COPD Analysis

Elena Batarseh

2025-01-21

```
# create url to census data
COPD_url <- url("https://raw.githubusercontent.com/khasenst/datasets_teaching/main/copd_data_project.csv

# load the census data into memory
copd <- read.csv(COPD_url)

# first few lines of dataset
head(copd)
```

```
##      sid visit_year visit_date visit_age gender  race height_cm weight_kg sysBP
## 1 10005Q       2008  1/15/2008      54.5 Female White     159.9      73.0   130
## 2 10006S       2008  1/15/2008      62.3 Female White     162.6      86.0   170
## 3 10010J       2008  1/15/2008      65.9 Female White     162.1      62.8    96
## 4 10015T       2008  2/15/2008      59.6   Male White     182.9     110.0   142
## 5 10017X       2008  6/15/2008      67.5   Male White     179.1      83.0   106
## 6 10022Q       2008  2/15/2008      69.8 Female White     158.8      78.0   122
##   diasBP hr O2_hours_day   bmi  asthma hay_fever bronchitis_attack pneumonia
## 1     80 87            0 28.55      No         0                No        No
## 2     80 81            8 32.53      No         0               Yes       Yes
## 3     63 66            0 23.90      No         0           unknown       Yes
## 4     88 75            0 32.88     Yes         1           unknown       Yes
## 5     72 72           10 25.88 unknown         0               Yes       Yes
## 6     78 87            0 30.93      No         3           unknown       Yes
##   chronic_bronchitis emphysema copd sleep_apnea SmokStartAge CigPerDaySmokAvg
## 1                 No        No   No          No           14               20
## 2                 No       Yes  Yes          No            8               20
## 3                Yes        No  Yes          No           25               15
## 4            unknown   unknown  Yes         Yes           16               20
## 5                 No       Yes  Yes          No           20               40
## 6            unknown       Yes  Yes          No           13               30
##   Duration_Smoking smoking_status total_lung_capacity pct_emphysema
## 1             40.5 Current smoker              5.6636      0.926851
## 2             52.0  Former smoker              5.2325     14.005900
## 3             40.9 Current smoker              5.1960      1.683760
## 4             28.0  Former smoker              6.3971      9.330450
## 5             35.0  Former smoker              7.8935     36.262400
## 6             30.0  Former smoker              5.1016     30.484400
##   functional_residual_capacity pct_gastrapping insp_meanatt exp_meanatt
## 1                       2.4766         6.80077     -830.343    -650.526
## 2                      -1.0000        -1.00000     -841.880      -1.000
## 3                       3.8993        41.34930     -833.429    -789.595
## 4                      -1.0000        -1.00000     -841.315      -1.000
## 5                       4.1043        46.17690     -887.947    -792.397
## 6                      -1.0000        -1.00000     -865.608      -1.000
##   FEV1_FVC_ratio  FEV1   FVC FEV1_phase2
## 1           0.77 2.921 3.805       2.622
## 2           0.43 1.288 3.022          NA
## 3           0.53 1.008 1.909       1.087
## 4           0.51 1.906 3.732       2.002
## 5           0.57 2.748 4.827       2.178
## 6           0.53 1.076 2.047       0.924
```

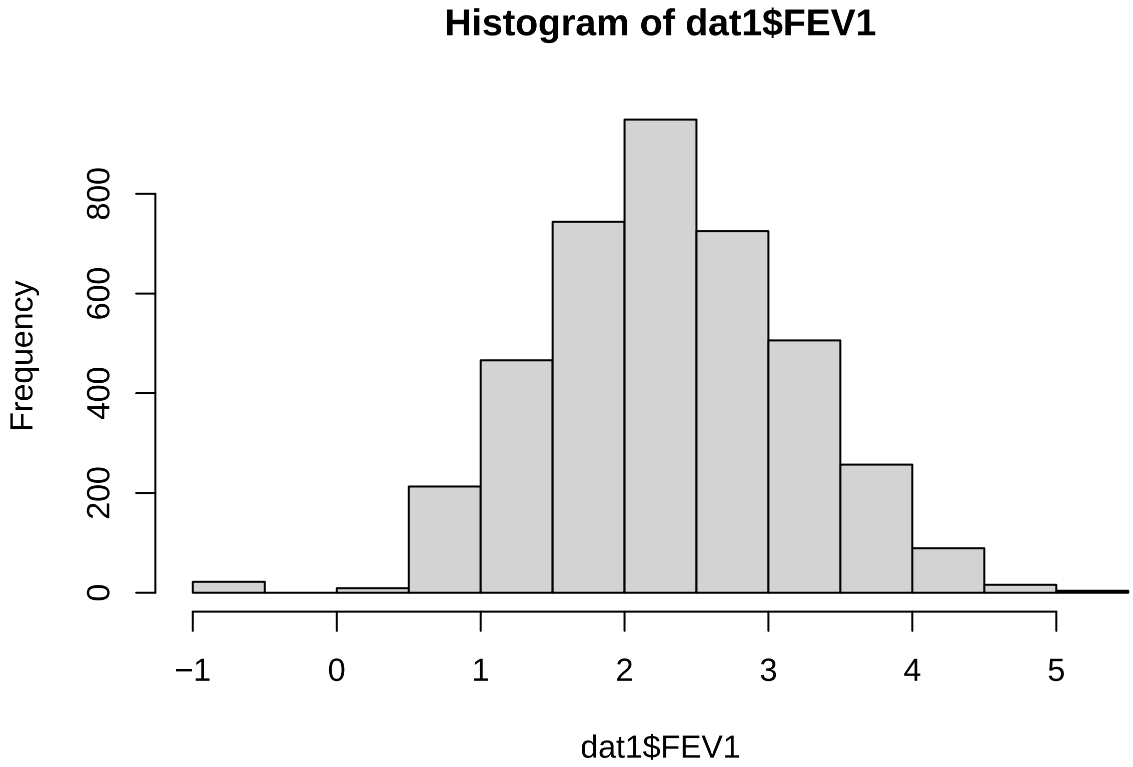
```
# remove NAs from dataset and store into dat1
dat1 <- na.omit(copd)

# number of rows in dat1
nrow(dat1)
```

```
## [1] 4000
```

```
# select rows where FEV1_phase2 is NA and store as dat2
dat2 <- dat1[is.na(copd$FEV1_phase2), ]

# number of rows in dat2
nrow(dat2)
```

```
## [1] 1747
```

```
# plot histogram of FEV1 from dat1
hist(dat1$FEV1)
```



```
# calculate mean and standard deviation
FEV1_mean <- mean(dat1$FEV1)
FEV1_SD <- sd(dat1$FEV1)
```

```r
# find the lower and upper bound within one standard deviation from mean
lowerbound1SD <- FEV1_mean - FEV1_SD
upperbound1SD <- FEV1_mean + FEV1_SD

# Subset FEV1 values that are within one standard deviation of the mean
FEV1within1SD <- dat1$FEV1[dat1$FEV1 >= lowerbound1SD & dat1$FEV1 <= upperbound1SD]

# calculate percentage
percentage1SD <- length(FEV1within1SD) / length(dat1$FEV1) * 100

percentage1SD
```

```
## [1] 67.575
```

