STA 138 Final Project

Investigate and Analysis of Byssinosis in North Carolina

Instructor: Andrew Farris

TA: Jiangshan Zhang

Group Member: Tianyu Lin

Yujie Zhong

Zhuoya Wang

I. Introduction

In 1973, a comprehensive study was conducted involving a prominent cotton textile company in North Carolina to explore the prevalence of byssinosis, The study encompassed data collected from 5,419 workers, including:

Type of work place [1 (most dusty), 2 (less dusty), 3 (least dusty)]

Employment, years [< 10, 10–19, 20–]

Smoking [Smoker, or not in last 5 years]

Sex [Male, Female]

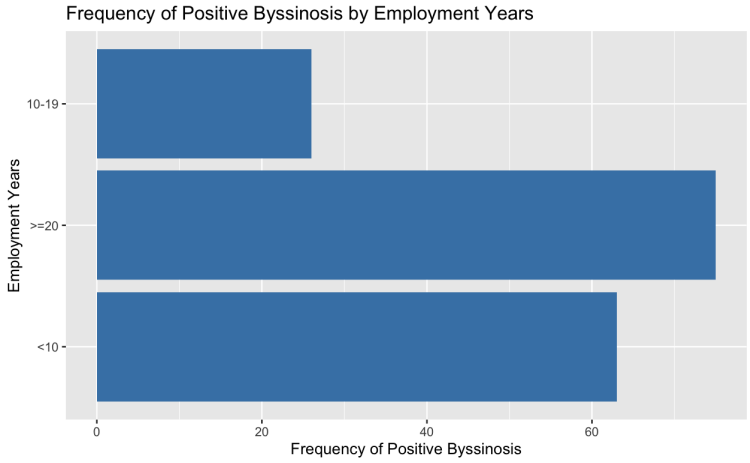
Race [White, Other]

Byssinosis [Yes, No]

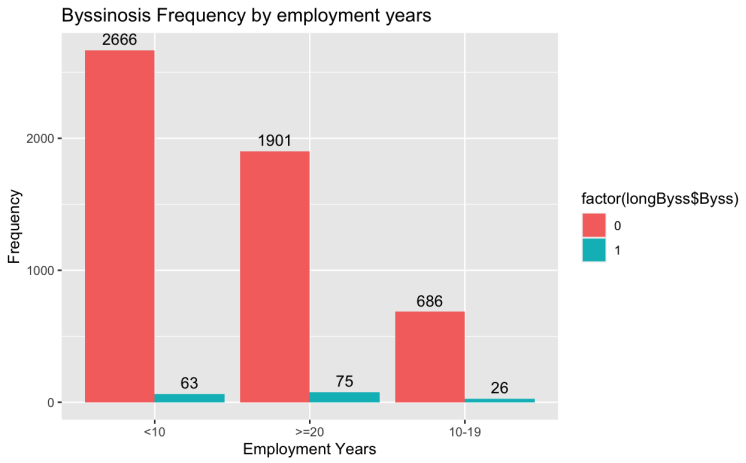
The primary objective of this project is to thoroughly investigate the relationships between byssinosis and various factors. To achieve this, our first step involves the initial establishment of plots for each variable and this ailment, aiming to observe potential relationships between them. In the second phase, we will use AIC and BIC to make inferences on which of the variables contribute to the presence of Byssinosis in order to find the best fitted statistical model.

II. Data Summary

Plot 1.



