CHRISTOPHER PRENER, PH.D.

FALL, 2017

WEEK 15

LECTURE 16

QUANTITATIVE ANALYSIS

ANOVA

AGENDA

1. Front Matter
2. ANOVA Theory
3. One-way ANOVA in `R`
4. ANOVA Assumptions
5. Back Matter

1 FRONT MATTER

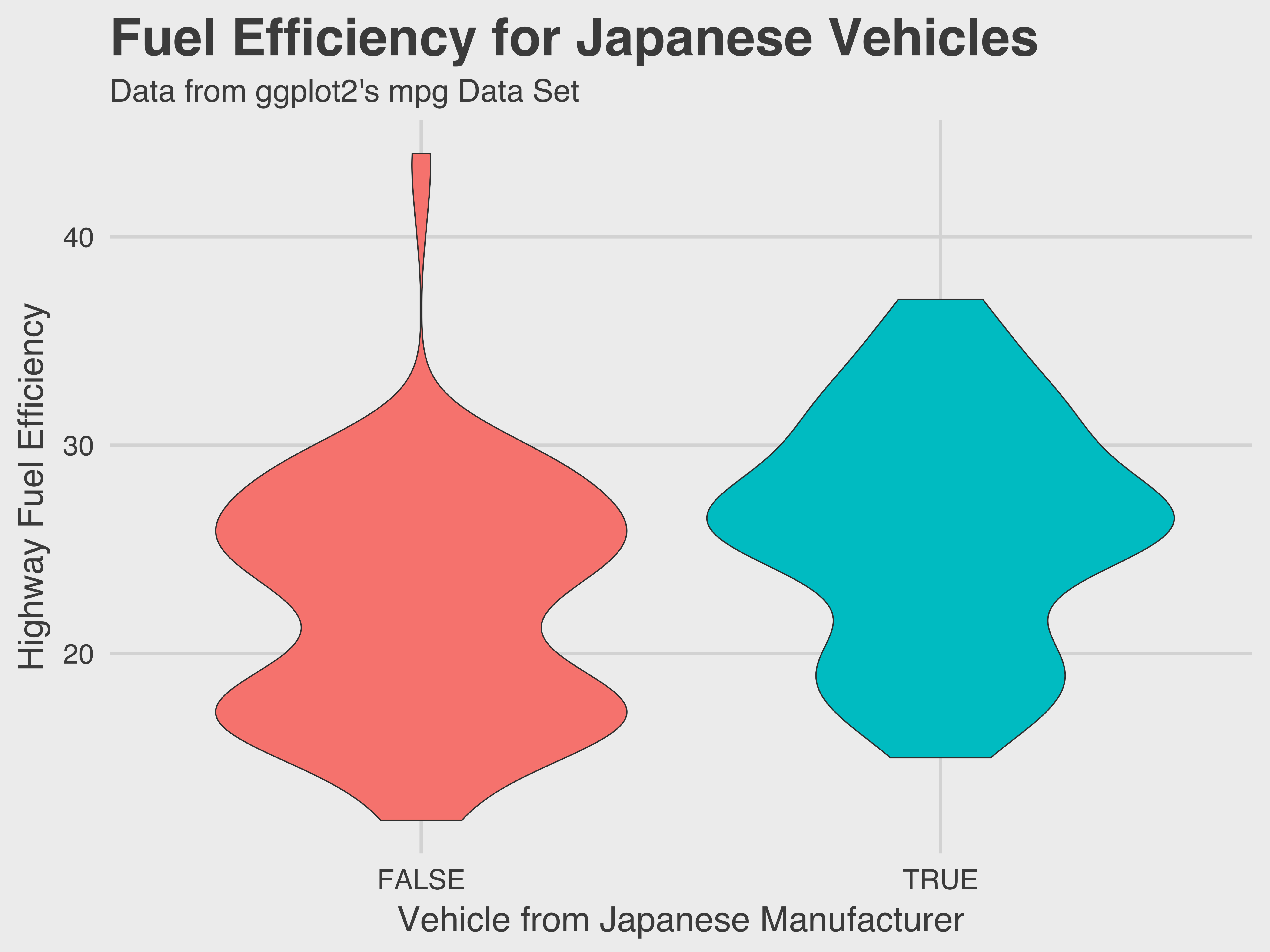
ANNOUNCEMENTS

