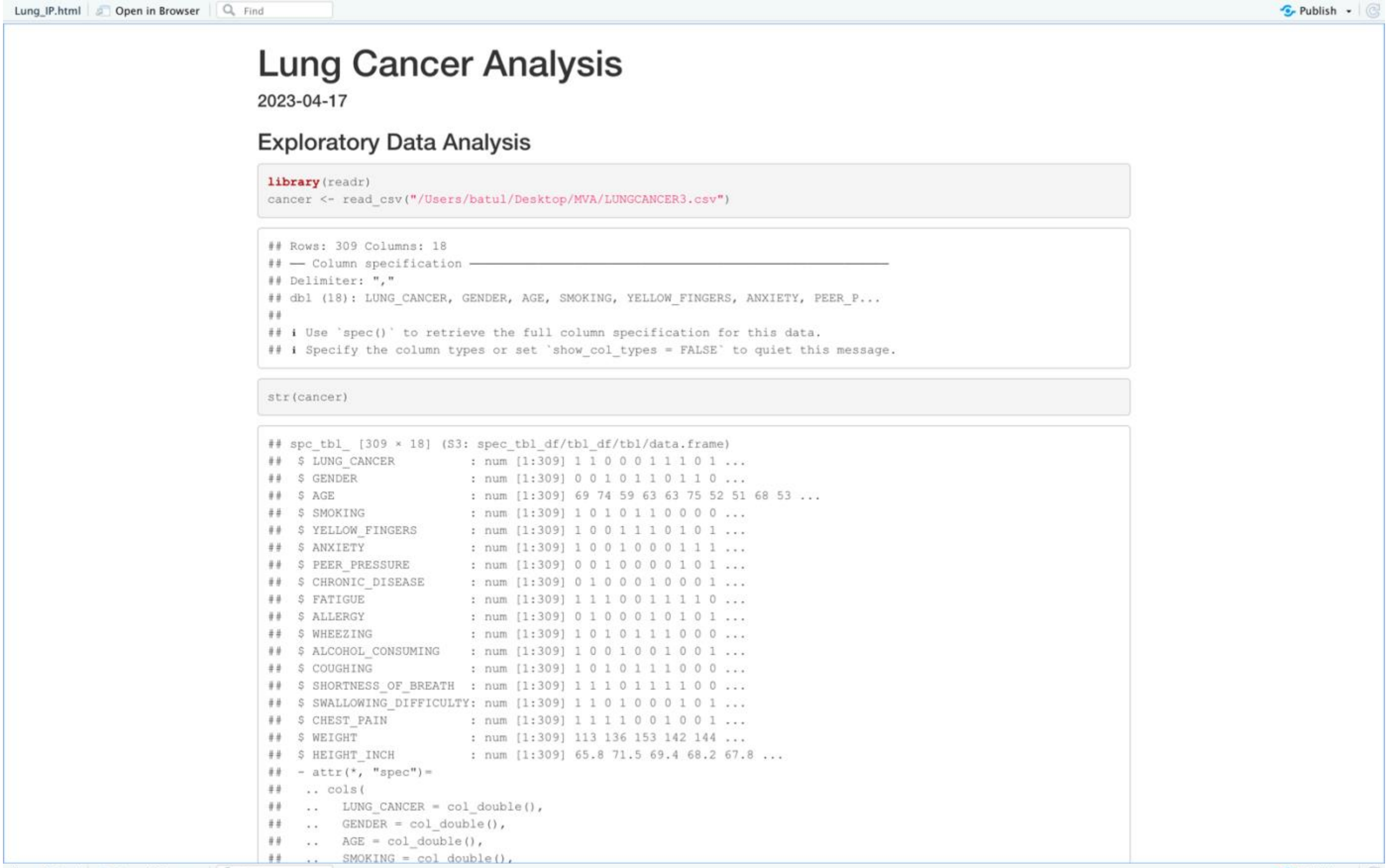
Individual Project

Performing Exploratory Data Analysis



Lung Cancer Analysis

2023-04-17

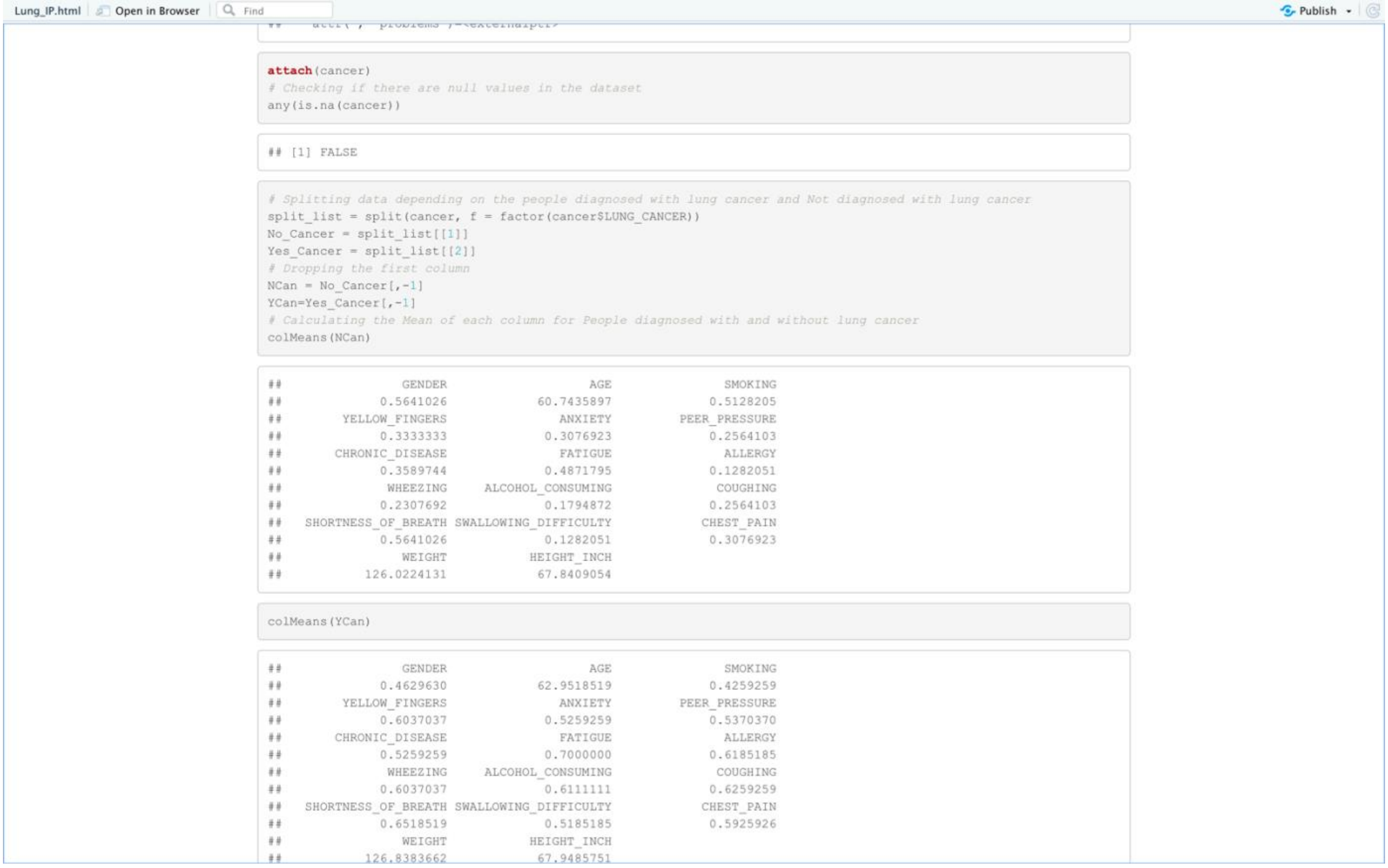
Exploratory Data Analysis

```
library(readr)
cancer <- read_csv("/Users/batul/Desktop/MVA/LUNGCANCER3.csv")
```

```
## Rows: 309 Columns: 18
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## dbl (18): LUNG_CANCER, GENDER, AGE, SMOKING, YELLOW_FINGERS, ANXIETY, PEER_P...
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
str(cancer)
```

```
## spc_tbl_ [309 × 18] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ LUNG_CANCER          : num [1:309] 1 1 0 0 0 1 1 1 0 1 ...
##  $ GENDER               : num [1:309] 0 0 1 0 1 1 0 1 1 0 ...
##  $ AGE                  : num [1:309] 69 74 59 63 63 75 52 51 68 53 ...
##  $ SMOKING              : num [1:309] 1 0 1 0 1 1 0 0 0 0 ...
##  $ YELLOW_FINGERS       : num [1:309] 1 0 0 1 1 1 0 1 0 1 ...
##  $ ANXIETY              : num [1:309] 1 0 0 1 0 0 0 1 1 1 ...
##  $ PEER_PRESSURE        : num [1:309] 0 0 1 0 0 0 0 1 0 1 ...
##  $ CHRONIC_DISEASE      : num [1:309] 0 1 0 0 0 1 0 0 0 1 ...
##  $ FATIGUE              : num [1:309] 1 1 1 0 0 1 1 1 1 0 ...
##  $ ALLERGY              : num [1:309] 0 1 0 0 0 1 0 1 0 1 ...
##  $ WHEEZING             : num [1:309] 1 0 1 0 1 1 1 0 0 0 ...
##  $ ALCOHOL_CONSUMING    : num [1:309] 1 0 0 1 0 0 1 0 0 1 ...
##  $ COUGHING             : num [1:309] 1 0 1 0 1 1 1 0 0 0 ...
##  $ SHORTNESS_OF_BREATH  : num [1:309] 1 1 1 0 1 1 1 1 0 0 ...
##  $ SWALLOWING_DIFFICULTY: num [1:309] 1 1 0 1 0 0 0 1 0 1 ...
##  $ CHEST_PAIN           : num [1:309] 1 1 1 1 0 0 1 0 0 1 ...
##  $ WEIGHT               : num [1:309] 113 136 153 142 144 ...
##  $ HEIGHT_INCH          : num [1:309] 65.8 71.5 69.4 68.2 67.8 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   LUNG_CANCER = col_double(),
##   ..   GENDER = col_double(),
##   ..   AGE = col_double(),
##   ..   SMOKING = col_double(),
```



```
##  - attr(*, "problems")=<externalptr>
```

```
attach(cancer)
# Checking if there are null values in the dataset
any(is.na(cancer))
```

