STAT200Project

January 11, 2025

1 STAT 200 Project (120 points)

JARED DE LA SERNA

For your project, you must analyze data from the COPDGene study using the concepts covered in this course. Please follow the directions below carefully:

- Tasks indicated in red must be completed to receive credit
- Include all code for your work
- Comment all code using the `#`. This is a must!
- Include markdown cells with written answers explaining your work when prompted
- **NOTE: Your completed project must be submitted to Canvas as a .ipynb file by the assigned due date**

1.1 Data

Data for the project is available here. A data dictionary describing the meaning of each of the variables in the dataset is available in the Project module in Canvas.

1.2 Introduction

Chronic obstructive pulmonary disease (COPD) affects over 16 million Americans and is the fourth leading cause of death in the United States behind heart disease, cancer, and accidental death. While COPD can result from various toxic inhalations or asthma, it is most commonly associated with cigarette smoking.

COPD severity is typically measured by a device called a spirometer. Patients forcefully exhale into the device and the volume of air exhaled is used as a measure for the severity of disease (less air exhaled ⇒ worse disease). Data collected by the COPDGene research group includes spirometry data on thousands of research participants.

Spirometry measures in the dataset: * The forced expiratory volume (`FEV1`) is the volume of air exhaled in 1 second * The forced vital capacity (`FVC`) is the total volume of air exhaled after a full breath * `FEV1_FVC_ratio` is the ratio between `FEV1` and `FVC` (smaller ⇒ worse disease) * `FEV1_phase2` is the `FEV1` of research participants 5 years later

You overall task in this project is to analyze the relationship between FEV1 at follow-up `FEV1_phase2` and other variables in the dataset. The project has been organized into a series of tasks to assist you with your analysis organization.

####Organized (5 points) * Answers should be organized in the following format for organization and readability

```
# 1.
This will be the code block to answer the first part of a task
```

1. This will be the text answer/explanation (if prompted) for the first task

```
# 2.
This will be the code block to answer the second part of a task in a separate code cell
```

2. This will be the text answer/explanation (if prompted) for the second task

Continue this format for the remaining subtasks

####Task 1 (5 points)

1. Load the COPDGene dataset and show the first few lines.
2. Remove the NAs from the dataset and store as `dat1`. How many rows does `dat1` contain? Answer in a complete sentence.
3. Select all rows in the dataset where `FEV1_phase2` is `NA` and store as `dat2`. How many rows does `dat2` contain? Answer in a complete sentence.

TASK 1.1

[ ]:
```
project <- read.csv("https://raw.githubusercontent.com/khasenst/
↪datasets_teaching/main/copd_data_project.csv")
```

4000

TASK 1.2

[ ]:
```
dat1 <- na.omit(project)
nrow(dat1)
```

TASK 1.2 Dat1 has 4000 rows.

[ ]:
```
dat2 <- project[is.na(project$FEV1_phase2),]
nrow(dat2)
```

1747

TASK 1.3

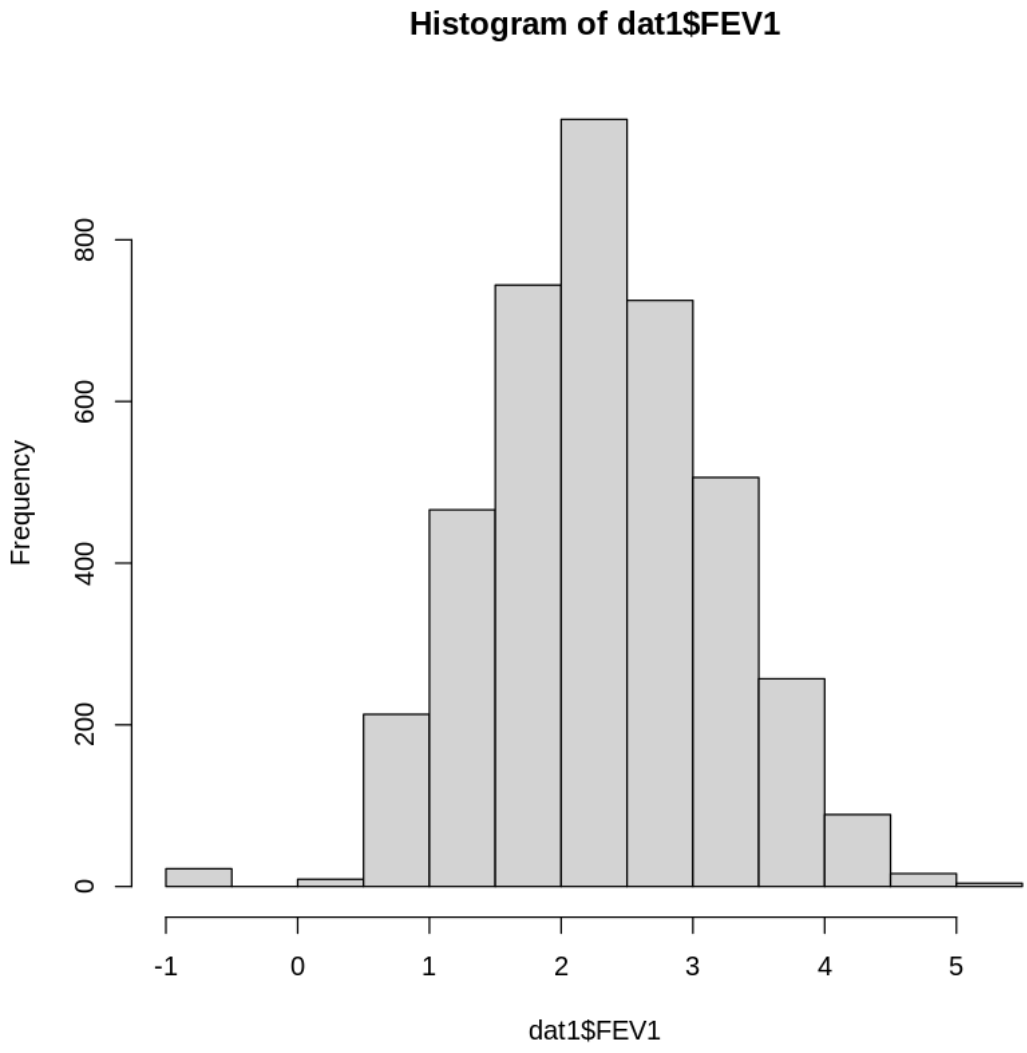
Dat2 has 1747 rows.

