# CS 660: Mathematical Foundations of Analytics

Dr. Francis Parisi

Pace University

Spring 2018

# Introduction

The focus of this course is on the mathematical foundations needed to conduct data analysis and the tools for statistical/machine learning

The books for this course are

*Introduction to Mathematical Statistics* $8^{th}$ ed. by Hogg, McKean, and Craig and

*Foundations of Predictive Analytics* by Wu and Coggeshall

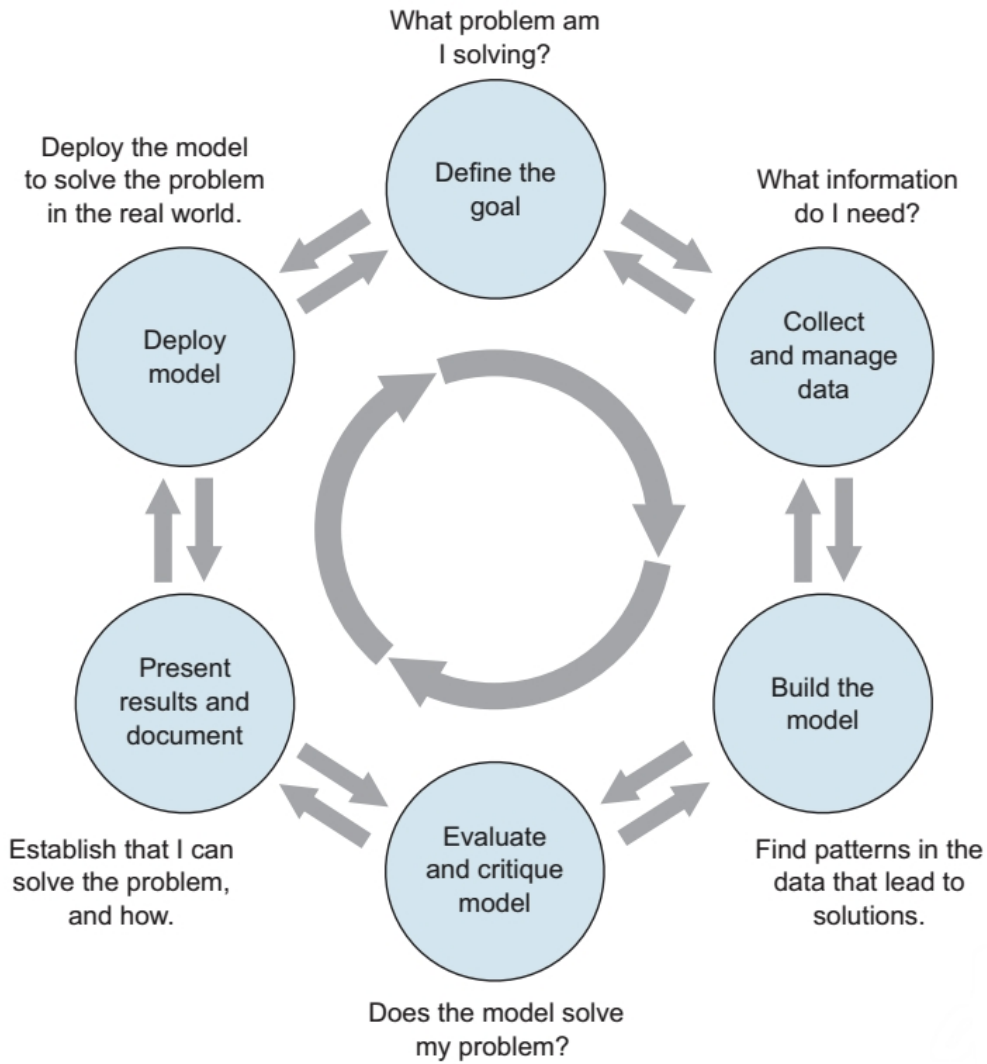
# Course Overview

- Data Science Concepts and Process
- Calculus and Linear Algebra Review
- Probability and Distributions
- Multivariate Distributions
- Special Distributions
- Statistical Inference
- Normal Models, Model Selection and Evaluation
- Non-parametric Methods & Robust Statistics
- Bayesian Statistics
- Time Series Analysis
- Other Methods and Topics - if time permits

# Data Science Concepts and Process

- Data science is more than statistical analysis
- Emphasis on collaboration and project definition
- Project roles
  - Project sponsor
  - Client or SME
  - Data scientist
  - Data architect
  - Operations
- Data science project life cycle . . .