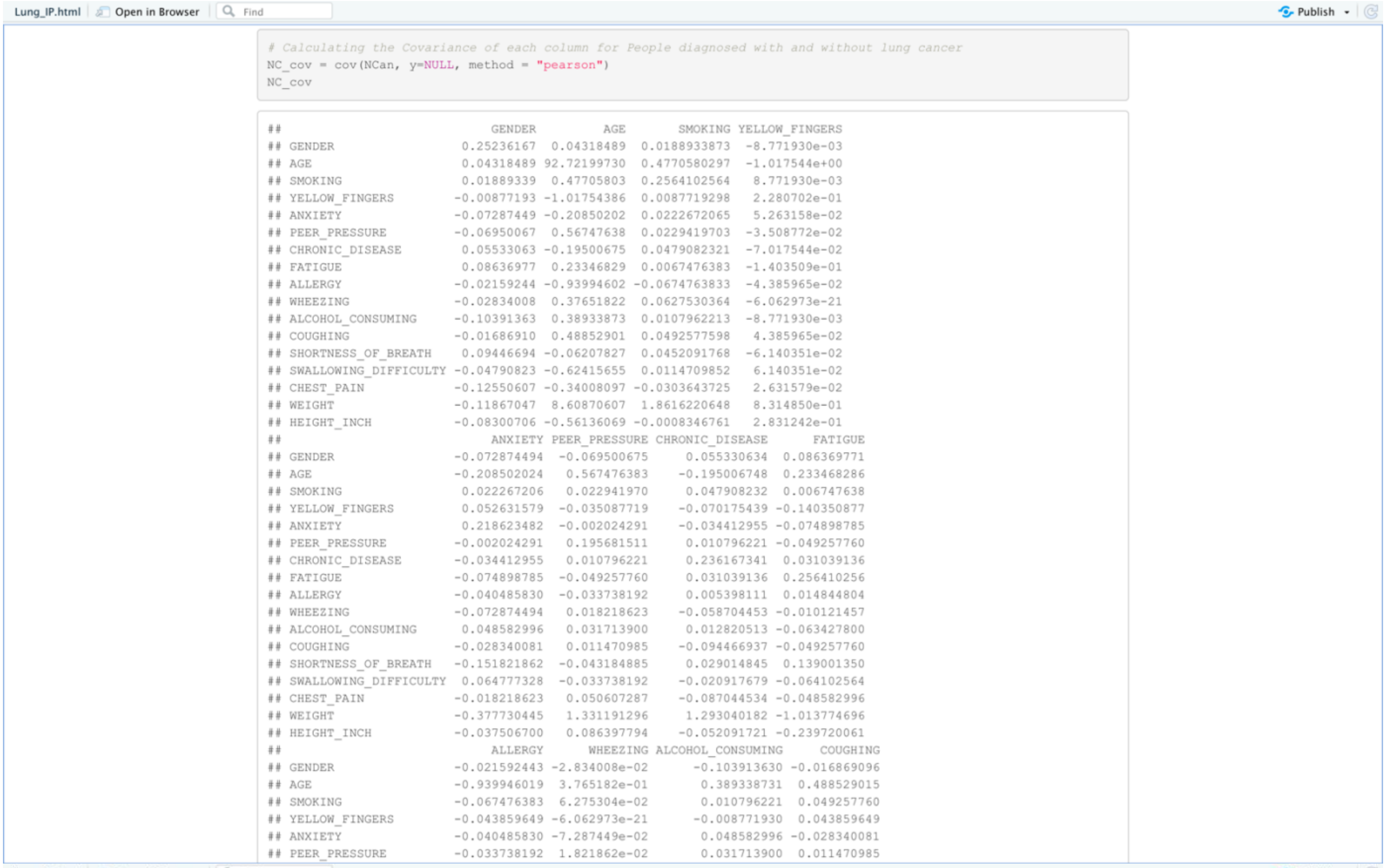
```
## [1] FALSE
```

```
# Splitting data depending on the people diagnosed with lung cancer and Not diagnosed with lung cancer
split_list = split(cancer, f = factor(cancer$LUNG_CANCER))
No_Cancer = split_list[[1]]
Yes_Cancer = split_list[[2]]
# Dropping the first column
NCan = No_Cancer[,-1]
YCan=Yes_Cancer[,-1]
# Calculating the Mean of each column for People diagnosed with and without lung cancer
colMeans(NCan)
```

```
##                GENDER                   AGE               SMOKING 
##             0.5641026            60.7435897             0.5128205 
##        YELLOW_FINGERS               ANXIETY         PEER_PRESSURE 
##             0.3333333             0.3076923             0.2564103 
##       CHRONIC_DISEASE               FATIGUE               ALLERGY 
##             0.3589744             0.4871795             0.1282051 
##              WHEEZING     ALCOHOL_CONSUMING              COUGHING 
##             0.2307692             0.1794872             0.2564103 
##   SHORTNESS_OF_BREATH SWALLOWING_DIFFICULTY            CHEST_PAIN 
##             0.5641026             0.1282051             0.3076923 
##                WEIGHT           HEIGHT_INCH 
##           126.0224131            67.8409054 
```

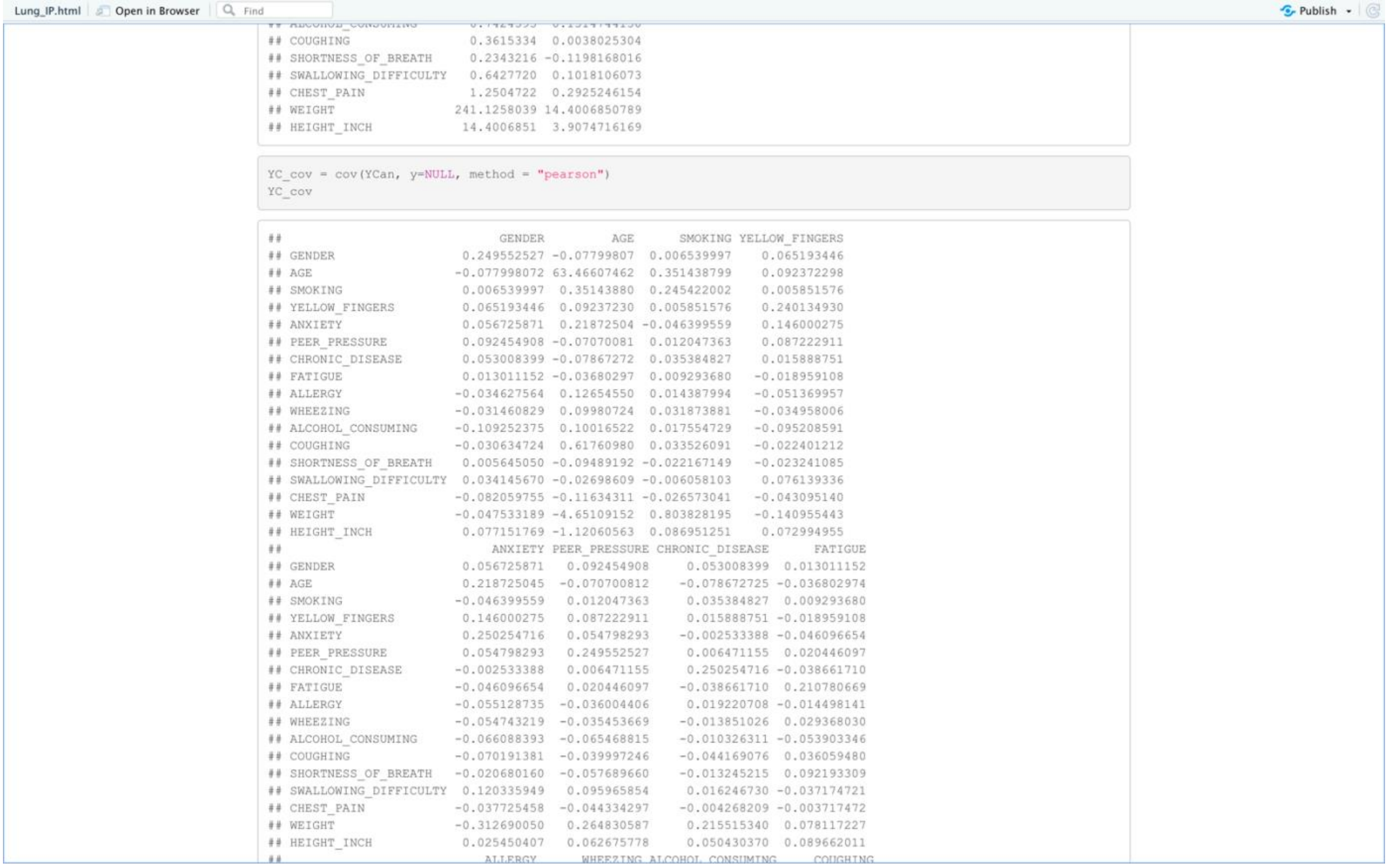
```
colMeans(YCan)
```

```
##                GENDER                   AGE               SMOKING 
##             0.4629630            62.9518519             0.4259259 
##        YELLOW_FINGERS               ANXIETY         PEER_PRESSURE 
##             0.6037037             0.5259259             0.5370370 
##       CHRONIC_DISEASE               FATIGUE               ALLERGY 
##             0.5259259             0.7000000             0.6185185 
##              WHEEZING     ALCOHOL_CONSUMING              COUGHING 
##             0.6037037             0.6111111             0.6259259 
##   SHORTNESS_OF_BREATH SWALLOWING_DIFFICULTY            CHEST_PAIN 
##             0.6518519             0.5185185             0.5925926 
##                WEIGHT           HEIGHT_INCH 
##           126.8383662            67.9485751 
```



```
# Calculating the Covariance of each column for People diagnosed with and without lung cancer
NC_cov = cov(NCan, y=NULL, method = "pearson")
NC_cov
```

```
##                           GENDER         AGE       SMOKING YELLOW_FINGERS
## GENDER                0.25236167  0.04318489  0.0188933873  -8.771930e-03
## AGE                   0.04318489 92.72199730  0.4770580297  -1.017544e+00
## SMOKING               0.01889339  0.47705803  0.2564102564   8.771930e-03
## YELLOW_FINGERS       -0.00877193 -1.01754386  0.0087719298   2.280702e-01
## ANXIETY              -0.07287449 -0.20850202  0.0222672065   5.263158e-02
## PEER_PRESSURE        -0.06950067  0.56747638  0.0229419703  -3.508772e-02
## CHRONIC_DISEASE       0.05533063 -0.19500675  0.0479082321  -7.017544e-02
## FATIGUE               0.08636977  0.23346829  0.0067476383  -1.403509e-01
## ALLERGY              -0.02159244 -0.93994602 -0.0674763833  -4.385965e-02
## WHEEZING             -0.02834008  0.37651822  0.0627530364  -6.062973e-21
## ALCOHOL_CONSUMING    -0.10391363  0.38933873  0.0107962213  -8.771930e-03
## COUGHING             -0.01686910  0.48852901  0.0492577598   4.385965e-02
## SHORTNESS_OF_BREATH   0.09446694 -0.06207827  0.0452091768  -6.140351e-02
## SWALLOWING_DIFFICULTY -0.04790823 -0.62415655  0.0114709852   6.140351e-02
## CHEST_PAIN           -0.12550607 -0.34008097 -0.0303643725   2.631579e-02
## WEIGHT               -0.11867047  8.60870607  1.8616220648   8.314850e-01
## HEIGHT_INCH          -0.08300706 -0.56136069 -0.0008346761   2.831242e-01
##                           ANXIETY PEER_PRESSURE CHRONIC_DISEASE      FATIGUE
## GENDER               -0.072874494  -0.069500675     0.055330634  0.086369771
## AGE                  -0.208502024   0.567476383    -0.195006748  0.233468286
## SMOKING               0.022267206   0.022941970     0.047908232  0.006747638
## YELLOW_FINGERS        0.052631579  -0.035087719    -0.070175439 -0.140350877
## ANXIETY               0.218623482  -0.002024291    -0.034412955 -0.074898785
## PEER_PRESSURE        -0.002024291   0.195681511     0.010796221 -0.049257760
## CHRONIC_DISEASE      -0.034412955   0.010796221     0.236167341  0.031039136
## FATIGUE              -0.074898785  -0.049257760     0.031039136  0.256410256
## ALLERGY              -0.040485830  -0.033738192     0.005398111  0.014844804
## WHEEZING             -0.072874494   0.018218623    -0.058704453 -0.010121457
## ALCOHOL_CONSUMING     0.048582996   0.031713900     0.012820513 -0.063427800
## COUGHING             -0.028340081   0.011470985    -0.094466937 -0.049257760
## SHORTNESS_OF_BREATH  -0.151821862  -0.043184885     0.029014845  0.139001350
## SWALLOWING_DIFFICULTY 0.064777328  -0.033738192    -0.020917679 -0.064102564
## CHEST_PAIN           -0.018218623   0.050607287    -0.087044534 -0.048582996
## WEIGHT               -0.377730445   1.331191296     1.293040182 -1.013774696
## HEIGHT_INCH          -0.037506700   0.086397794    -0.052091721 -0.239720061
##                           ALLERGY      WHEEZING ALCOHOL_CONSUMING     COUGHING
## GENDER               -0.021592443 -2.834008e-02      -0.103913630 -0.016869096
## AGE                  -0.939946019  3.765182e-01       0.389338731  0.488529015
## SMOKING              -0.067476383  6.275304e-02       0.010796221  0.049257760
## YELLOW_FINGERS       -0.043859649 -6.062973e-21      -0.008771930  0.043859649
## ANXIETY              -0.040485830 -7.287449e-02       0.048582996 -0.028340081
## PEER_PRESSURE        -0.033738192  1.821862e-02       0.031713900  0.011470985
```