####Task 2 (10 points)

1. From `dat1`, plot the histogram of `FEV1`
2. Calculate the percentage of `FEV1` values within one standard deviation of its mean. Answer in a complete sentence.
3. Calculate the percentage of `FEV1` values within 2 standard deviations of its mean. Answer in a complete sentence.
4. Use your answers from 2.1-2.3 to determine if the FEV1 variable is approximately normally distributed.

TASK 2.1

```
[ ]: hist(dat1$FEV1)
```



TASK 2.2

[ ]:
```
mean <- mean(dat1$FEV1)
sd <- sd(dat1$FEV1)
```

TASK 2.2

[ ]:
```
first_interval <- (mean + (-1 * sd))
second_interval <- (mean + (1 * sd))
paste("(", first_interval, "," , second_interval ,")")
allvalues <- dat1[dat1$FEV1 > first_interval & dat1$FEV1 < second_interval , ]
num <- nrow(allvalues)
den <- nrow(dat1)
percent <- num/den
percent
```

'(1.43701559163072 , 3.18428490836928)'

0.67575

TASK 2.2

The 68% confidence interval is between [1.4370 and 3.1843]. 68% is the percentage of FEV1 values with one standard deviation of its mean.

TASK 2.3

[ ]:
```
first_interval <- (mean + (-2 * sd))
second_interval <- (mean + (2 * sd))
paste("(", first_interval, "," , second_interval ,")")
allvalues <- dat1[dat1$FEV1 > first_interval & dat1$FEV1 < second_interval , ]
num <- nrow(allvalues)
den <- nrow(dat1)
percent <- num/den
percent
```

'(0.563380933261435 , 4.05791956673857)'

0.9675

TASK 2.3

The around 96% confidence interval is between [0.5634 and 4.0580]. 96% is the percentage of FEV1 values with two standard deviations of its mean.

TASK 2.4

The dat1$FEV1 in the histogram shows that it is normally distributed, but the values with one standard deviation are around 67%, and with two are around 96%, which are the values that are mostly normally distributed.

####Task 3 (10 points)

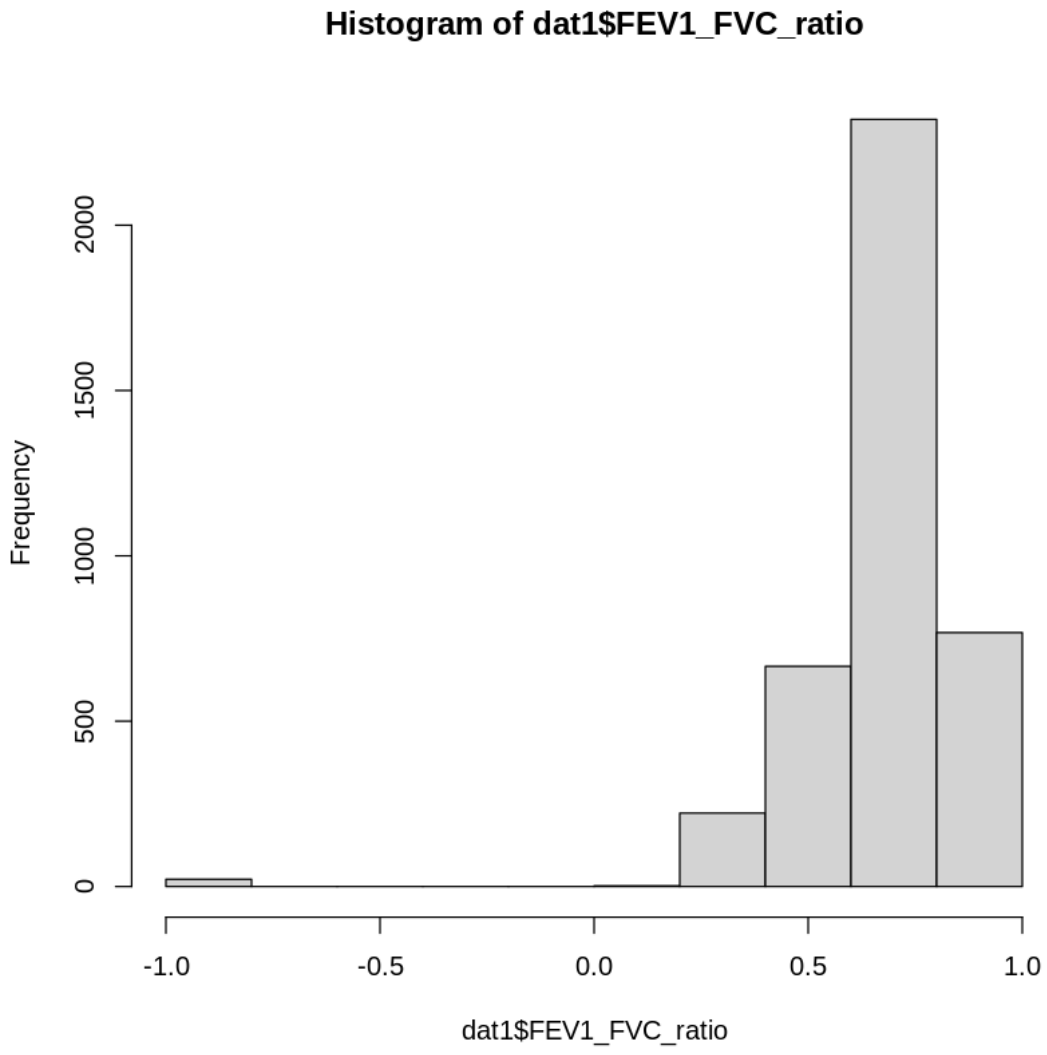
1. From `dat1`, find two other numeric variables and plot their histograms.
2. Describe the histogram of the FIRST variable you chose.

