Data Wrangling Project

Raymond Ogunjimi

2022-04-18

OVERVIEW

Answer the questions below. The datasets can be found on github and our google drive folder.

This project is due on April 17. Please submit your Rmd file on Blackboard.

TEXT WRANGLING

For the following questions, use `grepl()` function. This function can find patterns in text.

1. Read in the Hamster data. Are all of the hamsters named?
    a. Create a new column called `hname` searches for "NAME" in the `text` column and creates the value `TRUE` if it is there and `FALSE` if not
    b. Create a summary variable named `hamster_name` that sums and sorts the count each value in `hname` from highest count to lowest count
    c. Output `hamster_name`
    d. How many posts do not include the name of the hamster?
2. How many tagged users come from instagram? (hint: use the technique from the last problem)

```
hamster_data <- read.csv("/Users/ray/Downloads/mini_project/hamster_data.csv")
#hamster_data %>% head()

FALSE %in% grepl("(?<=HAMSTER NAME: )(.*?)(?=\\n)", hamster_data$text, ignore.case=TRUE, perl = TRUE)
```

```
## [1] TRUE
```

```
#is.element(FALSE, grepl("(?<=HAMSTER NAME: )(.*?)(?=\\n)", hamster_data$text, ignore.case=TRUE, perl =
#match(FALSE, grepl("(?<=HAMSTER NAME: )(.*?)(?=\\n)", hamster_data$text, ignore.case=TRUE, perl = TRUE
#which(grepl("(?<=HAMSTER NAME: )(.*?)(?=\\n)", hamster_data$text, ignore.case=TRUE, perl = TRUE), FALS

hamster_data <- hamster_data %>% mutate(hname = grepl("(NAME)", hamster_data$text, ignore.case=TRUE))

sum(hamster_data$hname == FALSE)
```

```
## [1] 181
```

```
sum(hamster_data$hname == TRUE)
```

```
## [1] 819
```

```
hamster_name_data <- hamster_data %>%
    group_by(user_id, hname) %>%
    summarise(hamster_name = sum(hname == TRUE)) %>%
    arrange(desc(hamster_name))
```

```
## `summarise()` has grouped output by 'user_id'. You can override using the `.groups` argument.
```

```
hamster_data <- hamster_data %>% mutate(hamster_insta = grepl("(instagram)", hamster_data$text, ignore.

sum(hamster_data$hamster_insta == FALSE)
```

```
## [1] 689
```

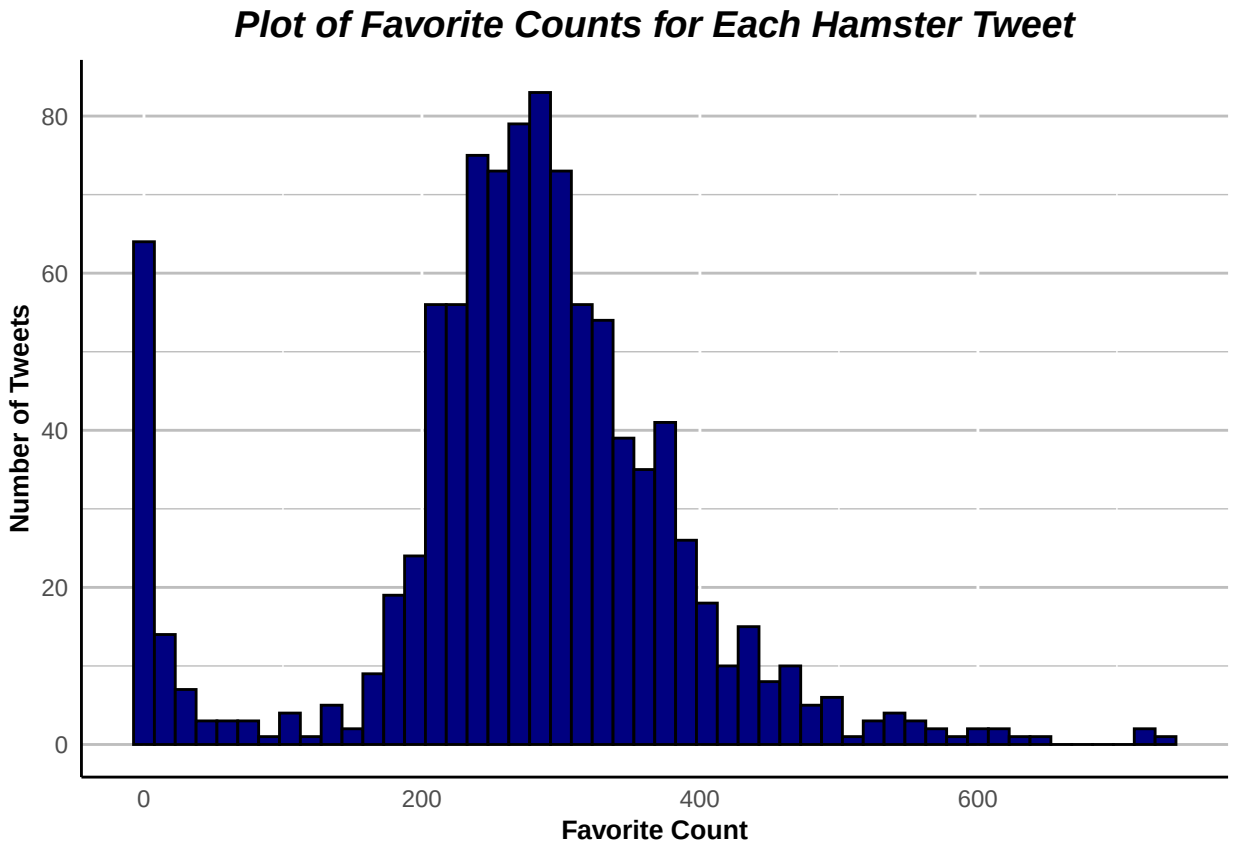
```
sum(hamster_data$hamster_insta == TRUE)
```

```
## [1] 311
```

