Lifestyle Choices and Level of Lung Cancer - A Statistical Inquiry

Group – 4 - Mansi Pagare, Osama Pervez Abdul Rehman, Vishra Shah, Zahra Al Zara

Indiana University-Purdue University

Introduction:

Lung cancer is a fatal disease that caused 1.59 million deaths in 2018, making it the primary cause of cancer-related deaths globally (Thandra et. al., 2021). Although smoking accounts for up to 85% of all cases, recent research indicates that air pollution can raise the risk of lung cancer even in non-smokers (CDC, 2020). Air pollution is responsible for approximately 7 million deaths worldwide each year, according to the World Health Organization (WHO, 2023), making it the largest environmental health risk. This study aims to examine the link between lifestyle factors and lung cancer level, analyzing 1000 unique data points to improve public health programs and preventive measures for lung cancer.

Hypothesis

Null hypothesis: There is no significant relationship between the variables depicting lifestyle choices and the level of lung cancer.
Alternate hypothesis: There is a significant relationship between the variables depicting lifestyle choices and the level of lung cancer.

Methodology:

Research question

Which lifestyle factor has the highest influence on the level of lung cancer?

Study Design

This research employs a quantitative correlation approach to measure and quantify the relationship between variables using numerical data analysis.

Data collection

Dataset used in the project was obtained from Kaggle, containing information on 1000 lung cancer patients from two regions in China with different pollution levels. Data was collected over six years via medical records, surveys, and environmental monitoring, and compiled into a CSV format. The dataset includes 26 variables, one being numerical and others categorical, scoring on scales of 1-7 or 1-8.

Data analysis:

We performed exploratory data analysis to understand data distribution, null and duplicate values, and the relationship between variables and the target variable. This helped us summarize key insights and interpret the hypothesis. After EDA, we found no null or duplicate values and converted categorical columns to categorical data to achieve the project goal.

```
## Checking for null values
```{r}
colSums(is.na(df))
```
```

```
index                Patient.Id                       Age                    Gender               Air.Pollution 
    0                         0                         0                         0                         0 
Alcohol.use               Dust.Allergy      OccuPational.Hazards              Genetic.Risk      chronic.Lung.Disease 
    0                         0                         0                         0                         0 
Balanced.Diet               Obesity                   Smoking           Passive.Smoker                Chest.Pain 
    0                         0                         0                         0                         0 
Coughing.of.Blood           Fatigue               Weight.Loss      Shortness.of.Breath                 Wheezing 
    0                         0                         0                         0                         0 
Swallowing.Difficulty Clubbing.of.Finger.Nails  Frequent.Cold               Dry.Cough                  Snoring 
    0                         0                         0                         0                         0 
Level 
    0 
```
Checking for duplicates
```{r}
df[duplicated(df),]
```
```

0 rows | 1-10 of 26 columns
```

Descriptive Statistics

Understanding of the essential elements of the dataset is necessary hence summary will help in making an informed decision about the data sample and its measurements.

```{r}
summary(df)
```
```
     index        Patient.Id             Age        Gender  Air.Pollution  Alcohol.use   Dust.Allergy OccuPational.Hazards Genetic.Risk
 Min.   :  0.0   Length:1000        Min.   :14.00   1:598   6      :326   2      :202   7      :405   7      :365          1: 40       
 1st Qu.:249.8   Class :character   1st Qu.:27.75   2:402   2      :201   8      :188   4      :133   3      :151          2:212       
 Median :499.5   Mode  :character   Median :36.00           3      :173   7      :167   5      :111   2      :132          3:173       
 Mean   :499.5                      Mean   :37.17           1      :141   1      :152   6      :110   5      :130          4: 40       
 3rd Qu.:749.2                      3rd Qu.:45.00           4      : 90   5      : 90   3      :101   4      :112          5:100       
 Max.   :999.0                      Max.   :73.00           7      : 30   3      : 80   2      : 70   1      : 50          6:108       
                                                            (Other): 39   (Other):121   (Other): 70   (Other): 60          7:327       
 chronic.Lung.Disease Balanced.Diet Obesity     Smoking    Passive.Smoker   Chest.Pain  Coughing.of.Blood    Fatigue     Weight.Loss
 1: 50                1: 40         1: 70   2      :222   2      :284    7      :296   7      :187       3      :212   2      :280  
 2:173                2:231         2:140   7      :207   7      :187    4      :191   4      :172       2      :211   7      :230  
 3:141                3:173         3:193   1      :181   4      :161    2      :181   3      :171       4      :180   3      :150  
 4:141                4: 61         4:191   3      :172   3      :140    3      :153   2      :121       1      :110   1      :121  
 5: 80                5: 40         5: 20   8      : 89   8      :108    1      : 80   8      :119       8      :109   5      :100  
 6:308                6:159         6: 30   6      : 60   1      : 60    6      : 40   1      : 71       5      : 89   4      : 60  
 7:107                7:296         7:356   (Other): 69   (Other): 60    (Other): 59   (Other):159       (Other): 89   (Other): 59  
 Shortness.of.Breath    Wheezing   Swallowing.Difficulty Clubbing.of.Finger.Nails Frequent.Cold Dry.Cough Snoring    Level     
 2      :243          2      :240   1      :221           2      :240            1:139         1:119     1:170   High   :365  
 6      :201          5      :171   4      :189           4      :220            2:192         2:251     2:300   Low    :303  
 3      :140          4      :163   2      :160           1      :131            3:230         3:101     3:211   Medium:332  
 4      : 90          1      :149   5      :110           5      :120            4:180         4:141     4:131                 
 7      : 89          7      :139   8      :110           3      :100            5: 20         5:131     5:139                 
 5      : 87          6      : 68   6      : 91           9      : 80            6:170         6: 89     6: 39                 
 (Other):150          (Other): 70   (Other):119           (Other):109            7: 69         7:168     7: 10                 
```

