Practice Prelim 2

AEM 2850 / AEM 5850 – Fall 2025

Answer Key

READ THESE NOTES FIRST:

- Prelim 2 will cover all content we covered in weeks 7 through 14
  - Prelim 2 will also rely on prerequisite knowledge of fundamental concepts we learned in the first part of the course, but the questions are not designed to test that knowledge directly
- These practice questions are intended as a study resource, not a comprehensive guide
- These practice questions are not exhaustive in terms of topics and question types
- These practice questions are not necessarily representative of the weight that different topics and question types will receive on Prelim 2

Preface

The goal of this prelim is to assess your understanding of data visualization concepts and facility with key visualization and programming tools covered in weeks 7 through 14.

Instructions

- You must complete Prelim 2 in person
- Prelim 2 is a closed-book paper prelim
- Manage your time carefully
- If you get stuck, move on and come back later as time allows

Additional notes

- There are 17 questions worth a total of 100 points. The total number of points per question is stated with each question
- We will give partial credit if your answers are incomplete, especially if you outline the logic of what you *would* do if you had more time

Multiple Choice: circle only one answer per question

1. [2 points] Which ggplot layer draws the data?

a. Aesthetics
b. Geometries
c. Scales
d. Theme

Solution

b. Geometries

2. [2 points] Which chart best compares proportions across different groups?

a. Line plot
b. Pie chart
c. Stacked bar
d. Side-by-side bar

Solution

c. Stacked bar

3. [2 points] Which statement about dual y-axes is most accurate?

a. They are never allowed
b. They are great for scatter plots
c. They are sometimes misleading
d. They are best for visualizing regressions

Solution

c. They are sometimes misleading

4. [2 points] What type of map might mislead by emphasizing land area over population?

a. Histogram
b. Choropleth
c. Maps with points
d. Cartogram

Solution

b. Choropleth

5. [2 points] Which of the following is not a common method for visualizing distributions?

a. Bar charts
b. Histograms
c. Box and whisker plots
d. Density plots

Solution

a. Bar charts

Multiple Choices: circle any number of answers per question

6. [8 points] Circle *any and all* of the following code snippets that will return the plot described below *without errors*.

We want to create a faceted scatterplot of highway mileage (`hwy`) versus engine displacement (`displ`) using the `mpg` dataset, with the following features:

- Points colored by drive type (`drv`)
- Facets split by drive type (`drv`)
- A linear regression line added to each facet

a.

```r
plot <- mpg |>
  ggplot(aes(x = displ, y = hwy)) +
  facet_wrap(vars(drv)) +
  geom_point(aes(color = drv)) +
  geom_smooth(method = "lm")
```

b.

```r
mpg |>
  ggplot(aes(x = displ, y = hwy, color = drv)) +
  geom_point() +
  geom_smooth(method = "lm") +
  facet_wrap(vars(drv))
```

c.

```r
facet_wrap(vars(drv)) +
  mpg |>
  ggplot(aes(x = displ, y = hwy, color = drv)) +
  geom_point() +
  geom_smooth(method = "lm")
```

d.

```r
mpg |>
  ggplot(aes(x = displ, y = hwy), color = drv) +
  geom_point() +
  geom_smooth(method = "lm") +
  facet_wrap(vars(drv))
```

Solution: b

Short Answer

7. [4 points] Explain, at a high level, how the `sf` package enables you to work with spatial data along with non-spatial attributes.

Solution

Correct answers should state that the sf package stores spatial data in a special geometry column of an sf data frame along with non-spatial attributes (which are in standard columns just like other data frames).

8. [6 points] What are two advantages of writing your own custom function to make the same type of plot many times for different inputs, as compared to copying and pasting the code to produce each plot?

Solution

Full credit given for any two of:

1. Avoid duplication (of lines of code)
2. Avoid mistakes (e.g., due to incorrect hard-coded values)
3. Easy to scale
4. Saves time/s (e.g., easier to troubleshoot, faster to call functions)

9. [6 points] Name one advantage and one disadvantage of tokenization and visualizing the frequencies of different tokens as a method of text analysis.

Solution

Advantage: identifies common tokens (e.g., words) without assuming or imposing any structure on a text. Helps discover patterns in unstructured text that are not evident due to its lack of structure. etc.

Disadvantage: raw token frequency is not always informative (case in point: stop words)

10. [6 points] We want to produce a map with data on median income by US county. We have data on median incomes by county in a csv file, and data on county borders in a shapefile. You can take for granted that there are no errors in the filenames and variable names, and that there are no missing values in either data source. Here is the code we wrote to produce the map:

```r
library(tidyverse)
library(sf)

# Import the data
county_incomes <- read_csv("county_income.csv")
county_borders <- read_sf("us_counties.shp")

# Merge the data
merged_data <- county_incomes |>
  inner_join(county_borders, join_by(county_code))

# Create map
merged_data |>
  ggplot(aes(fill = median_income)) +
  geom_sf()
```

Will we succeed? If yes, what will the output look like? If no, why not, and how can you fix it?

