Snowfall and connections to Latitude

Jesse Plummer

Participants (or Observations)

The data that I used was the **weather.csv** which included the following variables:

Variable Name	Type	Description
station	character	the geographic location of the station where the data was collected
state	character	the state (or territory) where the data was collected
latitude	numeric	the latitude of the station
longitude	numeric	the longitude of the station
elevation	numeric	the height of the station
date	numeric	the date when the data was collected (year, month, day)
temp_min	numeric	the minimumm temperature collected at the station
temp_max)	numeric	the maximum temperature collected at the station
temp_avg	numeric	the avergae temperature collected at the station
av_day_wind_spd	numeric	the average daily wind speed at the station
wi_dir_5sec	numeric	the wind direction in 5 second intervals (at the station)
wi_spd_5sec	numeric	the wind speed in 5 second intervals (at the station)
snow_fall	numeric	the amount of snowfall recorded at the station
snow_dep	numeric	the snow depth recorded at the station
precip	numeric	the amount precipitation recorded at the station

My hypothesis is that ***more snowfall would occur at stations above the 40th parallel.***

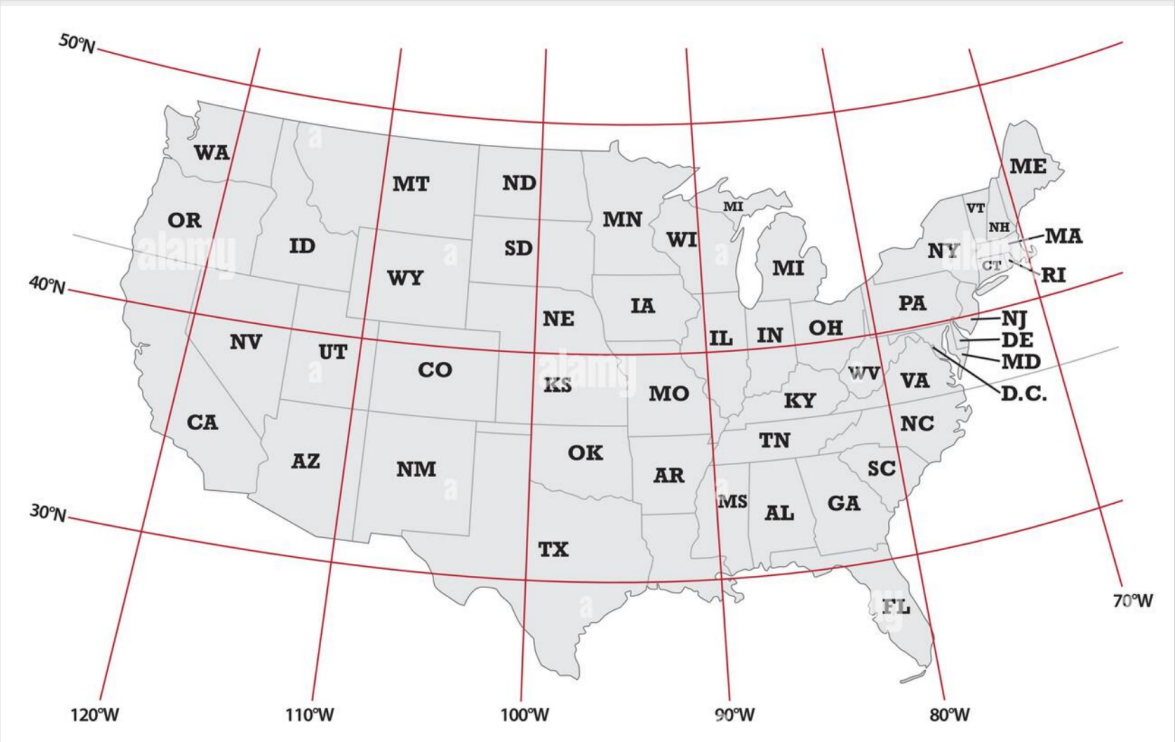


Figure 1: Map of the United States: Latitude & Longitude shown

The variables that I used in this analysis were:

- latitude: this would be used to create a clear division between data points
- snow_fall: one of the variables being analyzed that would show that the hypothesis could be verified or dismissed.
- temp_avg: another variable that demonstrates one possible explanation as to why the hypothesis could be true.