The percentage of FEV1 values within one standard deviation of its mean is 67.575%.

```r
# find the lower and upper bound within two standard deviation from mean
lowerbound2SD <- FEV1_mean - (2*FEV1_SD)
upperbound2SD <- FEV1_mean + (2*FEV1_SD)

# subset FEV1 values that are within two standard deviation of the mean
FEV1within2SD <- dat1$FEV1[dat1$FEV1 >= lowerbound2SD & dat1$FEV1 <= upperbound2SD]

# calculate percentage
percentage2SD <- length(FEV1within2SD) / length(dat1$FEV1) * 100

percentage2SD
```

```
## [1] 96.75
```

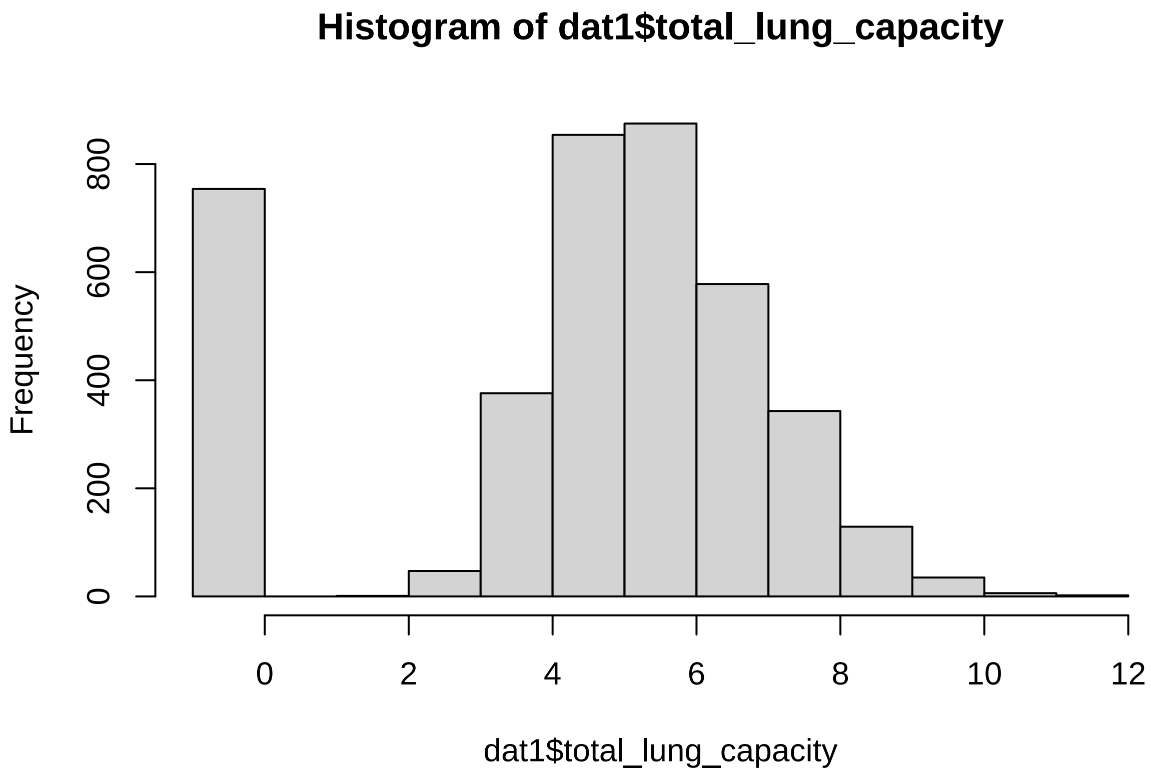
The percentage of FEV1 values within two standard deviation of its mean is 96.75%.

The empirical rule states that for a normal distribution, approximately 68% of the data falls within one standard deviation of the mean, 95% of the data falls within two standard deviations of the mean, and 99.7% of the data falls within three standard deviations of the mean.

