Final Project ETL

CIS 9440 -Data Warehousing for AnalyticsFinal Project Milestone 3

Due: April 24th,2021

Group Number - 21

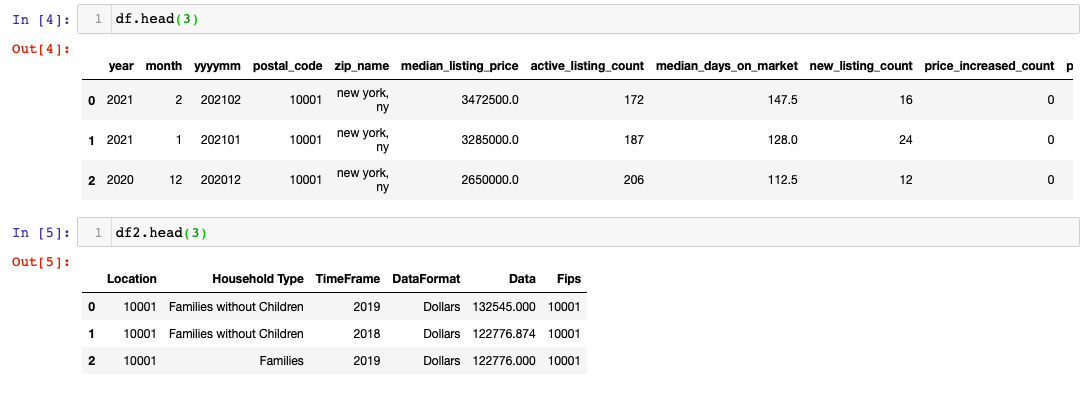
Student(s) - Leah Lu, Porfirio Mohabir, Eric Sedaghat, Tori Torsaengtong, Tshering Sherpa

**Utilized Jupyter Notebook and Pandas library to build ETL Pipeline**

**Extraction:**

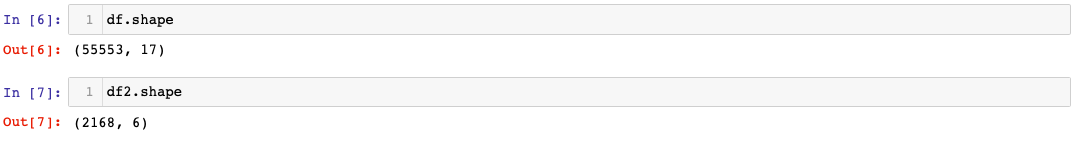
* Downloaded datasets on local computer from Google BigQuery
* Created two dataframes by reading in datasets “Real estate list.csv” (df) and "median incomes.csv" (df2) using read\_csv method



