Proyecto I: Introducción y Análisis Exploratorio de Datos (EDA)

I. Introducción

Ironhack Payments, fundada en 2020, se dedica a ofrecer adelantos de efectivo enfocándose en la gratuidad y la transparencia de sus servicios. Este análisis tiene por objetivo comprender el comportamiento de sus usuarios agrupados en cohortes definidos por el mes de creación de su primer adelanto en efectivo.

II. Área de estudio y datos

El estudio abarca el comportamiento de los usuarios de IronHack payments desde Noviembre de 2019 hasta Noviembre de 2020.

La información proviene de tres fuentes:

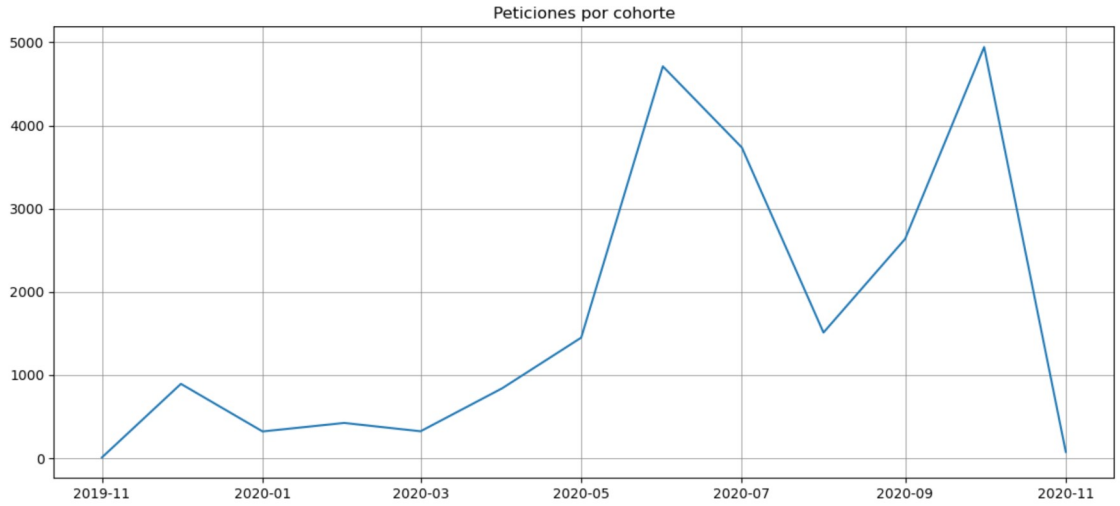
- cash_request.csv: Contiene los detalles de las solicitudes de efectivo
- fees.csv: Recoge información sobre las comisiones aplicadas
- lexique.csv: Proporciona la descripción de las variables de los archivos anteriores.

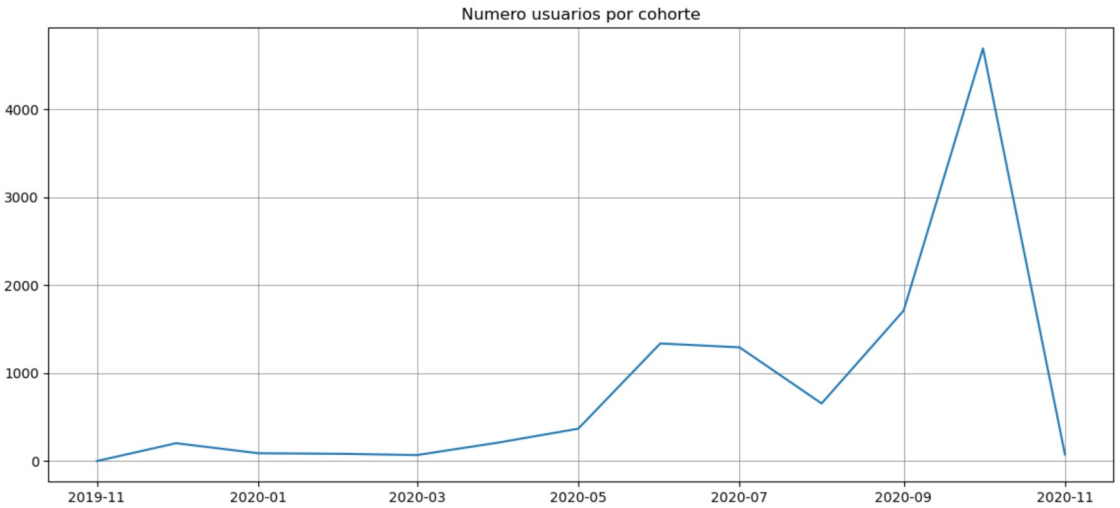
A continuación se describen las variables de trabajo. (Se considera que todas las variables tienen una posición de control al haber sido extraídas de la base de datos de la empresa y por tanto ser conocidas):