Statistical Testing:

Fisher's exact test is a statistical test used to analyze the association between two categorical variables. It determines if there is a significant association or dependency between the two variables by calculating the probability of obtaining the observed frequency of each category in a contingency table under the null hypothesis of independence. As our sample size is 1000 rows and has 26 variables of which all are categorical variables except age hence, we are doing fisher's exact test by doing a contingency table of each column with our target variable rather than chi-square test as it is applied on larger data.

```
# Create an empty list to store the test results
fisher_results <- list()

# Iterate over each column and perform Fisher's exact test
for (col in cols) {
  # Create a contingency table of the column and the target column
  cont_table <- table(df[, col], df$Level)
  # Perform Fisher's exact test
  fisher_result <- fisher.test(cont_table, simulate.p.value = TRUE, B = 1000)
  # Store the test result in the list
  fisher_results[[col]] <- fisher_result}
# Print the test results
fisher_results
```

```
$Age

        Fisher's Exact Test for Count Data with simulated p-value (based on 1000 replicates)

data:  cont_table
p-value = 0.000999
alternative hypothesis: two.sided


$Gender

        Fisher's Exact Test for Count Data with simulated p-value (based on 1000 replicates)

data:  cont_table
p-value = 0.000999
alternative hypothesis: two.sided


$Air.Pollution

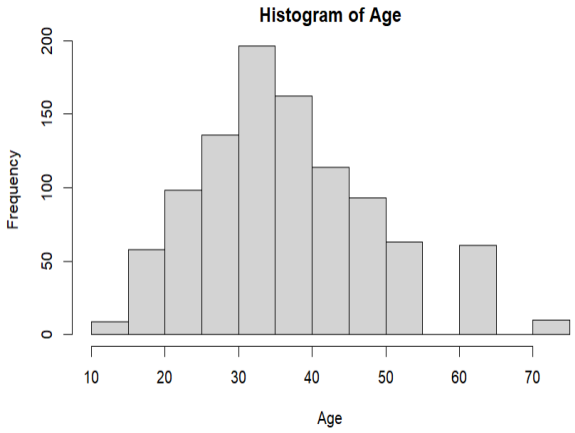
        Fisher's Exact Test for Count Data with simulated p-value (based on 1000 replicates)

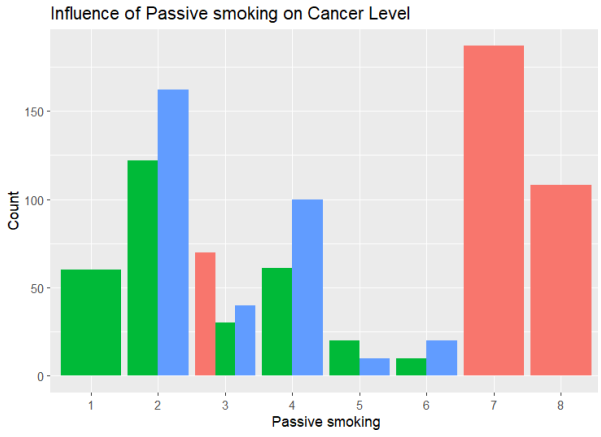
data:  cont_table
p-value = 0.000999
alternative hypothesis: two.sided
```

