**Análisis de Cohortes para los Pagos de Ironhack (Proyecto 1)**

**Mezcua, Carol & Tonzo, Vanina**

**Análisis Exploratorio de Datos (EDA)**

**1. Datos generales**

Ambos datasets tienen estructuras distintas pero relacionadas (por cash\_request\_id del datset fees y la columna id del datsset cash).

Muchos campos tienen valores nulos: category en fees tiene datos en solo ~10% de las filas.

En cash\_requests, columnas como cash\_request\_received\_date, money\_back\_date, send\_at, y recovery\_status tienen muchos valores ausentes o vacíos.

**2. Variables categóricas**

Status tiene valores como "accepted", "rejected" y "money\_back" en distintos contextos.

Type en fees muestra "instant\_payment", "incident" y otro tipo.

**3. Variables temporales**

Campos como created\_at, updated\_at, paid\_at, reimbursement\_date, etc. permiten hacer análisis temporal (por ejemplo, cuándo se crean solicitudes, cuánto demoran en pagarse, etc.).

Las fechas parecen estar bien formateadas, lo que facilita su conversión a tipo datetime. **Nos quedamos con aquellos registros creados a partir del 2020** (fecha en la que segun el enunciado se inicia esta metodologia), ya que los datos anteriores podrian haber sido gestionados bajo otros criterios y las metricas podrian verse afectadas.

**4. Datos numéricos**

Monto (total\_amount en fees, amount en cash\_requests) tiene distribuciones estrechas (principalmente 5 o 100 euros).

Hay pocas variaciones en los montos, lo que puede simplificar el análisis de ingresos o tasas de recuperación.

