Inhaltsverzeichnis

[EDA 2](#_Toc202192309)

[1. Introduction 2](#_Toc202192310)

[Project Objective 2](#_Toc202192311)

[2. Data Source Description 2](#_Toc202192312)

[3. Data Overview: 3](#_Toc202192313)

[Data Structure ofFrom the data set: extract - fees - data analyst.csv 3](#_Toc202192314)

[Data Structure of from the data set: extract - cash request - data analyst 5](#_Toc202192315)

[4. Summary Statistics (Mean, Median, Mode) 7](#_Toc202192316)

[Statistics **of** From the data set: extract - fees - data analyst 7](#_Toc202192317)

[Statistics from the data set: extract - cash request - data analyst 8](#_Toc202192318)

[5. Data Visualization: 10](#_Toc202192319)

[Histograms / Bar charts for distributions 10](#_Toc202192320)

[Data set: extract - fees - data analyst 10](#_Toc202192321)

[Data set: extract - cash request - data analyst 13](#_Toc202192322)

[Scatter plots for relationships 15](#_Toc202192323)

[Data set: extract - fees - data analyst 15](#_Toc202192324)

[Data set: extract - cash request – data analyst 17](#_Toc202192325)

[Box plots for outliers 18](#_Toc202192326)

[Data set: extract - fees - data analyst 18](#_Toc202192327)

[Data set: extract - cash request – data analyst 18](#_Toc202192328)

[6. Feature Analysis: 19](#_Toc202192329)

[Key Features Investigation 19](#_Toc202192330)

[Correlations 22](#_Toc202192331)

[7. Initial Insights: 27](#_Toc202192332)

[Patterns and Trends 27](#_Toc202192333)

[Anomalies and Outliers 30](#_Toc202192334)

[8. Next Steps: 32](#_Toc202192335)

[Hypotheses for Further Analysis 32](#_Toc202192336)

[Additional Data Needed 33](#_Toc202192337)

# EDA

## 1. Introduction

## Project Objective

The primary aim of this project is to conduct a cohort analysis for IronHack Payments. We will analyze user cohorts defined by the month in which they created their first cash advance. Our objective is to track the monthly evolution of key metrics for these cohorts. The ultimate goal is to gain valuable insights into user behavior and assess the performance of IronHack's financial services.

**Key Performance Indicators (KPIs)**

1. **Frequency of Service Usage**: Analyze usage patterns to understand how often users from each cohort utilize IronHack Payments' cash advance services over time.
2. **Incident Rate**: Determine the rate of payment incidents for each cohort, focusing on variations in incident rates among different cohorts.
3. **Revenue Generated by the Cohort**: Assess the financial impact of user behavior by calculating the total revenue generated by each cohort over time.
4. **New Relevant Metric**: Develop and calculate a new metric that offers additional insights into user behavior or the performance of IronHack Payments' services.

## 2. Data Source Description

The analysis leverages data from three key files that provide comprehensive information on fees and cash requests, along with additional descriptive context:

1. **Extract Fees Data Analyst CSV**:

The "Extract Fees Data Analyst CSV" file provides comprehensive information regarding the various fee types associated with IronHack Payments. It details each fee's unique identifier, its type—ranging from instant payment fees to those incurred for failed reimbursements or postponed transactions—and the status, whether confirmed, rejected, or accepted. Additionally, it explains the reasons for fee application, such as rejected direct debits or delays in payment, alongside timestamps for fee creation, updates, and payments. The file also includes information on the amount charged, the specific cash request linked to each fee, and the timing of when fees are to be charged, either before or after a cash request's reimbursement. This data is pivotal for analyzing the **financial operations** **and user interactions** within IronHack's financial services, offering insights into **incident** management and **service usage** **patterns**.

1. **Extract Cash Request Data Analyst**:

The "Extract Cash Request Data Analyst CSV" file provides detailed insights into each cash request processed by IronHack Payments. Each entry includes a unique identifier, the requested amount, and the current status of the cash request—ranging from approved to rejected, pending manual review, or awaiting user confirmation. It captures essential workflow timestamps, such as creation, updates, and manual review times, offering a chronological trail of the cash request’s lifecycle. User-specific details, such as the requester’s ID and recovery status, help trace payment incidents and resolutions. The file also records relevant dates like the planned reimbursement and when funds were sent or received, providing a comprehensive view of financial transactions. Additional fields like "transfer\_type" and "recovery\_status" offer context on how requests are managed and resolved, illustrating the system's adherence to regulatory guidelines like GDPR for user data management and fraud prevention. This information is crucial for understanding and **analyzing user behaviors**, **transaction flows**, and **incident management** within IronHack Payments' financial ecosystem.

