Introduction to ggplot2

Data Visualization with ggplot2

datasciencedojo

data science for everyone

About Arham Akheel

- Business Analyst at Data Science Dojo
- Started journey in Data Science last year after a career in engineering.
- Masters in Technology Management.
- Data enthusiast, enjoys data sleuthing.

Expectations

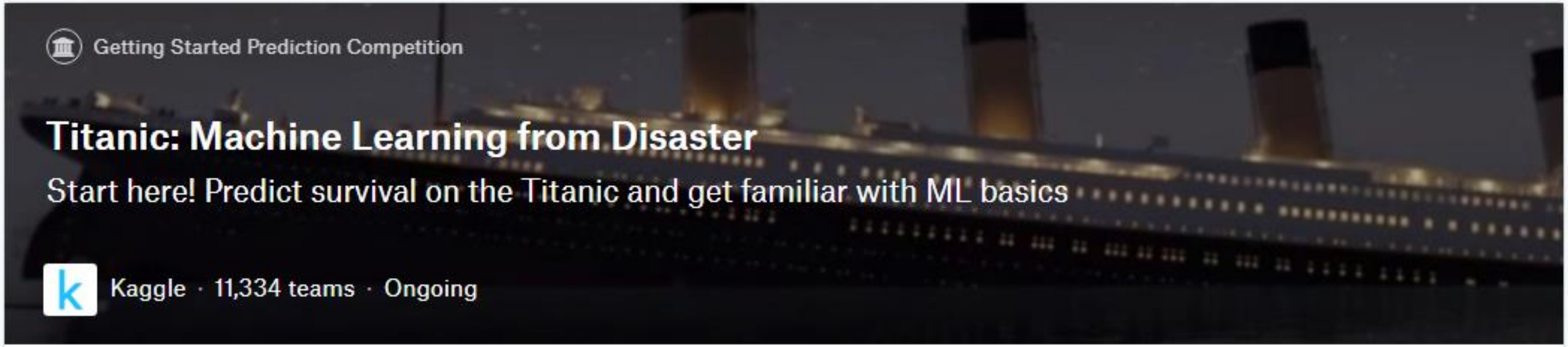
- You are experienced with R coding.
- You have some data visualization knowledge.
- You are interested to improve your data visualization skills with ggplot2.
- Focus will be on the 20% that is useful 80% of the time.

Prerequisites

- Install R & RStudio
- Install ggplot2 package on your R environment.
- The repository on GitHub has files for the source, data and slides.

URL: https://github.com/datasciencedojo/tutorials

The Data



Getting Started Prediction Competition

Titanic: Machine Learning from Disaster

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

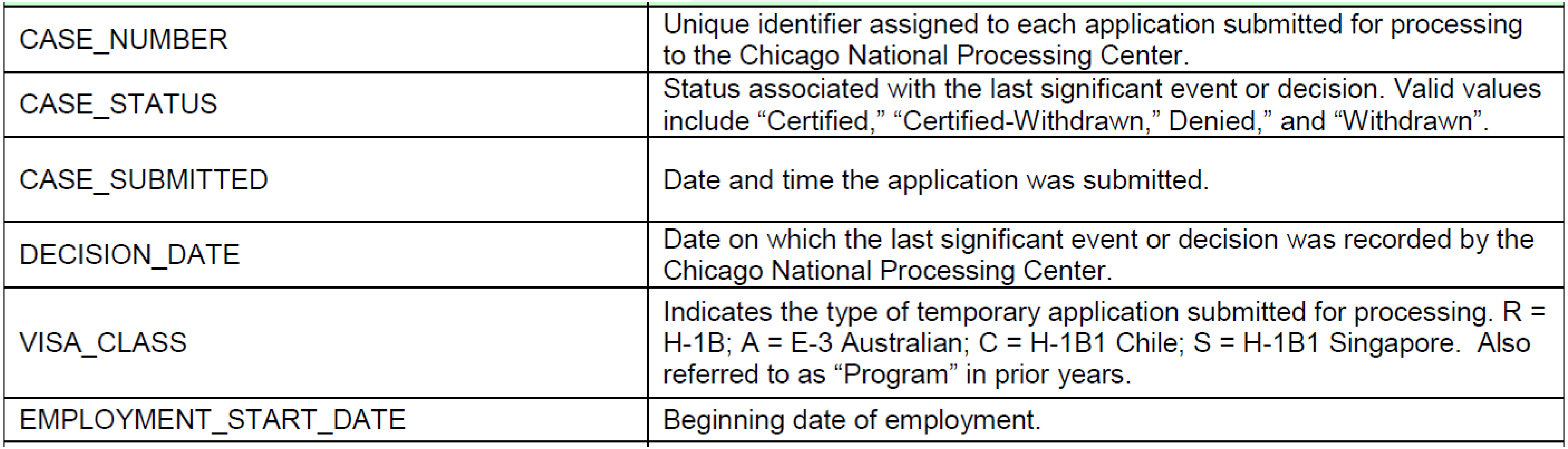
Kaggle · 11,334 teams · Ongoing

Why this dataset?

