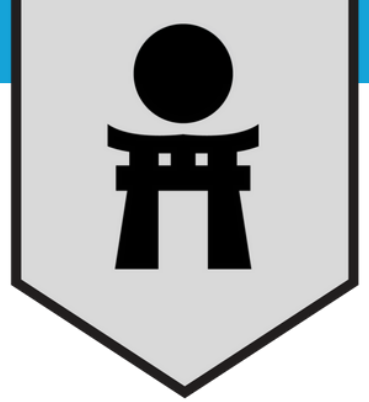


Data Exploration, Visualization, and Feature Engineering

Core R, lattice, ggplot2, and storytelling

datasciencedojo
data science for everyone

Agenda

- Why data exploration and visualization
- Exploration and visualization using R
  - Core R functionality – iris dataset
  - lattice package – mtcars dataset
  - ggplot2 package – diamonds dataset
- Story-telling with data
  - Titanic dataset

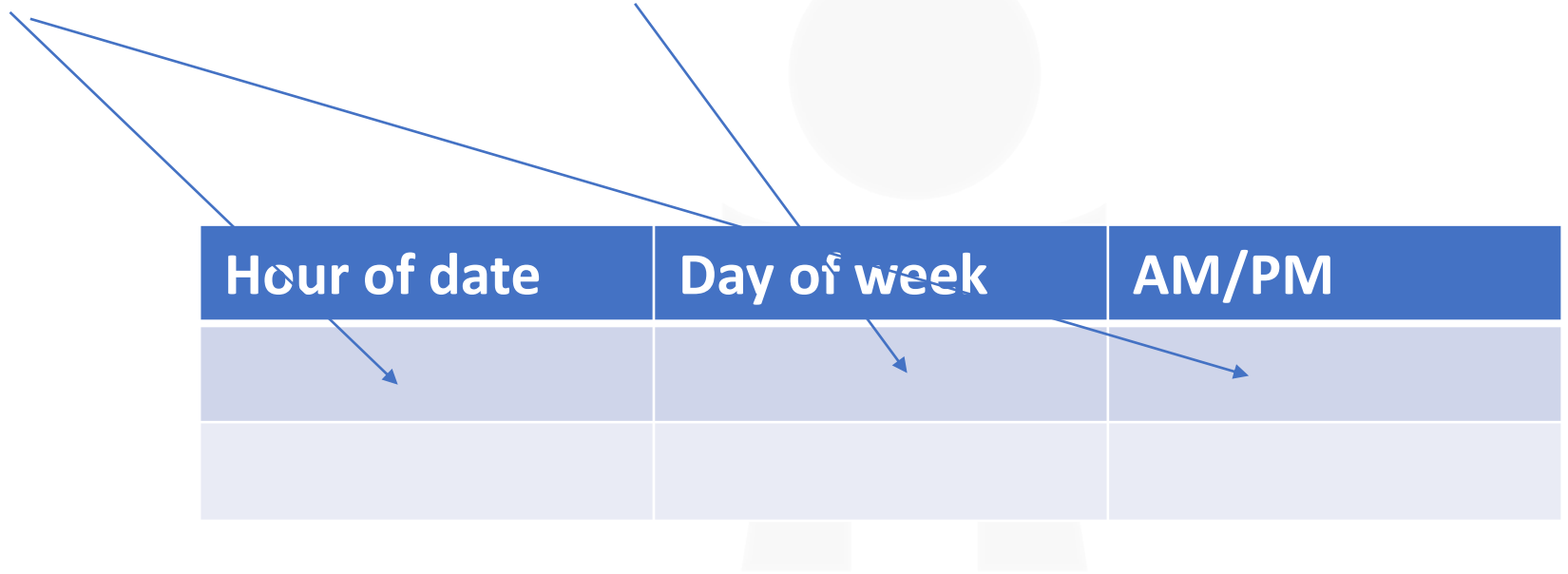
Why Data Exploration and Visualization

Data Beats Algorithm (Sort of)

- More data usually yields good generalization performance, even with a simple algorithm
- However, there are caveats
  - Amount of data may have diminishing returns
  - Data quality and variety matters
  - A decent performing learning algorithm is still needed
  - Most importantly, extracting useful features out of data is important

Is Datetime Stamp a Good Feature?

- 23:05:33 – 5 UTC, April 3, 2014



Dispelling Common Myths

- There is *NO* single ML algorithm that will take raw data and give you the best model

- You do *NOT* need to know a lot of machine learning algorithms to build robust predictive models