DISTRIBUTIONS

3. Create a histogram of favorite counts.
    a. Is the standard amount of bins appropriate? If not, assign the appropriate number of bins.
    b. Make the outline of the bars black and the color of the bars navy
    c. Ensure that the background of the graph is clear (or white)
    d. Remove any tick marks
    e. Remove any vertical grid lines. The horizontal grid lines should be light gray.
    f. Make sure the axes are labeled nicely
    g. Make sure there is a title on the plot
    h. Describe the distribution

```
ggplot(hamster_data) +
    scale_x_continuous() +
    scale_y_continuous() +
    aes(x=favorite_count) +
    geom_histogram(binwidth = 15, fill="navy", color="black") +
    ggtitle("Plot of Favorite Counts for Each Hamster Tweet") +
    xlab("Favorite Count") +
    ylab("Number of Tweets") +
    theme(plot.background = element_rect(fill = "white"),
        panel.background = element_rect(fill = "white"),
        axis.ticks.x=element_blank(),
        axis.ticks.y=element_blank(),
        panel.grid.minor.y=element_line(colour="gray"),
        panel.grid.major.y=element_line(colour="gray"),
        axis.line = element_line(size = 0.5, linetype = "solid", colour = "black"),
        plot.title = element_text(color="black", size=14, face="bold.italic", hjust=0.5),
        axis.title.x = element_text(color="black", size=10, face="bold"),
        axis.title.y = element_text(color="black", size=10, face="bold"))
```



The distribution looks like a bell curve centered around about 250-300 retweets for most posts (80 posts in a bin of size 15) with a lot of posts also having no (0) retweets at all.

4. Which hamster got the most favorites?

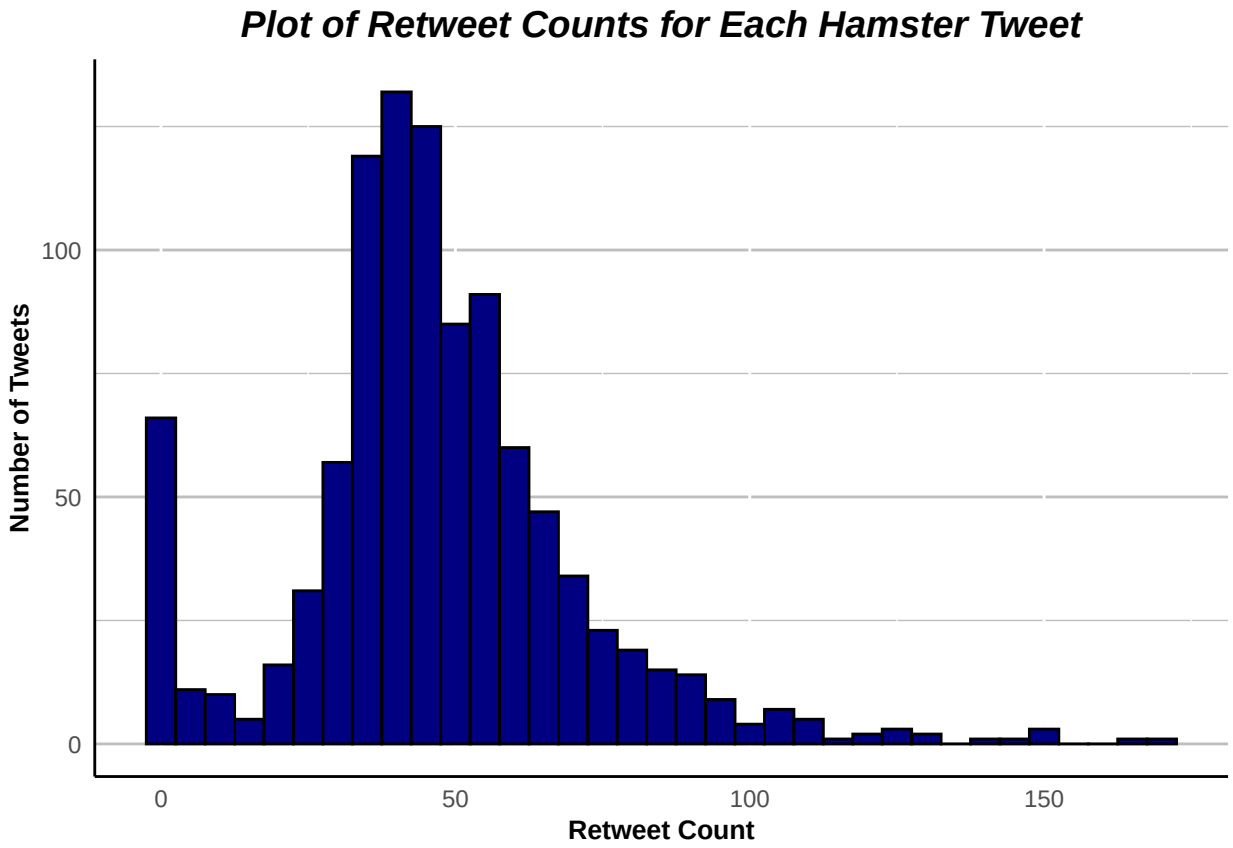
```
fav_sort_hamster <- hamster_data %>% arrange(desc(favorite_count))

regmatches(fav_sort_hamster$text,regexpr("(?<=HAMSTER NAME: )(.*?)(?=\\n)", fav_sort_hamster$text, perl
```

```
## [1] "Wolf"
```

5. Create a histogram of retweet counts.
    a. Is the standard amount of bins appropriate? If not, assign the appropriate number of bins.
    b. Make the outline of the bars white and the color of the bars purple
    c. Ensure that the background of the graph is light gray
    d. Change the color of any tick marks to purple
    e. Remove any vertical grid lines. The horizontal grid lines should be white.
    f. Make sure the axes are labeled nicely
    g. Make sure there is a title on the plot
    h. Describe the distribution

```
ggplot(hamster_data) +
    scale_x_continuous() +
    scale_y_continuous() +
    aes(x=retweet_count) +
    geom_histogram(binwidth = 5, fill="navy", color="black") +
    ggtitle("Plot of Retweet Counts for Each Hamster Tweet") +
    xlab("Retweet Count") +
    ylab("Number of Tweets") +
```

```
    theme(plot.background = element_rect(fill = "white"),
        panel.background = element_rect(fill = "white"),
        axis.ticks.x=element_blank(),
        axis.ticks.y=element_blank(),
        panel.grid.minor.y=element_line(colour="gray"),
        panel.grid.major.y=element_line(colour="gray"),
        axis.line = element_line(size = 0.5, linetype = "solid", colour = "black"),
        plot.title = element_text(color="black", size=14, face="bold.italic", hjust=0.5),
        axis.title.x = element_text(color="black", size=10, face="bold"),
        axis.title.y = element_text(color="black", size=10, face="bold"))
```



The distribution looks like a bell curve centered around about 50 retweets for most posts (150 posts in a bin of size 5) with a lot of posts also having no (0) retweets at all.

