Visualization Project on LUNG CANCER

REEWA MALIK-MDS202134

30/11/2021

#The data used for the project is a survey data, which has variables #"GENDER","AGE","SMOKING","YELLOW_FINGE
#"PEER_PRESSURE","CHRONIC.DISEASE","FATIGUE","ALLERGY" #"WHEEZING","ALCOHOL.CONSUMING","
#"SHORTNESS.OF.BREATH","SWALLOWING.DIFFICULTY","CHEST.PAIN","LUNG_CANCER#".The
idea is to study the relationship between LUNG_CANCER and other #variables, which are as follows-

#1.What is the percentage of people who didn't smoke(and other varaibles #too) but still has lung Cancer and what is the percentage of people who #smoke but still didnt get lung cancer?

#2.Is it common for a certain age group to have lung cancer? #3.What are the major factors that causes lung cancer and what are not? #4.Can the results obtained from the given data be generalised. #5.Do people in the dataset lead a healthy lifestyle?

```
library('ggplot2')
lung_cancer_data=read.csv(file.choose())
```

```
names(lung_cancer_data)
```

```
##  [1] "GENDER"                "AGE"                   "SMOKING"
##  [4] "YELLOW_FINGERS"        "ANXIETY"               "PEER_PRESSURE"
##  [7] "CHRONIC.DISEASE"       "FATIGUE"               "ALLERGY"
## [10] "WHEEZING"              "ALCOHOL.CONSUMING"     "COUGHING"
## [13] "SHORTNESS.OF.BREATH"   "SWALLOWING.DIFFICULTY" "CHEST.PAIN"
## [16] "LUNG_CANCER"
```

#For the Project, I am considering variables "AGE", #"SMOKING","ANXIETY", "CHRONIC.DISEASE" ,"ALCOHOL.CONSUMING" #,"COUGHING","SHORTNESS.OF.BREATH" ,"CHEST.PAIN","LUNG_CANCER".

Out of these variables, the variable AGE is a continuous random #variable and rest are all categorical and to be specific all are #Nominal, since they do not have any intrinsic order,its just Yes or No. #For some of these variables, 1 represents NO and 2 represents YES.

```
tab <- matrix(rep('Nominal', times=9), ncol=9, byrow=TRUE)
colnames(tab) <- c('AGE', 'SMOKING', 'ANXIETY', 'CHRONIC.DISEASE','ALCOHOL.CONSUMING','COUGHING','SHORT
    'CHEST.PAIN','LUNG_CANCER')
```

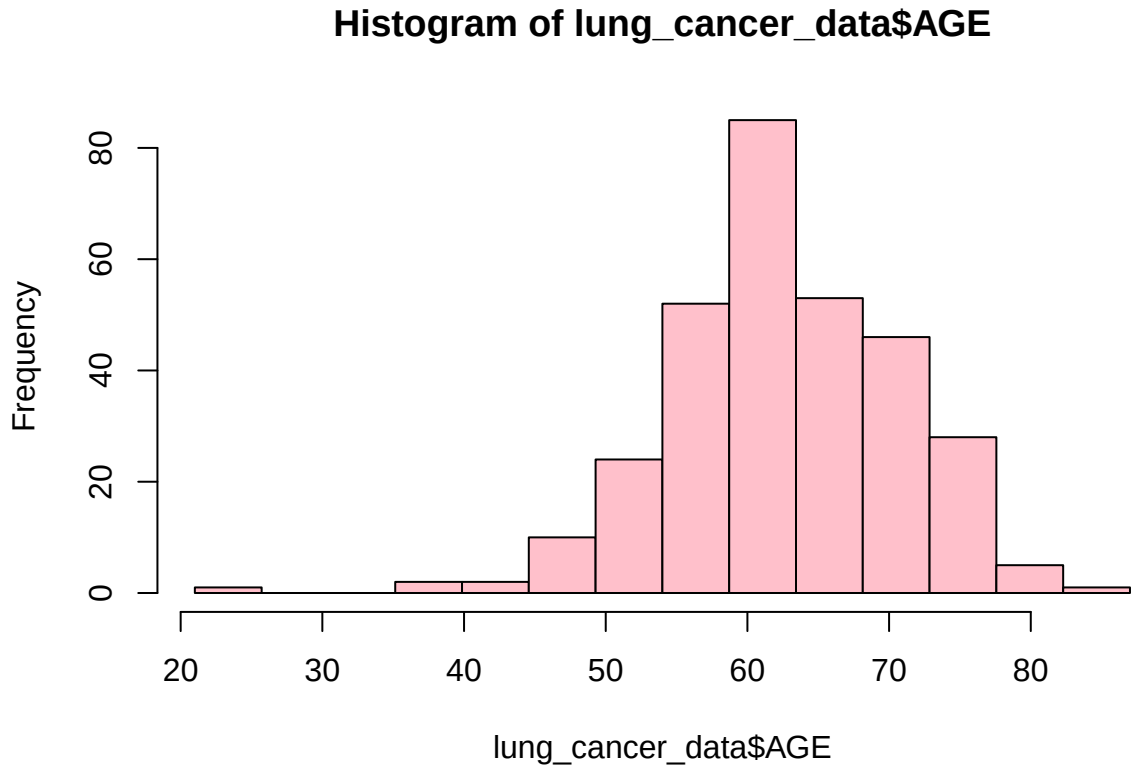
```
rownames(tab) <- c('TYpe of Variable')
tab[1,1]='Continuous'
tab
```

```
##                   AGE          SMOKING   ANXIETY   CHRONIC.DISEASE
## TYpe of Variable "Continuous" "Nominal" "Nominal" "Nominal"
##                   ALCOHOL.CONSUMING COUGHING  SHORTNESS.OF.BREATH CHEST.PAIN
## TYpe of Variable "Nominal"         "Nominal" "Nominal"           "Nominal"
##                   LUNG_CANCER
## TYpe of Variable "Nominal"
```

```
#The given table represents which considered variable is Continuous and #which is Nominal(Categorical)
```

#UNIVARIATE ANALYSIS OF DIFFERENT VARIABLES

```
#Making the histogram to check,data set has which age group in most #numbers.
hist(lung_cancer_data$AGE, col = 'pink', breaks = seq(min(lung_cancer_data$AGE), max(lung_cancer_data$A
```




```
table(lung_cancer_data$AGE)
```

```
##
## 21 38 39 44 46 47 48 49 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68
##  1  1  1  2  1  4  2  3  8  4  4  8 11 19  9 13 15 17 16 18 19 20  7  4 13  9
## 69 70 71 72 73 74 75 76 77 78 79 81 87
## 11 15 10 10  4  6  5  4  9  2  1  2  1
```

```r
summary(lung_cancer_data$AGE)
```

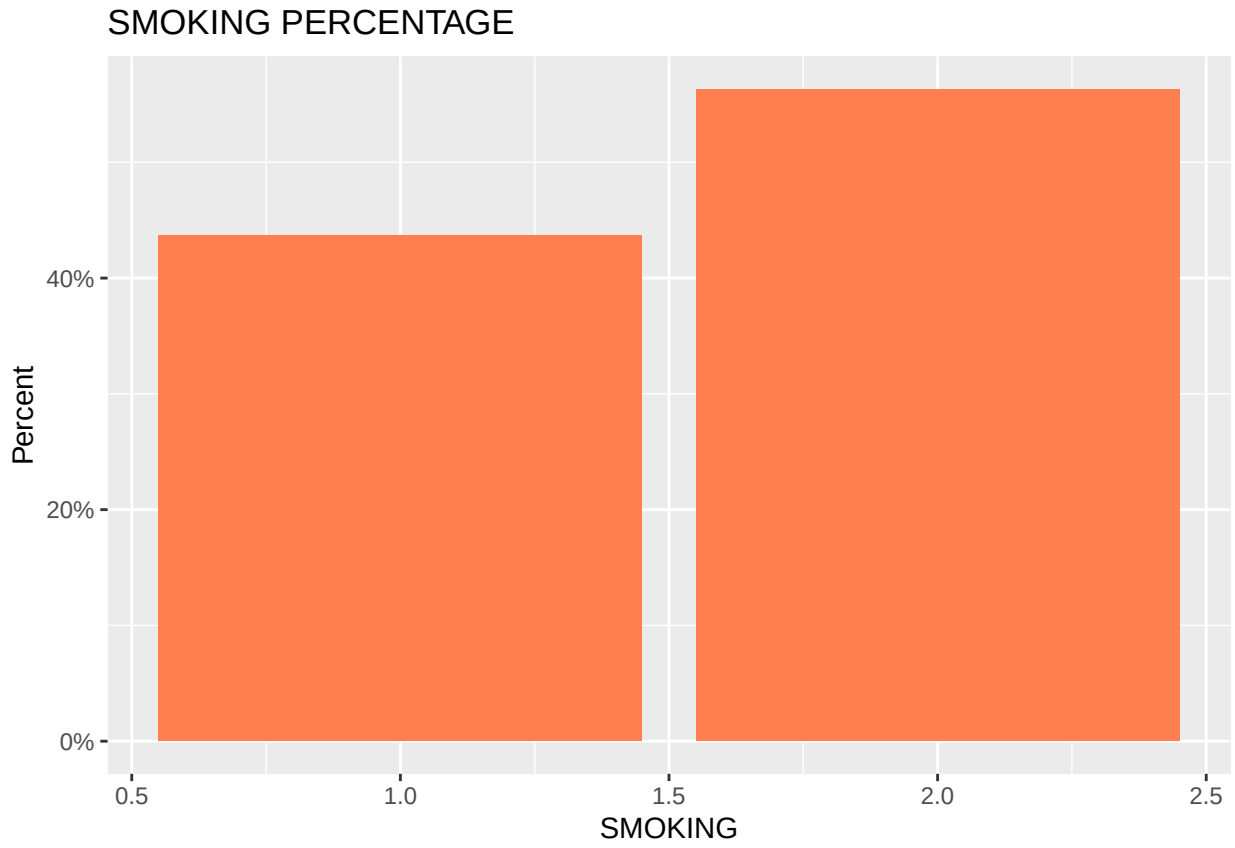
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   21.00   57.00   62.00   62.67   69.00   87.00
```

```r
#Clearly the data has majority of the people IN age group 54-72.So #whatever results we get from this d
```

```r
table(lung_cancer_data$SMOKING)#No of people who smoke and who don't
```

```
##
##   1   2
## 135 174
```

```r
ggplot(lung_cancer_data,
       aes(x = SMOKING,
           y = ..count.. / sum(..count..))) +
  geom_bar() +
  labs(x = "SMOKING",
       y = "Percent",
       title  = "SMOKING PERCENTAGE") +
  scale_y_continuous(labels = scales::percent)+geom_bar(fill = "coral")
```

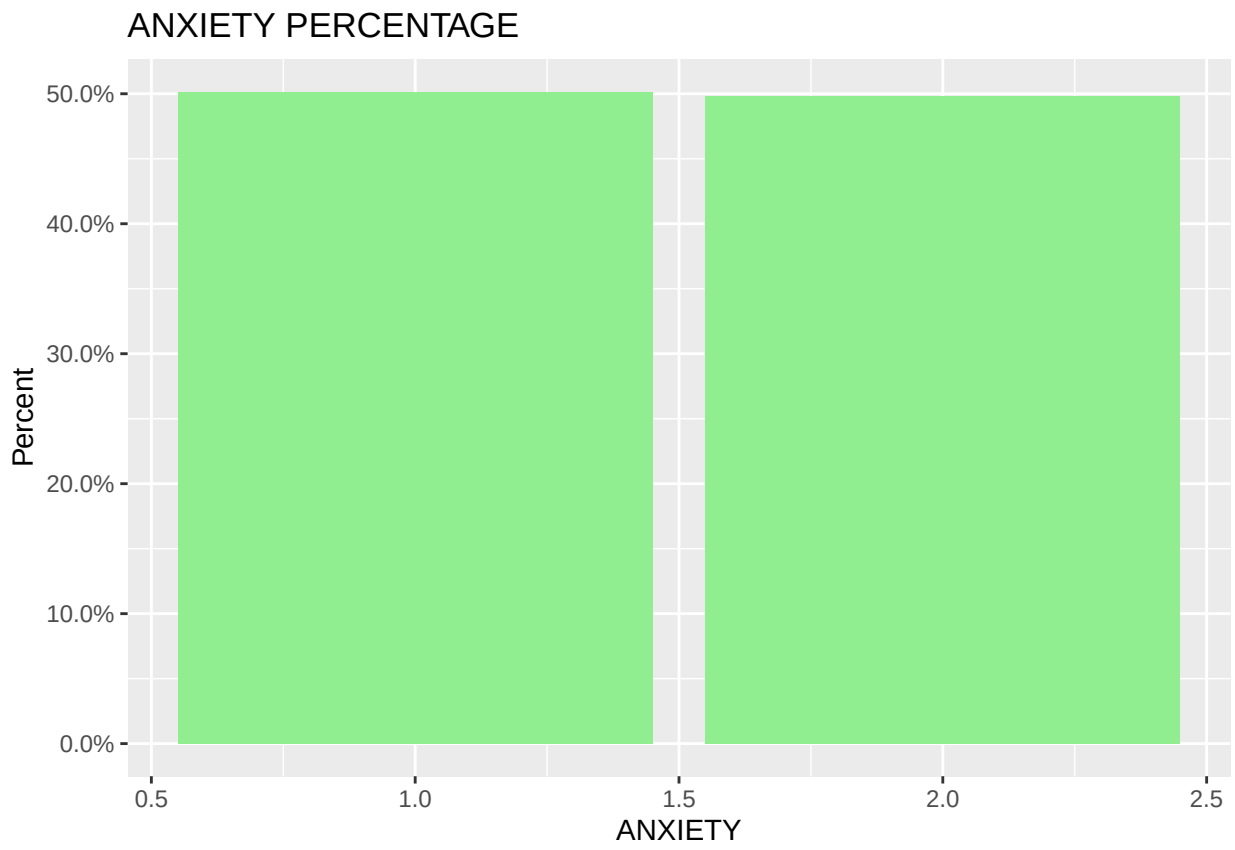