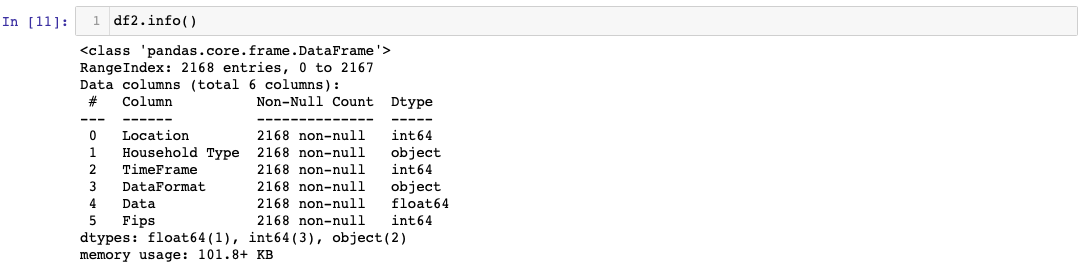
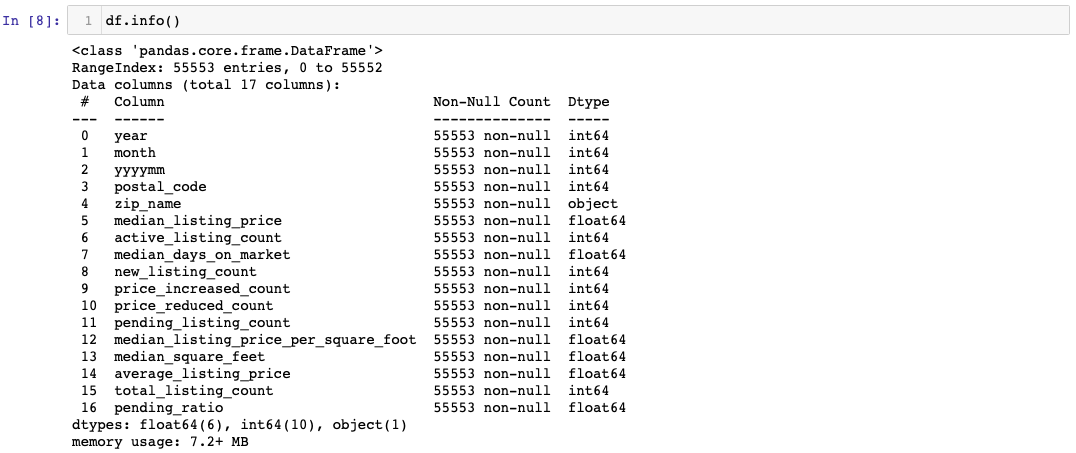


**Transformation:**

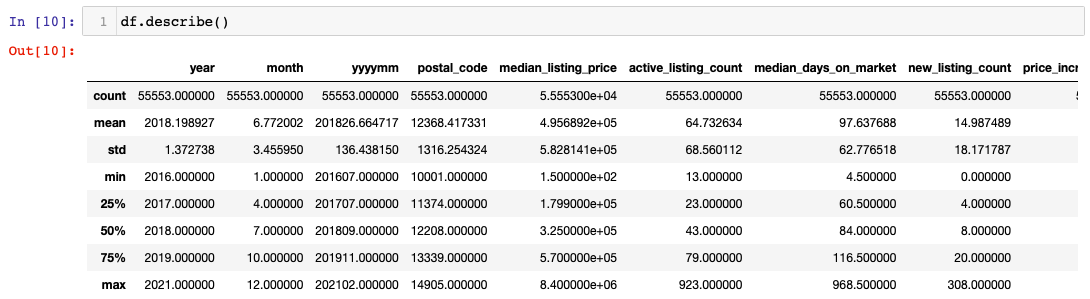
* Conducted Data Profiling by executing built-in Pandas methods on both dataframes
  + shape() → length of rows and columns

