HW1:607

Neil Shah

1/29/2020

#Neil Shah: DATA 607 HW1

Introduction:

This assignment is to test our our R transofmation/dataframe skills by playing around with data! I chose the We Watched 906 Foul Balls To Find Out Where The Most Dangerous Ones Land (https://fivethirtyeight.com/features/we-watched-906-foul-balls-to-find-out-where-the-most-dangerous-ones-land/), data set that covers foul balls.

The full data set is available here (https://github.com/fivethirtyeight/data/tree/master/foul-balls)

Loading data set

First I loaded the data into new dataframe

```
df <- read.csv('foul-balls.csv')
```

Exploratory Data Analysis

Now some basic exploration

```
r in df.head() : could not find function "df.head"
> head(df)
                           ï..matchup  game_date type_of_hit exit_velocity predicted_zone camera
_zone used_zone
1 Seattle Mariners VS Minnesota Twins 2019-05-18      Ground            NA              1
1         1
2 Seattle Mariners VS Minnesota Twins 2019-05-18         Fly            NA              4
NA         4
3 Seattle Mariners VS Minnesota Twins 2019-05-18         Fly          56.9              4
NA         4
4 Seattle Mariners VS Minnesota Twins 2019-05-18         Fly          78.8              1
1         1
5 Seattle Mariners VS Minnesota Twins 2019-05-18         Fly            NA              2
NA         2
6 Seattle Mariners VS Minnesota Twins 2019-05-18      Ground            NA              1
1         1
> dim(df)
[1] 906   7
> summary(df)
                                        ï..matchup       game_date             type_of_hit  exit
_velocity
 Baltimore Orioles VS Minnesota Twins        :113   2019-04-20:113   Batter hits self: 17   Min.
: 25.4
 Pittsburgh Pirates VS Milwaukee Brewers     :111   2019-06-01:111   Fly             :522   1st
Qu.: 69.7
 Oakland A's vs Houston Astros               :109   2019-06-02:109   Ground          :226   Medi
an : 75.7
 Seattle Mariners VS Minnesota Twins         :100   2019-05-18:100   Line            : 87   Mean
: 76.4
 Texas Rangers vs Toronto Blue Jays          : 87   2019-05-03: 87   Pop Up          : 54   3rd
Qu.: 81.7
 Los Angeles Dodgers vs Arizona Diamondsbacks: 86   2019-03-29: 86                          Max.
:110.6
 (Other)                                     :300   (Other)   :300                          NA's
:326
 predicted_zone   camera_zone      used_zone
 Min.   :1.000   Min.   :1.000   Min.   :1.000
 1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000
 Median :3.000   Median :1.000   Median :3.000
 Mean   :3.038   Mean   :2.369   Mean   :3.058
 3rd Qu.:5.000   3rd Qu.:4.000   3rd Qu.:5.000
 Max.   :7.000   Max.   :7.000   Max.   :7.000
                 NA's   :513
> names(df)
[1] "ï..matchup"     "game_date"      "type_of_hit"    "exit_velocity"  "predicted_zone" "camera
_zone"
[7] "used_zone"
```

This data set is a 906 X 7 matrix with 906 variables and 7 variables.
Matchup: categorical variable
Game-date: date-time varable (numeric)
type of hit: categorical
Exit velocity: continuous numerical variable

Predicted Zone: categorical variable
Camera Zone: categorical variable
Used zone: categorical variable

Renaming Columns

So first I'll rename game_date and exit_velocity just to make things a bit simpler

```
 names(df)[names(df) == "game_date"] <- "date"
> names(df)[names(df) == "exit_velocity"] <- "speed"
> names(df)
[1] "ï..matchup"     "date"           "type_of_hit"    "speed"          "predicted_zone" "camera
_zone"
[7] "used_zone"
```

Cleaning NA values

So I think it'd be interesting to see which type_of_hit has the fastest speed! However looking at the speed columns…

```
> head(df$speed)
[1]   NA   NA 56.9 78.8   NA   NA
```

