LUNG CANCER ANALYSIS

By Sean Volpi

BACKGROUND

- Study looked at data from 462,000 + people in China who were followed for an average of six years.

- The participants were divided into two groups: those who lived in areas with high levels of air pollution and those who lived in areas with low levels of air pollution.

- The researchers found that the people in the high-pollution group were more likely to develop lung cancer than those in the low-pollution group.

- Also found that the risk was higher in nonsmokers than smokers, and that the risk increased with age.

- While this study does not prove that air pollution causes lung cancer, it does suggest that there may be a link between the two.

ABOUT THE DATA SET

- Sample from study: 1000 Chinese patients all with lung cancer.
- **Predictor variables:**
  - Age
  - Gender
  - Air Pollution
  - Alcohol use
  - Dust allergies
  - Occupational hazards
  - Genetic risk
  - Diet
  - Obesity
  - Smoking
  - Passive smoking

Potential Causes

  - Chest pain
  - Coughing of blood
  - Fatigue
  - Shortness of breath
  - Wheezing
  - Swallowing difficulty
  - Fingernail clubbing

Potential Effects/Symptoms

- **Response variables:**
  - Chronic lung disease severity
  - “Level” (categorical)

ABOUT THE DATA SET

index	Patient Id	Age	Gender	Level
0	P1	33	**1**	**Low**
1	P10	17	**1**	**Medium**
2	P100	35	**1**	**High**
3	P1000	37	**1**	**High**
4	P101	46	**1**	**High**
5	P102	35	**1**	**High**
6	P103	52	**2**	**Low**
7	P104	28	**2**	**Low**
8	P105	35	**2**	**Medium**

ABOUT THE DATA SET

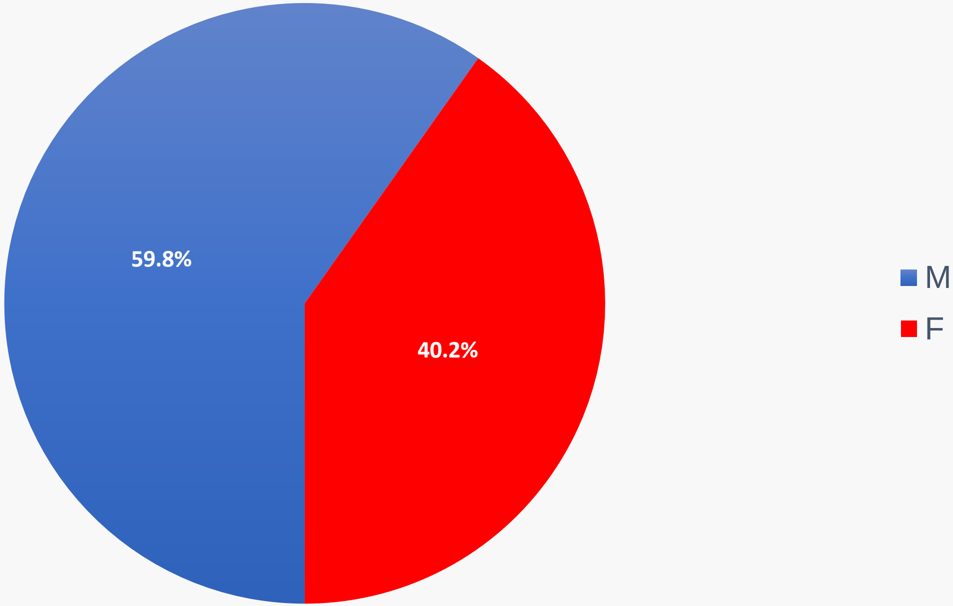
index	Patient Id	Age	Gender	Level
0	P1	33	**M**	**1**
1	P10	17	**M**	**2**
2	P100	35	**M**	**3**
3	P1000	37	**M**	**3**
4	P101	46	**M**	**3**
5	P102	35	**M**	**3**
6	P103	52	**F**	**1**
7	P104	28	**F**	**1**
8	P105	35	**F**	**2**

