Formative Piece 1

**Initial Data Quality Assessment**.

The dataset contains 418 records and 12 Fields. (A sample of data can be seen in Fig1)

Here are my observations regarding data quality:

**Completeness**

* The Age Field has 86 missing values.
* The Fare Field has 1 missing value.
* The Cabin Field is widely incomplete with multiple Missing Values and multiple values in the same record. (327 out of 418).

**Accuracy**

* The PassengerId, Survived, and Pclass fields appear consistent with expected integer values. The name Field often has 2 names recorded (possibly Husband and Wife for example under 1 record.

**Consistency**

* The Embarked field seems consistent (valid values likely "S", "C", "Q"). The Ticket Field is in Multiple formats.
* The name Field is inconsistent as there are often has 2 names recorded (possibly Husband and Wife for example) under 1 record.
* The Age Field is inconsistent as there are various records that are recorded to 1 decimal place while others are integers only
* The Ticket Field is very inconsistent as there are multiple Formats of ticket codes recorded.
* The Sex column has only two expected values ("male", "female").

**Uniqueness**

* The PassengerId field is unique (Ideal to be used as a primary key in a database).
* The Ticket Field may have duplicate values, as tickets were often shared among passengers.

**Timeliness**

* Not relevant in this dataset as it is historical Titanic passenger data.

**Data Cleansing**

In Order to Clean the data, I decided on the following actions;

* Age: Missing values were filled using the average age based the other records and values were rounded to the nearest whole year.
* Fare: The single missing value was replaced with the median fare and all Values were amended to 2 decimal places.
* Cabin data was removed due to excessive missing values (327 out of 418).
* Checked for Duplicate Tickets: Found 55 duplicate Ticket values, which suggests that some passengers shared tickets.

A Screenshot of the Cleaned Data can be seen at Fig 2.

**Database Schema Design**

After the Data Was cleansed, I decided upon the Schema Design as below.

* **Table**: Passengers

**Fields, Types & Constraints**:

* **PassengerId**: Primary key. Integer. Unique Identifier for Passenger
* **Name**: Passenger’s full name, non-null. String
* **Sex**: Limited to 'male' or 'female'. String
* **Age**: Integer.
* **Pclass**: Must be 1, 2, or 3. Integer. Class of Travel
* **SibSp**, **Parch**: representing family relationships. Integer.
* **Ticket**: String field.
* **Fare**: Floating-point value.
* **Embarked**: Single-character, limited to valid codes ('S', 'C', 'Q'). Code for Departure port
* **Survived**: indicating survival. Boolean Limited to (1,0)

To Create the Database & Table I used SQLiteonline.com as my DDL. Fig 3 Below shows the code I used to create the Table. Once the Table was created, I then used the “insert into” command to populate the database table. This can be seen in Fig 4 Below.

After the table had been populated, I then used the “Select \*” Command to confirm that the data had been populated Correctly. This code and the Populated Data can be seen in Fig 5 Below.