- Range, standard deviation, skewness or symmetry, mean, normally distributed, outlier observations etc.
- You may use R functions to help you answer and describe the distribution
- Answer in complete sentences in a paragraph

3. Describe the histogram of the SECOND variable you chose.
   - Range, standard deviation, skewness or symmetry, mean, normally distributed, outlier observations etc.
   - You may use R functions to help you answer and describe the distribution
   - Answer in complete sentences in a paragraph

TASK 3.1

[ ]:
```
hist(dat1$FEV1_FVC_ratio)
```



TASK 3.2

```
[ ]: sd(dat1$FEV1_FVC_ratio)
range(dat1$FEV1_FVC_ratio)
mean(dat1$FEV1_FVC_ratio)
```

0.187180763593818

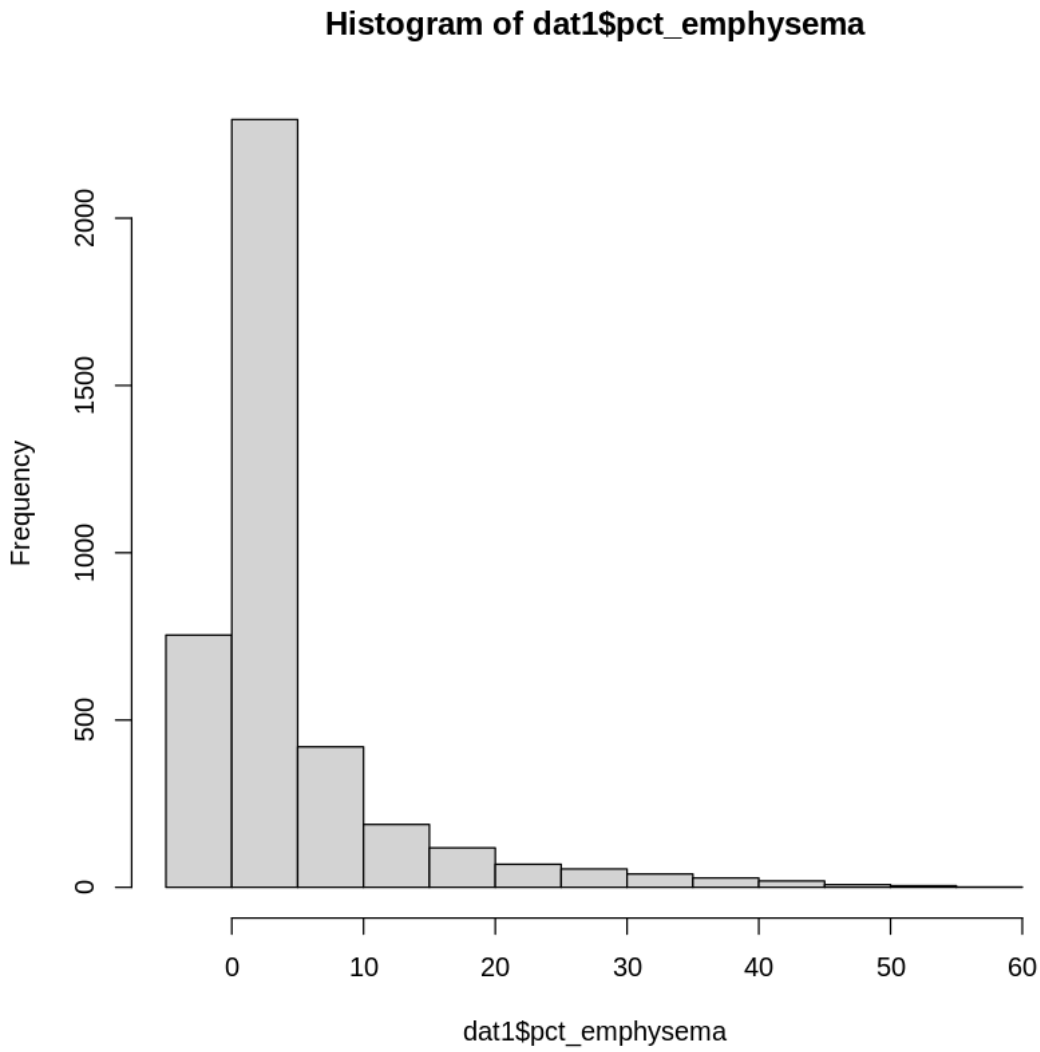
1. -1 2. 1

0.68047

The range of FEV1_FVC_Ratio from dat1 ranges from -1 to 1. The standard deviation is 0.18718. The mean is 0.68047. We can tell the graph is left-skewed and not normally distributed. We can't tell if there are outliers because the frequency is so high, but I assume not.

TASK 3.3

```
[ ]: hist(dat1$pct_emphysema)
```



[ ]:
```
sd(dat1$pct_emphysema)
range(dat1$pct_emphysema)
mean(dat1$pct_emphysema)
```

7.9201930799684

1. -1 2. 56.0577

4.2189148085855

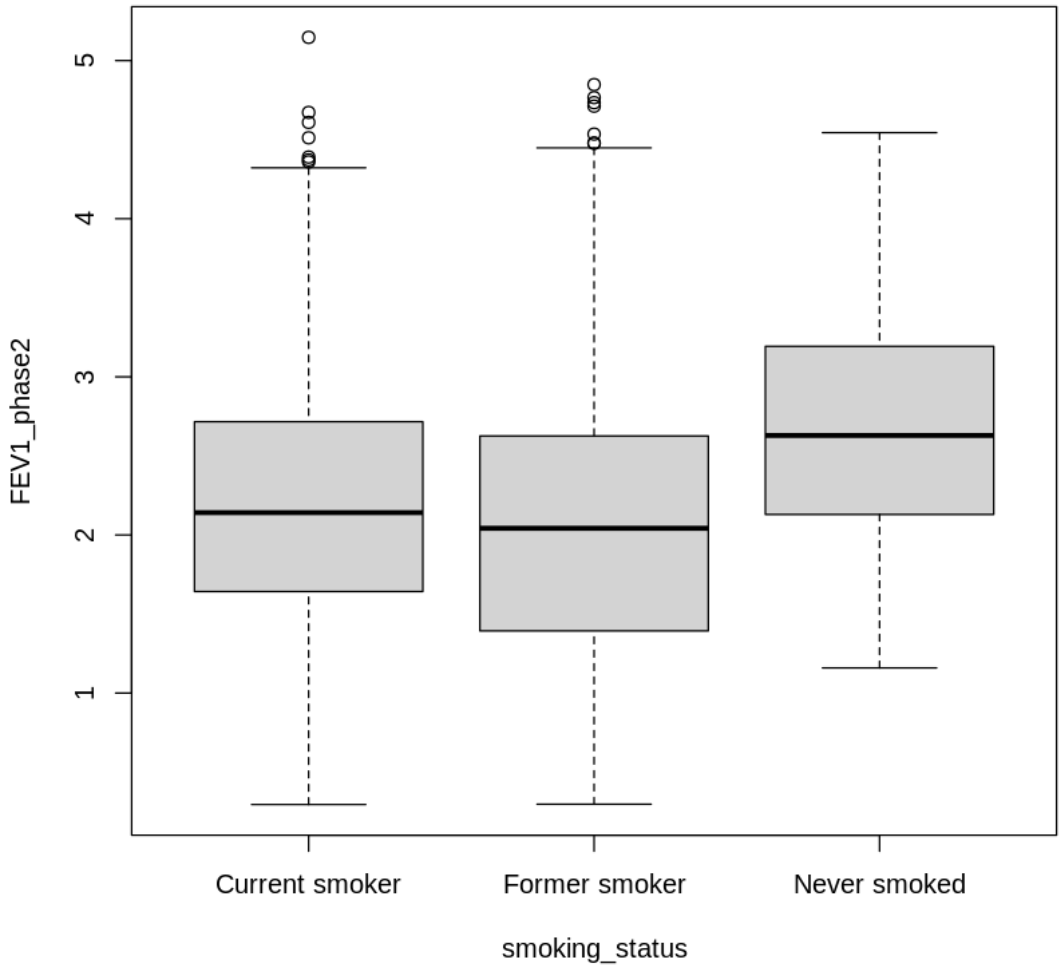
The range of PCT Emphysema from data1 is from -1 to 56.0577. The standard deviation is 7.92019. The mean is 4.21891. We can obvously tell that the graph is right-skewed. We can see that there is not outliers.

