**1. Initial Setup**

* Confirm understanding of project objectives:
  + Analyze user behavior and performance of financial services.
  + Focus on cohorts based on the month of their first cash advance.
* Gather required data from Ironhack Payments.
* Review the data-sharing protocols and permissions.

**2. Data Quality and Preprocessing**

* Conduct a **Data Quality Analysis**:
  + Check for missing values, inconsistencies, and duplicates.
  + Document data quality issues and the steps to resolve them.
* Clean the dataset:
  + Remove or address inconsistencies.
  + Handle missing values appropriately.
  + Verify accuracy of date fields and transaction values.
* Create derived columns for cohort analysis:
  + First cash advance month.
  + Service usage frequency metrics.
* Ensure data format is compatible with analysis tools (e.g., Python, Tableau).

**3. Exploratory Data Analysis (EDA)**

* Perform descriptive statistics on key metrics (e.g., usage frequency, revenue).
* Visualize data:
  + Monthly trends for each cohort.
  + Distribution of incident rates.
  + Revenue patterns over time.
* Identify outliers and trends that need further exploration.
* Document findings in an EDA report, including key patterns and any hypotheses.

**4. Cohort Analysis Execution**

* Define cohorts by their first cash advance month.
* Calculate:
  + **Frequency of Service Usage**: Monthly activity rates per cohort.
  + **Incident Rates**: Variability in payment incidents.
  + **Cohort Revenue**: Total revenue generated per cohort over months.
  + **Proposed New Metric**: Brainstorm, implement, and calculate (e.g., customer retention rates or average revenue per user).
* Visualize metrics over time for each cohort:
  + Line charts or heatmaps for service usage trends.
  + Revenue graphs by cohort.

**5. Reporting and Presentation**

* Prepare the following deliverables:
  + **Python Code**: Ensure it includes data loading, cleaning, cohort creation, metric calculation, and visualization.
  + **Tableau Dashboard**: Interactive views for the Ironhack Payments team.
  + **EDA Report**: Include visualizations and insights.
  + **Data Quality Report**: Document identified issues and resolutions.
  + **Presentation Slides** (maximum 4 slides):
    - Key findings and insights from cohort analysis.
    - Recommendations based on the findings.

**6. Review and Iteration**

* Share findings and visualizations with stakeholders for feedback.
* Iterate on analysis or visualizations based on feedback.

**7. Bonus (Optional Enhancements)**

* Create a **Streamlit App** for dynamic data interaction.
* Implement an operationalized .py script:
  + Include clear documentation and execution instructions.
  + Test script functionality in a terminal environment.
* Convert code to use an alternate strategy (e.g., object-oriented if the current is functional, or vice versa).

**8. Submission and Handover**

* Verify all deliverables are completed and uploaded.
* Prepare final documentation for the Ironhack Payments team.

EDA Notes.

Cash-request

================ DATA INFO ================

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 23970 entries, 0 to 23969

Data columns (total 16 columns):

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Column** | **Non-Null Count** | **Dtype** | **Missing Count** | **Missing %** |
| id | 23,970 | int64 | - | 0 |
| amount | 23,970 | float64 | - | 0 |
| status | 23,970 | object | - | 0 |
| created\_at | 23,970 | object | - | 0 |
| updated\_at | 23,970 | object | - | 0 |
| user\_id | 21,867 | float64 | 2,103 | 8.8 |
| moderated\_at | 16,035 | object | 7,935 | 33.1 |
| deleted\_account\_id | 2,104 | float64 | 21,866 | 91.2 |
| reimbursement\_date | 23,970 | object | - | 0.0 |
| cash\_request\_received\_date | 16,289 | object | 7,681 | 32.0 |
| money\_back\_date | 16,543 | object | 7,427 | 31.0 |
| transfer\_type | 23,970 | object | - | 0.0 |
| send\_at | 16,641 | object | 7,329 | 30.6 |
| recovery\_status | 3,330 | object | 20,640 | 86.1 |
| reco\_creation | 3,330 | object | 20,640 | 86.1 |
| reco\_last\_update | 3,330 | object | 20,640 | 86.1 |