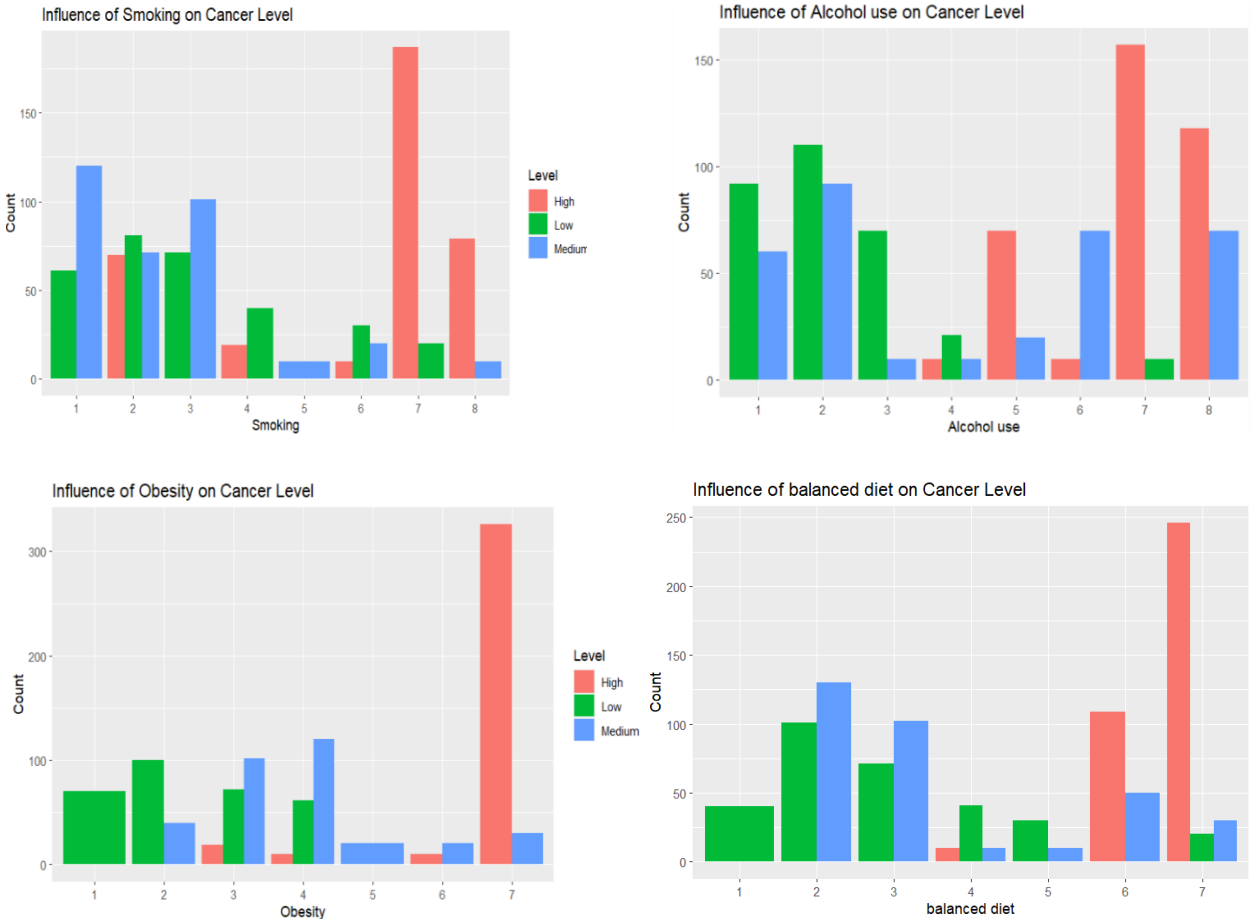
Interpretation of Fisher's Exact test - A simulated p-value has been generated based on the 1000 replicates for the dataset, in all the variables the simulated p-value is lesser than chosen significance 0.05 hence we can understand from this that there is a significant association between the variables and the target variable.

Data Visualization:

For understanding the effect of lifestyle variables on the level of lung cancer we will check the influence of Obesity, Alcohol Use, Smoking, Passive smoking, and Balanced diet on the level of cancer respectively by plotting some graphs which will help us understand the target variable's association with other lifestyle variables.







The histogram between age and frequency shows that our data is normally distributed and showcases the distribution of participants based on their age where the highest count lies between 30-35 years of age. Graph 2 represents the relationship between the level of Cancer and Exposure to Passive Smoking. High exposure to passive smoking increases the probability of cancer occurrence. Graph 3 represents the relationship between the level of cancer and Smoking. The relationship seems to be non-linear as we can see a high level of cancer for both lower and higher levels of smoking. Graph 4 represents the relationship between the level of cancer and Alcohol Consumption. The risk of cancer increases with increased consumption of alcohol. Graph 5 represents the relationship between the level of cancer and obesity. As obesity increases the risk of cancer increases. Graph 6 represents the relationship between the level of Cancer and Balanced Diet. Higher exposure to non-balanced diet can contribute to an increase in the level of cancer.

Logistic Regression Model:

Logistic regression was used to analyze categorical data by modeling the probability of an outcome based on predictor variables. This allowed us to test our hypotheses and draw statistical conclusions about the factors contributing to the risk of lung cancer.

Logistic Regression model is used to understand the relationship between lifestyle variables and cancer. Combinations of variables were tested to achieve an LR model with the lowest deviance and AIC value, resulting in the selected combination that helps to draw a conclusion for the project.

```
## Building LR model with the best AIC value
```{r}
library(nnet)
# create a multinomial logistic regression model
model <- multinom(Level ~ Obesity*Alcohol.use+Passive.Smoker, data = df)
# print the summary of the model
summary(model)
```
```

```
Residual Deviance: 0.0001064186
AIC: 132.0001
```

Result:

Fisher's exact test showed significant association between variables and the target variable. Logistic regression revealed that passive smoking had the highest association, while combining obesity, alcohol use, and passive smoking gave the lowest AIC score and residual deviance.

Conclusion:

We can reject the null hypothesis and prove that there is a significant association between the lifestyle variables (highest association with passive smoking) and the level of cancer thus answering the research question.

