**ETL-Project Technical Report**

**Uche Obianyimuo**

**Jack Ozcelik**

**04/27/2020**

**ETL stands for Extract, Transform, and Load.**

**The purpose of this project was to source and extract data sets, transform the data, and then load that transformed data onto a database.**

**Extraction**

**We obtained two datasets from**  [Opendata.gov](https://www.data.gov/open-gov/):

1. State Drug Utilization Data <https://catalog.data.gov/dataset/state-drug-utilization-data-2010-81ad0>

This dataset (csv format) comprises of information reported by states for covered outpatient drugs that are paid for by state Medicaid agencies since the start of the Medicaid Drug Rebate Program. The data includes state, drug name, National Drug Code, number of prescriptions and dollars reimbursed.

Data descriptions are available on Medicaid.gov: <https://www.medicaid.gov/medicaid/prescription-drugs/state-drug-utilizat>...

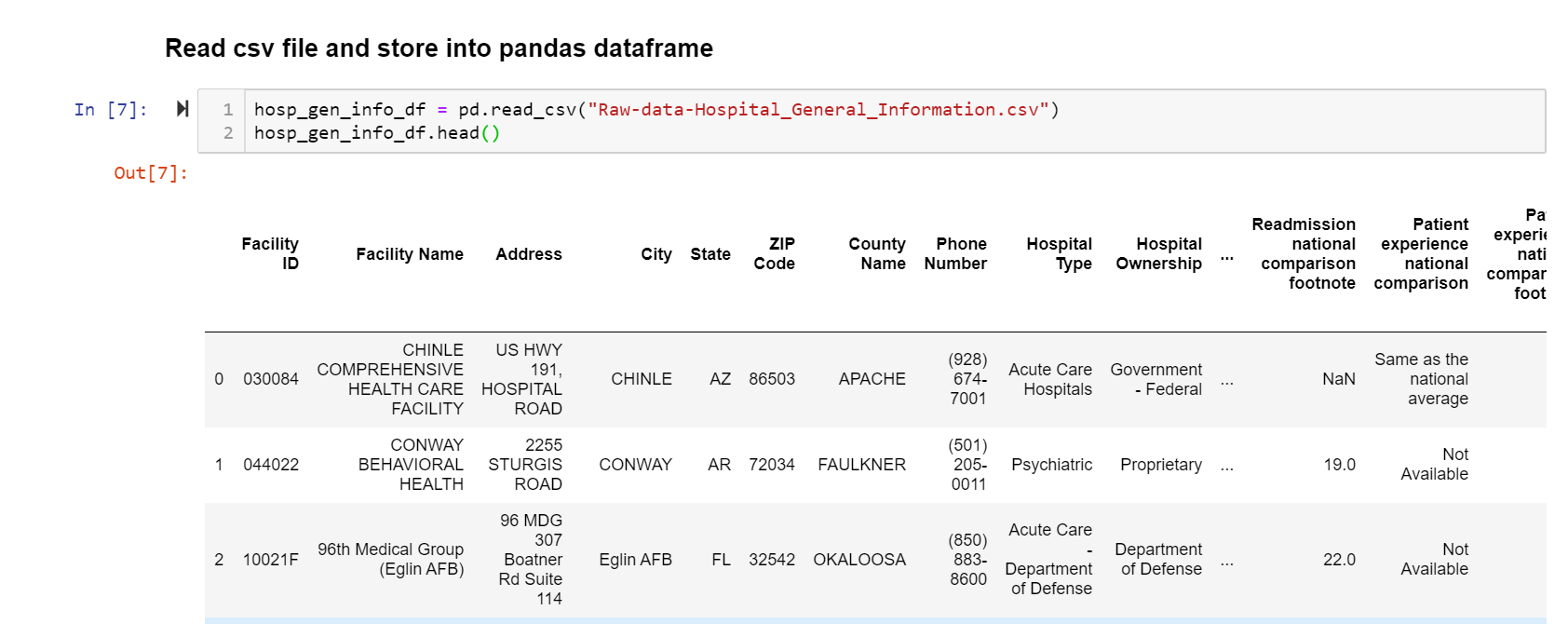
1. Hospital General Information <https://catalog.data.gov/dataset/hospital-general-information>