Var	Naturaleza				Nivel medición				Pos. Invest.*	
	Cuantitativa		Cualitativa		Cuantitativa		Cualitativa		Ind	Dep
Cash	Cont.	Discr.	Dicot.	Polit.	Razón	Intervalo	Ordinal	Nominal		
id		X			X				X	
amount		X			X				X	
status				X			X		X	
created_at	X					X				X
updated_at	X					X			X	
user_id		X			X				X	
moderated_at	X					X			X	
deleted_account_id		X			X				X	
reimbursement_date	X					X			X	
cash_request_debited_date	X					X			X	
cash_request_received_date	X					X			X	
money_back_date	X					X			X	
transfer_type			X		X				X	
send_at	X					X			X	
Fees										
id		X			X				X	

type				X	X				X	
status				X			X		X	
category			X		X				X	
reason				X	X			X		
created_at	X					X				X
updated_at	X					X			X	
paid_at	X					X			X	
from_date	X					X			X	
to_date	X					X			X	
cash_request_id		X			X				X	
total_amount		X			X				X	
charge_moment			X					X	X	

A. Limpieza





III. Metodología

1. Análisis exploratorio: Estadísticas básicas y visualización para entender la naturaleza y distribución de los datos
2. Limpieza de datos: Identificación y eliminación de outliers y datos no representativos
3. Selección de variables y creación de las cohortes: División en grupos de los usuarios según su primer mes de adelanto en efectivo e identificación de las variables significativas para el objeto de estudio
4. Exploración por cohortes : Evaluación del comportamiento de los grupos de usuarios definidos por su mes de primer adelanto

IV. Limpieza y preparación de datos

- Describe the steps taken to clean and preprocess the data.
- Mention any challenges faced and how they were addressed.
- Include code snippets or screenshots if necessary.

a. Limpieza de Cash_request:

- Eliminación de los user_id nulos en cash_request.csv

b. Limpieza de fees.csv:

- Eliminación de Outliers: Borrado del id 20604

```
fees["total_amount"].value_counts()
```

```
total_amount
5.0     21060
10.0        1
Name: count, dtype: int64
```

```
fees[ fees["total_amount"]==10]
```

	id	cash_request_id	type	status	category	total_amount	reason	created_at	updated_at
20604	15552	22799.0	instant_payment	accepted	NaN	10.0	Instant Payment Cash Request 22799	2020-10-21 13:01:52.493241+00	2021-01-21 15:42:51.372269+00

```
fees=fees[ fees["total_amount"]!=10]  # Eliminamos el registro porque es un outsider
fees["total_amount"].value_counts()
```

```
total_amount
5.0    21060
Name: count, dtype: int64
```

- **Reconstrucción de valores null en "cash_request_id" recuperando el id de la columna "reason"**

```
fees[fees["cash_request_id"].isnull()] # vemos que hay cash_request_id con null
```

	id	cash_request_id	type	status	category	total_amount	reason	created_at	updated_at	paid_at	from_date	to_date	charge_
1911	2990	NaN	instant_payment	cancelled	NaN	5.0	Instant Payment Cash Request 11164	2020-08-06 22:42:34.525373+00	2020-11-04 16:01:17.296048+00	NaN	NaN	NaN	
1960	3124	NaN	instant_payment	cancelled	NaN	5.0	Instant Payment Cash Request 11444	2020-08-08 06:33:06.244651+00	2020-11-04 16:01:08.332978+00	NaN	NaN	NaN	
4605	5185	NaN	instant_payment	cancelled	NaN	5.0	Instant Payment Cash Request 11788	2020-08-26 09:39:37.362933+00	2020-11-04 16:01:36.492576+00	NaN	NaN	NaN	
11870	3590	NaN	instant_payment	cancelled	NaN	5.0	Instant Payment Cash Request 12212	2020-08-12 14:20:06.657075+00	2020-11-04 16:01:53.106416+00	NaN	NaN	NaN	

