Project 1

# Cohort Analysis of Ironhack Payments Users

Ironhack Data Science and Machine Learning Bootcamp

Author: Ginosca Alejandro Dávila
Date: December 13, 2024



# Project Overview

- **Project:** Cohort Analysis of Ironhack Payments Users
- **Goal:** Understand user behavior over time, segmented by their first cash request month
- **About Ironhack Payments:** Ironhack Payments is a digital financial services company offering *short-term cash advances* with *transparent pricing*. Users can:
    - Request *regular (free)* or *instant (fee-based)* transfers
    - Repay via scheduled *SEPA direct debits*
    - Encounter *fees* for instant transfers, postponements, or failed payments

# Key Metrics Analyzed

## 📈 Frequency of Service Usage

Measure how often users from each cohort use Ironhack Payments' cash advance services over time.

## ⚠ Incident Rate

Track the rate of payment incidents (e.g. failed reimbursements) by cohort to identify trends or risk patterns.

## 💰 Revenue Generated

Calculate the total fees collected per cohort to assess the financial contribution of each user group.

## 🧠 New Relevant Metric

Includes an additional user behavior or performance metric tailored to reveal deeper insights beyond standard KPIs.

# Dataset Summary

### 🧾 Data Source

- Provided by Ironhacks Payments as raw `.csv` and `.xlsx` files
- Definitions and business context available in **Lexique - Data Analyst.xlsx**

### 📄 Cash Requests (`cash_df`):

- 23,970 rows × 16 columns
- Contains one row per cash advance request
- Includes: cash request ID, amount,status, timestamps, user IDs, transfer type, reimbursement info, recovery info

### 💸 Fees Dataset (`fees_df`)

- 21,061 rows × 13 columns
- Contains one row per fee applied to a cash request
- Includes: ,fee ID, cash request ID, fee type, status, reason, category, timestamps, amount and charge timing

# Data Quality Assessment

- ✅ Checked for **missing values**
  - **Mostly expected (e.g. `reimbursement_date` missing for requests that were rejected, canceled, or already reimbursed)**
  - **Some required deeper inspection** (e.g., `user_id`, `cash_request_id`)
- ✅ Verified **datetime columns**
  - **Inconsistent formats and some timezone-aware vs. naive values**
  - **No future-dated timestamps found beyond 2024-12-12**
- ✅ Validated **monetary fields**
  - **No negative or zero values in** `amount` or `total_amount`
- ✅ Checked **link integrity** between datasets
  - Found 4 `fees_df` rows with missing `cash_request_id`
- ✅ Duplicate analysis
  - No fully duplicated rows or duplicate primary keys found

# Data Cleaning Summary

- ✅ Converted all **date columns** to be timezone-naive (`datetime64[ns]`)
- ✅ Stripped inconsistent timezones for uniformity
- ✅ Unified `user_id` and `deleted_account_id` into `final_user_id`
- ✅ Removed 4 unlinkable rows from `fees_df`
- ✅ Standardized **categorical values** (e.g., `status`, `type`)
- ✅ Verified and preserved **monetary columns** (`amount`, `total_amount`)
- ✅ Reordered columns for analysis readability

# Exploratory Data Analysis Overview

## 🔎 EDA Objectives

- Understand request volumes, fee types, and user behavior
- Uncover time-based trends in requests, revenue, and incidents
- Prepare data for cohort definition and metric aggregation

