Final Project

Wyatt Workman, Jasper Tsai, Karl Jang, Jacob Herbstman, Aditi Goyal

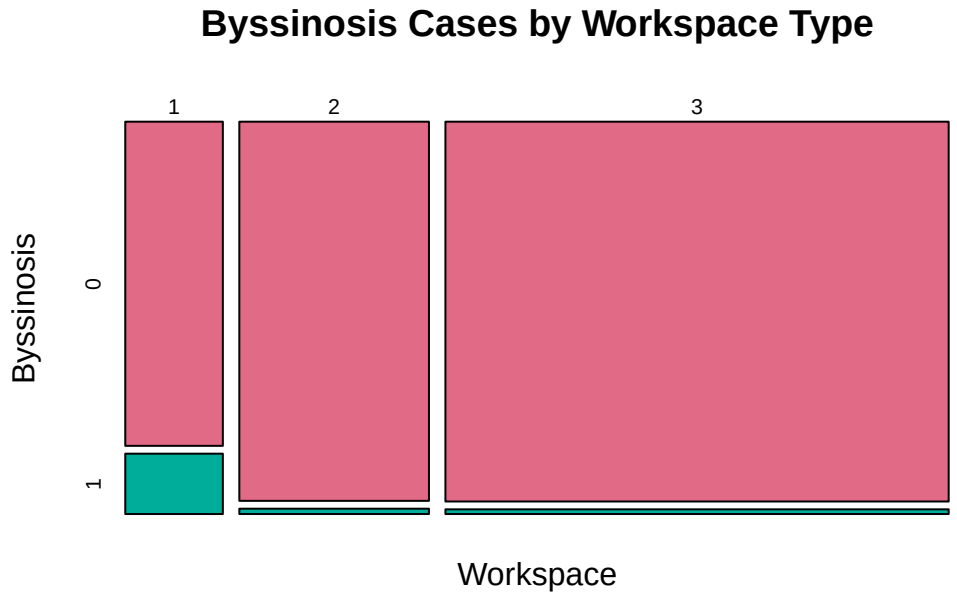
12/10/2021

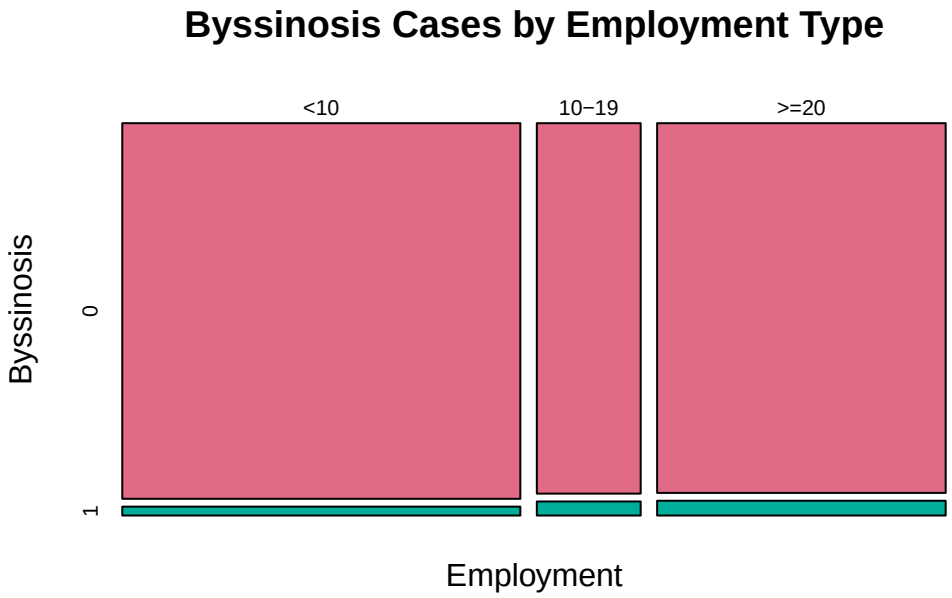
Introduction

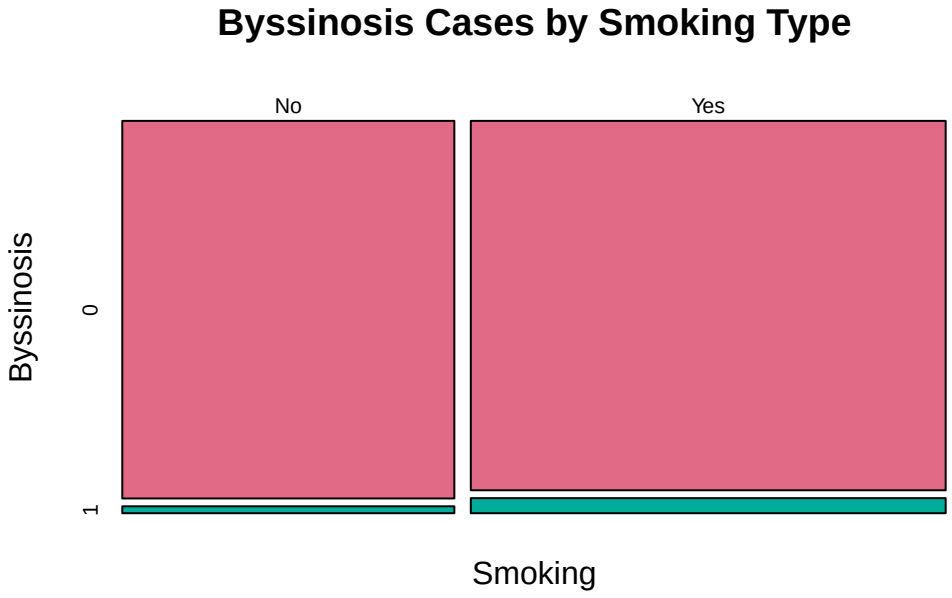
In this report, we examine some of the possible causes of byssinosis in cotton textile factories. Byssinosis is a pneumoconiosis disease, similar to pneumonia. Due to their exposure to small dust particles as a result of cotton processing. it is commonly found in cotton textile plant workers. We begin this report by exploring the data and uncovering any correlations between cases of byssinosis and various predictor variables. Then, in order to conduct inference on the most important predictors of byssinosis, we conduct model selection and fit the selected model.