GOALS

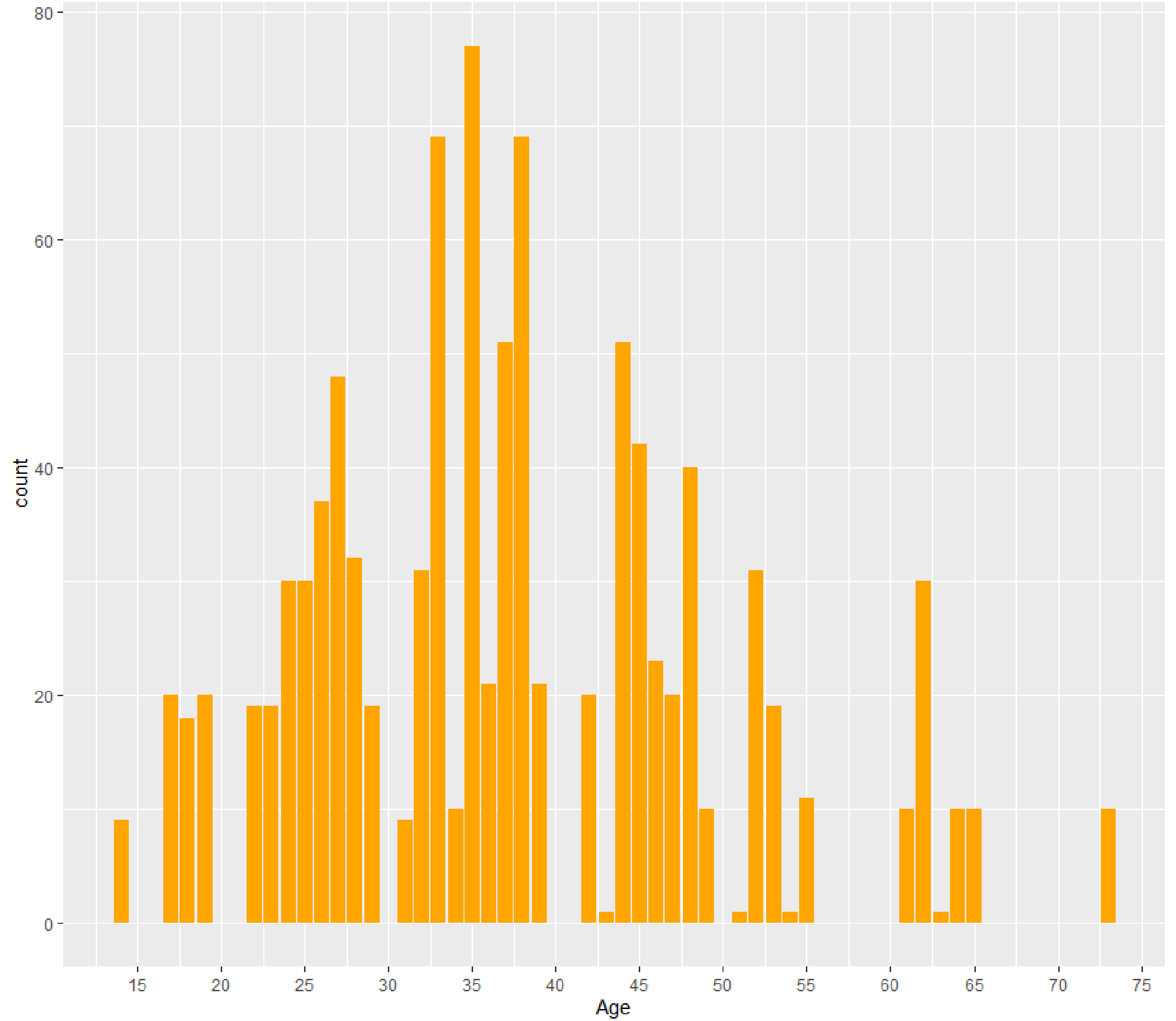
- Observe and identify general relationships between predictor variables and response variable(s).
  - Exploratory analysis.