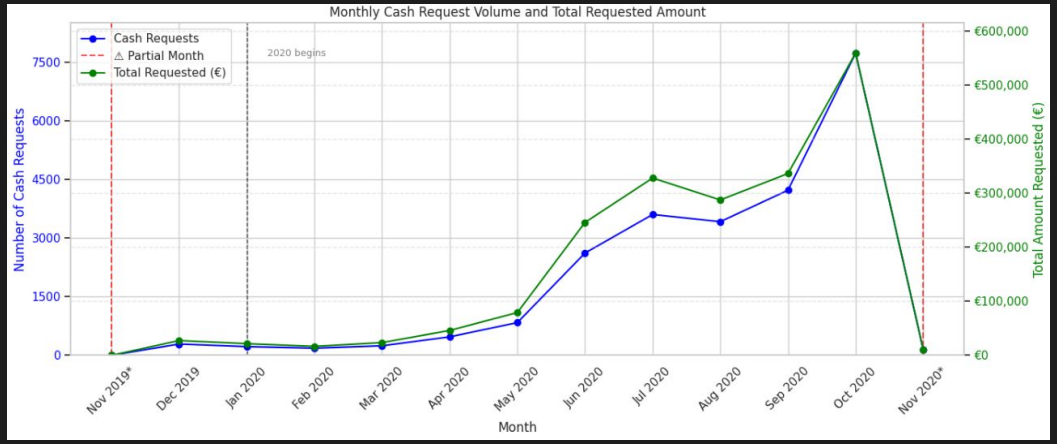
## 📁 Cleaned Datasets Used

- `clean_cash_requests.csv`: 23,970 rows
- `clean_fees.csv`: 21,057 rows

# Request and Revenue Growth Over Time

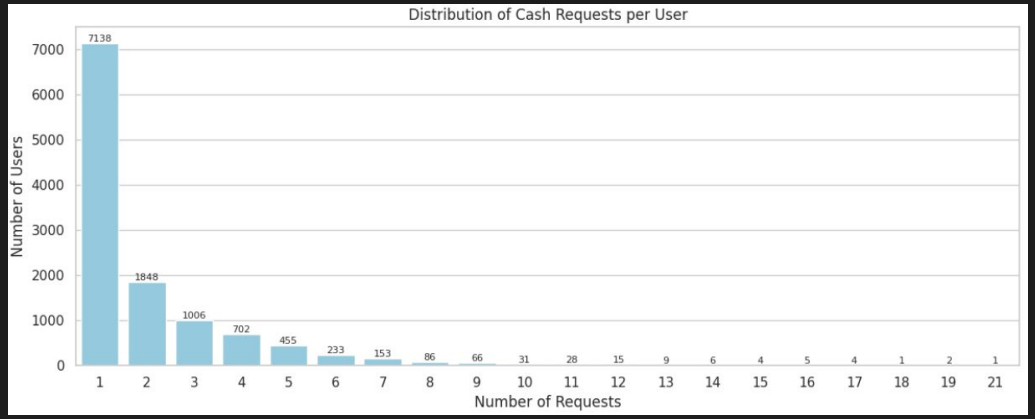
## 📈 Key Insights

- Strong growth began in May 2020, accelerating through the summer
- October 2020 marks the peak in both requests and cash volume
- Request volume and requested amounts mirror each other closely




📆 Note: Nov 2019 and Nov 2020 are partial months

# User Activity and Behavior



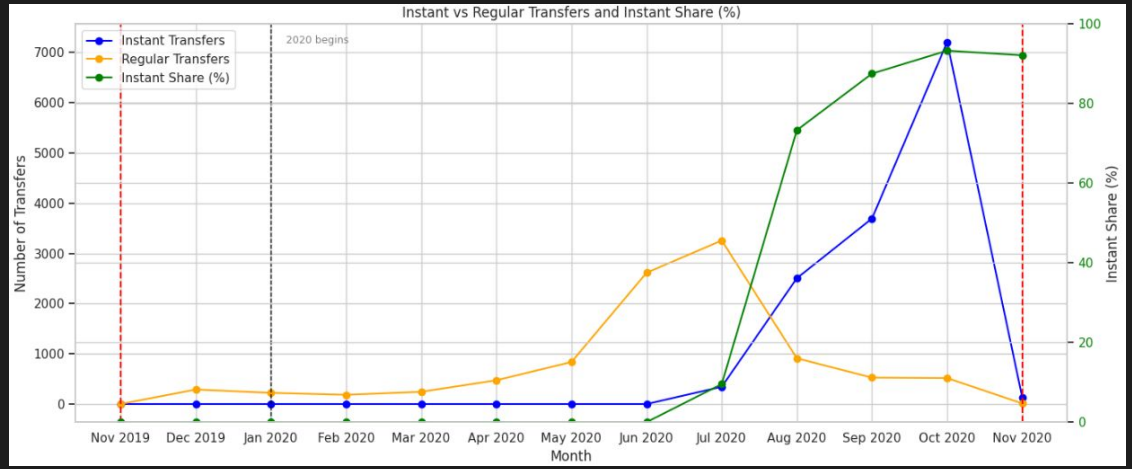

## 👤 User Insights

- 11,793 total unique users
- 60.5% used the service only once
- Remaining 39.5% used it multiple times

