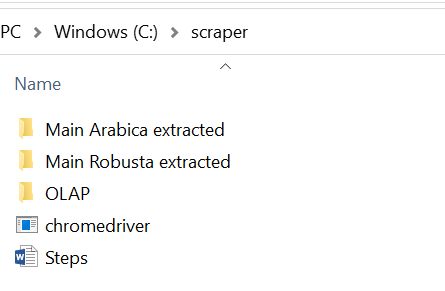
**Experian Coffee DB**

This document give the instructions I follow to create and fill the data in a local DB called Experian Coffee from Coffee Quality Institute website.

The folder **scraper** contain:

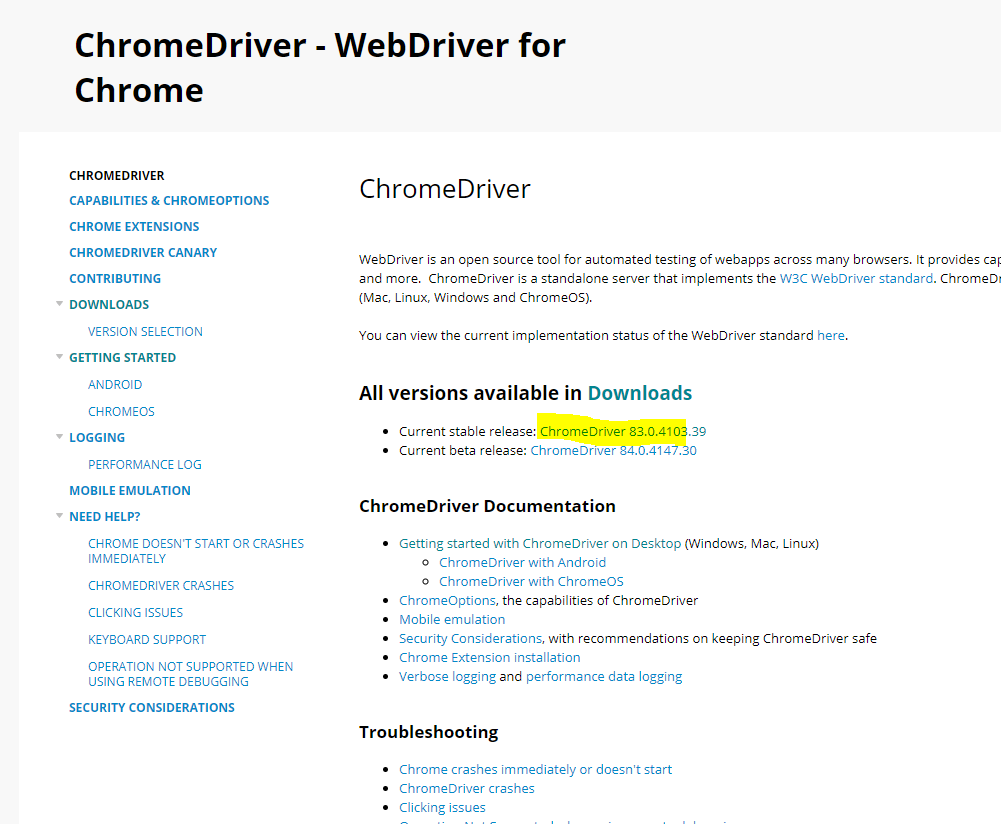


* Main Arabica extracted folder:
  + to all CSV extracted from website in Arabica section
  + to create a consolidated file called *df\_Arabica\_Detail*
  + to Python code for data extraction called *getData\_Arabica.py*
  + to Python code for consolidation step called *process\_table\_Arabica.py*
* Main Robusta extracted folder:
  + to all CSV extracted from website in Robusta section
  + to create a consolidated file called *df\_Robusta\_Detail*
  + to Python code for data extraction called *getData\_Robusta.py*
  + to Python code for consolidation step called *process\_table\_Robusta.py*
* OLAP folder:
  + to Analysis Services (SSAS solution)
  + to ETL Experian Coffee (SSIS solution)
  + to CSV files with data after data transformation
  + to Power BI file with the data transformation steps
  + to DDL for tables creation called *ExperianCoffee\_DDL.sql*
  + to testing data after ETL data loading
* *chromedriver.exe*
* *Steps file*

