

Midterm review

Fall 2025 Midterm 1

<https://live.ds306.org>

2026-10-13

Today's review

- We'll go through each question from last year's first midterm.
- Try an answer before we work it out together.
- Explain your reasoning, even if you don't remember the exact code.

Problem 1, Question 1: Install a package

Explain how and when to install a package.

```
# Write the command you would use in RStudio.
```

Problem 1, Question 2: Load a package

Explain how and when to load a package.

```
# Load a package.
```

Problem 1, Question 3: Get help

How do you access R's built-in help for a function or dataset?

```
# Get help for mean().
```

Problem 1, Question 4: Vectors

What is a vector?

```
# Create an example vector.
```

Problem 1, Question 5: Data frames

What is a data frame?

```
# Create a small example data frame.
```

Problem 1, Question 6: Variable types

Explain the difference between a categorical and a numerical variable.

Problem 1, Question 7: Two categorical variables

For two categorical variables, describe an appropriate plot.

The exam gives **heat map** as its worked example.

Problem 1, Question 8: One categorical variable

For one categorical variable, describe an appropriate plot to visualize its distribution.

Problem 1, Question 9: One numerical variable

For one numerical variable, describe an appropriate plot to visualize its distribution.

Problem 1, Question 10: Two numerical variables

For two numerical variables, describe an appropriate plot to visualize their joint distribution.

Problem 1, Question 11: Numerical and categorical

For one numerical variable and one categorical variable, describe an appropriate plot.

Problem 2, Question 12: Variable types

For each variable in `animals`, identify whether it is categorical or numerical.

```
# A tibble: 6 x 3
  species has_feathers weight
  <chr>    <lgl>         <dbl>
1 dog     FALSE         20.5
2 cat     FALSE         11.6
3 cat     FALSE           8
4 cat     FALSE           7
5 chicken TRUE          NA
6 chicken TRUE          2.5
```

Problem 2, Question 13: Rows and columns

How many rows and columns are in `animals`?

```
# A tibble: 6 x 3
  species has_feathers weight
  <chr>    <lgl>         <dbl>
1 dog     FALSE         20.5
2 cat     FALSE         11.6
3 cat     FALSE           8
4 cat     FALSE           7
5 chicken TRUE          NA
6 chicken TRUE          2.5
```

Problem 2, Question 14: Keep the cats

```
# A tibble: 6 x 3
  species has_feathers weight
<chr>    <lgl>         <dbl>
1 dog     FALSE        20.5
2 cat     FALSE        11.6
3 cat     FALSE         8
4 cat     FALSE         7
5 chicken TRUE         NA
6 chicken TRUE         2.5
```

Write a command that keeps only the rows corresponding to cats.

```
# Keep the cats.
animals
```

Problem 2, Question 15: Extract a value

```
# A tibble: 6 x 3
  species has_feathers weight
<chr>    <lgl>         <dbl>
1 dog     FALSE        20.5
2 cat     FALSE        11.6
3 cat     FALSE         8
4 cat     FALSE         7
5 chicken TRUE         NA
6 chicken TRUE         2.5
```

What does this command return? Predict before running it.

```
animals$species[5]
```

Problem 2, Question 16: Mean of logical values

A tibble: 6 x 3

	species	has_feathers	weight
	<chr>	<lgl>	<dbl>
1	dog	FALSE	20.5
2	cat	FALSE	11.6
3	cat	FALSE	8
4	cat	FALSE	7
5	chicken	TRUE	NA
6	chicken	TRUE	2.5

What does this command return? Predict before running it.

```
mean(animals$has_feathers)
```

Problem 2, Question 17: Distinct species

```
# A tibble: 6 x 3
  species has_feathers weight
  <chr>    <lgl>         <dbl>
1 dog     FALSE         20.5
2 cat     FALSE         11.6
3 cat     FALSE           8
4 cat     FALSE           7
5 chicken TRUE          NA
6 chicken TRUE          2.5
```

