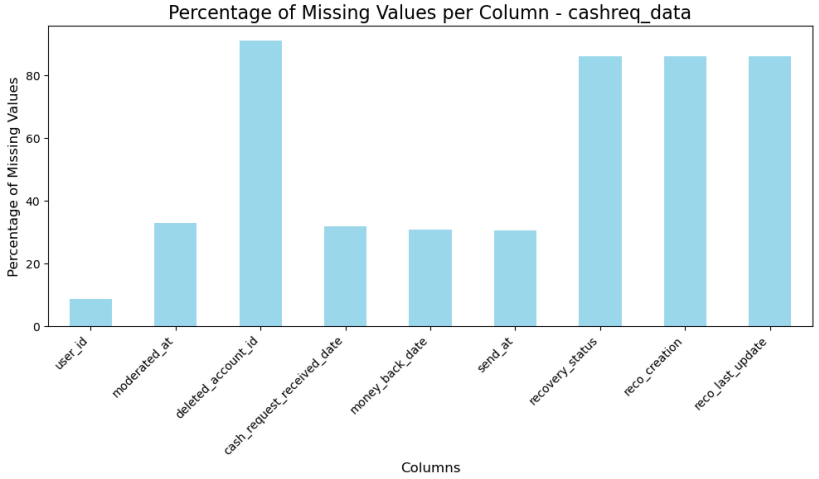
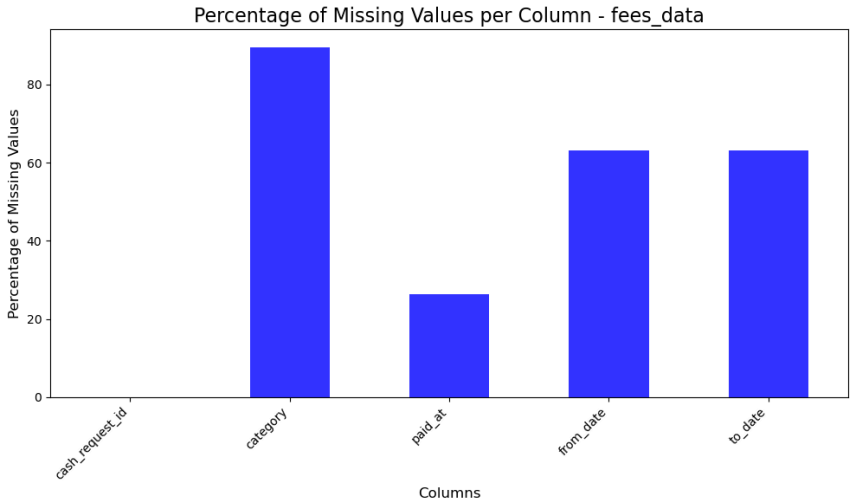
Data Quality Analysis Report:

1. Assessing Data Quality

a. Missing Values

- When reviewing the data to identify missing values, we discovered several columns with a significant amount of missing data. In the Cash Request dataframe, the following columns contained considerable missing data: deleted_accouny_id, recovery_status, reco_creation, and reco_last update. Similarly, in the Fees dataframe, some columns exhibited substantial missing data.
- Upon analyzing each column, we observed that some missing data resulted from conditions set by other variables. For example, the missing data in deleted_account_id was linked to the valid data on user_id. Also, the recovery columns with high missing data in Cash Request data are dependent on payment incidents. We understand that the missing data was not an issue at the moment of analyzing the data.
- Important columns were checked to verify no null rows would affect the analysis.



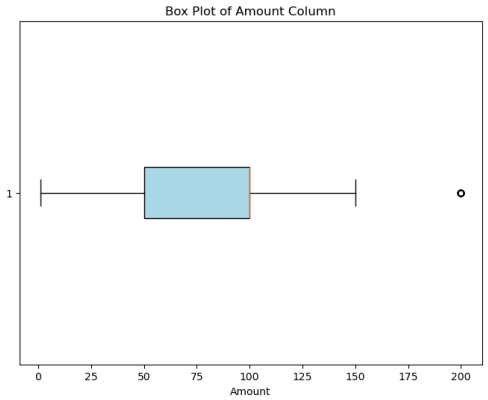


b. Data Inconsistencies

- There were no rows in the id columns that were duplicated.
- The data types of columns were checked and we found some needed type updating.
- There were some outliers, but we do not think they represented a problem for the data analysis. For example, we analyzed outliers in the amount column. If we removed the outliers there would be no significant change in the metrics.

Metrics Before Removing Outliers:
Mean: 82.72081768877764
Median: 100.0
Standard Deviation: 26.528065010796876
Min: 1.0
Max: 200.0
Count: 23970

Metrics After Removing Outliers:
Mean: 82.59837126748799
Median: 100.0
Standard Deviation: 26.269697346628416
Min: 1.0
Max: 150.0
Count: 23945



c. Potential Errors

- There was data inconsistent with the creation date of Ironhack Payments. The data that was dated before 2020, was deleted as it was not accurate and would damage the data analysis.
- Other values checked:
  - Columns were checked to ensure values fall within expected ranges.
  - We checked for logical inconsistencies to ensure related columns had consistent values.

2. Data Cleaning and Preprocessing

a. Handling Missing or inconsistent Values

- The data types were updated as needed. Some columns were set to string, while others were set to floats.
- In the created_at column, we dropped the data that was dated before 2020, as it was not accurate and would damage the data analysis.

b. Handling Outliers

- Outliers were analyzed. We agreed there was no need to manipulate data related to outliers.

3. Recommendations

Based on the data quality analysis we recommend Ironhack Payments to:

- Expand data collection efforts to include detailed user engagement metrics, transactions history, fees structures and external factors to accurately assess cohort performance.
- Document formulas to calculate the kpi's.