```
#Approx 57 percent people smoke and approx 43 percent people don't.
```

```
table(lung_cancer_data$ANXIETY)#No of people who have anxiety and who don't
```

```
##
##   1   2
## 155 154
```

```
ggplot(lung_cancer_data,
       aes(x = ANXIETY,
           y = ..count.. / sum(..count..))) +
  geom_bar() +
  labs(x = "ANXIETY",
       y = "Percent",
       title  = "ANXIETY PERCENTAGE") +
  scale_y_continuous(labels = scales::percent)+geom_bar(fill = "lightgreen")
```

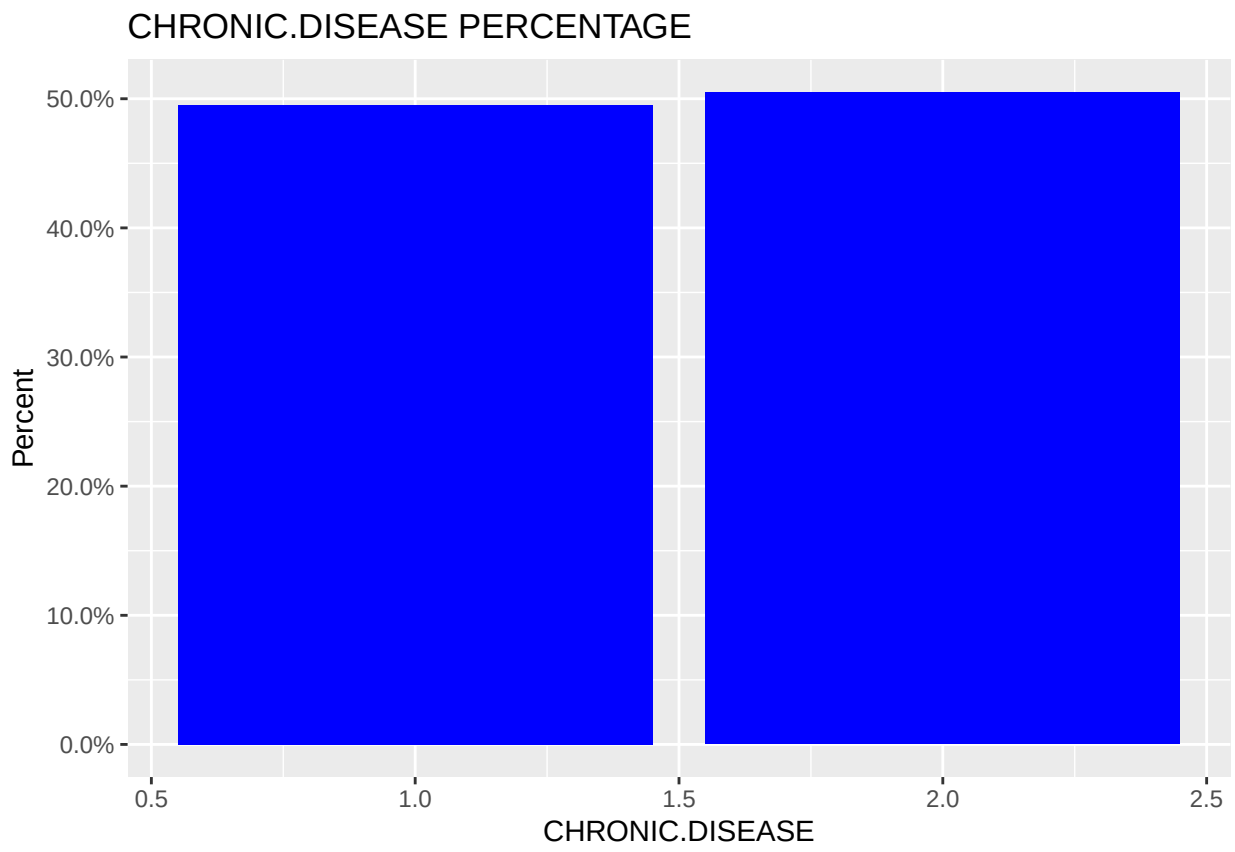



```
#Approx 50 percent people have anxiety and approx 50 percent people #don't.
```

```
table(lung_cancer_data$CHRONIC.DISEASE)#No of people who have chronic #disease and who don't
```

```
##
##   1   2
## 153 156
```

```
ggplot(lung_cancer_data,
       aes(x = CHRONIC.DISEASE,
           y = ..count.. / sum(..count..))) +
  geom_bar() +
  labs(x = "CHRONIC.DISEASE",
       y = "Percent",
       title  = "CHRONIC.DISEASE PERCENTAGE") +
  scale_y_continuous(labels = scales::percent)+geom_bar(fill = "blue")
```

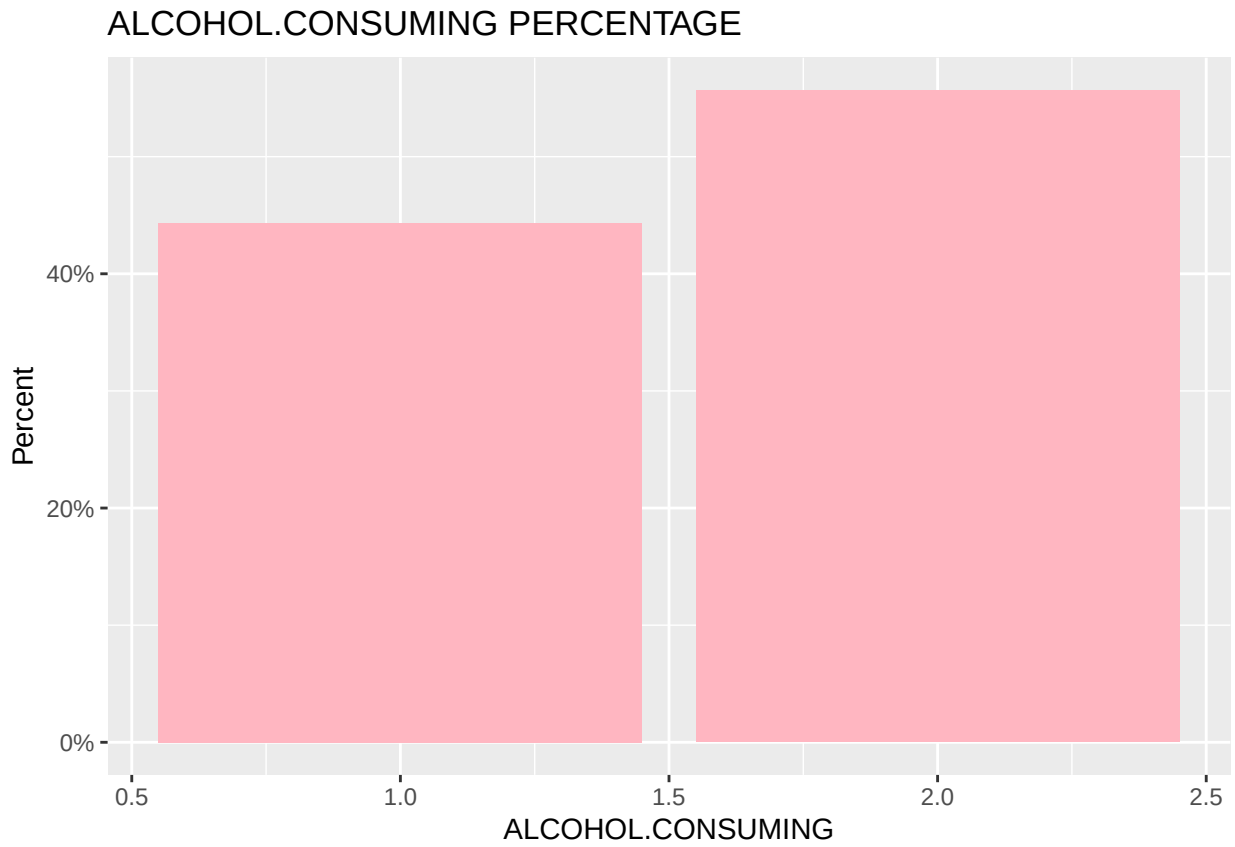



```
#Approx 51 percent people have chronic disease and approx 49 percent #people don't.
```

```
table(lung_cancer_data$ALCOHOL.CONSUMING)#No of people who consume #alcohol and who don't
```

```
##
##   1   2
## 137 172
```

```
ggplot(lung_cancer_data,
       aes(x = ALCOHOL.CONSUMING,
           y = ..count.. / sum(..count..))) +
  geom_bar() +
  labs(x = "ALCOHOL.CONSUMING",
       y = "Percent",
       title  = "ALCOHOL.CONSUMING PERCENTAGE") +
  scale_y_continuous(labels = scales::percent)+geom_bar(fill = "lightpink")
```

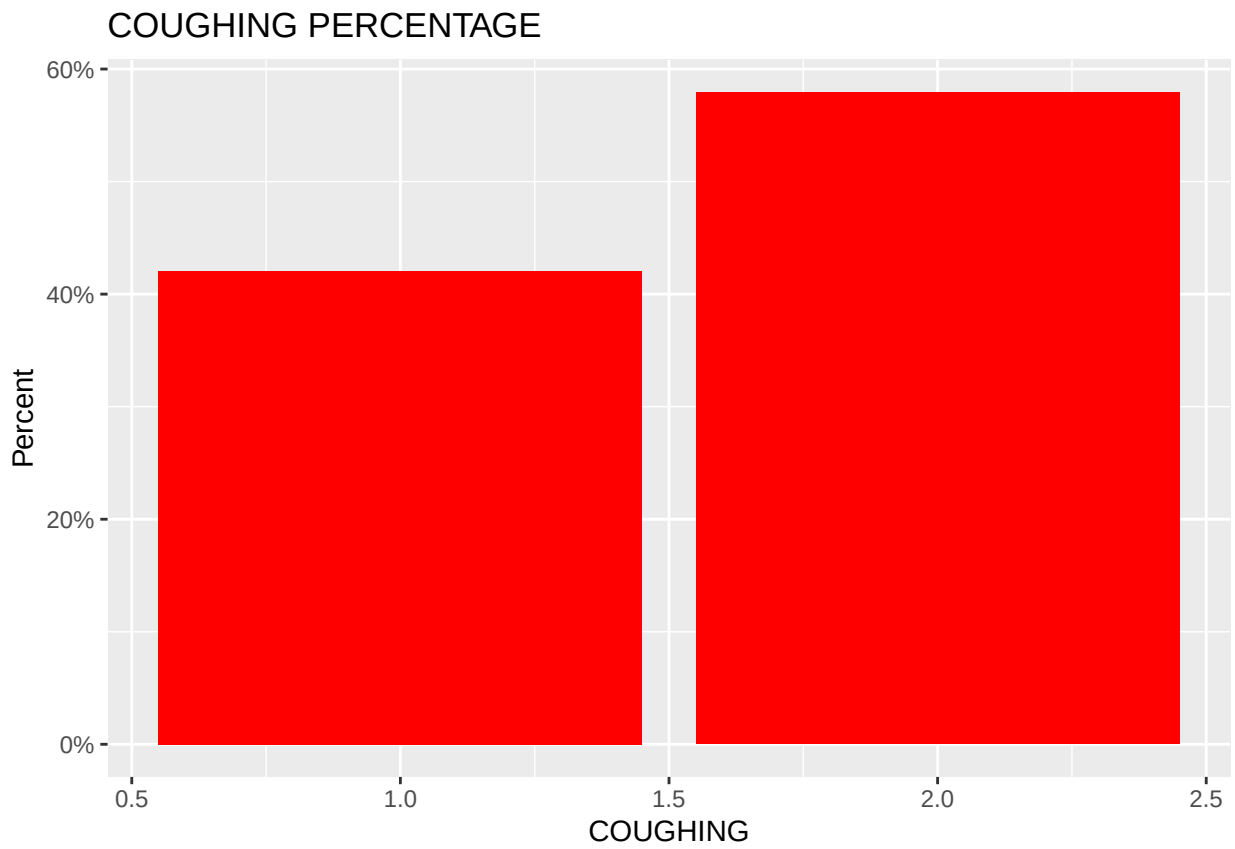



```
#Approx 57 percent consume alcohol and 43 percent don't consume alcohol.
```

```
table(lung_cancer_data$COUGHING)#No of people who have cough and who #don't
```