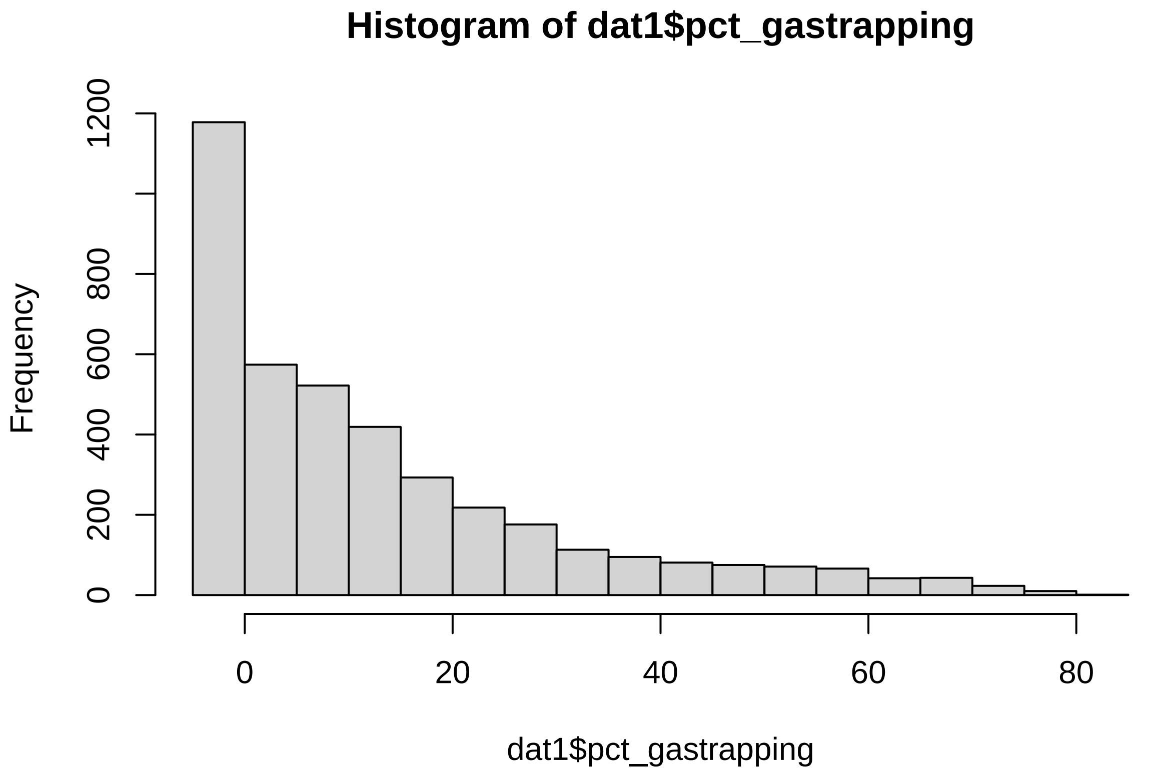
In our answer from 2.2, we found that 67.575% of the FEV1 values are within one standard deviation of the mean. In our answer from 2.3, we found that 96.75% of the FEV1 values are within two standard deviations of the mean.

These percentages are very close to 68% and 95%, respectively. Therefore, according to the empirical rule, we can conclude that the distribution of FEV1 values is approximately normal.

```r
# plot histogram of total_lung_capacity from dat1
hist(dat1$total_lung_capacity)
```



```r
# plot histogram of percentage of air trapped in lungs after exhaling in dat1
hist(dat1$pct_gastrapping)
```



```r
# find mean and standard deviation of total lung capacity
lung_capacity_mean <- mean(dat1$total_lung_capacity)
```

```
lung_capacity_SD <- sd(dat1$total_lung_capacity)

# print the range, mean, and standard deviation
cat("Range:", range(dat1$total_lung_capacity), "\n")
```

```
## Range: -1 11.7017
```

```
cat("Mean:",lung_capacity_mean, "\n")
```

```
## Mean: 4.299529
```

```
cat("Standard Deviation:", lung_capacity_SD, "\n")
```

```
## Standard Deviation: 2.853386
```

```
# determine if normally distributed - 1 standard deviation from mean

# find the lower and upper bound within one standard deviation from mean
lowerbound1SD_tlc <- lung_capacity_mean - lung_capacity_SD
upperbound1SD_tlc <- lung_capacity_mean + lung_capacity_SD

# subset total lung capacity values that are within one standard deviation of the mean
TLCwithin1SD <- dat1$total_lung_capacity[dat1$total_lung_capacity >= lowerbound1SD_tlc & dat1$total_lung

# calculate percentage
percentage1SD_tlc <- length(TLCwithin1SD) / length(dat1$FEV1) * 100

cat("Percentage 1 standard deviation from mean:" ,percentage1SD_tlc, "\n")
```

```
## Percentage 1 standard deviation from mean: 69.85
```

```
# determine if normally distributed - 2 standard deviation from mean

# find the lower and upper bound within one standard deviation from mean
lowerbound2SD_tlc <- lung_capacity_mean - (lung_capacity_SD*2)
upperbound2SD_tlc <- lung_capacity_mean + (lung_capacity_SD*2)

# subset total lung capacity values that are within one standard deviation of the mean
TLCwithin2SD <- dat1$total_lung_capacity[dat1$total_lung_capacity >= lowerbound2SD_tlc & dat1$total_lung

# calculate percentage
percentage2SD_tlc <- length(TLCwithin2SD) / length(dat1$FEV1) * 100

cat("Percentage 2 standard deviation from mean:" ,percentage2SD_tlc)
```

```
## Percentage 2 standard deviation from mean: 99.8
```

The range of total lung capacity is from -1 to 11.7017 liters, indicating that all data points fall within this interval. However, since negative lung capacity is not possible, any values of -1 are considered outliers in the dataset. The mean of total lung capacity is approximately 4.3 liters, indicating that the average of all values is 4.3. The standard deviation of total lung capacity is around 2.85 liters, suggesting high variability in the dataset, which is likely due to the presence of outliers. Approximately 69.85% of total lung capacity values fall within one standard deviation of the mean, and 99.8% of values fall within two standard deviations of the mean. This suggests that the data is approximately normal, as it roughly follows the empirical rule. However, the distribution is slightly skewed to the right, with the peak of the bell curve more to the left.

```
# find mean and standard deviation of percentage of air trapped
pct_gastrapping_mean <- mean(dat1$pct_gastrapping)
```

```
pct_gastrapping_SD <- sd(dat1$pct_gastrapping)

# print the range, mean, and standard deviation
cat("Range:", range(dat1$pct_gastrapping), "\n")
```

```
## Range: -1 81.269
```

```
cat("Mean:",pct_gastrapping_mean, "\n")
```