# Data Science Concepts and Process



# Data Science Concepts and Process

- Project goal – why are we doing this?
- Data collection, quality, sufficiency, and management
- Model development
- Model evaluation and sufficiency
- Presentation to stakeholders, project documentation, and reproducibility

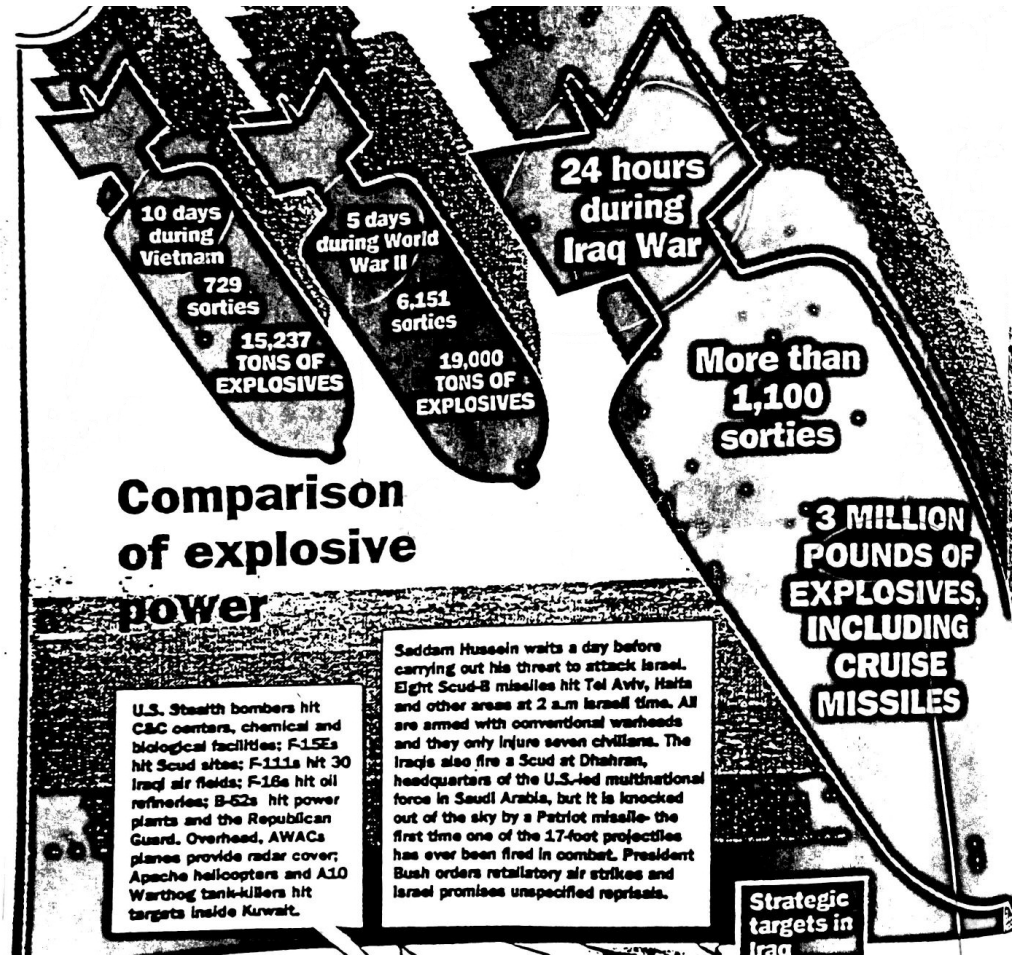
# Data Science Concepts and Process

*Communicating Results*

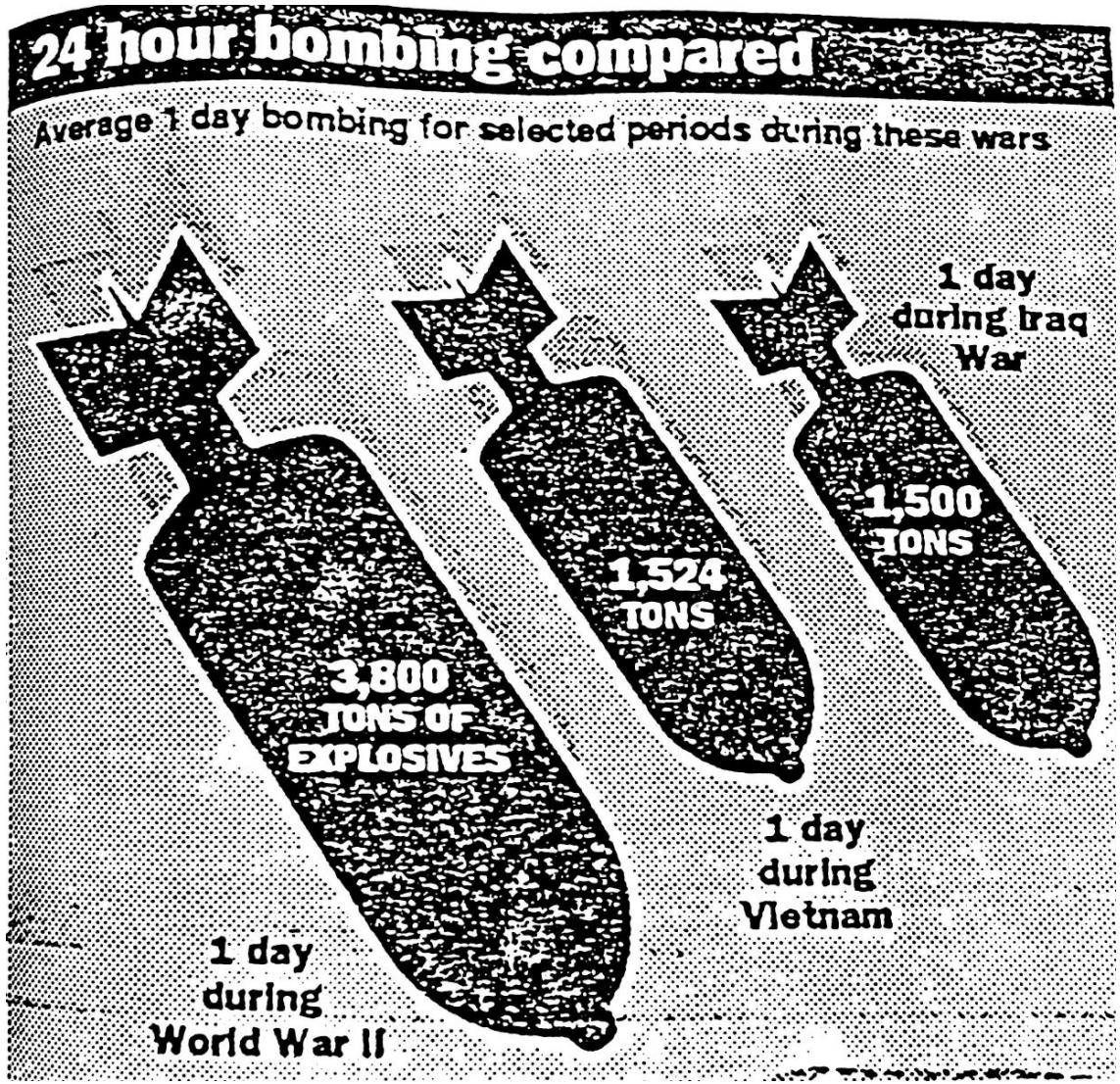
- You're telling a story
- What are the questions you're seeking to answer?
- Why are these interesting questions?

