Executive Summary

Angel Waters

05/27/2022

## Introduction

Lung capacity data was collected on 725 patients ages 3 to 19. In addition, their gender, birth status, and smoke status were collected.

Data was cleaned by reordering the data table so the measured variables for each patient came after their descriptive variables. Gender column was capitalized and the logical columns (columns with values as “yes” or “no”) were converted to “TRUE” and “FALSE” values.

## Key Findings

Age buckets were calculated by taking the difference of the minimum and maximum age and dividing it into 4 for four equal age groups. Frequency table shows that the age bucket with the most amount of data for both genders is in Bucket 3 (12 to 15 years) and the least amount of data is in Bucket 1 (3 to 7 years) (see **Table 1**). Both genders show the same trends for age bucket frequencies. A histogram of age by gender shows the distributions are slightly skewed left (see **Fig. 1**).

Summary statistics for the numerical data are shown in **Table 2**. Females tend to be shorter than males in this data set (see **Fig.** 2).

Age and height have a positive relationship, but the genders do show some separation; male patients are higher for height in the same age groups (see **Fig. 3**). There is a positive relationship between lung capacity and height (see **Fig 4**). An overlap between the genders at any given height for this data will show a similar lung capacity. For Lung Capacity vs Age, there is separation between the genders, males at any given age would have a higher lung capacity (see **Fig. 5**).

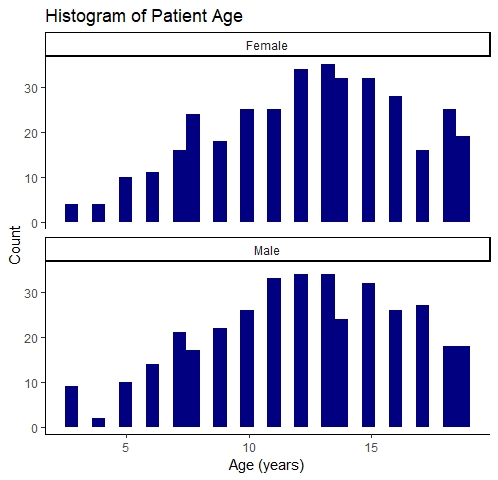
Smoking was looked at to understand how smokers’ lung capacity compares to non-smokers. The data shows an overlap between non-smokers and smokers; there is no abnormal trend in lung capacity shown for the smokers across age (see **Fig. 6**). The spread across the genders and ages were analyzed to understand the demographics of the smokers. Female patients were more frequent than male patients.

**Table 1.** Frequency table of Age by Gender and Age Bucket.

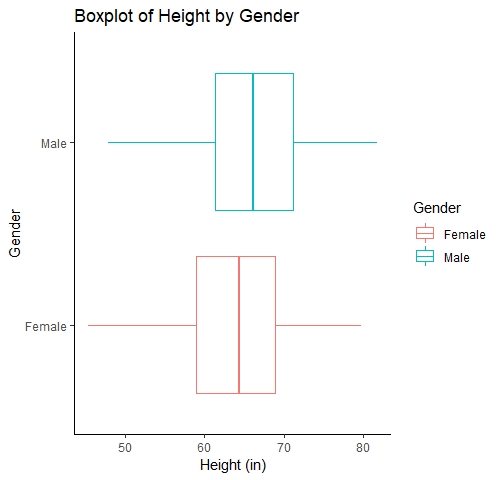
|  |  |  |  |
| --- | --- | --- | --- |
| **Gender** | **Age Bucket** | **Total** | **Frequency** |
| Female | B3 | 133 | 0.372 |
| Female | B2 | 92 | 0.257 |
| Female | B4 | 88 | 0.246 |
| Female | B1 | 45 | 0.126 |
| Male | B3 | 124 | 0.338 |
| Male | B2 | 98 | 0.267 |
| Male | B4 | 89 | 0.243 |
| Male | B1 | 56 | 0.153 |

