Week 9

Saher Manaseer

2022-09-20

Week 9

In this week, we are to continue working on the "Hitters" dataset. You will need to read the file as we did in week 8.

lets start by exploring the data set. Load the "explore" library then use the explore() function.

```
library(explore)
library(tidyverse)
```

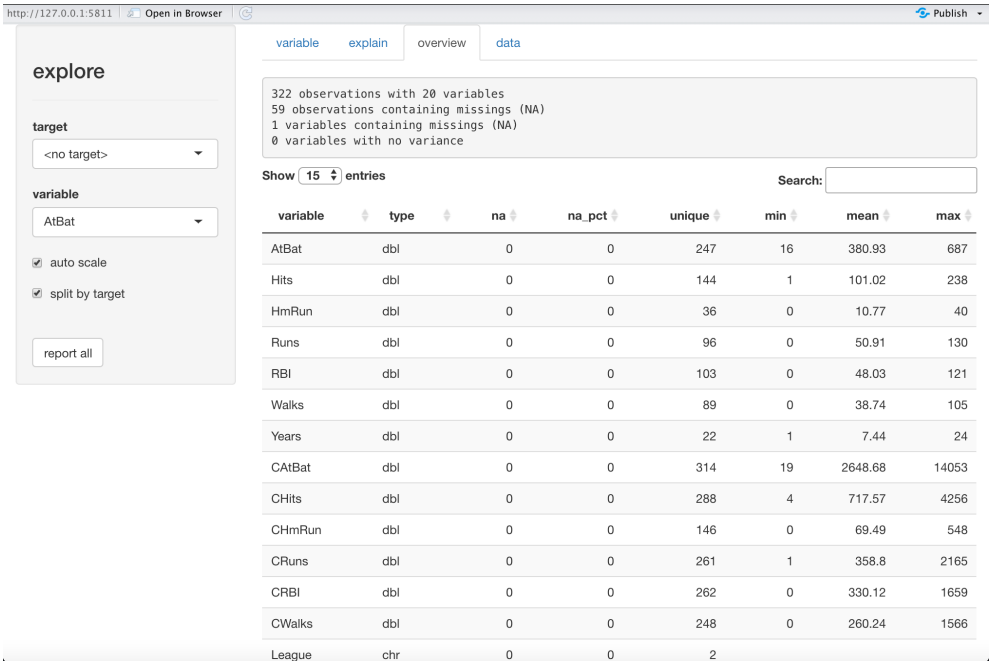
```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.8     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.1
## v readr   2.1.2     v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
df <- read_csv("Hitters.csv")
```

```
## Rows: 322 Columns: 20
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (3): League, Division, NewLeague
## dbl (17): AtBat, Hits, HmRun, Runs, RBI, Walks, Years, CAtBat, CHits, CHmRun...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
#explore(mtcars)
```

The follwoing window will popup



1. Click on Overview to explore the contents of the datasets.
2. After that, lets navigate to variables tab and explore relations between different variables.
3. find the variables with clear relations that can be used for modelling. Explain your answers.

Cricket Data

This tutorial explores cricket statistics. The 'fetch_cricinfo()'is used to fetch data on men's T20 cricket batting statistics.

```
# Load cricketdata
library(cricketdata)

# Fetch men's T20 batting data
MenT2 <- fetch_cricinfo("T20", "Men", "Batting")

# Filter for only Australia and India
MenT2_aus_ind <- MenT2 %>%
  filter(Country %in% c("India", "Australia"))
```

* How many rows and columns?

* What does each row represent?

* What function returns the top 7 rows.

```
# Convert MenT2_aus_ind to long form
MenT2_aus_ind_long <- MenT2_aus_ind %>%
  select(Player, Country, NotOuts, HighScore, Average, StrikeRate, Hundreds, Fifties, Ducks, Fours, Six
  gather(Bat_Stats, Value, -Player, -Country)

# Print MenT2_aus_ind_long
MenT2_aus_ind_long
```