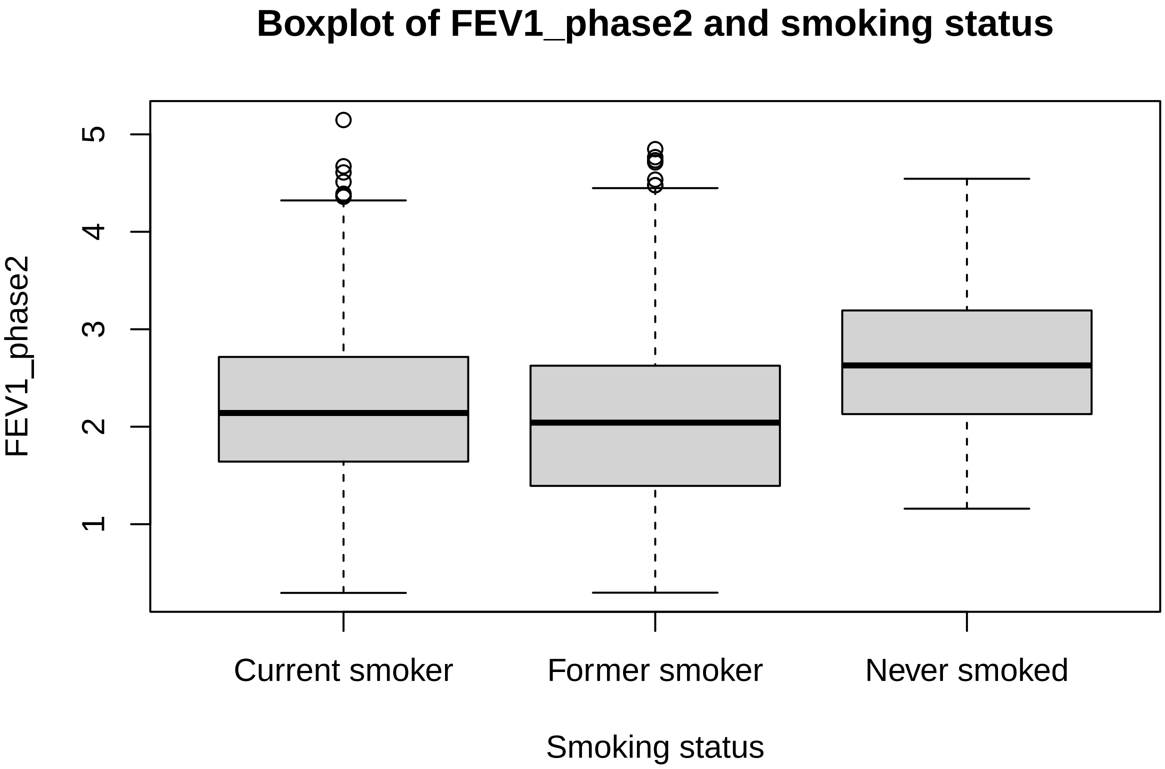
```
## Mean: 13.46015
```

```
cat("Standard Deviation:", pct_gastrapping_SD, "\n")
```

```
## Standard Deviation: 17.5218
```

The range of percentage of air trapped is from -1 to 81.269%, indicating that all data points fall within this interval. However, since negative percentage values are not possible, any values of -1 are considered outliers in the dataset. The mean of percentage of air trapped is approximately 13.46015%, indicating that the average of all values is 13.46015%. The standard deviation of percentage of air trapped is around 17.5218%, suggesting very high variability in the dataset, which may be due to the presence of outliers. The distribution shows a clear right skew, as most data points are concentrated to the left of the graph.

```
# create a boxplot between FEV1_phase2 and smoking status
boxplot(FEV1_phase2 ~ smoking_status, data = dat1,
        xlab = "Smoking status", ylab = "FEV1_phase2",
        main = "Boxplot of FEV1_phase2 and smoking status")
```



Based on the boxplots, the group with the best breathing capacity is the group that never smoked. The median of FEV1 phase two for the group that never smoked is higher than the groups Current Smoker and Former Smoker.

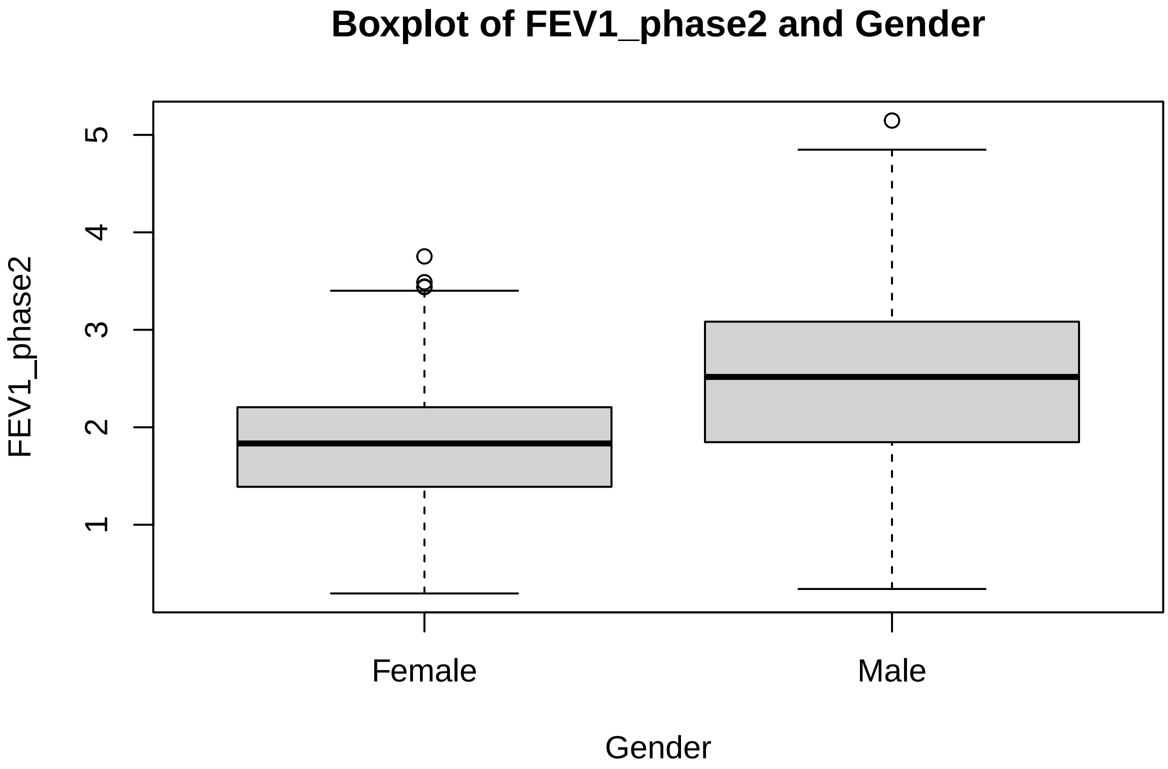
```
# 95% confidence interval
```

```r
subsetdat1 <- dat1[dat1$smoking_status %in% c("Current smoker", "Former smoker"), ]
t.test(FEV1_phase2 ~ smoking_status, data = subsetdat1, conf.level = 0.95)
```