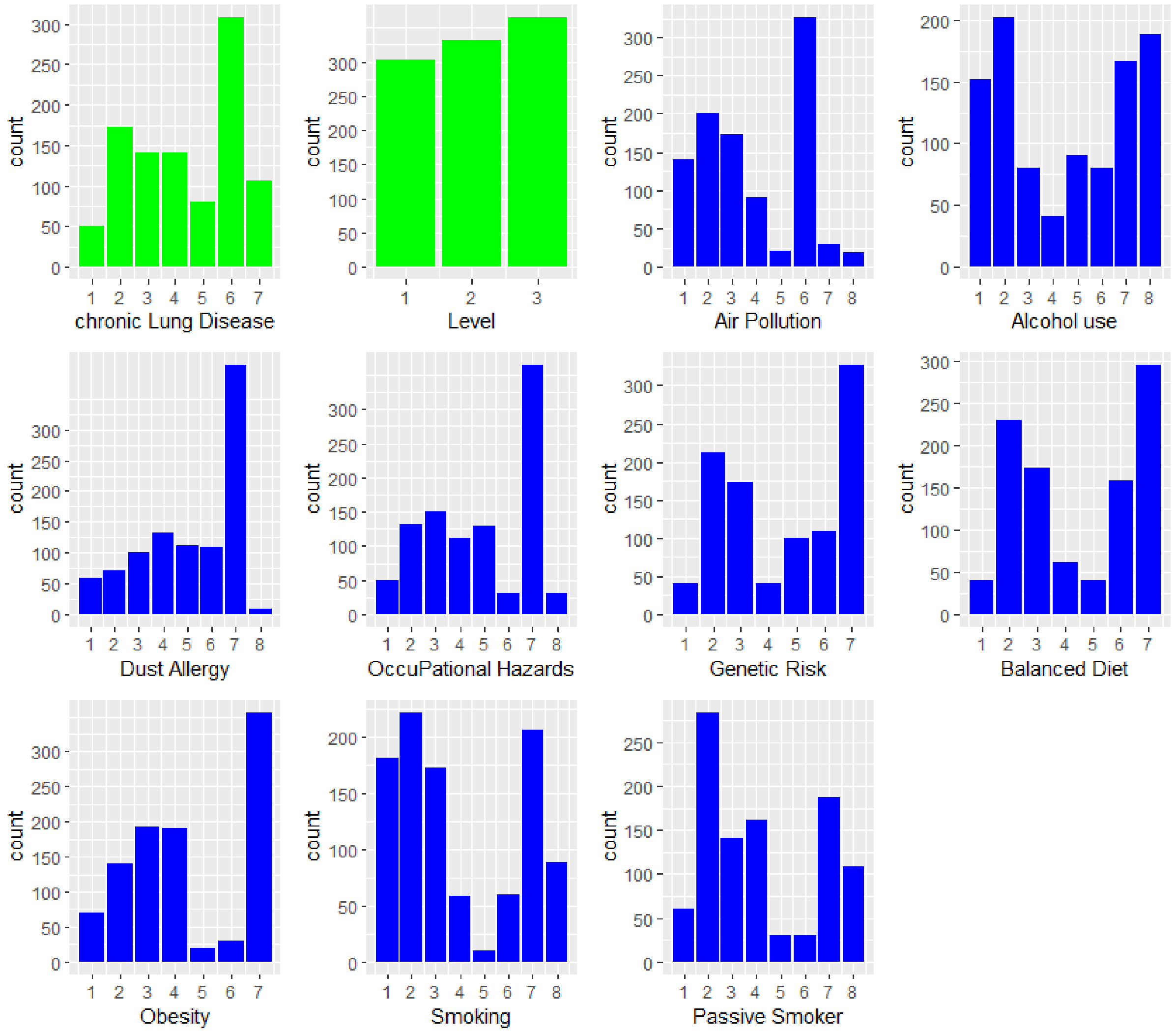
- Rather than focus on just air pollution, consider every predictor variable while model building to determine which are most impactful to lung cancer severity.
  - Create two models: one for potential causes and one for potential effects.
  - Hypothesis: smoking will be the most impactful.

