DSA2101

Essential Data Analytics Tools: Data Visualization

Yuting Huang

Week 13 (Friday) Review

Final exam

The final exam is worth 40% of your final grade.

- **Time: Friday April 28, 9-11am**
- **Venue: UTown AUD2**
- Open-book, open-notes, closed-internet.
  - `R` packages: `readxl`, `stringr`, `lubridate`, `tidyverse`.
  - Data files will be available five minutes before the exam.
- Exam submissions:
  - When the exam ends, stop working/typing. You must submit all your answers on Examplify on 11am.
  - From 11am, you have 15 minutes to upload your `Rmd` file to the Canvas assignment "Final Exam".

Question format

The exam consists of

- **Part I**: 20 Short answer questions (25 marks).
  - Answer questions directly on Examplify by clicking on the options/filling in the blanks.
- **Part II**: 2 Coding questions (15 marks).
  - Answer questions in a single `Rmd` file and submit it on both Examplify and Canvas.

Manage your time wisely during the exam.

Review

Today we review the topics covered in this semester.

- Steps performed in a data analytics project:
  - Data import
  - Data cleaning
  - Data transformation
  - data visualization
- Other important concepts.

Data import

We learned about how to import data into `R`

- CSV files: `read.csv()`.
- Excel files: `read_excel()` from the `readxl` package.
- `R`'s own data format, RDS files: `readRDS()`.

We should regularly use **relative file paths**, which specifies the location of a file starting from the current location.

Throughout the semester, we have been reading data files via paths like `"../data/mycsvfile.csv"`.

Best practices in data cleaning

Data cleaning is the process of fixing (or removing) incorrect, duplicated, or incorrectly formatted data within a data set.

- "Best practices" might be a bit dramatic. But here's a list of things we find important during data cleaning stages.

Check for	Possible action(s)
missing data	Use `summary()` to examine and decide how to handle the `NA` values.
duplicated data	Use `distinct()` from `tidyverse`.
variable types	Convert variable into appropriate types.
outliers	Use `summary()` or boxplots.
factor levels	Use `table()` to check the levels. Especially look out for typos or mis-labelled observations.

Data transformation

When working with data, particularly large data sets, you will encounter situations where you need to

- Subset the data so that it contains only those *observations* that you are interested in.
- Subset the data so that it contains only those *variables* that you are interested in.
- Create new variables, often through calculations based on variables in your data.
- ...