hkhsdjflsdjfdlj

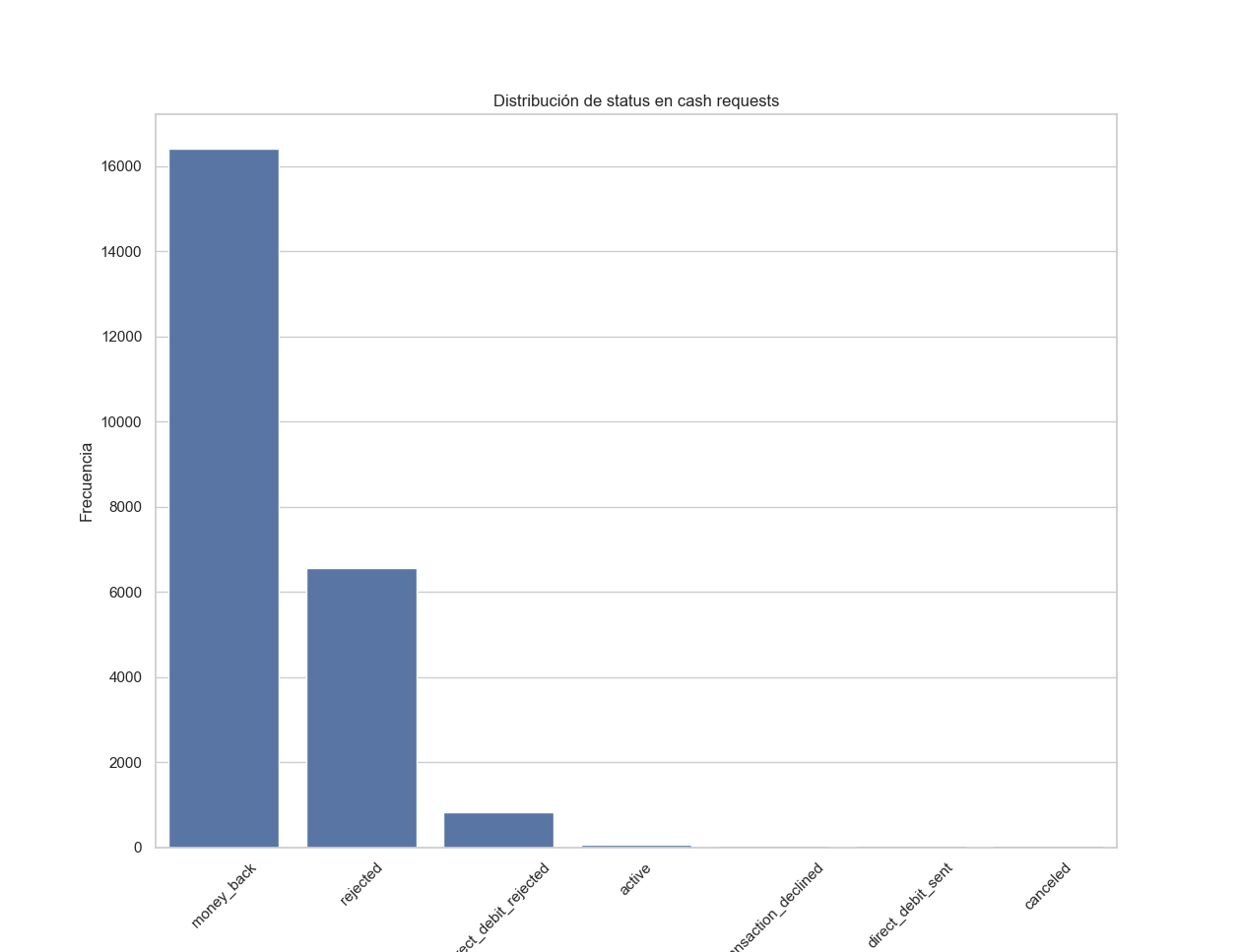

Figura 1. Histograma de la frecuencia de monto

Figura 2. Distribucion del estatus en el dataset cash

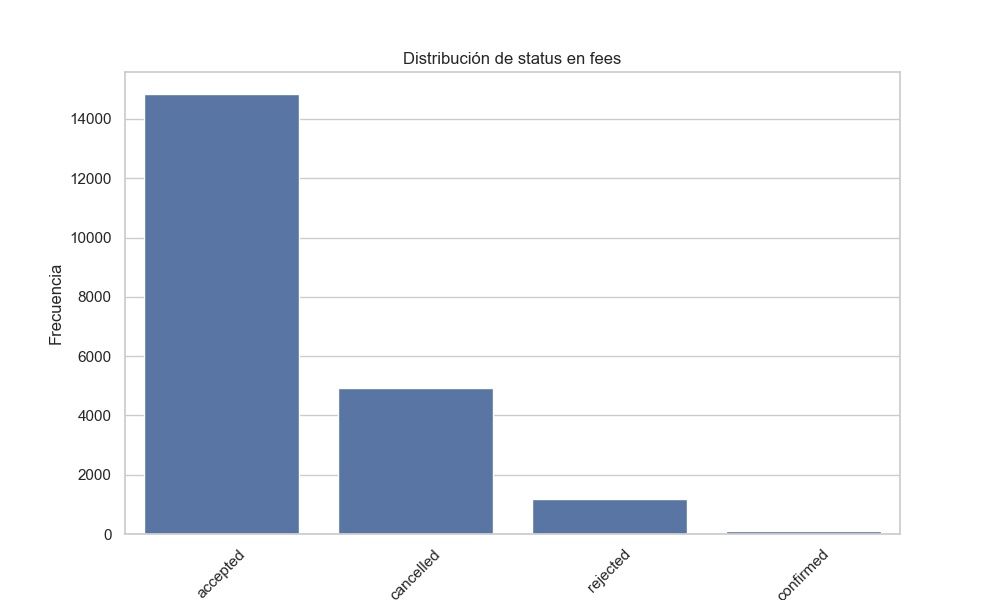


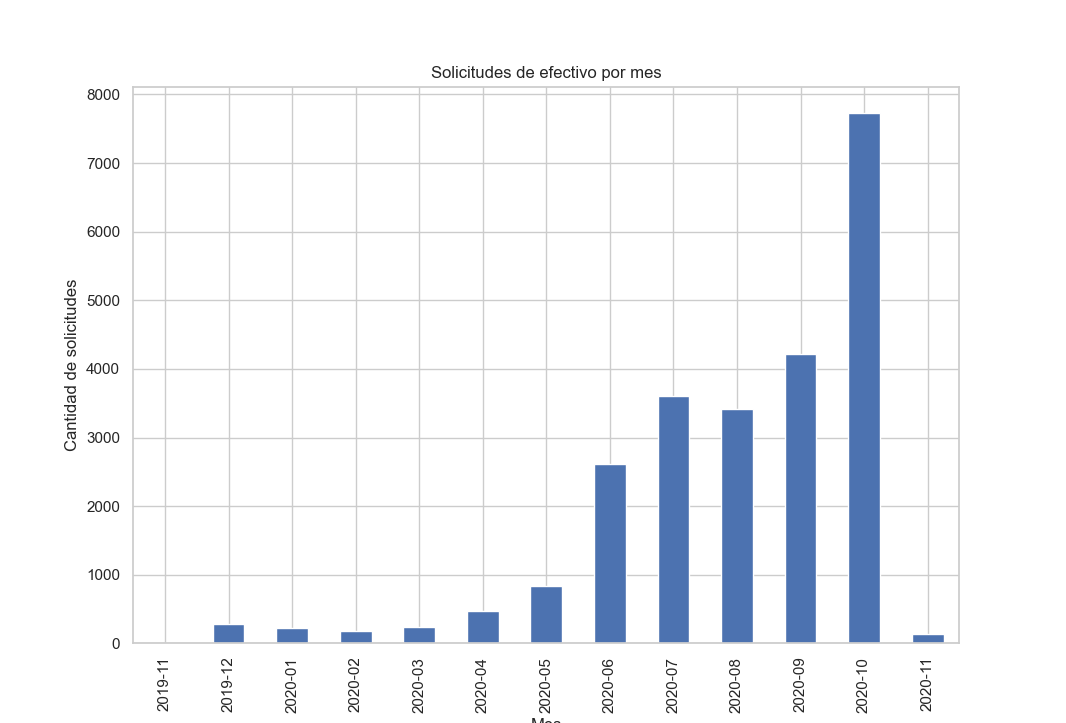
Figura 3. Distribucion del estatus en el datset Fees