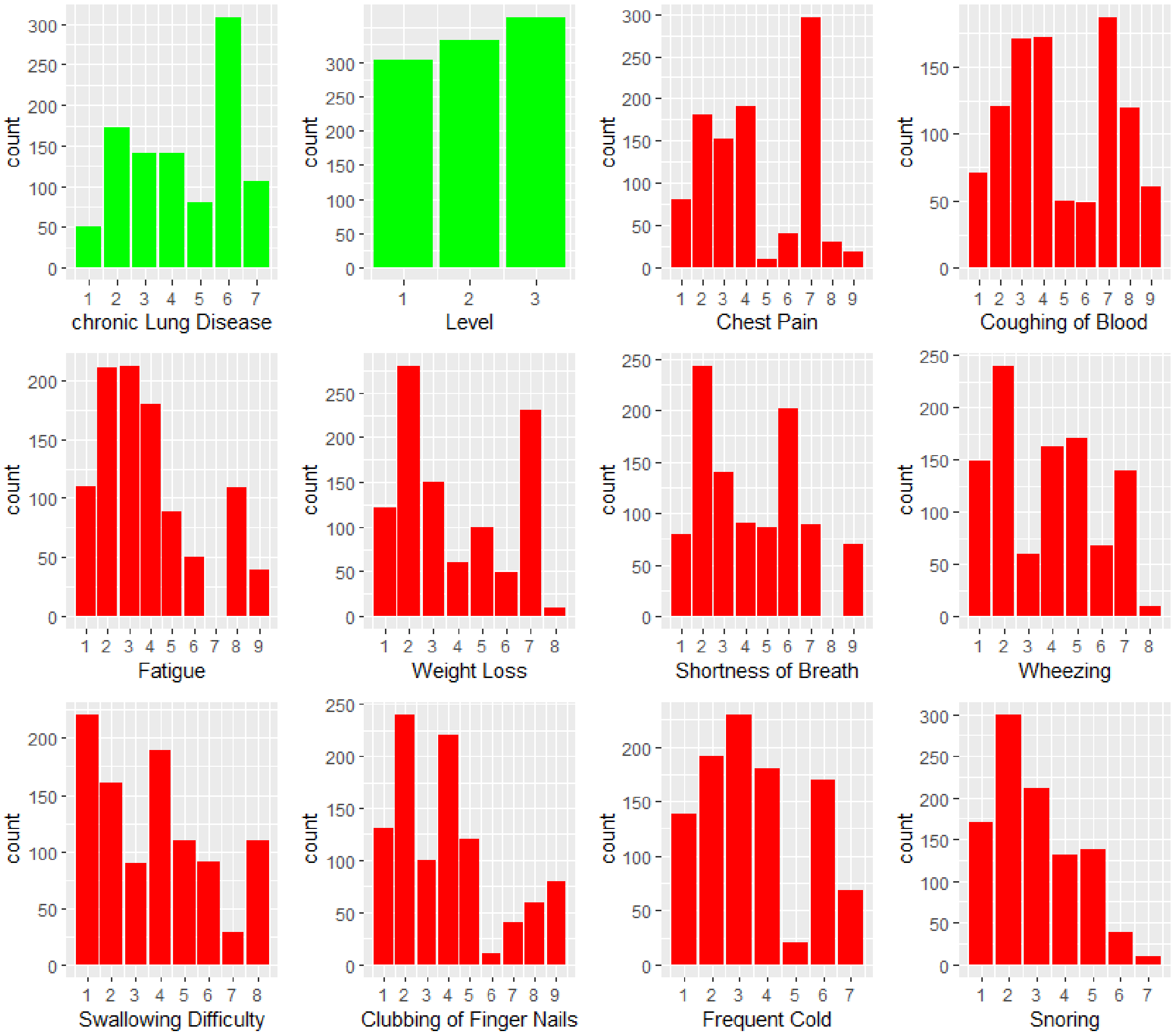
EXPLORATORY ANALYSIS

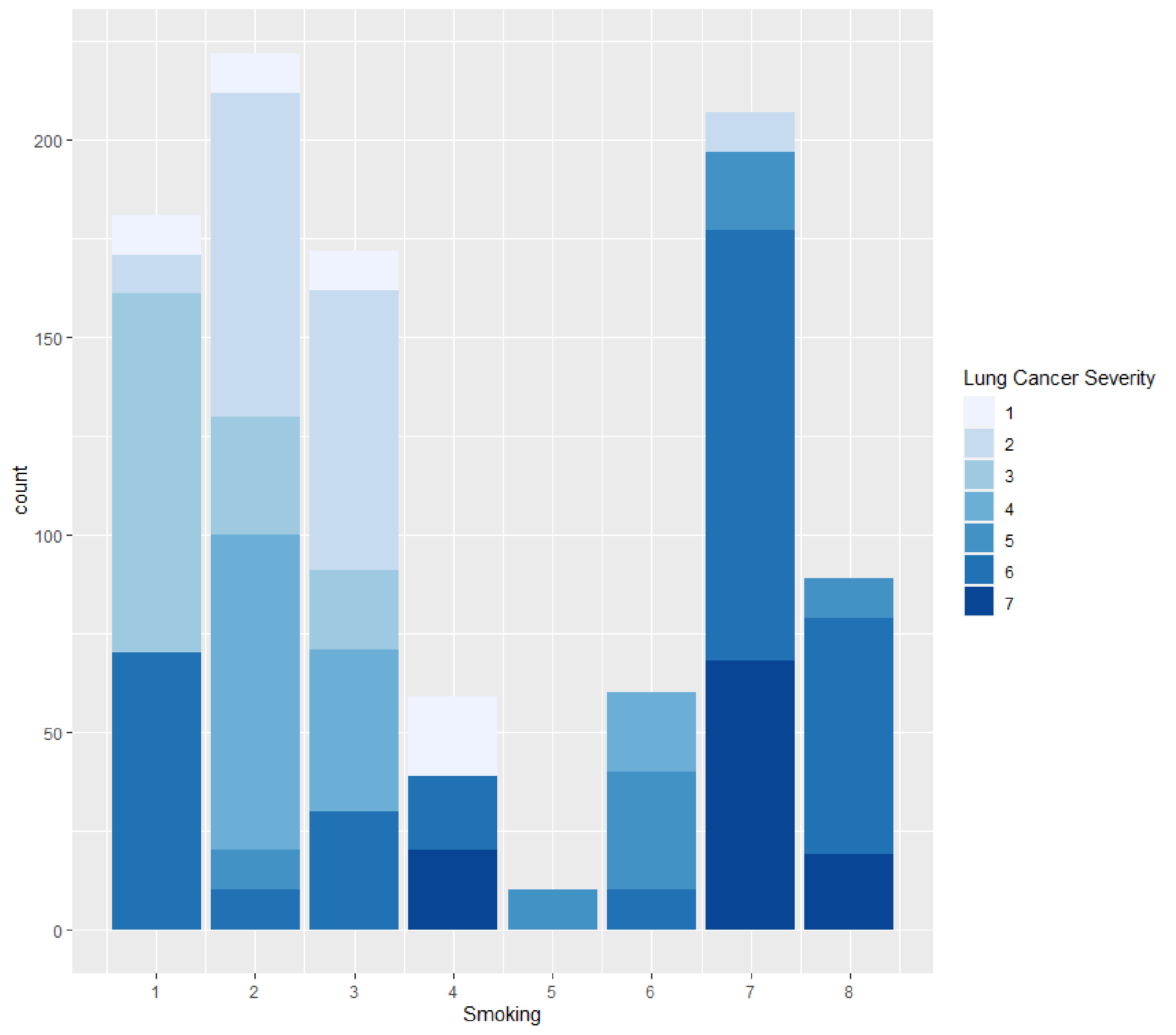


598 men & 402 women


80
60
40
20
0
count
15
20
25
30
35
40
45
50
55
60
65
70
75
Age


count
300
250
200
150
100
50
0
1 2 3 4 5 6 7
chronic Lung Disease
count
300
250
200
150
100
50
0
1 2 3
Level
count
300
250
200
150
100
50
0
1 2 3 4 5 6 7 8
Air Pollution
count
200
150
100
50
0
1 2 3 4 5 6 7 8
Alcohol use
count
300
250
200
150
100
50
0
1 2 3 4 5 6 7 8
Dust Allergy
count
300
250
200
150
100
50
0
1 2 3 4 5 6 7 8
OccuPational Hazards
count
300
250
200
150
100
50
0
1 2 3 4 5 6 7
Genetic Risk
count
300
250
200
150
100
50
0
1 2 3 4 5 6 7
Balanced Diet
count
300
250
200
150
100
50
0
1 2 3 4 5 6 7
Obesity
count
200
150
100
50
0
1 2 3 4 5 6 7 8
Smoking
count
250
200
150
100
50
0
1 2 3 4 5 6 7 8
Passive Smoker


count
300
250
