Capstone Project 1: Data Wrangling

The first step I did after downloading the data from the website of New Mexico Oil Conservation Division is to convert the files from XML to CSV format, appling xml.etree.ElementTree.iterparse function. The code written for this purpose will be included in the repository on GitHub.

Then I convert the CSVs to Pandas data frames, applying pd.read_csv function, and there are 16 data frames in total. After reviewing the contents of each data frame with the help of the data dictionary provided by OCD ('OCD Interface v1.1 Data Dictionary.xlsx', will be upload to GitHub), I choose 5 out of the 16 tables for this project to analyze the well completion designs, by ignoring the tables used to track the resource reports, acreage and spacing regulations, and records of a custody transfer off a property for oil and gas. The 5 tables I will use for exploratory data analysis are: *pool, wchistory, wellhistory, wcinjection, wcproduction* (explanation of the tables at the end of this document).

Next step is to evaluate each of the chosen tables:

First, remove the trailing spaces in:

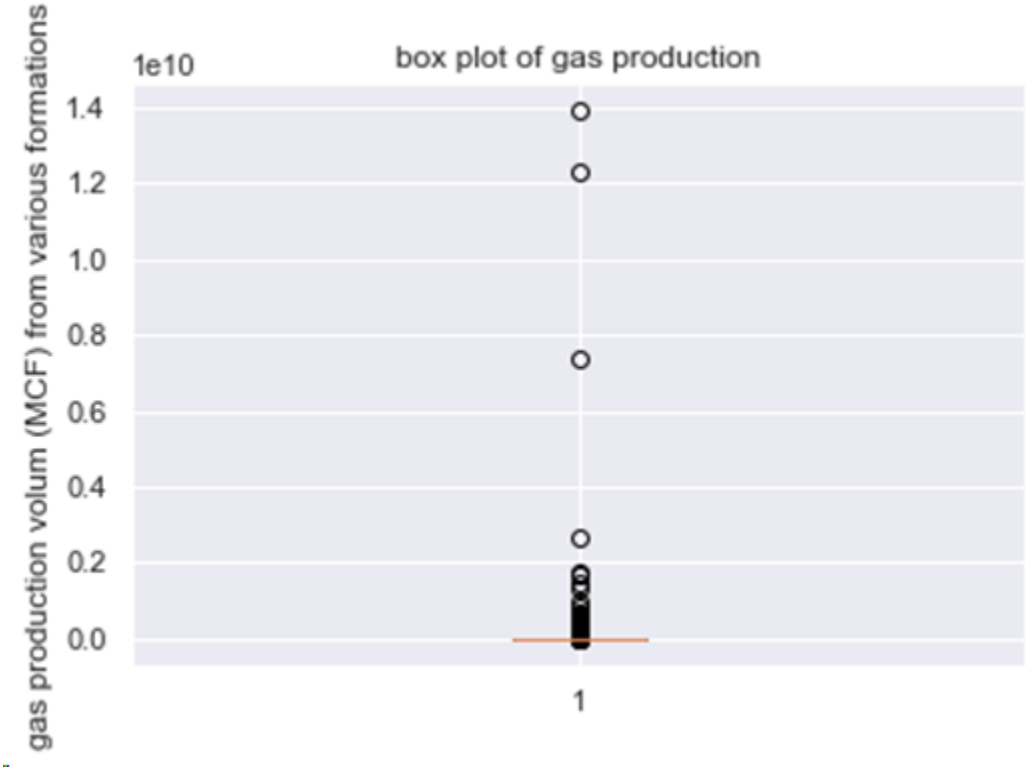
- column *prod_prop_nam* in table *properties*,
- column *inj_knd_cde* in table *wcinjection*,
- column *prd_knd_cde* in table *wcproduction*;

Then, convert the missing values into numpy.nan in tables *wchistory, wellhistory, and wcinjection*.

Some of the NaNs are ignorable because of the large data size, for example, only 1 NaN out of 338,720 rows in *well_typ_cde* in table *wchistory*. The NaNs in column *directional_status* in table *wellhistory* are forward filled after grouping by 'api', and the NaNs in column *prodn_meth_cde* are filled from the value of the same well (same *api*). However, there are still 40713 out of 257745 rows missing *direction_status* information, and 54208 out of 338720 rows missing *prodn_meth_cde* information. I will pay attention when use the column combined with other tables.

I need to point out that some of the columns may contain thousands of NaNs, but I'm not using those columns to solve the problem of this project.

Even though the boxplot of gas production in the figure below suggests the existence of outliers, those 'outliers' are good data points to analyze the correlation between production and well completion method, so they are included in my analysis.



Tables used in this project:

pool: producing formation or subdivision of producing formation properties: producing property
punevent: event resulting in possible changes to Production Unit Number for SLO
wchistory: well completion data
wellhistory: well data
wcinjection: injection data
wcproduction: production data