What does this pipeline return?

```
animals |>
  distinct(species) |>
  nrow()
```

Problem 2, Question 18: Grouped medians

```
# A tibble: 6 x 3
  species has_feathers weight
<chr>    <lgl>         <dbl>
1 dog     FALSE        20.5
2 cat     FALSE        11.6
3 cat     FALSE         8
4 cat     FALSE         7
5 chicken TRUE         NA
6 chicken TRUE         2.5
```

Draw the table returned by this pipeline.

```
animals |>
  group_by(species) |>
  summarize(median(weight))
```

Problem 2, Question 19: Count a new variable

```
# A tibble: 6 x 3
  species has_feathers weight
  <chr>    <lgl>         <dbl>
1 dog     FALSE         20.5
2 cat     FALSE         11.6
3 cat     FALSE           8
4 cat     FALSE           7
5 chicken TRUE          NA
6 chicken TRUE          2.5
```

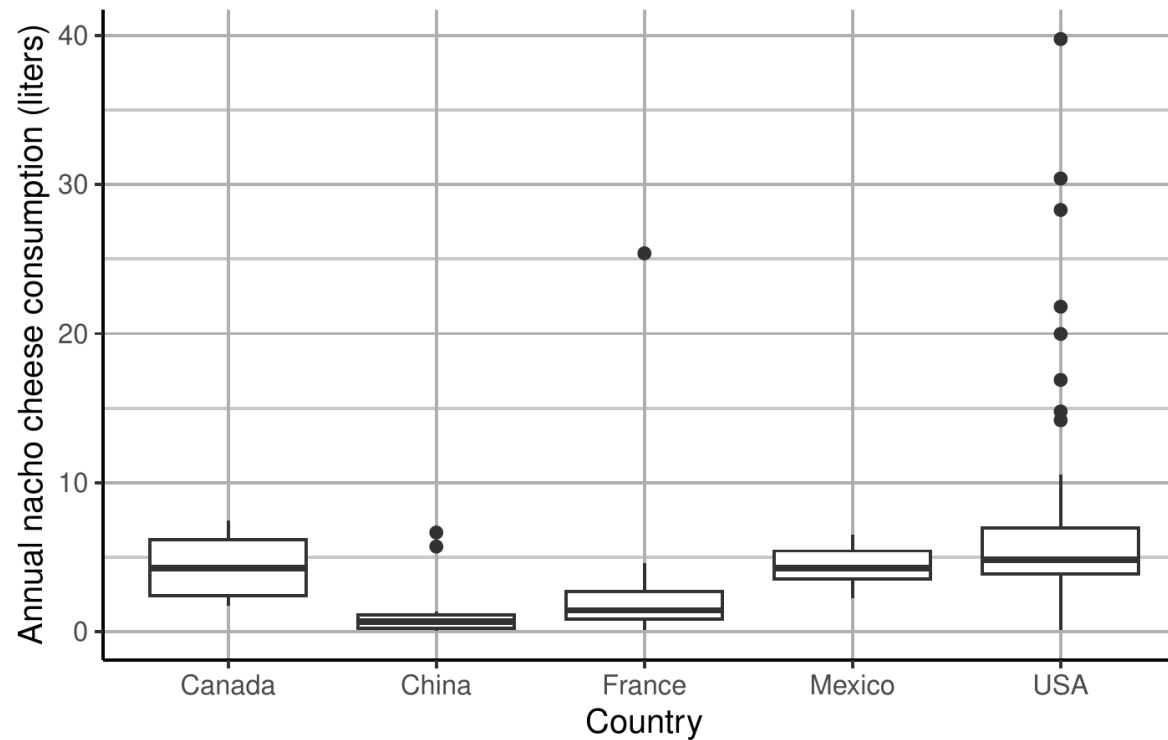
Draw the table returned by this pipeline.

```
animals |>
  mutate(is_fat = weight > 10) |>
  count(is_fat)
```

Problem 3, Question 20: Name the plot

A survey measured annual per-capita nacho cheese consumption in five countries.