####Task 4 (10 points)

1. Using `dat1`, create a boxplot to visualize the relationship between `FEV1_phase2` and `smoking_status`.
2. Based on the boxplot, which group tends to have better breathing capacity?
3. Select rows in `dat1` corresponding to a `smoking_status` of `Current smoker` or `Former smoker`. Create a 95% confidence interval to determine if average FEV1 for phase 2 is different between current or former smokers. Explain your answer in complete sentences.

TASK 4.1

[ ]:
```
boxplot(FEV1_phase2 ~ smoking_status, data=dat1)
```



TASK 4.2

Based on the boxplot above, people who never smoked tend to have a higher FEV1_phase2, the volume of air forcefully exhaled in one secon

TASK 4.3

[ ]:
```
smokers <- dat1[dat1$smoking_status == "Current smoker" | dat1$smoking_status␣
↪== "Former smoker", ]
```

[ ]:
```
t.test(FEV1_phase2 ~ smoking_status, data = smokers)
```

```
        Welch Two Sample t-test

data:  FEV1_phase2 by smoking_status
t = 5.0794, df = 3933.1, p-value = 3.962e-07
alternative hypothesis: true difference in means between group Current smoker␣
↪and group Former smoker is not equal to 0
95 percent confidence interval:
 0.08155305 0.18408473
sample estimates:
mean in group Current smoker  mean in group Former smoker
                    2.179463                     2.046644
```

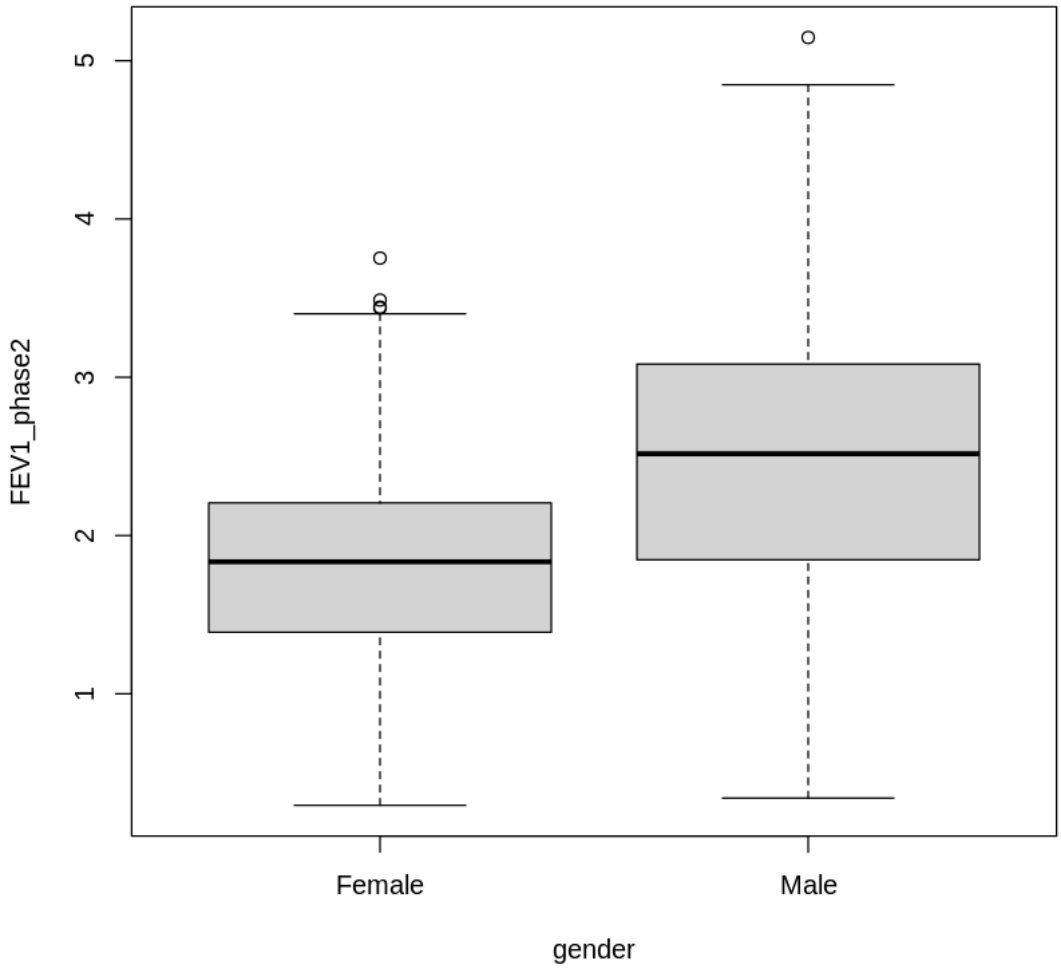
Our confidence interval is between [0.0816 and 0.1841]; zero is not between, so we can conclude that the difference average in FEV1_phase2 between smoker status is significant. Also, our p-value is less than 0.05, which makes it significant. There is enough evidence to reject the null. Making the difference significant.