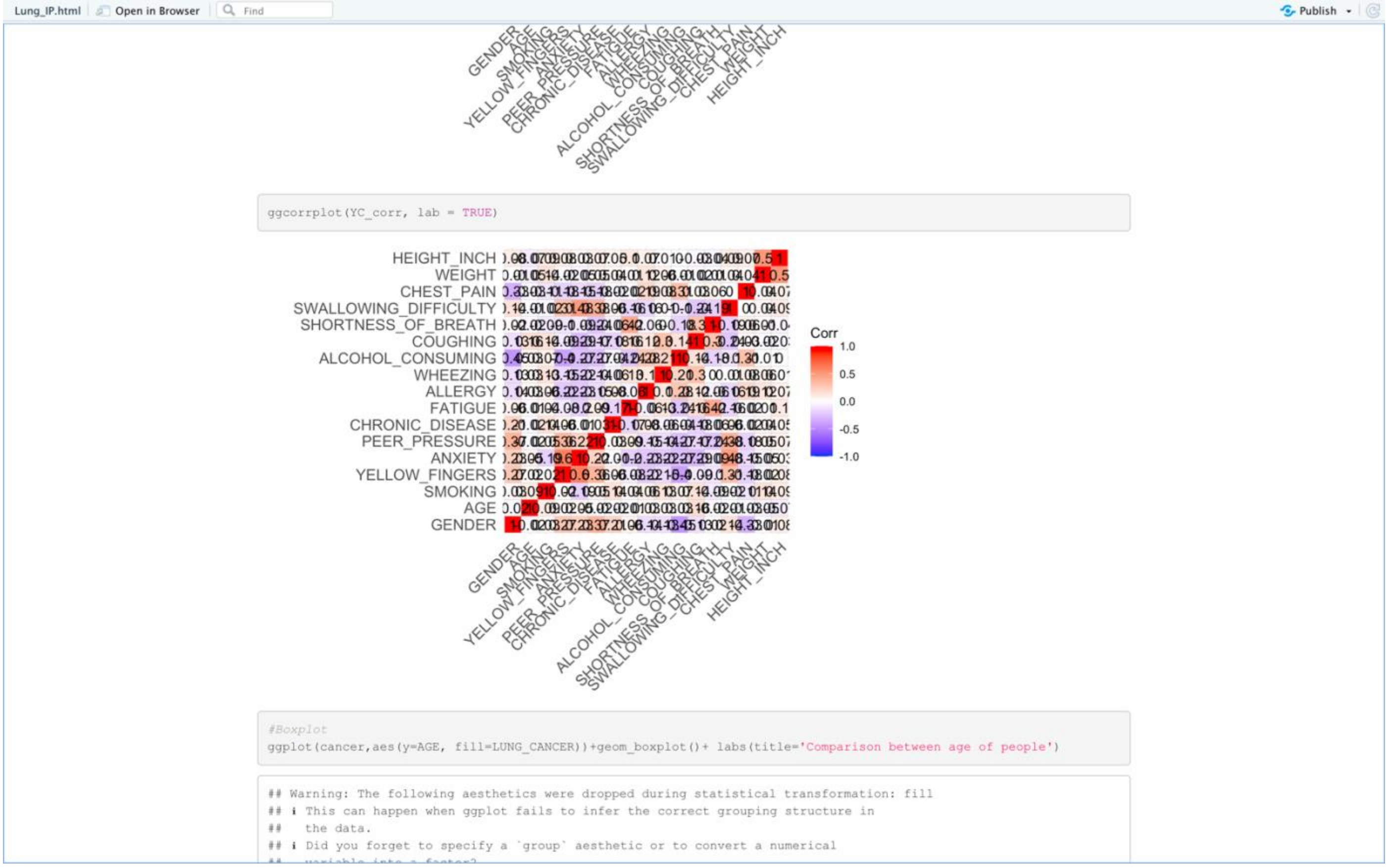


```
## ALCOHOL_CONSUMING
## COUGHING               0.3615334  0.0038025304
## SHORTNESS_OF_BREATH    0.2343216 -0.1198168016
## SWALLOWING_DIFFICULTY  0.6427720  0.1018106073
## CHEST_PAIN             1.2504722  0.2925246154
## WEIGHT               241.1258039 14.4006850789
## HEIGHT_INCH           14.4006851  3.9074716169
```

```
YC_cov = cov(YCan, y=NULL, method = "pearson")
YC_cov
```

```
##                            GENDER         AGE      SMOKING YELLOW_FINGERS
## GENDER                0.249552527 -0.07799807  0.006539997    0.065193446
## AGE                  -0.077998072 63.46607462  0.351438799    0.092372298
## SMOKING               0.006539997  0.35143880  0.245422002    0.005851576
## YELLOW_FINGERS        0.065193446  0.09237230  0.005851576    0.240134930
## ANXIETY               0.056725871  0.21872504 -0.046399559    0.146000275
## PEER_PRESSURE         0.092454908 -0.07070081  0.012047363    0.087222911
## CHRONIC_DISEASE       0.053008399 -0.07867272  0.035384827    0.015888751
## FATIGUE               0.013011152 -0.03680297  0.009293680   -0.018959108
## ALLERGY              -0.034627564  0.12654550  0.014387994   -0.051369957
## WHEEZING             -0.031460829  0.09980724  0.031873881   -0.034958006
## ALCOHOL_CONSUMING    -0.109252375  0.10016522  0.017554729   -0.095208591
## COUGHING             -0.030634724  0.61760980  0.033526091   -0.022401212
## SHORTNESS_OF_BREATH   0.005645050 -0.09489192 -0.022167149   -0.023241085
## SWALLOWING_DIFFICULTY 0.034145670 -0.02698609 -0.006058103    0.076139336
## CHEST_PAIN           -0.082059755 -0.11634311 -0.026573041   -0.043095140
## WEIGHT               -0.047533189 -4.65109152  0.803828195   -0.140955443
## HEIGHT_INCH           0.077151769 -1.12060563  0.086951251    0.072994955
##                           ANXIETY PEER_PRESSURE CHRONIC_DISEASE      FATIGUE
## GENDER                0.056725871   0.092454908     0.053008399  0.013011152
## AGE                   0.218725045  -0.070700812    -0.078672725 -0.036802974
## SMOKING              -0.046399559   0.012047363     0.035384827  0.009293680
## YELLOW_FINGERS        0.146000275   0.087222911     0.015888751 -0.018959108
## ANXIETY               0.250254716   0.054798293    -0.002533388 -0.046096654
## PEER_PRESSURE         0.054798293   0.249552527     0.006471155  0.020446097
## CHRONIC_DISEASE      -0.002533388   0.006471155     0.250254716 -0.038661710
## FATIGUE              -0.046096654   0.020446097    -0.038661710  0.210780669
## ALLERGY              -0.055128735  -0.036004406     0.019220708 -0.014498141
## WHEEZING             -0.054743219  -0.035453669    -0.013851026  0.029368030
## ALCOHOL_CONSUMING    -0.066088393  -0.065468815    -0.010326311 -0.053903346
## COUGHING             -0.070191381  -0.039997246    -0.044169076  0.036059480
## SHORTNESS_OF_BREATH  -0.020680160  -0.057689660    -0.013245215  0.092193309
## SWALLOWING_DIFFICULTY 0.120335949   0.095965854     0.016246730 -0.037174721
## CHEST_PAIN           -0.037725458  -0.044334297    -0.004268209 -0.003717472
## WEIGHT               -0.312690050   0.264830587     0.215515340  0.078117227
## HEIGHT_INCH           0.025450407   0.062675778     0.050430370  0.089662011
##                          ALLERGY      WHEEZING ALCOHOL_CONSUMING     COUGHING
```

Correlation matrix for lung cancer patients

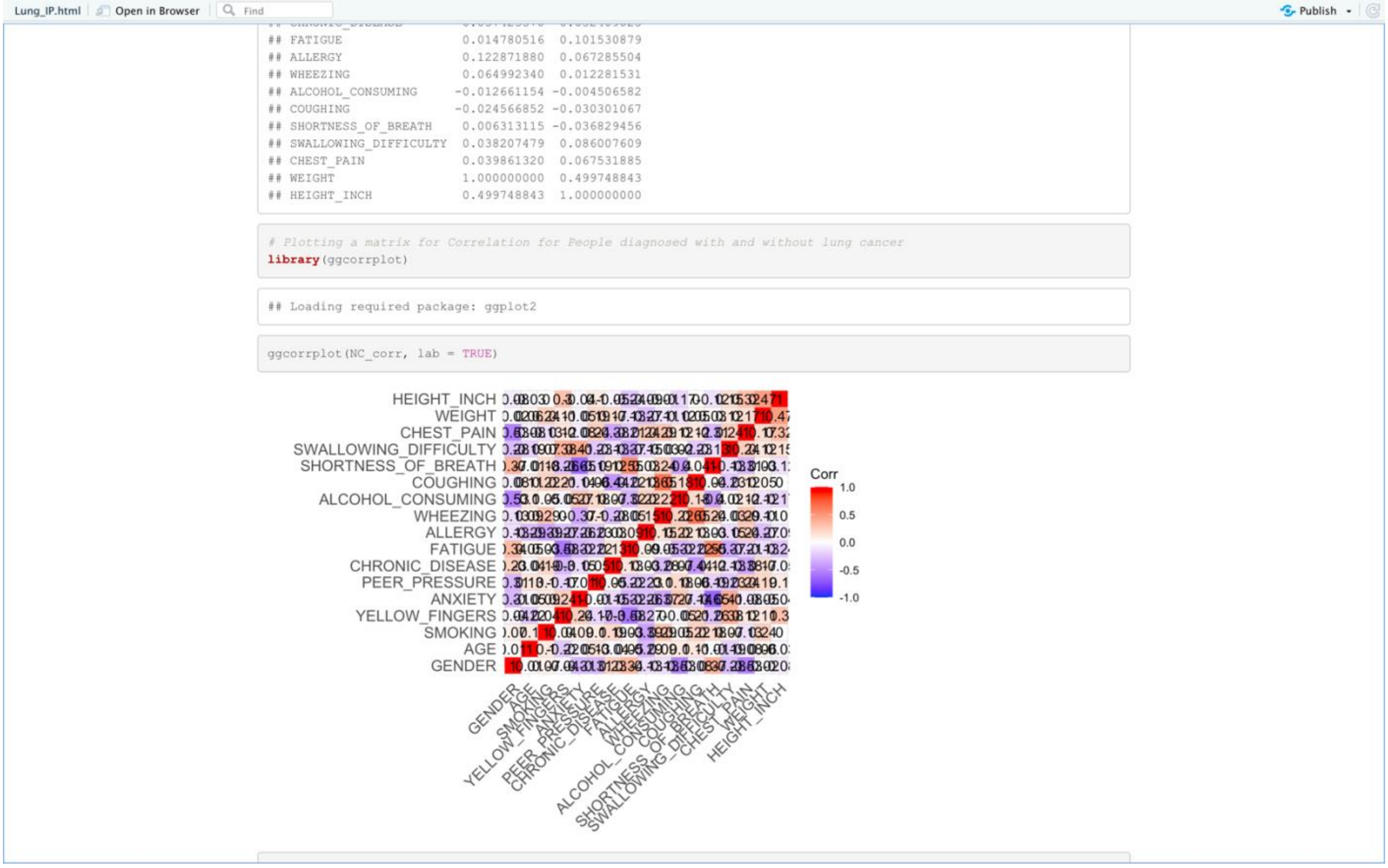


```
ggcorrplot(YC_corr, lab = TRUE)
```

```
#Boxplot
ggplot(cancer,aes(y=AGE, fill=LUNG_CANCER))+geom_boxplot()+ labs(title='Comparison between age of people')
```

```
## Warning: The following aesthetics were dropped during statistical transformation: fill
## ℹ This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## ℹ Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```

Correlation matrix for healthy people



