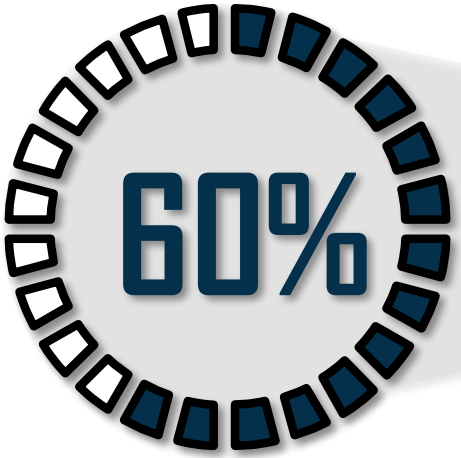
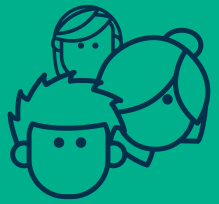
Hands-on Workshop: SAS® Data Cleansing Challenge

Welcome!





Data scientists spend **60%** of their time on *cleaning* and *organizing* data.

- CrowdFlower 2016 Survey of Data Scientists

SAS Data Cleansing Challenge

- Scenario
- Data and Desired Outcome
- The Challenge and Resources
- Get to Work!

Scenario

- You are a SAS programmer for the National Oceanic and Atmospheric Administration (NOAA).
- Your job is to prepare clean data sets for the analysts.

New project:

- Clean the **Earthquakes_dirty** data set.
- Create a ***valid*** data set and an ***invalid*** data set.

DATA

1. Earthquakes_dirty

Data Information

- Global Significant Earthquake Database (NOAA)
- National Geophysical Data Center / World Data Service (NGDC/WDS): Significant Earthquake Database. National Geophysical Data Center, NOAA.
- Data has been imported into SAS.

- 17 Variables
- 2,468 Observations

Earthquakes_dirty Overview

Variables:

- ID_REGIONCODE
- FLAG_TSUNAMI
- YEAR, MONTH, DAY, HOUR, MINUTE, SECONDS
- FOCAL_DEPTH
- EQ_MAG (6)
- COUNTRY and LOCATION_NAME

DESIRED
OUTCOMES

1. Clean the data.
2. Create a valid data set and an invalid data set.
3. Bonus

Clean the Data

ID and REGION_CODE

Create an **ID** variable and a **REGION_CODE** variable from **ID_REGIONCODE**. Drop the original variable.

Earthquakes_dirty

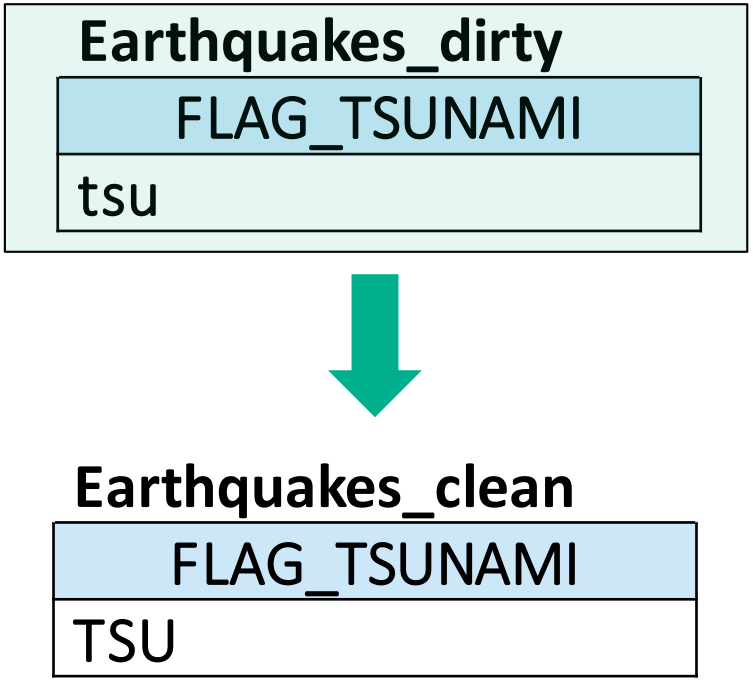
ID _REGIONCODE
3931-140

Earthquakes_clean

ID	REGION_CODE
3931	140

FLAG_TSUNAMI

All character values of **FLAG_TSUNAMI** should be converted to uppercase.



DATE_TIME

Create a SAS datetime variable named **DATE_TIME** from the date and time variables in the **Earthquakes_dirty** data set. Drop the original variables and format **DATE_TIME** using the SAS DATETIME format.

