Exploring

Tools for Exploratory Data Analysis

Data Exploration

> *Exploratory data analysis is detective work-numerical detective work-or counting detective work-or graphical detective work…Exploratory data analysis can never be the whole story, but nothing else can serve as the foundation stone--as the first step.* - John Tukey

EDA Starts With Specific Questions

- All exploration needs some direction
- Exploring aimlessly wastes time and rarely will get you where you need to be
- To avoid aimless exploration, you need to ask yourself what you are looking for? and what would be interesting to find?

Our Questions

In this tutorial we will use EDA to investigate some conventional wisdom in tennis. Here are 3 ideas commentators repeatedly say:

1. Second serve is more important than the first serve
2. Players who serve first have an advantage
3. The 7th game is the most important in a set

Match Data

To examine the first question we will use the `atp_matches` data set from the `deuce` package.

Load the data, limit to the years 2005 to 2015.

```r
library(deuce)

data(atp_matches)

atp_matches <- atp_matches %>%
  dplyr::filter(year >= 2005 & year <= 2015)
```

Variable Documentation

Use the `help` function to learn about the contents of the `atp_matches`. What do these data include?

```r
help("atp_matches", package = "deuce")
```

Second vs First Serve

The first question we will consider is the importance of the second serve versus the first serve.

- One implication of this statement is that we might expect the winner of a match to have outperformed on second serve compared to first serve
- This suggests focusing on the difference in winner and loser service stats

Service Points Won

For each match, we will calculate the proportion of second serve points won and first points won for the winner and loser of the match.

```
atp_matches <- atp_matches %>%
    dplyr::mutate(
      w_first = w_1stWon / w_svpt,
      w_second = w_2ndWon / w_svpt,
      l_first = l_1stWon / l_svpt,
      l_second = l_2ndWon / l_svpt
    )
```

Summarise First and Second Differential

Let's look at the difference in the winner and loser stats on each serve.

```
summary(with(atp_matches, w_first - l_first))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##   -0.50    0.02    0.08    0.09    0.14    0.88  224474
```

Filtering

One problem with the previous summary is that we have included some matches with NA stats and 0-values stats, as well as matches ending in retirement. How would you correct this?

Filtering

One problem with the previous summary is that we have included some matches with NA stats and 0-values stats, as well as matches ending in retirement. How would you correct this?

```
summary(with(
  subset(atp_matches, !is.na(w_svpt) & w_svpt != 0
         & !is.na(l_svpt) & l_svpt != 0 & !Retirement),
             w_first - l_first))
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
## -0.35260  0.01658  0.07684  0.08371  0.14330  0.61300
```

```
summary(with(
  subset(atp_matches, !is.na(w_svpt) & w_svpt != 0
         & !is.na(l_svpt) & l_svpt != 0 & !Retirement),
             w_second - l_second))
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
## -0.36840 -0.01673  0.03355  0.03709  0.08689  0.51210
```

Charting

Charting

- Most of our describing and exploration in `R` happens with graphics

Charting

- Most of our describing and exploration in `R` happens with graphics
- A powerful package for graphing is `ggplot2`

Charting

- Most of our describing and exploration in `R` happens with graphics
- A powerful package for graphing is `ggplot2`
- `ggplot2` provides a grammar for building univariate, bivariate, and lattice plots

Charting

- Most of our describing and exploration in `R` happens with graphics
- A powerful package for graphing is `ggplot2`
- `ggplot2` provides a grammar for building univariate, bivariate, and lattice plots
- Also, many specialty graphics packages build on `ggplot2`

Overview of `ggplot2`

To install, use `install.packages('ggplot2')`.

Function.Type	Description
aes	Relates variables to axes and aesthetic elements of our plot
geoms	Define how the variables will be displayed, that is, the type of chart
facet	Splits plot into rows and columns defined by a group variable
scales	Customizes the range and limits of aesthetics
themes	Further controls of the style and elements of the chart

Function.Type Description

Charting Serve Differentials

We could describe the difference in service stats with one of several univariate geoms: `geom_histogram`, `geom_density`, `geom_boxplot`.