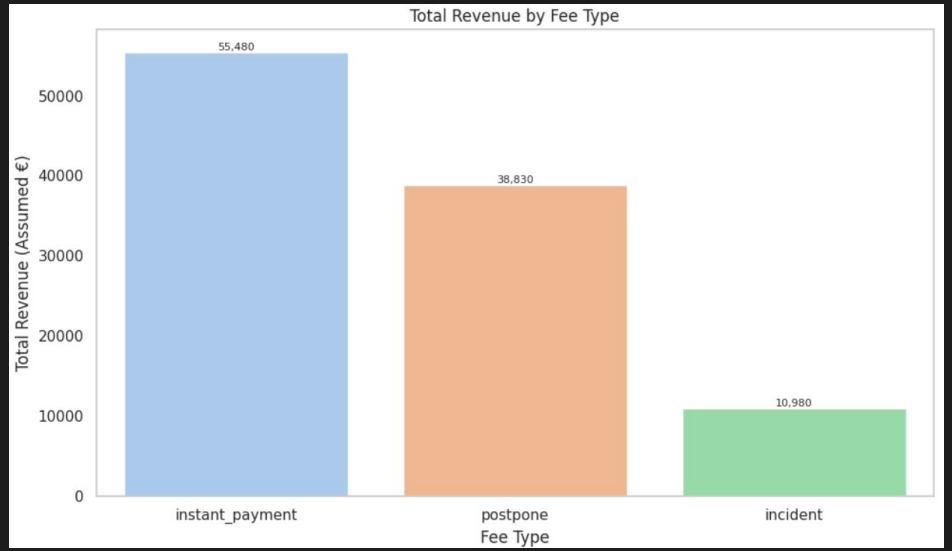
# Transfer Preferences Over Time

## 🚀 Instant vs Regular Transfers

- Instant transfers introduced around July 2020
- Instant share increased to 93% by October 2020
- Regular transfers declined as instant became dominant