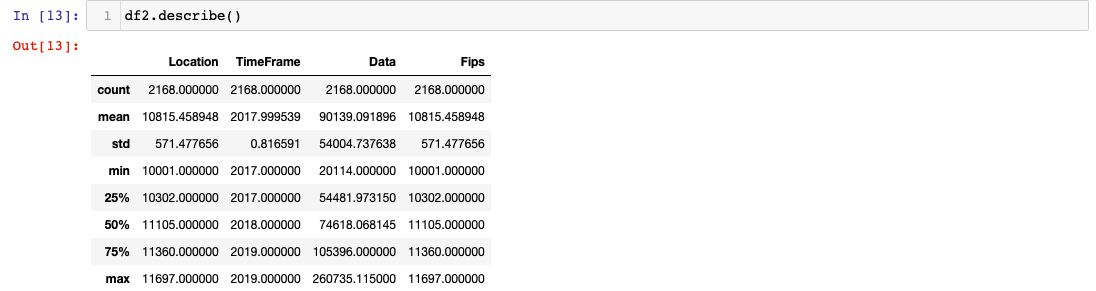


* + info() → results columns, data types, range, non\_null count, and memory

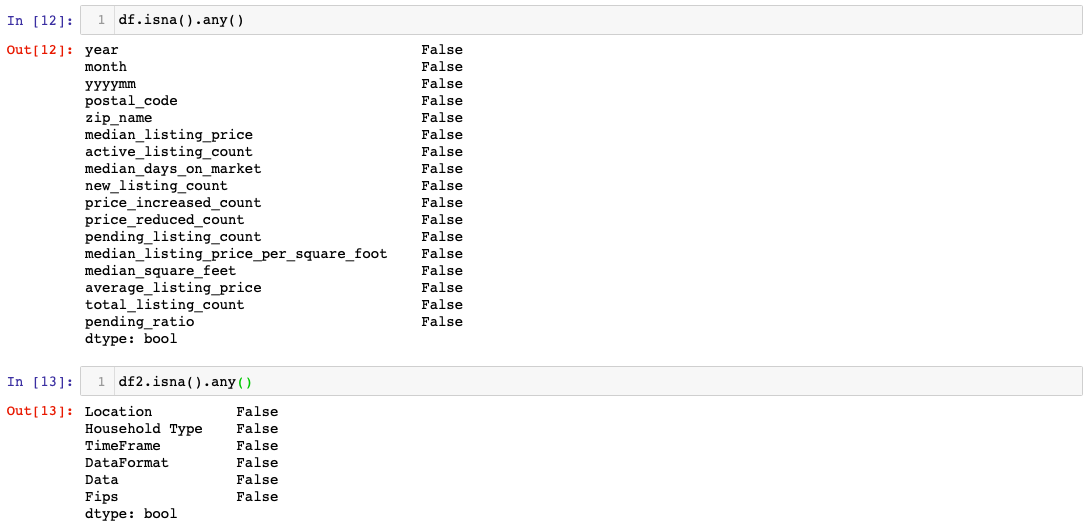


* + describe() → description of data by outputting statistics for each column

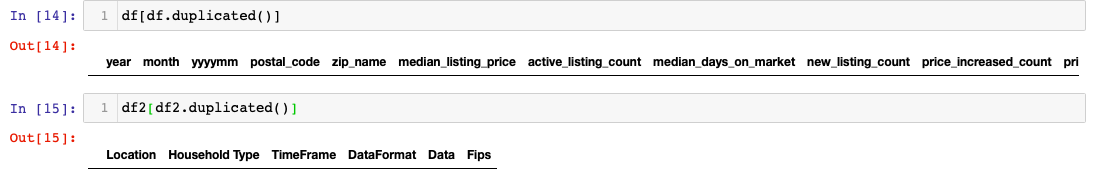




* Data Cleansing
  + Used .isna().any() to search for null values in both dataframes
    - Both df and df2 have **no null values**



* + Used .duplicated() to search for duplicate rolls in both dataframes
    - Both df and df2 have **no duplicate values**



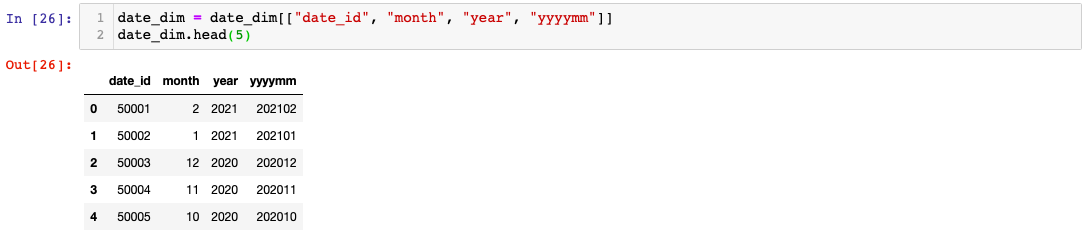
* Data Transformation: includes date, location, household dimensions along with property and income fact tables.
  + Created **date dimension**



* + - Created date\_id values starting from 50000



* + - Reordered to put it in a hierarchical order



* + - Added a month column to the "median income" dataset so that both "real estate" and "median income" datasets'



* + Created **location dimension**
    - Changed column name from RegionName to ZipCode and separated the location dimension columns from the main dataset.



