2.1.2.5 Conversion of Numeric Value to Logical Value*

Lastly, the values are checked to ensure they are within range as per the requirements from the previous section.

```
119 data <- data[data$Age >= 5 & data$Age <= 80, ]
120 data <- data[data$BMI >= 15 & data$BMI <= 40, ]
121 data <- data[data$PhysicalActivity >= 0 & data$PhysicalActivity <= 10, ]
122 data <- data[data$DietQuality >= 0 & data$DietQuality <= 10, ]
123 data <- data[data$SleepQuality >= 4 & data$SleepQuality <= 10, ]
124 data <- data[data$PollutionExposure >= 0 & data$PollutionExposure <= 10, ]
125 data <- data[data$PollenExposure >= 0 & data$PollenExposure <= 10, ]
126 data <- data[data$DustExposure >= 0 & data$DustExposure <= 10, ]
127
128 data <- data[data$LungFunctionFEV1 >= 1.0 & data$LungFunctionFEV1 <= 4.0, ]
129 data <- data[data$LungFunctionFVC >= 1.5 & data$LungFunctionFVC <= 6.0, ]
130
```

*Figure 2.1.2.6 Data Value Range Check*

After all these operations, the data that is left has been relatively cleaned. For the rows, that has some of its values went out of range, the entire row has been turnt to N.A as row 23 shown in Figure 2.1.2.7.

| | Age | Gender | Ethnicity | EducationLevel | BMI | Smoking | PhysicalActivity | DietQuality | Sleep |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 40 | Male | Caucasian | Bachelors's | 35.35373 | FALSE | 1.8305243 | 3.57016811 | |
| 2 | 19 | Female | Other | Higher | 26.81479 | FALSE | 9.0958218 | 2.35122471 | |
| 3 | 45 | Female | Caucasian | None | 18.83605 | FALSE | 4.4299395 | 0.43822037 | |
| 4 | 43 | Male | African American | High School | 20.54201 | FALSE | 6.0443706 | 0.03382623 | |
| 5 | 41 | Female | Caucasian | High School | 26.04526 | FALSE | 7.1543293 | 9.18062468 | |
| 6 | 30 | Male | Caucasian | Bachelors's | 16.95834 | FALSE | 3.3325443 | 4.74306642 | |
| 7 | 67 | Female | Caucasian | High School | 22.05611 | TRUE | 9.7590325 | 9.85616074 | |
| 8 | 70 | Male | Caucasian | High School | 35.18934 | FALSE | 1.2185108 | 5.04516038 | |
| 9 | 12 | Male | Caucasian | High School | 21.87412 | FALSE | 7.3729134 | 5.04516038 | |
| 10 | 31 | Male | Caucasian | Bachelors's | 31.82101 | TRUE | 8.5168352 | 3.53232765 | |
| 11 | 27 | Male | Caucasian | High School | 29.33705 | FALSE | 3.7769829 | 0.50220328 | |
| 12 | 61 | Female | African American | High School | 26.97392 | FALSE | 8.8702171 | 0.03874284 | |
| 13 | 74 | Female | Caucasian | Bachelors's | 22.82528 | FALSE | 7.1828668 | 4.89244346 | |
| 14 | 49 | Male | Caucasian | Bachelors's | 21.74180 | FALSE | 3.3549881 | 5.98760637 | |
| 15 | 77 | Male | Caucasian | None | 16.63423 | FALSE | 3.4622694 | 0.33234156 | |
| 16 | 13 | Female | Asian | None | 31.93345 | FALSE | 5.1118800 | 5.99269218 | |
| 17 | 16 | Male | Caucasian | High School | 28.00000 | FALSE | 5.7522036 | 6.73246237 | |
| 18 | 42 | Female | Caucasian | High School | 15.18003 | FALSE | 1.2881383 | 6.51434637 | |
| 19 | 73 | Male | African American | High School | 35.55208 | FALSE | 7.1237304 | 9.17509969 | |
| 20 | 77 | Male | Caucasian | High School | 36.09002 | FALSE | 4.4011743 | 5.97207956 | |
| 21 | 78 | Female | Caucasian | High School | 32.20171 | FALSE | 4.8368255 | 6.16333387 | |
| 22 | 43 | Female | Caucasian | Bachelors's | 29.05961 | FALSE | 3.0198535 | 6.11963734 | |
| 23 | NA | NA | NA | NA | NA | NA | NA | NA | |

*Figure 2.1.2.7 Post Cleaning*

## 2.1.3 Missing Value Handling

Rows with N.A values even after section 1 and section 2's preparation are deleted to ensure data integrity using the following snippet.

```
132  # clean out rows with undefined values
133  cleaned_row_data <- na.omit(data)
```

*Figure 2.1.3.1 Omitting NA Values after cleaning*

After that, only 4135 rows are left for analysis process.

## 2.2 Data Analysis

### 2.2.1 Objective 1: To compare the frequency of wheezing episodes reported by younger patients (ages 5-18) versus older patients (ages 19-8) - JONATHAN NG'UA MBAI(TP075128)

Wheezing is a common respiratory indication related with conditions such as asthma and hypersensitivities. Understanding the predominance of wheezing in distinctive age bunches can offer assistance healthcare suppliers tailor mediations and medicines. This investigation points to compare the recurrence of wheezing scenes between more youthful patients (ages 5-18) and more seasoned patients (ages 19-80) utilizing a dataset containing different health-related factors.

#### *2.2.1.1 Analysis: What is the overall prevalence of wheezing in the dataset?*

To begin any analysis, we must load the required libraries. For data processing and data visualization, 'ggplot2' is utilized together with the 'dplyr' package. **The library(dplyr)** loads the 'dplyr' package which is used in filtering and summarizing the data. On the other hand the **library(ggplot2) and (plotly)** is used in making data and interactive visualizations for the dataset.

```
library(dplyr)
library(ggplot2)
library(plotly)
library(ggpubr)
library(crayon)
library(rgl)
library(tidyr)
```

We load the dataset that is to be analysed. **The 'read.csv'** reads and opens the CSV file and loads it into the 'New_clean' data frame. The column names are in the first row, as indicated by the header = TRUE argument.

```
New_clean <- read.csv("C:/Users/Jonathan/Downloads/cleaned_dataset6.csv", header = TRUE)
New_clean
```

We check the data structure so as to understand its contents and ensure that the data types are appropriate for analysis. **The 'str'** function displays the data frame's structure, including column names and types. This helps to ensure that the Age and Wheezing columns are correctly formatted (numeric and logical, respectively).

```
# Checking the data structure
str(New_clean)
```

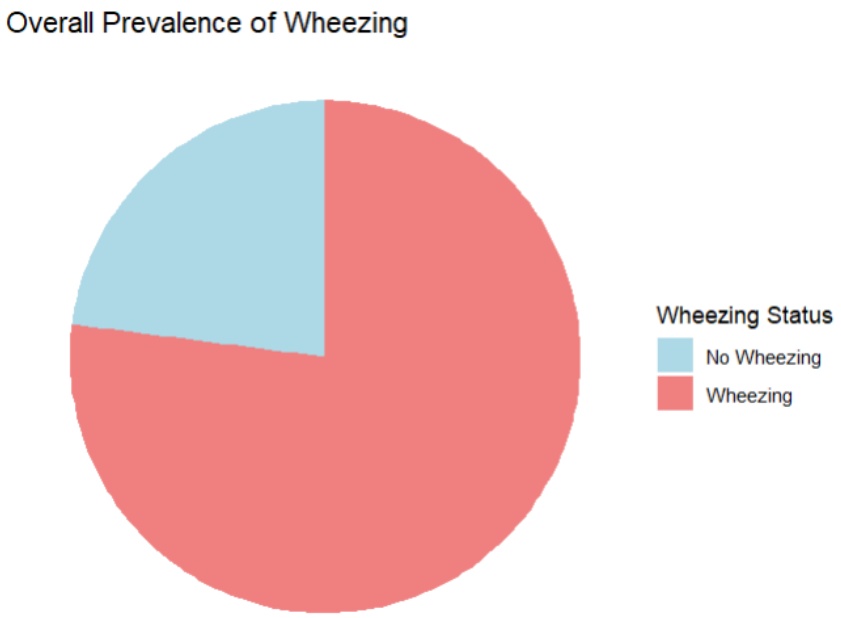
We then calculate the percentage wheezing frequency of the overall population. **'mean(New_clean$Wheezing)'**:This function calculates the mean (average) of the Wheezing

variable from the ‘New_clean’ dataset. The wheezing is coded as 0 (no wheezing) and 1 (wheezing).

```
# Calculate overall wheezing prevalence
overall_wheezing_pct <- mean(New_clean$Wheezing) * 100

cat(sprintf("Overall prevalence of wheezing: %.2f%%\n", overall_wheezing_pct))
```

```
> overall_wheezing_pct <- mean(New_clean$Wheezing) * 100
>
> cat(sprintf("Overall prevalence of wheezing: %.2f%%\n", overall_wheezing_pct))
Overall prevalence of wheezing: 77.05%
```

**Results**: The overall prevalence of wheezing in the dataset is 77.05%.




The above calculation and visualization provides a baseline understanding of wheezing frequency across all patients, which can be compared to the specific age groups later in the analysis.

**Filtering Data**

We then filter the data to get the required findings for the analysis. As per our objective we filter the data according to the specified age groups (younger patients (ages 5-18) and older patients (ages 19-80)). This differentiation allows us to assess and compare wheezing frequencies within each age group, which is the major goal of the study.

Filtering the data ensures that the calculations are performed on the appropriate subsets of patients, preventing any skewing of the results due to the inclusion of irrelevant age groups.

```
# Filter younger and older patients
younger_patients <- New_clean %>% filter(Age >= 5 & Age <= 18)
older_patients <- New_clean %>% filter(Age >= 19 & Age <= 80)
```

The **'filter'** function from the 'dplyr' package creates two new data frames: younger_patients and older_patients. The **'%>%'** operator allows for chaining commands, making the code more readable.

*2.2.1.2 Analysis: What is the distribution of wheezing frequencies within each age group?*

After filtering the data we now calculate the frequency of wheezing between the two age groups and get their results.

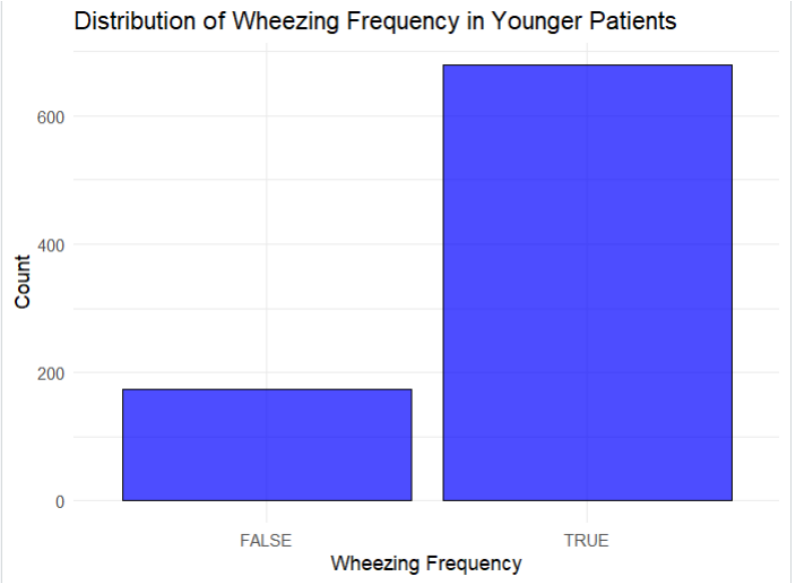
```
# Calculate wheezing frequency
younger_wheezing_pct <- mean(younger_patients$Wheezing) * 100
older_wheezing_pct <- mean(older_patients$Wheezing) * 100

# Print results
cat(sprintf("Wheezing frequency in younger patients: %.2f%%\n", younger_wheezing_pct))
cat(sprintf("Wheezing frequency in older patients: %.2f%%\n", older_wheezing_pct))
```

**The 'mean'** function calculates the average of the Wheezing column, which is treated as a binary variable (TRUE/FALSE). Multiplying the average by 100 returns a percentage.
**The 'cat'** outputs the results to the console. The **'sprintf'** method prepares the output using two decimal places.

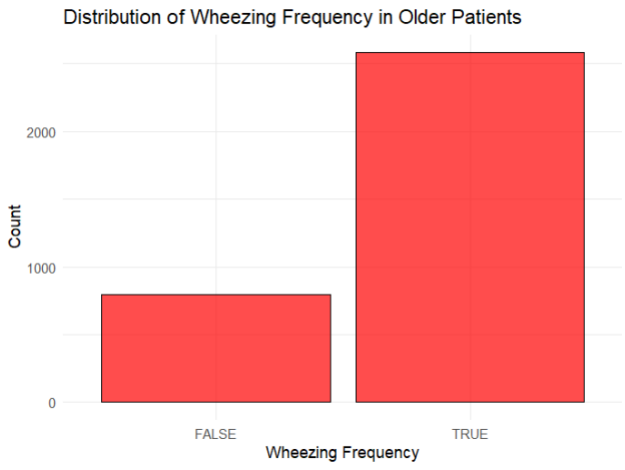
```
> cat(sprintf("Wheezing frequency in younger patients: %.2f%%\n", younger_wheezing_pct))
Wheezing frequency in younger patients: 79.69%
> cat(sprintf("Wheezing frequency in older patients: %.2f%%\n", older_wheezing_pct))
Wheezing frequency in older patients: 76.38%
```

Wheezing frequency in younger patients (ages 5-18): 79.69%
Wheezing frequency in older patients (ages 19-80): 76.38%

```r
# Create bar plot for younger patients
ggplot(younger_patients, aes(x = Wheezing)) +
  geom_bar(fill = "blue", color = "black", alpha = 0.7) +
  labs(title = "Distribution of Wheezing Frequency in Younger Patients",
       x = "Wheezing Frequency",
       y = "Count") +
  theme_minimal()
```




```r
# Create bar plot for older patients
ggplot(older_patients, aes(x = Wheezing)) +
  geom_bar(fill = "red", color = "black", alpha = 0.7) +
  labs(title = "Distribution of Wheezing Frequency in Older Patients",
       x = "Wheezing Frequency",
       y = "Count") +
  theme_minimal()
```




The visualizations provide insights into the respiratory health of the two age groups. As from the above the False value in older patients is higher than the True value in younger patients indicating and confirming that the percentage frequency of wheezing is higher in younger patients than younger patients.