Limitations:

1) Our dataset is smaller in size for the research to be conducted for the benefit of the health public initiative hence it does not represent accurately the larger population and hence reduces the generalizability of the statistical testing.
2) The data is collected based on the participants' self-reports, which can lead to recall bias and social desirability bias.
3) Data collection is done from a single demographic region which cannot become a representation for overall world's population as every demographic factors can affect the results of the statistical testing.

References

Thandra, K. C., Barsouk, A., Saginala, K., Aluru, J. S., & Barsouk, A. (2021). Epidemiology of lung cancer. *Contemporary oncology (Poznan, Poland)*, *25*(1), 45–52. https://doi.org/10.5114/wo.2021.103829

CDC. (2020, September 22). What Are the Risk Factors for Lung Cancer?. *Center for Disease Control and Prevention*. https://www.cdc.gov/cancer/lung/basic_info/risk_factors.htm

World Health Organization. (2023). Air Pollution. *World Health Organization: WHO*. https://www.who.int/health-topics/air-pollution#tab=tab_1

Lung Cancer Prediction. (n.d.). *Www.kaggle.com*. https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-new-link

Appendix

Reading the dataset

```{r}
df <- read.csv("~/cancer patient data sets.csv")
head(df)
```

Description: df [6 × 26]

	index <int>	Patient.Id <chr>	Age <int>	Gender <int>	Air.Pollution <int>	Alcohol.use <int>	Dust.Allergy <int>	OccuPational.Hazards <int>	Genetic.Risk <int>
1	0	P1	33	1	2	4	5	4	3
2	1	P10	17	1	3	1	5	3	4
3	2	P100	35	1	4	5	6	5	5
4	3	P1000	37	1	7	7	7	7	6
5	4	P101	46	1	6	8	7	7	7
6	5	P102	35	1	4	5	6	5	5

6 rows | 1-10 of 26 columns

Converting the categorical columns into categorical data

```{r}
# Convert columns 4 to n to factors
df[, 4:(ncol(df))] <- lapply(df[, 4:(ncol(df))], factor)
# Print the structure of the dataframe
str(df)
```

```
'data.frame':   1000 obs. of  26 variables:
 $ index                   : int  0 1 2 3 4 5 6 7 8 9 ...
 $ Patient.Id              : chr  "P1" "P10" "P100" "P1000" ...
 $ Age                     : int  33 17 35 37 46 35 52 28 35 46 ...
 $ Gender                  : Factor w/ 2 levels "1","2": 1 1 1 1 1 1 1 2 2 2 1 ...
 $ Air.Pollution           : Factor w/ 8 levels "1","2","3","4",..: 2 3 4 7 6 4 2 3 4 2 ...
 $ Alcohol.use             : Factor w/ 8 levels "1","2","3","4",..: 4 1 5 7 8 5 4 1 5 3 ...
 $ Dust.Allergy            : Factor w/ 8 levels "1","2","3","4",..: 5 5 6 7 7 6 5 4 6 4 ...
 $ OccuPational.Hazards    : Factor w/ 8 levels "1","2","3","4",..: 4 3 5 7 7 5 4 3 5 2 ...
 $ Genetic.Risk            : Factor w/ 7 levels "1","2","3","4",..: 3 4 5 6 7 5 3 2 6 4 ...
 $ chronic.Lung.Disease    : Factor w/ 7 levels "1","2","3","4",..: 2 2 4 7 6 4 2 3 5 3 ...
 $ Balanced.Diet           : Factor w/ 7 levels "1","2","3","4",..: 2 2 6 7 7 6 2 4 5 3 ...
 $ Obesity                 : Factor w/ 7 levels "1","2","3","4",..: 4 2 7 7 7 7 4 3 5 3 ...
 $ Smoking                 : Factor w/ 8 levels "1","2","3","4",..: 3 2 2 7 8 2 3 1 6 2 ...
 $ Passive.Smoker          : Factor w/ 8 levels "1","2","3","4",..: 2 4 3 7 7 3 2 4 6 3 ...
 $ Chest.Pain              : Factor w/ 9 levels "1","2","3","4",..: 2 2 4 7 7 4 2 3 6 4 ...
 $ Coughing.of.Blood       : Factor w/ 9 levels "1","2","3","4",..: 4 3 8 8 9 8 4 1 5 4 ...
 $ Fatigue                 : Factor w/ 8 levels "1","2","3","4",..: 3 1 7 4 3 7 3 3 1 1 ...
 $ Weight.Loss             : Factor w/ 8 levels "1","2","3","4",..: 4 3 7 2 2 7 4 2 4 2 ...
 $ Shortness.of.Breath     : Factor w/ 8 levels "1","2","3","4",..: 2 7 8 3 4 8 2 2 3 4 ...
 $ Wheezing                : Factor w/ 8 levels "1","2","3","4",..: 2 8 2 1 1 2 2 4 2 6 ...
 $ Swallowing.Difficulty   : Factor w/ 8 levels "1","2","3","4",..: 3 6 1 4 4 1 3 2 4 5 ...
 $ Clubbing.of.Finger.Nails: Factor w/ 9 levels "1","2","3","4",..: 1 2 4 5 2 4 1 2 6 4 ...
 $ Frequent.Cold           : Factor w/ 7 levels "1","2","3","4",..: 2 1 6 6 4 6 2 3 2 2 ...
 $ Dry.Cough               : Factor w/ 7 levels "1","2","3","4",..: 3 7 7 7 2 7 3 4 4 1 ...
 $ Snoring                 : Factor w/ 7 levels "1","2","3","4",..: 4 2 2 5 3 2 4 3 1 5 ...
 $ Level                   : Factor w/ 3 levels "High","Low","Medium": 2 3 1 1 1 1 2 2 3 3 ...
```