1. **Lexique - Data Analyst**: This document contains two informative tabs:
   * The first tab explains the context around fees, offering descriptions of categories and parameters.
   * The second tab provides context for cash requests, detailing the associated categories and parameters.

## Data Overview:

* + Data Structure (Rows, Columns, Data Types)

### Data Structure ofFrom the data set: extract - fees - data analyst.csv

* **Dataset Shape**: The dataset comprises **21,061 rows** and **13 columns**.
* **Columns and Data Types**:
  + **id**: int64
  + **cash\_request\_id**: float64
  + **type**: object
  + **status**: object
  + **category**: object
  + **total\_amount**: float64
  + **reason**: object
  + **created\_at**: object
  + **updated\_at**: object
  + **paid\_at**: object
  + **from\_date**: object
  + **to\_date**: object
  + **charge\_moment**: object

#### Overview of Missing Data

**Null Values**:

* **cash\_request\_id**:
  + **Null Values**: 4
  + **Impact**: The missing data is negligible for the overall analysis. However, these gaps could represent missing linkages between fees and cash requests, which the company should investigate further.
* **category**:
  + **Null Values**: 18,865
  + **Impact**: The large number of missing values implies that categorical analysis might not be reliable unless these are addressed. Contextually, a missing category often indicates that no incident fees were applicable since **category** describes the reason for the incident fee. The two possible values (**rejected\_direct\_debit**, **month\_delay\_on\_payment**) focus on specific incident descriptions.
* **paid\_at, from\_date, to\_date**:
  + **Null Values**: **paid\_at** - 5,530; **from\_date** and **to\_date** - 13,295 each
  + **Impact**: While these columns have a considerable number of missing entries, their absence might reflect the non-applicability of certain processes (e.g., postponement of fees). Given the context, these can often be ignored; however, it's crucial to evaluate whether their missingness affects any time-based analyses or KPIs related to cash advance timelines.

### Data Structure of from the data set: extract - cash request - data analyst

The dataset "Extract Cash Request Data Analyst CSV" provides valuable insights into IronHack's cash request processes. This comprehensive dataset contains **23,970 rows** and **16 columns**.

* **Columns and Data Types**:
  + **id**: **int64** - Unique identifier for each cash request.
  + **amount**: **float64** - The requested cash amount.
  + **status**: **object** - Current status of the cash request (e.g., approved, rejected).
  + **created\_at**: **object** - Timestamp for when the cash request was created.
  + **updated\_at**: **object** - Timestamp for the last update to the cash request.
  + **user\_id**: **float64** - Identifier for the user initiating the cash request. (2,103 missing values)
  + **moderated\_at**: **object** - Timestamp for manual review, if undertaken. (7,935 missing)
  + **deleted\_account\_id**: **float64** - Substitute ID used when an account is deleted due to GDPR compliance. (21,866 missing)
  + **reimbursement\_date**: **object** - Scheduled date for reimbursement.
  + **cash\_request\_received\_date**: **object** - Date the cash request is recorded as received. (7,681 missing)
  + **money\_back\_date**: **object** - Date when the amount is seen as returned. (7,427 missing)
  + **transfer\_type**: **object** - Denotes if the transaction is instant or regular.
  + **send\_at**: **object** - Date funds were transferred to the requester. (7,329 missing)
  + **recovery\_status**: **object** - Status of payment incidents, showing if there's an unresolved issue, or it’s been resolved. (20,640 missing)
  + **reco\_creation**: **object** - Timestamp for when recovery actions started. (20,640 missing)
  + **reco\_last\_update**: **object** - Last update on recovery status. (20,640 missing)

#### Summary of Missing Values

The dataset contains significant missing values in several columns, which necessitates careful consideration during analysis:

* **user\_id**: 2,103 missing values, potentially, the user who requested the cash advance, so 2,103 dont want that advnace method
* **moderated\_at**: 7,935 missing values, indicative of cases not subjected to manual review.
* **deleted\_account\_id**: 21,866 missing values, reflecting active user accounts.
* **cash\_request\_received\_date**: 7,681 missing values, highlighting delays or processing discrepancies.
* **money\_back\_date**: 7,427 missing values, often pending further financial processing or resolution.
* **send\_at**: 7,329 missing values, indicating pending fund transfers.
* **recovery\_status**, **reco\_creation**, **reco\_last\_update**: Each missing nearly 20,640 entries, suggesting that payment incidents were not common or these steps were not needed.