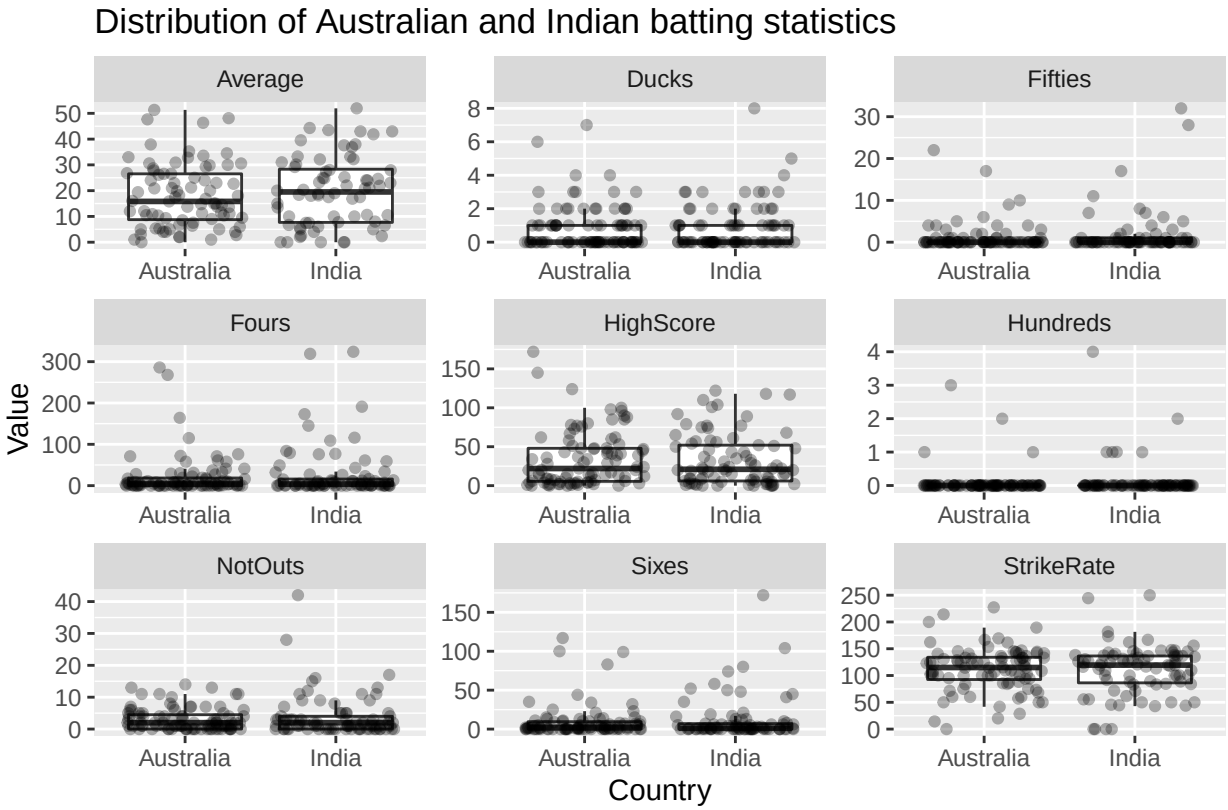
```
## # A tibble: 1,755 x 4
##    Player        Country   Bat_Stats Value
##    <chr>         <chr>     <chr>     <dbl>
##  1 "RG Sharma "  India     NotOuts      16
##  2 "V Kohli "    India     NotOuts      28
##  3 "AJ Finch "   Australia NotOuts      11
##  4 "DA Warner "  Australia NotOuts      11
##  5 "GJ Maxwell " Australia NotOuts      13
##  6 "KL Rahul "   India     NotOuts       8
##  7 "S Dhawan "   India     NotOuts       3
##  8 "MS Dhoni "   India     NotOuts      42
##  9 "SK Raina "   India     NotOuts      11
## 10 "SR Watson "  Australia NotOuts       6
## # ... with 1,745 more rows
```

Let have some box plots of the data. Why? What can you conclude from the boxes produced?

```
# Box plots of countries by batting statistics
MenT2_aus_ind_long %>%
  ggplot(aes(x = Country,  y = Value)) +
  geom_boxplot(outlier.alpha = 0) + # hide the outliers
  geom_jitter(alpha = 0.3) +
  facet_wrap(~ Bat_Stats, scales = "free") +
  labs(title = "Distribution of Australian and Indian batting statistics",
       caption = "Data source: https://github.com/ropenscilabs/cricketdata")
```

```
## Warning: Removed 237 rows containing non-finite values (stat_boxplot).
```

```
## Warning: Removed 237 rows containing missing values (geom_point).
```



