**ETL Project Report – Find Price Per Plate Based on Zip Code**

Team Members: Chinedu, Ed, Amanda

**Introduction – Extract Data Sources**

The aim of the project was to compare restaurant information (cost of meals) to the median income of people based on Zipcode. To achieve this, Yelp and US Census data were identified as primary sources of data. The required data was to be extracted from Yelp using an API key and downloading the required data from the census database.

**Transform – Cleaning Process:**

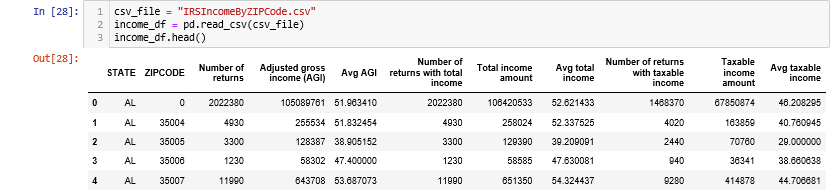
It was difficult extracting the required data from Yelp, and as a result, Zomato Restaurants Data was identified as a secondary source of data, and information was downloaded in a CSV file. The census data was also downloaded from the site in a CSV file containing all the information for all the states in the US.

Census Data: The census data was manipulated in the following ways;

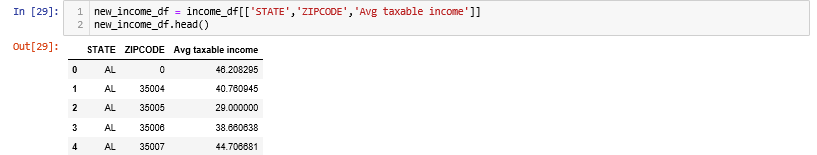
* A Jupyter notebook file (ETLProject.ipynb) was created to enable the importation of the information in the CSV file into Pandas



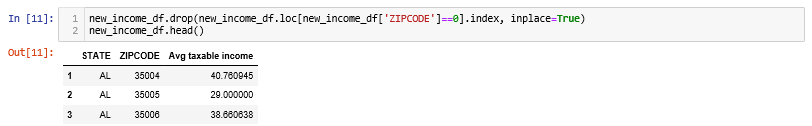
* The dependencies were imported and then the CSV file was read into a Pandas Dataframe



* The unnecessary columns in the dataframe were dropped to leave only the “State”, “ZIPCODE” and “Avg taxable income” columns



* There was a row with a Zipcode value of “0” was dropped by using the column property of the row to call it



* A database was created on postgresql and a database connection was established



Zomato Data: The Zomato data was manipulated in the following ways;

* A Jupyter notebook file (greenplates.ipynb) was created to enable the importation of the information in the CSV file into Pandas
* The dependencies were imported and then the CSV file was read into a Pandas Dataframe



* The unnecessary columns in the dataframe were dropped to leave only the “Zip Code”, “City”, “Cuisines”, “Average Cost for two”, “Price range”



* The “Zip Code” was renamed to “ZIPCODE” in line with the names of the columns from the census dataframe