WP-16, Lab-14, Lab-15, and PS-10 are due next Monday

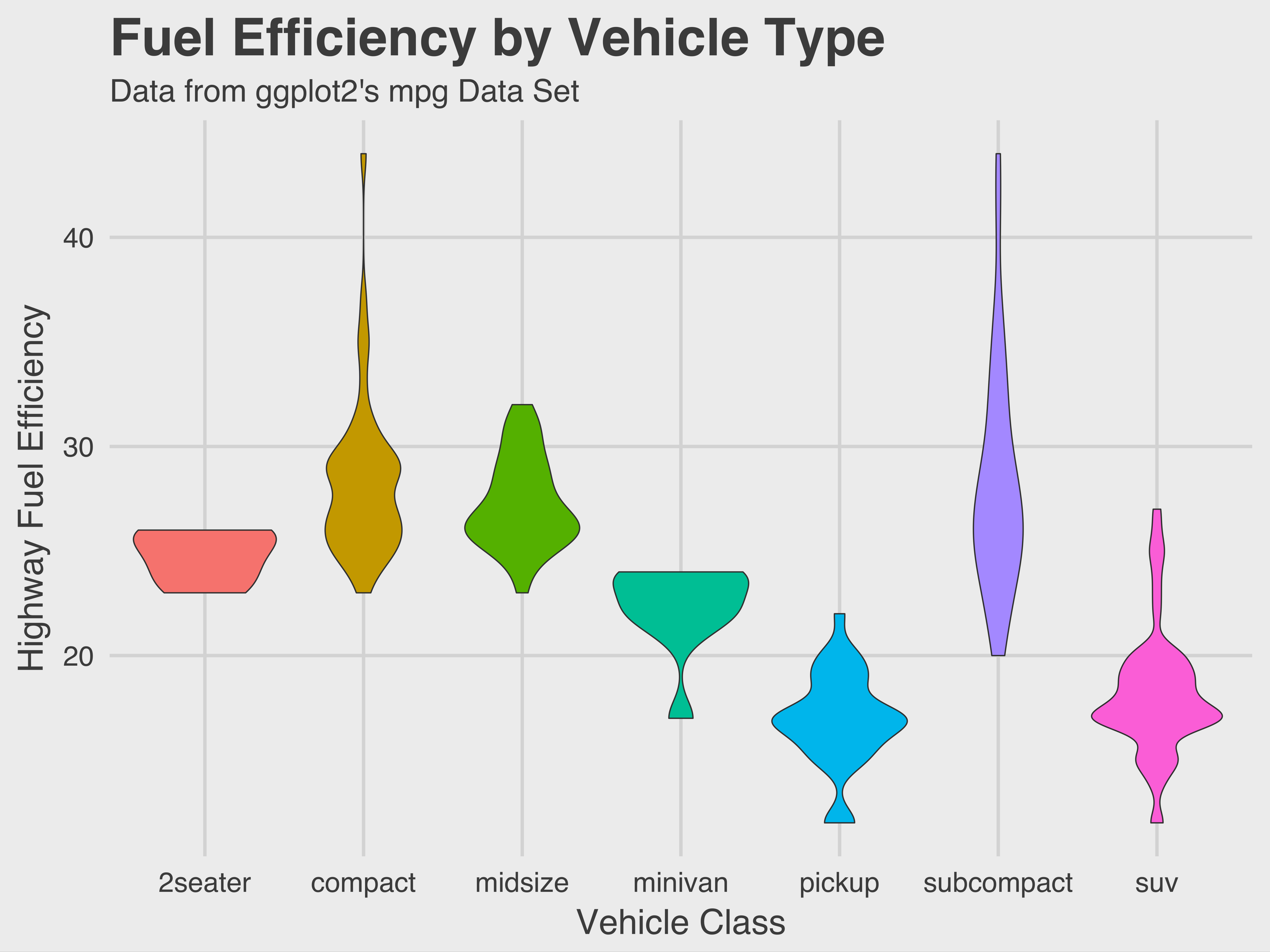
Final project presentations begin at *4pm* on Monday, December 18th in 2718 Morrissey. I will have a laptop there, or you can present using your own device.