Earthquakes_dirty

YEAR	MONTH	DAY	HOUR	MINUTE	SECONDS
1950	1	19	17	27	0

Earthquakes_clean

DATE_TIME
19JAN1950:17:27:00

EQ_PRIMARY

Create an **EQ_PRIMARY** variable that selects the first nonmissing value from the **EQ_MAG** measurement variables in the following order. Drop the original variables and format **EQ_PRIMARY** with one decimal point.

Earthquakes_dirty

EQ_MAG_MW	EQ_MAG_MS	EQ_MAG_MB	EQ_MAG_ML	EQ_MAG_MFA	EQ_MAG_UNK
.	7	7.3	.	.	7.2

EQ_PRIMARY

Create an **EQ_PRIMARY** variable that selects the first nonmissing value from the **EQ_MAG** measurement variables in the following order. Drop the original variables and format **EQ_PRIMARY** with one decimal point.

Earthquakes_dirty

EQ_MAG_MW	EQ_MAG_MS	EQ_MAG_MB	EQ_MAG_ML	EQ_MAG_MFA	EQ_MAG_UNK
.	7	7.3	.	.	7.2

EQ_PRIMARY

Create an **EQ_PRIMARY** variable that selects the first nonmissing value from the **EQ_MAG** measurement variables in the following order. Drop the original variables and format **EQ_PRIMARY** with one decimal point.

Earthquakes_dirty

EQ_MAG_MW	EQ_MAG_MS	EQ_MAG_MB	EQ_MAG_ML	EQ_MAG_MFA	EQ_MAG_UNK
.	7	7.3	.	.	7.2

EQ_PRIMARY

Create an **EQ_PRIMARY** variable that selects the first nonmissing value from the **EQ_MAG** measurement variables in the following order. Drop the original variables and format **EQ_PRIMARY** with one decimal point.

Earthquakes_dirty

EQ_MAG_MW	EQ_MAG_MS	EQ_MAG_MB	EQ_MAG_ML	EQ_MAG_MFA	EQ_MAG_UNK
.	7	7.3	.	.	7.2

Earthquakes_clean

EQ_PRIMARY
7.0

Desired Output **Earthquakes_clean**

Eight Variables

Partial **Earthquakes_clean**

ID	REGION _CODE	FLAG _TSUNAMI	DATE_TIME	EQ _PRIMARY	FOCAL _DEPTH	COUNTRY	LOCATION _NAME
3931	140		19JAN1950:17:27:00	.	.	IRAN	IRAN: BUSHIRE

Create a Valid and an Invalid Data Set

Determining Invalid Observations

- Duplicate **ID**
- Invalid **REGION_CODE**
- **FLAG_TSUNAMI** must be blank or *TSU*
- Missing **DATE_TIME**
- Missing **EQ_PRIMARY** or out of the ***defined range***
- Missing **FOCAL_DEPTH** or out of the ***defined range***

Bonus

Using the **Invalid** data set, create a new variable named **INVALID_DESCRIPTION** to determine the reason (or reasons) for an invalid observation.

Partial **Invalid**

DATE_TIME	EQ _PRIMARY	FOCAL _DEPTH	COUNTRY	LOCATION_NAME	INVALID_DESCRIPTION
19JAN1950:17:27:00	.	.	IRAN	IRAN: BUSHIRE	EQ Primary, Focal Depth
02FEB1950:19:33:00	7.0	.	CHINA	CHINA: YUNNAN PROVINCE	Focal depth

The Challenge and Resources

Files ⇨ C:\Workshop\Challenge\DataCleansing

Challenge Overview

01 **Cleansing Challenge**

Open the starter program.

02 **Earthquakes_clean**

Clean the data.

03 **Earthquakes_valid and Invalid**

Create the final data sets.

SAS Data Cleansing Challenge Document

- Introduction
- Data Layout
- Challenge Issues
  - Be sure to answer the validation questions at the end of the section by running the provided validation code at the bottom of the **Cleansing Challenge.sas** program.
- Challenge Hints (*HTML file in your challenge folder*)
- Suggested Answer

Recap

- Using any interface, clean the **Earthquakes_dirty** SAS data set.
- Create an **Earthquakes_valid** data set and an **Invalid** data set from the cleaned data.
- Files ⇨ **C:\Workshop\Challenge\DataCleansing**.
- You can use the **Challenge Issues** section as a guide to clean the data set.
- If you are stuck, use the **Challenge Hints** section or the HTML file.
- Answer the questions at the bottom of the **Challenge Issues** section of your document to participate in the end-of-class trivia!

SAS Data Cleansing Challenge

Download the following:

- Challenge PDF
- Data Set
- SAS Program