**ETL Project**

**Project Team Members: Sara McLean, Danica Quero, Lindsey Giron, Alex Corona**

**Project Summary**

Problem Statement: Is the advancement in technology changing sources in which people communicate with one another?

We collected data from data.worldbank.org to set up analysis of data usage trends from 2005 to 2018 time period. This data includes the following data sets – fixed telephone subscriptions (landline use) and mobile cellular subscriptions.

Data Sets Used –

* Metadata Country API IT: features global landline use 1960-2019
* Metadata Indicator API IT: features global mobile phone use 1960-2019
  + Mobile\_cellular\_subscriptions.csv
  + Fixed\_telephone\_subscriptions.csv

The goal for future analysts is to compare and contrast trends in data usage over time, beginning from landline use to mobile cellular use and finally to app-based data use. Is the rise of mobile cell subscriptions causing the decline in landline use? The data will help to provide analysts data for this goal.

**ETL Objectives**

* Extract data table CSV sets in JupyterLab
* Create new PostgresSQL database tables in PgAdmin 4 called fixed\_telephone and mobile\_cellular
* Transform the data sets to DataFrames to match to schema for PostgreSQL fixed\_telephone and mobile\_cellular databases
* Load transformed DataFrames to PostgreSQL

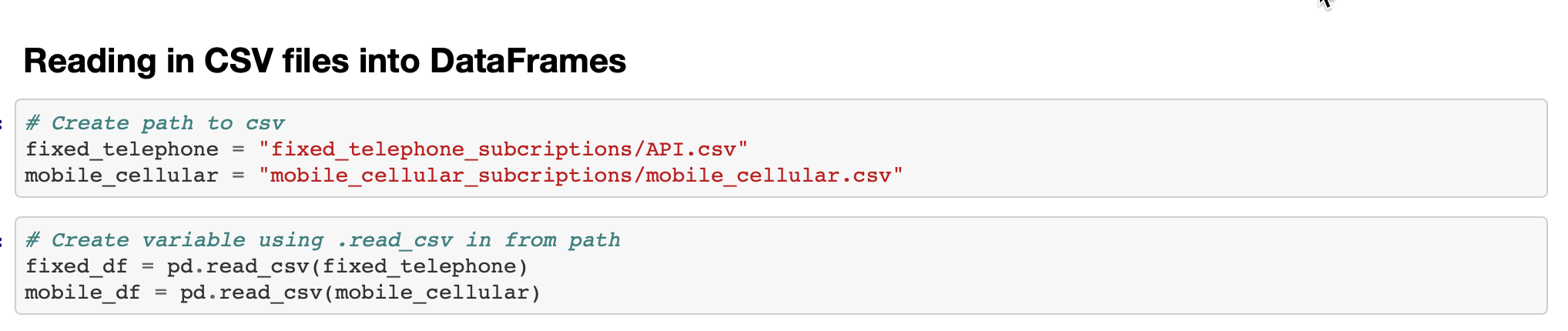
**Extract**

The data extraction process included the following:

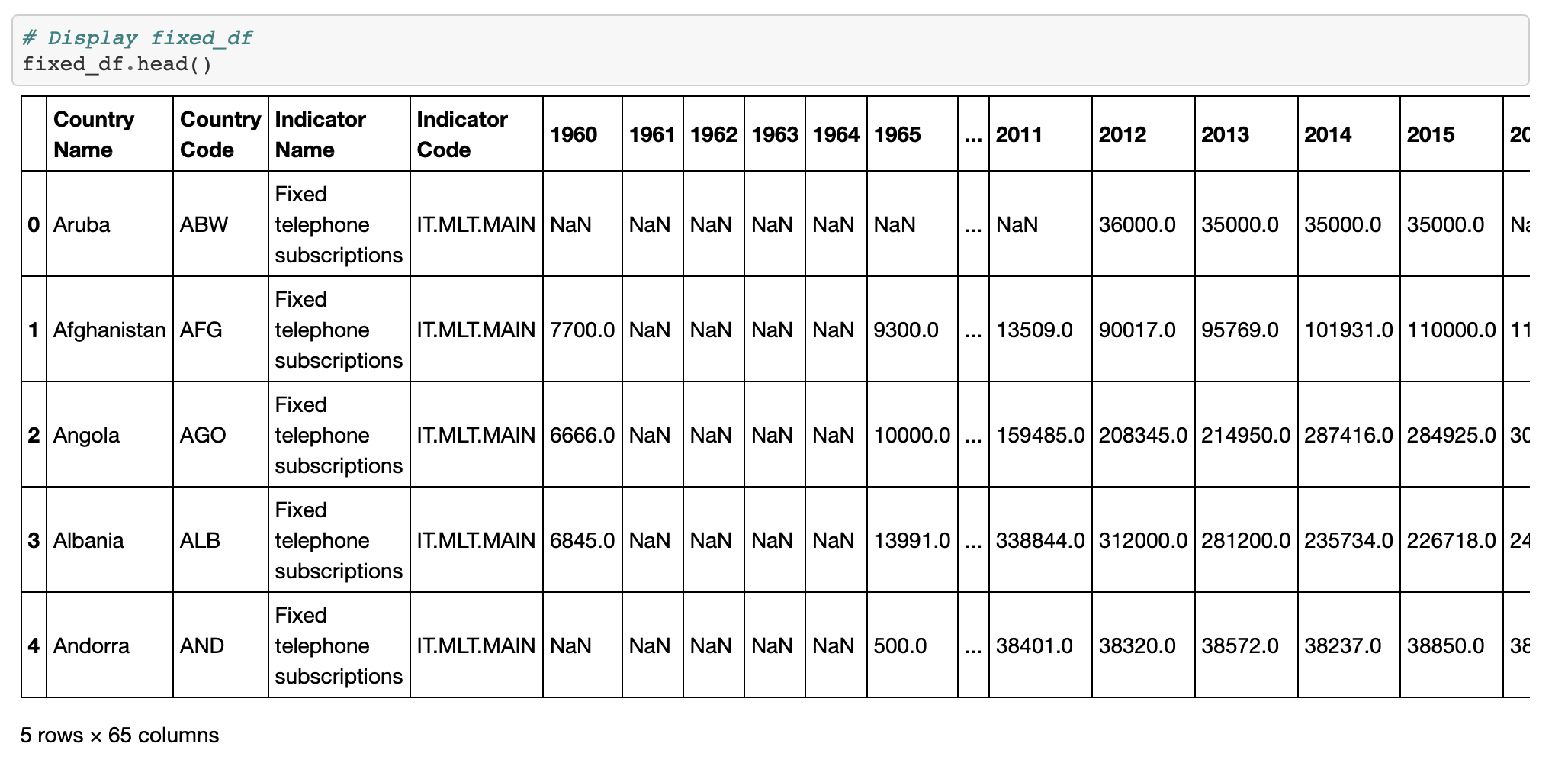
* JupyterLab created for Project called ETL\_project.ipynb
* Dependencies:
  + Pandas
  + SQLAlchemy
  + Config/pg\_password