```
## FATIGUE                0.014780516  0.101530879
## ALLERGY                0.122871880  0.067285504
## WHEEZING               0.064992340  0.012281531
## ALCOHOL_CONSUMING     -0.012661154 -0.004506582
## COUGHING              -0.024566852 -0.030301067
## SHORTNESS_OF_BREATH    0.006313115 -0.036829456
## SWALLOWING_DIFFICULTY  0.038207479  0.086007609
## CHEST_PAIN             0.039861320  0.067531885
## WEIGHT                 1.000000000  0.499748843
## HEIGHT_INCH            0.499748843  1.000000000
```

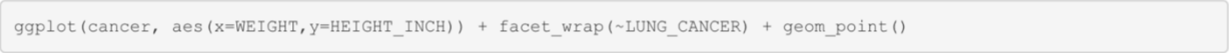
```
# Plotting a matrix for Correlation for People diagnosed with and without lung cancer
library(ggcorrplot)
```

```
## Loading required package: ggplot2
```

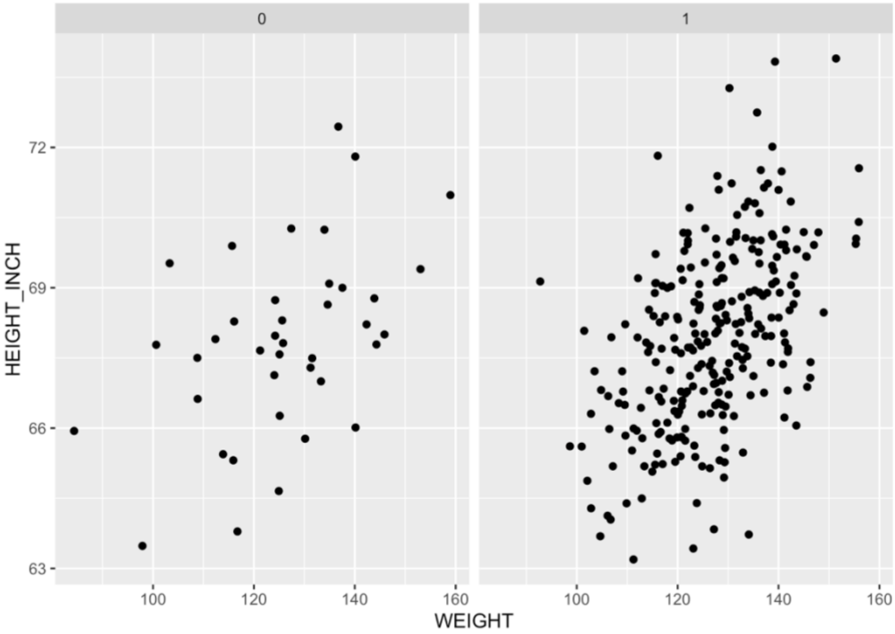
```
ggcorrplot(NC_corr, lab = TRUE)
```

GGplot of weight and height of lung cancer patients and healthy people

WEIGHT

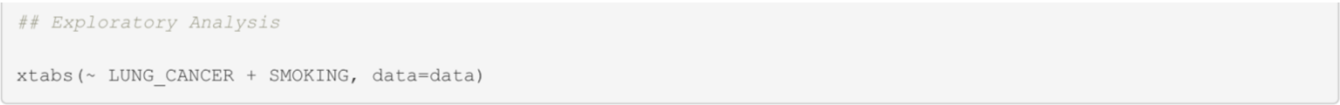


```
ggplot(cancer, aes(x=WEIGHT,y=HEIGHT_INCH)) + facet_wrap(~LUNG_CANCER) + geom_point()
```

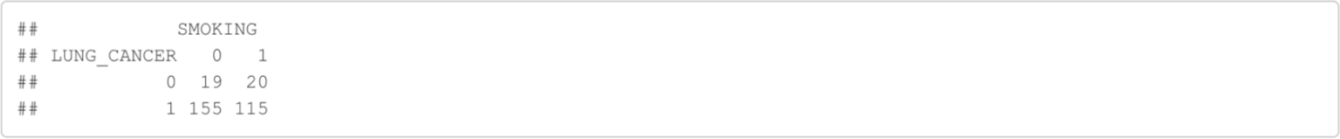


Logistic Regression:

Q) How many factors should be taken into consideration for lung cancer prediction? How to predict if someone is at a risk of getting lung cancer or not based on their symptoms?



```
## Exploratory Analysis

xtabs(~ LUNG_CANCER + SMOKING, data=data)
```



```
##             SMOKING
## LUNG_CANCER   0   1
##           0  19  20
##           1 155 115
```

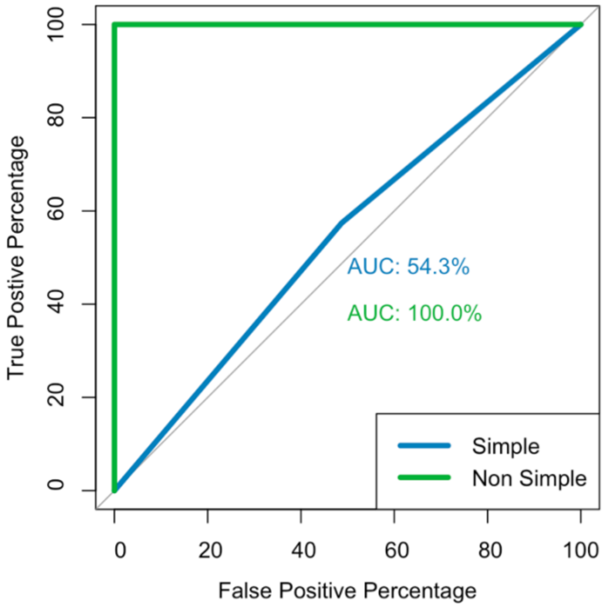
There are 19 observations that are non-smokers and do not have lung cancer, 20 observations that are non-smokers and have lung cancer, 155 observations that are smokers and do not have lung cancer, and 115 observations that are smokers and have lung cancer. This information can be useful to understand the relationship between smoking and lung cancer and to understand if there is a significant association between the two variables. Similarly, we can get a table for other variables like anxiety, fatigue etc and understand if there is a significant association between those variables.

```
logistic_simple <- glm(LUNG_CANCER ~ SMOKING, data=data, family="binomial")
summary(logistic_simple)
```

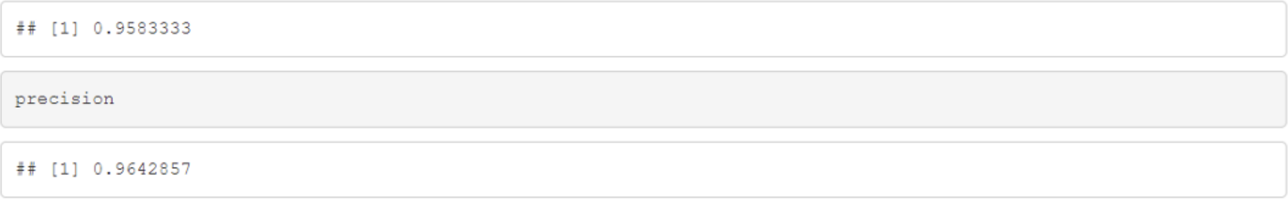
```
##
## Call:
## glm(formula = LUNG_CANCER ~ SMOKING, family = "binomial", data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1046   0.4809   0.4809   0.5663   0.5663
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   2.0990     0.2431   8.636   <2e-16 ***
## SMOKING1     -0.3498     0.3432  -1.019    0.308
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 234.30  on 308  degrees of freedom
## Residual deviance: 233.26  on 307  degrees of freedom
## AIC: 237.26
##
## Number of Fisher Scoring iterations: 4
```

The slope estimate for SMOKING 1 is -0.3498, indicating that for every one-unit increase in SMOKING (i.e., from 0 to 1), the log odds of LUNG_CANCER decreases by 0.3498 units. The p-value for SMOKING 1 is 0.308, which is not statistically significant at the conventional alpha level of 0.05. The residual deviance is 233.26 on 307 degrees of freedom, indicating that the model provides a good fit to the data. The AIC value of 237.26 suggests that this simple logistic model is a good fit for predicting the probability of LUNG CANCER based on SMOKING.

```
legend("bottomright", legend=c("Simple", "Non Simple"), col=c("#377eb8", "#4daf4a"), lwd=4)
```



The simple line on the graph is generated when the model is a random classifier, which means that it predicts classes at random. The non-simple line on the graph is generated when the model is a non-random classifier, which means that it makes informed predictions based on the features of the data.The TPP is the percentage of actual positive instances that the model correctly identifies as positive, while the FPP is the percentage of actual negative instances that the model incorrectly identifies as positive. An AUC of 100% for the non-simple line indicates that it has correctly classified all the samples in the dataset, while an AUC of 54.3% for the simple line indicates that it has not performed much better than randomly assigning class labels. This suggests that the non-simple line is a more reliable classifier than the simple line for this particular dataset. This means that the non-simple model captures more of the variation in the data and provides better discrimination between the two classes. Therefore, the goal to accurately classify data, the non-simple model would be preferred over the simple model.



```
## [1] 0.9583333
```

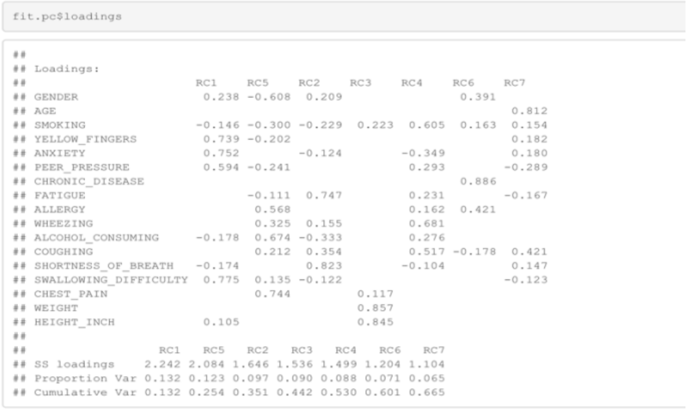
```
precision
```

```
## [1] 0.9642857
```

The accuracy of the model is 95% and it is 96% precise.

EFA

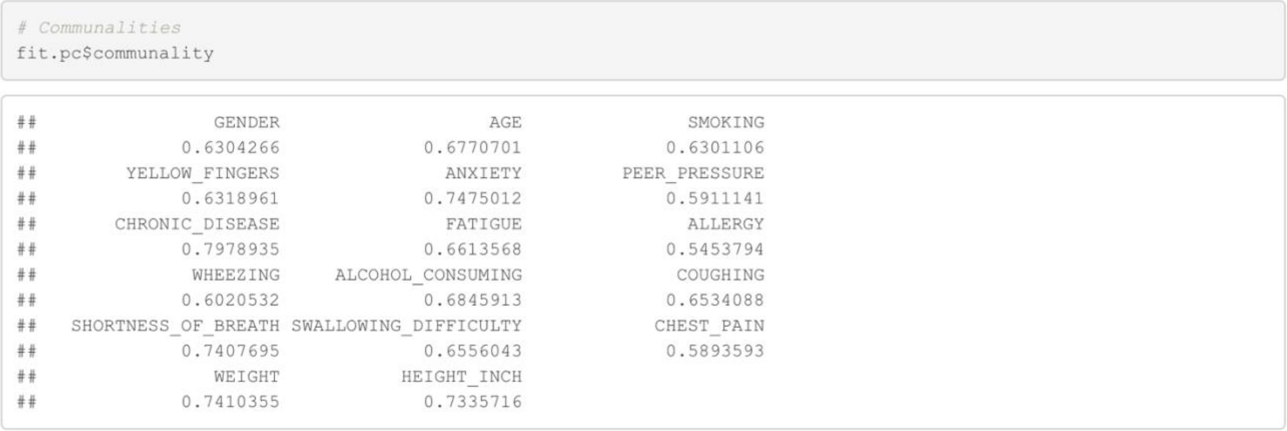
Q) What are some of the underlying factors that cause lung cancer?