================ DUPLICATE ROWS ================

Total duplicate rows: 0

================ COLUMN STATISTICS ================

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Metric** | **id** | **amount** | **user\_id** | **deleted\_account\_id** |
| **Count** | 23970 | 23970 | 21867 | 2104 |
| **Mean** | 13910.966 | 82.721 | 32581.251 | 9658.755 |
| **Std. Dev.** | 7788.117 | 26.528 | 27618.566 | 7972.743 |
| **Min** | 3 | 1 | 34 | 91 |
| **25th Percentile (Q1)** | 7427.25 | 50 | 10804 | 3767 |
| **Median (Q2)** | 14270.5 | 100 | 23773 | 6121.5 |
| **75th Percentile (Q3)** | 20607.75 | 100 | 46965 | 16345 |
| **Max** | 27010 | 200 | 103719 | 30445 |

|  |  |  |
| --- | --- | --- |
| **Column** | **Skewness** | **Kurtosis** |
| **id** | -0.11 | -1.16 |
| **amount** | -0.92 | -0.22 |
| **user\_id** | 0.94 | -0.22 |
| **deleted\_account\_id** | 0.86 | -0.68 |

Categorical Summary:

|  |  |
| --- | --- |
| **Status Category** | **Count** |
| **money\_back** | 16397 |
| **rejected** | 6568 |
| **direct\_debit\_rejected** | 831 |
| **active** | 59 |
| **transaction\_declined** | 48 |
| **direct\_debit\_sent** | 34 |
| **canceled** | 33 |

EDA Notes.

fees

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 id 21061 non-null int64

1 cash\_request\_id 21057 non-null float64

2 type 21061 non-null object

3 status 21061 non-null object

4 category 2196 non-null object

5 total\_amount 21061 non-null float64

6 reason 21061 non-null object

7 created\_at 21061 non-null datetime64[ns, UTC]

8 updated\_at 21061 non-null datetime64[ns, UTC]

9 paid\_at 15438 non-null datetime64[ns, UTC]

10 from\_date 6749 non-null datetime64[ns, UTC]

11 to\_date 6512 non-null datetime64[ns, UTC]

12 charge\_moment 21061 non-null object

================ MISSING DATA ================

Missing Count Missing %

id 0 0.000000

cash\_request\_id 4 0.018992

type 0 0.000000

status 0 0.000000

category 18865 89.573145

total\_amount 0 0.000000

reason 0 0.000000

created\_at 0 0.000000

updated\_at 0 0.000000

paid\_at 5623 26.698637

from\_date 14312 67.954988

to\_date 14549 69.080291

charge\_moment 0 0.000000

================ DUPLICATE ROWS ================

Total duplicate rows: 0

================ COLUMN STATISTICS ================

Numerical Summary:

id cash\_request\_id total\_amount

count 21061.000000 21057.000000 21061.000000

mean 10645.355111 16318.449162 5.000237

std 6099.315256 6656.149949 0.034453

min 1.000000 1456.000000 5.000000

25% 5385.000000 11745.000000 5.000000

50% 10652.000000 17160.000000 5.000000

75% 15925.000000 21796.000000 5.000000

max 21193.000000 27010.000000 10.000000

Skewness and Kurtosis of Numerical Columns:

id: Skewness=-0.01, Kurtosis=-1.20

cash\_request\_id: Skewness=-0.40, Kurtosis=-0.78

total\_amount: Skewness=145.12, Kurtosis=21061.00

Categorical Summary:

Column: type

type

instant\_payment 11099

postpone 7766

incident 2196

Name: count, dtype: int64

Column: status

status

accepted 14841

cancelled 4938

rejected 1194

confirmed 88

Name: count, dtype: int64

Column: category

category

rejected\_direct\_debit 1599

month\_delay\_on\_payment 597

Name: count, dtype: int64

Column: reason

reason

rejected direct debit 1599

month delay on payment - 9/2020 283

month delay on payment - 8/2020 203

month delay on payment - 10/2020 102

Postpone Cash Request 12225 34

...

