STA 141 Worksheet 8

Richard McCormick

November 7, 2023

Due Date: Thursday, November 16, 2023 before 11:00am.

Instructions

Worksheets must be turned in as a PDF file through Canvas. The worksheet is worth a total of **15 points**, which is 3 percent of your overall grade.

Exercises

Begin by running the following code block to add the packages we need to use to our library.

Exercise 1

(a) This question will use the `MouseBrain` dataset from the `Stat2Data` package.

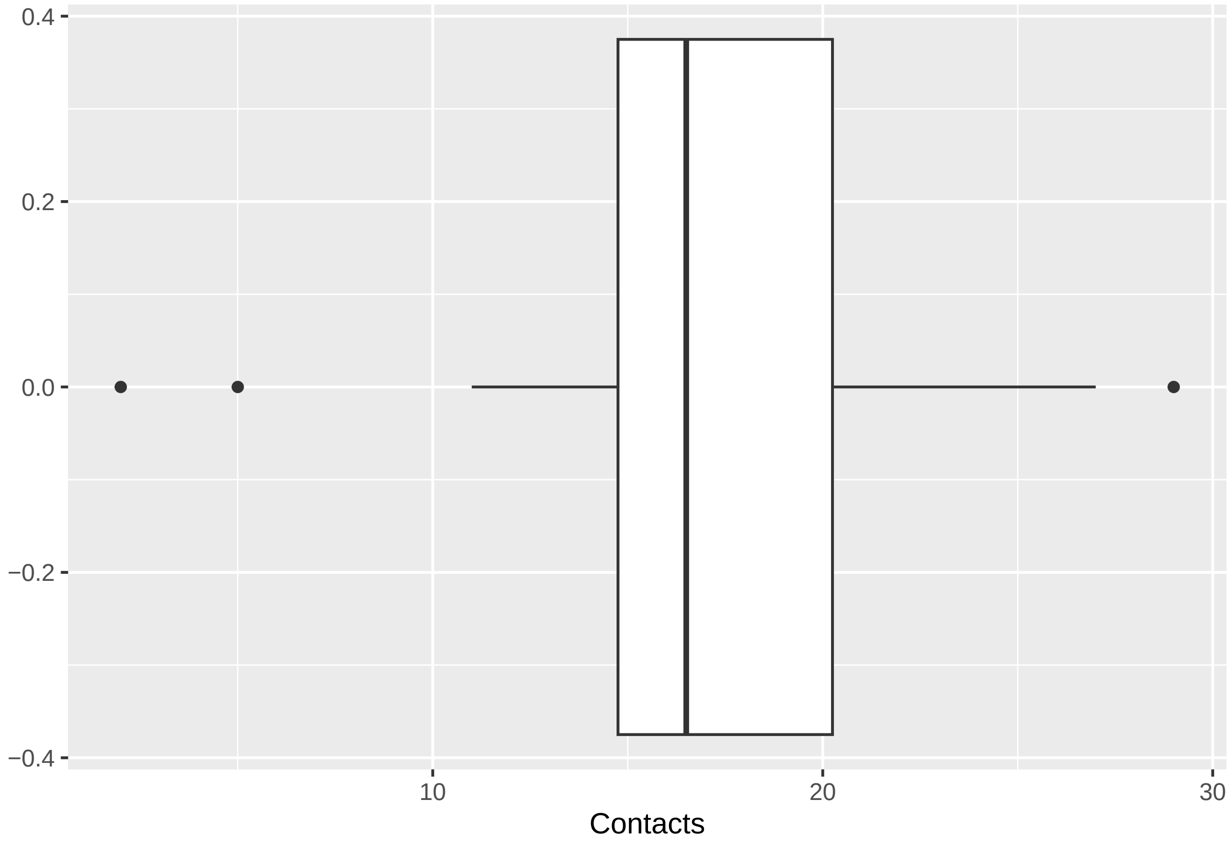
```
data("MouseBrain")
my.mice <- MouseBrain
```

Use the help documents and commands we've learned to understand what is being shown in the dataset. Describe the data in one or two sentences below:

The data appears to be a long dataset showing behavior in mice. The number of social contacts, the sex, and the genotype of each mouse make up a row, with each column describing one of these categories.

(b) Create a boxplot of the `Contacts` variable.

