ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

BUSINESS ANALYTICS
Master of Science

# COVID-19 Vaccinations in the United States

M.Sc.: Business Analytics

Course: Data Management and Business Intelligence

Professor: Damianos Chatziantoniou

---

Students: Lakkas-Pyknis Evangelos

Melosora Stamatoula

# DESCRIPTION OF DATASET

Source: Centers for Disease Control and Prevention (CDC).

This dataset, captures information from the years 2020 to 2023, describes vaccination trends across the U.S., offering a granular view that spans counties and age demographics (initial number of rows approximately 1.8 million rows).

Dataset's issues:

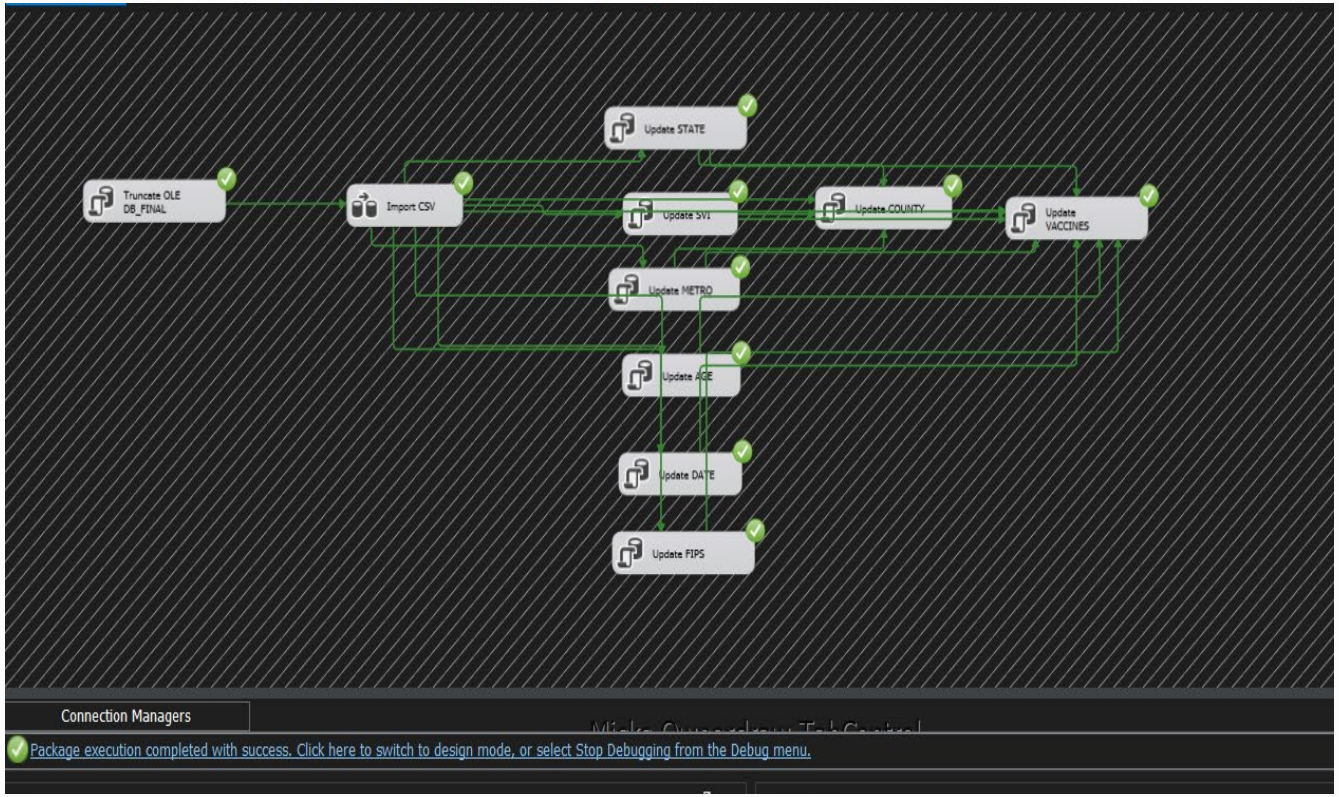
- **Insignificant observations** for years 2020 and 2023, as vaccination had not started yet or had been completed in large scale
- **Not well-defined age groups** that could not be used for a meaningful analysis
- **Missing values** in columns that would be used as measures (e.g. Census)

# TRANSFORMATION & CLEANING PROCESS

The transformation and cleaning of the dataset has been executed through Python, R and SQL.