200
150
100
50
0
1 2 3 4 5 6 7
chronic Lung Disease
count
300
250
200
150
100
50
0
1 2 3
Level
count
300
250
200
150
100
50
0
1 2 3 4 5 6 7 8 9
Chest Pain
count
150
100
50
0
1 2 3 4 5 6 7 8 9
Coughing of Blood
count
200
150
100
50
0
1 2 3 4 5 6 7 8 9
Fatigue
count
250
200
150
100
50
0
1 2 3 4 5 6 7 8
Weight Loss
count
250
200
150
100
50
0
1 2 3 4 5 6 7 8 9
Shortness of Breath
count
250
200
150
100
50
0
1 2 3 4 5 6 7 8
Wheezing
count
200
150
100
50
0
1 2 3 4 5 6 7 8
Swallowing Difficulty
count
250
200
150
100
50
0
1 2 3 4 5 6 7 8 9
Clubbing of Finger Nails
count
200
150
100
50
0
1 2 3 4 5 6 7
Frequent Cold
count
300
250
200
150
100
50
0
1 2 3 4 5 6 7
Snoring


count
200
150
100
50
0
1
2
3
4
5
6
7
8
Smoking
Lung Cancer Severity
1
2
3
4
5
6
7

MODEL BUILDING

- Most of the data is ordinal: cumulative logit regression model.