🡺 Thus the missing data is negligible for the overall analysis.

However, in the data set: extract - cash request – data analyst “**reason” the**  object: is Missing.

## Summary Statistics (Mean, Median, Mode)

### Statistics **of** From the data set: extract - fees - data analyst

1. **Count:** Are the number of non-null entries in the **total\_amount** column, which is 21,061. It matches the number of rows in your dataset, indicating no missing values in this column.
2. **Mean (Average):** The average value of **total\_amount** is approximately 5.000237. This suggests that the typical cash request amount revolves around this value.
3. **Standard Deviation (std):** This measures the variation or spread of the **total\_amount** values. A std of 0.034453 indicates very little variation around the mean. In simpler terms, most of the **total\_amount** values are very similar to each other.
4. **Minimum (min):** The smallest value in the **total\_amount** column is 5.000000. This is likely indicative of a base or default fee.
5. **25th Percentile (25%):** 25% of the entries have a **total\_amount** value of 5.000000 or less. This suggests that a large portion of the fees are at this base level.
6. **50th Percentile (Median - 50%):** Half of the entries have a **total\_amount** value of 5.000000 or less. This value being the same as the 25th percentile indicates a concentration of values at this point.
7. **75th Percentile (75%):** 75% of the entries have a **total\_amount** value of 5.000000 or less. This reinforces the observation about the concentration at the minimal value.
8. **Maximum (max):** The highest value in the **total\_amount** column is 10.000000. This suggests that the typical range of total amounts does not vary much from the base.

**Interpreting the Summary**

The **total\_amount** data tells you that most cash requests are likely for a very uniform or standard fee, often precisely the base fee of 5.000000 in this dataset. The presence of a maximum at 10.000000 suggests there might only be a few instances where the fee increased, which could be exceptional cases or a specific category.

### Statistics from the data set: extract - cash request - data analyst

id

* Count: 23,970 entries, which matches the total rows, indicating each cash request has a unique identifier.
* Mean: 13,910.966 – It's an average identifier number and doesn't usually offer much analysis value.
* Standard Deviation (std): 7,788.117 – This reflects variance in identifier distribution, again mostly useful for ensuring non-repetitive IDs.
* Min: 3.000 – The lowest id number.
* 25th Percentile (25%): 7,427.250 – 25% of IDs are less than this value.
* 50th Percentile (Median - 50%): 14,270.500 – Middle value splitting IDs in half.
* 75th Percentile (75%): 20,607.750 – 75% of IDs fall under this value.
* Max: 27,010 – The highest ID registered.

amount

* Count: 23,970, indicating no missing data for requested amounts.
* Mean: 82.720818 – This is the average requested cash amount.
* Standard Deviation (std): 26.528065 – Shows variability in requested amounts; moderate deviation suggesting some diversity in amounts.
* Min: 1.000 – Smallest cash request made.
* 25th Percentile (25%): 50.000 – 25% of requests ask for this amount or less.
* 50th Percentile (Median - 50%): 100.000 – Half of the requests are 100 or less.
* 75th Percentile (75%): 100.000 – Suggests a common cap on requests; roughly 75% have this amount or less.
* Max: 200.000 – Largest cash request recorded.

user\_id

* Count: 21,867 – There are 2,103 missing user IDs, meaning not all entries have a user\_id linked.
* Mean: 32,581.250789 – Average user ID, significant only for ensuring a non-repeating sequence and user management.
* Standard Deviation (std): 27,618.565773 – The high std indicates quite a variation in user IDs.
* Min: 34.000 – Smallest user ID.
* 25th Percentile (25%): 10,804.000 – 25% of IDs are below this value.
* 50th Percentile (Median - 50%): 23,773.000 – Splits user\_id such that 50% are higher, 50% are lower.
* 75th Percentile (75%): 46,965.000 – 75% of user IDs fall below this range.
* Max: 103,719.000 – Maximum user ID registered.

