Lung Cancer Prediction

Healy Li

Introduction

The data using in this analysis is from kaggle and it is originally collected from a online lung cancer predict system. This project is trying to fit this data with a model that can best predict the outcome: incident of lung cancer, by using the given variable including age, smoking, gender, etc.

loading data

Loading the data and make sure there is no missing values

```
## [1] 0
```

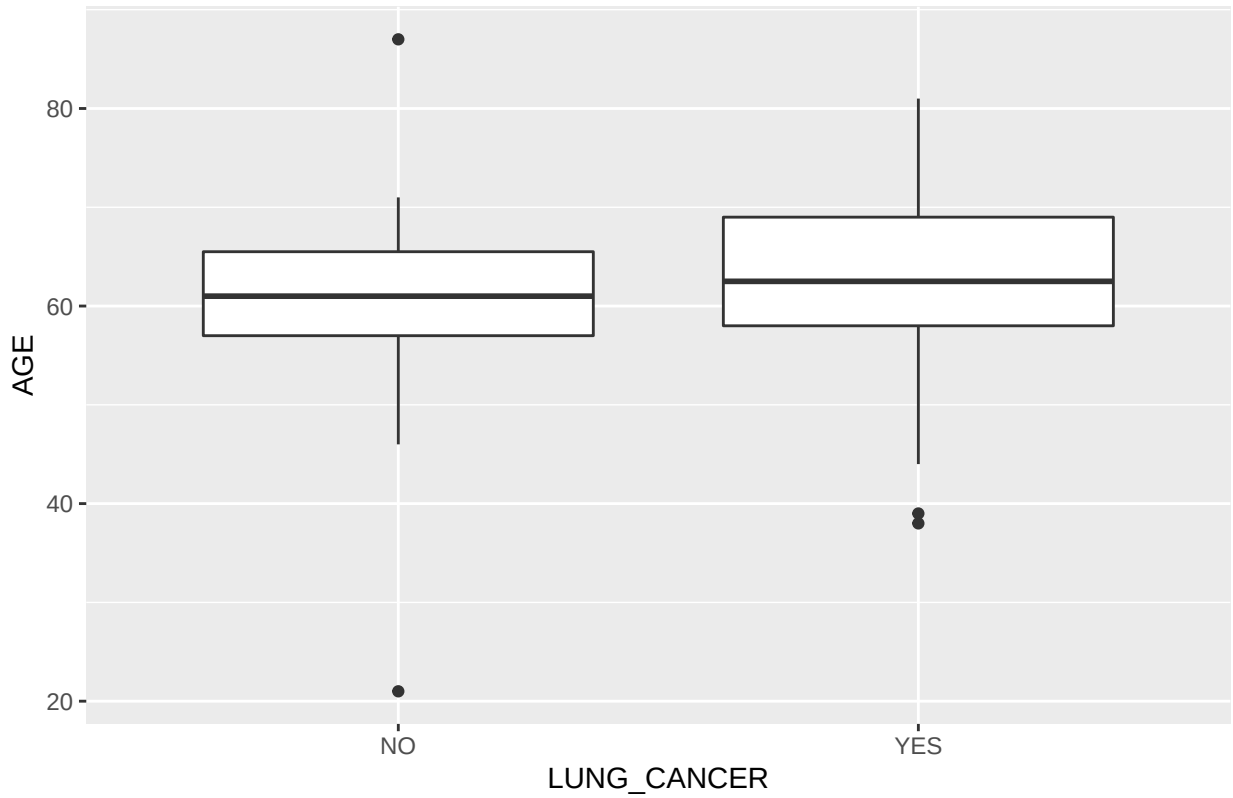
investigate the variables

Looking into the levels of different variables. It shows that beside the variable `AGE` is discrete numeric variable, others are all binary variables with only two levels, notice that the variable `GENDER` is in the type of characteristic. Hence I decided to investigate this two variables at first. there is an age difference between the groups with and without lung cancer but not very large (Graph 1). The two-sample t-test shows that the mean age of this two groups are not significantly different (p-value = 0.18), which means age and the incident of lung cancer is not associated in this data set. The variable `GENDER` and the outcome variable `LUNG_CANCER` are both binary variable hence chi-square test applied. The result shows that they are not significantly associated (p-value = 0.31). There is also no significant association between age and gender (Graph 2). As for the other binary variables, all of them are not significantly correlated with each other other than `ANXIETY` and `YELLOW_FINGERS`. They have a slightly positive correlation and will be further investigate when fitting model (Graph 3 & Table 1).