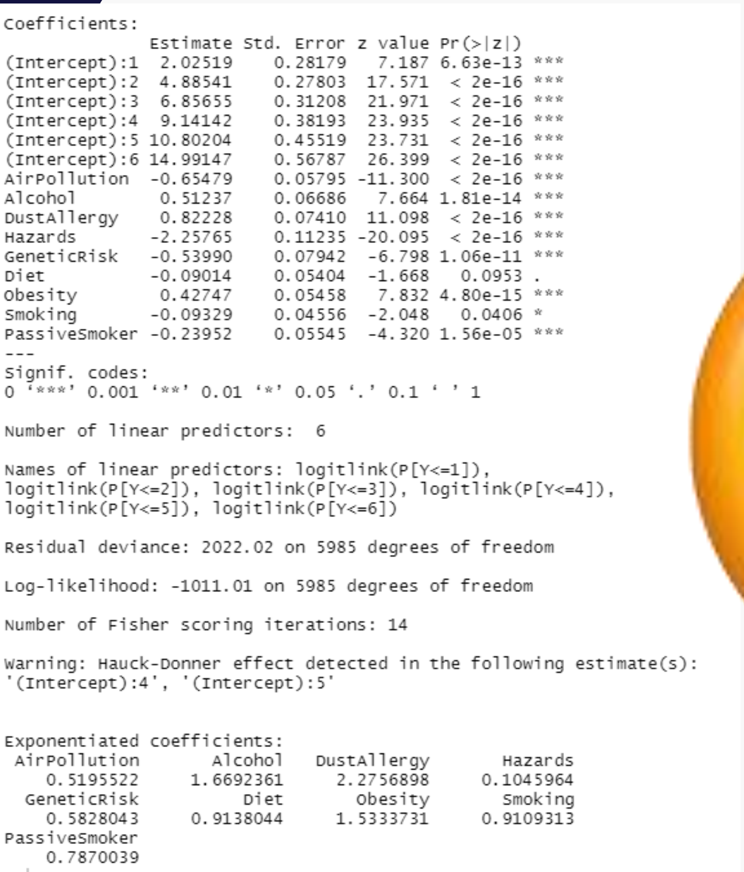
- Proportional odds assumption (cumulative logit slopes are the same, intercepts differ).

- As said before, building two models.

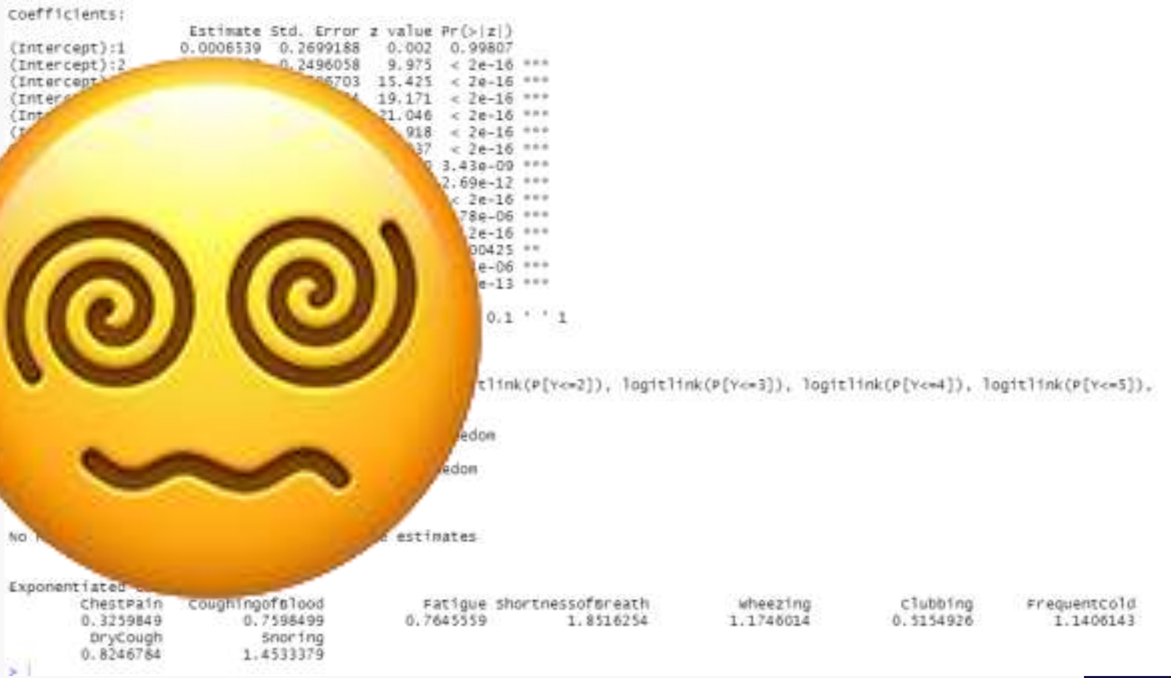
WHAT KIND OF MODEL IS THIS ANYWAY?

