# An Introduction to Data Visualization with R and ggplot2

datasciencedojo
data science for everyone

# Agenda

- Meet the speaker
- Presentation (60 minutes)
- Question and answer (15 minutes)
- Wrap-up

- GitHub URL:
  https://github.com/datasciencedojo/IntroDataVisualizationWithRAndGgplot2

# Meet the Speaker

- 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.
  - Current interests are text analytics, event log mining, and mathematical programming.
  - Passion for teaching others data science – more tutorials on YouTube!
- Connect with Dave via:
  - LinkedIn
  - YouTube
  - Twitter

# Expectation Setting

- I am assuming the following:
  - You are experienced with R coding - not an expert, but you can hack.
  - You have some data visualization knowledge (e.g., what is a histogram).
  - You are interested in how ggplot2 can accelerate and improve your data visualizations.

- This is a quick intro to data visualization with ggplot2:
  - I will gloss over a lot of things (e.g., multiple layers).
  - The focus will be on the 20% that is useful 80% of the time.
  - More in-depth coverage is available via resources I will mention later.

- My goal is to make you excited about ggplot2!

# THE SCENARIO

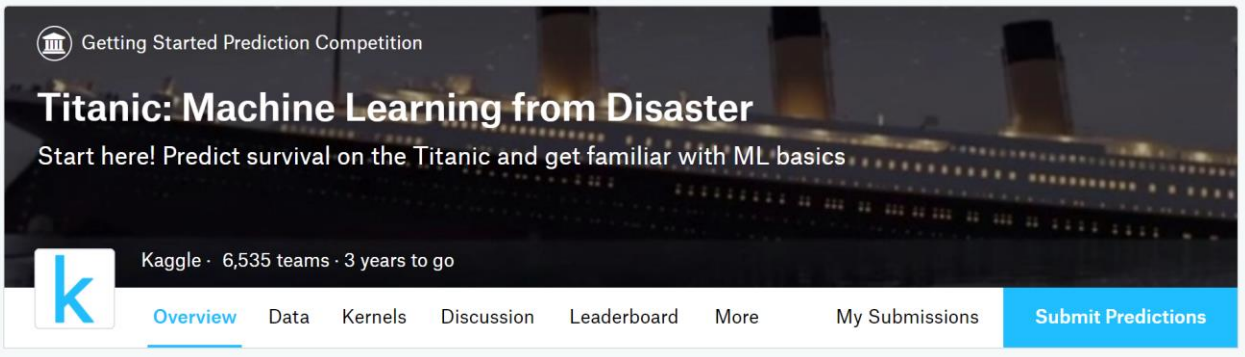
# Prerequisites

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

- You will need the ggplot2 package installed in your R environment.

- The GitHub repo has files for the source, data, and slides.

# The Data



Getting Started Prediction Competition

Titanic: Machine Learning from Disaster

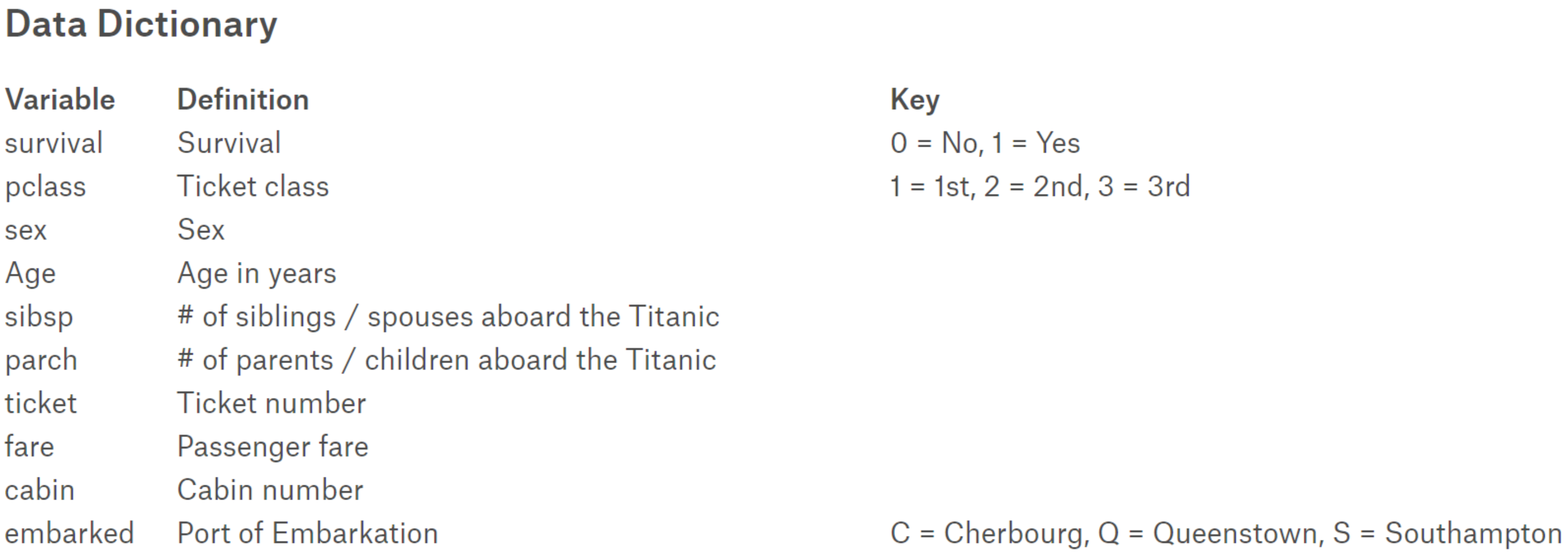
Start here! Predict survival on the Titanic and get familiar with ML basics

Kaggle · 6,535 teams · 3 years to go

Why use this dataset?