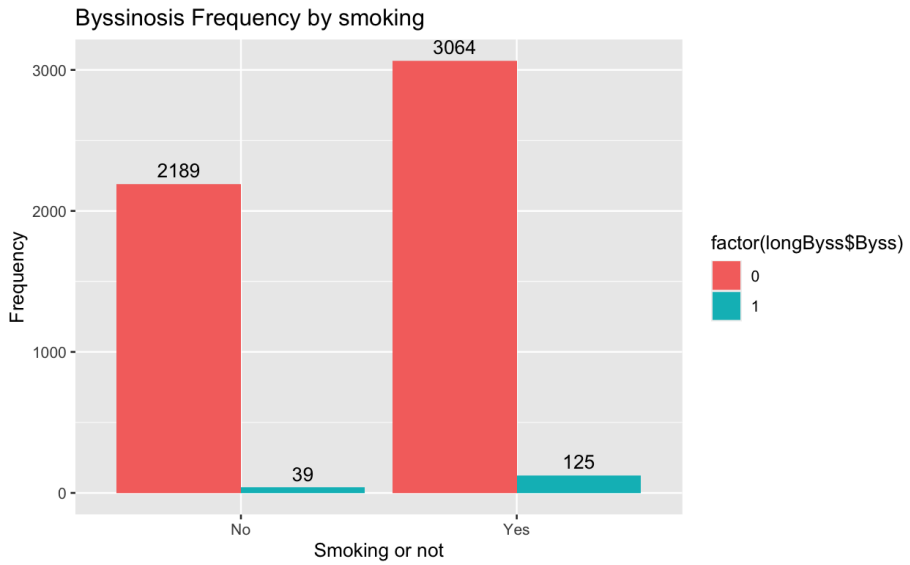
Plot 1.A



Plot 1.B

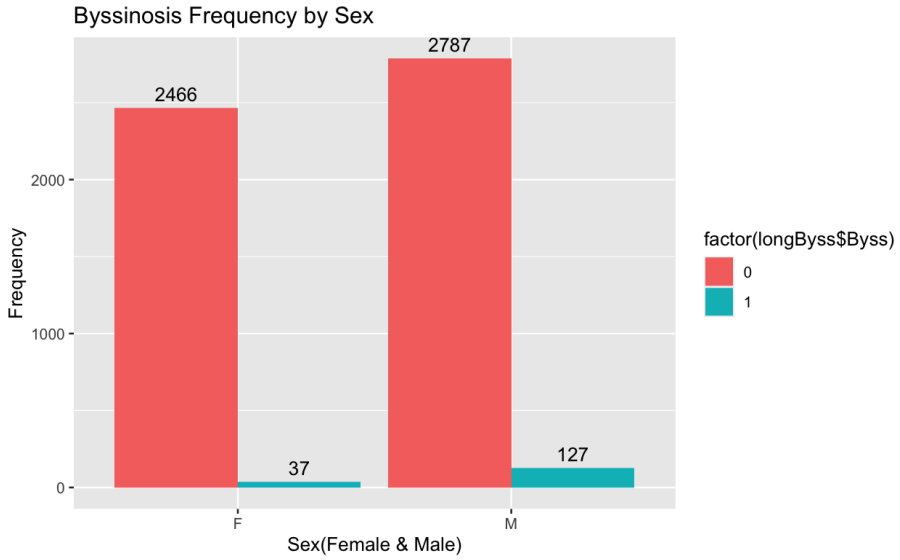
From the above plots, we found that those people who work more than 20 years have the largest number of people infected with Byssinosis.

Plot 2.



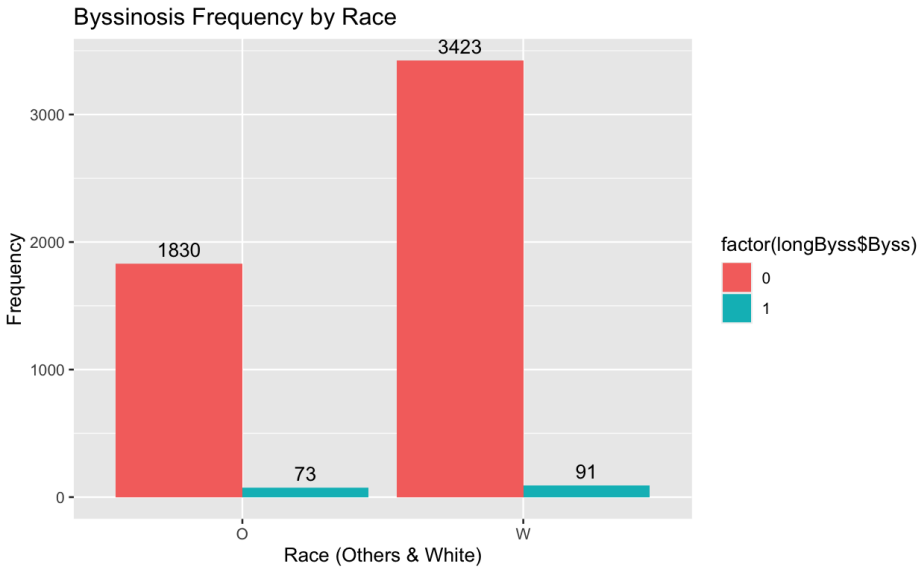
In the non-smoking group, the majority of individuals (2189 people) did not have the disease, with only a small number (39 people) being affected. Similarly, in the smoking group, the majority of individuals (3064 people) did not have the disease, but a relatively higher number (125 people) were affected.