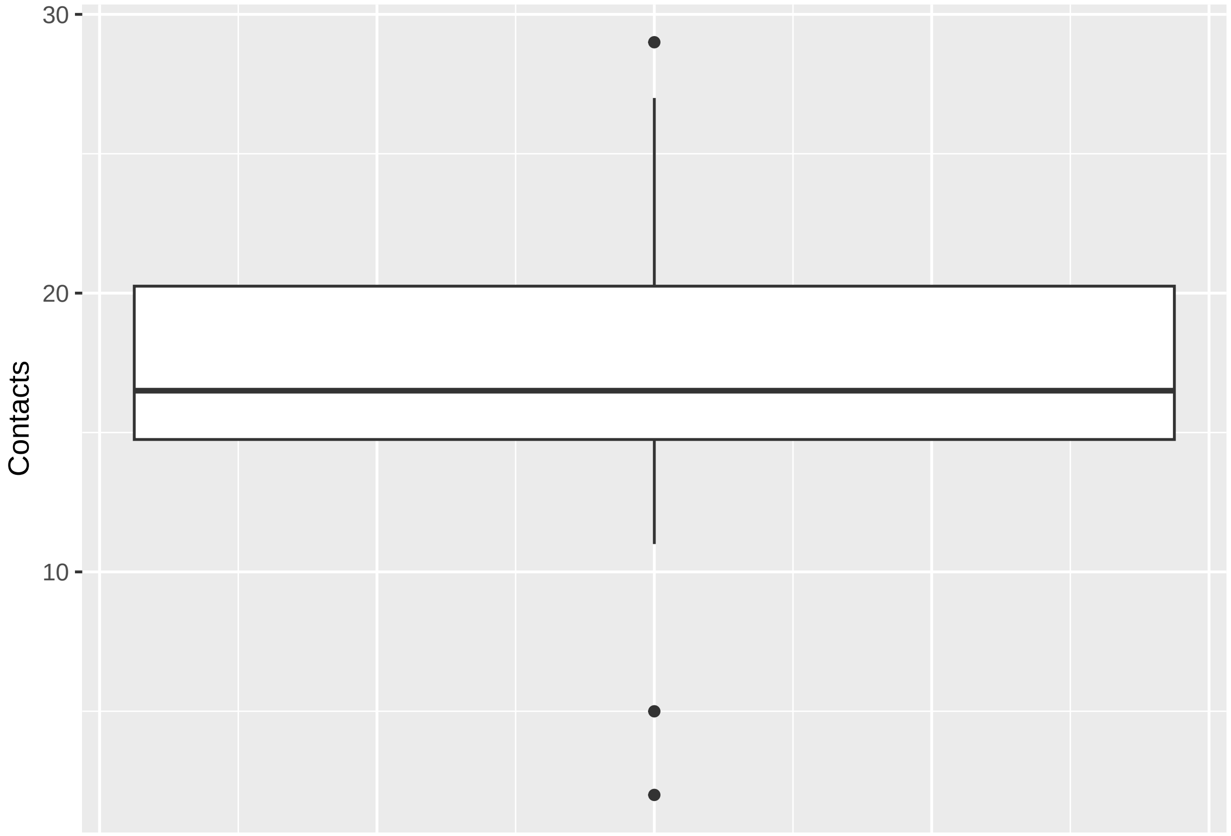
```
ggplot( data=my.mice, aes( x=Contacts ) ) +
  geom_boxplot()
```



(c) Change the orientation so that we have a vertical boxplot.

Note: to remove all the unnecessary ticks and marks add the following to the end of your command + theme(axis.ticks.x = element_blank(), axis.text.x = element_blank()).

```
ggplot( data=my.mice, aes( y=Contacts ) ) +
  geom_boxplot() +
  theme(axis.ticks.x = element_blank(), axis.text.x = element_blank())
```

