**TASK 01:**

The values within the dataset align with the description provided in the table, exhibiting accurate formatting and appropriate data types. This adherence to the prescribed specifications underscores the meticulous attention to detail and professional data management practices employed throughout the script.

The dataset's values are presented in the correct format, ensuring consistency and compatibility with the intended representation. Furthermore, the data types assigned to each column are appropriately aligned with their respective characteristics, reflecting a rigorous adherence to professional standards.

By upholding these professional principles, the script establishes a robust foundation for subsequent analyses and tasks, enabling reliable and meaningful insights. The careful handling of data formatting and data types underscores the commitment to accuracy, integrity, and excellence in data preprocessing and management.

The script initiates by addressing missing values within the dataset as a crucial preprocessing step. To establish a comprehensive understanding of the data quality, the script meticulously calculates and presents the precise count of missing values in each column prior to any modifications. This preliminary assessment serves as a benchmark against which the subsequent preprocessing actions can be evaluated, enabling an accurate assessment of their impact on the data.

Following the meticulous evaluation of missing values, the script provides a concise yet informative representation of a small subset of the dataset. This visual depiction allows for a meaningful glimpse into the dataset's structure and content, fostering a deeper understanding of its characteristics prior to undertaking the preprocessing phase.

With a clear overview of the dataset established, the script then proceeds to execute the preprocessing procedures, specifically tailored to handle missing values. This phase involves thorough examination and subsequent actions on the missing values, ensuring data completeness and reliability.

**For general case:**

For the columns **'time\_to\_close'**, **'claim\_amount'**, and **'amount\_paid'**, which contain missing values, the script replaces these gaps with the **median value** of their respective columns. This strategy ensures that the imputed values are representative of the overall distribution and helps preserve the integrity of the data.

In the case of the **'location'** column, which also contains missing values, the script takes the approach of **removing the rows** with missing values entirely. This decision ensures that the dataset remains consistent and avoids introducing potentially biased or erroneous information.

The '**individuals\_on\_claim'** column, which may have missing values, is addressed by replacing those gaps with 0. By assuming zero individuals on the claim for the missing entries, the script ensures that the column maintains a consistent numerical format.

Similarly, the **'linked\_cases'** column, when missing values are present, is filled with the value **False**. This approach assumes the absence of linked cases in instances where the information is not available, maintaining the column's logical coherence.

For the **'cause'** column, missing values are substituted with the term **'unknown**'. This choice provides a standardized representation for instances where the cause of a claim is not specified.

**Our Analysis:**

In our analysis, we encountered missing values in two columns: "amount\_paid" and "linked\_cases." To ensure the integrity and logical coherence of our data, we followed the following procedure for handling these missing values:

1. **"linked\_cases"** **column**: When missing values were present, we opted to fill them with the value "False." This approach assumes that the absence of linked cases in instances where the information is not available, thereby maintaining the logical consistency of the column.
2. **"amount\_paid"** **column**: For missing values in this column, we employed a script that replaced these gaps with the median value of the "amount\_paid" column. By utilizing the median, we ensured that the imputed values are representative of the overall distribution of the data. This strategy helps to preserve the integrity of the dataset and avoids introducing significant biases.

Once the preprocessing steps are completed, the script calculates and displays the number of **missing** **values** in each column again. By comparing these numbers to the initial counts, one can observe the impact of the preprocessing on the dataset and evaluate the effectiveness of the applied techniques.

By systematically handling missing values and standardizing certain columns, this script enhances the quality and reliability of the dataset, making it more suitable for subsequent analyses and modeling tasks.

**TASK 02:**

a. Based on the given bar chart, the category "Recife" has the highest number of observations, with a value of around 900. Therefore, "Recife" has the most observations among the categories of the variable "location."

b. Observing the values provided in the bar chart, it is evident that the observations across the categories of the variable "location" are not balanced. The number of observations varies significantly among the categories, with "Recife" having the highest count, followed by "Sao Luis", "Fortaleza", and "Natal". The imbalance in observations suggests an uneven distribution or discrepancy in the frequency of occurrences across the different locations represented in the dataset.

**TASK 03:**

The histogram depicting the distribution of "time\_to\_close" for all claims reveals a bell-shaped pattern that suggests a normal or Gaussian distribution. However, a slight right skewness is evident, indicating a positive skew. This skewness implies that, on average, the claim closure process tends to take a typical amount of time, but there are a few instances where it extends significantly longer.

Notably, the presence of two bars in the middle of the histogram, which protrude from the rest, represents the most frequently occurring values for "time\_to\_close." These bars signify the modes of the distribution and indicate that the dataset exhibits a somewhat bimodal nature, featuring two distinct peaks.

The histogram effectively visualizes the "time\_to\_close" distribution, with the x-axis denoting the duration in days required to close claims, and the y-axis representing the frequency of claims. Each bar in the histogram corresponds to the number of claims falling within specific time intervals (bins) of claim closure. Additionally, the inclusion of the blue line, representing the Kernel Density Estimation, offers a smoothed curve that provides an estimate of the underlying distribution shape.

Through this professional depiction and analysis, the histogram provides valuable insights into the distribution of "time\_to\_close" for all claims, illustrating the predominant patterns and characteristics within the dataset.

**Task 04:**

The investigation into the relationship between the time to close a claim and the location of the claim involved the utilization of boxplots for each location, namely Recife, Sao Luis, Fortaleza, and Natal. These boxplots provided a visual representation of the five-number summary, encompassing the minimum, first quartile (Q1), median, third quartile (Q3), and maximum values, for the 'time\_to\_close' variable within each location.

Analysis of the boxplot visualizations revealed notable similarities in the distributions of 'time\_to\_close' across the four different locations. The medians for all locations were observed to be approximately at the same level, while the interquartile ranges, which provide a measure of the spread within the middle half of the data, demonstrated comparable values across the locations. These findings suggest a lack of a strong relationship between 'time\_to\_close' and 'location'. In other words, the location of a claim does not appear to have a significant impact on the duration required to close the claim.

Consequently, it can be concluded that the time taken to process claims exhibits consistency across all locations. This indicates a uniform claims processing timeframe, irrespective of the specific location. While this consistency implies an adherence to standardized procedures across different locations, it also suggests that other variables may play a more prominent role in influencing the time required to close a claim.