1. Everyone is familiar with the problem domain.
2. It is a good proxy for common business data – for example, customer profile data.

# The Data



## Data Dictionary

| Variable | Definition | Key |
|---|---|---|
| survival | Survival | 0 = No, 1 = Yes |
| pclass | Ticket class | 1 = 1st, 2 = 2nd, 3 = 3rd |
| sex | Sex | |
| Age | Age in years | |
| sibsp | # of siblings / spouses aboard the Titanic | |
| parch | # of parents / children aboard the Titanic | |
| ticket | Ticket number | |
| fare | Passenger fare | |
| cabin | Cabin number | |
| embarked | Port of Embarkation | C = Cherbourg, Q = Queenstown, S = Southampton |

# The Scenario

- You are a consulting data scientist and have been hired to analyze the Titanic data.

- The goal of the analysis is to explain patterns of survival in the data:
  - NOTE – The audience is decidedly non-data savvy!

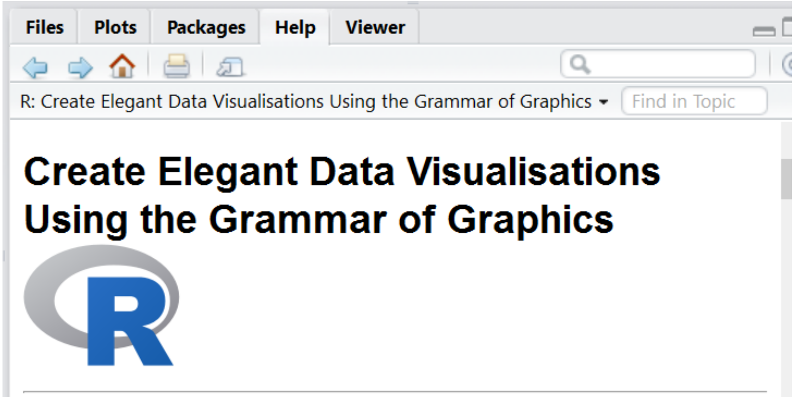
- This scenario has many real world analogs:
  - Customer churn, fraud detection, conversions, etc.

# INTRODUCTION TO GGPLOT2

# ggplot2

- De facto standard visualization package in R.

- Designed for print-quality graphics in seconds.



R: Create Elegant Data Visualisations Using the Grammar of Graphics

**Create Elegant Data Visualisations Using the Grammar of Graphics**

- Fine-grained control via an API (i.e., “the grammar”) for layering graphical elements to build visualizations.

# The Grammar

Every visualization in ggplot2 is composed of the following:

- **Data** – The raw material of your visualization.
- **Layers** – What you see on the plots (e.g., points, lines, etc.).
- **Scales** – Maps the data to graphical output.
- **Coordinates** – The visualization’s perspective (e.g., a grid).
- **Faceting** – Provides “visual drill-down” into the data.
- **Themes** – Controls the details of the display (e.g., fonts).

# Working with the Grammar

While ggplot2 is designed with a rich grammar, using ggplot2 in practice is quite simple. Each ggplot2 visualization has three required components:

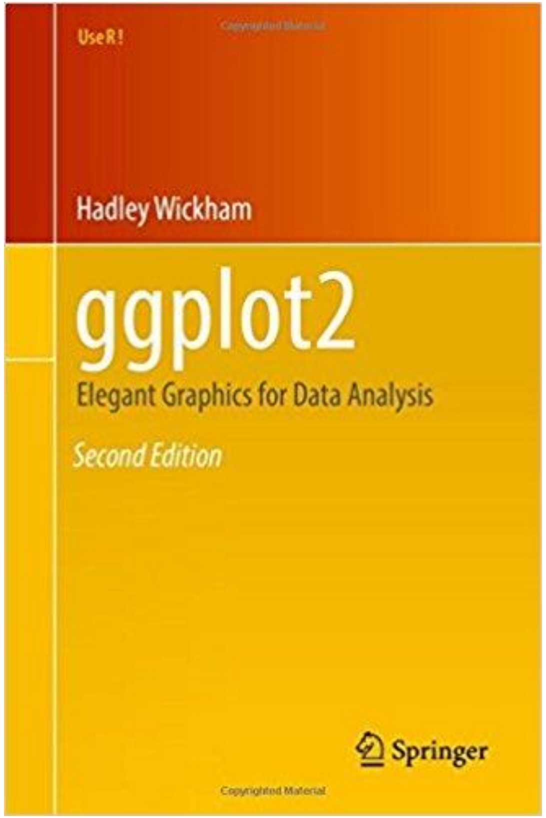
- **Data** – The raw material of your visualization.
- **Aesthetics** – The mappings of your data to the visualization. For example, mapping the value of Titanic passenger ages to the y-axis of a graph.
- **Layers** – A visualization requires at least one layer to render the data and aesthetics to the screen. These layers typically take the form of a ggplot2 *geom* function – for example, a simple scatter plot.

# ggplot2 – The Book

- Single best resource for learning ggplot2.

- Written by the author of the ggplot2 package!

- Excellent introductory resource – good for all skill/experience levels.



# R CODE!

# QUESTIONS

# Want More?

- Follow us Facebook, Twitter, & LinkedIn
- More tutorials available via the Data Science Dojo YouTube channel:
  - https://www.youtube.com/user/DataScienceDojo
- Hear what our students say about our bootcamp:
  - https://datasciencedojo.com/reviews

# THANK YOU!

datasciencedojo

data science for everyone