

Homework 8

Zahlen Zbinden

2023-11-21

```
library(tidyverse)
library(nycflights13)
```

Warning: package 'nycflights13' was built under R version 4.3.2

Tasks that require an answer are bolded (inside ****** in the .qmd file). For any task that includes a question (i.e. it ends with “?”), you should also answer the question in sentence form.

Logical Vectors

The first part of the homework uses the `flights` and `airports` data in the `nycflights13` package.

1.

(2 pts)

For each of the following, filter `flights` to **find the number of flights that match the criteria**. An easy way to return just the number of flights is to pipe your resulting table into `nrow()`.

a) **Flights that flew from from JFK to PDX.**

```
flights |>
  filter(origin == "JFK" & dest == "PDX") |>
  nrow()
```

[1] 783

- b) **Flights that arrive early, but have a arrival time that is greater than the scheduled arrival time.**

```
flights |>
  filter(arr_delay < 0 & (arr_time > sched_arr_time)) |>
  nrow()
```

[1] 777

- c) **Flights that depart in a odd numbered month.**

```
flights |>
  filter(month %% 2 != 0) |>
  nrow()
```

[1] 168901

- d) **Flights that departed on Friday the 13th.** *(In 2013, Friday the 13th occurred in September and December).*

```
flights |>
  filter(month == 9 | month == 12) |>
  filter(day == 13) |>
  nrow()
```

[1] 1966

2. How many flights a month went Hawaii?

(2 pts)

First, you'll need to find all the airport codes for airports in Hawaii. **Filter the airports data to airports where the tz column has the value -10, and then, pull() the faa column. Store the results in hawaii_codes.**

```
hawaii_codes <- airports |>
  filter(tz == -10) |>
  pull(faa)
```

```

flights |>
  filter(dest %in% hawaii_codes) |>
  group_by(month) |>
  summarise(count = n())

```

A tibble: 12 x 2

	month	count
	<int>	<int>
1	1	62
2	2	56
3	3	62
4	4	60
5	5	62
6	6	60
7	7	62
8	8	62
9	9	55
10	10	52
11	11	55
12	12	59

Extract the rows in `flights` with a destination airport in the `hawaii_codes` vector. Store the result in `hawaii_flights`. Hint: the `%in%` operator is a logical operator (like `==` or `<`) that can be used to check if a value is equal to one of the values in a vector.

```

hawaii_flights <- flights |>
  filter(dest %in% hawaii_codes)

```

Summarise `hawaii_flights` by counting the number of flights in each month, and then, produce a scatterplot of the number of flights by month.

```

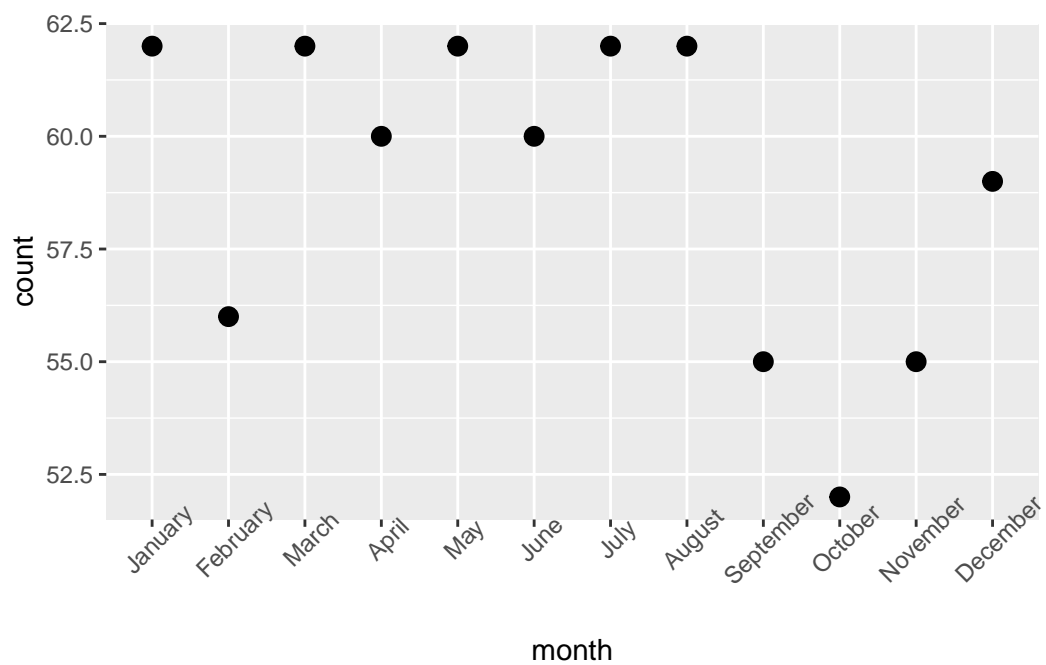
hawaii_flights |>
  group_by(month) |>
  summarise(count = n()) |>
  mutate(month = month.name[month]) |>
  mutate(month = factor(month)) |>
  mutate(
    month = fct_relevel(
      month,
      c(
        "January",