**Reading in Dependencies**

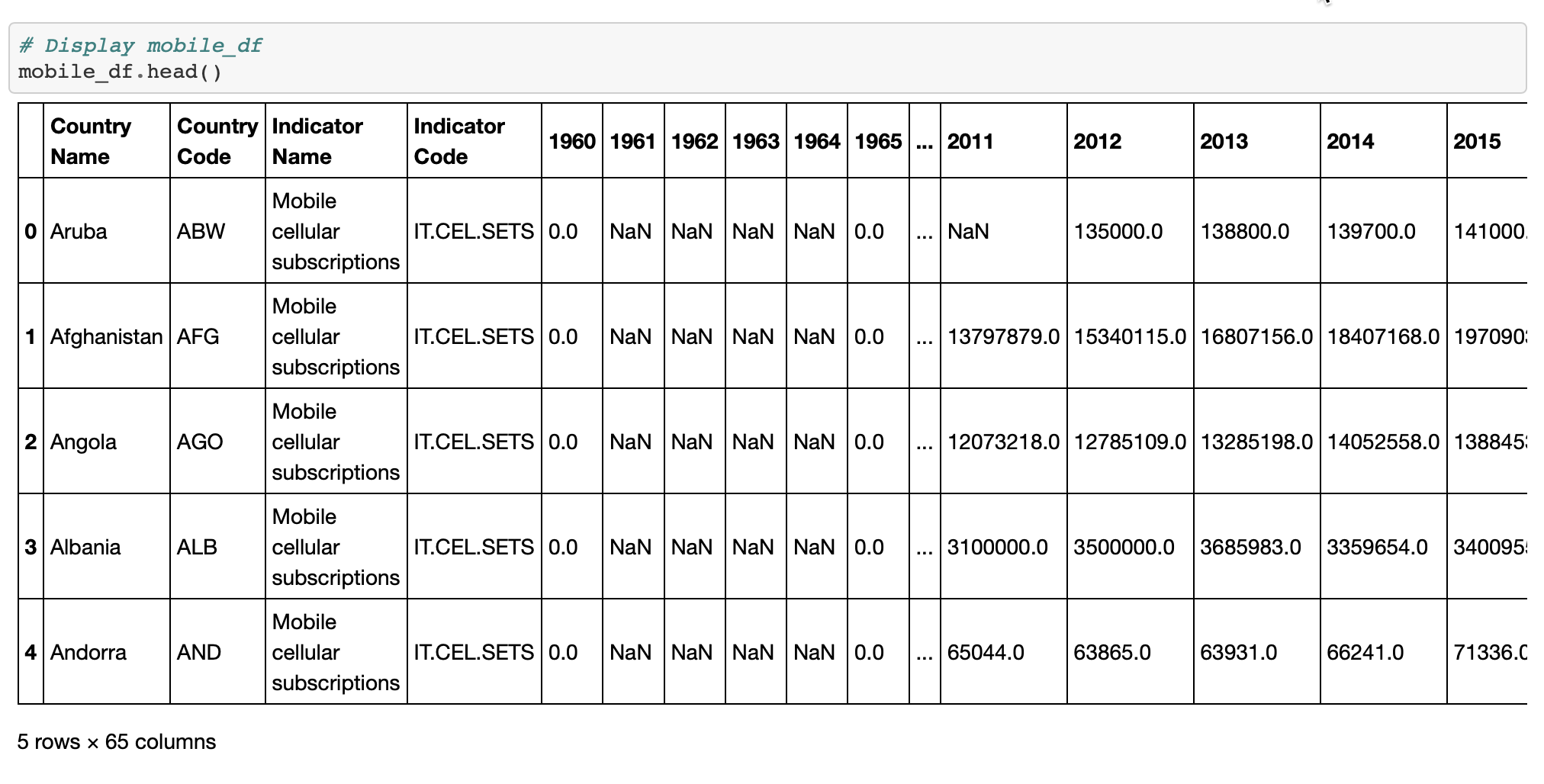




Displaying Fixed DataFrame



Displaying Mobile DataFrame



**Transform**

The goal of the data set transformation was to focus on the United States for the years 2005 – 2018 in highlighting future analysis in digital communications.

To transform the data through cleaning and filtering, the process included:

* Converting to DataFrames from .csv files
* Filtered out Null Values from Data
* Merging tables using Panda DataFrame using join statements
* Importing data to pdAdmin 4 using SQLAlchemy (relational database) connections
* Used relational database to link the data by our common identifier Country Name as Primary Key

