rpractice markdown

Angel Waters

2022-05-28

##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.

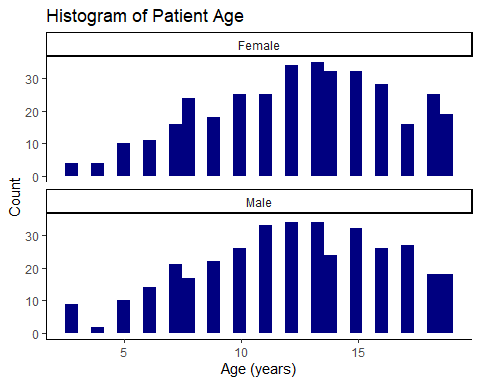
gen\_Age <- gen\_Age %>% mutate(Frequency = Total / sum(Total))  
female\_freq <- gen\_Age %>% filter(Gender=="Female") %>% arrange(desc(Frequency),by\_group= Gender)  
male\_freq <- gen\_Age %>% filter(Gender=="Male") %>% arrange(desc(Frequency),by\_group= Gender)

##Data Visualization

Histograms for the age by gender.

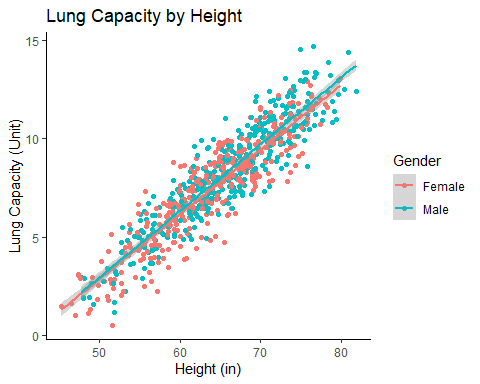
age\_hist <- ggplot(lung\_cap)+  
 geom\_histogram(mapping=aes(Age), fill="navy")+  
 facet\_wrap(vars(Gender), nrow=2)+  
 theme\_classic()+  
 labs(x= "Age (years)", y="Count", title="Histogram of Patient Age")  
age\_hist

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

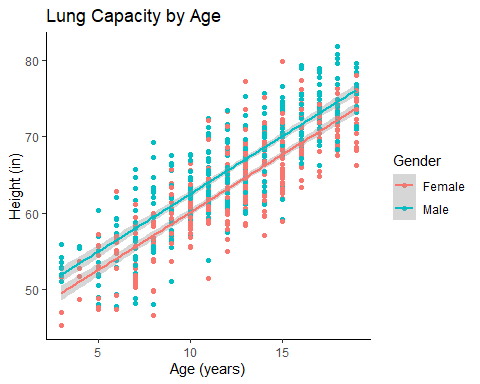


Relationships between the continuous variables.

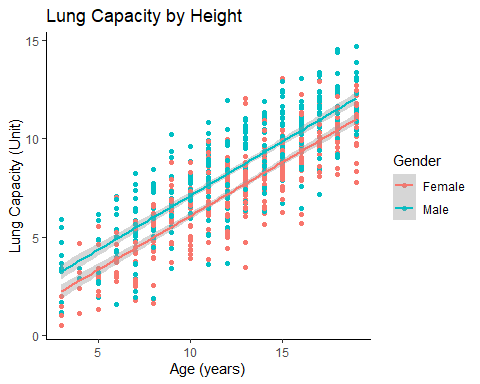
LCvH\_Scatter <- ggplot(lung\_cap, mapping=aes(x=Height, y=LungCap, color = Gender))+  
 geom\_point()+  
 geom\_smooth(method = lm, formula=y~x)+  
 theme\_classic()+  
 labs(x="Height (in)", y="Lung Capacity (Unit)", title="Lung Capacity by Height")  
LCvH\_Scatter



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

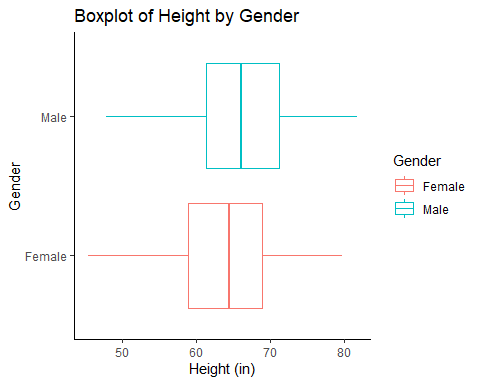


LCvA\_Scatter <- ggplot(lung\_cap, mapping=aes(x=Age, y=LungCap, color = Gender))+  
 geom\_point()+  
 geom\_smooth(method = lm, formula=y~x)+  
 theme\_classic()+  
 labs(x="Age (years)", y="Lung Capacity (Unit)", title="Lung Capacity by Height")  
LCvA\_Scatter



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

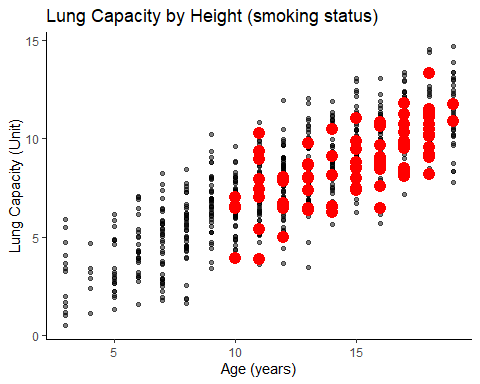


#Smokers analysis Separate the smokers and non-smokers data.

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

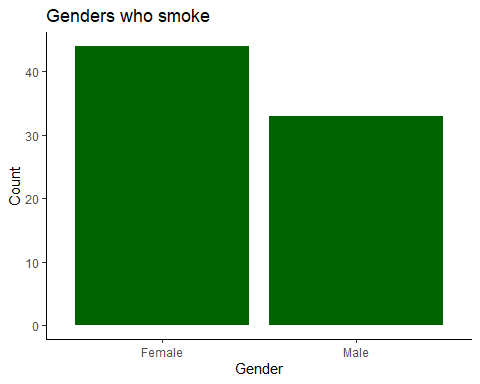


Distributions of Age and Gender for smokers in the dataset.

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



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).