Fig 1: Raw Data

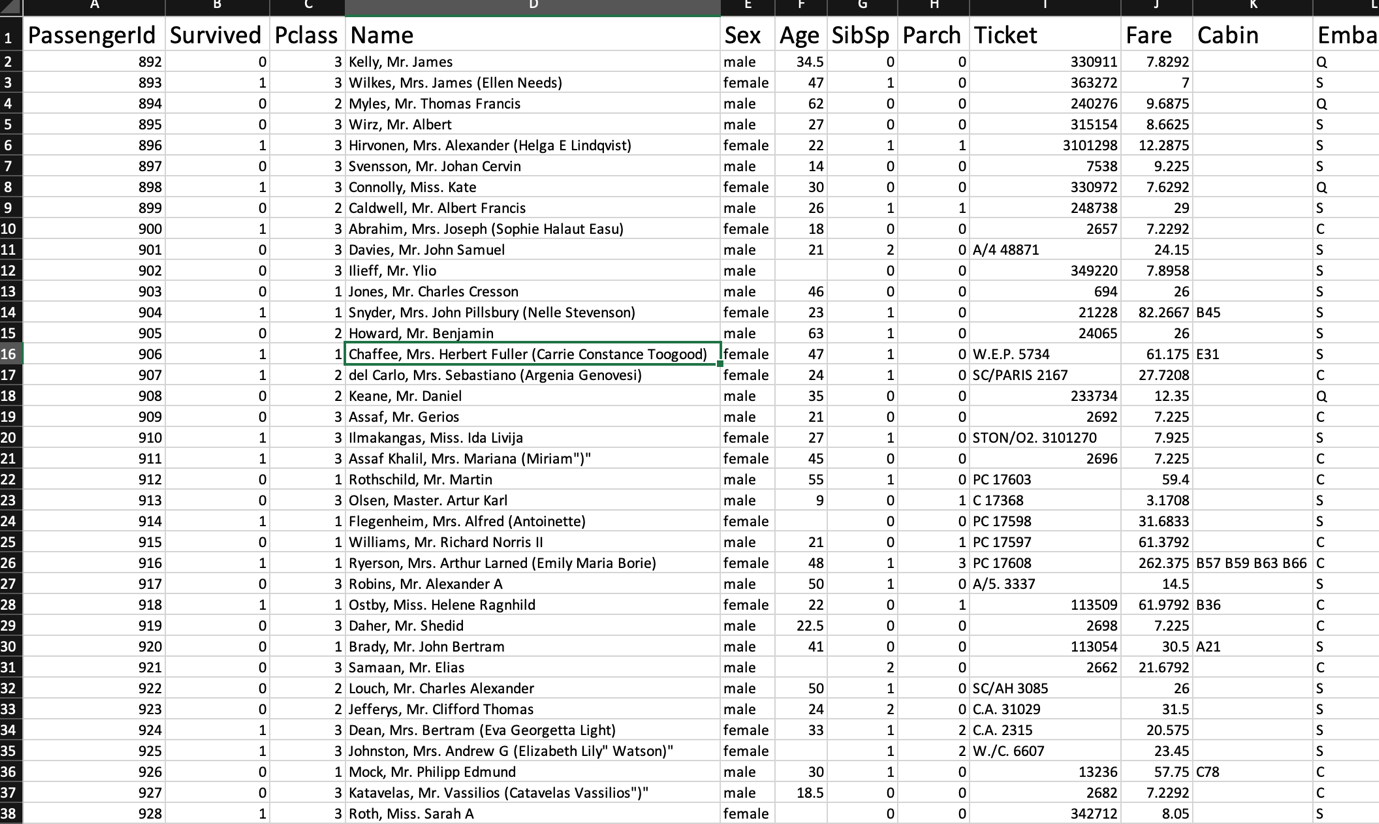


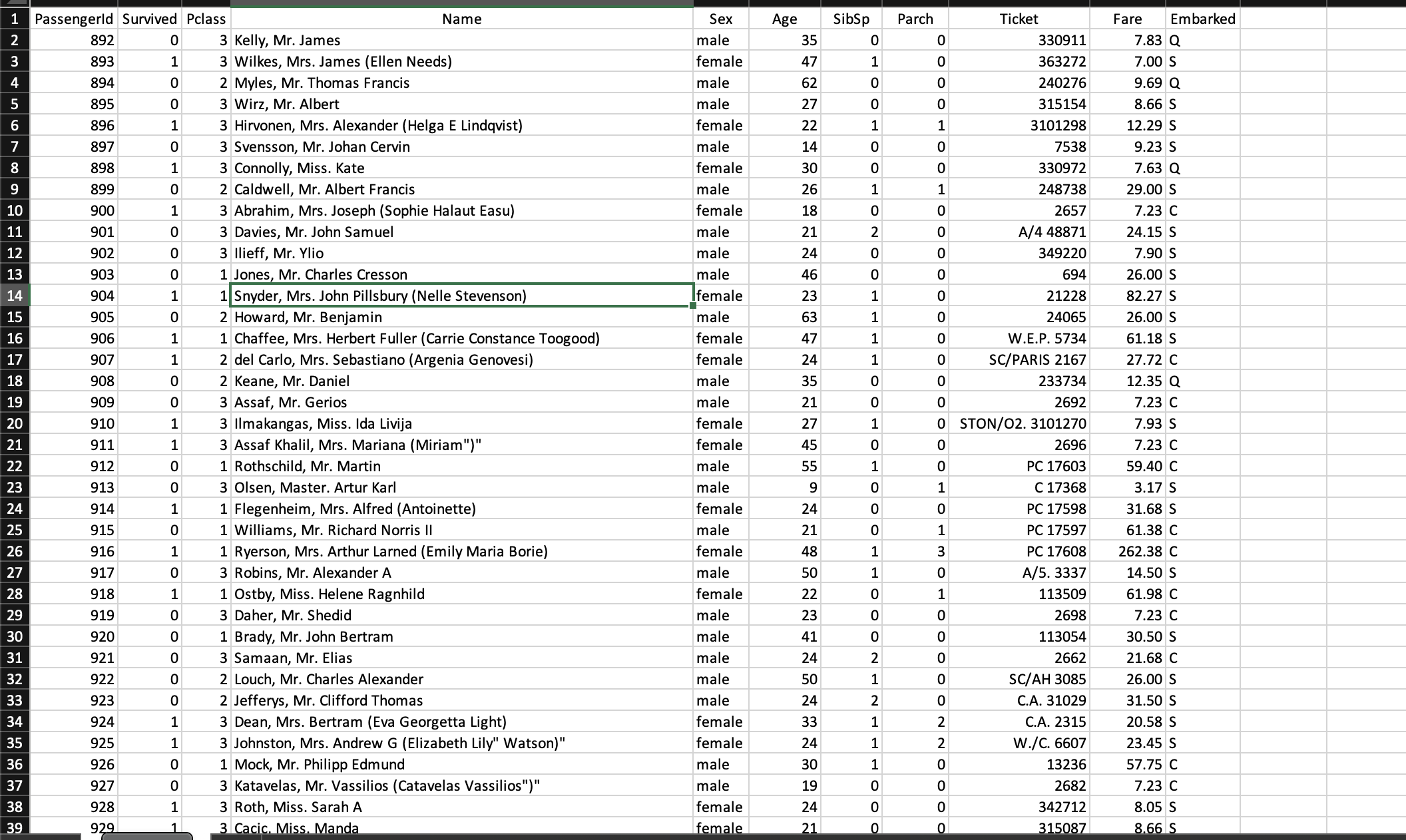
