Module 2 R Practice

Angel Waters

2022-06-05

Descriptive Statistics

Descriptive statistics were calculated using various R functions, and then plotted to visually understand the relationships between the data.

Loading the necessary R packages for functions used throughout this report.

```
library(tidyverse)

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

## v ggplot2 3.3.5     v purrr   0.3.4
## 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

## -- Conflicts ------------------------------------------
tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

library(ggplot2)
library(Hmisc)

## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:dplyr':
##
##     src, summarize

## The following objects are masked from 'package:base':
##
##     format.pval, units

library(skimr)
```

To understand the Lung Capacity data, first it needs to be loaded into the R document. Understanding what is in the data table can help understand what type of analyses need to be performed on it.

```
lung <- read_csv("LungCapDataCSV.csv")

## Rows: 725 Columns: 6
## -- Column specification ----------------------------------------------------
------
## Delimiter: ","
## chr (3): Smoke, Gender, Caesarean
## dbl (3): LungCap, Age, Height
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this
message.

#Visualizing the columns headers and data types.  (NOTE: changing yes/no
columns to type=Logical, and capitalizing values)
lung <- mutate(lung, Smoke = as.logical(ifelse(Smoke=="no", FALSE, TRUE)),
                   Caesarean = as.logical(ifelse(Caesarean=="no", FALSE,
TRUE)),
                   Gender = ifelse(Gender=="male", "Male", "Female"))
names(lung)

## [1] "LungCap"   "Age"       "Height"    "Smoke"     "Gender"
"Caesarean"

str(lung)

## tibble [725 x 6] (S3: tbl_df/tbl/data.frame)
##  $ LungCap  : num [1:725] 6.47 10.12 9.55 11.12 4.8 ...
##  $ Age      : num [1:725] 6 18 16 14 5 11 8 11 15 11 ...
##  $ Height   : num [1:725] 62.1 74.7 69.7 71 56.9 58.7 63.3 70.4 70.5 59.2
...
##  $ Smoke    : logi [1:725] FALSE TRUE FALSE FALSE FALSE FALSE ...
##  $ Gender   : chr [1:725] "Male" "Female" "Female" "Male" ...
##  $ Caesarean: logi [1:725] FALSE FALSE TRUE FALSE FALSE FALSE ...
```

Numerical Data: Summary

Looking at all the data as a whole can help show what the population may be doing without confounding variables (we will look at the categorical data and how it interacts with the numerical data later). These function list all summary statistics (numerical and categorical) into table formats.

```
summary(lung)

##     LungCap            Age            Height        Smoke
##  Min.   : 0.507   Min.   : 3.00   Min.   :45.30   Mode :logical
##  1st Qu.: 6.150   1st Qu.: 9.00   1st Qu.:59.90   FALSE:648
```

```
##  Median : 8.000   Median :13.00   Median :65.40   TRUE :77
##  Mean   : 7.863   Mean   :12.33   Mean   :64.84
##  3rd Qu.: 9.800   3rd Qu.:15.00   3rd Qu.:70.30
##  Max.   :14.675   Max.   :19.00   Max.   :81.80
##     Gender          Caesarean
##  Length:725         Mode :logical
##  Class :character   FALSE:561
##  Mode  :character   TRUE :164
##
##
##

skim(lung)
```

Data summary

Name	lung
Number of rows	725
Number of columns	6

Column type frequency:	
character	1
logical	2
numeric	3

Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
Gender	0	1	4	6	0	2	0

Variable type: logical

skim_variable	n_missing	complete_rate	mean	count
Smoke	0	1	0.11	FAL: 648, TRU: 77
Caesarean	0	1	0.23	FAL: 561, TRU: 164

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
LungCap	0	1	7.86	2.66	0.51	6.15	8.0	9.8	14.68	▁▂▅▇▆▁
Age	0	1	12.3	4.0	3.00	9.00	13.	15.	19.0	▂▅▇▇▅