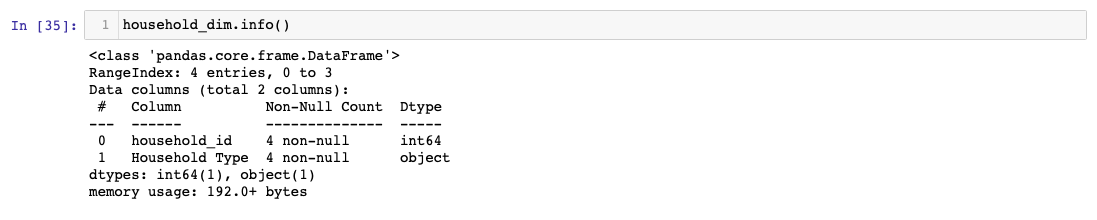
* + - Created a hierarchy for the location dimension. Therefore, we separated the citystate column to a city and a state column.



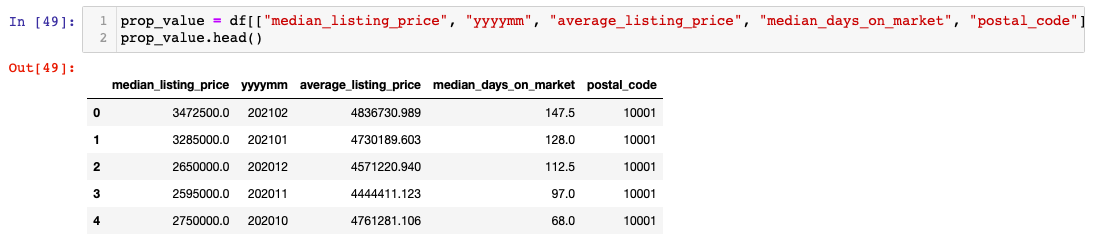
* + - Created location\_id for the location dimension.

****

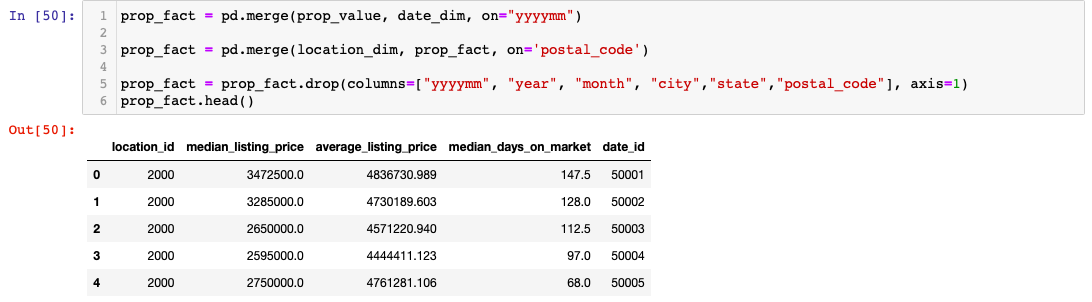
* + Created **household dimension**
    - Created household\_id for unique household types.

****

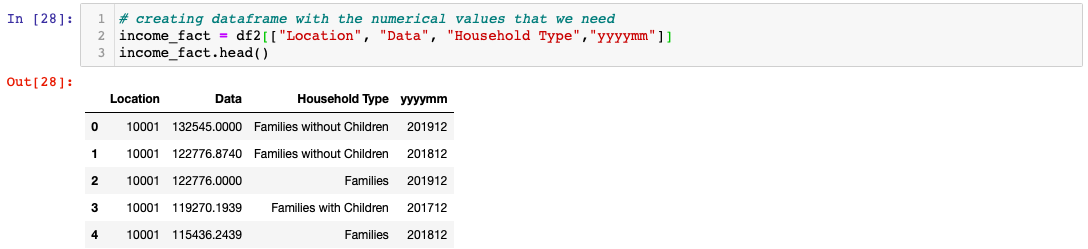
* + Created **property fact table**
    - Created a dataframe with the numerical values needed.

