

An Introduction to SQL Server, R0DBC, and ggplot2 with PITCHf/x Data

Sam Justice Tyler Olson Alex Zajicheck

Computing in Statistics
Fall 2015

Table of Contents

1 Introduction

- Objectives
- The Data

2 ggplot2

- ggplot2: An Introduction
- The Power of ggplot2
- Components
- Example 1

3 References

Objectives

- To demonstrate the basics of SQL Server, open database connectivity (via RODBC), and ggplot2 using an interesting Major League Baseball (MLB) dataset
 - Focus on data management and data visualization

The Data

- PITCHf/x Data: collected for each individual pitch in an MLB game
 - Cameras in stadium record release point, velocity, movement, spin, and location of pitch
 - Data have only been collected since 2006
- We will work with data from a single game
 - October 6, 2009: American League Central Division tie-breaker between Detroit Tigers and Minnesota Twins
 - Winner crowned AL Central champion, advanced to playoffs

ggplot2: An Introduction

- Data visualization package for R
- Intended to improve upon base and lattice graphics packages
 - Takes care of annoying details (e.g., making legends) for user
 - Enables user to focus on producing graphics that best bring out the data
 - Facilitates easy creation of publication-quality graphics

The Power of ggplot2

- Deep underlying grammar (“The Grammar of Graphics”)
 - Composed of independent components that can be combined in many ways
 - Simple set of core principles that can be tailored to specific problem at hand
 - Not limited to set of pre-designed graphics
- Creation of graphics proceeds in layered fashion
 - First layer simply involves raw data
 - Additional layers allow user to summarize data in more illuminating ways

Components

1 ggplot() object

- Basic plot structure
- Holds the information that is to be visualized
- Doesn't create a physical plot

2 Layers

- Geometric layers
- Statistics layers

Example 1

- Code:

```
library(ggplot2)
```


Example 1

- Code:

```
library(ggplot2)
ggplot(data=pitch, aes(x=x,y=y))
```

Example 1

- Code:

```
library(ggplot2)
ggplot(data=pitch, aes(x=x,y=y))
```

- Output:

Error: No layers in plot

Example 1

- Code:

```
library(ggplot2)  
ggplot(data=pitch, aes(x=x,y=y))
```

- Output:

Error: No layers in plot

How do we want to visualize?

Example 1

- Code: (second try with layers)

```
ggplot(data=pitch, aes(x=x,y=y))
```

Example 1

- Code: (second try with layers)

```
ggplot(data=pitch, aes(x=x,y=y)) +  
geom_point(aes(color=inning))
```

Example 1

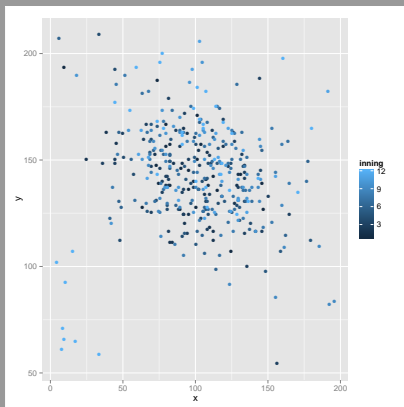
- Code: (second try with layers)

```
ggplot(data=pitch, aes(x=x,y=y)) +  
geom_point(aes(color=inning))
```

OR

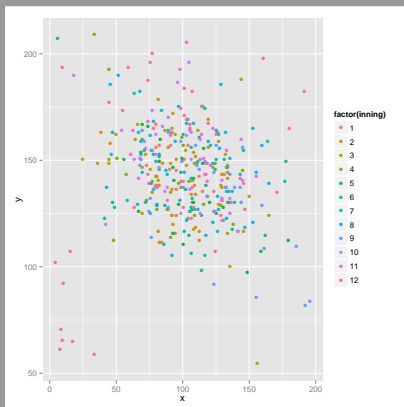
```
ggplot(data=pitch, aes(x=x,y=y,color=inning)) +  
geom_point()
```

Example 1: Output



Example 1: Alternate

```
ggplot(data=pitch, aes(x=x,y=y)) +  
geom_point(aes(color=factor(inning)))
```



References

- 1 Albert, J. and Marchi, M. (2014). *Analyzing Baseball Data with R*. CRC Press.
- 2 Ripley, B. and Lapsley, M. (2015). *RODBC: ODBC Database Access*. R package version 1.3-12.
- 3 Sievert, C. (2015). *pitchRx: Tools for Harnessing MLBAM Gameday Data and Visualizing PITCHf/x*. R package version 1.8.1.
- 4 Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*. Springer New York.