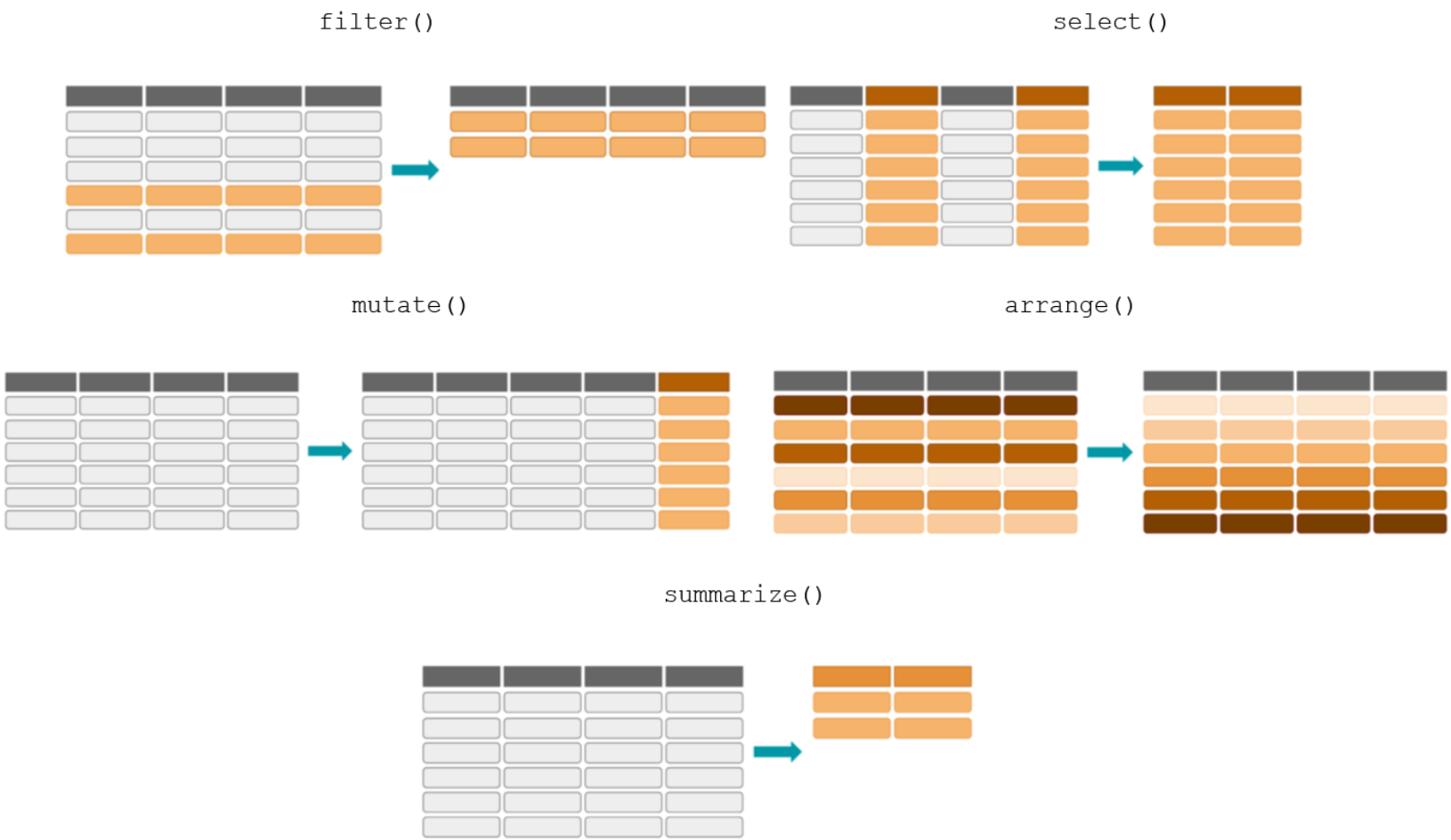
Data transformation

To achieve these goals, you will need functions from the `tidyverse` package.

Function	Action
`filter()`	Keep/drop rows.
`select()`	Keep/drop variables.
`mutate()`	Create new variables.
`arrange()`	Sort values from smallest to largest.
`summarize()`	Summarize all observations in the data frame.
`group_by()`	Group a data frame so subsequent operations are performed by group.

The pipe operator `%>%` chains `tidyverse` operations. It takes the output of a function and passes it into the argument of the subsequent function.

Data transformation



Tidy (reshaping) data

Tidy data follows the three rules:

- Each variable has its own column.
- Each observation has its own row.
- Each value has its own cell.

Many of the tools in `tidyverse` expect data to be formatted as a tidy data frame.

Tidy (reshaping) data

`gather()`

country	year	cases
Afghanistan	1999	745
Afghanistan	2000	2666
Brazil	1999	37737
Brazil	2000	80488
China	1999	212258
China	2000	213766

country	1999	2000
Afghanistan	745	2666
Brazil	37737	80488
China	212258	213766

table4

`spread()`

country	year	key	value
Afghanistan	1999	cases	745
Afghanistan	1999	population	19987071
Afghanistan	2000	cases	2666
Afghanistan	2000	population	20595360
Brazil	1999	cases	37737
Brazil	1999	population	172006362
Brazil	2000	cases	80488
Brazil	2000	population	174504898
China	1999	cases	212258
China	1999	population	1272915272
China	2000	cases	213766
China	2000	population	1280428583

table2

country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

`separate()`

country	year	rate
Afghanistan	1999	**745** / 19987071
Afghanistan	2000	**2666** / 20595360
Brazil	1999	**37737** / 172006362
Brazil	2000	**80488** / 174504898
China	1999	**212258** / 1272915272
China	2000	**213766** / 1280428583

table3

country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

`unite()`

country	year	rate
Afghanistan	**1999**	745 / 19987071
Afghanistan	**2000**	2666 / 20595360
Brazil	**1999**	37737 / 172006362
Brazil	**2000**	80488 / 174504898
China	**1999**	212258 / 1272915272
China	**2000**	213766 / 1280428583