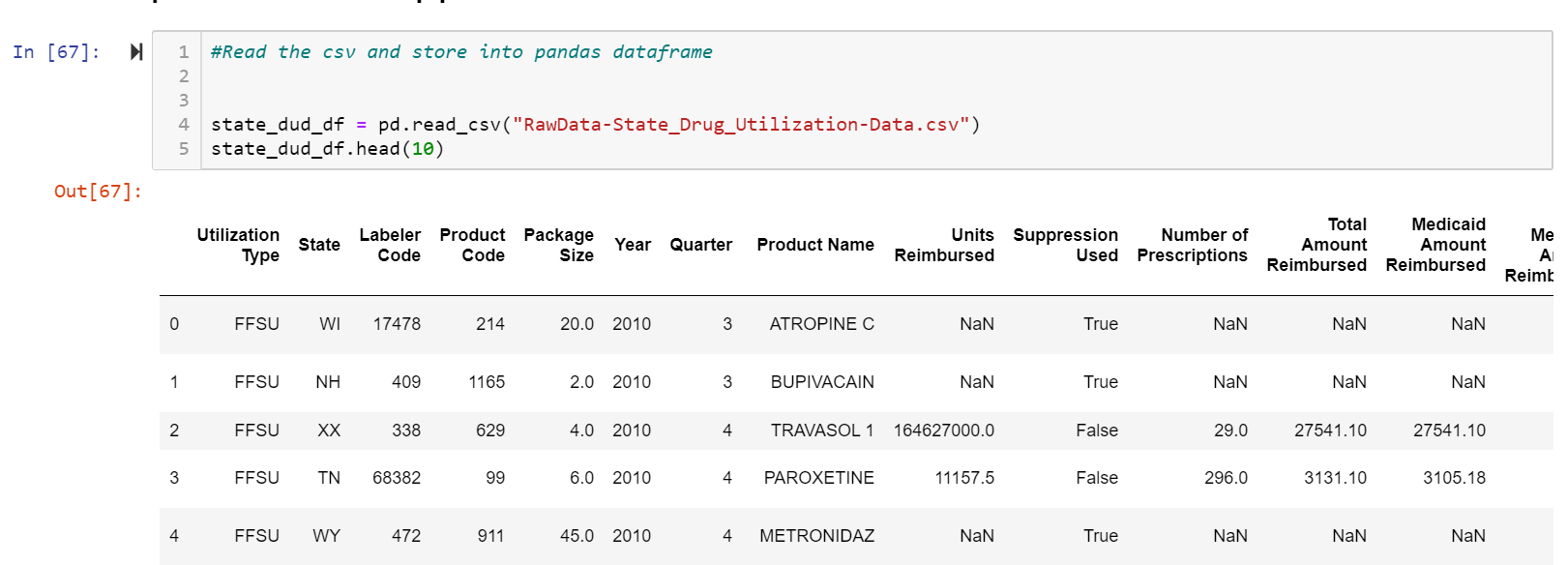
This is a list of all hospitals that have been registered with Medicare. The list includes addresses, phone numbers, and hospital type. This list was compiled in a csv format.

**These 2 sets of data were downloaded, and we proceeded to take a look at them. We discussed on what relevant information to keep and decided on the what database to utilize.**

**Transformation**

**The size of the State Drug Utilization dataset was too large to be migrated into the project repository. The file had more that a million rows but was finally reduced to about half of the original file. Both sets of data were read and stored into pandas data frames for easy readability and manipulation.**