Checking the data types of all columns

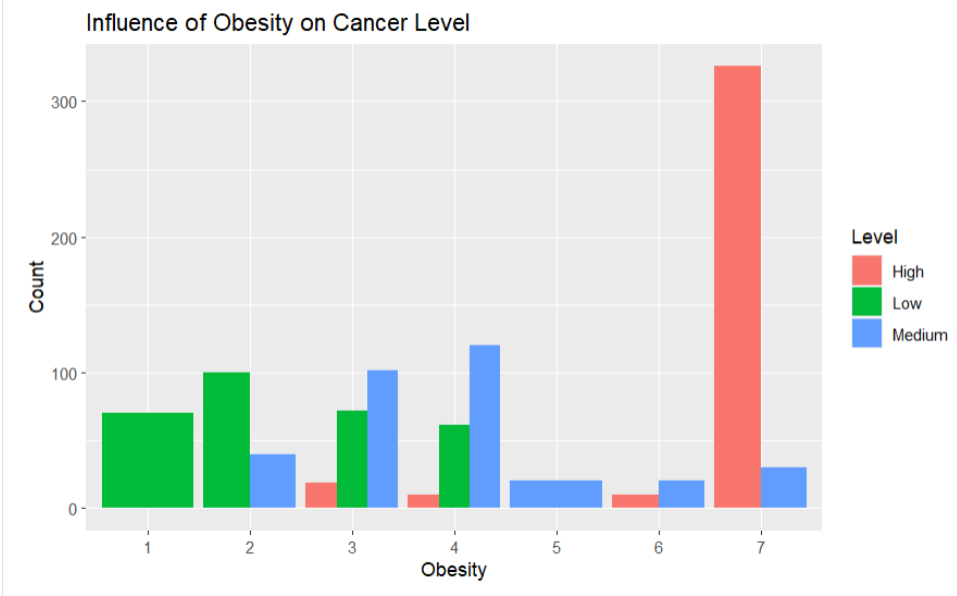
```{r}
str(df)
```

```
'data.frame':   1000 obs. of  26 variables:
 $ index                   : int  0 1 2 3 4 5 6 7 8 9 ...
 $ Patient.Id              : chr  "P1" "P10" "P100" "P1000" ...
 $ Age                     : int  33 17 35 37 46 35 52 28 35 46 ...
 $ Gender                  : int  1 1 1 1 1 1 1 2 2 2 1 ...
 $ Air.Pollution           : int  2 3 4 7 6 4 2 3 4 2 ...
 $ Alcohol.use             : int  4 1 5 7 8 5 4 1 5 3 ...
 $ Dust.Allergy            : int  5 5 6 7 7 6 5 4 6 4 ...
 $ OccuPational.Hazards    : int  4 3 5 7 7 5 4 3 5 2 ...
 $ Genetic.Risk            : int  3 4 5 6 7 5 3 2 6 4 ...
 $ chronic.Lung.Disease    : int  2 2 4 7 6 4 2 3 5 3 ...
 $ Balanced.Diet           : int  2 2 6 7 7 6 2 4 5 3 ...
 $ Obesity                 : int  4 2 7 7 7 7 4 3 5 3 ...
 $ Smoking                 : int  3 2 2 7 8 2 3 1 6 2 ...
 $ Passive.Smoker          : int  2 4 3 7 7 3 2 4 6 3 ...
 $ Chest.Pain              : int  2 2 4 7 7 4 2 3 6 4 ...
 $ Coughing.of.Blood       : int  4 3 8 8 9 8 4 1 5 4 ...
 $ Fatigue                 : int  3 1 8 4 3 8 3 3 1 1 ...
 $ Weight.Loss             : int  4 3 7 2 2 7 4 2 4 2 ...
 $ Shortness.of.Breath     : int  2 7 9 3 4 9 2 2 3 4 ...
 $ Wheezing                : int  2 8 2 1 1 2 2 4 2 6 ...
 $ Swallowing.Difficulty   : int  3 6 1 4 4 1 3 2 4 5 ...
 $ Clubbing.of.Finger.Nails: int  1 2 4 5 2 4 1 2 6 4 ...
 $ Frequent.Cold           : int  2 1 6 6 4 6 2 3 2 2 ...
 $ Dry.Cough               : int  3 7 7 7 2 7 3 4 4 1 ...
 $ Snoring                 : int  4 2 2 5 3 2 4 3 1 5 ...
 $ Level                   : chr  "Low" "Medium" "High" "High" ...
```

Checking the influence of Obesity on the level of cancer

```{r}
library(ggplot2)

ggplot(df, aes(x = Obesity, fill = Level)) +
  geom_bar(position = "dodge") +
  labs(title = "Influence of Obesity on Cancer Level",
       x = "Obesity",
       y = "Count",
       fill = "Level")
```