Final project rubrics will be posted tomorrow. Please check them as you work on your projects!

2 ANOVA THEORY



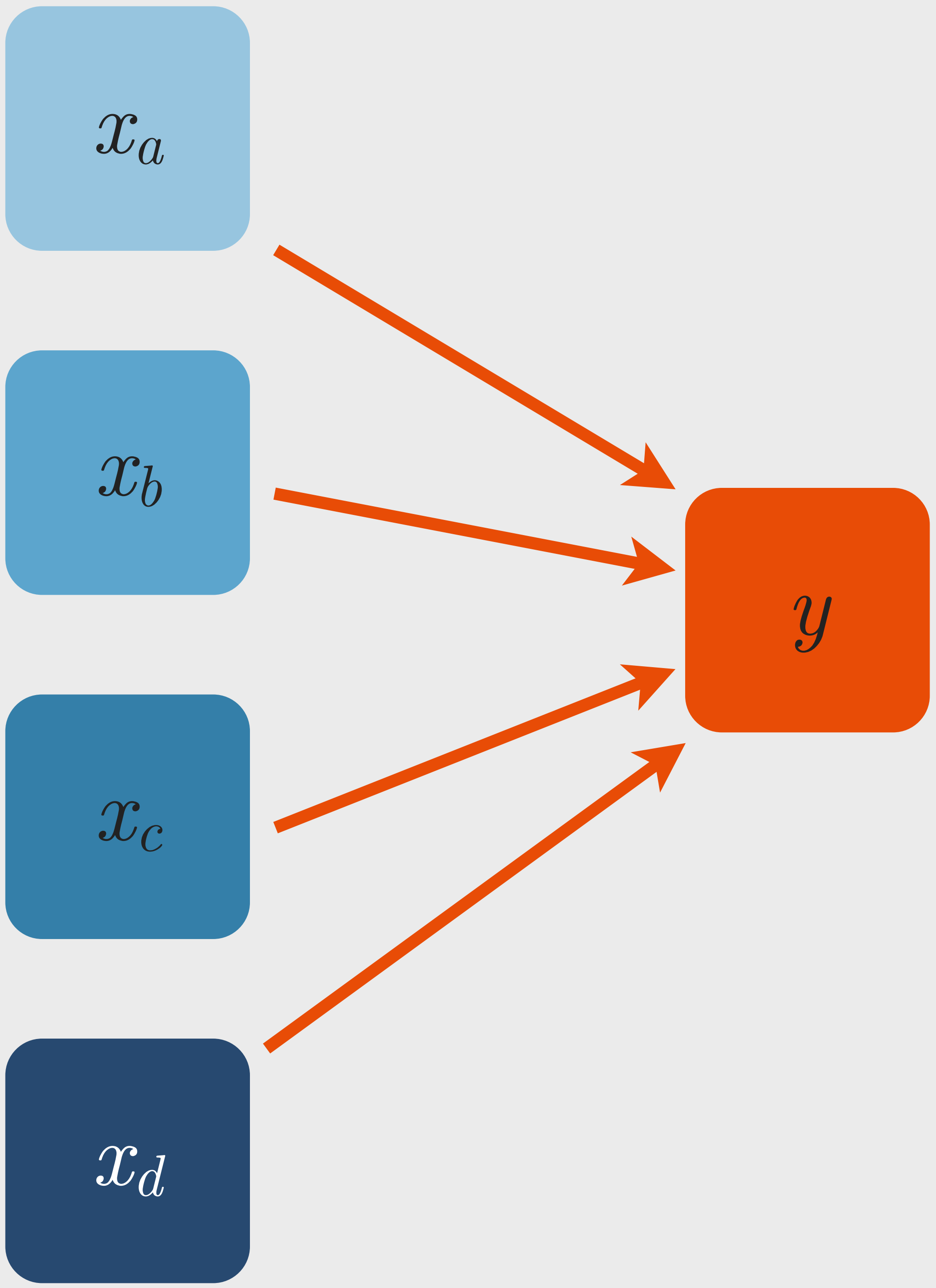
Fuel Efficiency for Japanese Vehicles
Data from ggplot2's mpg Data Set
Highway Fuel Efficiency
40
30
20
FALSE
TRUE
Vehicle from Japanese Manufacturer



