**ETL Project Final Report**

**Group:**

Arian Sadri

Nguyen Nguyen

Ozkar Agoz

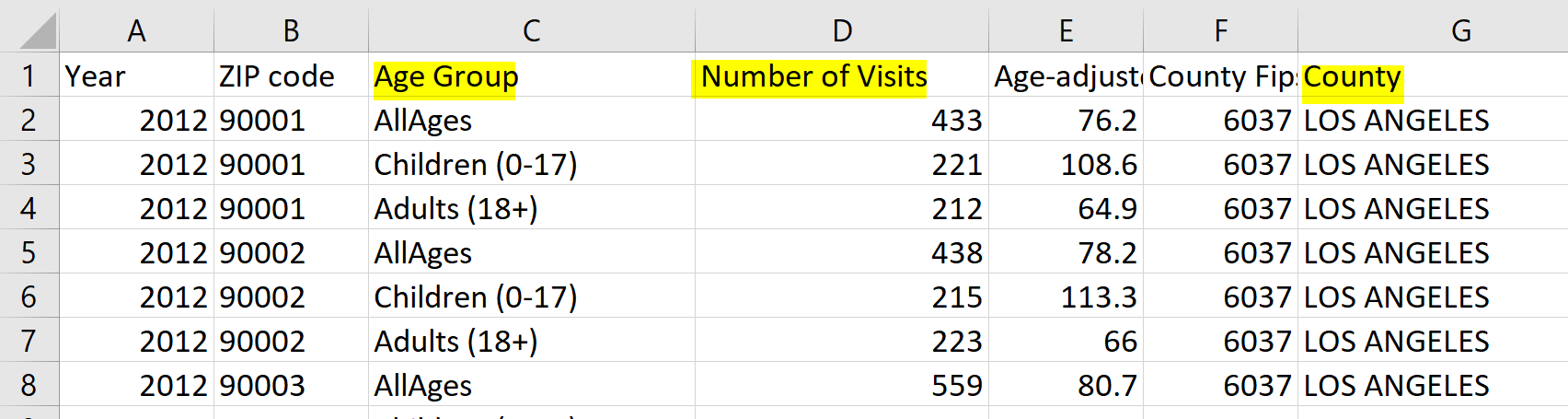
**ASTHMA ER VISITS VS AIR QUALITY INDEX**

**BY COUNTY IN CALIFORNIA IN 2012**

1. Extraction: Downloaded two csv files from two different sources:

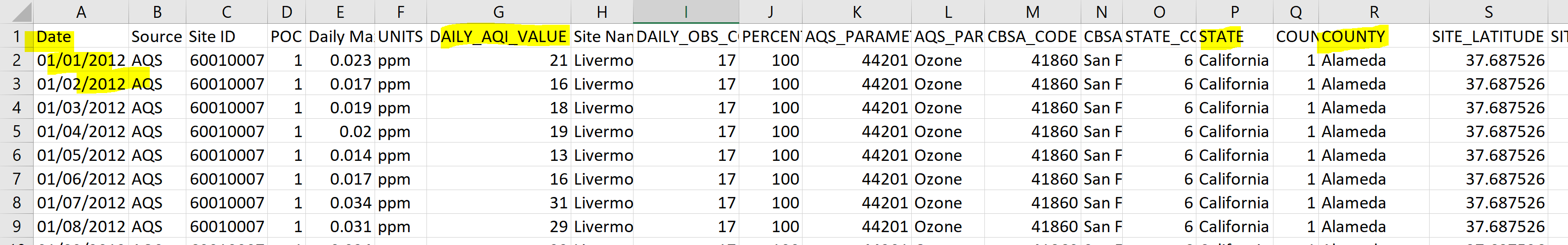
<https://data.world/healthdata/asthma-ed-visit-rates-by-zip>

Get Number of Visits for Children group by COUNTY



<https://www.epa.gov/outdoor-air-quality-data/download-daily-data>

Get Daily AQL Value by County



1. Transformation:

* In Python:
  + Read both csv files to Pandas
  + Put both files to two dataframes named “ashma\_visit\_by\_zip\_df” and “Air\_Quality\_by\_Zip\_df”
  + Cleaned up two dataframes:
    - Renamed columns to eliminate space in name for “Number of Visits”
    - Grouped by country, average “Daily\_AQI\_VALUE” and average “Number of Visits”
    - Lowercased for column County to make it unify with county’s format from the other dataframe
    - Picked only needed columns: County, Number of Visits and Daily\_AQI\_VALUE from two dataframes

1. Load:

* Use Relational Database: Join, load to Postgres.
  + Loaded two clean dataframes to PgAdmin to two new tables named “ashma\_updated” and “aqi\_by\_mean”
  + Joined two tables by County to get the average AQI and average Number of ER Ashma Visits.

**Final thoughts:**

Purpose of the project is to visualize a correlation, if any, between the air quality and numbers of visit for ashma patients in California. However, there is no obvious correlation being seen. Thus, more data will be needed to draw meaningful conclusions.