####Task 5 (10 points)

1. Using `dat1`, create two more boxplots to visualize the relationship between `FEV1_phase2` and two other categorical variables.
2. For the first variable, use a 95% confidence interval to determine if the average `FEV1_phase2` is different between two categories.
3. For the second variable, use a 95% confidence interval to determine if the average `FEV1_phase2` in phase 2 is different between two categories.

TASK 5.1

[ ]:
```
boxplot(FEV1_phase2 ~ gender, data=dat1)
```



TASK 5.1

[ ]:
```
boxplot(FEV1_phase2 ~ race, data=dat1)
```

TASK 5.2

[ ]:
```
t.test(dat1$FEV1_phase2 ~ dat1$gender)
```

```
        Welch Two Sample t-test

data:  dat1$FEV1_phase2 by dat1$gender
t = -26.837, df = 3591.6, p-value < 2.2e-16
```

```
alternative hypothesis: true difference in means between group Female and group␣
 ↪Male is not equal to 0
95 percent confidence interval:
 -0.6918855 -0.5976740
sample estimates:
mean in group Female   mean in group Male
            1.796581             2.441361
```

According to our confidence interval between [-0.6918855 and -0.5876740], zero is not between, so we can conclude that the difference average in FEV1_phase2 between genders is significant. Also, our p-value is less than 0.05, which makes it significant. There is enough evidence to reject the null. Making the difference significant.

TASK 5.3

[ ]:
```
t.test(dat1$FEV1_phase2 ~ dat1$race)
```

```
        Welch Two Sample t-test

data:  dat1$FEV1_phase2 by dat1$race
t = -2.5896, df = 2514.5, p-value = 0.009663
alternative hypothesis: true difference in means between group Black or African␣
 ↪American and group White is not equal to 0
95 percent confidence interval:
 -0.12506982 -0.01728048
sample estimates:
mean in group Black or African American                     mean in group White
                                2.070865                                2.142040
```

Our confidence interval is between [-0.1251 and 0.0.017], and 0 is not between. However, the p-value is barely above 0.009663, making the average between FEV1_phase2 and race not significant.

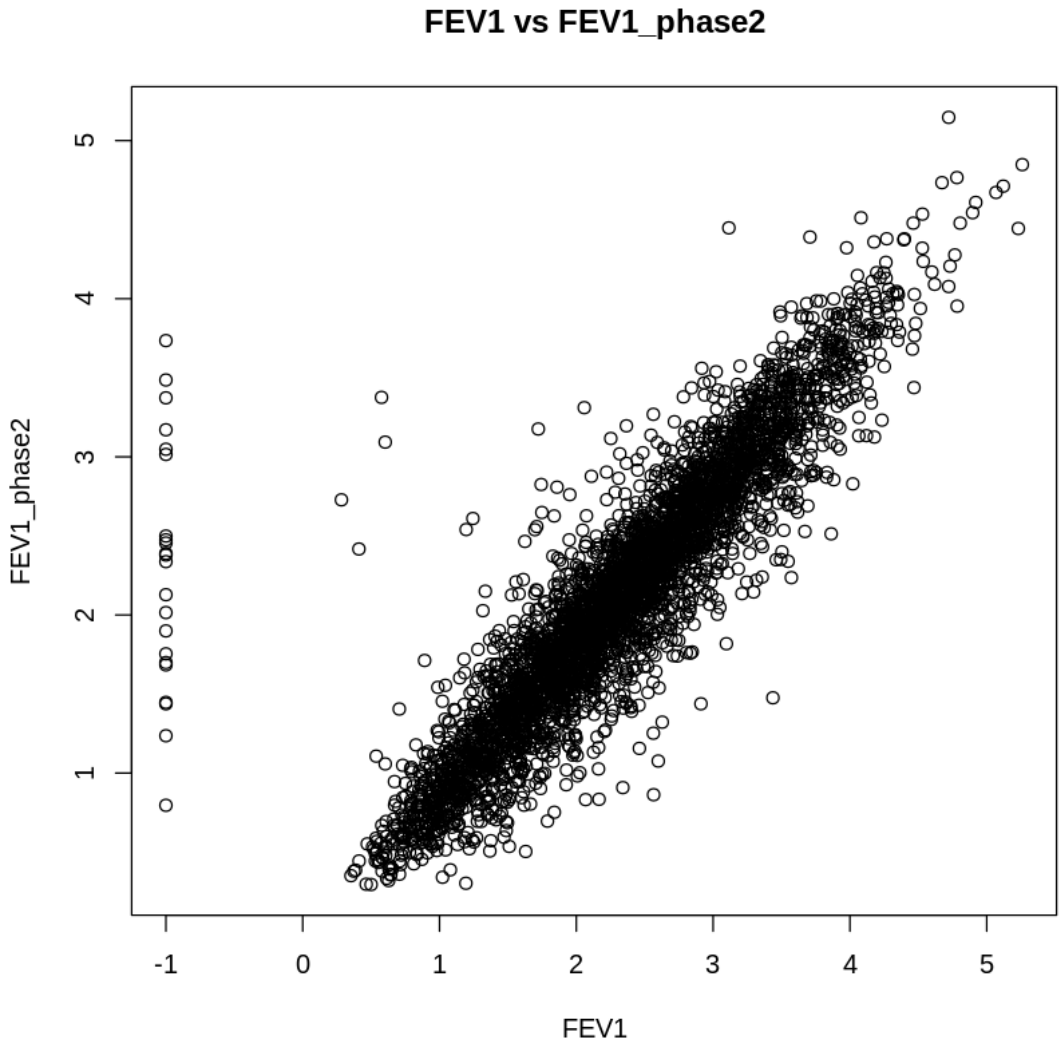
####Task 6 (20 points)

1. Using `dat1`, create a scatterplot visualizing the relationship between `FEV1_phase2` (y-axis) and `FEV1` (x-axis). Note that values that are -1 are missing observations. You may ignore this for this class.
2. What relationship do you observe based on the scatterplot?
3. Fit a simple linear regression model by regressing `FEV1_phase2` (Y) on `FEV1` (X). Summarize the regression.
4. Describe the R^2 of the regression in the context of its definition.
5. What is the slope and its interpretation in the context of `FEV1` and `FEV1_phase2`? Explain in complete sentences
6. What is the intercept and its interpretation in the context of `FEV1` and `FEV1_phase2`? Explain in complete sentences

7. Use a 95% confidence interval to determine if the slope is significantly less than 1. Explain what this means in terms of breathing health of the patients.

