# Data cleansing

The actual process of data cleansing involve removing typographical errors and correcting values against a known list of entities.

The process may be strict, or fuzzy (such as correcting records that partially match existing, known records).

There are also

1. Data enhancement: adding data to make it more complete.
2. Harmonization: bringing together data of varying file formats, naming conventions, and columns.

# Customer Demographic

## ID

No duplication of ID is found.

No ID is missing.

## DOB

Customer ID (34): DOB is 1843-12-21, which is not valid, because his DOB makes him 177 years old.

There are 86 customers whose DOB value is missing.

## GENDER

There is one person with F instead of Female.

There are 88 people with no gender data associate with them.

All the customers with the gender undetermined also do not have their DOB data, except for customer ID 34, who also has his DOB wrong.

## Default values

It is corrupted by shady players.

For example: for user id = 3958, default value is **1;DROP TABLE users**, which is SQL injection attack.

## Job title

There are 505 people with job title missing. Other data seem to be reliable.

## Job industry category

There are 655 people with n/a as job industry category.

## Past 3 years bike related purchases

Seem to be correct.

## Tenure

There are no meta data explaining what the field indicated.

There are also 86 customer data missing.

Data quality dimensions:

* Correct values
  + Need to use zip code to look up external databases to help match up zip codes to geographical locations and also help verify that street addresses within these zip codes actually exists.
* Data fields with values
  + Should put data constrain on the data input.
* Values free from contradiction
  + The field of first date purchased shows contradiction
* Values up to date
* Data items with value meta-data
* Data containing allowable values
* Records that are duplicated
* Inconsistency: decide what data source to keep. (for example, the more up-to-date ones)

Accuracy, completeness and duplicate values

# Customer address

## State

The states are not consistent. Some are full names, and some are abbreviations.

## Property valuation

The property valuations do not have units associated with them.

## Customer ID

There are 3 customer ID missing.

# Transaction

## Transaction ID

There are no transactions missing.

## Product ID

The product ID doesn’t seem to correspond well with the item brand and product line.

For example, produce 100 correspond to Norco Bicycles and Trek Bicycles.

## Customer ID

Customer ID is fine, there is no missing IDs.

## Transaction date

Everything is OK with transaction date as well.

Order dates start from 1/1/2017 to 12/30/2017

## Online purchase

There are about 359 order missing whether the order is online information.

## Brand

There are 196 purchases with **Brand, product line, product class, product size, standard cost and product first sold date** information missing.

## Product first sold date

This column is not accurate as well.

Same product can end up having two different first sold dates.