The findings might lead to more study into possible risk factors for wheeze, such as environmental exposures, lifestyle variables, or comorbidities that vary by age group.

*2.2.1.3 Analysis: Are there any environmental or lifestyle factors (e.g., smoking, pollution exposure) associated with higher wheezing frequencies in either age group?*

Regression analysis is a useful approach for comparing wheezing rates across age groups. It enables the identification of predictors, measurement of impact sizes, correction for confounding factors, investigation of interactions, prediction of future outcomes, and, finally, the provision of evidence-based recommendations for healthcare choices and policies.

```
# Perform regression analysis
younger_model <- glm(Wheezing ~ Smoking + PollutionExposure, data = younger_patients, family = "binomial")
older_model <- glm(Wheezing ~ Smoking + PollutionExposure, data = older_patients, family = "binomial")

summary(younger_model)
summary(older_model)
```

```
> summary(younger_model)

Call:
glm(formula = Wheezing ~ Smoking + PollutionExposure, family = "binomial",
    data = younger_patients)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        1.46839    0.19801   7.416 1.21e-13 ***
SmokingTRUE        0.02989    0.19536   0.153    0.878
PollutionExposure -0.02067    0.03210  -0.644    0.520
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 859.85  on 851  degrees of freedom
Residual deviance: 859.40  on 849  degrees of freedom
AIC: 865.4

Number of Fisher Scoring iterations: 4
```

```
> summary(older_model)

Call:
glm(formula = Wheezing ~ Smoking + PollutionExposure, family = "binomial",
    data = older_patients)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       1.170576   0.091507  12.792   <2e-16 ***
SmokingTRUE      -0.249291   0.108199  -2.304   0.0212 *
PollutionExposure 0.008624   0.016116   0.535   0.5926
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3698.6  on 3382  degrees of freedom
Residual deviance: 3693.1  on 3380  degrees of freedom
AIC: 3699.1

Number of Fisher Scoring iterations: 4
```

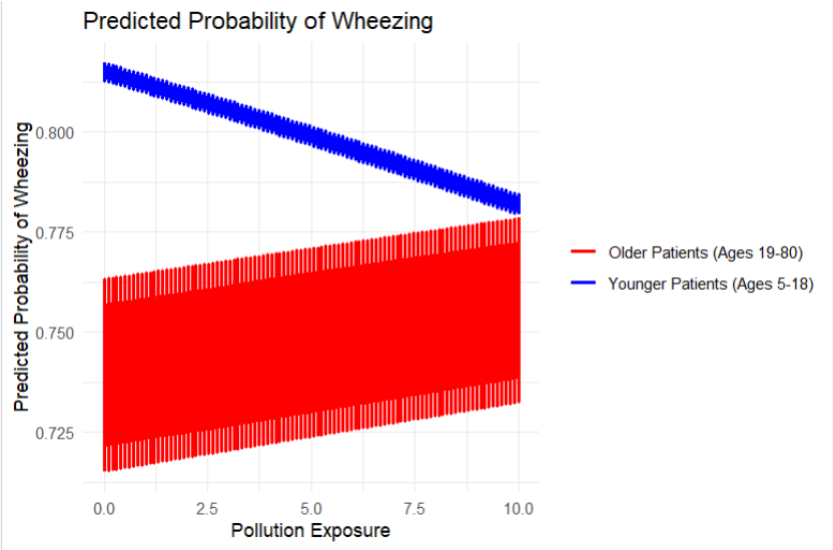
**Younger Patients Model:**

1. **Intercept:** The intercept (1.46839) reflects the log-odds of wheeze when both smoking and pollution exposure are zero. This implies that, even in the absence of these conditions, younger patients have a relatively high baseline risk of wheezing.
2. **Smoking:** The coefficient for smoking (0.02989) is positive but not statistically significant (p-value = 0.878). This indicates that while smoking may slightly increase the odds of wheezing in younger patients, the effect is not strong enough to be considered significant.
3. **Pollution Exposure**: The coefficient for pollution exposure (-0.02067) is negative but not statistically significant (p = 0.520). This implies that pollution does not have a major effect on the incidence of wheezing in children.

**Older Patients Model:**

1. **Intercept:** The intercept (1.170576) reflects the log-odds of wheeze when both smoking and pollution exposure are zero. This shows that, even without these characteristics, older people have a somewhat significant baseline risk of wheezing.
2. **Smoking:** The coefficient for smoking (-0.249291) is negative and statistically significant (p-value = 0.0212). This indicates that smoking significantly decreases the odds of wheezing in older patients. This counterintuitive result may be due to the complex interplay of factors affecting respiratory health in older individuals.
3. **Pollution Exposure:** The coefficient for pollution exposure (0.008624) is positive but not statistically significant (p-value = 0.5926). This suggests that pollution exposure does not have a significant impact on the likelihood of wheezing in older patients.

```r
# Plot for younger patients
ggplot(prediction_data, aes(x = PollutionExposure)) +
  geom_line(aes(y = younger_prob, color = "Younger Patients (Ages 5-18)"), size = 1) +
  geom_line(aes(y = older_prob, color = "Older Patients (Ages 19-80)"), size = 1) +
  labs(title = "Predicted Probability of Wheezing",
       x = "Pollution Exposure",
       y = "Predicted Probability of Wheezing") +
scale_color_manual(values = c("Younger Patients (Ages 5-18)" = "blue", "Older Patients (Ages 19-80)" = "red")
theme_minimal() +
theme(legend.title = element_blank())
```




From the above regression analysis and its visualization we can conclude that:

- Older patients are generally at a higher risk of wheezing compared to younger patients, particularly in relation to pollution exposure.
- Smoking appears to be a significant predictor of wheezing in both age groups, indicating the need for targeted smoking cessation programs.
- Pollution exposure is a critical factor affecting wheezing frequency, highlighting the importance of addressing environmental health issues to improve respiratory outcomes.

These findings can help healthcare practitioners and governments create tailored treatments to minimize wheezing and enhance respiratory health, particularly in susceptible groups like older persons and smokers.

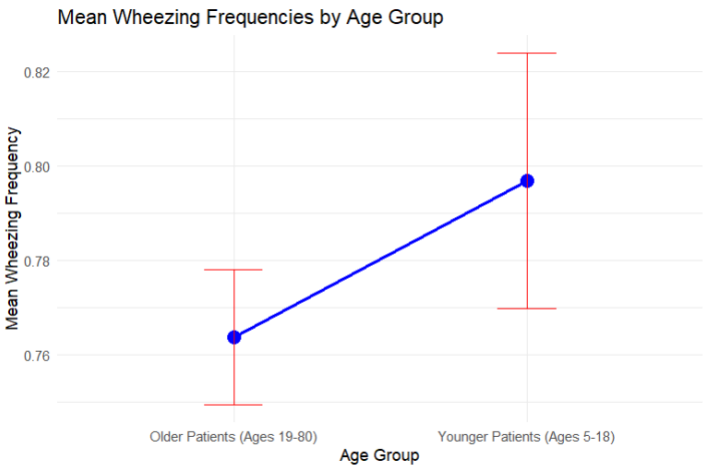
*2.2.1.4 Analysis: Is there a statistically significant difference in wheezing frequency between younger and older patients?*

```
t.test(younger_patients$Wheezing, older_patients$Wheezing)
```

```
> t.test(younger_patients$Wheezing, older_patients$Wheezing)

        Welch Two Sample t-test

data:  younger_patients$Wheezing and older_patients$Wheezing
t = 2.1231, df = 1368.3, p-value = 0.03393
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.002518217 0.063740306
sample estimates:
mean of x mean of y
0.7969484 0.7638191
```

The mean wheezing frequency is higher in younger patients (0.7969) compared to older patients (0.7638). This finding may indicate that younger individuals in the study are more likely to experience wheezing episodes than older individuals.
The 95% confidence interval (0.0025 to 0.0637) suggests that the true difference in means is likely positive, reinforcing the conclusion that younger patients experience more wheezing episodes.

```
# Create the line graph with confidence intervals
ggplot(mean_data, aes(x = AgeGroup, y = MeanWheezing)) +
geom_line(aes(group = 1), color = "blue", size = 1) +  # Line connecting the means
geom_point(size = 4, color = "blue") +  # Points for the means
geom_errorbar(aes(ymin = CI_Lower, ymax = CI_Upper), width = 0.2, color = "red") +
labs(title = "Mean Wheezing Frequencies by Age Group",
     x = "Age Group",
     y = "Mean Wheezing Frequency") +
theme_minimal()
```

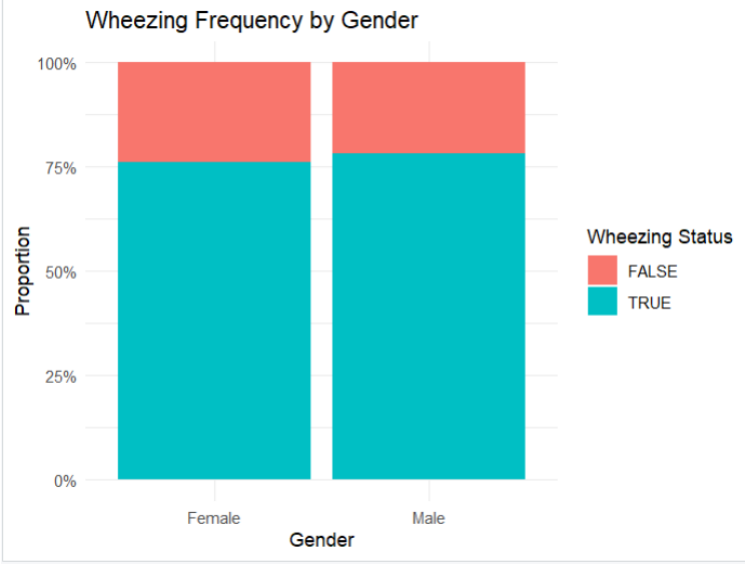


The t-test findings show a statistically significant difference in wheeze frequency between younger and older individuals. Specifically, younger patients are more likely to experience wheezing than older individuals. This data may suggest that age-related variables contribute to a higher prevalence of wheezing in younger people, necessitating more research into potential underlying causes and specific healthcare interventions for younger patients.

### *2.2.1.5 Analysis: How do other demographic factors (e.g., gender, ethnicity, BMI) correlate with wheezing frequency in both age groups?*

Between gender and wheezing we have used a bar plot to represent the correlation. The bar plot shows the proportion of patients who reported wheezing (True/False) for each gender. Each bar depicts the distribution of wheezing status among male and female groups.

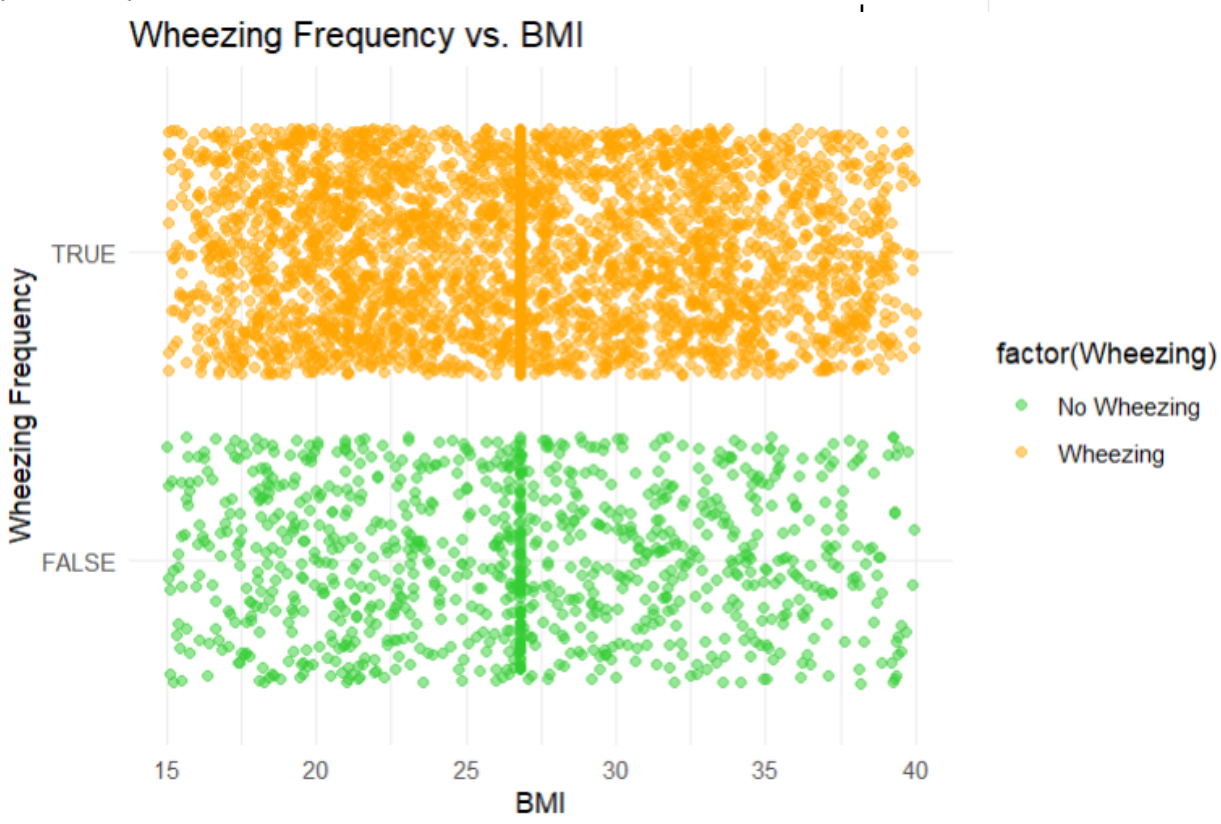
```
# Create a bar plot for wheezing frequency by gender
gender_plot <- ggplot(New_clean, aes(x = Gender, fill = factor(Wheezing))) +
geom_bar(position = "fill") +
labs(title = "Wheezing Frequency by Gender",
      x = "Gender",
      y = "Proportion",
      fill = "Wheezing Status") +
scale_y_continuous(labels = scales::percent) +
theme_minimal()

print(gender_plot)
```




From the above analysis it shows that men record  slightly higher wheezing frequences than women indicating that men are likely to be at higher risks for respiratory issues. Professionals in healthcare can opt for gender-specific therapies or education initiatives to address the increased frequency of wheezing in men.