```
##
##   1   2
## 130 179
```

```
ggplot(lung_cancer_data,
       aes(x = COUGHING,
           y = ..count.. / sum(..count..))) +
  geom_bar() +
  labs(x = "COUGHING",
       y = "Percent",
       title  = "COUGHING PERCENTAGE") +
  scale_y_continuous(labels = scales::percent)+geom_bar(fill = "red")
```

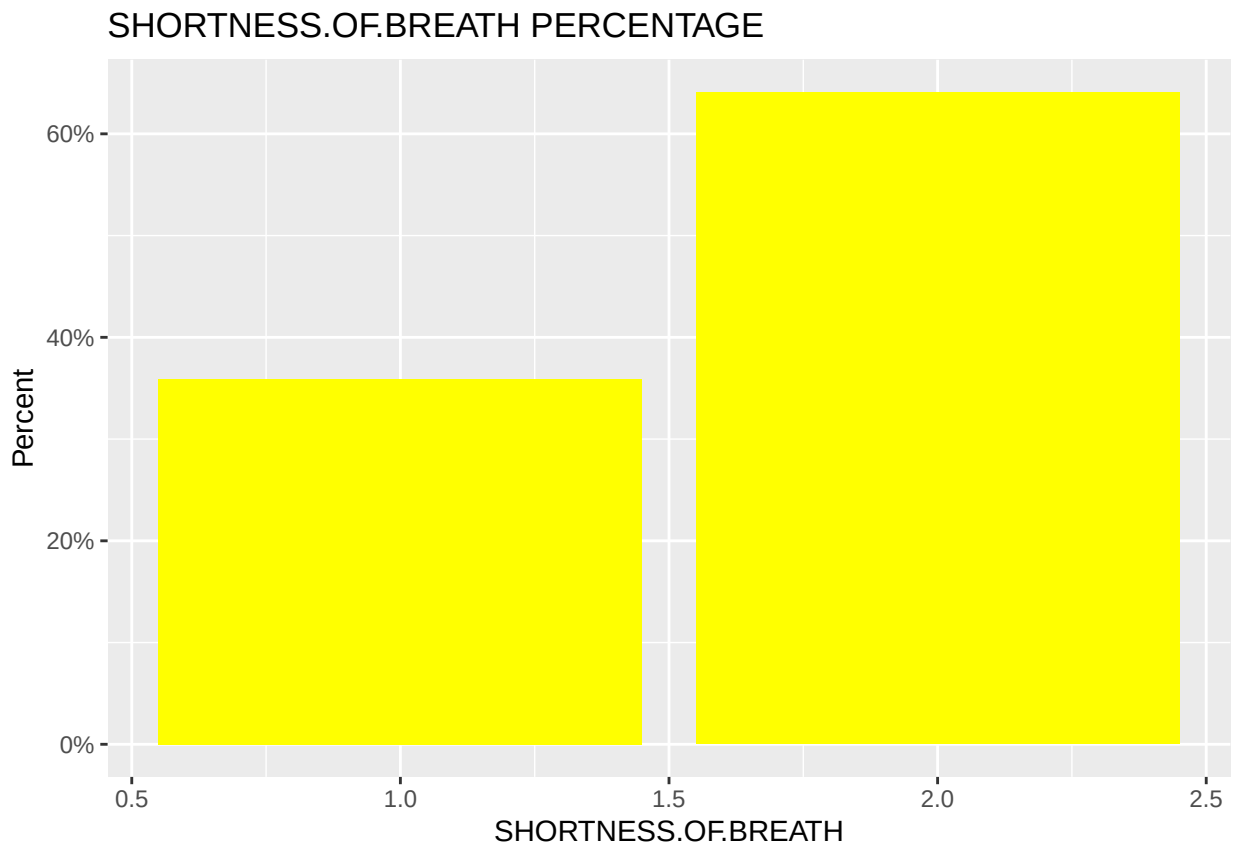



```
#Approx 58 percent have cough and 42 percent don't have cough.
```

```
table(lung_cancer_data$SHORTNESS.OF.BREATH)#No of people who have #shortness of breath and who don't
```

```
##
##   1   2
## 111 198
```

```
ggplot(lung_cancer_data,
       aes(x = SHORTNESS.OF.BREATH,
           y = ..count.. / sum(..count..))) +
  geom_bar() +
  labs(x = "SHORTNESS.OF.BREATH",
       y = "Percent",
       title  = "SHORTNESS.OF.BREATH PERCENTAGE") +
  scale_y_continuous(labels = scales::percent)+geom_bar(fill = "yellow")
```

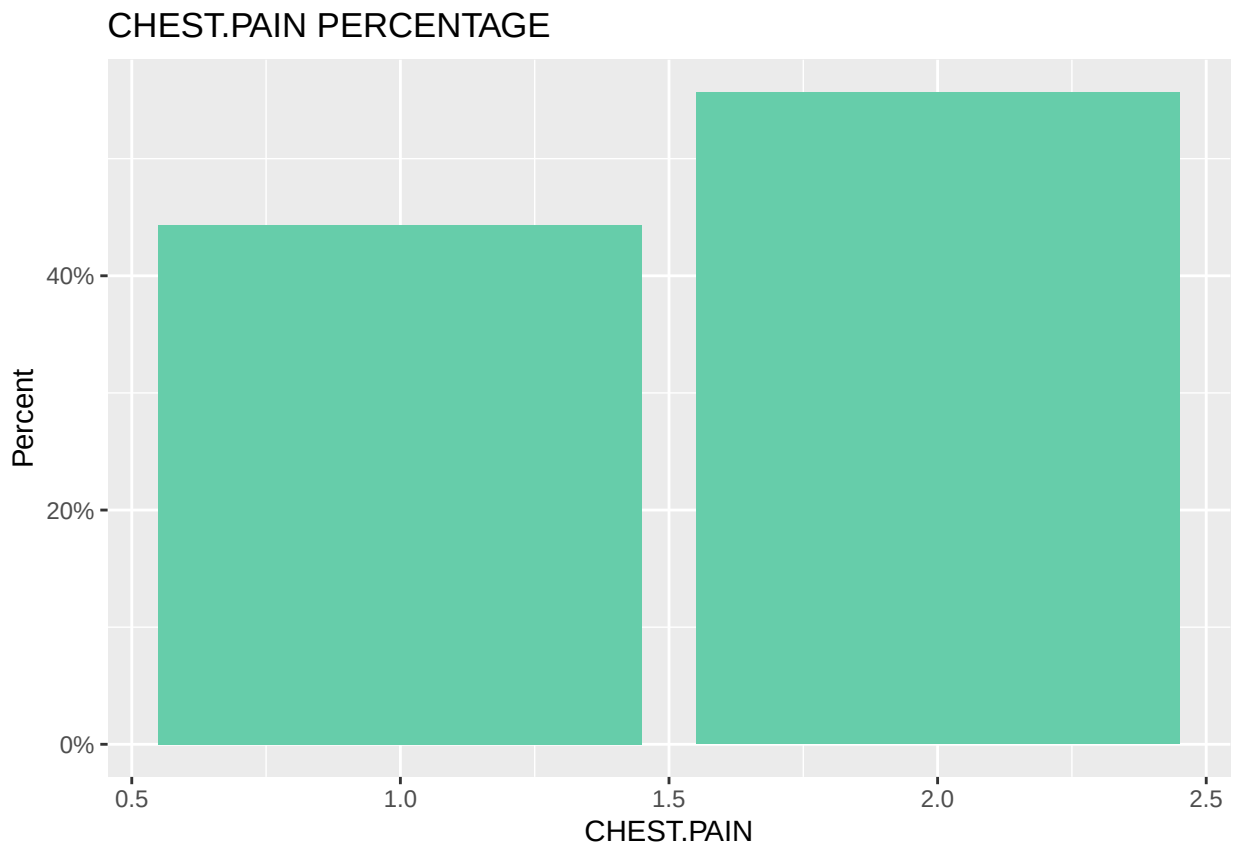



```
#Approx 64 percent have SHORTNESS.OF.BREATH and 36 percent don't have #SHORTNESS.OF.BREATH.
```

```
table(lung_cancer_data$CHEST.PAIN)#No of people who have chest pain and #who don't
```

```
##
##   1   2
## 137 172
```

```
ggplot(lung_cancer_data,
       aes(x = CHEST.PAIN,
           y = ..count.. / sum(..count..))) +
  geom_bar() +
  labs(x = "CHEST.PAIN",
       y = "Percent",
       title  = "CHEST.PAIN PERCENTAGE") +
  scale_y_continuous(labels = scales::percent)+geom_bar(fill = "#66CDAA")
```

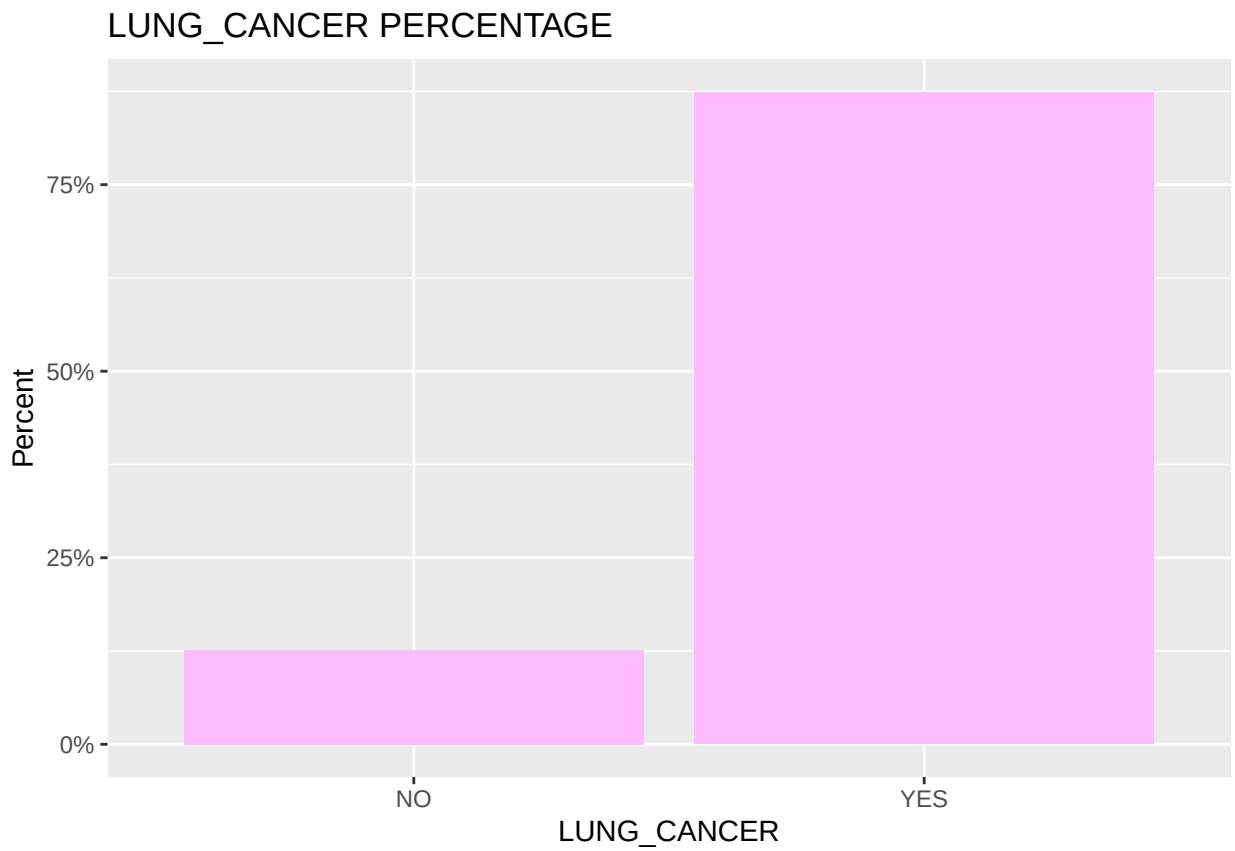



```
#Approx 56 percent have CHEST.PAIN and 44 percent don't have CHEST.PAIN.
```

```
table(lung_cancer_data$LUNG_CANCER)#No of people who have lung cancer #and who don't
```

```
##
##  NO YES
##  39 270
```

```
ggplot(lung_cancer_data,
       aes(x = LUNG_CANCER,
           y = ..count.. / sum(..count..))) +
  geom_bar() +
  labs(x = "LUNG_CANCER",
       y = "Percent",
       title  = "LUNG_CANCER PERCENTAGE") +
  scale_y_continuous(labels = scales::percent)+geom_bar(fill = "#FFBBFF")
```




```
#Approx 87.5 percent have LUNG_CANCER and 12.5 percent don't have #LUNG_CANCER.
```

#BIVARIATE ANALYSIS

```
my_tab<-table(lung_cancer_data$SMOKING,lung_cancer_data$LUNG_CANCER)
my_tab#No of people who have Lung Cancer and who Smoke
```