****

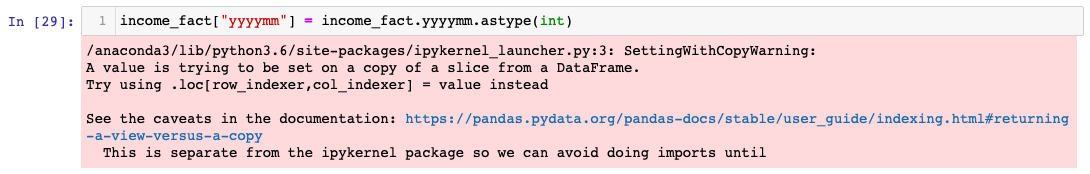
* + - Merged the data and dropped columns that are not being used.

****

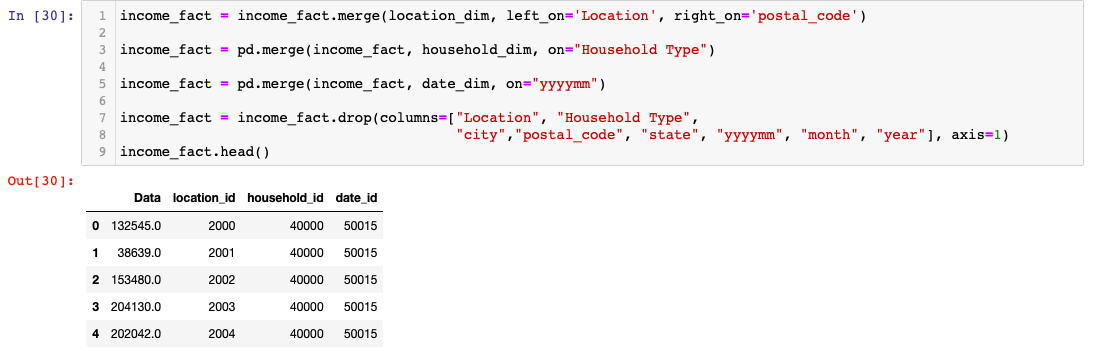
* + Created **income fact table**
    - Created a dataframe with the numerical values needed.



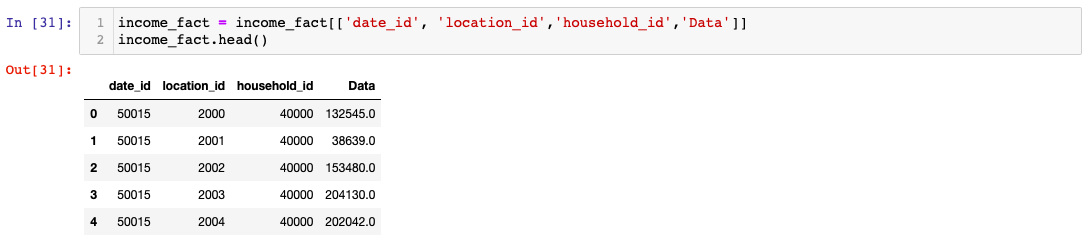
* + - Change the datatype for "yyyymm '' column to match with date\_dim for merging.



* + - Merged the data and dropped columns that are not being used.

