Chicago Dataset Statistics

The Chicago Dataset is analysed and visualised using a variety of data points, all of which are included in this report's analytical summary. It is vital to comprehend the structure and quality of the dataset in order to utilise it efficiently inside the business model.   
Over 26,791 observations, 17 variables are included in the dataset. This size suggests a rather extensive set of data points appropriate for in-depth examination.

**Data Quality:** Data Quality: Approximately 1.8% of the dataset is comprised of 8,399 missing cells. This indicates a high-quality dataset in terms of completeness, as there is a very small amount of missing data. Interestingly, there aren't any duplicate rows, indicating that each observation's uniqueness was preserved during the data collecting procedure.

**Memory Usage:** Memory Usage: The dataset has an average record size of 144.0 B and a total size of 3.7 MiB in memory. These numbers imply that most contemporary computer systems should be able to handle the dataset without the requirement for specialised data processing methods. Five of the variables are numeric, six are text, five are categorical, and one is datetime. Due to the diversity of data formats, several data pretreatment techniques will be required to guarantee that the dataset is prepared appropriately for any processing or analysis.

**Possible Consequences:** It is unlikely that the small proportion of missing data will have a major effect on the reliability of machine learning models or statistical analysis. Nevertheless, depending on the analysis's needs, attention should still be made to deal with these missing values effectively, either via imputation or exclusion. It is advantageous because there are no duplicate rows, which lessens the requirement for initial data cleansing.  
It is advised to look into the pattern of the missing data before moving on with the analysis to see if it is a random omission or if there is a systematic problem that needs to be fixed. If appropriate, think about filling in the missing data with imputation techniques. Furthermore, confirm that each variable's data type corresponds to the format needed for the planned analysis.

**ETL:**

**Conclusion:** In summary, the dataset seems to be of excellent quality, with few missing values and no duplicate entries. Because of its moderate size, analysis on standard computing systems shouldn't encounter any major difficulties. The dataset is ready for a comprehensive examination or usage in machine learning applications after the few missing values have been handled correctly.