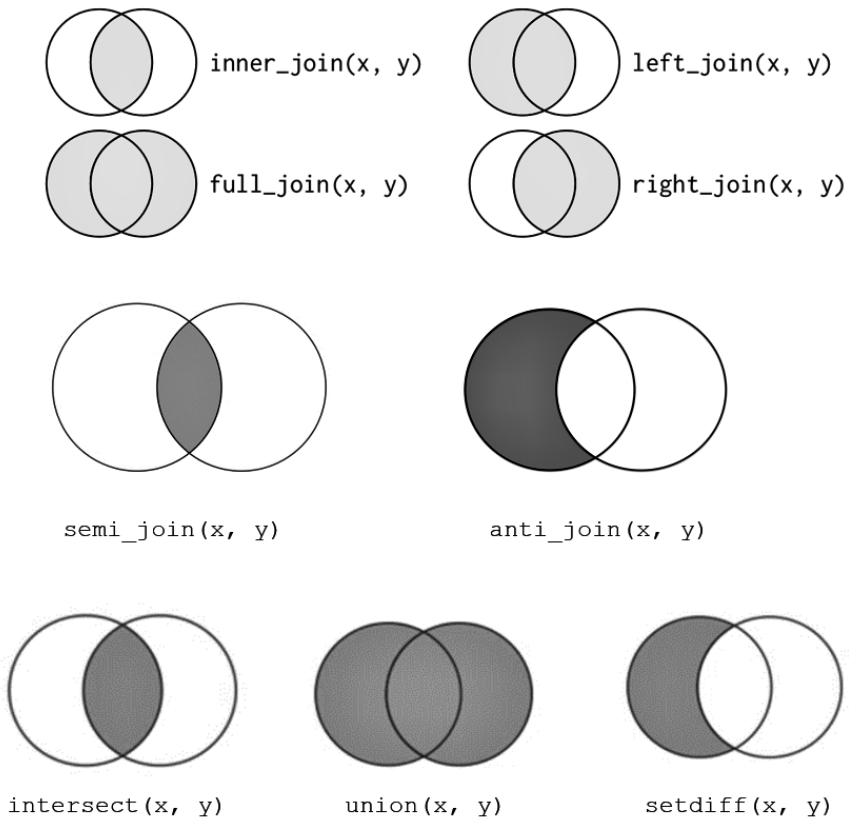
country	century	year	rate
Afghanistan	19	**99**	745 / 19987071
Afghanistan	20	**0**	2666 / 20595360
Brazil	19	**99**	37737 / 172006362
Brazil	20	**0**	80488 / 174504898
China	19	**99**	212258 / 1272915272
China	20	**0**	213766 / 1280428583

Relational data

- **Mutating joins**:
  Match by key variables and keep columns of both inputs.

- **Filtering joins**:
  Match by key variables and keep columns of the first input.

- **Set operations**:
  Expect column names to be the same in two inputs and compare values of every row.



