# FinalProject

Eeman, Linda, Sofia

2025-11-06

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.1.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
dat <- read_csv("cancer patient data sets 2.csv")
```

```
## Rows: 1000 Columns: 26
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (2): Patient Id, Level
## dbl (24): index, Age, Gender, Air Pollution, Alcohol use, Dust Allergy, Occu...
##
## 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.
```

```
library(MASS)
```

```
##
## Attaching package: 'MASS'
##
## The following object is masked from 'package:dplyr':
##
##     select
```

```
dat <- read_csv("cancer patient data sets 2.csv")
```

```
## Rows: 1000 Columns: 26
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (2): Patient Id, Level
## dbl (24): index, Age, Gender, Air Pollution, Alcohol use, Dust Allergy, Occu...
##
## 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.
```

```r
names(dat) <- make.names(names(dat))

names(dat)
```

```
##  [1] "index"                    "Patient.Id"
##  [3] "Age"                      "Gender"
##  [5] "Air.Pollution"            "Alcohol.use"
##  [7] "Dust.Allergy"             "OccuPational.Hazards"
##  [9] "Genetic.Risk"             "chronic.Lung.Disease"
## [11] "Balanced.Diet"            "Obesity"
## [13] "Smoking"                  "Passive.Smoker"
## [15] "Chest.Pain"               "Coughing.of.Blood"
## [17] "Fatigue"                  "Weight.Loss"
## [19] "Shortness.of.Breath"      "Wheezing"
## [21] "Swallowing.Difficulty"    "Clubbing.of.Finger.Nails"
## [23] "Frequent.Cold"            "Dry.Cough"
## [25] "Snoring"                  "Level"
```

```r
dat$Level <- factor(dat$Level,
                    levels = c("Low","Medium","High"),
                    ordered = TRUE)

dat <- dat |>
  mutate(High = if_else(Level == "High", 1, 0))



m1causes <- glm(High ~ Gender + Genetic.Risk + Age + Smoking + Air.Pollution + Obesity, data = dat, fami

m1symptoms <- glm(High ~  Chest.Pain + Coughing.of.Blood + Shortness.of.Breath + Weight.Loss + Frequent

summary(m1causes)
```

```
##
## Call:
## glm(formula = High ~ Gender + Genetic.Risk + Age + Smoking +
##     Air.Pollution + Obesity, family = "binomial", data = dat)
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -16.26976    1.79617  -9.058  < 2e-16 ***
## Gender          1.92822    0.47620   4.049 5.14e-05 ***
## Genetic.Risk   -0.25830    0.14304  -1.806 0.070940 .
## Age            -0.05281    0.01424  -3.708 0.000209 ***
## Smoking         0.79419    0.10285   7.722 1.15e-14 ***
## Air.Pollution   1.22708    0.13725   8.941  < 2e-16 ***
## Obesity         1.59079    0.14919  10.663  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1312.48  on 999  degrees of freedom
## Residual deviance:  236.77  on 993  degrees of freedom
```

```
## AIC: 250.77
##
## Number of Fisher Scoring iterations: 8
```

```
summary(m1symptoms)
```

```
##
## Call:
## glm(formula = High ~ Chest.Pain + Coughing.of.Blood + Shortness.of.Breath +
##     Weight.Loss + Frequent.Cold, family = "binomial", data = dat)
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -18.2853     1.7624 -10.375  < 2e-16 ***
## Chest.Pain            0.8487     0.1487   5.708 1.14e-08 ***
## Coughing.of.Blood     1.3555     0.1470   9.220  < 2e-16 ***
## Shortness.of.Breath  -0.1467     0.1118  -1.312 0.189549
## Weight.Loss           1.1024     0.1756   6.278 3.42e-10 ***
## Frequent.Cold         0.3682     0.1038   3.548 0.000389 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1312.5  on 999  degrees of freedom
## Residual deviance:  296.4  on 994  degrees of freedom
## AIC: 308.4
##
## Number of Fisher Scoring iterations: 8
```