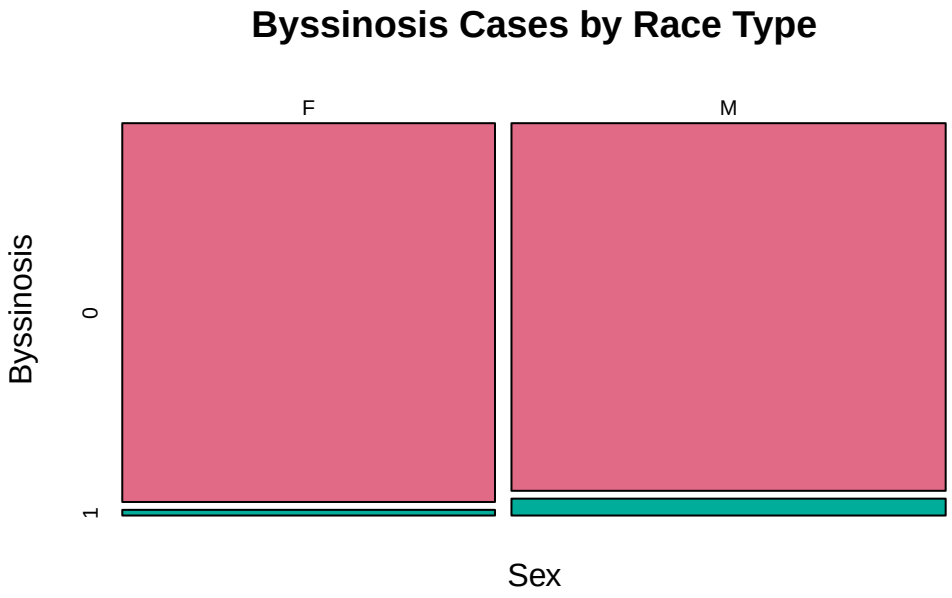
Data Visualization and Exploration

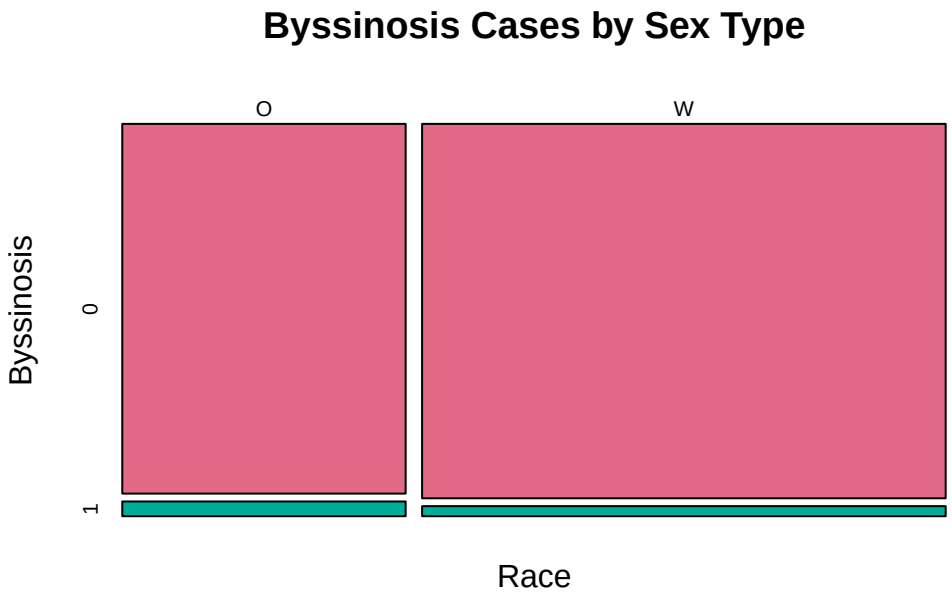
First, we plot each covariate and corresponding factor level to visualize how many cases of byssinosis occurred in each factor level.