Procedure

How I did the analysis:

1. Load the required libraries into RStudio

```
#|label: load-packages
#|code-summary: Packages required for analysis
```

```
#|message: false
#|include: false
#|warning: false

library(tidyverse)
```

```
Warning: package 'tidyverse' was built under R version 4.2.3

Warning: package 'ggplot2' was built under R version 4.2.3

Warning: package 'tibble' was built under R version 4.2.3

Warning: package 'tidyr' was built under R version 4.2.3

Warning: package 'readr' was built under R version 4.2.3

Warning: package 'purrr' was built under R version 4.2.3

Warning: package 'dplyr' was built under R version 4.2.3

Warning: package 'stringr' was built under R version 4.2.3

Warning: package 'forcats' was built under R version 4.2.3

Warning: package 'lubridate' was built under R version 4.2.3

-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.4.4     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.0
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to beco
```

```
  library(lme4)
```

```
Warning: package 'lme4' was built under R version 4.2.3
```

```
Loading required package: Matrix
```

```
Warning: package 'Matrix' was built under R version 4.2.3
```

```

Attaching package: 'Matrix'
```

```
The following objects are masked from 'package:tidyr':

    expand, pack, unpack
```

```
  library(lmerTest)
```

```
Warning: package 'lmerTest' was built under R version 4.2.3
```

```

Attaching package: 'lmerTest'
```

```
The following object is masked from 'package:lme4':

    lmer
```

```
The following object is masked from 'package:stats':

    step
```

```
  library(dplyr)
```

(Bates et al. 2015)

(Wickham et al. 2019)

(Kuznetsova, Brockhoff, and Christensen 2017)

1. (Wickham et al. 2023)
2. Move the data into RStudio then read the csv file into a tibble

```
#|label: Load data file
#|code-summary: read csv into a tibble
#| output: false

#load("weather.csv")
weather <- read_csv('weather.csv')
```

```
Rows: 416937 Columns: 15
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (2): station, state
dbl (13): latitude, longitude, elevation, date, temp_min, temp_max, temp_avg...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

3. Add a column count to the new tibble and save it to a data frame (df0)

```
#|label: Create data frame
#|code-summary: create data frame with count column
#|output: false

df0 <- weather %>%
add_count(station, name = 'station_cnt') %>%

  add_count(state, name = 'state_cnt') %>%

  add_count(state, name = 'state_cnt') %>%

  add_count(latitude, name = 'latitude_cnt') %>%

  add_count(longitude, name = 'longitude_cnt') %>%

  add_count(elevation, name = 'elevation_cnt') %>%

  add_count(date, name = 'date_cnt') %>%

  add_count(temp_min, name = 'temp_min_cnt') %>%

  add_count(temp_max, name = 'temp_max_cnt') %>%
```

```
  add_count(temp_avg, name = 'temp_ave_cnt') %>%

  add_count(av_day_wi_spd, name = 'av_day_wi_spd_cnt') %>%

  add_count(wi_dir_5sec, name = 'wi_dir_5sec_cnt') %>%

  add_count(wi_spd_5sec, name = 'wi_spd_5sec_cnt') %>%

  add_count(snow_fall, name = 'snow_fall_cnt') %>%

  add_count(snow_dep, name = 'snow_dep_cnt') %>%

  add_count(precip, name = 'precip_cnt')
```

4. Make sure data is clean (remove any NAs), add a new column to the df0 data frame that differentiates above or below the 40th parallel, and save that to a new data frame, data_for_analysis.