Building LR model between Obesity and level of cancer

```{r}
library(nnet)

# create a multinomial logistic regression model
model <- multinom(Level ~ Obesity, data = df)

# print the summary of the model
summary(model)


```

```
# weights:  24 (14 variable)
initial  value 1098.612289
iter  10 value 549.337082
iter  20 value 540.902570
iter  30 value 540.760559
iter  30 value 540.760558
final  value 540.760558
converged
Call:
multinom(formula = Level ~ Obesity, data = df)

Coefficients:
       (Intercept) Obesity2    Obesity3   Obesity4  Obesity5     Obesity6   Obesity7
Low      48.622453  8.35696 -47.2907281 -46.814371 -41.82667 -100.2646659 -58.769988
Medium    1.044304 55.01916   0.6363283   1.440646  43.90001   -0.3511818  -3.429844

Std. Errors:
       (Intercept)  Obesity2  Obesity3  Obesity4    Obesity5     Obesity6  Obesity7
Low      1.9894325 0.9071667 1.9980858 2.0044687 1.65405e-13 1.934666e-14 6.8636945
Medium   0.2148083 0.9071667 0.2889442 0.3325593 5.93432e-14 3.755526e-01 0.2627627

Residual Deviance: 1081.521
AIC: 1109.521
```

Building LR model between Alcohol.use and level of cancer

```{r}
library(nnet)

# create a multinomial logistic regression model
model <- multinom(Level ~ Alcohol.use, data = df)

# print the summary of the model
summary(model)


```

```
# weights:  27 (16 variable)
initial  value 1098.612289
iter  10 value 580.564270
iter  20 value 554.218435
iter  30 value 553.369081
final  value 553.368099
converged
Call:
multinom(formula = Level ~ Alcohol.use, data = df)

Coefficients:
       (Intercept) Alcohol.use2 Alcohol.use3 Alcohol.use4 Alcohol.use5 Alcohol.use6 Alcohol.use7 Alcohol.use8
Low       18.95791    -3.430968    -3.546743    -18.21597    -46.73226    -33.09842    -21.71158    -37.68378
Medium    18.53047    -3.182217    -5.065211    -18.53047    -19.78323    -16.58455    -34.06716    -19.05266

Std. Errors:
       (Intercept) Alcohol.use2 Alcohol.use3 Alcohol.use4 Alcohol.use5 Alcohol.use6 Alcohol.use7 Alcohol.use8
Low       58.21500     198.1427     231.2007     58.21597 2.825502e-04     348.96589     58.21578     4.411072
Medium    58.21497     198.1427     231.2008     58.21632 5.821544e+01     58.21569    184.22022    58.215136

Residual Deviance: 1106.736
AIC: 1138.736
```

Building LR model between Smoking and level of cancer

```{r}
library(nnet)

# create a multinomial logistic regression model
model <- multinom(Level ~ Smoking, data = df)

# print the summary of the model
summary(model)


```

```
# weights:  27 (16 variable)
initial  value 1098.612289
iter  10 value 690.699600
iter  20 value 670.738086
iter  30 value 670.433416
final  value 670.433028
converged
Call:
multinom(formula = Level ~ Smoking, data = df)

Coefficients:
       (Intercept)  Smoking2  Smoking3  Smoking4   Smoking5  Smoking6  Smoking7  Smoking8
Low       20.52885 -20.38289 -3.099091 -19.78440 -14.331537 -19.43023 -22.76422 -37.24543
Medium    21.20547 -21.19129 -3.423273 -35.03639   4.287428 -20.51232 -38.77456 -23.27233

Std. Errors:
       (Intercept) Smoking2 Smoking3  Smoking4     Smoking5 Smoking6  Smoking7  Smoking8
Low       61.34649 61.34666 12.60226  61.34695 7.389782e-06 61.34727  61.34686 271.57497
Medium    61.34649 61.34667 12.60227 219.86679 9.227678e-04 61.34738 286.80132  61.34724

Residual Deviance: 1340.866
AIC: 1372.866
```

Building LR model between Passive smoking and level of cancer

```{r}
library(nnet)

# create a multinomial logistic regression model
model <- multinom(Level ~ Passive.Smoker, data = df)

# print the summary of the model
summary(model)


```

```
# weights:  27 (16 variable)
initial  value 1098.612289
iter  10 value 499.421839
iter  20 value 484.283449
iter  30 value 483.888740
final  value 483.888338
converged
Call:
multinom(formula = Level ~ Passive.Smoker, data = df)

Coefficients:
       (Intercept) Passive.Smoker2 Passive.Smoker3 Passive.Smoker4 Passive.Smoker5 Passive.Smoker6 Passive.Smoker7 Passive.Smoker8
Low      28.245308       -10.61677      -29.092606       -13.15657       -7.650585       -8.428063       -45.33670       -61.96155
Medium   -6.954208        24.86632        6.394592        22.53724       26.855784       27.464601       -14.31672       -25.02818

Std. Errors:
       (Intercept) Passive.Smoker2 Passive.Smoker3 Passive.Smoker4 Passive.Smoker5 Passive.Smoker6 Passive.Smoker7 Passive.Smoker8
Low      100.34931        22.48496       100.34945        191.0484        25.46411        25.46124     305.9281232    1.906292e-06
Medium    81.17122        30.94434        81.17133        165.5426        25.52957        25.53247       0.6267271    8.002591e-06

Residual Deviance: 967.7767
AIC: 999.7767
```

