Into to R Visualization with Power BI

March 15th, 2017

datasciencedojo
data science for everyone

Who Am I?

- Dave Langer, VP of Data Science – Data Science Dojo

- 20+ years in technology:
  - Roles in development, architecture, & BI/DW/analytics.
  - Last job – Sr. Director, BI & Analytics @ Microsoft.

- Hooked on Data Science 5 years ago:
  - Extensive background in data and analytics.
  - Learned Machine Learning from 2nd place Netflix Prize winner.
  - #1 Data Scientist on YouTube.

- Joined Data Science Dojo to democratize Data Science.

datasciencedojo
data science for everyone

Motivation

- Power BI is core to Microsoft's analytics strategy:
  - Sustained Engineering investment.
  - Cloud, On-Premises, and Hybrid.
  - Growing community of contributors.

- Support for R Visualizations!

Why Power BI?

- Highly productive environment for:
  - Data exploration and analysis.
  - Executive Scorecards.
  - Operational Dashboards.

- Excellent solution for descriptive analytics.

- Power BI Desktop is free*!

* At the time of this presentation.

Expectation Setting

- I am assuming the following:
  - You are familiar with Power BI
  - You are familiar with R
  - You prefer R over DAX

- This is not a Power BI tutorial.
- This is not a R programming tutorial.

- We will cover practical aspects of using R visualizations from within Power BI.

datasciencedojo
data science for everyone

Prerequisites

- To follow along you will need the following:
  - Power BI Desktop
  - R
  - RStudio

- The following R packages are required to follow along:
  - dplyr, lubridate, ggplot2, scales, qcc

- The GitHub repo has source, data, and slide files.

datasc*ence*dojo
data science for everyone

The Scenario

- You are an Analyst for World Wide Importers and have been asked to conduct an analysis on WWI Customers.

- You have access to the WWI Database.

- You have access to Power BI, but are more comfortable using R than DAX in your work.

datasciencedojo

— data science for everyone —

THE DATA

POWER BI DESKTOP

USING R WITH POWER BI

Gotchas

- Power BI limits data to R at 150,000 rows.

- Power BI automatically drops duplicate rows:
  - Leveraging PKs is a good idea!

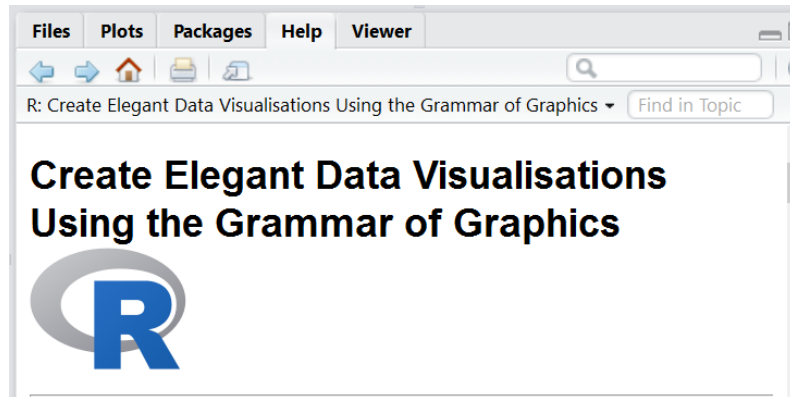- Power BI allows for very permissive column names.

Some Best Practices

- Prefer native Power BI visuals:
  - https://app.powerbi.com/visuals/

- Change columns names to be R-friendly.

- Use R Studio, Visual Studio R Tools, etc. to develop your visuals:
  - Coding & debugging R in Power BI is painful.

datasciencedojo
data science for everyone

GGPLOT2 & DPLYR

ggplot2

- De facto standard visualization package in R.

- Designed for print-quality graphics.

- Fine-grained control via an API focusing on layering graphical element to build visualization.

ggplot2

The collection of data that we're working with.

The aesthetic – how data is mapped to visuals.