Plot 3.



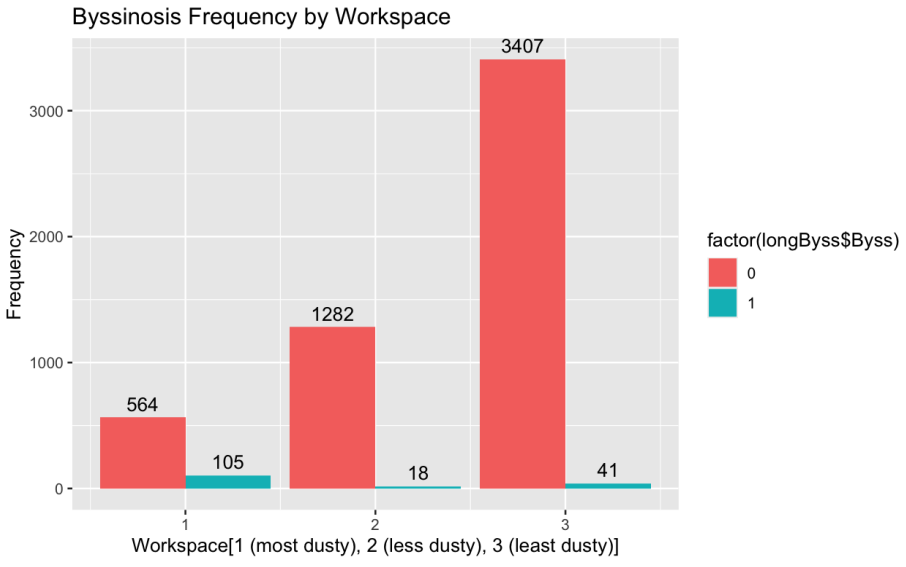
In plot 3, we can observe that there are 321 more male employees than female employees, and there are also 90 more affected individuals among males. Therefore, we cannot conclusively determine the relationship between gender and the probability of contracting the disease; further research is needed.

Plot 4.



In plot 4, we observe that a small portion of individuals working in this factory are of white race, with 73 people affected by the disease, while in other racial groups, 91 people are affected. Therefore, we do not find a clear association between race and the incidence of the disease.

Plot 5.



In this plot, we can clearly see that among those working in the most dusty environment, there are only 564 individuals, but the number of affected individuals is as high as 105. In contrast, among those working in less dusty environments (1283) and the least dusty environments (3407), only 18 and 41 individuals are affected, respectively. Therefore, we believe that there might be a significant relationship between the workplace and the probability of contracting the disease.

In the following steps, we will compute the AIC and BIC for each model as inference to determine the most appropriate model that can present the relationship between the chance of having Byssinosis and predictors.

III. Model Selection

We first fit all data into the empty model and the full model with interaction terms to study the relationship between Byssinosis and an individual's smoking status, sex, race, length of employment, smoking, and dustiness of the workplace. As a result, we have the largest AIC and BIC in the empty model, which contains no predictors. It does not explain any of the variability in the dependent variable, leading to a large residual sum of squares. Both AIC and BIC penalize poor fit, leading to higher values. A full model with interaction terms, which contains all available predictors, also might not be the best fitted model due to overfitting, where it captures noise rather than the underlying pattern, leading to poor performance on new data. It increases the risk of type I error and causes high penalties in AIC and BIC because of the complexity, suggesting a less optimal balance between model fit and simplicity.

Thus, we are going to utilize the stepwise model selection process to find the model(s) with lowest AIC and/or BIC in between empty and full model by computing it in three directions: **forward**, **backward** or **both simultaneously.**

