# Wrangle Report

## Data Analysis Nano Degree – DAND

**Prepared by:**

Abdulmajeed Alharbi

alharbimabd@gmail.com

Data Analysis Nano Degree Program – 23 July of 2020

# Data wrangling definition

Data wrangling is one of the most important process in data analysis, data wrangling make your data and result very clear and valid, most of analyst spend 80% of his effort on data wrangling because it's take more time and have many changes. data wrangling is three main things: Gather, assess and cleaning.

In this course we learning how to Gather the data from multiple type of sources, for example ( HTML,CSV,TSV and JSON), all these type of data have special way to gathering it in your workspace, we need to know how to gather data from multiple sources because of sometimes our data storing in different format and comes as unstructured data, not always our data become as excel sheet or databases, it's depends in your business or your question.

# Data wrangling Project

This project we did was about wrangling a Twitter archive containing about more than 5000, the goal of the project was to take the current data and make it good, clean and trustworthy so we can perform analyses and visualizations.

In this project I divided the work as 3 main steps as well we learned:

1. **Gathering**: we gathering data from multiple format ( HTML, CSV and JSON ) and try to fetch it into data frame in python, also we using tweepy querying the twitter API and we use also the request to downloading the image dataset.

2. **Assessing**: after gather all data into data frame we assessing our data by trying to find mistakes or issue, we have to find 8 quality issue and 2 tidiness issue, we can assessing our data visually by looking into the data and see it to find something interesting or programming by using panda functions lilke info(), describe()…etc, then wee noted the issue and make it clear for us ( document it ) to make it easy to use it in next step.

3. **Cleaning**: The last and important steps in data wrangling is to clean the data, after we assessing our data visually and programming and noted the issues, here we will make a 3 steps to cleaning the data: define means define the issue that we noted before and make it as task or steps, code means solve the issue and write your own code and last step is test means evaluate and check your job and correct it if still not resolved.