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
			3	0			0	0	0	
Height	0	1	64.84	7.20	45.30	59.90	65.4	70.3	81.80	▁▅▆▃▂

```
describe(lung)

## lung
##
##  6  Variables      725  Observations
## ---------------------------------------------------------------------------
------
## LungCap
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      725        0      342        1    7.863    3.021    2.965    4.250
##      .25      .50      .75      .90      .95
##    6.150    8.000    9.800   11.205   12.030
##
## lowest :  0.507  1.025  1.125  1.175  1.325, highest: 13.375 13.875 14.375
14.550 14.675
## ---------------------------------------------------------------------------
------
## Age
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      725        0       17    0.995    12.33    4.574      5.0      7.0
##      .25      .50      .75      .90      .95
##      9.0     13.0     15.0     18.0     18.8
##
## lowest :  3  4  5  6  7, highest: 15 16 17 18 19
##
## Value          3     4     5     6     7     8     9    10    11    12
13
## Frequency     13     6    20    25    37    41    40    51    58    68
69
## Proportion 0.018 0.008 0.028 0.034 0.051 0.057 0.055 0.070 0.080 0.094
0.095
##
## Value         14    15    16    17    18    19
## Frequency     56    64    54    43    43    37
## Proportion 0.077 0.088 0.074 0.059 0.059 0.051
## ---------------------------------------------------------------------------
------
## Height
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      725        0      274        1    64.84    8.215    52.00    55.14
##      .25      .50      .75      .90      .95
##    59.90    65.40    70.30    74.00    75.78
##
## lowest : 45.3 46.6 47.0 47.4 47.7, highest: 79.6 79.8 80.3 80.8 81.8
```

```
## ---------------------------------------------------------------------------
------
## Smoke
##        n  missing distinct
##      725        0        2
##
## Value      FALSE  TRUE
## Frequency    648    77
## Proportion 0.894 0.106
## ---------------------------------------------------------------------------
------
## Gender
##        n  missing distinct
##      725        0        2
##
## Value      Female   Male
## Frequency     358    367
## Proportion  0.494  0.506
## ---------------------------------------------------------------------------
------
## Caesarean
##        n  missing distinct
##      725        0        2
##
## Value      FALSE  TRUE
## Frequency    561   164
## Proportion 0.774 0.226
## ---------------------------------------------------------------------------
------
```

To make tables of the population parameters (mu, sigma, min, max, and median), R was used to calculate specific statistics and combine all into a data frame.

```
#Creating variables for the summary statistics.
mu_lc <- mean(lung$LungCap)
sigma_lc <- sd(lung$LungCap)
max_lc <- max(lung$LungCap)
min_lc <- min(lung$LungCap)
med_lc <- median(lung$LungCap)

mu_a <- mean(lung$Age)
sigma_a <- sd(lung$Age)
max_a <- max(lung$Age)
min_a <- min(lung$Age)
med_a <- median(lung$Age)

mu_h <- mean(lung$Height)
sigma_h <- sd(lung$Height)
max_h <- max(lung$Height)
min_h <- min(lung$Height)
```

```
med_h <- median(lung$Height)

#Turning the variables into vectors that will populate the data frame.
Column <- c("LungCap", "Age", "Height")
mu <- c(mu_lc, mu_a, mu_h)
sigma <- c(sigma_lc, sigma_a, sigma_h)
Maximum <- c(max_lc, max_a, max_h)
Minimum <- c(min_lc, min_a, min_h)
Median <- c(med_lc, med_a, med_h)

DescStat_Num <- data.frame(Column,mu, sigma, Maximum, Minimum, Median)
view(DescStat_Num)
```

Data appears to be normally distributed because the means and medians do not vary greatly for all the numerical values. Visualizing the data later will help verify this claim.

Categorical Data: Summary

Understanding the quantity of values in each bucket for categorical data can help understand what the frequency of each attribute is in the population. Frequency tables were made to understand the quantity of data within each bucket for the factors, character, and logical data.

```
table(lung$Smoke)

##
## FALSE  TRUE
##   648    77

table(lung$Gender)

##
## Female   Male
##    358    367

table(lung$Caesarean)

##
## FALSE  TRUE
##   561   164

DescStat_Gender <- data.frame(Gender=table(lung$Gender))
DescStat_Smoke <- data.frame(Smokers=table(lung$Smoke))
DescStat_Caesarean <- data.frame(Caesarean=table(lung$Caesarean))
```

Each categorical column had two possible outcomes. The gender had an even split for the two outcomes: male or female. Smokers and caesarean were not even, there was significantly smaller population if the result was TRUE.