6. Which hamster got the most retweets?

```
rt_sort_hamster <- hamster_data %>% arrange(desc(retweet_count))
regmatches(rt_sort_hamster$text,regexpr("(?<=HAMSTER NAME: )(.*?)(?=\\n)", rt_sort_hamster$text, perl =
```

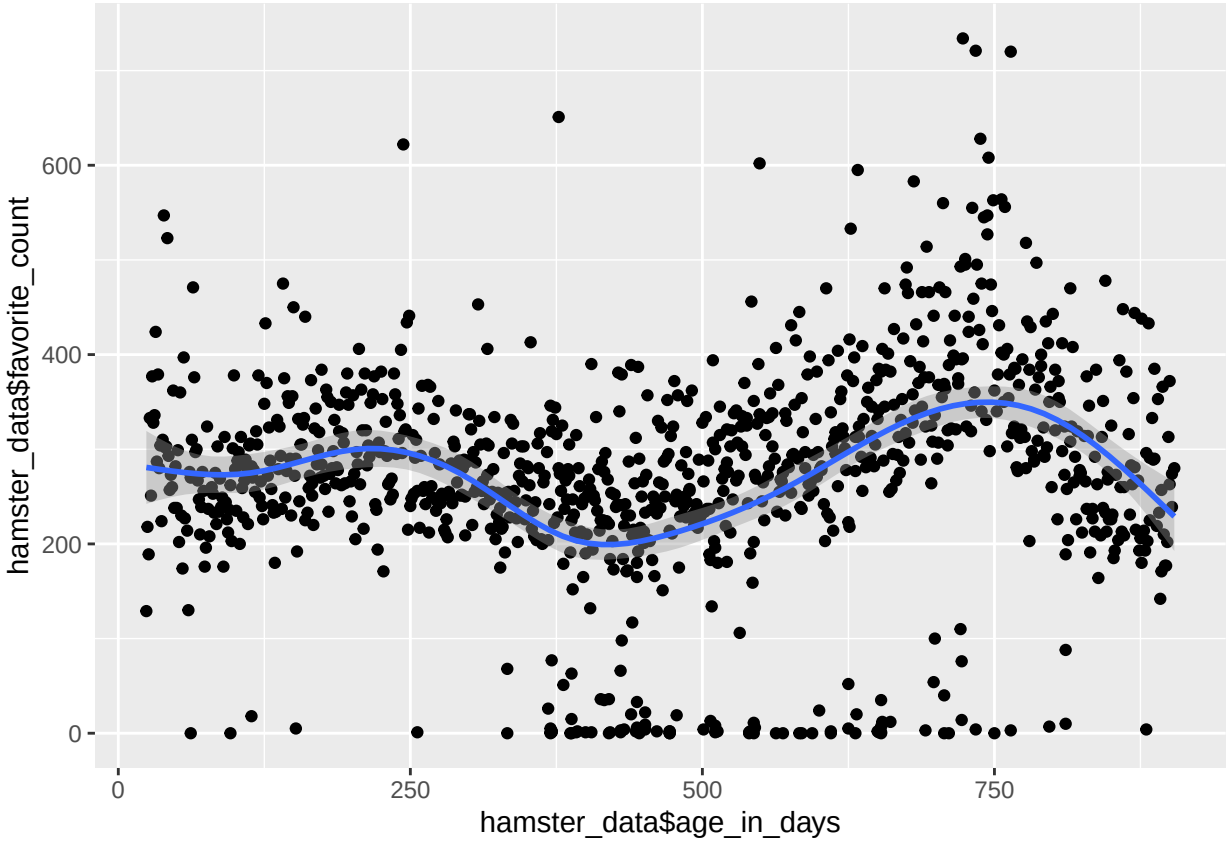
```
## [1] "Theodore "
```

CORRELATION

7. Is there a correlation between age of tweet (in days) and favorite count? You will need to create a new column that counts the age of the tweet. I am asking if the longer a tweet is up, does it get more likes? Or is it just a standard 1 day of visibility or something like that? (hint: remember the `lubridate` package)

```
hamster_data <- hamster_data %>%
    mutate(age_in_days = interval(ymd_hms(hamster_data$created_at),now()) %/% days(1))


ggplot() +
    aes(x = hamster_data$age_in_days, y = hamster_data$favorite_count)+
    geom_point()+
    geom_smooth(method = NULL)
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



There doesn't seem to be any obvious correlation. The favorite counts are about evenly distributed over all ages.

OTHER DATA MINING

8. What is the first day in the dataset? What is the last day in the dataset?

```
hamster_data %>%
    arrange(ymd_hms(hamster_data$created_at)) %>%
    head(1)["created_at"]
```

```
## [1] "2019-10-27 17:35:45"
```

```
#hamster_data %>%
#    arrange(desc(ymd_hms(hamster_data$created_at))) %>%
#    head(1) %>%
#    select("created_at")
```

```
hamster_data %>%
    arrange(desc(ymd_hms(hamster_data$created_at))) %>%
    head(1) %>%
    pluck("created_at")
```

```
## [1] "2022-03-25 09:54:00"
```

9. Are there any days with multiple posts? How many days?

```
hamster_data %>%
    group_by(date(ymd_hms(hamster_data$created_at))) %>%
    summarize(count = n()) %>%
    arrange(desc(count)) %>%
    count(count > 1)
```

```
## # A tibble: 2 x 2
##   `count > 1`     n
##   <lgl>       <int>
## 1 FALSE         788
## 2 TRUE           89
```

WEATHER DATA

10. Read in the weather data.
    a. What are the first and last day in the dataset?
    b. How many weather stations are there?
    c. Do all the stations have the same amount of entries?
    d. Which station has the most entries?