**Table 2.** Summary statistics of Age, Height and Lung Capacity of patients.

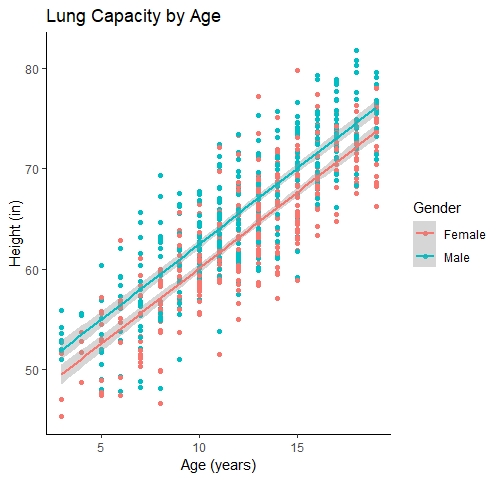
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Min** | **Max** | **Median** | **Mean** | **Standard Deviation** |
| Age (Years) | 3.00 | 19.00 | 13.00 | 12.33 | 4.00 |
| Height (in) | 45.30 | 81.80 | 65.40 | 64.84 | 7.20 |
| Lung Capacity | 0.507 | 14.675 | 8.000 | 7.863 | 2.662 |



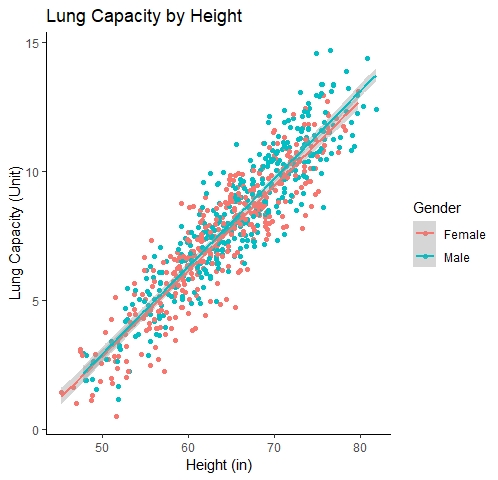
**Figure 1.** Histogram of Patient Age by gender.



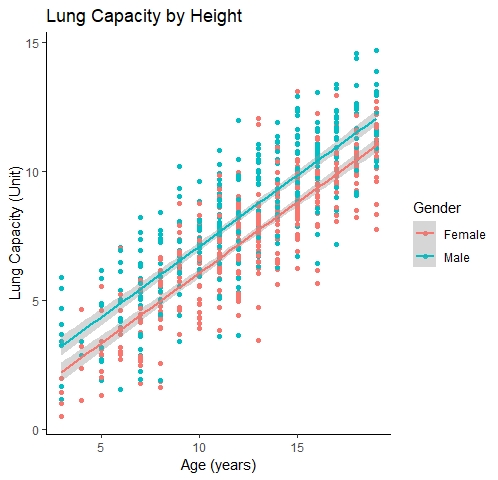
**Figure 2.** Boxplot of Height by Gender.



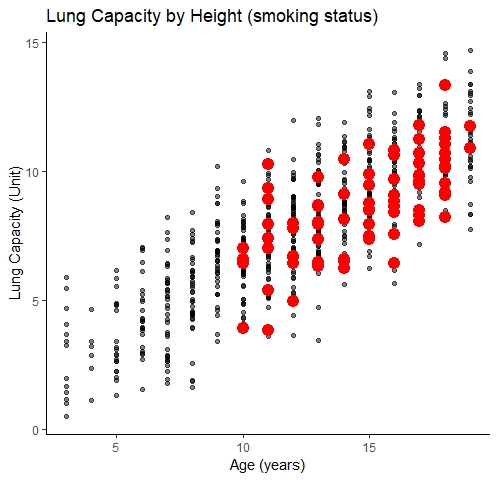
**Figure 3.** Scatterplot of Age by Height, grouped by gender.



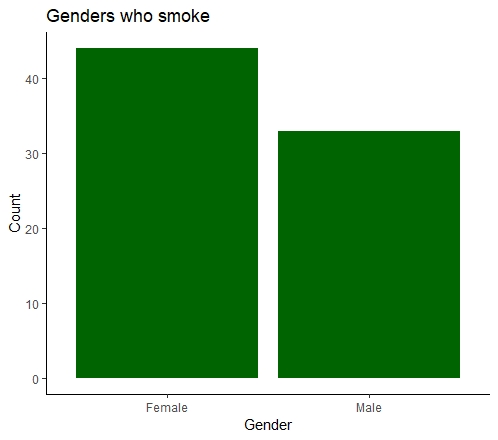
**Figure 4.** Scatterplot of Lung Capacity and Height grouped by gender.