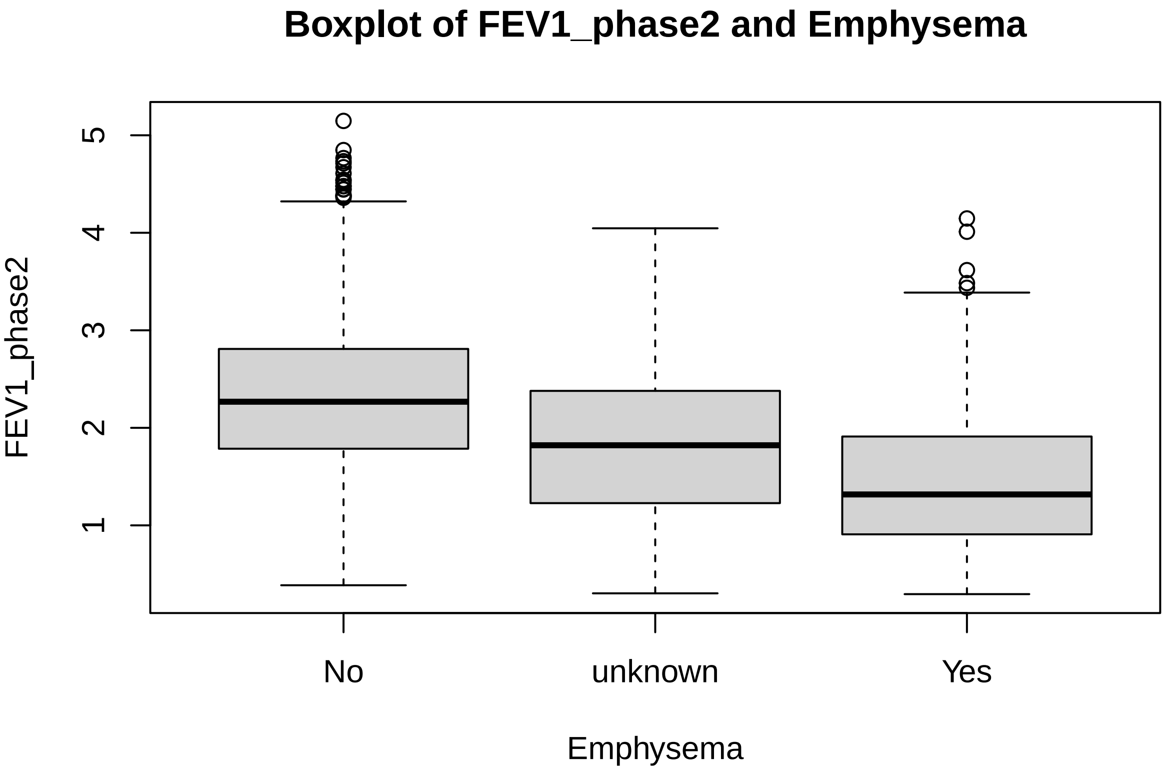
```
##
##  Welch Two Sample t-test
##
## data:  FEV1_phase2 by smoking_status
## t = 5.0794, df = 3933.1, p-value = 3.962e-07
## alternative hypothesis: true difference in means between group Current smoker and group Former smoke
## 95 percent confidence interval:
##  0.08155305 0.18408473
## sample estimates:
## mean in group Current smoker  mean in group Former smoker
##                     2.179463                     2.046644
```

In a two-sample t-test, if the p-value is less than 0.05, then that result is said to be statistically significant. If a p-value is greater than 0.05, then the result is considered not statistically significant at the 0.05 level. In this case, the p-value is 3.962eˆ-7, which is much less than 0.05. This means that the differences between the means are statistically significant, and that there is a significant difference in FEV1 Phase 2 in the Current Smoker and Former Smoker categories.

```r
# create a boxplot of FEV1_phase2 and gender
boxplot(FEV1_phase2 ~ gender, data = dat1,
        xlab = "Gender", ylab = "FEV1_phase2",
        main = "Boxplot of FEV1_phase2 and Gender")
```



```r
# create boxplot of FEV1_phase2 and emphysema
boxplot(FEV1_phase2 ~ emphysema, data = dat1,
        xlab = "Emphysema", ylab = "FEV1_phase2",
        main = "Boxplot of FEV1_phase2 and Emphysema")
```



```
# 95% confidence interval for FEV1_phase 2 and gender
subset2dat1 <- dat1[dat1$gender %in% c("Female", "Male"), ]
t.test(FEV1_phase2 ~ gender, data = subset2dat1, conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  FEV1_phase2 by 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
## 95 percent confidence interval:
##  -0.6918855 -0.5976740
## sample estimates:
## mean in group Female   mean in group Male
##             1.796581             2.441361
```

A p-value less than the significance level of 0.05 suggests that the null hypothesis can be rejected and that there is evidence to support the alternative hypothesis. In this case, the p-value of 2.2x10ˆ-16 is less than 0.05, indicating strong evidence to reject the null hypothesis and accept that there is a significant difference between the two groups.

The interval [-0.692, -0.598] is a confidence interval for the true mean difference between the two groups. A 95% confidence level means that if we were to repeat the study many times, 95% of the time, the true mean difference would fall within this interval. Since this interval does not contain zero, we can say with 95% confidence that the mean difference is significantly different from zero, which supports the conclusion that the FEV1 Phase 2 for the male group is significantly greater than the FEV1 Phase 2 for the female group.