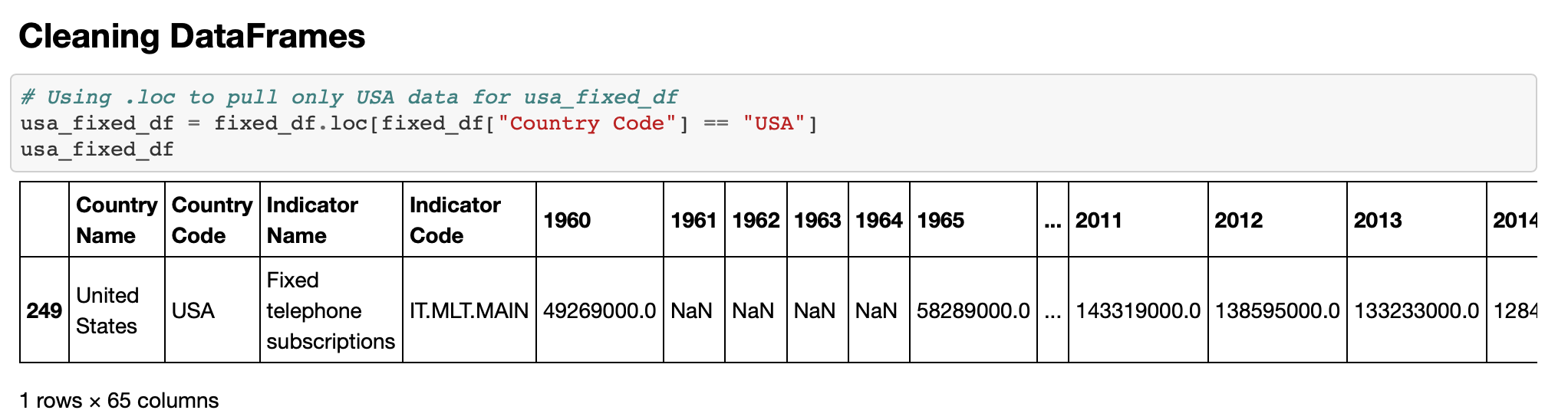


Image: fixed\_telephone Transformation

* Converting data
* Converting column
* Replacing NaN values
* Renaming columns
* Ensure column headers match PostgreSQL schema