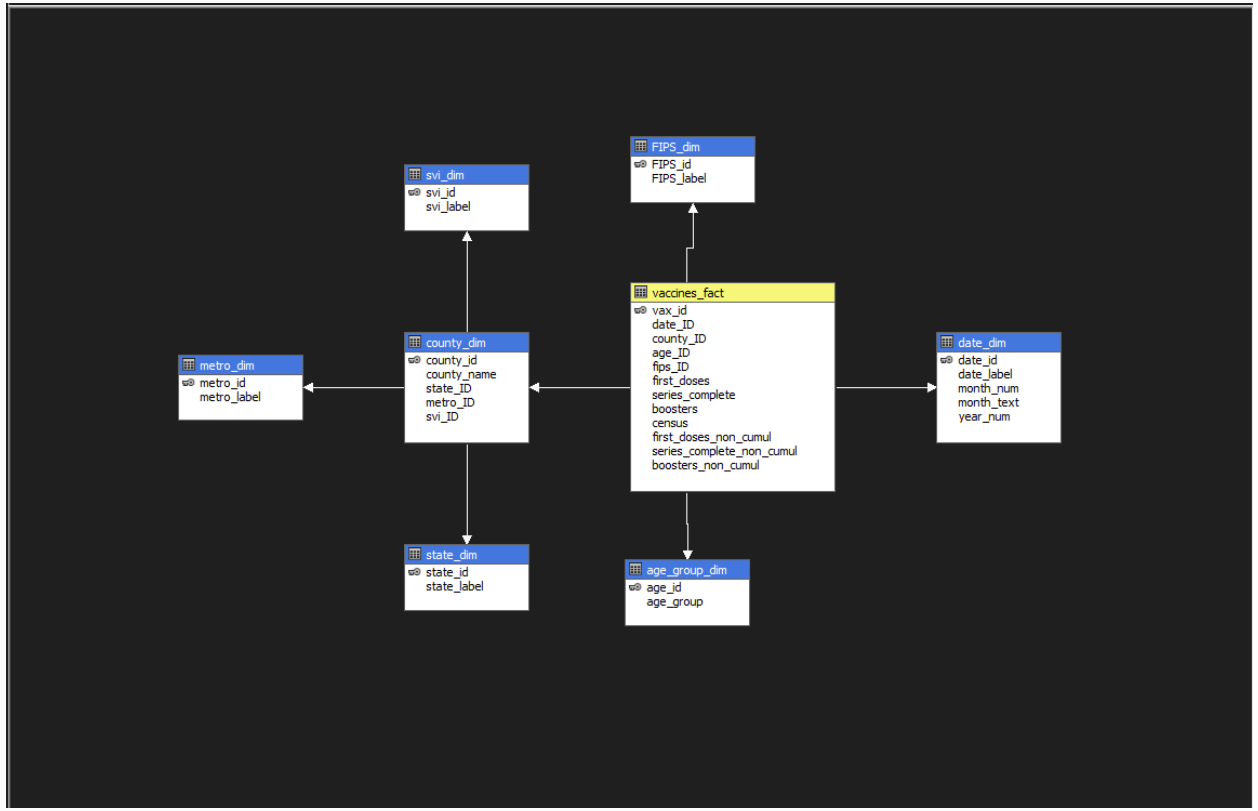
- Dropping rows containing values for years 2020 and 2023
- Creation of well-structured age groups in R (to use them as a dimension) by performing operations with the existing columns referred to age
- Filling in missing values mainly for booster vaccines in Python (setting their values to 0 for dates before their first administration) and census in SQL (using COALESCE function to acquire values for each county from other dates)
- Unpivot columns referred to age groups into rows in SQL after importing the data to create proper values for the age group dimension and better manipulation (final number of rows approximately 9 million rows).

# DATA IMPORTING & SSIS




- Connection of a Flat File Source and OLE DB Destination for data insertion to the database
- Execution of Data Flow Task
- Creation of SQL tasks
- Connection between the SQL tasks and the data flow and with each other to import and update the data in the dimension and the fact tables

# CUBE




## Snowflake Schema

### Fact Table "Vaccines":

- Foreign Keys
  - Date
  - County
  - Age
  - FIPS
- Measures
  - First doses
  - Series Complete
  - Booster
  - Census

### Dimensions:

- FIPS
- Date
- Age Group
- County

*Foreign Keys from other dimensions*

  - State
  - SVI
  - Metro

# CUBE

County and Date dimensions have natural hierarchies:

- County is connected to state label from State dimension
- Date's hierarchy consists of the attribute relationships: Date label – Month – Year

Created calculated measures:

- First Doses per Census pct
- Series Complete per Census pct
- Boosters per Census pct

# Thank you for your time!