AIC:

```
## Start:  AIC=1472.14
## Byss ~ 1
##
##               Df Deviance     AIC
## + Workspace   2   1216.2 1222.2
## + Sex         1   1429.5 1433.5
## + Smoking     1   1447.7 1451.7
## + Employment  2   1460.3 1466.3
## + Race        1   1463.8 1467.8
## <none>              1470.1 1472.1
##
## Step:  AIC=1222.22
## Byss ~ Workspace
##
##               Df Deviance     AIC
## + Smoking     1   1203.4 1211.4
## + Employment  2   1201.7 1211.7
## + Race        1   1213.9 1221.9
## <none>              1216.2 1222.2
## + Sex         1   1215.7 1223.7
##
## Step:  AIC=1211.39
## Byss ~ Workspace + Smoking
##
##                     Df Deviance     AIC
## + Employment        2   1189.5 1201.5
## + Smoking:Workspace  2   1198.2 1210.2
## + Race              1   1201.3 1211.3
## <none>                  1203.4 1211.4
## + Sex               1   1203.4 1213.4
##
## Step:  AIC=1201.46
## Byss ~ Workspace + Smoking + Employment
##
##                           Df Deviance     AIC
## + Smoking:Workspace       2   1184.8 1200.8
## <none>                    1189.5 1201.5
## + Sex                     1   1189.0 1203.0
## + Race                    1   1189.1 1203.1
## + Employment:Smoking      2   1187.2 1203.2
## + Employment:Workspace    4   1184.8 1204.8
##
```

```
## Step:  AIC=1200.77
## Byss ~ Workspace + Smoking + Employment + Workspace:Smoking
##
##                           Df Deviance     AIC
## <none>                    1184.8 1200.8
## + Employment:Smoking     2   1182.0 1202.0
## + Race                   1   1184.5 1202.5
## + Sex                    1   1184.6 1202.6
## + Employment:Workspace  4   1179.9 1203.9
```

After AIC selection, we obtain that the least value of AIC is 1200.77 with predictors Employment, Smoking status,Workplace, and the interaction term of Workspace and Smoking from the forward selection.

BIC:

```
## Start:  AIC=1478.74
## Byss ~ 1
##
##              Df Deviance    AIC
## + Workspace   2   1216.2 1242.0
## + Sex         1   1429.5 1446.7
## + Smoking     1   1447.7 1464.9
## <none>            1470.1 1478.7
## + Race        1   1463.8 1481.0
## + Employment  2   1460.3 1486.1
##
## Step:  AIC=1242.01
## Byss ~ Workspace
##
##              Df Deviance    AIC
## + Smoking     1   1203.4 1237.8
## <none>            1216.2 1242.0
## + Employment  2   1201.7 1244.7
## + Race        1   1213.9 1248.3
## + Sex         1   1215.7 1250.1
##
## Step:  AIC=1237.78
## Byss ~ Workspace + Smoking
##
```

```
##                       Df Deviance    AIC
## <none>                     1203.4 1237.8
## + Employment           2   1189.5 1241.0
## + Race                 1   1201.3 1244.3
## + Sex                  1   1203.4 1246.4
## + Smoking:Workspace    2   1198.2 1249.8
```

After BIC selection, we obtain the same model from selection in three directions, and the least value of BIC is 1237.78 with predictors Smoking and Workplace without any interaction terms.

Best fitted model based on AIC:

```
##
## Call:
## glm(formula = longByss$Byss ~ Workspace + Smoking + Employment +
##     Workspace:Smoking, family = binomial(), data = longByss)
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -2.7085     0.2644 -10.243  < 2e-16 ***
## Workspace2              -1.7756     0.4176  -4.251 2.12e-05 ***
## Workspace3              -2.4241     0.3812  -6.358 2.04e-10 ***
## SmokingYes               0.9569     0.2750   3.479 0.000503 ***
## Employment>=20           0.6503     0.1825   3.563 0.000367 ***
## Employment10-19          0.4984     0.2501   1.993 0.046238 *
## Workspace2:SmokingYes   -1.1816     0.5489  -2.152 0.031360 *
## Workspace3:SmokingYes   -0.4074     0.4411  -0.924 0.355666
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1470.1  on 5416  degrees of freedom
## Residual deviance: 1184.8  on 5409  degrees of freedom
## AIC: 1200.8
##
## Number of Fisher Scoring iterations: 7
```

$logit(\hat{\pi})$ = **– 2.7085 + 0.9569*SmokingYes – 1.7756*Workspace with less dusty – 2.4241*Workspace with least dusty + 0.6503*Employment{ years >=20} + 0.4984*Employment{years 10-19} – 1.1816*Workspace with less dusty:SmokingYes – 0.4074*Workspace with least dusty:SmokingYes**