```
weather_data <- read.csv("/Users/ray/Downloads/mini_project/weather_data.csv")
weather_data %>% head(20)
```

```
##        STATION                    NAME LATITUDE LONGITUDE ELEVATION       DATE
## 1  US1PABK0018 LANGHORNE 2.8 NE, PA US 40.20922 -74.88604      54.9 2019-01-03
## 2  US1PABK0018 LANGHORNE 2.8 NE, PA US 40.20922 -74.88604      54.9 2019-01-04
## 3  US1PABK0018 LANGHORNE 2.8 NE, PA US 40.20922 -74.88604      54.9 2019-01-05
## 4  US1PABK0018 LANGHORNE 2.8 NE, PA US 40.20922 -74.88604      54.9 2019-01-06
## 5  US1PABK0018 LANGHORNE 2.8 NE, PA US 40.20922 -74.88604      54.9 2019-01-07
## 6  US1PABK0018 LANGHORNE 2.8 NE, PA US 40.20922 -74.88604      54.9 2019-01-08
## 7  US1PABK0018 LANGHORNE 2.8 NE, PA US 40.20922 -74.88604      54.9 2019-01-09
## 8  US1PABK0018 LANGHORNE 2.8 NE, PA US 40.20922 -74.88604      54.9 2019-01-10
## 9  US1PABK0018 LANGHORNE 2.8 NE, PA US 40.20922 -74.88604      54.9 2019-01-11
## 10 US1PABK0018 LANGHORNE 2.8 NE, PA US 40.20922 -74.88604      54.9 2019-01-12
## 11 US1PABK0018 LANGHORNE 2.8 NE, PA US 40.20922 -74.88604      54.9 2019-01-13
## 12 US1PABK0018 LANGHORNE 2.8 NE, PA US 40.20922 -74.88604      54.9 2019-01-16
## 13 US1PABK0018 LANGHORNE 2.8 NE, PA US 40.20922 -74.88604      54.9 2019-01-17
## 14 US1PABK0018 LANGHORNE 2.8 NE, PA US 40.20922 -74.88604      54.9 2019-01-23
## 15 US1PABK0018 LANGHORNE 2.8 NE, PA US 40.20922 -74.88604      54.9 2019-01-24
## 16 US1PABK0018 LANGHORNE 2.8 NE, PA US 40.20922 -74.88604      54.9 2019-01-25
## 17 US1PABK0018 LANGHORNE 2.8 NE, PA US 40.20922 -74.88604      54.9 2019-01-26
## 18 US1PABK0018 LANGHORNE 2.8 NE, PA US 40.20922 -74.88604      54.9 2019-01-27
## 19 US1PABK0018 LANGHORNE 2.8 NE, PA US 40.20922 -74.88604      54.9 2019-01-28
## 20 US1PABK0018 LANGHORNE 2.8 NE, PA US 40.20922 -74.88604      54.9 2019-01-30
##    AWND AWND_ATTRIBUTES DAPR DAPR_ATTRIBUTES DASF DASF_ATTRIBUTES MDPR
## 1    NA                 13             ,,N   NA                2.54
```

```
## 2    NA                  NA                  NA                  NA
## 3    NA                  NA                  NA                  NA
## 4    NA                  NA                  NA                  NA
## 5    NA                  NA                  NA                  NA
## 6    NA                  NA                  NA                  NA
## 7    NA                  NA                  NA                  NA
## 8    NA                  NA                  NA                  NA
## 9    NA                  NA                  NA                  NA
## 10   NA                  NA                  NA                  NA
## 11   NA                  NA                  NA                  NA
## 12   NA                  NA                  NA                  NA
## 13   NA                  NA                  NA                  NA
## 14   NA                   6             ,,N   NA                1.54
## 15   NA                  NA                  NA                  NA
## 16   NA                  NA                  NA                  NA
## 17   NA                  NA                  NA                  NA
## 18   NA                  NA                  NA                  NA
## 19   NA                  NA                  NA                  NA
## 20   NA                  NA                  NA                  NA
##    MDPR_ATTRIBUTES MDSF MDSF_ATTRIBUTES PGTM PGTM_ATTRIBUTES PRCP
## 1              ,,N   NA                  NA                  NA
## 2                    NA                  NA                0.00
## 3                    NA                  NA                0.34
## 4                    NA                  NA                0.17
## 5                    NA                  NA                0.00
## 6                    NA                  NA                0.09
## 7                    NA                  NA                0.19
## 8                    NA                  NA                0.00
## 9                    NA                  NA                0.00
## 10                   NA                  NA                0.00
## 11                   NA                  NA                0.06
## 12                   NA                  NA                0.00
## 13                   NA                  NA                0.00
## 14             ,,N   NA                  NA                  NA
## 15                   NA                  NA                0.15
## 16                   NA                  NA                0.94
## 17                   NA                  NA                0.00
## 18                   NA                  NA                0.00
## 19                   NA                  NA                0.00
## 20                   NA                  NA                0.24
##    PRCP_ATTRIBUTES SNOW SNOW_ATTRIBUTES SNWD SNWD_ATTRIBUTES TAVG
## 1                    NA                  NA                  NA
## 2              ,,N   NA                  NA                  NA
## 3              ,,N   NA                  NA                  NA
## 4              ,,N   NA                  NA                  NA
## 5              ,,N  0.0             ,,N   NA                  NA
## 6              ,,N   NA                  NA                  NA
## 7              ,,N   NA                  NA                  NA
## 8              ,,N   NA                  NA                  NA
## 9              ,,N   NA                  NA                  NA
## 10             ,,N  0.0             ,,N   NA                  NA
## 11             ,,N  0.5             ,,N   NA                  NA
## 12             ,,N   NA                  NA                  NA
## 13             ,,N   NA                  NA                  NA
```

