**Credit Card Fraud Detection**

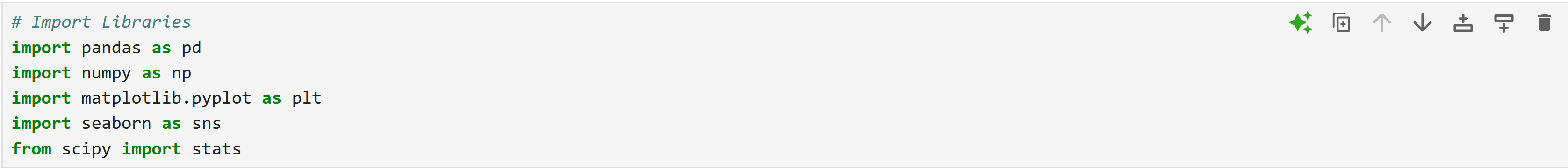
This dataset simulates credit card transactions, including both legitimate and fraudulent activities, recorded between January 1, 2019, and December 31, 2020. It includes transactions from 1,000 customers across a network of 800 merchants. The data is divided into two subsets: Train and Test, both of which share the same column structure, allowing them to be seamlessly merged into a single dataset. Below is a brief description of each column.

* **Unnamed: 0** – An index column, likely unnecessary for analysis.
* **trans\_date\_trans\_time** – The date and time of the transaction.
* **cc\_num** – The credit card number used for the transaction.
* **merchant** – The merchant where the transaction occurred.
* **category** – The category of the transaction (e.g., travel, personal care, health).
* **amt** – The transaction amount in USD.
* **first** – The first name of the cardholder.
* **last** – The last name of the cardholder.
* **gender** – The gender of the cardholder.
* **street** – The street address of the cardholder.
* **city** – The city of the cardholder.
* **state** – The state where the cardholder resides.
* **zip** – The ZIP code of the cardholder.
* **lat** – The latitude of the cardholder’s home location.
* **long** – The longitude of the cardholder’s home location.
* **city\_pop** – The population of the city where the cardholder lives.
* **job** – The job title of the cardholder.
* **dob** – The date of birth of the cardholder.
* **trans\_num** – A unique identifier for each transaction.
* **unix\_time** – The timestamp of the transaction in Unix format.
* **merch\_lat** – The latitude of the merchant’s location.
* **merch\_long** – The longitude of the merchant’s location.
* **is\_fraud** – A binary indicator (0 = not fraud, 1 = fraud) for fraudulent transactions.

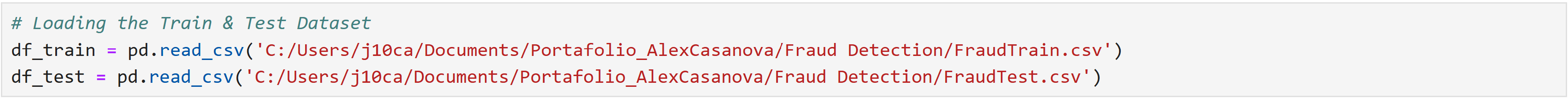
First, the two datasets are imported and merged into a single, unified dataset. Then, data cleaning and preprocessing are performed to ensure consistency and improve interpretability, providing a clearer understanding of each variable. Additionally, necessary transformations are applied to structure the data more effectively and enhance its analytical value.

**Import & Clean.**

Import libraries.