Byssinosis Cases by Employment Type
<10
10−19
>=20
Byssinosis
0
1
Employment

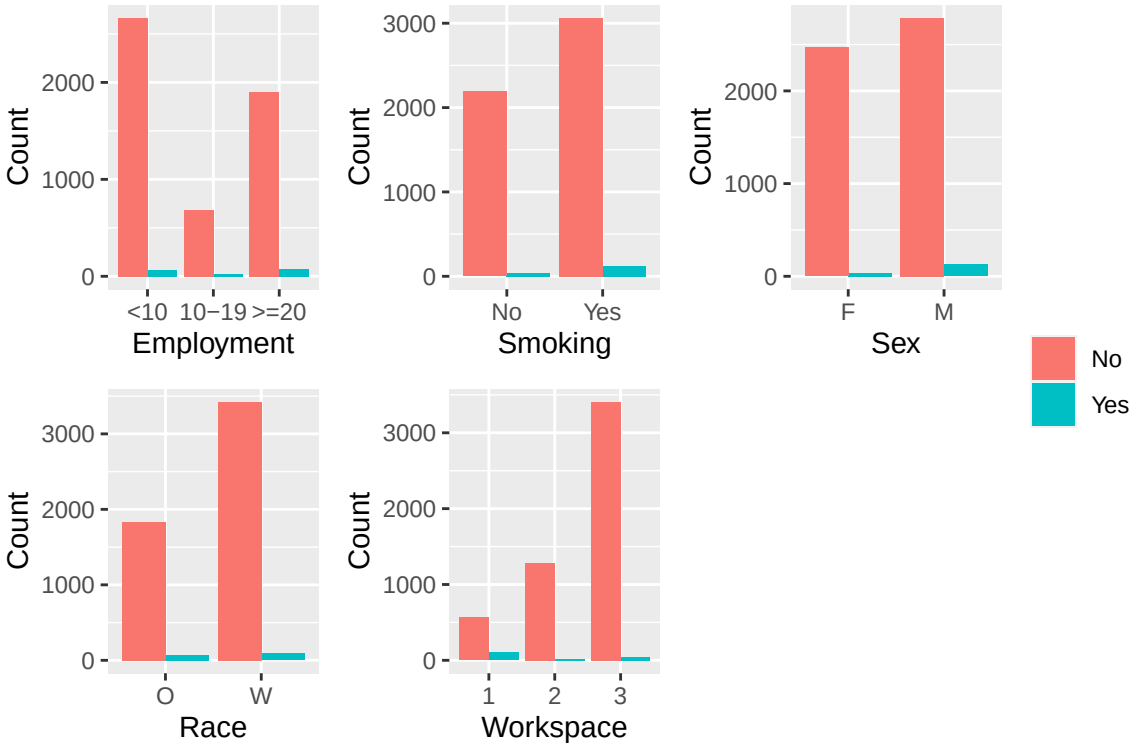

Byssinosis Cases by Smoking Type
No
Yes
Byssinosis
0
1
Smoking


Byssinosis Cases by Race Type
F
M
Byssinosis
0
1
Sex


Byssinosis Cases by Sex Type
O
W
Byssinosis
0
1
Race

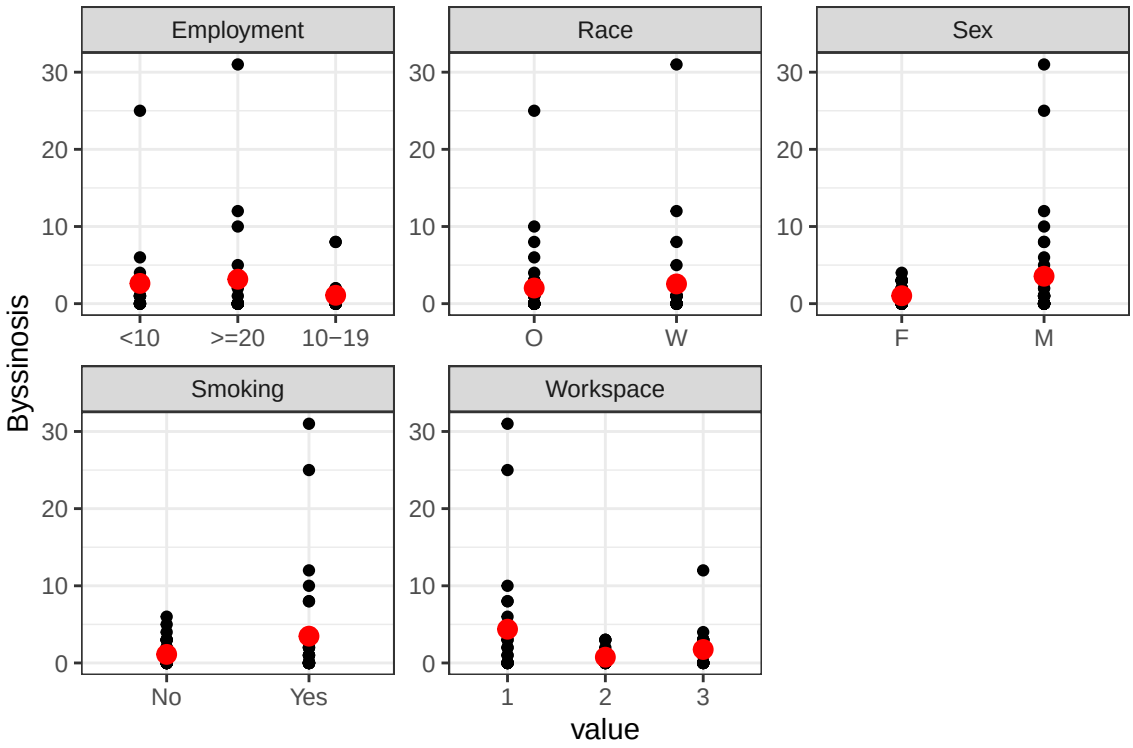
We generated several mosaic plots, one for each predictor, to examine the relationship between each one and our response variable. Our mosaic plots display the proportion of 0s and 1s for our response by each predictor factor level. From these, we can see that workspace type seems to be relevant to whether someone contracts byssinosis or not. Along with this, we see that smoking and sex could also matter, however sex could simply be a selection effect from a mostly male workforce in these factories. Overall, we would expect workspace, smoking, and possibly sex to be significant in a logistic regression framework. Type of employment and race seem to have more even proportions across each factor, so we would expect those to potentially matter less.

