Week 5: The Importance of Exploratory Analysis

09/26/2019

Jake Campbell

What is Exploratory Analysis?

- Exploratory analysis is an initial investigative approach to our data prior to more complex analysis
- Helps us get a good understanding of any general patterns in our data
- Very visual-based
    - Lots of plotting!

Why is Exploratory Analysis Important

- Machine learning models are great, but they do not have the eye that humans do
    - An important variable to the model may just be the result of an error in our data that we would've captured with exploratory analysis
- By doing some initial investigation with our data, we can build better models down the road
- Important to get a feel for our data
    - You may not have a ton of knowledge about the topic you're analyzing, but it's important to build some understanding!

Summary Statistics

- Summary statistics give us a quick overview of different aspects of our data (how spread out it is, what's the typical value, etc.)
- Good for quick insight, but don't always tell the whole story
    - Ex: `mean` of a skewed variable

Common Summary Statistics

- The mean is the average of a vector of values
  - `R` function is `mean()`
- The median is the middle value in a vector of values
  - `R` function is `median()`

```
mean(iris$Sepal.Length)
```

```
## [1] 5.843333
```

```
median(iris$Sepal.Length)
```

```
## [1] 5.8
```

Common Summary Statistics

- Quantiles divide observations into bins
  - The median is a quantile (splits the data in two)
  - R function is `quantile()`

```
quantile(iris$Sepal.Length)
```

```
##   0%  25%  50%  75% 100%
##  4.3  5.1  5.8  6.4  7.9
```

```
# Specify where the splits are
quantile(iris$Sepal.Length, probs = c(0, .33, .66, 1))
```

```
##    0%   33%   66%  100%
## 4.300 5.400 6.234 7.900
```

Common Summary Statistics

- Variance is the average squared difference of a set of observations from the mean
  - $\frac{\sum(x_i - mean(x))^2}{n-1}$
  - Commonly represented as σ^2
  - `R` function is `var()`
- Standard deviation is square root of the variance
  - Commonly represented as σ
  - `R` function is `sd()`
  - Puts variance measure on the level of the data

Common Summary Statistics

```
var(iris$Sepal.Length)
```

```
## [1] 0.6856935
```

```
sd(iris$Sepal.Length)
```

```
## [1] 0.8280661
```

Common Summary Statistics

- Covariance combines the variance of two variables, `x` and `y` and says whether they have a positive or negative relationship
  - $\frac{\sum(x_i - mean(x))(y_i - mean(y))}{n-1}$
  - `R` function is `cov()`
- Correlation divides the covariance of `x` and `y` by the standard deviations of `x` and `y`
  - $\frac{cov(xy)}{sd(x)sd(y)}$
  - Puts the covariance on a scale from `-1` to `1`
  - `R` function is `cor()`

Common Summary Statistics

```
cov(iris$Sepal.Length, iris$Sepal.Width)
```

```
## [1] -0.042434
```

```
cor(iris$Sepal.Length, iris$Sepal.Width)
```

```
## [1] -0.1175698
```

Visualizing Our Data

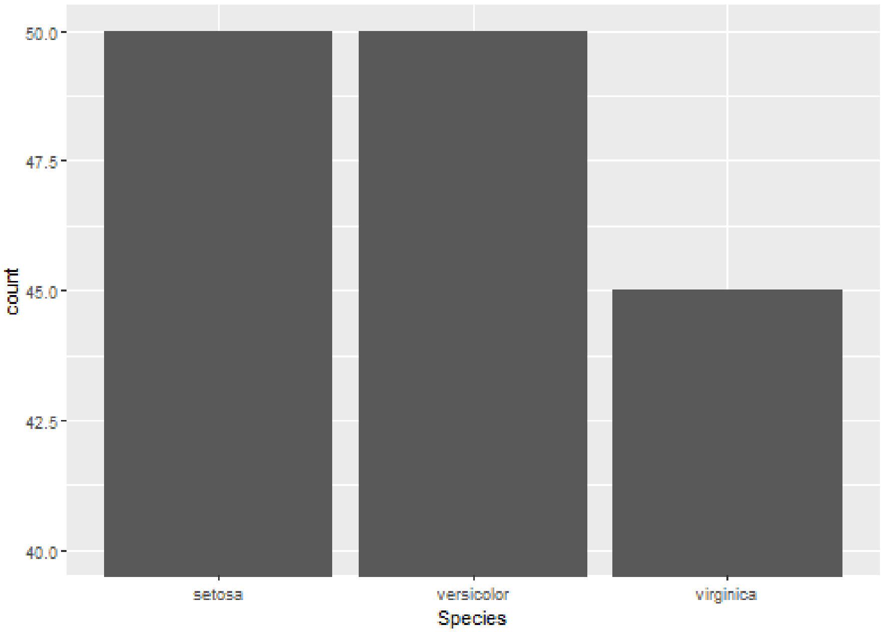
- The best method for us to investigate our data is visually
- We can easily identify patterns or notice issues with our data that we would miss otherwise
- It's important that we use plots correctly!