Sustituimos los cash_request_id nulos por el valor del id que recuperamos de la columna reason

```
fees["reason"].unique()
```

```
array(['Instant Payment Cash Request 14941', 'rejected direct debit',
       'Instant Payment Cash Request 23371', ...,
       'Instant Payment Cash Request 25331',
       'Instant Payment Cash Request 23628',
       'Instant Payment Cash Request 20982'], dtype=object)
```

```
#sustituimos el valor
fees['cash_request_id'] = fees.apply(lambda x: x['reason'].split()[-1] if pd.isnull(x['cash_request_id']) else x['cash_request_id'], axis=1)
```

c. Acotación del periodo de estudio y cohortes:

- En noviembre 2019 hay un solo usuario por lo que se elimina por falta de datos suficientes para evaluar el mes.

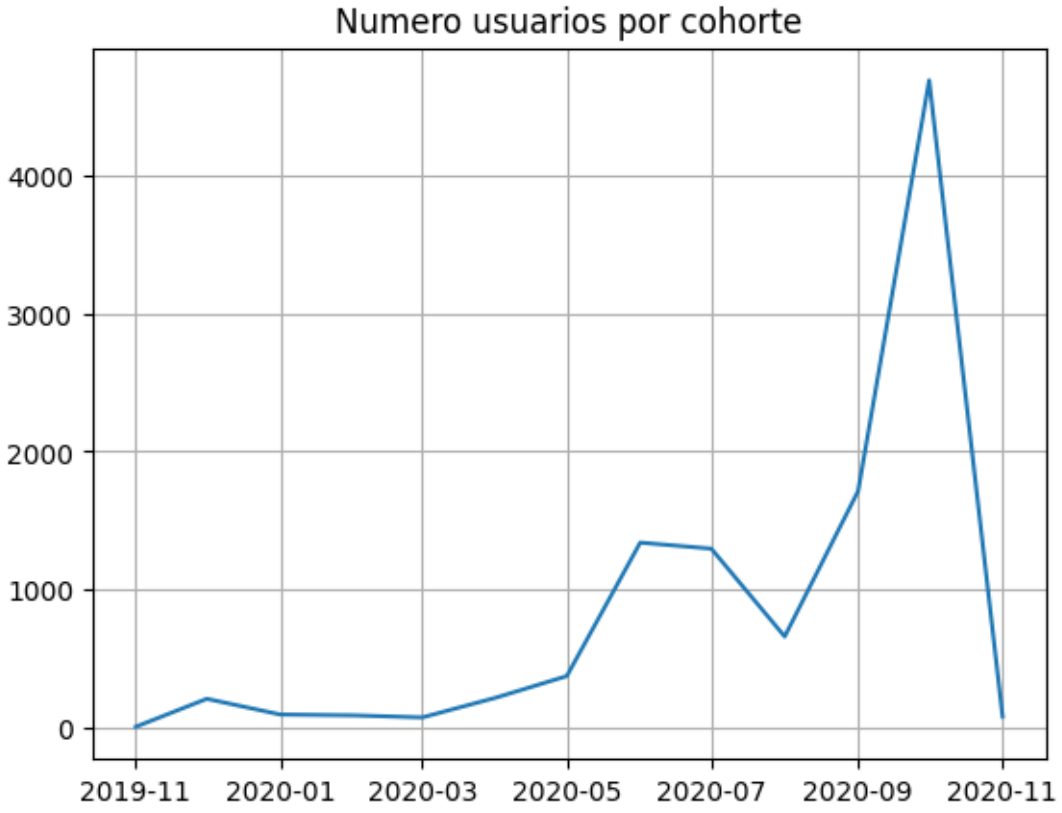
- En noviembre 2020 hay 75 usuarios que corresponden únicamente al 1/11, por lo que no es representativo del mes y no han tenido tiempo de hacer más solicitudes, por lo que también los eliminaremos