****

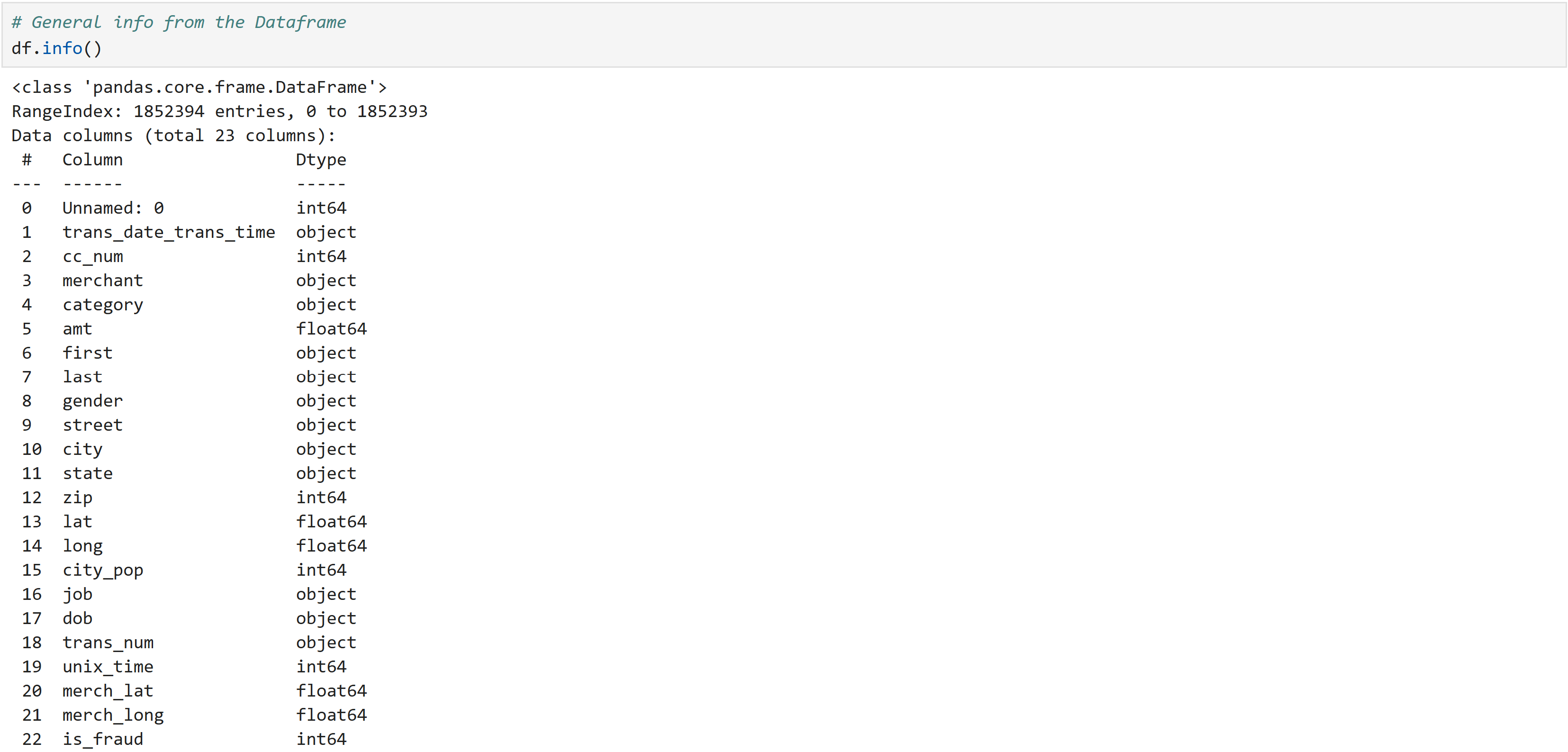
Load the train and test dataset for this fraud dataset.



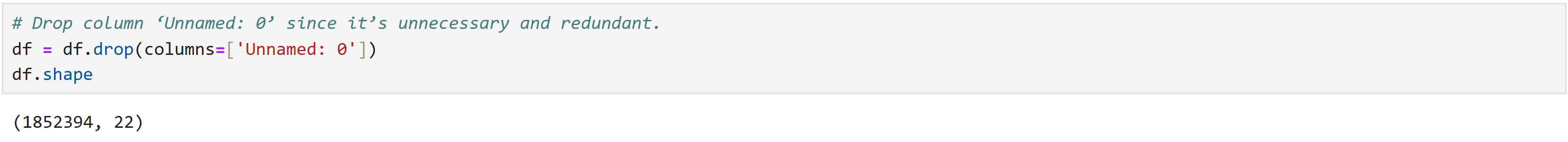
Merge the two datasets together and confirm the shape.



Let’s review the general information from the data frame ‘**df**’.