Fig 2: Cleaned Data

Fig 3:Schema creation

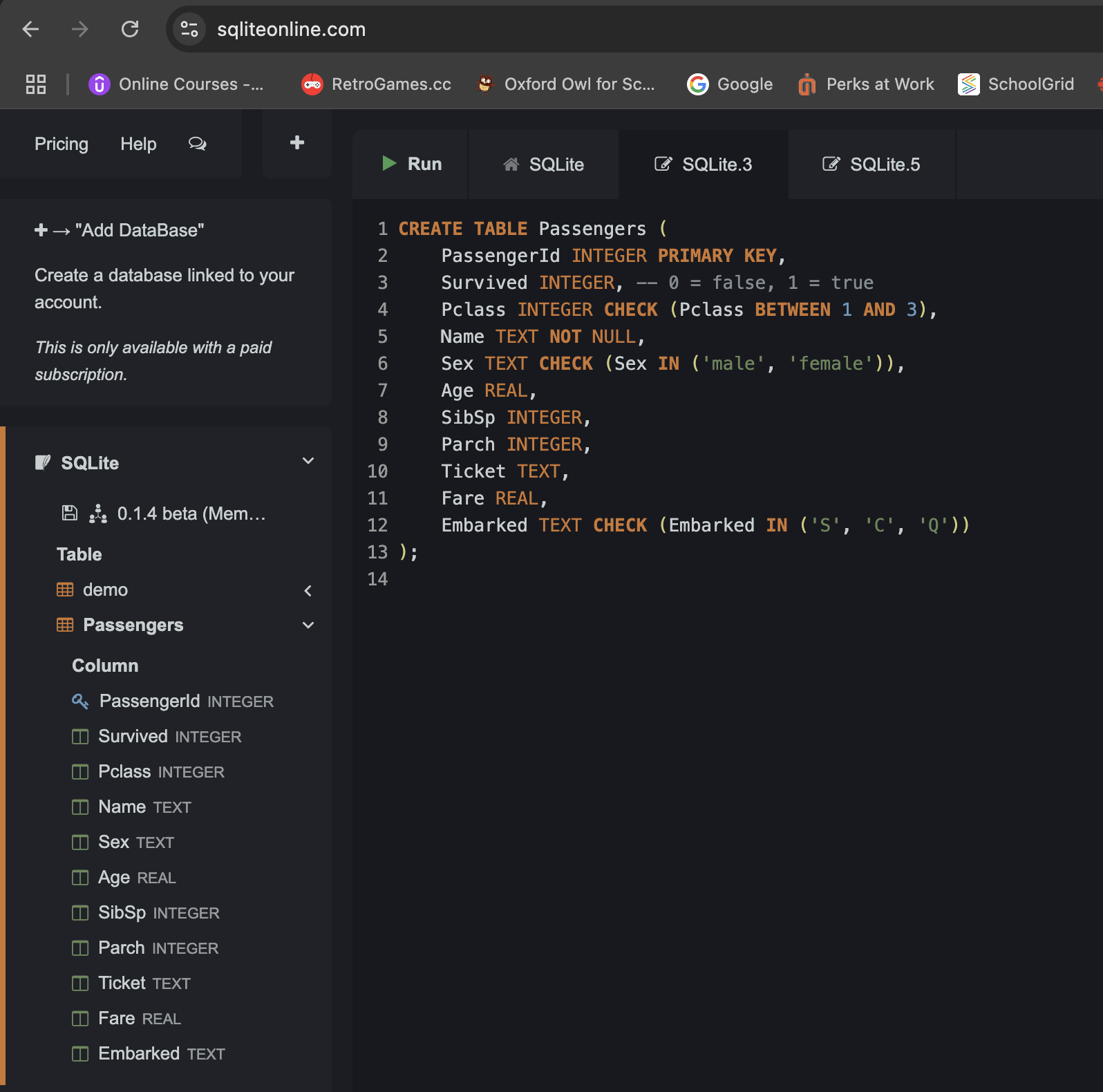