When to Use What Plot

- Scatterplots are great for showing the relationship between continuous variables
- Histograms and density plots help us visualize the distribution of a variable
- Boxplots can show the relationship between a continuous and categorical variable
- Barplots are good for giving count data of a categorical variable

Common Plot No-no's

- Don't mess with the limits of your plot
- This can easily be manipulated to show relationships that aren't there



Common Plot No-no's

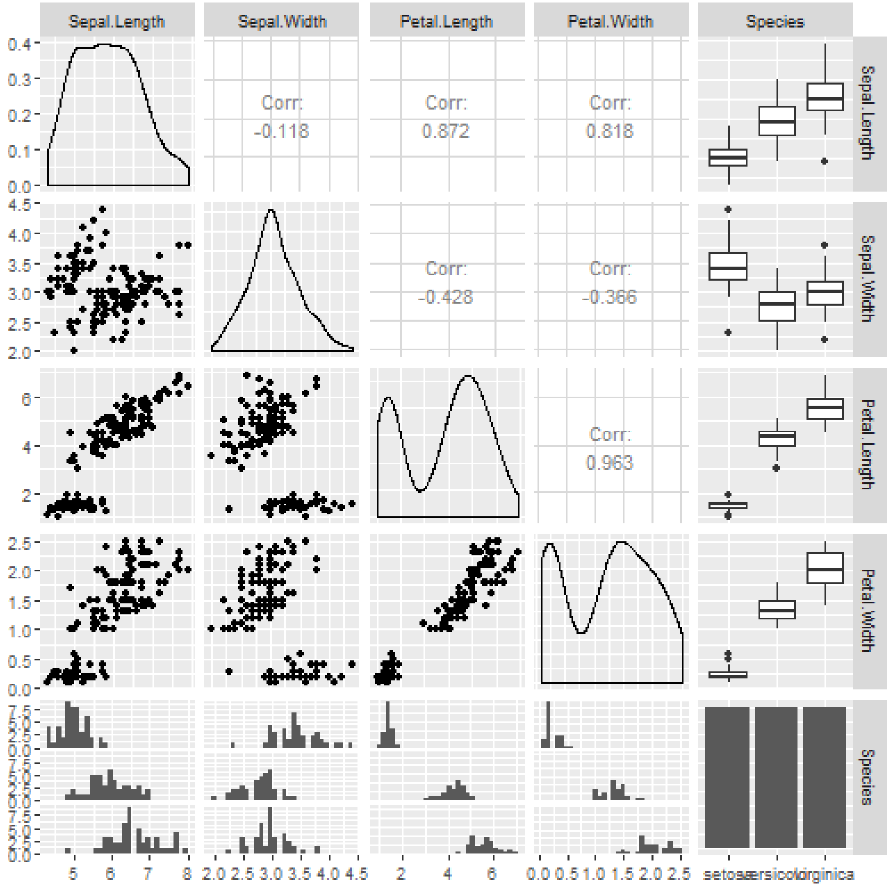
- Avoid pie charts
- Although very popular, hard to interpret exact differences between groups
- Much easier to visualize in a bar chart

Creating Multiple Plots at Once

- We can use the `GGally` package to create several extensions off of `ggplot2`
- One of these extensions is a plot matrix
  - Creates a matrix of plots, each showing a relationship between different variables
  - Function to use is `ggpairs`

ggpairs

```
ggpairs(iris)
```


Sepal.Length
Sepal.Width
Petal.Length
Petal.Width
Species
Corr:
Corr:
Corr:
Corr:
Corr:
Corr:
Sepal.Length
Sepal.Width
Petal.Length
Petal.Width
Species
setosaversicolorvirginica