Best fitted model based on BIC:

```
##
## Call:
## glm(formula = longByss$Byss ~ Smoking + Workspace, family =
binomial(),
##     data = longByss)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.1559     0.1818 -11.857  < 2e-16 ***
## SmokingYes    0.6574     0.1914   3.434 0.000594 ***
## Workspace2   -2.5306     0.2607  -9.709  < 2e-16 ***
## Workspace3   -2.6951     0.1902 -14.167  < 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: 1470.1  on 5416  degrees of freedom
## Residual deviance: 1203.4  on 5413  degrees of freedom
## AIC: 1211.4
##
## Number of Fisher Scoring iterations: 7
```

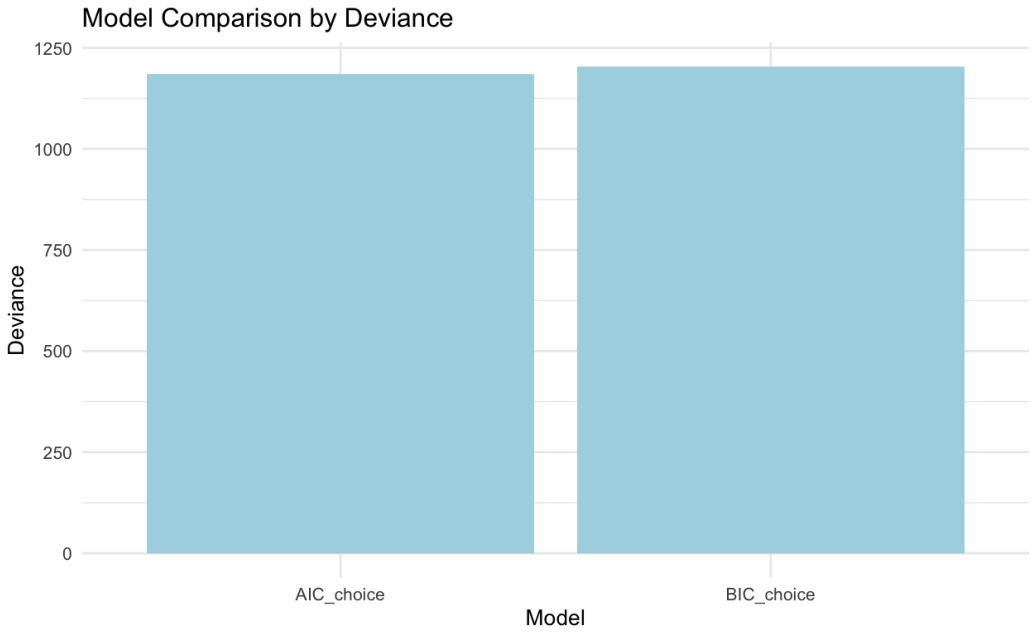
$logit(\hat{\pi})$ = **– 2.1559 + 0.6574*SmokingYes – 2.5306*Workspace with less dusty – 2.6951* Workspace with least dusty**

Since different fitted models based on AIC and BIC are given, we are supposed to do more interpretation and analysis to choose a better one.

IV. Model Interpretation

Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are used in model selection. $AIC = 2k - 2log(\hat{L})$, where k is the number of parameters estimated by a model and $\hat{L}$ is the maximized value of likelihood function of a model. $BIC = klog(n) - 2log(\hat{L})$, where k is the number of parameters estimated by a model, n is the number of observations, and $\hat{L}$ is the maximized value of likelihood function of a model. Both AIC and BIC aim to balance model fit and complexity. AIC is preferred when picking the most appropriate model for unknown data based on the known dataset, while BIC is used when choosing the true model.

Both AIC and BIC have their own preference, so deviance should be considered alongside other criteria like AIC or BIC to determine the better fitted model.



Plot6. The deviance of the AIC chosen model is slightly lower than the deviance of the BIC chosen model.

Deviance is a measure of error, and it quantifies how well a model fits the data. A lower deviance means better fit to data, while a higher deviance means a poorer fits. Based on the above deviance comparison barplot, we can observe that the AIC chosen model has a lower deviance compared to the BIC chosen model. Thus, the AIC chosen model has a better fit.

V. Conclusion

We aim to explore the relationships between Byssinosis and various factors, including workplace conditions, smoking habits, and employment years. Initially, we created vertical bar charts to visually examine potential direct relationships among these five factors and the likelihood of contracting the disease. We observed that there might be a connection between workplace conditions, smoking habits, and the incidence of Byssinosis. To further confirm this, we fit all data into both an empty model and a full model with interaction terms.

