# Exam 02 (Open Notes)

**DSST289: Introduction to Data Science**

Erik Fredner

2024-11-10

## Table of contents

### Deadline

Monday, October 28 *before* the start of class.

### Honor Pledge

> "I pledge that I will neither give nor receive unauthorized assistance during the completion of this work."

For this exam, you may use class notes, notebooks, and slides. Any other resource (e.g., non-class websites, ChatGPT, etc.) is unauthorized.

**Signature (type your full name)**:

**UR email:**

**Section start time:**

## Setup

1. Navigate to Blackboard > Course Documents > Exams > Exam 02
2. Download this exam from Blackboard: `exam02_open.qmd`
3. Download the data from Blackboard:

- `storms.csv`
- `storm_gender.csv`
- `storm_codes.csv`

4. Move the exam to the `nb` folder in your `DSST289` folder, just as we do when working on new notebooks in class: (`...DSST289/nb/exam02_open.qmd`)
5. Move the data to `...DSST289/data/`.

## Instructions

1. Although there are multiple ways of producing the results requested in each question, I expect to see you use patterns and techniques that we have discussed.

2. If you are unable to complete a question, explain your attempt to maximize partial credit.

3. If you encounter R or RStudio errors that you cannot resolve on your own, contact me ASAP. I can help you with configuration issues, but will not help you answer questions. If you run into any issues with your personal computer, use the computers in the library to complete the exam.

4. When you have finished the exam, **render** your `.qmd` file to `.pdf`. If rendering fails, upload the `.qmd` file.

5. Go to Blackboard > Assignments > Exam 02 (open notes). Upload your **rendered** document there.

## Data: storms

The data for this exam consists of information about tropical storms in the Atlantic Ocean between 1950 and 2020.

```r
library(tidyverse)

storms <- read_csv("../data/storms.csv")

storms |>
  slice_sample(n = 5)
```

```
# A tibble: 5 x 10
   year name   letter   doy  hour   lat   lon status category  wind
  <dbl> <chr>  <chr>  <dbl> <dbl> <dbl> <dbl> <chr>     <dbl> <dbl>
1  1985 Henri  H        267     6  38.2 -74   TS            0    35
2  1965 Anna   A        235    18  37.5 -52   HU            1    75
3  1984 Diana  D        260     6  46   -57.8 EX            0    60
4  1984 Arthur A        243    18  14   -57.8 TS            0    35
5  1985 Kate   K        321     0  20.7 -66   HU            1    75
```

`storms` contains one row for each time a particular storm was measured. Storms are generally measured once every six hours.

### Features

| Variable | Description |
|---|---|
| `year` | The year in which the storm was recorded |
| `name` | A common name for the storm. Names can be reused for different storms in different years. |
| `letter` | The first letter of the name; storms are (usually) named in alphabetical order |
| `doy` | The day of the year (1-365) of the record |
| `hour` | The hour of the day (0-23) of the record in Eastern time |
| `lat` | Latitude of the record in degrees |
| `lon` | Longitude of the record in degrees |
| `status` | A two-digit status code of the storm system; see `storm_codes.csv` for full names |
| `category` | For hurricanes (`status == "HU"`), a number giving the category of the storm from 0-5 |
| `wind` | The observed sustained wind speed in miles per hour |

### Metadata

In addition to the main storms table, there are two metadata tables. storm_gender provides an automatically determined estimate of whether storm's name is male or female. Its prob column gives a confidence score for the accuracy of the gender determination. A higher score indicates a higher confidence.

```
storm_gender <- read_csv("../data/storm_gender.csv")

storm_gender |>
  filter(prob < 1) |>
  arrange(desc(prob)) |>
  slice_head(n = 3)
```

```
# A tibble: 3 x 3
  name  gender  prob
  <chr> <chr>  <dbl>
1 Anna  female 0.999
2 Grace female 0.999
3 Julia female 0.999
```

```
storm_gender |>
  arrange(desc(prob)) |>
  slice_tail(n = 3)
```

```
# A tibble: 3 x 3
  name    gender  prob
  <chr>   <chr>  <dbl>
1 Nana    female 0.688
2 Charley female 0.642
3 Joan    female 0.510
```