Fuel Efficiency by Vehicle Type
Data from ggplot2's mpg Data Set
Highway Fuel Efficiency
40
30
20
2seater
compact
midsize
minivan
pickup
subcompact
suv
Vehicle Class

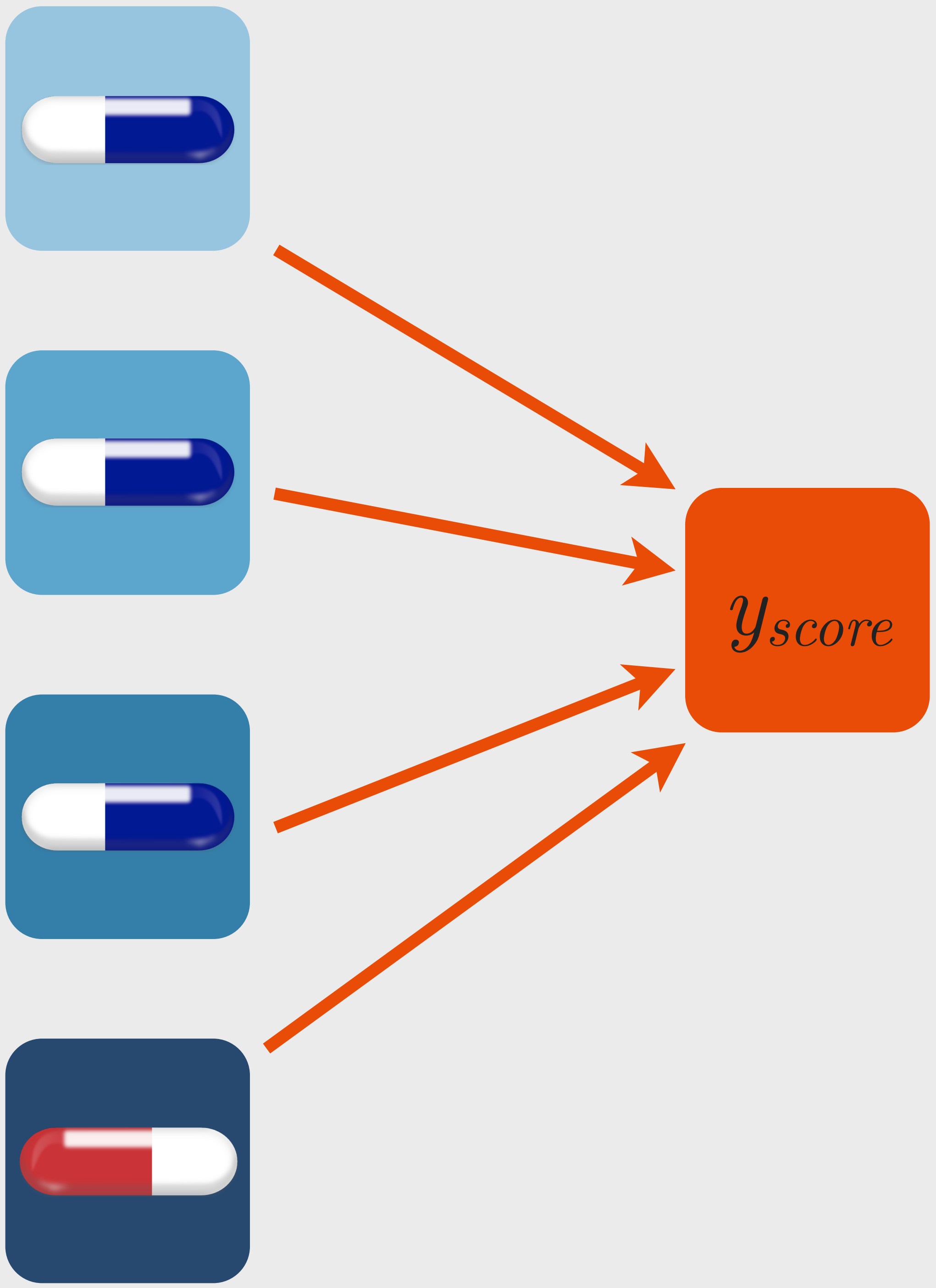
ANOVA

- Both ANOVA and regression are special cases of the generalized linear model
- ANOVAs are primarily used in experimental settings
- ANOVAs share some characteristics with t-tests in that mean comparisons are being made



ANOVA

- Both ANOVA and regression are special cases of the generalized linear model
- ANOVAs are primarily used in experimental settings
- ANOVAs share some characteristics with t-tests in that mean comparisons are being made



GROUPING VALUES

f(x)

`group_by(dataFrame, varName)`

Parameters:

- *data*
- *var* ... eted "by

Both functions in section available in `dplyr`
Download via CRAN alone or as part of `tidyverse`

GROUPING OBSERVATIONS

f(x)

```
group_by(dataFrame, varName)
```

Parameters:

- *dataFrame* is the data frame or tibble to be modified
- *varName* is the grouping variable that you want operations completed "by group"

GROUPING OBSERVATIONS

f(x)

`group_by(dataFrame, varName)`

Using the `class` variable from `ggplot2`'s `mpg` data:

```
> group_by(mpg, class)
```