In the empty model, we obtained the highest AIC and BIC values, indicating that this model, which lacks any predictors, does not explain any variability in the dependent variable. Therefore, we plan to employ the stepwise model selection process in three directions: forward, backward, and both simultaneously. After AIC selection, the model with the lowest AIC (1200.77) includes predictors such as Employment, Smoking status, Workplace, and the interaction term of Workspace and Smoking, obtained through forward selection. Similarly, after BIC selection, the same model with predictors Smoking and Workplace (without any interaction terms) has the lowest BIC value of 1237.78.We found that the best-fitted model based on AIC is:

$logit(\hat{\boldsymbol{\pi}})$ = – 2.7085 + 0.9569*SmokingYes – 1.7756*Workspace with less dusty – 2.4241*Workspace with least dusty + 0.6503*Employment{ years >=20} + 0.4984*Employment{years 10-19} – 1.1816*Workspace with less dusty:SmokingYes – 0.4074*Workspace with least dusty:SmokingYes

However, the model based on BIC is: $logit(\hat{\boldsymbol{\pi}})$ = – 2.1559 + 0.6574*SmokingYes – 2.5306*Workspace with less dusty – 2.6951* Workspace with least dusty

We observed that these two models are different, but based on our deviance comparison plot, we found that the AIC model has a lower deviance. Therefore, we believe that the model based on AIC is more appropriate, which indicates that the chance of getting byssinosis has a strong relationship with **length of employment, workplace dustiness, and smoking status**. We can also conclude that there is a strong relationship between smoking status and workplace dustiness because of the interaction term in the final fitted model. The negative coefficient of workplace presents a negative relationship with chance of having byssinosis, implying that workplace dustiness does contribute to the chance of byssinosis.