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

The Data

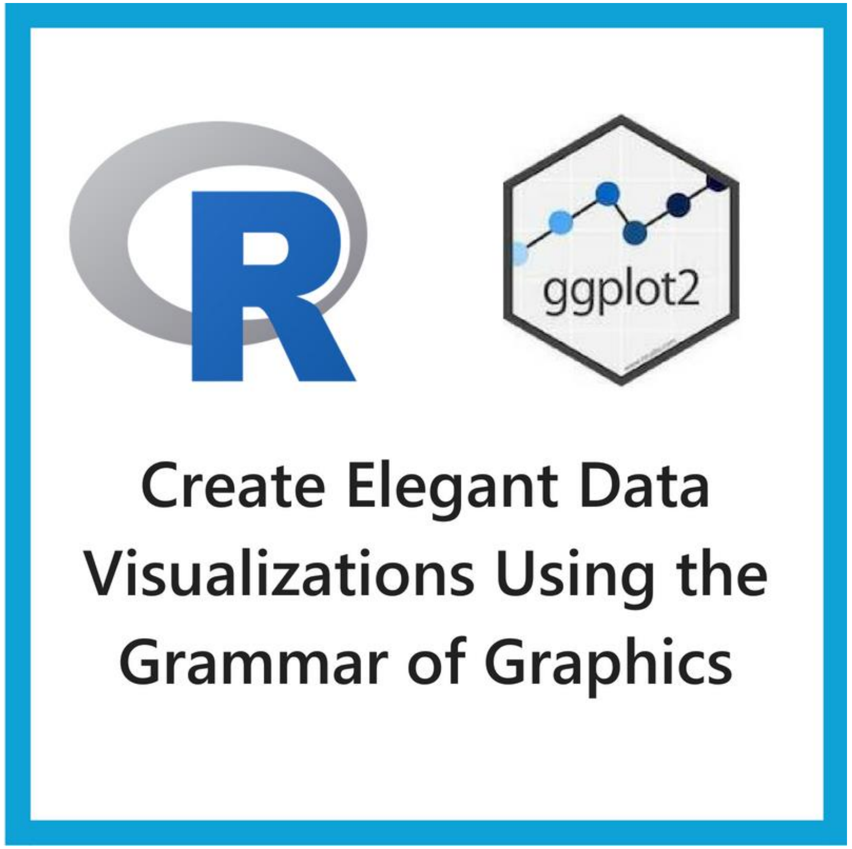
- H1B data from U.S. Department of Labor for 2018.



CASE_NUMBER	Unique identifier assigned to each application submitted for processing to the Chicago National Processing Center.
CASE_STATUS	Status associated with the last significant event or decision. Valid values include “Certified,” “Certified-Withdrawn,” Denied,” and “Withdrawn”.
CASE_SUBMITTED	Date and time the application was submitted.
DECISION_DATE	Date on which the last significant event or decision was recorded by the Chicago National Processing Center.
VISA_CLASS	Indicates the type of temporary application submitted for processing. R = H-1B; A = E-3 Australian; C = H-1B1 Chile; S = H-1B1 Singapore. Also referred to as “Program” in prior years.
EMPLOYMENT_START_DATE	Beginning date of employment.

ggplot2

- Standard visualization package in R
- Designed for print-quality graphics in seconds.
- Fine-grained control via an API for layering graphical elements to build visualizations.



Create Elegant Data Visualizations Using the Grammar of Graphics

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).

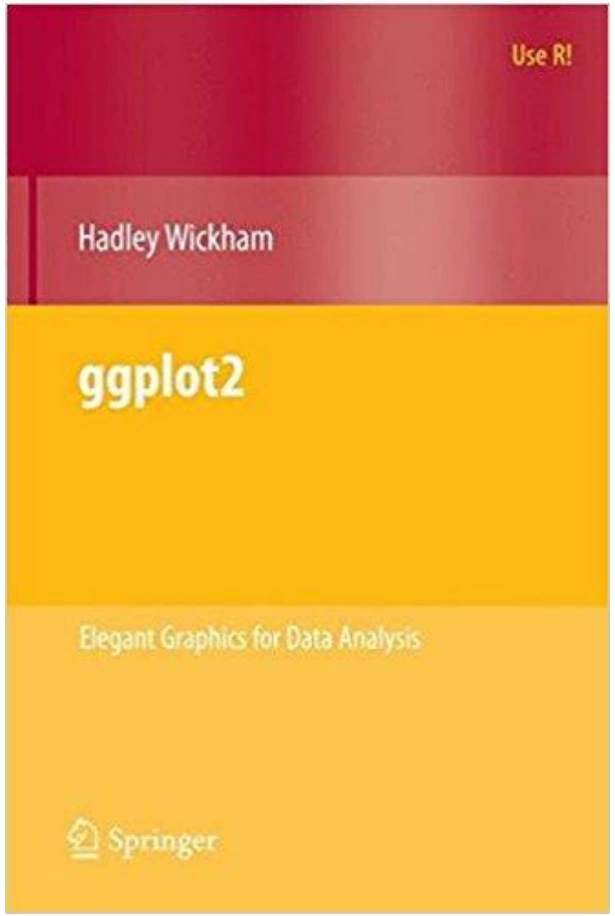
Working with the Grammar

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.
- **Layers** – A visualization requires at least once 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

- Resource for learning ggplot2.
- Written by the author of the ggplot2 package!
- Excellent introductory resource – good for all skill/experience levels.



R CODE!

QUESTIONS

THANK YOU!

Want More?

- Follow us on Facebook, Twitter, & LinkedIn
- More tutorials available on https://tutorials.datasciencedojo.com/.
- Hear what our students say about our bootcamp on https://datasciencedojo.com/bootcamp/reviews/