There is a column in storms called status that describes the type of storm with a two letter code. storm_codes provides a full name for each of these codes:

```
storm_codes <- read_csv("../data/storm_codes.csv")
storm_codes
```

```
# A tibble: 9 x 2
  status status_name
  <chr>  <chr>
1 TD     tropical depression
2 TS     tropical storm
3 HU     hurricane
4 EX     extratropical cyclone
5 SD     subtropical depression
6 SS     subtropical storm
7 LO     low
8 WV     tropical wave
9 DB     disturbance
```

## Questions

### Max wind speed over hurricane lifetime

Output a table with one row for each storm in the data set that provides the maximum wind speed the storm achieved over its lifetime.

```
storms |>
  group_by(year, name) |>
  summarize(wind_max = max(wind))
```

```
# A tibble: 761 x 3
# Groups:   year [71]
    year name    wind_max
   <dbl> <chr>      <dbl>
 1  1950 Able         110
 2  1950 Baker         90
 3  1950 Charlie       95
 4  1950 Dog          125
 5  1950 Easy         105
 6  1950 Fox          120
 7  1950 George        95
 8  1950 How           40
 9  1950 Item          90
10  1950 Jig          100
# i 751 more rows
```

### Average speed by hurricane category

Hurricanes get assigned one of six different categories based on their sustained wind speed. When a hurricane is covered on the news, you may hear it described as a “Category 3” storm, for example.

Create a new data set that has one row for each hurricane category in each year that shows the average wind speed of hurricanes in that category during that year.

> Tip
>
> Not every storm in `storms` is a hurricane.

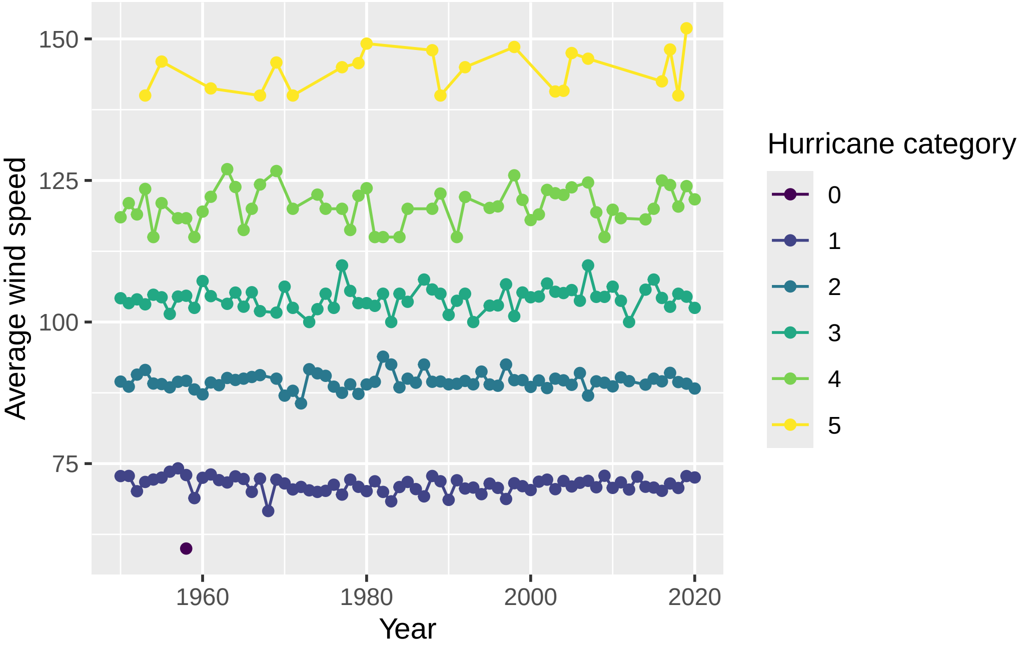
```
storms |>
  filter(status == "HU") |>
  group_by(year, category) |>
  summarize(wind_avg = mean(wind))
```

```
# A tibble: 281 x 3
# Groups:   year [71]
    year category wind_avg
   <dbl>    <dbl>    <dbl>
 1  1950        1     72.8
 2  1950        2     89.5
 3  1950        3    104.
 4  1950        4    118.
 5  1951        1     72.8
 6  1951        2     88.6
 7  1951        3    103.
 8  1951        4    121
 9  1952        1     70.1
10  1952        2     90.7
# i 271 more rows
```

