**Data Quality Analysis Report - Overview of "extract - cash request - data analyst.csv"**

The dataset consists of 23,970 rows and 16 columns, with various data types including integers, floats, and objects (strings/dates). The columns represent financial transaction information related to cash requests.

**2. Identifying Data Quality Issues**

* **Missing Values**:
  + **user\_id**: 1,103 missing values (approximately 4.6% of the data).
  + **moderated\_at**: 7,935 missing values (approximately 33.2% of the data).
  + **deleted\_account\_id**: 22,866 missing values (approximately 95.6% of the data).
  + **cash\_request\_received\_date**: 7,681 missing values (approximately 32% of the data).
  + **money\_back\_date**: 7,427 missing values (approximately 31% of the data).
  + **send\_at**: 7,329 missing values (approximately 30.5% of the data).
  + **recovery\_status**: 20,640 missing values (approximately 86.2% of the data).
  + **reco\_creation**: 20,640 missing values (approximately 86.2% of the data).
  + **reco\_last\_update**: 20,640 missing values (approximately 86.2% of the data).
* **Inconsistent Data**:
  + **Data Types**: The columns created\_at, updated\_at, moderated\_at, reimbursement\_date, cash\_request\_received\_date, money\_back\_date, send\_at, reco\_creation, and reco\_last\_update are stored as objects (strings), but they should ideally be in datetime format for proper analysis.
* **Potential Errors**:
  + **Deleted Account**: The deleted\_account\_id has mostly missing values, but some entries may indicate missing relationships or account deletions.
  + **Recovery Columns**: The recovery-related columns (recovery\_status, reco\_creation, reco\_last\_update) are highly sparse, potentially indicating a lack of relevant data or incomplete records.

**3. Data Cleaning summary**

We can take the following steps to address the issues mentioned:

* **Missing Values**: Imputed missing values for user\_id with the median and for deleted\_account\_id with 0 (indicating no deleted account). Forward-filled date columns with missing values (e.g., cash\_request\_received\_date, money\_back\_date, send\_at).
* **Datetime Conversion**: Converted object columns representing dates to datetime format for accurate time-based analysis.
* **Dropped Sparse Columns**: Removed columns with over 85% missing data (recovery\_status, reco\_creation, reco\_last\_update), as they would not provide significant insights.

**Data Quality Analysis Report- Overview of “**extract - fees - data analyst - .csv”

The data consists of multiple columns, with information related to cash requests, statuses, payment types, amounts, and timestamps. The dataset has missing values, especially in category, from\_date, and to\_date, and some fields need to be converted to proper data types (e.g., datetime for timestamp fields). There are no duplicate records, and the total\_amount is fairly uniform, with values primarily being 5 or 10.

**1. Missing Values**

The dataset contains several missing values across different columns:

* **cash\_request\_id**: 4 missing values.
* **category**: 18,865 missing values.
* **paid\_at**: 5,530 missing values.
* **from\_date and to\_date**: Both columns have 13,295 missing values.

**2. Data Types**

* **cash\_request\_id**: This is stored as a float64, but it should be an int64. This could be due to the presence of missing values.
* **created\_at, updated\_at, and paid\_at**: These are stored as object (string) types. They should be converted to datetime .
* **from\_date and to\_date**: Also stored as object types, they should ideally be in datetime format .
* **total\_amount**: This column is stored as float64, for numerical analysis.
* **category**, **reason**, **charge\_moment**, **type**, and **status**: These are all stored as object types, for categorical data.

**3. Duplicates**

* There are no duplicate records based on the combination of id and cash\_request\_id.

**4. Unique Values in Categorical Columns**

* **type**: Values include "instant\_payment", "incident", and "postpone". These appear consistent but might require checking if "postpone" is used correctly.
* **status**: The unique values are "rejected", "accepted", "cancelled", and "confirmed". This column is consistent.
* **charge\_moment**: The values are "after" and "before". These are consistent but could require validation based on the use case.

**5. Outliers in total\_amount**

* **total\_amount**: The values range from 5 to 10, with all data points being either 5 or 10, making this column fairly uniform. The mean is 5, with a small standard deviation of 0.03, suggesting very few deviations.

**Data Cleaning Summary**

1. **Handling Missing Values**:
   * **cash\_request\_id**: Rows with missing values in this column were **dropped** to ensure that all requests have a valid identifier.
   * **category**: Missing values were **replaced** with "Unknown" to ensure no gaps in this column.
   * **from\_date and to\_date**: These columns had **excessive missing values** (over 13,000 missing), so they were **dropped** from the dataset as they appeared irrelevant.
   * **paid\_at**: The missing values in this column remain, indicating that some payments are still pending.
2. **Converting Data Types**:
   * **Timestamps** (created\_at, updated\_at, and paid\_at) were **converted to datetime** format to facilitate time-based operations and analysis.
   * **cash\_request\_id** was **converted to Int64** to align with its role as an integer identifier.
3. **Handling Outliers**:
   * **total\_amount**: There were no outliers. The values in this column were found to be consistently either 5 or 10, which is expected based on the dataset's nature.
4. **Ensuring Consistency in Categorical Columns**:
   * **type, status, and charge\_moment**: All categorical columns were **standardized to lowercase** to avoid inconsistencies like "Accepted" vs. "accepted" or "After" vs. "after".