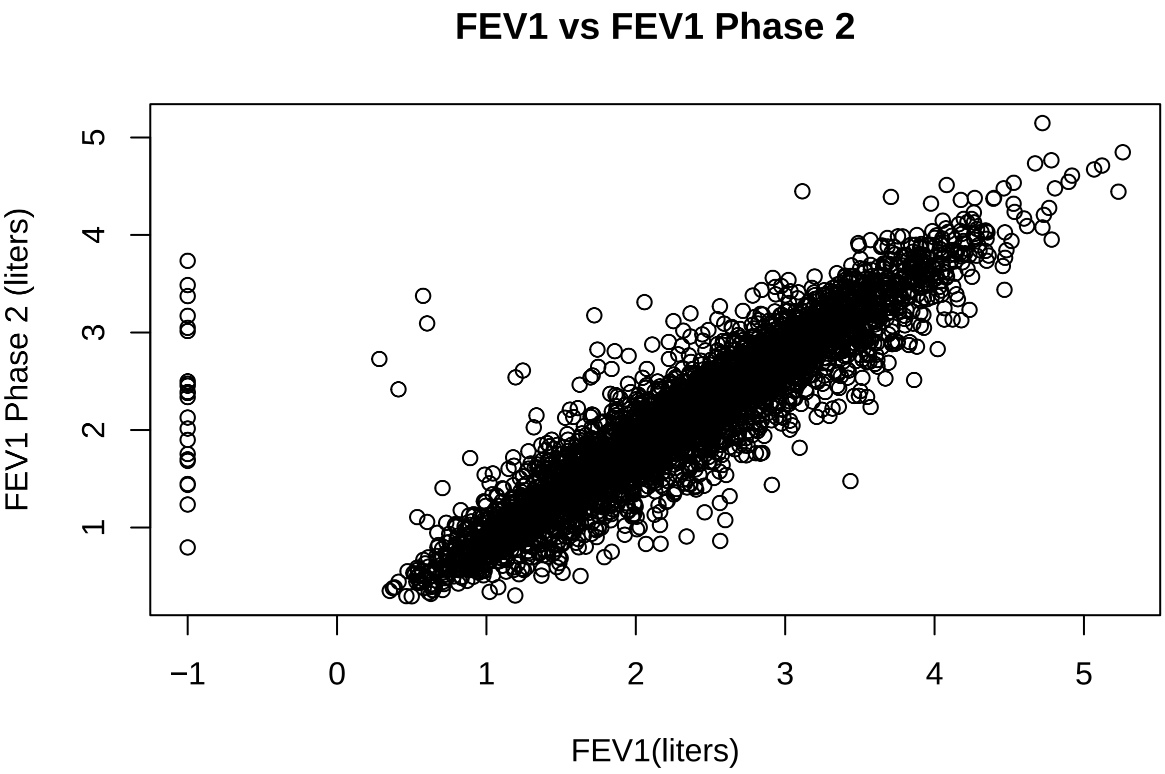
```
# 95% confidence interval for FEV1_phase 2 and emphysema
subset3dat1 <- dat1[dat1$emphysema %in% c("No", "Yes"), ]
t.test(FEV1_phase2 ~ emphysema, data = subset3dat1, conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  FEV1_phase2 by emphysema
## t = 28.474, df = 1202.9, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group No and group Yes is not equal to 0
## 95 percent confidence interval:
##  0.7907560 0.9077912
## sample estimates:
##  mean in group No mean in group Yes
##          2.311681          1.462408
```

A p-value less than the significance level of 0.05 suggests that the null hypothesis can be rejected and that there is evidence to support the alternative hypothesis. In this case, the p-value of 2.2x10ˆ-16 is less than 0.05, indicating strong evidence to reject the null hypothesis and accept that there is a significant difference between the two groups.

The interval [0.791, 0.908] is a confidence interval for the true mean difference between the two groups. A 95% confidence level means that if we were to repeat the study many times, 95% of the time, the true mean difference would fall within this interval. Since this interval does not contain zero, we can say with 95% confidence that the mean difference is significantly different from zero, which supports the conclusion that the FEV1 Phase 2 for the group without emphysema is significantly greater than the FEV1 Phase 2 for the group with emphysema.

```
# scatter plot between FEV1 and FEV1 Phase 2
plot(dat1$FEV1, dat1$FEV1_phase2,
     xlab = "FEV1(liters)",
     ylab = "FEV1 Phase 2 (liters)",
     main = "FEV1 vs FEV1 Phase 2")
```



Based on the scatterplot, there is a strong positive linear correlation between FEV1 and FEV1 Phase 2.

```r
# linear regression model FEV1_Phase 2 on FEV1
fit1 <- lm(FEV1_phase2 ~ FEV1, data = dat1)
summary(fit1)
```

```
##
## 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
```

Multiple R-squared(coefficient of determination) is the percentage of variance in the outcome that is explained by the predictor variables. In the context of the problem, the coeffecient of determination is 0.7897. This means that ~79% of the variation in FEV1 Phase 2 is explained by variation in FEV1.

The slope of the regression is 0.84. The slope indicates the rate of change in the dependent variable "y" for a one-unit increase in the independent variable "x" For every 1 liter icrease in FEV1, FEV1 Phase two will increase by 0.84 liters on average.

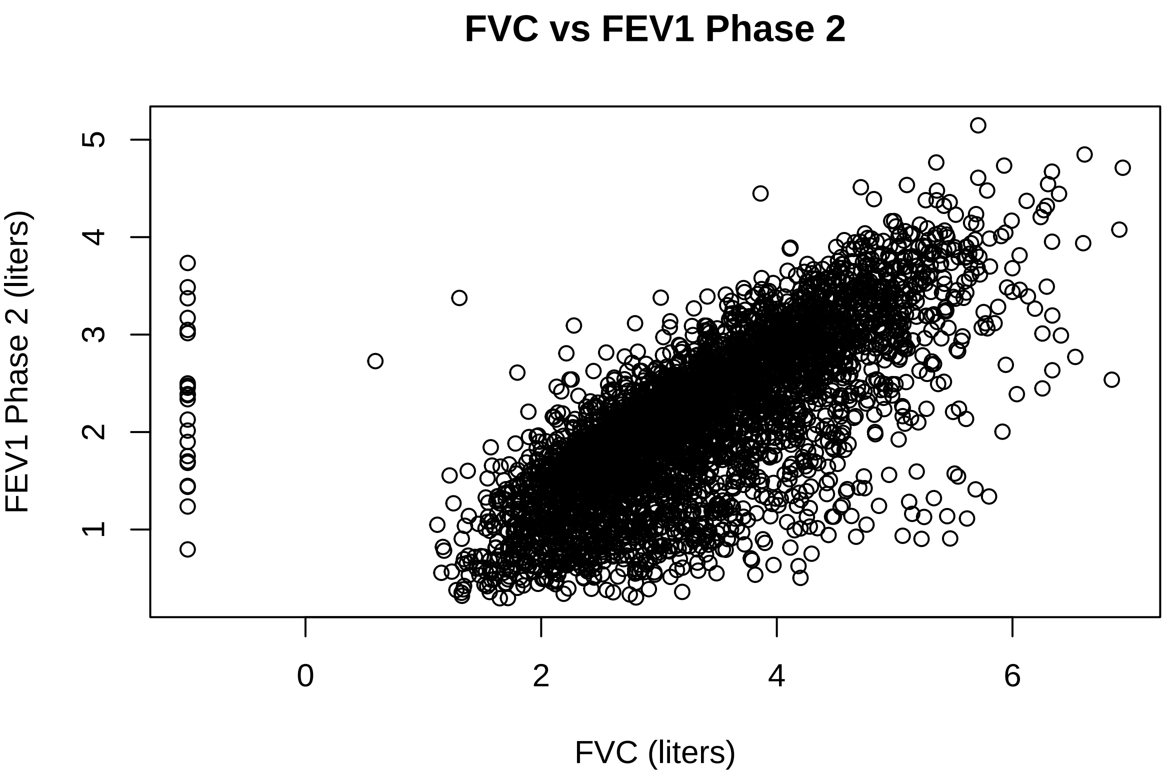
The intercept is 0.178336. The y intercept is the value of the dependent variable(y) when the independent variable (x) is equal to 0. So when FEV1 is 0 liters, FEV1 Phase 2 is 0.178336 liters.

```r
# 95% conf int to determine if the slope is significantly less than 1
confint(fit1, conf.level = 0.95)
```