It looks like we habe some NA values, that's no good. Let's cleanup and drop them.

```
> df <- na.omit(df)
> df$speed
  [1]  78.8  76.0  95.9  69.9  84.9 104.6  74.6  76.1  72.2 100.8  78.6  88.1  73.4  78.9  85.2
76.4  77.6  96.8  84.7
 [20]  94.0  63.7  94.8 100.7  79.2  87.3  77.6  76.9  98.6  81.7  85.8  67.6  74.3 106.6 110.6
105.3  74.9  98.4  42.0
 [39]  61.2  98.7  79.6  73.9  85.3  80.1 106.2  66.3  85.5  92.7  66.3 108.5  53.3 108.5  78.1
60.3  73.3  74.4  71.0
 [58]  80.2  96.6  62.8 101.4  85.3  70.9  74.6  81.7  68.8  88.7  66.4  69.1  96.6  89.2  48.8
80.0  68.1  77.2  76.7
 [77]  82.4  68.3  66.7  78.3  76.6  75.0  74.2  74.9  74.0  96.0  85.1  78.7  53.8 102.3  98.0
79.6 107.0  74.8  77.9
 [96]  77.8  72.2  65.5  58.6  68.7  65.8  80.0  78.9  84.6  66.8  68.7  65.2  60.4  81.7  87.4
53.7  72.0  79.7  82.0
[115] 101.8  89.5  69.9  74.6  69.8  65.6  57.4  72.0  94.0  76.9  85.1  88.2  67.1  78.1  76.8
83.8  78.5 105.3  76.9
[134]  76.0  83.9  91.9  95.7  79.2  94.4  83.1  95.9  87.1  91.7  70.2  79.1  85.6  73.7 100.7
58.9  39.9  77.9  92.4
[153]  91.1  79.0  89.7  80.3  86.1  58.4  84.6  82.4  82.6  47.3  81.3  53.1  83.7  79.9  58.8
78.4  83.5  84.7  77.1
[172]  92.7  71.9 103.0  76.8  65.3  96.0  78.7  99.9  69.7  57.5  80.7  79.8  72.8  84.1  85.0
86.8  80.5  96.5  73.9
[191]  78.4  75.9  78.0  61.8  84.4  76.0  71.5  76.8  75.9  91.5  25.4  69.5  73.5  71.7  86.8
80.7  85.5 103.3  79.0
[210] 101.7  83.5  80.5 100.0  64.2  74.2  73.1  87.5  76.8  81.7  77.6  81.4  74.9  74.2  68.7
79.3  67.5  75.8  91.5
[229]  76.7  90.7  73.4  78.9  72.1  70.4  73.0  65.1  85.0  73.6  91.9  73.6 102.3  81.3  93.6
84.6  77.5  75.5  77.3
[248]  52.0  68.5

> dim(df)
[1] 249   7
```

Ok so much better–Looks like we eliminated almost half the data!

Speed statistics

First let's aggregate by simple geometric mean

```
> summarise(group_by(df, type_of_hit), mean(speed))
# A tibble: 5 x 2
  type_of_hit      `mean(speed)`
  <fct>                    <dbl>
1 Batter hits self          69.4
2 Fly                       81.6
3 Ground                    75.5
4 Line                      82.1
5 Pop Up                    77.9
```

Interesting–so line hits and fly hits have a very similar speed, following by pop-up, ground and then batter hits self (which I have no clue what that is).

What about the median?

```
> summarise(group_by(df, type_of_hit), median(speed))
# A tibble: 5 x 2
  type_of_hit      `median(speed)`
  <fct>                      <dbl>
1 Batter hits self            68.3
2 Fly                         79.1
3 Ground                      74.8
4 Line                        82.6
5 Pop Up                      77.5
```

Ok now we have some separation and this shows the true divide between the Line (the fastest it seems) and the other tyoe of hits.

What about max?

```
summarise(group_by(df, type_of_hit), max(speed))
# A tibble: 5 x 2
  type_of_hit      `max(speed)`
  <fct>                   <dbl>
1 Batter hits self         82
2 Fly                     108.
3 Ground                  107
4 Line                    111.
5 Pop Up                   90.7
```