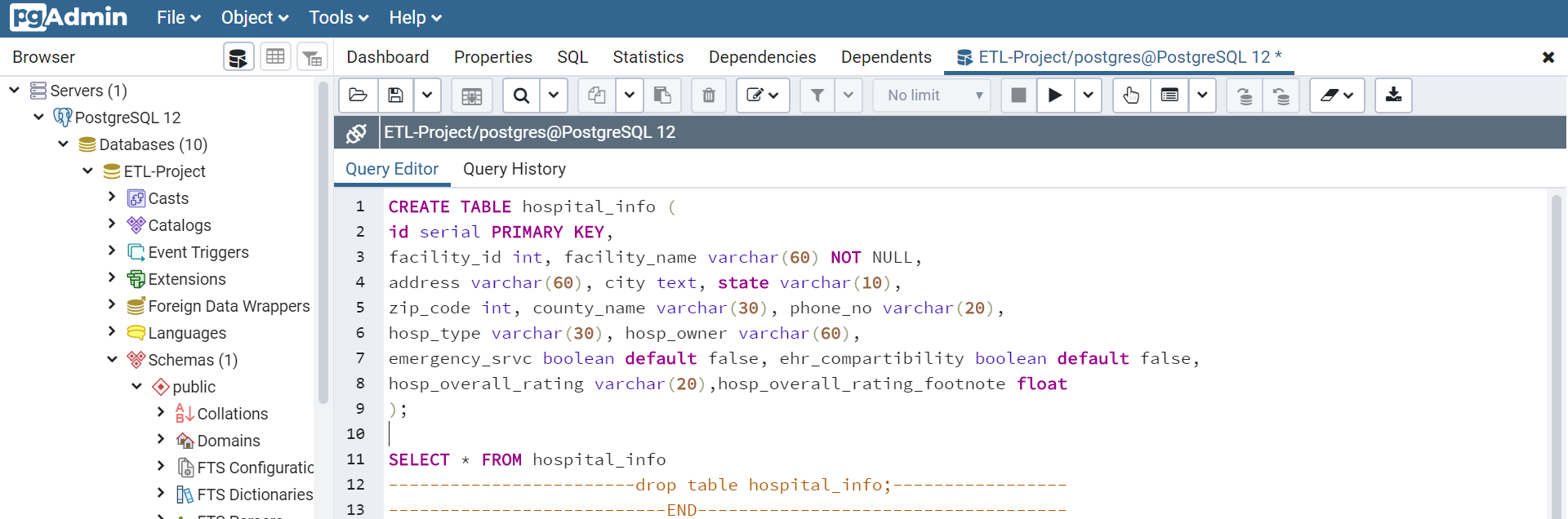
 **Fig. 1**

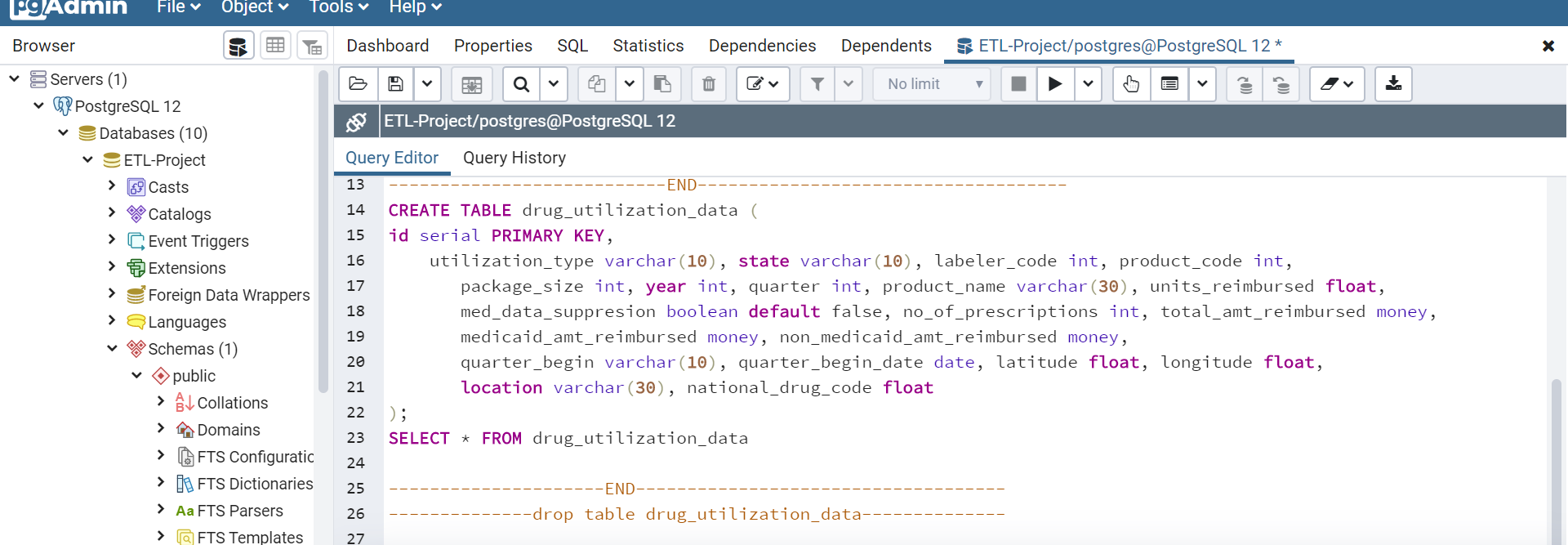
 **Fig. 2**

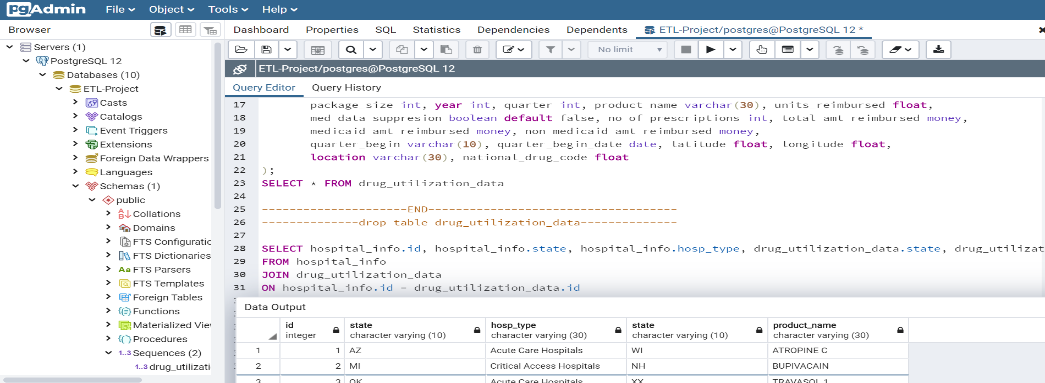
**Irrelevant columns were removed and the remaining ones were renamed to make it easier to understand (refer to repository for detailed transformation of the data).**

**Load**

**Postgres database was used for this project. A project database was created (ETL-Project) within Postgres and tables were created to accommodate the newly transformed sets of data.**