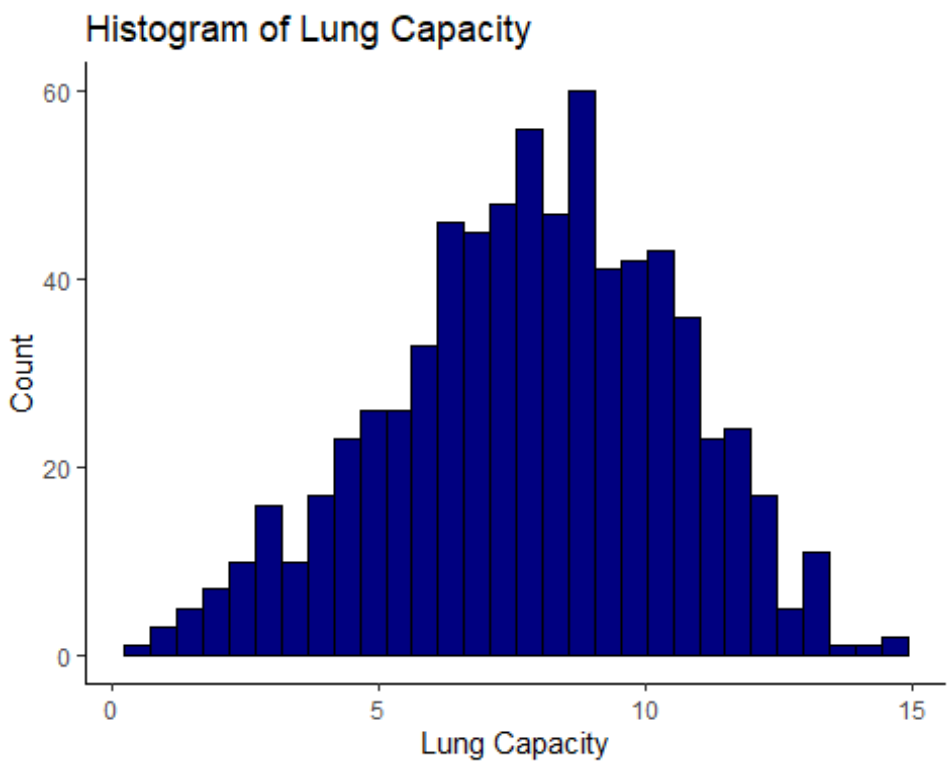
Three line tables

Three line tables help visualize categorical data by using multiple categories and counting the n in each category for specific groups.

Data Visualization

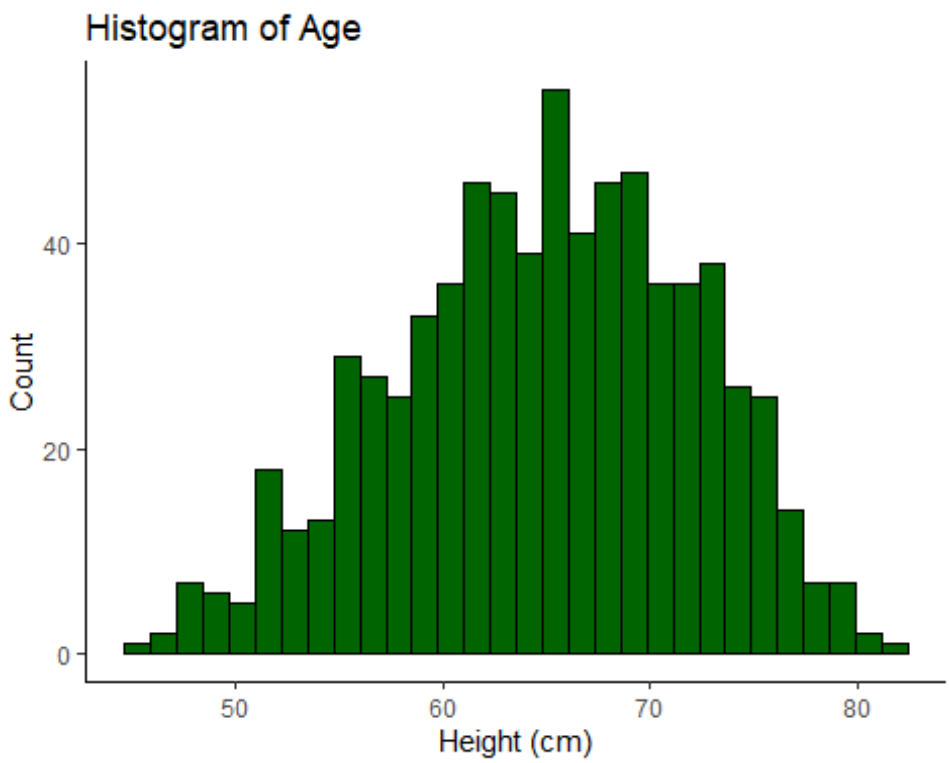
To see the spread of the numerical data, histograms were plotted in R.

```
hist_lc <- ggplot(lung)+
  geom_histogram(mapping=aes(LungCap), fill="Navy", color="Black")+
  theme_classic()+
  labs(title= "Histogram of Lung Capacity", x= "Lung Capacity", y="Count")
hist_lc

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
#Binwidth adjusted to 1 to have the appropriate bar width for the level of
significant figures the data provided.
hist_a <- ggplot(lung)+
  geom_histogram(mapping=aes(Age), fill="Maroon", color="Black", binwidth =
1)+
  theme_classic()+
  labs(title= "Histogram of Age", x= "Age (years)", y="Count")
hist_a
```