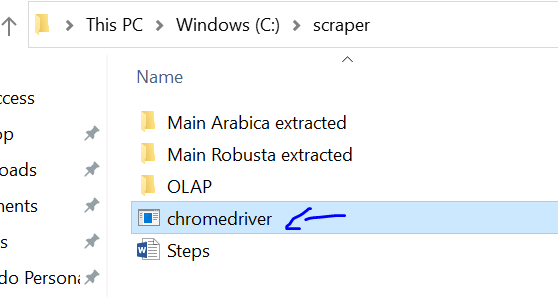
**Setup tools:**

1. Download the Chrome WebDriver

<https://chromedriver.chromium.org/>

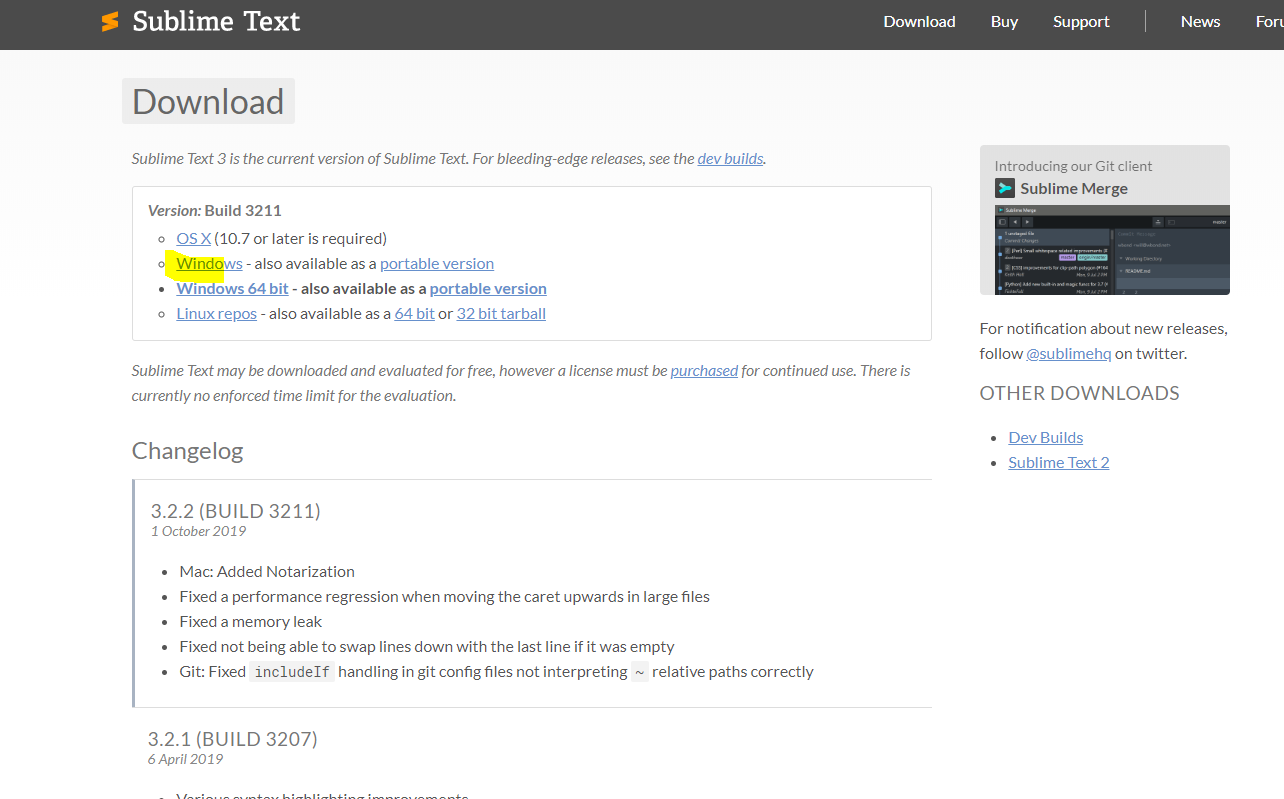


1. Copy the .exe file to the local folder



1. Download and install **Sublime** **text** app to work with Python.

<https://www.sublimetext.com/>

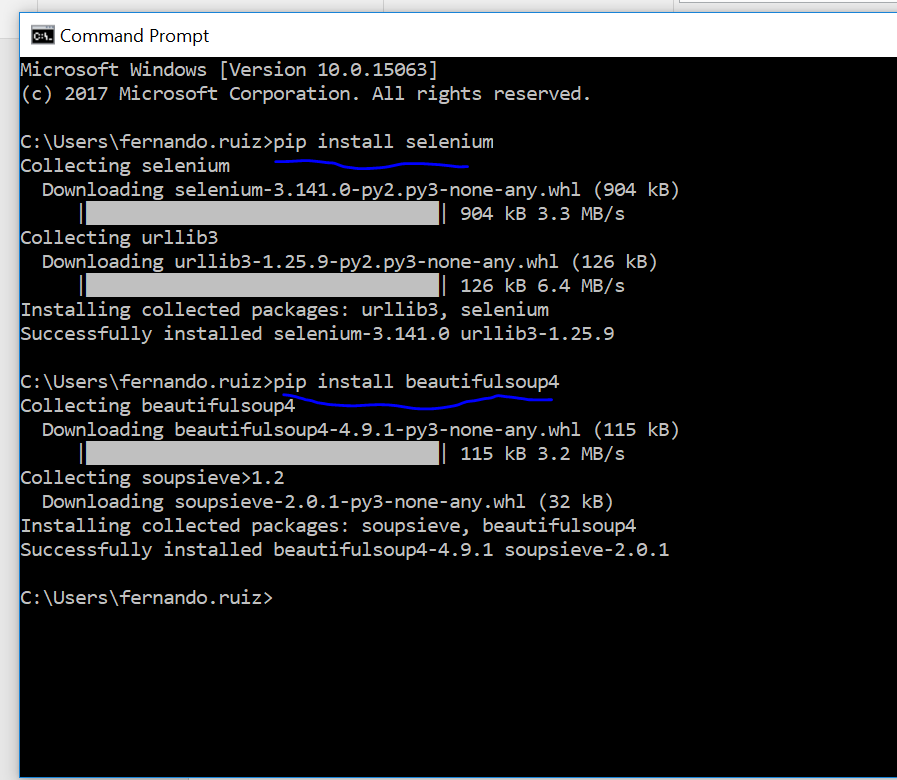


1. In CMD Install:

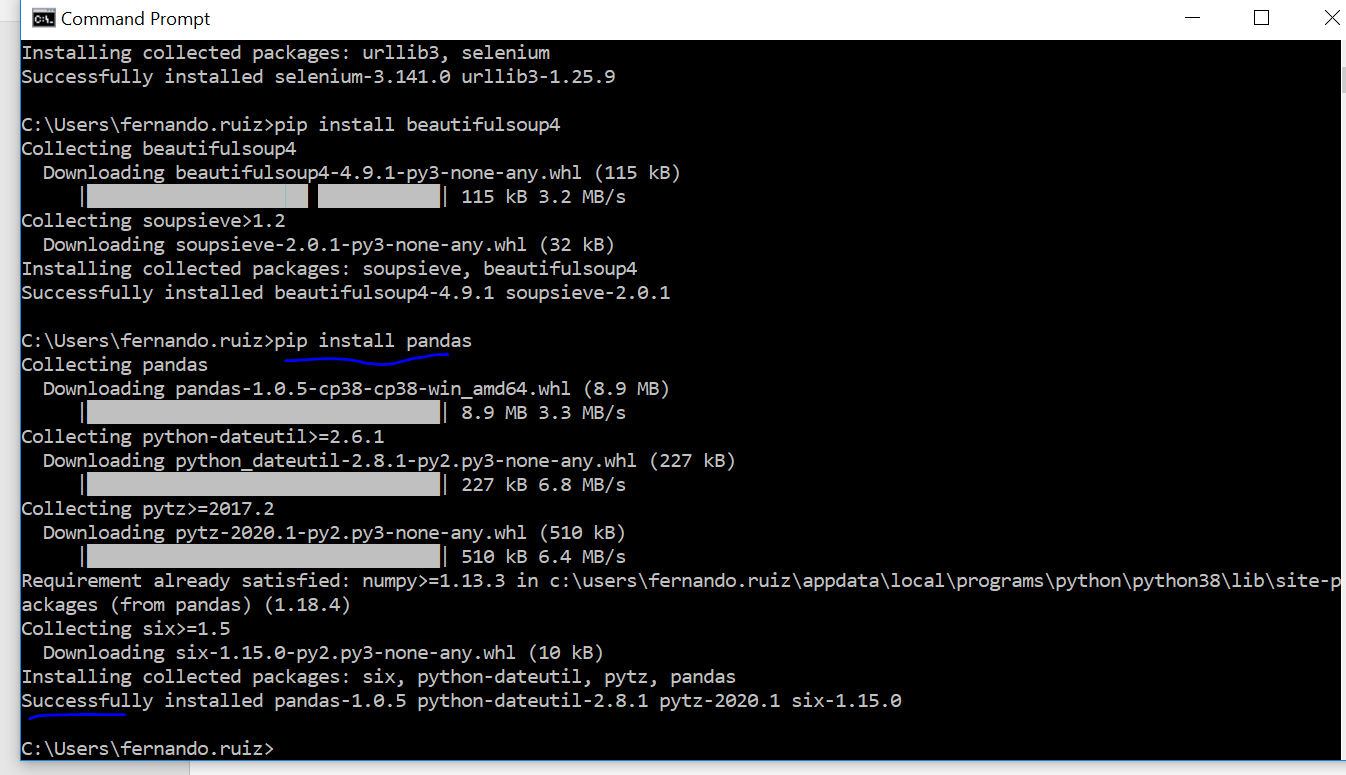
* **pip install selenium**



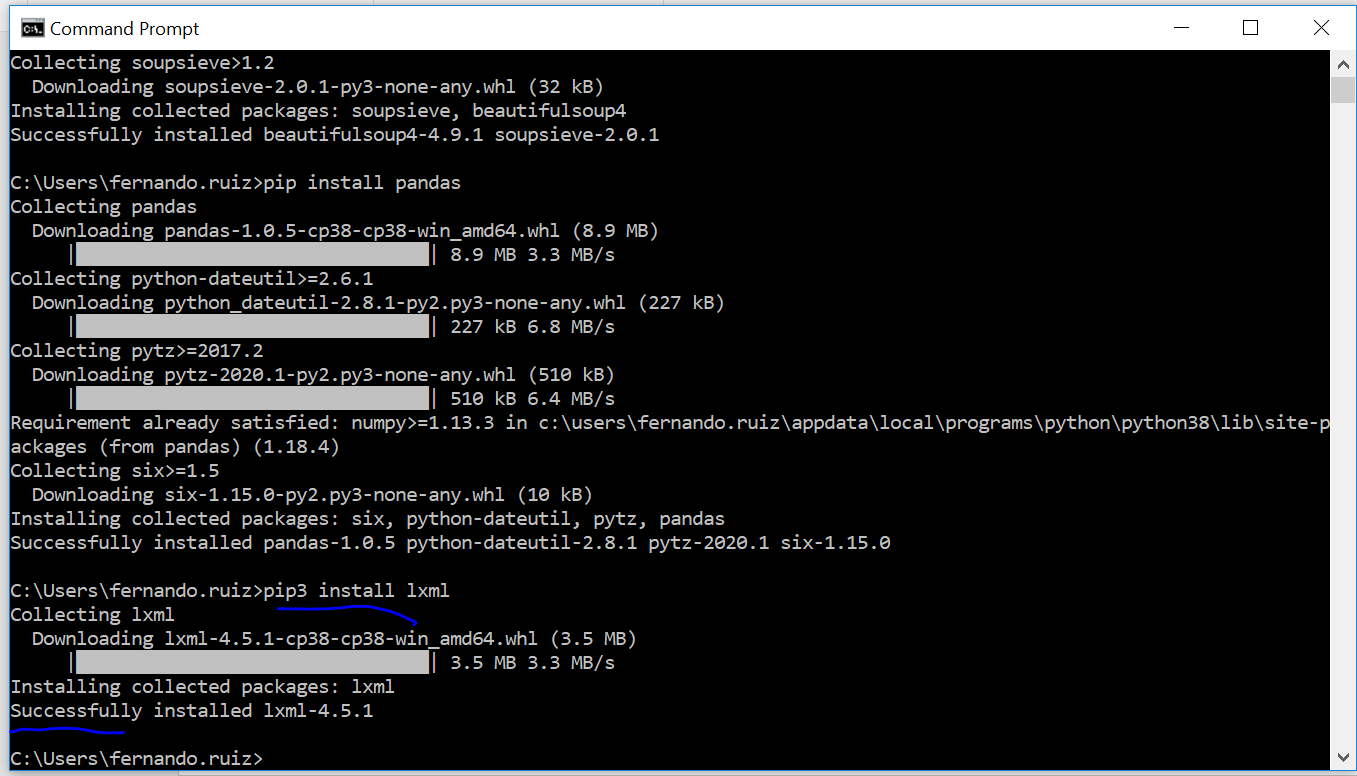
* **pip install beautifulsoup4**



* **pip install pandas**

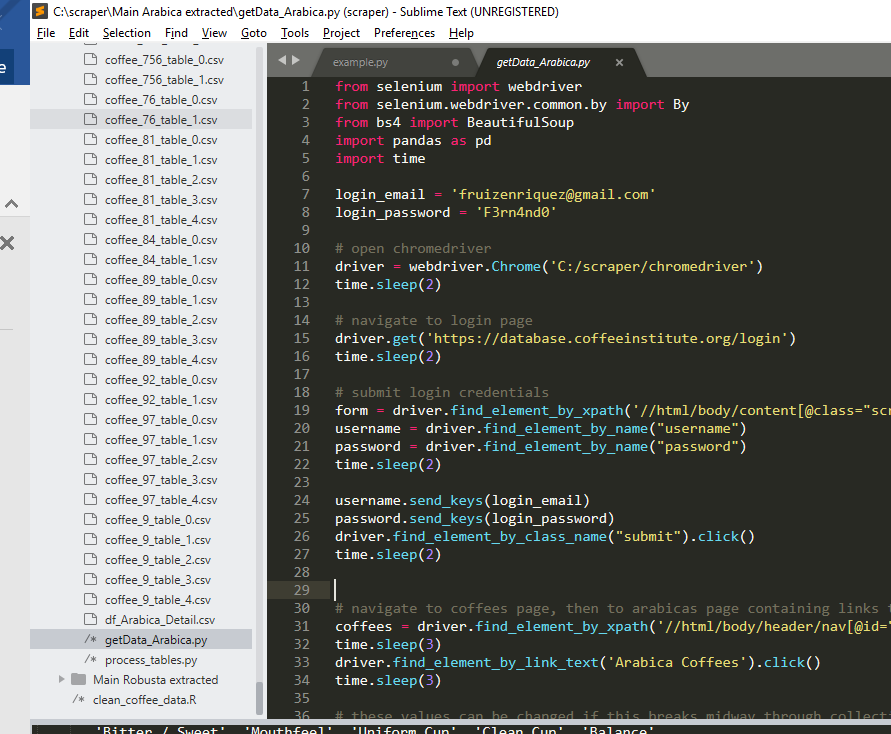


* **pip3 install lxml**



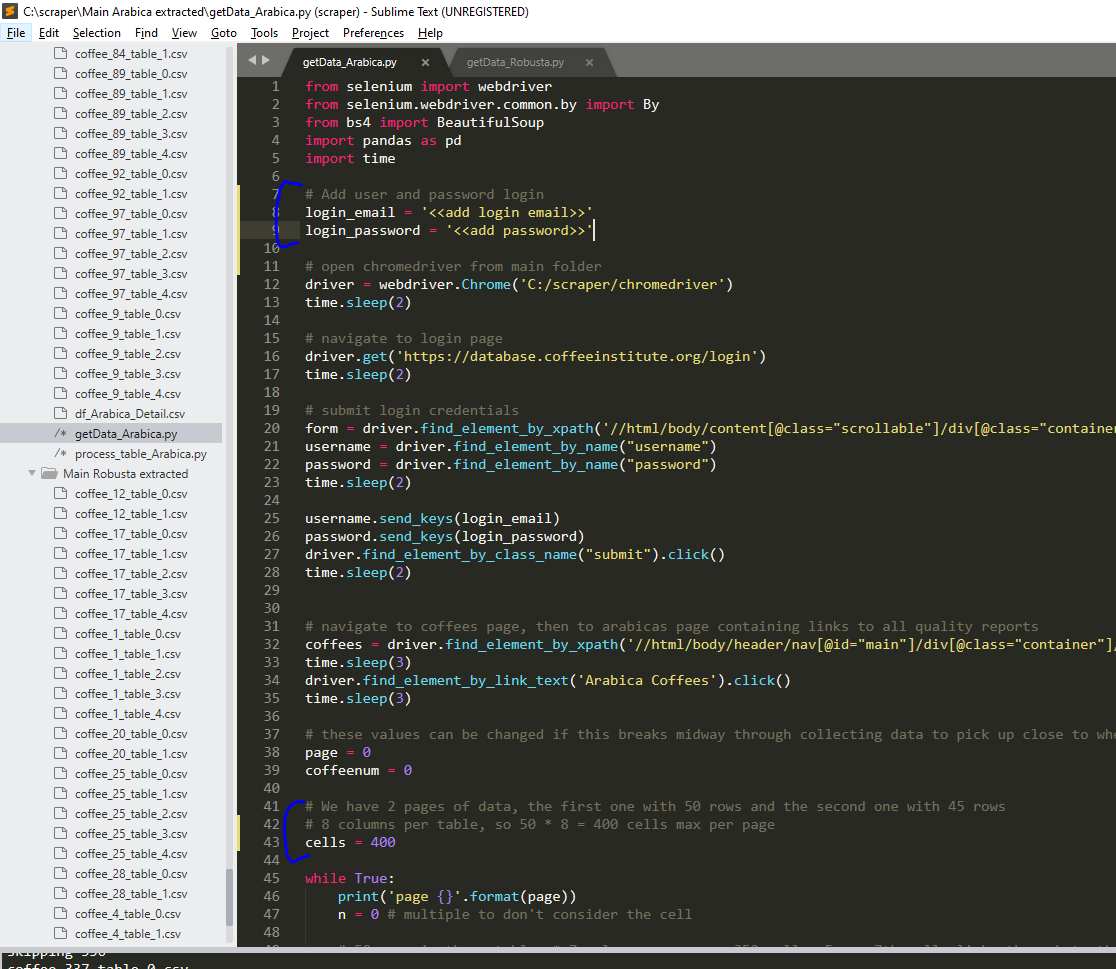
**Extract and consolidate data from website:**