Solution

No, we will not succeed, because `merged_data` does not inherit the `sf` class from `county_borders`. As a result, `geom_sf()` will fail. To make it work, `county_borders` needs to be the first argument (`x`) given to `inner_join()`.

11. [7 points] Consider the function `get_type()` that is designed to return a message to the user based on whether a number is positive, negative, or neither.

```
get_type <- function(x) {
  if(x>0) {
    str_glue("{x} is positive.")
  } else if (x<0) {
    str_glue("{x} is negative.")
  } else if (is.na(x)) {
    str_glue("{x} is a missing value.")
  } else {
    str_glue("{x} is neither positive nor negative.")
  }
}
```

Separately for each part below, write the message that the function call will return. If the function call will not return a message, explain why.

a. `get_type(1)` **Solution:** 1 is positive.

b. `get_type(-1)` **Solution:** -1 is negative.

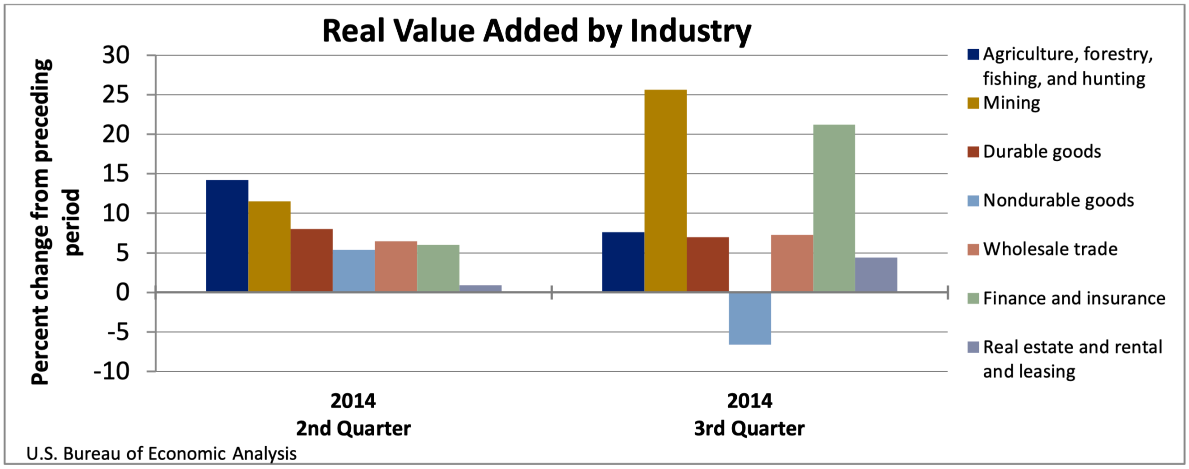
c. `get_type(NA)` **Solution:** Error, because `NA > 0` evaluates to `NA`, and `if()` requires either `TRUE` or `FALSE`. *Note: if the condition `if(is.na(x))` came first within the function, the function would return "NA is a missing value" without an error, since `is.na(NA)` evaluates to `TRUE`.*

d. `get_type(0)` **Solution:** 0 is neither positive nor negative.

e. `get_type(TRUE)` **Solution:** TRUE is positive.

12. [8 points] The plot below comes from a press release from the U.S. Bureau of Economic Analysis about changes in Real Value Added by Industry over time. Among other things, the press release stated:

- *Finance and insurance real value added increased 21.2 percent in the third quarter, after increasing 6.0 percent in the second quarter.*
- *Mining increased 25.6 percent, after increasing 11.5 percent.*
- *Real estate and rental and leasing increased 4.4 percent, after increasing 0.9 percent.*



U.S. Bureau of Economic Analysis

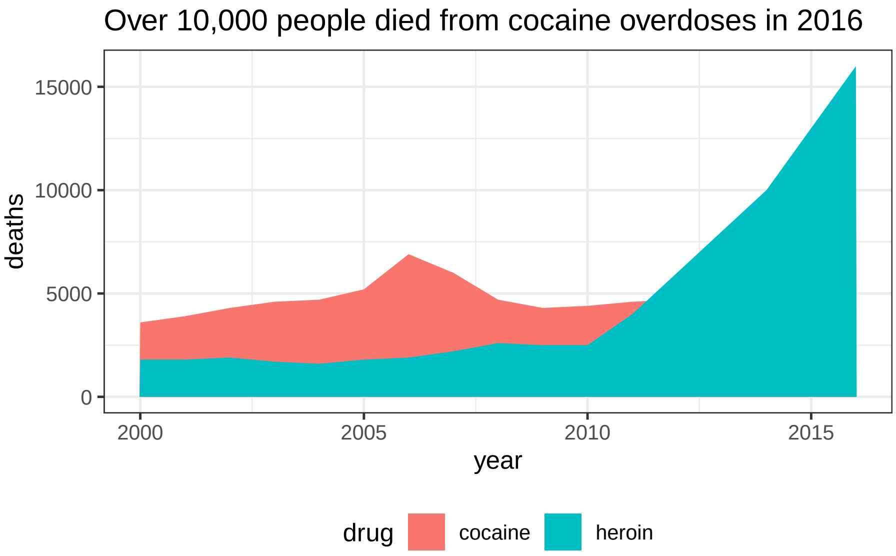
Describe two changes you could make to this visualization to more clearly and effectively illustrate the points made in the text above, and explain what issue each change is meant to address.