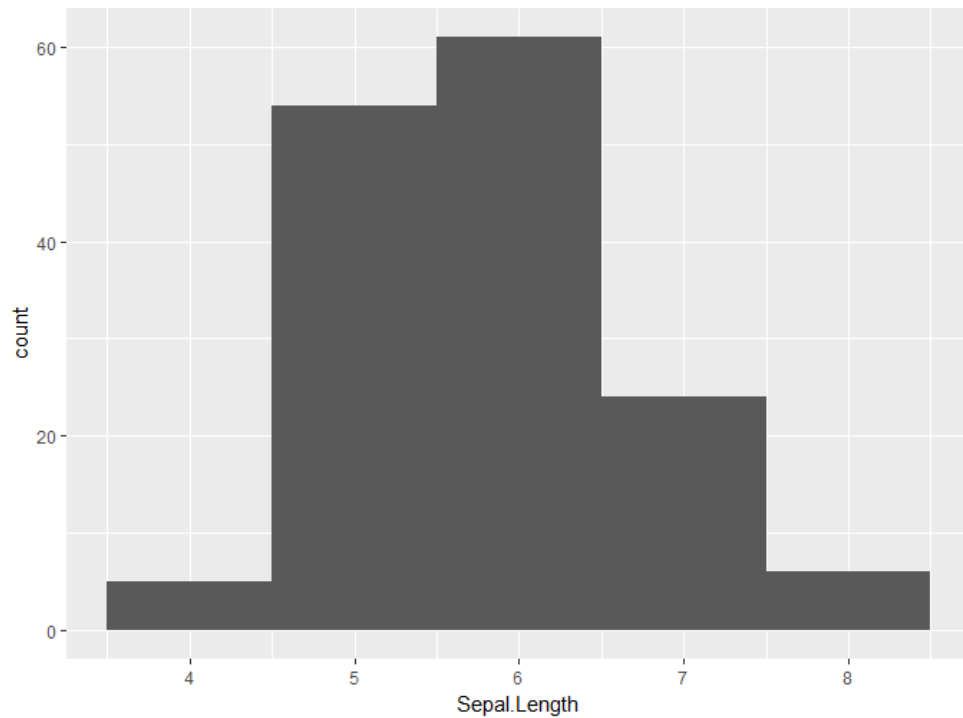Main function – the starting point.

The aesthetic – how data is mapped to visuals.

```
1  library(ggplot2)
2
3  data("iris")
4
5  ggplot(iris, aes(x = Sepal.Length)) +
6      geom_histogram(binwidth = 1)
```

A visual to layer on to the canvas.

Visual parameters – control the look.

datasciencedojo

data science for everyone

ggplot2

ggplot2

Incrementally build more powerful visualizations

```r
library(ggplot2)

data("iris")

ggplot(iris, aes(x = Sepal.Length, fill = Species)) +
  theme_bw() +
  geom_histogram(binwidth = 1) +
  labs(x = "Sepal Length",
       y = "Count of Observations",
       title = "Distribution of Iris Sepal Length by Species")
```

datasciencedojo
data science for everyone

ggplot2

Distribution of Iris Sepal Length by Species

dplyr

- Popular package for data wrangling.

- Similar to Pig – data flows and pipelines.

- Very intuitive to those with familiarity with SQL – just about everything you can do with SQL you can do with dplyr.
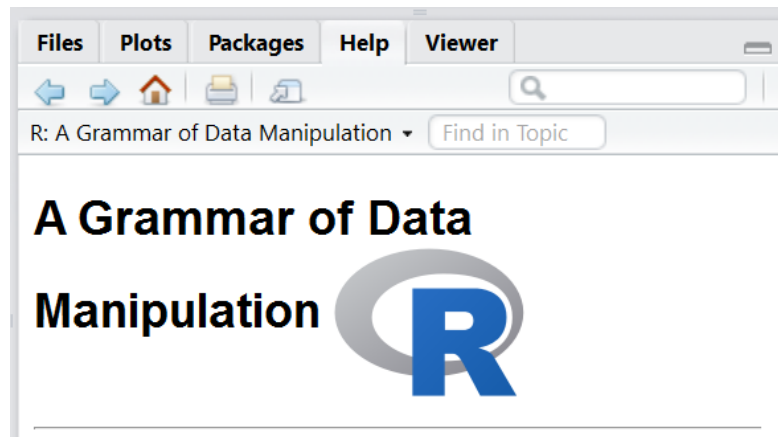
dplyr

The data we're working with.

Pipe data to the next operation.

Filter data down.

Group into categories.

```
14  library(dplyr)
15
16  iris.stats <- iris %>%
17    filter(Species == "setosa" |
18           Species == "virginica") %>%
19    group_by(Species) %>%
20    summarize(Sepal.Length.Min = min(Sepal.Length),
21              Sepal.Length.Max = max(Sepal.Length),
22              Sepal.Length.Mean = mean(Sepal.Length),
23              Sepal.Length.Median = median(Sepal.Length),
24              Sepal.Length.SD = sd(Sepal.Length))
25  View(iris.stats)
```

Summarize groups.

datasciencedojo
data science for everyone

dplyr

	Species	Sepal.Length.Min	Sepal.Length.Max	Sepal.Length.Mean	Sepal.Length.Median	Sepal.Length.SD
1	setosa	4.3	5.8	5.006	5.0	0.3524897
2	virginica	4.9	7.9	6.588	6.5	0.6358796

R CODE!

QUESTIONS

APPENDIX

Get the Files!

- GitHub Repo:
  - https://github.com/datasciencedojo/meetup/tree/master/r_visualization_with_power_bi

- World Wide Importers DB:
  - https://github.com/Microsoft/sql-server-samples/releases/tag/wide-world-importers-v1.0