V. Resultados

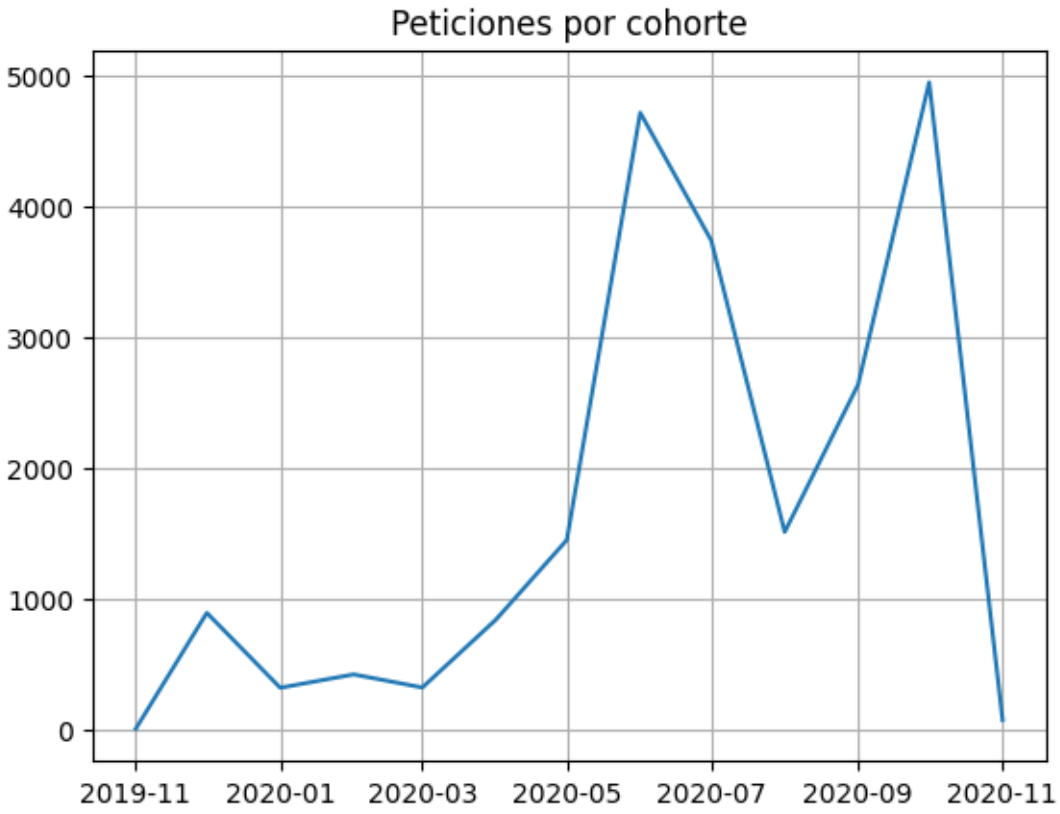
*

a. Frecuencia de uso del servicio

Los gráficos a continuación muestran cómo ha evolucionado el número de solicitudes de adelanto de efectivo por cohorte desde diciembre de 2019 hasta octubre 2020.



Los datos indican un aumento significativo del uso del servicio, especialmente a partir de mayo de 2020 y alcanzando su punto máximo en octubre del mismo año.



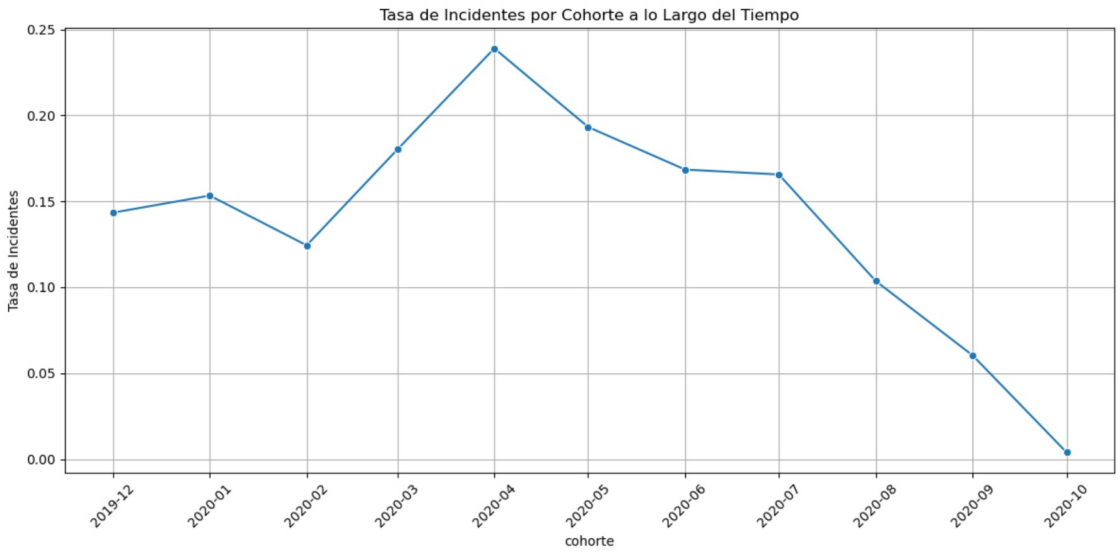
Este incremento constante puede explicarse por la expansión significativa de la base de usuarios a lo largo del tiempo, como puede observarse en el gráfico a continuación.