Using the table you just created, create a line plot with a points layer showing the average wind speed over time by hurricane category. Color the points and lines by hurricane category using a colorblind-friendly scale. Label the axes and legend.

```
storms |>
  filter(status == "HU") |>
  group_by(year, category) |>
  summarize(wind_avg = mean(wind)) |>
```

```
ggplot(aes(year, wind_avg, color = as_factor(category))) +
geom_line() +
geom_point() +
scale_color_viridis_d() +
labs(
  x = "Year",
  y = "Average wind speed",
  color = "Hurricane category"
)
```



### Days of the year with midnight hurricanes

Create a new table where the unit of observation is the day of the year. Count the total number of hurricanes observed at midnight on each day of the year.

Tip

There are days of the year without hurricanes observed at midnight. You do **not** need rows for those days.

```
storms |>
  filter(status == "HU", hour == 0) |>
  count(doy)
```

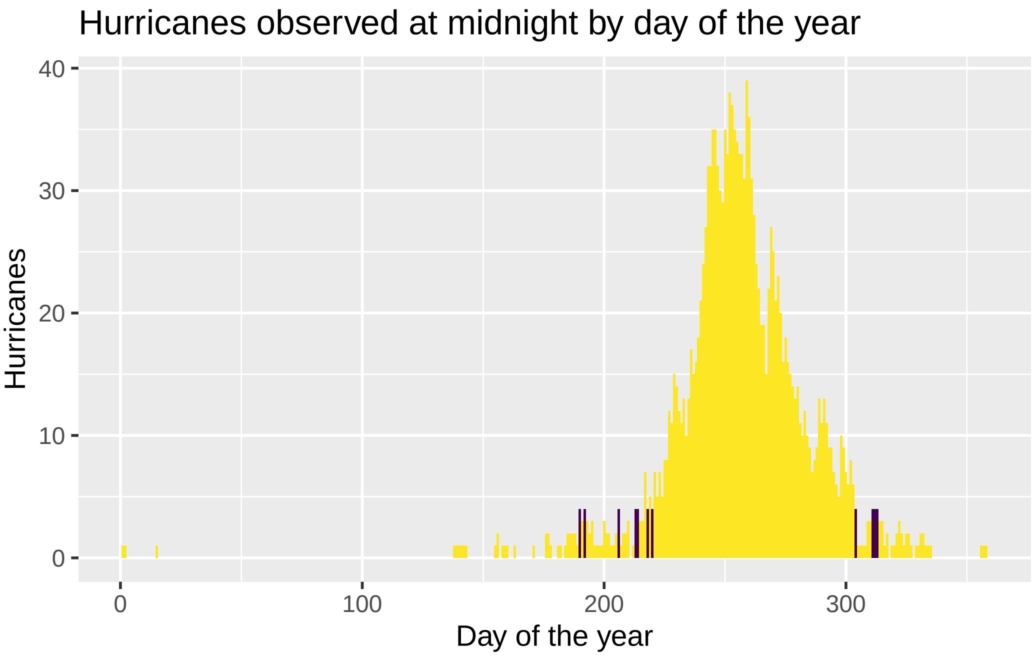
```
# A tibble: 173 x 2
     doy     n
   <dbl> <int>
 1     1     1
 2     2     1
 3    15     1
 4   138     1
 5   139     1
 6   140     1
 7   141     1
 8   142     1
 9   143     1
10   155     1
# i 163 more rows
```

Using the table you just created, make a bar plot that shows the number of hurricanes at midnight on each day of the year, with the day of the year on the x-axis and the number of hurricanes observed at midnight on that day on the y-axis.

Add a layer on top of that bar plot that highlights the days of the year with the *median* number of hurricanes observed at midnight. Color the bars for the median days `"#440154"` and the non-median days `"#fde725"`. Label the axes and title the plot.

```
median_days <- storms |>
  filter(status == "HU", hour == 0) |>
  count(doy) |>
  filter(n == median(n))

storms |>
  filter(status == "HU", hour == 0) |>
  count(doy) |>
  ggplot(aes(doy, n)) +
  geom_col(fill = "#fde725") +
  geom_col(data = median_days, fill = "#440154") +
  labs(
    x = "Day of the year",
    y = "Hurricanes",
    title = "Hurricanes observed at midnight by day of the year"
  )
```



#### Last letter of the year

Storms are named in alphabetical order, with the first storm of the year starting with the letter A, the second with the letter B, and so on.