```
## $GENDER
## [1] "F" "M"
##
## $AGE
##  [1] "21" "38" "39" "44" "46" "47" "48" "49" "51" "52" "53" "54" "55" "56" "57"
## [16] "58" "59" "60" "61" "62" "63" "64" "65" "66" "67" "68" "69" "70" "71" "72"
## [31] "73" "74" "75" "76" "77" "78" "79" "81" "87"
##
## $SMOKING
## [1] "1" "2"
##
## $YELLOW_FINGERS
## [1] "1" "2"
##
```

```
## $ANXIETY
## [1] "1" "2"
##
## $PEER_PRESSURE
## [1] "1" "2"
##
## $CHRONIC.DISEASE
## [1] "1" "2"
##
## $FATIGUE
## [1] "1" "2"
##
## $ALLERGY
## [1] "1" "2"
##
## $WHEEZING
## [1] "1" "2"
##
## $ALCOHOL.CONSUMING
## [1] "1" "2"
##
## $COUGHING
## [1] "1" "2"
##
## $SHORTNESS.OF.BREATH
## [1] "1" "2"
##
## $SWALLOWING.DIFFICULTY
## [1] "1" "2"
##
## $CHEST.PAIN
## [1] "1" "2"
##
## $LUNG_CANCER
## [1] "NO"  "YES"
```

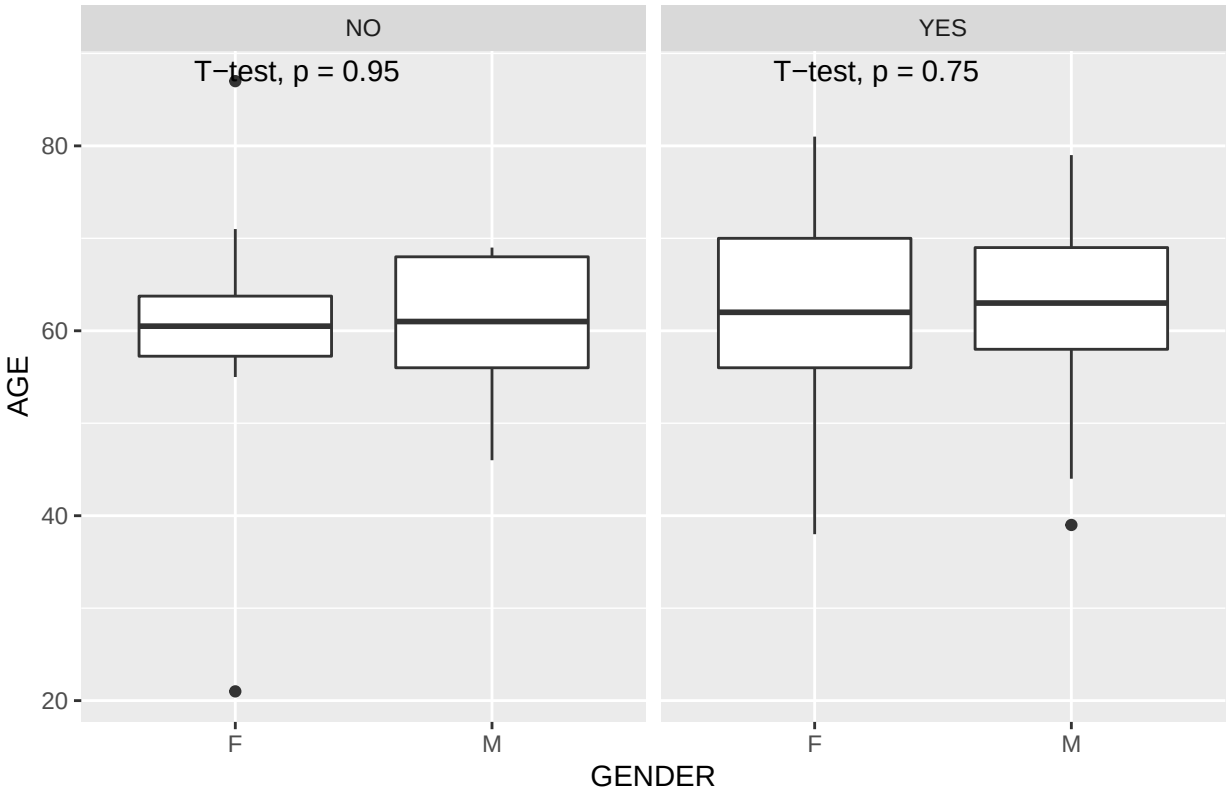
Graph 1




```
##
##  Welch Two Sample t-test
##
## data:  lung$AGE[lung$LUNG_CANCER == "YES"] and lung$AGE[lung$LUNG_CANCER == "NO"]
## t = 1.3662, df = 45.822, p-value = 0.1785
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.045597  5.462121
## sample estimates:
## mean of x mean of y
##  62.95185  60.74359
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(lung$GENDER, lung$LUNG_CANCER)
## X-squared = 1.0215, df = 1, p-value = 0.3122
```