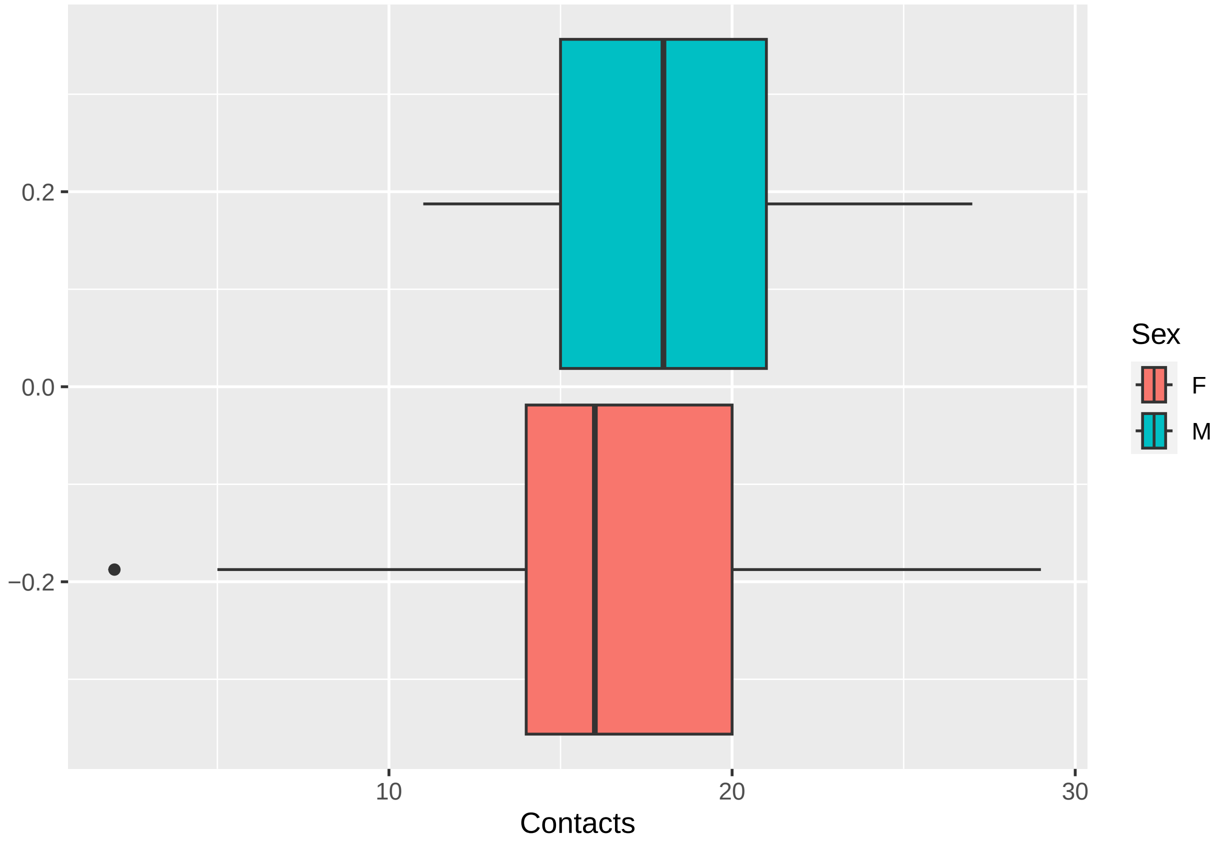


(d) Use a boxplot to compare the social contacts of Female and Male mice.

Note: If you add an argument width={value} inside your geom_histogram() function it will adjust the width of the box(es).

```
ggplot( data=my.mice, aes( x=Contacts ) ) +
  geom_boxplot( aes( group_by=Sex, fill=Sex ) )
```

```
## Warning in geom_boxplot(aes(group_by = Sex, fill = Sex)): Ignoring unknown
## aesthetics: group_by
```



(e) Is the median number of contacts of Female or Male mice greater? Calculate the medians using `group_by` and `summarize` to find the difference.

The median number of contacts is slightly higher for Male mice over Female mice.

```
my.mice %>%
  group_by( by=Sex ) %>%
  summarize( Median.Contacts=median( Contacts ) )
```