Output a table with two columns: `letter` and `n`. `n` should indicate the number of years in which each letter was the *last* letter used to name a storm in that year. For example, in 1972 and 1983, the last storm of the year started with "D."

> **ℹ Note**
>
> The data has been filtered to exclude some storms, such as those with Greek letters, so do not expect these results to exactly match other sources.

```
storms |>
  group_by(year) |>
  arrange(desc(letter)) |>
  slice(1) |>
  ungroup() |>
  count(letter)
```

```
# A tibble: 17 x 2
   letter     n
   <chr>  <int>
 1 D          2
 2 E          3
 3 F          7
 4 G          8
 5 H          9
 6 I          3
 7 J          4
 8 K          8
 9 L          5
10 M          4
11 N          3
12 O          5
13 P          1
14 R          1
15 S          3
16 T          3
17 W          2
```

#### Average max storm wind speed by storm name gender

Create a table with two rows showing the average maximum wind speeds of storms with male or female names.

> ℹ Note
>
> By "average maximum," I mean that you should first compute each storm's maximum wind speed, *then* take the average of these maximum values.

```r
storms |>
  group_by(year, name) |>
  summarize(wind_max = max(wind)) |>
  inner_join(storm_gender, by = "name") |>
  group_by(gender) |>
  summarize(wind_max_average = mean(wind_max))
```

```
# A tibble: 2 x 2
  gender wind_max_average
  <chr>             <dbl>
```

```
1 female              75.6
2 male                75.5
```

Not all storm names appear in `storm_gender`. Write code that returns an alphabetical list of the unique names that appear in `storms` but do not have a `storm_gender`.

```
storms |>
  anti_join(storm_gender, by = "name") |>
  select(name) |>
  distinct() |>
  arrange(name)
```

```
# A tibble: 20 x 1
   name
   <chr>
 1 Babe
 2 Dog
 3 Dottie
 4 Easy
 5 Fernand
 6 Fifi
 7 Flossie
 8 Flossy
 9 Fran
10 Francelia
11 Gerda
12 Gert
13 Hermine
14 Hortense
15 How
16 Isbell
17 Item
18 Jig
19 Shary
20 Sixteen
```

### Storms by status

`storms` contains codes describing the status of the storm at the point of observation. Produce a table containing the number of distinct storms observed with each status. This table should have two columns: One containing the full name of the storm status (*not* the abbreviated code), and the other containing the count of observed storms, sorted by the most frequent status name.

> Tip
>
> The same storm can have multiple different statuses across different observations.

```
storms |>
  select(year, name, status) |>
  distinct() |>
  count(status) |>
  left_join(storm_codes, by = "status") |>
  select(status_name, n) |>
  arrange(desc(n))
```

```
# A tibble: 9 x 2
  status_name                n
  <chr>                  <int>
1 tropical storm           753
2 tropical depression      674
3 hurricane                436
4 extratropical cyclone    362
5 low                      173
6 subtropical storm         59
7 subtropical depression    38
8 disturbance               22
9 tropical wave             14
```

### Max wind speed by first letter per year

Use the `storms` data set to create a table, the *first four* rows and columns of which look like the following:

| year | A | B | C |
|---|---|---|---|
| 1950 | 110 | 90 | 95 |
| 1951 | 80 | 50 | 115 |
| 1952 | 85 | 95 | 105 |
| 1953 | 60 | 80 | 140 |
| 1954 | 95 | 50 | 100 |

Your table must contain one column for all of the years and one column for each letter in the `storms` data set.

