Data Science Series: Exploratory Data Analysis

Natya Hans

2024-04-25

SET UP

```r
# Create a vector of package names
all.lib<-c("tidyverse","ggplot2", "tidyr",
           "dplyr","modelr")

# install packages
#install.packages(all.lib)

# Load packages
lapply(all.lib,require,character.only=TRUE)
```

```
## [[1]]
## [1] TRUE
##
## [[2]]
## [1] TRUE
##
## [[3]]
## [1] TRUE
```

Generate questions and hypothesis about the data.

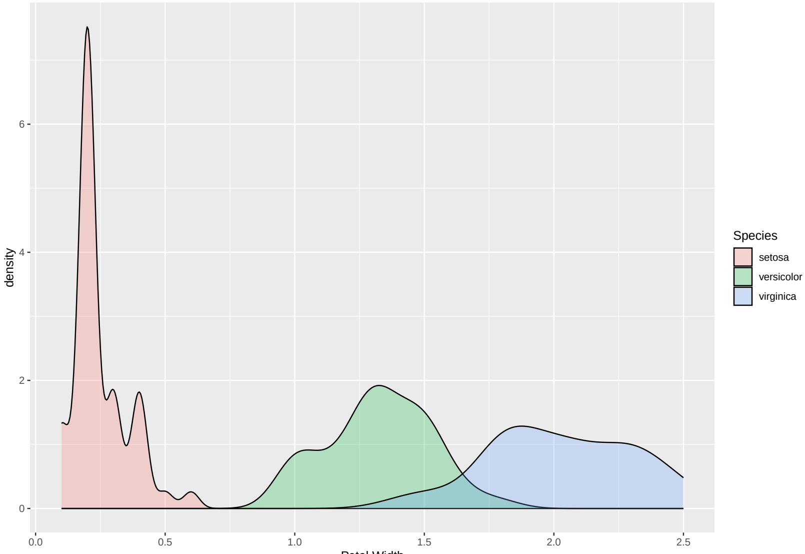
- Understand your data
- Read the metadata if the data is not yours
- Think about the analysis plan led by questions
- Make sure your hypothesis-driven studies are clearly stated
- Multiple questions are often better

Load your data and explore

```
#ncol()
#nrow()
#dim()
#str()
#summary()
#head()
#tail()
#table()
```

Look for answers and patterns in the data by using visualization techniques

- For example Iris dataset:

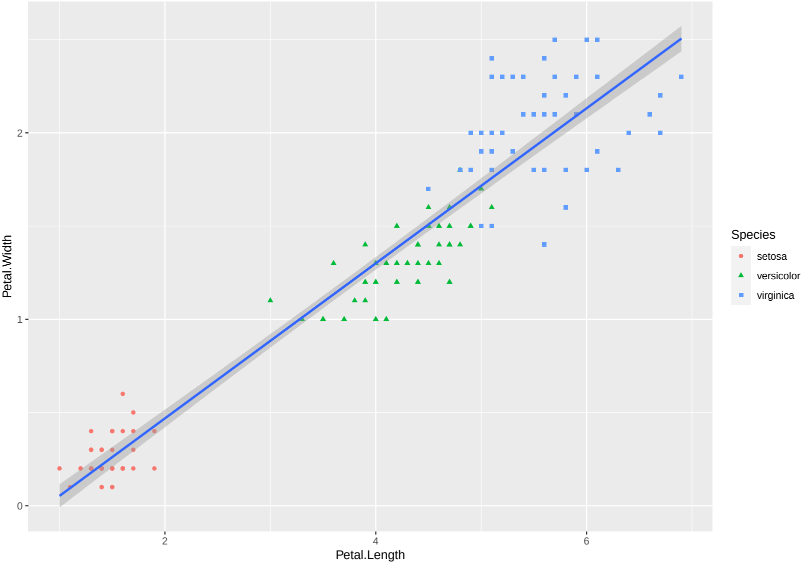


Transformations of data

- Some common functions:

```
# mutate()
# group_by()
# summarize()
# arrange()
# glimpse()
# select()
# filter()
```

Modelling the data



Refine the questions based on what you learn and repeat the process

Most common questions:

- Variation in data
- Covariation within variables in data
- Univariate Analysis
- Multivariate Analysis

Definitions

- A variable is a quantity, quality, or property that you can measure.
- A value is the state of a variable when you measure it.
- An observation is a set of measurements made under similar conditions.
- Tabular data is a set of values, each associated with a variable and an observation.

Note for tidy data:

- Each column is a variable
- Each row is an observation

Resources

1. R Cookbook http://www.cookbook-r.com/
2. ggplot2 tutorials https://r-graph-gallery.com/ggplot2-package.html
3. Interactively learn R https://www.programiz.com/r
4. ggplot2 https://ggplot2.tidyverse.org/