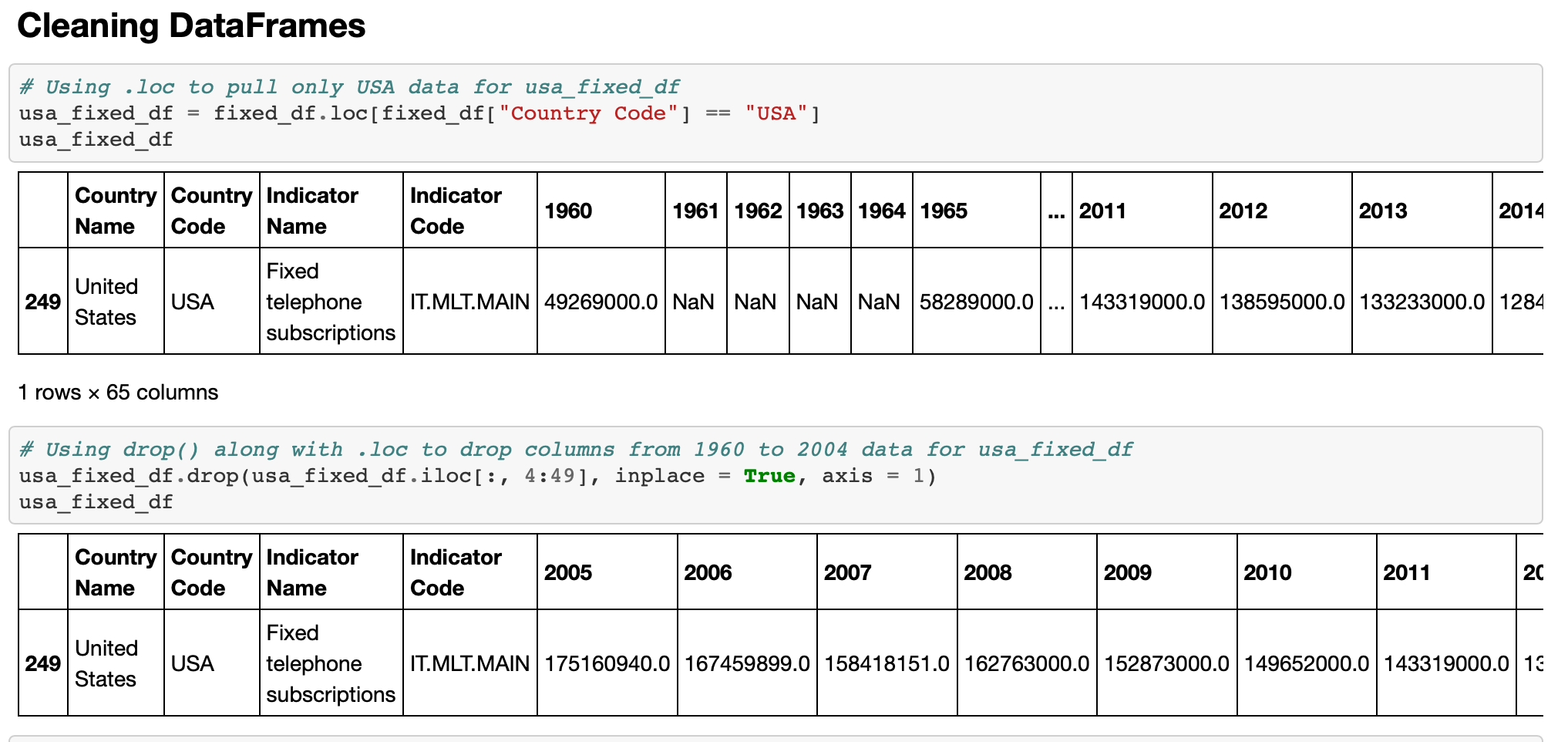
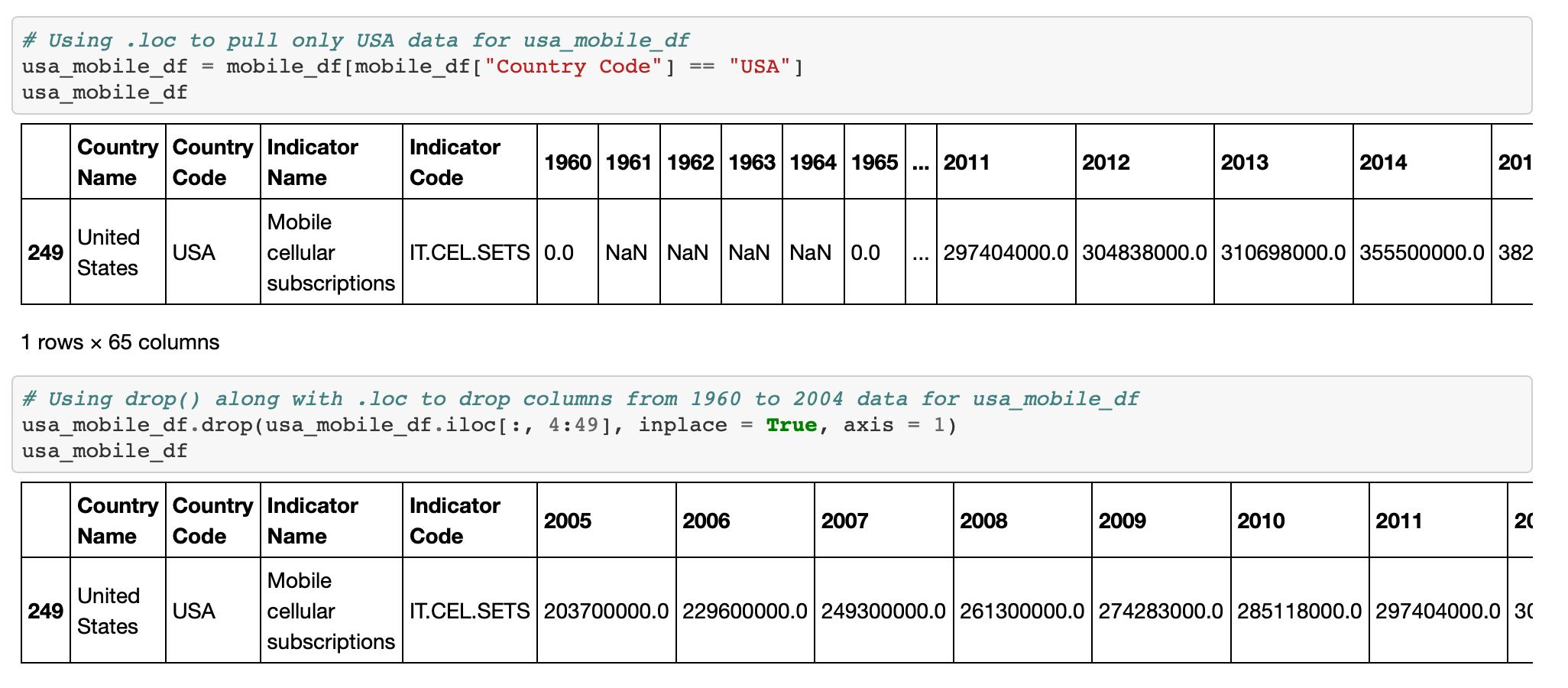