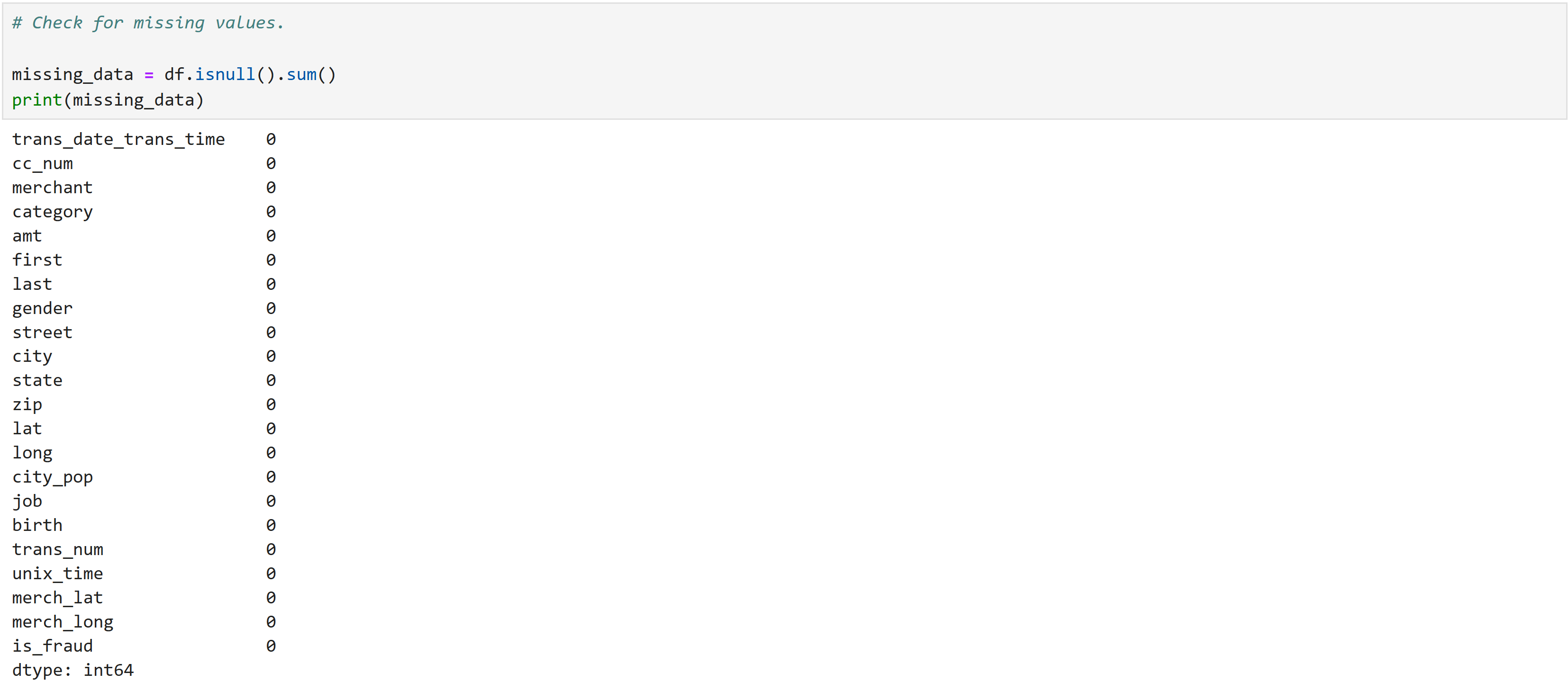
Drop column ‘**Unnamed: 0**’ since it’s unnecessary and redundant.

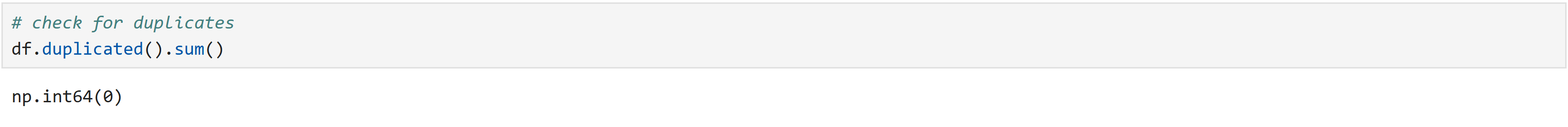


Rename column ‘**dob**’ (Date of Birth) to ‘**birth**’.



Let’s confirm if our data frame ‘**df**’ requires extra steps due to missing values and duplicates.