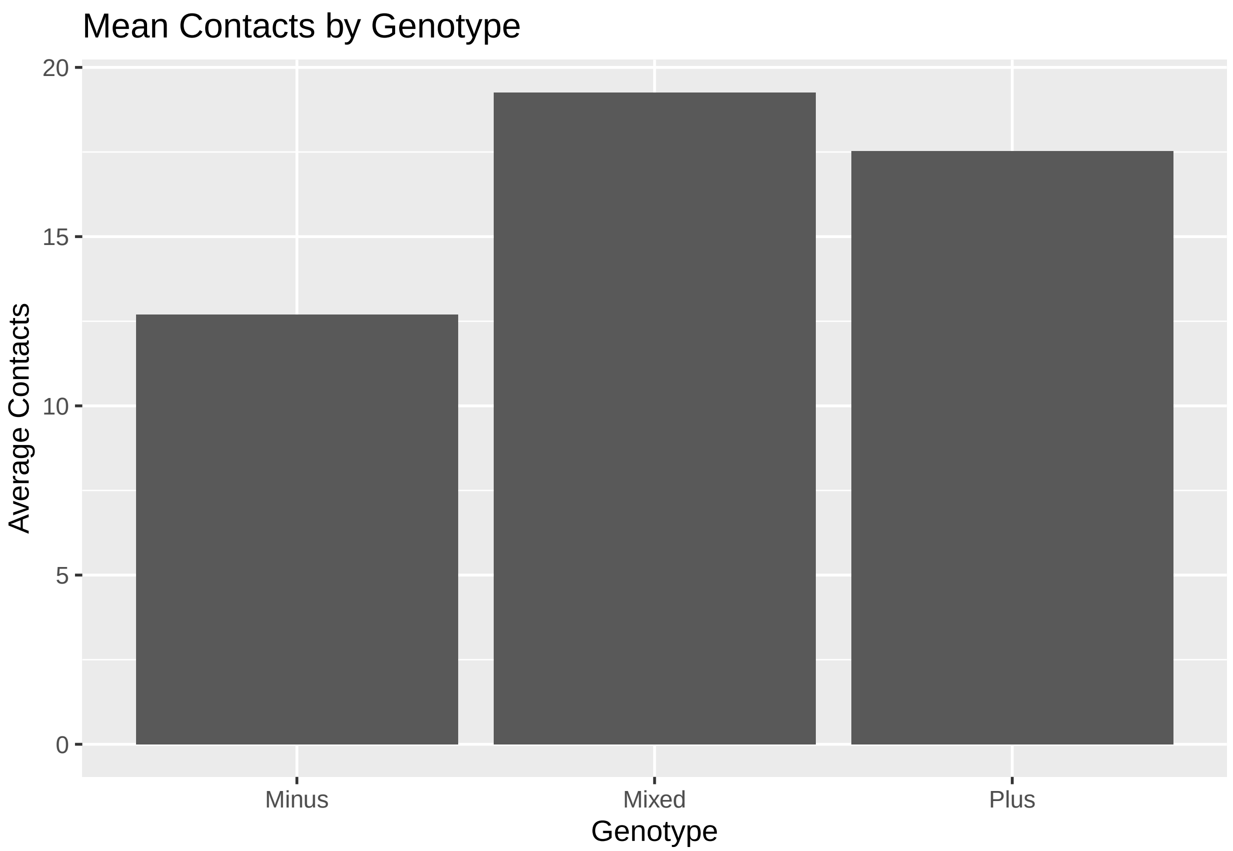
```
## # A tibble: 2 x 2
##   by    Median.Contacts
##   <fct>           <int>
## 1 F                  16
## 2 M                  18
```

(f) Which Genotype has the highest mean number of social contacts? Use a bar graph to verify your answer.

The 'mixed' genotype had the highest number of average contacts, followed closely by the Plus genotype. Minus genotypes had the lowest overall number of average contacts.

```
mean.mice <- my.mice %>%
  group_by( by=Genotype ) %>%
  summarize( Mean.Contacts=mean( Contacts ) )

ggplot( data=mean.mice, aes( x=by, y=Mean.Contacts ) ) +
  geom_bar( stat='identity' ) +
  labs( title="Mean Contacts by Genotype", x="Genotype", y="Average Contacts" )
```



Exercise 2

This question will use the `HousesNY` dataset from the `Stat2Data` package.