```
##
##      NO YES
##   1  20 115
##   2  19 155
```

```
#1-NO,2-YES for SMOKING
```

```
my_tab<-table(lung_cancer_data$ANXIETY,lung_cancer_data$LUNG_CANCER)
my_tab#No of people who have Lung Cancer and who have Anxiety
```

```
##
##      NO YES
##   1  27 128
##   2  12 142
```

```
#1-NO,2-YES for ANXIETY
```

```
my_tab<-table(lung_cancer_data$CHEST.PAIN,lung_cancer_data$LUNG_CANCER)
my_tab#No of people who have Lung Cancer and who have chest pain
```

```
##
##      NO YES
##   1  27 110
##   2  12 160
```

```
#1-NO,2-YES for CHEST.PAIN
```

```
my_tab<-table(lung_cancer_data$ALCOHOL.CONSUMING,lung_cancer_data$LUNG_CANCER)
my_tab#No of people who have Lung Cancer and who consume alcohol
```

```
##
##      NO YES
##   1  32 105
##   2   7 165
```

```
#1-NO,2-YES for ALCOHOL.CONSUMING
```

```
my_tab<-table(lung_cancer_data$SHORTNESS.OF.BREATH,lung_cancer_data$LUNG_CANCER)
my_tab#No of people who have Lung Cancer and who have SHORTNESS.OF.BREATH
```

```
##
##      NO YES
##   1  17  94
##   2  22 176
```

```
#1-NO,2-YES for SHORTNESS.OF.BREATH
```

```
my_tab<-table(lung_cancer_data$COUGHING,lung_cancer_data$LUNG_CANCER)
my_tab#No of people who have Lung Cancer and who have COUGHING
```

```
##
##      NO YES
##   1  29 101
##   2  10 169
```

```
#1-NO,2-YES for COUGHING
```

```
my_tab<-table(lung_cancer_data$CHRONIC.DISEASE,lung_cancer_data$LUNG_CANCER)
my_tab#No of people who have Lung Cancer and who have CHRONIC.DISEASE
```

```
##
##      NO YES
##   1  25 128
##   2  14 142
```

```
#1-NO,2-YES for CHRONIC.DISEASE
```

```
library(ggplot2)
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
#Lets find out what percentage of people have lung cancer and also a  #YES for considered Categorical v
a=nrow(filter(lung_cancer_data, LUNG_CANCER == 'YES',SMOKING==2))
b=nrow(filter(lung_cancer_data, LUNG_CANCER == 'YES',ANXIETY==2))
c=nrow(filter(lung_cancer_data, LUNG_CANCER == 'YES',CHEST.PAIN==2))
d=nrow(filter(lung_cancer_data, LUNG_CANCER == 'YES',COUGHING==2))
e=nrow(filter(lung_cancer_data, LUNG_CANCER == 'YES',ALCOHOL.CONSUMING==2))
f=nrow(filter(lung_cancer_data, LUNG_CANCER == 'YES', CHRONIC.DISEASE==2))
g=nrow(filter(lung_cancer_data, LUNG_CANCER == 'YES',SHORTNESS.OF.BREATH==2))
n=nrow(lung_cancer_data)
smoking_positive_cancer_positive_percentage=(a/n)*100
smoking_positive_cancer_positive_percentage#Approx 50% people who smoke #have Lung Cancer.
```

```
## [1] 50.16181
```

```
anxiety_positive_cancer_positive_percentage=(b/n)*100
anxiety_positive_cancer_positive_percentage#Approx 46% people who suffer #from Anxiety have Lung Cancer
```

```
## [1] 45.95469
```

```
chest.pain_positive_cancer_positive_percentage=(c/n)*100
chest.pain_positive_cancer_positive_percentage#Approx 52% people who #suffer from Chest Pain have Lung
```

```
## [1] 51.77994
```

```
coughing_positive_cancer_positive_percentage=(d/n)*100
coughing_positive_cancer_positive_percentage#Approx 55% people who #suffer from Coughing problem have L
```

```
## [1] 54.69256
```

```
ALCOHOL.CONSUMING_cancer_positive_positive_percentage=(e/n)*100
ALCOHOL.CONSUMING_cancer_positive_positive_percentage#Approx 53% people #who consume Alcohol  have Lung
```

```
## [1] 53.39806
```

```
CHRONIC.DISEASE_cancer_positive_positive_percentage=(f/n)*100
CHRONIC.DISEASE_cancer_positive_positive_percentage#Approx 46% people #who suffer from some Chronic dis
```

```
## [1] 45.95469
```

```
SHORTNESS.OF.BREATH_cancer_positive_positive_percentage=(g/n)*100
SHORTNESS.OF.BREATH_cancer_positive_positive_percentage#Approx 57% #people who suffer from Shortness of
```

```
## [1] 56.95793
```

```
#Lets find out what percentage of people don't have lung cancer and also #a  YES for one of the conside
a=nrow(filter(lung_cancer_data, LUNG_CANCER == 'NO',SMOKING==2))
b=nrow(filter(lung_cancer_data, LUNG_CANCER == 'NO',ANXIETY==2))
c=nrow(filter(lung_cancer_data, LUNG_CANCER == 'NO',CHEST.PAIN==2))
d=nrow(filter(lung_cancer_data, LUNG_CANCER == 'NO',COUGHING==2))
e=nrow(filter(lung_cancer_data, LUNG_CANCER == 'NO',ALCOHOL.CONSUMING==2))
f=nrow(filter(lung_cancer_data, LUNG_CANCER == 'NO', CHRONIC.DISEASE==2))
g=nrow(filter(lung_cancer_data, LUNG_CANCER == 'NO',SHORTNESS.OF.BREATH==2))
n=nrow(lung_cancer_data)
smoking_positive_cancer_negative_percentage=(a/n)*100
smoking_positive_cancer_negative_percentage#Approx 6% people  smoke but #do not have Lung Cancer.
```

```
## [1] 6.148867
```

```
anxiety_positive_cancer_negative_percentage=(b/n)*100
anxiety_positive_cancer_negative_percentage#Approx 4% people have #anxiety issues but do not have Lung
```

```
## [1] 3.883495
```

```
chest.pain_positive_cancer_negative_percentage=(c/n)*100
chest.pain_positive_cancer_negative_percentage#Approx 4% people have #Chest pain but do not have Lung C
```

```
## [1] 3.883495
```

```
coughing_positive_cancer_negative_percentage=(d/n)*100
coughing_positive_cancer_negative_percentage#Approx 3% people have #coughing problem but do not have Lu
```

```
## [1] 3.236246
```

```
ALCOHOL.CONSUMING_positive_cancer_negative_percentage=(e/n)*100
ALCOHOL.CONSUMING_positive_cancer_negative_percentage#Approx 2% people #smoke but do not have Lung Canc
```

```
## [1] 2.265372
```

```
CHRONIC.DISEASE_positive_cancer_negative_percentage=(f/n)*100
CHRONIC.DISEASE_positive_cancer_negative_percentage#Approx 5% people #have Chronic disease but do not h
```

```
## [1] 4.530744
```

```
SHORTNESS.OF.BREATH_positive_cancer_negative_percentage=(g/n)*100
SHORTNESS.OF.BREATH_positive_cancer_negative_percentage#Approx 7% people #have shortness of breath prob
```

```
## [1] 7.119741
```

```
#As indicated by all the precentages, it is quite a rare chance that any #person who smoke,has anxiety
```

```
#Lets find out what percentage of people don't have lung cancer and also #a NO for one of the considere
a=nrow(filter(lung_cancer_data, LUNG_CANCER == 'NO',SMOKING==1))
b=nrow(filter(lung_cancer_data, LUNG_CANCER == 'NO',ANXIETY==1))
c=nrow(filter(lung_cancer_data, LUNG_CANCER == 'NO',CHEST.PAIN==1))
d=nrow(filter(lung_cancer_data, LUNG_CANCER == 'NO',COUGHING==1))
e=nrow(filter(lung_cancer_data, LUNG_CANCER == 'NO',ALCOHOL.CONSUMING==1))
f=nrow(filter(lung_cancer_data, LUNG_CANCER == 'NO', CHRONIC.DISEASE==1))
g=nrow(filter(lung_cancer_data, LUNG_CANCER == 'NO',SHORTNESS.OF.BREATH==1))
n=nrow(lung_cancer_data)
smoking_negative_cancer_negative_percentage=(a/n)*100
smoking_negative_cancer_negative_percentage#There are Approx 6% people  #who don't smoke and do not hav
```

```
## [1] 6.472492
```

```
anxiety_negative_cancer_negative_percentage=(b/n)*100
anxiety_negative_cancer_negative_percentage#There are Approx 9% people  #who don't have anxiety and do
```

```
## [1] 8.737864
```

```
chest.pain_negative_cancer_negative_percentage=(c/n)*100
chest.pain_negative_cancer_negative_percentage#There are Approx 9% #people  who don't have chest pain a
```

```
## [1] 8.737864
```

```
coughing_negative_cancer_negative_percentage=(d/n)*100
coughing_positive_cancer_negative_percentage#There are Approx 3% people #who don't have coughing proble
```

```
## [1] 3.236246
```

```
ALCOHOL.CONSUMING_negative_cancer_negative_percentage=(e/n)*100
ALCOHOL.CONSUMING_negative_cancer_negative_percentage#There are Approx #10% people  who don't consume a
```

```
## [1] 10.35599
```

```
CHRONIC.DISEASE_negative_cancer_negative_percentage=(f/n)*100
CHRONIC.DISEASE_negative_cancer_negative_percentage#There are Approx 8% #people  who don't have chronic
```

```
## [1] 8.090615
```

```
SHORTNESS.OF.BREATH_negative_cancer_negative_percentage=(g/n)*100
SHORTNESS.OF.BREATH_negative_cancer_negative_percentage#There are Approx #5% people  who don't have sho
```

```
## [1] 5.501618
```

```
#As indicated by all the precentages, it is quite a rare chance that any #person who don't smoke,do not
```

```
#Lets find out what percentage of people have lung cancer and also a NO #for one of the Categorical var
a=nrow(filter(lung_cancer_data, LUNG_CANCER == 'YES',SMOKING==1))
b=nrow(filter(lung_cancer_data, LUNG_CANCER == 'YES',ANXIETY==1))
c=nrow(filter(lung_cancer_data, LUNG_CANCER == 'YES',CHEST.PAIN==1))
d=nrow(filter(lung_cancer_data, LUNG_CANCER == 'YES',COUGHING==1))
e=nrow(filter(lung_cancer_data, LUNG_CANCER == 'YES',ALCOHOL.CONSUMING==1))
f=nrow(filter(lung_cancer_data, LUNG_CANCER == 'YES', CHRONIC.DISEASE==1))
g=nrow(filter(lung_cancer_data, LUNG_CANCER == 'YES',SHORTNESS.OF.BREATH==1))
n=nrow(lung_cancer_data)
smoking_negative_cancer_positive_percentage=(a/n)*100
smoking_negative_cancer_positive_percentage#There are Approx 37% people #who don't smoke and have Lung
```

```
## [1] 37.21683
```

```
anxiety_negative_cancer_positive_percentage=(b/n)*100
anxiety_negative_cancer_positive_percentage#There are Approx 41% people #who don't have anxiety and hav
```

```
## [1] 41.42395
```

```
chest.pain_negative_cancer_positive_percentage=(c/n)*100
chest.pain_negative_cancer_positive_percentage#There are Approx 36% #people  who don't have chest pain
```

```
## [1] 35.59871
```

```
coughing_negative_cancer_positive_percentage=(d/n)*100
coughing_negative_cancer_positive_percentage#There are Approx 33%
```

```
## [1] 32.68608
```

```
#people who don't have coughing problem and have Lung Cancer
ALCOHOL.CONSUMING_negative_cancer_positive_percentage=(e/n)*100
ALCOHOL.CONSUMING_negative_cancer_positive_percentage#There are Approx #34% people  who don't consume a
```