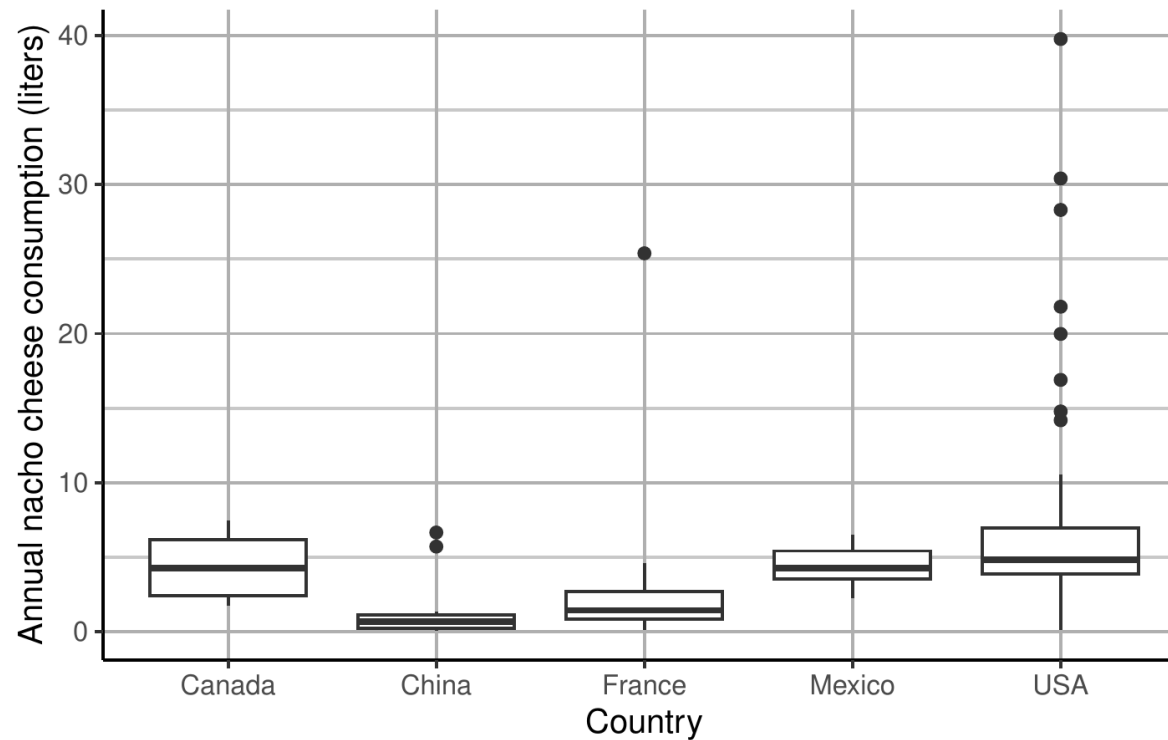
What is this type of plot called?