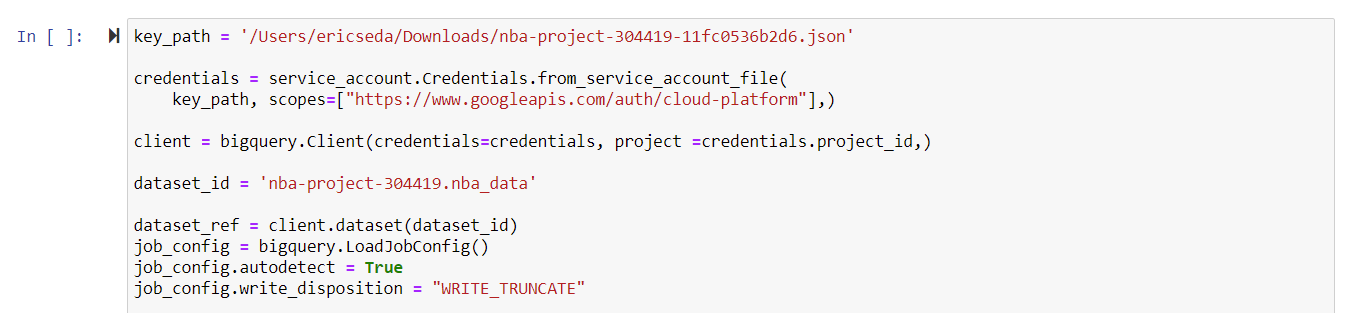


* + - Reordered the columns so that the foreign key columns come before others.

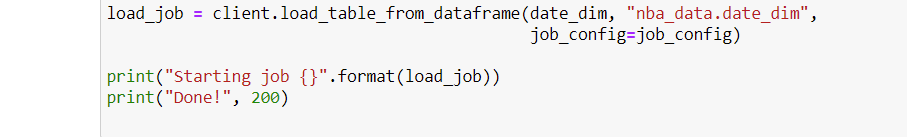


**Load:**

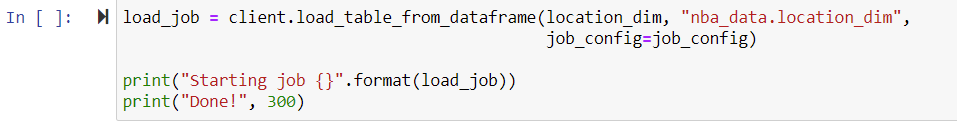
* Connect to Google BiqQuery to begin loading dimensions and fact tables.
  + Assigned dataset ID to variable dataset\_id
  + Created a working job using loadJobConfig() method
  + Set autodetect (instance variable) of object “job\_config” to Boolean True
  + Set write\_disposition (instance variable) of object “job\_config” to String “Write\_Truncate”

****

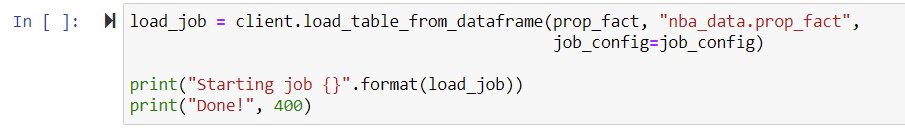
* Loading Date Dimension



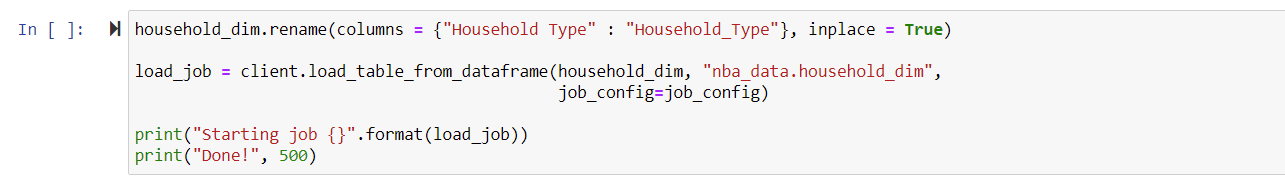
* Loading Location Dimension

****

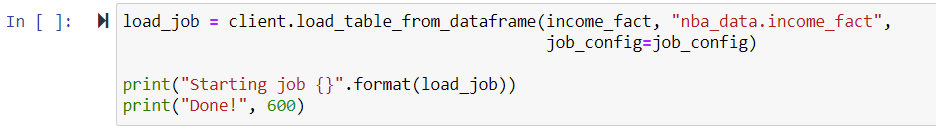
* Loading Property Fact Table

****

* Loading Household Dimension
  + Renamed “Household Type” → “Household\_Type”