```
## [1] 33.98058
```

```
CHRONIC.DISEASE_negative_cancer_positive_percentage=(f/n)*100
CHRONIC.DISEASE_negative_cancer_positive_percentage#There are Approx 41% #people  who don't have chroni
```

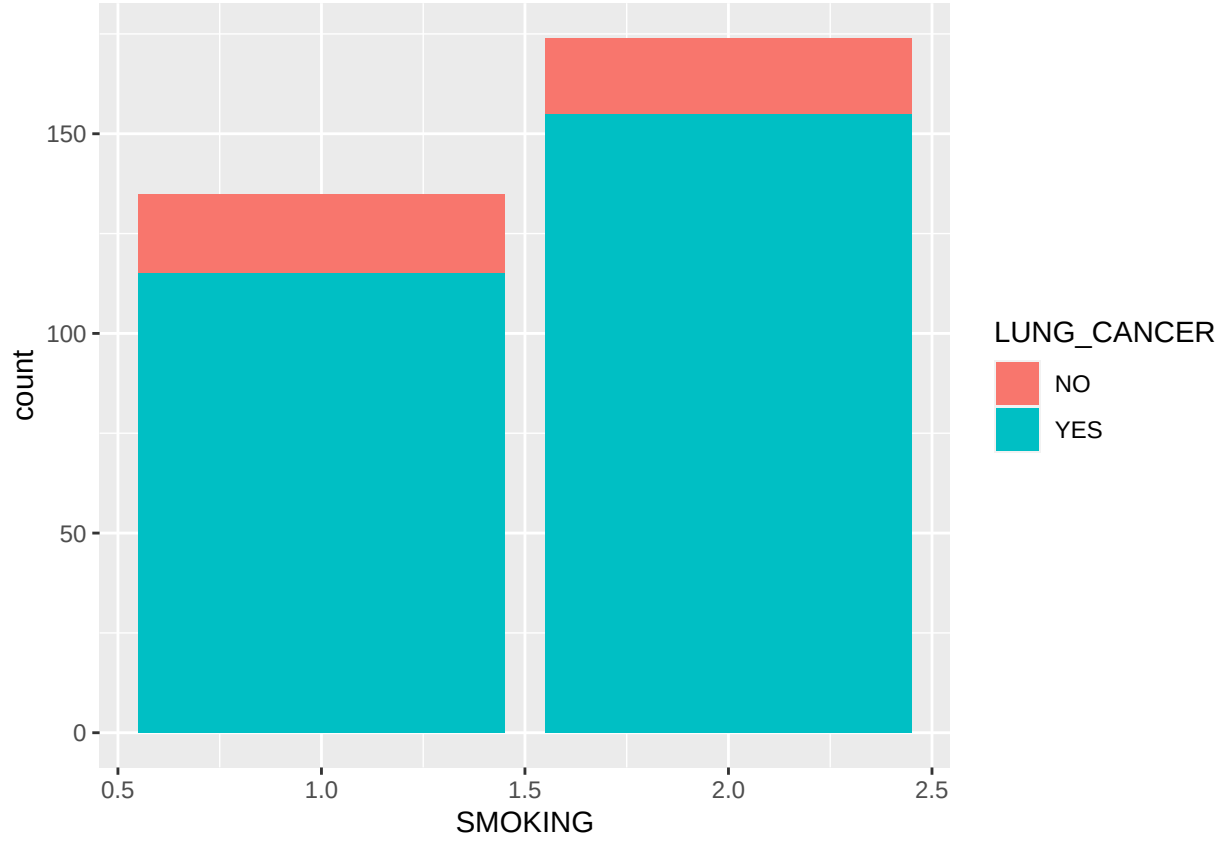
```
## [1] 41.42395
```

```
SHORTNESS.OF.BREATH_negative_cancer_positive_percentage=(g/n)*100
SHORTNESS.OF.BREATH_negative_cancer_positive_percentage#There are Approx #30% people  who don't have sh
```

```
## [1] 30.42071
```

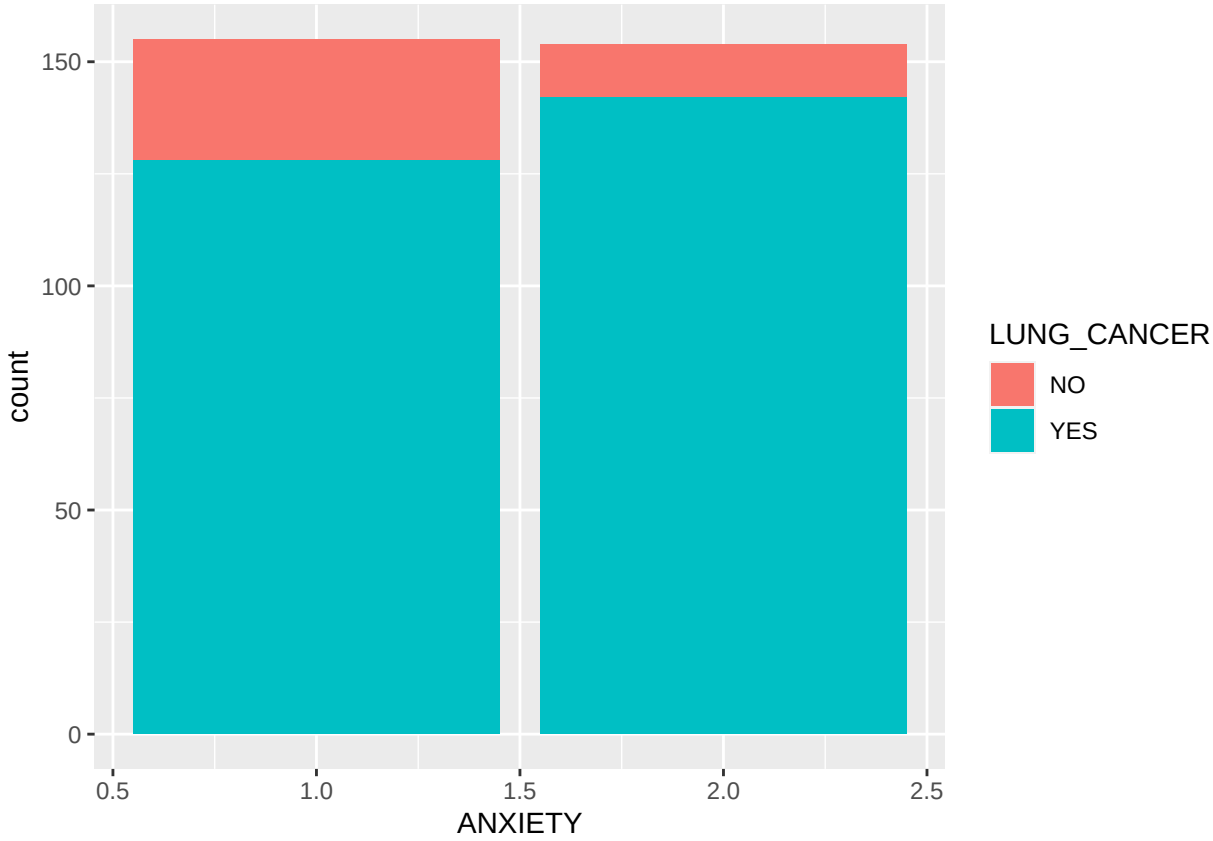
```
#As indicated by all the precentages, there is high chance that any #person who don't smoke,do not have
```

```
library('ggplot2')
ggplot(lung_cancer_data, aes(x =SMOKING, fill =LUNG_CANCER )) +
  geom_bar(position = "stack")
```



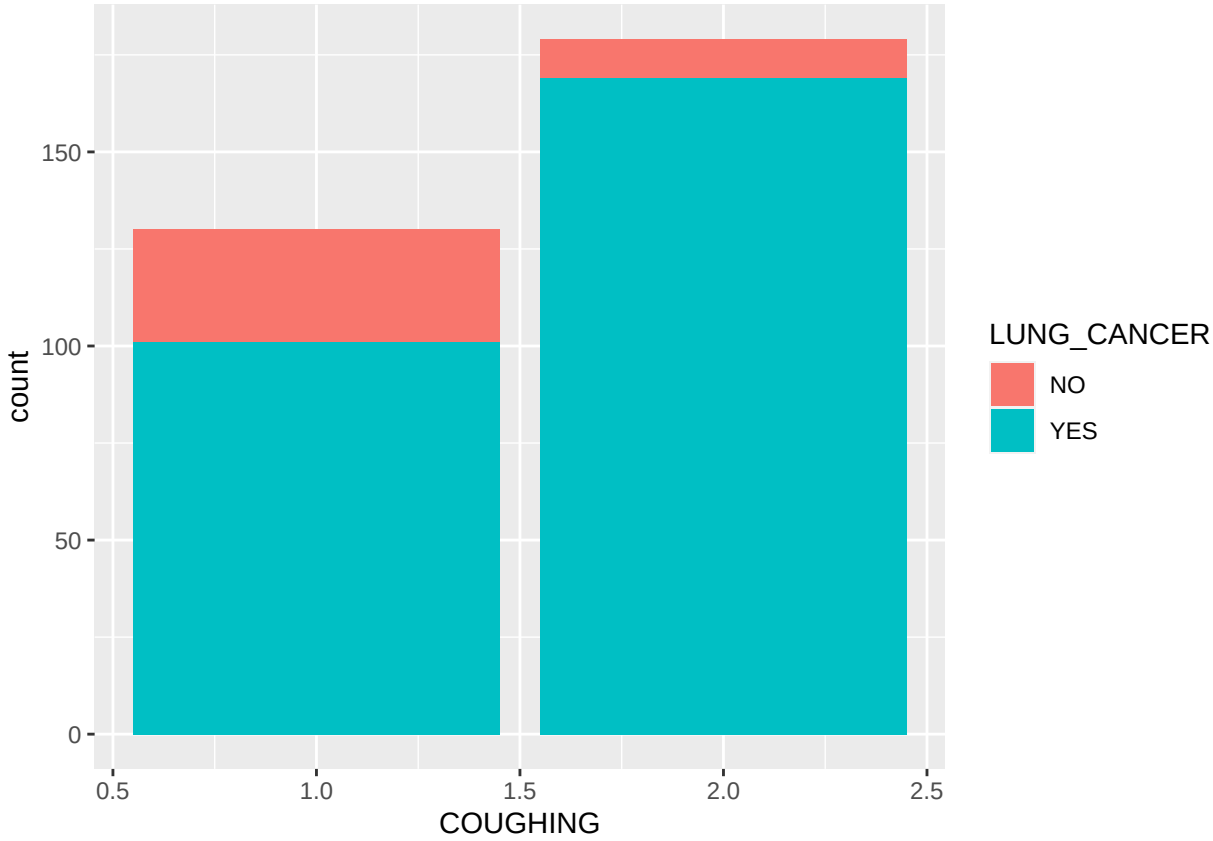

```
#Bar plot representing no of people who have cancer and who don't out
#of those who smoke and don't smoke.
```

```
ggplot(lung_cancer_data,
       aes(x =ANXIETY,
           fill =LUNG_CANCER )) +
  geom_bar(position = "stack")
```



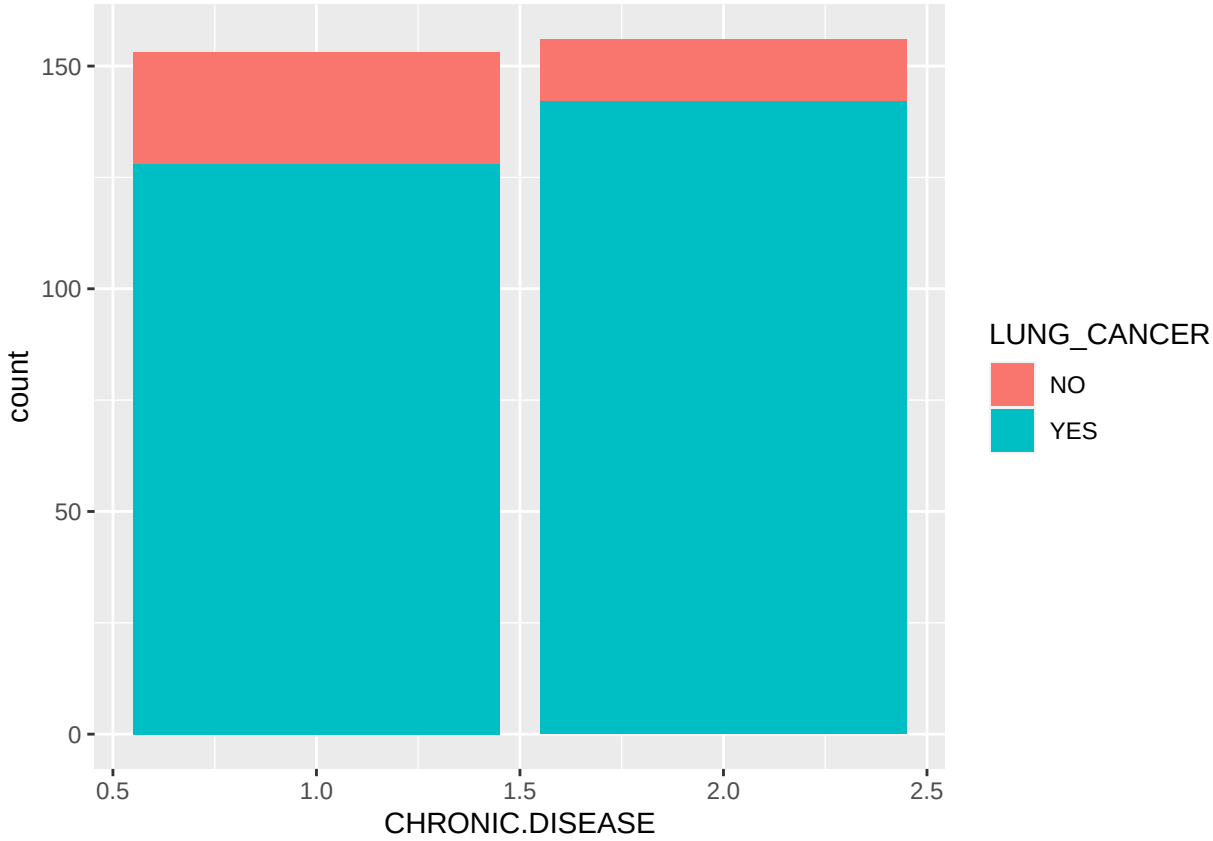

```
#Bar plot representing no of people who have cancer and who don't out
#of those who have anxiety and  who don't have anxiety.
```

