# Transforming and Cleaning Data

Matthew Renze

@matthewrenze | www.matthewrenze.com

# Overview

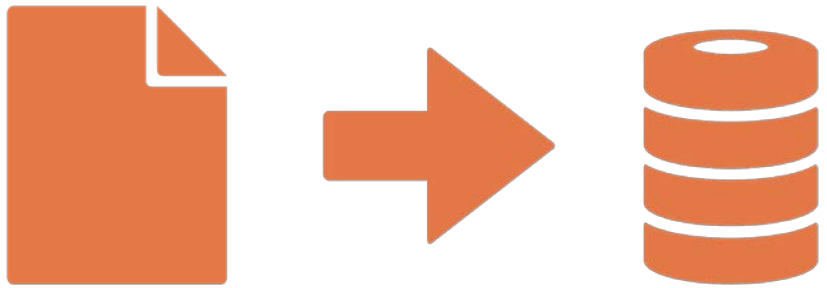
- Introduction
- Loading Data
- Cleaning Data
- Exporting Data
- Demo

# Data Munging

Transforming data

Raw data to usable data

Data must be cleaned first



# Data Munging Tasks

Renaming variables

Data type conversion

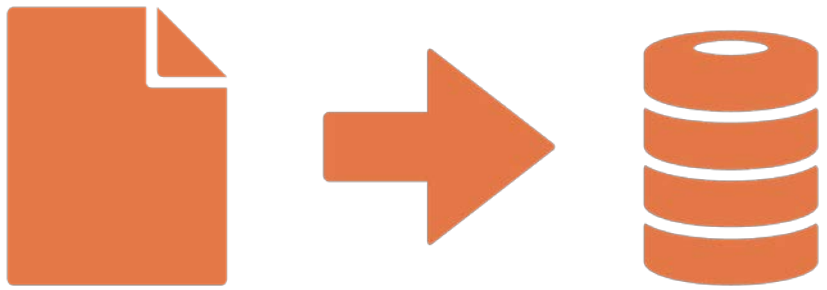
Encoding values

Merging data sets

Converting units

Handling missing data

Handling anomalous data



# Loading Data into R

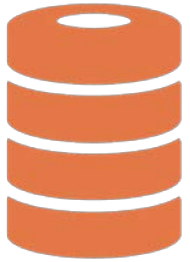
File-based data

Web-based data

Databases

Statistical data





# Cleaning Data

This step is often the:

- Most difficult

- Most time consuming

Record all steps

# Clean Data

Single type of observation

Variables in columns

Column names are readable

Observations in rows

Rows are uniquely identified

| ID | Date | Customer | Product | Quantity |
|---|---|---|---|---|
| 1 | 2015-08-27 | John | Pizza | 2 |
| 2 | 2015-08-27 | John | Soda | 2 |
| 3 | 2015-08-27 | Jill | Salad | 1 |
| 4 | 2015-08-27 | Jill | Milk | 1 |
| 5 | 2015-08-28 | Miko | Pizza | 3 |
| 6 | 2015-08-28 | Miko | Soda | 2 |
| 7 | 2015-08-28 | Sam | Pizza | 1 |
| 8 | 2015-08-28 | Sam | Milk | 1 |

# Clean Data

No errors

No missing values

Properly encoded

Internally consistent

| ID | Date | Customer | Product | Quantity |
|---|---|---|---|---|
| 1 | 2015-08-27 | John | Pizza | 2 |
| 2 | 2015-08-27 | Johnx | Soda | 2 |
| 3 | 2015-08-27 | Jill | Salad | 1 |
| 4 | 2015-08-27 | Jill | | 1 |
| 5 | 2015-08-28 | Miko | Pizza | 3 |
| 6 | 08/28/2015 | Miko | Soda | 2 |
| 7 | 2015-08-28 | Sam | Pizza | 1 |
| 8 | 2015-08-28 | Sam | Milk | 1.5 |

# Clean Data

| ID | Date | Customer | Product | Quantity |
|---|---|---|---|---|
| 1 | 2015-08-27 | John | Pizza | 2 |
| 2 | 2015-08-27 | John | Soda | 2 |
| 3 | 2015-08-27 | Jill | Salad | 1 |
| 4 | 2015-08-27 | Jill | Milk | 1 |
| 5 | 2015-08-28 | Miko | Pizza | 3 |
| 6 | 2015-08-28 | Miko | Soda | 2 |
| 7 | 2015-08-28 | Sam | Pizza | 1 |
| 8 | 2015-08-28 | Sam | Milk | 1 |