Figura 4. Evolucion en el tiempo de las solicitudes de prestamos

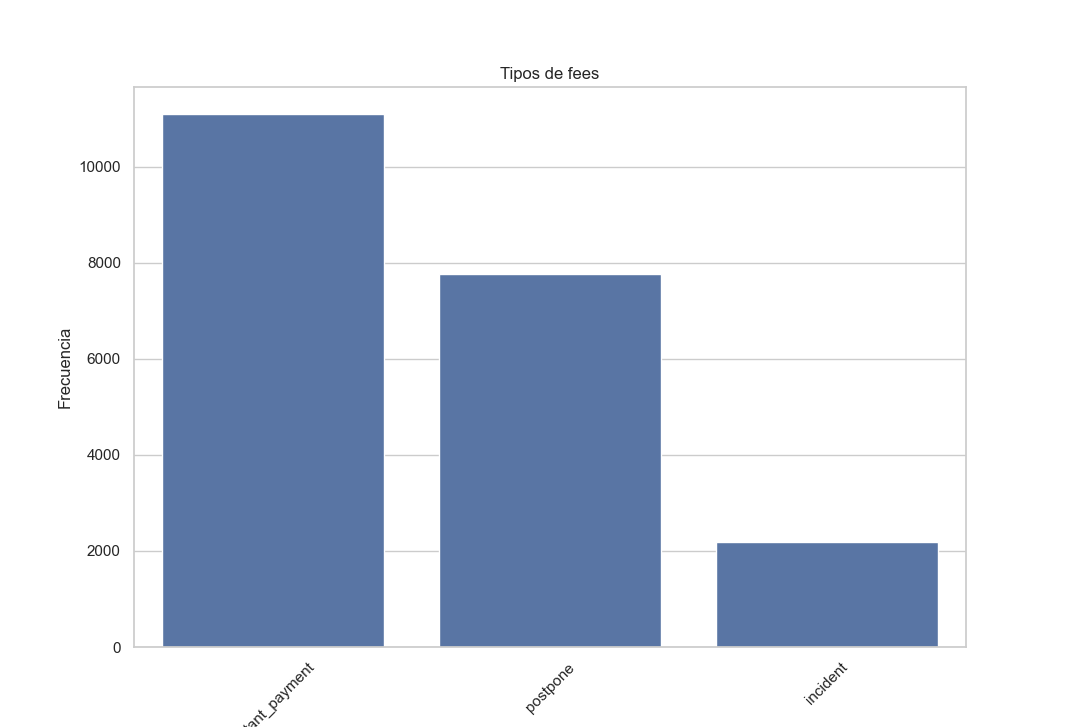


Figura 5. Frecuencia del tipo de comision o fee

**No eliminaremos registros de la variable amount del dataset cash porque no consideramos outliers los pagos de 1 a 200**

**Análisis de la calidad de los datos y métricas**

**Matriz de frecuencia de uso por cohorte y mes relativo**

**A screenshot of a graph

AI-generated content may be incorrect.**

**Descripción:**

* Filas: cohorte (mes del primer request por parte del usuario)
* Columnas: meses transcurridos desde el primer uso
* Valores: número de operaciones de cash en cada mes-cohorte (frecuencia de uso)
* Color: intensidad del uso del servicio por cada cohorte

**Relación con calidad de datos:**

* La asignación correcta de la cohorte (tras unificar user\_id y deleted\_account\_id, y filtrar por cash\_request\_received\_date válida) asegura que cada usuario esté en la cohorte que le corresponde, y solo se cuentan operaciones reales.
* La transformación a datetime permitió calcular correctamente el mes relativo y la evolución temporal.