```
ggplot(lung_cancer_data,
       aes(x =COUGHING,
           fill =LUNG_CANCER )) +
  geom_bar(position = "stack")
```



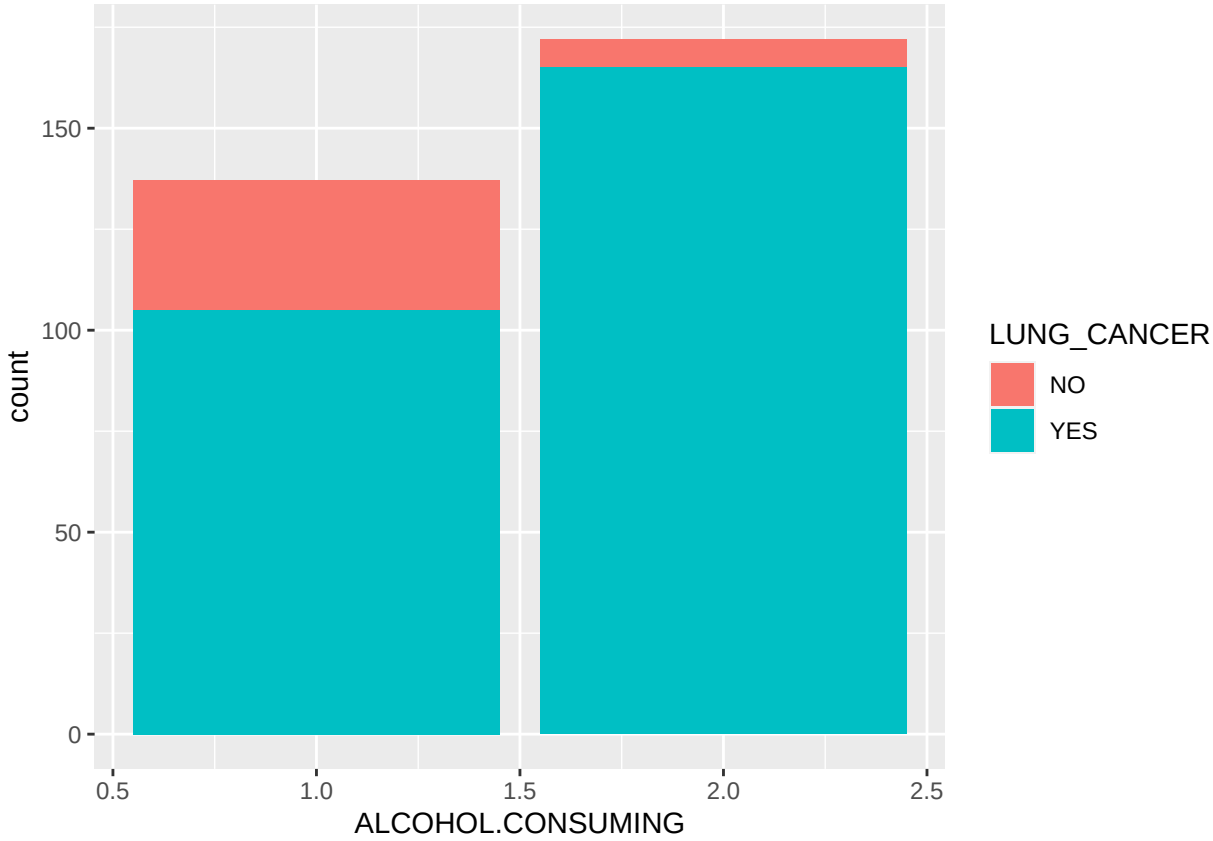

```
#Bar plot representing no of people who have cancer and who don't out
#of those who have cough and don't.
```

```
ggplot(lung_cancer_data,
       aes(x =CHRONIC.DISEASE,
           fill =LUNG_CANCER )) +
  geom_bar(position = "stack")
```



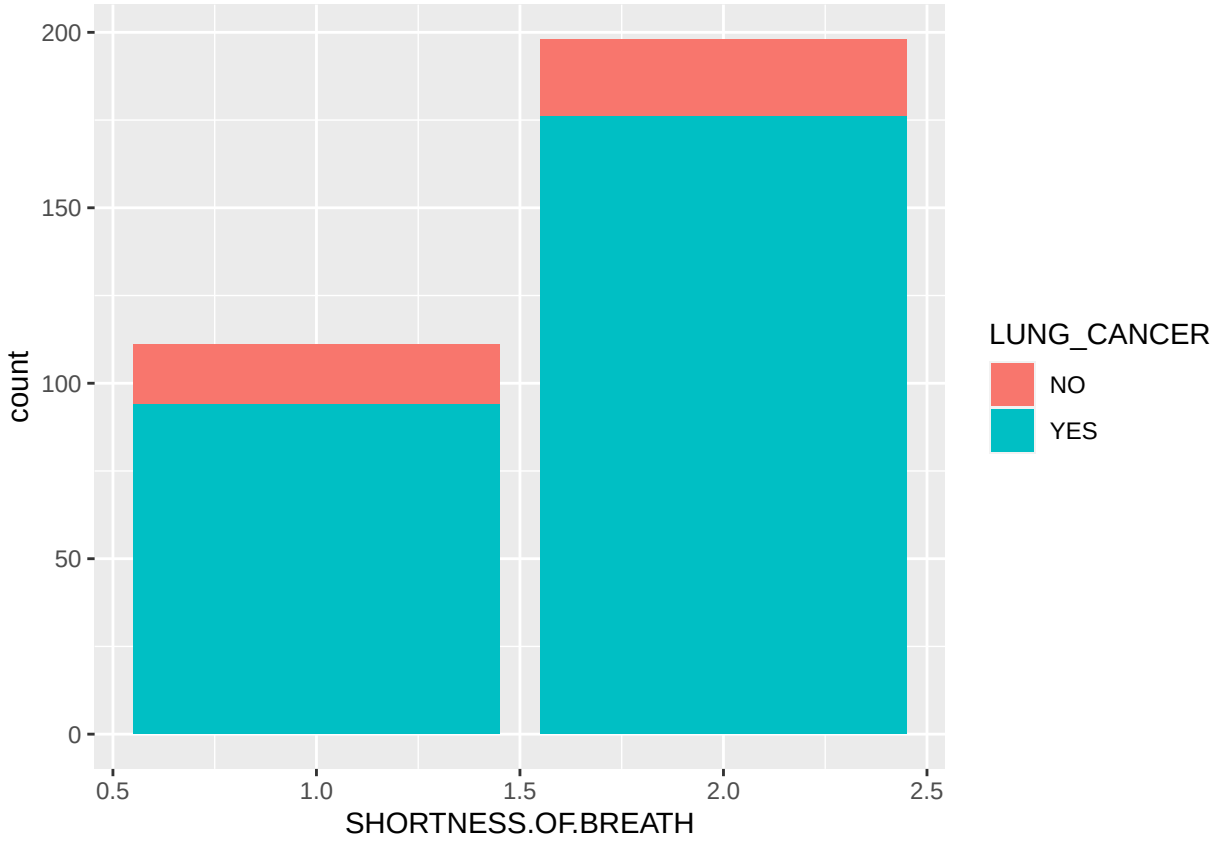

```
#Bar plot representing no of people who have cancer and who don't out
#of those who have chronic disease and who don't.
```

```
ggplot(lung_cancer_data,
       aes(x =ALCOHOL.CONSUMING,
           fill =LUNG_CANCER )) +
  geom_bar(position = "stack")
```



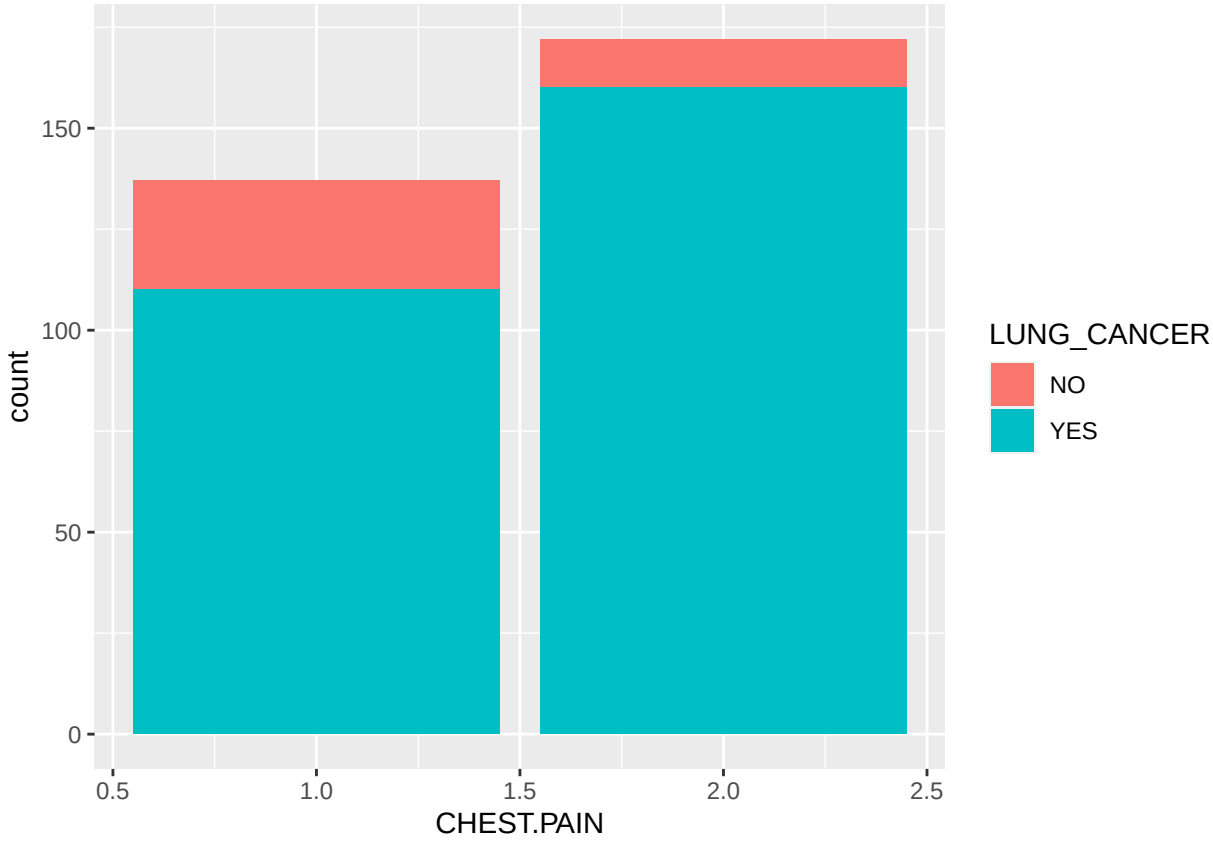

```
#Bar plot representing no of people who have cancer and who don't out
#of those who consume alcohol and who don't.
```

```
ggplot(lung_cancer_data,
       aes(x =SHORTNESS.OF.BREATH,
           fill =LUNG_CANCER )) +
  geom_bar(position = "stack")
```



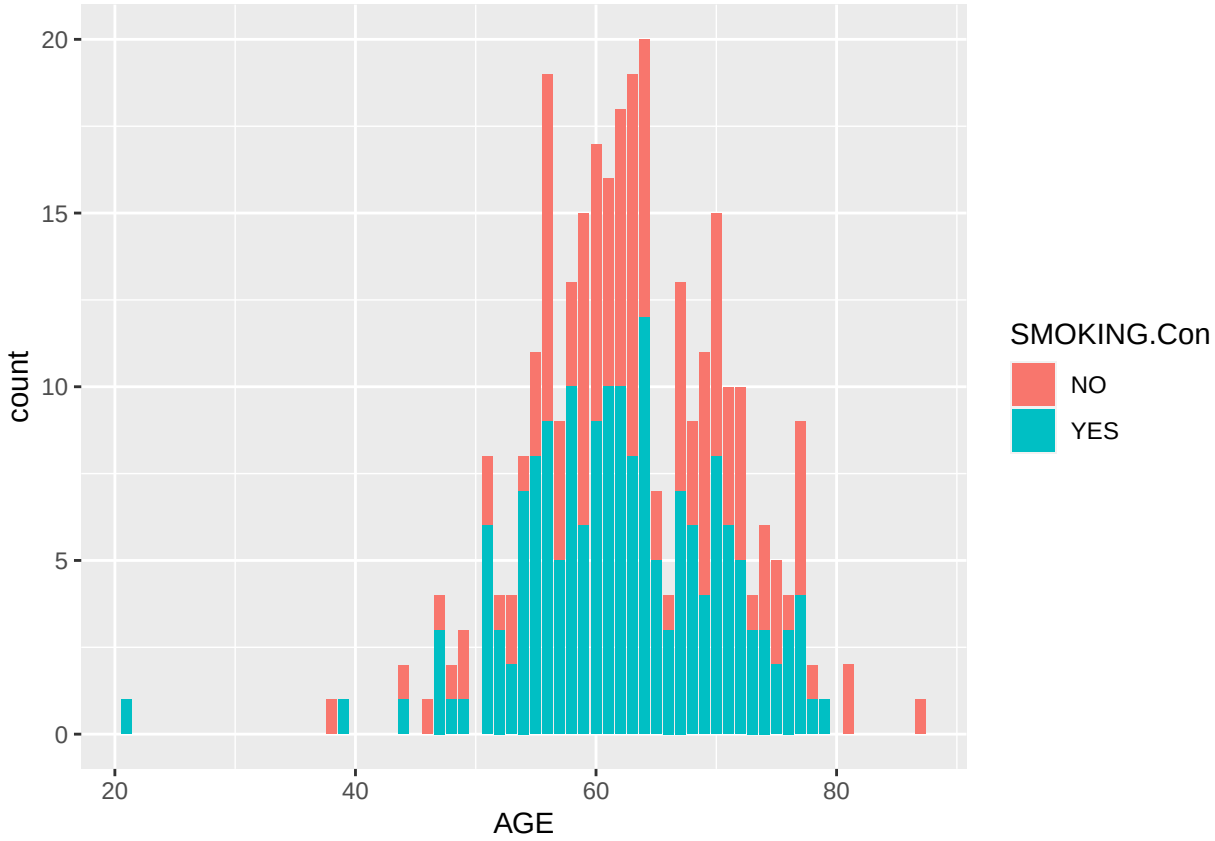