# Exporting Data

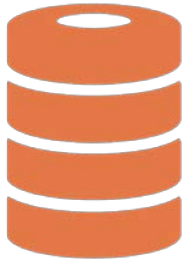
File-based data

Web-based data

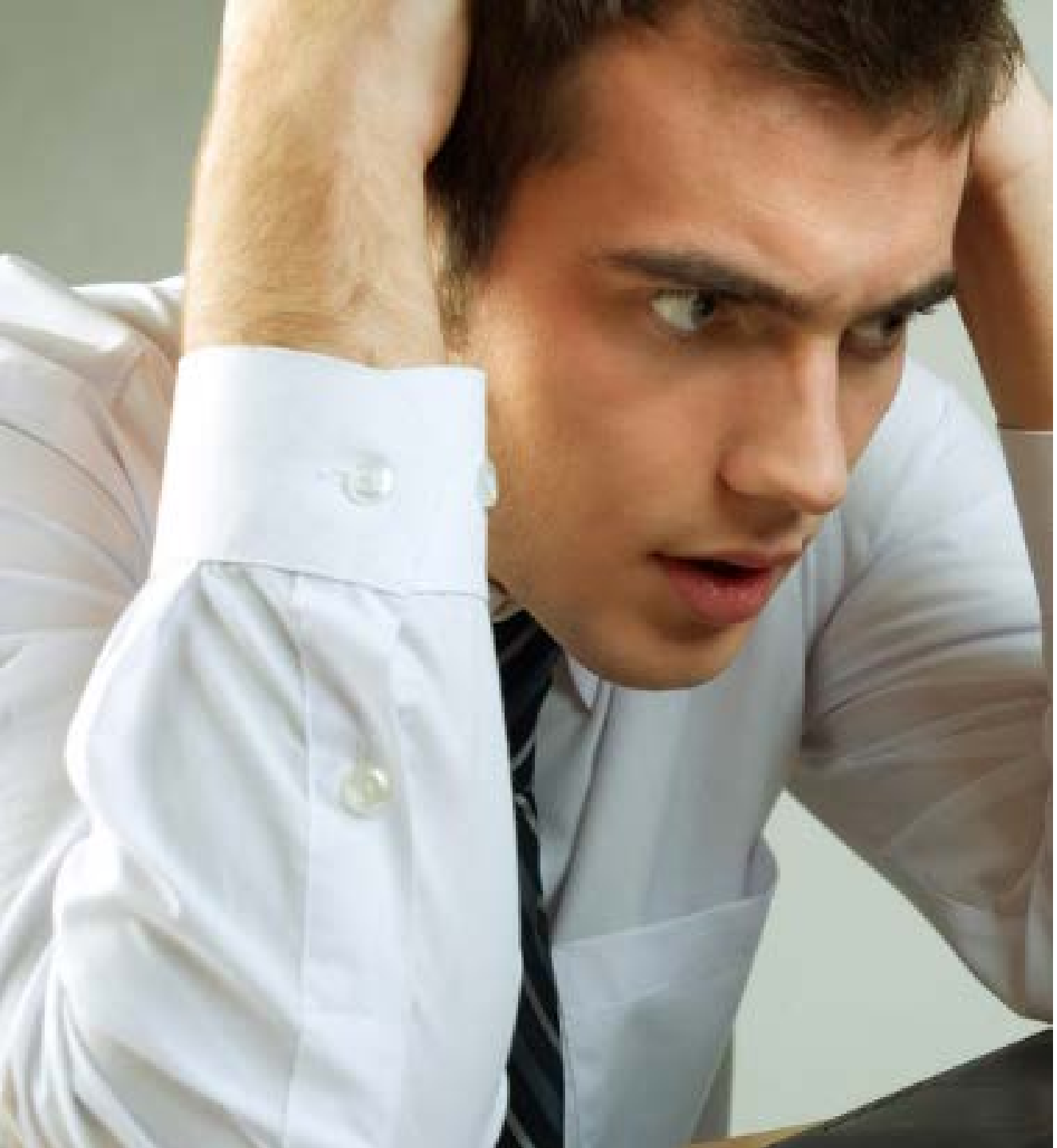
Databases

Statistical data






PROD. NO.
SCENE
TAKE
ROLL



Column with wrong name

Rows with missing values

Runtime column has units

Revenue in multiple scales

Wrong file format

# Source File

# Set Working Directory

# Load the Data

# Inspect the Data

# Inspect Column Names

# Problem 1: Column Name

# Problem 2: Missing Values

# Problem 3: Remove Units

# Problem 4: Multiple Units

# Export File

# Target File

# Summary

Introduction

Loading Data

Cleaning Data

Exporting Data

Demo