The scatter plot depicts the association between BMI and wheeze frequency, with each point representing an individual patient and coloured according to their wheezing status (Yes/No).

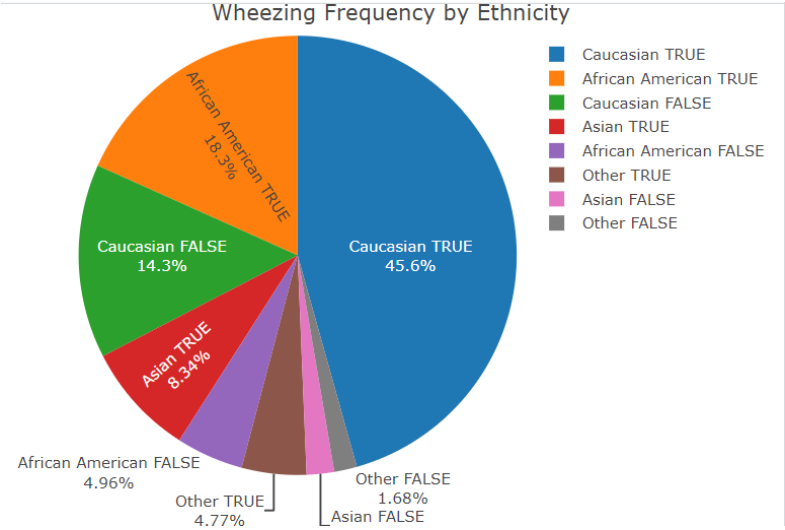
```
# Create a scatter plot for wheezing frequency vs. BMI
bmi_plot <- ggplot(New_clean, aes(x = BMI, y = Wheezing, color = factor(Wheezing))) +
geom_jitter(alpha = 0.5) +
labs(title = "Wheezing Frequency vs. BMI",
     x = "BMI",
     y = "Wheezing Frequency") +
scale_color_manual(values = c("limegreen", "orange"), labels = c("No Wheezing", "Wheezing")) +
theme_minimal()

print(bmi_plot)
```




From the above analysis its concluded that wheezing frequencies varies across the various different BMI levels. This shows us that BMI levels are not necessarily in relation to wheezing frequencies and can not be necessarily used to analyze the various causes of wheezing frequencies.

Finally for the correlation between the wheezing frequencies and the Ethnicity, a 3D pie chart has been used. The pie chart displays slices representing the proportion of patients who report wheezing (True) versus those who do not (False) for each ethnic group.

```r
# Calculate the counts of wheezing and non-wheezing by ethnicity
wheezing_by_ethnicity <- New_clean %>%
  group_by(Ethnicity, Wheezing) %>%
  summarise(Count = n(), .groups = 'drop')


# Create a 3D pie chart for wheezing by ethnicity
plot_ly(wheezing_by_ethnicity, labels = ~paste(Ethnicity, Wheezing), values = ~Count, type = 'pie',
        textinfo = 'label+percent', insidetextorientation = 'radial') %>%
layout(title = 'Wheezing Frequency by Ethnicity',
        showlegend = TRUE)
```




From the above analysis we get to see that a high number of Caucasian people report high wheezing frequency compared to other ethnicities. This can call for health campaigns and thorough health interventions in the region to address the specified issue.

The visualizations together emphasize the significance of demographic variables in understanding wheeze prevalence. By evaluating gender, ethnicity, and BMI, we may identify at-risk groups and personalize healthcare interventions appropriately. The findings highlight the need of focused public health policies that address the particular problems that different demographic groups confront, with the ultimate goal of improving respiratory health outcomes for the entire community. More study may be needed to investigate the root reasons of the reported differences, such as environmental exposures, lifestyle variables, and access to healthcare services.

### *2.2.1.6 Conclusion of Objective 1:*

In conclusion, the analysis clearly shows that younger patients (ages 5-18) had more wheezing frequencies than older patients (years 19-80). This large variation emphasizes the necessity of addressing respiratory health concerns in younger people and advises that more study into the underlying causes of wheezing in various age groups is needed. The findings from this investigation can help healthcare practitioners create focused therapies and improve overall respiratory health outcomes.

Moreover, older patients are also at risk of wheezing caused by different factors for example matters related to environmental pollution. This calls for a thorough health intervention for the older generation and also calls for environmental education in relation to health.

#### *2.2.1.7 Additional Features:*

1. **'Plotly'** is a R utility for creating interactive web-based visualizations. It is especially excellent for constructing sophisticated plots like 3D charts, pie charts, and scatter plots that the viewer can readily control (for example, zooming and hovering for details). It works nicely with R and offers a great degree of customisation for visualizations.

2. **'ggpubr'** is a R package that provides simple methods for producing publication-ready charts using the **'ggplot2'** library. It streamlines the process of adding statistical annotations, integrating numerous graphs, and modifying themes. It is very effective for improving the visual attractiveness of plots and preparing them for scholarly publishing.

3. **'Filter'** is a function in the '**dplyr'** package that allows you to subset rows in a data frame depending on certain criteria. It is widely used to extract specific subsets of data, for as choosing patients between a specified age range or with specific traits.

4. **'Cat'** is a basic R function for concatenating and printing things. It is frequently used to display messages or results in the console without the extra formatting that comes with the 'print()' method. It's handy for producing personalized output messages.

5. **'Sprintf'** is a R function that formats strings using provided format codes. It is frequently used to generate structured output, such as presenting figures with a certain number of decimal places.

6. **'Geom_bar'** is a function in the 'ggplot2' package that generates bar charts. It may be used to show the number of occurrences of a category variable. It may also be configured to display proportions or stacked bars based on other grouping parameters.

7. The ggplot2 package includes the function **'theme_minimal()'**, which adds a minimalistic theme to a plot. This theme lowers visual clutter by eliminating backdrop grids and extraneous features, making the data stand out more clearly.

8. **'glm'** stands for **Generalized Linear Model**. It is a R function for fitting generalized linear models that may handle a variety of response variables (for example, binary and count). It is often used in logistic regression when the response variable is binary, such as wheezing(True/ False).

9. **‘t.test’** is a R function that uses a t-test to compare the means of two groups. It is used to see if there is a statistically significant difference between the two groups' means, which is important for comparing wheezing frequencies in younger and older individuals.

10. **‘geom_jitter’** adds jittering to scatter plot points. This helps to reduce overplotting by introducing random noise to the location of points, making it easier to discern the distribution of data points, especially when there are numerous overlapping points.

11. **‘plot_ly’** is a function in the **‘plotly’** package that generates interactive plots. It enables users to create a variety of visualizations, such as scatter plots, line charts, and pie charts, with interactive features like tooltips and zooming.

12. The **‘geom_errorbar’** function in the **‘ggplot2’** package is used to add error bars to a plot. The error bars can be used to represent confidence intervals, standard errors, or standard deviations.

13. **‘ggpubr':** This library provides easy-to-use functions for creating publication-ready plots based on the ‘ggplot2’ package. It simplifies the process of adding statistical annotations, combining multiple plots, and customizing themes, making it particularly useful for enhancing the visual appeal of plots in academic publications.

### 2.2.2 Objective 2: To analyse correlation between asthma diagnosis and the frequency of coughing episodes reported by younger patients (ages 5-18) compared to older patients (ages 19-80)

To effectively analyse the correlation between coughing in younger patients and older patients, and asthma diagnosis, a 4-step analysis process is employed:

1. Descriptive Analysis on Coughing Frequency, Age Group and Asthma Diagnosis

2. Diagnostic Analysis on Coughing Frequency, Age Group and Asthma Diagnosis

3. Predictive Analysis on Coughing Frequency, Age Group and Asthma Diagnosis

4. Prescriptive Analysis on Coughing Frequency, Age Group and Asthma Diagnosis

#### 2.2.2.1 Descriptive Analysis on Coughing Frequency, Age Group and Asthma Diagnosis

To perform descriptive analysis, un-necessary columns are dropped and only the values in question: Age, Coughing and Diagnosis, are left behind. A new column called “Age Group” is also added to perform additional visualization information.

```
df <- data[, c("Age", "Coughing", "Diagnosis")]
View(df)

df$"Age Group" <- ifelse(df$Age < 18, "Child", "Adult")
View(df)
```

*Figure 2.2.2.1*

Using the dataset generated by the code snippet in Figure 2.2.2.1, some summarizations are done.

##### *2.2.2.1.1 Summarization of Age Groups*

Code:

```
ggplot(df, aes(x = `Age Group`)) +
  geom_bar(fill = "blue", color = "black") +
  labs(title = "Age Group Distribution", x = "Age Group", y = "Count")
```

*Figure 2.2.2.2*

Observation:

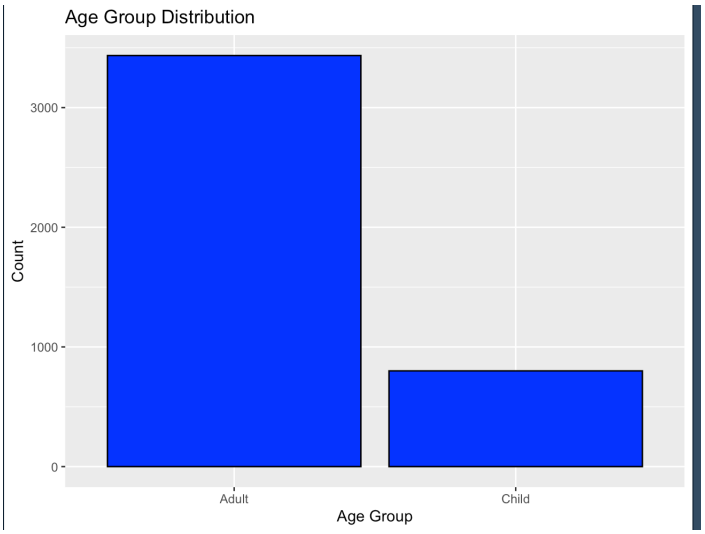



*Figure 2.2.2.3*

Upon summarization, it is found that there are over **3435** adults, and **800** children within the cleaned dataset.

### *2.2.2.1.2 Summarization of Coughing Distribution*

Code:

```
ggplot(df, aes(x = `Coughing`)) +
  geom_bar(fill = "lightgreen", color = "black") +
  labs(title = "Coughing Distribution", x = "Coughing (Yes, No)", y = "Count")
```

*Figure 2.2.2.4*

Observation:

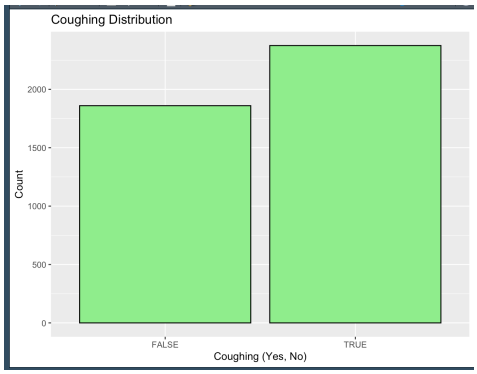



*Figure 2.2.2.5*

Upon summarization, it is found that there are over **2375** patients who display coughing-like symptoms, with **1860** patients showing no symptoms.