```
#Bar plot representing no of people who have cancer and who don't out
#of those who have shortness of breath and who don't.
```

```
ggplot(lung_cancer_data,
       aes(x =CHEST.PAIN,
           fill =LUNG_CANCER )) +
  geom_bar(position = "stack")
```



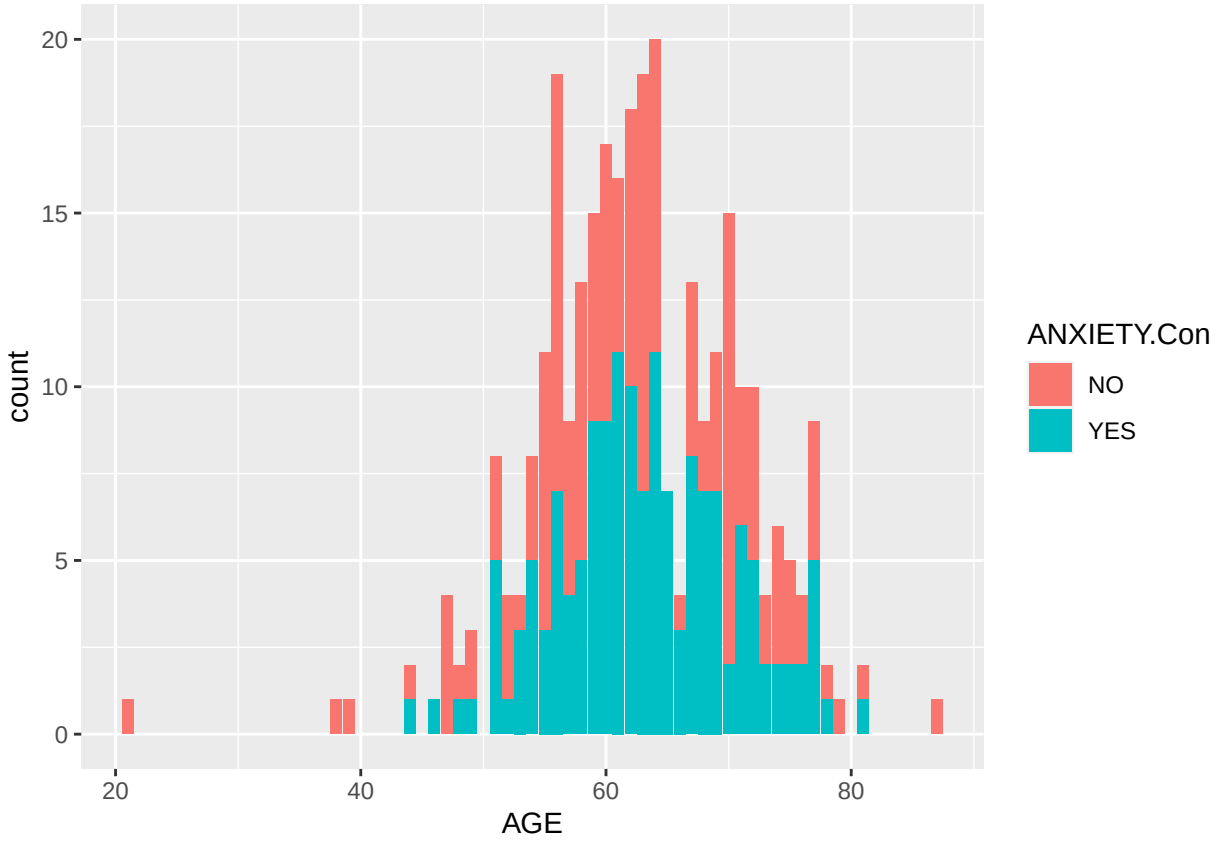

```
#Bar plot representing no of people who have cancer and who don't out
#of those who have chest pain and who don't.
```

```
lung_cancer_data_con <- within(lung_cancer_data, {
  SMOKING.Con<-NA
  SMOKING.Con[SMOKING==1] <- "NO"
  SMOKING.Con[SMOKING==2] <- "YES"

   })
ggplot(lung_cancer_data_con,
       aes(x =AGE,
           fill =SMOKING.Con )) +
  geom_bar(position = "stack")
```



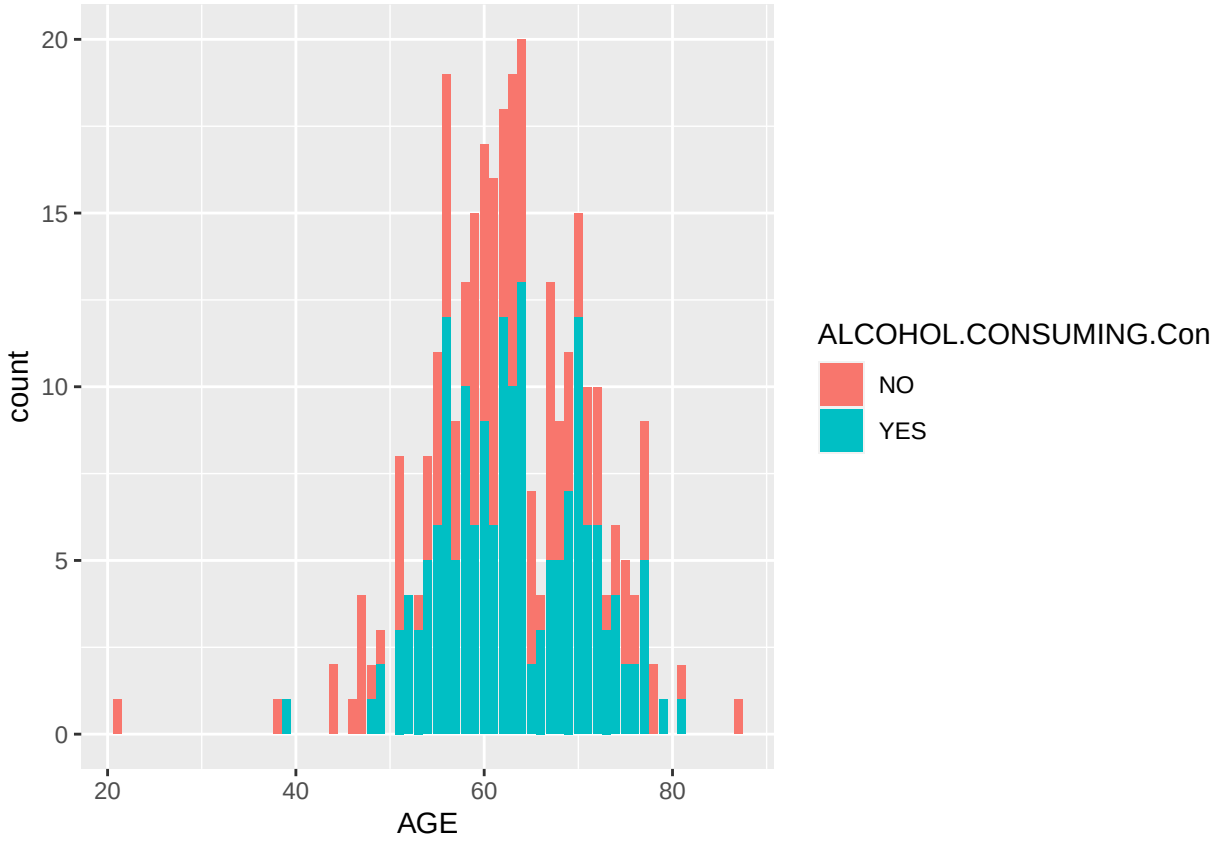

```
#bar plot representing people of particular age smoking and not smoking.
```

```
lung_cancer_data_con1 <- within(lung_cancer_data, {
  ANXIETY.Con<-NA
  ANXIETY.Con[ANXIETY==1] <- "NO"
 ANXIETY.Con[ANXIETY==2] <- "YES"

   })
ggplot(lung_cancer_data_con1,
       aes(x =AGE,
           fill =ANXIETY.Con )) +
  geom_bar(position = "stack")
```



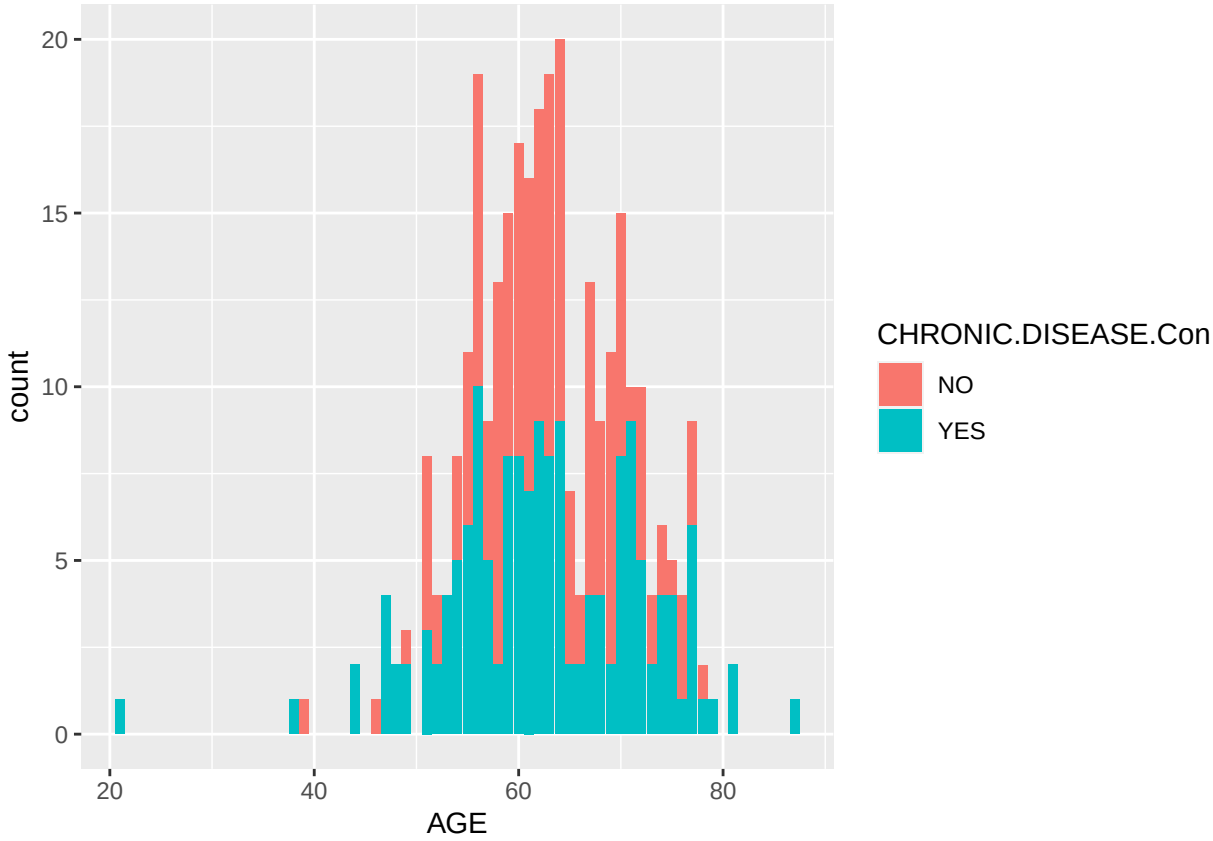

```
#bar plot representing people of particular age having anxiety or not.
```

```
lung_cancer_data_con2 <- within(lung_cancer_data, {
  ALCOHOL.CONSUMING.Con<-NA
  ALCOHOL.CONSUMING.Con[ALCOHOL.CONSUMING==1] <- "NO"
 ALCOHOL.CONSUMING.Con[ALCOHOL.CONSUMING==2] <- "YES"