1. Go to the *Main Arabica extracted* folder and run the Python code in Sublime called “*getData\_Arabica.py*”



**Note before to run:**

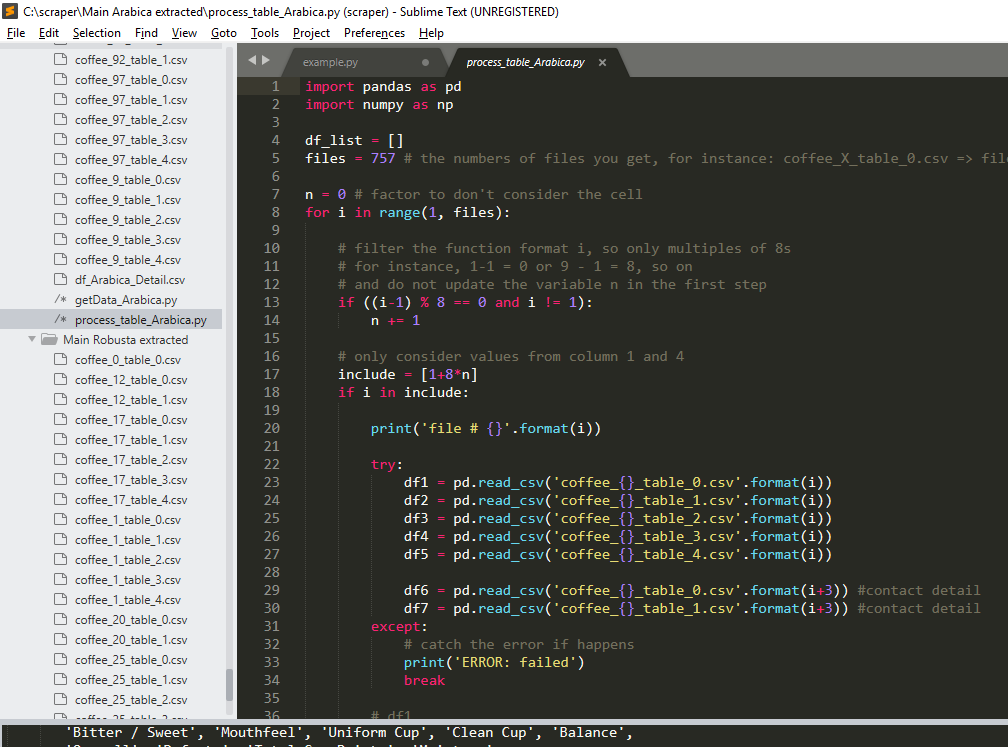
* Add user login and password
* Update the number of cells according total rows per page. For instance, cells = **400**
* Tools > Build System > Python
* Tools > Build



**Note**: it takes approximately 40 min to be ready.

Then to consolidate the files:

* Open the Python code called “*process\_table\_Arabica.py*”
* Run the code Tools > Build
* Will get the Arabica dataframe called “*df\_Arabica\_Detail.csv*”.



**Note to investigate:**

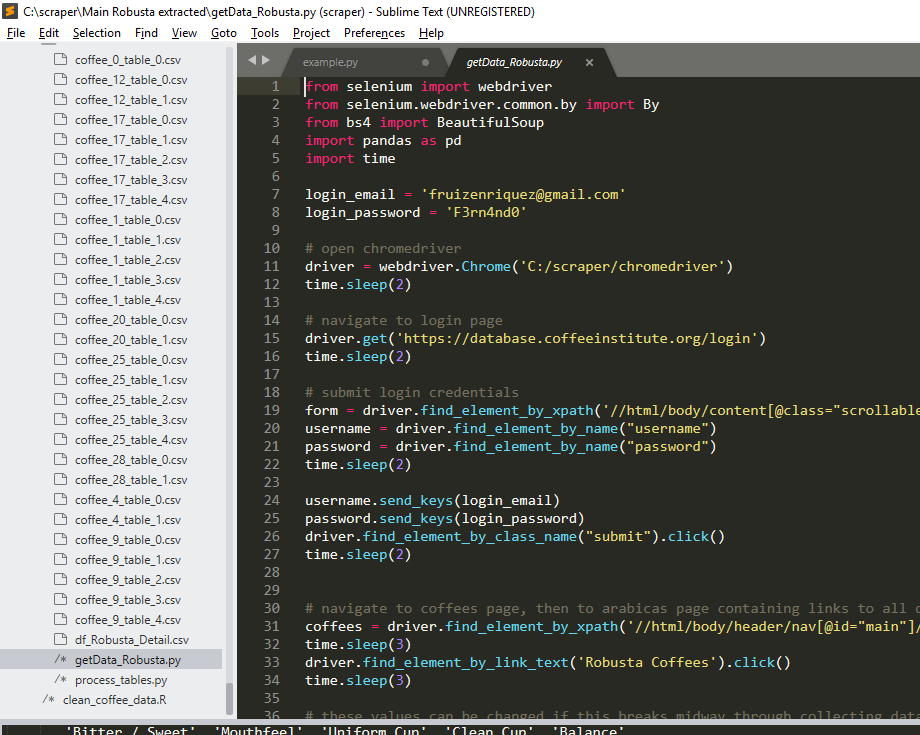
The files coffee\_481\_table\_3 and coffee\_481\_table\_4 comes duplicate due to an issue in the webpage where the tables are split.

**Manually fix the issue:**

* Remove the *coffee\_481\_table\_4* file
* Rename the file *coffee\_481\_table\_5* as *coffee\_481\_table\_4*
* Repeat the same step for the file *coffee\_721\_table\_4*
* Run again

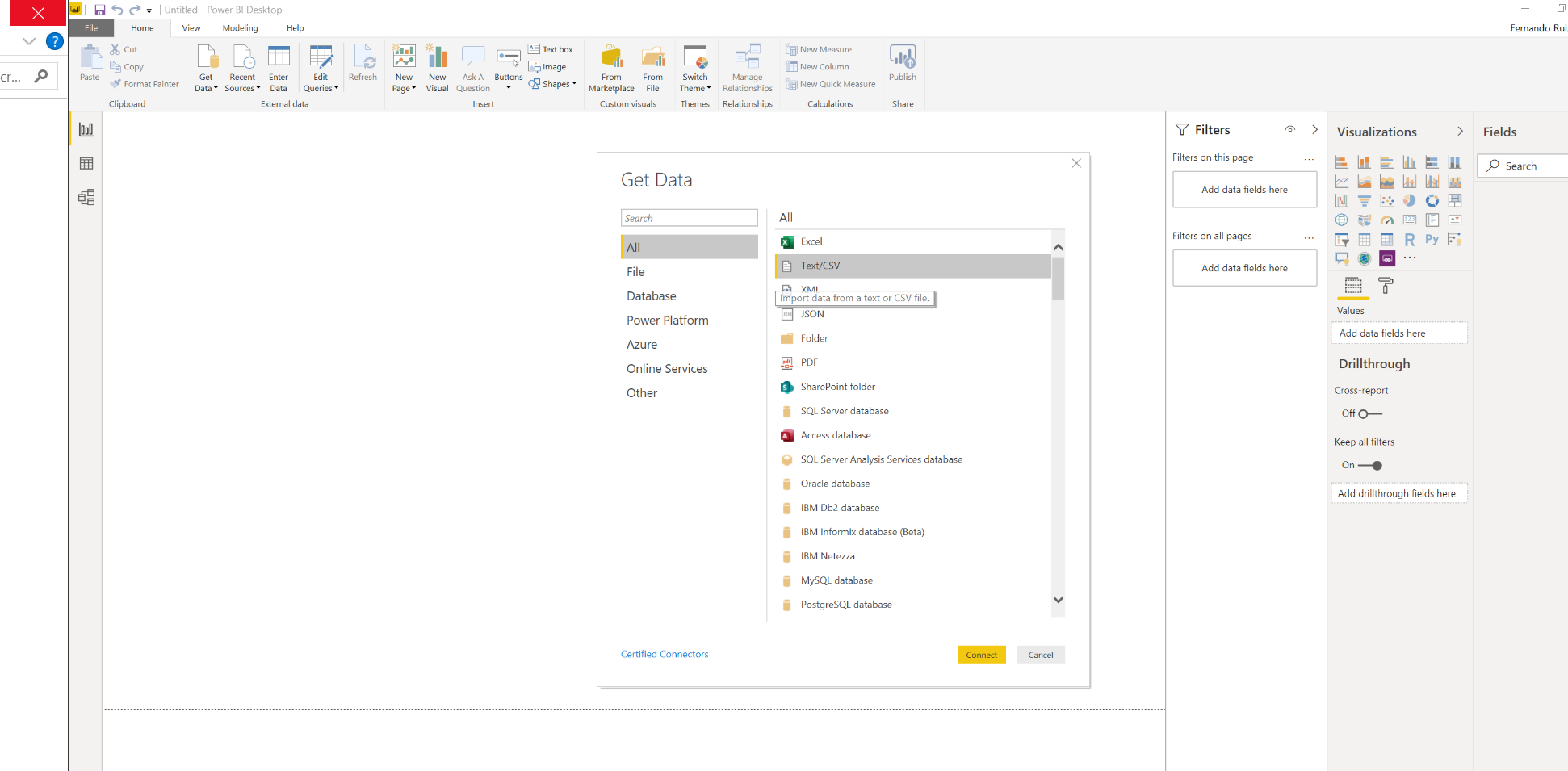
1. Repeat the same steps from **point 5** for Robusta in the “*Main Robusta extracted*” folder

* run the Python code called “*getData\_Robusta.py*”
* run the Python code called “*process\_table\_Robusta.py*”
* We will get the Robusta dataframe called “*df\_Robusta\_Detail.csv*”.



**Data cleaning:**

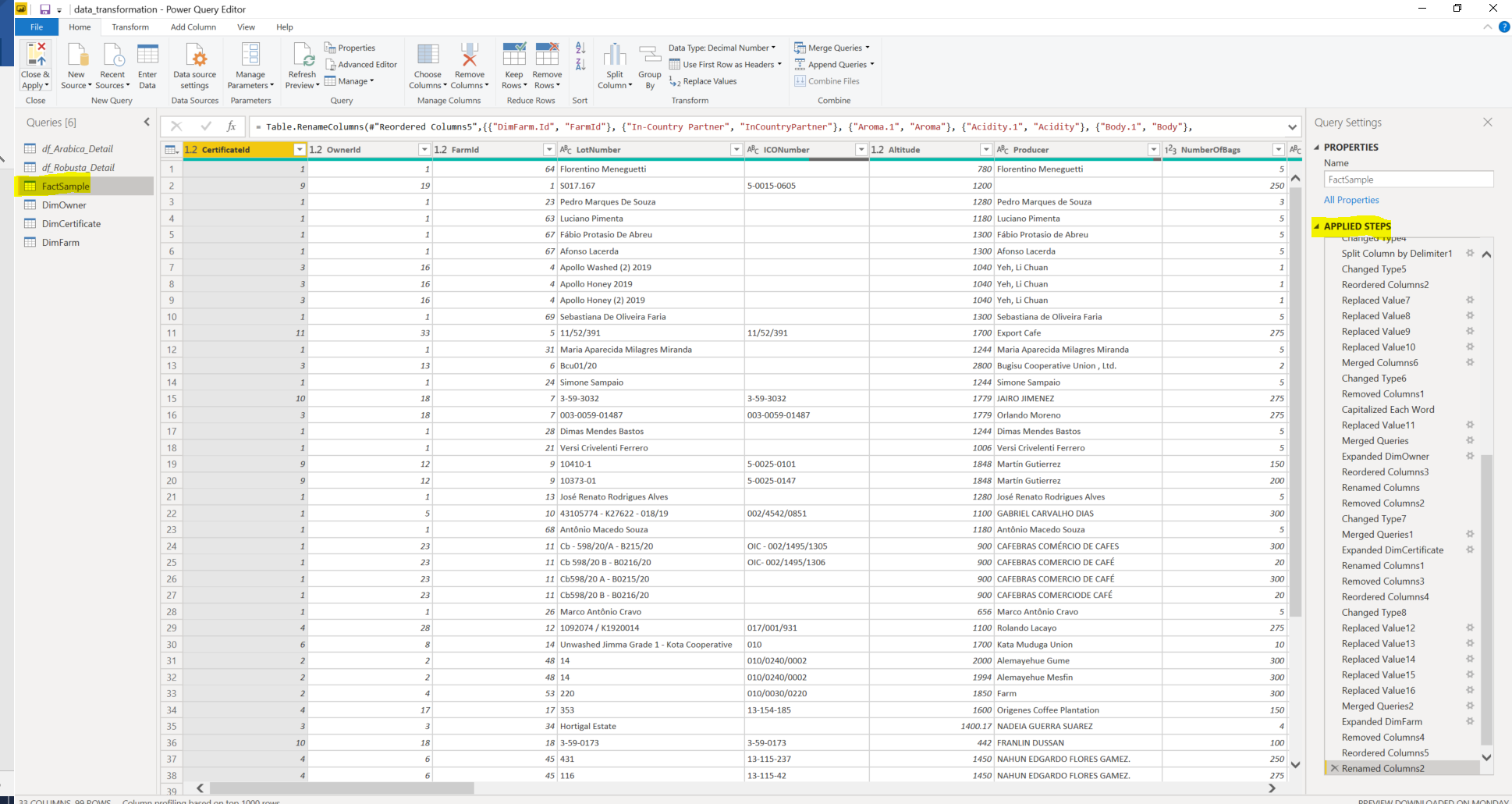
1. Working in Power BI to clean/transform the data