deleted\_account\_id

* Count: 2,104 entries – Reflects rare instances where accounts were deleted (due to GDPR or other reasons).
* Mean: 9,658.755228 – Average deleted account ID, mainly for administrative tracking.
* Standard Deviation (std): 7,972.743249 – Shows large variation, indicating scattered @deleted account IDs.
* Min: 91.000 – Smallest deleted account ID noted.
* 25th Percentile (25%): 3,767.000 – 25% of deleted account IDs are below this number.
* 50th Percentile (Median - 50%): 6,121.500 – Midway point for deleted IDs.
* 75th Percentile (75%): 16,345.000 – 75% of deleted account IDs fall below this value.
* Max: 30,445.000 – Highest recorded deleted account ID.

Interpretation

This summary presents a basic sense of the dataset's structure and distribution. Here's a quick interpretation:

* amount: Majority of cash requests tend to be around 100, with a max of 200. This may suggest typical limits placed on cash requests.
* user\_id: While the dataset is missing a substantial number of user IDs, it generally shows that the user IDs vary widely.
* deleted\_account\_id: Mostly corresponding to compliance or account management, these are scarce and exhibit large spacing between IDs.
* Missing user\_id: Our dataset reports 2,103 missing user\_id entries. These correspond to instances where users have opted to delete their accounts.
* Count of deleted\_account\_id: There are 2,104 entries in the deleted\_account\_id column.
* This discrepancy of one record suggests a possible mismatch, such as a deletion error or a duplicate entry, which should be thoroughly investigated during the data preparation phase. Ensuring data integrity and accuracy here is crucial to maintaining the reliability of the dataset and subsequent analyses.

## Data Visualization:

### Histograms / Bar charts for distributions

### Data set: extract - fees - data analyst

Histogram for numerical distribution with amount a column of interest:

Ein Bild, das Text, Screenshot, Diagramm, Display enthält.

Automatisch generierte Beschreibung

Bar chart for categorical distribution

Ein Bild, das Screenshot, Text, Diagramm, Rechteck enthält.

Automatisch generierte Beschreibung

Distribution of Charge Moment

Ein Bild, das Text, Screenshot, Diagramm enthält.

Automatisch generierte Beschreibung

Distribution of Type of Fee

Ein Bild, das Text, Screenshot, Diagramm, Rechteck enthält.

Automatisch generierte Beschreibung

Distribution of category of the reason of the incident fee

Ein Bild, das Text, Screenshot, Diagramm enthält.

Automatisch generierte Beschreibung

### Data set: extract - cash request - data analyst

Distribution of Amount of the Cash Request

Ein Bild, das Text, Screenshot, Diagramm, Reihe enthält.

Automatisch generierte Beschreibung

Distribution of Status CR

Ein Bild, das Screenshot, Text, Diagramm, Reihe enthält.

Automatisch generierte Beschreibung

Distribution of Status Recovery

Ein Bild, das Text, Screenshot, Rechteck, Diagramm enthält.

Automatisch generierte Beschreibung

## Scatter plots for relationships

### Data set: extract - fees - data analyst

Ein Bild, das Text, Screenshot, Rechteck, Reihe enthält.

Automatisch generierte Beschreibung Ein Bild, das Text, Screenshot, Display, Diagramm enthält.

Automatisch generierte Beschreibung

Ein Bild, das Text, Screenshot, Diagramm, Wasser enthält.

Automatisch generierte Beschreibung

### Data set: extract - cash request – data analyst

Ein Bild, das Text, Screenshot, Zahl, Reihe enthält.

Automatisch generierte BeschreibungEin Bild, das Text, Screenshot, Zahl, Reihe enthält.

Automatisch generierte Beschreibung

## Box plots for outliers

### Data set: extract - fees - data analyst

Ein Bild, das Text, Screenshot, Rechteck enthält.

Automatisch generierte Beschreibung

Ein Bild, das Text, Screenshot, Rechteck enthält.

Automatisch generierte Beschreibung

### Data set: extract - cash request – data analyst

Ein Bild, das Diagramm, Screenshot, Rechteck, Reihe enthält.