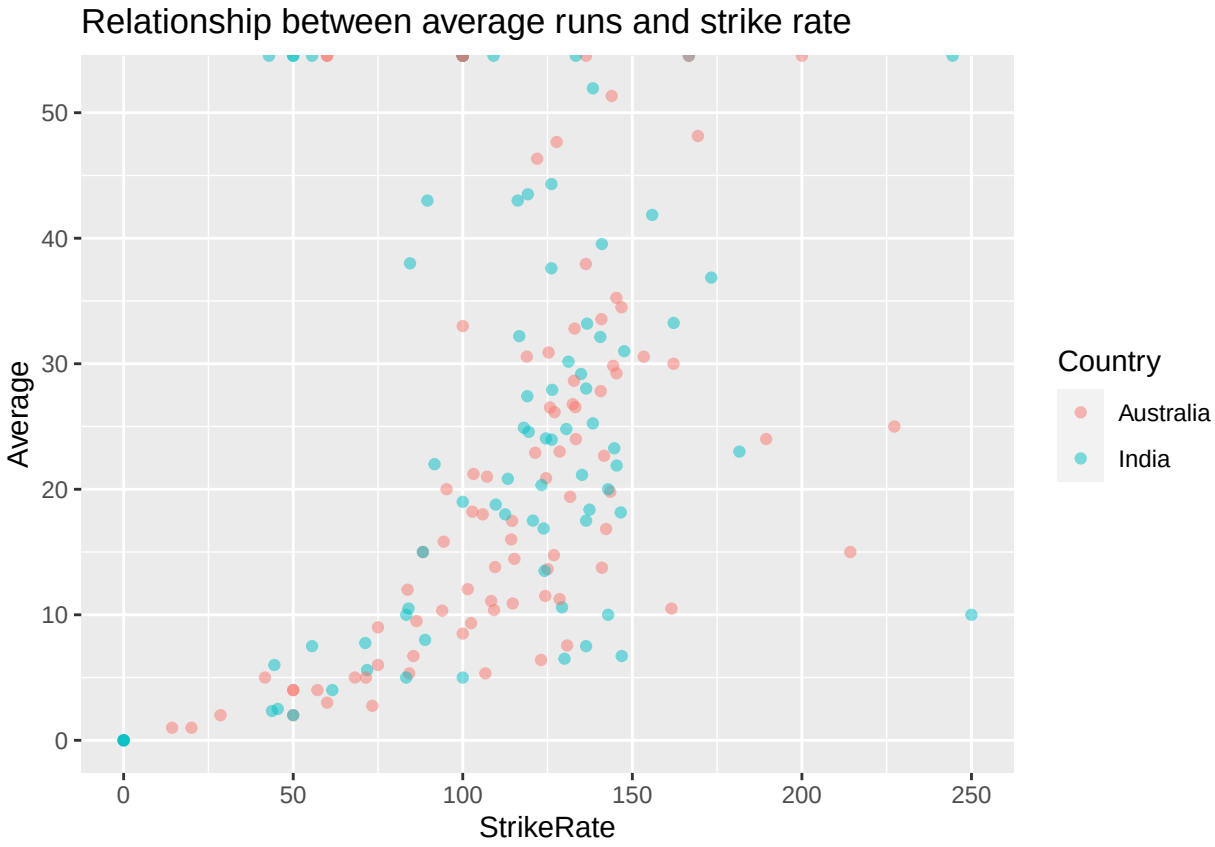
Performing group statistics

```
# Compute ratio of total runs divided by total matches
MenT2_aus_ind %>%
  group_by(Country) %>%
  summarise(RunsTotal = sum(Runs, na.rm = TRUE),
            MatchesTotal = sum(Matches, na.rm = TRUE),
            Runs_matches_ratio = RunsTotal/MatchesTotal) %>%
  ungroup()
```

```
## # A tibble: 2 x 4
##   Country   RunsTotal MatchesTotal Runs_matches_ratio
##   <chr>         <int>        <int>               <dbl>
## 1 Australia     22753         1731                13.1
## 2 India         25868         1931                13.4
```

```
MenT2_aus_ind %>%
  ggplot(aes(x = StrikeRate, y = Average, colour = Country)) +
  geom_point(alpha = 0.5) +
  labs(title = "Relationship between average runs and strike rate")
```

```
## Warning: Removed 29 rows containing missing values (geom_point).
```


Relationship between average runs and strike rate
Average
StrikeRate
0
10
20
30
40
50
0
50
100
150
200
250
Country
Australia
India