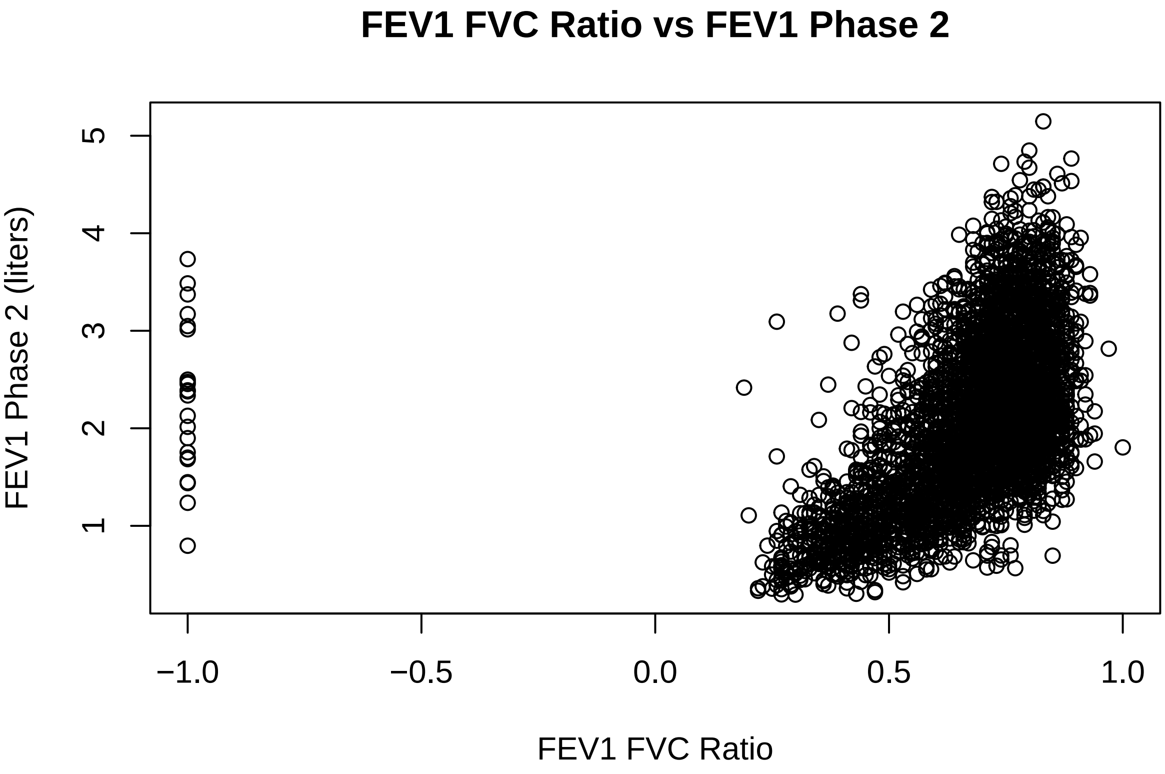
```
##                 2.5 %    97.5 %
## (Intercept) 0.1451177 0.2115549
## FEV1        0.8269759 0.8538708
```

We are 95% confident that the slope in the population is between 0.827 and 0.854. For every one liter increase in FEV1, the predicted value of FEV1 Phase 2 increases between 0.827 and 0.854 liters. Since this confidence interval does not contain the value 1, we can conclude that there is a statistically significant association between FEV1 and FEV1 Phase 2. In terms of breathing health, this shows that there is a strong correlation between FEV1 and FEV1 Phase 2. If you have low FEV1, you are more likely to have low FEV1 Phase 2.

```r
# scatter plot between FVC and FEV1_phase2
plot(dat1$FVC, dat1$FEV1_phase2,
     xlab = "FVC (liters)",
     ylab = "FEV1 Phase 2 (liters)",
     main = "FVC vs FEV1 Phase 2")
```



```
# scatter plot between FEV1_FCV_ratio and FEV1_phase2
plot(dat1$FEV1_FVC_ratio, dat1$FEV1_phase2,
     xlab = "FEV1 FVC Ratio",
     ylab = "FEV1 Phase 2 (liters)",
     main = "FEV1 FVC Ratio vs FEV1 Phase 2")
```



```r
# fit linear regression model
fit2 <- lm(FEV1_phase2 ~ FEV1 + FVC + FEV1_FVC_ratio, data = dat1)
summary(fit2)
```

```
##
## Call:
## lm(formula = FEV1_phase2 ~ FEV1 + FVC + FEV1_FVC_ratio, data = dat1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.69650 -0.17107 -0.00349  0.16521  2.56454
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.31653    0.02751   47.86   <2e-16 ***
## FEV1            1.54224    0.01620   95.17   <2e-16 ***
## FVC            -0.46215    0.01137  -40.64   <2e-16 ***
## FEV1_FVC_ratio -1.79506    0.04080  -44.00   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3029 on 3996 degrees of freedom
## Multiple R-squared:  0.8657, Adjusted R-squared:  0.8656
## F-statistic:  8584 on 3 and 3996 DF,  p-value: < 2.2e-16
```

```r
# 95% conf int to determine if the slope is significantly different from 0
confint(fit2, conf.level = 0.95)
```

```
##                     2.5 %     97.5 %
## (Intercept)     1.2625983  1.3704668
## FEV1            1.5104702  1.5740093
## FVC            -0.4844493 -0.4398581
## FEV1_FVC_ratio -1.8750469 -1.7150695
```

