# How to conduct data science project based on a business context

*Presented by Jiaqiang Yi*

*github: https://github.com/Datajacker/fuel_efficiency_analysis*

*LinkedIn: https://www.linkedin.com/in/jiaqiang-yi/*

# Self-Introduction

Education:

***MSc in Chemical Engineering*** Sep 2017 – Nov 2020
University of Alberta, Edmonton AB

***BSc* in Chemistry** Sep 2013 - Jun 2017
Renmin University of China, Beijing CN

Data Science:

***Data Science Certificate*** Winter 2020
NAIT, Edmonton AB

***Data Scientist with Python*** Dec 2018 – Oct 2019
DataCamp, Online

# General process of a data science project

https://blog.dataiku.com/2019/07/04/fundamental-steps-data-project-success

- Define the problem
- Collect the data
- Clean the data
- Enrich the data
- Find insights and visualize
- Machine learning model
- Iterate

# Define the problem

- 1. Which manufacturer produces the most fuel-efficient fleet of cars (type_1 & type_2)?

- 2. Build a model to predict city mpg (variable “UCity” in column BG).

Some questions can be answered without machine learning modelling

# Collect the data

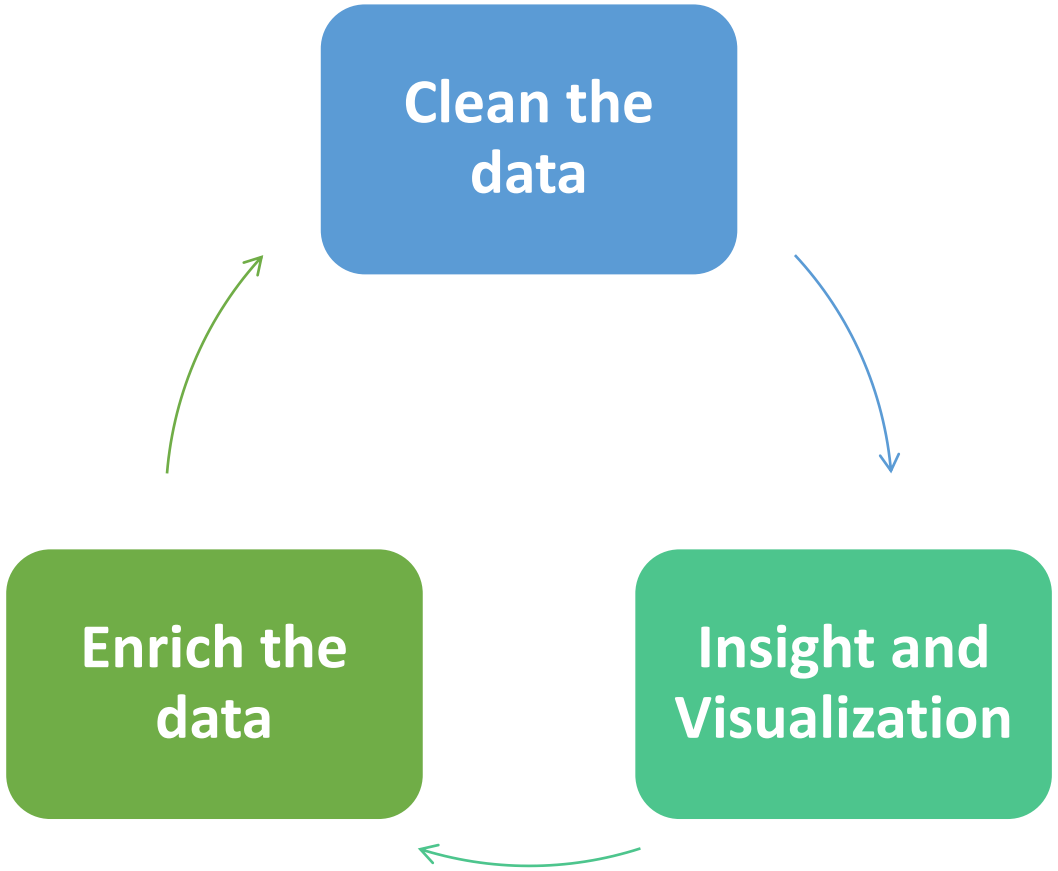
Data source: Vehicle

The fuel economy data is directly taken from the Fuel Economy office from the U.S. Department of Energy.

A shape of 40081 * 83

Data format:

- Bool: 1
- Float64: 31
- Int64: 27
- Object: 23

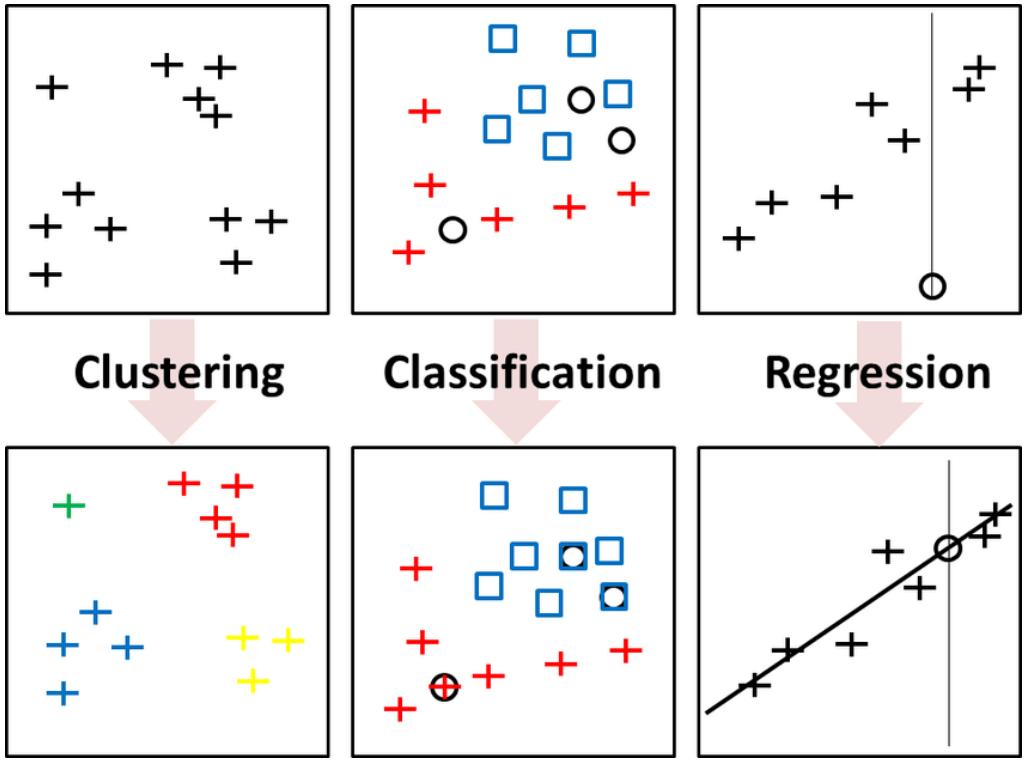



**Methods to clean data**

1. **Remove Irrelevant Values**
2. **Get rid of Duplicate Values**
3. **Avoid Typos (and similar errors)**
4. **Convert Data Types**
5. **Take Care of Missing Values**

Based on the **business context**

# Machine learning model




scikit learn

TensorFlow

Keras

PyTorch

# Iterate

Define the problem

Collect the data

Clean the data

Enrich the data

Find insights and visualize

Machine learning model

# Thank you