Needs a second function to perform "grouped by" operations; can be used in a pipe with the *`dataFrame`* omitted

GROUPING OBSERVATIONS

f(x)

`group_by(dataFrame, varName)`

Using the `class` variable from `ggplot2`'s `mpg` data:

```
> group_by(mpg, class)
```

Data can also be un-grouped using `group_by()`'s compliment, `ungroup(dataFrame)`

SUMMARIZING OBSERVATIONS

f(x)

```
summarize(dataFrame, newVar = sumFun)
```

Parameters:

- `dataFrame` is the data frame or tibble to be modified *that has grouped data*
- `newVar` is the new variable to be created that stores the results of the operation performed
- `sumFun` is one of the available summary functions, including `first()`, `last()`, `nth()`, `n()`, `IQR()`, `min()`, `max()`, `median()`, `mean()`, `var()`, and `sd()`

SUMMARIZING OBSERVATIONS

f(x)

```
summarize(dataFrame, newVar = sumFun)
```

Using the `ggplot2`'s `mpg` data:

```
> summarize(mpg, count = n())
```

Will give you a count of the number of observations in `mpg`

SUMMARIZING OBSERVATIONS

f(x)

```
summarize(dataFrame, newVar = sumFun)
```

Using the `hwy` variable from `ggplot2`'s `mpg` data:

```
> summarize(mpg, meanHwy = mean(hwy))
```

Will give you the mean of the variable `hwy`, but it will not be grouped unless `group_by()` has already be used!