```
#|label: Clean data and add new column
#|code-summary: omit any NAs from df0, add another column

df0 %>%
  na.omit() %>%
  mutate(lat_Abv_40 = ifelse(latitude > 40, "abv", "bel")) -> data_for_model
```

5. Create a data model that will show if we may accept or reject the null hypothesis

```
#|label: For a conclusion on the hypothesis based on p-value
#|code_summary: using lmer find a p-value

high_lat_model <- lmer(snow_fall ~ lat_Abv_40 + (1|station),
                       data= data_for_model)
summary(high_lat_model)
```

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method [
lmerModLmerTest]
Formula: snow_fall ~ lat_Abv_40 + (1 | station)
   Data: data_for_model

REML criterion at convergence: 97063.1

Scaled residuals:
```

```
   Min     1Q Median     3Q    Max
-0.637 -0.185 -0.039 -0.008 59.716

Random effects:
 Groups   Name        Variance Std.Dev.
 station  (Intercept) 0.002445 0.04945
 Residual             0.268421 0.51809
Number of obs: 63536, groups:  station, 272

Fixed effects:
                Estimate Std. Error         df t value Pr(>|t|)
(Intercept)     0.114634   0.005588 207.643785   20.52   <2e-16 ***
lat_Abv_40bel  -0.101701   0.007462 208.526977  -13.63   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
            (Intr)
lt_Abv_40bl -0.749
```

Based on this p value we can reject the null hypothesis, and conclude that stations above the 40th parallel (latitudes above 40 degrees) have more snowfall that stations below the 40th parallel.

6. Pearson Correlation Test: The next thing I thought about was to to give a reason why my prediction could be true. Generally speaking (through common knowledge) the higher the parallel (the higher the latitude) the lower the average temperature would be. To show this correlation, I ran a Pearson's correlation, looking for an r value that shows a correlation between the average temperature and latitude.

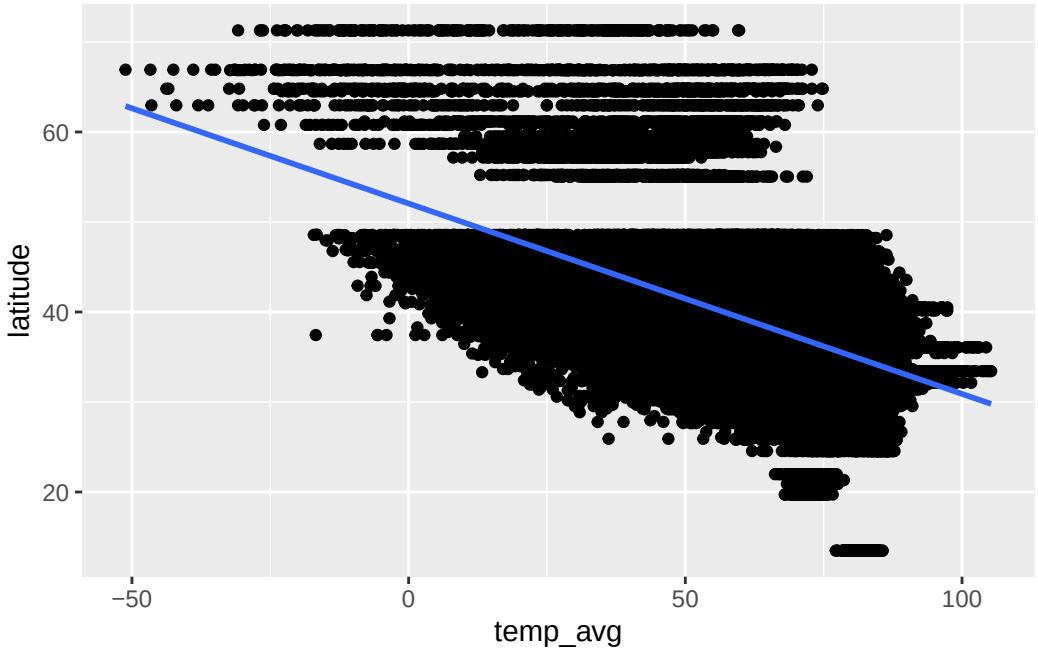
```
#|label: Pearson Correlation
#|code_summary: running a pearson correlation on data_for_model
data_cor_0 <- cor(data_for_model$latitude, data_for_model$temp_avg,
                  method = c("pearson"))
summary(data_cor_0)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-0.507  -0.507  -0.507  -0.507  -0.507  -0.507
```