Fig 4:Data population

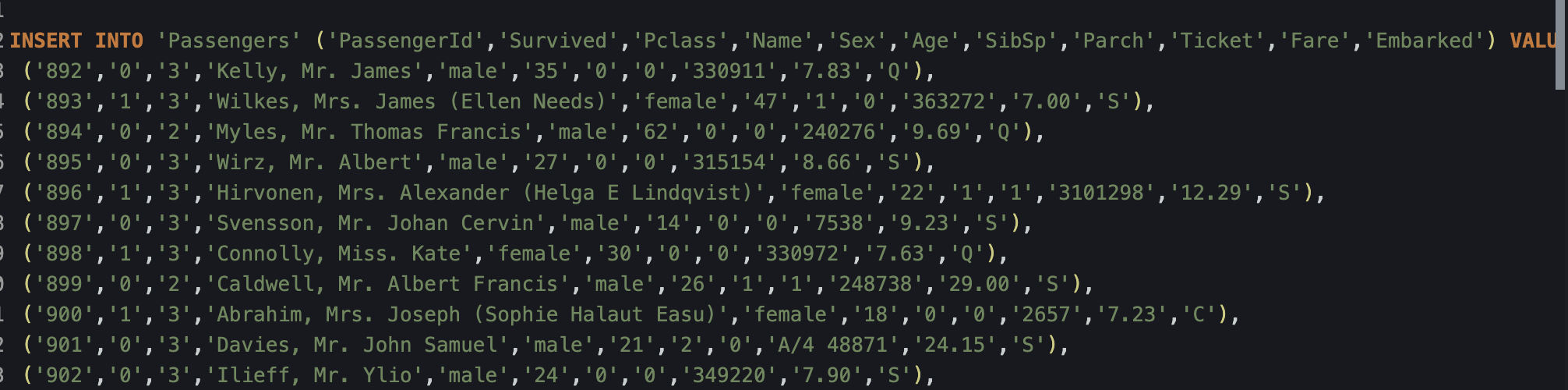


Fig 5: Check of Data Poplulation