```
## 14                    NA                    NA                    NA
## 15              ,,N   NA                    NA                    NA
## 16              ,,N   NA                    NA                    NA
## 17              ,,N  0.0              ,,N   NA                    NA
## 18              ,,N  0.0              ,,N   NA                    NA
## 19              ,,N   NA                    NA                    NA
## 20              ,,N   NA                    NA                    NA
##    TAVG_ATTRIBUTES TMAX TMAX_ATTRIBUTES TMIN TMIN_ATTRIBUTES TOBS
## 1                     NA                    NA                    NA
## 2                     NA                    NA                    NA
## 3                     NA                    NA                    NA
## 4                     NA                    NA                    NA
## 5                     NA                    NA                    NA
## 6                     NA                    NA                    NA
## 7                     NA                    NA                    NA
## 8                     NA                    NA                    NA
## 9                     NA                    NA                    NA
## 10                    NA                    NA                    NA
## 11                    NA                    NA                    NA
## 12                    NA                    NA                    NA
## 13                    NA                    NA                    NA
## 14                    NA                    NA                    NA
## 15                    NA                    NA                    NA
## 16                    NA                    NA                    NA
## 17                    NA                    NA                    NA
## 18                    NA                    NA                    NA
## 19                    NA                    NA                    NA
## 20                    NA                    NA                    NA
##    TOBS_ATTRIBUTES WDF2 WDF2_ATTRIBUTES WDF5 WDF5_ATTRIBUTES WESD
## 1                     NA                    NA                    NA
## 2                     NA                    NA                    NA
## 3                     NA                    NA                    NA
## 4                     NA                    NA                    NA
## 5                     NA                    NA                    NA
## 6                     NA                    NA                    NA
## 7                     NA                    NA                    NA
## 8                     NA                    NA                    NA
## 9                     NA                    NA                    NA
## 10                    NA                    NA                    NA
## 11                    NA                    NA                    NA
## 12                    NA                    NA                    NA
## 13                    NA                    NA                    NA
## 14                    NA                    NA                    NA
## 15                    NA                    NA                    NA
## 16                    NA                    NA                    NA
## 17                    NA                    NA                    NA
## 18                    NA                    NA                    NA
## 19                    NA                    NA                    NA
## 20                    NA                    NA                    NA
##    WESD_ATTRIBUTES WESF WESF_ATTRIBUTES WSF2 WSF2_ATTRIBUTES WSF5
## 1                     NA                    NA                    NA
## 2                     NA                    NA                    NA
## 3                     NA                    NA                    NA
## 4                     NA                    NA                    NA
```

```
## 5                    NA                    NA                    NA
## 6                    NA                    NA                    NA
## 7                    NA                    NA                    NA
## 8                    NA                    NA                    NA
## 9                    NA                    NA                    NA
## 10                   NA                    NA                    NA
## 11                   NA                    NA                    NA
## 12                   NA                    NA                    NA
## 13                   NA                    NA                    NA
## 14                   NA                    NA                    NA
## 15                   NA                    NA                    NA
## 16                   NA                    NA                    NA
## 17                   NA                    NA                    NA
## 18                   NA                    NA                    NA
## 19                   NA                    NA                    NA
## 20                   NA                    NA                    NA
##    WSF5_ATTRIBUTES WT01 WT01_ATTRIBUTES WT02 WT02_ATTRIBUTES WT03
## 1                    NA                    NA                    NA
## 2                    NA                    NA                    NA
## 3                    NA                    NA                    NA
## 4                    NA                    NA                    NA
## 5                    NA                    NA                    NA
## 6                    NA                    NA                    NA
## 7                    NA                    NA                    NA
## 8                    NA                    NA                    NA
## 9                    NA                    NA                    NA
## 10                   NA                    NA                    NA
## 11                   NA                    NA                    NA
## 12                   NA                    NA                    NA
## 13                   NA                    NA                    NA
## 14                   NA                    NA                    NA
## 15                   NA                    NA                    NA
## 16                   NA                    NA                    NA
## 17                   NA                    NA                    NA
## 18                   NA                    NA                    NA
## 19                   NA                    NA                    NA
## 20                   NA                    NA                    NA
##    WT03_ATTRIBUTES WT04 WT04_ATTRIBUTES WT05 WT05_ATTRIBUTES WT06
## 1                    NA                    NA                    NA
## 2                    NA                    NA                    NA
## 3                    NA                    NA                    NA
## 4                    NA                    NA                    NA
## 5                    NA                    NA                    NA
## 6                    NA                    NA                    NA
## 7                    NA                    NA                    NA
## 8                    NA                    NA                    NA
## 9                    NA                    NA                    NA
## 10                   NA                    NA                    NA
## 11                   NA                    NA                    NA
## 12                   NA                    NA                    NA
## 13                   NA                    NA                    NA
## 14                   NA                    NA                    NA
## 15                   NA                    NA                    NA
## 16                   NA                    NA                    NA
```

```
## 17                   NA                   NA                   NA
## 18                   NA                   NA                   NA
## 19                   NA                   NA                   NA
## 20                   NA                   NA                   NA
##    WT06_ATTRIBUTES WT08 WT08_ATTRIBUTES WT09 WT09_ATTRIBUTES WT11
## 1                   NA                   NA                   NA
## 2                   NA                   NA                   NA
## 3                   NA                   NA                   NA
## 4                   NA                   NA                   NA
## 5                   NA                   NA                   NA
## 6                   NA                   NA                   NA
## 7                   NA                   NA                   NA
## 8                   NA                   NA                   NA
## 9                   NA                   NA                   NA
## 10                  NA                   NA                   NA
## 11                  NA                   NA                   NA
## 12                  NA                   NA                   NA
## 13                  NA                   NA                   NA
## 14                  NA                   NA                   NA
## 15                  NA                   NA                   NA
## 16                  NA                   NA                   NA
## 17                  NA                   NA                   NA
## 18                  NA                   NA                   NA
## 19                  NA                   NA                   NA
## 20                  NA                   NA                   NA
##    WT11_ATTRIBUTES
## 1
## 2
## 3
## 4
## 5
## 6
## 7
## 8
## 9
## 10
## 11
## 12
## 13
## 14
## 15
## 16
## 17
## 18
## 19
## 20
```

```
weather_data <- subset(weather_data, !is.na(TAVG))

weather_data %>%
    arrange(ymd(weather_data$DATE)) %>%
    head(1)["DATE"]
```

```
## [1] "2019-01-01"
```

```
weather_data %>%
    arrange(desc(ymd(weather_data$DATE))) %>%
    head(1)["DATE"]
```

```
## [1] "2022-03-23"
```

```
length(unique(weather_data$NAME))
```

```
## [1] 2
```

```
weather_data %>%
    group_by(NAME) %>%
    tally()
```

```
## # A tibble: 2 x 2
##   NAME                                          n
##   <chr>                                     <int>
## 1 PHILADELPHIA INTERNATIONAL AIRPORT, PA US  1178
## 2 WILMINGTON NEW CASTLE CO AIRPORT, DE US     632
```

```
#weather_data %>% count(NAME)
```

JOINING DATA

11. Join the hamster data to the weather data for the top station from the previous problem (if there are any ties, pick your favorite). What issues do you have and what choices did you make to join this data together?