In order to further visualize the relationships between the response and predictors, we produce the following bar plot, which shows the number of byssinosis and non-byssinosis cases for each predictor.



To verify our intuition from the mosaic plots, we also made side-by-side bar plots. A similar picture from the variables is detected in that sense, with variation across the factors being the most prominent in smoking, workspace, and sex.

To further demonstrate the strength of these observations, we also produce a plot to show these relations for the data in the wide format.



This plot shows the number of positive byssinossis cases for each set of conditions. By looking at the data, we know that for every combination of people sampled, there were a certain number of positive, and certain number of negative cases. This graph only illustrates the positive cases. The red dot represents the mean number of cases per category This plot informs us briefly about the spread of the data, and further highlights the differences in the averages

To gain a more thorough understanding of how the covariates are related to each other, we examine the Cramér's V value of each covariate to the response variable. Cramér's V is a numerical value calculated from the Pearson's Chi-Square test for the goodness of fit, and quantifies the relation between two variables. It is synonymous with correlation values in the linear regression setting (such as Pearson's correlation coefficient). Values of Cramér's V that are greater than 0.2 in magnitude are said to be moderately related. The following table contains the Cramér's V between the response and each covariate:

```
##     variable correlation_coefficient
## 1       Race              0.03275291
## 2        Sex              0.08339847
## 3    Smoking              0.05980021
## 4 Employment              0.04330082
## 5  Workspace              0.27633990
```

We can see that Workspace is the most correlated factor to the response. This shows us that workplace is a very important predictor of byssinosis cases, and that we should expect to see this predictor in our logistic model.

Model Selection

Next, to find the best model, we use bidirectional stepwise selection on the training dataset. We use the Bayesian Information Criterion (BIC) as the selection criterion. Using BIC typically yields a simpler model when compared to the Akaike Information Criterion (AIC), enabling easier inference and interpretations, which is the objective of this report.

```
## Start:  AIC=817.82
## y ~ 1
##
##              Df Deviance    AIC
## + Workspace   2   661.18 684.89
## + Sex         1   783.75 799.56
## + Smoking     1   796.06 811.86
## <none>            809.91 817.82
## + Race        1   807.86 823.67
## + Employment  2   805.06 828.77
##
## Step:  AIC=684.89
## y ~ Workspace
##
##              Df Deviance    AIC
## + Smoking     1   652.01 683.63
## <none>            661.18 684.89
## + Race        1   659.25 690.86
## + Sex         1   660.69 692.31
## + Employment  2   655.17 694.69
## - Workspace   2   809.91 817.82
##
## Step:  AIC=683.63
## y ~ Workspace + Smoking
##
##                      Df Deviance    AIC
## <none>                    652.01 683.63
## - Smoking             1   661.18 684.89
## + Race                1   650.25 689.77
## + Sex                 1   652.00 691.53
## + Smoking:Workspace   2   644.40 691.83
## + Employment          2   646.49 693.92
## - Workspace           2   796.06 811.86
```

The best model according to bidirectional stepwise selection, on the basis of BIC, is: $y = \beta_1 Workspace + \beta_2 Smoking$. As expected, workplace is one of the predictors, as well as smoking. This is somewhat surprising, as smoking had a low Cramér's V value. Next, we fit this model on the test dataset, and interpret the slope coefficients.