Annual per-capita nacho cheese consumption in five countries

Problem 3, Question 21: Read a quartile

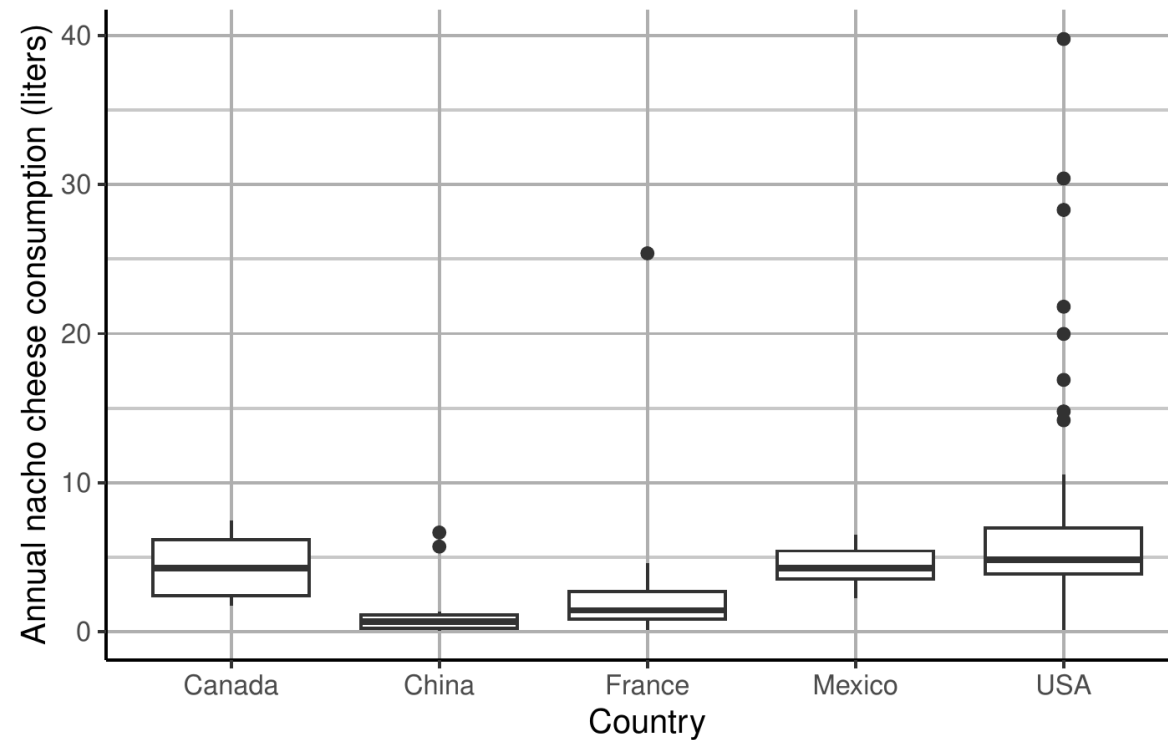
True or false: a quarter of the people surveyed in the USA consume less than 5 liters of nacho cheese per year.



Annual per-capita nacho cheese consumption in five countries

Problem 3, Question 22: Describe Canada

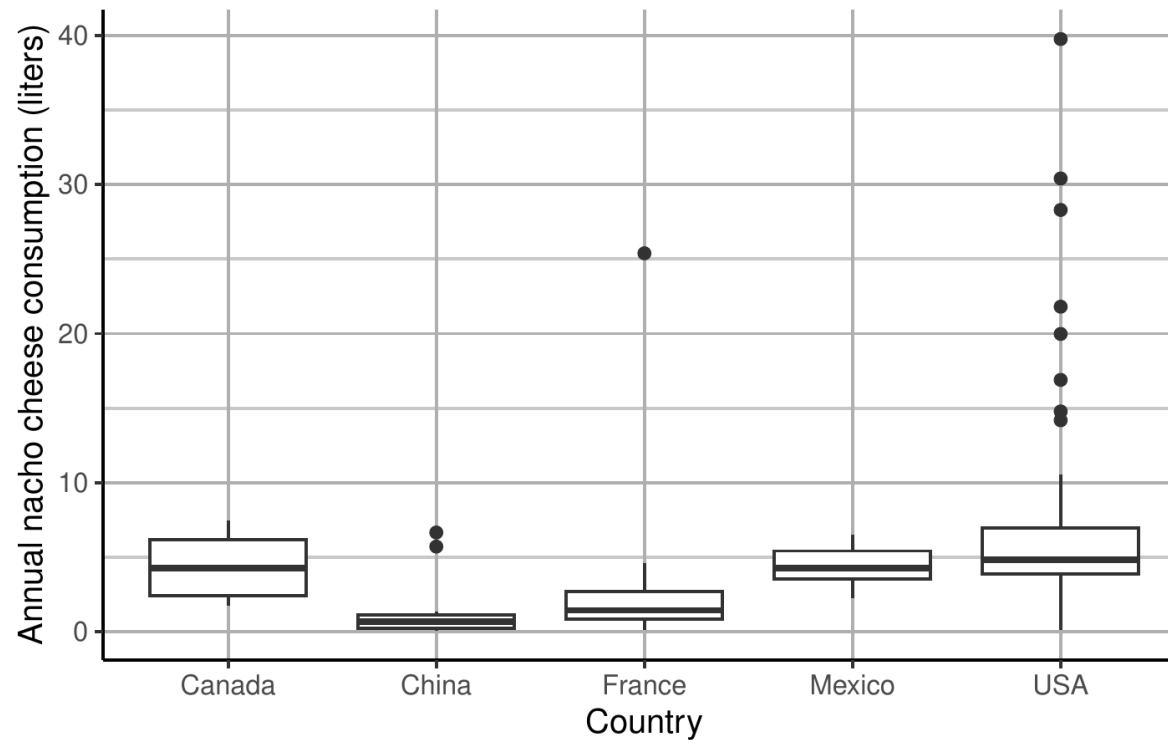
Explain everything the plot tells us about consumption in the **Canadian portion** of the sample.