b. Tasa de incidentes

La tasa de incidentes se calcula como el ratio entre el número total de incidentes por cohorte, entre el número total de solicitudes. Se consideran incidentes aquellas transacciones que resultaron rechazadas o fallidas.

Los datos muestran unas tasas de incidentes considerablemente altas entre diciembre de 2019 y marzo de de 2020, con un aumento notable a partir de febrero.

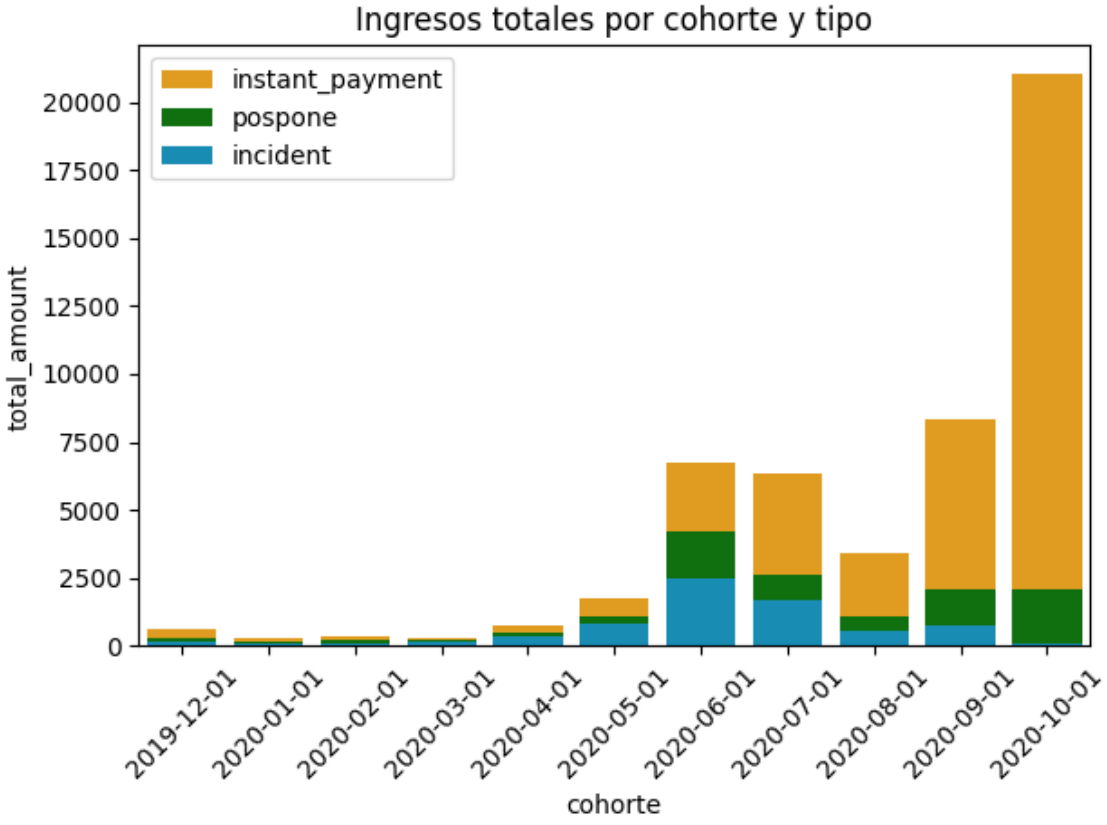
Sin embargo, pese al aumento en el volumen de solicitudes, la tasa de incidentes disminuye drásticamente a partir de abril, alcanzando su mínimo histórico en octubre de 2020 que es el último mes analizado. A este respecto, se puede apreciar por lo tanto una mejora muy significativa en el proceso de solicitud, con una reducción constante de las solicitudes rechazadas o fallidas notable a partir de abril de 2020.