Model Fit and Inference

Next, we fit this chosen model on the test data in order to perform statistical tests on the significance of the model predictor variables in the population.

```
##               Estimate Std. Error   z value     Pr(>|z|)
## (Intercept) -2.2074129  0.2665977 -8.279939 1.232385e-16
## Workspace2  -2.3485360  0.3739818 -6.279813 3.389802e-10
## Workspace3  -2.6178742  0.2830212 -9.249746 2.250263e-20
## SmokingYes   0.4989459  0.2782615  1.793082 7.295975e-02
```

Conclusions

Using the summary output above, we can use multiple Wald test to examine whether or not the slopes are significant in the population. We also use a Bonferroni correction of $\frac{0.05}{3} = 0.0167$, in order to control for type 1 error while conducting simultaneous tests on each slope coefficient. Note that we are not particularly interested in alpha, so we will not test for it. Thus, g = 3 when computing the Bonferroni correction.

Based on the p-values for the Wald tests, we can conclude that both workspaces 2 and 3 are statistically significant with 95% confidence, while smoking is not statistically significant.

We thus conclude that, when compared to those in workspace 1 (most dusty) employees in workspaces 2 and 3 are -2.348536 less likely to contract byssinosis. We also conclude that, when compared to those in workspace 1, employees at other workspaces are -2.6178742 less likely to contract byssinosis. Lastly, since the coefficient of workspace 3 is less than that of workspace 2, hence minimizing the risk of byssinosis, we conclude that workplace 3 is the safest workspace in terms of byssinosis.