The r-value shows a strong negative correlation between the average temperature and latitude, as shown in the plot below. This means that as the latitude increases, the temperature decreases. This meets my prediction and expectation in support of the hypothesis.

```
#|label: Plot of correlation
#|code_summary: ggplot to plot all data points, add a line showing correlation
#|fig-cap: Figure-1 Pearson correlation of average temperature and latitude
data_for_model %>%
  ggplot(aes(x = temp_avg, y = latitude)) + geom_point() +
  geom_smooth(method="lm")
```

```
`geom_smooth()` using formula = 'y ~ x'
```



Note: each dot represents a station.

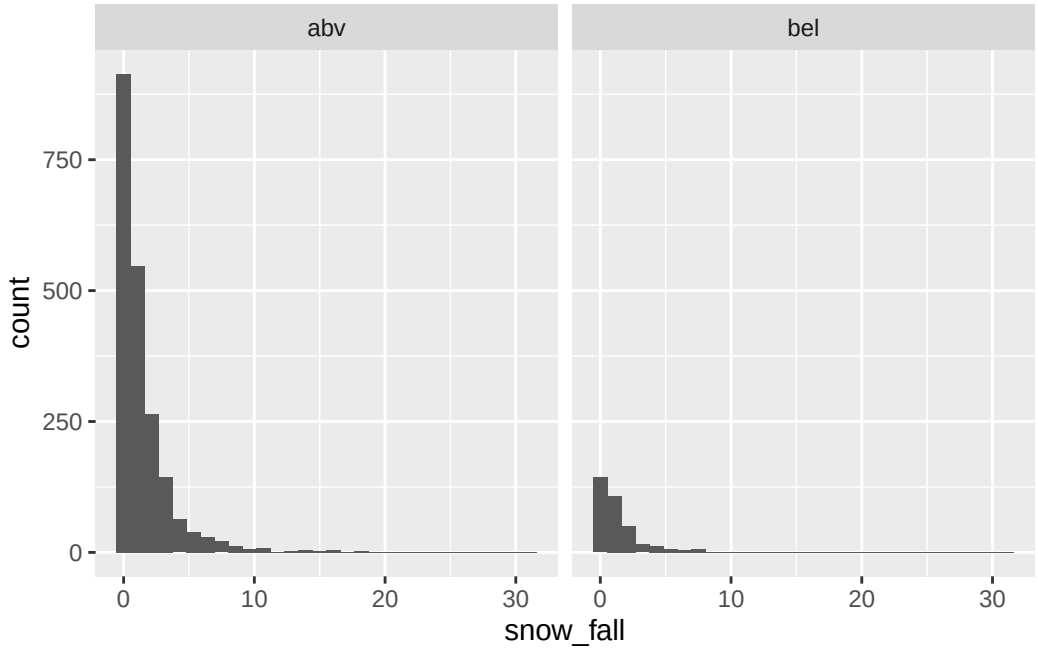
In order to show the snowfall in a different way, I cleaned the data once again and displayed it in a histogram

```
#|label: Clean data for use with histogram
#|code_summary: filter out stations without snowfall
#|fig-cap: Histogram showing only stations that received snowfall
```

```
data_for_model %>%
  filter(snow_fall != 0) -> data_only_snow_fall
```

```
#|label: Plot hisgtoram
#| code_summary: create histogram
#| fig-cap: Histogram showing both stations that did and did not receive snowfall
data_only_snow_fall %>%
  ggplot(aes(x = snow_fall)) + geom_histogram() + facet_wrap(~lat_Abv_40)
```