# Revenue Breakdown by Fee Type



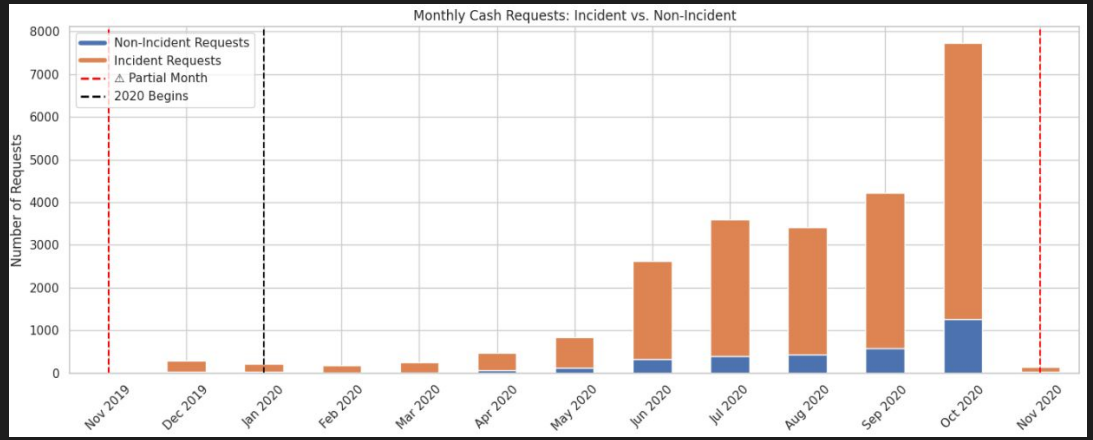

## 💸 Revenue Sources

- Instant fees: €55,480
- Postpone fees: €38,830
- Incident fees: €10,980

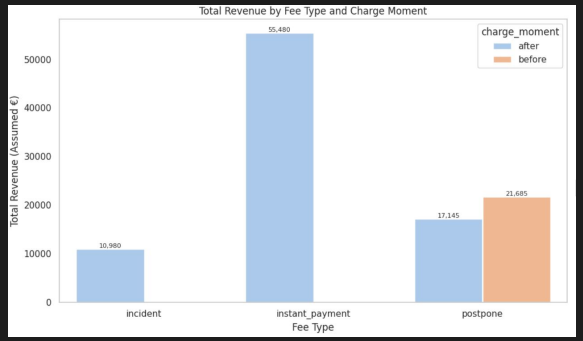
# Incidents and Risk Trends

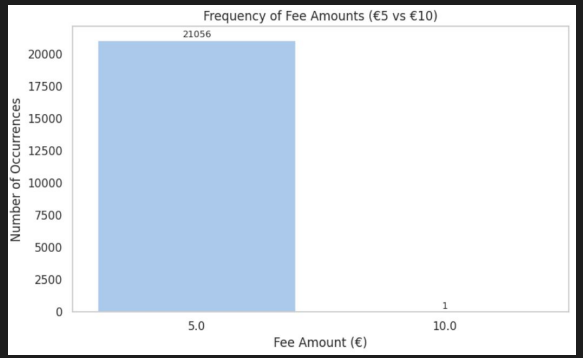
## ⚠Incident Insights

- 86% of requests had no incidents
- Peak incident volume in October 2020
- Incident rate remained stable around 10–16%



# Revenue Timing and Fee Consistency




## Key Insights

- Instant and incident fees are always charged after request
- Postpone fees can be charged before or after
- 99.99% of fees are flat €5 (only one €10 outlier)

# Linking Requests to Revenue

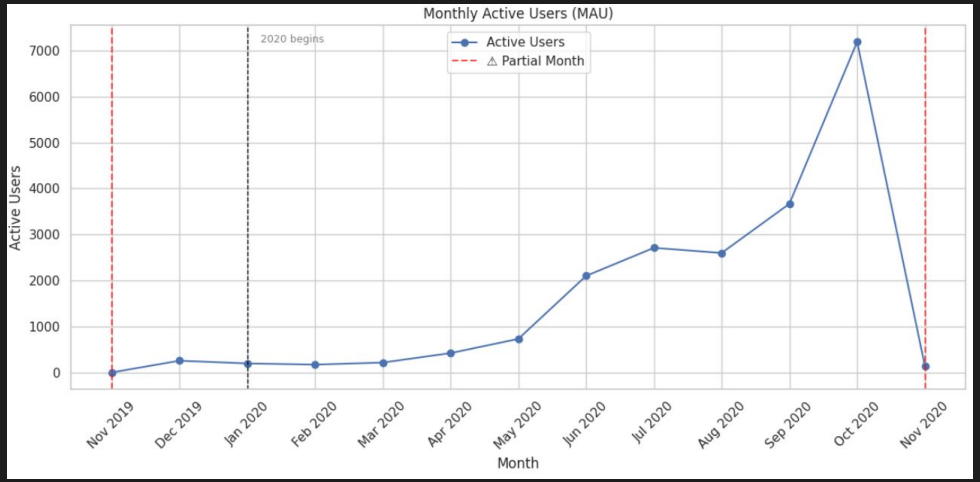
## 🔗 Merged Dataset Overview

- Merged on `cash_request_id`
- 23,970 cash requests, 12,933 had fees (~54%)
- Resulting dataset: 32,094 rows x 33 columns