Appendix:

```
knitr::opts_chunk$set(fig.align='center', message = FALSE, warning = FALSE)
library(ggplot2)
library(dplyr)
#install.packages("ggpubr")
library(ggpubr)
library(MASS)
library(reshape2)
#install.packages("reshape")
library(reshape)
library(data.table)
### raw data
byss <- read.csv("~/Desktop/Fall 2023/138/final project/Byssinosis.csv")
n <- nrow(byss)
countColumns <- c(which(names(byss) == "Byssinosis"),
                  which(names(byss) == "Non.Byssinosis"))
longByss <- rbind(
  cbind(Byss=1,
        byss[rep(1:n, byss["Byssinosis"]), -countColumns]),
  cbind(Byss=0,
        byss[rep(1:n, byss["Non.Byssinosis"]), -countColumns])
)
row.names(longByss) <- 1:nrow(longByss)
longByss$Workspace = as.factor(longByss$Workspace)
### Employment years & Y-- PLot 1

ggplot(longByss, aes(x = longByss$Employment, fill = factor(longByss$Byss))) +
  geom_bar(position = position_dodge(width = 0.9), stat = "count") +
  geom_text(stat = 'count', aes(label = ..count..),
            vjust = -0.5, position = position_dodge(width = 0.9)) +
  labs(title = "Byssinosis Frequency by employment years",
            x = "Employment Years",
            y = "Frequency")

##### For positive Byss

byssinosis_count <- longByss %>%
  filter(Byss == 1) %>%
  group_by(Employment) %>%
  summarise(Frequency = n())

# Now create the horizontal bar plot
ggplot(byssinosis_count, aes(x = Frequency, y = Employment)) +
  geom_bar(stat = "identity", fill = "steelblue") +
```

```
  labs(title = "Frequency of Positive Byssinosis by Employment Years",
           x = "Frequency of Positive Byssinosis",
           y = "Employment Years")

### Smoking & Byss --plot 2

ggplot(longByss, aes(x = longByss$Smoking, fill = factor(longByss$Byss))) +
  geom_bar(position = position_dodge(width = 0.9), stat = "count") +
  geom_text(stat = 'count', aes(label = ..count..),
           vjust = -0.5, position = position_dodge(width = 0.9)) +
  labs(title = "Byssinosis Frequency by smoking", x = "Smoking or not",
           y = "Frequency")
#### Sex & Byss --- plot3
ggplot(longByss, aes(x = longByss$Sex, fill = factor(longByss$Byss))) +
  geom_bar(position = position_dodge(width = 0.9), stat = "count") +
  geom_text(stat = 'count', aes(label = ..count..),
           vjust = -0.5, position = position_dodge(width = 0.9)) +
  labs(title = "Byssinosis Frequency by Sex", x = "Sex(Female & Male)",
           y = "Frequency")
#### Race & Byss --- plot4
ggplot(longByss, aes(x = longByss$Race, fill = factor(longByss$Byss))) +
  geom_bar(position = position_dodge(width = 0.9), stat = "count") +
  geom_text(stat = 'count', aes(label = ..count..),
           vjust = -0.5, position = position_dodge(width = 0.9)) +
  labs(title = "Byssinosis Frequency by Race", x = "Race (Others & White)",
           y = "Frequency")
#### Race & Byss --- plot5
ggplot(longByss, aes(x = longByss$Workspace, fill = factor(longByss$Byss))) +
  geom_bar(position = position_dodge(width = 0.9), stat = "count") +
  geom_text(stat = 'count', aes(label = ..count..),
           vjust = -0.5, position = position_dodge(width = 0.9)) +
  labs(title = "Byssinosis Frequency by Workspace",
           x = "Workspace[1 (most dusty), 2 (less dusty), 3 (least dusty)]",
           y = "Frequency")
empty_model  = glm(longByss$Byss ~ 1, data = longByss, family = binomial())
summary(empty_model)
BIC(empty_model)

in_full_model <- glm(longByss$Byss~.^2,data=longByss,
           family = binomial())
summary(in_full_model)

in_forward.model.AIC = stepAIC(empty_model,
           scope = list(lower = empty_model, upper= in_full_model),k = 2,
```

```
        direction = "forward", trace = FALSE)

in_backward.model.AIC = stepAIC(in_full_model,
        scope = list(lower = empty_model, upper= in_full_model),
        k = 2,direction = "backward",trace = FALSE)

in_both.model.AIC = stepAIC(in_full_model,
        scope = list(lower = empty_model, upper= in_full_model),
        k = 2,direction = "both",trace = FALSE)

summary(in_forward.model.AIC)
summary(in_backward.model.AIC)
summary(in_both.model.AIC)

in_forward.model.AIC$aic
in_backward.model.AIC$aic
in_both.model.AIC$aic
in_forward.model.BIC = stepAIC(empty_model,
                scope = list(lower = empty_model, upper= in_full_model),
                k = log(n_new),direction = "forward", trace = FALSE)

in_backward.model.BIC = stepAIC(in_full_model, scope = list(lower = empty_model,
        upper= in_full_model), k = log(n_new),direction = "backward",trace = FALSE)

in_both.model.BIC = stepAIC(in_full_model, scope = list(lower = empty_model,
        upper= in_full_model), k = log(n_new),direction = "both",trace = FALSE)

summary(in_forward.model.BIC)
summary(in_backward.model.BIC)
summary(in_both.model.BIC)

in_forward.model.BIC$aic
in_backward.model.BIC$aic
in_both.model.BIC$aic


anova(in_forward.model.AIC, in_backward.model.AIC,in_both.model.AIC )
anova(in_forward.model.BIC,in_backward.model.BIC,in_both.model.BIC )

BIC_choice = glm(formula = longByss$Byss ~ Smoking + Workspace, family = binomial(),
        data = longByss)

AIC_choice = glm(formula = longByss$Byss ~ Workspace + Smoking + Employment +
        Workspace:Smoking, family = binomial(), data = longByss)
```

```r
library(ggplot2)
deviance_values <- c(deviance(AIC_choice), deviance(BIC_choice))
model_names <- c("AIC_choice", "BIC_choice")

plot_data <- data.frame(Model = model_names, Deviance = deviance_values)

ggplot(plot_data, aes(x = Model, y = Deviance)) +
  geom_bar(stat = "identity", fill = "lightblue") +
  theme_minimal() +
  labs(title = "Model Comparison by Deviance", x = "Model", y = "Deviance")

result_BIC_FOR <- step(glm(Byss~1, binomial, longByss),

            scope = ~Employment*Smoking*Sex*Race*Workspace,

            k=log(5417),

            trace=0,

            direction = "forward")

result_AIC_FOR <- step(glm(Byss~1, binomial, longByss),

            scope = ~Employment*Smoking*Sex*Race*Workspace,

            k=2,

            trace=0,

            direction = "forward")
```