Note: Do not worry about what "real value added" means. We are looking for targeted data visualization suggestions based on the text, but they do not depend on the specifics of the outcome plotted on the y axis.

Solution: Full credit given for any two of:

1. Use **paired side-by-side bars** that are **grouped by industry** rather than time period, enabling direct comparisons of each pair of data points that is discussed in the bullet points
2. **Facet by industry** to separate the plots out by industry are also an acceptable answer since they make it easier to compare over time within industry
3. Creating a **dot plot that shows the two data points on changes between time periods** along one axis and **industry labels** along another axis

13. [9 points] Here is a graph of drug overdoses over time in the U.S.:



Describe three *different* ways you could revise this plot to ensure the visualization supports the message in its title *without removing any data from the plot*.

Note: For this question, we are not looking for generic graphical improvements *– they have to be specific to addressing the inconsistency between the plot's contents and its title. Do not write code for this part – just describe the changes you would make.*

Solution:

1. Use lines instead of filled areas

2. Facet by drug / separate the current graph into two different (sub-)graphs, one for cocaine and one for heorin

3. Increase transparency of the areas

14. [6 points] You gained access to the data frame `drug_data` used for the plot from the previous question. It has the following structure:

```
drug_data |> head(2)
```

```
# A tibble: 2 x 3
   year drug    deaths
  <int> <chr>    <dbl>
1  2000 cocaine   3600
2  2000 heroin    1800
```

Write a complete code snippet to implement one of your proposals from the previous question for how to ensure the visualization supports the title.

Note: Keep your code simple to avoid errors. Your code will be graded on whether it successfully addresses the inconsistency between the original plot's contents and its title. You will not be rewarded for making generic graphical improvements (e.g., axis labels, etc.).

Solution: Multiple answers would receive full credit:

```
# lines
drug_data |>
  ggplot(aes(x = year, y = deaths, color = drug)) +
  geom_line()

# facets
drug_data |>
  ggplot(aes(x = year, y = deaths, fill = drug)) +
  geom_area() +
  facet_wrap(vars(drug))

# transparency
drug_data |>
  ggplot(aes(x = year, y = deaths, fill = drug, alpha = 0.5)) +
  geom_area(position = "identity")
```

15. [15 points] Complete the table below to outline the three key steps and requisite functions to scrape data from a single table on a server-side website using the `rvest` package and convert it into a data frame. In the conceptual step column, please use bullet points to describe each step at a high level, including its input and output (in terms of concepts, not code). In the key function column, you only need to name the function, not its arguments.

Solution

	Conceptual step	**Key function**
1.	Concept: Retrieve html code from website Input: a url Output: an html document	read_html()
2.	Concept: find relevant element in html document Input: html document Output: html element	html_element()
3.	Concept: Convert html table to data frame Input: html element Output: data frame	html_table()

16. [10 points] The office of student services as a renowned university wants to understand how a student's performance in an exam relates to the number of hours they spent studying in the prior week. They have given you access to a dataset with information on 80 students enrolled in a programming course. You imported this data into R as a data frame and assigned it to the name `df`. The two variables in `df` required for our analysis are:

- `score`: a student's score in the course's exam (this is the dependent variable)
- `hours_study`: the number of hours spent studying in the week prior to the exam

The office asks you to analyze the relationship between the two variables, telling you to treat it as a linear relationship. Describe two ways in which you could analyze and convey your findings about the linear relationship between the two variables. For each one, describe your approach, name the key functions you would use, and explain what the output would look like.

Note: We will award credit based on the logic of your approaches first and foremost, followed by your understanding of key functions needed to implement the approach. You are not expected to write a complete code snippet that will run without errors (and will not receive any extra credit if you do).

Solution

1. Linear regression using the function `lm()`, which would produce a regression table with point estimates, standard errors, etc.

2. Plotting a linear fit using `ggplot()` with the geometry `geom_smooth(method = "lm")`, which would produce a plot of the regression line with `score` on the y-axis and `hours_study` on the x-axis.

17. [5 points] The dataset from the previous problem also contains information on the number of hours each student slept on the night before the exam in a variable `hours_sleep`. The office of student services is worried that student sleep habits may affect their performance. Explain how you would modify one of the two approaches you described in the previous problem to understand the linear relationship between `score` and `hours_study` along with this new variable `hours_sleep`.

Solution

Modify the linear regression model to be `lm(score ~ hours_study + hours_sleep, data = df)`. Because this is now a bivariate regression, you would interpret the coefficient for `hours_sleep` as how much `score` would change on average given an additional one hour of sleep holding constant the number of hours studied.

Similar answers that include both regressors in the linear model but present the results as a coefficient plot rather than regression table would also receive full credit.