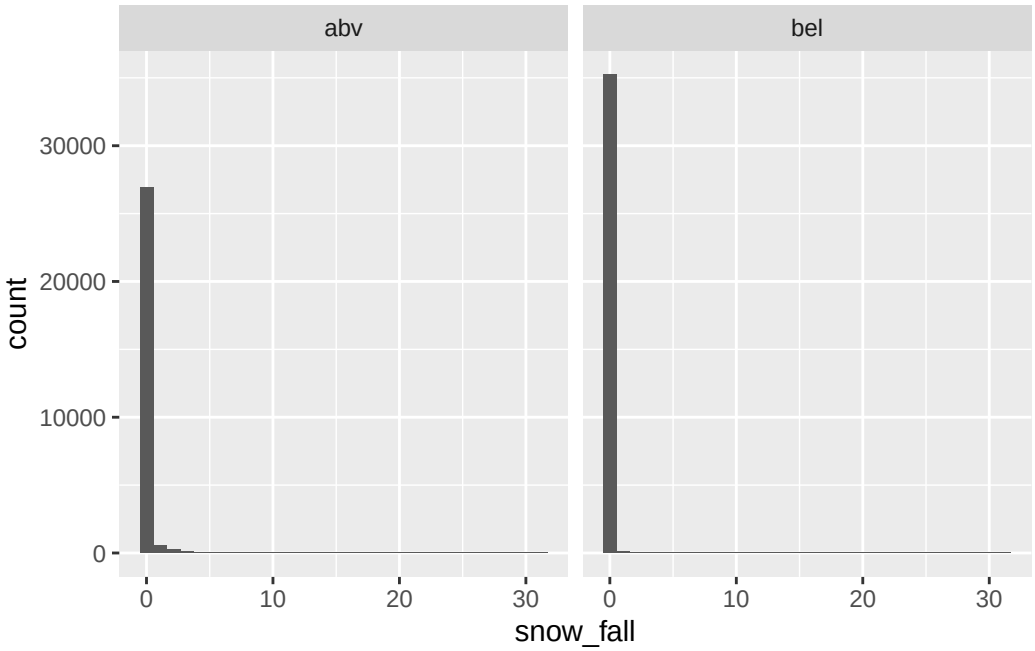
```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



For context, this is what the data looks like with the stations without snowfall

```
#|label: Plot histogram (with no snowfall datapoints)
#|code_summary: create a second histogram
data_for_model %>%
  ggplot(aes(x = snow_fall)) + geom_histogram() + facet_wrap(~lat_Abv_40)
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



More difficult to see what the results are.

The relevant statistics that support the conclusion of the hypothesis were Section 5: the data model. The p-value allowed us to reject the null hypothesis and conclude that the amount of snowfall increases as the latitude increases.

A second relevant statistic supporting the conclusion of the hypothesis was from Section 6: correlating the average temperature with latitude. It is common knowledge that at the Equator (latitude 0) the temperature is extremely high. As one travels away from the Equator, the average temperature decreases. The r value showed that the latitude and temperature are negatively correlated; meaning that as the latitude increase, the temperature decreases. This was more clearly shown in Figure 1 with the blue line.

The analysis is appropriate for my data-set because snowfall prediction is an important, not only for forecasting reasons but considering where large amounts of water naturally occur that are released slowly over time (snow melt). Climate change is altering weather patterns, and increasing temperatures, and establishing baselines and convincing data is important to building sustainable plans on how the resources we have are being utilized, how much need to be set aside for the ecosystems that currently exist, and if there is a possible change from the trends of the past to the present.

In checking my assumptions the distribution for snowfall is skewed. In spite of this, the data set has met the assumptions of my statistical tests.

What my outcome means is that there is more snow falling above the 40th parallel than below it. This needs to be taken into consideration for natural resource planning, ecosystem preservation, urban planning, farming, etc. This is an essential guide post for people who want to ask questions with implications that rely on latitude and snowfall.

The implications of the relationship are that is is far more likely that one will find snow accumulated above the 40th parallel when compared to below the 40th parallel. What this means in the real world is that people need to consider geography when taking into consideration conditions they are to be living in. This analysis could have impacts on land use, urban planning, ecosystem and resource management.