```
fit.pc$loadings
```

```
##
## Loadings:
##                       RC1    RC5    RC2    RC3    RC4    RC6    RC7   
## GENDER                 0.238 -0.608  0.209                0.391       
## AGE                                                              0.812
## SMOKING               -0.146 -0.300 -0.229  0.223  0.605  0.163  0.154
## YELLOW_FINGERS         0.739 -0.202                              0.182
## ANXIETY                0.752        -0.124        -0.349         0.180
## PEER_PRESSURE          0.594 -0.241                0.293        -0.289
## CHRONIC_DISEASE                                          0.886       
## FATIGUE                      -0.111  0.747         0.231        -0.167
## ALLERGY                       0.568                0.162  0.421       
## WHEEZING                      0.325  0.155         0.681              
## ALCOHOL_CONSUMING     -0.178  0.674 -0.333         0.276              
## COUGHING                      0.212  0.354         0.517 -0.178  0.421
## SHORTNESS_OF_BREATH   -0.174         0.823        -0.104         0.147
## SWALLOWING_DIFFICULTY  0.775  0.135 -0.122                      -0.123
## CHEST_PAIN                    0.744         0.117                     
## WEIGHT                                      0.857                     
## HEIGHT_INCH            0.105                0.845                     
##
##                  RC1   RC5   RC2   RC3   RC4   RC6   RC7
## SS loadings    2.242 2.084 1.646 1.536 1.499 1.204 1.104
## Proportion Var 0.132 0.123 0.097 0.090 0.088 0.071 0.065
## Cumulative Var 0.132 0.254 0.351 0.442 0.530 0.601 0.665
```

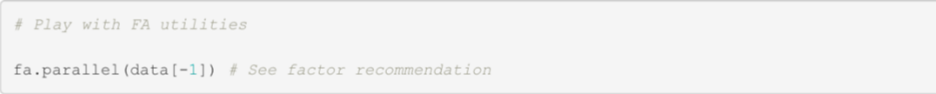
The loadings table shows how strongly each item is associated with each factor, with larger absolute values indicating stronger associations. For example, in the first row we see that GENDER has a loading of 0.238 on the first factor (RC1), -0.608 on the fifth factor (RC5), 0.209 on the second factor (RC2), and no significant loading on the other factors.



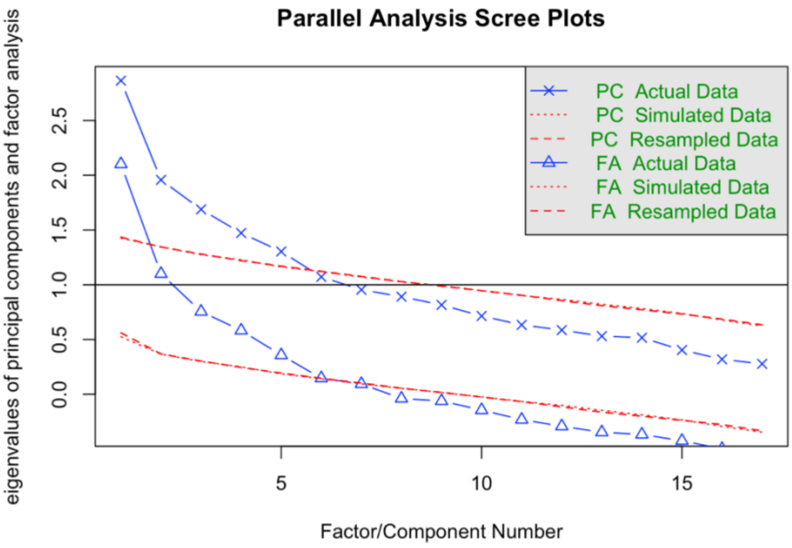
```
# Communalities
fit.pc$communality
```

```
##                GENDER                   AGE               SMOKING 
##             0.6304266             0.6770701             0.6301106 
##        YELLOW_FINGERS               ANXIETY         PEER_PRESSURE 
##             0.6318961             0.7475012             0.5911141 
##       CHRONIC_DISEASE               FATIGUE               ALLERGY 
##             0.7978935             0.6613568             0.5453794 
##              WHEEZING     ALCOHOL_CONSUMING              COUGHING 
##             0.6020532             0.6845913             0.6534088 
##   SHORTNESS_OF_BREATH SWALLOWING_DIFFICULTY            CHEST_PAIN 
##             0.7407695             0.6556043             0.5893593 
##                WEIGHT           HEIGHT_INCH 
##             0.7410355             0.7335716
```

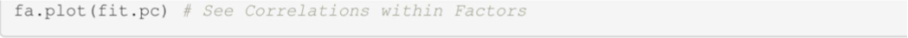
The variable GENDER, the communality estimate is 0.6304266. This means that 63.04% of the variance in GENDER is accounted for by the factors in the analysis. Similarly, for AGE, the communality estimate is 0.6770701, which means that 67.71% of the variance in AGE is accounted for by the factors.



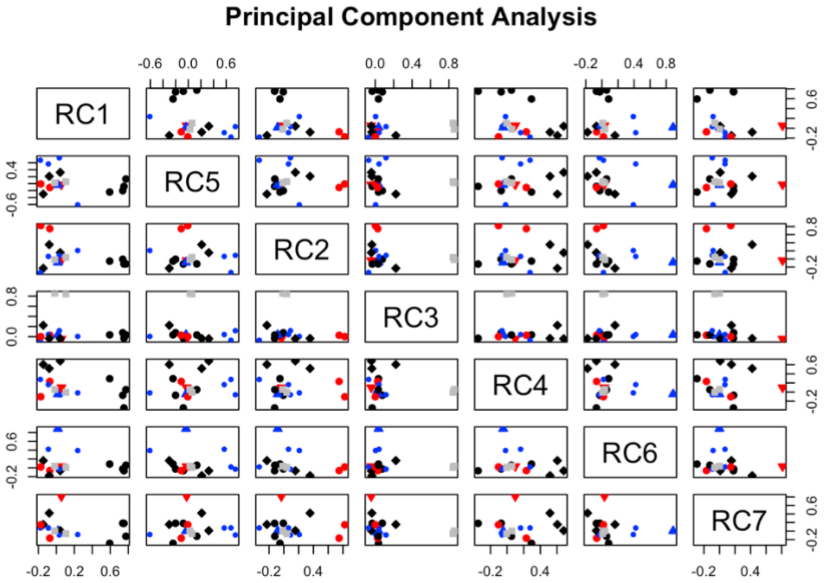
```
# Play with FA utilities

fa.parallel(data[-1]) # See factor recommendation
```

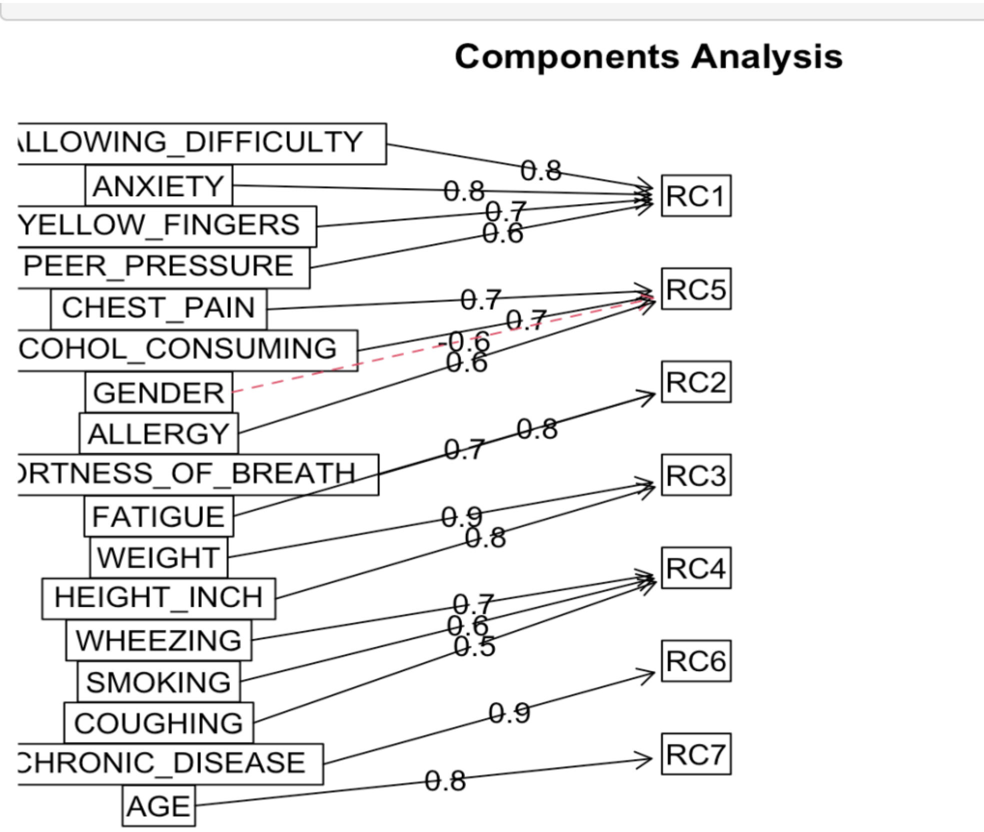


Parallel analysis in scree plots is used to help determine the optimal number of factors/components to retain in a factor analysis by comparing the eigenvalues obtained from the data to those obtained from randomly generated data sets. Here the ideal number of factors to be considered are 7.



```
fa.plot(fit.pc) # See Correlations within Factors
```





Some of the underlying factors that show symptoms of lung cancer are difficulty in swallowing, anxiety, yellow fingers, chest pain, alcohol consumption, allergy, shortness of breath and fatigue.

PCA

Q) Can we identify the most important risk factors for lung cancer? And what is the relationship between the factors?