Annual per-capita nacho cheese consumption in five countries

Problem 3, Question 23: Population means

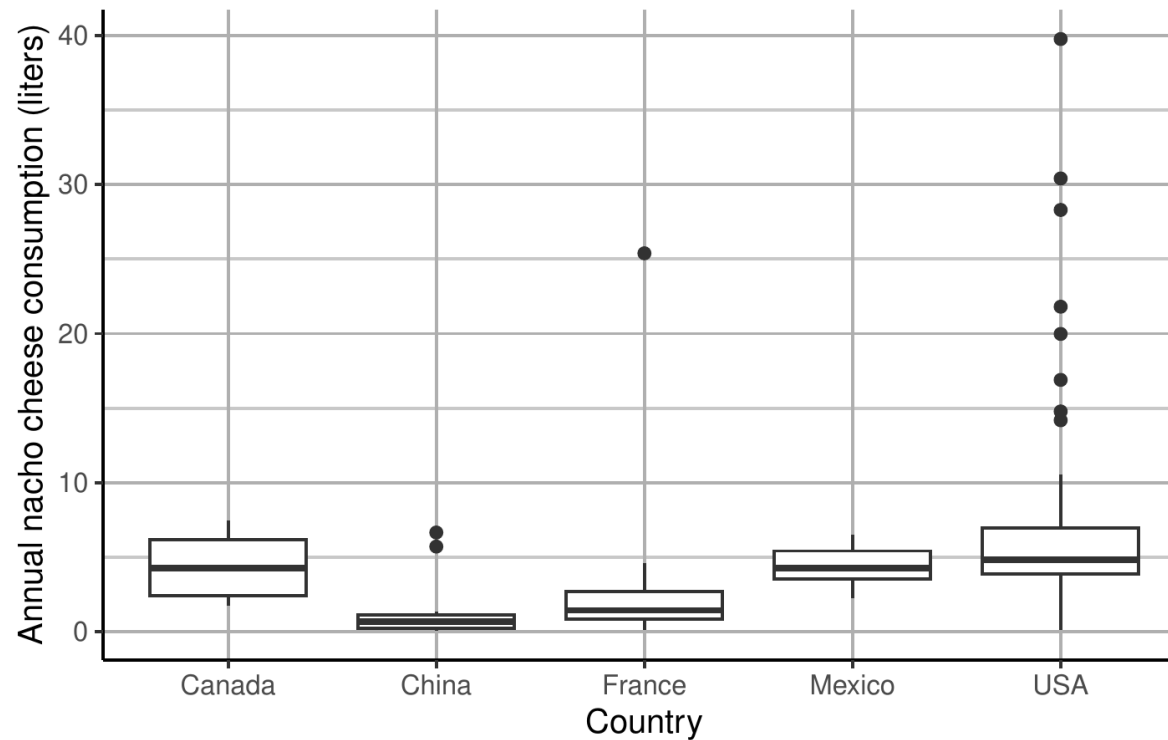
Can you conclude that the **population** of the USA consumes more nacho cheese on average than the population of France? Explain.