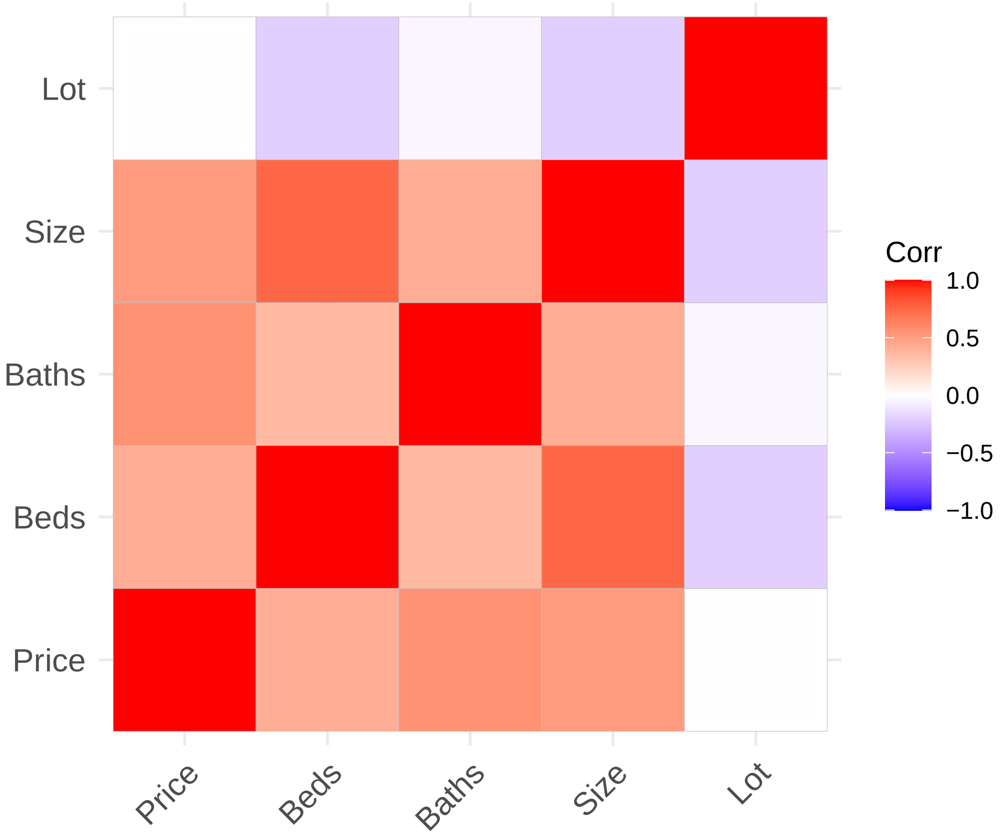
```
data("HousesNY")
my.houses <- HousesNY
```

The question will also require the use of the `ggcorrplot` package. You will need to run `install.packages("ggcorr`
before doing anything else.

(a) Find the correlation between each pair of numerical variables and create a correlation plot using the `ggcorrplot()` function.

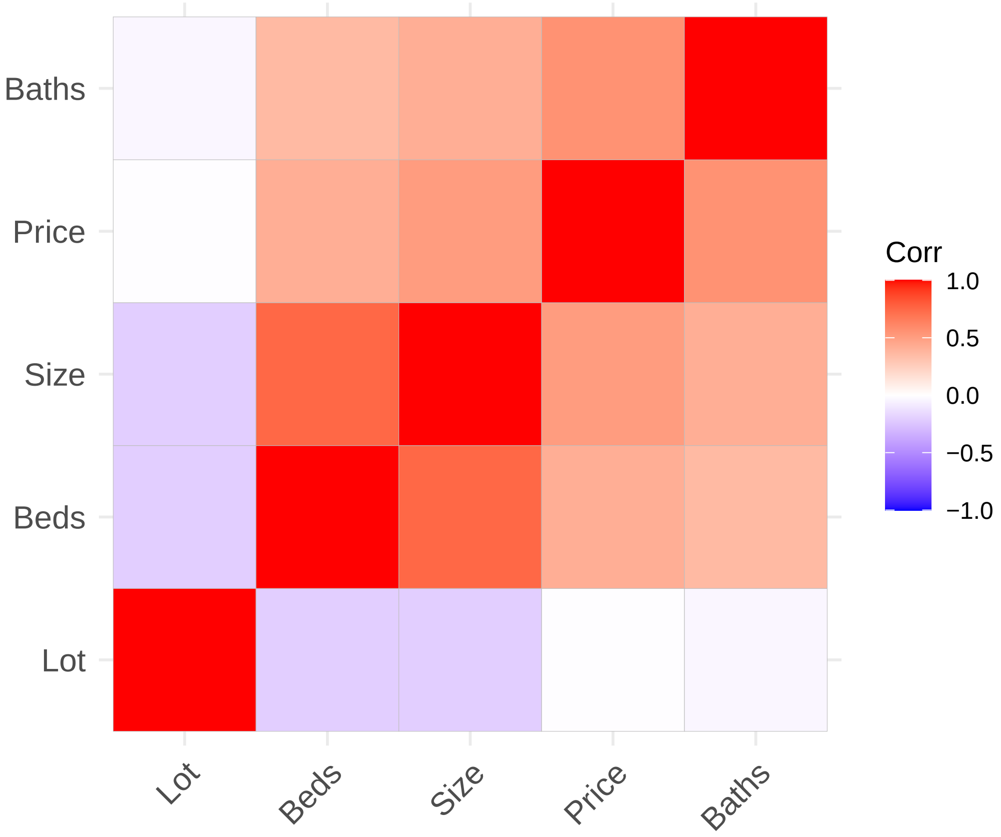
```
houses.corr <- cor( my.houses )

ggcorrplot( houses.corr )
```



(b) Work out how to add a `hc.order = TRUE` argument to command above. This should order the variables by their correlation.

```
ggcorrplot( houses.corr, hc.order=TRUE)
```



(c) Is the size of the house and its price correlated in NYC? What about the size of the lot and its price?