Graph 2



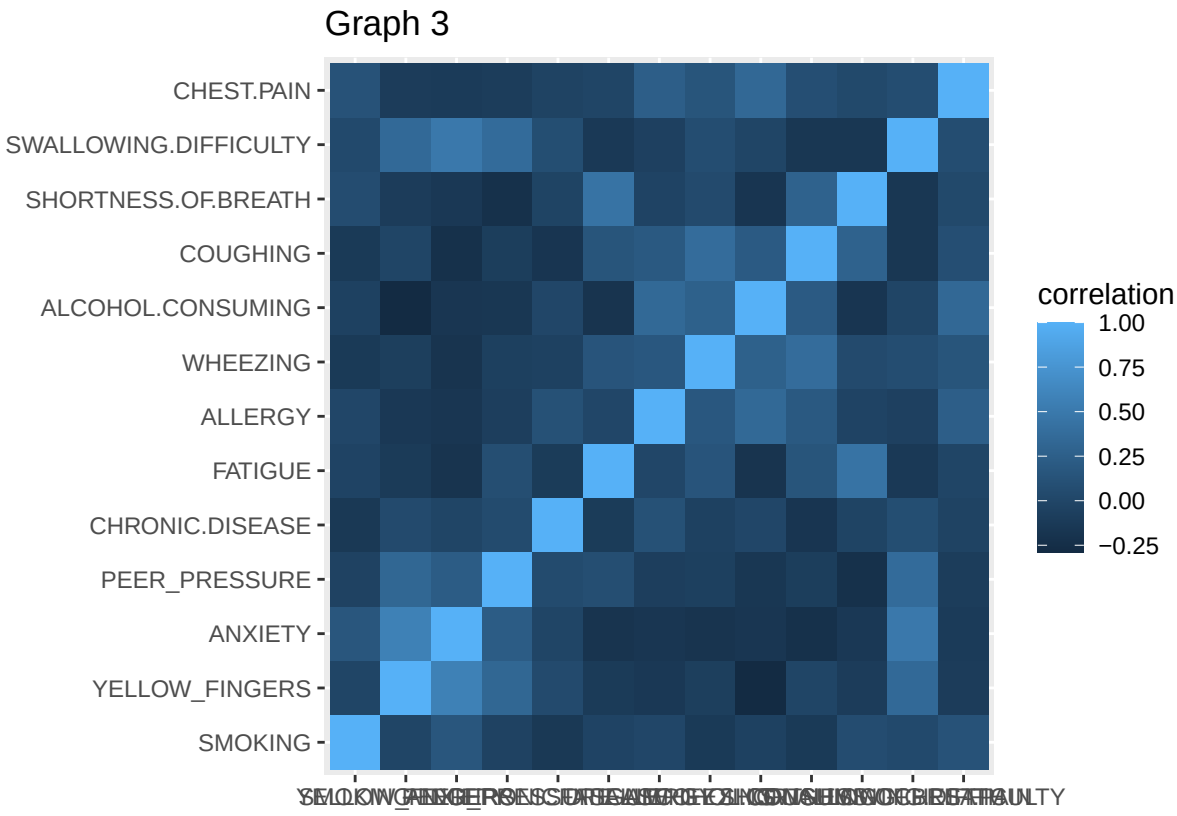


Table 1: correlated variables

Var1	Var2	value
ANXIETY	YELLOW_FINGERS	0.5658293
YELLOW_FINGERS	ANXIETY	0.5658293

Now look into the association between `LUNG_CANCER` and the binary variables. According to Graph 4, `SHORTNESS.OF.BREATH` and `SMOKING` is not a possible strong predictor of incident of lung cancer. Chi-square test additionally told that `CHRONIC.DISEASE` is not significantly associated with `LUNG_CANCER` as well (p-value = 0.07)

