Data Wrangling I

The questions below refer to the data frame called `day_at_cal`, with the habits of six students during a day at Cal including the day of the week (`weekday`), their study spot that day (`study_spot`), the number of total hours that they used AI in the course of their work (`ai`), and the number of steps walked, in thousands (`steps_k`). To check your answers in R, you can copy and paste the code that creates this data frame from the last slide from class.

```
day_at_cal
```

```
# A tibble: 6 × 4
  weekday study_spot ai_hours steps_k
  <ord>   <fct>         <dbl>   <dbl>
1 Mon     Moffitt         1.2     8.5
2 Mon     Doe             0.4     6.2
3 Tue     Cory            2.1     9.1
4 Tue     MLK             1       7
5 Wed     CITRIS          0.8     5.8
6 Wed     Soda            1.6    10.3
```

Write a single `dplyr` command that will produce each of the following.

1. A data frame containing the 2nd and 5th students.

2. A data frame without the 3rd student.

3. A data frame with only the AI hours column.

4. A data frame with only the factor columns[1].

5. A data frame containing only the students with at least one hour of AI use.

6. A data frame with students who study in either Moffitt or Doe.

7. A data frame with a fifth column called `ai_min` with the AI use in minutes.

8. A data frame with a fifth (logical) column called `active_day` where the number of steps exceeds 8,000.

[1] There is a clever way to do this that avoids having to type the names of the columns that are factors. See the help file for `select()`, particularly about predicate functions.

9. The same data frame sorted in descending order of number of hours of AI.

10. The same data frame sorted according to the day of the week. Within rows with the same day of the week, sort in descending order of the number of steps.

11. A data frame with the overall mean hours of AI use. Name that statistic `mean_ai`.

12. A data frame with the minimum steps walked (`min_steps`), the maximum steps walked (`max_steps`), and the standard deviation of steps walked (`sd_steps`).

13. Skim the documentation for the `tibble` package (https://tibble.tidyverse.org/) then write down three ways in which tibbles behave differently than data frames.

14. If you're interested in reading more about the design philosophy behind the Tidyverse, read *The Tidy Tools Manifesto* by Hadley Wickham, https://cran.r-project.org/web/packages/tidyverse/vignettes/manifesto.html, (it's just two pages).

Data Wrangling II

For this problem set, you'll be working with the `storms` tibble built into `dplyr`. This is a larger data frame than previous, so you may need to inspect it more carefully in R before proceeding with these questions. Please answer the question in common english and provide the code needed to produce the output used in your answer.

On the *challenging* exercises, a helpful approach is to start by sketching the data frame that would answer the question. What are the dimensions? What does each row refer to? Which columns are needed? Then piece together your pipeline to lead from `storms` to that data frame.

```
library(dplyr)
storms
```

```
# A tibble: 19,537 × 13
   name   year month   day  hour   lat  long status      category  wind pressure
   <chr> <dbl> <dbl> <int> <dbl> <dbl> <dbl> <fct>          <dbl> <int>    <int>
 1 Amy    1975     6    27     0  27.5 -79   tropical d…       NA    25     1013
 2 Amy    1975     6    27     6  28.5 -79   tropical d…       NA    25     1013
 3 Amy    1975     6    27    12  29.5 -79   tropical d…       NA    25     1013
 4 Amy    1975     6    27    18  30.5 -79   tropical d…       NA    25     1013
 5 Amy    1975     6    28     0  31.5 -78.8 tropical d…       NA    25     1012
 6 Amy    1975     6    28     6  32.4 -78.7 tropical d…       NA    25     1012
 7 Amy    1975     6    28    12  33.3 -78   tropical d…       NA    25     1011
 8 Amy    1975     6    28    18  34   -77   tropical d…       NA    30     1006
 9 Amy    1975     6    29     0  34.4 -75.8 tropical s…       NA    35     1004
10 Amy    1975     6    29     6  34   -74.8 tropical s…       NA    40     1002
# ℹ 19,527 more rows
# ℹ 2 more variables: tropicalstorm_force_diameter <int>,
#   hurricane_force_diameter <int>
```

15. What are the dimensions of the data frame? What does each row correspond to (i.e. what is the unit of observation)?

16. Calculate the average wind speed and pressure for each category of storm.

17. Find the number of observations and the maximum wind speed for each combination of storm status and category. Sort the results by maximum wind speed in descending order[2].

18. Use the `.by` shortcut to calculate the average pressure for each year. Repeat the calculation using `group_by()`. Do you get the same answer?

19. For each storm name, calculate the z-score of wind speed (standardized within each storm) and store it in a new column called `wind_z`. Return the data frame that

20. *Challenging*. For each storm category, arrange the storms in descending order by their maximum wind speed. Retain only the top three storms in each category[3].