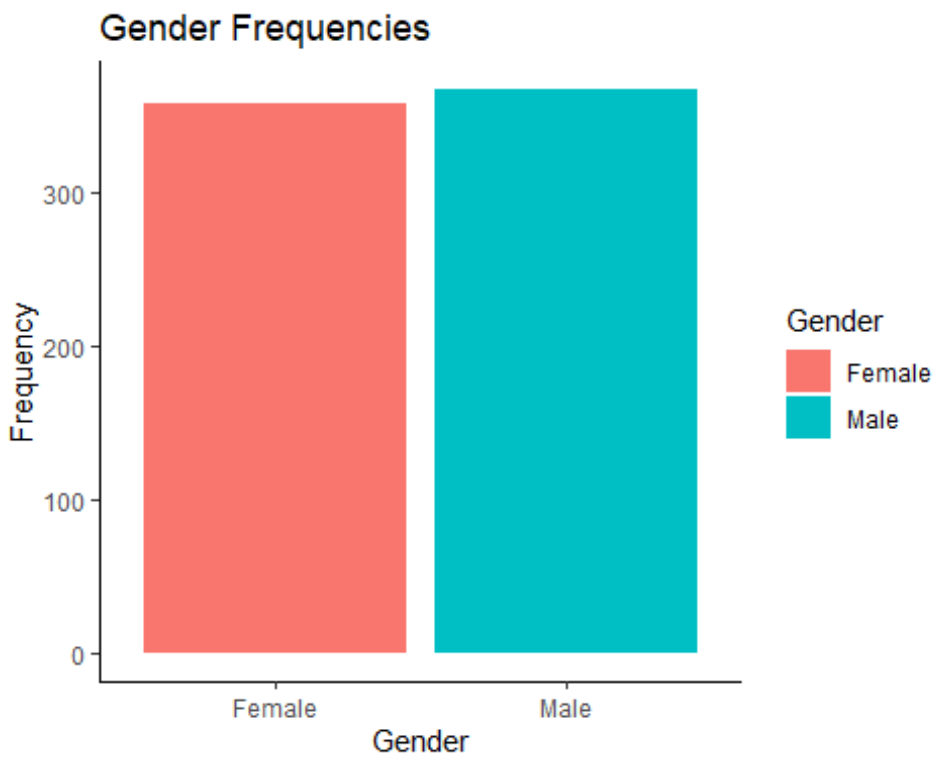

```
hist_h <- ggplot(lung)+
  geom_histogram(mapping=aes(Height), fill="Dark green", color="Black")+
  theme_classic()+
  labs(title= "Histogram of Age", x= "Height (cm)", y="Count")
hist_h

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



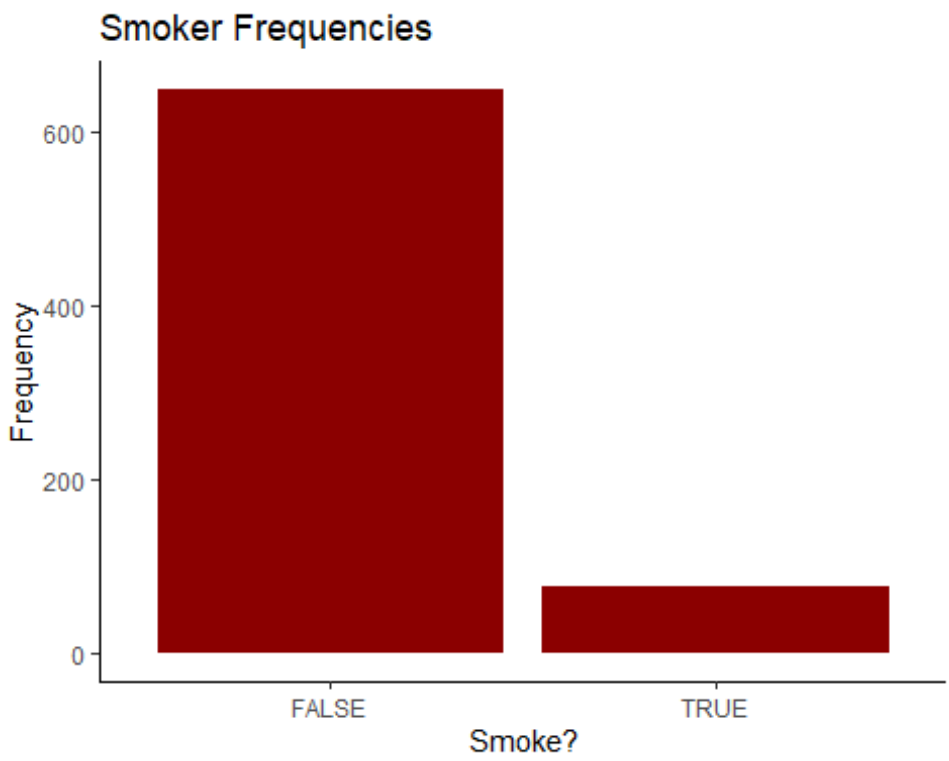
As stated above, the data appears to be normally distributed with