 **Fig. 3**

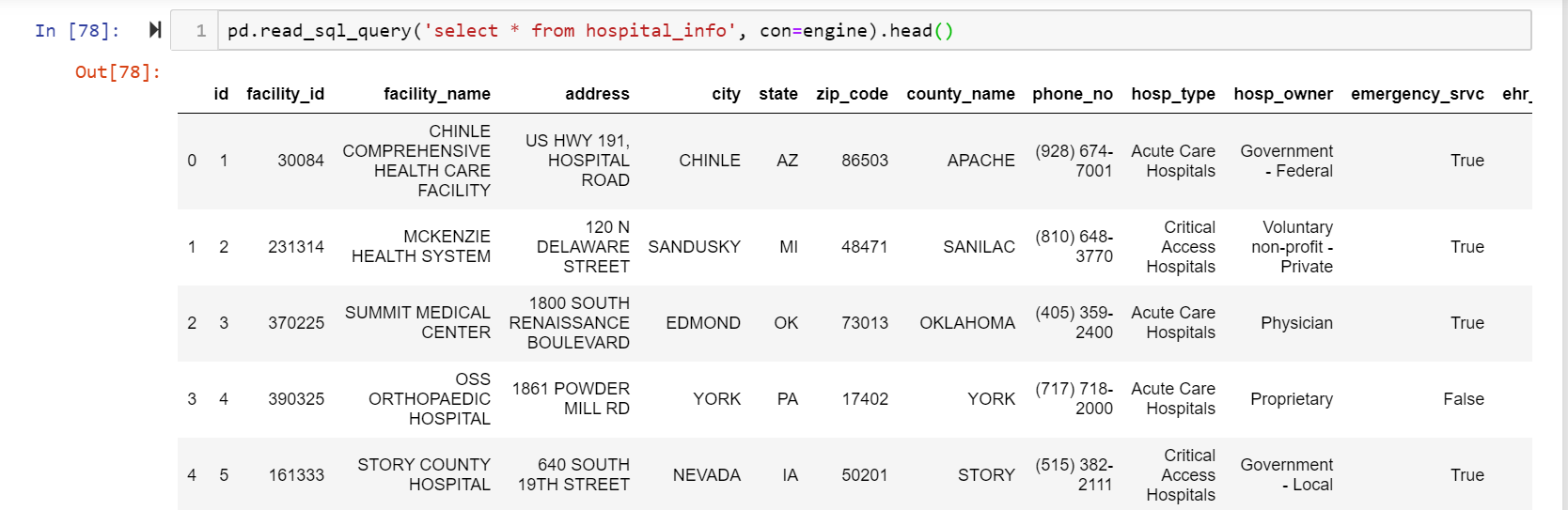
 **Fig. 4**

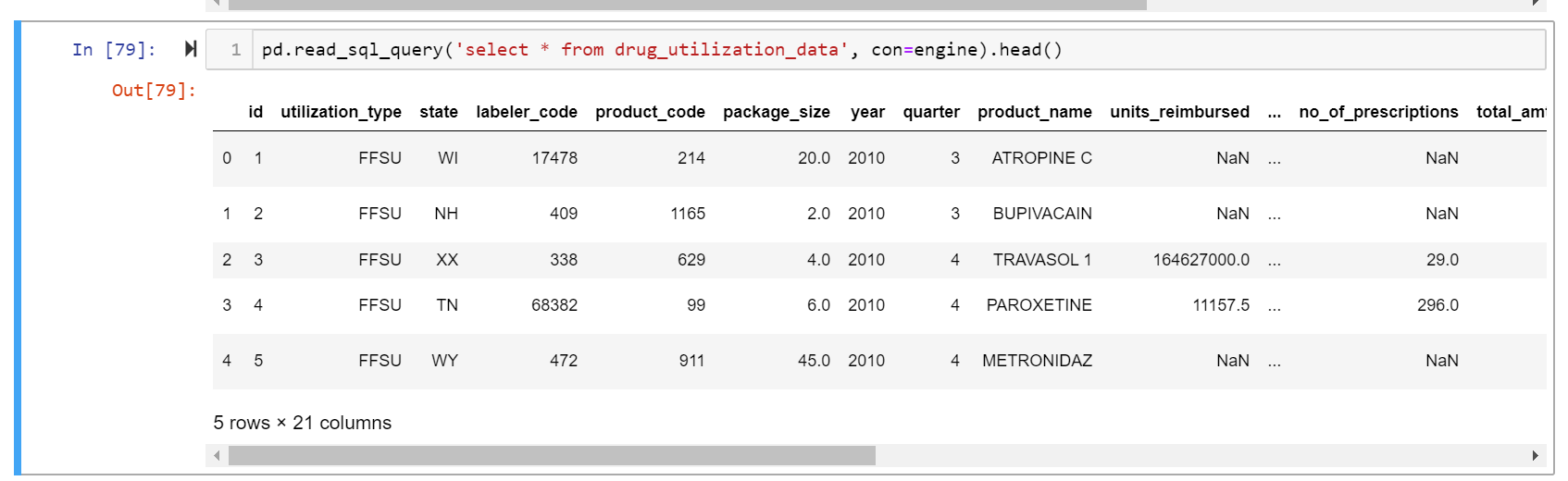


**These datasets were then loaded onto the ETL-Project database through pandas:**

 **Fig.5**

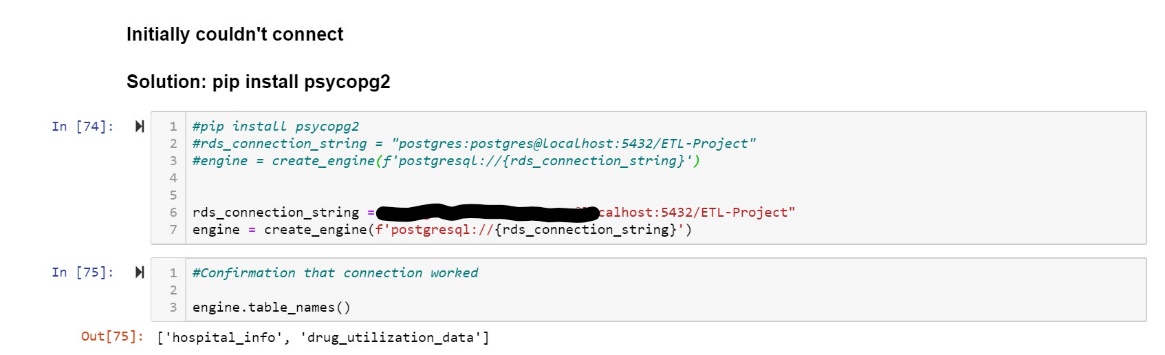
**And the database was queried:**

 **Fig. 6**

**Fig. 7**

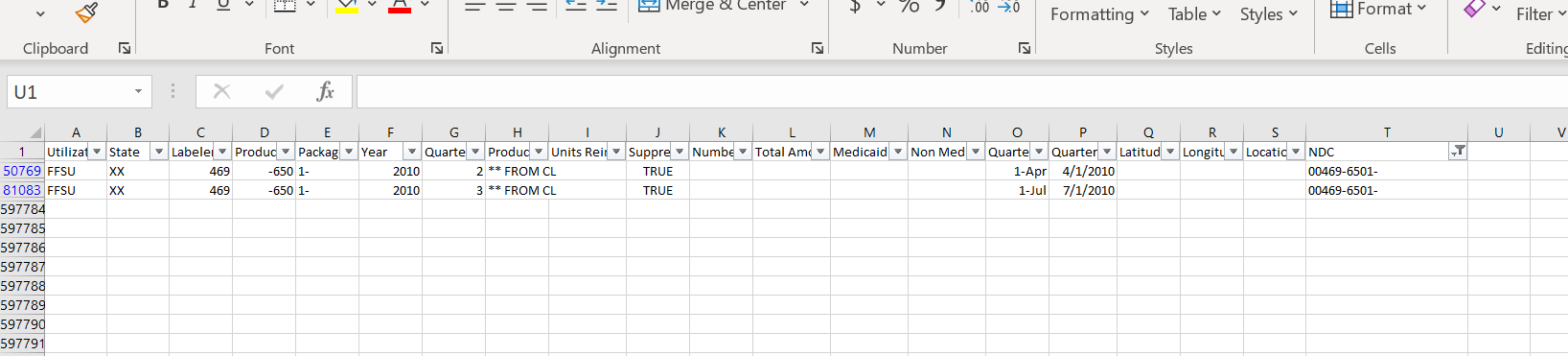
**Problems Encountered:**

* **We initially could not connect to the database. We had to ‘pip install psycopg2’ to proceed.**

**Fig. 8**

* **We also found it difficult to load the cleaned ‘State Drug Utilization Data’ file. It was discovered that a column within the file had 2 rows with ‘non-integer’ values.**

**The rows were removed by applying a filter on the column in the csv file and manually deleting the rows.**

 **Fig. 9**

**The data was eventually loaded successfully.**

Tools used:

* Postgres
* Jupyter Notebook
* Python

Please refer to <ETL-Project-Repo> for the final database.