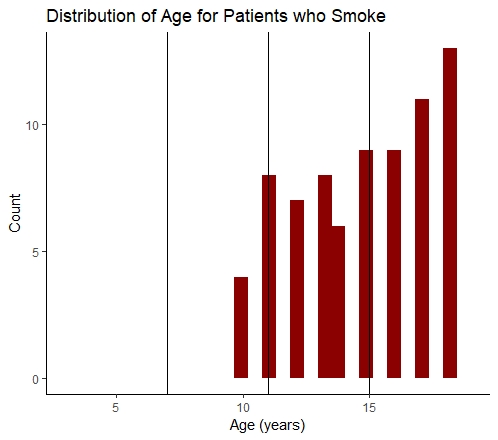
**Figure 5.** Scatterplot of Lung Capacity by Age, grouped by gender.



**Figure 6.** Scatterplot of Lung Capacity by Age, highlighted in red are the smokers of the data collection.



**Figure 8.** Smoking by gender.



**Figure 9.** Distribution of Age for patients who smoke. Age buckets are denoted by the vertical lines.

## Conclusion

Lung capacity for the patients collected seems to be related to height and age and increases for both. Between the genders, the data shows that at any given height, both genders will have comparable lung capacity, but not at any given age. This is possibly because there was a separation in height across the ages for males over females. Height may affect lung capacity more than age. Smokers for this data set have similar lung capacities to non-smokers as seen in the analysis above. They are mostly female and can be found in all age buckets except bucket 1.

## Appendix

### Data Cleanup

Loading the libraries used in this script.

library(ggplot2)  
library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v tibble 3.1.6 v dplyr 1.0.8  
## v tidyr 1.2.0 v stringr 1.4.0  
## v readr 2.1.2 v forcats 0.5.1  
## v purrr 0.3.4

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

Loading the csv needed and viewing properties of the data set.

lung\_cap <- read.csv("LungCapDataCSV.csv", header = TRUE)  
str(lung\_cap)

## 'data.frame': 725 obs. of 6 variables:  
## $ LungCap : num 6.47 10.12 9.55 11.12 4.8 ...  
## $ Age : int 6 18 16 14 5 11 8 11 15 11 ...  
## $ Height : num 62.1 74.7 69.7 71 56.9 58.7 63.3 70.4 70.5 59.2 ...  
## $ Smoke : chr "no" "yes" "no" "no" ...  
## $ Gender : chr "male" "female" "female" "male" ...  
## $ Caesarean: chr "no" "no" "yes" "no" ...

view(lung\_cap)

Manipulating columns and creating new columns for analysis. Also rearranging the columns that makes logical sense.

lung\_cap <- mutate(lung\_cap, Smoke = as.logical(ifelse(Smoke=="no", FALSE, TRUE)),  
 Caesarean = as.logical(ifelse(Caesarean=="no", FALSE, TRUE)),  
 Gender = ifelse(Gender=="male", "Male", "Female"))  
  
lung\_cap <- lung\_cap %>% transmute(Gender, Age, Height, LungCap, Smoke, Caesarean)  
lung\_cap <- lung\_cap %>%   
 mutate(AgeBuckets = ifelse(Age<=7, "B1",   
 ifelse(between(Age, 8, 11), "B2",   
 ifelse(between(Age, 12, 15), "B3", "B4"))))

### **Summary Statistics**

Pulling the descriptive statistics for this data set.

summary(lung\_cap)

## Gender Age Height LungCap   
## Length:725 Min. : 3.00 Min. :45.30 Min. : 0.507   
## Class :character 1st Qu.: 9.00 1st Qu.:59.90 1st Qu.: 6.150   
## Mode :character Median :13.00 Median :65.40 Median : 8.000   
## Mean :12.33 Mean :64.84 Mean : 7.863   
## 3rd Qu.:15.00 3rd Qu.:70.30 3rd Qu.: 9.800   
## Max. :19.00 Max. :81.80 Max. :14.675   
## Smoke Caesarean AgeBuckets   
## Mode :logical Mode :logical Length:725   
## FALSE:648 FALSE:561 Class :character   
## TRUE :77 TRUE :164 Mode :character   
##   
##   
##

sd(lung\_cap$Age)

## [1] 4.00475

sd(lung\_cap$Height)

## [1] 7.202144

sd(lung\_cap$LungCap)