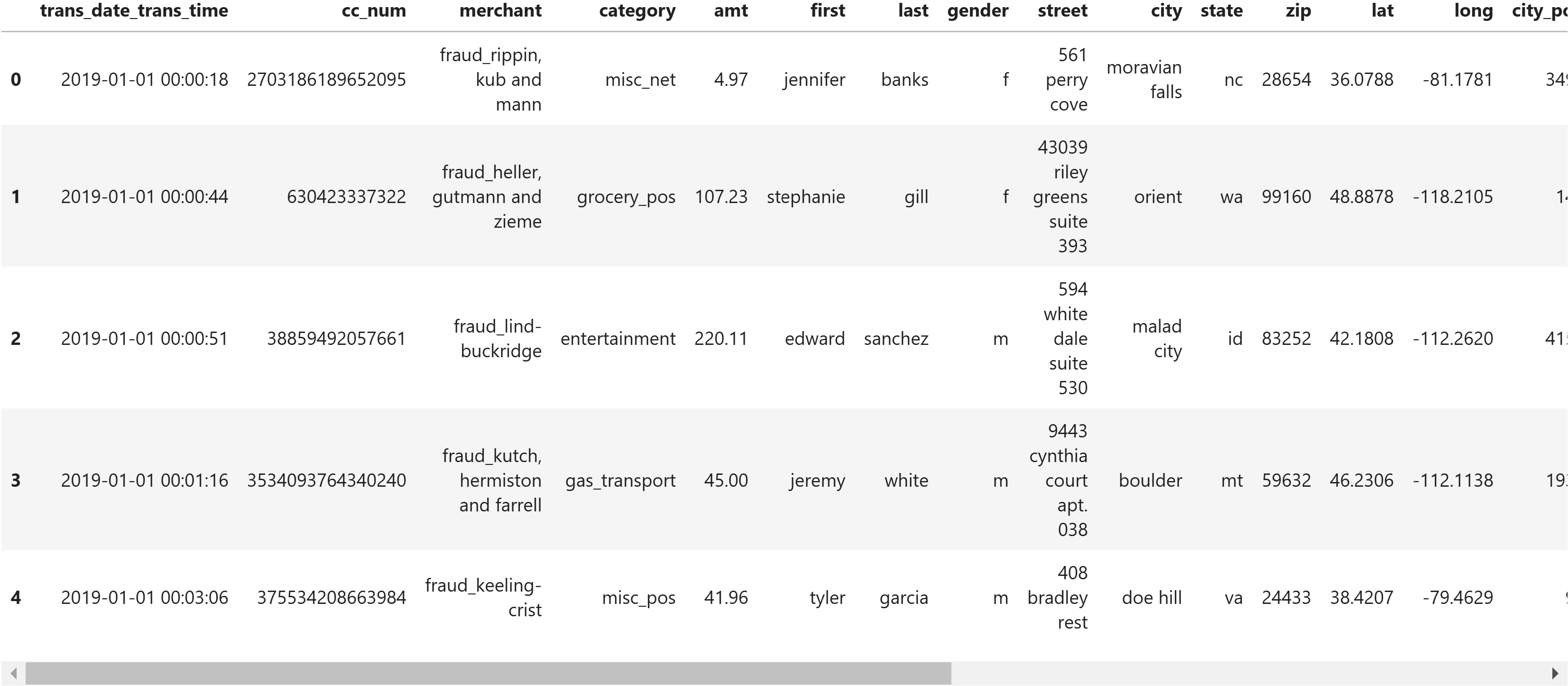


**Column Transformations.**

Now that we’ve cleaned up the column titles and removed unnecessary columns, we can take a closer look at the columns we plan to analyze.

Let’s start by converting categorical columns to lowercase and strip spaces, this will ensure data standardization by preventing variations in capitalization and spaces from being treated as different values (e.g., "Amazon " vs. "amazon"). This transformation improves data quality by reducing inconsistencies in category names and facilitates searches and analysis by enabling accurate comparisons without format discrepancies.