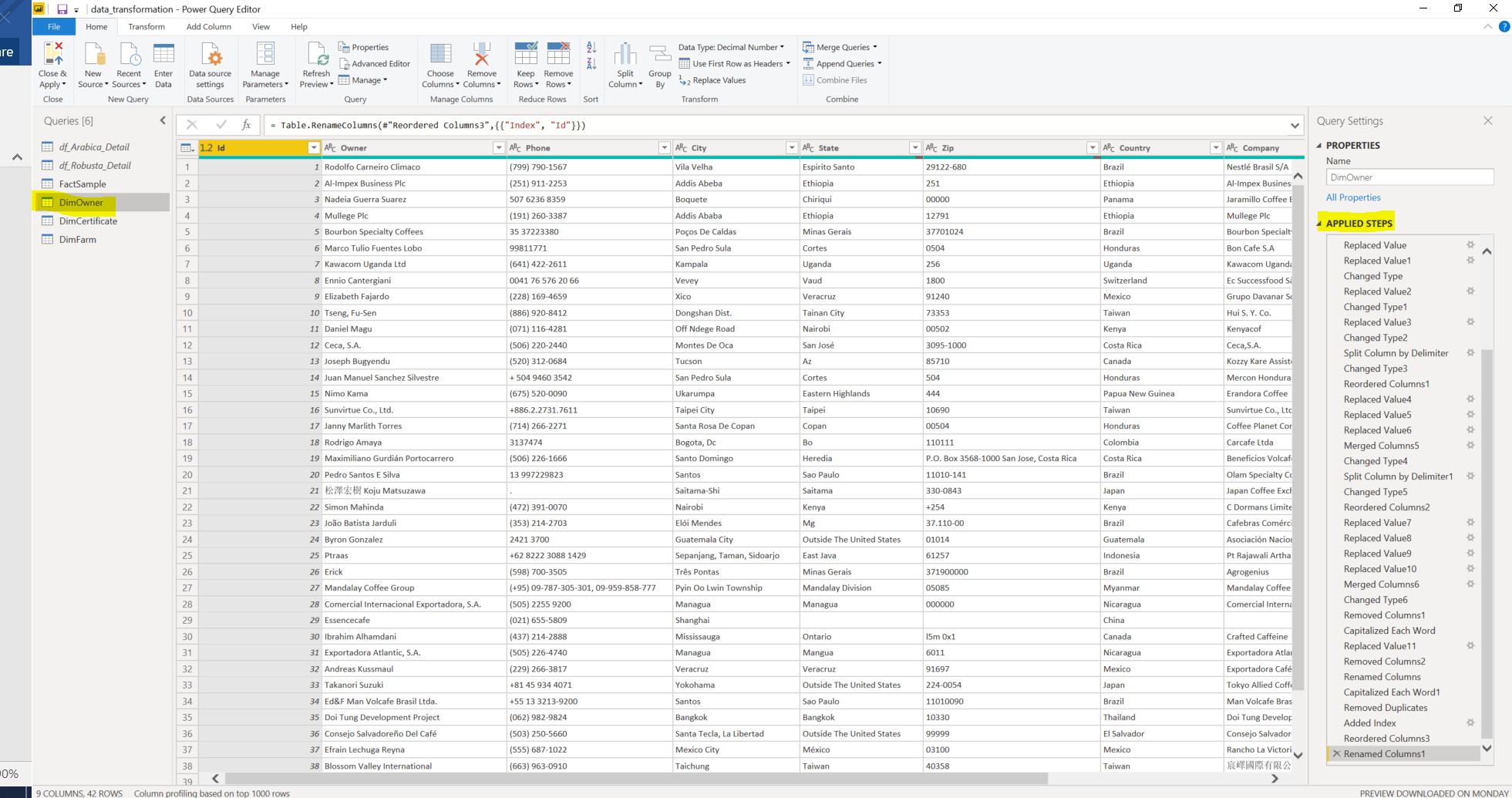
When import make sure to choose File Origin **Unicode UTF-8** to allow Power BI recognize Asian and Arabic characters, like Japanese language.