## [1] 2.662008

Creating frequency tables bucketed by gender and age buckets.

gen\_Age <- lung\_cap %>% group\_by(Gender, AgeBuckets) %>% summarise(Total = n())

## `summarise()` has grouped output by 'Gender'. You can override using the  
## `.groups` argument.

gen\_Age <- gen\_Age %>% mutate(Frequency = Total / sum(Total))  
female\_freq <- gen\_Age %>% filter(Gender=="Female") %>% arrange(desc(Frequency),by\_group= Gender)  
male\_freq <- gen\_Age %>% filter(Gender=="Male") %>% arrange(desc(Frequency),by\_group= Gender)

### **Data Visualization**

Histograms for the age by gender.

age\_hist <- ggplot(lung\_cap)+  
 geom\_histogram(mapping=aes(Age), fill="navy")+  
 facet\_wrap(vars(Gender), nrow=2)+  
 theme\_classic()+  
 labs(x= "Age (years)", y="Count", title="Histogram of Patient Age")  
age\_hist

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

Chart, bar chart

Description automatically generated

Relationships between the continuous variables.

LCvH\_Scatter <- ggplot(lung\_cap, mapping=aes(x=Height, y=LungCap, color = Gender))+  
 geom\_point()+  
 geom\_smooth(method = lm, formula=y~x)+  
 theme\_classic()+  
 labs(x="Height (in)", y="Lung Capacity (Unit)", title="Lung Capacity by Height")  
LCvH\_Scatter

Chart, scatter chart

Description automatically generated

HvA <- ggplot(lung\_cap, mapping=aes(x=Age, y=Height, color = Gender))+  
 geom\_point()+  
 geom\_smooth(method = lm, formula=y~x)+  
 theme\_classic()+  
 labs(x="Age (years)", y="Height (in)", title="Lung Capacity by Age")  
HvA

Chart, scatter chart

Description automatically generated

LCvA\_Scatter <- ggplot(lung\_cap, mapping=aes(x=Age, y=LungCap, color = Gender))+  
 geom\_point()+  
 geom\_smooth(method = lm, formula=y~x)+  
 theme\_classic()+  
 labs(x="Age (years)", y="Lung Capacity (Unit)", title="Lung Capacity by Height")  
LCvA\_Scatter

Chart, scatter chart

Description automatically generated

Boxplot of height by gender.

height\_boxplot <- ggplot(lung\_cap)+  
 geom\_boxplot(mapping=aes(Height, Gender, color=Gender))+  
 theme\_classic()+  
 labs(x="Height (in)", y="Gender", title = "Boxplot of Height by Gender")  
height\_boxplot

Chart, box and whisker chart

Description automatically generated

Smokers analysis: Separate the smokers and non-smokers data.

smokers <- filter(lung\_cap, Smoke==TRUE)  
nonsmokers <- filter(lung\_cap, Smoke==FALSE)

Non-smokers scatterplot overlayed by smoker data.

LCvA\_smokers <- ggplot()+  
 geom\_point(nonsmokers, mapping=aes(Age, LungCap), alpha=0.5)+  
 geom\_point(smokers, mapping=aes(Age, LungCap), color="red", size=4)+  
 theme\_classic()+  
 labs(x="Age (years)", y="Lung Capacity (Unit)", title="Lung Capacity by Height (smoking status)")  
LCvA\_smokers

Chart, scatter chart

Description automatically generated

Distributions of Age and Gender for smokers in the dataset.

smokers\_MF <- ggplot(smokers)+  
 geom\_histogram(mapping=aes(Gender), stat="count", fill = "Dark Green")+  
 theme\_classic()+  
 labs(y="Count", title = "Genders who smoke")

## Warning: Ignoring unknown parameters: binwidth, bins, pad

smokers\_MF

Chart, bar chart

Description automatically generated

smokers\_Age <- ggplot(smokers)+  
 geom\_histogram(mapping=aes(Age), fill="dark red")+  
 xlim(3, 19)+  
 geom\_vline(xintercept = 7)+  
 geom\_vline(xintercept= 11)+  
 geom\_vline(xintercept=15)+  
 theme\_classic()+  
 labs(x= "Age (years)", y="Count", title="Distribution of Age for Patients who Smoke")  
smokers\_Age

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 2 rows containing missing values (geom\_bar).