# Data Science Concepts and Process

**Use graphs to clarify not confuse or mislead** What is wrong with this graphic?

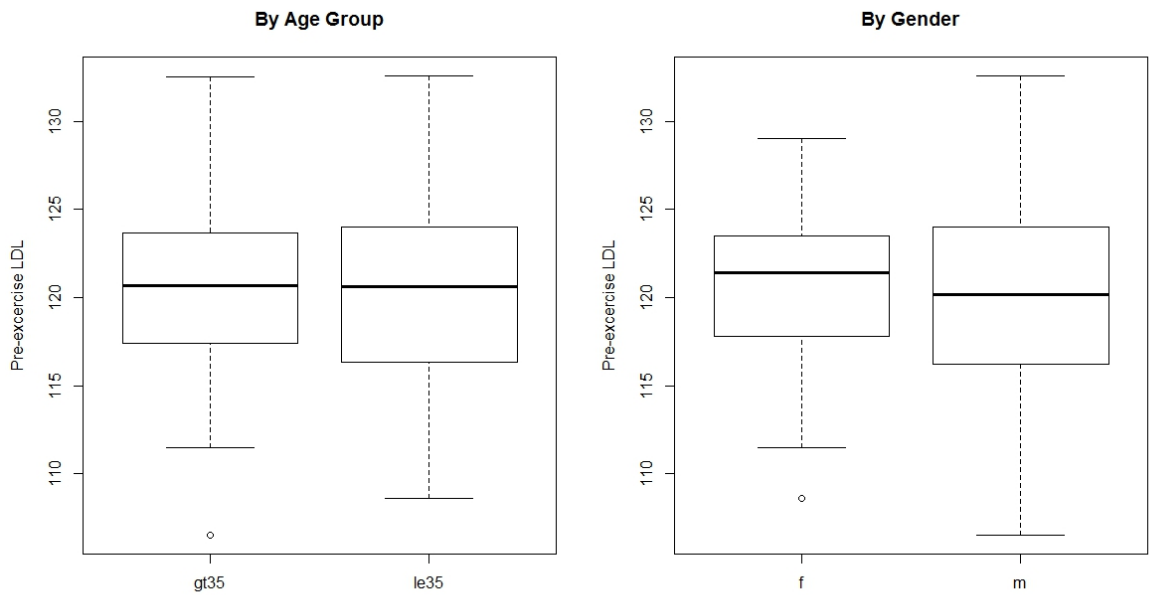


# Data Science Concepts and Process



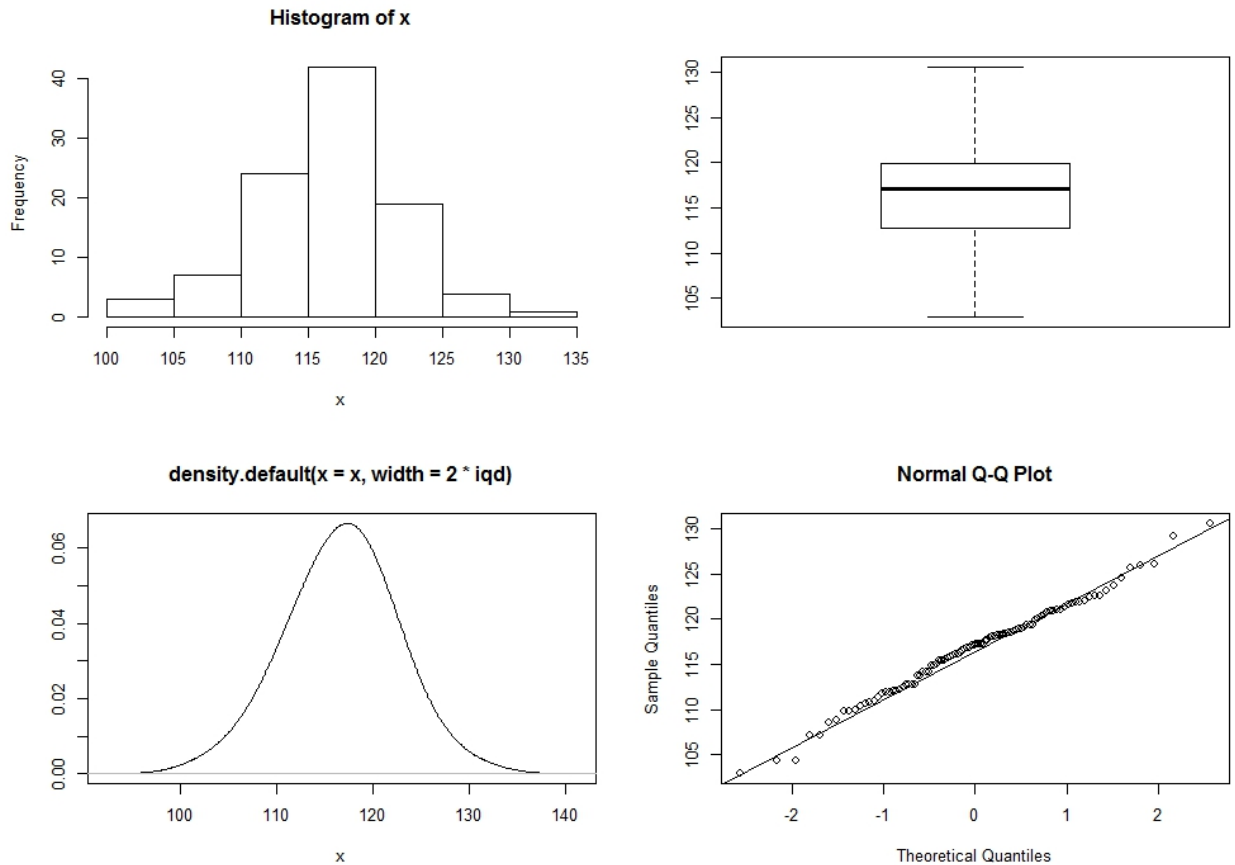
# Exploratory Data Analysis

- EDA allows us to get a sense for what data we have if/and how they are related
- Summary statistics provide a snap shot of the data in a few key measures
- Visualizing data gives us the ability to see what relationships if any exist among the variables



# Exploratory Data Analysis

The histogram and density plot give you the best picture of the distribution shape, while the boxplot and normal qq-plot give the clearest display of outliers.



# Exploratory Data Analysis

- Identifying data problems through data summaries and visualization
  - Missing data
  - Invalid data and outliers