Building LR model between Balanced Diet and level of cancer

```{r}
library(nnet)

# create a multinomial logistic regression model
model <- multinom(Level ~ Balanced.Diet, data = df)

# print the summary of the model
summary(model)


```

```
# weights:  24 (14 variable)
initial  value 1098.612289
iter  10 value 632.065297
iter  20 value 617.757539
iter  30 value 617.451738
iter  30 value 617.451737
final  value 617.451737
converged
Call:
multinom(formula = Level ~ Balanced.Diet, data = df)

Coefficients:
       (Intercept) Balanced.Diet2 Balanced.Diet3 Balanced.Diet4 Balanced.Diet5 Balanced.Diet6 Balanced.Diet7
Low       41.41700      -28.16453      -32.22668      -40.00604      -6.898585      -83.87628      -43.92654
Medium   -15.74085       29.24567       25.29343       15.74069      49.160878       14.96146       13.63676

Std. Errors:
       (Intercept) Balanced.Diet2 Balanced.Diet3 Balanced.Diet4 Balanced.Diet5 Balanced.Diet6 Balanced.Diet7
Low       14.98650       60.32640       17.55607       14.98904       1.082252   2.703842e-10       14.98764
Medium    12.84564       62.42862       16.01270       12.85091       1.082252   1.284643e+01       12.84665

Residual Deviance: 1234.903
AIC: 1262.903
```