Code Appendix:

```
knitr::opts_chunk$set(echo = TRUE)
knitr::opts_chunk$set(fig.width = 6, fig.height = 4)
# Read in data:
dat = read.csv("Byssinosis.csv")
# convert categorical variables into factors to enable conducting model
# selection:
dat$Employment = factor(dat$Employment, levels = c("<10", "10-19", ">=20"), ordered = T)
dat$Smoking = as.factor(dat$Smoking)
dat$Sex = as.factor(dat$Sex)
dat$Race = as.factor(dat$Race)
dat$Workspace = as.factor(dat$Workspace)
# Convert data to long format, to fit a model more accurately:
library(tidyr)
library(dplyr)
long_dat = dat %>%
    pivot_longer(dat, cols = c("Byssinosis", "Non.Byssinosis"), names_to = "disease_type",
        values_to = "response") %>%
    uncount() %>%
    mutate(y = ifelse(disease_type == "Byssinosis", 1, 0))
# Extract important columns: long_dat = subset(long_dat, select =
# c('Employment', 'Smoking', 'Sex', 'Race', 'Workspace', 'y'))

# Subset test and training data:
set.seed(123456789)
samp <- sample(nrow(long_dat), 2709)
train = long_dat[samp, ]
test <- long_dat[-samp, ]

colors <- colorspace::qualitative_hcl(2)

tab1 = table(long_dat$Workspace, long_dat$y)
plot1 <- mosaicplot(tab1, xlab = "Workspace", ylab = "Byssinosis", main = "Byssinosis Cases by Workspac
    col = colors)

tab2 = table(long_dat$Employment, long_dat$y)
plot2 <- mosaicplot(tab2, xlab = "Employment", ylab = "Byssinosis", main = "Byssinosis Cases by Employm
    col = colors)

tab3 = table(long_dat$Smoking, long_dat$y)
plot3 <- mosaicplot(tab3, xlab = "Smoking", ylab = "Byssinosis", main = "Byssinosis Cases by Smoking Ty
    col = colors)

tab4 = table(long_dat$Sex, long_dat$y)
plot4 <- mosaicplot(tab4, xlab = "Sex", ylab = "Byssinosis", main = "Byssinosis Cases by Race Type",
    col = colors)
tab5 = table(long_dat$Race, long_dat$y)
plot5 <- mosaicplot(tab5, xlab = "Race", ylab = "Byssinosis", main = "Byssinosis Cases by Sex Type",
    col = colors)
# data for barplots: before pivot_longer gives you, yes/no on separate column
Employment.dat <- dat %>%
    group_by(Employment) %>%
    summarise(Yes = sum(Byssinosis), No = sum(Non.Byssinosis)) %>%
```

```
    pivot_longer(., -Employment, names_to = "Byssinosis", values_to = "Count")

Smoking.dat <- dat %>%
    group_by(Smoking) %>%
    summarise(Yes = sum(Byssinosis), No = sum(Non.Byssinosis)) %>%
    pivot_longer(., -Smoking, names_to = "Byssinosis", values_to = "Count")

Sex.dat <- dat %>%
    group_by(Sex) %>%
    summarise(Yes = sum(Byssinosis), No = sum(Non.Byssinosis)) %>%
    pivot_longer(., -Sex, names_to = "Byssinosis", values_to = "Count")

Race.dat <- dat %>%
    group_by(Race) %>%
    summarise(Yes = sum(Byssinosis), No = sum(Non.Byssinosis)) %>%