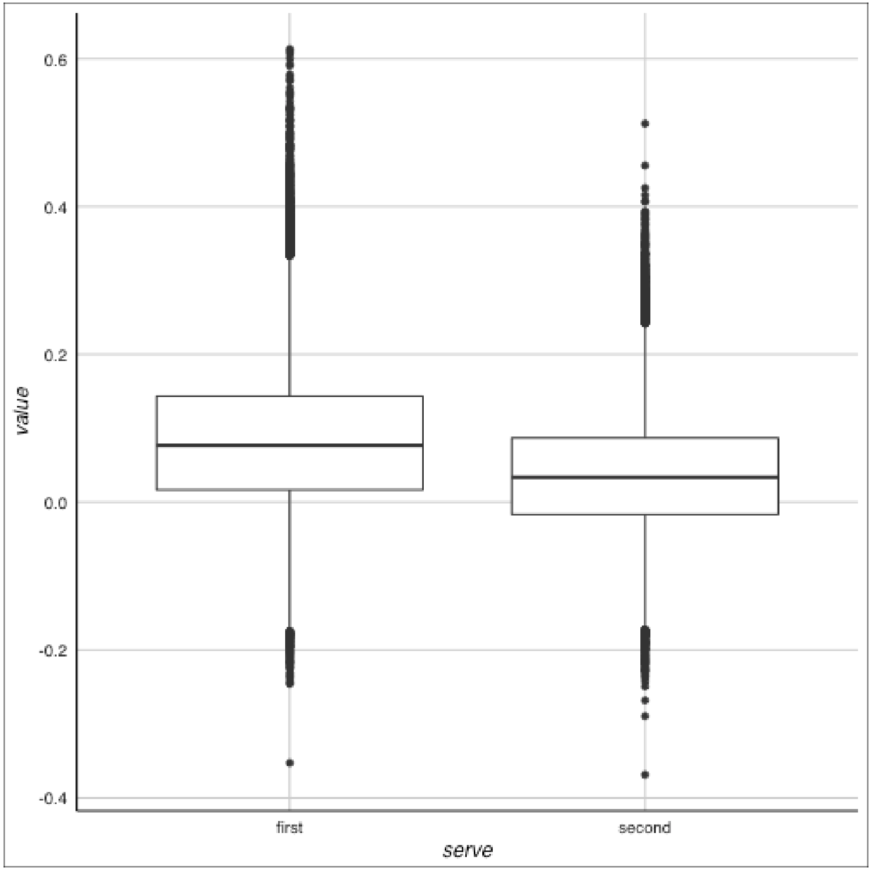
1. Reshape the data to long format with first and second serve differences stacked length-wise
2. Then use `geom_boxplot` to look at the difference in the differentials for the first and second.
3. What do you conclude about the comparative importance of first and second serve?

Solution: Charting Serve Differentials

```
# Prepare long format
serve_stats <- atp_matches %>%
  filter(!is.na(w_svpt), w_svpt != 0,
         !is.na(l_svpt), l_svpt != 0, !Retirement) %>%
  dplyr::mutate(
    first = w_first - l_first,
    second = w_second - l_second
  ) %>%
  gather("serve", "value", first, second)
```

Solution: Charting Serve Differentials

```
library(ggplot2)
library(ggthemes) # Extra themes

serve_stats %>% # We can use pipes with ggplot2!
  ggplot(aes(y = value, x = serve)) +
  geom_boxplot() +
  theme_gdocs()
```


0.6
0.4
0.2
0.0
-0.2
-0.4
value
first
second
serve

Interpretation

- We see a greater gap in the stats on the first serve compared to the second
- This suggests that it is a stronger differentiator between winning and losing

Serving First

- Next, let's consider the importance of serving first and what advantage that has for winning matches

- For this question, we will make use of the `gs_point_by_point` data which includes point-level data for several years of Grand Slam matches

```
data("gs_point_by_point")

gs_point_by_point <- gs_point_by_point %>%
  filter(Tour == "atp")
```

Practice: Exploring Serve Advantage

1. Sort the data by match and point number
2. Determine the percentage of players who served first that won the match
3. Determine the percentage who served second and won the match
4. Chart your results using a bar chart