TASK 6.1

```
[ ]: plot(dat1$FEV1, dat1$FEV1_phase2, main = "FEV1 vs FEV1_phase2",
         xlab = "FEV1" ,
         ylab = "FEV1_phase2")
```



TASK 6.2

If we ignore the -1 values in the relationship between FEV1(x) and FEV1_phase2(y), we see a strong positive linear relationship.

TASK 6.3

[ ]:
```
model <- lm(FEV1_phase2 ~ FEV1, data=dat1)
summary(model)
```

```
Call:
lm(formula = FEV1_phase2 ~ FEV1, data = dat1)

Residuals:
    Min      1Q  Median      3Q     Max
-1.5909 -0.1880 -0.0144  0.1609  4.3971

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.178336   0.016943   10.53   <2e-16 ***
FEV1        0.840423   0.006859  122.53   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3789 on 3998 degrees of freedom
Multiple R-squared:  0.7897,        Adjusted R-squared:  0.7897
F-statistic: 1.501e+04 on 1 and 3998 DF,  p-value: < 2.2e-16
```

TASK 6.4

The R-SQUARED. is 0.7897, meaning that including FEV1 explains the 78.97% variability in FEV1_phase2.

TASK 6.5

For every unit that FEV1 increases, levels of FEV1_phase2 increase by about 0.840423.

TASK 6.6

When the FEV1 value is 0, the average value for FEV1_phase is 0.178336. Saying when FEV1 is nothing, the FEV1_phase is already at 0.178336.

TASK 6.7

[ ]:
```
confint(model)
```

A matrix: 2 × 2 of type dbl

	2.5 %	97.5 %
(Intercept)	0.1451177	0.2115549
FEV1	0.8269759	0.8538708

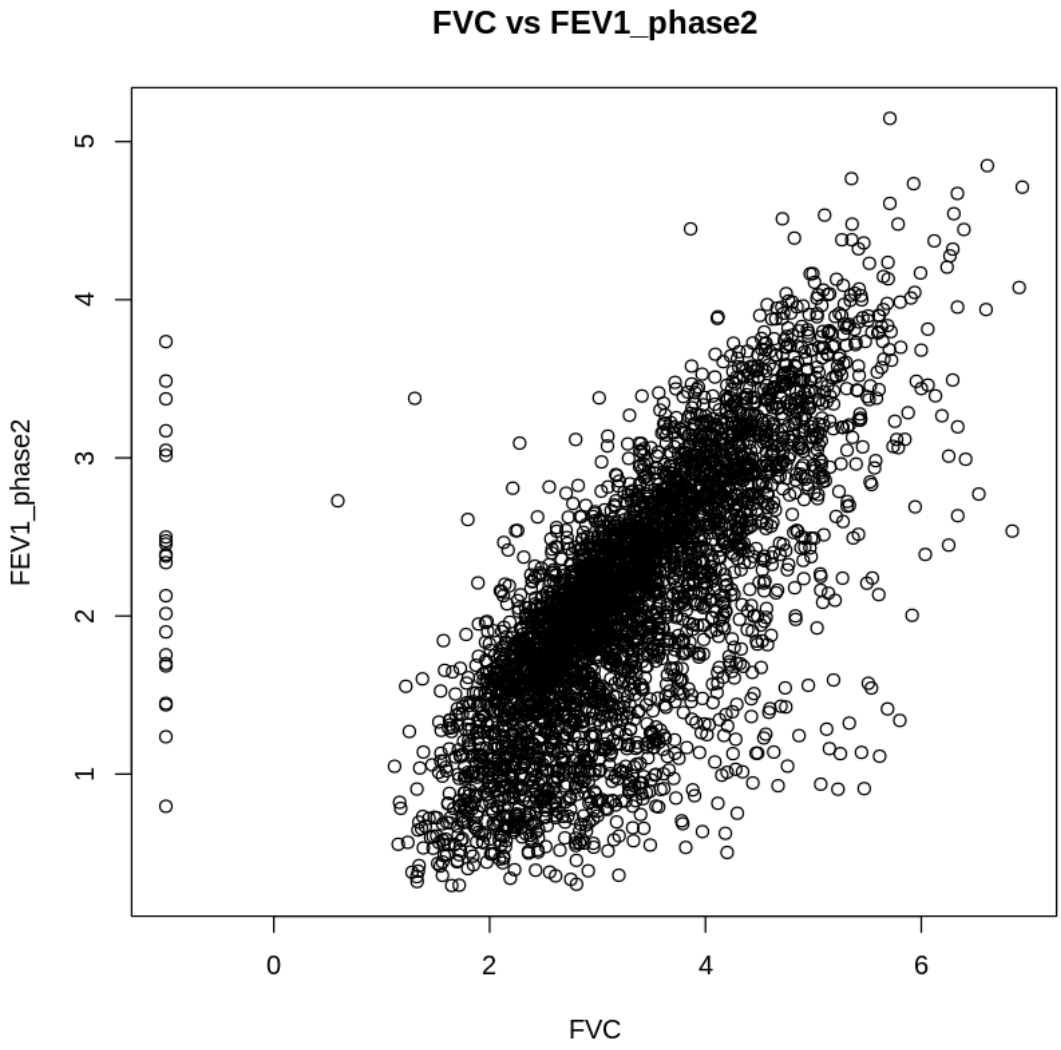
The confidence interval is between [0.8270 and 0.8539], which doesn't include 0, which makes it significant. Also, Looking at the p-value for FEV1, which is <2e-16, which is significantly less than 0.05, which means that the more FEV1 in patients increases significantly FEV1_phase2.

####Task 7 (10 points)

1. Identify two other variables in `dat1` that you think would be related to breathing health, as measured by `FEV1_phase2`. Plot these variables using scatterplots (`FEV1_phase2` on Y axis, variable 1 or variable 2 on X axis). What do you observe? Explain in complete sentences.
2. Include these two variables (categorical or numeric), along with `FEV1`, in your regression from Task 6. Show the regression summary. Keep `FEV1_phase2` as Y.
3. Using 95% confidence intervals, are the slopes for the two variables you selected significantly different from 0? Explain and interpret in complete sentences.