$$\log\left(\frac{P(Y \le j)}{P(Y > j)}\right) = \log\left(\frac{P(Y \le j)}{1 - P(Y \le j)}\right) = \log\left(\frac{\pi_1 + \ldots + \pi_j}{\pi_{j+1} + \ldots + \pi_J}\right)$$

$$L_{J-1} \quad = \quad \beta_{0,J-1} + \beta_{1,J-1}x_1 + \cdots + \beta_{p,J-1}x_p$$



```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)    
(Intercept):1  2.02519    0.28179   7.187 6.63e-13 ***
(Intercept):2  4.88541    0.27803  17.571  < 2e-16 ***
(Intercept):3  6.85655    0.31208  21.971  < 2e-16 ***
(Intercept):4  9.14142    0.38193  23.935  < 2e-16 ***
(Intercept):5 10.80204    0.45519  23.731  < 2e-16 ***
(Intercept):6 14.99147    0.56787  26.399  < 2e-16 ***
AirPollution  -0.65479    0.05795 -11.300  < 2e-16 ***
Alcohol        0.51237    0.06686   7.664 1.81e-14 ***
DustAllergy    0.82228    0.07410  11.098  < 2e-16 ***
Hazards       -2.25765    0.11235 -20.095  < 2e-16 ***
GeneticRisk   -0.53990    0.07942  -6.798 1.06e-11 ***
Diet          -0.09014    0.05404  -1.668   0.0953 .  
Obesity        0.42747    0.05458   7.832 4.80e-15 ***
Smoking       -0.09329    0.04556  -2.048   0.0406 *  
PassiveSmoker -0.23952    0.05545  -4.320 1.56e-05 ***
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Number of linear predictors:  6 

Names of linear predictors: logitlink(P[Y<=1]), 
logitlink(P[Y<=2]), logitlink(P[Y<=3]), logitlink(P[Y<=4]), 
logitlink(P[Y<=5]), logitlink(P[Y<=6])

Residual deviance: 2022.02 on 5985 degrees of freedom

Log-likelihood: -1011.01 on 5985 degrees of freedom

Number of Fisher scoring iterations: 14 

Warning: Hauck-Donner effect detected in the following estimate(s):
'(Intercept):4', '(Intercept):5'


Exponentiated coefficients:
 AirPollution       Alcohol   DustAllergy       Hazards 
    0.5195522     1.6692361     2.2756898     0.1045964 
  GeneticRisk          Diet       Obesity       Smoking 
    0.5828043     0.9138044     1.5333731     0.9109313 
PassiveSmoker 
    0.7870039 
```

