Report

By: Tyler McElroy

The purpose of this project is to identify anomalies from transaction amounts in the dataset titled "Financial Anomalies." Within the dataset there are 216,960 records with transaction dates between 1/1/2023 and 5/31/2023. For the transaction amounts, there are 481 records that do not have values, representing 0.22% of the population; as such, I removed them since the missing values are immaterial to the analysis I am performing.

The summary statistics for the data can be found listed below.

count	216960.000000
mean	50090.025108
std	29097.905016
min	10.510000
25%	25061.242500
50%	50183.980000
75%	75080.460000
max	978942.260000

The summary statistics listed above indicate a slight rightward skew in the data. The value returned from the "skew" function is 0.4, which confirms there is a slight rightward skew. Note: the skew function comes from the scipy.stats library.

Since there is only a slight rightward skew in the data, it is appropriate to use z-score for identifying outliers. Z-scores are used to indicate how many standard deviations a data point is away from the mean of a distribution. Typically, values beyond a z-score threshold of 2 or 3 are considered outliers. For this analysis, a z-score threshold of 3 is used to identify outliers.

After implementing logic to obtain z-scores for each transaction, I filtered the data to return only the transactions that are outliers – that is, transactions that have a z-score above 3. The result contains 11 records (See: notebook for individual records).

In the financial services industry, it is common to investigate outliers as part of an ongoing monitoring plan. At a real company, these transactions may be forwarded to a team for review and adjudication. Performing due diligence on transactions aids in identifying fraud, data quality issues, and ensuring compliance with various regulations.

Please refer to the Jupyer Notebook file to view the code and output referenced in this report.