The values in each cell other than `year` should represent the max wind speed attained by a storm in that year with a name starting with the corresponding letter. For example, the 1950 storm named Baker had a maximum wind speed of 90.

```r
storms |>
  group_by(year, letter) |>
  summarize(wind_max = max(wind)) |>
  pivot_wider(names_from = letter, values_from = wind_max)
```

```
# A tibble: 71 x 22
# Groups:   year [71]
    year     A     B     C     D     E     F     G     H     I     J     K     L
   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
 1  1950   110    90    95   125   105   120    95    40    90   100   115    70
 2  1951    80    50   115    80   130   100    50    85    55    65    NA    NA
 3  1952    85    95   105    60    90   125    NA    NA    NA    NA    NA    NA
 4  1953    60    80   140    65   100   100    70    75    NA    NA    NA    NA
 5  1954    95    50   100    75   110    55    60   115    NA    NA    NA    NA
 6  1955    80    60   120    90    85    90    65   105   120   150    95    NA
 7  1956    75   105    60    50    50    80    85    NA    NA    NA    NA    NA
 8  1957   110    55   120    35    55    75    NA    NA    NA    NA    NA    NA
 9  1958    55    60   120   115    95    75    50   130    95    85    NA    NA
10  1959    55    60    65    75    50    65   115   105    40    75    NA    NA
# i 61 more rows
# i 9 more variables: M <dbl>, N <dbl>, O <dbl>, P <dbl>, R <dbl>, S <dbl>,
#   T <dbl>, V <dbl>, W <dbl>
```

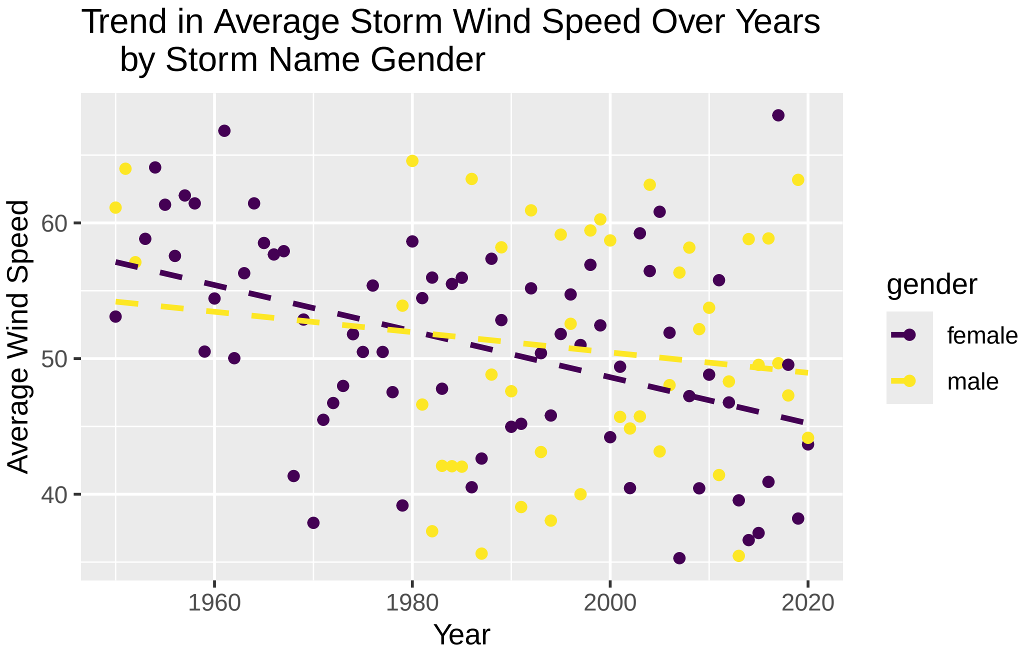
### Trend in Average Wind Speed Over Time by Storm Name Gender

Determine if there is a trend in the average wind speed of storms over time by the gender of the storm's name.

First, calculate the average wind speed for all storms in each year by storm name gender. Then, create a scatter plot of the average wind speed per year by gender, and add linear trend lines *within* each group.

```r
storms |>
  inner_join(storm_gender, by = "name") |>
  group_by(year, gender) |>
```

```
summarize(avg_wind = mean(wind)) |>
ggplot(aes(year, avg_wind, color = gender)) +
geom_point() +
geom_smooth(method = "lm", se = FALSE, linetype = "dashed") +
scale_color_viridis_d() +
labs(
  x = "Year",
  y = "Average Wind Speed",
  title = "Trend in Average Storm Wind Speed Over Years
  by Storm Name Gender"
)
```


Trend in Average Storm Wind Speed Over Years
by Storm Name Gender
Average Wind Speed
Year
gender
female
male