Annual per-capita nacho cheese consumption in five countries

Problem 3, Question 24: An upper bound?

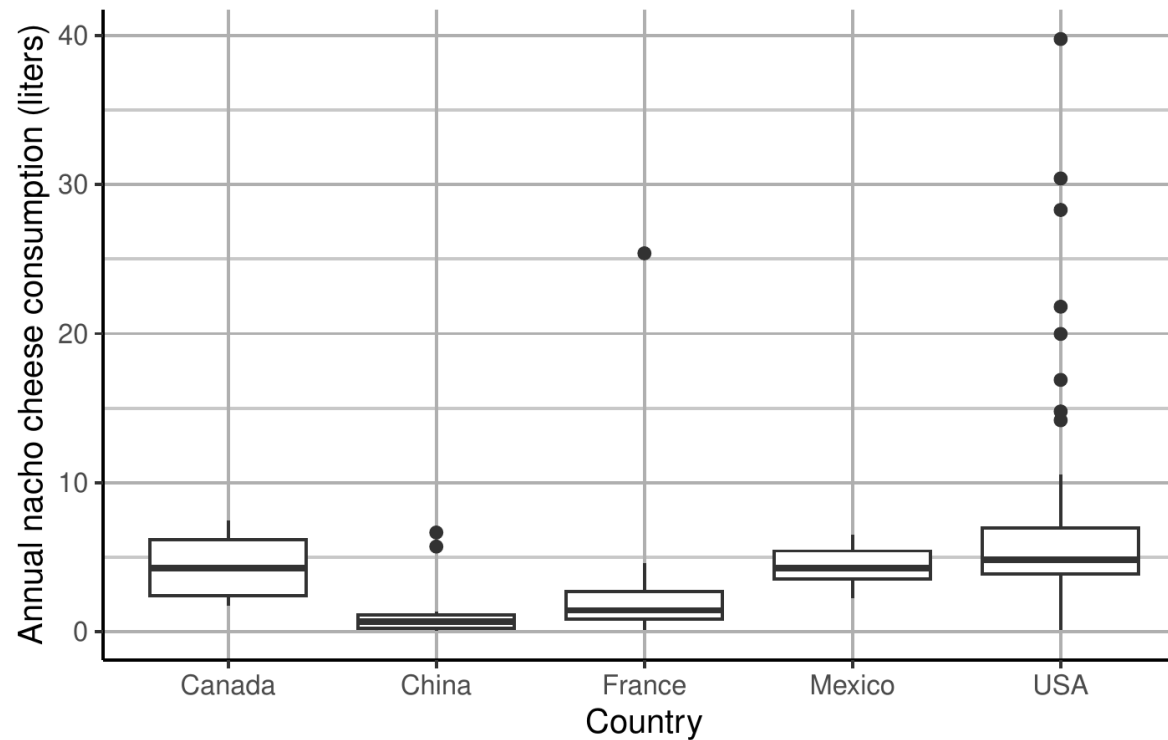
Can you conclude that average consumption in the USA portion of the sample is at most 5 liters per year? Explain.



Annual per-capita nacho cheese consumption in five countries

Problem 3, Question 25: A lower bound?

Can you conclude that average consumption in the USA portion of the sample is at least 2 liters per year? Explain.



Annual per-capita nacho cheese consumption in five countries

Problem 4, Question 26: Response

The `heights` table contains self-reported student height in inches and sex.