```
summary(cancer_pca)
```

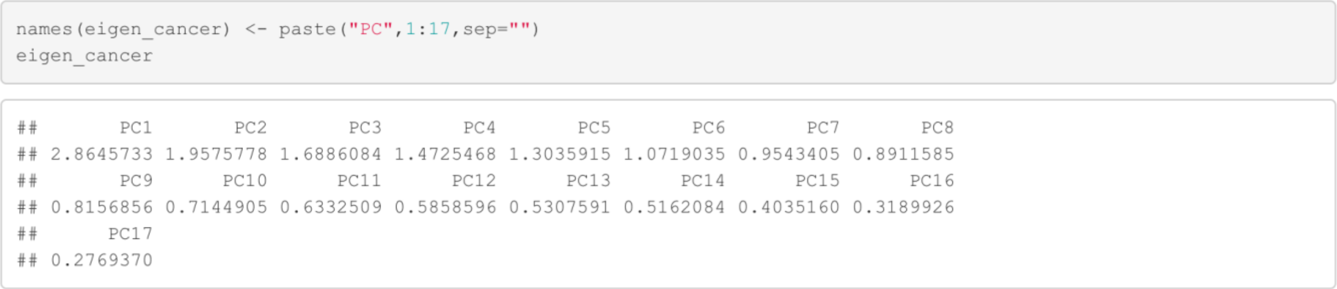
```
## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     1.6925 1.3991 1.29946 1.21349 1.14175 1.03533 0.97690
## Proportion of Variance 0.1685 0.1152 0.09933 0.08662 0.07668 0.06305 0.05614
## Cumulative Proportion  0.1685 0.2837 0.38299 0.46961 0.54629 0.60934 0.66548
##                            PC8     PC9    PC10    PC11    PC12    PC13    PC14
## Standard deviation     0.94401 0.90315 0.84528 0.79577 0.76541 0.72853 0.71848
## Proportion of Variance 0.05242 0.04798 0.04203 0.03725 0.03446 0.03122 0.03037
## Cumulative Proportion  0.71790 0.76588 0.80791 0.84516 0.87962 0.91084 0.94121
##                           PC15    PC16    PC17
## Standard deviation     0.63523 0.56479 0.52625
## Proportion of Variance 0.02374 0.01876 0.01629
## Cumulative Proportion  0.96495 0.98371 1.00000
```

Standard deviation: The standard deviation indicates how much of the variance in the original data is captured by each principal component. The first principal component (PC1) has the highest standard deviation of 1.6925, which means it accounts for the most variation in the original data. As we move down the list of principal components, the standard deviation decreases, indicating that each subsequent component captures less and less of the overall variation.

Proportion of variance: The proportion of variance indicates the amount of variance in the original data that is explained by each principal component. For example, the first principal component (PC1) explains 16.85% of the total variance in the data, while the second component (PC2) explains an additional 11.52% of the variance. As we move down the list of principal components, the proportion of variance explained by each component tends to decrease.

Cumulative proportion: The cumulative proportion indicates the total amount of variance in the data that is explained by each principal component and all of the preceding components. For example, the first principal component (PC1) captures

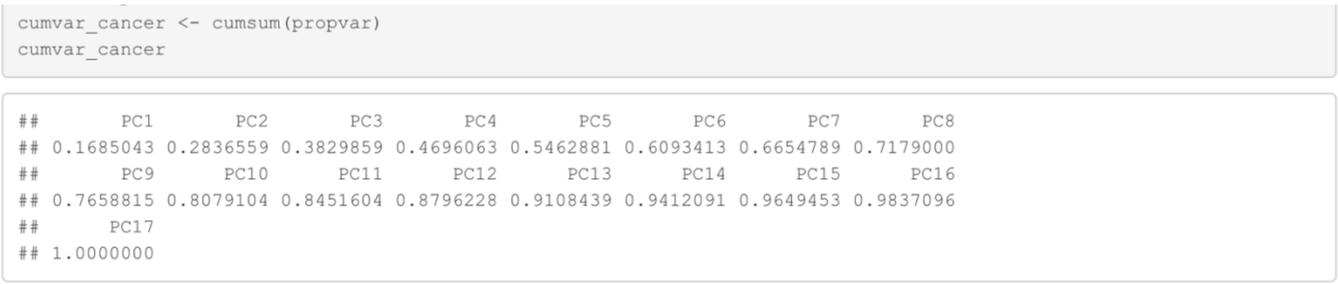
16.85% of the variance in the data, while PC1 and PC2 together capture 28.37% of the variance. The cumulative proportion can be useful in determining how many principal components to retain for further analysis.



```
names(eigen_cancer) <- paste("PC",1:17,sep="")
eigen_cancer
```

```
##       PC1       PC2       PC3       PC4       PC5       PC6       PC7       PC8 
## 2.8645733 1.9575778 1.6886084 1.4725468 1.3035915 1.0719035 0.9543405 0.8911585 
##       PC9      PC10      PC11      PC12      PC13      PC14      PC15      PC16 
## 0.8156856 0.7144905 0.6332509 0.5858596 0.5307591 0.5162084 0.4035160 0.3189926 
##      PC17 
## 0.2769370
```

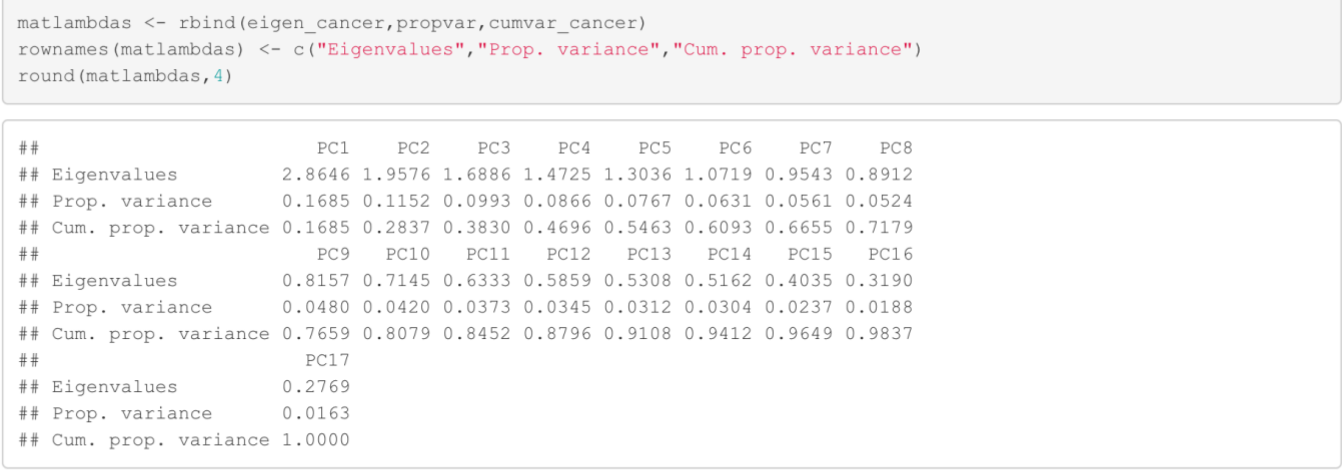
The number 2.8645733 in the first row and first column means that the first principal component explains 2.8645733 units of variance in the data. The number 0.2769370 in the last row means that the 17th principal component explains 0.2769370 units of variance in the data. We can see that the first 7 principal components explain a cumulative proportion of variance of 0.66548, so we might choose to retain those 7 components.



```
cumvar_cancer <- cumsum(propvar)
cumvar_cancer
```

```
##       PC1       PC2       PC3       PC4       PC5       PC6       PC7       PC8 
## 0.1685043 0.2836559 0.3829859 0.4696063 0.5462881 0.6093413 0.6654789 0.7179000 
##       PC9      PC10      PC11      PC12      PC13      PC14      PC15      PC16 
## 0.7658815 0.8079104 0.8451604 0.8796228 0.9108439 0.9412091 0.9649453 0.9837096 
##      PC17 
## 1.0000000
```

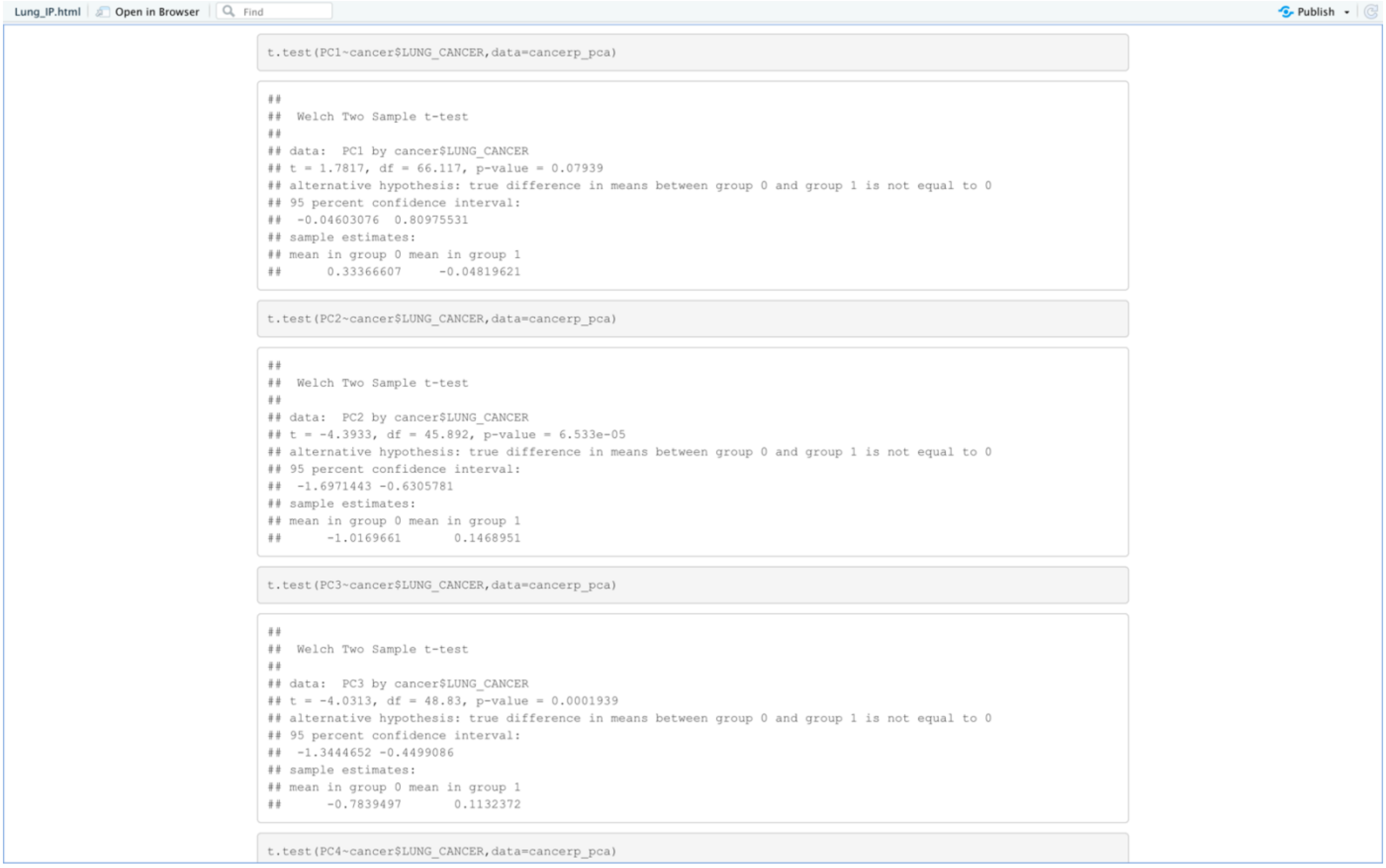
Here, PC1 alone explains 16.85% of the variance, while the first two principal components combined (PC1 and PC2) explain 28.37% of the variance, and so on. This cumulative proportion of variance can be useful for determining how many principal components to include in further analysis. We can see that the first 7 principal components explain around 71.79% of the total variance.



```
matlambdas <- rbind(eigen_cancer,propvar,cumvar_cancer)
rownames(matlambdas) <- c("Eigenvalues","Prop. variance","Cum. prop. variance")
round(matlambdas,4)
```

```
##                        PC1    PC2    PC3    PC4    PC5    PC6    PC7    PC8
## Eigenvalues         2.8646 1.9576 1.6886 1.4725 1.3036 1.0719 0.9543 0.8912
## Prop. variance      0.1685 0.1152 0.0993 0.0866 0.0767 0.0631 0.0561 0.0524
## Cum. prop. variance 0.1685 0.2837 0.3830 0.4696 0.5463 0.6093 0.6655 0.7179
##                        PC9   PC10   PC11   PC12   PC13   PC14   PC15   PC16
## Eigenvalues         0.8157 0.7145 0.6333 0.5859 0.5308 0.5162 0.4035 0.3190
## Prop. variance      0.0480 0.0420 0.0373 0.0345 0.0312 0.0304 0.0237 0.0188
## Cum. prop. variance 0.7659 0.8079 0.8452 0.8796 0.9108 0.9412 0.9649 0.9837
##                       PC17
## Eigenvalues         0.2769
## Prop. variance      0.0163
## Cum. prop. variance 1.0000
```

Here, the first PC has an eigenvalue of 2.8646, which explains 16.85% of the total variance in the data. The first two PCs combined explain 28.37% of the variance, and the first seven PCs combined explain 71.79% of the variance.