[ ]:
```
TASK 7.1
```

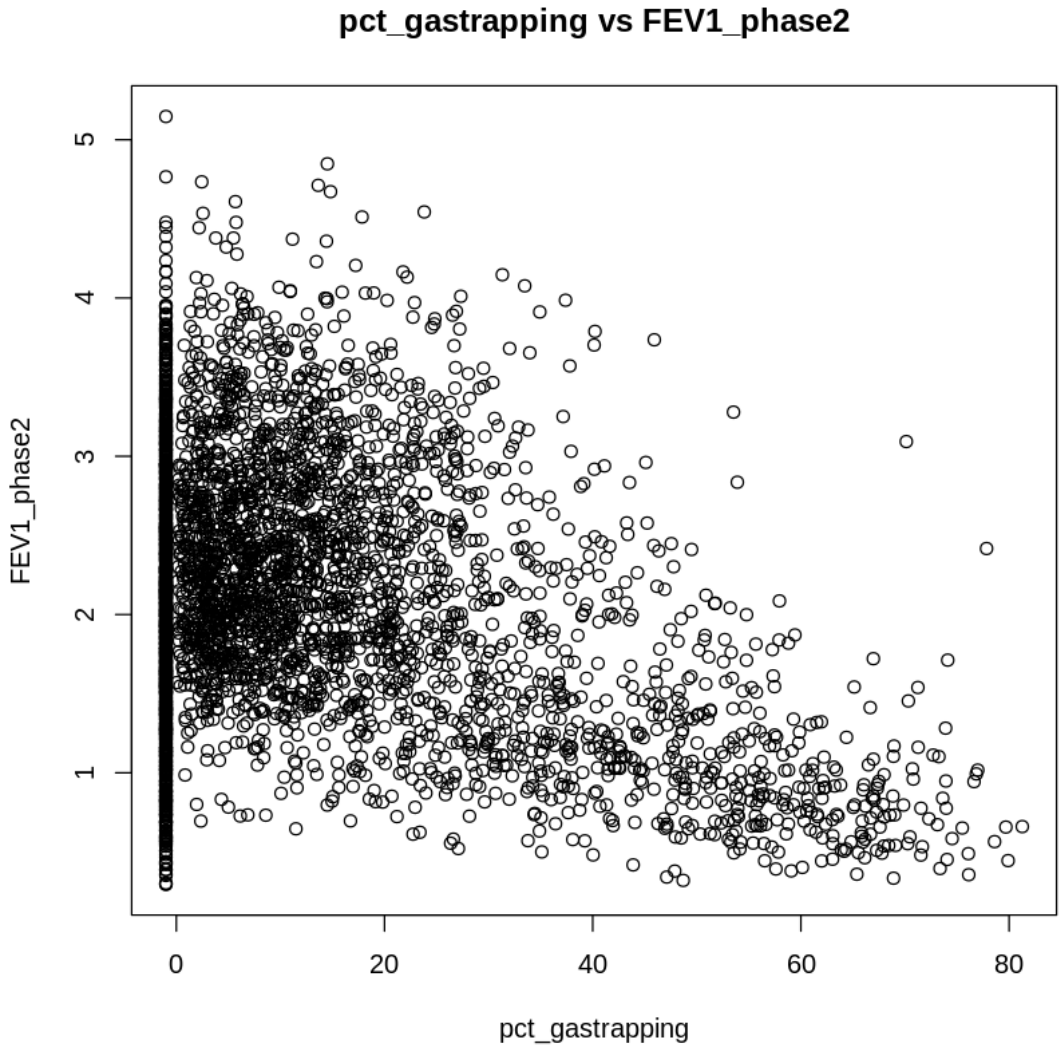
[ ]:
```
plot(dat1$FVC, dat1$FEV1_phase2,
main = "FVC vs FEV1_phase2",
    xlab = "FVC" ,
    ylab = "FEV1_phase2")
```



If we ignore the -1 values, I observe a strong positive correlation between FEV1_phase2 and FVC.

TASK 7.1

```
[ ]: plot(dat1$pct_gastrapping, dat1$FEV1_phase2,
     main = "pct_gastrapping vs FEV1_phase2",
         xlab = "pct_gastrapping" ,
         ylab = "FEV1_phase2")
```



If we ignore the -1 values, I observe a strong negative correlation between FEV1phase2 and pct_gastrapping.

TASK 7.2

[ ]:
```
lr <- lm(FEV1_phase2 ~ FEV1 + FVC + pct_gastrapping, data=dat1)
summary(lr)
```

```
Call:
lm(formula = FEV1_phase2 ~ FEV1 + FVC + pct_gastrapping, data = dat1)

Residuals:
    Min      1Q  Median      3Q     Max
-1.5882 -0.1867 -0.0254  0.1533  4.2137

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.3640558  0.0207171  17.573  < 2e-16 ***
FEV1             0.9741320  0.0159747  60.980  < 2e-16 ***
FVC             -0.1431752  0.0129131 -11.088  < 2e-16 ***
pct_gastrapping -0.0013438  0.0004107  -3.272  0.00108 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3686 on 3996 degrees of freedom
Multiple R-squared:  0.8011,        Adjusted R-squared:  0.801
F-statistic:  5366 on 3 and 3996 DF,  p-value: < 2.2e-16
```

TASK 7.3

[ ]:
```
fit <- lm(FEV1_phase2 ~ pct_gastrapping + FEV1, data = dat1)

confint(fit, level = 0.95)
```

A matrix: 3 × 2 of type dbl

	2.5 %	97.5 %
(Intercept)	0.246530222	0.323990384
pct_gastrapping	-0.004360018	-0.002951687
FEV1	0.801322349	0.829568137

Also, all 2 confidence intervals from pct_gastrapping [-0.004 to -0.003] and FEV1 [0.8013 to -0.82396], none of the 95% confidence intervals have 0 showing a significant relationship with FEV1_phase2. Look at the P -values for FEV1 and FVC P-values is <2e-16 which is significantly less than 0.05 Higher FEV1 significantly increases FEV1_phase2. Higher FEV1 significantly reduces FEV1_phase2. Even though the pct_gastrapping P-value is not significantly less than 0, it is less than <0.05, which still makes it significant, so we can say that pct_gastrapping significantly reduces FEV1_phase2.

####Task 8 (5 points)

1. Using your regression model with three variables from Task 7, predict the `FEV1_phase2`.
2. Calculate the root mean squared error. Is this error large or small? Explain your answer.

```
# Hint
y <- dat1$FEV1_phase2
y_predicted <- fitted(lm(...))

# root mean squared error
rmse <- sqrt(mean((y - y_predicted)^2))
```

TASK 8.1

[ ]:
```
y <- dat1$FEV1_phase2
```

[ ]:
```
y_predicted <- fitted(lm(FEV1_phase2 ~ FEV1 + FVC + pct_gastrapping, data=dat1))
```

TASK 8.2

[ ]:
```
rmse <- sqrt(mean((y - y_predicted)^2))
rmse
```

0.368401110563264

The root mean squared error is 0.3684, which can also say that the mean squared error is .1351, which can be considered low because we are dealing with values from 0 to 100, representing a small amount of error.

