IMPUTATION AND CLEANING

1. Kept only P value
2. Outlier amount was removed

MERCH STATE:

In the first step, missing values are filled by mapping "Merch zip" with "state_id" from the "zip_codes" dataframe using a dictionary. In the second step, missing values are filled by creating a dictionary of "Merchnum" and the mode value of the "Merch state" for that "Merchnum". The mode value is the value that appears most frequently in the "Merch state" column for a given "Merchnum". The same approach is used in the third and fourth steps, where missing values are filled using dictionaries of "Merch description" and "Cardnum", respectively.

MERCH ZIP:

For "Merch zip", the imputation process involves two steps. In the first step, missing values are filled by creating a dictionary with "Merch state" and mode values of zip code. In the second step, missing values are filled by creating a dictionary with "Merch description" and mode values of zip code. Additionally, any "Merch zip" values that correspond to adjustment transactions ("RETAIL CREDIT ADJUSTMENT" and "RETAIL DEBIT ADJUSTMENT") are filled with "Unknown" using the mask() method.

MERCH NUM:

For "Merch num", the imputation process involves four steps, like that of "Merch state". The first step matches "Merch state" values with zip codes by creating a dictionary. The second step fills "Merch state" values by creating a dictionary of "Merchnum" and the mode value of "Merch state". The same approach is used in the third and fourth steps, where missing values are filled using dictionaries of "Merch description" and "Cardnum", respectively.

Finally, any remaining missing values are filled with "Unknown". Additionally, outlier values are removed from the dataset.

Overall, this imputation and cleaning process helps ensure that the dataset is completer and more accurate, which can improve the quality of any analyses or models that are performed using the data.