  - Data ranges and comparable units

# Cleaning & Manipulating Data

- Data transformations
  - It is often helpful/necessary to transform data for analysis
  - As we saw earlier, the transformed income data helped in visualization
  - When fitting a linear model a $\log$ transform will reduce nonlinearity
- Invalid data values
- Data ranges and units

# Cleaning & Manipulating Data

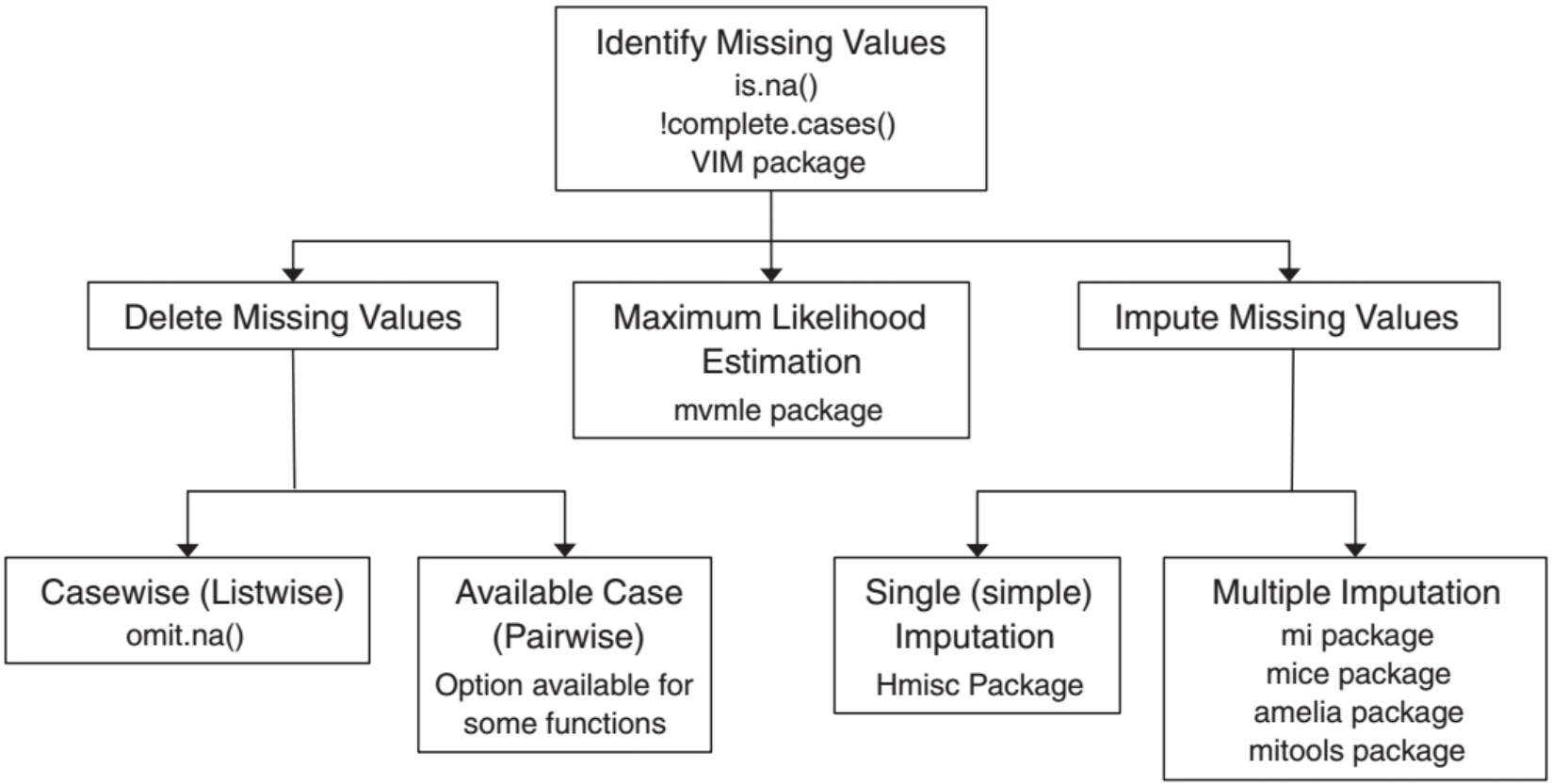
- Handling missing data - just deleting is not always good
- Data may be missing for a number of reasons
- It is important to understand what data are missing and why
  - Identifying missing data
  - Visualizing missing data patterns
- We can remove the observations with missing data (avoid if possible)
  - Complete-case analysis – listwise deletion
  - Pair-wise deletion
- We can replace the missing values
  - Simple imputation
  - Multiple imputation