Building LR model with the best AIC value

```{r}
library(nnet)
# create a multinomial logistic regression model
model <- multinom(Level ~ Obesity*Alcohol.use+Passive.Smoker, data = df)
# print the summary of the model
summary(model)
```

```
# weights:  192 (126 variable)
initial  value 1098.612289
iter  10 value 39.736520
iter  20 value 0.254536
iter  30 value 0.054253
iter  40 value 0.014627
iter  50 value 0.006378
iter  60 value 0.002288
iter  70 value 0.000343
final  value 0.000053
converged
Warning: NaNs producedCall:
multinom(formula = Level ~ Obesity * Alcohol.use + Passive.Smoker,
    data = df)

Coefficients:
       (Intercept)  Obesity2   Obesity3  Obesity4  Obesity5  Obesity6   Obesity7 Alcohol.use2 Alcohol.use3 Alcohol.use4 Alcohol.use5 Alcohol.use6
Low       275.8987 -181.1207   7.449124 -127.0576 -82.38104 -185.8409 -126.48358   -25.810043   -53.274739     21.02724   -131.95468  -61.5547608
Medium   -271.0347  218.1575 -70.203855  107.0235 126.90453  147.6832   80.19865     4.209199     8.844643    -11.44146     63.45679   -0.4983999
       Alcohol.use7 Alcohol.use8 Passive.Smoker2 Passive.Smoker3 Passive.Smoker4 Passive.Smoker5 Passive.Smoker6 Passive.Smoker7 Passive.Smoker8
Low        45.88816     -100.3075       -144.1007      0.07784804       -111.2311       -139.1129       -40.81522       -286.30764       -229.03104
Medium     50.07927      100.8446        229.7059     97.59189475        303.9139        144.4611        117.94351        -22.87267        -10.65449
       Obesity2:Alcohol.use2 Obesity3:Alcohol.use2 Obesity4:Alcohol.use2 Obesity5:Alcohol.use2 Obesity6:Alcohol.use2 Obesity7:Alcohol.use2
Low                146.4954             -112.7190              101.3771                     0                     0                     0
Medium            -181.8890              199.6543             -127.6404                     0                     0                     0
       Obesity2:Alcohol.use3 Obesity3:Alcohol.use3 Obesity4:Alcohol.use3 Obesity5:Alcohol.use3 Obesity6:Alcohol.use3 Obesity7:Alcohol.use3
Low                17.39137             -369.2219              208.1664                     0                     0                     0
Medium            -68.31761              317.3870             -238.4680                     0                     0                     0
       Obesity2:Alcohol.use4 Obesity3:Alcohol.use4 Obesity4:Alcohol.use4 Obesity5:Alcohol.use4 Obesity6:Alcohol.use4 Obesity7:Alcohol.use4
Low                        0                     0              180.2751                     0            -159.24790                     0
Medium                     0                     0              -36.9554                     0              25.51393                     0
       Obesity2:Alcohol.use5 Obesity3:Alcohol.use5 Obesity4:Alcohol.use5 Obesity5:Alcohol.use5 Obesity6:Alcohol.use5 Obesity7:Alcohol.use5
Low                        0                     0                     0           -28.11711            -26.59295             -77.24462
Medium                     0                     0                     0            31.90572            122.16922             -90.61815
       Obesity2:Alcohol.use6 Obesity3:Alcohol.use6 Obesity4:Alcohol.use6 Obesity5:Alcohol.use6 Obesity6:Alcohol.use6 Obesity7:Alcohol.use6
Low                31.37326                     0             -88.83727             -54.26392                     0              50.17317
Medium             67.37114                     0            -117.29950              94.99880                     0             -45.56884
       Obesity2:Alcohol.use7 Obesity3:Alcohol.use7 Obesity4:Alcohol.use7 Obesity5:Alcohol.use7 Obesity6:Alcohol.use7 Obesity7:Alcohol.use7
Low               138.16378                     0                     0                     0                     0             -92.27562
Medium            -86.16221                     0                     0                     0                     0             136.24148
       Obesity2:Alcohol.use8 Obesity3:Alcohol.use8 Obesity4:Alcohol.use8 Obesity5:Alcohol.use8 Obesity6:Alcohol.use8 Obesity7:Alcohol.use8
Low                        0            -247.9375             154.76651                     0                     0             -7.136512
Medium                     0             100.7293             -80.02892                     0                     0             80.144160

Std. Errors:
       (Intercept) Obesity2     Obesity3 Obesity4     Obesity5     Obesity6     Obesity7 Alcohol.use2 Alcohol.use3 Alcohol.use4 Alcohol.use5
Low       9371.841 2400.495 1.064348e-09 11697.18 2.132204e-11          NaN 1.063737e-02 1.307496e-16     2400.495 1.623419e-67 1.190800e-11
Medium   11350.859 2891.300 4.572753e-22 11697.18 3.043912e-16 1.153708e-42 1.504163e+04 5.187005e-17     2891.300 1.153708e-42 3.043912e-16
       Alcohol.use6 Alcohol.use7 Alcohol.use8 Passive.Smoker2 Passive.Smoker3 Passive.Smoker4 Passive.Smoker5 Passive.Smoker6 Passive.Smoker7
Low      0.01064082 1.618637e-40    11697.179        11697.18     3.671773e-06      0.01063737    1.507266e-19        2400.495     3.586013e-65
Medium   0.01064082 6.738173e+03     8634.235        11697.18     3.671777e-06      0.01063737    2.925298e-27        2891.300     5.187240e+04
       Passive.Smoker8 Obesity2:Alcohol.use2 Obesity3:Alcohol.use2 Obesity4:Alcohol.use2 Obesity5:Alcohol.use2 Obesity6:Alcohol.use2
Low       1.618637e-40          1.307497e-16          6.797467e-25          9.075894e-28                     0                     0
Medium    3.757392e+04          5.187005e-17          8.969291e-25          9.075893e-28                     0                     0
       Obesity7:Alcohol.use2 Obesity2:Alcohol.use3 Obesity3:Alcohol.use3 Obesity4:Alcohol.use3 Obesity5:Alcohol.use3 Obesity6:Alcohol.use3
Low                        0             2400.495          1.577276e-82          2.215777e-55                     0                     0
Medium                     0             2891.300          4.567722e-22         2.122176e-126                     0                     0
       Obesity7:Alcohol.use3 Obesity2:Alcohol.use4 Obesity3:Alcohol.use4 Obesity4:Alcohol.use4 Obesity5:Alcohol.use4 Obesity6:Alcohol.use4
Low                        0                     0                     0          1.623419e-67                     0         7.983574e-132
Medium                     0                     0                     0          1.623419e-67                     0          1.153708e-42
       Obesity7:Alcohol.use4 Obesity2:Alcohol.use5 Obesity3:Alcohol.use5 Obesity4:Alcohol.use5 Obesity5:Alcohol.use5 Obesity6:Alcohol.use5
Low                        0                     0                     0                     0          1.654608e-19          5.292023e-112
Medium                     0                     0                     0                     0          3.043912e-16          1.812284e-64
       Obesity7:Alcohol.use5 Obesity2:Alcohol.use6 Obesity3:Alcohol.use6 Obesity4:Alcohol.use6 Obesity5:Alcohol.use6 Obesity6:Alcohol.use6
Low            1.190800e-11          3.671777e-06                     0          2.358432e-86          5.374667e-54                     0
Medium         4.000867e-37          3.671777e-06                     0        1.371735e-112          2.925298e-27                     0
       Obesity7:Alcohol.use6 Obesity2:Alcohol.use7 Obesity3:Alcohol.use7 Obesity4:Alcohol.use7 Obesity5:Alcohol.use7 Obesity6:Alcohol.use7
Low              0.01063737         2.430725e-103                     0                     0                     0                     0
Medium           0.01063737         2.430764e-103                     0                     0                     0                     0
       Obesity7:Alcohol.use7 Obesity2:Alcohol.use8 Obesity3:Alcohol.use8 Obesity4:Alcohol.use8 Obesity5:Alcohol.use8 Obesity6:Alcohol.use8
Low            1.618637e-40                     0        1.843189e-138              11697.18                     0                     0
Medium         6.738173e+03                     0         1.516616e-55              11697.18                     0                     0
       Obesity7:Alcohol.use8
Low            7.329741e-67
Medium         9.362783e+03

Residual Deviance: 0.0001064186
AIC: 132.0001
```