R, MongoDB, and ggplot2

Research Slides

Luke Roy

R

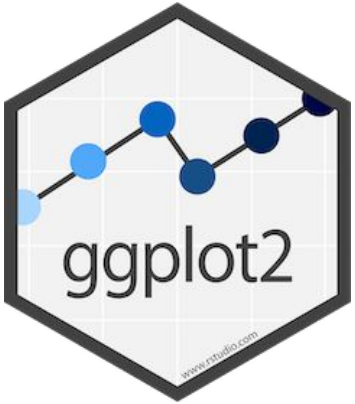
- Designed by Ross Ihaka & Robert Gentleman.
- R works well with data and excels in data analysis, data visualization, and data science.
- Invented explicitly for statistical computing.
- Widely used among statisticians and data scientists.
- Technically a programming language it is easier to think of R as a tool to understand data.

MongoDB

- An open-source NoSQL database, that uses JSON-like documents (BSON) with optional schemas.
- Very Flexible, does not require schemas, and uses internal memory making it faster in accessing data.
- BSON - “Binary JSON”, in a modern database JSON can be less than ideal as it is text based, making text parsing very slow, not very space efficient, and JSON only supports a limited number of basic data types.
- BSON was invented as sort of a hybrid to optimize speed, space and flexibility as it is a binary representation to store data in JSON format.
- Has GUI (MongoDB Compass).

RDBMS	MongoDB
Database	Database
Table	Collection
Row	Document
Column	Field
Primary Key	Primary Key
Table Join	Embedded Documents

ggplot2



- A data visualization package, designed with the grammar of graphics in mind, hence the gg (grammar of graphics).
- Three main basic grammatical units:
  - Data - Information you want visualized
  - Geometries - Describes the shapes you want data in (dots, bars, lines, etc.)
  - Aesthetics - Visual attributes of the plot
- Two main libraries used when using ggplot2:
  - readr - Reads “rectangular data” such as a csv file
  - dplyr - Contains set of functions to enable dataframe manipulation

Ggplot2 example

```
bar_graph <- ggplot(data = airbnb_data, aes(neighbourhood_group)) + geom_bar(aes(fill=neighbourhood_group)) +
labs(title="Airbnb Data for NYC", subtitle="From Kaggle (2019)", x = "NY Borough", y = "# of Airbnbs")
```