What is the response variable in this model?

Call:

```
lm(formula = height ~ sex, data = heights)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	64.9394	0.2363	274.82	<2e-16
sexMale	4.3753	0.2687	16.28	<2e-16

Residual standard error: 3.645 on 1048 df

Multiple R-squared: 0.2019

Problem 4, Question 27: Predictor

What is the predictor variable in this model?

Call:

```
lm(formula = height ~ sex, data = heights)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	64.9394	0.2363	274.82	<2e-16
sexMale	4.3753	0.2687	16.28	<2e-16

Residual standard error: 3.645 on 1048 df

Multiple R-squared: 0.2019

Problem 4, Question 28: Fitted equation

Call:

```
lm(formula = height ~ sex, data = heights)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	64.9394	0.2363	274.82	<2e-16
sexMale	4.3753	0.2687	16.28	<2e-16

Residual standard error: 3.645 on 1048 df

Multiple R-squared: 0.2019

Write the equation of the fitted regression line.

```
# Record the fitted equation or calculate its coefficients
```

Problem 4, Question 29: Sample size

Call:

```
lm(formula = height ~ sex, data = heights)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	64.9394	0.2363	274.82	<2e-16
sexMale	4.3753	0.2687	16.28	<2e-16

Residual standard error: 3.645 on 1048 df

Multiple R-squared: 0.2019

How many observations were used to fit this model?

```
# Determine the number of observations.
```

Problem 4, Question 30: Predict height

Call:

```
lm(formula = height ~ sex, data = heights)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	64.9394	0.2363	274.82	<2e-16
sexMale	4.3753	0.2687	16.28	<2e-16

Residual standard error: 3.645 on 1048 df

Multiple R-squared: 0.2019

A student is male and 170 cm tall. What height does the model predict for this student? Round to the nearest inch.

```
# Calculate the predicted height.
```

Problem 4, Question 31: Test equal means

Call:

```
lm(formula = height ~ sex, data = heights)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	64.9394	0.2363	274.82	<2e-16
sexMale	4.3753	0.2687	16.28	<2e-16

Residual standard error: 3.645 on 1048 df

Multiple R-squared: 0.2019

State a test of whether average male and female heights are equal. Explain how to use this output and state your conclusion.

```
# Use height_model or fit a test.
```

Problem 4, Question 32: Correlation

Call:

```
lm(formula = height ~ sex, data = heights)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	64.9394	0.2363	274.82	<2e-16
sexMale	4.3753	0.2687	16.28	<2e-16

Residual standard error: 3.645 on 1048 df

Multiple R-squared: 0.2019

What is the correlation between height and an indicator that equals 1 when the student is female?

```
# Calculate the correlation.
```

Problem 5, Question 33: Average dog weight

```
# A tibble: 6 x 3
  species has_feathers weight
  <chr>    <lgl>         <dbl>
1 dog     FALSE        20.5
2 cat     FALSE        11.6
3 cat     FALSE         8
4 cat     FALSE         7
5 chicken TRUE         NA
6 chicken TRUE         2.5
```

This code is intended to compute average dog weight. Explain the bug and fix it.

```
animals |>
  select(species == "dog") |>
  summarize(mean(weight))
```

Problem 5, Question 34: Count each species

```
# A tibble: 6 x 3
  species has_feathers weight
  <chr>    <lgl>         <dbl>
1 dog     FALSE         20.5
2 cat     FALSE         11.6
3 cat     FALSE           8
4 cat     FALSE           7
5 chicken TRUE          NA
6 chicken TRUE          2.5
```

This code is intended to count each species.
Explain the bug and fix it.