Limitations

I cannot conclude that there is no snowfall below the 40th parallel, or that there may be other factors involved in higher snowfall accumulation.

Conclusions

The main takeaway from this study is that snowfall is more likely at or above the 40th parallel. Is it impossible to find it elsewhere? No, but less likely than above the 40th parallel.

```
#|label: Citations & References
#|code_summary: used citation() to get information to add to references.bib
citation("tidyverse")
```

```
To cite package 'tidyverse' in publications use:

  Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R,
  Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller
  E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V,
  Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). "Welcome to
  the tidyverse." _Journal of Open Source Software_, *4*(43), 1686.
  doi:10.21105/joss.01686 <https://doi.org/10.21105/joss.01686>.

A BibTeX entry for LaTeX users is

  @Article{,
    title = {Welcome to the {tidyverse}},
    author = {Hadley Wickham and Mara Averick and Jennifer Bryan and Winston Chang and Lucy
    year = {2019},
```

```
    journal = {Journal of Open Source Software},
    volume = {4},
    number = {43},
    pages = {1686},
    doi = {10.21105/joss.01686},
  }
```

```
citation("lme4")
```

```
To cite lme4 in publications use:

  Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker (2015).
  Fitting Linear Mixed-Effects Models Using lme4. Journal of
  Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.

A BibTeX entry for LaTeX users is

  @Article{,
    title = {Fitting Linear Mixed-Effects Models Using {lme4}},
    author = {Douglas Bates and Martin M{\"a}chler and Ben Bolker and Steve Walker},
    journal = {Journal of Statistical Software},
    year = {2015},
    volume = {67},
    number = {1},
    pages = {1--48},
    doi = {10.18637/jss.v067.i01},
  }
```

```
citation("lmerTest")
```

```
To cite lmerTest in publications use:

  Kuznetsova A, Brockhoff PB, Christensen RHB (2017). "lmerTest
  Package: Tests in Linear Mixed Effects Models." _Journal of
  Statistical Software_, *82*(13), 1-26. doi:10.18637/jss.v082.i13
  <https://doi.org/10.18637/jss.v082.i13>.

A BibTeX entry for LaTeX users is
```

```
  @Article{,
    title = {{lmerTest} Package: Tests in Linear Mixed Effects Models},
    author = {Alexandra Kuznetsova and Per B. Brockhoff and Rune H. B. Christensen},
    journal = {Journal of Statistical Software},
    year = {2017},
    volume = {82},
    number = {13},
    pages = {1--26},
    doi = {10.18637/jss.v082.i13},
  }
```

```
citation("dplyr")
```

```
To cite package 'dplyr' in publications use:

  Wickham H, François R, Henry L, Müller K, Vaughan D (2023). _dplyr: A
  Grammar of Data Manipulation_. R package version 1.1.4,
  <https://CRAN.R-project.org/package=dplyr>.

A BibTeX entry for LaTeX users is

  @Manual{,
    title = {dplyr: A Grammar of Data Manipulation},
    author = {Hadley Wickham and Romain François and Lionel Henry and Kirill Müller and Davis
    year = {2023},
    note = {R package version 1.1.4},
    url = {https://CRAN.R-project.org/package=dplyr},
  }
```

```
#citation("ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/readme.txt")
#maybe try adding a citation using zotero or something else...?
```


[@ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/readme.txt]

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. "Fitting Linear Mixed-Effects Models Using**lme4**." *Journal of Statistical Software* 67 (1). https://doi.org/10.18637/jss.v067.i01.

Kuznetsova, Alexandra, Per B. Brockhoff, and Rune H. B. Christensen. 2017. "**lmerTest** Package: Tests in Linear Mixed Effects Models." *Journal of Statistical Software* 82 (13). https://doi.org/10.18637/jss.v082.i13.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the Tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. "Dplyr: A Grammar of Data Manipulation." https://CRAN.R-project.org/package=dplyr.