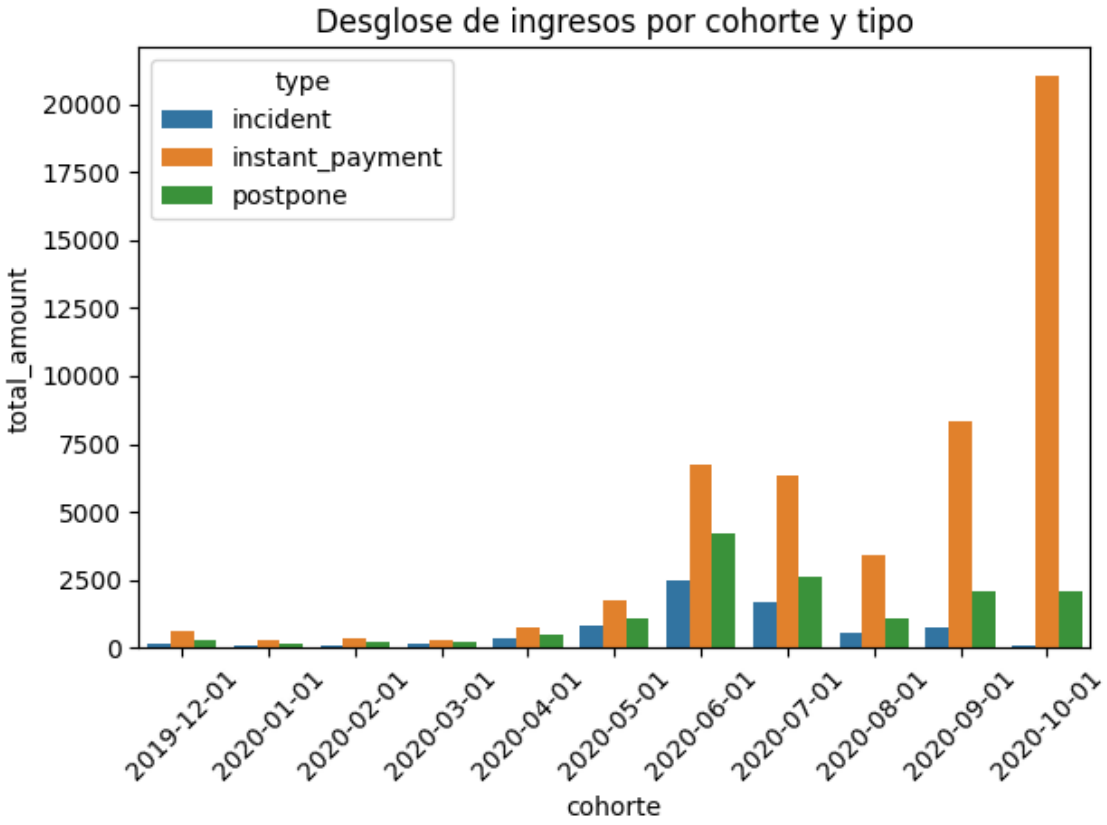
c. Ingresos generados por cohorte

Del análisis del total de ingresos generados por comisiones se observa que el aumento de solicitudes y usuarios descrito anteriormente va de la mano, como cabría esperar, de un aumento constante de los ingresos generados por cohorte.

INGRESOS POR COHORTE Y TIPO DE SOLICITUD



Si prestamos atención al tipo de ingreso por cohorte y tipo de solicitud, observamos asimismo un aumento sostenido de los ingresos provenientes de solicitudes de efectivo instantáneas frente a las solicitudes postpuestas.



Las solicitudes pospuestas lideran los ingresos desde diciembre de 2019 hasta mayo de 2020, cuando alcanza su punto álgido, para después disminuir progresivamente. Mientras que los ingresos por solicitudes instantáneos tienden a aumentar a lo largo del rango de tiempo objeto de estudio, con un aumento notable a partir de mayo y un crecimiento exponencial en Octubre de 2020.

En cuanto a los ingresos relacionados con solicitudes con incidencias, su fluctuación va de la mano de la reducción de las mismas estudiado con detalle anteriormente. Se observa un aumento constante de los ingresos de solicitudes con incidencia hasta mayo de 2020 y una reducción progresiva a partir de ese mes de inflexión, para finalmente pasar a ser ingresos residuales en el último mes de estudio.