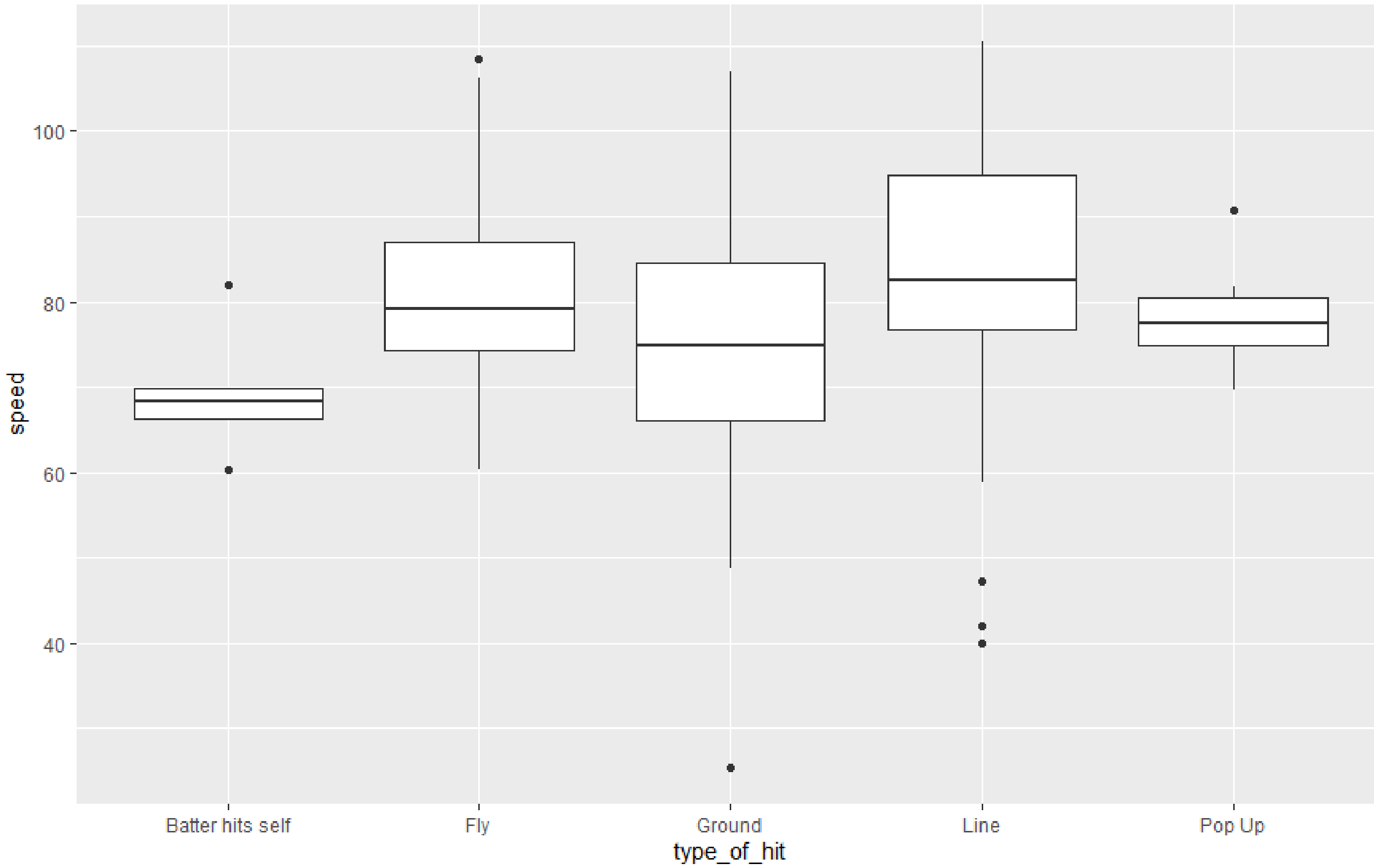
Interesting–so it seems that the Fly/Ground have similar top speeds–but once again the Line is the overall fastest (max speed) wise. Yet from the median data set–Pop Up had a higher median speed than Fly, yet here the Fly has a higher top speed. Very cool! I wonder if we can make a new metric to see what % of max speed possible speed each type of hit is–iike an efficiency?

```
type_of_hit      `mean(speed)/max(speed)`
  <fct>                               <dbl>
1 Batter hits self                    0.846
2 Fly                                 0.752
3 Ground                              0.706
4 Line                                0.742
5 Pop Up                              0.859
```

So at first this seems confusing–since Line has a lower rating but this makes sense due to it's really high top speed. Pop-up's data shows that most pop up hits travel at approximately 86% of it's max speed–this might be due to it being hit off the bat a certain way. Fly and Ground operate near 75% of their max speed.

A Final Boxplot

Let's put this all together to visualizat the data



Boxplot

Conclusions

This was a quick and dirty way to look at a data-set in R, but it shows the power of exploratory data-analysis and grouping functions.

The main observation were the different stratifications of speeds for the type of hit–particularly Line being the fastest. We can improve on this study by the following analysis

-Plot variation of speed by matches; Maybe there is a team with strong batters?
-See how speed evolved over time–did batters get stronger?