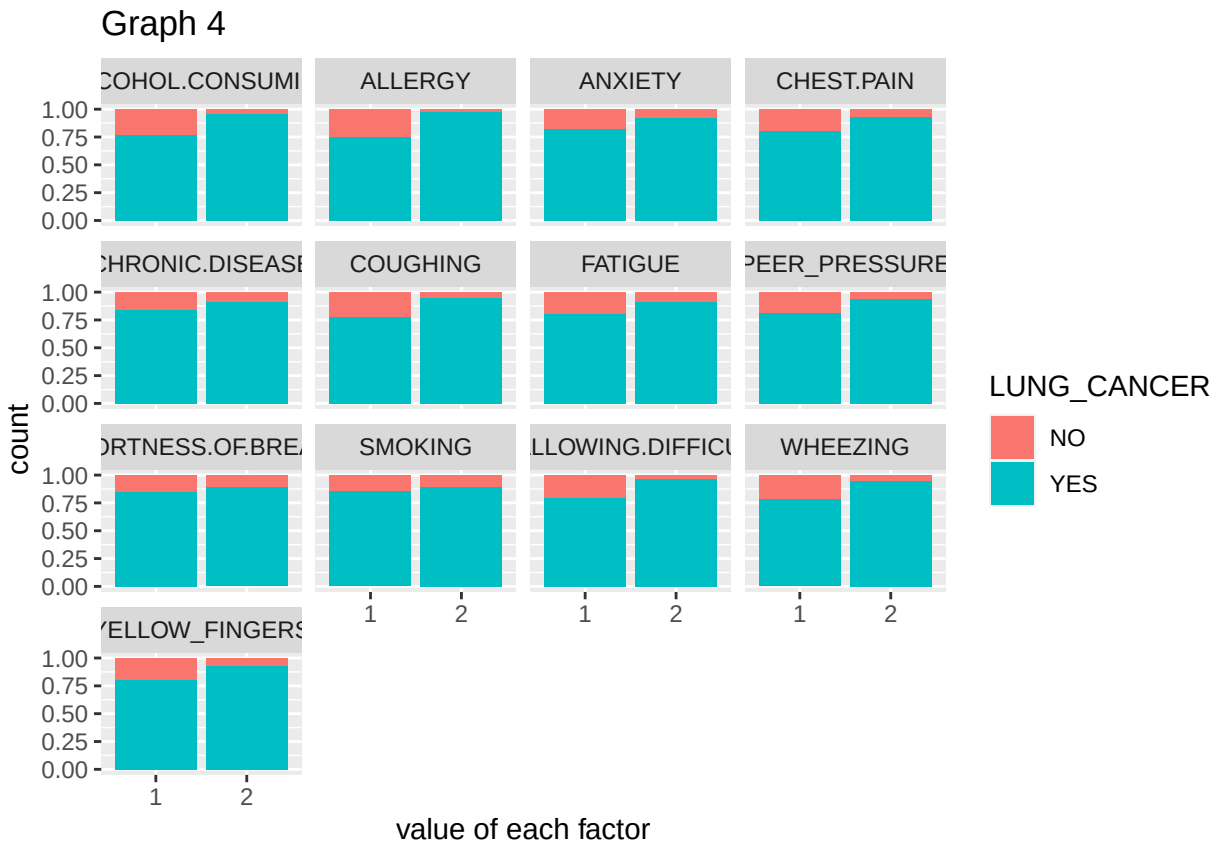


Table 2: Association Between Lung Cancer and Factor

	p.value	correlation
alcohol	9.60655896246563e-07	correlated
allergy	2.28142192226923e-08	correlated
anxiety	0.0174714072968476	correlated
chest.pain	0.00149627458178166	correlated
chronic.disease	0.0754077231120523	not-correlated
coughing	2.71712298549256e-05	correlated
fatigue	0.0136635630316486	correlated
peer.pressure	0.00190220081881552	correlated
swallowing.difficulty	1.11281404355328e-05	correlated
wheezing	2.55505496690905e-05	correlated
yellow.fingers	0.00257265867676361	correlated