```

```

    "February",
    "March",
    "April",
    "May",
    "June",
    "July",
    "August",
    "September",
    "October",
    "November",
    "December"
  )
)
) |>
ggplot(aes(x = month, y = count)) +
  geom_point(
    size = 3
  ) +
  theme(
    axis.text.x = element_text(angle = 45)
  )

```



3. What proportion of flights went to Hawaii?

(2 pts)

A criticism of the previous plot might be that it just shows the absolute number of flights to Hawaii, not what proportion of all flights head to Hawaii.

Rather than filtering flights to flights heading to Hawaii, add a column called `hawaii` to `flights` that is `TRUE` if the flight has a destination in Hawaii.

```
hawaii_flights <- flights |>
  mutate(
    hawaii = ifelse(
      dest %in% hawaii_codes,
      TRUE,
      FALSE
    )
  )
```

Summarise flights by computing the `mean()` of the `hawaii` column in each month. What does this value represent?

```
plt_summary <- hawaii_flights |>
  group_by(month) |>
  mutate(month = month.name[month]) |>
  mutate(month = factor(month)) |>
  mutate(
    month = fct_relevel(
      month,
      c(
        "January",
        "February",
        "March",
        "April",
        "May",
        "June",
        "July",
        "August",
        "September",
        "October",
        "November",
        "December"
      )
    )
  )
```

```

    )
  ) |>
  summarise(hawaii_prop = mean(hawaii))

```

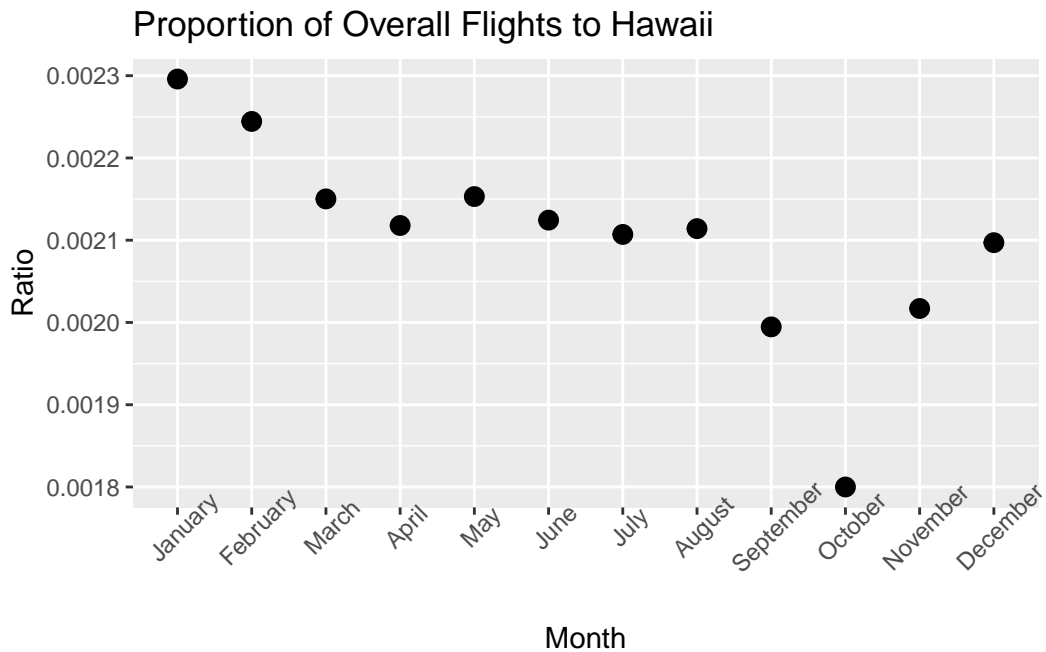
Create a scatterplot of the `mean()` of the `hawaii` column by month. Does it tell a different story to your previous plot?