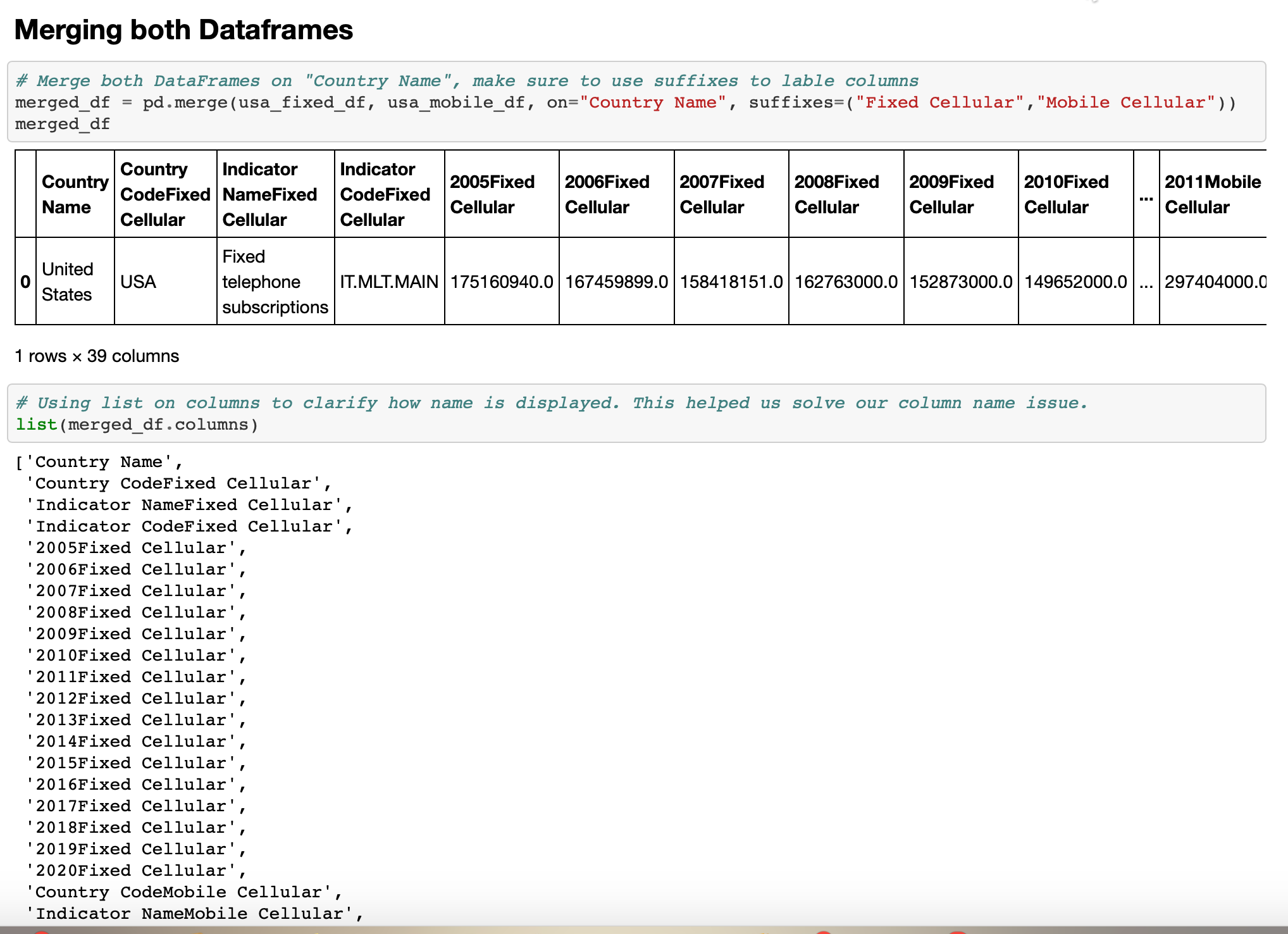