Solution: Exploring Serve Advantage

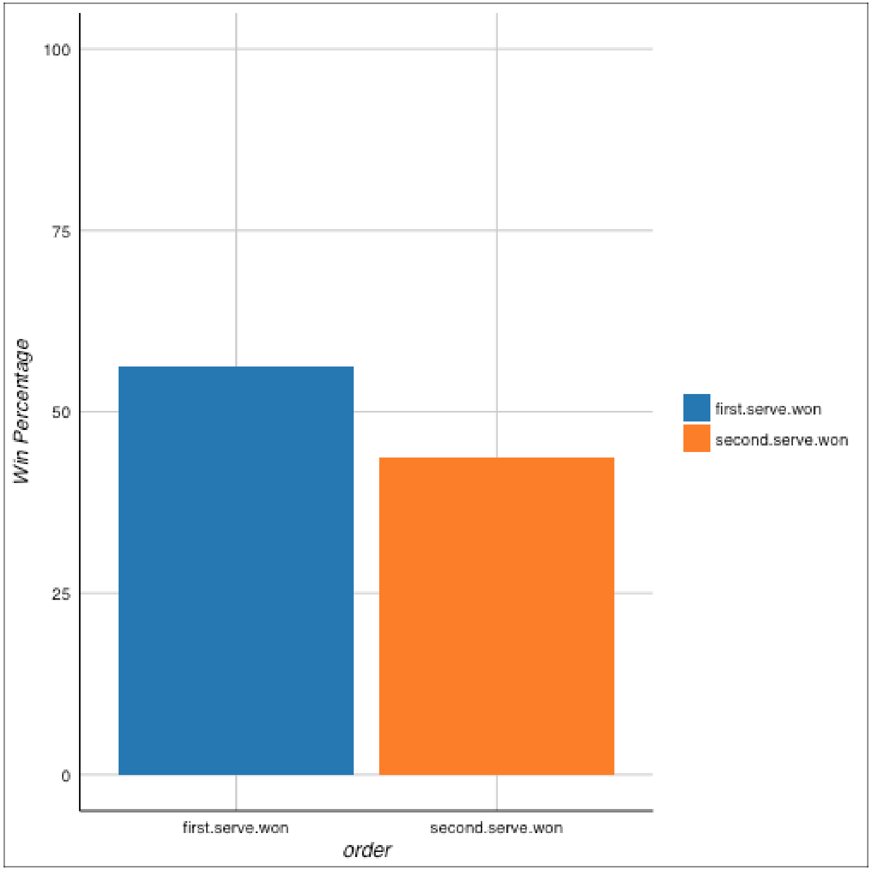
Here we sort and calculate the outcome with respect to the first and second server

```
# Sort data
gs_point_by_point <-
  gs_point_by_point[order(gs_point_by_point$match_id, gs_point_by_po

serve_advantage <- gs_point_by_point %>%
  group_by(match_id) %>%
  dplyr::summarise(
    first.serve.won = PointServer[1] == PointWinner[n()],
    second.serve.won = ifelse(PointServer[1] == 1, 2, 1) == PointWinn
  )
```

Solution: Exploring Serve Advantage

Now we summarise and reshape the data to prepate for charting.

```
serve_advantage <- serve_advantage %>%
  dplyr::summarise(
    first.serve.won = mean(first.serve.won),
    second.serve.won = mean(second.serve.won)
  ) %>%
  gather("order", "win", first.serve.won, second.serve.won)

serve_advantage %>%
  ggplot(aes(y = win * 100, x = order, fill = order)) +
  geom_bar(stat = "identity") +
  scale_y_continuous("Win Percentage", lim = c(0, 100)) +
  scale_fill_tableau(name = "") +
  theme_gdocs()
```



We find a near 10 percentage point advantage with serving first!

Practice: Game 7

Using the same point-level dataset and similar methods, investigating whether the "all-important game 7" is really that important for winning a set.

1. Determine how often the winner of the 7th game of a set won the set

2. Limit your analysis to "close" sets, which we will define as sets with 10 more games

3. Plot the percentage difference in set wins for winner's and loser's of game 7

Solution: Game 7

First we prepare the variables for summarising the set wins.

```
# Determine max games
game7 <- gs_point_by_point %>%
  group_by(match_id, SetNo) %>%
  dplyr::mutate(
    MaxGame = max(GameNo),
    SetWinner = PointWinner[n()]
)

# Filter and determine game 7 winner
game7 <- game7 %>%
  filter(MaxGame >= 10) %>%
  group_by(match_id, SetNo) %>%
  dplyr::mutate(
    Game7 = PointWinner[max(which(GameNo == 7))]
)
```

Solution: Game 7

Now, we summarise the set win percentages by game 7 status.

```
game7 <- game7 %>%
  group_by(match_id, SetNo) %>%
  dplyr::summarise(
    game7.winner = Game7[1] == SetWinner[1],
    game7.loser = ifelse(Game7[1] == 1, 2, 1) == SetWinner[1]
)

# Summarise and put in long format
game7 <- as.data.frame(game7) %>%
  dplyr::summarise(
    game7.winner = mean(game7.winner),
    game7.loser = mean(game7.loser)
  ) %>%
  gather("gamewinner", "setwin", game7.winner, game7.loser)
```

Solution: Game 7

Now we are ready to chart our findings.

```
game7 %>%
  ggplot(aes(y = setwin * 100, x = gamewinner, fill = gamewinner)) +
  geom_bar(stat = "identity") +
  scale_y_continuous("Set Win Percentage", lim = c(0, 100)) +
  scale_fill_solarized(name = "") +
  theme_gdocs()
```

Summary

- Using common tools for data manipulation and graphics we have investigated 3 commonly held views in tennis
- Our observations provide support for only 1 of these 3 beliefs: the advantage of serving first

Resources

- ggplot2
- ggthemes
- Tukey and EDA
- EDA by Roger Peng