We are 95% confident that the slope between FEV1 Phase 2 and FEV1 is between 1.510 and 1.574. For every one liter increase in FEV1, the predicted value of FEV1 Phase 2 increases between 1.510 and 1.574 liters. Since this confidence interval doesn't contain the value 0, we can conclude that there is a statistically significant association between FEV1 and FEV1 Phase 2. This also means there is a less than 5% chance that the difference is zero.

We are 95% confident that the slope between FEV1 Phase 2 and FVC is between -0.484 and -0.440. For every one liter increase in FVC, the predicted value of FEV1 Phase 2 decreases between 0.440 and 0.484 liters. Since this confidence interval doesn't contain the value 0, we can conclude that there is a statistically significant association between FVC and FEV1 Phase 2. This also means there is a less than 5% chance that the difference is zero.

We are 95% confident that the slope between FEV1 Phase 2 and FEV1 FVC Ratio is between -1.875 and -1.715. For every one unit increase in FEV1 FVC Ratio, the predicted value of FEV1 Phase 2 decreases between 1.715 and 1.875 liters. Since this confidence interval doesn't contain the value 0, we can conclude that there is a statistically significant association between FEV1 FVC Ratio and FEV1 Phase 2. This also means there is a less than 5% chance that the difference is zero.

```r
# predict
y <- dat1$FEV1_phase2
y_predicted <- fitted(lm(fit2))
```

```r
# find rmse
rmse <- sqrt(mean((y - y_predicted)^2))
rmse
```

```
## [1] 0.3027762
```

In the context of the study, a root mean squared error (RMSE) of 0.302776 means that on average, the predicted values of FEV1 Phase 2 from the regression model are off by 0.302776 liters compared to the actual values. Relative to the study, this may be a large root mean squared error and this may not be the best way to predict FEV1 Phase 2 from other variables.

Building a model that best predicts the FEV1__phase2 variable in the dat2 dataframe.

```r
# randomly sample 300 rows from the dataset
samp <- sample(1:nrow(dat1), 3000)
train <- dat1[samp, ]
nrow(train)
```

```
## [1] 3000
```

```r
# use the remaining rows as validation
valid <- dat1[-samp, ]
nrow(valid)
```

```
## [1] 1000
```

```r
# load randomForest library
#install.packages("randomForest")
#library(randomForest)
```

```r
#mean squared error function
mse <- function(true, pred) {
  return(mean((true - pred)^2))
}
```

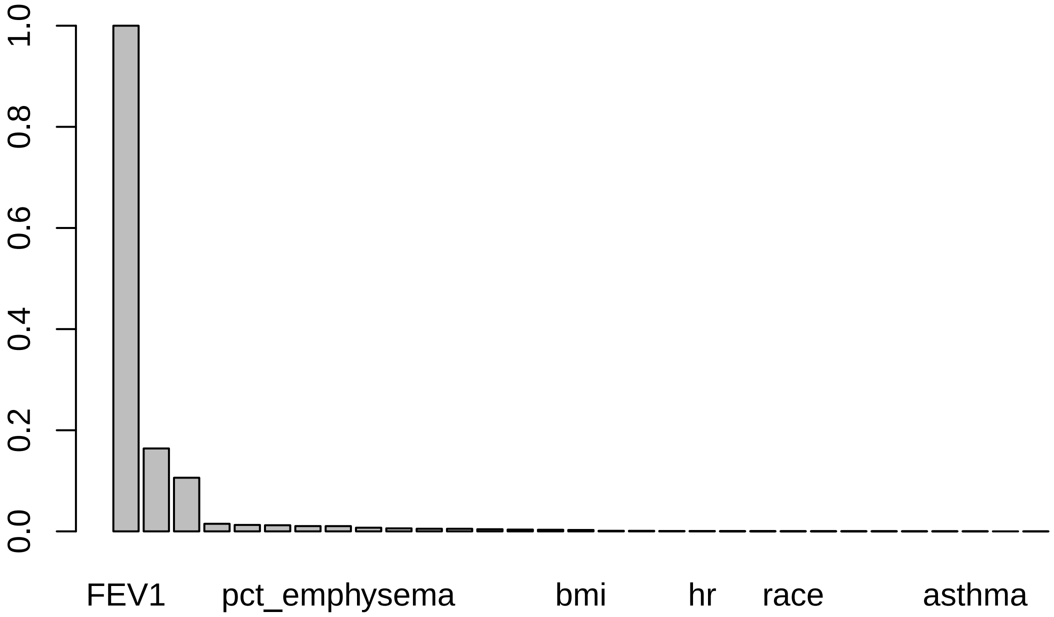
```r
# random forest
fit3 <- randomForest(FEV1_phase2 ~ .,
                     data = train [, -(1:3)],
                     importance = TRUE,

                     # hyperparameters (change to improve predictions)
                     ntree    = 850,  # number of trees to fit
                     mtry     = 13,   # number of variables to sample per tree
                     nodesize = 10,    # minimum size of terminal nodes
                     maxnodes = NULL, # maximum number of terminal nodes a tree can have
                     )
# calculate MSE on testing dataset
mse(valid$FEV1_phase2, predict(fit3, newdata = valid))
```

```
## [1] 0.08399584
```

```r
# find predictions using random forest
FEV1_phase2_predictions <- predict(fit3, dat2)
preds <- data.frame(sid = dat2$sid, FEV1_phase2_predictions)
write.csv(preds, 'copd_predictions.csv')
```

```r
options(repr.plot.width = 60, repr.plot.height = 8)
```

```
importance <- fit3$importance[, 1]
importance <- sort(importance, decreasing = TRUE) / max(importance)
barplot(importance)
```


FEV1
pct_emphysema
bmi
hr
race
asthma