### *2.2.2.1.3 Summarization of Diagnosis Distribution*

Code:

```
ggplot(df, aes(x = `Diagnosis`)) +
  geom_bar(fill = "red", color = "black") +
  labs(title = "Coughing Distribution", x = "Diagnosis (Yes, No)", y = "Count")
```

*Figure 2.2.2.6*

Observation:

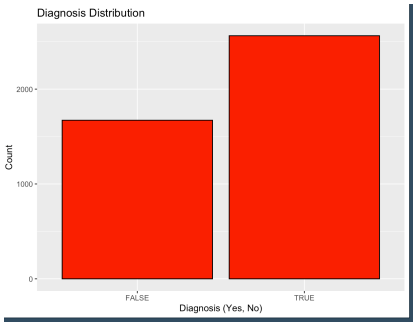



*Figure 2.2.2.7*

Upon summarization, it is found that 1672 patients have not been diagnosed with asthma, while 2563 patients have been.

### *2.2.2.1.4 Summarization of Relationship between Asthma Age Group and Asthma Diagnosis*

Code:

```
ggplot(df, aes(x = `Age Group`, fill = Diagnosis)) +
  geom_bar(position = "fill", color = "black") +
  labs(title = "Asthma Proportion by Age Group", x = "Age Group", y = "Proportion") +
  scale_fill_manual(values = c("lightblue", "orange"))
```

*Figure 2.2.2.8*

Observation:

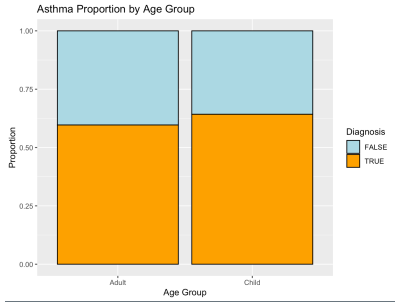



*Figure 2.2.2.9*

With 58% of adults of being contracted and 63% of children being contracted, it can be concluded that both age groups are equally vulnerable to be contracted with Asthma.

### *2.2.2.1.5 Summarization of Relationship between Coughing and Asthma*

Code:

```
ggplot(df, aes(x = `Coughing`, fill = Diagnosis)) +
  geom_bar(position = "fill", color = "black") +
  labs(title = "Asthma Proportion by Coughing", x = "Coughing", y = "Proportion") +
  scale_fill_manual(values = c("lightblue", "orange"))
```

*Figure 2.2.2.10*

Observation:

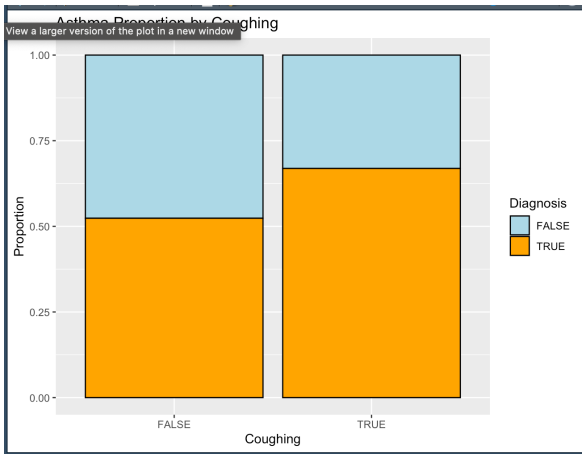



*Figure 2.2.2.11*

From the visualization, it can be concluded that patients with coughing symptoms are much likely to also display symptoms of asthma.

### *2.2.2.1.6 Summarization of Relationship between Age and Asthma, and Age and Coughing with Box plot*

Code:

```
ggplot(df, aes(x = Diagnosis, y = Age, fill = Diagnosis)) +
  geom_boxplot() +
  labs(title = "Age Distribution by Asthma Status", x = "Asthma (Yes/No)", y = "Age") +
  scale_fill_manual(values = c("lightblue", "orange"))

ggplot(df, aes(x = Coughing, y = Age, fill = Coughing)) +
  geom_boxplot() +
  labs(title = "Age Distribution by Coughing Status", x = "Coughing (Yes/No)", y = "Age") +
  scale_fill_manual(values = c("lightgreen", "salmon"))
```

*Figure 2.2.2.12*

Visualization:

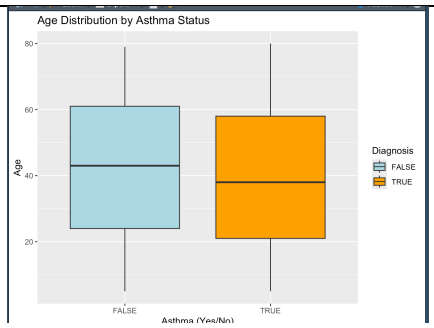



*Figure 2.2.2.13*

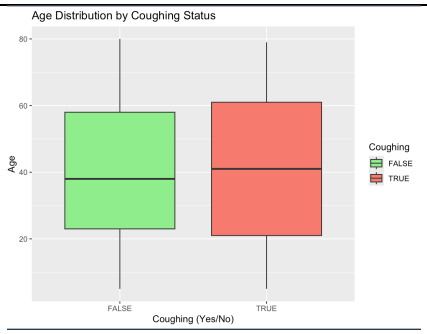



*Figure 2.2.2.14*

Observation:
Figure 2.2.2.13 and Figure 2.2.2.14 clearly shows that although there are outliers and minor differences, the major overlaps in age distribution suggests that age alone is not a definitive factor in asthma diagnosis

## 2.2.2.2 Diagnostic Analysis on Coughing Frequency, Age Group and Asthma Diagnosis

To further understand why the correlation between different variables exist, diagnostic analysis is done.

### *2.2.2.2.1 Chi Square Tests between Age Group and Coughing*

Null Hypothesis: There is no relationship between age group and coughing
Alternative Hypothesis: There is a relationship between age group and coughing

A contingency table is generated with the following results:

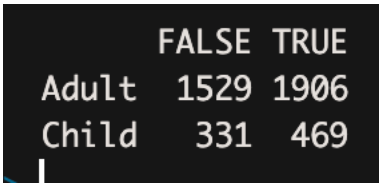



*Figure 2.2.2.15*

| Age Group | Coughing (Yes) | Coughing (No) | Total |
|---|---|---|---|
| Adult | 1529 | 1906 | 3435 |
| Child | 331 | 469 | 800 |
| Total | 1860 | 2375 | 4235 |

*Figure 2.2.2.16 Contingency Table*

| | |
|---|---|
| Expected Frequency of Coughing in Children | 351.35 |

| | |
|---|---|
| Expected Frequency of Coughing in Adult | 1508.64 |
| Expected Frequency of not Coughing in Children | 448.64 |
| Expected Frequency of not Coughing in Adults | 1926.35 |

*Figure 2.2.2.17 Expected Frequency Table*

The formula for Chi-Squared Statistic is:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

*Figure 2.2.2.18 Chi-Squared Formula*

Chi-Squared is calculated with relative ease with the use of R:

```
age_group_coughing_contingency_table <- table(df$"Age Group", df$Coughing)
age_group_chi_sq_test <- chisq.test(age_group_coughing_contingency_table)
print(age_group_chi_sq_test)
```

*Figure 2.2.2.19*

Results:

```
data:  age_group_coughing_contingency_table
X-squared = 2.4673, df = 1, p-value = 0.1162
```

*Figure 2.2.2.20*

Observation:

Since p-value is > 0.05 critical value, null hypothesis cannot be rejected. Hence, from this analysis, it is concluded that there is no relationship between age-group and coughing.

#### *2.2.2.2.2 Chi-Square Tests between Age Group and Asthma Diagnosis*

Null Hypothesis: There is no relationship between Age Group and Asthma Diagnosis
Alternative Hypothesis: There is relationship between Age Group and Asthma Diagnosis

A contingency table is generated with the following results:

```
       FALSE TRUE
Adult   1386 2049
Child    286  514
```

*Figure 2.2.2.21*

Code:

```
age_group_diagnosis_contingency_table <- table(df$"Age Group", df$Diagnosis)
print(age_group_diagnosis_contingency_table)
age_group_diagnosis_chi_sq_test <- chisq.test(age_group_diagnosis_contingency_table)
print(age_group_diagnosis_chi_sq_test)
```

*Figure 2.2.2.22*

Results:

```
        Pearson's Chi-squared test with Yates' continuity correction

data:  age_group_diagnosis_contingency_table
X-squared = 5.5539, df = 1, p-value = 0.01844
```

*Figure 2.2.2.23*

Observation:

Since p-value of 0.01844 is much lesser than the critical value of 0.05, it can be concluded that there is a statistically large correlation between age group and diagnosis. Hence, null hypothesis has been rejected, and alternative hypothesis has been accepted.

#### *2.2.2.2.3 Chi-Square Test between Coughing and Asthma Diagnosis*

Null Hypothesis: There is no relationship between coughing and asthma diagnosis
Alternative Hypothesis: There is relationship between coughing and asthma diagnosis

A contingency table is generated with the following results:

```
       FALSE TRUE
FALSE    886  974
TRUE     786 1589
```

*Figure 2.2.2.24*

Code:

```
coughing_diagnosis_contingency_table <- table(df$Coughing, df$Diagnosis)
print(coughing_diagnosis_contingency_table)
coughing_diagnosis_chi_sq_test <- chisq.test(coughing_diagnosis_contingency_table)
print(coughing_diagnosis_chi_sq_test)
```

*Figure 2.2.2.25*

Results:

```
        Pearson's Chi-squared test with Yates' continuity correction

data:  coughing_diagnosis_contingency_table
X-squared = 91.682, df = 1, p-value < 2.2e-16
```

*Figure 2.2.2.26*

Observation:

Since the p-value of 2.2e-16 is far below the 0.05 critical value, it can be confidently concluded that there is high correlation between coughing symptoms and asthma diagnosis. Hence, the alternative hypothesis is accepted.

### *2.2.2.2.4 Measure of Association between Age Group, Coughing and Diagnosis*

To understand a bit further on the diagnosis results based on the other two values, a measure of association is done in R

Code:

```
assocstats(table(df$`Age Group`, df$Coughing, df$Diagnosis))
```

*Figure 2.2.2.27*

Results:

```
$`:FALSE`
                    X^2 df P(> X^2)
Likelihood Ratio 0.50317  1  0.47811
Pearson          0.50245  1  0.47842

Phi-Coefficient   : 0.017
Contingency Coeff.: 0.017
Cramer's V        : 0.017

$`:TRUE`
                   X^2 df P(> X^2)
Likelihood Ratio 4.7559  1 0.029197
Pearson          4.7003  1 0.030158

Phi-Coefficient   : 0.043
Contingency Coeff.: 0.043
Cramer's V        : 0.043
```

*Figure 2.2.2.28*

Observation:

When asthma diagnosis is ‘FALSE’, there is no significant association between the values, but there is significant association when diagnosis is ‘TRUE’. As with the chi-squared tests earlier, age-group does not seem to highly affect the diagnosis results compared to coughing, and coughing might have been caused by other factors.

### *2.2.2.2.5 Chi-Square Test between Age Group and Coughing for two different subsets based on Diagnosis*

Null Hypothesis 1: Age Group and Coughing have no relationship when Diagnosis is true
Null Hypothesis 2: Age Group and Coughing have no relationship when Diagnosis is false
Alternative Hypothesis 1: Age Group and Coughing have relationship when Diagnosis is true
Alternative Hypothesis 2: Age Group and Coughing have relationship when Diagnosis is false

To perform this operation, the dataset is first splitted into two subsets based on whether if the patient has asthma or not.

```
df_true <- df[df$Diagnosis == TRUE, ]
df_false <- df[df$Diagnosis == FALSE, ]
```

*Figure 2.2.2.29*

After that, Chi-Squared test is performed.

Code:

```
chisq_test_true <- chisq.test(table(df_true$`Age Group`, df_true$Coughing))
print(chisq_test_true)
chisq_test_false <- chisq.test(table(df_false$`Age Group`, df_false$Coughing))
print(chisq_test_false)
```

*Figure 2.2.2.30*

Results:

```
> chisq_test_true <- chisq.test(table(df_true$`Age Group`, df_true$Coughing))
> print(chisq_test_true)

        Pearson's Chi-squared test with Yates' continuity correction

data:  table(df_true$`Age Group`, df_true$Coughing)
X-squared = 4.4825, df = 1, p-value = 0.03424

> chisq_test_false <- chisq.test(table(df_false$`Age Group`, df_false$Coughing))
> print(chisq_test_false)

        Pearson's Chi-squared test with Yates' continuity correction

data:  table(df_false$`Age Group`, df_false$Coughing)
X-squared = 0.41445, df = 1, p-value = 0.5197
```

*Figure 2.2.2.31*

Observation:

When diagnosis is `TRUE`, p-value is less than 0.05, hence alternative hypothesis has been accepted. However, when diagnosis is `FALSE`, p-value is greater than 0.05, hence null hypothesis is favored instead. This result aligns with the observation found in the previous analysis.