# Ready for Cohort Analysis

## 📅 Cohort Preparation Steps

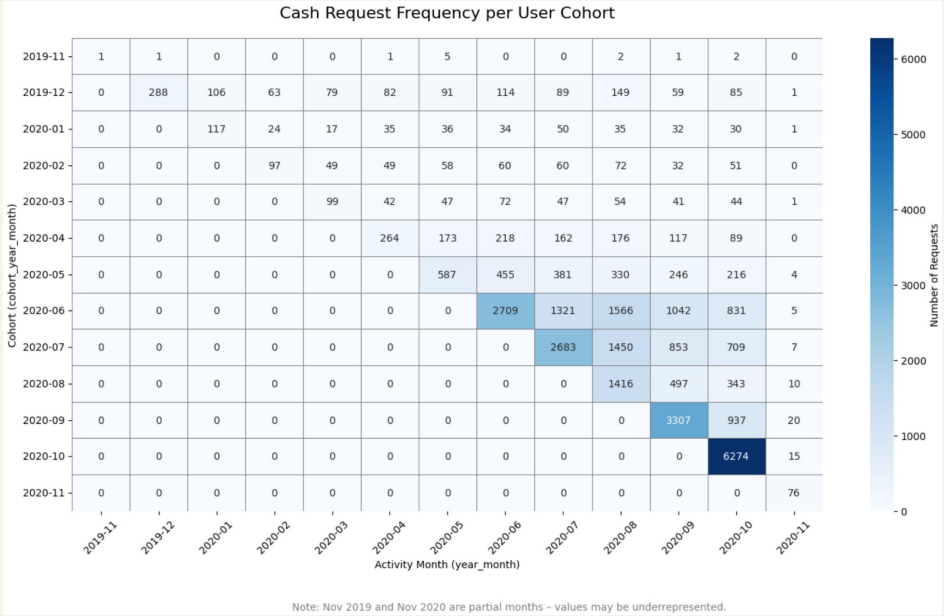
- Extracted first request date per user
- Assigned each user a cohort month
- Tracked Monthly Active Users (MAU)



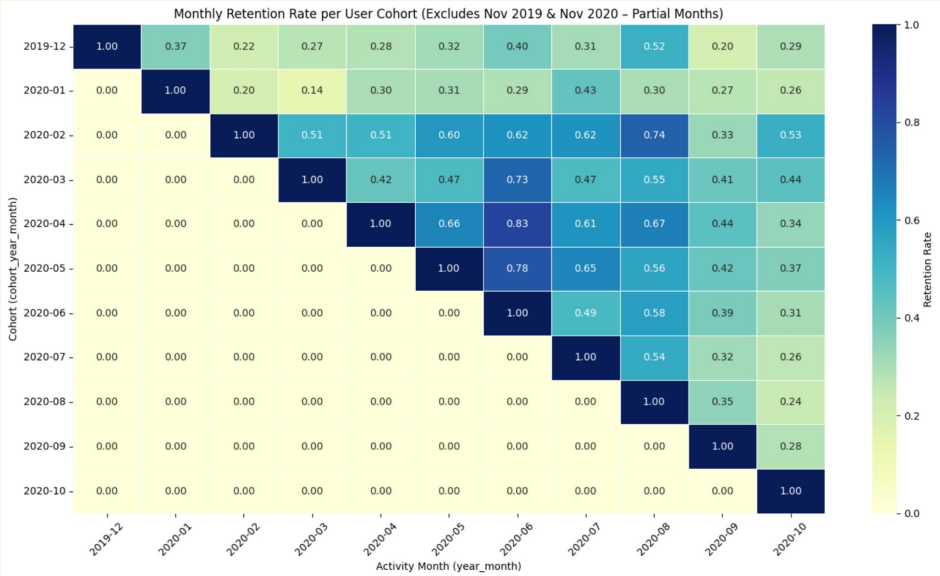
# Cohort Usage Frequency

## 🧮 Cash Request Frequency per Cohort

- Platform activity surged starting April 2020, peaking in October 2020 with 6,274.
- Other high-activity months include June 2020 (2,709), July 2020 (2,683) and September 2020 (3,307).
- Earlier cohorts like Feb-Apr 2020 exhibit sustained multi-month engagement.
- Later cohorts (e.g. Aug-Oct 2020) show strong initial use but rapid drop-off, suggesting short-term engagement.