Yes this tells a much different story, as we can see the number of flights to hawaii in proportion to all other flights is decreasing, as opposed to the previous plot that just made it look like the flights to hawaii were pretty stagnant in level.

```

plt_summary |>
  ggplot(aes(x = month, y = hawaii_prop)) +
    geom_point(size = 3) +
    labs(
      title = "Proportion of Overall Flights to Hawaii",
      x = "Month",
      y = "Ratio"
    ) +
    theme(
      axis.text.x = element_text(angle = 45)
    )

```



Lists

4.

(2 pts)

You saw that tibbles are also considered lists, so this means you can use the same subsetting tools ([, [[, \$) on them as you can on lists.

Extract the 1st element of `starwars` with [.

```
starwars[1]
```

```
# A tibble: 87 x 1
  name
  <chr>
1 Luke Skywalker
2 C-3PO
3 R2-D2
4 Darth Vader
5 Leia Organa
6 Owen Lars
7 Beru Whitesun lars
8 R5-D4
9 Biggs Darklighter
10 Obi-Wan Kenobi
# i 77 more rows
```

Extract the 1st element of `starwars` with [[.

```
starwars[[1]]
```

[1] "Luke Skywalker"	"C-3PO"	"R2-D2"
[4] "Darth Vader"	"Leia Organa"	"Owen Lars"
[7] "Beru Whitesun lars"	"R5-D4"	"Biggs Darklighter"
[10] "Obi-Wan Kenobi"	"Anakin Skywalker"	"Wilhuff Tarkin"
[13] "Chewbacca"	"Han Solo"	"Greedo"
[16] "Jabba Desilijic Tiure"	"Wedge Antilles"	"Jek Tono Porkins"
[19] "Yoda"	"Palpatine"	"Boba Fett"
[22] "IG-88"	"Bossk"	"Lando Calrissian"
[25] "Lobot"	"Ackbar"	"Mon Mothma"
[28] "Arvel Crynyd"	"Wicket Systri Warrick"	"Nien Nunb"

[31]	"Qui-Gon Jinn"	"Nute Gunray"	"Finis Valorum"
[34]	"Jar Jar Binks"	"Roos Tarpals"	"Rugor Nass"
[37]	"Ric Olié"	"Watto"	"Sebulba"
[40]	"Quarsh Panaka"	"Shmi Skywalker"	"Darth Maul"
[43]	"Bib Fortuna"	"Ayla Secura"	"Dud Bolt"
[46]	"Gasgano"	"Ben Quadinaros"	"Mace Windu"
[49]	"Ki-Adi-Mundi"	"Kit Fisto"	"Eeth Koth"
[52]	"Adi Gallia"	"Saesee Tiin"	"Yarael Poof"
[55]	"Plo Koon"	"Mas Amedda"	"Gregar Typho"
[58]	"Cordé"	"Cliegg Lars"	"Poggle the Lesser"
[61]	"Luminara Unduli"	"Barriss Offee"	"Dormé"
[64]	"Dooku"	"Bail Prestor Organa"	"Jango Fett"
[67]	"Zam Wesell"	"Dexter Jettster"	"Lama Su"
[70]	"Taun We"	"Jocasta Nu"	"Ratts Tyerell"
[73]	"R4-P17"	"Wat Tambor"	"San Hill"
[76]	"Shaak Ti"	"Grievous"	"Tarfful"
[79]	"Raymus Antilles"	"Sly Moore"	"Tion Medon"
[82]	"Finn"	"Rey"	"Poe Dameron"
[85]	"BB8"	"Captain Phasma"	"Padmé Amidala"