*2.2.2.2.6 Visualization of Coughing vs Age Group by Diagnosis*

Code

```
ggplot(df, aes(x = `Age Group`, fill = Coughing)) +
  geom_bar(position = "dodge") +
  facet_wrap(~ Diagnosis) +
  labs(title = "Coughing vs Age Group by Diagnosis")
```

*Figure 2.2.2.32*

Results:

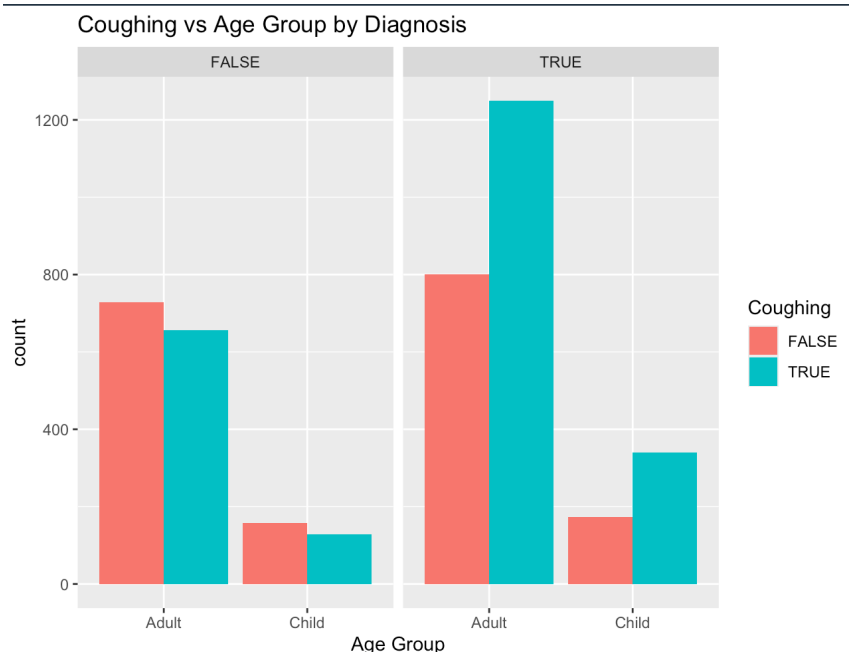



*Figure 2.2.2.33*

Observation:

From the visualization, it can be seen that for both Adult and Children population, coughing is a good indicator of having Asthma.

### 2.2.2.3 Predictive Analysis on Coughing Frequency, Age Group and Asthma Diagnosis

#### *2.2.2.3.1 Logistic Regression on Age Group and Coughing to predict Asthma Diagnosis*

To train the predictive model, the data is first splitted into training set, and testing set using the following snippet

```
set.seed(123)
train_indices <- sample(1:nrow(df), size = 0.7 * nrow(df))
train_data <- df[train_indices, ]
test_data <- df[-train_indices, ]
```

*Figure 2.2.2.33*

Code:

```
age_group_and_coughing_to_diagnosis_model <- glm(Diagnosis ~ `Age Group` +
summary(age_group_and_coughing_to_diagnosis_model)

# Predict on test data
predictions <- predict(age_group_and_coughing_to_diagnosis_model, newdata

# Convert probabilities to binary outcomes (0.5 as threshold)
predicted_classes <- ifelse(predictions > 0.5, TRUE, FALSE)

# Confusion Matrix
table(Predicted = predicted_classes, Actual = test_data$Diagnosis)

# Accuracy
accuracy <- mean(predicted_classes == test_data$Diagnosis)
print(accuracy)
```

*Figure 2.2.2.34*

```
roc_curve <- roc(df$Diagnosis, predictions)
plot(roc_curve)
auc(roc_curve)
```

*Figure 2.2.2.35*

Results:

```
> accuracy <- mean(predicted
> print(accuracy)
[1] 0.5837923
>
```

*Figure 2.2.2.36*

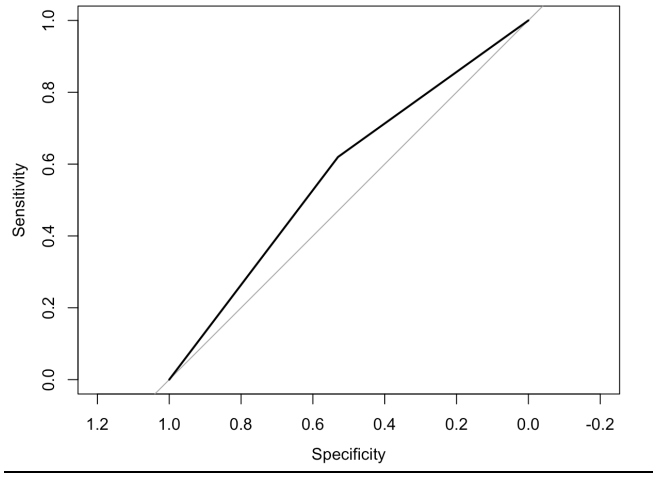



*Figure 2.2.2.37*

Observation:

With an accuracy of 0.583, it can be concluded that logistic regression model with only those two variables is not suitable to predict the results with Coughing and Age Group in an accurate manner. This is also supported by the fact that generated ROC Curve which is not skewed towards the top as much.

## Conclusion of the objective

Albeit there is a high correlation between Age Group and Asthma Diagnosis, and Coughing and Asthma, it can be concluded that those variables alone cannot reliably predict whether if Diagnosis is TRUE or FALSE.

It is essential for doctors to consider other variables and symptoms as those two variables alone cannot be used to reliably diagnosed whether if a patient has asthma or not.

## 2.2.3 Objective 3: To evaluate the correlation between age and severity of asthma symptons, including nightime symptoms and exercise induced-symptoms

### *2.2.3.1 Analysis 1: Logistic regression model to study the correlation between age and nighttime asthma symptoms*

**Analysis Method:** The approach is easier since logistic regression and scatter plots are simple yet useful tools for investigating the link between age and nocturnal symptoms. Logistic regression effectively predicts binary outcomes, such as having or not experiencing nighttime symptoms, but the scatter plot graphically depicts this association, making the results simple to grasp. These approaches are simple, delivering clear insights without adding excessive complication, making them excellent for quickly and visually understanding how age effects the incidence of nighttime symptoms.

**Code:**

```
# Create logistic regression model
model_nighttime <- glm(NighttimeSymptoms ~ Age, family=binomial, data=asthma_data)

# Summarize the logistic regression
summary(model_nighttime)

# Visualize the regression using scatter plot
scatterplot(asthma_data$Age, asthma_data$NighttimeSymptoms,
            xlab="Age", ylab="Nighttime Symptoms",
            main="Age vs Nighttime Symptoms")
```

*Figure 2.2.3.1*

**Vindication**: There is a moderate positive correlation between age and nighttime symptoms, meaning older individuals tend to experience more nighttime symptoms.

**Screenshot of Code Output**:

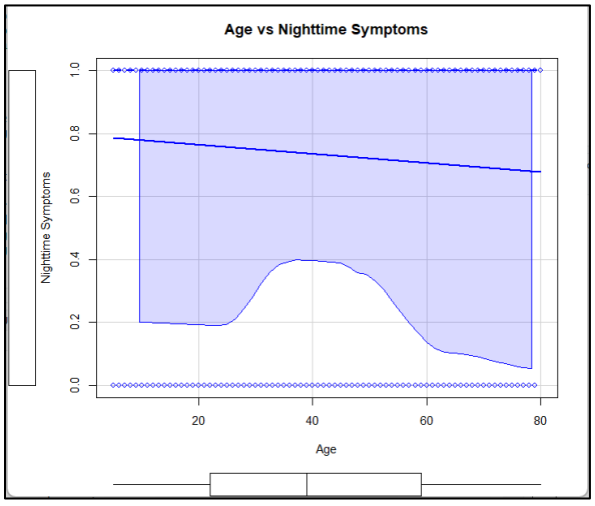


*Figure 2.2.3.2*

**Observations**: Nighttime symptoms are an important marker of asthma severity. Understanding this relationship helps clinicians focus on nighttime symptom control, especially for older patients who may experience more frequent nighttime issues.

### *2.2.3.2 Analysis 2: Logistic regression to examine the relationship between age and exercise-induced asthma symptoms.*

**Analysis Method:** The research used logistic regression to investigate the association between age and the risk of having exercise-induced symptoms, using a binary result (yes or no). This method assists in predicting how age effects the occurrence of symptoms. In addition, a scatter plot with a logistic regression line is utilised to graphically represent this association, demonstrating how the likelihood of symptoms changes with age. These approaches were chosen because they successfully describe binary outcomes and give a clear, understandable data visualisation, making the results accessible and interpretable.

**Code:**

```
# Logistic regression for Exercise-Induced Symptoms
model_exercise <- glm(ExerciseInduced ~ Age, family=binomial, data=asthma_data)

# Summary of the model
summary(model_exercise)

# Visualization using ggplot2
ggplot(asthma_data, aes(x=Age, y=ExerciseInduced)) +
  geom_point() +
  stat_smooth(method="glm", method.args=list(family=binomial),
              se=FALSE, color="blue") +
  labs(title="Age vs Exercise-Induced Symptoms", x="Age",
       y="Exercise-Induced Symptoms")
```

*Figure 2.2.3.3*

**Vindication**: Exercise-induced symptoms can limit activity in asthma patients. Understanding how these symptoms evolve with age helps in developing tailored exercise plans for older individuals.

**Screenshot of Code Output**:

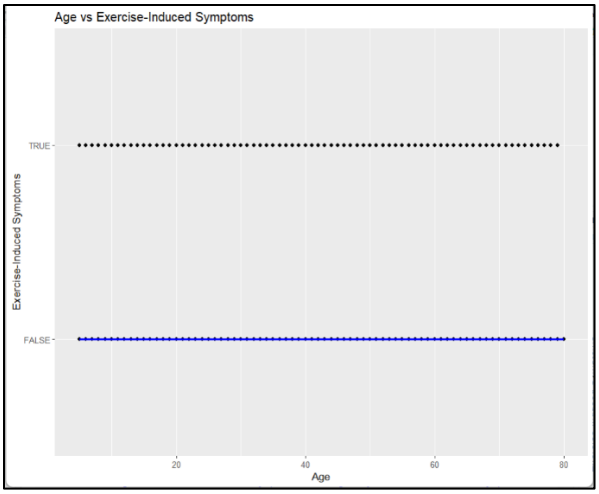



*Figure 2.2.3.4*

**Observations**: The regression indicates a weak positive correlation between age and exercise-induced symptoms, meaning older individuals may be more prone to symptoms during physical activity.

### *2.2.3.3 Analysis 3: These scatter plots visualize the relationship between age and the symptoms of wheezing and shortness of breath.*

**Analysis Method:** The study employs lattice scatter plots to visually investigate the link between age and the occurrence of wheezing and shortness of breath, with the symptoms represented as binary outcomes (0 = No, 1 = Yes). various plots assist to demonstrate trends in how the chance of experiencing various symptoms vary with age, and regression lines are used to emphasise overall patterns. The use of scatter plots simplifies the visualisation of this data, making it easier to grasp how asthma symptoms increase or decrease in frequency as people age.

**Code:**

```
# Scatter plot for Age vs Wheezing
xyplot(Wheezing ~ Age, data=asthma_data,
       xlab="Age", ylab="Wheezing (0=No, 1=Yes)",
       main="Correlation between Age and Wheezing",
       type=c("p", "r"))

# Scatter plot for Age vs Shortness of Breath
xyplot(ShortnessOfBreath ~ Age, data=asthma_data,
       xlab="Age", ylab="Shortness of Breath (0=No, 1=Yes)",
       main="Correlation between Age and Shortness of Breath",
       type=c("p", "r"))
```

*Figure 2.2.3.5*

**Vindication**: Wheezing and shortness of breath are key asthma symptoms. Tracking these across age groups helps identify populations that may need more intensive monitoring and treatment.

**Screenshot of Code Output:**

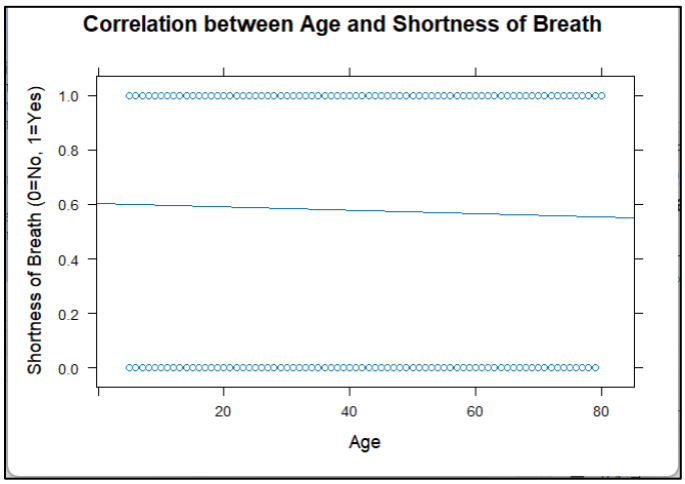



*Figure 2.2.3.6*

**Observations**: Both plots indicate a slight increase in wheezing and shortness of breath with age. Older individuals seem more prone to these symptoms.

### *2.2.3.4 Analysis 4: These plots visualize the relationship between age and lung function measures (FEV1 and FVC) using interactive plotly charts*

**Analysis Method:** The study employs a box plot and a scatter plot to depict the association between age and lung function, notably FEV1 (Forced Expiratory Volume) and FVC. The box plot compares FEV1 across age groups, demonstrating that lung performance declines with age. The scatter plot for FVC contains a regression line to demonstrate the downward trend in lung capacity as age increases. Both charts are interactive using plotly, allowing users to explore the data dynamically. These visualisation approaches were chosen because they are simple and effective in demonstrating how age affects lung function, making the trends easy to grasp and explore interactively.

**Code:**

```r
# Box plot for Age vs LungFunctionFEV1
p1 <- ggplot(asthma_data, aes(x=factor(Age), y=LungFunctionFEV1)) +
  geom_boxplot() +
  labs(title="Lung Function FEV1 across Age Groups", x="Age Groups"
       , y="Lung Function FEV1")

# Convert to interactive plot with plotly
ggplotly(p1)

# Scatter plot for Age vs LungFunctionFVC
p2 <- ggplot(asthma_data, aes(x=Age, y=LungFunctionFVC)) +
  geom_point() +
  geom_smooth(method="lm", color="red") +
  labs(title="Correlation between Age and Lung Function FVC",
       x="Age", y="Lung Function FVC")

# Convert to interactive plot with plotly
ggplotly(p2)
```

*Figure 2.2.3.7*

**Vindication:** Lung function declines naturally with age, but it's particularly important to monitor in asthma patients. Interactive charts help clinicians and patients explore these trends in a more intuitive way.