```
animals |>
  summarize(n = sum(species))
```

Problem 5, Question 35: Sort the animals

```
# A tibble: 6 x 3
  species has_feathers weight
  <chr>    <lgl>         <dbl>
1 dog     FALSE         20.5
2 cat     FALSE         11.6
3 cat     FALSE           8
4 cat     FALSE           7
5 chicken TRUE          NA
6 chicken TRUE          2.5
```

Sort alphabetically by species, breaking ties by putting the heaviest animals first. Explain the bug and fix it.

```
animals |>
  sort(-weight, species)
```

Problem 5, Question 36: Average height

The `heights` table has columns `sex` and `height`.

Height is measured in inches.

This code is intended to compute average student height. Explain the bug and fix it.

```
heights |> mean()
```

Problem 5, Question 37: Monthly temperature

The weather table has one row per date.

- date is the observation date.
- tmax is daily maximum temperature in °C.
- Some temperatures are missing.

Compute average daily maximum temperature for each month. Explain the bugs and fix them.

```
a2weather |>
  group_by(month(date)) |>
  summarize(avg_max_temp = mean(max(temp)))
```

Problem 6: the starting table

Daily Ann Arbor weather measurements, as used in last year's exam.

```
# A tibble: 48,636 x 12
  date      station  tmin  tmax  prcp  snow  snwd
<date>    <chr>    <dbl> <dbl> <dbl> <dbl> <dbl>
1 1891-10-01 USC00200230   7.8  20.6  NA    NA    NA
2 1891-10-02 USC00200230  13.9  26.7  NA    NA    NA
3 1891-10-03 USC00200230  16.1  26.1  NA    NA    NA
4 1891-10-04 USC00200230  11.1  22.8   7.6  NA    NA
5 1891-10-05 USC00200230   6.7  13.9  NA    NA    NA
6 1891-10-06 USC00200230    5   14.4  NA    NA    NA
# 5 more variables: tmax_flag, tmin_flag, prcp_flag,
# snow_flag, snwd_flag
```

Problem 6: the table we want

Average daily maximum temperature by year and month, 2010–2020, in °C.

year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2010	-0.9	0.5	11.5	19.4	22.3	26.9	29.9	29.0	23.4	19.0	10.5	-0.8
2011	-2.5	0.7	6.3	14.0	21.6	27.3	32.3	27.8	22.1	16.9	11.6	4.6
2012	2.9	4.1	16.3	16.0	25.2	28.8	32.9	28.7	24.1	16.2	9.0	4.9
2013	2.4	0.3	4.9	13.6	24.0	25.6	27.7	26.8	24.0	16.9	7.2	0.3
2014	-4.1	-2.7	3.7	15.9	22.0	26.5	26.1	27.1	23.3	16.2	5.6	3.9
2015	-2.1	-4.5	7.1	15.7	24.1	25.3	27.6	26.8	26.4	17.5	12.6	7.6
2016	0.9	4.0	11.7	14.5	21.8	28.2	30.3	29.7	25.4	19.0	13.1	1.0
2017	2.3	8.4	7.4	18.1	20.7	27.9	28.9	26.8	25.4	19.2	8.6	0.1
2018	0.1	3.7	5.9	11.4	25.4	26.7	29.6	28.6	24.9	15.7	4.3	3.6
2019	-0.4	1.8	6.2	15.6	21.1	25.3	30.1	27.4	24.8	16.9	5.4	4.8
2020	2.7	2.4	9.7	13.7	20.1	28.1	30.5	28.4	23.4	15.5	12.7	3.6

Problem 6, Question 38: Build the monthly table

We need:

- Years **2010** through **2020**.
- One row per year.
- One column per month, Jan–Dec.
- Average daily maximum temperature in each cell.

Describe the steps to make the table on the previous slide.

Write code or a plan precise enough for an experienced R user to implement.

```
# Build the table one step at a time.  
a2weather
```

Before the exam

What would you like to go over again?

Week 7, Lecture 13: Midterm review

Closed October 16, 2026 at 12:22 PM

0 class questions • 0 anonymous responses • 0 student questions

The following slides preserve what students submitted during class. No names or login information are included.