####Task 9 Group Task (15 points)

1. Using the statistical/machine learning concepts from class, build a model (regression or random forest) that best predicts the `FEV1_phase2` variable in the `dat2` dataframe.
2. Submit your predictions as a csv file in the format presented in the `copd_predictions.csv` file on Canvas.

- The group with the lowest prediction error will receive 10 points extra credit on their overall project grade.
- The group with the second lowest prediction error will receive 5 points extra credit on their overall project grade.
- The group with the third lowest prediction error will receive 2 points extra credit on their overall project grade.

TASK 9.1

[ ]:
```
install.packages("randomForest")
library(randomForest)
```

```
Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)

randomForest 4.7-1.1

Type rfNews() to see new features/changes/bug fixes.
```

[ ]:
```
samp <- sample(1:nrow(dat1), 3000)
train <- dat1[samp, ]
valid <- dat1[-samp, ]
```

[ ]:
```
fit <- randomForest(FEV1_phase2 ~ FEV1 + FVC + pct_gastrapping +␣
 ↪total_lung_capacity + height_cm + exp_meanatt + gender + pct_emphysema +␣
 ↪hay_fever + bmi, # MY BEST MODEL
                    data = train,
                    importance = TRUE,

                    # hyperparameters (change to improve predictions)
                    ntree    = 1000,  # number of trees to fit
                    mtry     = 10,    # number of variables to sample per tree
                    nodesize = 1,     # minimum size of terminal nodeslm
                    )
#the training is always better!
```

[ ]:
```
mse <- function(true, pred) {
  return(mean((true - pred)^2))
}
```

[ ]:
```
# calculate MSE on testing dataset
mse(valid$FEV1_phase2, predict(fit, newdata = valid))

# calculate RMSE on testing dataset
sqrt(mse(valid$FEV1_phase2, predict(fit, newdata = valid)))
```

0.081202616136111

0.284960727357492

[ ]:
```
mse(train$FEV1_phase2, predict(fit, newdata = train))
sqrt(mse(train$FEV1_phase2, predict(fit, newdata = train)))
mse(valid$FEV1_phase2, predict(fit, newdata = valid))
sqrt(mse(valid$FEV1_phase2, predict(fit, newdata = valid)))
```

0.0131233334427677

0.11455711869093

0.081202616136111

0.284960727357492

TASK 9.2

[ ]:
```
FEV1_phase2_predictions <-  predict(fit, newdata = dat2)
preds <- data.frame(sid = dat2$sid, FEV1_phase2_predictions)
write.csv(preds, 'copd_predictions.csv')
```

####Task 10 (20 points)

Write a 500-750 word abstract describing your analysis. Note this is a formal writeup and should be written with proper spelling and grammar. Imagine you are submitting this abstract to a conference for review. The writeup should include the following:

- Introduction to the topic of COPD and motivation for the analysis
- Description of the purpose of the study/report (what you were interested in finding)
- Briefly describe the data used to perform the analysis
- Description of the variables and methods used to conduct the analysis (data visualization, t-tests, regression, random forests, etc.)
- Description of the major results (estimates, significance, etc.)
- Major conclusions from the analysis in the context of the original application.

COPD stands for Chronic Obstructive Pulmonary Disease, which affects over 15 million people in the United States. The main concern of COPD is its cause; the cause seems uncertain. Even though there are a few factors correlated with COPD, such as smoking and genetics, the purpose of the study is to determine what is the leading cause of COPD; discovering the cause of COPD can help the research of developing a solution to control the disease. The study consists of 20 clinical sites and 10,000 subjects, hoping to expand the information on COPD. The data consists of 32 variables, excluding their patient number, visit year, and visit data. We compared multiple variables individually and collectively to see if there was a correlation between FEV1 Phase 2 and the air volume forcefully exhaled in 1 second after 5 years. The first comparison that we did was FEV1_phase2 between smoking status, which was divided into three categories: current smoker, former smoker, and never smoked. We didn't see a correlation between people who had smoked, but surprisingly, we noticed the people who never smoked tended to have a higher FEV1_phase 2. We did see a correlation between gender and FEV1_phase2; according to our two-sample t-tests, our p-value was less than 0.05, making a significant difference between levels of FEV1_phase between females and males. We also did a boxplot comparison, and the FEV1_phase2 was noticeably larger for males than females. A scatterplot between FEV1 and FEV1_phase2 showed a strong positive linear correlation. To prove this, we did a linear regression between FEV1_phase2, and the p-value was less than 0.05, making it significant, and the adjusted R squared was 79%, which is pretty high, proving the correlation. We also wanted to compare FVC and FEV1_phase2, which have a strong positive relationship. We also wanted to see if pct_gastrapping correlated with FEV1_phase, and we say a negative relationship, which means the more percentage of air trapped in lungs after exhaling [%], the more likely to have fewer levels of FEV1_phase2. With all this data that we collected, we wanted a linear regression between Y-variable FEV1_phase 2 and our explanatory variables, which were FEV1, FVC, and pct_gastrappings. The results were all variables were significantly correlated to FEV1_phase2. The residual standard error was around 0.3686, which can be considered high, but our multiple R-square was above 80%, which explains that our three explanatory variables can explain the variability of FEV1_phase2. This also can be proven by analyzing our 95% confidence interval for pct_gastrapping and FEV1; none of them included 0, making them significant. Also FEV1. After all this data, we generated a forest tree with a specific algorithm. Our response variable was FEV1_phase, and our explanatory variables were FEV1, FVC, pct gas trapping, total lung capacity, height, average lung density, gender, emphysema, fever, and BMI which was ur best model to make sure what combinations of variables showed a significant correlation to FEV1_phase with the knowledge what variables are correlated we can classify them as a cause. Our mean standard error on our testing dataset was 0.081, which is incredibly low. This number can predict FEV1_phase2, which is such a precession. This was our

prediction of FEV1_phase with our algorithm in our data with no missing values in FEV1_phase2. The primary conclusion that our team came up with was that the cause of COPD doesn't just rely on smoking status how taught; we tested explanatory variables that are considered genetic and those we saw the most correlation. Our research concludes that there are genetic factors that are correlated with high or low FEV1_phase2 levels.