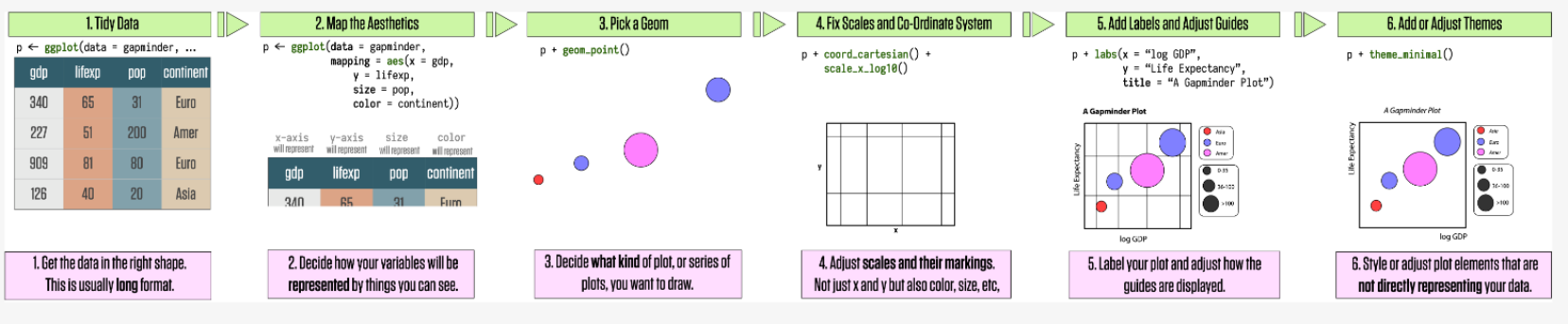
Data visualization

We learned about producing high-quality graphs with `ggplot()`.

- `ggplot()` can be described as a combination of the 7 parameters:

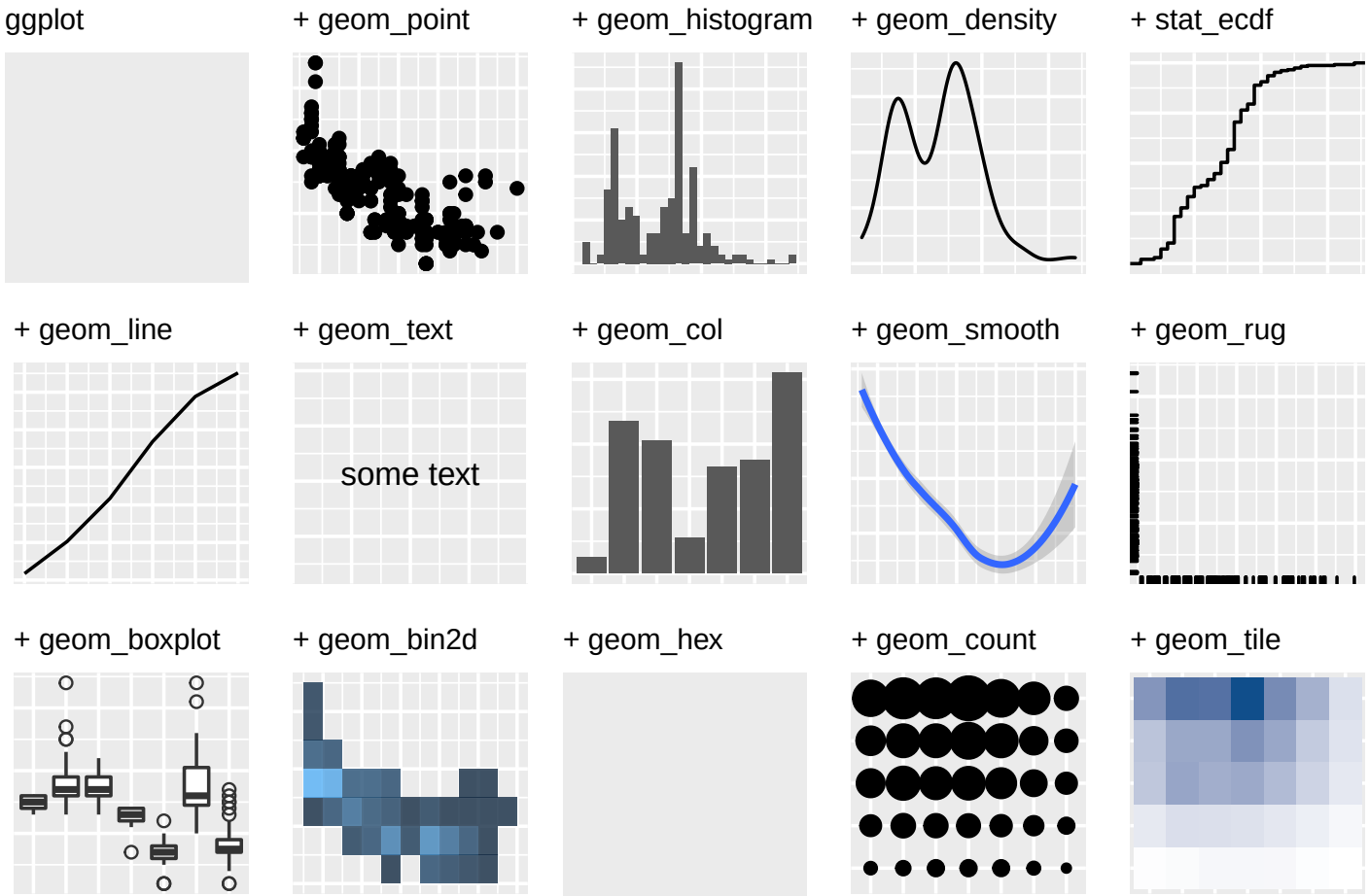
```
ggplot(data = <DATA>) +
  <GEOM_FUNCTION>(mapping = aes(<MAPPINGS>),
                  stat = <STAT>,
                  position = <POSITION>) +
  <COORDINATE_FUNCTION> +
  <FACET_FUNCTION>
```