[2]Hint 1: `n()` is a function that returns the row count when used inside `summarize()`. Hint 2: You can group by more than one variable.

[3]Hint: If you group by multiple variables, a call to `summarize()` will peel off the outermost grouping (the last argument in `group_by()`) and keep the others

21. *Challenging*: Create a summary that shows, for each year: the number of unique storms, the average duration (in hours) of storms, and the proportion of observations that are category 4 or 5 storms.

22. You may have noticed that your results for the answers in this problem set contain a value of `NA`. By default, most statistical summaries in R (`mean()`, `max()`, etc.) with return `NA` if any of the values are `NA`. If you prefer, you can calculate the mean after dropping the `NA` observations. One way to do this is to use the `na.rm = TRUE` option for functions that allow it (see `?mean`). The other way is to use the `drop_na()` function in `dplyr` (see `?drop_na()`).

    Revisit your code above and, if you haven't already, drop the `NA` values in a column before calculating summary statistics.

Intro to Data Visualization

Today's exercises bear on the first chapter of *Data Visualization* by Kieran Healy.

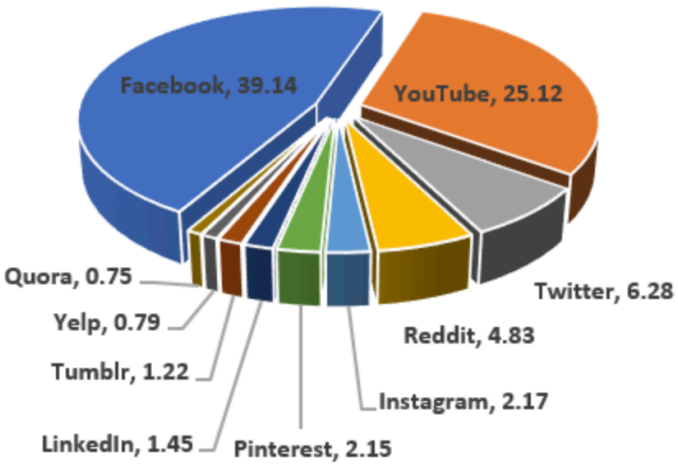
https://socviz.co/lookatdata.html#lookatdata

Healy splits bad data visualization into three categories of failure: aesthetic, substantive, and perceptual.

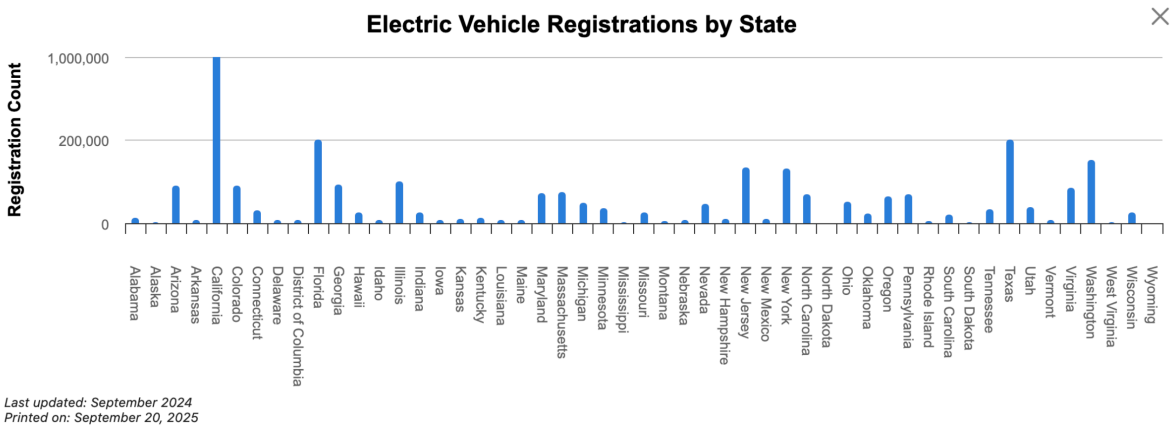
For each of the following visualizations, describe whether the graphic falls into any of those categories.

23.

24.



25.



For each of the following variables with data types, describe which color palette would be most appropriate: sequential, diverging, or qualitative.

26. Annual income (double vector).

27. Temperature change (double vector), containing positive and negative changes.

28. Country (character vector or factor).

29. Highest educational degree (ordered factor).

30. Perceptual research shows that human can more accurately interpret data visualizations when they utilize particular channels. Which common plot types use channels that are known to be the most accurate? Which plot types use channels known to be inaccurate?