# Cohort Retention Trends



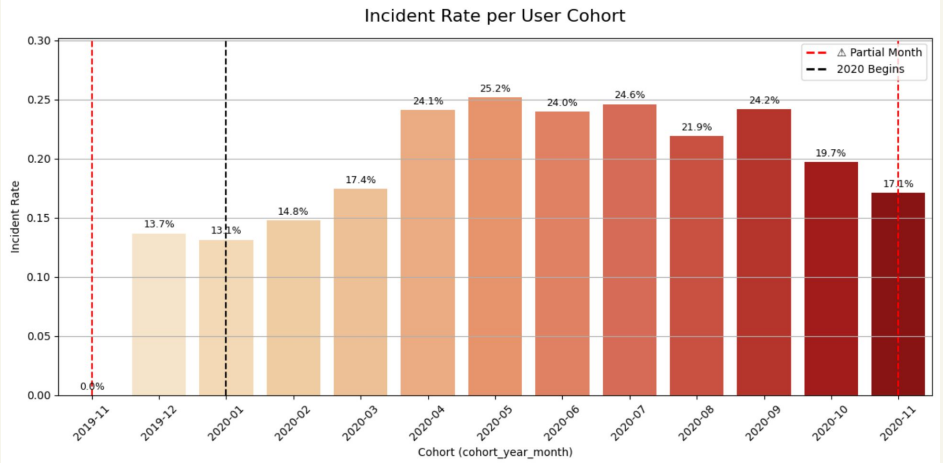

## 🔁 Retention Insights

- Retention typically drops after month 1
- April 2020 cohort retained 82.6% in month 2 and stayed above 60% through month 4
- February 2020 peaked at 74.2% by month 6
- Later cohorts showed faster drop-off

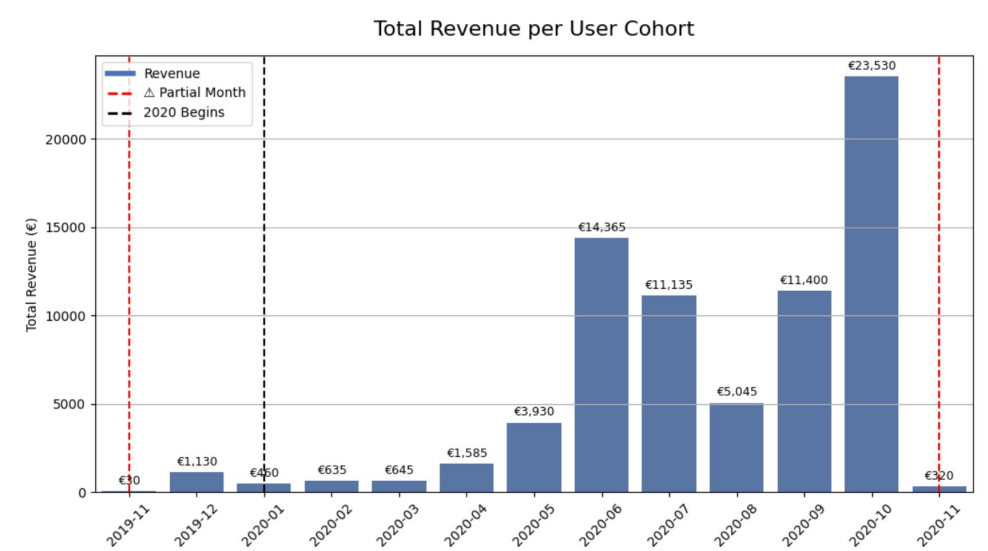
# Incident Rates by Cohort

## ⚠ Incident Trends

- Incident rates peaked around 25% between April-July 2020
- Early 2020 cohorts (e.g. Jan-Feb 2020) had lower risk, with rates around 13%
- Incident volume closely followed user growth, indicating operational pressure.