the exception of age, where there is a slight negative skew.

Plotting the frequencies for the categorical data. geom_bar was used because the heights of the bars represent the number of values.

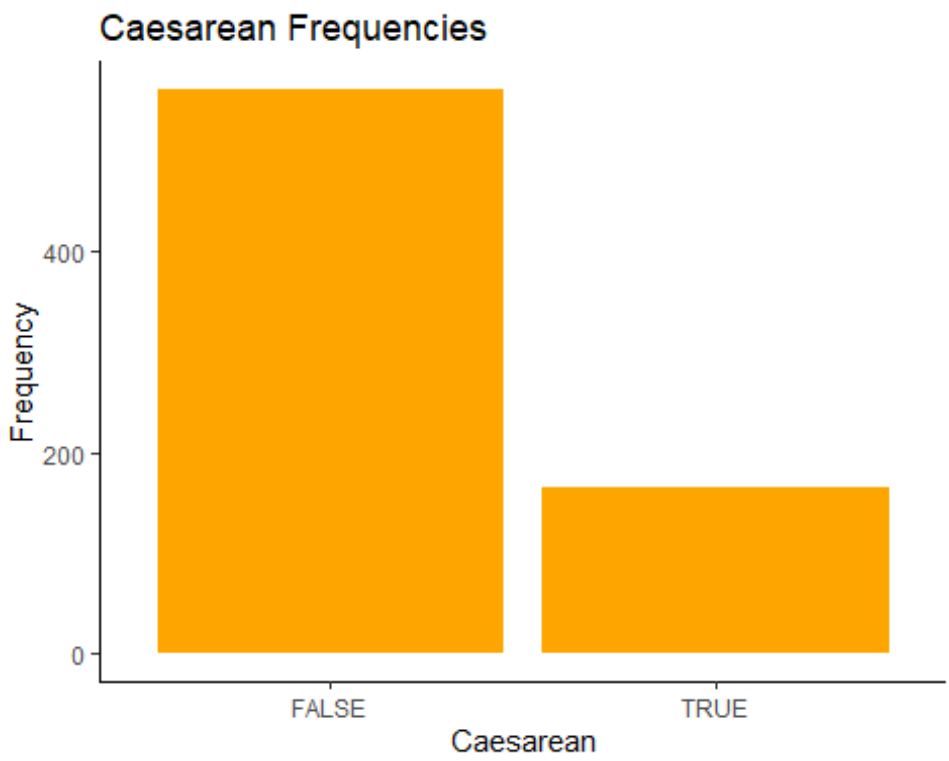
```
bar_gen <- ggplot(lung)+
  geom_bar(mapping=aes(Gender, fill=Gender))+
  theme_classic()+
  labs(title = "Gender Frequencies", y="Frequency")
bar_gen
```




