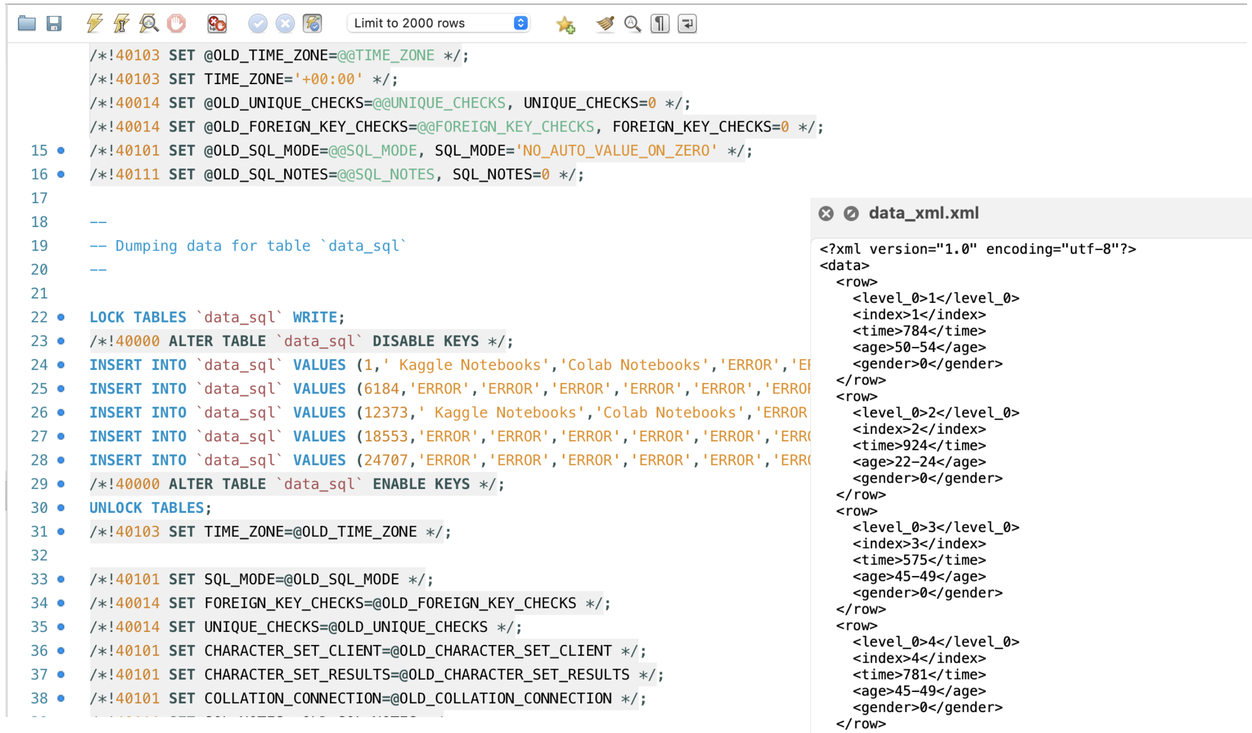
P R E S E N T S

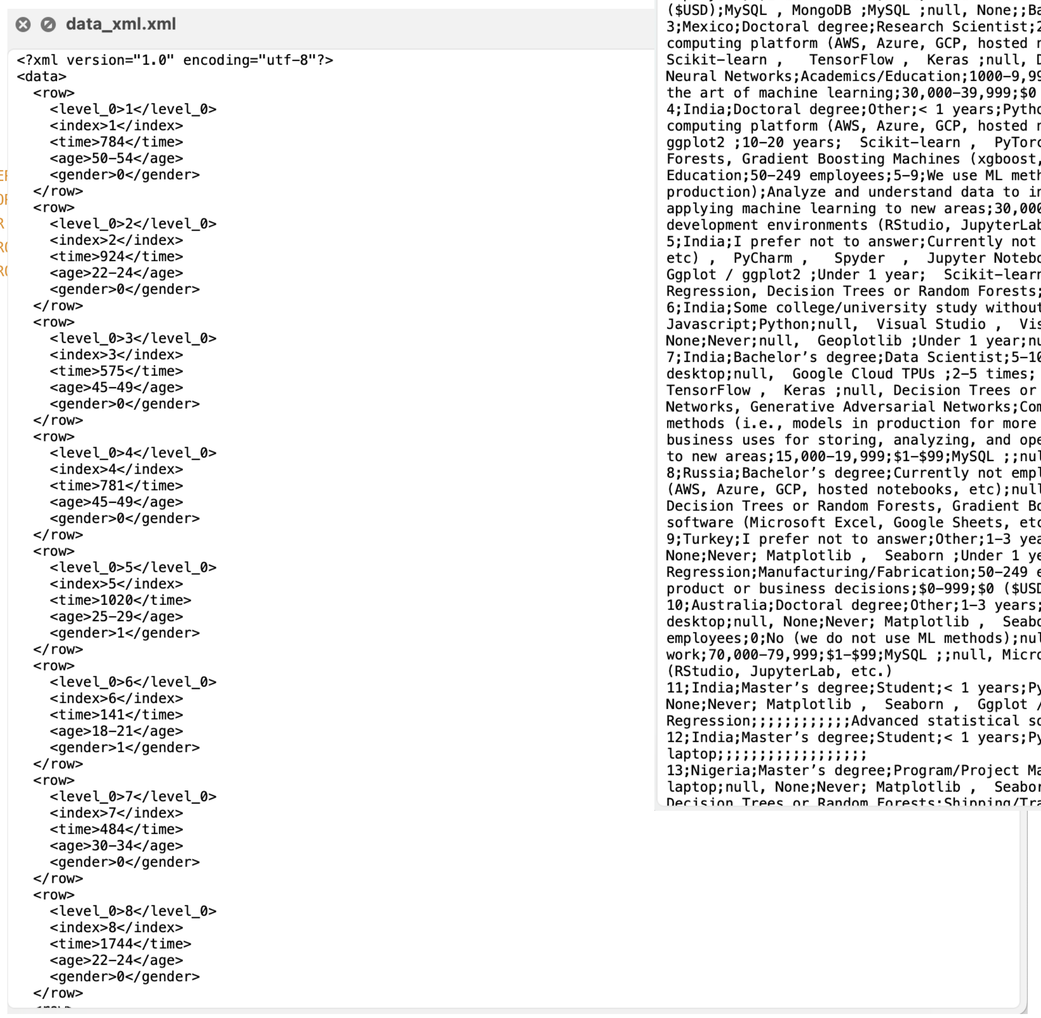
Data base creation project

Data base creation and automatization with .sql, .xml and .txt files

OBJETIVO: El objetivo de este proyecto era devolverle al cliente (P. O.)
una base de datos limpia y unificada a partir de tres archivos que él nos proporcionaba,
así como automatizar la limpieza e introducción de futuros datos.

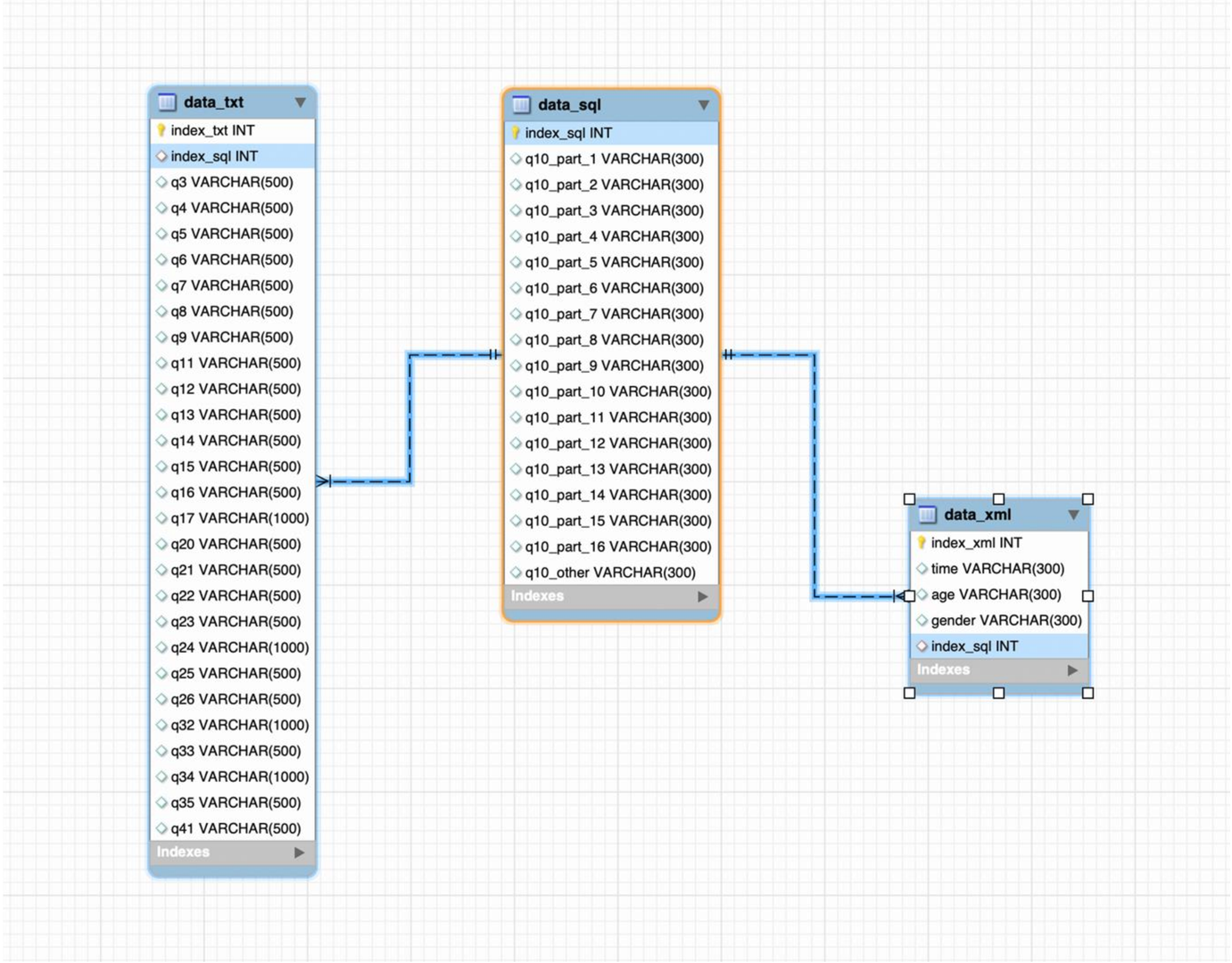


Archivo .SQL



Archivo .TXT

Archivo .XML



Características de la Base de Datos:

- Nombre de la Base de Datos: “project1”:
- Nombre de las Tablas: “data_sql", "data_xml", "data_txt".

Las especificaciones que deben cumplir las tablas son:

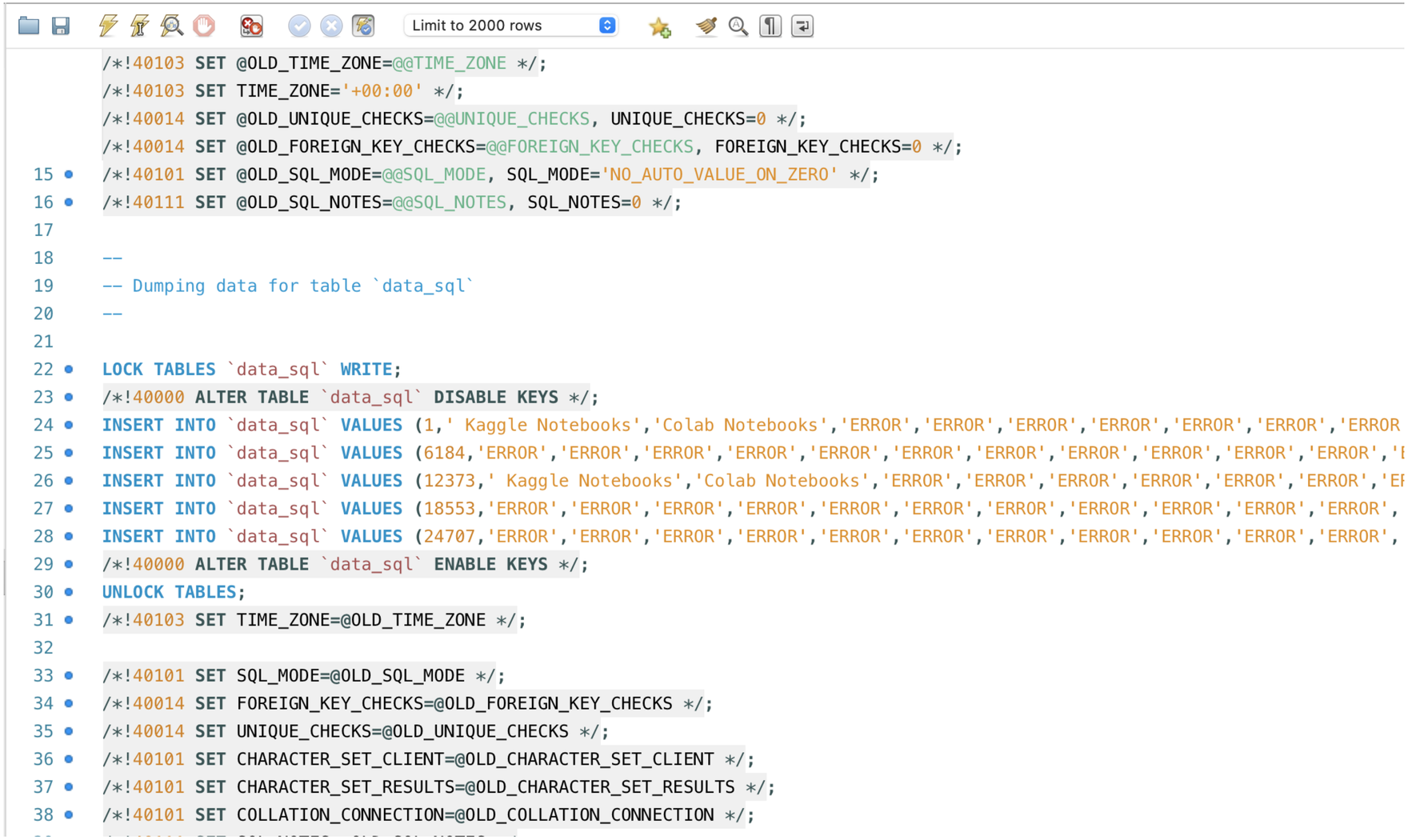
- data_sql: tabla madre. Clave primaria index_sql (INT).

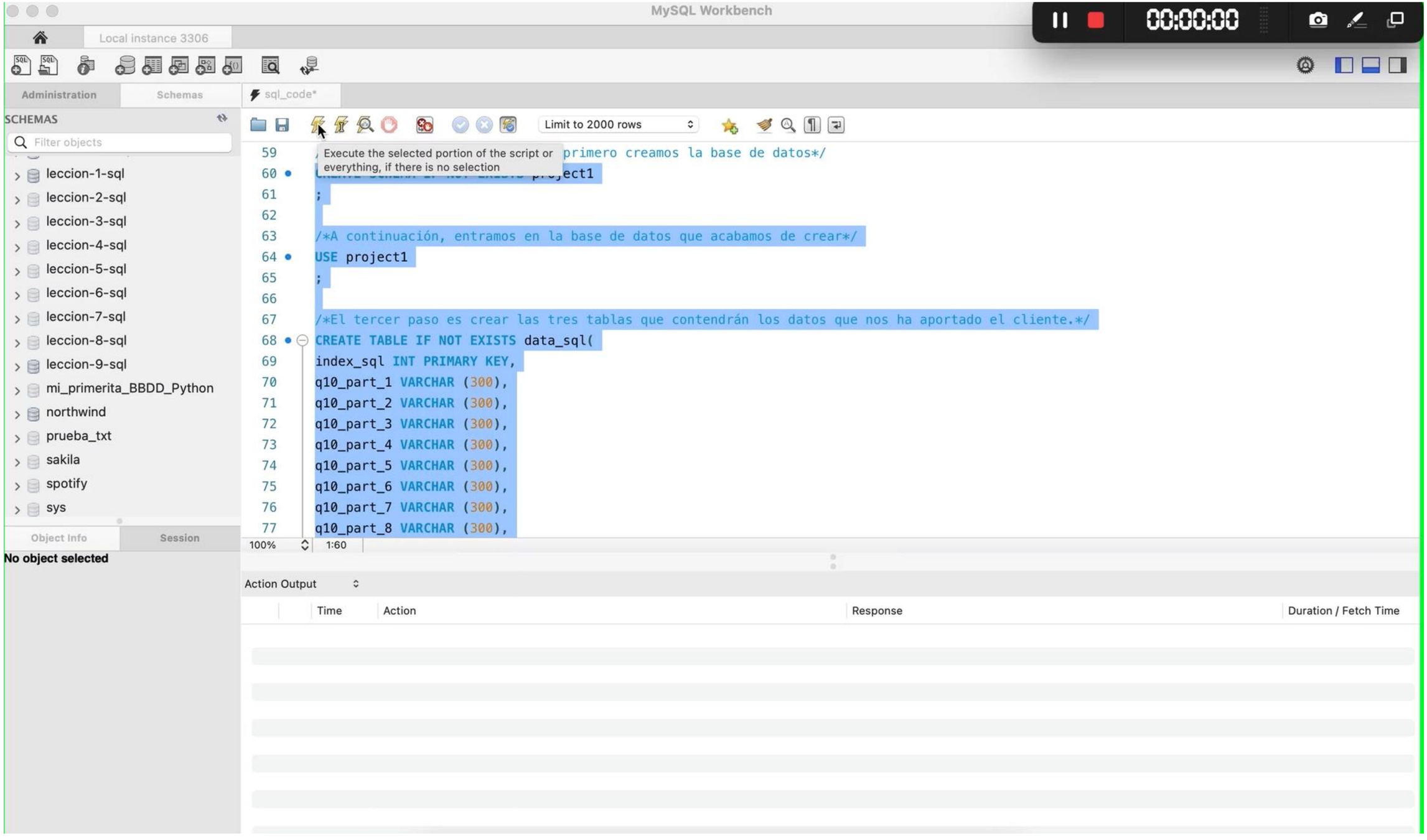
- data_xml: Su clave primaria es index_xml y su clave foránea es la columna index_sql, ambas de tipo numérico.

- data_txt: Su clave primaria es index_txt y su clave foránea es la columna index_sql, ambas de tipo numérico.

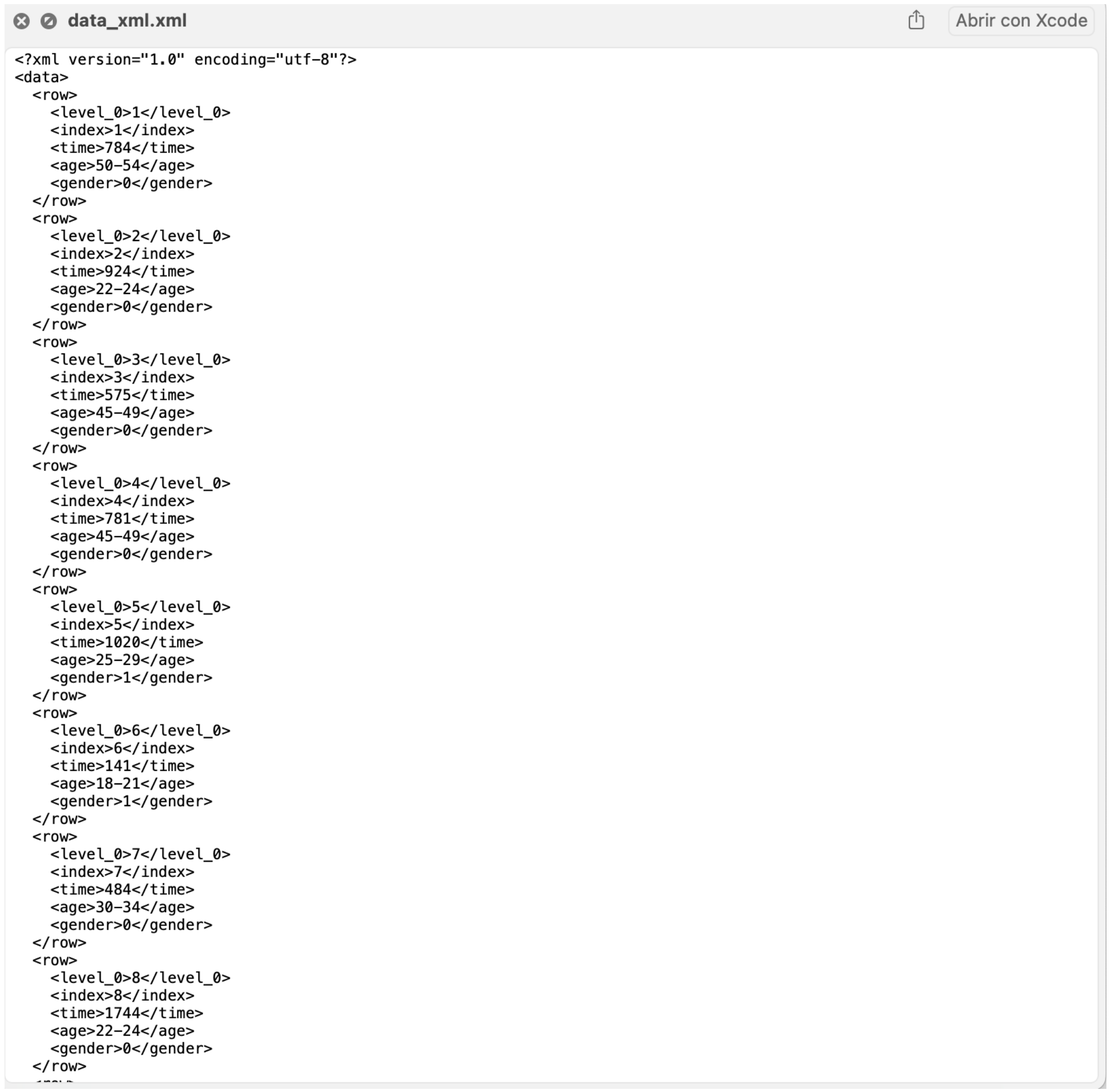
Las tres tablas se relacionarán mediante la columna index_sql.

Características del archivo .sql




MySQL Workbench
Local instance 3306
Administration
Schemas
sql_code*
SCHEMAS
Filter objects
leccion-1-sql
leccion-2-sql
leccion-3-sql
leccion-4-sql
leccion-5-sql
leccion-6-sql
leccion-7-sql
leccion-8-sql
leccion-9-sql
mi_primerita_BBDD_Python
northwind
prueba_txt
sakila
spotify
sys
Object Info
Session
No object selected
Limit to 2000 rows
Execute the selected portion of the script or everything, if there is no selection
primero creamos la base de datos*/
ject1
;
/*A continuación, entramos en la base de datos que acabamos de crear*/
USE project1
;
/*El tercer paso es crear las tres tablas que contendrán los datos que nos ha aportado el cliente.*/
CREATE TABLE IF NOT EXISTS data_sql(
index_sql INT PRIMARY KEY,
q10_part_1 VARCHAR (300),
q10_part_2 VARCHAR (300),
q10_part_3 VARCHAR (300),
q10_part_4 VARCHAR (300),
q10_part_5 VARCHAR (300),
q10_part_6 VARCHAR (300),
q10_part_7 VARCHAR (300),
q10_part_8 VARCHAR (300),
100%
1:60
Action Output
Time
Action
Response
Duration / Fetch Time

data_xml.xml — Abrir con Xcode



```
<?xml version="1.0" encoding="utf-8"?>
<data>
  <row>
    <level_0>1</level_0>
    <index>1</index>
    <time>784</time>
    <age>50-54</age>
    <gender>0</gender>
  </row>
  <row>
    <level_0>2</level_0>
    <index>2</index>
    <time>924</time>
    <age>22-24</age>
    <gender>0</gender>
  </row>
  <row>
    <level_0>3</level_0>
    <index>3</index>
    <time>575</time>
    <age>45-49</age>
    <gender>0</gender>
  </row>
  <row>
    <level_0>4</level_0>
    <index>4</index>
    <time>781</time>
    <age>45-49</age>
    <gender>0</gender>
  </row>
  <row>
    <level_0>5</level_0>
    <index>5</index>
    <time>1020</time>
    <age>25-29</age>
    <gender>1</gender>
  </row>
  <row>
    <level_0>6</level_0>
    <index>6</index>
    <time>141</time>
    <age>18-21</age>
    <gender>1</gender>
  </row>
  <row>
    <level_0>7</level_0>
    <index>7</index>
    <time>484</time>
    <age>30-34</age>
    <gender>0</gender>
  </row>
  <row>
    <level_0>8</level_0>
    <index>8</index>
    <time>1744</time>
    <age>22-24</age>
    <gender>0</gender>
  </row>
```

Características del archivo .xml

Objetivos en la depuración del fichero XML:

- Omitir columnas con contenido redundante.
- Recodificar la variable de género (‘Man’, ‘Woman’, ‘Non-binary’, etc.)

Novedades respecto a la depuración:

- Se ha omitido la columna con contenido redundante (level _ 0)
- Variable género: Numérica → String (conjunto de caracteres)
- Relacionada con la tabla ‘data_sql’ mediante la columna ‘index_sql’

ANTES

level_0
index_xml
time
age
gender

DESPUÉS

index_xml	INT AI PK
time	VARCHAR(300)
age	VARCHAR(300)
gender	VARCHAR (300)
index_sql	INT FK

DEF_limpieza_xml.ipynb ●

+ Código + Markdown | Ejecutar todo | Borrar resultados de todas las celdas | Reiniciar | Execute Group 1 | Execute Group 2 | Variables | Esquema ··· base (Python 3.9.13

```
import xml.etree.ElementTree as ET
import mysql.connector
from mysql.connector import errorcode

class XML:
    def __init__(self, archivoxml):
        self.tree = ET.parse(archivoxml)
        self.root = self.tree.getroot()
        self.archivoxml = archivoxml

    def limpieza_xml(self):
        def clean_gender(gender: str):
            if (gender == '0'):
                return 'Man'
            if (gender == '1'):
                return 'Woman'
            if (gender == '2'):
                return 'Nonbinary'
            if (gender == '3'):
                return 'Prefer not to say'
            if (gender == '4'):
                return 'Prefer to self-describe'

        for row in self.root.iter('row'):
            level_column = row.find('level_0')
            row.remove(level_column)
            gender_column = row.find('gender')
            gender_column.text = clean_gender(gender_column.text)


    def insertar_xml_sql(self):
        # Para evitar que nuestro código se pare utilizamos un try y except al hacer la conexión con
        # la base de datos de "project1".
        try:
            cnx = mysql.connector.connect(user='root', password='AlumnaAdalab',
                                          host='127.0.0.1',
                                          database='project1')
            print('Conexión con la base de datos realizada')
        except mysql.connector.Error as err:
```

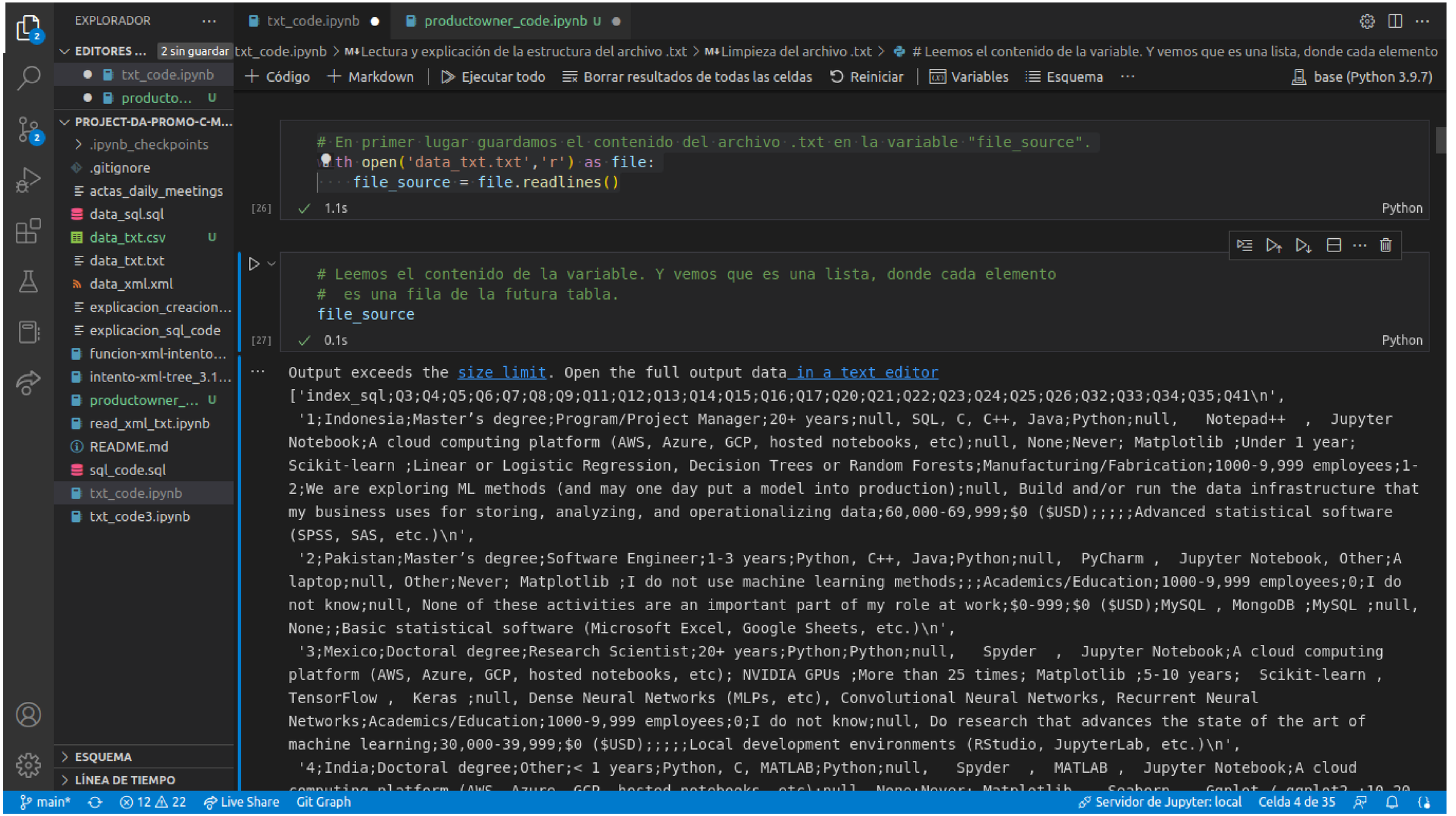
Características del archivo .txt

data_txt.txt — Abrir con TextEdit

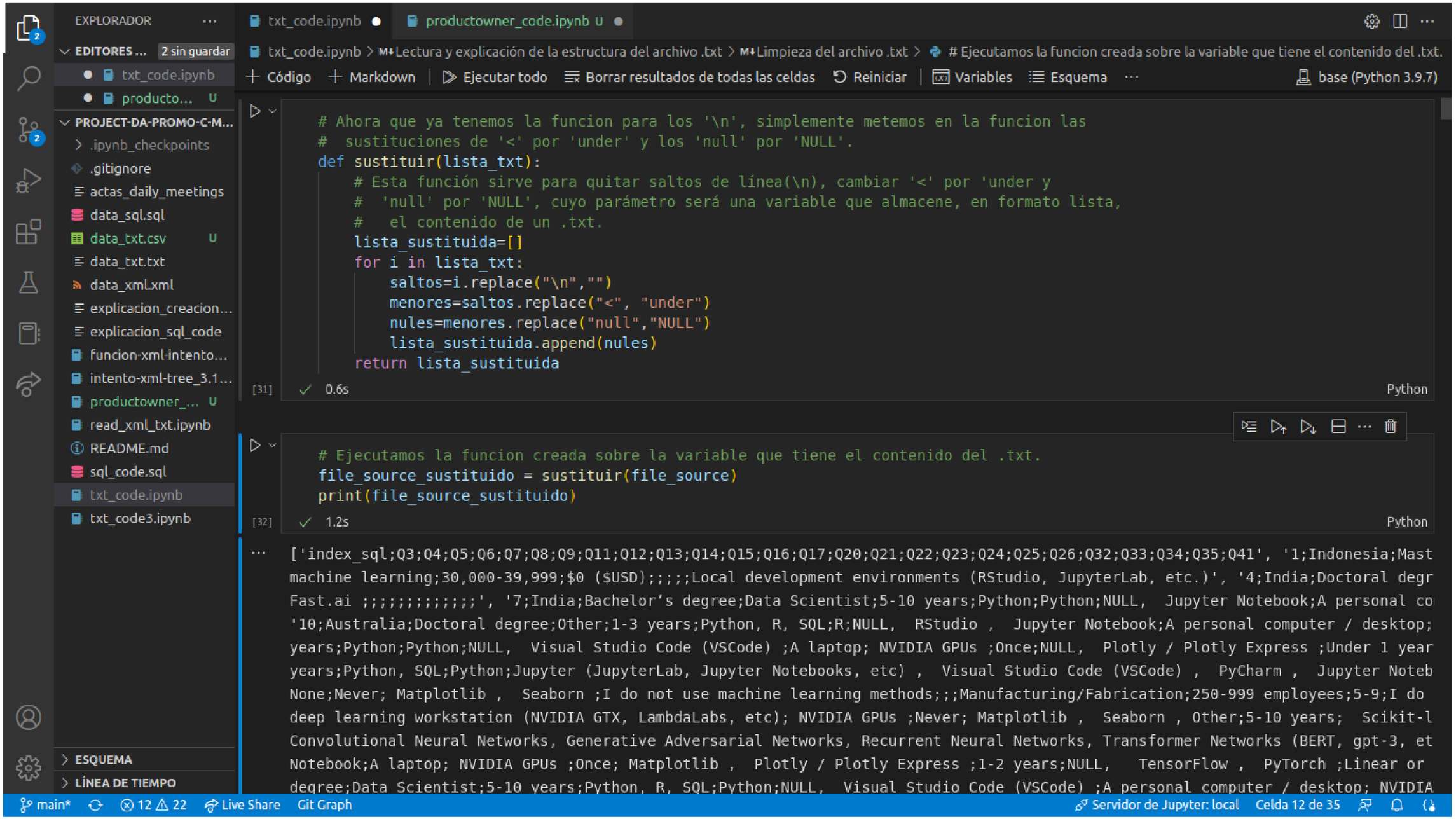
```
index_sql;Q3;Q4;Q5;Q6;Q7;Q8;Q9;Q11;Q12;Q13;Q14;Q15;Q16;Q17;Q20;Q21;Q22;Q23;Q24;Q25;Q26;Q32;Q33;Q34;Q35;Q41
1;Indonesia;Master’s degree;Program/Project Manager;20+ years;null, SQL, C, C++, Java;Python;null,   Notepad++  ,
Jupyter Notebook;A cloud computing platform (AWS, Azure, GCP, hosted notebooks, etc);null, None;Never;
Matplotlib ;Under 1 year;  Scikit-learn ;Linear or Logistic Regression, Decision Trees or Random Forests;Manufacturing/
Fabrication;1000-9,999 employees;1-2;We are exploring ML methods (and may one day put a model into production);null,
Build and/or run the data infrastructure that my business uses for storing, analyzing, and operationalizing
data;60,000-69,999;$0 ($USD);;;;;Advanced statistical software (SPSS, SAS, etc.)
2;Pakistan;Master’s degree;Software Engineer;1-3 years;Python, C++, Java;Python;null,  PyCharm ,  Jupyter Notebook,
Other;A laptop;null, Other;Never; Matplotlib ;I do not use machine learning methods;;;Academics/Education;1000-9,999
employees;0;I do not know;null, None of these activities are an important part of my role at work;$0-999;$0
($USD);MySQL , MongoDB ;MySQL ;null, None;;Basic statistical software (Microsoft Excel, Google Sheets, etc.)
3;Mexico;Doctoral degree;Research Scientist;20+ years;Python;Python;null,   Spyder  ,  Jupyter Notebook;A cloud
computing platform (AWS, Azure, GCP, hosted notebooks, etc); NVIDIA GPUs ;More than 25 times; Matplotlib ;5-10 years;
Scikit-learn ,   TensorFlow ,  Keras ;null, Dense Neural Networks (MLPs, etc), Convolutional Neural Networks, Recurrent
Neural Networks;Academics/Education;1000-9,999 employees;0;I do not know;null, Do research that advances the state of
the art of machine learning;30,000-39,999;$0 ($USD);;;;;Local development environments (RStudio, JupyterLab, etc.)
4;India;Doctoral degree;Other;< 1 years;Python, C, MATLAB;Python;null,   Spyder  ,  MATLAB ,  Jupyter Notebook;A cloud
computing platform (AWS, Azure, GCP, hosted notebooks, etc);null, None;Never; Matplotlib ,  Seaborn ,  Ggplot /
ggplot2 ;10-20 years;  Scikit-learn ,  PyTorch ,  LightGBM ;Linear or Logistic Regression, Decision Trees or Random
Forests, Gradient Boosting Machines (xgboost, lightgbm, etc), Bayesian Approaches, Evolutionary Approaches;Academics/
Education;50-249 employees;5-9;We use ML methods for generating insights (but do not put working models into
```

- Una lista de strings.
- Presenta un indice .sql para cada fila.
- Limpieza:
  - Eliminar espacios innecesarios.
  - Cambiar “<“ por “under”.
  - Cambiar “null” por “NULL”.
  - Eliminar “\n “ al final de cada elemento.

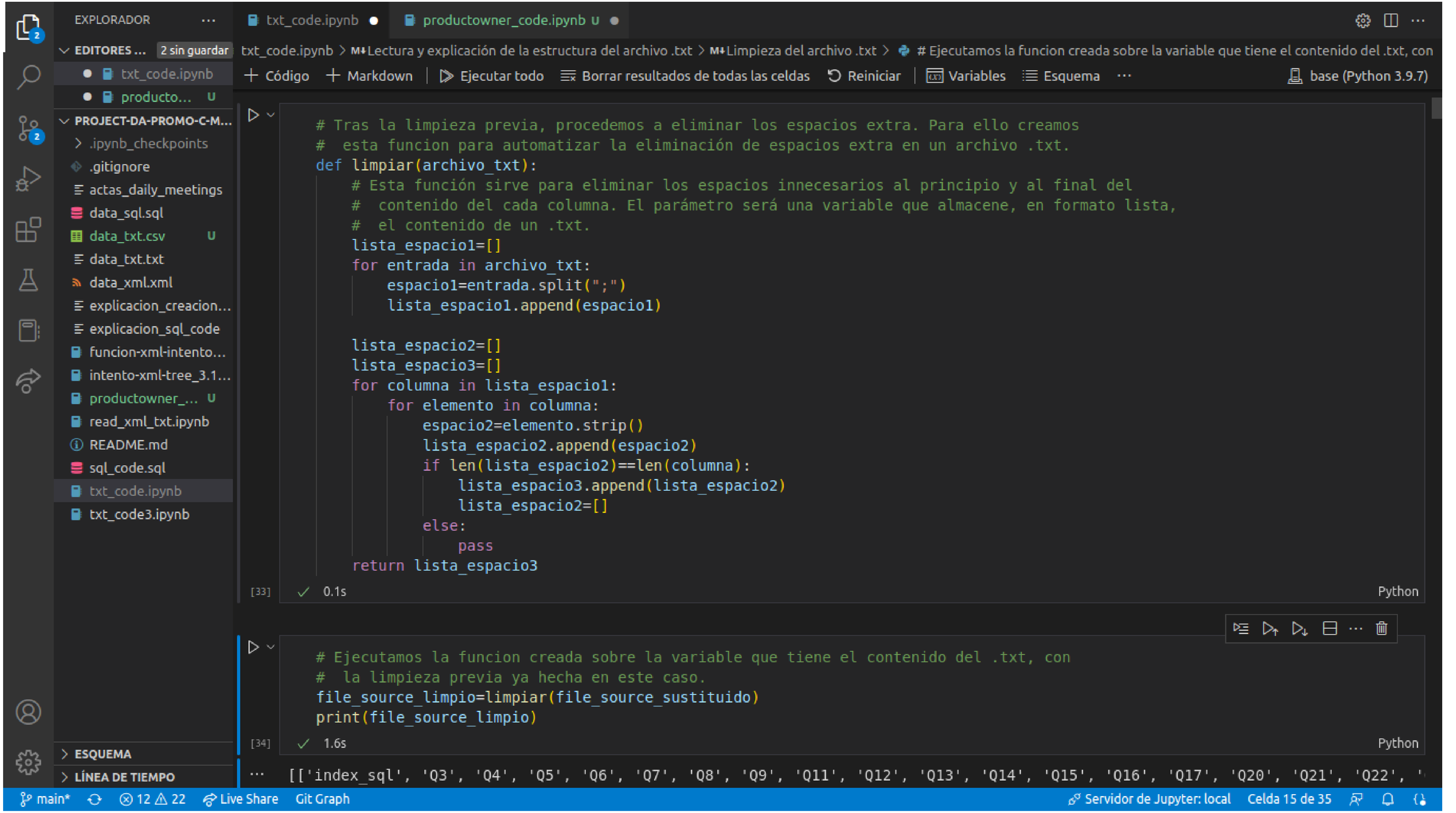
¿Que apariencia tiene el archivo que nos llega?



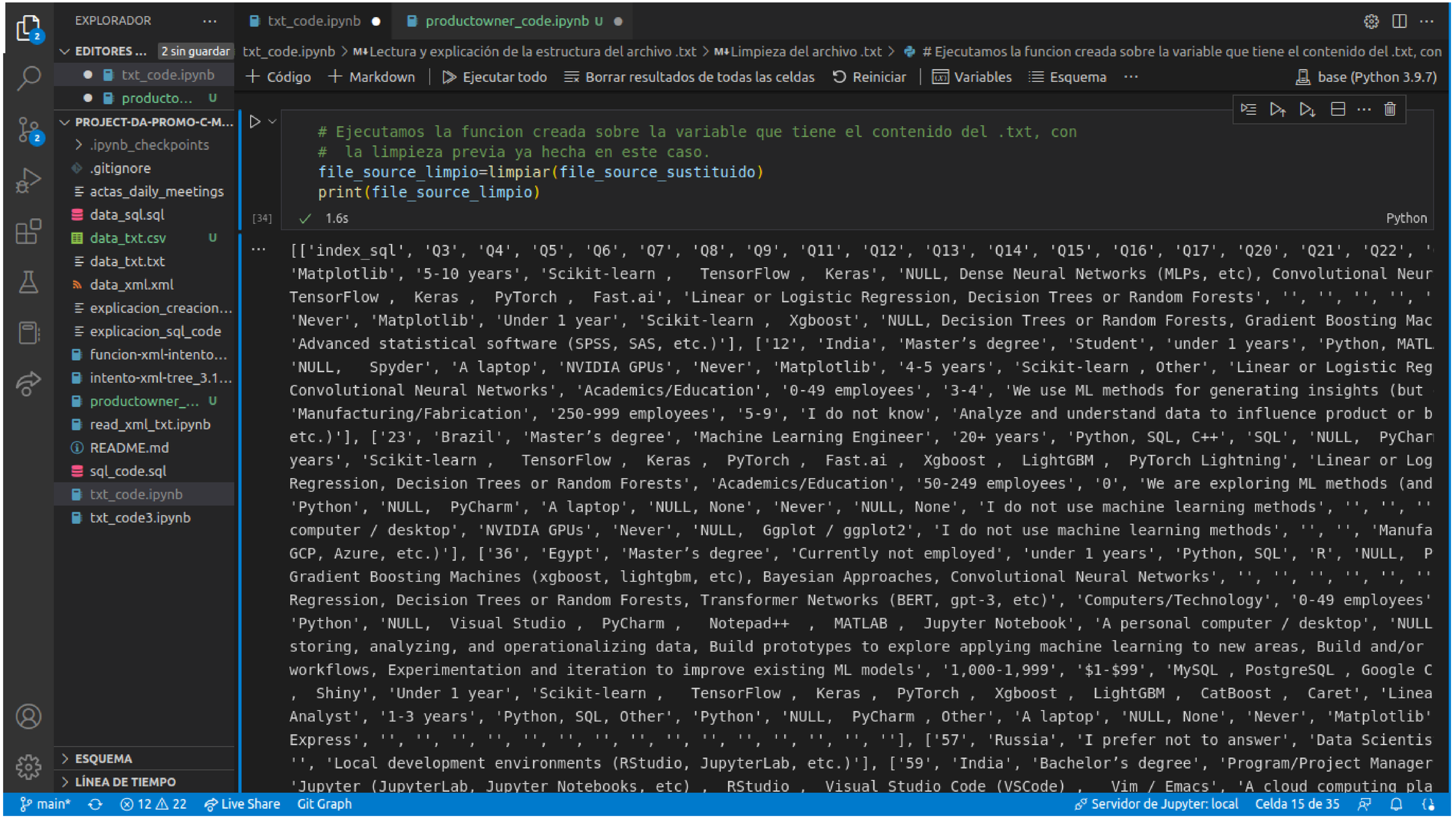
Desarrollo de funciones.



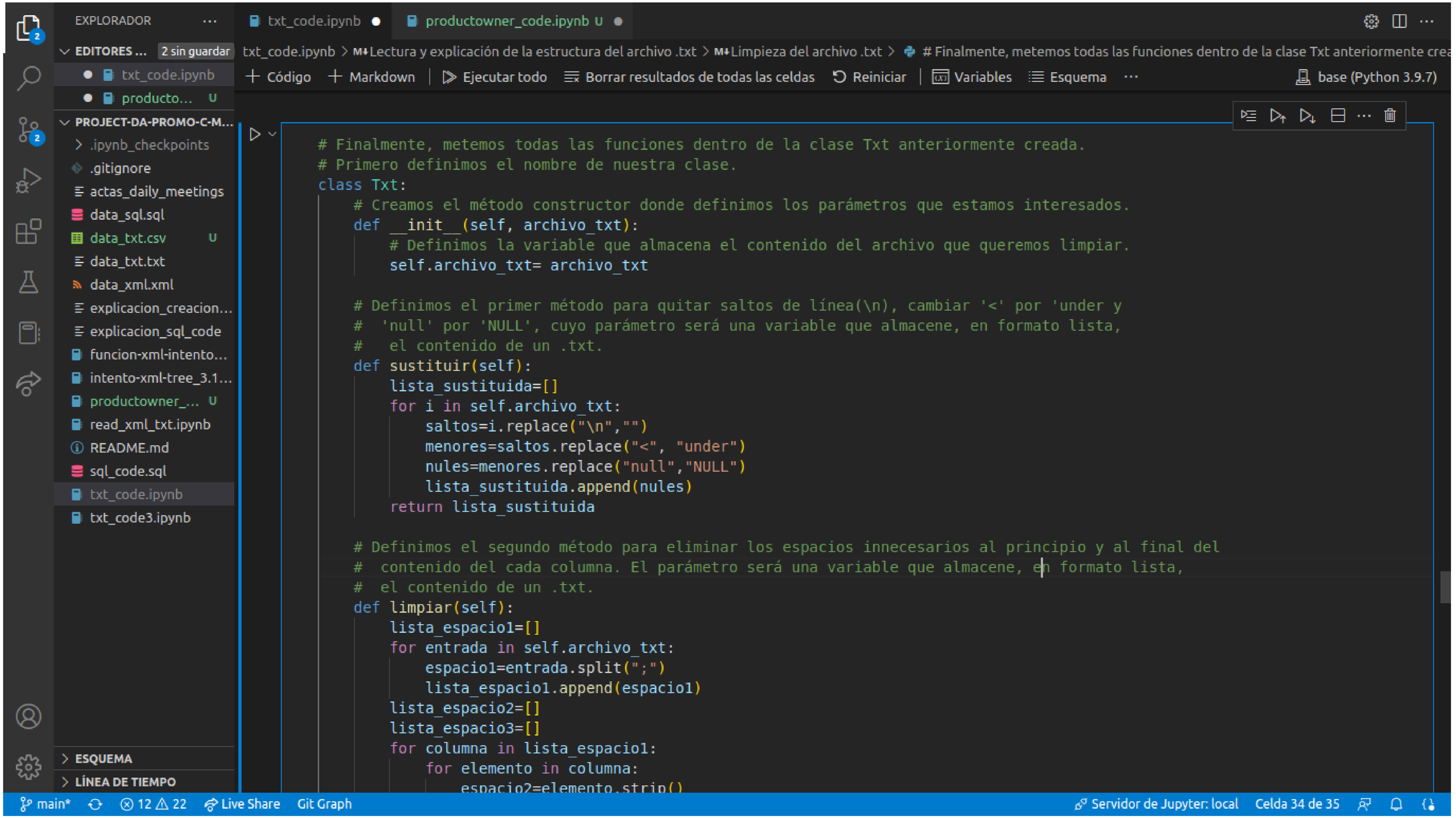
Desarrollo de funciones.



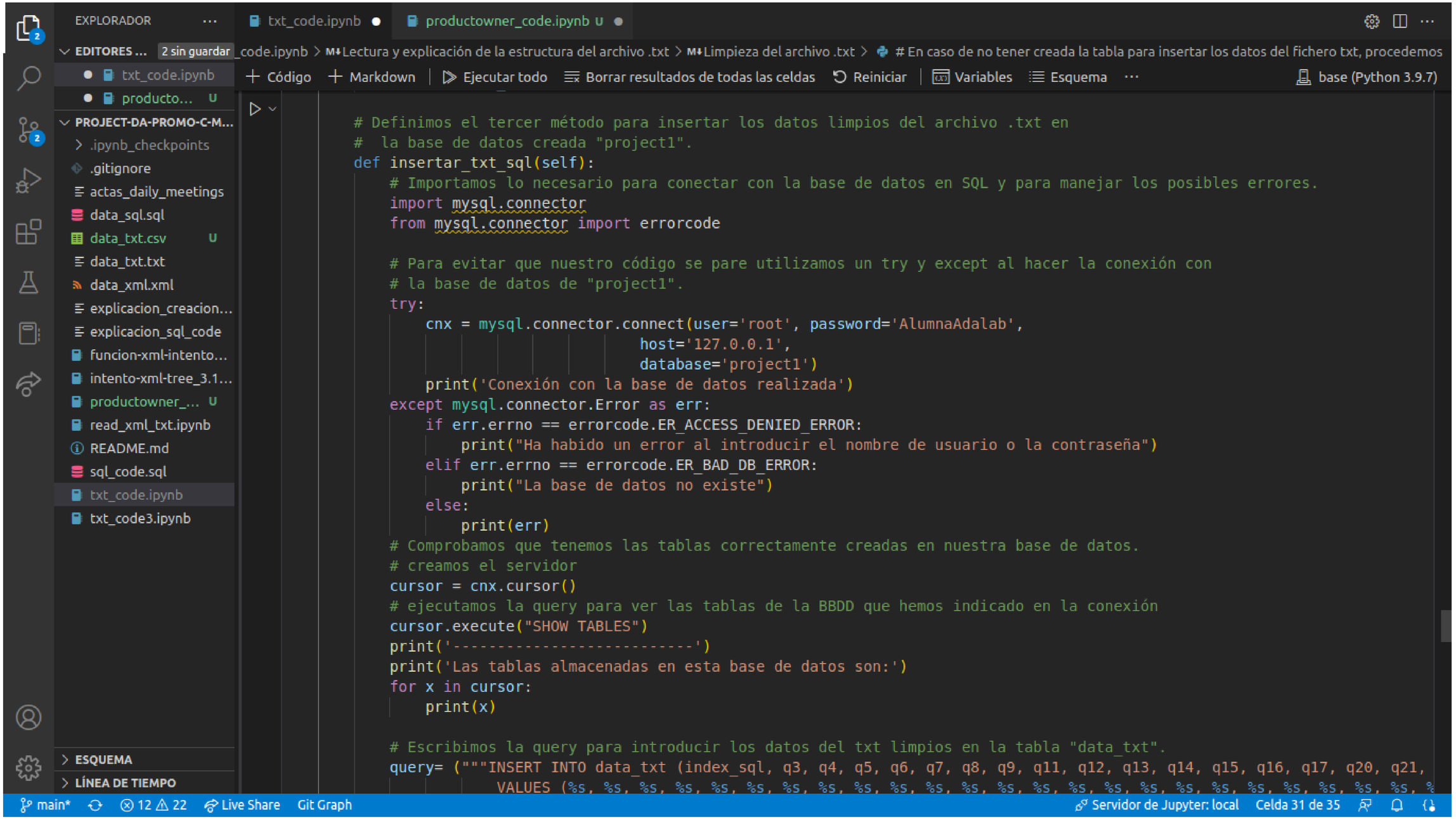
Resultado preliminar.



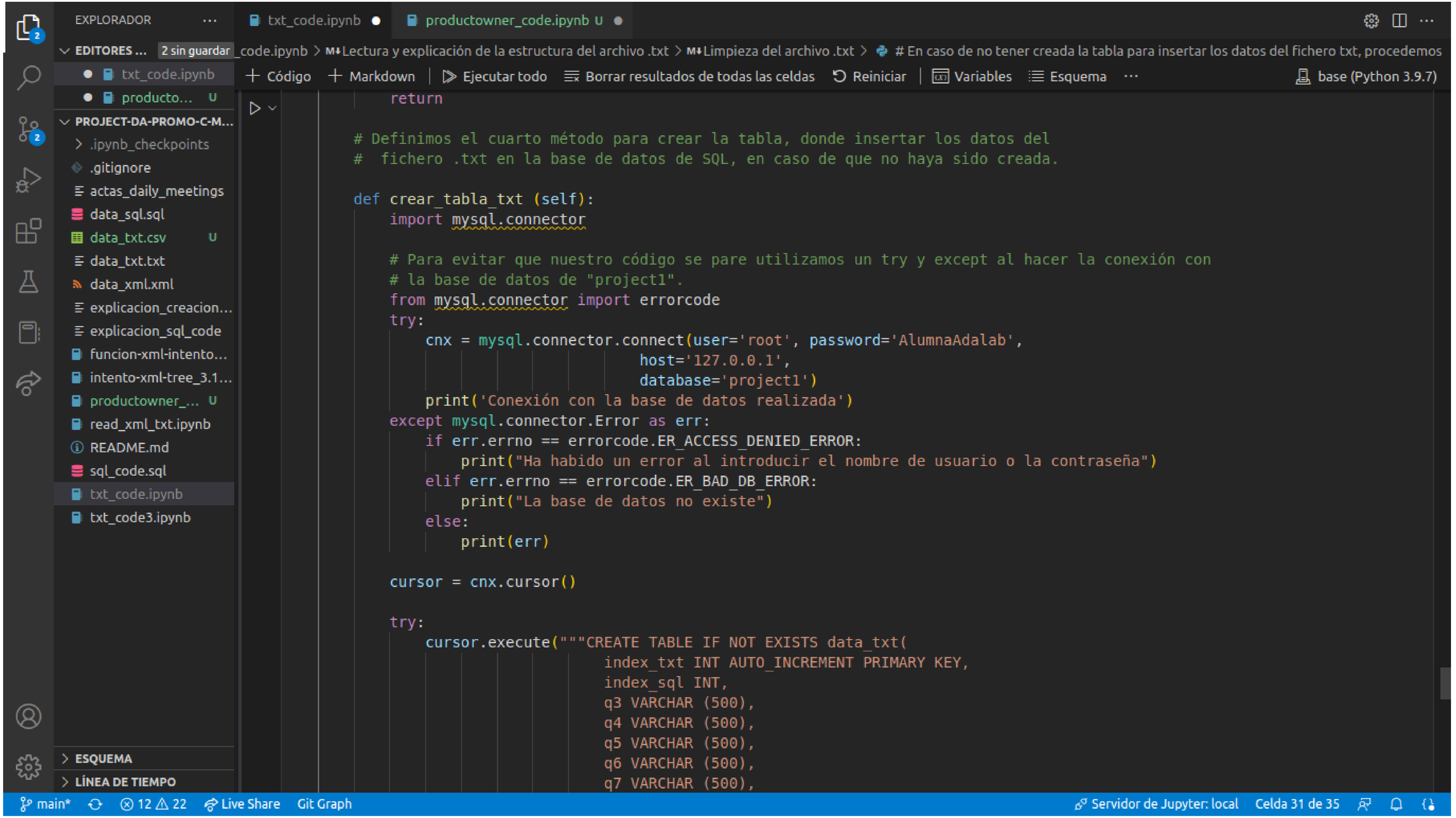
Desarrollo de clase.



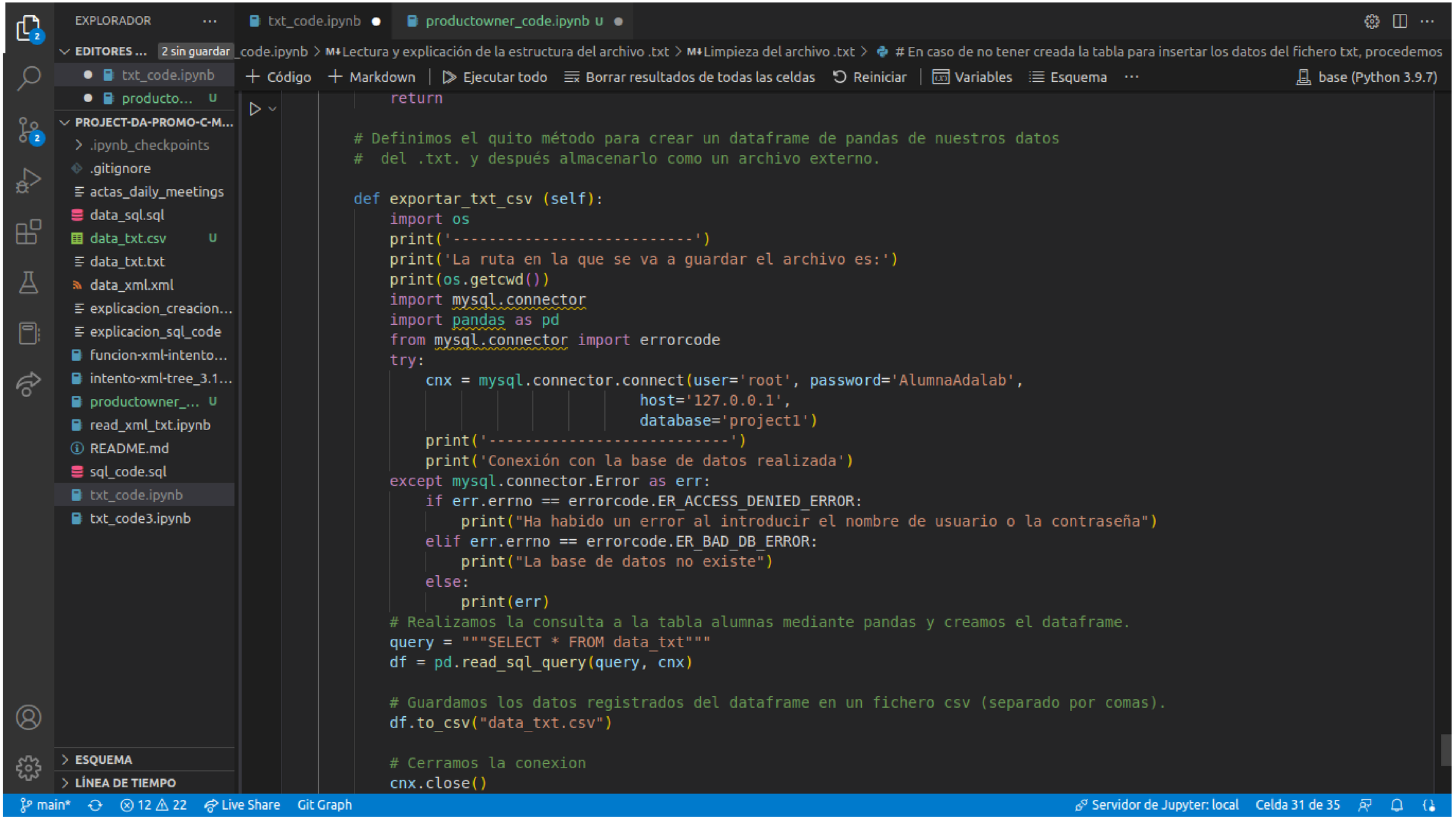
Desarrollo de clase.



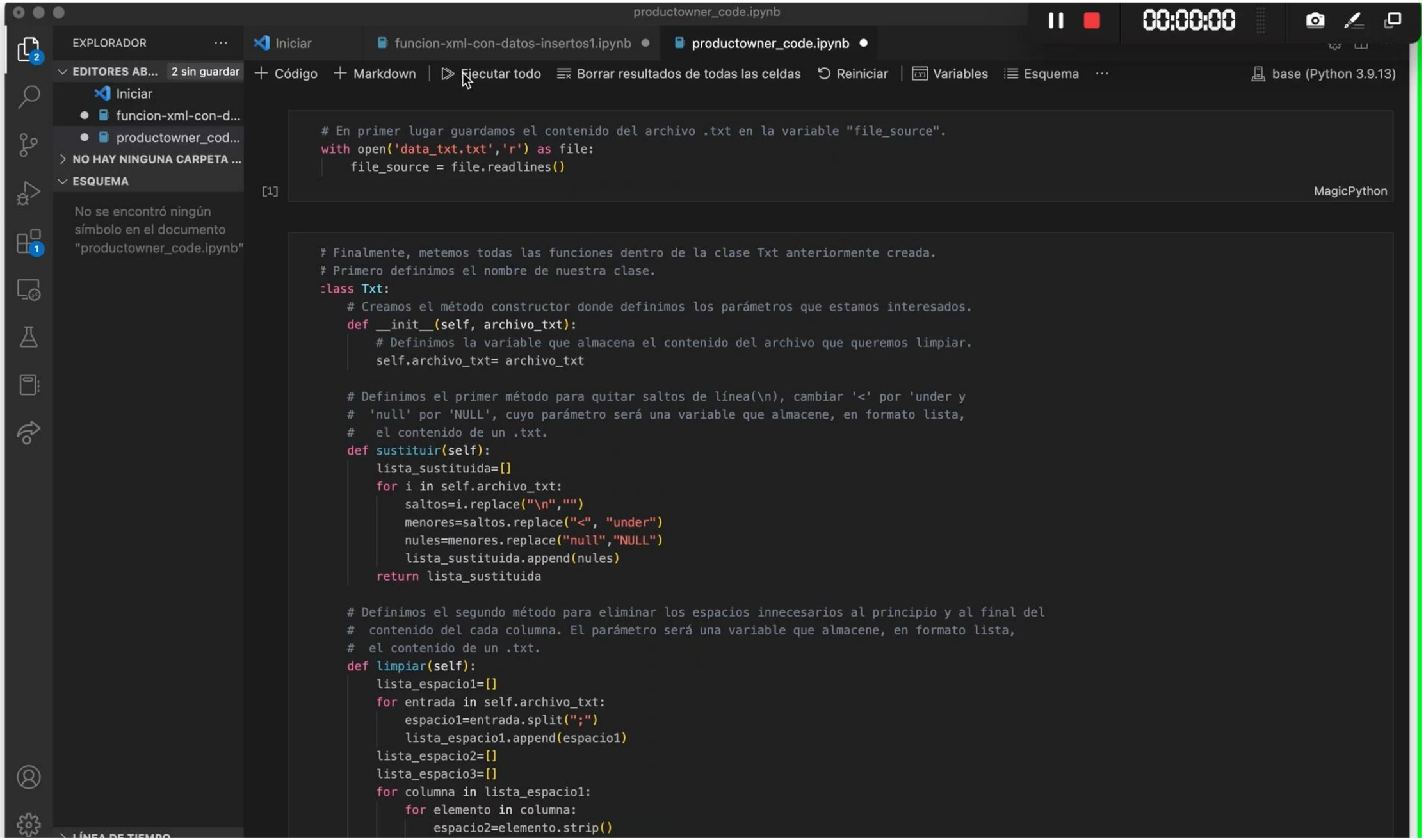
Desarrollo de clase.



Desarrollo de clase.



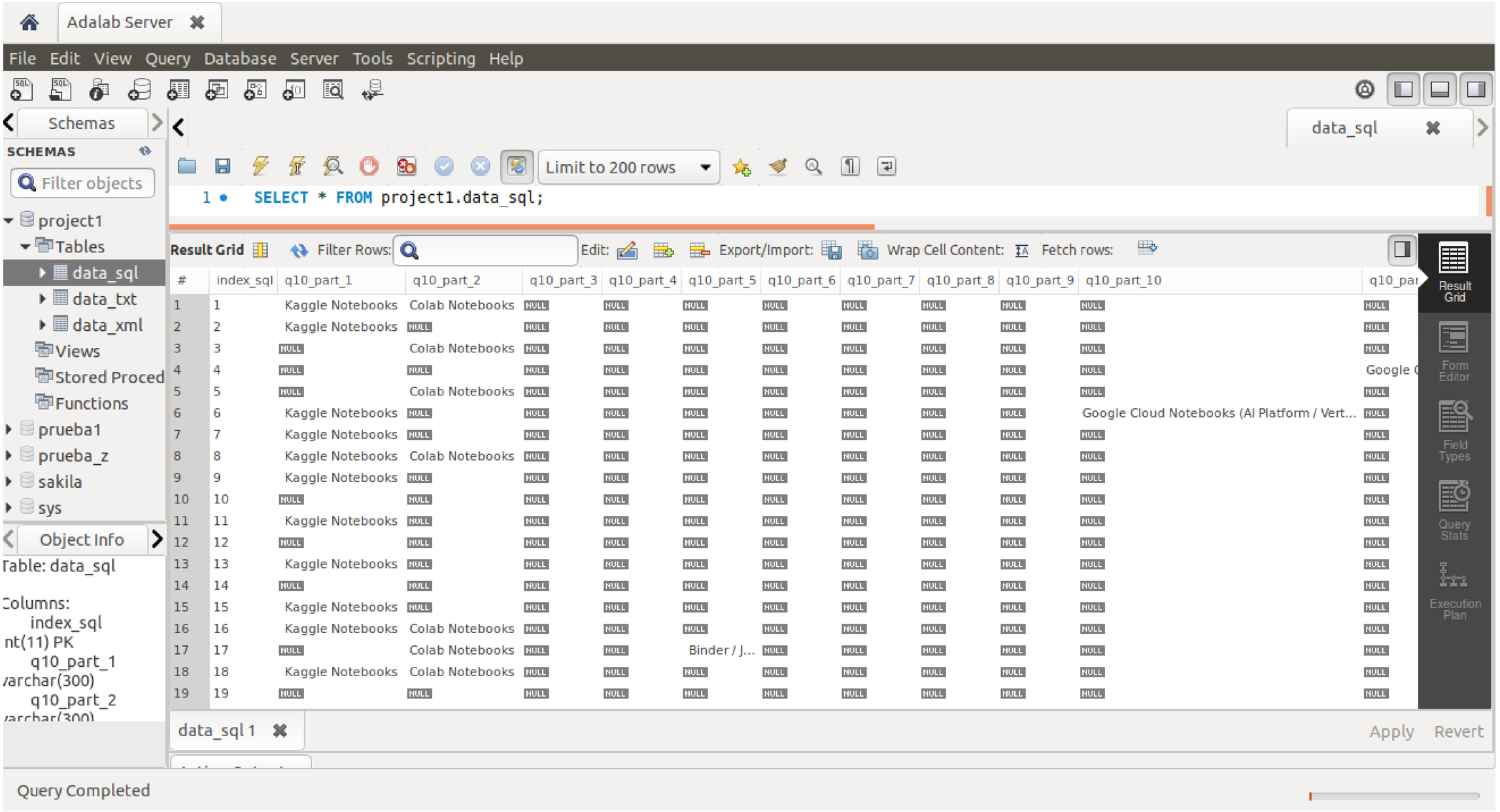
Código en ejecución.



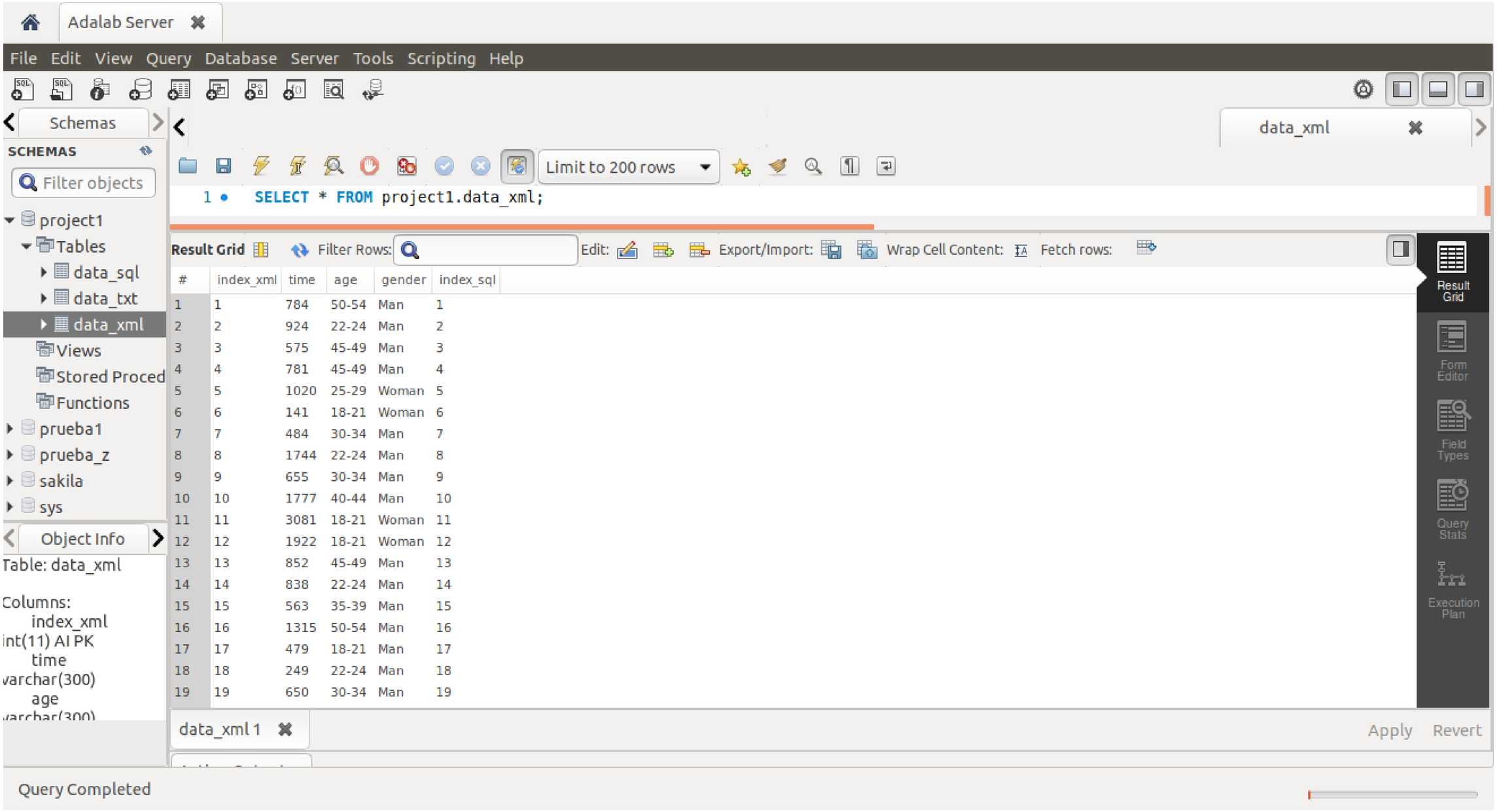
Datos TXT en CSV

	index_txt	index_sql	q3	q4	q5	q6	q7	q8
0	1	1	Indonesia	Master’s degree	Program/Project Manager	20+ years	NULL, SQL, C, C++, Java	Python
1	2	2	Pakistan	Master’s degree	Software Engineer	1-3 years	Python, C++, Java	Python
2	3	3	Mexico	Doctoral degree	Research Scientist	20+ years	Python	Python
3	4	4	India	Doctoral degree	Other	under 1 years	Python, C, MATLAB	Python
4	5	5	India	I prefer not to answer	Currently not employed	under 1 years	Python	Python
5	6	6	India	Some college/university study without earning a bachelor’s degree	Student	1-3 years	NULL, C++, Java, Javascript	Python
6	7	7	India	Bachelor’s degree	Data Scientist	5-10 years	Python	Python
7	8	8	Russia	Bachelor’s degree	Currently not employed	3-5 years	Python, SQL	Python
8	9	9	Turkey	I prefer not to answer	Other	1-3 years	Python, SQL	SQL
9	10	10	Australia	Doctoral degree	Other	1-3 years	Python, R, SQL	R
10	11	11	India	Master’s degree	Student	under 1 years	Python, R, C++	R
11	12	12	India	Master’s degree	Student	under 1 years	Python, MATLAB	Python
12	13	13	Nigeria	Master’s degree	Program/Project Manager	5-10 years	Python, SQL	Python
13	14	14	Nigeria	Bachelor’s degree	Other	under 1 years	Python	Python
14	15	15	Greece	Doctoral degree	Research Scientist	10-20 years	Python, C, C++, MATLAB	Python
15	16	16	Belgium	Bachelor’s degree	Data Analyst	20+ years	Python, SQL	Python
16	17	17	Pakistan	Bachelor’s degree	Data Scientist	1-3 years	Python, SQL	Python
17	18	18	Japan	Master’s degree	Software Engineer	3-5 years	Python, SQL, C, Java, Javascript	Python
18	19	19	Egypt	Bachelor’s degree	Other	under 1 years	NULL, None	R
19	20	20	Singapore	Bachelor’s degree	Other	under 1 years	Python	Python
20	21	21	Turkey	Bachelor’s degree	Data Scientist	3-5 years	Python, R, SQL, C++	R
21	22	22	Indonesia	Master’s degree	Student	1-3 years	NULL, R	R
22	23	23	Brazil	Master’s degree	Machine Learning Engineer	20+ years	Python, SQL, C++	SQL
23	24	24	India	Bachelor’s degree	Student	1-3 years	Python, R, SQL, C, C++, MATLAB	Python
24	25	25	Poland	Master’s degree	Machine Learning Engineer	3-5 years	Python, C++	Python
25	26	26	Brazil	Doctoral degree	Research Scientist	under 1 years	Python, R	Python

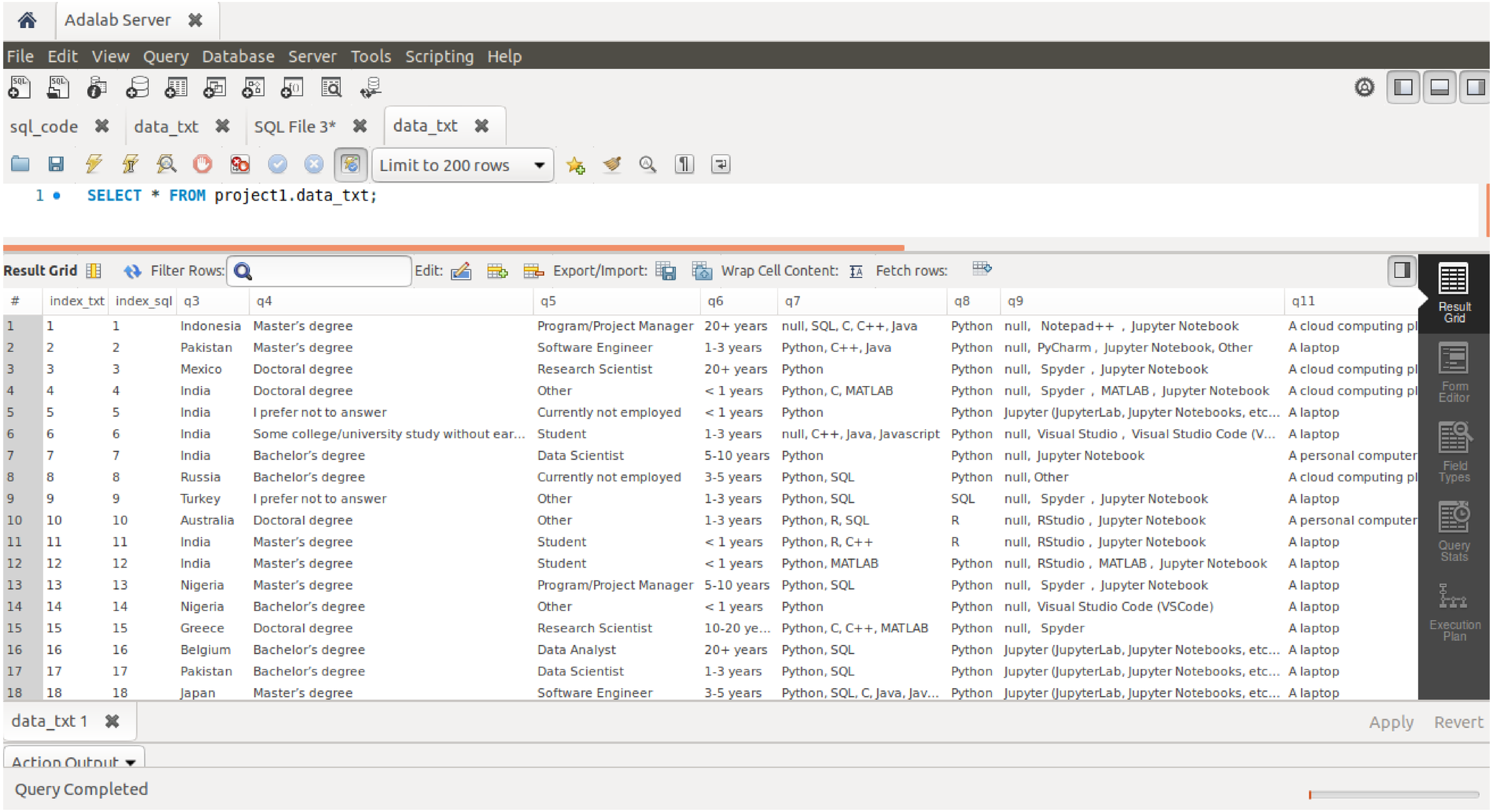
Archivo SQL en SQL



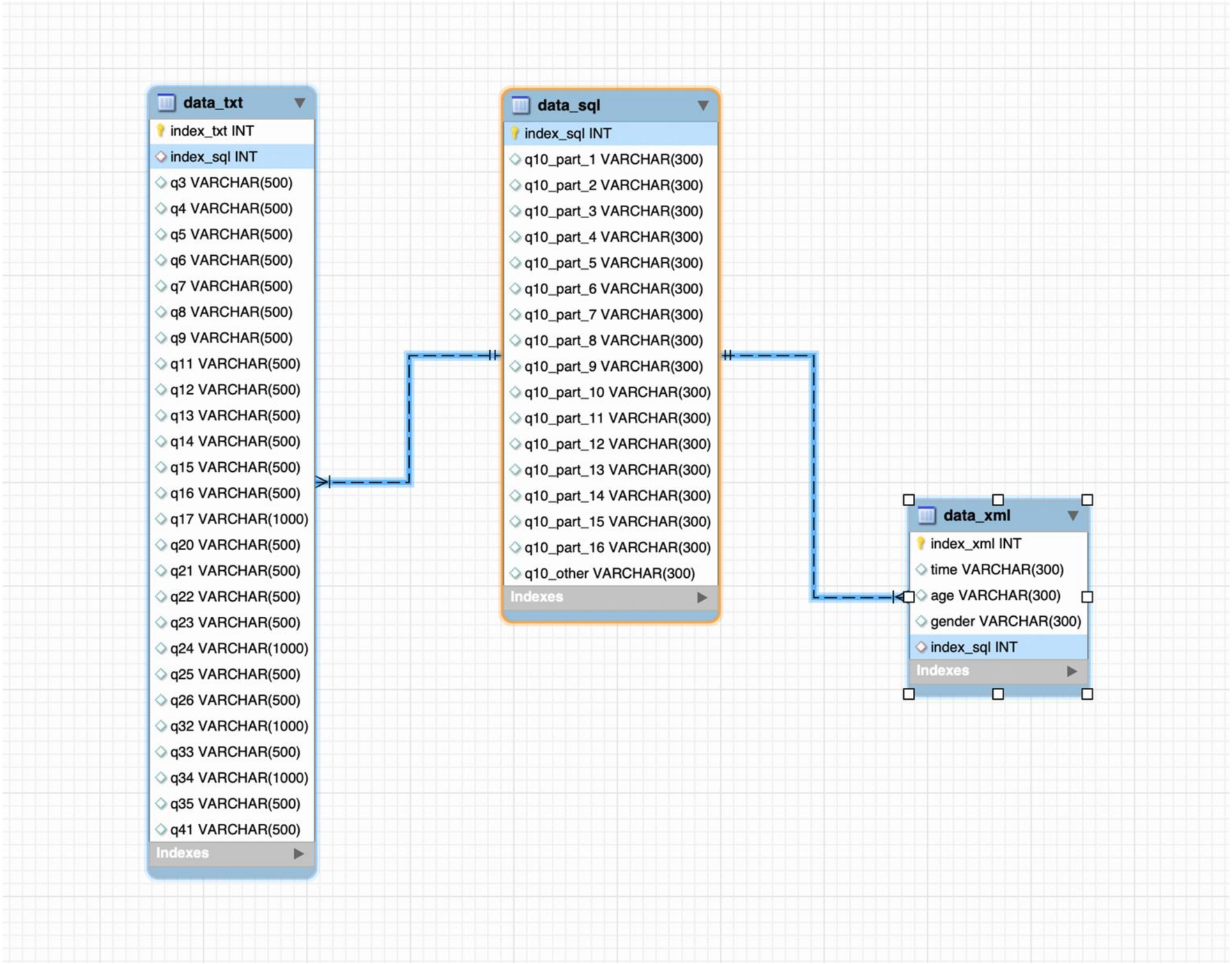
Archivo XML en SQL



Archivo TXT en SQL



#	index_txt	index_sql	q3	q4	q5	q6	q7	q8	q9	q11
1	1	1	Indonesia	Master's degree	Program/Project Manager	20+ years	null, SQL, C, C++, Java	Python	null, Notepad++ , Jupyter Notebook	A cloud computing pl
2	2	2	Pakistan	Master's degree	Software Engineer	1-3 years	Python, C++, Java	Python	null, PyCharm , Jupyter Notebook, Other	A laptop
3	3	3	Mexico	Doctoral degree	Research Scientist	20+ years	Python	Python	null, Spyder , Jupyter Notebook	A cloud computing pl
4	4	4	India	Doctoral degree	Other	< 1 years	Python, C, MATLAB	Python	null, Spyder , MATLAB , Jupyter Notebook	A cloud computing pl
5	5	5	India	I prefer not to answer	Currently not employed	< 1 years	Python	Python	Jupyter (JupyterLab, Jupyter Notebooks, etc...	A laptop
6	6	6	India	Some college/university study without ear...	Student	1-3 years	null, C++, Java, Javascript	Python	null, Visual Studio , Visual Studio Code (V...	A laptop
7	7	7	India	Bachelor's degree	Data Scientist	5-10 years	Python	Python	null, Jupyter Notebook	A personal computer
8	8	8	Russia	Bachelor's degree	Currently not employed	3-5 years	Python, SQL	Python	null, Other	A cloud computing pl
9	9	9	Turkey	I prefer not to answer	Other	1-3 years	Python, SQL	SQL	null, Spyder , Jupyter Notebook	A laptop
10	10	10	Australia	Doctoral degree	Other	1-3 years	Python, R, SQL	R	null, RStudio , Jupyter Notebook	A personal computer
11	11	11	India	Master's degree	Student	< 1 years	Python, R, C++	R	null, RStudio , Jupyter Notebook	A laptop
12	12	12	India	Master's degree	Student	< 1 years	Python, MATLAB	Python	null, RStudio , MATLAB , Jupyter Notebook	A laptop
13	13	13	Nigeria	Master's degree	Program/Project Manager	5-10 years	Python, SQL	Python	null, Spyder , Jupyter Notebook	A laptop
14	14	14	Nigeria	Bachelor's degree	Other	< 1 years	Python	Python	null, Visual Studio Code (VSCode)	A laptop
15	15	15	Greece	Doctoral degree	Research Scientist	10-20 ye...	Python, C, C++, MATLAB	Python	null, Spyder	A laptop
16	16	16	Belgium	Bachelor's degree	Data Analyst	20+ years	Python, SQL	Python	Jupyter (JupyterLab, Jupyter Notebooks, etc...	A laptop
17	17	17	Pakistan	Bachelor's degree	Data Scientist	1-3 years	Python, SQL	Python	Jupyter (JupyterLab, Jupyter Notebooks, etc...	A laptop
18	18	18	Japan	Master's degree	Software Engineer	3-5 years	Python, SQL, C, Java, Jav...	Python	Jupyter (JupyterLab, Jupyter Notebooks, etc...	A laptop


data_txt
index_txt INT
index_sql INT
q3 VARCHAR(500)
q4 VARCHAR(500)
q5 VARCHAR(500)
q6 VARCHAR(500)
q7 VARCHAR(500)
q8 VARCHAR(500)
q9 VARCHAR(500)
q11 VARCHAR(500)
q12 VARCHAR(500)
q13 VARCHAR(500)
q14 VARCHAR(500)
q15 VARCHAR(500)
q16 VARCHAR(500)
q17 VARCHAR(1000)
q20 VARCHAR(500)
q21 VARCHAR(500)
q22 VARCHAR(500)
q23 VARCHAR(500)
q24 VARCHAR(1000)
q25 VARCHAR(500)
q26 VARCHAR(500)
q32 VARCHAR(1000)
q33 VARCHAR(500)
q34 VARCHAR(1000)
q35 VARCHAR(500)
q41 VARCHAR(500)
Indexes
data_sql
index_sql INT
q10_part_1 VARCHAR(300)
q10_part_2 VARCHAR(300)
q10_part_3 VARCHAR(300)
q10_part_4 VARCHAR(300)
q10_part_5 VARCHAR(300)
q10_part_6 VARCHAR(300)
q10_part_7 VARCHAR(300)
q10_part_8 VARCHAR(300)
q10_part_9 VARCHAR(300)
q10_part_10 VARCHAR(300)
q10_part_11 VARCHAR(300)
q10_part_12 VARCHAR(300)
q10_part_13 VARCHAR(300)
q10_part_14 VARCHAR(300)
q10_part_15 VARCHAR(300)
q10_part_16 VARCHAR(300)
q10_other VARCHAR(300)
Indexes
data_xml
index_xml INT
time VARCHAR(300)
age VARCHAR(300)
gender VARCHAR(300)
index_sql INT
Indexes

Thank you

FOLLOW ME ON :

My portfolio: solkiria.github.io/solkiria/
e-mail : sonia.ruiz.p31@gmail.com
LinkedIn : linkedin.com/in/sonia-ruiz-perez/
GitHub : github.com/solkiria
Tableau : public.tableau.com/app/profile/solkiria