```
hamster_joinable <- hamster_data %>% mutate(join_date = date(ymd_hms(hamster_data$created_at)))
weather_joinable <- weather_data %>% mutate(join_date = date(ymd(weather_data$DATE)))
joined <- merge(x=weather_joinable,y=hamster_joinable,by="join_date")
joined <- within(joined, rm("DATE","created_at"))
```

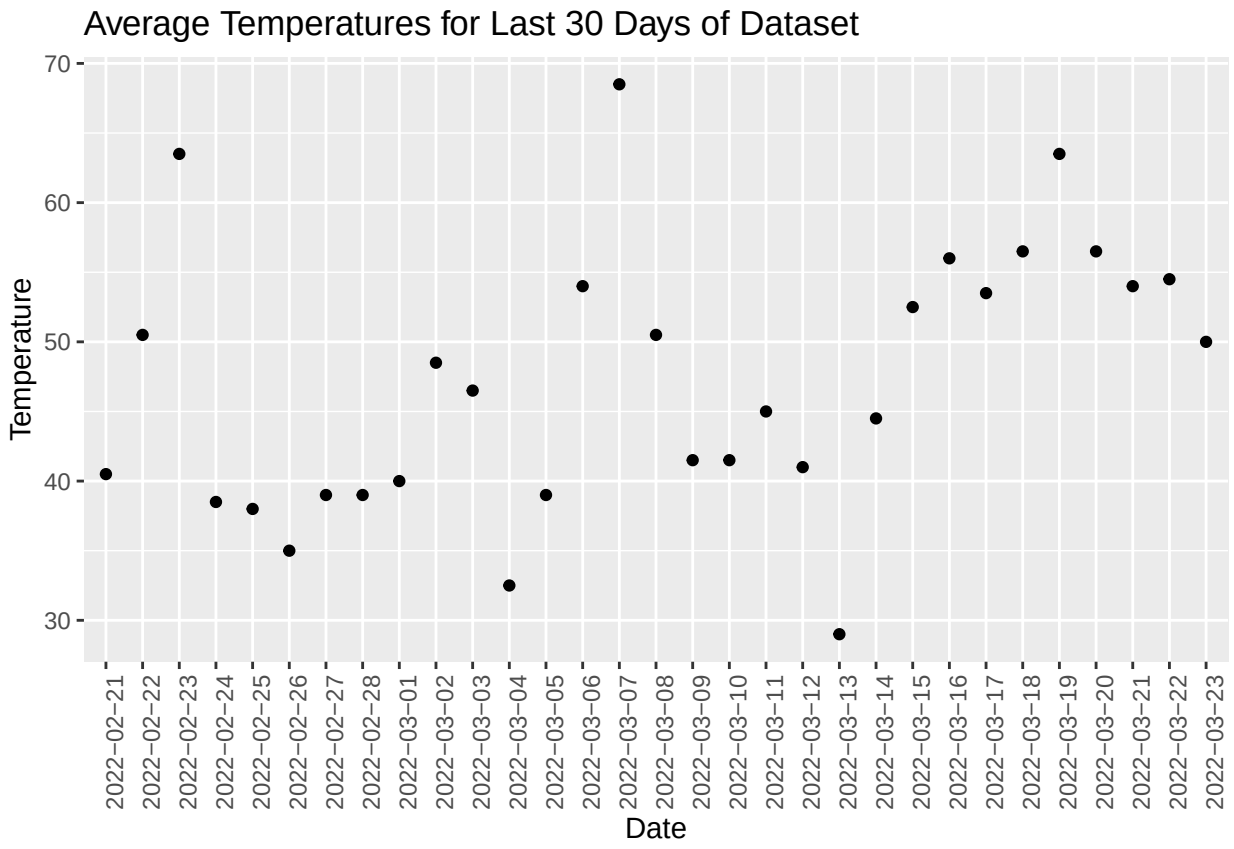
The type of join was one choice (date) and the tie to join by was another (inner/natural).

PLOTS AND LABELS

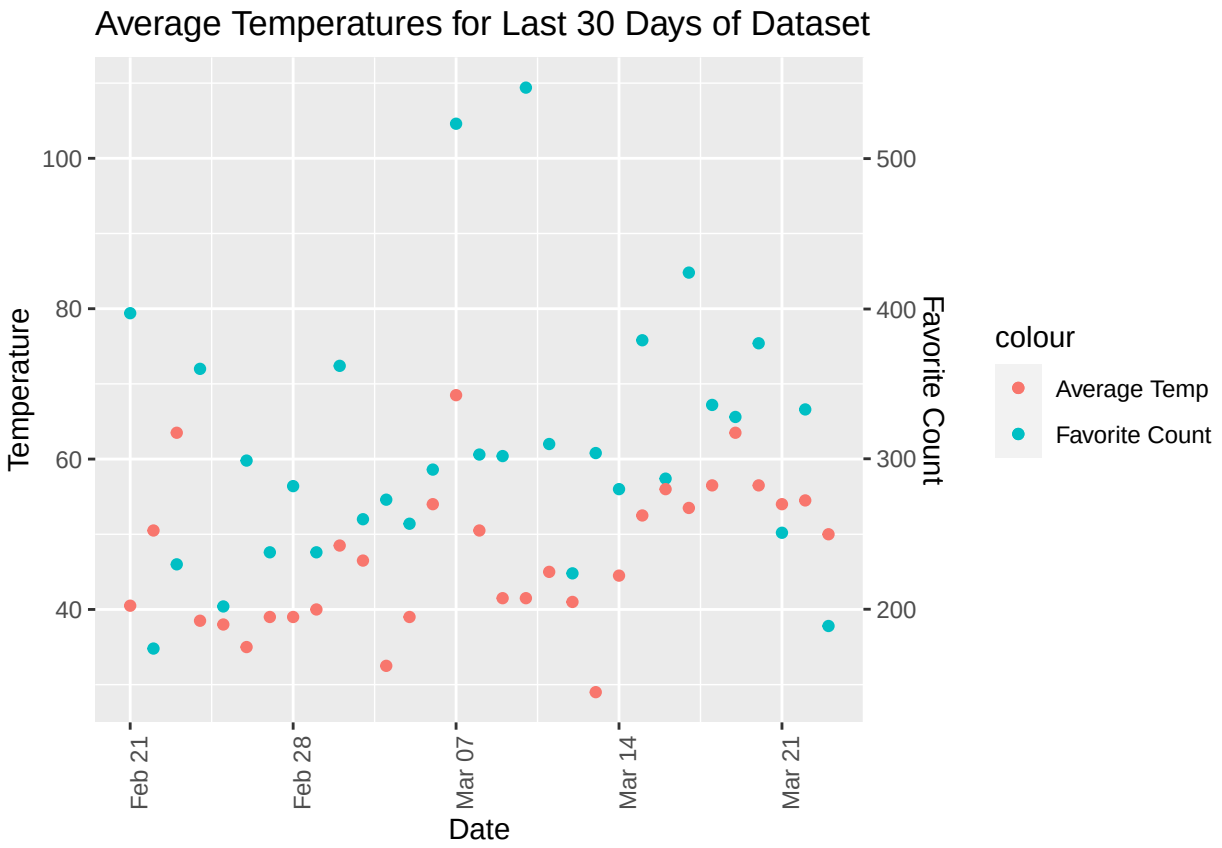
12. Plot the daily average temperature from the last 30 days of the dataset. Create a data label that adds the favorite count of the daily hamster to the plot (if there are multiple, you may choose one or paste all together).