    pivot_longer(., -Race, names_to = "Byssinosis", values_to = "Count")

Workspace.dat <- dat %>%
    group_by(Workspace) %>%
    summarise(Yes = sum(Byssinosis), No = sum(Non.Byssinosis)) %>%
    pivot_longer(., -Workspace, names_to = "Byssinosis", values_to = "Count")
library(ggpubr)
Employment_hist <- ggplot(Employment.dat, aes(x = Employment, y = Count, fill = Byssinosis)) +
    geom_bar(stat = "identity", position = position_dodge()) + theme(legend.title = element_blank()) +
    theme(legend.position = "none")

Smoking_hist <- ggplot(Smoking.dat, aes(x = Smoking, y = Count, fill = Byssinosis)) +
    geom_bar(stat = "identity", position = position_dodge()) + theme(legend.title = element_blank()) +
    theme(legend.position = "none")

Sex_hist <- ggplot(Sex.dat, aes(x = Sex, y = Count, fill = Byssinosis)) + geom_bar(stat = "identity",
    position = position_dodge()) + theme(legend.title = element_blank()) + theme(legend.position = "none

Race_hist <- ggplot(Race.dat, aes(x = Race, y = Count, fill = Byssinosis)) + geom_bar(stat = "identity"
    position = position_dodge()) + theme(legend.title = element_blank()) + theme(legend.position = "none

Workspace_hist <- ggplot(Workspace.dat, aes(x = Workspace, y = Count, fill = Byssinosis)) +
    geom_bar(stat = "identity", position = position_dodge()) + theme(legend.title = element_blank()) +
    theme(legend.position = "none")

ggarrange(Employment_hist, Smoking_hist, Sex_hist, Race_hist, Workspace_hist, ncol = 3,
    nrow = 2, common.legend = TRUE, legend = "right")
dat %>%
    gather(-Byssinosis, -Non.Byssinosis, key = "var", value = "value") %>%
    ggplot(aes(x = value, y = Byssinosis)) + geom_point() + stat_summary(fun = mean,
    geom = "point", size = 3, color = "red", fill = "red") + facet_wrap(~var, scales = "free") +
    theme_bw()
library(lsr)
t1 = cramersV(table(long_dat$Race, long_dat$y))

t2 = cramersV(table(long_dat$Sex, long_dat$y))

t3 = cramersV(table(long_dat$Smoking, long_dat$y))
```

```
t4 = cramersV(table(long_dat$Employment, long_dat$y))

t5 = cramersV(table(long_dat$Workspace, long_dat$y))

variable = c("Race", "Sex", "Smoking", "Employment", "Workspace")
correlation_coefficient = c(t1, t2, t3, t4, t5)
corr_values = data.frame(variable, correlation_coefficient)
corr_values
library(MASS)
null_model = glm(y ~ 1, data = train, family = binomial)
full_model = glm(y ~ Employment * Smoking * Sex * Race * Workspace, data = train,
    family = binomial)

best_model = stepAIC(null_model, direction = "both", scope = list(lower = null_model,
    upper = full_model), k = log(2709))
model = glm(y ~ Workspace + Smoking, data = test, family = binomial)
summary(model)$coefficients
```