Automatisch generierte Beschreibung

## Feature Analysis:

### Key Features Investigation

#### Frequency of Service Usage CR:

* **Columns to Focus On:** **created\_at** (df1 and df2), **status** (df1 and df2).
* **Why:** The **created\_at** column helps determine how frequently the service is used over time. The **status** column helps see the disposition of these requests.

created\_month

2019-11 1

2019-12 289

2020-01 223

2020-02 184

2020-03 244

2020-04 473

2020-05 837

2020-06 2615

2020-07 3601

2020-08 3417

2020-09 4221

2020-10 7725

2020-11 140

**Distribution of statuses - df1:**

Distribution of statuses - df1

status\_distribution\_df1 = df1['status'].value\_counts()

print(status\_distribution\_df1)

#### Incident Rate:

* **Columns to Focus On:** **type** and **category** (df1), **status**, **recovery\_status** (df2).
* **Why:** Analyzing these columns helps identify the rate and type of incidents and their resolution status across cohorts.

**Incident rate for payments fees:**

category

rejected\_direct\_debit 0.728142

month\_delay\_on\_payment 0.271858

**Incident handling in CR:**

Recovery\_status

completed 0.741141

pending 0.253754

pending\_direct\_debit 0.004805

cancelled 0.000300

#### Revenue Generated by the Cohort

* **Columns to Focus On:** **total\_amount** (df1), **amount** (df2).
* **Why:** Understanding how cohorts generate revenue over time provides insights into user monetary activities.

**Revenue generated from fees:**

created\_at

2020-05 80.0

2020-06 3845.0

2020-07 8095.0

2020-08 15260.0

2020-09 22860.0

2020-10 53835.0

2020-11 1335.0

**Revenue from CR requests:**

created\_at

2019-11 1.0

2019-12 27297.0

2020-01 21587.0

2020-02 16653.0

2020-03 23549.0

2020-04 46093.0

2020-05 79236.0

2020-06 246026.0

2020-07 328187.0

2020-08 287633.0

2020-09 336507.0

2020-10 559339.0

2020-11 10710.0

#### New Relevant Metric Proposal

* **Proposed Metric:** Average Fee Per Transaction.
* **Columns to Focus On:** **total\_amount** (df1), **amount**, **status** (df2).
* **Why:** This metric helps understand the average fee applied per successful transaction, shedding light on service affordability and revenue efficiency.

Average Fee Per Transaction - df1: 5.000336904521259

Average amount per successful transaction - df2: 86.52542372881356

### Correlations

**Fees Correlation Matrix:**

id cash\_request\_id total\_amount

id 1.000000 0.884152 0.005544

cash\_request\_id 0.884152 1.000000 0.006709

total\_amount 0.005544 0.006709 1.000000

Ein Bild, das Text, Screenshot, Rechteck, Quadrat enthält.

Automatisch generierte Beschreibung

Analysis and Interpretation:

1. id and cash\_request\_id

* Correlation: 0.884152
  + Interpretation: There is a strong positive correlation between id and cash\_request\_id, implying that as one increases, the other tends to increase as well. This relationship suggests that the IDs are closely sequential or related, but since they are unique identifiers, this correlation might not hold much practical significance in terms of user behavior or financial metrics.

2. total\_amount and id

* Correlation: 0.005544
  + Interpretation: There is effectively no correlation between total\_amount and id. This indicates that the fee amount associated with each ID appears to be independent of the sequence or indexing of the requests. This makes sense given IDs are primarily identifiers and do not inherently convey financial metrics.

3. total\_amount and cash\_request\_id

* Correlation: 0.006709
  + Interpretation: Similarly, there's no meaningful correlation between total\_amount and cash\_request\_id. This further supports the idea that the fee amounts are not systematically related to the sequential or structural properties of cash request identifiers.

Objectives-Relevant Interpretation:

* Frequency of Service Usage:
  + The correlation between id and cash\_request\_id does not directly inform usage frequency. However, understanding the voluminous and sequential nature of requests (high correlation) could be hypothesis for high engagement or a procedural artifact like batch processing of requests.
* Incident Rates:
  + The low correlation values with total\_amount suggest that incident-related features may not be captured in these numeric columns. The actual incident rates should be analyzed using categorical variables like status or category combined with numerical insights.
* Revenue Generated by the Cohort:
  + total\_amount would be a key driver of revenue, but its lack of correlation with other IDs signifies that individual amounts might be distributed independently, reflecting diverse user behaviors rather than systematic differences in cohorts.
* Propose New Relevant Metric:
  + Given the lack of substantial insights from numeric correlations, a new metric might involve incorporating datetime fields or categorical attributes (e.g., rate of fee application post-creation).