Instant Payment Cash Request 21481 1

Instant Payment Cash Request 21140 1

Instant Payment Cash Request 25010 1

Postpone Cash Request 16749 1

Instant Payment Cash Request 20982 1

Name: count, Length: 15149, dtype: int64

Column: charge\_moment

charge\_moment

after 16724

before 4337

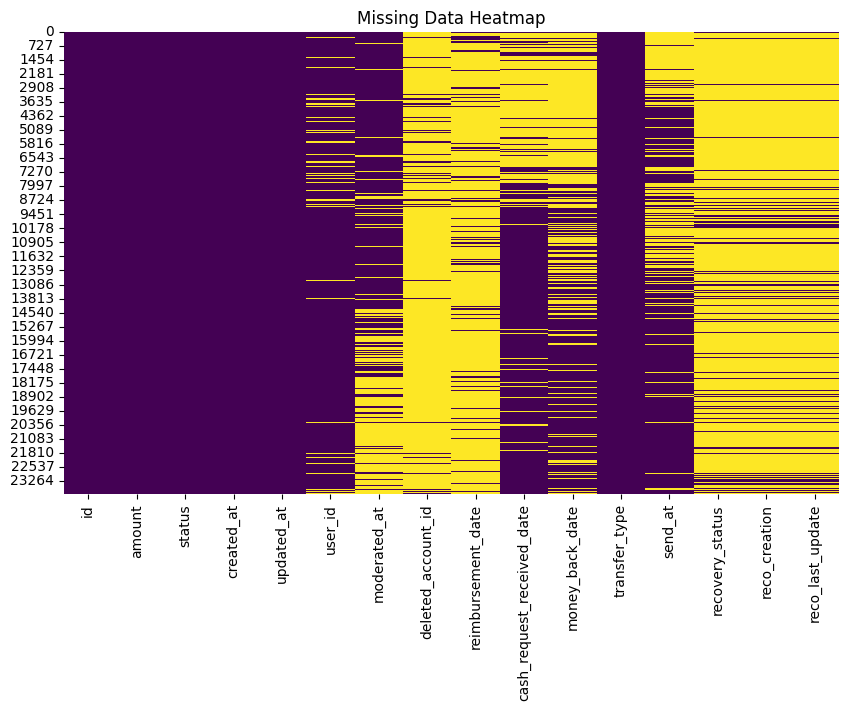
Name: count, dtype: int64

================ DATE VERIFICATION ================

* + Missing column ‘reason’ in cash request file.
  + Date format **ISO 8601**

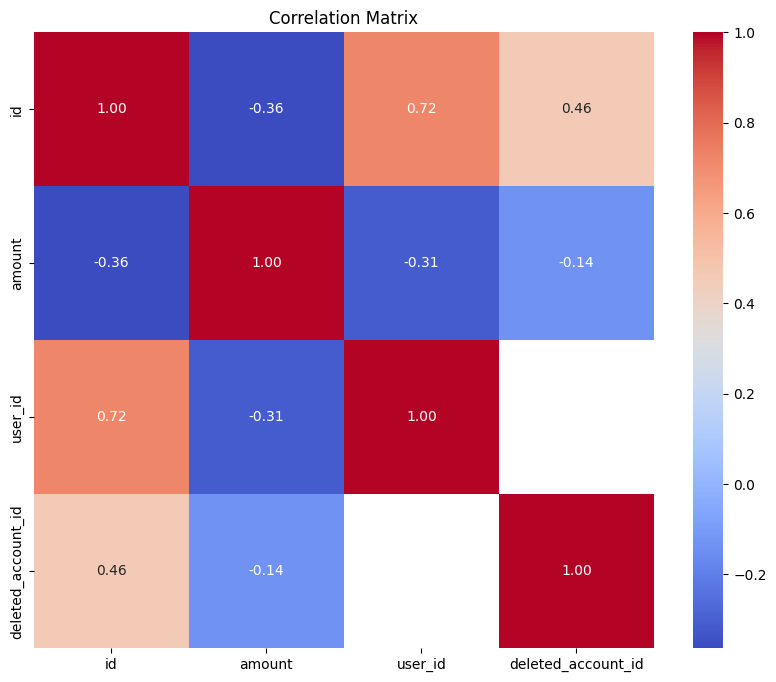
### ****Structure of the Format****

1. **Date**: YYYY-MM-DD (e.g., 2019-12-10)
   * YYYY: Year
   * MM: Month
   * DD: Day
2. **Time**: hh:mm:ss.sssssss+zz:zz
   * hh: Hours (24-hour format)
   * mm: Minutes
   * ss.sssssss: Seconds (with microseconds or milliseconds precision)
   * +zz:zz: Timezone offset from UTC (e.g., +00:00 means UTC)



