**Data Cleaning**

In order to make the “Geolocation” field more usable, I wanted to do two things: remove the enclosing parentheses, and split it into two fields, “Latitude” and “Longitude”.

**Removing the parentheses**

UPDATE medicare

SET geolocation = LTRIM(geolocation, '(');

UPDATE medicare

SET geolocation = RTRIM(geolocation, ')');

**Adding and updating the “Latitude” column**

ALTER TABLE medicare

ADD Latitude varchar(255);

UPDATE medicare

SET Latitude = SPLIT\_PART(geolocation, ',', 1);

**Adding and updating the “Longitude” column**

ALTER TABLE medicare

ADD Longitude varchar(255);

UPDATE medicare

SET Longitude = SPLIT\_PART(geolocation, ',', 2);

**--Checking for null values**

SELECT COUNT(\*) FROM Medicare WHERE \*field name\* IS NULL;

529 entries contain null values at: Data\_Value, LowConfidenceLimit, and HighConfidenceLimit.

These entries also contain a description in Data\_Value\_Footnote that reads “Statistically unstable estimates not presented [numerator <10].”

820 entries contained null values in the Latitude and Longitude fields. In the LocationID field, these entries were marked with “59”.

The only entries in the Data\_Value\_Footnote\_Symbol and Data\_Value\_Footnote fields that were not null were the 529 entries mentioned above.

**Checking for duplicates**

SELECT (medicare.\*)::text, COUNT(\*) FROM medicare

GROUP BY medicare.\* HAVING COUNT(\*) > 1;

No duplicates found.

**Removing unused columns**

ALTER TABLE Medicare

DROP COLUMN Geolocation (was split into ‘Latitude’ and ‘Longitude’ columns)

DROP COLUMN CategoryID (all values were ‘C1’)

DROP COLUMN Data\_Value\_TypeID (all values were ‘Crude’)

DROP COLUMN Data\_Value\_Alt;; (all values were the same as ‘Data\_Value’)