# Revenue and Monetization



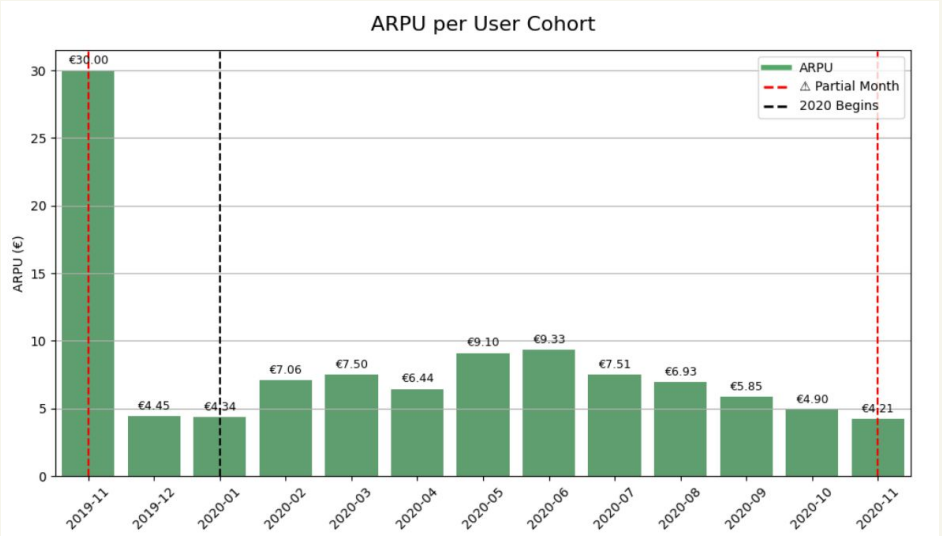

## 💰 Revenue by Cohort

- October 2020 cohort generated the highest revenue (€23.5k)
- June and July 2020 together contributed over €25k
- These cohorts also retained up to 49%-54% by the first month, showing stronger early engagement
- October 2020 cohort's retention may be underrepresented due to partial November 2020 data

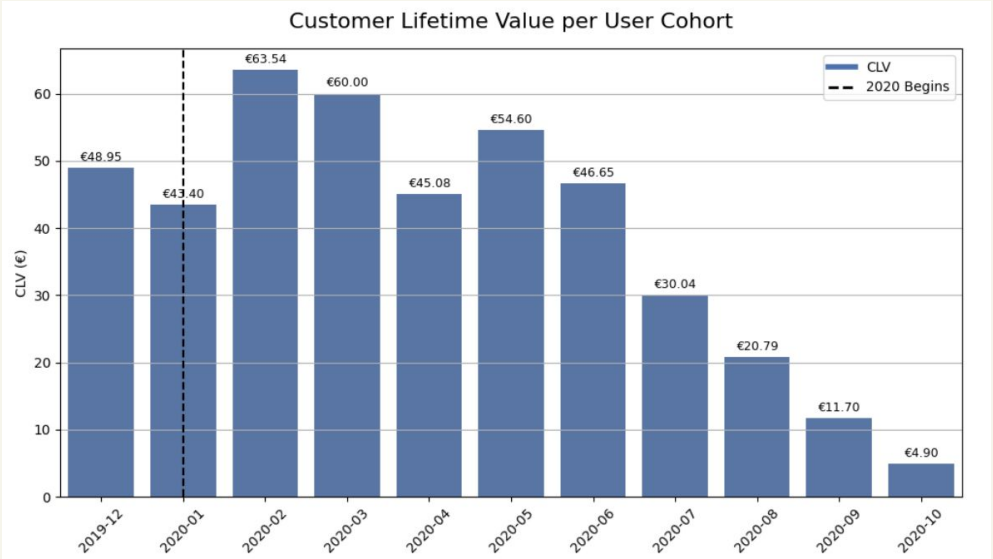
# ARPU (Average Revenue Per User) by Cohort

💸 **Monetization Effectiveness**

- June 2020 had the highest ARPU at €9.33
- May 2020 and July 2020 also performed strongly (€9.10 and €7.51)
- Early cohorts had lower ARPU but smaller user bases
- ARPU declined in late 2020 despite user growth



# CLV (Customer Lifetime Value) by Cohort




## 💡 Long-Term Value Insights

- February 2020 cohort had the highest CLV at €63.54
- March and May 2020 also showed strong CLV (€60.00 and €54.60)
- High CLV combines strong ARPU with sustained retention
- CLV declines in later cohorts due to faster churn

# Summary of Cohort Analysis

- Sharp user growth from April-July 2020, peaking in June
- Retention peaks early but was stronger for Feb-May 2020 cohorts
- Incident rates rose in May-July 2020, aligned with high activity
- June and July 2020 brought in over €25k combined revenue
- Feb 2020 had highest CLV (€63.54); June 2020 had the highest ARPU (€9.33)

# Business Recommendations

- Prioritize early engagement and onboarding
- Improve incident handling during growth surges
- Re-engage high-value cohorts (e.g., Feb-May 2020)
- Increase ARPU via upsells or loyalty programs
- Establish ongoing cohort monitoring

# Thank You!

## Questions?