# Cleaning & Manipulating Data

**Classifying Missing Data**

- Missing completely at random (MCAR)
- Missing at Random (MAR)
- Not Missing at Random (NMAR)

# Cleaning & Manipulating Data




Source: Kabacoff (2015)

# Presenting Results

- Clear communication of results
- Target presentation to your audience
  - Project sponsor or company executives
  - End-users
  - Other data scientists, analysts or researchers

# Presenting Results

**Project sponsor or company executives**

- Summarize the project's goals motivation for doing it
- State the results
- Provide details as needed
- Discuss recommendations or future work

Think Journalistic Style . . .

# Presenting Results

**End-users**

- Summarize the project's goals and motivation for doing it
- Show how the model fits into the users' workflow and *improves* the workflow
- Show how to use the model

# Presenting Results

**Other data scientists, analysts or researchers**

- Introduce the problem
- Discuss related work
- Discuss your approach
- Results and findings
- Discuss future work

This reflects the structure of a research article in a journal

# Building Models

## What is a Model?

- A model is a representation of something
- Models are usually simpler representations of more complicated systems
- Models help us understand the real world, how things interact, or to predict what may happen as things change or evolve
- Algorithmic models (a set of rules and/or equations) can be first principles or statistical
  1. First principles models reflect rules we observe and "hard code"
  2. Statistical models are necessary when the real world is too complex to write down the rules
- In a statistical model we have parameters that we "train" until we find the parameters that best fit our data

# Building Models

**The Modeling Process**

1. Define the goals
2. Gather data
3. Decide the model structure
4. Prepare the data
5. Select and eliminate variables
6. Build candidate models
7. Finalize the model
8. Implement and monitor
9. Finally, *avoid the pitfalls*

# Building Models

**Characteristics of "Good" Modelers**

1. Technical competence
2. Curiosity
3. Common sense and perspective
4. Passion for data
5. Tenacity
6. Creativity
7. Communication skills
8. Ability to work in teams

# References


Hogg, R. V., McKean, J. W. and Craig, A. T. (2018).
*Introduction to Mathematical Statistics*.
Pearson, Boston, eighth edition.

Kabacoff, R. I. (2015).
*R in Action*.
Manning, Shelter Island, NY, second edition.

Lander, J. P. (2014).
*R for Everyone*.
Addison-Wesley, Upper Saddle River.

Wu, J. and Coggeshall, S. (2012).
*Foundations of Predictive Analytics*.
CRC Press, Boca Raton, FL.

Zumel, N. and Mount, J. (2014).
*Practical Data Science with R*.
Manning, Shelter Island, NY, second edition.