```
weather_data_30 <- weather_data %>%
    mutate(days_since_last = interval(ymd(weather_data$DATE),max(ymd(weather_data$DATE))) %/% days(1)) %
    arrange(DATE)

weather_data_30_toplot <- subset(weather_data_30, days_since_last <= 30)

weather_data_30_toplot <- aggregate(x=weather_data_30_toplot$TAVG, by=list(weather_data_30_toplot$DATE)
#weather_data_30_toplot <- aggregate(weather_data_30_toplot$TAVG ~ weather_data_30_toplot$DATE, weather

ggplot(weather_data_30_toplot) +
    aes(x=Group.1, y = x) +
    geom_point() +
    ggtitle("Average Temperatures for Last 30 Days of Dataset") +
    xlab("Date") +
```

```
    ylab("Temperature") +
    theme(axis.text.x = element_text(angle = 90))
```




```
weather_data_30_wit_hams <- joined %>%
    mutate(days_since_last = interval(ymd(joined$join_date),max(ymd(joined$join_date))) %/% days(1)) %>%
    arrange(join_date)

weather_data_30_wit_hams <- subset(weather_data_30_wit_hams, days_since_last <= 30)

weather_data_30_wit_hams <- setNames(aggregate(x=list(weather_data_30_wit_hams$TAVG,weather_data_30_wit_

ggplot(weather_data_30_wit_hams) +
    geom_point(aes(x = join_date, y = fav_count/5, color = "Favorite Count")) +
    geom_point(aes(x = join_date, y = avg_temp, color = "Average Temp")) +
    scale_y_continuous(name = "Temperature", sec.axis = sec_axis( trans=~.*5, name="Favorite Count")) +
    ggtitle("Average Temperatures for Last 30 Days of Dataset") +
    xlab("Date") +
    theme(axis.text.x = element_text(angle = 90))
```



MINING THE WEATHER DATA

13. What is the average precipitation (PRCP) at "PHILADELPHIA INTERNATIONAL AIRPORT, PA US"?

```
phl_weather <- weather_data %>% filter(weather_data$NAME == "PHILADELPHIA INTERNATIONAL AIRPORT, PA US"
mean(phl_weather$PRCP, na.rm=TRUE)
```

```
## [1] 0.1261597
```

14. What is the overall average (TAVG), minimum (TMIN), and maximum (TMAX) temperature of this data at "PHILADELPHIA INTERNATIONAL AIRPORT, PA US"?

```
mean(phl_weather$TAVG, na.rm = TRUE)
```

```
## [1] 56.52207
```

```
min(phl_weather$TMIN, na.rm = TRUE)
```

```
## [1] 5
```

```
max(phl_weather$TMAX, na.rm = TRUE)
```

```
## [1] 98
```

15. What is the average, minimum, and maximum temperature from this data by month at "PHILADELPHIA INTERNATIONAL AIRPORT, PA US"?

```
phl_weather %>% group_by(month(ymd(DATE))) %>%
    summarize(month = month(ymd(DATE)),
```

```
        temp_avg = mean(TAVG, na.rm = TRUE),
        temp_min = min(TMIN, na.rm = TRUE),
        temp_max = max(TMAX, na.rm = TRUE)) %>%
    unique()
```

```
## `summarise()` has grouped output by 'month(ymd(DATE))'. You can override using the `.groups` argumen
```

```
## # A tibble: 12 x 5
## # Groups:   month(ymd(DATE)) [12]
##    `month(ymd(DATE))` month temp_avg temp_min temp_max
##                 <dbl> <dbl>    <dbl>    <int>    <int>
##  1                  1     1     35.6        5       67
##  2                  2     2     37.9       11       71
##  3                  3     3     46.9       19       83
##  4                  4     4     55.2       29       87
##  5                  5     5     63.6       35       92
##  6                  6     6     74.2       51       97
##  7                  7     7     80.2       62       98
##  8                  8     8     77.6       61       96
##  9                  9     9     71         47       92
## 10                 10    10     61.3       34       95
## 11                 11    11     47.5       23       76
## 12                 12    12     41.2       19       68
```

16. What is the average, minimum, and maximum from this data by year at "PHILADELPHIA INTERNATIONAL AIRPORT, PA US"?