```
bar_smo <- ggplot(lung)+
  geom_bar(mapping=aes(Smoke), fill="dark red")+
  theme_classic()+
  labs(title = "Smoker Frequencies", y="Frequency", x="Smoke?")
bar_smo
```




```
bar_cae <- ggplot(lung)+
  geom_bar(mapping=aes(Caesarean), fill="orange")+
  theme_classic()+
  labs(title = "Caesarean Frequencies", y="Frequency")
bar_cae
```



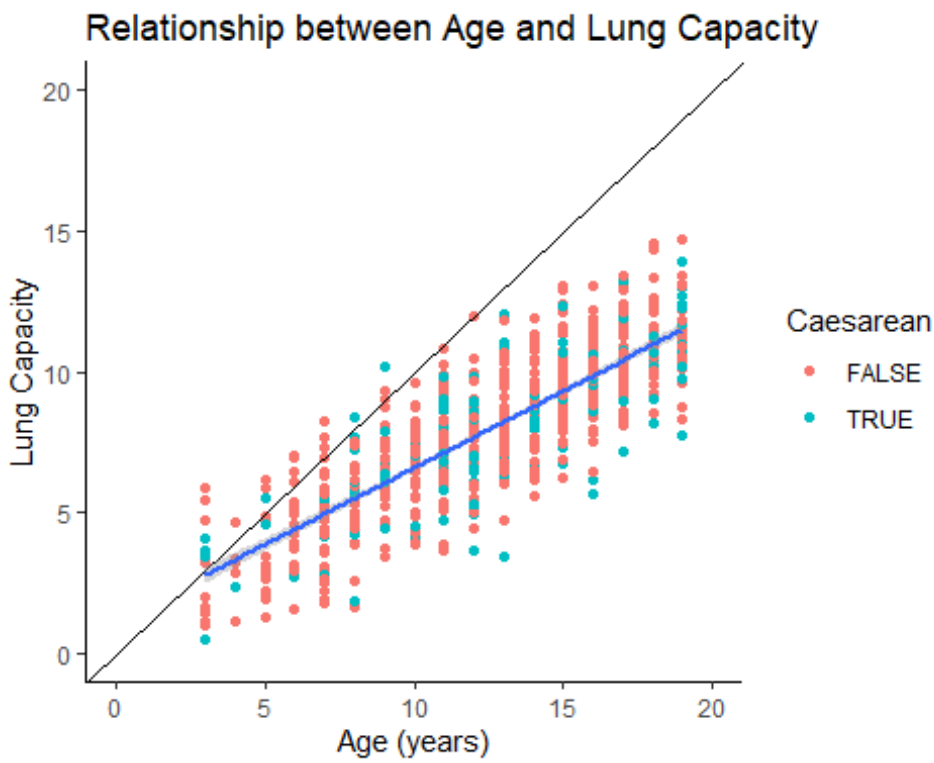
The visualizations show the same conclusions from the descriptive stats.

Relationships

Now understand the relationships between the categorical and numerical data.

To look at the relationship between age and lung capacity by Caesarean status, plotted against a y=x reference line.

```
scatter_avlc <- ggplot(lung, mapping=aes(Age, LungCap))+
  geom_point(mapping=aes(color=Caesarean))+
  geom_abline()+
  geom_smooth(method=lm, formula=y~x)+
  xlim(0,20)+
  ylim(0,20)+
  theme_classic()+
  labs(title= "Relationship between Age and Lung Capacity", y= "Lung
Capacity",
       x= "Age (years)")
scatter_avlc
```

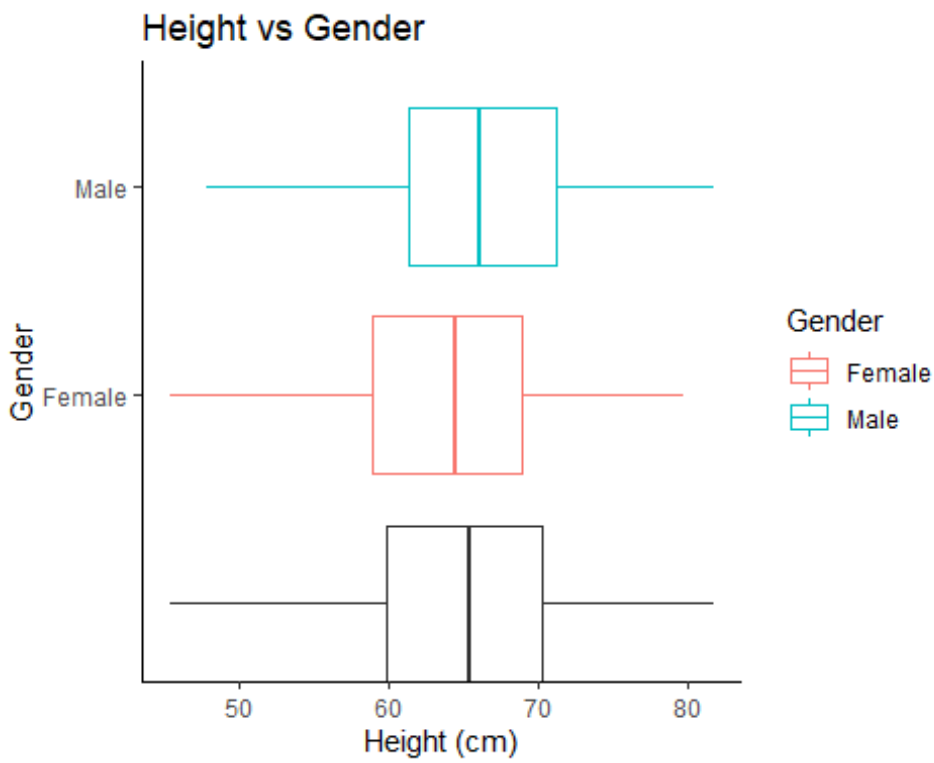



