**Exploratory Data Analysis (EDA) Summary of extract - cash request - data analyst.csv:**

The dataset consists of **23,970** entries and **13 columns**. Below are some key insights based on the initial data exploration:

1. **Columns Overview**:
   * **Numeric Columns**:
     + amount: Cash request amount, ranging from 1 to 200.
     + user\_id: User ID, with values ranging from 34 to 103,719.
     + deleted\_account\_id: Represents deleted accounts, mostly containing 0.
   * **Object Columns**:
     + status: Status of the cash request (e.g., approved, rejected, etc.).
     + created\_at, updated\_at, moderated\_at, reimbursement\_date, cash\_request\_received\_date, money\_back\_date, send\_at: Date-related columns in string format.
     + transfer\_type: The type of transfer (e.g., regular).
2. **Missing Values**:
   * moderated\_at has missing values (approximately 33%).
   * cash\_request\_received\_date, money\_back\_date, and send\_at have some missing values.
3. **Key Statistics**:
   * The average amount is **82.72**, with a range from **1** to **200**.
   * user\_id has a wide spread, with the mean user ID around **31,808**.

**Visualizations:**

1. **Distribution of Amount**:
   * The distribution of amount is slightly skewed, with most cash requests falling between **1** and **100**, and a few larger requests pushing the upper limit to **200**.
2. **Distribution of User ID**:

(Frequency of Service Usage: Understand how often users from each cohort utilize IronHack Payments' cash advance services over time.)

* + The user\_id variable has a wide range, with several distinct users, but the distribution is quite spread out. The central tendency appears to be around **31,808**.
* **Repeatedness of User IDs**: The **majority of users** only request cash once, while some users have **multiple requests**. The top repeated user\_id values are as follows:
  + Other frequent users have made between **16 and 19 requests**.

This indicates that there are a few highly frequent users (potentially power users), while the rest make fewer requests.

1. **Status Distribution**:

* The **status** of cash requests shows a **dominant frequency of rejected requests**. There are significantly fewer approved requests, and a small number of other statuses (e.g., pending or others).
* The majority of requests have a status of "money\_back" (16,397 occurrences).
* Other notable statuses include "rejected" (6,568 occurrences), "direct\_debit\_rejected" (831 occurrences), and very few instances of "active", "transaction\_declined", "direct\_debit\_sent", and "canceled".

This distribution can guide us in understanding the acceptance/rejection rate of cash requests.

**Exploratory Data Analysis (EDA) Summary of extract - fees - data analyst - .csv**

**1. General Statistics:**

* **total\_amount**:
  + Mean: 5.00
  + Standard Deviation: 0.03
  + Range: All values are either 5 or 10, showing little variation.
* **cash\_request\_id**:
  + Range: From 1456 to 27010.
  + Most IDs are within a fairly wide range of values, with some clustering around certain values (e.g., 17160 as the median).

**2. Categorical Distributions:**

* **Payment Types (type)**:
  + Most common: "instant\_payment" (11,099 occurrences), followed by "postpone" (7,766), and "incident" (2,196).
* **Status (status)**:
  + Most common: "accepted" (14,841), followed by "cancelled" (4,938), "rejected" (1,194), and "confirmed" (88).
* **Charge Moment (charge\_moment)**:
  + Most common: "after" (16,724), followed by "before" (4,337).
* **Category (category)**:
  + Most common: "rejected\_direct\_debit" (1,599), followed by "month\_delay\_on\_payment" (597).

**3. Visualizations:**

* **total\_amount**: The values are concentrated around 5, with a slight presence of 10, indicating a very limited range.
* **type**: The "instant\_payment" type dominates the dataset.
* **status**: "Accepted" status is the most frequent, but there is also a notable portion of "cancelled" statuses.
* **charge\_moment**: The "after" charge moment is more common than "before."

**(Incident Rate: Determine the incident rate, specifically focusing on payment incidents, for each cohort. Identify if there are variations in incident rates among different cohorts.)**

calculate the **incident rate** if there are variations across different cohorts based on created\_at. This will give us insights into incident patterns over time.

**Incident Rate Analysis by Cohort (Monthly)**

From the **incident rate by cohort** plot and the data, we can make the following observations:

1. **Incident Rate Over Time**:
   * The incident rate shows a steady increase over the months, with a noticeable jump starting from **September 2020**.
   * The incident rate in **May 2020** is relatively low (around 0.26%), but it rises sharply in subsequent months, peaking at **6.23% in September 2020**.
2. **Insights**:
   * There are variations in incident rates across cohorts, suggesting that payment incidents (e.g., rejected requests) become more frequent as time progresses.
   * The rise in incidents may indicate changes in user behavior, system errors, or other external factors that affect payment acceptance.

The **incident rate** analysis shows how the rate of rejected payments evolves over time. The dataset has **significant variability** in payment incidents, which could warrant further investigation to identify the causes behind the increasing rejection rate.

(Revenue Generated by the Cohort: Calculate the total revenue generated by each cohort over months to assess the financial impact of user behavior.)

**Revenue Generated by Each Cohort** are:

1. **Group by Cohorts**: We'll extract the cohort month from the created\_at column. To determine the cohort (based on the month).
2. **Calculate Total Revenue**: We'll sum the total\_amount for each cohort. To calculate the total revenue.

**Graph Overview:**

* **X-Axis**: This represents the **cohort months**. Each point on the x-axis corresponds to a specific month when a cohort was created or when users first interacted with the service. The months are shown as time periods (e.g., 2020-01, 2020-02), which help identify trends in revenue over time.
* **Y-Axis**: This represents the **total revenue** generated by each cohort in that particular month. The values on the y-axis reflect the sum of total\_amount for all users in that cohort. The higher the value, the more revenue was generated in that cohort during the corresponding month.
* **Line/Markers**: The line with markers represents the **total revenue per cohort**. The line moves from left to right, showing how the revenue for each cohort evolves over time. Each data point represents the total revenue for that cohort in a given month.

**Summary of Insights :**

1. **Revenue Growth**: If revenue increases steadily, it indicates growth in the user base or transaction volume. Fluctuations may signal events like promotions, market changes, or seasonality.
2. **Cohort Performance**:
   * **High Revenue Cohorts**: Some cohorts may show stronger revenue generation, reflecting higher user engagement or spending.
   * **Low Revenue Cohorts**: These may indicate areas for improvement, such as newer or less engaged users.
3. **Seasonality**: Predictable revenue dips or spikes at certain times of year (e.g., holidays) could suggest seasonal trends, helping with forecasting and resource planning.
4. **User Behavior**: Higher revenue cohorts likely have more engagement or larger transactions, while lower revenue cohorts may need targeted strategies to improve performance.
5. **Financial Impact**: Understanding which cohorts contribute more revenue helps prioritize strategies to retain or expand high-performing groups while addressing underperforming ones.
6. **Targeting Opportunities**: Identifying cohorts with low revenue allows for targeted marketing campaigns to increase spending, such as offering promotions or loyalty rewards.