This correlation matrix shows that while there are structural relationships between IDs, the financial aspects represented by total\_amount operate independently of these identifiers. The focus should be on how fees correlate with categorical features or time-based trends rather than the raw numerical correlations provided.

**CRs Correlation Matrix:**

id amount user\_id deleted\_account\_id

id 1.000000 -0.363045 0.699699 -0.040324

amount -0.363045 1.000000 -0.287400 -0.017842

user\_id 0.699699 -0.287400 1.000000 -0.249670

deleted\_account\_id -0.040324 -0.017842 -0.249670 1.000000

Ein Bild, das Text, Screenshot, Rechteck, Quadrat enthält.

Automatisch generierte Beschreibung

**Analysis and Interpretation:**

**1. id and amount**

* **Correlation: -0.363045**
  + **Interpretation:** There is a moderate negative correlation between **id** and **amount**. This suggests that as the **id** increases (possibly reflecting newer entries in your dataset), the **amount** requested tends to decrease, or vice versa. This might indicate changing user behavior or policy over time affecting the requested amounts.

**2. id and user\_id**

* **Correlation: 0.699699**
  + **Interpretation:** There is a strong positive correlation, indicating that **user\_id** tends to increase with **id**. This could suggest that both evolve together, perhaps as sequential records in the dataset. However, it's important to note that such structural correlations might not provide direct insights into behavior or value-related objectives.

**3. id and deleted\_account\_id**

* **Correlation: -0.040324**
  + **Interpretation:** This weak negative correlation suggests almost no linear relationship, meaning the deletion of accounts is likely independent of the sequential assignment of **id**.

**4. amount and user\_id**

* **Correlation: -0.287400**
  + **Interpretation:** A moderate negative correlation suggests that as user IDs increase, the amounts requested tend to decrease slightly. This could hint at existing users requesting higher amounts or newer cohorts being more conservative in their requests.

**5. amount and deleted\_account\_id**

* **Correlation: -0.017842**
  + **Interpretation:** Virtually no correlation, indicating that the amounts requested have no meaningful linear relationship with the users who deleted their accounts.

**6. user\_id and deleted\_account\_id**

* **Correlation: -0.249670**
  + **Interpretation:** There is a mild negative correlation, suggesting that as **user\_id**s increase, the likelihood of accounts being deleted decreases slightly.

**Objectives-Relevant Interpretation:**

* **Frequency of Service Usage:**
  + The evolution of **id** and **user\_id** indicates structural usage patterns that might point to increasing or varying engagement over time, though actual frequency analysis would need time-based or categorical insights.
* **Incident Rates:**
  + The weak relationships here suggest these numeric correlations do not directly capture incident behavior. Correlations among categorical variables related to incidents may offer more insight.
* **Revenue Generated by the Cohort:**
  + Although **amount** correlations with **id** provide some behavioral insights, actionable conclusions on revenue would benefit from combining this numeric analysis with categorical or ordinal factors like **status**.
* **Propose New Relevant Metric:**
  + A new metric combining user engagement features (e.g., request frequency, average amount over time) might provide a comprehensive overview of revenue influences and incident propensity.

**Conclusion:**

The correlations hint at certain alignments and trends between numerical features but suggest that fully understanding user behavior, service usage, and financial impacts necessitates a multi-faceted approach incorporating behavioral and categorical data.

## Initial Insights:

### Patterns and Trends

**Patterns in Service Usage**

Ein Bild, das Diagramm, Screenshot, Reihe enthält.

Automatisch generierte Beschreibung

Month-period to track how often cash requests are utilized over time, giving insights into service demand trends.

**Incident Rate by Cohort**

**Ein Bild, das Text, Diagramm, Software, Screenshot enthält.

Automatisch generierte Beschreibung**

This plot visualizes the number of incidents in different categories, helping to identify which types are more frequent, and could influence service improvements.