- Here's the whole process from start to finish:



Data visualization

Some of the graphs we covered in class.



Practice questions

- Short-answer questions — No `Rmd` submission required
  - Questions with data
    - `volcano_practice.csv`
    - `continents_practice.csv`
  - Other questions
- `R` Coding questions — `Rmd` submission required
  - Questions with data
    - `retail_practice.rds`

Part I Short answer questions

These include multiple-choice and fill-in-the-blank questions.

- Answer these questions directly on Examplify by clicking on the options/filling in the blanks.
- Do not submit any code for this part.

Question with data

We will analyze data on volcano eruptions. Here are the variables available in the `volcano_practice.csv` data set.

```
volcano = read.csv("../data/volcano_practice.csv")
region = read.csv("../data/continents_practice.csv")
names(volcano)
```

```
##  [1] "volcano_number"           "volcano_name"
##  [3] "primary_volcano_type"     "last_eruption_year"
##  [5] "country"                  "region"
##  [7] "subregion"                "latitude"
##  [9] "longitude"                "elevation"
## [11] "tectonic_settings"        "evidence_category"
## [13] "major_rock_1"             "major_rock_2"
## [15] "major_rock_3"             "major_rock_4"
## [17] "major_rock_5"             "minor_rock_1"
## [19] "minor_rock_2"             "minor_rock_3"
## [21] "minor_rock_4"             "minor_rock_5"
## [23] "population_within_5_km"   "population_within_10_km"
## [25] "population_within_30_km"  "population_within_100_km"
```

We begin our analyses with numerical summaries of our data.

1. How many countries are there in this data set?
   Your answer: _____.

2. How many unique types of volcanoes are there in the data?
   Your answer: _____.

If we simply count the distinct volcano types in the raw data, we would find that there are 25 categories.

- Read closely the volcano types, we will see that some types should probably be merged. For example, Shield & Shield(s), Stratovolcano & Stratovolcano(es).
- Here is a code which merge the different types.

```
volcano = volcano %>%
  mutate(type = str_remove(primary_volcano_type, "\\(s\\)|\\(es\\)|\\?"))
```

- After cleaning up this variable, there are _____ distinct types of volcano in the data.