SUMMARIZING OBSERVATIONS

f(x) `summarize(dataFrame, newVar = sumFun)`

Using multiple arguments from `ggplot2`'s `mpg` data:

```
> summarize(mpg, count = n(), meanHwy = mean(hwy))
```

Will give you the mean of the variable `hwy`, but it will not be grouped unless `group_by()` has already be used!

SUMMARIZING OBSERVATIONS

```
> mpg %>%
      group_by(class) %>%
      summarise(count = n(), meanHwy = mean(hwy))
# A tibble: 7 x 3
        class count  meanHwy
        <chr> <int>    <dbl>
1     2seater     5 24.80000
2     compact    47 28.29787
3     midsize    41 27.29268
4     minivan    11 22.36364
5      pickup    33 16.87879
6 subcompact    35 28.14286
7         suv    62 18.12903
```

3 ONE-WAY ANOVA IN R

ANOVA

f(x) `aov(yvar ~ xvar, data = dataFrame)`

Parameters:

- *yva*
- *xva*
- *dataFrame* is a data frame or tibble

Both functions in section available in `stats`
Included in standard distributions of R

ANOVA

f(x)

```
aov(yvar ~ xvar, data = dataFrame)
```

Parameters:

- *yvar* is the dependent variable
- *xvar* is the factor-formatted independent variable
- *dataFrame* is a data frame or tibble

ANOVA

f(x)

```
aov(yvar ~ xvar, data = dataFrame)
```

Using the `hwy` and `class` variables from `ggplot2`'s `mpg` data:

```
> aov(hwy ~ class, data = mpg)
```

<<<<< OUTPUT OMITTED >>>>>

Save the model output to an object for reference later!

ANOVA

```
> model <- aov(hwy ~ class, data = mpg)

> summary(model)
             Df Sum Sq Mean Sq F value Pr(>F)
class         6   5683   947.2   83.39 <2e-16 ***
Residuals   227   2578    11.4
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

ANOVA

```
> model <- aov(hwy ~ class, data = mpg)

> summary(model)
             Df Sum Sq Mean Sq F value Pr(>F)
class         6   5683   947.2   83.39 <2e-16 ***
Residuals   227   2578    11.4
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

How would you interpret this result?

ANOVA

```
> model <- aov(hwy ~ class, data = mpg)

> summary(model)
             Df Sum Sq Mean Sq F value Pr(>F)
class         6   5683   947.2   83.39 <2e-16 ***
Residuals   227   2578    11.4
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The model's results ($f = 83.39,\ df = 6,\ p < .001$) suggest that there is meaningful variation between the mean highway fuel efficiency of vehicles from different classes.

TUKEY HONEST SIGNIFICANT DIFFERENCES

f(x) `TukeyHSD(model)`

Parameters:

- *model* is an ANOVA model object

TUKEY HONEST SIGNIFICANT DIFFERENCES

f(x) `TukeyHSD(model)`

Using the `model` object created from `ggplot2`'s `mpg` data:

```
> TukeyHSD(model)
```

<<<<< OUTPUT OMITTED >>>>>

Will calculate ever permutation of combinations and test them to see if the mean difference for each is statistically significant.

TUKEY HONEST SIGNIFICANT DIFFERENCES

```
> TukeyHSD(model)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = hwy ~ class, data = mpg)

$class
                         diff         lwr       upr     p adj
compact-2seater      3.4978723  -1.2185908  8.214335 0.2962191
midsize-2seater      2.4926829  -2.2568476  7.242213 0.7070356
minivan-2seater     -2.4363636  -7.8442474  2.971520 0.8321849
pickup-2seater      -7.9212121 -12.7329120 -3.109512 0.0000377
subcompact-2seater   3.3428571  -1.4507195  8.136434 0.3713580
```

<<<<< OUTPUT TRUNCATED >>>>>>

TUKEY HONEST SIGNIFICANT DIFFERENCES

```
> TukeyHSD(model)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = hwy ~ class, data = mpg)