**Screenshot of Code Output**:

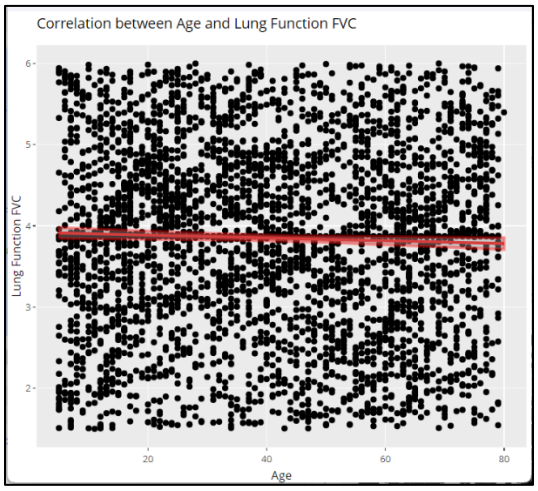



*Figure 2.2.3.8*

**Observations**: The box plot and scatter plot show a clear decline in lung function (FEV1 and FVC) with age, confirming that lung capacity diminishes in older age groups.

### *2.2.3.5.Analysis 5: Logistic regression predicting asthma diagnosis based on age, evaluated with a ROC curve.*

**Analysis Method:** The study used a logistic regression model to investigate the association between age and the risk of an asthma diagnosis. The model estimates the likelihood of an asthma diagnosis depending on age. To assess the efficacy of this prediction model, a ROC (Receiver Operating Characteristic) curve is created, which visually depicts the trade-off between true and false positive rates. The AUC (Area Under the Curve) value is produced to assess the model's predictive capacity, with larger values suggesting greater discrimination between people with and without asthma. These approaches are used because logistic regression is appropriate for binary outcomes (diagnosis: yes/no), and the ROC curve with AUC gives a clear indicator of the model's accuracy and efficacy in predicting asthma diagnoses based on age.

**Code:**

```
# Logistic regression model for Age predicting Diagnosis
logit_model <- glm(Diagnosis ~ Age, family="binomial", data=asthma_data)

# ROC Curve
roc_curve <- roc(asthma_data$Diagnosis, fitted(logit_model))
plot(roc_curve, main="ROC Curve for Age predicting Asthma Diagnosis")

# Calculate AUC
auc_value <- auc(roc_curve)
auc_value
```

*Figure 2.2.3.9*

**Vindication**: Identifying how well age predicts asthma diagnosis helps guide early interventions and screenings for high-risk populations, improving overall asthma management.

**Screenshot of Code Output**:

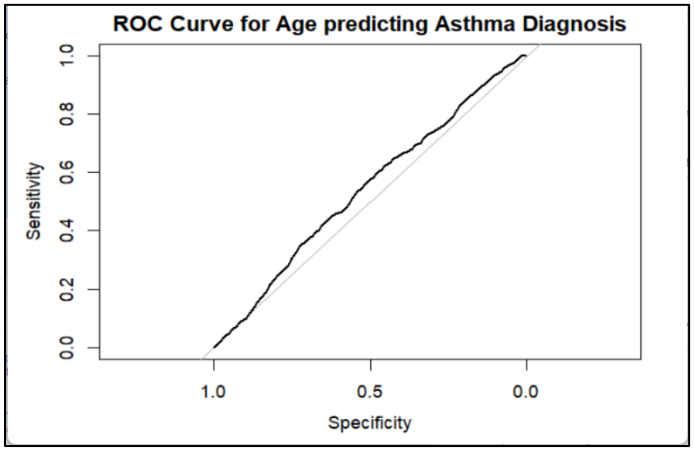



*Figure 2.2.3.10*

**Observations**: The ROC curve and AUC value indicate that age is a moderately good predictor of asthma diagnosis. A higher AUC means the model has a better predictive ability.

### *2.2.3.6 Analysis 6: A box plot to examine how FEV1 (lung function) varies across age groups.*

**Analysis Method:** This study used a box plot to depict the link between age and Lung Function FEV1 (Forced Expiratory Volume in 1 second) across various age groups. The box plot depicts the distribution of lung function by age group, demonstrating how FEV1 fluctuates as people age. Box plots are useful for comparing distributions since they provide the median, quartiles, and outliers of lung function for each age group. This approach was chosen for its simplicity and clarity in depicting disparities in lung function across age groups, allowing for a clear view of the general pattern of lung function reduction as age increases.

**Code:**

```
# Box plot for Age vs LungFunctionFEV1
ggplot(asthma_data, aes(x=factor(Age), y=LungFunctionFEV1)) +
  geom_boxplot() +
  labs(title="Lung Function FEV1 across Age Groups", x="Age Groups",
       y="Lung Function FEV1")
```

*Figure 2.2.3.11*

**Vindication**: FEV1 is a critical measure of lung health in asthma patients. Tracking its decline across age groups helps clinicians monitor lung function deterioration in older populations.

**Screenshot of Code Output**:

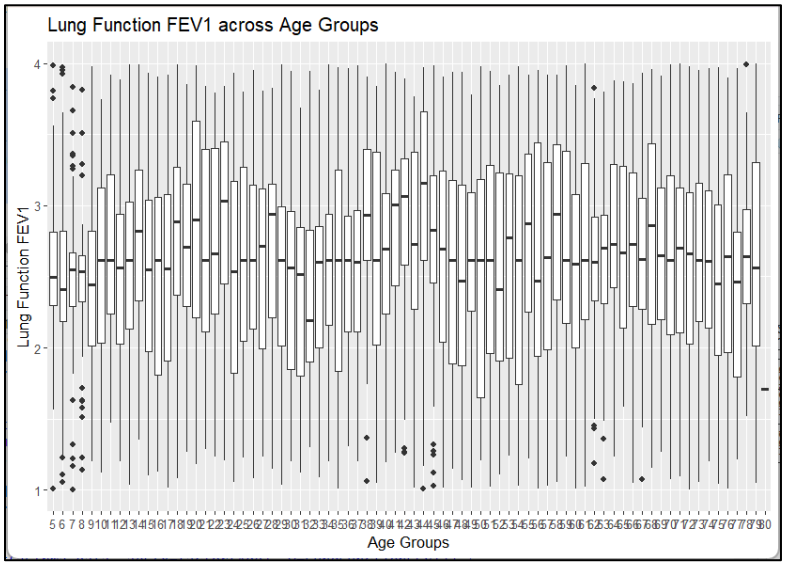



*Figure 2.2.3.12*

**Observations**: FEV1 decreases significantly in older age groups, showing reduced lung function as individuals age.

### *2.2.3.7 Analysis 7: Scatter plot showing the relationship between age and BMI.*

**Analysis Method:** This research use a scatter plot to investigate the association between age and BMI (Body Mass Index). Each point represents an individual's age and BMI, and a linear regression line (in purple) is used to depict the overall trend. The regression line shows how BMI changes with age, with the slope showing whether BMI increases or decreases as people become older. This approach was chosen because it is simple and effective in graphically representing how age affects BMI, allowing for easy interpretation of the general trend in the data.

**Code:**

```
# Scatter plot for Age vs BMI
ggplot(asthma_data, aes(x=Age, y=BMI)) +
  geom_point() +
  geom_smooth(method="lm", color="purple") +
  labs(title="Age vs BMI", x="Age", y="BMI")
```

*Figure 2.2.3.13*

**Vindication**: Higher BMI is associated with worse asthma outcomes. Monitoring BMI changes with age helps target interventions that focus on weight management for asthma patients.

**Screenshot of Code Output**:

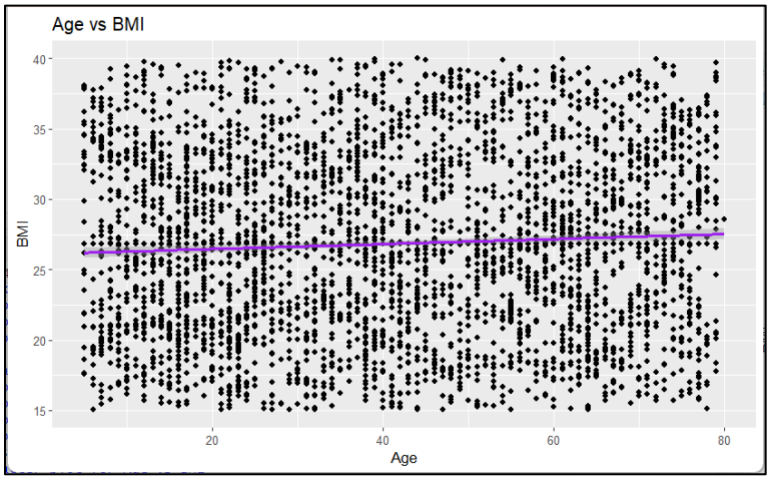



*Figure 2.2.3.14*

**Observations**: BMI tends to increase with age, especially in middle-aged individuals, although there is considerable variability across different age groups.

### *2.2.3.8 Analysis: Bar plot showing the proportion of smokers across different age groups.*

**Analysis Method:** This research use a bar plot to depict the proportion of smokers in various age groups. The bars indicate age categories, and each bar is filled with the percentage of smokers and nonsmokers in that group. The position="fill" parameter stacks the bars by percentage, allowing you to easily compare the prevalence of smoking across age groups. This approach is useful for showing categorical data, allowing for a clear comparison of smoking behaviours across age groups. It quickly determines whether age groups have greater or lower smoking rates, which is crucial for directing smoking-related therapies in asthma care.

**Code:**

```
# Bar plot for Age vs Smoking
ggplot(asthma_data, aes(x=factor(Age), fill=Smoking)) +
  geom_bar(position="fill") +
  labs(title="Proportion of Smokers across Age Groups", x="Age",
       y="Proportion of Smokers")
```

*Figure 2.2.3.15*

**Vindication**: Smoking is a known trigger for asthma. Understanding the distribution of smokers across age groups helps design targeted interventions to reduce smoking-related asthma complications.

**Screenshot of Code Output**:

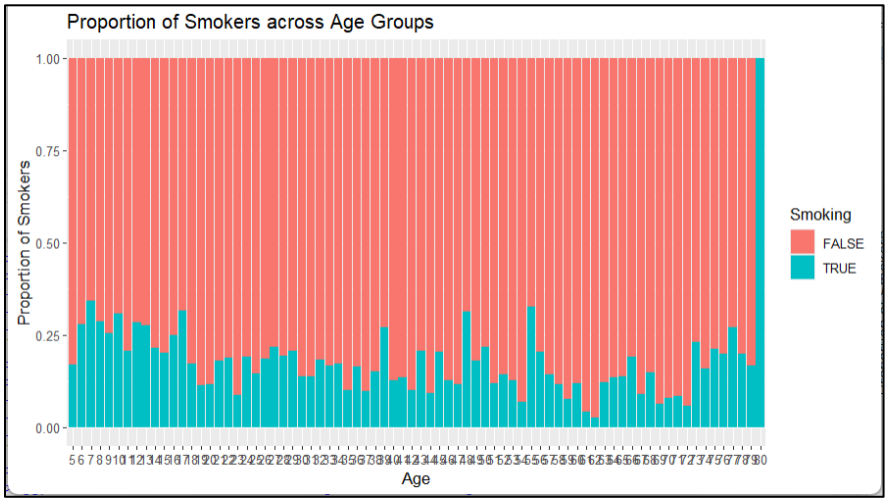



*Figure 2.2.3.16*

**Observations**: Smoking is more prevalent in middle-aged individuals, with lower rates among younger and older populations.

### *2.2.3.9 Analysis: Interactive scatter plot using plotly to visualize the relationship between age and physical activity.*

**Analysis Method:** This study used a scatter plot to investigate the association between age and physical activity, with a linear regression line added to highlight the trend. The plot is turned to an interactive plot using plotly, allowing viewers to zoom in, hover over data points, and examine the connection in a more dynamic manner. The scatter plot depicts how physical activity levels vary with age, but the regression line indicates the general pattern, such as whether physical activity rises or falls as people age. This approach was chosen for its simplicity and greater interactivity, which allows for more detailed and user-friendly data analysis and exploration.

**Code:**

```
# Scatter plot for Age vs Physical Activity using plotly
p1 <- ggplot(asthma_data, aes(x=Age, y=PhysicalActivity)) +
  geom_point() +
  geom_smooth(method="lm", color="blue") +
  labs(title="Age vs Physical Activity", x="Age", y="Physical Activity")

# Convert ggplot2 plot to interactive plotly plot
ggplotly(p1)
```

*Figure 2.2.3.17*

**Vindication**: Maintaining physical activity is crucial for asthma management. Understanding how activity declines with age helps clinicians develop appropriate activity recommendations for older patients.