3. What is the **second most common** volcano type in the data?

A. Stratovolcano

B. Volcanic field

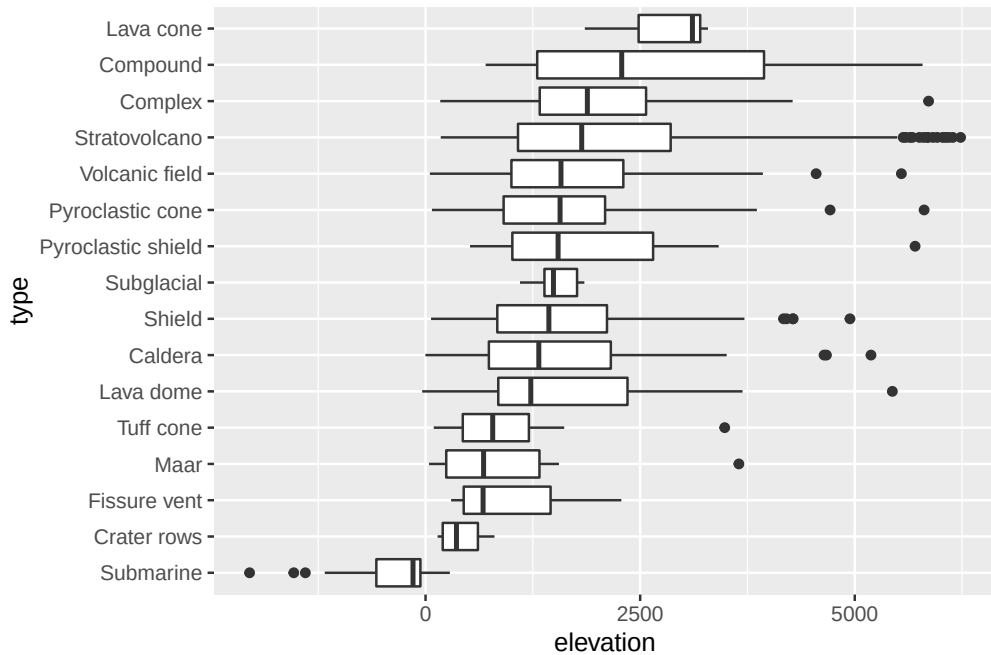
C. Shield

D. Caldera

E. Pyroclastic cone

4. Fill in the blanks to create the plot below that compares the distribution of elevation of different volcano types.

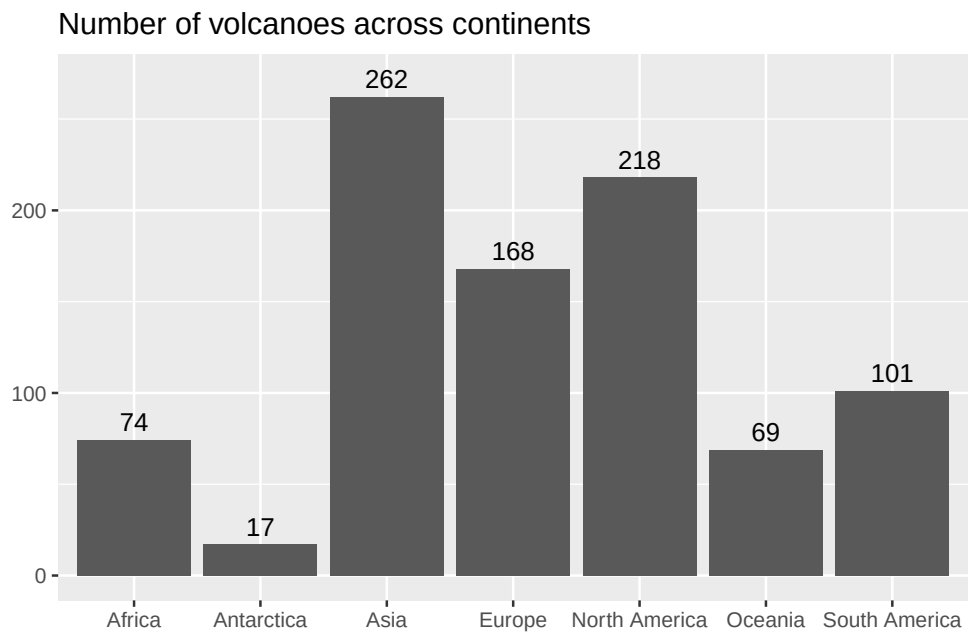
```
volcano %>%
  mutate(type = reorder(______, ______, median)) %>%
  ggplot(aes(x = type, y = elevation)) +
  geom_boxplot() +
  coord_flip()
```



5. Next, let us examine the number of volcanoes across continents. The file `continents_practice.csv` classifies countries into continents. We've read it in as `region`.

