Explore R’s capabilities for statistical analysis

Gahyeon

2023-04-13

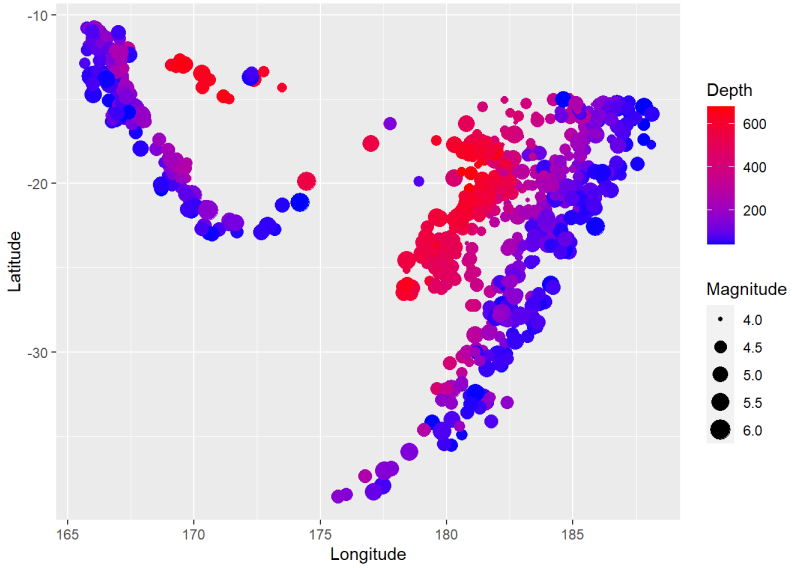
Exercise 1

The ggplot2 built-in R data set quakes gives the locations of earthquakes off of Fiji in the 1960’s. Create a plot of the locations of these earthquakes, showing depth with color and magnitude with size.

```
library(ggplot2)

# Load the quakes dataset
data(quakes)

# Create the plot
ggplot(quakes, aes(x = long, y = lat)) +
  geom_point(aes(color = depth, size = mag)) +
  scale_color_gradient(low = "blue", high = "red") +
  labs(x = "Longitude", y = "Latitude", color = "Depth", size = "Magnitude")
```



Exercise 2

The data set storms is included in the dplyr package. It contains information about 425 tropical storms in the Atlantic. Produce a plot showing the position track of each storm from 2014 (use long for x and lat for y). Color your points by the name of the storm so you can distinguish the seven storm tracks. Which storm in 2014 made it the furthest North?

*Answer: The highest latitude(the furthest North) is “Cristobal” as below graph.

```
library(dplyr)
```

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

```
##
## Attaching package: 'dplyr'
```

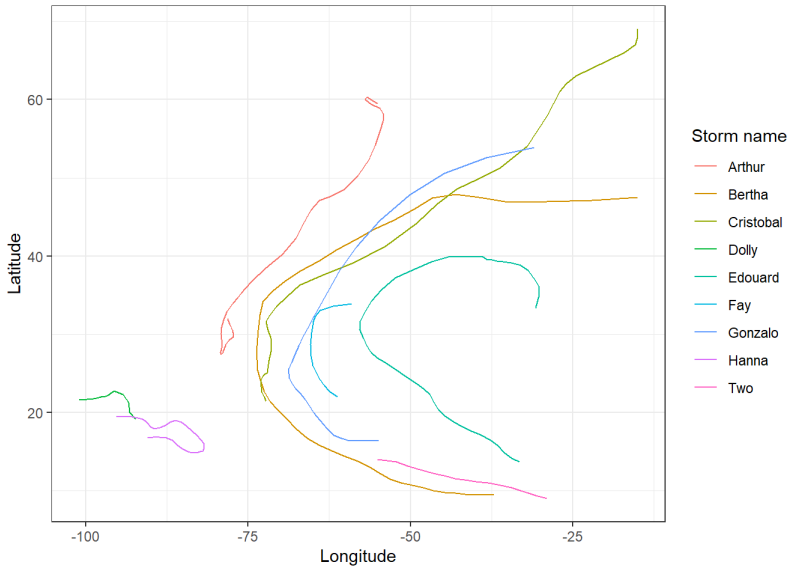
```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(ggplot2)

# Filter the storms dataset to include only storms from 2014
storms2014 <- storms %>% filter(year == 2014)

# Create the plot
# On the updated data, there are 9 different storm
ggplot(storms2014, aes(x = long, y = lat, color = name)) +
  geom_path() +
  labs(x = "Longitude", y = "Latitude", color = "Storm name") +
  theme_bw()
```




```r
# Filter the storms dataset to include only storms from 2014
storms2014 <- storms %>% filter(year == 2014)

# Find the storm with the highest latitude value
storms2014[which.max(storms2014$lat), "name"]
```

name
<chr>
Cristobal

1 row

Exercise 3

Use the Batting data set from the Lahman package. This gives the batting statistics of every player who has played baseball from 1871 through the present day. Create boxplots for total runs (variable R) scored per year in the AL and the NL from 1969 to the present.

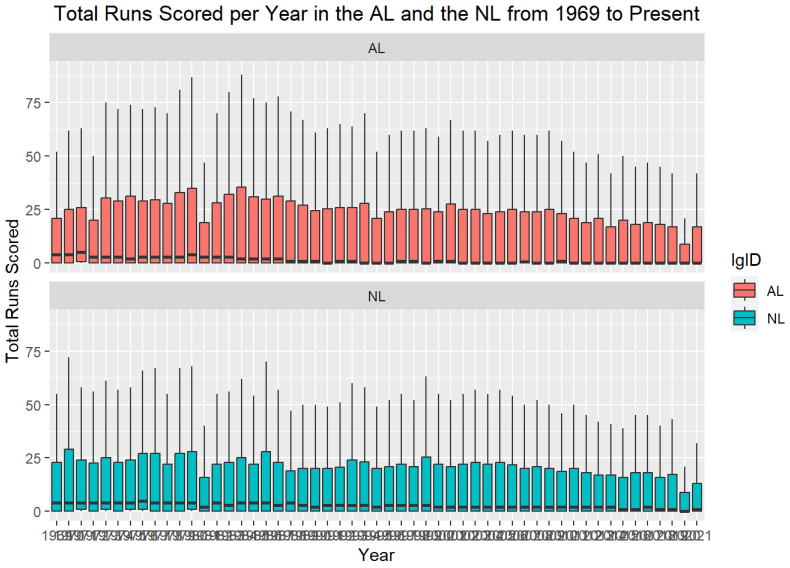
```r
library(Lahman)
```

```
## Warning: package 'Lahman' was built under R version 4.2.3
```

```r
library(dplyr)
library(ggplot2)

# Subset the Batting dataset to include only the years 1969 to present
Batting_subset <- Batting %>%
  filter(yearID >= 1969)

# Create separate boxplots for total runs (variable R) scored per year in the AL and the NL
ggplot(Batting_subset, aes(x=as.factor(yearID), y=R)) +
  geom_boxplot(aes(fill=lgID), outlier.shape=NA) +
  facet_wrap(~lgID, ncol=1, scales="free_y") +
  labs(title = "Total Runs Scored per Year in the AL and the NL from 1969 to Present",
       x = "Year", y = "Total Runs Scored") +
  theme(plot.title = element_text(hjust = 0.5)) +
  coord_cartesian(ylim=c(0, 90))
```


Total Runs Scored per Year in the AL and the NL from 1969 to Present
AL
NL
Total Runs Scored
0
25
50
75
Year
lgID
AL
NL