```
phl_weather %>% group_by(year(ymd(DATE))) %>%
    summarize(year = year(ymd(DATE)),
        temp_avg = mean(TAVG, na.rm = TRUE),
        temp_min = min(TMIN, na.rm = TRUE),
        temp_max = max(TMAX, na.rm = TRUE)) %>%
    unique()
```

```
## `summarise()` has grouped output by 'year(ymd(DATE))'. You can override using the `.groups` argument
```

```
## # A tibble: 4 x 5
## # Groups:   year(ymd(DATE)) [4]
##   `year(ymd(DATE))`  year temp_avg temp_min temp_max
##               <dbl> <dbl>    <dbl>    <int>    <int>
## 1              2019  2019     57.2        5       98
## 2              2020  2020     58.0       14       97
## 3              2021  2021     58.1       19       97
## 4              2022  2022     39.9       11       77
```

17. What is the average, minimum, and maximum from this data by year and month at "PHILADELPHIA INTERNATIONAL AIRPORT, PA US"?

```
phl_weather %>% group_by(year(ymd(DATE)), month(ymd(DATE))) %>%
    summarize(year = year(ymd(DATE)),
        month = month(ymd(DATE)),
        temp_avg = mean(TAVG, na.rm = TRUE),
        temp_min = min(TMIN, na.rm = TRUE),
        temp_max = max(TMAX, na.rm = TRUE)) %>%
    unique()
```

```
## `summarise()` has grouped output by 'year(ymd(DATE))', 'month(ymd(DATE))'. You can override using the
```

```
## # A tibble: 39 x 7
## # Groups:   year(ymd(DATE)), month(ymd(DATE)) [39]
##    `year(ymd(DATE))` `month(ymd(DATE))`  year month temp_avg temp_min temp_max
##                <dbl>               <dbl> <dbl> <dbl>    <dbl>    <int>    <int>
##  1              2019                   1  2019     1     34.0        5       61
##  2              2019                   2  2019     2     36.8       11       67
##  3              2019                   3  2019     3     43.1       19       77
##  4              2019                   4  2019     4     58.6       32       82
##  5              2019                   5  2019     5     65.6       46       90
##  6              2019                   6  2019     6     73.5       53       94
##  7              2019                   7  2019     7     80.5       64       98
##  8              2019                   8  2019     8     76.9       61       93
##  9              2019                   9  2019     9     72.0       53       92
## 10              2019                  10  2019    10     61.0       41       95
## # ... with 29 more rows
```

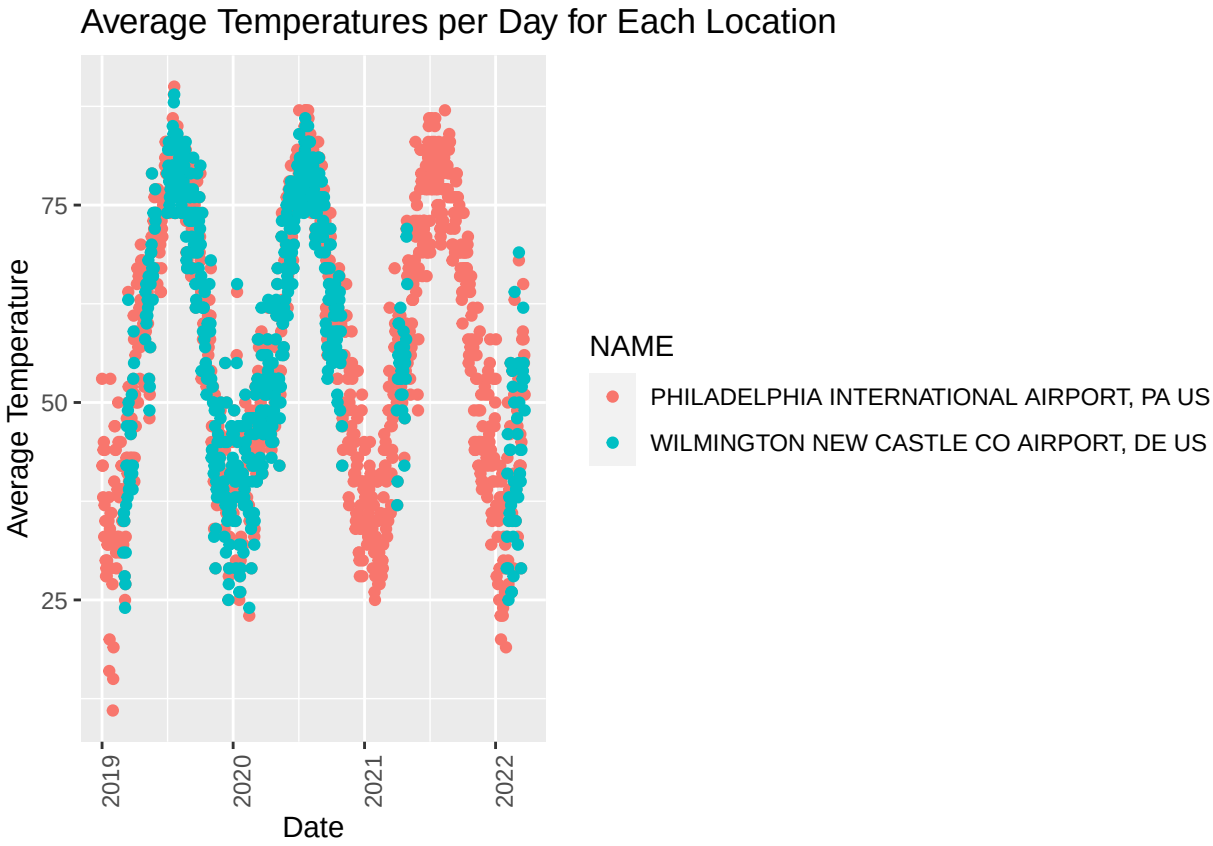
MAKING SENSE OF MANY GROUPS

18. Plot the average daily temperature for the entire dataset grouped by station. What issues do you see with the plot? Make it the clearest possible.

```
daily_temps <- weather_data %>% group_by(NAME, DATE) %>%
    summarize(temp_avg = mean(TAVG, na.rm = TRUE)) %>%
    unique()
```

```
## `summarise()` has grouped output by 'NAME'. You can override using the `.groups` argument.
```

```
ggplot(daily_temps) +
    geom_point(aes(x = ymd(DATE), y = temp_avg, color = NAME)) +
    ggtitle("Average Temperatures per Day for Each Location") +
    xlab("Date") +
    ylab("Average Temperature") +
    theme(axis.text.x = element_text(angle = 90))
```



There are a lot of data points on this plot. That is the biggest issue but overall it shows a clear trend of temperatures as expected in the DE/PA region