$class
                          diff         lwr       upr     p adj
compact-2seater      3.4978723  -1.2185908  8.214335 0.2962191
midsize-2seater      2.4926829  -2.2568476  7.242213 0.7070356
minivan-2seater     -2.4363636  -7.8442474  2.971520 0.8321849
pickup-2seater      -7.9212121 -12.7329120 -3.109512 0.0000377
subcompact-2seater   3.3428571  -1.4507195  8.136434 0.3713580
```

? How would you interpret this result?

TUKEY HONEST SIGNIFICANT DIFFERENCES

```
> TukeyHSD(model)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = hwy ~ class, data = mpg)

$class
                          diff         lwr       upr     p adj
compact-2seater      3.4978723  -1.2185908  8.214335 0.2962191
midsize-2seater      2.4926829  -2.2568476  7.242213 0.7070356
minivan-2seater     -2.4363636  -7.8442474  2.971520 0.8321849
pickup-2seater      -7.9212121 -12.7329120 -3.109512 0.0000377
subcompact-2seater   3.3428571  -1.4507195  8.136434 0.3713580
```

Of the comparisons with “two-seater” sports cars, the only mean difference that was statistically significant based on the Tukey post-hoc test was the relationship with pickup trucks ($p < .001$).

4 ANOVA ASSUMPTIONS

ASSUMPTIONS

- y should be normally distributed
  - Use standard techniques to evaluate normality
- the categories within x should have equal (homogeneous) variance
- There should be no significant outliers
  - Use the Bonferonni test (`car::outlierTest()`) discussed in Week-14

HOMOGENEITY OF VARIANCE

f(x)

```
bartlett.test(yvar ~ xvar, data = dataFrame)
```

Parameters:

- *yvar*
- *xvar*
- *dataFrame* is a data frame or tibble

Available in `stats`

Included in standard distributions of `R`

HOMOGENEITY OF VARIANCE

f(x)

```
bartlett.test(yvar ~ xvar, data = dataFrame)
```

Parameters:

- *yvar* is the dependent variable
- *xvar* is the factor-formatted independent variable
- *dataFrame* is a data frame or tibble

HOMOGENEITY OF VARIANCE

f(x)

```
bartlett.test(yvar ~ xvar, data = dataFrame)
```

Using the `hwy` and `class` variables from `ggplot2`'s `mpg` data:

```
> bartlett.test(hwy ~ class, data = mpg)
```

<<<<< OUTPUT OMITTED >>>>>

!

The null and alternative hypotheses are the same as the Levene's test (see Week-07 and Week-08)

HOMOGENEITY OF VARIANCE

```
> bartlett.test(hwy ~ class, data = mpg)

    Bartlett test of homogeneity of variances

data:  hwy by class
Bartlett's K-squared = 50.523, df = 6, p-value = 3.692e-09
```

HOMOGENEITY OF VARIANCE

```
> bartlett.test(hwy ~ class, data = mpg)

	Bartlett test of homogeneity of variances

data:  hwy by class
Bartlett's K-squared = 50.523, df = 6, p-value = 3.692e-09
```

How would you interpret this result?

HOMOGENEITY OF VARIANCE

```
> bartlett.test(hwy ~ class, data = mpg)

	Bartlett test of homogeneity of variances

data:  hwy by class
Bartlett's K-squared = 50.523, df = 6, p-value = 3.692e-09
```

The results of the Bartlett Test ($k^2 = 50.523,\ df = 6,\ p < .001$) indicate that these data do not meet the homogeneity of variance assumption for ANOVA.

5 BACK MATTER

WHAT WE COVERED TODAY

2. ANOVA Theory
3. One-way ANOVA in R
4. ANOVA Assumptions

REMINDERS

WP-16, Lab-14, Lab-15, and PS-10 are due next Monday

Final project presentations begin at *4pm* on Monday, December 18th in 1705 Morrissey

Final project rubrics will be posted tomorrow. Please check them as you work on your projects!