   Fill in the blanks in the code to create the bar chart below.

```
volcano = volcano %>%
  left_join(region, by = c(_______))
ggplot(data = volcano, aes(x = Continent)) +
  geom_bar() +
  geom_text(aes(label = _______), stat = "count", nudge_y = 10) +
  labs(x = "", y = "", title = "Number of volcanoes across continents")
```



6. Based on the bar plot in Question 7, to the closest 5%, what proportion of volcanoes are there in Asia?

A. 25%

B. 30%

C. 35%

D. 40%

Other questions

7. When reshaping data with the `gather()` function, what does the `key =` argument do?

A. Name a column that stores information previously spread across columns.

B. Name a column that stores values previously reside in entries

C. Name a column to join the current data frame with another table.

D. None of the above.

8. Complete the code to return the output:

```
speed = c("medium", "slow", "fast")
factor_speed = factor(speed, ____________,
                      levels = c("slow", "medium", "fast"))
factor_speed
```

```
## [1] medium slow   fast
## Levels: slow medium fast
```

A. `ordered = TRUE`

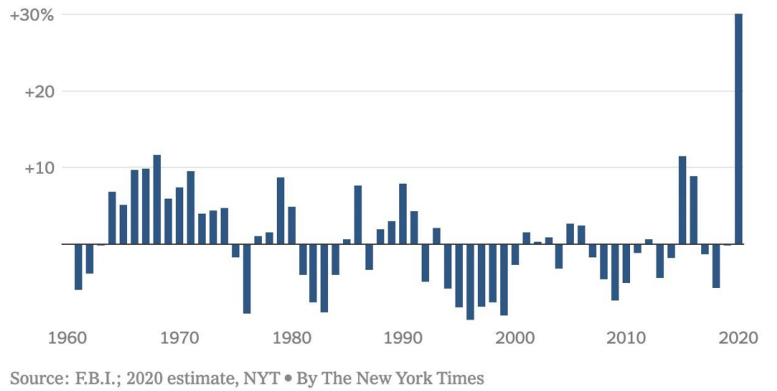
B. `ordered = FALSE`

C. `levels = TRUE`

D. `levels = FALSE`

9. Which of the following is **FALSE** about the two plots below?

Plot 1



Plot 2

A. Plot 1 shows the percentage changes in murders from one year to the next, while Plot 2 shows the rate of murders per 100,000 people across the years.

B. The spike after 2020 appears much larger in Plot 1 compared to Plot 2.

C. When the values in Plot 1 decrease, those in Plot 2 increase.

D. All of the above are true.

Part II Coding questions

- Answer these questions in a single `R` Markdown file.
- Make sure your `Rmd` file can knit to `HTML`.
- At the end of the exam, you are required to submit your answer to both Examplify and Canvas.

Data description

The file `retail_practice.rds` contains information on transactions made with an UK-based online retailer. The retailer mainly sells all-occasion gifts.

Here are some information about the variables in the data set.

- `InvoiceNo`: A number uniquely assigned to each transaction.
- `StockCode`: Product code.
- `Description`: Product name.
- `Quantity`: The quantity of each product in each transaction.
- `InvoiceDate`: Invoice date and time.
- `UnitPrice`: Unit price in Sterling Pounds.
- `CustomerID`: A number uniquely assigned to each customer.
- `Country`: The name of the country where each customer resides.