We firstly tried the logistic regression model. All the possible predictor are included and the fitted model has a AIC of 76.942, which is quiet small. However the p-value of the factors `WHEEZING` and `CHEST.PAIN` is not significant enough (Table 3). Hence they are excluded in the following models. The result shows that the model with `CHEST.PAIN` excluded has the smallest AIC (Table 4). Therefore, we choose it as our final logistic regression model, and the accuracy of its prediction compare to the original data is 0.9016.

Table 3: P-value for Different Factors

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.7968370	0.1346505	-5.917816	0.0000000
ALLERGY	0.1638732	0.0335155	4.889482	0.0000017
ALCOHOL.CONSUMING	0.1806641	0.0378190	4.777072	0.0000028
SWALLOWING.DIFFICULTY	0.1052813	0.0387361	2.717913	0.0069542
WHEEZING	0.0497221	0.0353614	1.406114	0.1607322
COUGHING	0.0994768	0.0353449	2.814458	0.0052112
CHEST.PAIN	0.0358715	0.0333804	1.074628	0.2834108
PEER_PRESSURE	0.0760099	0.0346217	2.195443	0.0289027
YELLOW_FINGERS	0.1145707	0.0409212	2.799786	0.0054475
FATIGUE	0.1602115	0.0355069	4.512129	0.0000092
ANXIETY	0.0849299	0.0422085	2.012150	0.0451037

Table 4: Comparison of AIC

	AIC
All Factors	76.94203
No Wheezing	76.98540
No Chest.Pain	76.13717
No Both	76.21427

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0  16   6
##          1  23 264
##
##                Accuracy : 0.9061
##                  95% CI : (0.868, 0.9362)
##     No Information Rate : 0.8738
##     P-Value [Acc > NIR] : 0.047640
##
##                   Kappa : 0.477
##
##  Mcnemar's Test P-Value : 0.002967
##
##             Sensitivity : 0.41026
##             Specificity : 0.97778
##          Pos Pred Value : 0.72727
##          Neg Pred Value : 0.91986
##              Prevalence : 0.12621
##          Detection Rate : 0.05178
##    Detection Prevalence : 0.07120
##       Balanced Accuracy : 0.69402
##
##        'Positive' Class : 0
##
```

Other than logistic regression model, other models including Linear Discriminant Analysis (LDA), Classification and Regression Trees (CART), k-Nearest Neighbors (KNN), Support Vector Machines (SVM) with a linear kernel and Random Forest (RF) were also tried with the same predictors using the the logistic regression model. Among them, the SVM model is best fitted (Table 5) and the accuracy of its prediction compare to the original data is 0.8997, which is lower than the accuracy of logistic regression model. Hence we decided that the logistic regression model with predictor `ALLERGY`, `ALCOHOL.CONSUMING`, `SWALLOWING.DIFFICULTY`, `WHEEZING`, `COUGHING`, `PEER_PRESSURE`, `YELLOW_FINGERS`, `FATIGUE`and `ANXIETY` is the best model for predicting lung cancer.

	Accuracy	Kappa
cart	0.8956522	0.5860828
lda	0.8571429	0.4707620
knn	0.8928571	0.4794733
rf	0.8990826	0.6462083
svm	0.9310345	0.6534727

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0  15   7
##          1  24 263
##
##                Accuracy : 0.8997
##                  95% CI : (0.8606, 0.9308)
##     No Information Rate : 0.8738
##     P-Value [Acc > NIR] : 0.096646
##
##                   Kappa : 0.4409
##
##  Mcnemar's Test P-Value : 0.004057
##
##             Sensitivity : 0.38462
##             Specificity : 0.97407
##          Pos Pred Value : 0.68182
##          Neg Pred Value : 0.91638
##              Prevalence : 0.12621
##          Detection Rate : 0.04854
##    Detection Prevalence : 0.07120
##       Balanced Accuracy : 0.67934
##
##        'Positive' Class : 0
##
```