El análisis de los datos muestra por tanto una tendencia evidente por par parte de los usuarios hacia las transferencias instantáneas, de donde provienen la mayor parte de los ingresos recientes.

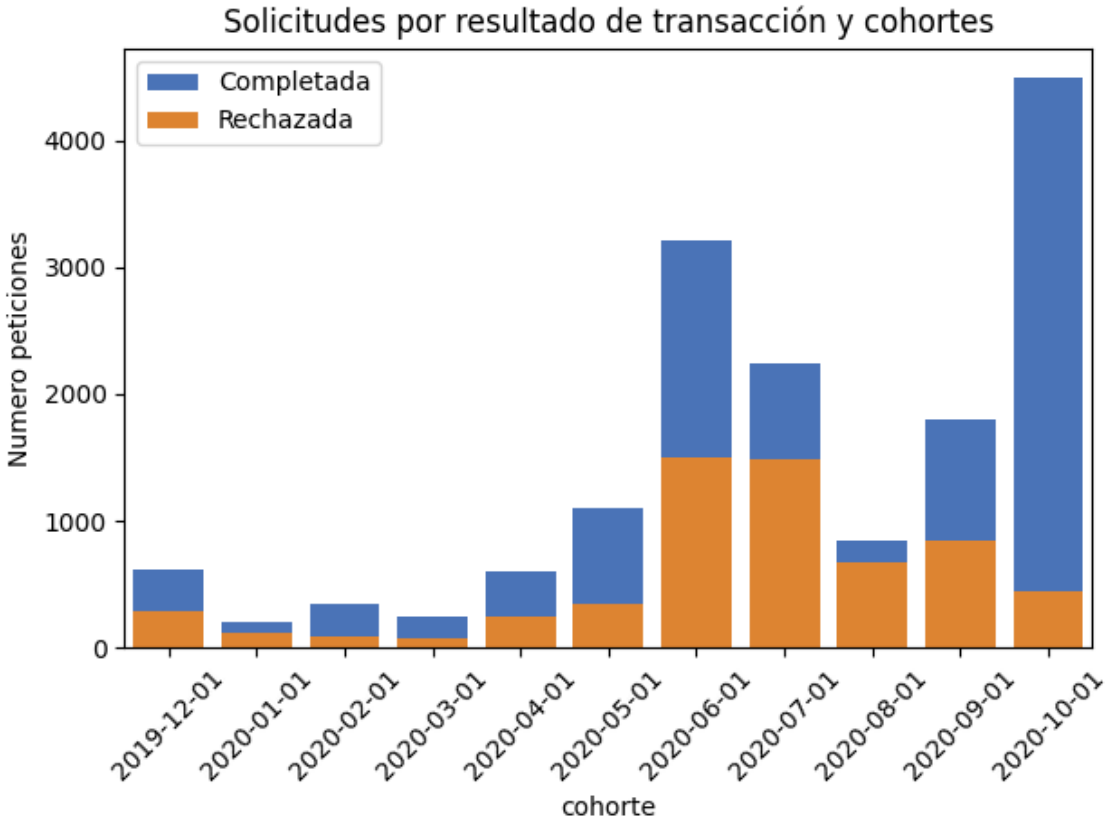
d. Desglose de solicitudes por resultado de la transacción

El desglose de comisiones por tipo de transacción muestra cómo se distribuyen los ingresos entre diferentes categorías de tarifas.

El desglose de solicitudes por resultado de la transacción muestra cómo se distribuyen las solicitudes en función de su resultado. Para facilitar su análisis, estas se han clasificado en tres tipos:

- Completadas: Incluye las transacciones con el estatus "accepted", "money sent", "active"," money_back" y "direct_debit_sent"
- Rechazadas: Estatus "Rejected", "Transaction_declined", "Canceled" y "Direct_debit_rejected"
- Las transacciones pendientes ("waiting_user_confirmation" y "waiting_reimbursement") son nulas y por lo tanto no se reflejan en el gráfico.

Cómo se puede apreciar en el gráfico, las solicitudes rechazadas representan más de un cuarto de las solicitudes totales en el periodo estudiado.



Sin embargo, si estudiamos la evolución del resultado de las solicitudes por cohorte, vemos que a partir de junio de 2020, hay una clara tendencia de reducción de las solicitudes rechazadas con respecto al de completadas. En particular, observamos un aumento muy notable de las solicitudes completadas con éxito en el mes de Octubre, en el que las rechazadas pasar a ser minoritarias.