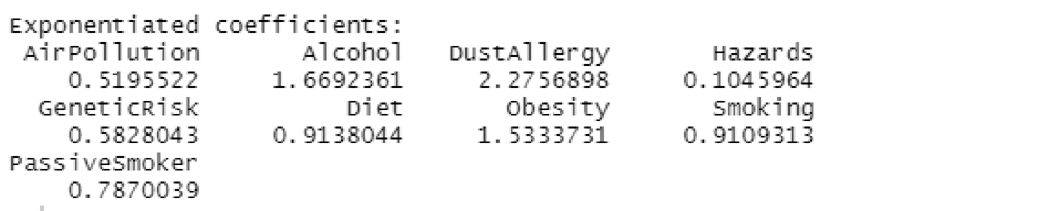


```
Coefficients:
                   Estimate Std. Error z value Pr(>|z|)    
(Intercept):1     0.0006539  0.2699188   0.002  0.99807    
(Intercept):2                0.2496058   9.975  < 2e-16 ***
(Intercept)                      6703  15.425  < 2e-16 ***
(Interc                                19.171  < 2e-16 ***
(Int                                   21.046  < 2e-16 ***
(                                         918  < 2e-16 ***
                                           37  < 2e-16 ***
                                             3.43e-09 ***
                                             2.69e-12 ***
                                              < 2e-16 ***
                                               78e-06 ***
                                               2e-16 ***
                                               00425 **
                                                e-06 ***
                                                e-13 ***

                                                0.1 ‘ ’ 1


                                                tlink(P[Y<=2]), logitlink(P[Y<=3]), logitlink(P[Y<=4]), logitlink(P[Y<=5]),

                                               edom

                                              edom


No                                        estimates


Exponentiated 
      ChestPain   CoughingofBlood          Fatigue ShortnessofBreath          Wheezing          Clubbing      FrequentCold 
      0.3259849         0.7598499        0.7645559         1.8516254         1.1746014         0.5154926         1.1406143 
       DryCough           Snoring 
      0.8246784         1.4533379 
> 
```

IN SIMPLER TERMS…

- For the potential causes and effects, for a one unit increase in ____________ there is a ______ multiplicative change in the odds of a being at a lower lung cancer severity level.

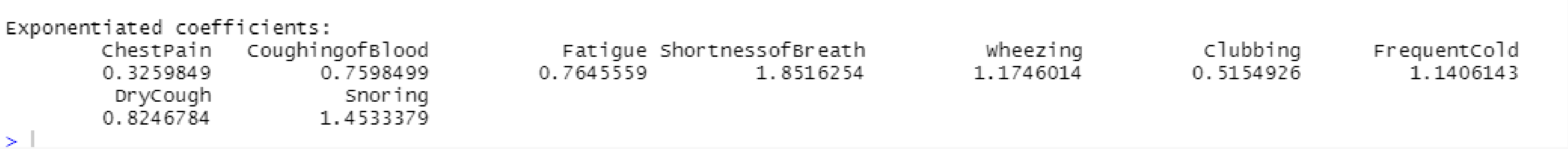
FINDINGS + CONCLUSIONS



```
Exponentiated coefficients:
 AirPollution       Alcohol   DustAllergy       Hazards
    0.5195522     1.6692361     2.2756898     0.1045964
  GeneticRisk          Diet       Obesity       Smoking
    0.5828043     0.9138044     1.5333731     0.9109313
PassiveSmoker
    0.7870039
```

- **Potential causes:**
  - A 1 level increase in alcohol, dust allergies, or obesity is associated with a higher odds of being at a lower lung cancer severity level.
    - **Most influential:** Dust allergies
  - A 1 level increase in air pollution, genetic risk, smoking, passive smoking, diet, or work hazards is associated with a higher odds of being at a higher lung cancer severity level.
    - **Most influential:** Work hazards

FINDINGS + CONCLUSIONS



```
Exponentiated coefficients:
       ChestPain   CoughingofBlood           Fatigue ShortnessofBreath         Wheezing         Clubbing      FrequentCold
       0.3259849         0.7598499         0.7645559         1.8516254        1.1746014        0.5154926         1.1406143
        DryCough           Snoring
       0.8246784         1.4533379
> 
```

- **Potential effects/symptoms:**
  - A 1 level increase in snoring, shortness of breath, wheezing, or frequent colds is associated with a higher odds of being at a lower lung cancer severity level.
    - **Most influential:** Shortness of breath
  - A 1 level increase in chest pain, dry cough, coughing of blood, fatigue, or fingernail clubbing is associated with higher odds of being at a higher lung cancer severity level.
    - **Most influential:** Chest pain

IF I HAD MORE TIME…

- Explore different relationships.
- Different model type?
- Verify and check my model.
  - Complicated
- Additional research?

THANK YOU FOR LISTENING!

References

https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-new-link

https://online.stat.psu.edu/stat504/book/export/html/793