**Revenue**

Ein Bild, das Text, Screenshot, Diagramm, Reihe enthält.

Automatisch generierte Beschreibung

* Understanding which types of fees generate the most revenue can inform strategic decisions to optimize profits.

**New Relevant Metric: Average Processing Time?**

A new metric could be the average processing time from created\_at to updated\_at to evaluate operational efficiency.

Average Processing Time: 75.38135168961202 daysEin Bild, das Text, Diagramm, Reihe, Screenshot enthält.

Automatisch generierte Beschreibung

This metric offers insights into the service speed, highlighting potential areas for process optimization.

### Anomalies and Outliers

**a. Analyze total\_amount in fees**

Ein Bild, das Text, Screenshot, Rechteck, Zahl enthält.

Automatisch generierte Beschreibung

A box plot shows the distribution and its extremes. Outliers are values that lie beyond the whiskers of the box, representing atypical fee amounts.

**Analyze amount in CRs**

Ein Bild, das Screenshot, Text, Rechteck, Reihe enthält.

Automatisch generierte Beschreibung

**Investigate Frequency of Service Usage**

Ein Bild, das Screenshot, Reihe, Diagramm, Design enthält.

Automatisch generierte Beschreibung

## Next Steps:

### Hypotheses for Further Analysis

1. **Frequency of Service Usage:**
   * Hypothesis: Certain cohorts use cash advance services more frequently during specific times of the month or year.
   * Approach: Analyze the **created\_at** field in **CR** to identify usage patterns tied to temporal factors.
2. **Incident Rate:**
   * Hypothesis: Specific cohorts are more prone to payment incidents, influenced by their payment behavior and service type.
   * Approach: Utilize **status** and **category** in **Fees** to determine variations in incident rates among different cohorts.
3. **Revenue Generated by the Cohort:**
   * Hypothesis: Revenue generation from specific cohorts peaks during certain periods due to marketing campaigns or economic factors.
   * Approach: Use **type** and **total\_amount** in **Fees** to assess revenue trends over time.
4. **New Relevant Metric:**
   * Hypothesis: The processing time for cash requests affects user satisfaction and repeat use of services.
   * Approach: Calculate the average processing time using **created\_at** and **updated\_at** in **Crs** and analyze its correlation with repeat service usage.

### Additional Data Needed

For **1. Frequency of Service Usage**

Current data allows us to understand usage over time using the **created\_at** field in **df2**. However, cohort identification might require you to have a user characteristic (e.g., **user\_id**) to group by user characteristics rather than just timestamps.

**Additional Data Needed:**

* If cohort is based on user demographics or signup source, you would need access to user profile data not present here.

For 2. Incident Rate

You can assess incident rates with **status** and **category** in **df1**. This gives you an overview of what incidents occur, but not detailed reasons or user characteristics to understand cohort-specific trends.

**Additional Data Needed:**

* More comprehensive cohort labeling, such as demographic data or cohort assignment, would strengthen this analysis.

**3. Revenue Generated by Cohort**

Revenue can be assessed through the **total\_amount** in **df1**, segmented by **type**. However, understanding this by different user groups or cohorts requires additional labeling.

**Additional Data Needed:**

* Cohort definitions based on user socio-demographics, acquisition channels, etc., would aid in deep segmentation.

**4. New Relevant Metric**

The current data allows for process efficiency analysis, such as measuring processing time (**updated\_at** vs. **created\_at** in **df2**). However, understanding the impact on user experience requires user feedback or additional satisfaction measures.

**Additional Data Needed:**

* User feedback scores or qualitative data post-interaction to correlate efficiency with user satisfaction.

**Conclusion**

The existing data provides a foundation to begin analysis on service usage, incident rates, and revenue. However, for a deep-dive analysis, you will need additional data, particularly relating to user demographics, acquisition channels, and potentially user feedback or satisfaction to draw richer insights. These could facilitate more precise cohort analysis, detailed incident impact evaluations, and a comprehensive understanding of revenue streams relative to user characteristics

**Checklist for EDA:**

* Understand the project's objective.
* Obtain and describe the dataset.
* Examine data structure and basic stats.
* Visualize data distributions and relationships.
* Investigate main features and correlations.
* Summarize initial findings and insights.
* Define the next steps for analysis.