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 tab
le definition, their column constraint are
"NOT NULL". These information can be fetched from the website listed
in the records.
2. Some zip codes are 9 digit and some zip code are 5 digits only. B
ased on the definition 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 sinc
e the format from raw data is character.
4. Identify duplicated records by contractor_id since contractor_id
is primary key in destination table, and keep the record with contra
ctor_version =2 for contractor_id in (139,140,228,236,238).
5. Identify duplicated records by contractor_number since
contractor_number is unique identifier for each contractor in real w
orld. Among 8 duplicated contractor_number, except  5 duplicated rec
ords overlapped with duplicated records identified by
contractor_id,the records with contractor_number 11004,11201,11202 h
ave different last_updated date. In this case, I will only keep the
```

rows with most recent last_updated date and remove duplicated row wi
th contractor_id in (373,374,378).
6. Remove 'Attn:' in address1 columns and replace '-' with blank.
7. Row 122-129' with contractor_id in (382,383,384,385,386,387) have unrecognized symbol '?' in column contractor_bus_name. Since default UTF8 Encoding in postgresql DB can not recognize it, this symbol sho uld be removed, or it will cause data error when loading data to dat abase.

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 g et state code for each contractor. Then,
rename "state_abbrev" from state_lookup table to "state_code".
3. contractor_bus_name is not unique per contractor_number in origin al data. Original contractor_bus_name and
contractor_number column should be merged with separator '-' in the middle, into a unique contractor_bus_name
per record.