   })
ggplot(lung_cancer_data_con2,
       aes(x =AGE,
           fill =ALCOHOL.CONSUMING.Con )) +
  geom_bar(position = "stack")
```



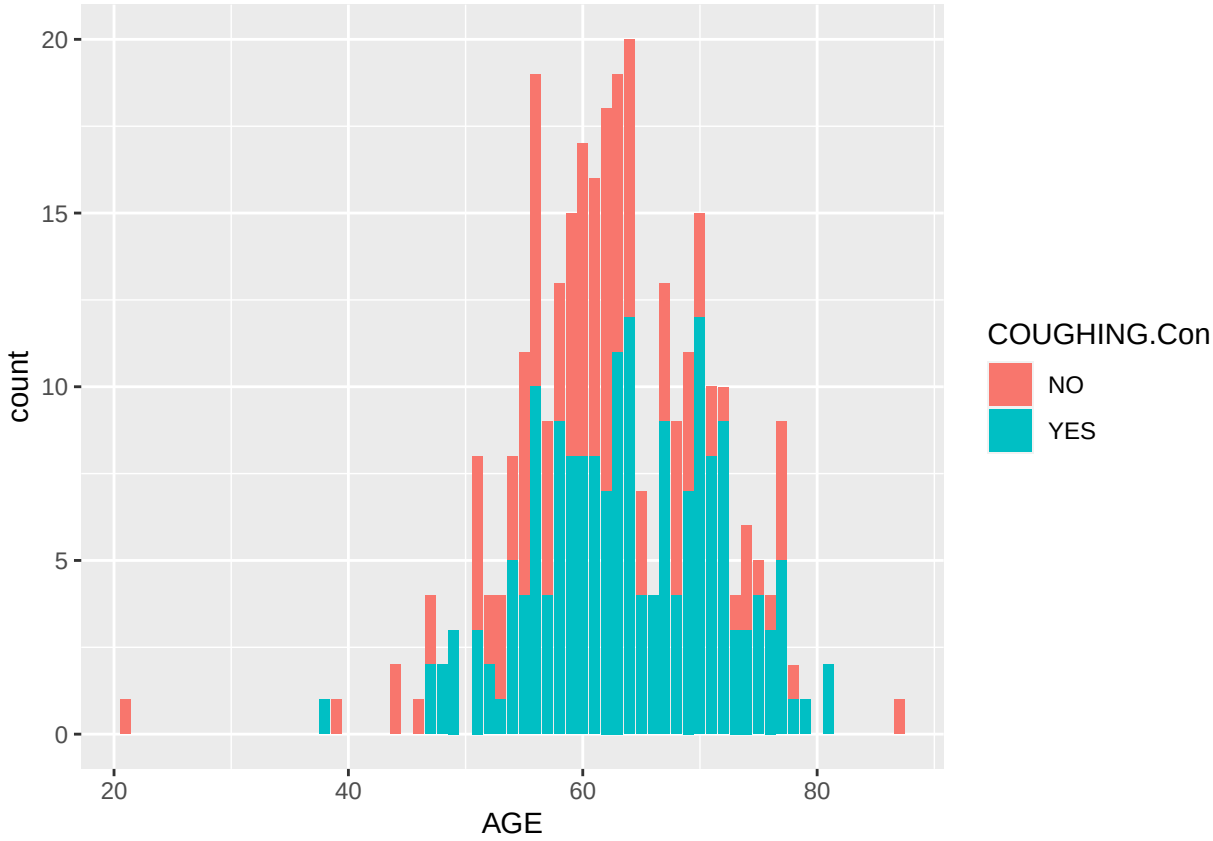

```
#bar plot representing people of particular age consuming alcohol or #not.
```

```
lung_cancer_data_con3 <- within(lung_cancer_data, {
  CHRONIC.DISEASE.Con<-NA
  CHRONIC.DISEASE.Con[CHRONIC.DISEASE==1] <- "NO"
CHRONIC.DISEASE.Con[CHRONIC.DISEASE==2] <- "YES"

   })
ggplot(lung_cancer_data_con3,
       aes(x =AGE,
           fill =CHRONIC.DISEASE.Con )) +
  geom_bar(position = "stack")
```



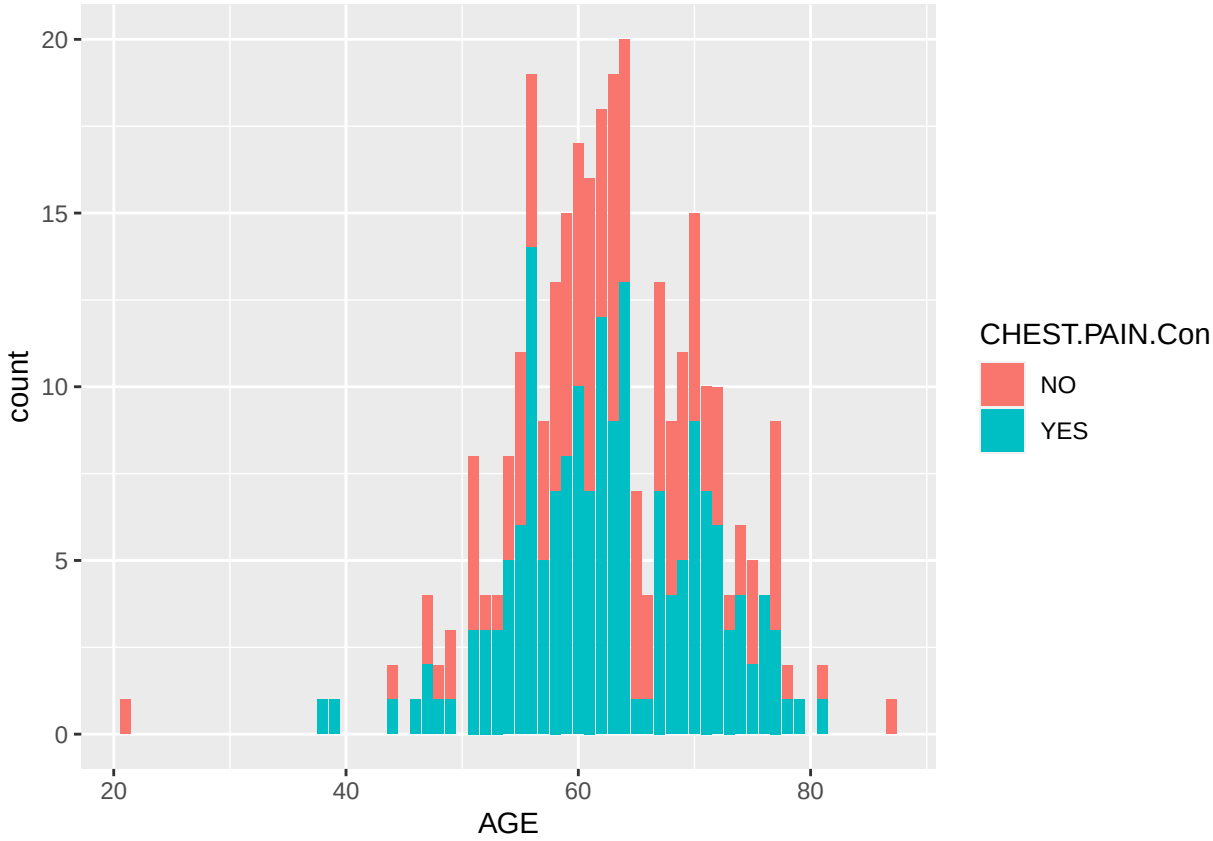

```
#bar plot representing people of particular age having chronic disease #or not.
```

```
lung_cancer_data_con4 <- within(lung_cancer_data, {
  COUGHING.Con<-NA
  COUGHING.Con[COUGHING==1] <- "NO"
COUGHING.Con[COUGHING==2] <- "YES"

   })
ggplot(lung_cancer_data_con4,
       aes(x =AGE,
           fill =COUGHING.Con )) +
  geom_bar(position = "stack")
```



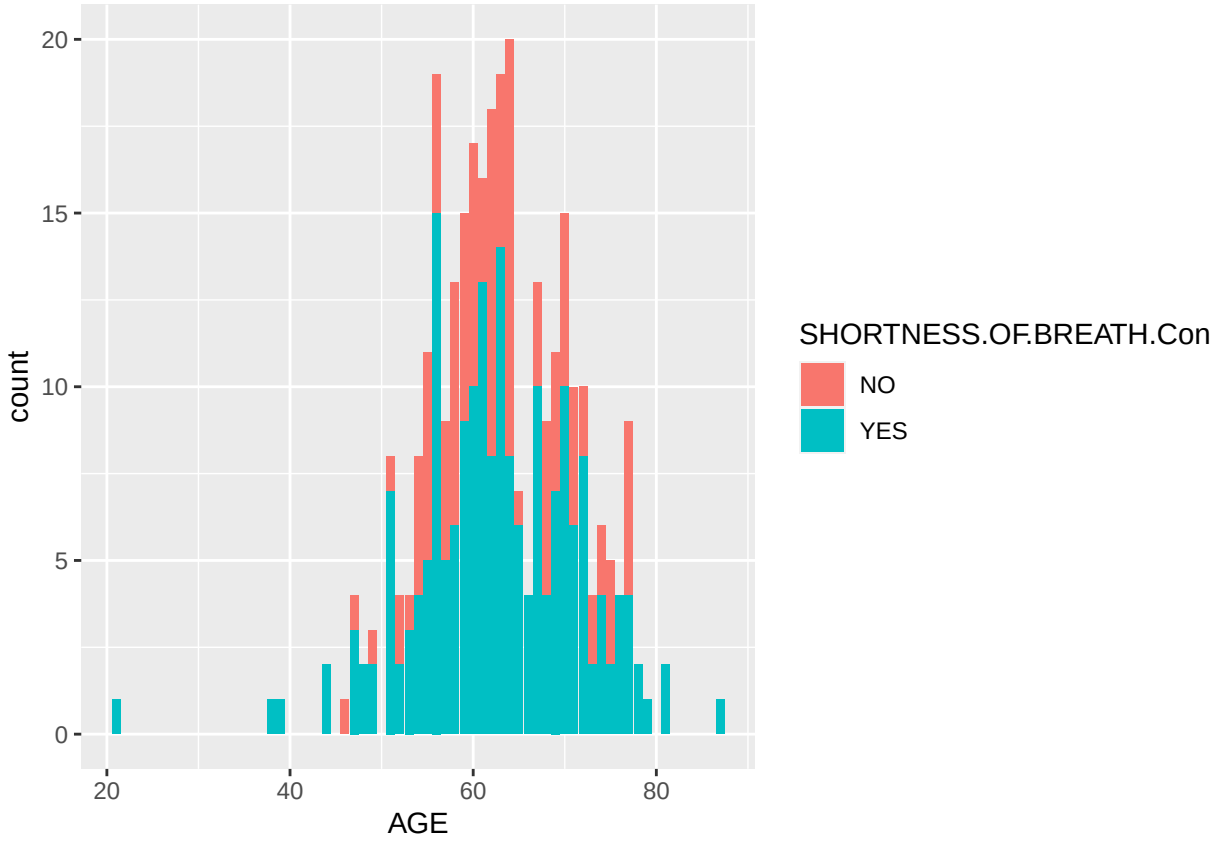

```
#bar plot representing people of particular age having cough or
#not.
```

```
lung_cancer_data_con5 <- within(lung_cancer_data, {
  CHEST.PAIN.Con<-NA
  CHEST.PAIN.Con[CHEST.PAIN==1] <- "NO"
 CHEST.PAIN.Con[CHEST.PAIN==2] <- "YES"

   })
ggplot(lung_cancer_data_con5,
       aes(x =AGE,
           fill =CHEST.PAIN.Con )) +
  geom_bar(position = "stack")
```




```
#bar plot representing people of particular age having chest pain or #not.
```

```
lung_cancer_data_con6 <- within(lung_cancer_data, {
  SHORTNESS.OF.BREATH.Con<-NA
  SHORTNESS.OF.BREATH.Con[SHORTNESS.OF.BREATH==1] <- "NO"
 SHORTNESS.OF.BREATH.Con[SHORTNESS.OF.BREATH==2] <- "YES"

   })
ggplot(lung_cancer_data_con6,
       aes(x =AGE,
           fill =SHORTNESS.OF.BREATH.Con )) +
  geom_bar(position = "stack")
```




```
#bar plot representing people of particular age having shortness of #breath or not.
```

#Summary #After doing the analysis of the given data set, I have come across the #fact that this data set lacks information about every age group, it #majorly consists of the people between 55 to 70 year of age, which #makes it difficult to reach at any of the objective that I initially #planned to achieve. But still for this age group I have figured out #that the people who smoke,have shortness of breathness, and coughing #problems, consume alcohol,have higher chances of Lung Cancer, but also #on the other hand there is a significant set of people(although the #percentage is less) who do not smoke, don't have shortness of #breathness and coughing problems and don't consume alcohol also might #get Lung cancer.

#Addition to this I got to know that from the given data set, from #almost every age group, there are more than 50% people who smoke,have #shortness of breathness, and coughing problems, consume alcohol,have #chest pain, anxiety and chronic disease and less than 50% don't, so we #can say that not majority of the people in the given data set are not #living a healthy life style. Also the results obtained reflects that #these are not the only factors which causes lung cancer, beacause there #are people who didnt smoke or have any other issues, but still got lung #cancer. So we can interpret that these are not the only causes of Lung #cancer, they are but not the only ones.