ST 308 Final Project Review Session

Carter Hall

2023-11-30

Welcome

Hello, everyone! This review session is geared to be “interactive” – I have questions that I will ask you and then promptly go over them together. Think of this as a guided, larger ICA that has *some* overlap with the final project! (Any additional time will be normal office hours!)

Repository for this Review Session

Please see the link for the dataset and associated .Rmd file. **It is advised that you work alongisde everyone!**

https://github.com/halljc76/ST308_ReviewSession

How To Use This

1. Go to the link.
2. Click the green button that says 'Code'
3. Scroll down and hit 'Download ZIP'
4. Extract the ZIP file to a particular directory, somewhere in your computer. (Alternatively, if you just want to create a new .Rmd file and get the full filepath to the .csv file to read in, after extracting, that is fine!)
5. **Highly suggested, but not required for this nor the final project.** Click the R Project logo in the top right of RStudio, click 'New Project'. Then, click 'Existing Directory', 'Browse', and find the extracted folder. (This will... or should... ensure filepaths are all in-order!)

Data Import

Something we've done a thousand times, but make it 1001!

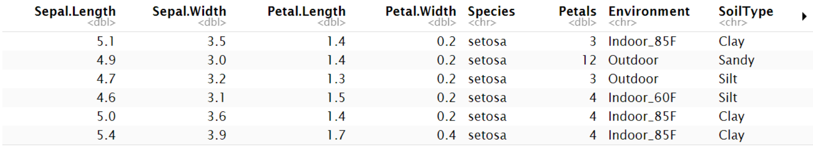
How to? Note: This is a .csv file; so, within the tidyverse package, we should use. . .

Data Import

Something we've done a thousand times, but make it 1001!

How to? Note: This is a .csv file; so, within the tidyverse package, we should use... `read_csv()`!

```
review_data <- read_csv("modified_iris.csv")
head(review_data)  # Look at first few rows of the
data
```



Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species	Petals	Environment	SoilType
<dbl>	<dbl>	<dbl>	<dbl>	<chr>	<dbl>	<chr>	<chr>
5.1	3.5	1.4	0.2	setosa	3	Indoor_85F	Clay
4.9	3.0	1.4	0.2	setosa	12	Outdoor	Sandy
4.7	3.2	1.3	0.2	setosa	3	Outdoor	Silt
4.6	3.1	1.5	0.2	setosa	4	Indoor_60F	Silt
5.0	3.6	1.4	0.2	setosa	4	Indoor_85F	Clay
5.4	3.9	1.7	0.4	setosa	4	Indoor_85F	Clay

Chaining Operator, %>%

Arguably the G.O.A.T of operators in any non-base-R package. For our data, let's

1. Remove any observations where Species is equal to `virginica`.
2. Create a new variable `MoreThan10` that indicates if a flower has more than 10 petals.
3. Remove the `Sepal.Length` variable.

Chaining Operator, %>%

Arguably the G.O.A.T of operators in any non-base-R package. For our data, let's remove any observations where

```
review_data <- review_data %>% filter(Species !=
"virginica") %>% mutate(MoreThan10 = Petals > 10) %>%
select(-Sepal.Length)
```

Remember: Approach these problems **step-by-step**! Break things down. :)

kable in the knitr Package

Install the package if necessary. `install.packages("knitr")`

```
kable(review_data[1:10,1:6]) # I use other parameters
to fit it into the slide set!
```

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species	Petals
5.1	3.5	1.4	0.2	setosa	3
4.9	3.0	1.4	0.2	setosa	12
4.7	3.2	1.3	0.2	setosa	3
4.6	3.1	1.5	0.2	setosa	4
5.0	3.6	1.4	0.2	setosa	4
5.4	3.9	1.7	0.4	setosa	4
4.6	3.4	1.4	0.3	setosa	6
5.0	3.4	1.5	0.2	setosa	3
4.4	2.9	1.4	0.2	setosa	5
4.9	3.1	1.5	0.1	setosa	3

Summary Statistics + Categorical Variables

Think back to Oral Exam 2!

The sentence in your final project looks something like (this is my copy):

> *Produce the following summary statistics about the SalePrice, BsmtUnfSF, and MoSold variables (and no other summary statistics) at every level of the Exterior2nd variable.*

What are the key parts of the sentence that inform what code we write?

Summary Statistics + Categorical Variables

Think back to Oral Exam 2!

The sentence in your final project looks something like (this is my copy):

> *Produce the following summary statistics about the SalePrice, BsmtUnfSF, and MoSold variables (and no other summary statistics) at every level of the Exterior2nd variable.*

1. The variable(s) for which we produce summary statistics.
2. "At every level of the __________ variable."

What `dplyr` functions do we think this uses? `group_by()` and `summarize()`!

Summary Statistics + Categorical Variables

We're going to do the following:

> *Produce the following summary statistics about the `Sepal.Width` and `Petals` variables (and no other summary statistics) at every level of the `SoilType` variable:*
>
> *Sample Mean, Sample 1st Quartile*

```
review_data %>% group_by(SoilType_______) %>%
summarize(meanSepalWidth =mean(S....)______, meanPetals =
_______, firstQuartSepalWidth = quantile(Sepal.Width, 0.25)_____,
firstQuartPetals = ______)
```

Summary Statistics + Categorical Variables

We're going to do the following:

> *Produce the following summary statistics about the `Sepal.Width` and `Petals` variables (and no other summary statistics) at every level of the `SoilType` variable:*
>
> *Sample Mean, Sample 1st Quartile*

```
review_data %>% group_by(SoilType) %>%
summarize(meanSepalWidth = mean(Sepal.Width),
meanPetals = mean(Petals), firstQuartSepalWidth =
quantile(Sepal.Width, 0.25), firstQuartPetals =
quantile(Petals, 0.25))
```

Plotting

Remember the `ggplot2` package!

