

Data Wrangling Report

The main goal of this project is to transform contractor and state_lookup table into contractor with the following

Column	DataType	Constraint
contractor_id	INT	PRIMARY KEY,
contractor_bus_name	VARCHAR(200)	NOT NULL UNIQUE,
address1	varchar(200)	NOT NULL,
address2	varchar(200)	NULL,
city	varchar(200)	NOT NULL,
state_code	char(2)	NOT NULL,
zip_part1	char(5)	NOT NULL,
zip_part2	char(4)	NULL

column contractor_id is primary key, column contractor_bus_name has "UNIQUE" constraint. address1, city, state_code, and zip_part1 column have "NOT NULL" constraint. address2 and zip_part2 columns allows NULL. Though accessing the destination table definition, I made some key action plan as following:

- 1.zipcode column should be spilited into zip_part1 and zip_part2
- 2.state_code should be fetched by joining the table
- 3.access column contractor_bus_name to see if value are unique
- 4.check missing value in address1 column

Data Quality Issue:

1. Column city and address1 have missing data and in destination table definition, their column restraint are "Not NULL". These information can be fetched from the website listed .
2. Some zip code are 9 digit and some zip code are 5 digits only. Based on the defition of destination table Zip_code columns should be transformed into two column - zip_part1 and zip_part2.
3. Last_updated column should be transformed to datetime format since the format from raw data is character.
4. Identify duplicated record by contractor_id, and keep the record with contractor_version =2 for contractor_id in (139,140,228,236,238). Contractor_id is the unique identifier, assigned for table for each contractor and contractor_number is real unique identifier for each contractor.
5. For duplicated records by contractor_number and contractor_number 11004,11201,11202, they have same information,but last_updated datetime are different. In this case, I will only keep the rows with most recent

will only keep the rows with most recent last_updated as datetime format and remove row with contractor_id in (373,374,378).

6. Remove 'Attn:' in address1 columns and replace '-' with blank value.

7. Row 122-129' with contractor_id in (382,383,384,385,386,387) have unrecognized symbol '?' in column contractor_bus_name. This symbol should be removed, or it will cause data error when loading data to sql database.

Data Tidiness Issue:

1. Column fax, address3, email and ignore should be dropped because these columns are empty and not used in destination table.

2. Join the contractor table and state_lookup table by state_id to get state code for each contractor. Then, rename state_abbrev from state_lookup table to state_code, and drop unnecessary columns.

3. Combine original contractor_bus_name and contractor_number column, separated with '-', combine them into a unique contractor_bus_name per record.