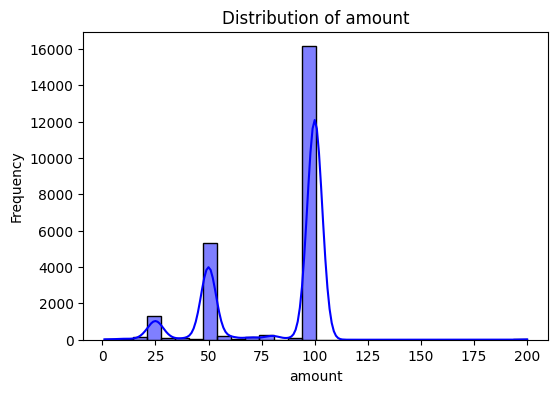
**user\_id, reimbursement\_date, deleted\_account\_id have significant missing data**

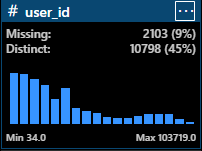
id and amount are almost entirely complete

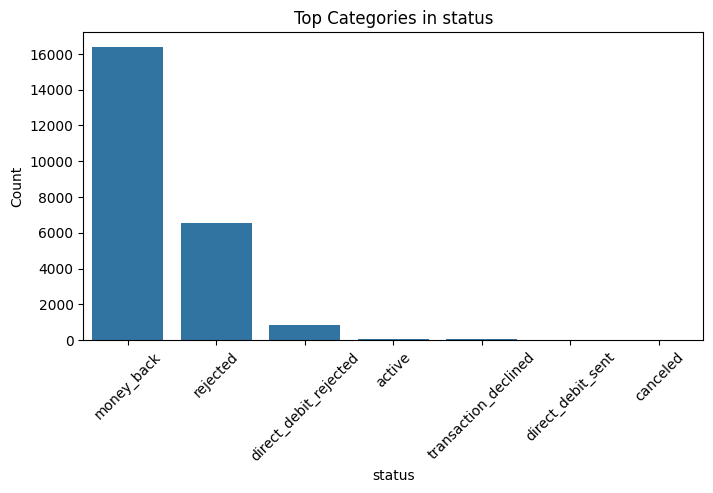


**id and user\_id**:

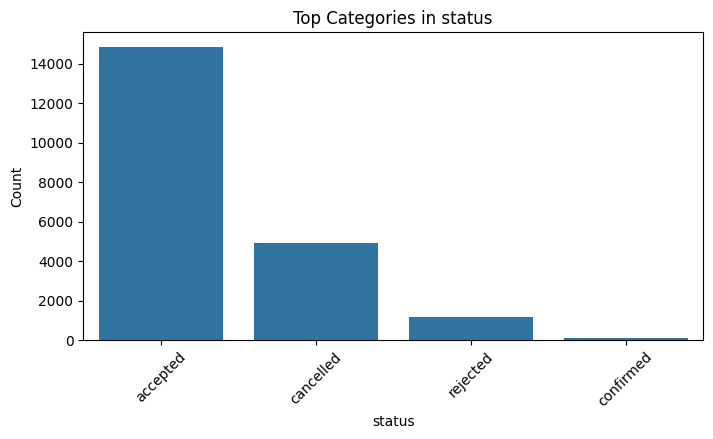
* Correlation = **0.72**: A strong positive correlation, suggesting that the id and user\_id are likely related or follow a sequential pattern.

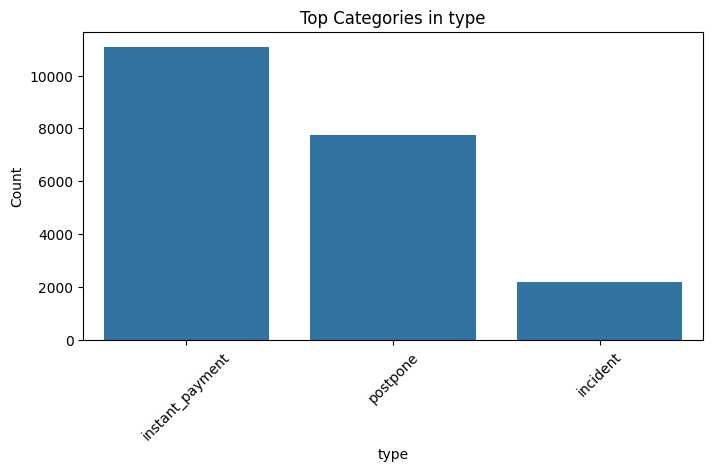






Fees





Cohorts

A graph of a bar graph

Description automatically generated with medium confidence

A graph of a number of red bars

Description automatically generated

A graph of a number of red bars

Description automatically generated with medium confidence