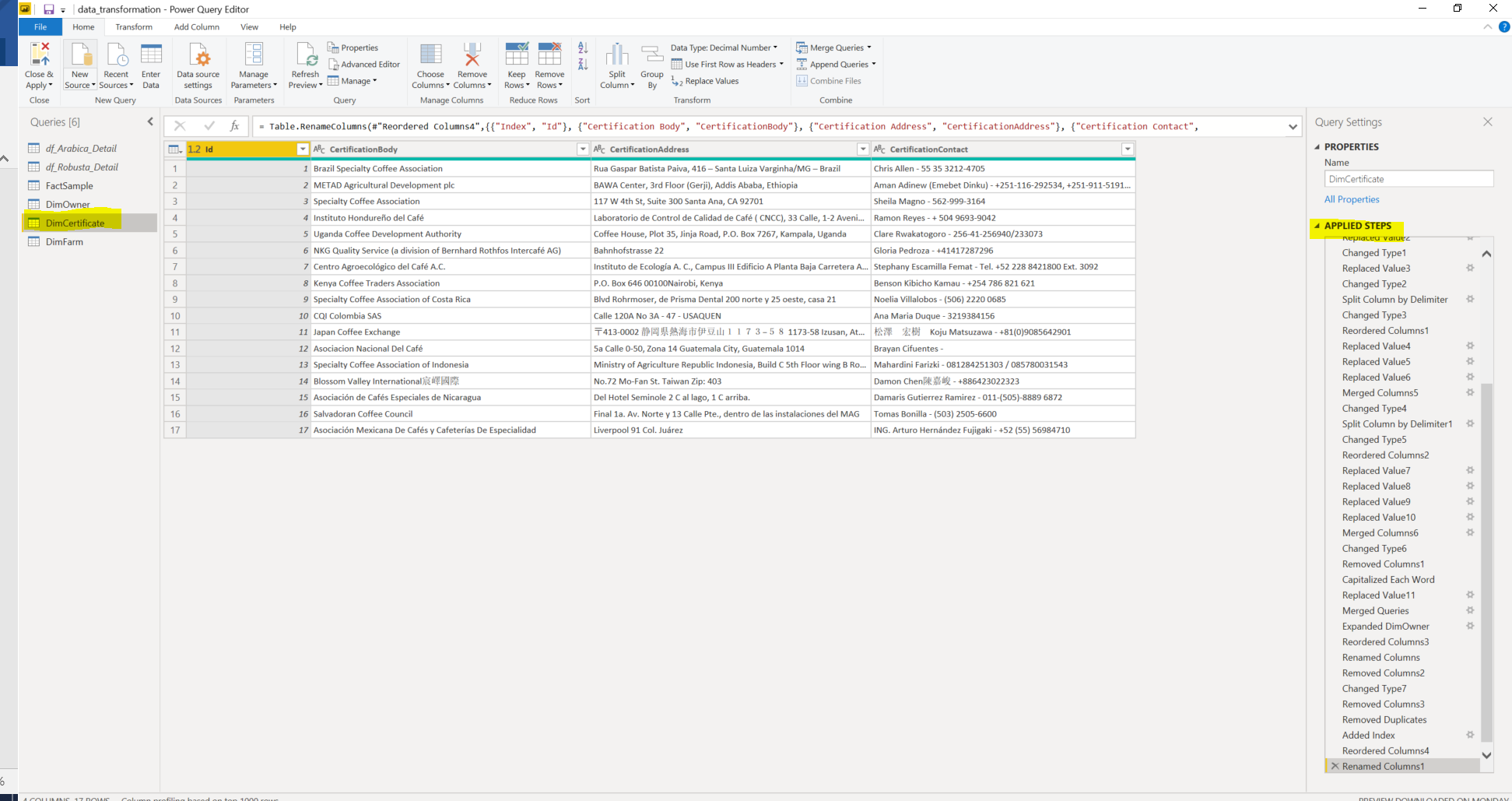
* FactSample



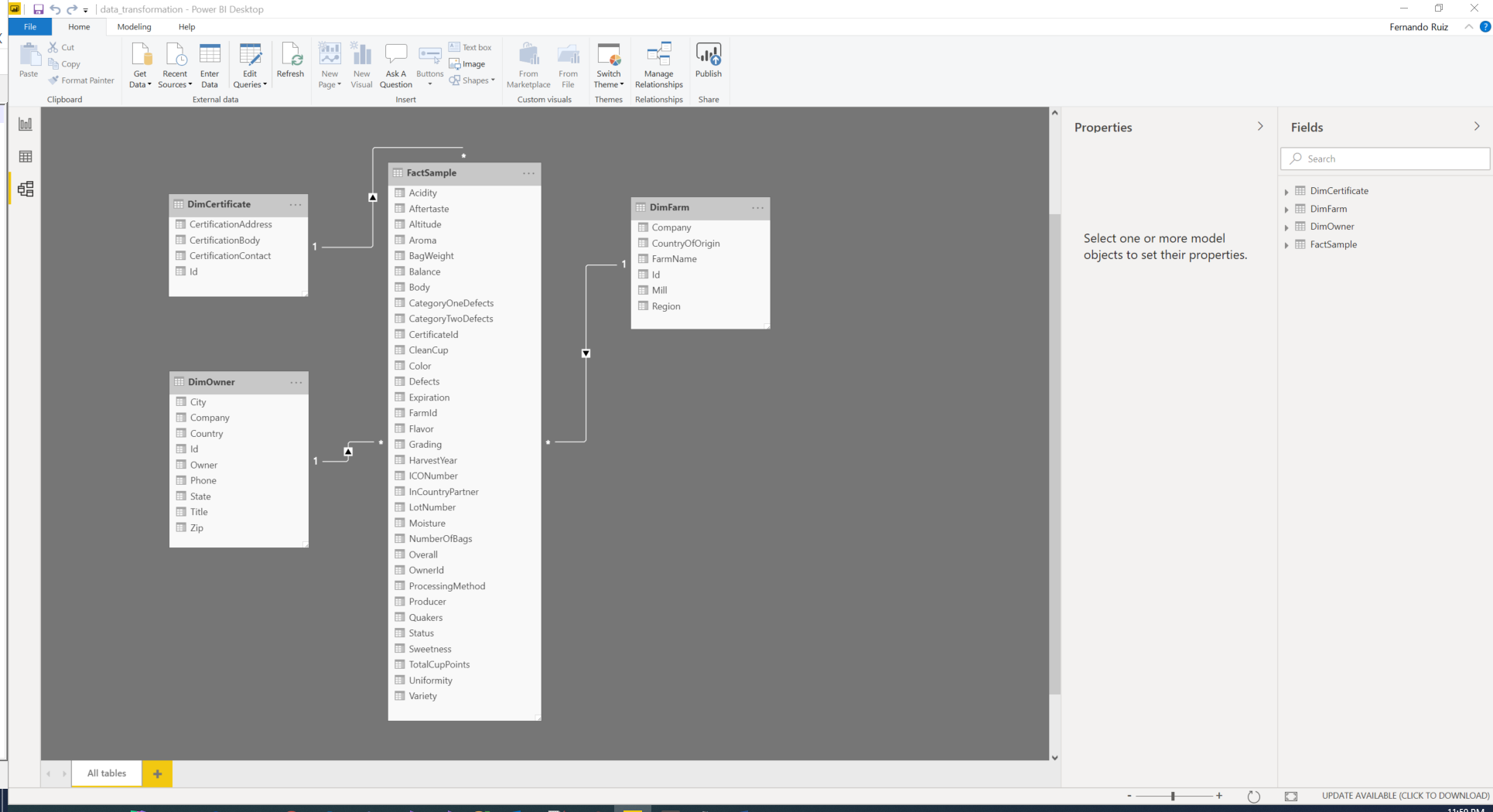
* DimOwner



* DimCertificate

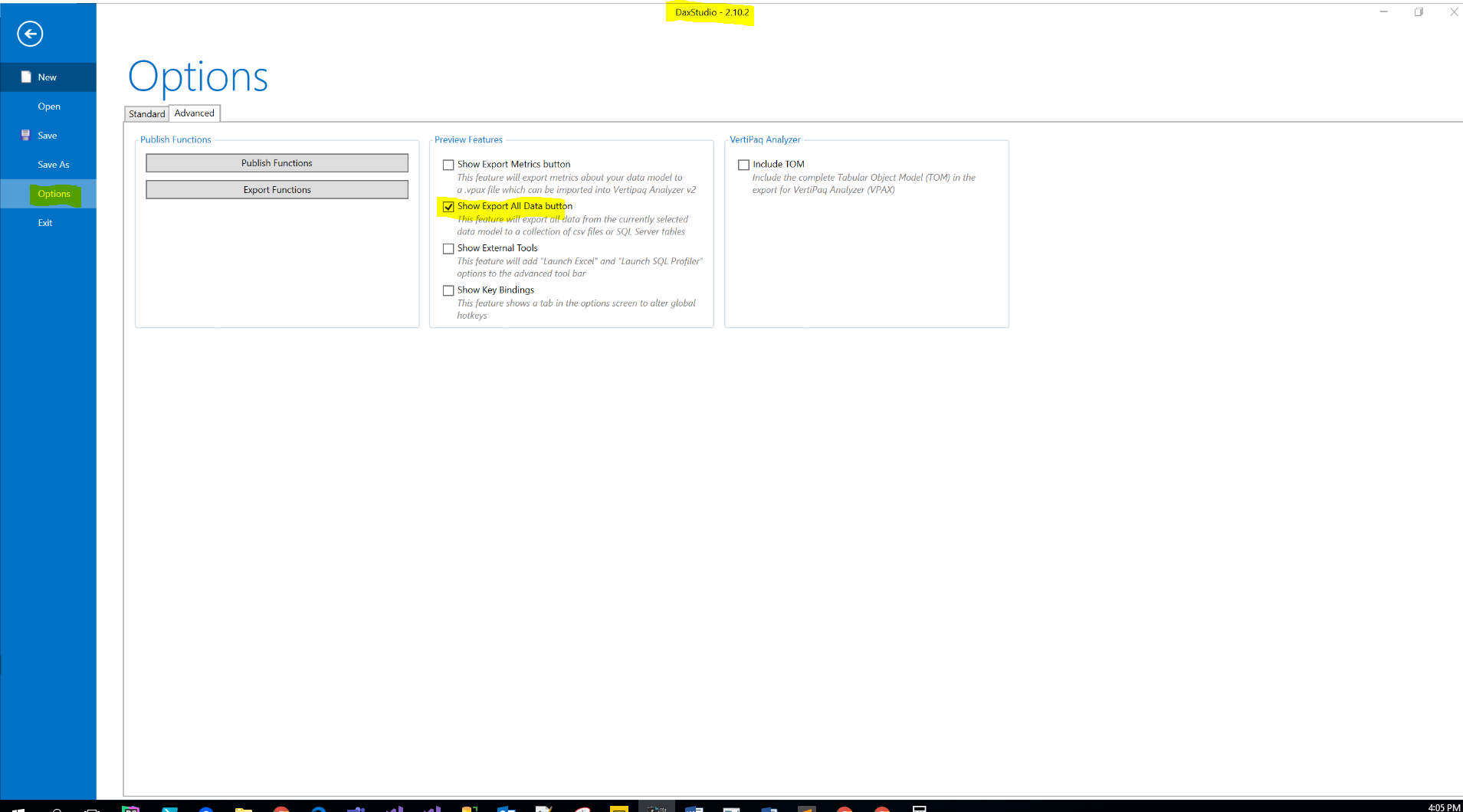


* Diagram DB

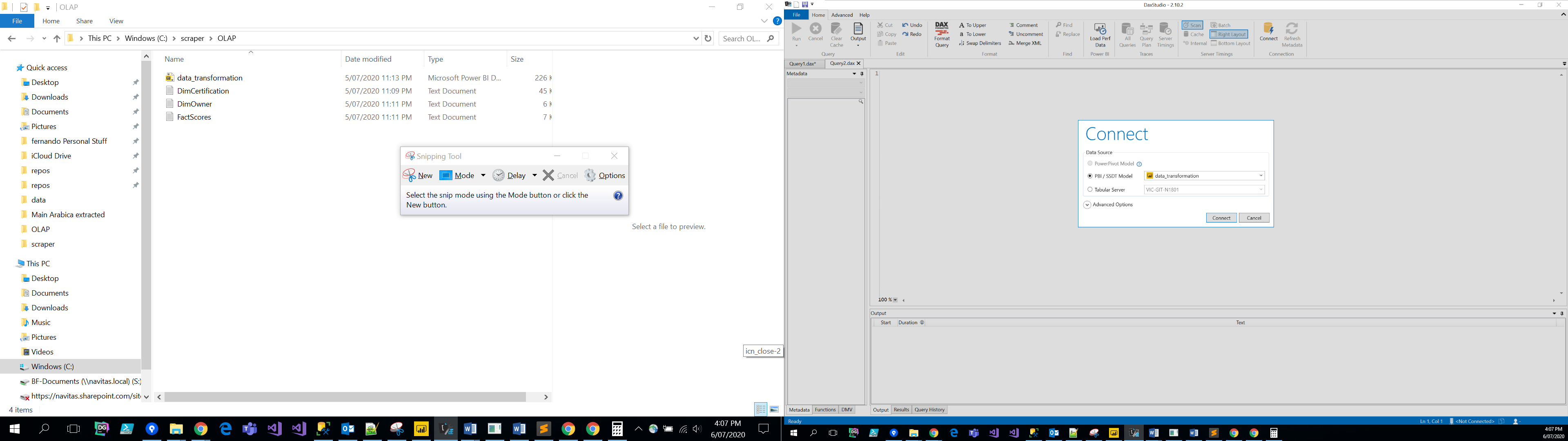