```
r_avlc <- cor(lung$Age, lung$LungCap)
r_avlc

## [1] 0.8196749

r2_avlc <- r_avlc^2
r2_avlc

## [1] 0.6718669
```

Because the units are not 1 for 1, the reference line may not be indicative of a 1 for 1 relationship. However there is a linear relationship between age and lung capacity as shown with the regression line in blue. The correlation between them is week when looking at the correlation coefficients (R=0.82, R^2=0.67).

Looking at gender vs height through boxplots to visualize the data and identify any outliers by gender.

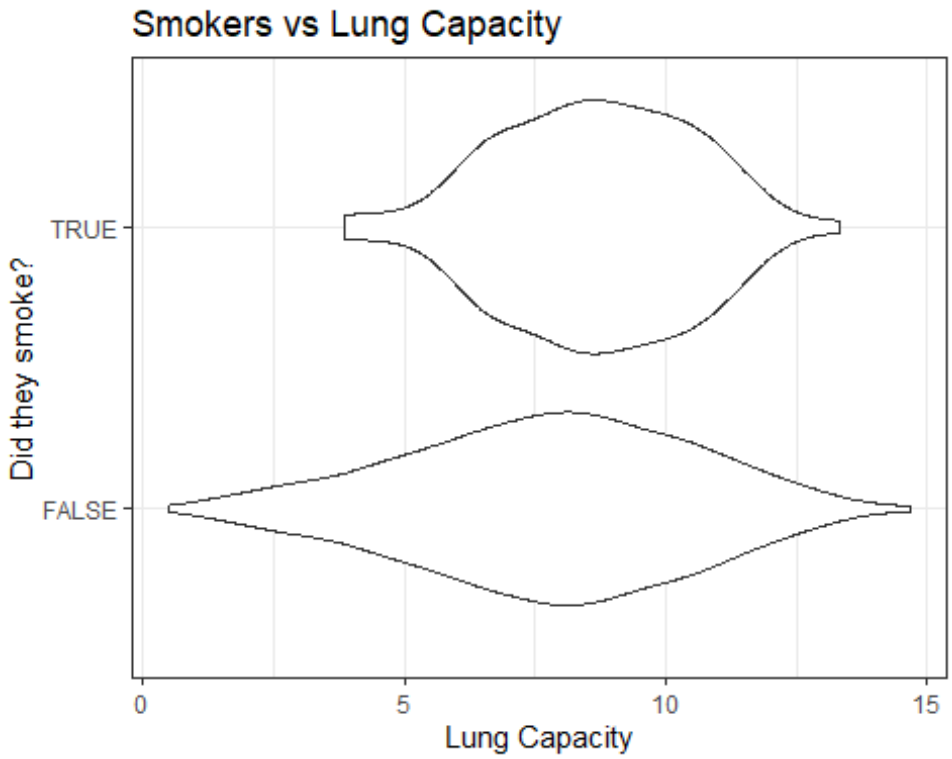
```
boxplot_gvsh <- ggplot(lung)+
  geom_boxplot(mapping=aes(y=Gender, x=Height, color=Gender))+
  geom_boxplot(mapping=aes(x=Height))+
  theme_classic()+
  labs(title="Height vs Gender", x="Height (cm)")
boxplot_gvsh
```



The overall data is shown in the black boxplot. There appears to be no outlier points shown in the boxplots for either gender. Females appear to have a slightly smaller median than males.

To understand what the lung capacity looks like for smokers vs non-smokers.

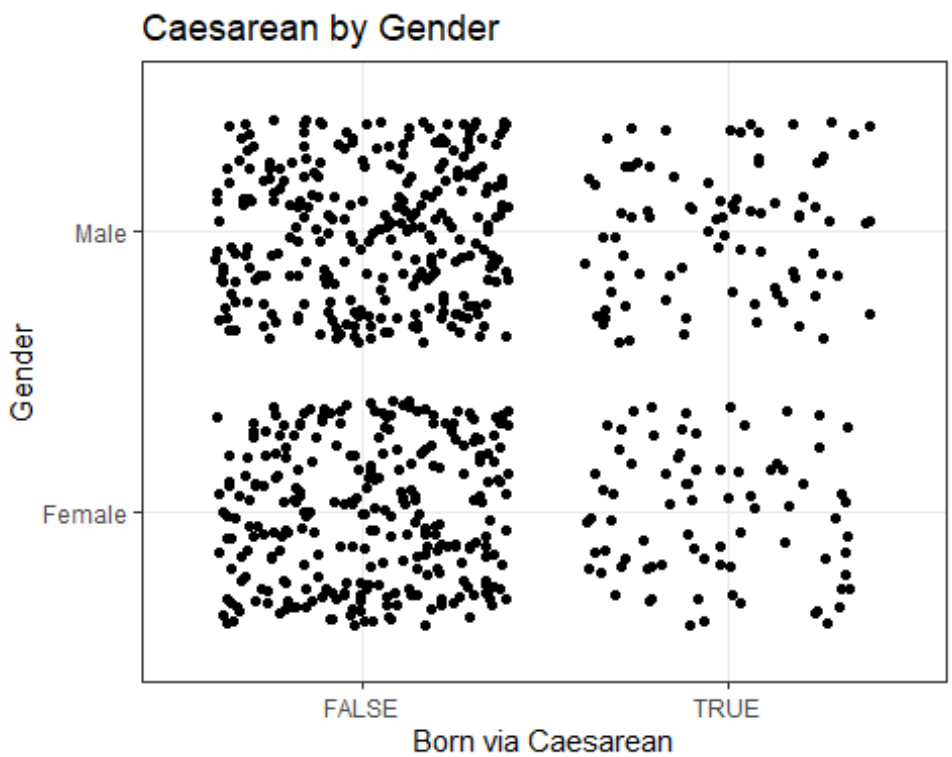
```
violin_lcvS <- ggplot(lung)+
  geom_violin(mapping=aes(x=LungCap, y=Smoke))+
  theme_bw()+
  labs(title= "Smokers vs Lung Capacity", x="Lung Capacity", y="Did they
smoke?")
