

Statistical Computing - Exercises 10 - HURDAT and dplyr

The code below loads the HURDAT data, which records the best guess of the position and state of tropical cyclones in the Atlantic Ocean basin. A tropical cyclone is a strong rotating storm. A hurricane is a tropical cyclone whose winds exceed 74 mph or 64 knots. Information about the data can be found here:

<https://www.nhc.noaa.gov/data/hurdat/hurdat2-format-atl-1851-2021.pdf>

```
hurdat_raw <- read.csv(
  "../datasets/hurdat2-atl-02052024.txt", header = FALSE, strip.white = TRUE
)
```

Next, we process the data to add columns for the identifier and name. You did this on a previous assignment.

```
# Find the identifier rows, and create an object which has n + 1 appended (for looping)
id_rows <- grep("AL", hurdat_raw[,1] )
first_row <- c( id_rows, nrow(hurdat_raw)+1 )

# append an id column on the left, then loop over first_row
hurdat_raw <- cbind( V0 = NA, name = NA, hurdat_raw )

for( j in 1:( length(first_row) - 1 ) ){

  # get the indicies matching this storm
  inds <- first_row[j]:( first_row[j+1] - 1 )

  # assign identifier (now in column 3)
  hurdat_raw[inds,1] <- hurdat_raw[ first_row[j], 3 ]

  # assign name (now in column 4)
  hurdat_raw[inds,2] <- hurdat_raw[ first_row[j], 4 ]

}

# assign column names
colnames(hurdat_raw) <- c(
  "id", "name", "date", "utc_time", "note", "status", "lat", "lon", "max_speed", "min_pres",
  "rad_34kt_NE", "rad_34kt_SE", "rad_34kt_SW", "rad_34kt_NW",
  "rad_50kt_NE", "rad_50kt_SE", "rad_50kt_SW", "rad_50kt_NW",
  "rad_64kt_NE", "rad_64kt_SE", "rad_64kt_SW", "rad_64kt_NW",
  "rad_max_wind"
)
```

Exercises

1. Use dplyr to create a new object, `hurdat`, that discards the identifier rows of `hurdat_raw`, and converts the data frame to a tibble.
2. Re-assign the -999 values in `hurdat` as NA
3. Use dplyr mutate to convert the `date` column to the Date class and the `utc_time` to numeric.
4. Use dplyr to select only the `id`, `date`, `utc_time`, `note`, and `status` columns. Keep `hurdat` unchanged.
5. Use dplyr to create a tibble that has a row for each unique storm, and columns showing the id, the name, the first date, and first time. Keep `hurdat` unchanged. Hint: use `select` and `distinct`, and read the documentation for `distinct`.

6. Use `group_by` and other related functions to create a list called `rows_for_id` that allows you to quickly return the rows of `hurdat` for a specific storm, e.g. `hurdat[rows_for_id[["AL052005"]],]` would return a tibble with data only for storm `AL052005`.
7. Assign a new tibble with a row for each unique storm, and columns for various summary statistics of the storm. You'll need to use `group_by` and `summarize` for this part. Here are the summary statistics:
 - The id
 - The name of the storm
 - The first date that the storm appears in the dataset
 - The maximum wind speed that the storm reached
 - The minimum pressure that the storm reached
 - TRUE/FALSE for whether the storm became a hurricane.
 - TRUE/FALSE for whether the storm made landfall
8. Print out the top ten most severe storms in terms of maximum speed
9. Print out the top ten most severe storms in terms of minimum pressure
10. Create a new tibble with a row for each year (do not skip any years!), and a column counting the total number of storms each year, the total number of hurricanes each year, and the total number of landfalling hurricanes each year.