**Screenshot of Code Output**:

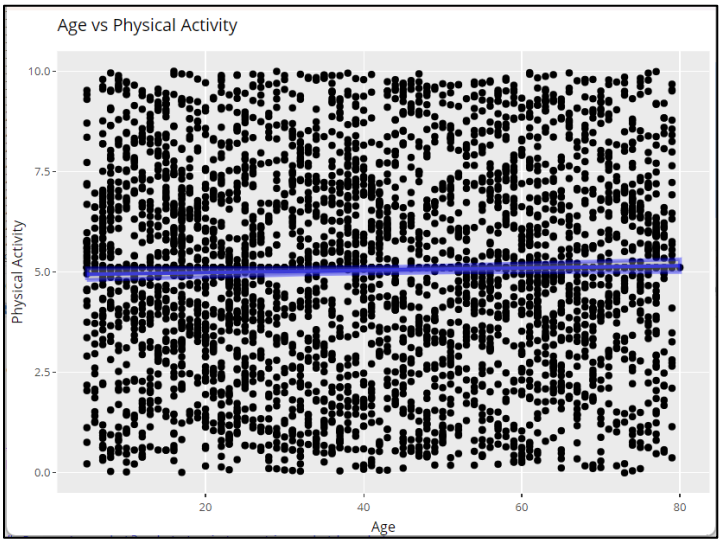



*Figure 2.2.3.18*

**Observations**: Physical activity declines with age, with older individuals generally reporting less physical activity.

### *2.2.3.10 Analysis 10: Scatter plot using lattice to analyze the relationship between age and sleep quality.*

**Analysis Method**: This research investigates the association between age and sleep quality using a scatter plot and the lattice library. Each point represents an individual's age and sleep quality score, with a regression line (type "r") given to demonstrate the data's trend. This graphic reveals how sleep quality changes with age, making it simpler to identify patterns, such as whether older people have poorer sleep quality. The lattice method gives an organised, understandable visual representation, making it useful for finding patterns in the link between age and sleep quality.

**Code**:

```
# Scatter plot for Age vs Sleep Quality using lattice
xyplot(SleepQuality ~ Age, data=asthma_data,
       xlab="Age", ylab="Sleep Quality",
       main="Age vs Sleep Quality",
       type=c("p", "r"))
```

*Figure 2.2.3.19*

**Vindication**: Sleep quality is directly linked to asthma control. Monitoring changes in sleep quality with age helps guide interventions to improve sleep, especially in older patients.

**Screenshot of Code Output**:

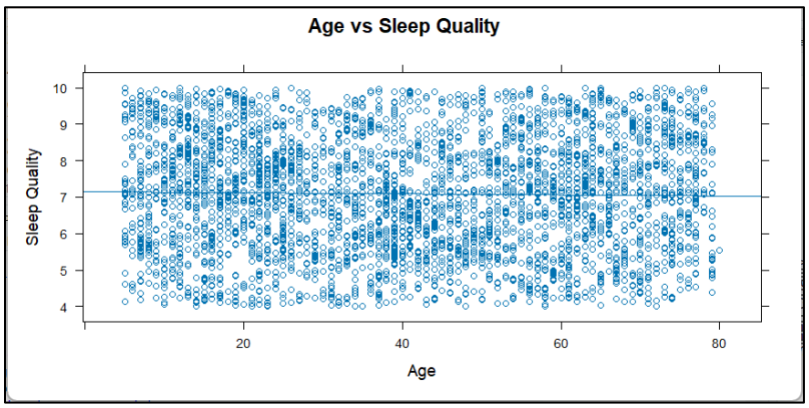



*Figure 2.2.3.20*

**Observations**: Sleep quality slightly declines with age, indicating that older individuals may experience more sleep disturbances.

*2.2.3.11 Analysis 11: The scatter plot shows the age-related relationship with diet quality, using the viridis color scale for better interpretation, and a linear regression line illustrating the general trend.*

**Analysis Method:** This research used a scatter plot to investigate the association between age and food quality, with the viridis colour scale used for improved visual contrast. Each point represents an individual's age and diet quality, with colour intensity denoting the degree of diet quality. A linear regression line is added to depict the overall trend, revealing how food quality varies with age. The viridis colour scale was chosen for its visual clarity and colorblind-friendliness, which make it simpler to analyse data patterns. This approach is useful for determining if nutrition quality improves or diminishes with age in a visually appealing and understandable style.

**Code:**

```
# Scatter plot for Age vs Diet Quality using ggplot2 and viridis color scale
ggplot(asthma_data, aes(x=Age, y=DietQuality)) +
  geom_point(aes(color=DietQuality), size=3) +
  geom_smooth(method="lm", color="red") +
  scale_color_viridis(option="C") +
  labs(title="Age vs Diet Quality", x="Age", y="Diet Quality")
```

*Figure 2.2.3.21*

**Vindication**: Diet quality significantly impacts asthma management, with healthcare providers designing interventions to improve nutrition, particularly for older individuals with lower diet quality.

**Screenshot of Code Output**:

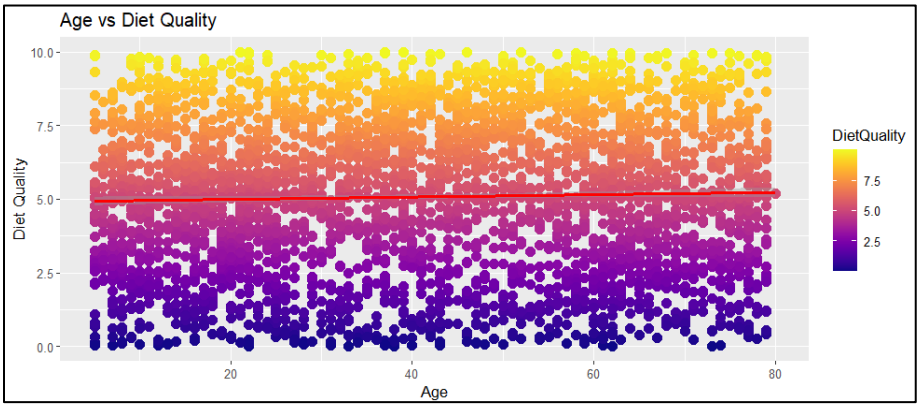



*Figure 2.2.3.22*

**Observations**: The plot indicates that as individuals age, their diet quality slightly decreases, with a weaker relationship observed in older individuals.

*2.2.3.12 Analysis: The first plot (A) depicts the correlation between smoking and exercise-induced symptoms, with each bar representing smokers and non-smokers, and the color indicating whether they experience asthma symptoms. The second scatter plot (B) reveals a correlation between age and exercise-induced symptoms in smokers, using a linear regression line to highlight the trend.*

**Analysis Method:** This analysis combines two visualizations to explore the relationship between smoking, exercise-induced symptoms, and age.

**1. Bar Plot** (Smoking versus Exercise-Induced Symptoms): The first plot is a bar chart that compares the number of smokers and nonsmokers who suffer from exercise-induced symptoms. The bars are separated to illustrate how many people in each group (smokers and nonsmokers) had these symptoms, providing a clear picture of the link between smoking and exercise-induced symptoms.

**Code:**

```
# Create a bar plot for Smoking vs Exercise-Induced Symptoms
p1 <- ggplot(asthma_data, aes(x=Smoking, fill=factor(ExerciseInduced))) +
  geom_bar(position="dodge") +
  labs(title="Smoking vs Exercise-Induced Symptoms", x="Smoking", y="Count")

# Create a scatter plot for Age vs Exercise-Induced Symptoms for smokers only
p2 <- ggplot(asthma_data %>% filter(Smoking == TRUE), aes(x=Age, y=ExerciseInduced)) +
  geom_point() +
  geom_smooth(method="lm", color="red") +
  labs(title="Age vs Exercise-Induced Symptoms (Smokers)", x="Age", y="Exercise-Induced Symptoms")

# Combine the two plots into a single layout
combined_plot <- plot_grid(p1, p2, labels=c("A", "B"), ncol=2)
print(combined_plot)
```

*Figure 2.2.3.23*

**Vindication:** This plot illustrates smoking's impact on exercise-induced asthma, highlighting age's role in severity, and smoking as a known risk factor.

**Screenshot of Code Output**:

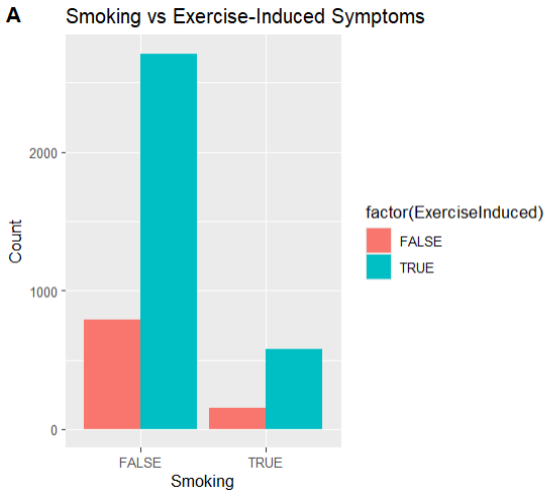



*Figure 2.2.3.24*

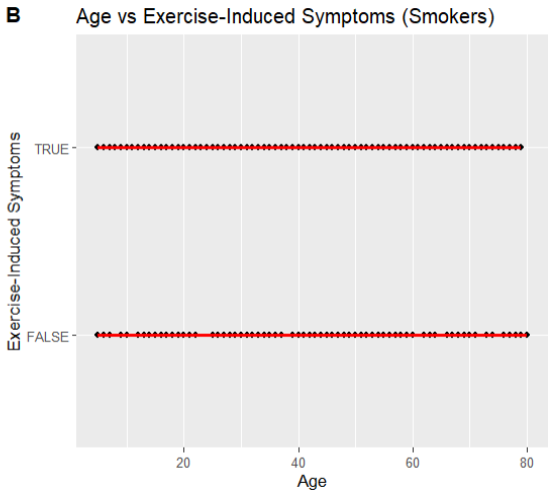



*Figure 2.2.3.25*

**Observations:** Bar Plot (A): Smokers have a higher count of exercise-induced symptoms compared to non-smokers, suggesting that smoking is linked to increased exercise-induced symptoms.

Scatter Plot (B): Among smokers, the scatter plot shows that as age increases, exercise-induced symptoms become more frequent, as indicated by the positive slope of the regression line.

#### *2.2.3.13 Conclusion for the objective*

In this R Studio analysis, we looked at how age affects several asthma-related symptoms and health parameters, including nighttime symptoms, exercise-induced symptoms, lung function, BMI, food quality, and physical activity. The findings clearly demonstrate that age is a critical factor in influencing the severity of asthma symptoms and associated health effects.

Older people were shown to have a greater loss in lung function, a higher risk of suffering nocturnal and exercise-induced symptoms, and a higher BMI, all of which can worsen asthma. Furthermore, smoking was observed to exacerbate exercise-induced symptoms, especially in older people.

This research gives useful information for creating age-appropriate asthma management methods. Healthcare providers can assist reduce the detrimental effects of asthma in ageing populations by focussing on improving lifestyle variables such as food, physical exercise, and weight control, as well as controlling symptom aggravation in older patients.

*2.2.4 Objective 4 : To Investigate the relationship between age and the overall number of asthma symptons (wheezing, coughing, shortness of breath, chest tightness) reported by patients in the dataset*

*2.2.4.1 Analysis 1: Perform a Fisher's Exact Test to analyze the association between age groups and the presence of the Wheezing symptom.*

**Code**:

```
#Using Fisher's Exact Test for Wheezing
fisher_test_wheezing ← fisher.test(table(asthma_dataset$Age_Group, asthma_dataset$Wheezing))
print(fisher_test_wheezing)
```

*Figure 2.2.4.1*

The objective of this test is to determine if the distribution of the Wheezing symptom is independent of the variable of age.

The Fisher's Exact Test will then calculates the probability of obtaining a distribution of values as extreme as, or more extreme than, the observed distribution.

To do this, the test produces a p-value, where if it is below 0.05, it determines a significant association between the symptom and the age group. And if it is equal to or above 0.05, there is no significant association between the symptom and the age group.

**Result:**

```
> fisher_test_wheezing ← fisher.test(table(asthma_dataset$Age_Group, asthma_dataset$Wheezing))
> print(fisher_test_wheezing)

        Fisher's Exact Test for Count Data

data:  table(asthma_dataset$Age_Group, asthma_dataset$Wheezing)
p-value = 0.5346
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.9052326 1.2120015
sample estimates:
odds ratio
  1.047429
```

*Figure 2.2.4.2*

According to the p-value, which is 0.5346, which is above 0.05, this proves that age does not play a significant role in determining the presence of the Wheezing symptom.

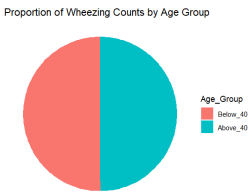



*Figure 2.2.4.3*

**Code to generate Pie Chart:**

```
#Building Pie Chart for Wheezing Counts against Age Groups
wheezing_pie ← melt(symptom_counts, id.vars = "Age_Group", measure.vars = "wheezing_count")
ggplot(wheezing_pie, aes(x = "", y = value, fill = Age_Group)) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar("y") +
  labs(title = "Proportion of Wheezing Counts by Age Group") +
  theme_void()
```

*Figure 2.2.4.4*

*2.2.4.2 Analysis 2: Perform a Chi-square Test to analyze the association between age and the presence of the Coughing symptom.*

**Code:**

```
#Using Chi-square Test for Coughing
asthma_dataset$Coughing ← factor(asthma_dataset$Coughing, levels = c(FALSE, TRUE))
chi_square_coughing ← chisq.test(table(asthma_dataset$Age_Group, asthma_dataset$Coughing))
print(chi_square_coughing)
```

*Figure 2.2.4.5*

The objective of this test is to determine if the distribution of the Coughing symptom is independent of the variable of age.