Image: mobile\_cellular Transformation

* Converting data
* Converting column
* Replacing NaN values
* Renaming columns
* Ensure column headers match PostgreSQL schema



Merging Both DataFrames and Confirmed Columns





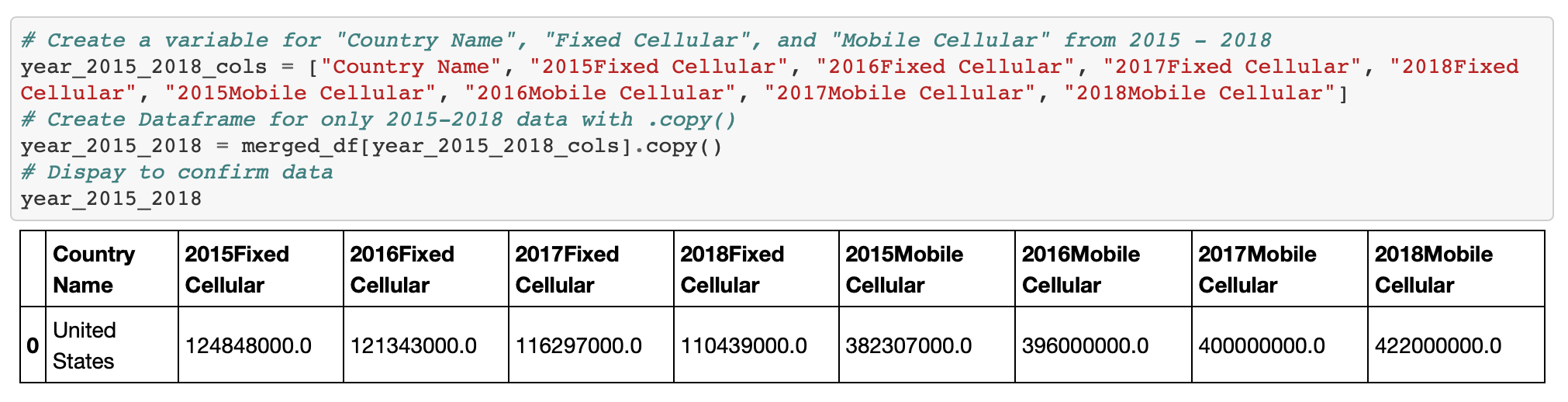


Image: PostgreSQL Database and schema

* Table statements were created using 5 year increments for analysts to review trend data
* Tables set up include query formulas for both fixed\_cellular\_data and mobile\_cellular\_data
  + Fixed
    - 2005 – 2009
    - 2011 – 2014
    - 2015 -- 2018
  + Mobile
    - 2005 – 2009
    - 2010 – 2014
    - 2015 -- 2018