```
t.test(PC1~cancer$LUNG_CANCER,data=cancerp_pca)
```

```
##
##  Welch Two Sample t-test
##
## data:  PC1 by cancer$LUNG_CANCER
## t = 1.7817, df = 66.117, p-value = 0.07939
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.04603076  0.80975531
## sample estimates:
## mean in group 0 mean in group 1
##      0.33366607     -0.04819621
```

```
t.test(PC2~cancer$LUNG_CANCER,data=cancerp_pca)
```

```
##
##  Welch Two Sample t-test
##
## data:  PC2 by cancer$LUNG_CANCER
## t = -4.3933, df = 45.892, p-value = 6.533e-05
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -1.6971443 -0.6305781
## sample estimates:
## mean in group 0 mean in group 1
##      -1.0169661       0.1468951
```

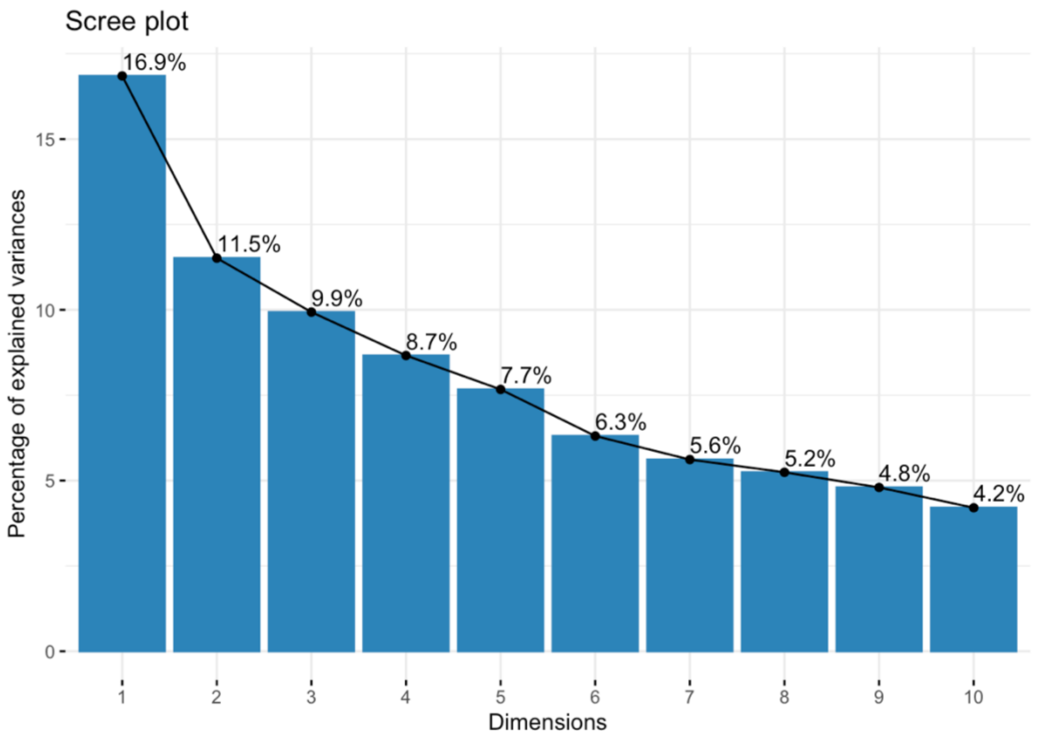
```
t.test(PC3~cancer$LUNG_CANCER,data=cancerp_pca)
```

```
##
##  Welch Two Sample t-test
##
## data:  PC3 by cancer$LUNG_CANCER
## t = -4.0313, df = 48.83, p-value = 0.0001939
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -1.3444652 -0.4499086
## sample estimates:
## mean in group 0 mean in group 1
##      -0.7839497       0.1132372
```

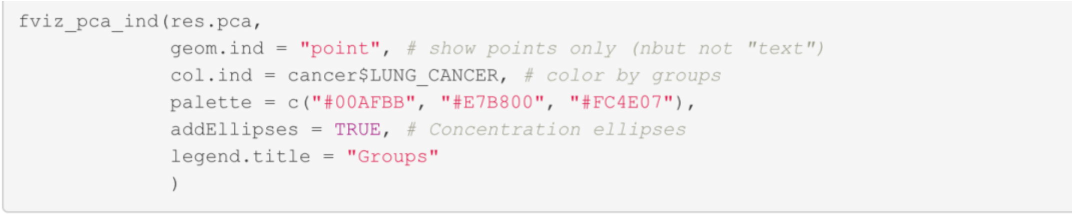
```
t.test(PC4~cancer$LUNG_CANCER,data=cancerp_pca)
```

The output shows that the p-value is 0.07939, which is above the significance level of 0.05, indicating that we cannot reject the null hypothesis of no difference in means between the two groups at the 5% significance level for PC1. Similarly we can do this for other PC's as well.

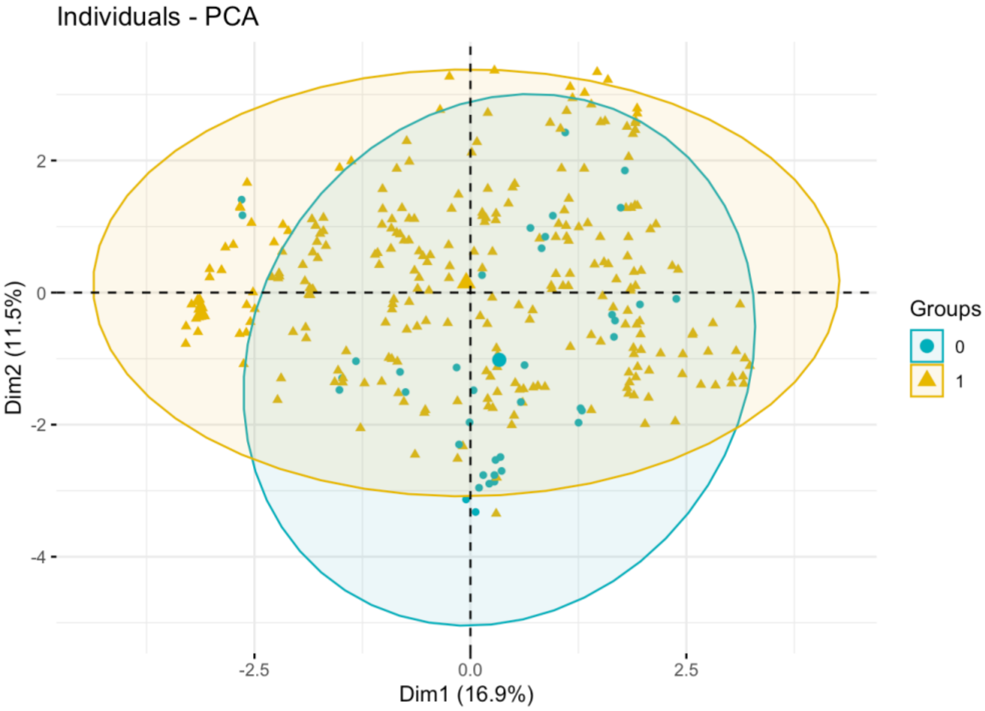
```
fviz_eig(cancer_pca, addlabels = TRUE)
```



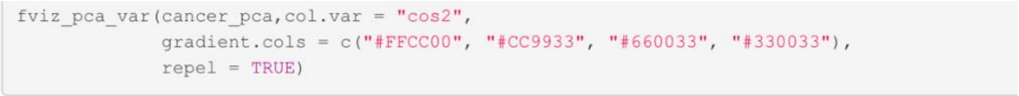
In this scree plot, we will consider the first 7 PC's as it explains almost 70% of the variance in the dataset.



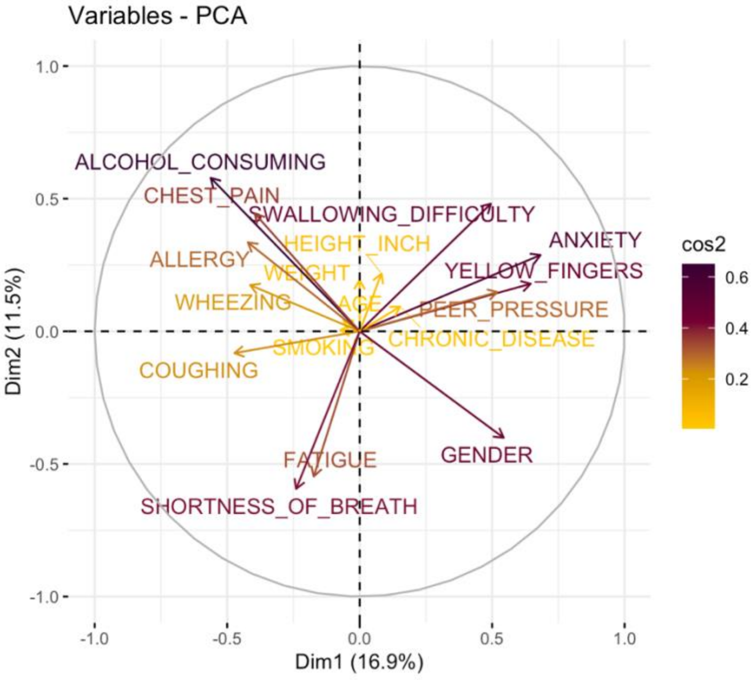
```
fviz_pca_ind(res.pca,
             geom.ind = "point", # show points only (nbut not "text")
             col.ind = cancer$LUNG_CANCER, # color by groups
             palette = c("#00AFBB", "#E7B800", "#FC4E07"),
             addEllipses = TRUE, # Concentration ellipses
             legend.title = "Groups"
             )
```



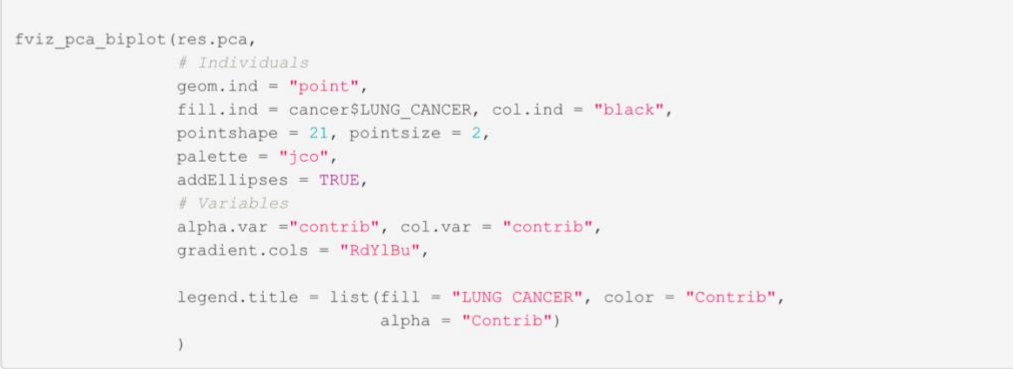
Each ellipse represents the range of scores for each group in the first two principal components. The center of the ellipse represents the centroid of each group in the first two principal components. If the ellipses of two groups overlap significantly, it means that they have similar scores in the first two principal components, and thus are not easily distinguishable based on those components. On the other hand, if the ellipses of two groups do not overlap, it means that they have different scores in the first two principal components, and thus are easily distinguishable based on those components. In the graph, some part of the ellipse is overlapping while some part isn't. So, some of the scores for PC1 and PC2 are same while some are different.



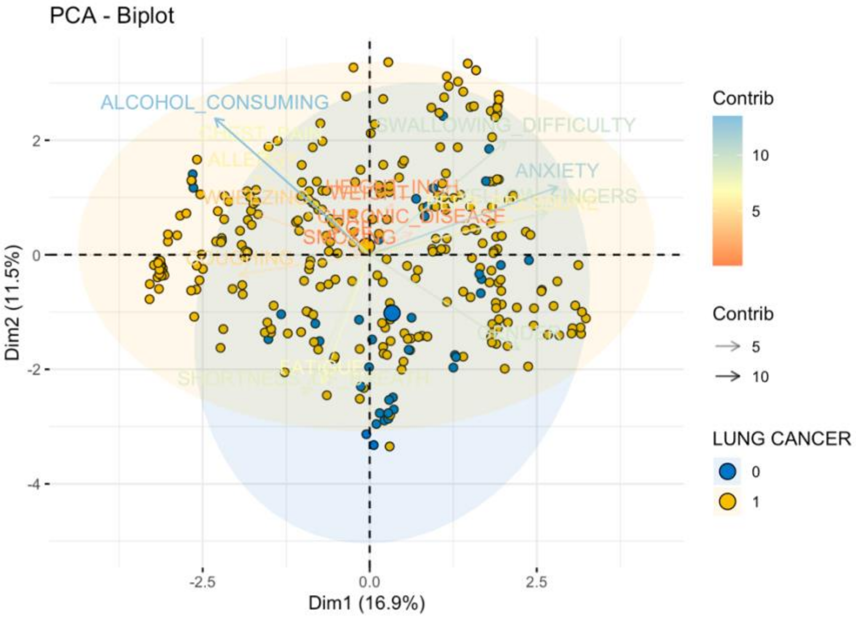
```
fviz_pca_var(cancer_pca,col.var = "cos2",
             gradient.cols = c("#FFCC00", "#CC9933", "#660033", "#330033"),
             repel = TRUE)
```



The graph is a visualization of the principal components of a PCA analysis. Each point on the plot represents a variable in the original dataset, and the position of the point represents the contribution of the variable to the principal components. The size of the point represents the variability explained by the variable, and the color represents the squared cosine of the variable, which indicates the correlation between the variable and the principal component. The gradient of colors indicates the strength of the correlation, with darker colors indicating stronger correlation. From this graph, we can identify which variables are most important in explaining the variation in the data, and which variables are most strongly correlated with each principal component. Here, alcohol consumption, swallowing difficulty, anxiety, yellow fingers, chest pain, shortness of breath are some of the variables

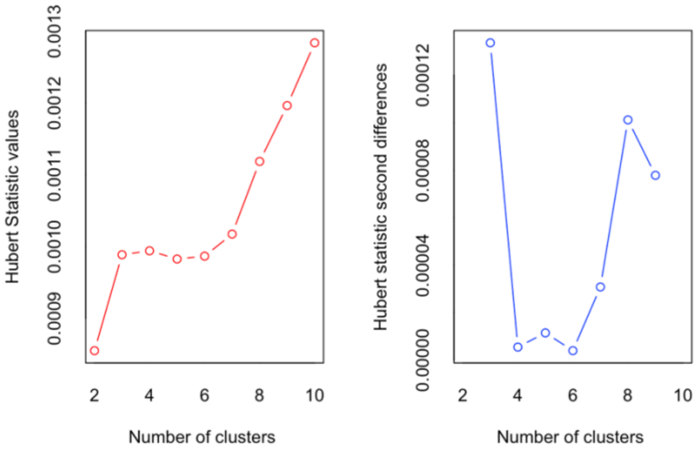


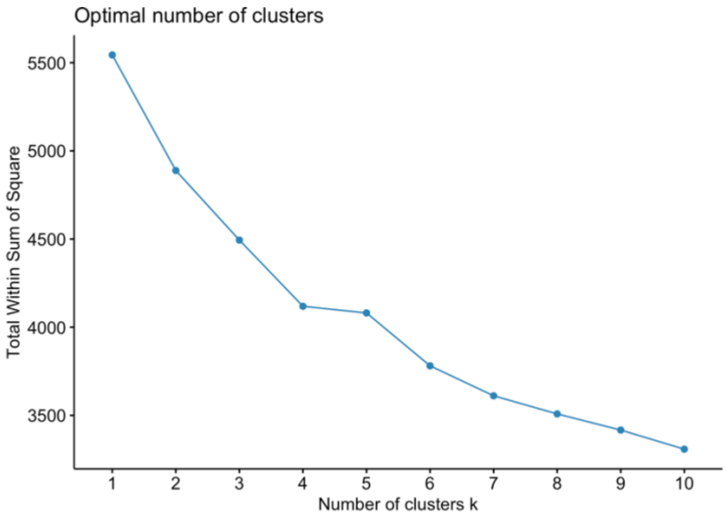
```
fviz_pca_biplot(res.pca,
                # Individuals
                geom.ind = "point",
                fill.ind = cancer$LUNG_CANCER, col.ind = "black",
                pointshape = 21, pointsize = 2,
                palette = "jco",
                addEllipses = TRUE,
                # Variables
                alpha.var ="contrib", col.var = "contrib",
                gradient.cols = "RdYlBu",