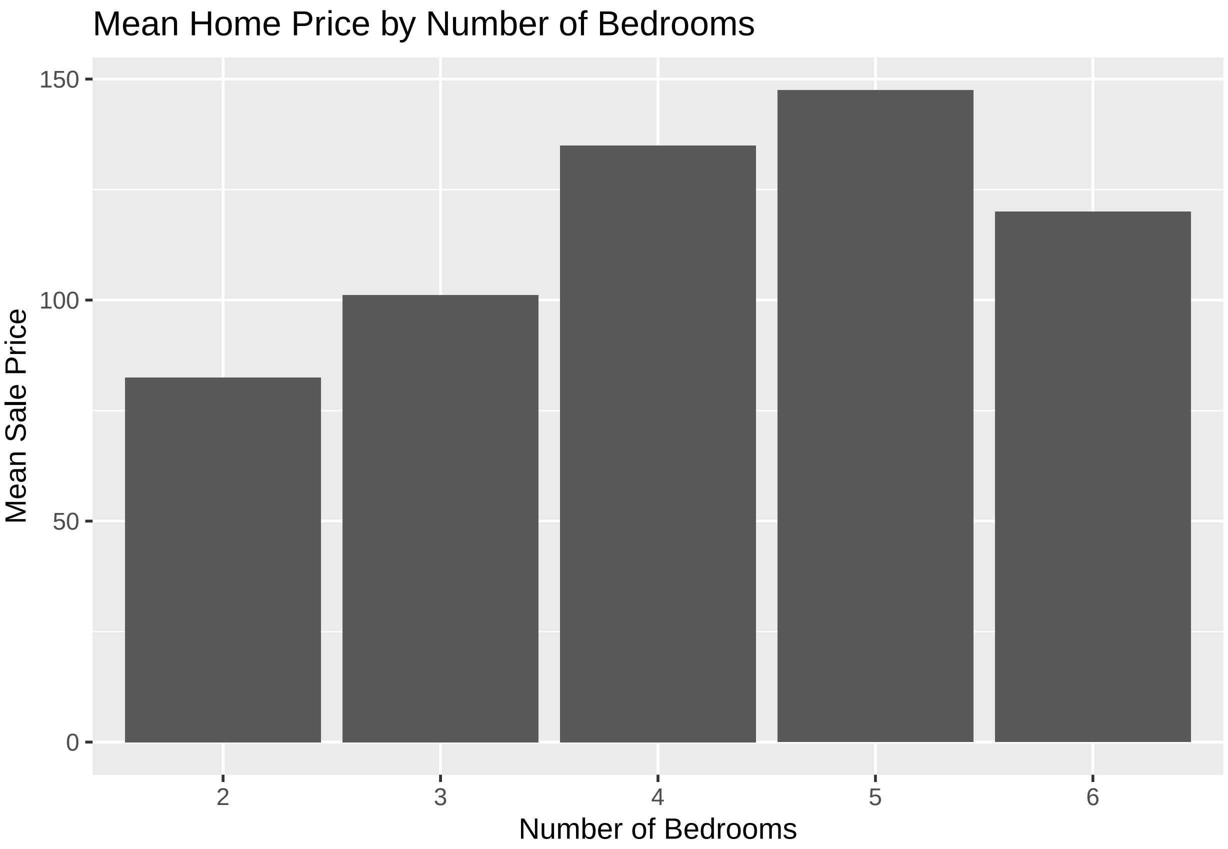
The size of a house is moderately correlated to its price. On the other hand, the size of the lot is almost completely uncorrelated to its price.

(d) Create a bar chart of the mean price of a house against the number of bedrooms in NYC. Note that you will need to make number of bedrooms a factor, if you haven't already.

```r
my.houses$Beds <- as.factor( my.houses$Beds )

my.houses.mean <- my.houses %>%
  group_by( Beds ) %>%
  summarize( mean.Price = mean( Price ) )

ggplot( data=my.houses.mean, aes( x=Beds, y=mean.Price ) ) +
  geom_bar( stat='identity' ) +
  labs( title="Mean Home Price by Number of Bedrooms",
        x="Number of Bedrooms", y="Mean Sale Price" )
```


Mean Home Price by Number of Bedrooms
Mean Sale Price
Number of Bedrooms