```
exp(7.6069)
```

```
## [1] 2012.031
```

```
exp(4.1667)
```

```
## [1] 64.50224
```

```
exp(6.3718)
```

```
## [1] 585.1101
```

```
exp(-0.7692)
```

```
## [1] 0.4633836
```

```
exp( 10.4882)
```

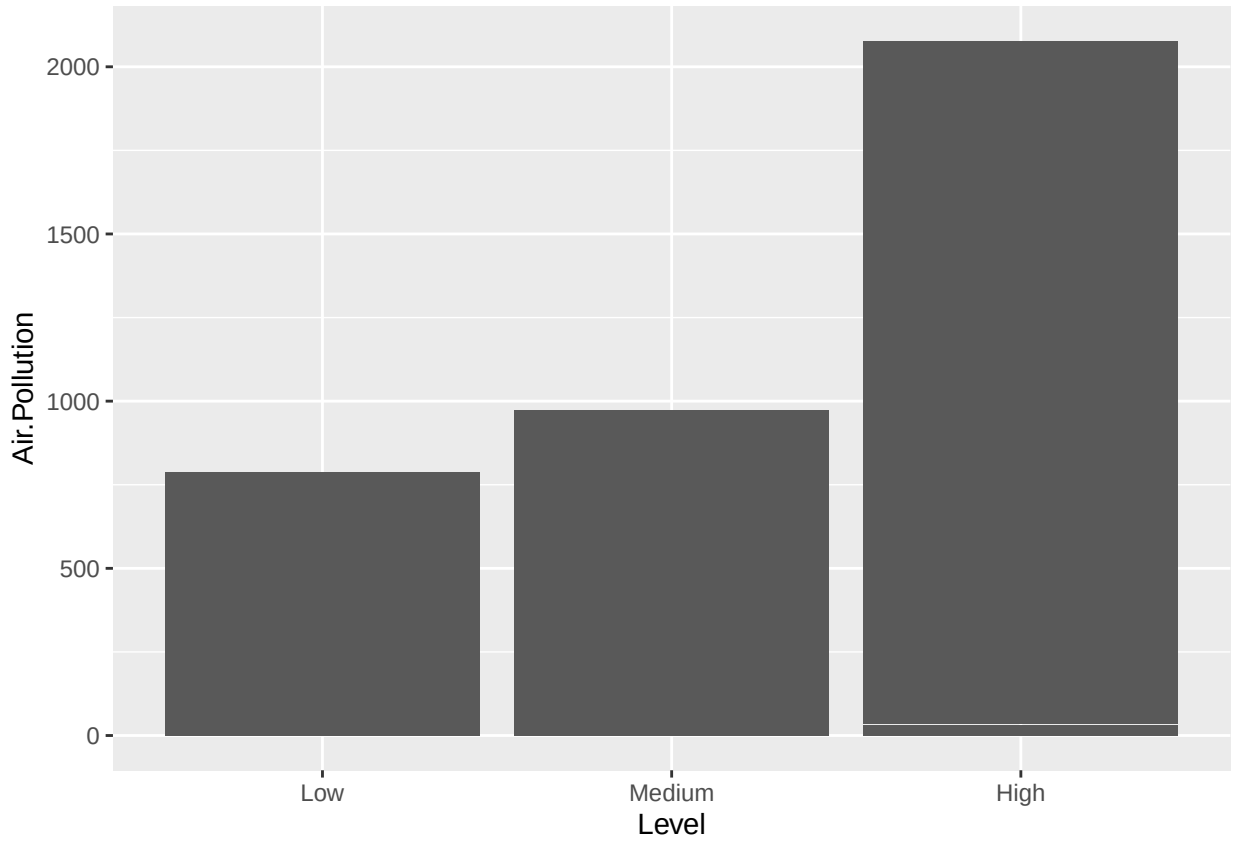
```
## [1] 35889.5
```

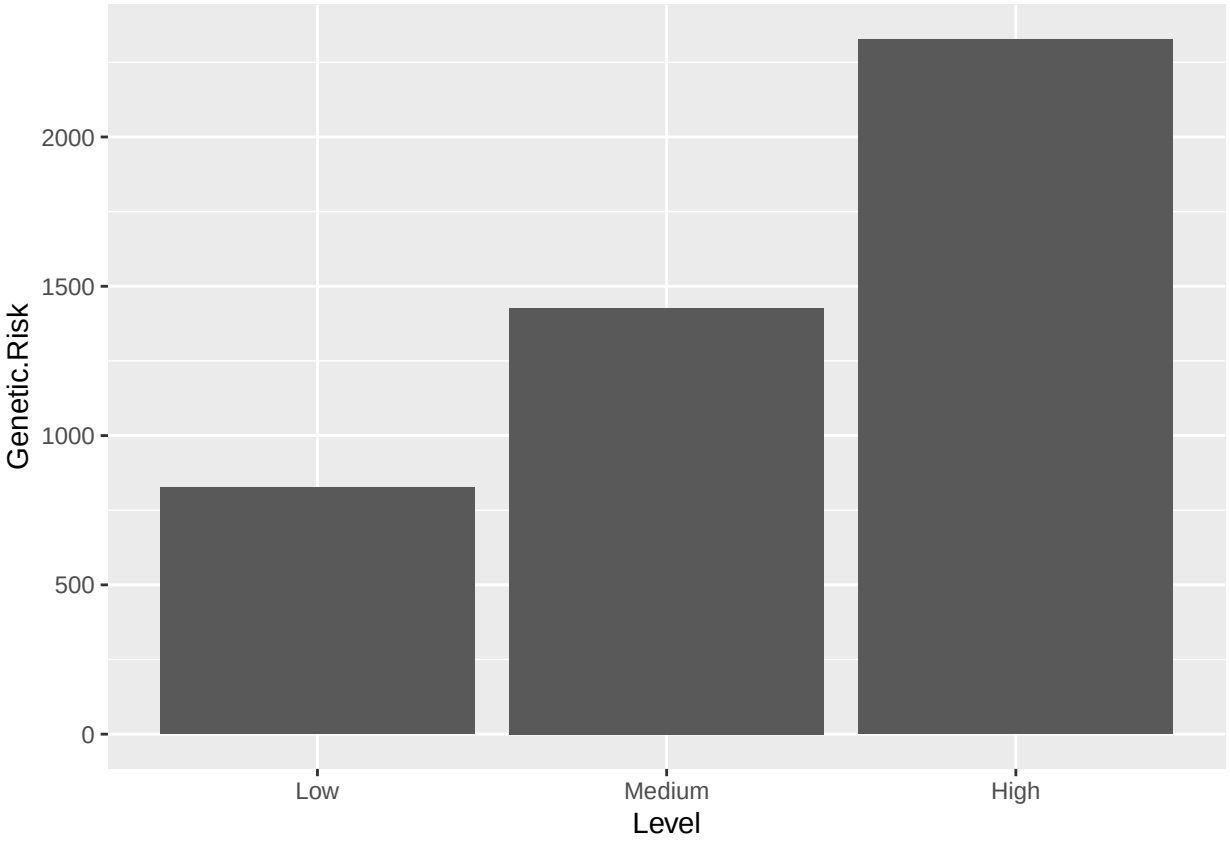
Logression Coefficients For Cancer Markers

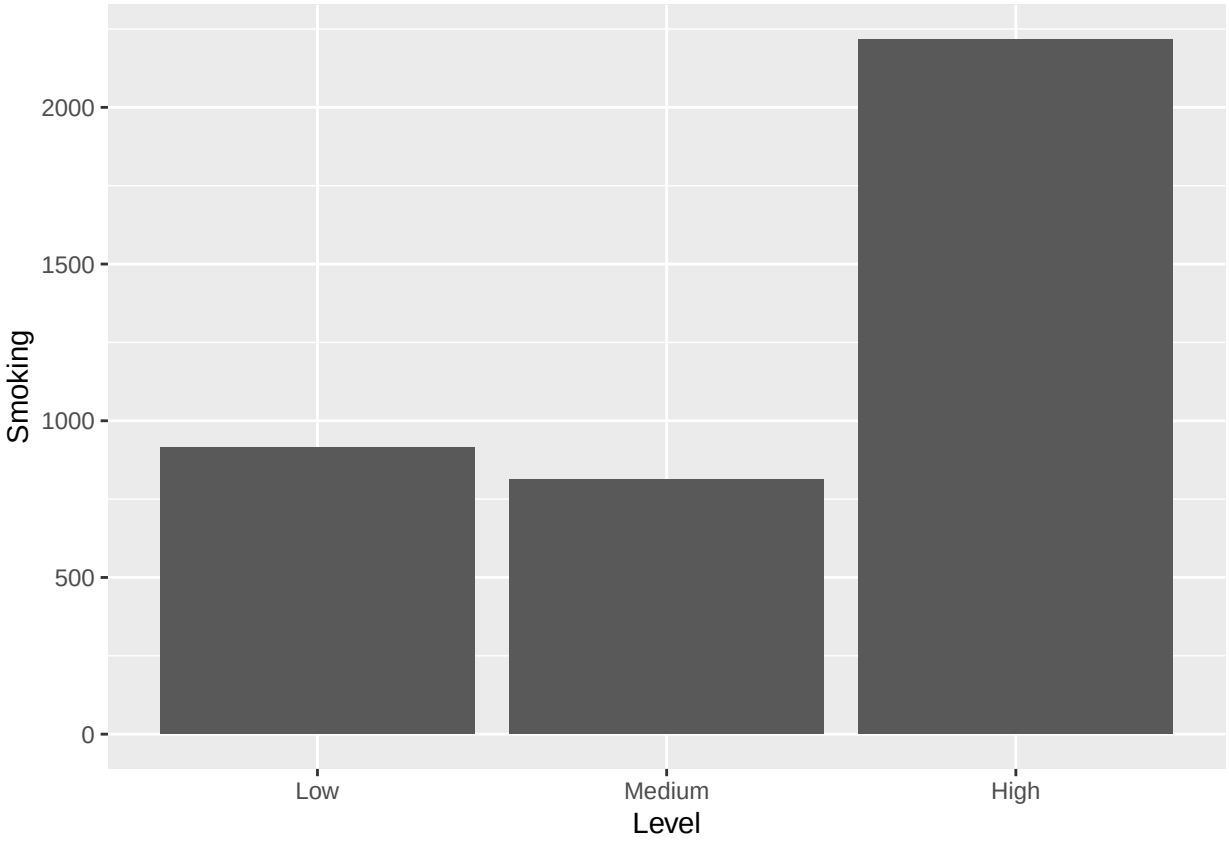