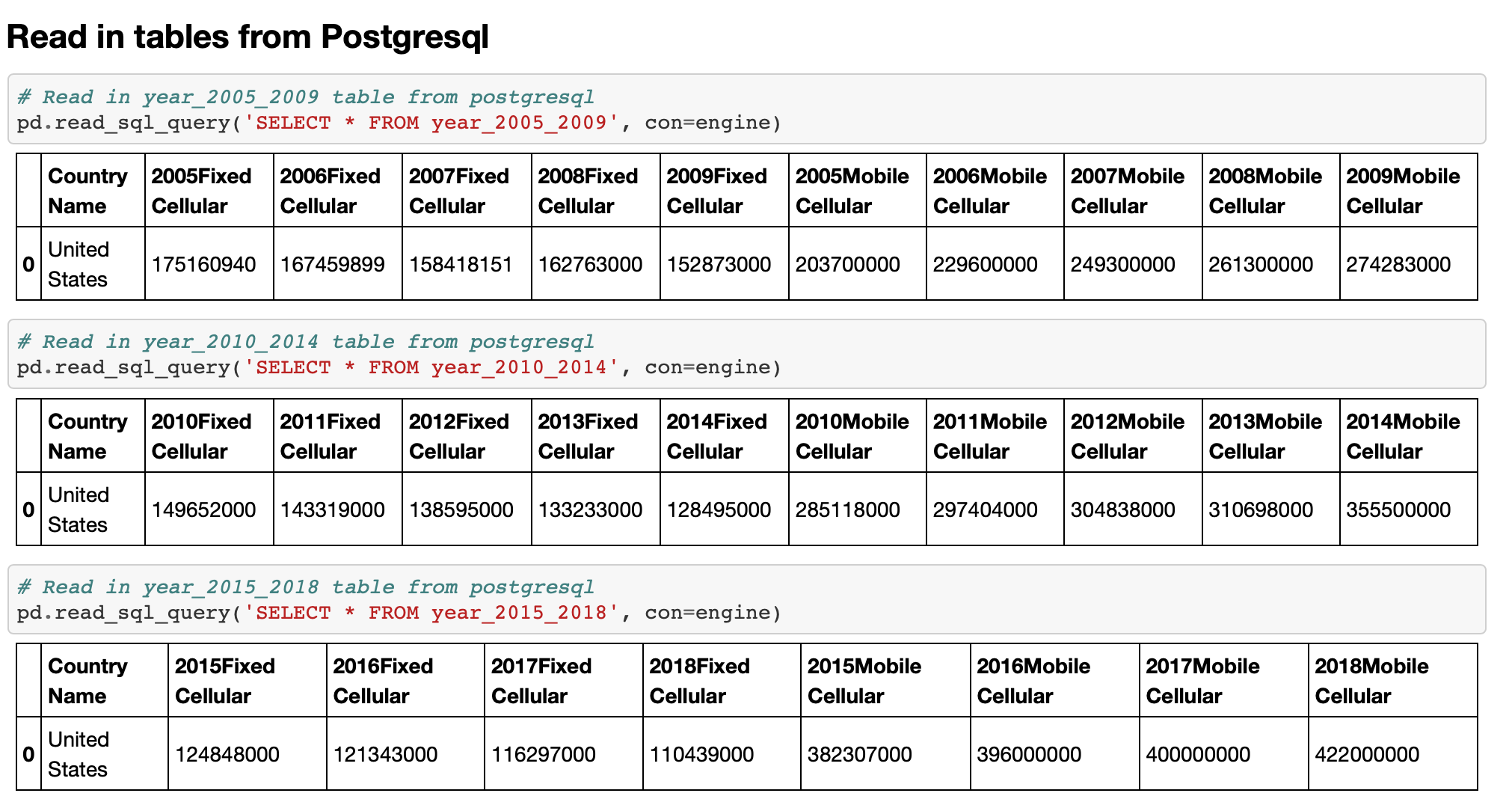
**Load**

A database connection to PostgreSQL was set up then the tables were confirmed as available. The final step was to load transformed DataFrames into PostgreSQL fixed\_cellular\_data

Loading Process



Reading in Tables from PostgreSQL



**Conclusions**

The data has been extracted, transformed and loaded. Data analysts can now use the data to review trends in digital communications within the USA.