                legend.title = list(fill = "LUNG CANCER", color = "Contrib",
                                    alpha = "Contrib")
                )
```

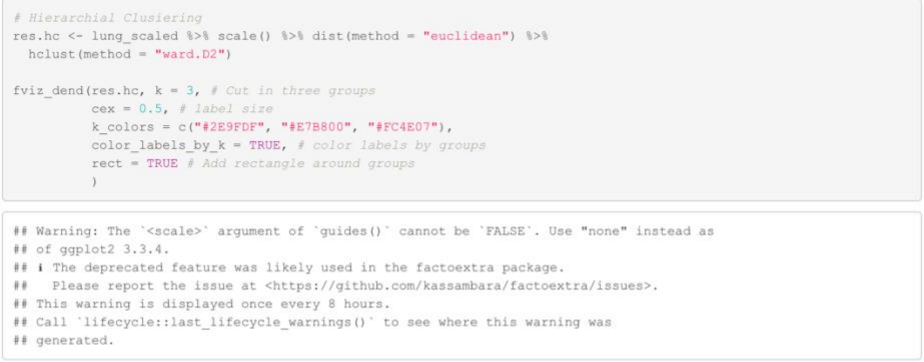


Here we can understand that the closer the variables are to a principal component, the stronger the relationship. Variables that are close together on the biplot are positively correlated, while variables that are far apart are negatively correlated. Here, peer pressure, anxiety, swallowing difficulty, yellow fingers are strongly correlated with each other. Also, chest pain, allergy, wheezing and coughing are correlated to each other. These can be considered as some of the important factors that affect lung cancer.

Cluster Analysis:

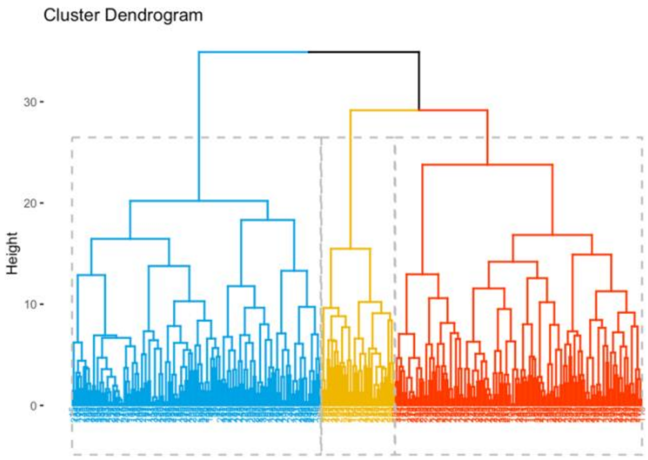




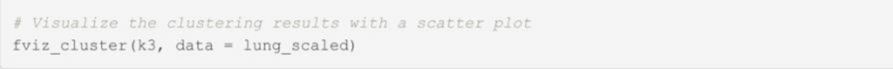


```
# Hierarchial Clusiering
res.hc <- lung_scaled %>% scale() %>% dist(method = "euclidean") %>%
  hclust(method = "ward.D2")

fviz_dend(res.hc, k = 3, # Cut in three groups
          cex = 0.5, # label size
          k_colors = c("#2E9FDF", "#E7B800", "#FC4E07"),
          color_labels_by_k = TRUE, # color labels by groups
          rect = TRUE # Add rectangle around groups
          )
```

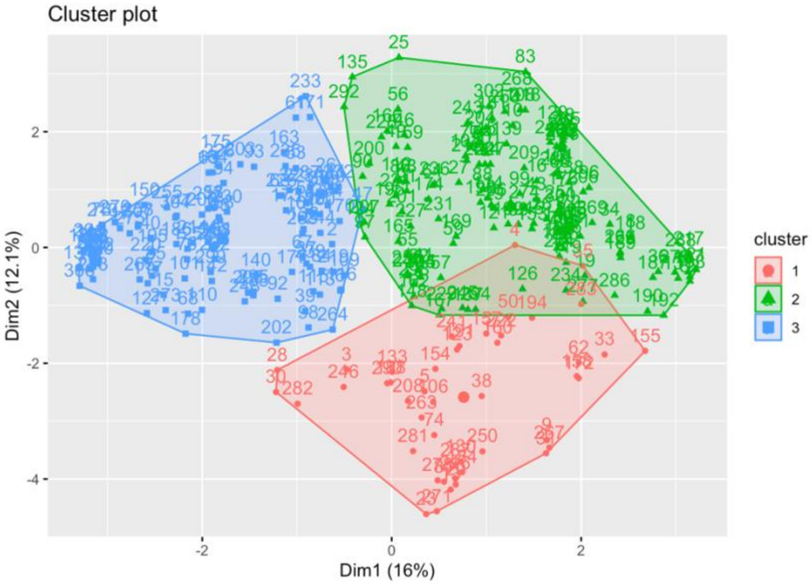
```
## Warning: The `<scale>` argument of `guides()` cannot be `FALSE`. Use "none" instead as
## of ggplot2 3.3.4.
## ℹ The deprecated feature was likely used in the factoextra package.
##   Please report the issue at <https://github.com/kassambara/factoextra/issues>.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



In the plot of Hubert index, we seek a significant knee that corresponds to a significant increase of the value of the measure i.e the significant peak in Hubert index second differences plot. Here we take 3 clusters from analyzing the graphs. The dendrogram also suggest that we should take 3 clusters. The dotted line represents the break point.



```
# Visualize the clustering results with a scatter plot
fviz_cluster(k3, data = lung_scaled)
```



Data Dictionary:

Attribute information:

1) Gender: 1(male), 0(female)
2) Age: Age of the people
3) Smoking: YES=1 , NO=0.
4) Yellow fingers: YES=1 , NO=0.
5) Anxiety: YES=1 , NO=0.
6) Peer_pressure: YES=1 , NO=0.
7) Chronic_Disease: YES=1 , NO=0.
8) Fatigue: YES=1 , NO=0.
9) Allergy: YES=1 , NO=0.
10) Wheezing: YES=1, NO=0.
11) Alcohol: YES=1 , NO=0.
12) Coughing: YES=1, NO=0.
13) Shortness_of_Breath: YES=1 , NO=0.
14) Swallowing_Difficulty: YES=1 , NO=0.
15) Chest_pain: YES=1, NO=0.
16) Lung_Cancer: 1(yes) , 0(no)
17) Weight: Weight of the people in lb
18) Height_inch: Height of the people in inches