Extract the 1st element of starwars with \$. (You'll have to figure out what this element is called to use \$).

```
starwars$name
```

[1]	"Luke Skywalker"	"C-3P0"	"R2-D2"
[4]	"Darth Vader"	"Leia Organa"	"Owen Lars"
[7]	"Beru Whitesun lars"	"R5-D4"	"Biggs Darklighter"
[10]	"Obi-Wan Kenobi"	"Anakin Skywalker"	"Wilhuff Tarkin"
[13]	"Chewbacca"	"Han Solo"	"Greedo"
[16]	"Jabba Desilijic Tiure"	"Wedge Antilles"	"Jek Tono Porkins"
[19]	"Yoda"	"Palpatine"	"Boba Fett"
[22]	"IG-88"	"Bossk"	"Lando Calrissian"
[25]	"Lobot"	"Ackbar"	"Mon Mothma"
[28]	"Arvel Crynyd"	"Wicket Systri Warrick"	"Nien Nunb"
[31]	"Qui-Gon Jinn"	"Nute Gunray"	"Finis Valorum"
[34]	"Jar Jar Binks"	"Roos Tarpals"	"Rugor Nass"
[37]	"Ric Olié"	"Watto"	"Sebulba"
[40]	"Quarsh Panaka"	"Shmi Skywalker"	"Darth Maul"
[43]	"Bib Fortuna"	"Ayla Secura"	"Dud Bolt"
[46]	"Gasgano"	"Ben Quadinaros"	"Mace Windu"
[49]	"Ki-Adi-Mundi"	"Kit Fisto"	"Eeth Koth"

[52]	"Adi Gallia"	"Saesee Tiin"	"Yarael Poof"
[55]	"Plo Koon"	"Mas Amedda"	"Gregar Typho"
[58]	"Cordé"	"Cliegg Lars"	"Poggle the Lesser"
[61]	"Luminara Unduli"	"Barriss Offee"	"Dormé"
[64]	"Dooku"	"Bail Prestor Organa"	"Jango Fett"
[67]	"Zam Wesell"	"Dexter Jettster"	"Lama Su"
[70]	"Taun We"	"Jocasta Nu"	"Ratts Tyerell"
[73]	"R4-P17"	"Wat Tambor"	"San Hill"
[76]	"Shaak Ti"	"Grievous"	"Tarfful"
[79]	"Raymus Antilles"	"Sly Moore"	"Tion Medon"
[82]	"Finn"	"Rey"	"Poe Dameron"
[85]	"BB8"	"Captain Phasma"	"Padmé Amidala"

Compare the subsetting above to the tidyverse tools you know. Which correspond to `select()` and which to `pull()`?

`starwars[1]` and `select` work the same by returning a tibble 87X1 with all the names, `starwars[[1]]`, `starwars$name`, and `pull(name)` all work the same by returning a list of all the names.

5.

(2 pts)

Data frames are also considered lists. Read the section comparing [Tibbles vs. data.frame in R for Data Science](#). Pay close attention to the Extracting variables, and Subsetting sections.

How do data frames behave differently to tibbles in terms of extracting or subsetting variables? Tibbles are more strict, they never do partial matching, and will generate a warning if the column you are trying to access does not exist.