We're going to do the following:

> *Produce a scatter plot with `Sepal.Width` on the Y-Axis, `Petal.Length` on the X-Axis, and color the points by `Species`.*

```
ggplot(data = review_data) +
geom_[somefunction](aes(x = ___, y = ____,
color=as.factor(____)))
```

Plotting

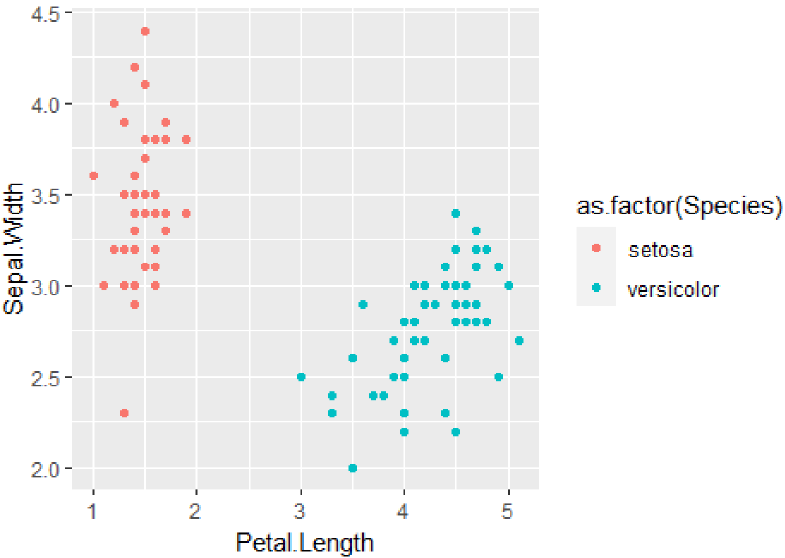
Remember the `ggplot2` package!

We're going to do the following:

> *Produce a scatter plot with `Sepal.Width` on the Y-Axis, `Petal.Length` on the X-Axis, and color the points by `Species`.*

```
ggplot(data = review_data) + geom_point(aes(x =
Petal.Length, y = Sepal.Width,
color=as.factor(Species)))
```

Plotting


4.5
4.0
3.5
3.0
2.5
2.0
Sepal.Width
1
2
3
4
5
Petal.Length
as.factor(Species)
setosa
versicolor

Plotting II

Now, let's make these plots for **each** level of the `MoreThan10` variable!

What function do we need?

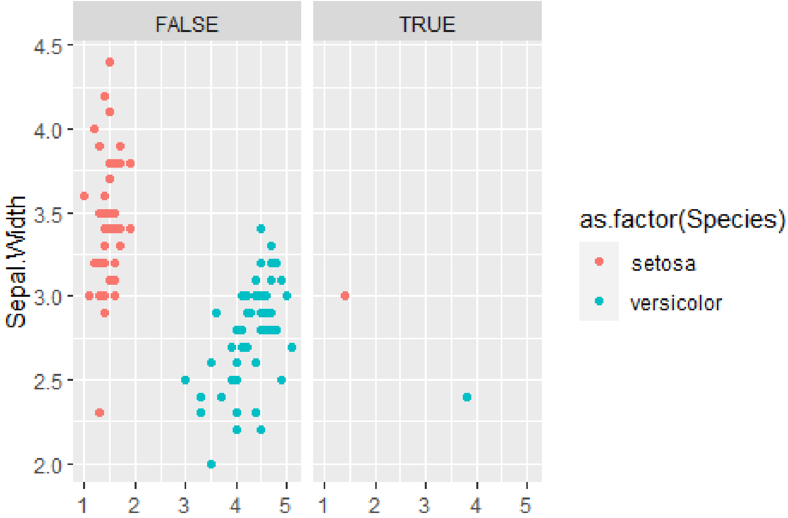
Plotting II

Now, let's make these plots for **each** level of the `MoreThan10` variable!

What function do we need? `facet_wrap()` or `facet_grid()`

Plotting II

```
ggplot(data = review_data) + geom_point(aes(x =
Petal.Length, y = Sepal.Width,
color=as.factor(Species))) + facet_wrap(~MoreThan10)
```



Creating Functions

What `keyword` in R defines functions?

Creating Functions

What `keyword` in R defines functions? `function`

We're going to do the following:

> *Write a function, passing in a dataframe and vector of columns as arguments, that returns the mean and standard deviation of all numeric columns.*

Key parts:

1. "Write a function" – obviously, implies use of `function`.
2. "Passing in a dataframe … as arguments" – Parameters!
3. "That returns the mean and st. dev. of numeric columns" – Body of function!

Creating Functions

This problem took me a good while to complete! (Get started on the project ASAP!)

Helpful tips on this question:

1. I break this up into three individual steps:

   - Create vectors, using a for loop, that split the parameter representing the columns on which we want summary statistics into a vector for *numerical* and a vector for *categorical* columns.
   - For each vector, subset the **original** data according to this vector independently, and either use `table()` (categorical) or `summarize()` (numerical). **Follow the syntax of the example below (adapted from the `?across` documentation):

```
iris %>%
  group_by(Species) %>%
  summarise(across(all_of(cols), list(mean = ~ mean(.),
  sd = ~ sd(.), ...")))
```

the ,..." should be removed -- oops! :)

Creating Functions

- **There is surely a better way to do this problem! Try things out!**

2. The parameter for the grouping variable is going to be a string. If you write `!!sym("Variable"_____)` and replace the blank with your variable name, you can use this as the argument to group_by to group by the variable!

Q/A

Ask me any questions you have! (The remaining time will be just normal OH. Thanks, and good luck!! :))