This Chi-square test will build a contingency table to summarize the frequency of Coughing occurrences and the age, and it will compute the Chi-Square statistic by comparing the observed frequencies to the expected frequencies using the formula:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

*Figure 2.2.4.6*

where $O^i$ is the observed frequency and $E^i$ is the expected frequency for each cell.

The result from that equation will then be compared to a Chi-square distribution to obtain the p-value, where if it is below 0.05, it determines a significant association between the symptom and the age group. And if it is equal to or above 0.05, there is no significant association between the symptom and the age group.

**Result:**

```
> asthma_dataset$Coughing ← factor(asthma_dataset$Coughing, levels = c(FALSE, TRUE))
> chi_square_coughing ← chisq.test(table(asthma_dataset$Age_Group, asthma_dataset$Coughing))
> print(chi_square_coughing)

        Pearson's Chi-squared test with Yates' continuity correction

data:  table(asthma_dataset$Age_Group, asthma_dataset$Coughing)
X-squared = 7.1343, df = 1, p-value = 0.007562
```

*Figure 2.2.4.7*

According to the p-value obtained from the test, which is 0.007562, this proves that age plays quite a significant role in the presence of the Coughing symptom, where people above the age of 40 experience coughing more than those who are below the age of 40.

**Bar chart showing the relationship between age and the presence of the Coughing symptom:**

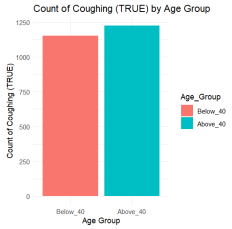



*Figure 2.2.4.8*

**Code to generate the Bar Chart:**

```
#Building Bar Chart for Coughing Counts against Age Groups
ggplot(symptom_counts, aes(x = Age_Group, y = coughing_count, fill = Age_Group)) +
  geom_bar(stat = "identity") +
  labs(title = "Count of Coughing (TRUE) by Age Group", x = "Age Group", y = "Count of Coughing (TRUE)") +
  theme_minimal()
```

*Figure 2.2.4.9*

### *2.2.4.3 Analysis 3: Perform a Logistic Regression Analysis to predict the probability of the presence of the Shortness of Breath symptom based on age.*

**Code:**

```
#Using Logistic Regression Analysis for Shortness of Breath
logistic_model_short_breath ← glm(ShortnessOfBreath ~ Age, data = asthma_dataset, family = binomial)
summary(logistic_model_short_breath)
```

*Figure 2.2.4.9*

The objective of this test is to determine if age causes a higher occurrence of the Shortness of Breath symptom.

The Logistic Regression Analysis is suitable as it is built for tests where the outcome variable is binary (eg. True or False).

By analyzing the coefficients and predicting probabilities, this test helps us understand the strength and direction of the relationship between age and the presence of the Shortness of Breath symptom.

Through the summary of this test, if it shows a positive coefficient value, then it suggests that as age develops, the likelihood of having shortness of breath increases.

**Result:**

```
> #Using Logistic Regression Analysis for Shortness of Breath
> logistic_model_short_breath ← glm(ShortnessOfBreath ~ Age, data = asthma_dataset, family = binomial)
> summary(logistic_model_short_breath)

Call:
glm(formula = ShortnessOfBreath ~ Age, family = binomial, data = asthma_dataset)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.421748   0.067059   6.289 3.19e-10 ***
Age         -0.002617   0.001462  -1.789   0.0735 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5766.7  on 4234  degrees of freedom
Residual deviance: 5763.5  on 4233  degrees of freedom
AIC: 5767.5

Number of Fisher Scoring iterations: 4
```

*Figure 2.2.4.10*

According to the coefficient value, which is -0.002617, this test predicts that development in age does not affect the presence of shortness of breath.

**Heat Map showing the Relationship between Age and the Shortness of Breath symptom:**

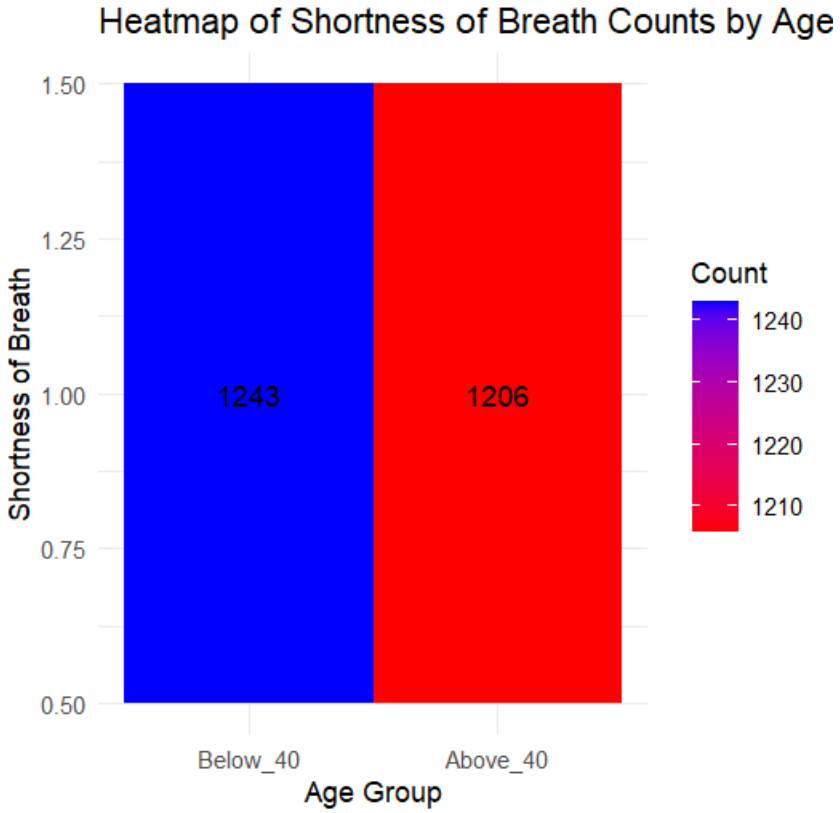



*Figure 2.2.4.11*

**Code to generate the Heat Map:**

```
#Building Heat Map for Shortness of Breath Counts against Age Groups
short_breath_matrix ← matrix(symptom_counts$short_breath_count, nrow = 1, dimnames = list(NULL, symptom_counts$Age_Group))
short_breath_melted ← melt(short_breath_matrix)
ggplot(short_breath_melted, aes(x = Var2, y = Var1, fill = value)) +
  geom_tile() +
  geom_text(aes(label = value), color = "black") +
  scale_fill_gradient(low = "red", high = "blue") +
  labs(title = "Heatmap of Shortness of Breath Counts by Age Group", x = "Age Group", y = "Shortness of Breath", fill = "Count") +
  theme_minimal()
```

*Figure 2.2.4.12*

*Figure 2.2.4.4. Analysis 4: Perform a T-Test to compare the means of individuals with and without chest tightness. Doing this helps find out if there is a difference in their average ages.*

**Code:**

```
#Using T-Test for Chest Tightness
asthma_dataset$Chest_Tightness ← as.numeric(asthma_dataset$ChestTightness)
t_test_chest_tightness ← t.test(Age ~ ChestTightness, data = asthma_dataset)
print(t_test_chest_tightness)
```

*Figure 2.2.4.13*

The objective of this test is to determine if age plays a determining role in the presence of chest tightness among individuals.

This T-Test calculates the T-statistic, which helps to measure the size in the difference in both means relative to the variability. The result is then compared to a cricitical value from the T-distribution to determine a p-value.

The T-Test uses the following formula:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

*Figure 2.2.4.14*

where $X_1$ and $X_2$ are the means of the two groups, $s^2_1$ and $s^2_2$ are the variances, and $n_1$ and $n_2$ are the sample sizes.

**Result:**

```
> #Using T-Test for Chest Tightness
> asthma_dataset$Chest_Tightness ← as.numeric(asthma_dataset$ChestTightness)
> t_test_chest_tightness ← t.test(Age ~ ChestTightness, data = asthma_dataset)
> print(t_test_chest_tightness)

        Welch Two Sample t-test

data:  Age by ChestTightness
t = -2.7421, df = 3936.3, p-value = 0.006132
alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
95 percent confidence interval:
 -3.1045598 -0.5159683
sample estimates:
mean in group FALSE  mean in group TRUE
           39.40751            41.21777
```

*Figure 2.2.4.15*

According to this test, the p-value is 0.006132, which means that there is a significant difference between the means of individuals with chest tightness, and those who do not. From this, we can conclude that age does play a somewhat significant role in the presence of chest tightness.

**Line Chart for Chest Tightness against Age:**

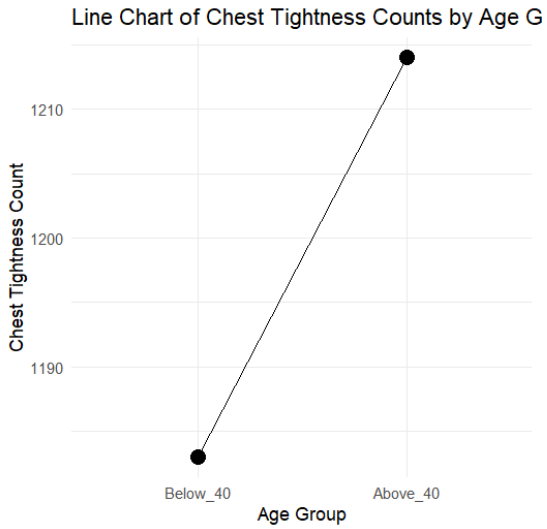



*Figure 2.2.4.16*

**Code to generate Line Chart:**

```
#Building a Line Chart for Chest Tightness counts against Age Groups.
line_chart_data ← symptom_counts %>%
  select(Age_Group, chest_tightness_count)
ggplot(line_chart_data, aes(x = Age_Group, y = chest_tightness_count, group = 1)) +
  geom_line() +
  geom_point(size = 4) +
  labs(title = "Line Chart of Chest Tightness Counts by Age Group",
       x = "Age Group",
       y = "Chest Tightness Count") +
  theme_minimal()
```

*Figure 2.2.4.17*

### *2.2.4.5 Conclusion of the Objective*

In conclusion, the above tests show us that age plays a role in the presence of chest tightness and coughing, but not so much when it comes to wheezing and shortness of breath.

This R Studio analysis shows that age is not a significant factor in causing the presence of asthmatic symptoms, but they do play a small part in causing some of them, such as coughing and chest tightness.

This research can help give useful and knowledgeable information about asthma and its symptoms, as well as help healthcare providers tackle this issue that has been growing throughout our community for a long time.

**Extra Features**

Heatmap - (LEHWIKS RAJ A/L GUNASELAN THANGARAJ | TP061519)
t.test - (LEHWIKS RAJ A/L GUNASELAN THANGARAJ | TP061519)
Logistic Regression Test - (LEHWIKS RAJ A/L GUNASELAN THANGARAJ | TP061519)

library(reshape2) - (LEHWIKS RAJ A/L GUNASELAN THANGARAJ | TP061519)
fisher.test - (LEHWIKS RAJ A/L GUNASELAN THANGARAJ | TP061519)
chisq.test - (LEHWIKS RAJ A/L GUNASELAN THANGARAJ | TP061519)
melt - (LEHWIKS RAJ A/L GUNASELAN THANGARAJ | TP061519)
matrix - (LEHWIKS RAJ A/L GUNASELAN THANGARAJ | TP061519)
geom_bar - (LEHWIKS RAJ A/L GUNASELAN THANGARAJ | TP061519)
geom_tile - (LEHWIKS RAJ A/L GUNASELAN THANGARAJ | TP061519)
geom_point - (LEHWIKS RAJ A/L GUNASELAN THANGARAJ | TP061519)

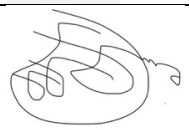
## Conclusion

This comprehensive analysis highlights the multifaceted relationship between age and asthma-related symptoms. It reveals that younger patients (ages 5-18) experience higher frequencies of wheezing, emphasizing the need for early intervention in respiratory health among youth. Conversely, older patients (ages 19-80) are more prone to deteriorations in lung function, higher BMI, and nocturnal or exercise-induced symptoms, which worsen asthma severity. Environmental factors, such as pollution and smoking, further contribute to worsening conditions in older populations.

Despite the clear correlation between age, coughing, and asthma diagnosis, these two variables alone are not enough to reliably predict an asthma diagnosis. This means that the healthcare providers need to consider a broader range of symptoms and contributing factors when diagnosing asthma.

The findings provide valuable insights for developing tailored, age-specific asthma management strategies. For younger patients, targeted therapies addressing frequent wheezing are essential, while for older patients, interventions should focus on lifestyle modifications, such as diet, physical activity, weight management, and minimizing environmental exposure. By considering these factors holistically, healthcare providers can better mitigate the impacts of asthma across different age groups and improve overall patient outcomes

# Workload Matrix

| Student Name | Student ID | Tasks | Percentage | Signature |
|---|---|---|---|---|
| HTET AUNG HLAING | TP075706 | Objective 2, File Compiling, Project Management, Data Cleaning, Document Compliation | 25% | HTET AUNG HLAING |
| JONATHAN NG'UA MBAI | TP075128 | Objective 1 | 25% | JONATHAN |
| LEHWIKS RAJ A/L GUNASELAN THANGARAJ | TP061519 | Objective 4 | 25% |  |
| FARES ADEL OMER BA MOHRES | TP071376 | Objective 3 | 25% |  |