**Conclusiones extraídas del gráfico:**

* Es importante tener en cuenta que los datos se cortan a fecha 2020-11-01, por lo que los datos del cohorte 2020-11 podrían no ser representativos. Los cohortes 2020-10 y 2020-11 no han tenido tiempo suficiente para mostrar su evolución en meses posteriores.
* Se observa una adopción muy fuerte el mes 0 para algunas cohortes recientes, y caída en el mes siguiente.
* Los cohortes antiguos muestran menor volumen de operaciones que los cohortes creados a partir del 2020-06, quizás por un cambio de enfoque por parte de negocio o implementación de campañas de marketing.
* Cohortes como 2020-10 y 2020-09 tienen un pico muy alto en el mes 0 (5046 y 2179 usos).
* Muchos cohortes muestran una caída muy abrupta de uso tras el primer mes. El caso más notable se presenta en 2020-10: de 5046 a 16 en el mes siguiente. Esto podría indicar:
  + Falta de valor percibido tras el primer uso.
  + Problemas de satisfacción.
  + Que el servicio se use solo una vez por necesidad puntual.

**Matriz de tasa de incidentes por cohorte y mes relativo**

**A graph with red squares

AI-generated content may be incorrect.**

**Descripción**:

* Filas: cohorte
* Columnas: meses transcurridos desde la primera request del usuario
* Valores: % de operaciones que acaban en incidente
* Color: rojo intenso = más incidentes

**Relación con calidad de datos**:

* Solo se incluyen incidentes de fechas posteriores al 01-01-2020, que es cuando Ironhack creó este sistema.

**Interpretación**:

* **Cohortes de abril-julio 2020** muestran tasas de incidente más altas, lo que podría reflejar problemas operativos o perfiles con más riesgo.
* **Cohortes recientes** tienen tasas mucho más bajas.
* Cohortes como **2020-04**, **2020-05**, **2020-06** y **2020-07** presentan tasas de incidentes muy elevadas, alcanzando **0.20 (20%)** en algunos meses.
* **La tasa tiende a concentrarse** en los primeros meses del ciclo de vida del usuario.

**Ingresos generados por cohorte**

A graph with blue bars

AI-generated content may be incorrect.

**Descripción**:

* Eje X: cohorte (mes del primer request del usuario)
* Eje Y: ingresos totales generados (suma de fees)

**Interpretación**:

* El cohorte 2020-10 es el que presenta ingresos más elevados, seguido del cohorte del 2020-06 y 2020-07.
* Los cohortes más antigüos han generado menos ingresos que los más recientes pese a que la fecha de corte es el 01-11-2020. Esto puede deberse a:
  + Mejor adquisición de usuarios.
  + Mejor producto u oferta.
  + Mayor volumen o recurrencia de uso.
* El cohorte 2020-11 no presenta la información completa de un mes y no tiene la antigüedad suficiente para poder proceder a un análisis completo.

**Proporción de usuarios problemáticos por cohorte**

A graph of blue bars

AI-generated content may be incorrect.

**Descripción**:

* Eje X: cohorte (mes del primer request del usuario)
* Eje Y: Proporción de usuarios preocupantes dentro de cada cohorte, es decir, usuarios con más de 3 incidencias.
* Altura de la barra: porcentaje de usuarios problemáticos sobre el total de esa cohorte.

**Interpretación**:

* El pico de usuarios problemáticos se concentra en 2020-04 a 2020-06. En 2020-04 la tasa de usuarios preocupantes supera el 40%, y sigue alta en 2020-05 y 2020-06.
* A partir de 2020-08 el porcentaje cae, probablemente porque la fecha de corte es el 01-11-2020.

**Conclusiones Finales**

* El análisis realizado sobre los cohortes ha permitido identificar patrones de uso, incidentes e ingresos a lo largo de 11 meses.
* La limpieza y revisión de la calidad de los datos han sido esenciales para asegurar la fiabilidad de los resultados y evitar sesgos o interpretaciones erróneas.
* Aunque se han observado diferencias notables entre cohortes, es recomendable seguir monitorizando su evolución y profundizar en el análisis conforme se disponga de más datos y madure el histórico.