| | Estimate | Standard.Error | pValue |
|---|---|---|---|
| Intercept | -16.270 | 1.796 | 2.00e-16 |
| Gender | 1.928 | 0.476 | 5.14e-05 |
| Genetic Risk | -0.258 | 0.142 | 7.10e-02 |
| Age | -0.052 | 0.014 | 2.00e-04 |
| Smoking | 0.794 | 0.102 | 1.15e-14 |
| Air Pollution | 1.227 | 0.137 | 2.00e-16 |
| Obesity | 1.590 | 0.149 | 2.00e-16 |

Logression Coefficients For Cancer Symptoms

| | Estimate | Standard.Error | pValue |
|---|---|---|---|
| Intercept | -18.285 | 1.762 | 2.00e-16 |
| Chest Pain | 0.848 | 0.148 | 1.14e-08 |
| Coughing of Blood | 1.356 | 0.147 | 2.00e-16 |
| Shortness of Breath | -0.147 | 0.112 | 1.90e-01 |
| Weight Loss | 1.102 | 0.176 | 3.42e-10 |
| Frequent Cold | 0.368 | 0.104 | 3.89e-04 |

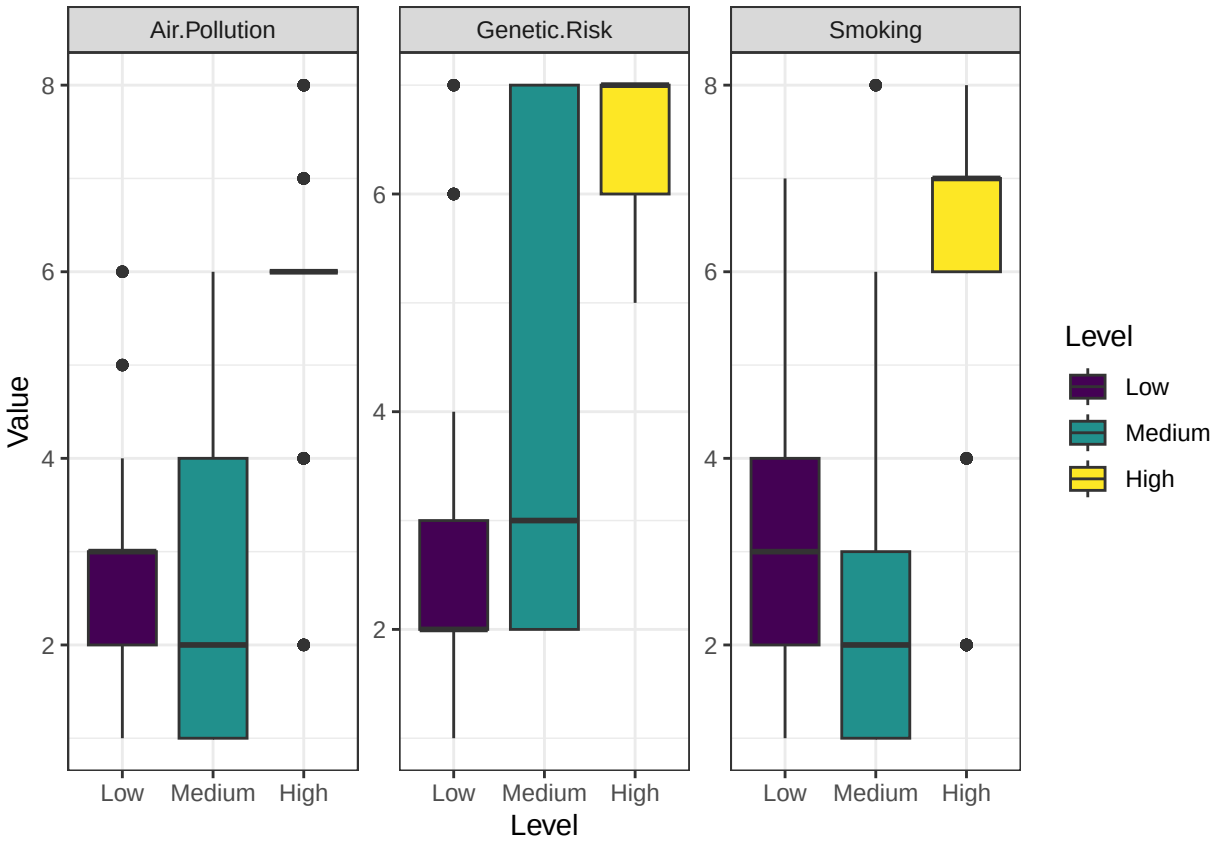



2000
1500
1000
500
0
Genetic.Risk
Low
Medium
High
Level




```
##
##      Low Medium High
##   1  40      0    0
##   2 121     91    0
##   3  92     81    0
##   4  20     20    0
##   5   0     20   80
##   6  20     20   68
##   7  10    100  217
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(dat$Genetic.Risk, dat$Level)
## X-squared = 632.14, df = 12, p-value < 2.2e-16
```


Air.Pollution
Genetic.Risk
Smoking
Value
Level
Low
Medium
High
2
4
6
8
Level
Low
Medium
High