1. Export the information using DAX STUDIO

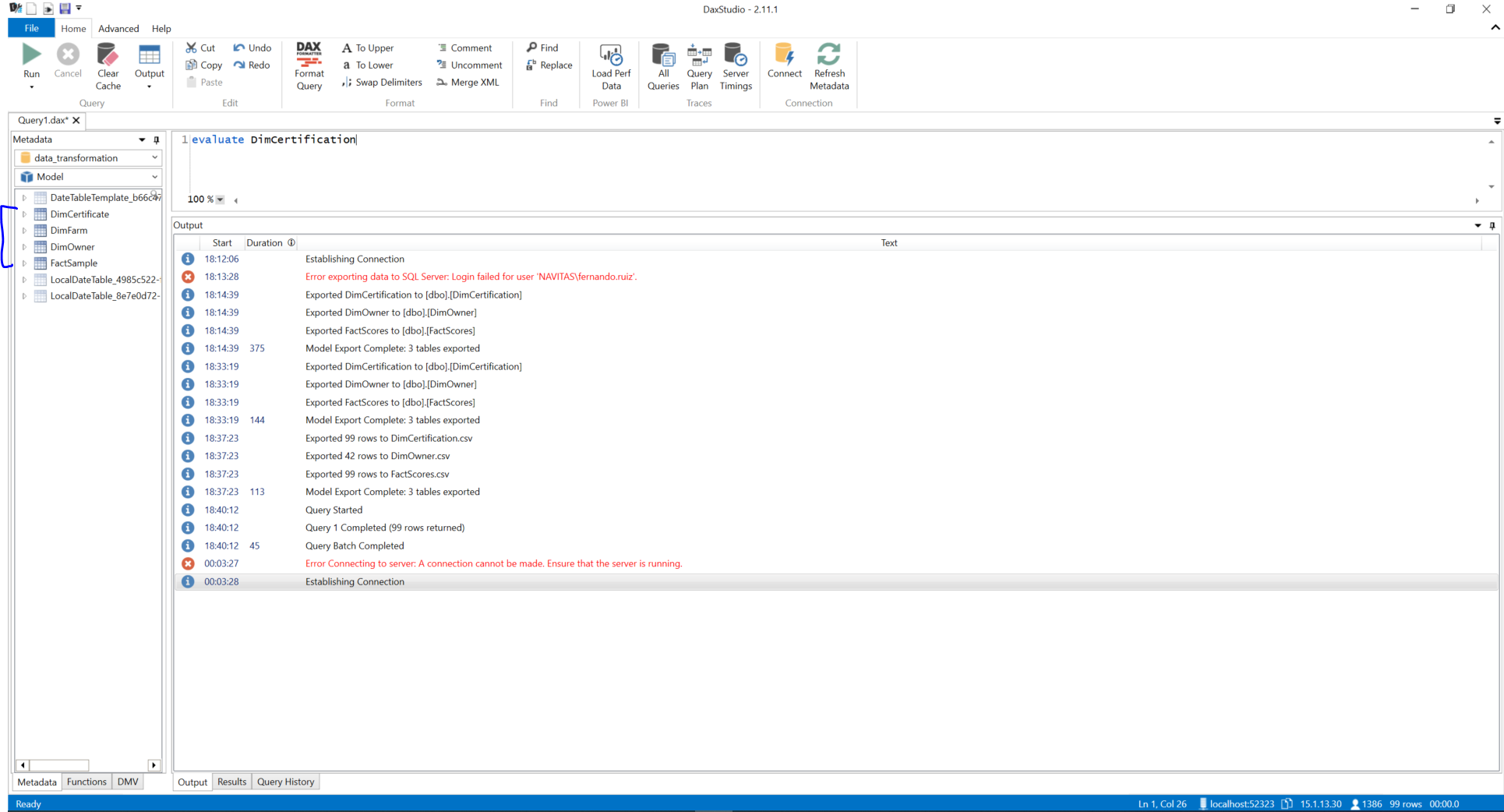
* Update the configuration



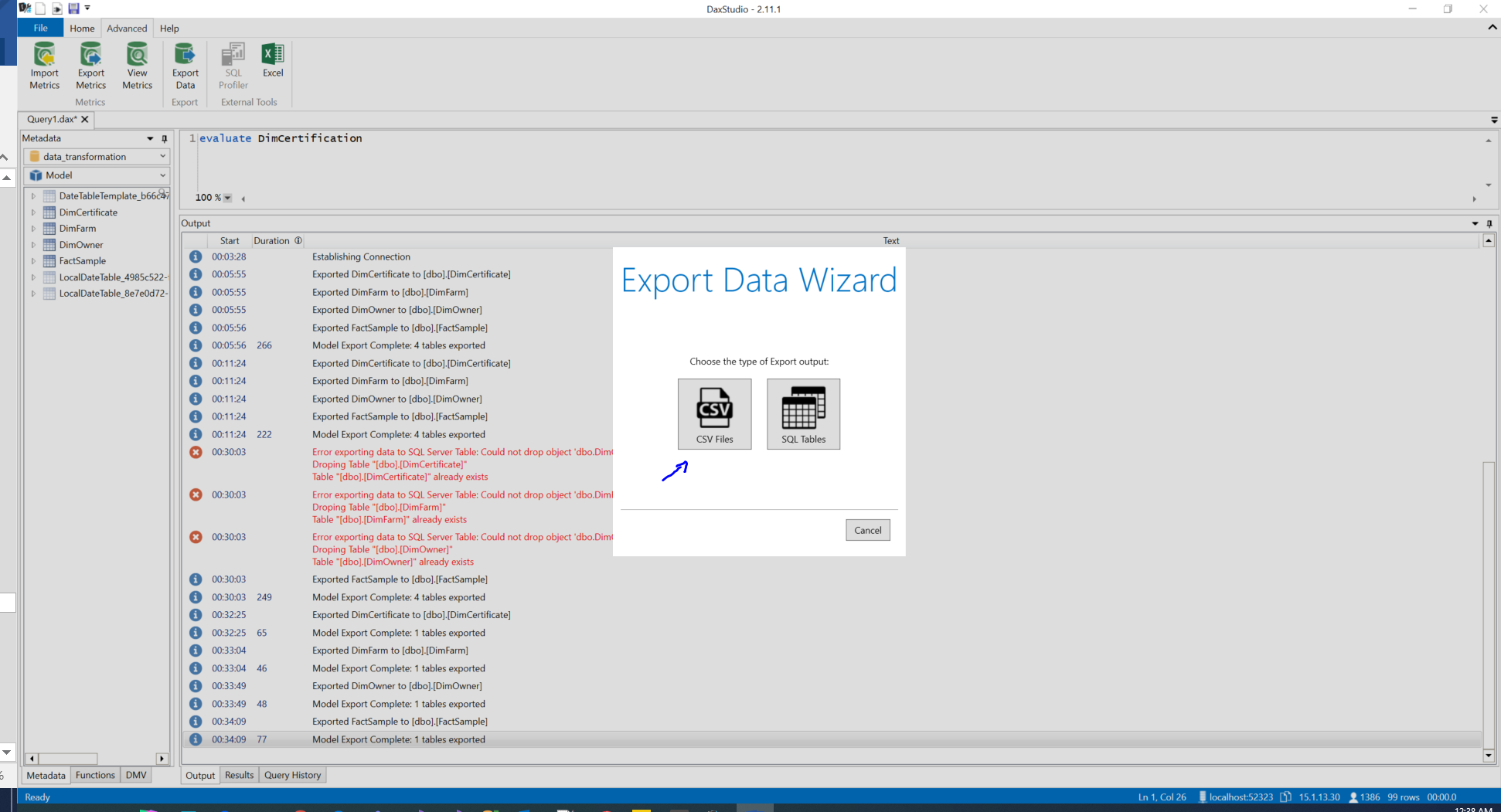
* Select the data model worked on Power BI in the previous step.



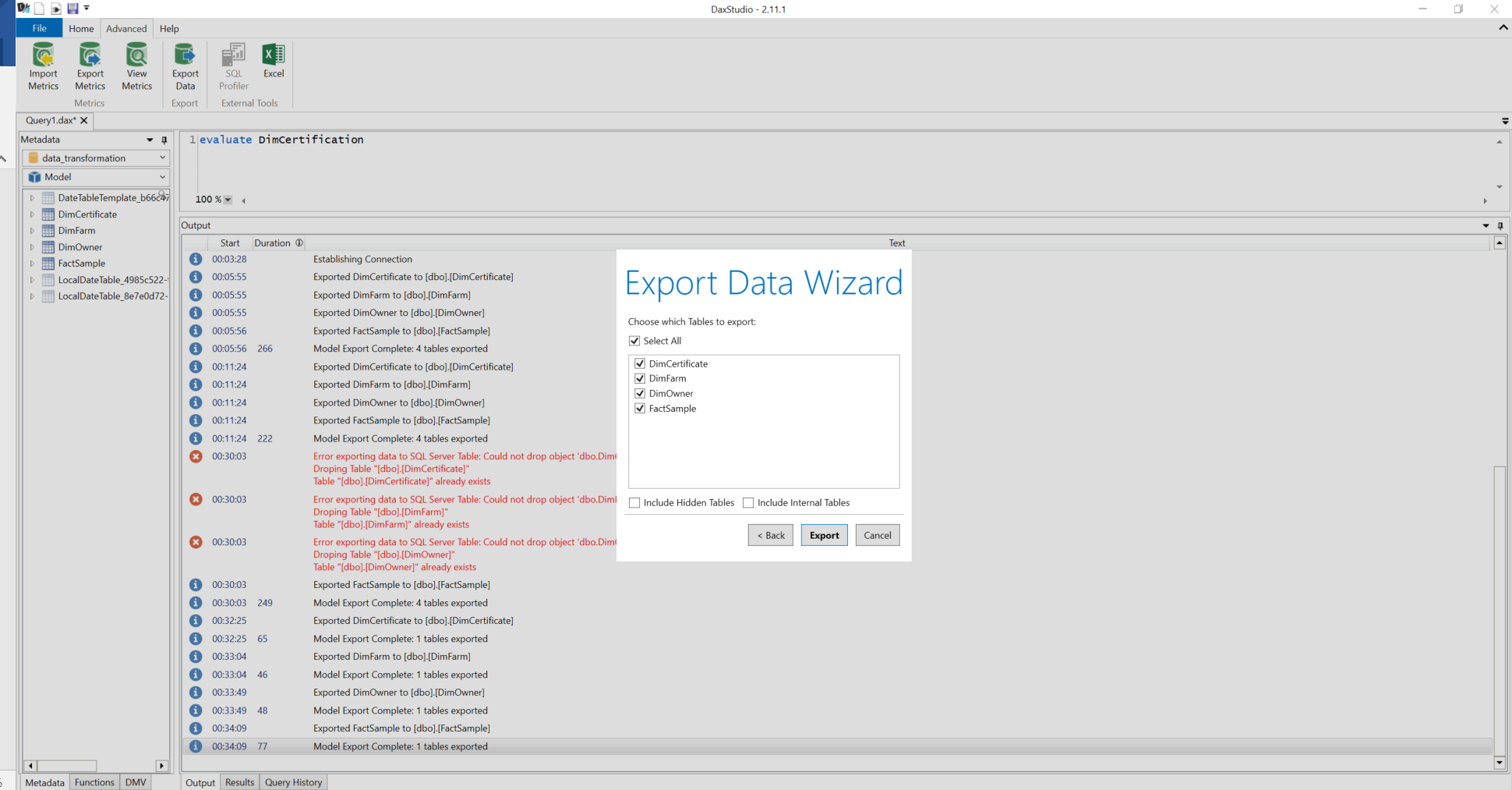
1. Export data

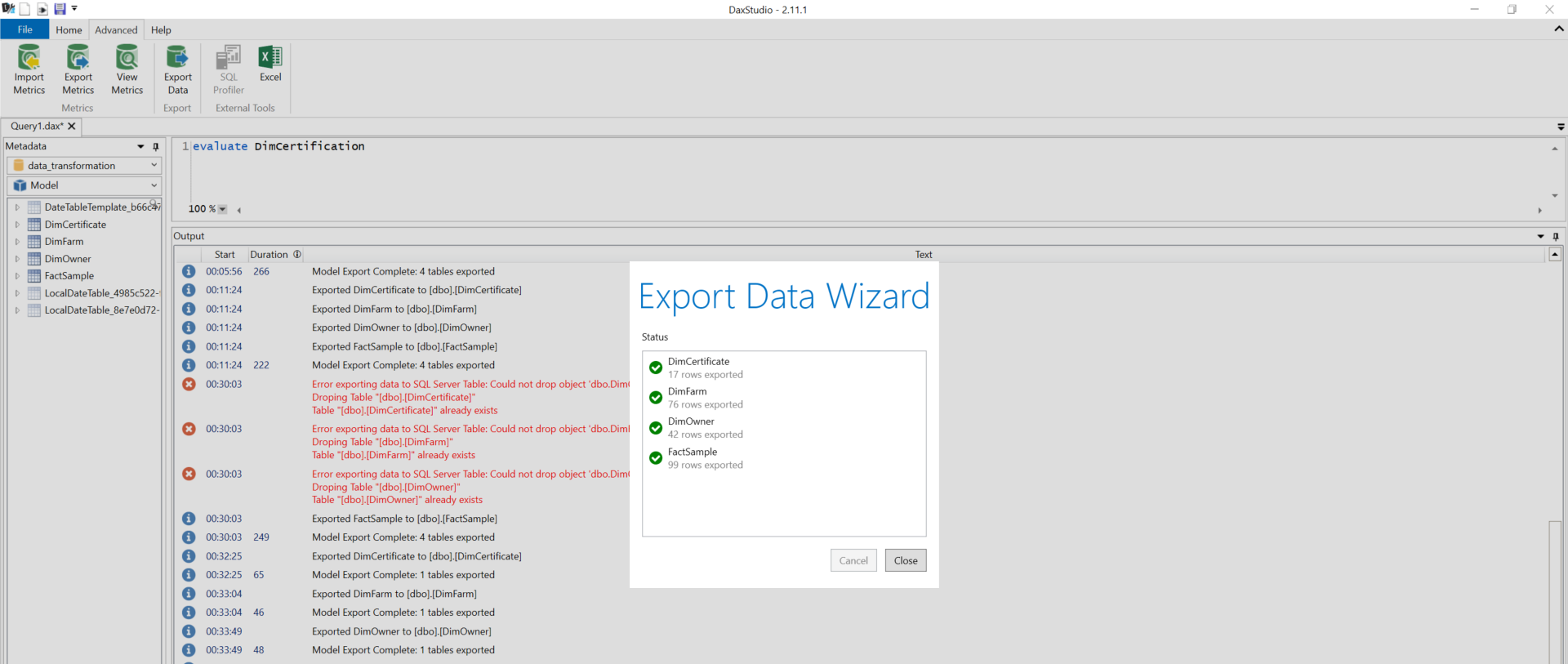


* Select CSV files



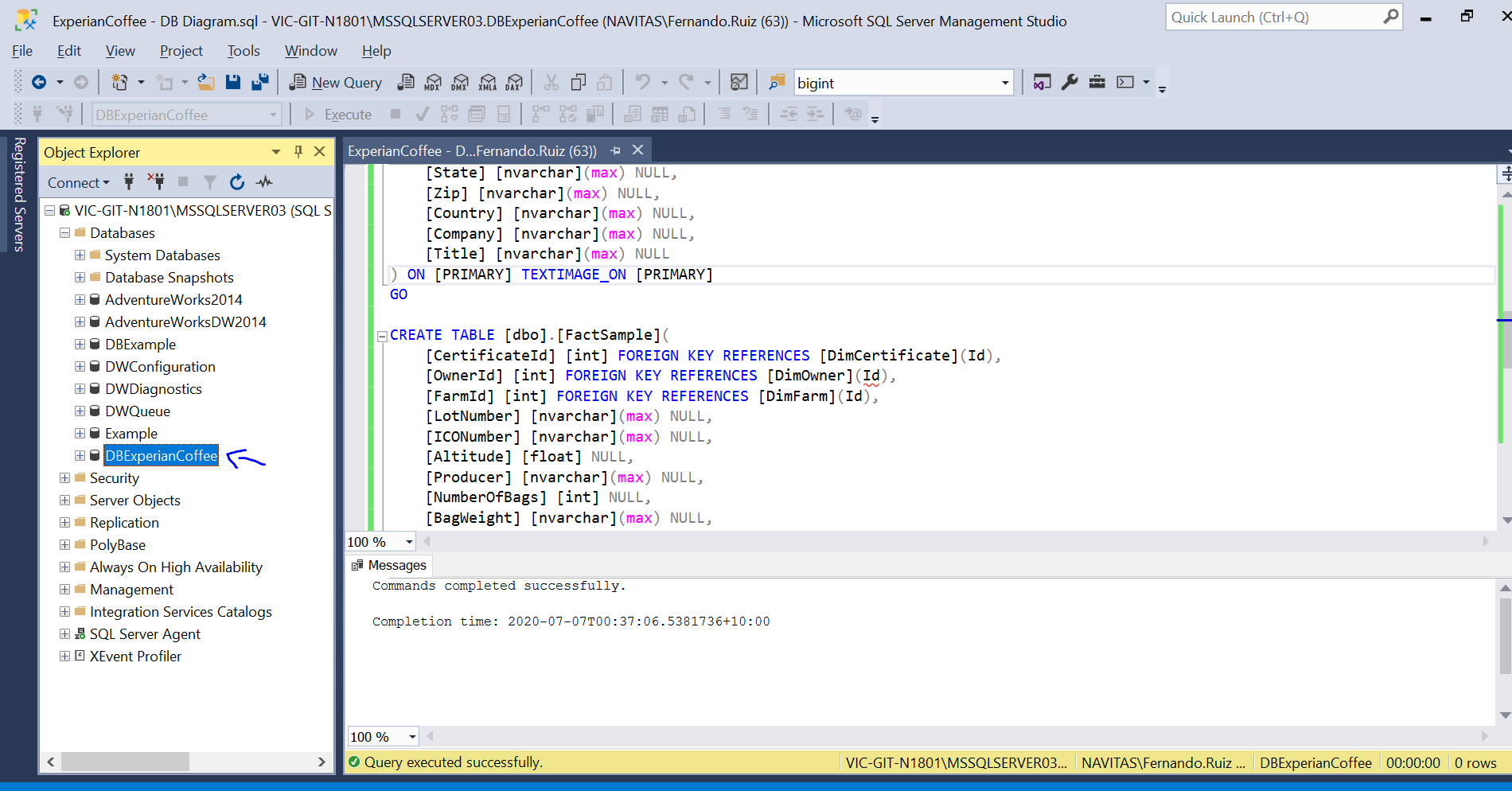
* Select the location and tables to be exported.
  + Location: \scraper\OLAP



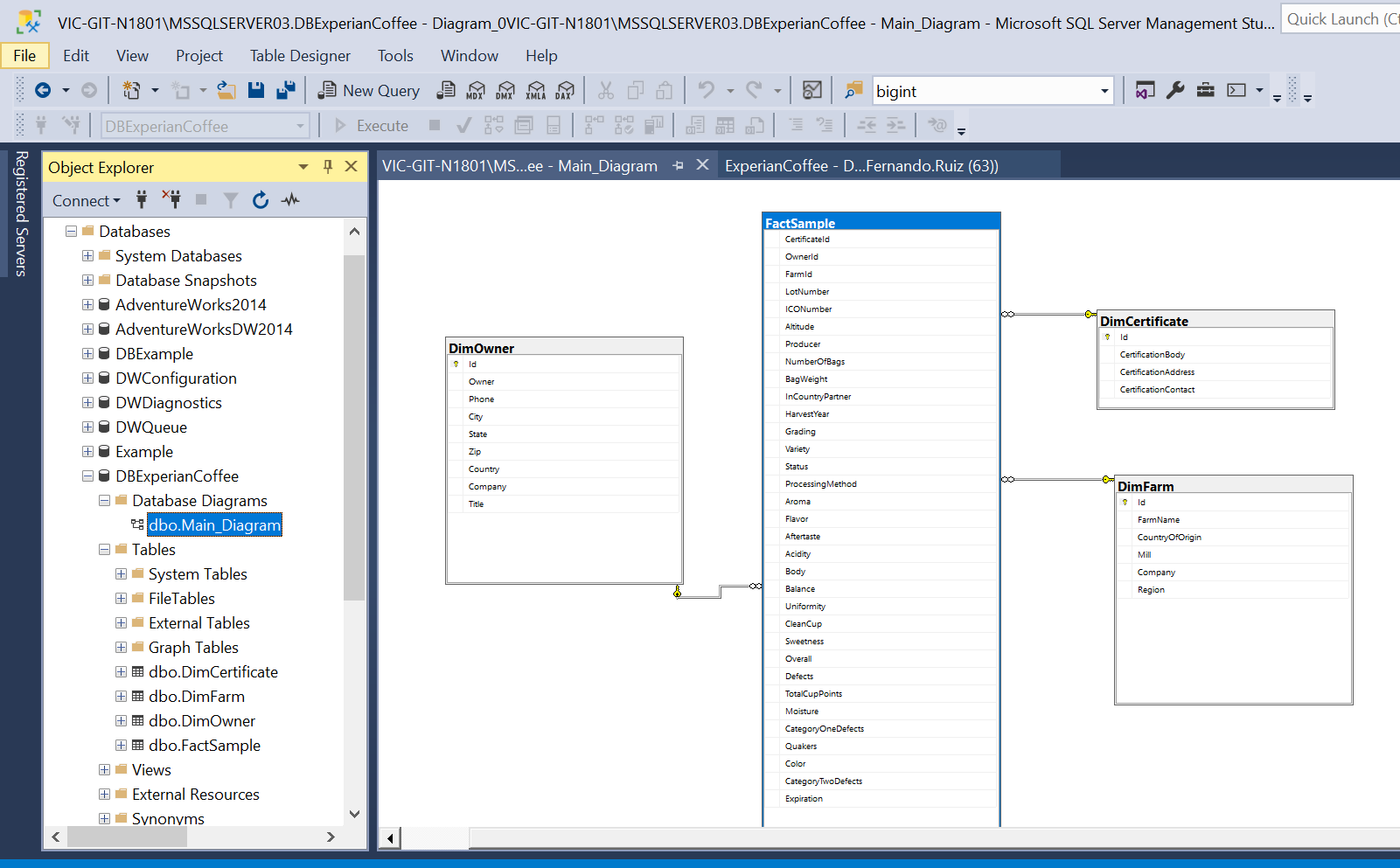


* Create DB in SQL, for this topic was called **DBExperianCoffee**

Note: SQL Server name is VIC-GIT-N1801\MSSQLSERVER03



* Run here the “*ExperianCoffee\_DDL.sql*” script to create tables and FKs constraints

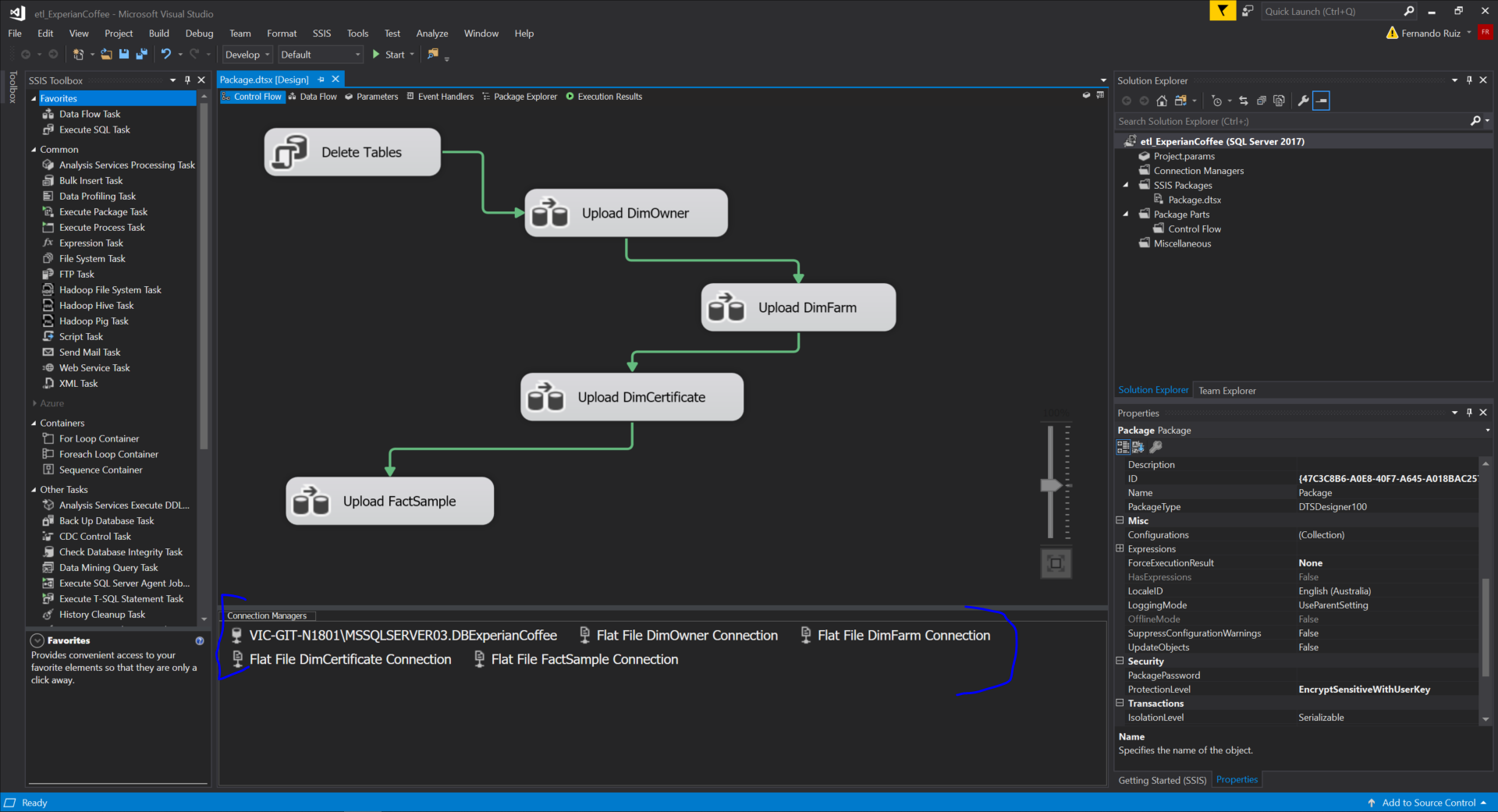


Note: SQL connection

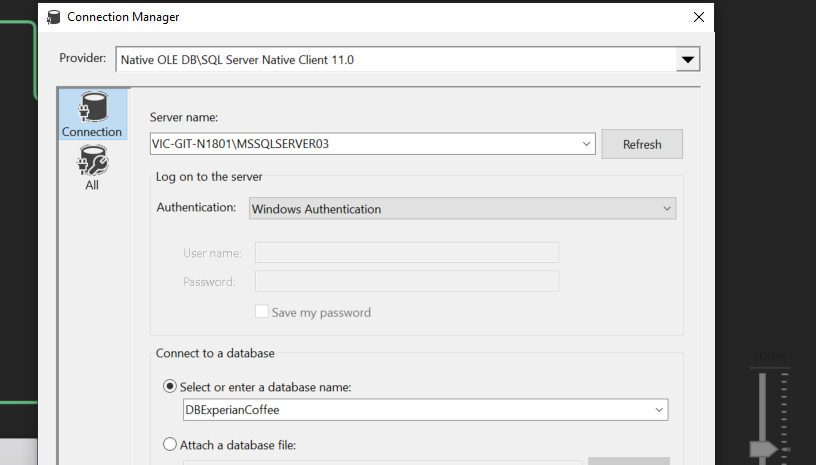
**Server=VIC-GIT-N1801\MSSQLSERVER03;Database=DBExperianCoffee;Trusted\_Connection=True;**

**Load the data by SSIS:**

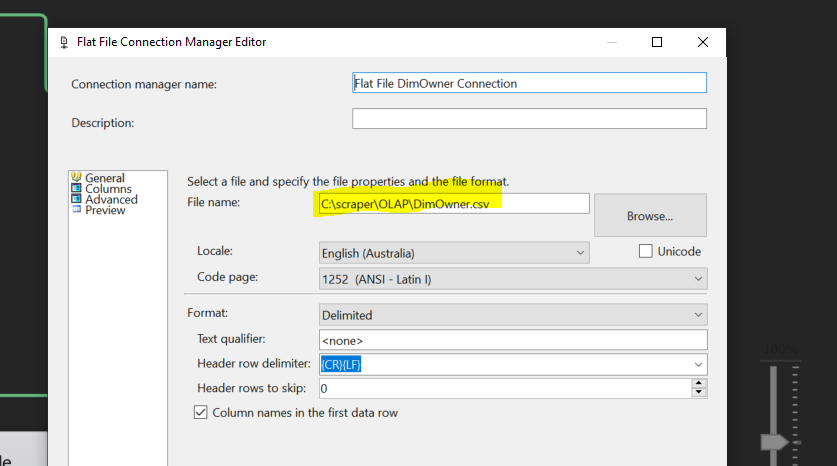
1. Open the solution located in \scraper\OLAP\etl\_ExperianCoffee



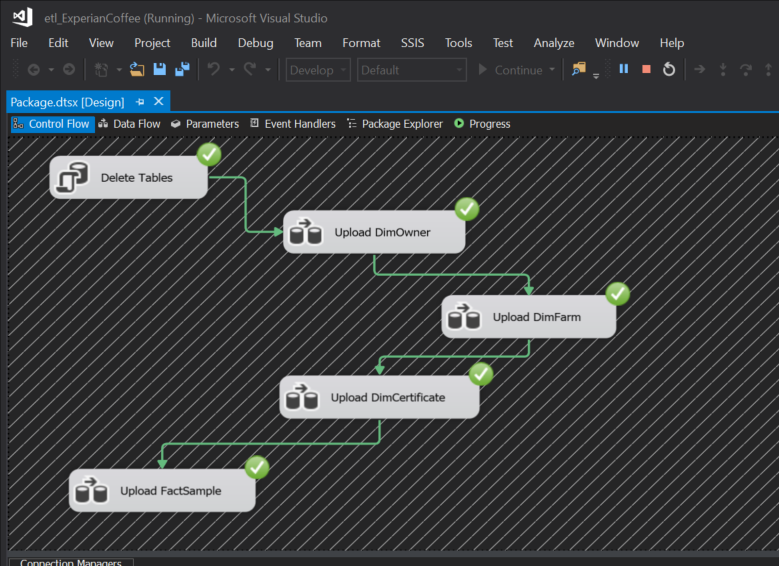
The OLEDB connection should point our local DB



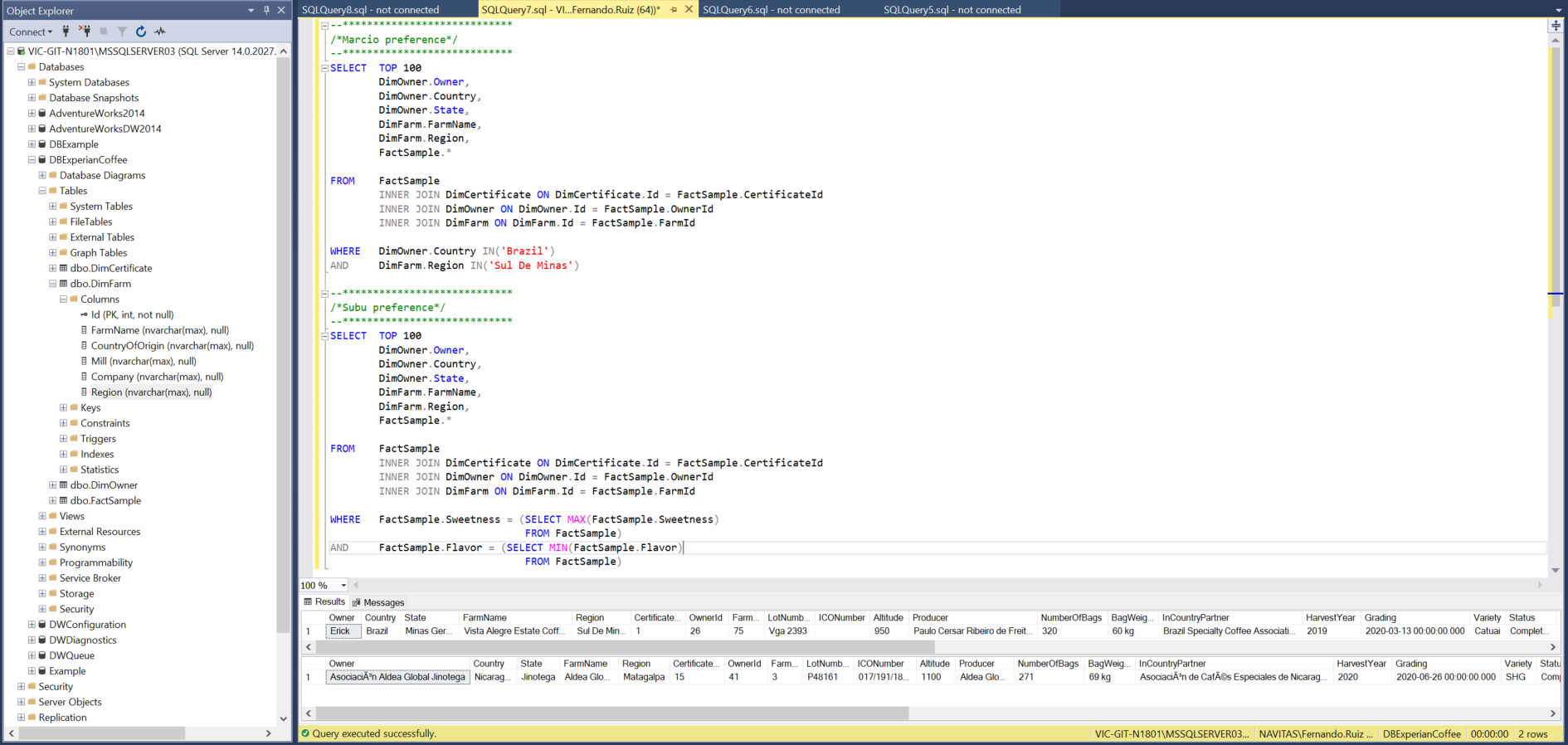
The FLATFILE connection should point the CVS files from \scraper\OLAP folder



Run the package

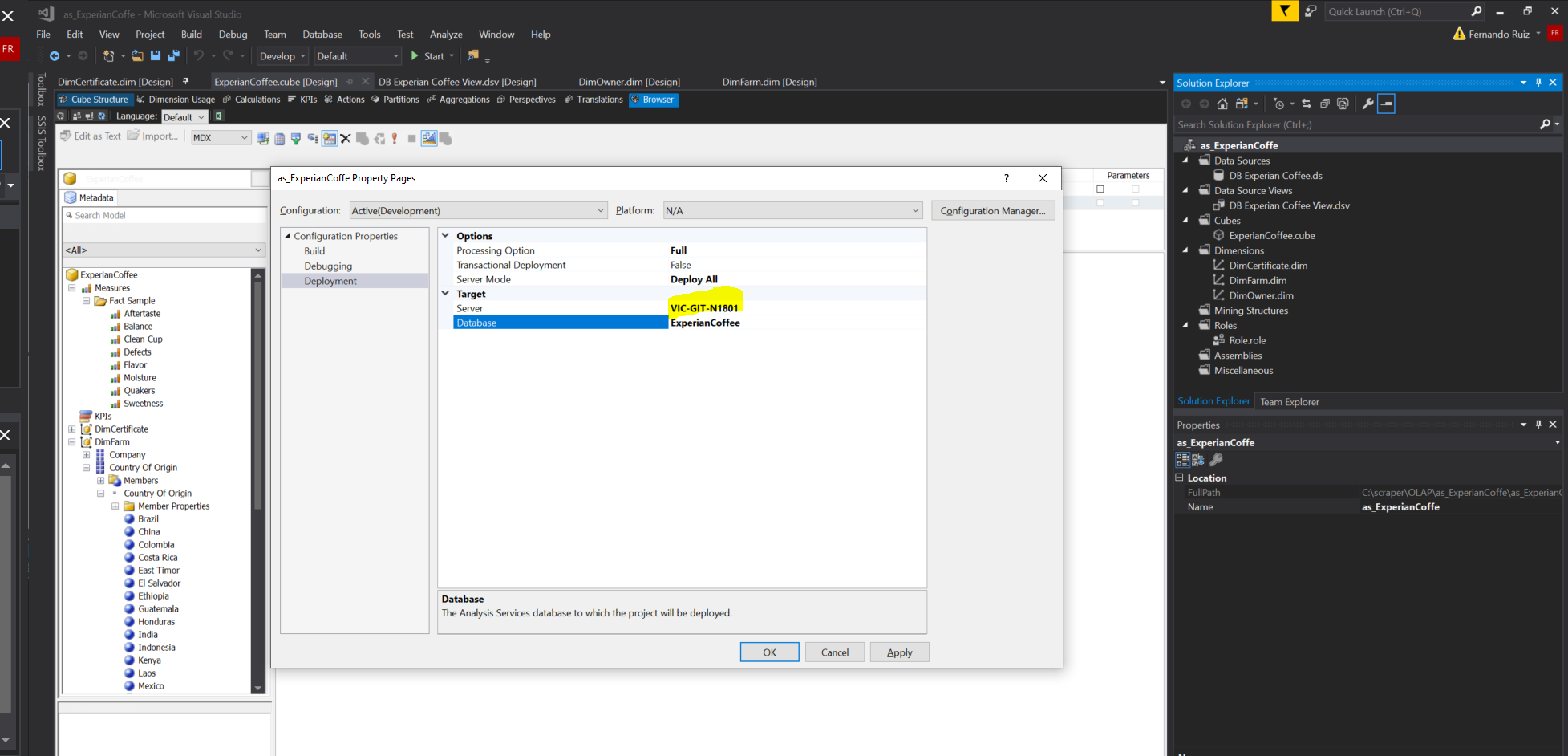


Test the data in SSMS running the “*Testing\_Data.sql*” script



**Analysis Services by SSAS**

1. Open the SSAS file *as\_ExperianCoffe* and add in property the target multidimensional mode server: VIC-GIT-N1801



The multidimensional service mode has been created in a