* A database was created on postgresql and a database connection was established

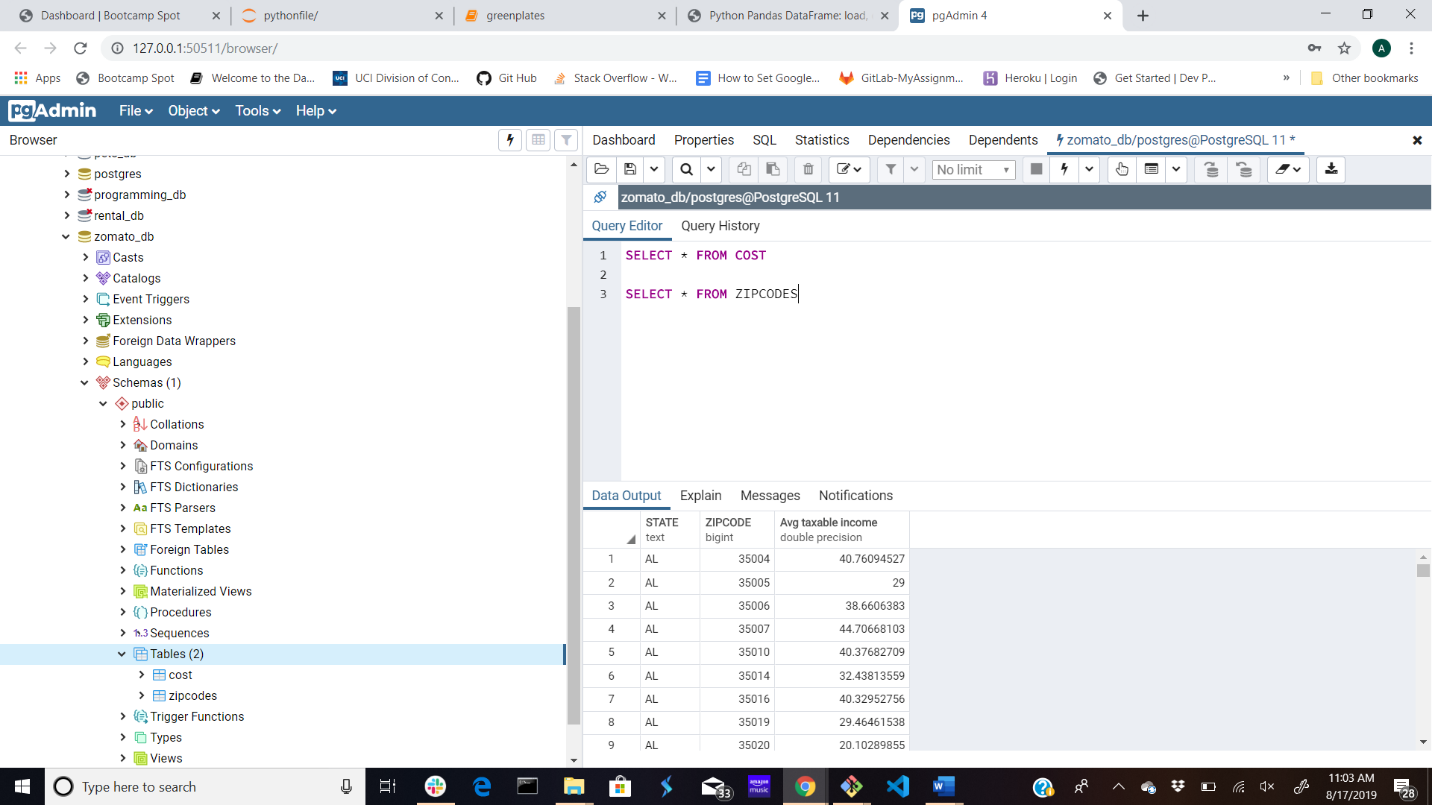


**Load – Make the Connection:**

A database called zomato\_db was created on postgresql and one database connection was established:

* Using the greenplates.ipynb file, the zomato\_df was loaded as a table called ‘cost’ into the zomato\_db database using SQL Alchemy.
* Using the ETLProject.ipynb file, the new\_income\_df was loaded as a table called ‘zipcodes’ into the zomato\_db database using SQL Alchemy.

See below for screenshot of pgAdmin showing zomato\_db with query editor and table schema:



**Application and Future Use:**

The process by which the database was created will be a pattern to copy or a template for future applications.

* The database would be renamed to more generic name that does not use “zomato” as zomato was a source for this iteration. Due to time restraints, the database was not dropped recreated for this purpose.
* The “zipcodes” table would not change, it would be constant, showing the relationship between zipcodes and average taxable income.
* The “cost” table from zomato could be dropped or enhanced with additional data from more extensive sources. The zomato\_df was limited in scope as fewer than 900 restaurants across the United States were listed.
* The “cost” table contains the restaurant, location by zipcode and average price per plate as well as price range from 1 – 5 with 1 being the least expensive and 5 being the most expensive.
* With additional time for development, the Yelp API could be converted into a dataframe and then into a “newcost” table that contains the restaurant, location by zipcode and price range using dollar signs [$] from $ – $$$$$ with 1 ‘$’ being the least expensive and 5 ‘$$$$$” being the most expensive.
* Inferences could be made from the “cost” table to create a variable relationship between cost per plate and price range. This relationship could be applied to the price range within the “newcost” table to infer the cost per plate of the restaurants listed within the “newcost” table, since the Yelp API source is more extensive and includes more restaurants than the Zomato source.
* The application could synthesize the tables by merging the two data sources. The user interface would require a zipcode to find the restaurant within the desired price range.