violin_lcvS
```



The smokers appear to have data on the higher end of the lung capacity scale whereas non-smokers span the range. The data is slightly confounded by age and the sample size is smaller for the smokers.

To visualize two discrete variables, jitter was used to create clouds of points for each category and subcategory.

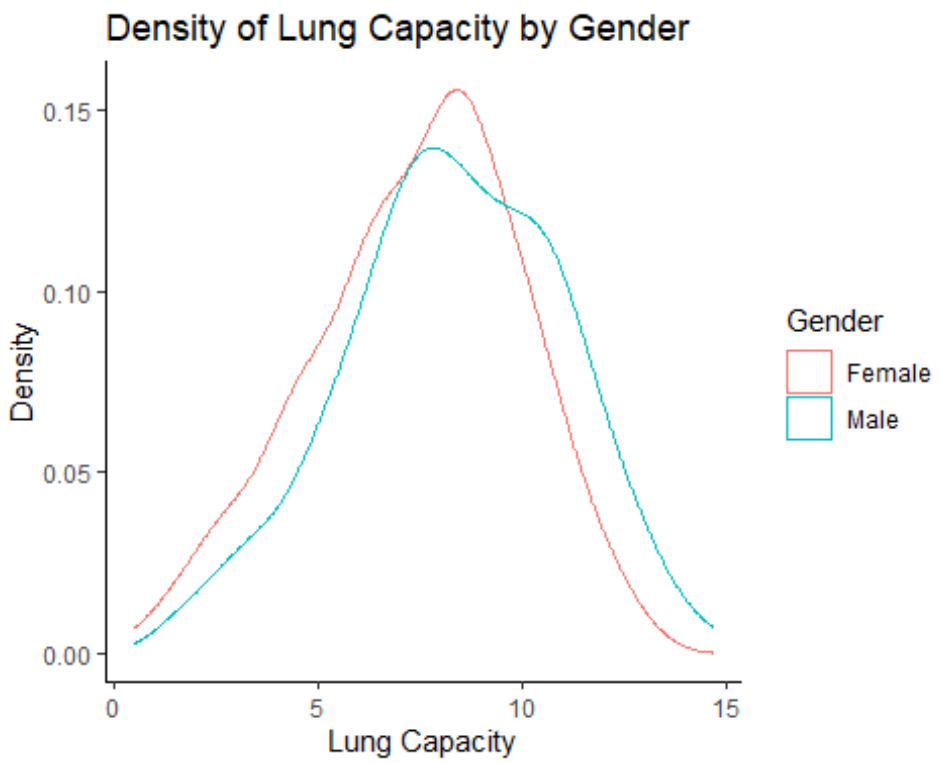
```
jitter_cvsg <- ggplot(lung)+
  geom_jitter(mapping=aes(Caesarean, Gender))+
  theme_bw()+
  labs(title= "Caesarean by Gender", x="Born via Caesarean", y="Gender")
jitter_cvsg
```



Majority of the data, despite gender, were not born via a Caesarean. The genders do show a similar trend and have approximately even spread within each category.

To show the spread of the lung capacity data by gender, a density plot was made using ggplot.

```
density_lcvg <- ggplot(lung)+
  geom_density(mapping=aes(LungCap, color=Gender))+
  theme_classic()+
  labs(title="Density of Lung Capacity by Gender", y="Density", x="Lung
Capacity")
density_lcvg
```



Peaks are located in a similar area however more females are found at their respective most dense area of the data. The males density chart shows a slight positive offset in the data, meaning there were more males at the higher end than females for this data set.