Data Wrangling: Walkthrough with a practical example

# Data Wrangling definition

Wrangling is just another fancy word for data preparation. It is the processing of data to make further exploration and analysis straightforward and less error prone. Sometimes it is as simple as opening an Excel formatted file in MS Excel, and you are well on your way. Most times, as we will surely experience, it is not that simple. Most data in the world are inherently messy or ‘untidy’, if you will. Sometimes, the data is arranged in rows, but you want them in the form of columns to make the next step of your analysis possible. Most times, this is just one step of the metaphorical staircase that is data cleaning.

Here we will follow a typical data wrangling framework for cleaning some part of a sample dataset. This is a real-world dataset, as you will find out soon, that will help us in getting our hands dirty with the first step of the complete data analysis process.

# Background of dataset used

The dataset that we are using can be found on Kaggle. It contains over 19000 job posting in the period between 2004 to 2015. The data was mined using text mining techniques from an Armenian HR portal. Though the heavy lifting of data wrangling is already been done, it still needs a little work to make it useable. It should be noted that our objective is not to completely and exhaustively clean the dataset, but to get a first hand taste of what data cleaning is and how it can be done in the real world. The objective is to use a consistent and reproducible framework deliberately designed to aid in systematizing data wrangling.

# Framework for data wrangling

# Gather

# Assess

# Clean

## Define

## Code

## Test