1. The company wants to figure out when people begin Christmas shopping, so that they can start preparing for it.

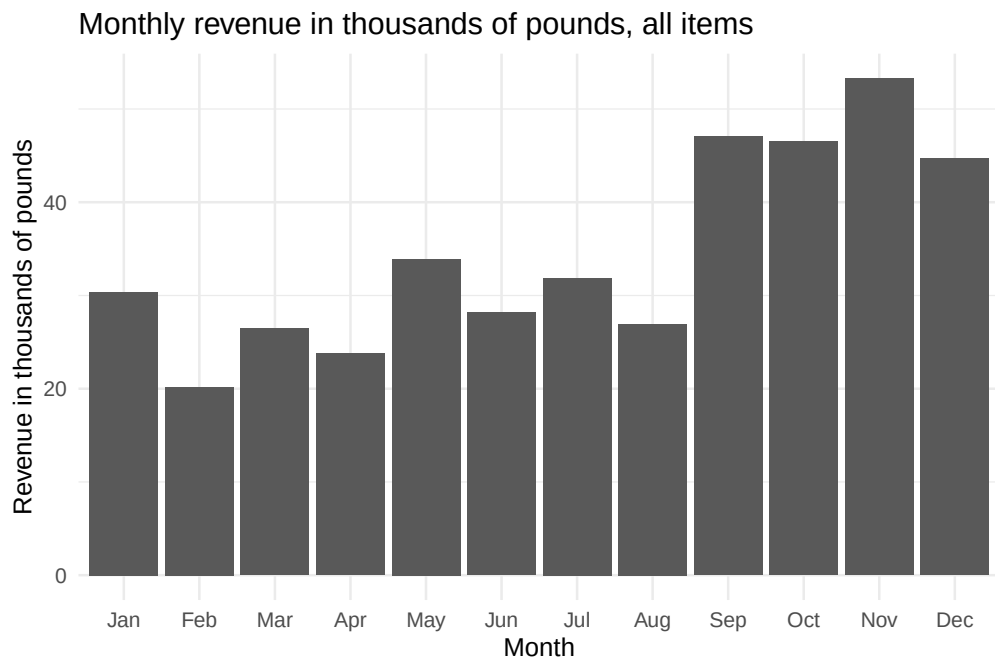
- Write the `R` code to create an object named `qn1` that contain the number of Christmas-related items sold in each month. Specifically, look for the following keywords in the `Description` column: CHRISTMASS, XMAS, REINDEER, or SANTA.

  The first four rows of `qn1` would look like this:

```
qn1 %>% head(4)
```

```
## # A tibble: 4 x 3
##    year month     n
##   <dbl> <ord> <int>
## 1  2010 Dec      80
## 2  2011 Jan       3
## 3  2011 Feb       2
## 4  2011 Mar       1
```

2. For each month, the revenue can be computed using `sum(UnitPrice*Quantity)`. An analyst in the company tried to visualize the monthly revenue for all items using the following plot. Is this plot accurate?
   - If you feel that there is any miskate(s) in this plot, produce a corrected one.
   - If you feel that this plot cannot be corrected or improved, create a variation of it that also shows the monthly revenue of the company.



At the end of exam

- When the exam ends, stop working/typing. You must submit all your answers on Examplify at 11am sharp.
- From 11am, you have **15 minutes** to upload your `Rmd` file to the Canvas assignment "Final Exam".
- Make sure you submit the correct file.

Do not modify your code in your `Rmd` submission file. Any difference (except for indentation differences) found between Examplify and Canvas submissions will be penalized.

Additional practices

See if you can answer a few more questions about the NYC flights data set. Most of these questions can be answered in one `tidyverse` chain.

1. What destination received most flights from `JFK` in December?
   Answer: LAX

2. Which carrier had the greatest average distance per flight?
   HA, or Hawaiian Airlines Inc

3. What day had the largest average arrival delay for all flights?
   July 10th

4. What was the average number of seats (round to the second decimal place) on the planes that left from airports in the New York City on July 4th?
   140.66

5. Which scheduled departure hour had the largest proportion of flights delayed longer than 15 minutes?
21

6. What plane (`tailnum`) traveled the most times to `JFK`? Plot the number of trips per week over the year for that plane. Your plot should look similar to the following: