

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

```

/*!40103 SET @OLD_TIME_ZONE=@@TIME_ZONE */;
/*!40103 SET TIME_ZONE='+00:00' */;
/*!40014 SET @OLD_UNIQUE_CHECKS=@UNIQUE_CHECKS, UNIQUE_CHECKS=0 */;
/*!40014 SET @OLD_FOREIGN_KEY_CHECKS=@FOREIGN_KEY_CHECKS, FOREIGN_KEY_CHECKS=0 */;
15 • /*!40101 SET @OLD_SQL_MODE=@SQL_MODE, SQL_MODE='NO_AUTO_VALUE_ON_ZERO' */;
16 • /*!40111 SET @OLD_SQL_NOTES=@SQL_NOTES, SQL_NOTES=0 */;

17 --
18 --
19 -- Dumping data for table `data_sql`
20 --
21 
22 • LOCK TABLES `data_sql` WRITE;
23 • /*!40000 ALTER TABLE `data_sql` DISABLE KEYS */;
24 • INSERT INTO `data_sql` VALUES (1,' Kaggle Notebooks','Colab Notebooks','ERROR','E
25 • INSERT INTO `data_sql` VALUES (6184,'ERROR','ERROR','ERROR','ERROR','ERR
26 • INSERT INTO `data_sql` VALUES (12373,' Kaggle Notebooks','Colab Notebooks','ERROR
27 • INSERT INTO `data_sql` VALUES (18553,'ERROR','ERROR','ERROR','ERROR','ERR
28 • INSERT INTO `data_sql` VALUES (24707,'ERROR','ERROR','ERROR','ERROR','ERR
29 • /*!40000 ALTER TABLE `data_sql` ENABLE KEYS */;
30 • UNLOCK TABLES;
31 • /*!40103 SET TIME_ZONE=@OLD_TIME_ZONE */;
32 
33 • /*!40101 SET SQL_MODE=@OLD_SQL_MODE */;
34 • /*!40014 SET FOREIGN_KEY_CHECKS=@OLD_FOREIGN_KEY_CHECKS */;
35 • /*!40014 SET UNIQUE_CHECKS=@OLD_UNIQUE_CHECKS */;
36 • /*!40101 SET CHARACTER_SET_CLIENT=@OLD_CHARACTER_SET_CLIENT */;
37 • /*!40101 SET CHARACTER_SET_RESULTS=@OLD_CHARACTER_SET_RESULTS */;
38 • /*!40101 SET COLLATION_CONNECTION=@OLD_COLLATION_CONNECTION */;
```

Archivo .XML

```

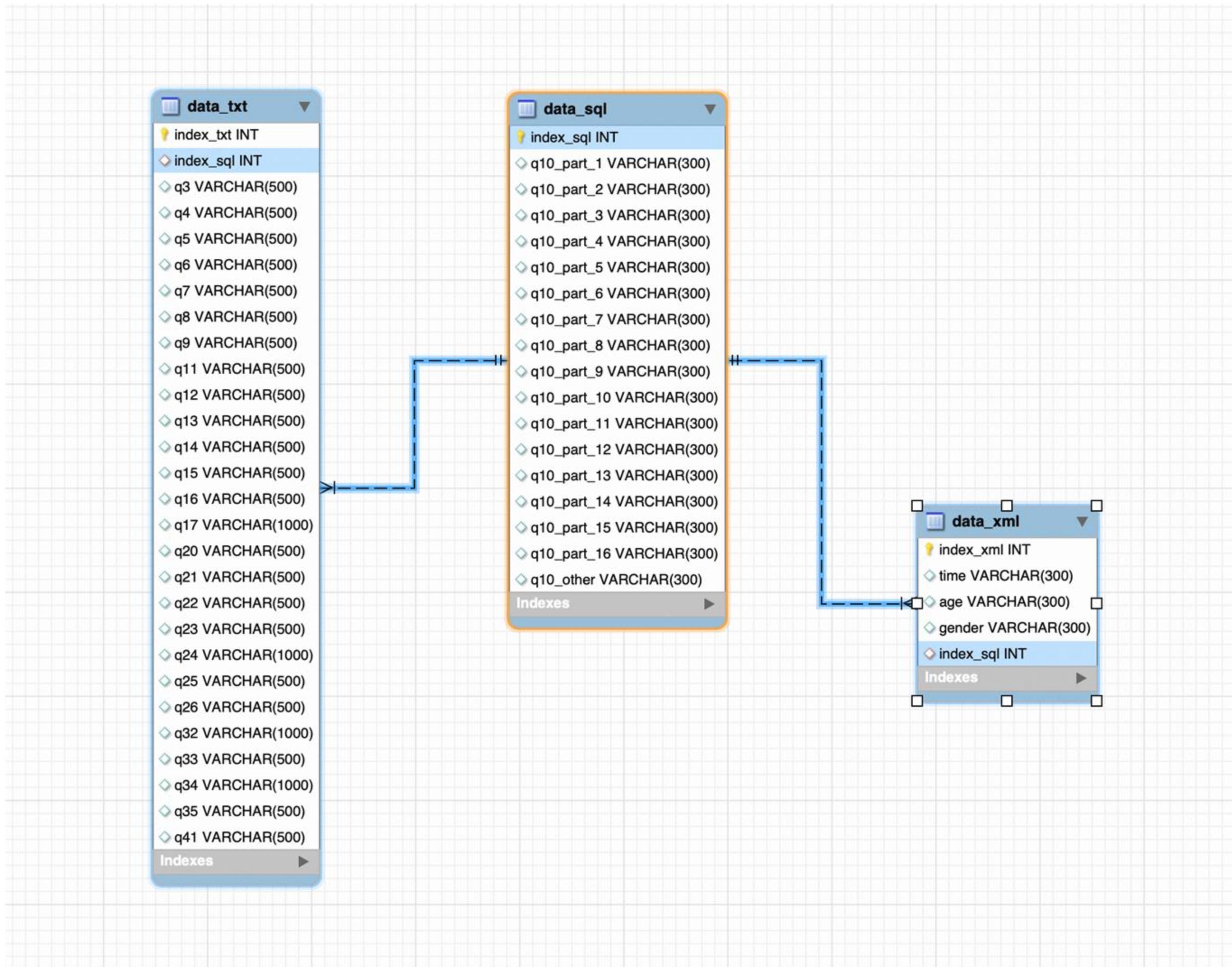
<?xml version="1.0" encoding="utf-8"?>
<data>
<row>
<level_0>1</level_0>
<index>1</index>
<time>784</time>
<age>50-54</age>
<gender>0</gender>
</row>
<row>
<level_0>2</level_0>
<index>2</index>
<time>924</time>
<age>22-24</age>
<gender>0</gender>
</row>
<row>
<level_0>3</level_0>
<index>3</index>
<time>575</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>
<row>
```

Archivo .XML

Archivo .TXT

```

x 0 data_txt.txt
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, None;Never;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 (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 production);Analyze and understand data to influence product or business decisions, Build prototypes to explore applying machine learning to new areas;30,000-39,999;$1000-$9,999;null, None;;null, Microsoft Power BI;;Local development environments (RStudio, JupyterLab, etc.)5;India;I prefer not to answer;Currently not employed;< 1 years;Python;Python;JupyterLab, Jupyter Notebooks, etc , PyCharm , Spyder , Jupyter Notebook;A laptop;null, Google Cloud TPUs ;2-5 times; Matplotlib , Seaborn , Ggplot / ggplot2 ;Under 1 year; Scikit-learn , TensorFlow , Keras , PyTorch , Fast.ai ;Linear or Logistic Regression, Decision Trees or Random Forests;;;;;Local development environments (RStudio, JupyterLab, etc.)6;India;Some college/university study without earning a bachelor's degree;Student;1-3 years;null, C++, Java, JavaScript;Python;null, Visual Studio , Visual Studio Code (VSCode) , Jupyter Notebook;A laptop;null, None;Never;null, Geoplotlib ;Under 1 year;null, Fast.ai ;;;;;;;7;India;Bachelor's degree;Data Scientist;5-10 years;Python;Python;null, Jupyter Notebook;A personal computer / desktop;null, Google Cloud TPUs ;2-5 times; Matplotlib , Plotly Express ;2-3 years; Scikit-learn , TensorFlow , Keras ;null, Decision Trees or Random Forests, Dense Neural Networks (MLPs, etc), Convolutional Neural Networks, Generative Adversarial Networks;Computers/Technology;10,000 or more employees;3-4;We have well established ML methods (i.e., models in production for more than 2 years);null, Build and/or run the data infrastructure that my business uses for storing, analyzing, and operationalizing data, Build prototypes to explore applying machine learning to new areas;15,000-19,999;$1-$99;MySQL ;null, None;Basic statistical software (Microsoft Excel, Google Sheets, etc.)8;Russia;Bachelor's degree;Currently not employed;3-5 years;Python, SQL;Python;null, Other;A cloud computing platform (AWS, Azure, GCP, hosted notebooks, etc);null, None;Never; Matplotlib ;Under 1 year; Scikit-learn , Xgboost ;null, Decision Trees or Random Forests, Gradient Boosting Machines (xgboost, lightgbm, etc);;;;;;Basic statistical software (Microsoft Excel, Google Sheets, etc.)9;Turkey;I prefer not to answer;Other;1-3 years;Python, SQL;SQL;null, Spyder , Jupyter Notebook;A laptop;null, None;Never; Matplotlib , Seaborn ;Under 1 year; Scikit-learn , TensorFlow ;Linear or Logistic Regression;Manufacturing/Fabrication;50-249 employees;1-2;I do not know;Analyze and understand data to influence product or business decisions;$0-999;$0 ($USD);;;Business intelligence software (Salesforce, Tableau, Spotfire, etc.)10;Australia;Doctoral degree;Other;1-3 years;Python, R, SQL;R;null, RStudio , Jupyter Notebook;A personal computer / desktop;null, None;Never; Matplotlib , Seaborn , Ggplot / ggplot2 ;I do not use machine learning methods;;;Other;0-49 employees;0;No (we do not use ML methods);null, None of these activities are an important part of my role at work;70,000-79,999;$1-$99;MySQL ;null, Microsoft Power BI, Tableau, Alteryx ;Tableau;Local development environments (RStudio, JupyterLab, etc.)11;India;Master's degree;Student;< 1 years;Python, R, C++;R;null, RStudio , Jupyter Notebook;A laptop;null, None;Never; Matplotlib , Seaborn , Ggplot / ggplot2 ;Under 1 year; Scikit-learn ;Linear or Logistic Regression;;;;;Advanced statistical software (SPSS, SAS, etc.)12;India;Master's degree;Student;< 1 years;Python, MATLAB;Python;null, MATLAB , Jupyter Notebook;A laptop;;;;;;;13;Nigeria;Master's degree;Program/Project Manager;5-10 years;Python, SQL;Python;null, Spyder , Jupyter Notebook;A laptop;null, None;Never; Matplotlib , Seaborn ;1-2 years; Scikit-learn , Xgboost ;Linear or Logistic Regression, Decision Trees or Random Forests;Shipping/Transportation;1000-9,999 employees;10-14;We are exploring ML methods (and
```



## Características de la Base de Datos:

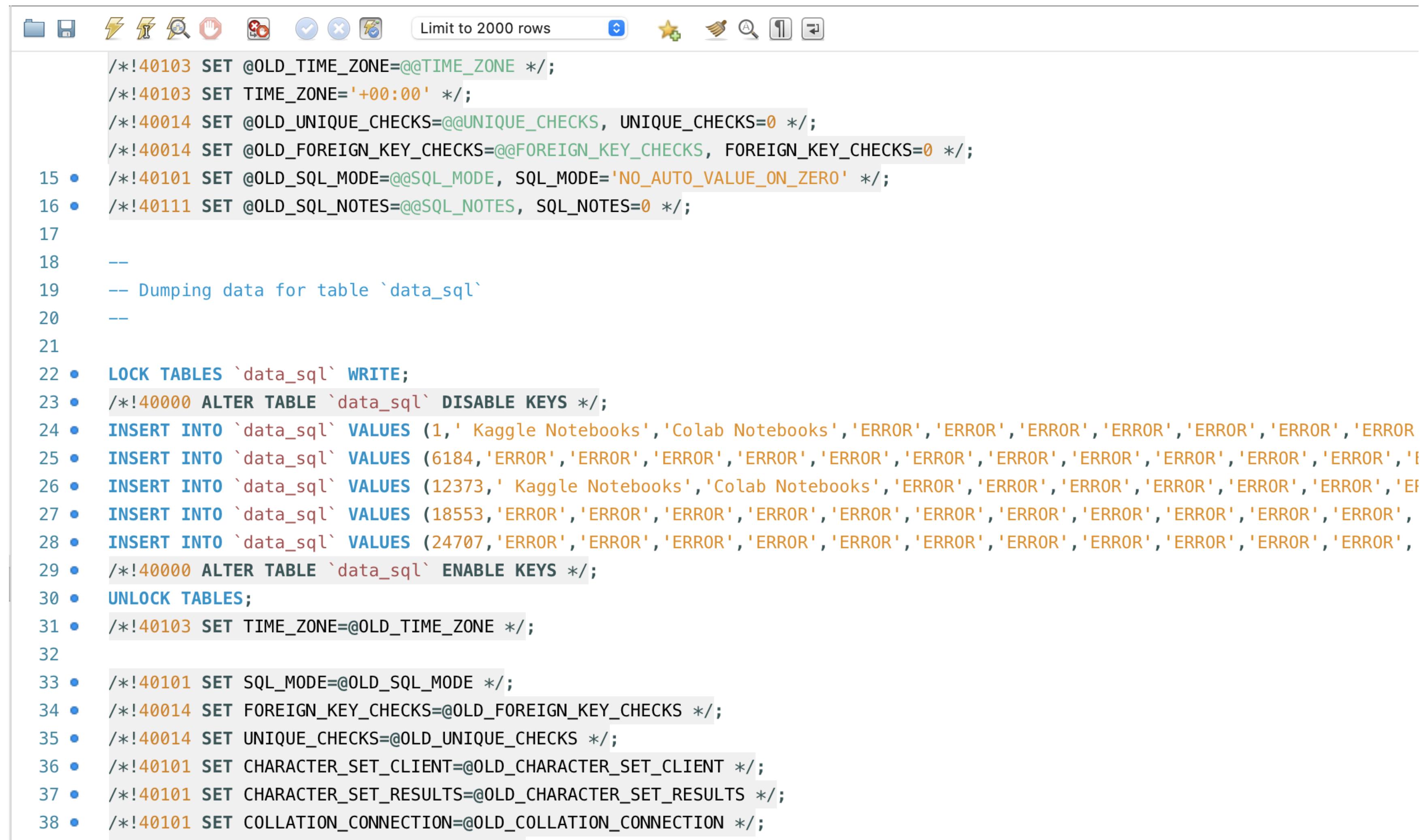
- Nombre de la Base de Datos: "project1":
- Nombre de las Tablas: "data\_sql", "data\_xml", "data\_txt".

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.

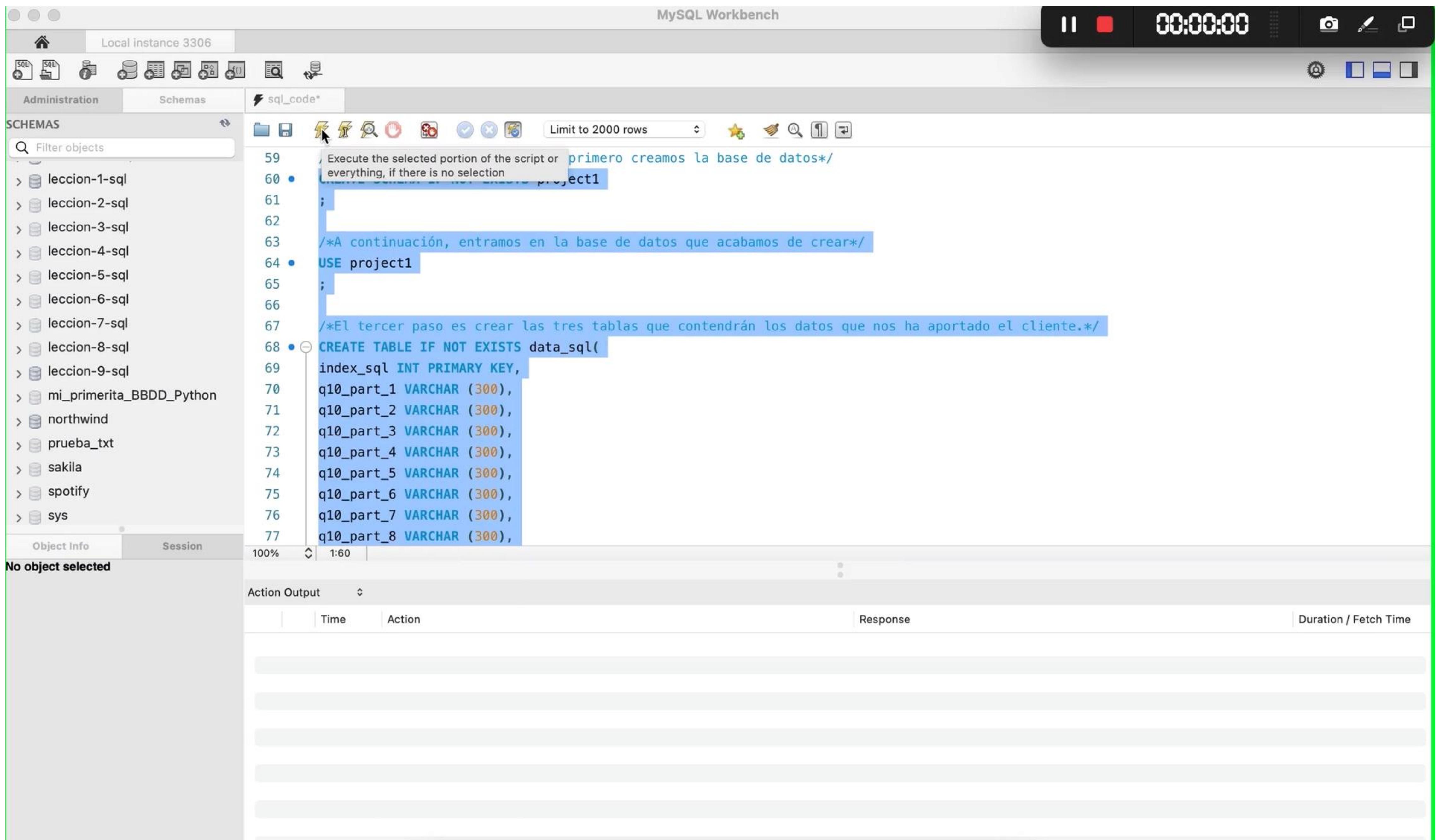
## Características del archivo .sql



```
/*!40103 SET @OLD_TIME_ZONE=@@TIME_ZONE */;
/*!40103 SET TIME_ZONE='+00:00' */;
/*!40014 SET @OLD_UNIQUE_CHECKS=@@UNIQUE_CHECKS, UNIQUE_CHECKS=0 */;
/*!40014 SET @OLD_FOREIGN_KEY_CHECKS=@@FOREIGN_KEY_CHECKS, FOREIGN_KEY_CHECKS=0 */;
15 • /*!40101 SET @OLD_SQL_MODE=@@SQL_MODE, SQL_MODE='NO_AUTO_VALUE_ON_ZERO' */;
16 • /*!40111 SET @OLD_SQL_NOTES=@@SQL_NOTES, SQL_NOTES=0 */;

17
18 --
19 -- Dumping data for table `data_sql`
20 --
21
22 • LOCK TABLES `data_sql` WRITE;
23 • /*!40000 ALTER TABLE `data_sql` DISABLE KEYS */;
24 • INSERT INTO `data_sql` VALUES (1,' Kaggle Notebooks','Colab Notebooks','ERROR','ERROR','ERROR','ERROR','ERROR','ERROR',
25 • INSERT INTO `data_sql` VALUES (6184,'ERROR','ERROR','ERROR','ERROR','ERROR','ERROR','ERROR','ERROR','ERROR','ERROR','ERROR','E
26 • INSERT INTO `data_sql` VALUES (12373,' Kaggle Notebooks','Colab Notebooks','ERROR','ERROR','ERROR','ERROR','ERROR','ERROR','ERROR','E
27 • INSERT INTO `data_sql` VALUES (18553,'ERROR','ERROR','ERROR','ERROR','ERROR','ERROR','ERROR','ERROR','ERROR','ERROR','ERROR','E
28 • INSERT INTO `data_sql` VALUES (24707,'ERROR','ERROR','ERROR','ERROR','ERROR','ERROR','ERROR','ERROR','ERROR','ERROR','ERROR','E
29 • /*!40000 ALTER TABLE `data_sql` ENABLE KEYS */;
30 • UNLOCK TABLES;
31 • /*!40103 SET TIME_ZONE=@OLD_TIME_ZONE */;

32
33 • /*!40101 SET SQL_MODE=@OLD_SQL_MODE */;
34 • /*!40014 SET FOREIGN_KEY_CHECKS=@OLD_FOREIGN_KEY_CHECKS */;
35 • /*!40014 SET UNIQUE_CHECKS=@OLD_UNIQUE_CHECKS */;
36 • /*!40101 SET CHARACTER_SET_CLIENT=@OLD_CHARACTER_SET_CLIENT */;
37 • /*!40101 SET CHARACTER_SET_RESULTS=@OLD_CHARACTER_SET_RESULTS */;
38 • /*!40101 SET COLLATION_CONNECTION=@OLD_COLLATION_CONNECTION */;
```



```
<?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>
<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

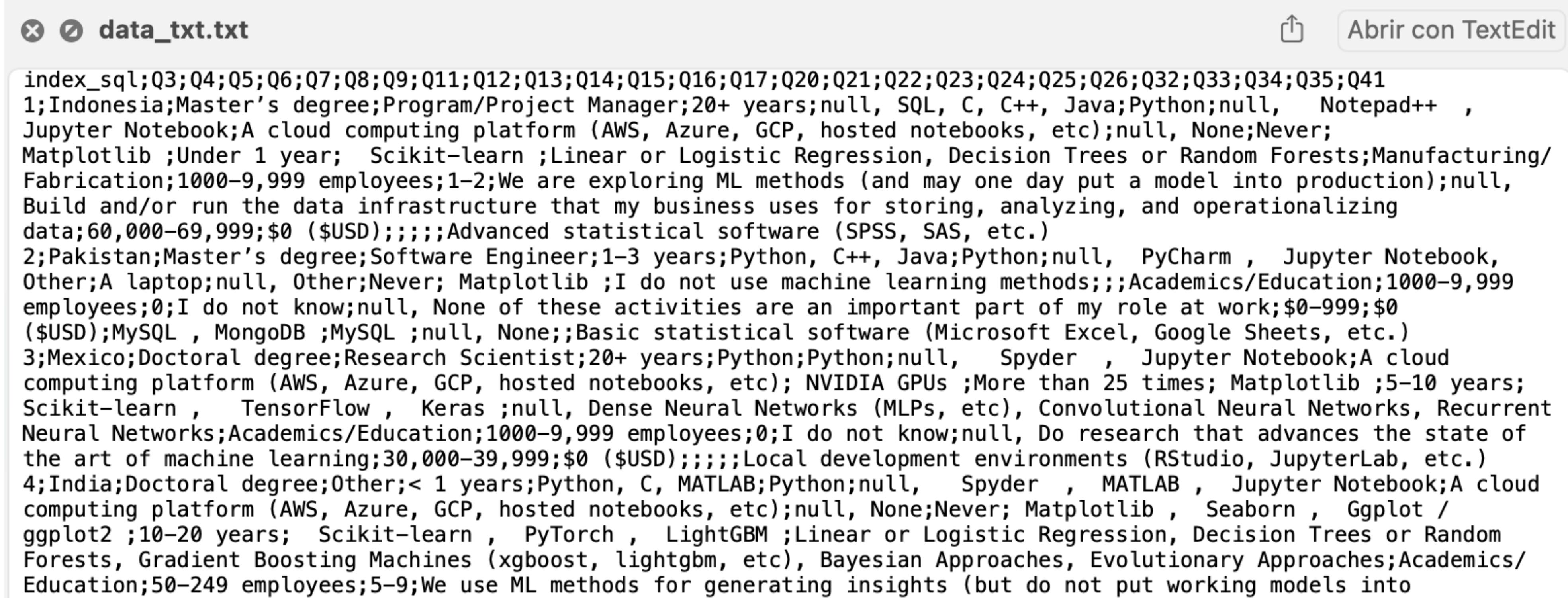
class XML:
    def __init__(self, archivoxml):
        self.tree = ET.parse(archivoxml)
        self.root = self.tree.getroot()
        self.archivoxml = archivoxml

    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



```
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?

The screenshot shows a Jupyter Notebook interface with the following details:

- EXPLORADOR**: Shows a list of files and notebooks in the current project:
  - txt\_code.ipynb
  - productowner\_code.ipynb
  - txt\_code.ipynb
  - producto... u
  - .ipynb\_checkpoints
  - .gitignore
  - actas\_daily\_meetings
  - data\_sql.sql
  - data\_txt.csv u
  - data\_txt.txt
  - data\_xml.xml
  - explicacion\_creacion...
  - explicacion\_sql\_code
  - funcion-xml-intento...
  - intento-xml-tree\_3.1...
  - productowner... u
  - read\_xml\_txt.ipynb
  - README.md
  - sql\_code.sql
  - txt\_code.ipynb
  - txt\_code3.ipynb
- EDITORES**: Shows two code cells:

  - Cell 26**:

```
# En primer lugar guardamos el contenido del archivo .txt en la variable "file_source".  
with open('data_txt.txt', 'r') as file:  
    file_source = file.readlines()
```

Output: ✓ 1.1s
  - Cell 27**:

```
# Leemos el contenido de la variable. Y vemos que es una lista, donde cada elemento  
# es una fila de la futura tabla.  
file_source
```

Output: ✓ 0.1s

- Output**: Shows the full content of the `file\_source` list, which is a large string of data separated by newlines.
- Bottom Bar**: Includes icons for file operations, a search bar, and status information: "main\*", "Live Share", "Git Graph", "Servidor de Jupyter: local", "Celda 4 de 35", and a progress bar.

## Desarrollo de funciones.

The screenshot shows a Jupyter Notebook environment with the following details:

- EXPLORADOR**: Shows files like `txt_code.ipynb`, `productowner_code.ipynb`, `txt_code.ipynb`, and `producto... U`.
- EDITORES...**: Shows two code cells:

  - Cell 31 (Python)**:

```
# Ahora que ya tenemos la funcion para los '\n', simplemente metemos en la funcion las
# sustituciones de '<' por 'under' y los 'null' por 'NULL'.
def sustituir(lista_txt):
    # Esta función sirve para quitar saltos de linea(\n), cambiar '<' por 'under' y
    # 'null' por 'NULL', cuyo parámetro será una variable que almacene, en formato lista,
    # el contenido de un .txt.
    lista_sustituida=[]
    for i in lista_txt:
        saltos=i.replace("\n","");
        menores=saltos.replace("<","under")
        nules=menores.replace("null","NULL")
        lista_sustituida.append(nules)
    return lista_sustituida
```
  - Cell 32 (Python)**:

```
# Ejecutamos la funcion creada sobre la variable que tiene el contenido del .txt.
file_source_sustituido = sustituir(file_source)
print(file_source_sustituido)
```

- RESULTADOS**: Shows the output of Cell 32, which is a long list of strings representing the cleaned content of the text file.
- BOTONES**: Includes buttons for executing cells, refreshing, and other notebook operations.
- BAR DE HERRAMIENTAS**: Includes buttons for main, live share, git graph, and a footer bar with server information.

## Desarrollo de funciones.

The screenshot shows a Jupyter Notebook environment with the following details:

- EXPLORADOR**: Shows the project structure with files like `txt_code.ipynb`, `productowner_code.ipynb`, `.ipynb_checkpoints`, `.gitignore`, `actas_daily_meetings`, `data_sql.sql`, `data_txt.csv`, `data_txt.txt`, `data_xml.xml`, `explicacion_creacion...`, `explicacion_sql_code`, `funcion-xml-intento...`, `intento-xml-tree_3.1...`, `productowner...`, `read_xml_txt.ipynb`, `README.md`, `sql_code.sql`, `txt_code.ipynb` (selected), and `txt_code3.ipynb`.
- EDITORES ...**: Displays two code cells:

  - Cell 33 (Python)**: Contains the function `limpiar` which takes a parameter `archivo_txt`. It splits the input into a list of columns, strips whitespace from each column, and then checks if the length of the resulting list matches the original column. If so, it appends the list to `lista_espacio3` and clears `lista_espacio2`. Otherwise, it appends the original column to `lista_espacio2`. Finally, it returns `lista_espacio3`. The cell status is `0.1s`.
  - Cell 34 (Python)**: Contains the code to execute the `limpiar` function on the variable `file_source_sustituido` and print the result. The cell status is `1.6s`.

- TOOLBAR**: Includes icons for search, refresh, and other Jupyter functions.
- STATUS BAR**: Shows the current notebook (`main*`), cell count (12), and live share status.

## Resultado preliminar.

The screenshot shows a Jupyter Notebook interface with the following details:

- Left Sidebar:** Includes icons for Explorador, Editores, Projecto, Archivos, Compartir, README.md, SQL Code, and two .ipynb files.
- Top Bar:** Shows the file names "txt\_code.ipynb" and "productowner\_code.ipynb" with status indicators (● or ○), and a "base (Python 3.9.7)" label.
- Toolbar:** Includes buttons for Código, Markdown, Ejecutar todo, Borrar resultados de todas las celdas, Reiniciar, Variables, Esquema, and a set of icons for copy/paste, refresh, and other operations.
- Code Cell:** Contains the following Python code:

```
# Ejecutamos la funcion creada sobre la variable que tiene el contenido del .txt, con
# la limpieza previa ya hecha en este caso.
file_source_limpio=limpiar(file_source_sustituido)
print(file_source_limpio)
```
- Output Cell:** Shows the result of the execution: "[34] ✓ 1.6s" followed by a large block of text representing the cleaned content of the .txt file.
- Bottom Status Bar:** Displays "main\* ○ ⊞ 12 ▲ 22 Live Share Git Graph" and "Servidor de Jupyter:local Celda 15 de 35" along with other small icons.

## Desarrollo de clase.

The screenshot shows a Jupyter Notebook interface with the following details:

- EXPLORADOR**: Shows the file structure of the project. The current file is `txt_code.ipynb`.
- EDITORES ...**: Shows two code cells. The active cell contains the following Python code:

```
# Finalmente, metemos todas las funciones dentro de la clase Txt anteriormente creada.
# Primero definimos el nombre de nuestra clase.
class Txt:
    # Creamos el método constructor donde definimos los parámetros que estamos interesados.
    def __init__(self, archivo_txt):
        # Definimos la variable que almacena el contenido del archivo que queremos limpiar.
        self.archivo_txt= archivo_txt

    # Definimos el primer método para quitar saltos de línea(\n), cambiar '<' por 'under' y
    # 'null' por 'NULL', cuyo parámetro será una variable que almacene, en formato lista,
    # el contenido de un .txt.
    def sustituir(self):
        lista_sustituida=[]
        for i in self.archivo_txt:
            saltos=i.replace("\n","");
            menores=saltos.replace("<","under")
            nules=menores.replace("null","NULL")
            lista_sustituida.append(nules)
        return lista_sustituida

    # Definimos el segundo método para eliminar los espacios innecesarios al principio y al final del
    # contenido de cada columna. El parámetro será una variable que almacene, en formato lista,
    # el contenido de un .txt.
    def limpiar(self):
        lista_espaciol=[]
        for entrada in self.archivo_txt:
            espacio1=entrada.split(";");
            lista_espaciol.append(espacio1)
        lista_espacio2=[]
        lista_espacio3=[]
        for columna in lista_espaciol:
            for elemento in columna:
                espacio2=elemento.strip()
                lista_espacio2.append(espacio2)
            lista_espacio3.append(lista_espacio2)
        return lista_espacio3
```

The notebook has tabs for `txt_code.ipynb` and `productowner_code.ipynb`. The top bar includes buttons for Código, Markdown, Ejecutar todo, Borrar resultados de todas las celdas, Reiniciar, Variables, Esquema, and base (Python 3.9.7). The bottom bar shows navigation icons and status information: main\*, 12 ▲ 22, Live Share, Git Graph, Servidor de Jupyter: local, Celda 34 de 35, and a few other small icons.

# Desarrollo de clase.

# Desarrollo de clase.

## Desarrollo de clase.

The screenshot shows a Jupyter Notebook interface with the following details:

- EXPLORADOR**: Shows a file tree with several .ipynb files and other project files like .gitignore and README.md.
- EDITORES ...**: Displays two notebooks: "txt\_code.ipynb" and "productowner\_code.ipynb".
- CELLS**: The current cell is numbered 31 of 35.
- CONTENIDO**: The cell content is a Python script for exporting data from a MySQL database to a CSV file.

```
txt_code.ipynb  ●  productowner_code.ipynb u  ●
txt_code.ipynb > M+Lectura y explicación de la estructura del archivo .txt > M+Limpieza del archivo .txt > # En caso de no tener creada la tabla para insertar los datos del fichero txt, procedemos
+ Código + Markdown | Ejecutar todo Borrar resultados de todas las celdas Reiniciar Variables Esquema ...
base (Python 3.9.7)

return

# Definimos el quito método para crear un dataframe de pandas de nuestros datos
# del .txt. y después almacenarlo como un archivo externo.

def exportar_txt_csv (self):
    import os
    print('-----')
    print('La ruta en la que se va a guardar el archivo es:')
    print(os.getcwd())
    import mysql.connector
    import pandas as pd
    from mysql.connector import errorcode
    try:
        cnx = mysql.connector.connect(user='root', password='AlumnaAdalab',
                                      host='127.0.0.1',
                                      database='project1')
        print('-----')
        print('Conexión con la base de datos realizada')
    except mysql.connector.Error as err:
        if err.errno == errorcode.ER_ACCESS_DENIED_ERROR:
            print("Ha habido un error al introducir el nombre de usuario o la contraseña")
        elif err.errno == errorcode.ER_BAD_DB_ERROR:
            print("La base de datos no existe")
        else:
            print(err)
    # Realizamos la consulta a la tabla alumnas mediante pandas y creamos el dataframe.
    query = """SELECT * FROM data_txt"""
    df = pd.read_sql_query(query, cnx)

    # Guardamos los datos registrados del dataframe en un fichero csv (separado por comas).
    df.to_csv("data_txt.csv")

    # Cerramos la conexión
    cnx.close()
```

- FOOTER**: Includes navigation icons (Back, Forward, Home, etc.) and status information: "Servidor de Jupyter: local" and "Celda 31 de 35".

## Código en ejecución.

productowner\_code.ipynb

funcion-xml-con-datos-insertos1.ipynb ● productowner\_code.ipynb ●

00:00:00

EXPLORADOR ... Iniciar + Código + Markdown ▶ Ejecutar todo Borrar resultados de todas las celdas Reiniciar Variables Esquema ... base (Python 3.9.13)

EDITORES AB... 2 sin guardar

Iniciar ● funcion-xml-con-d... ● productowner\_cod...

> NO HAY NINGUNA CARPETA ...

ESQUEMA

No se encontró ningún símbolo en el documento "productowner\_code.ipynb"

[1] MagicPython

```
# En primer lugar guardamos el contenido del archivo .txt en la variable "file_source".
with open('data_txt.txt','r') as file:
    file_source = file.readlines()

# Finalmente, metemos todas las funciones dentro de la clase Txt anteriormente creada.
# Primero definimos el nombre de nuestra clase.
class Txt:
    # Creamos el método constructor donde definimos los parámetros que estamos interesados.
    def __init__(self, archivo_txt):
        # Definimos la variable que almacena el contenido del archivo que queremos limpiar.
        self.archivo_txt= archivo_txt

        # Definimos el primer método para quitar saltos de línea(\n), cambiar '<' por 'under y
        # 'null' por 'NULL', cuyo parámetro será una variable que almacene, en formato lista,
        # el contenido de un .txt.
    def sustituir(self):
        lista_sustituida=[]
        for i in self.archivo_txt:
            saltos=i.replace("\n","");
            menores=saltos.replace("<","under")
            nules=menores.replace("null","NULL")
            lista_sustituida.append(nules)
        return lista_sustituida

    # Definimos el segundo método para eliminar los espacios innecesarios al principio y al final del
    # contenido de cada columna. El parámetro será una variable que almacene, en formato lista,
    # el contenido de un .txt.
    def limpiar(self):
        lista_espacio1=[]
        for entrada in self.archivo_txt:
            espacio1=entrada.split(";");
            lista_espacio1.append(espacio1)
        lista_espacio2=[]
        lista_espacio3=[]
        for columna in lista_espacio1:
            for elemento in columna:
                espacio2=elemento.strip()
```

## Datos TXT en CSV

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

## Archivo SQL en SQL

Adalab Server

File Edit View Query Database Server Tools Scripting Help

Schemas

SCHEMAS

Filter objects

project1

Tables

data\_sql

data\_txt

data\_xml

Views

Stored Proced

Functions

prueba1

prueba\_z

sakila

sys

Object Info

Table: data\_sql

Columns:

- index\_sql nt(11) PK
- q10\_part\_1 varchar(300)
- q10\_part\_2 varchar(300)

data\_sql 1

data\_sql

Limit to 200 rows

1 • SELECT \* FROM project1.data\_sql;

Result Grid

Filter Rows:

Edit: Export/Import: Wrap Cell Content: Fetch rows:

#	index_sql	q10_part_1	q10_part_2	q10_part_3	q10_part_4	q10_part_5	q10_part_6	q10_part_7	q10_part_8	q10_part_9	q10_part_10	q10_part_11
1	1	Kaggle Notebooks	Colab Notebooks	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
2	2	Kaggle Notebooks	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
3	3	NULL	Colab Notebooks	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
4	4	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	Google Col...
5	5	NULL	Colab Notebooks	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
6	6	Kaggle Notebooks	NULL	NULL	NULL	NULL	NULL	NULL	NULL	Google Cloud Notebooks (AI Platform / Ver...	NULL	NULL
7	7	Kaggle Notebooks	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
8	8	Kaggle Notebooks	Colab Notebooks	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
9	9	Kaggle Notebooks	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
10	10	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
11	11	Kaggle Notebooks	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
12	12	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
13	13	Kaggle Notebooks	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
14	14	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
15	15	Kaggle Notebooks	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
16	16	Kaggle Notebooks	Colab Notebooks	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
17	17	NULL	Colab Notebooks	NULL	NULL	Binder / J...	NULL	NULL	NULL	NULL	NULL	NULL
18	18	Kaggle Notebooks	Colab Notebooks	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
19	19	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL

Apply Revert

Query Completed

## Archivo XML en SQL

Adalab Server

File Edit View Query Database Server Tools Scripting Help

Schemas

SCHEMAS

Filter objects

project1

Tables

- data\_sql
- data\_txt
- data\_xml

Views

Stored Proced

Functions

prueba1

prueba\_z

sakila

sys

Object Info

Table: data\_xml

Columns:

- index\_xml int(11) AI PK
- time varchar(300)
- age varchar(300)

data\_xml 1

data\_xml

1 • SELECT \* FROM project1.data\_xml;

Result Grid

#	index_xml	time	age	gender	index_sql
1	1	784	50-54	Man	1
2	2	924	22-24	Man	2
3	3	575	45-49	Man	3
4	4	781	45-49	Man	4
5	5	1020	25-29	Woman	5
6	6	141	18-21	Woman	6
7	7	484	30-34	Man	7
8	8	1744	22-24	Man	8
9	9	655	30-34	Man	9
10	10	1777	40-44	Man	10
11	11	3081	18-21	Woman	11
12	12	1922	18-21	Woman	12
13	13	852	45-49	Man	13
14	14	838	22-24	Man	14
15	15	563	35-39	Man	15
16	16	1315	50-54	Man	16
17	17	479	18-21	Man	17
18	18	249	22-24	Man	18
19	19	650	30-34	Man	19

Result Grid

Form Editor

Field Types

Query Stats

Execution Plan

Apply Revert

Query Completed

## Archivo TXT en SQL

Adalab Server

File Edit View Query Database Server Tools Scripting Help

sql\_code data\_txt SQL File 3\* data\_txt

Limit to 200 rows

```
1 • SELECT * FROM project1.data_txt;
```

Result Grid Filter Rows: Export/Import: Wrap Cell Content: Fetch rows:

#	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

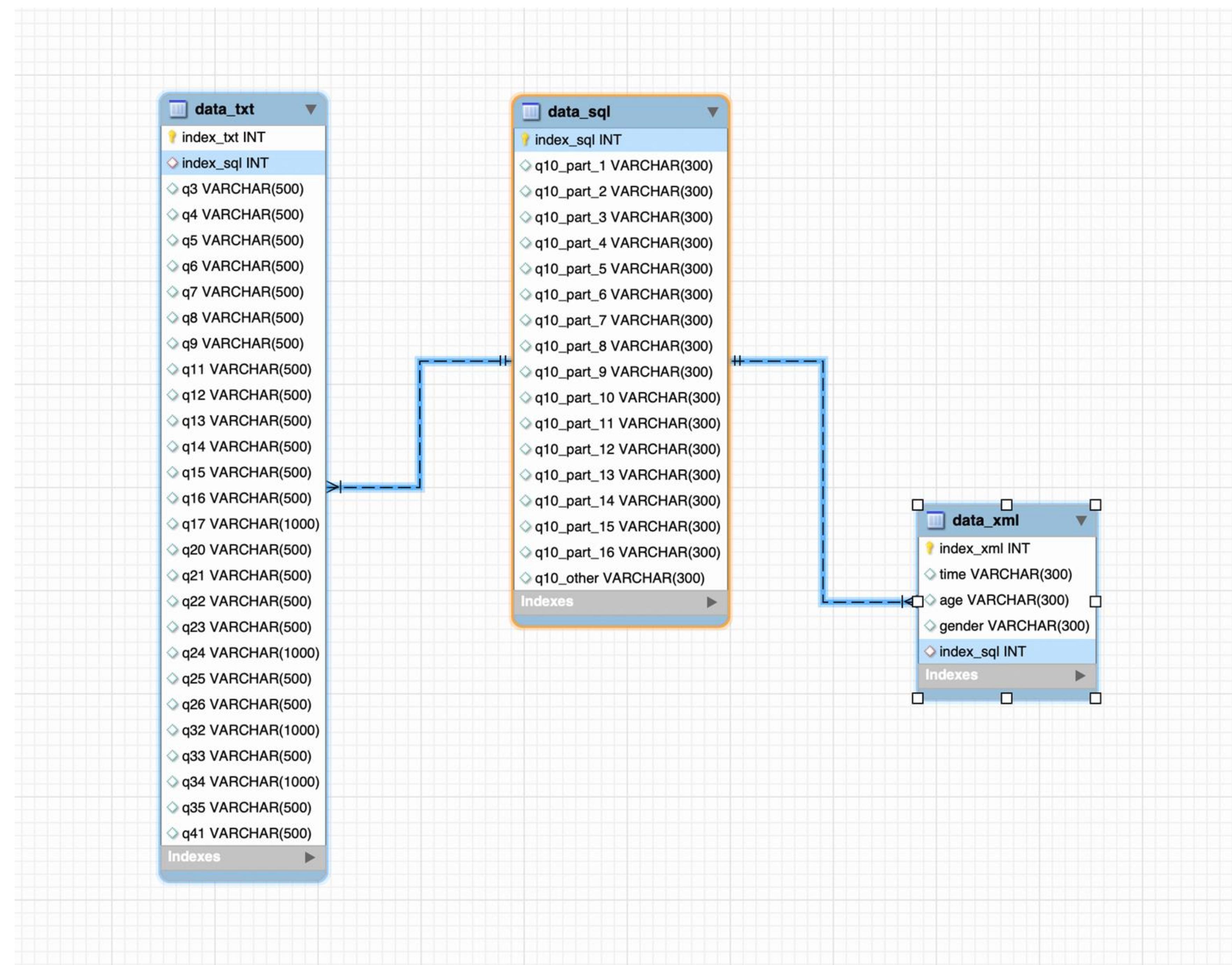
data\_txt 1

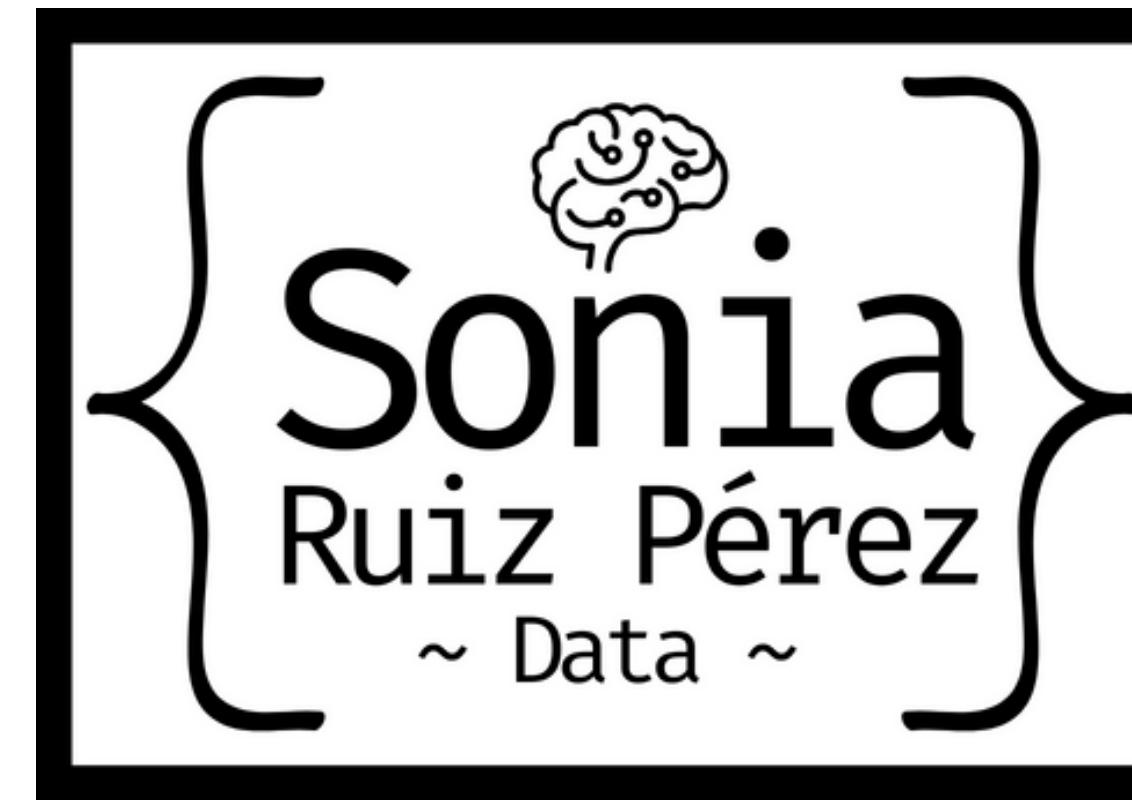
Apply Revert

Action Output

Query Completed

Result Grid Form Editor Field Types Query Stats Execution Plan





# Thank you

FOLLOW ME ON :

**My portfolio:** [solkiria.github.io/solkiria/](https://solkiria.github.io/solkiria/)

**e-mail :** [sonia.ruiz.p31@gmail.com](mailto:sonia.ruiz.p31@gmail.com)

**LinkedIn :** [linkedin.com/in/sonia-ruiz-perez/](https://linkedin.com/in/sonia-ruiz-perez/)

**GitHub :** [github.com/solkiria](https://github.com/solkiria)

**Tableau :** [public.tableau.com/app/profile/solkiria](https://public.tableau.com/app/profile/solkiria)