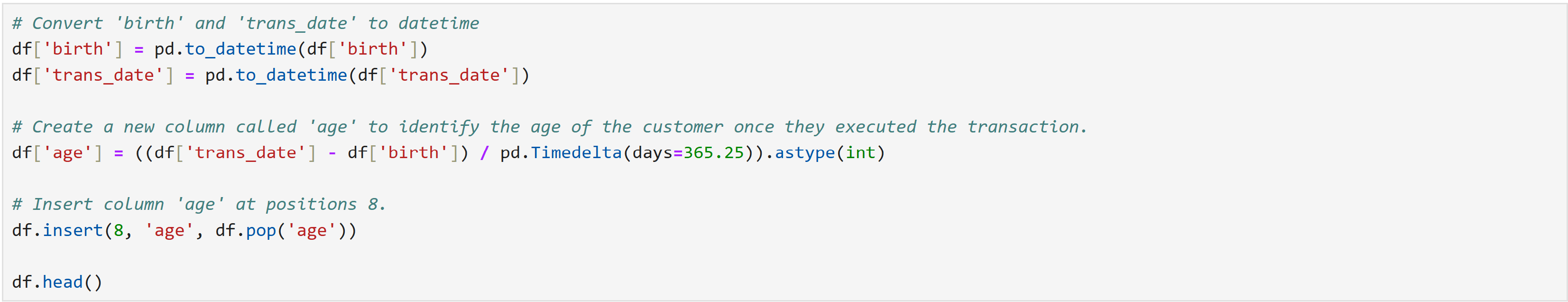


Now, we can split the ‘**trans\_date\_trans\_time**’ column into two columns. One column called ‘**trans\_date**’ and a second one called ‘**trans\_time**’.





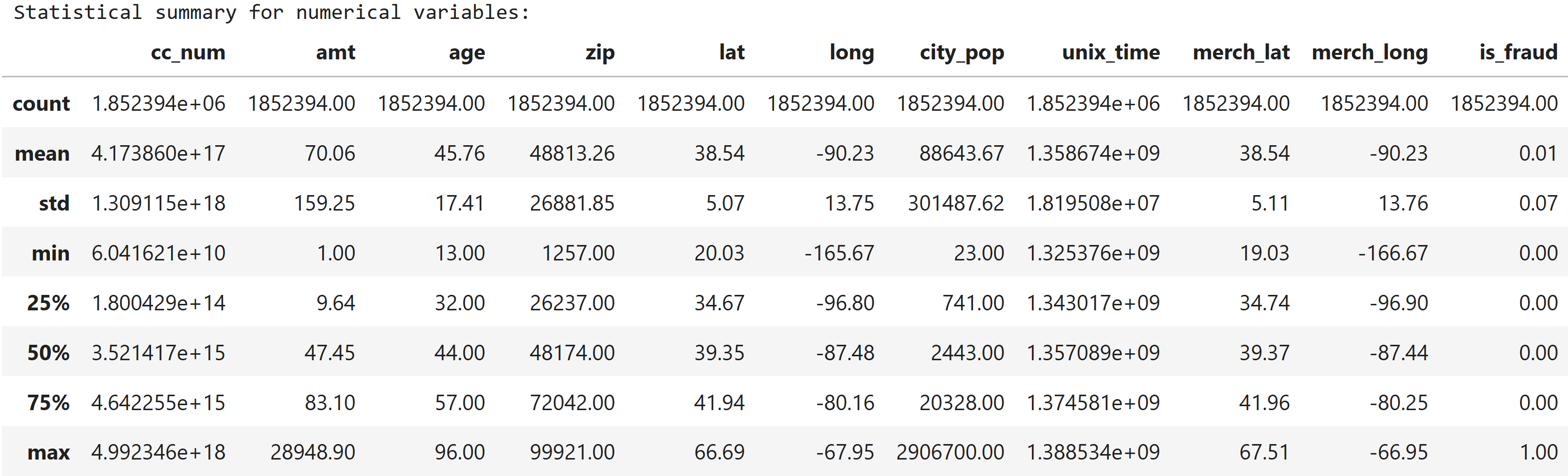
With new column ‘**trans\_date**’ and column ‘**birth**’, we can determine the age of each customer when they executed the transaction. First, we need to convert **'birth**' and '**trans\_date**' to datetime then we can create our new column called ‘**age**’.





In the final steps of our data cleaning and transformation project, we will generate a statistical summary of the numerical columns in the DataFrame ‘**df**’. This analysis will provide deeper insights into the data, helping us better understand its characteristics before further analysis.





Summary statistics of numerical columns:

* The average transaction is **$70.06**, but there's a large spread with a maximum transaction of **$28,948.90**.
* The average cardholder age is **45.76 years**, ranging from **13 to 96 years**.
* The median city population is **88,643.67**, indicating that many transactions happen in smaller towns.
* Fraud cases are quite rare, with a mean value of **0.01**, indicating only about **1%** of transactions are fraudulent.