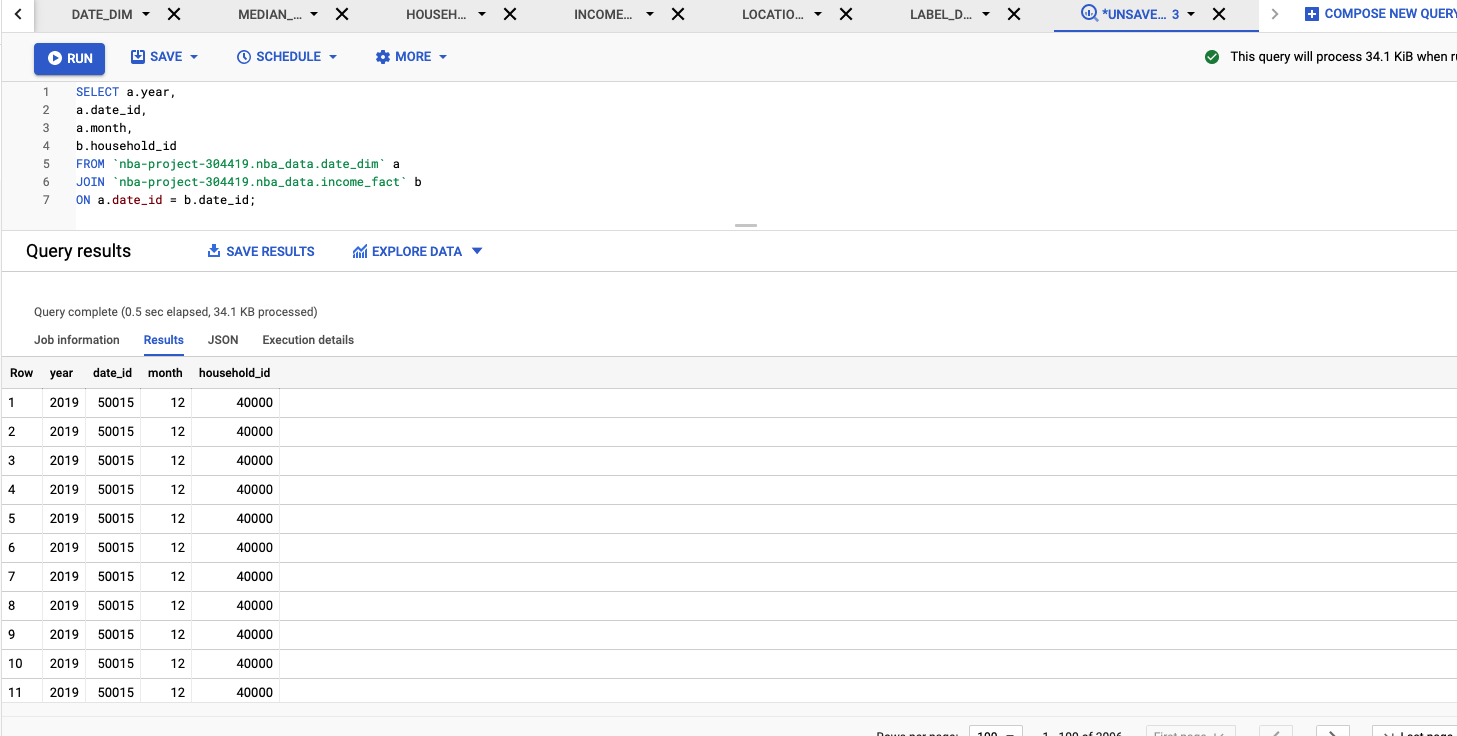
****

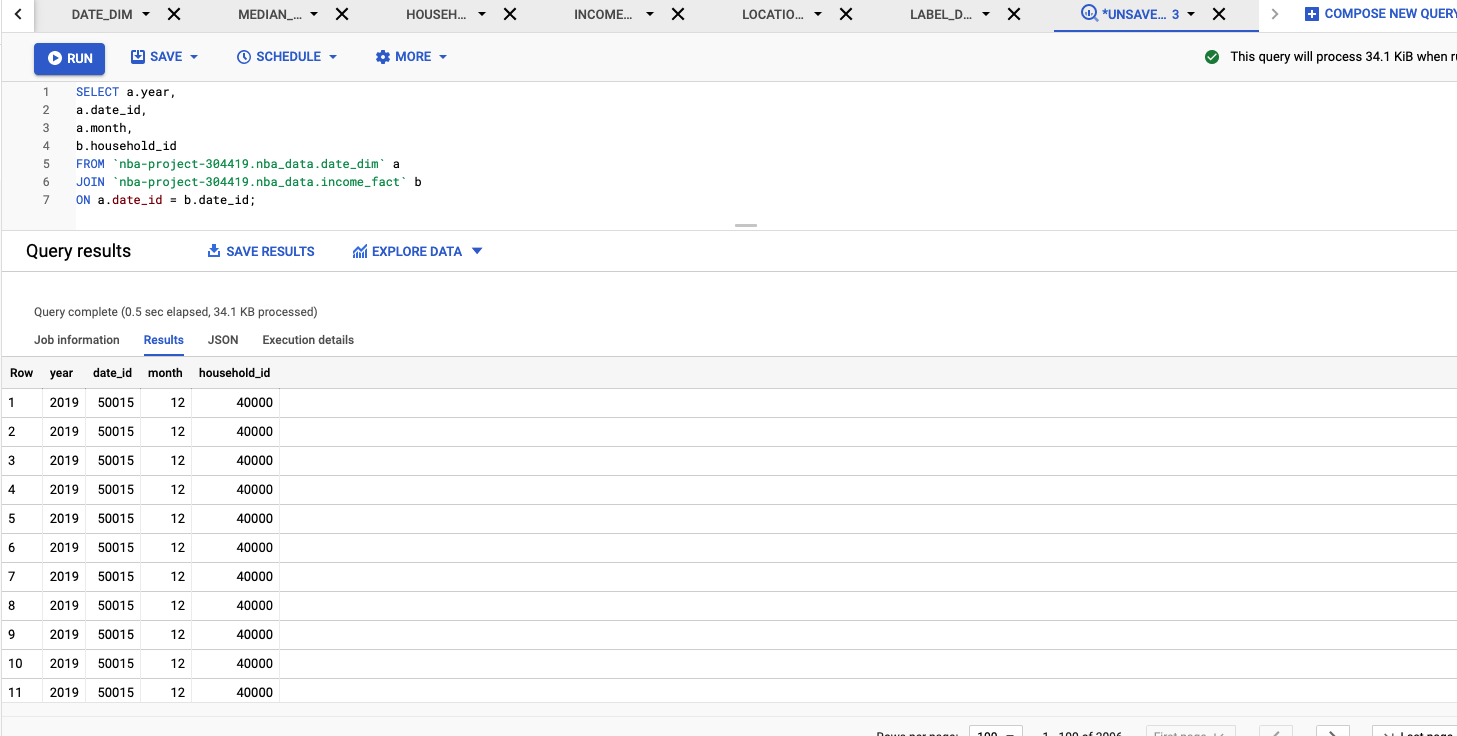
* Loading Income Fact Table



**JOIN STATEMENT**

* Wrote a join statement (to check if the tables link up and work) after uploading the most updated tables to google big query.



**