Preparation steps

- Read the problem
- Write the pseudocode
    - Data importing and checking
    - Dataset investigation
    - Data cleaning
    - Data manipulation
    - Analysis steps
    - Visualizations

Dataset Exploration

- Check dataset info() and type()
- Check for missing values or NaN
- Check for duplicates

Data Cleaning

- Replace NaN values with 0 for columns that made sense
- Convert columns with “wrong” data type with the ones I need
- Check if the changes were made
- Merging tables
- Creating requested columns
- Exclude userID = 0 from the merged table

Data Deep Dive

- Analyse main statistics metrics using describe()
- Cohort size
- Unique cash request ids
- Frequency of service by cohort
- Incident Rate total and per cohort
- Total amount requested and received in fees
- ...

Data quality issues

Converting all Date columns from object to datetime type at the same time:

Because some columns had different format, no function would work and would give an error.
Solution: picked only the date columns I would be needing.

Left merge would duplicate rows:

Because the right table has ids duplicated, it would bring those duplications to the left table.
Solution: Removed duplicates and summed up fees/amount requested.

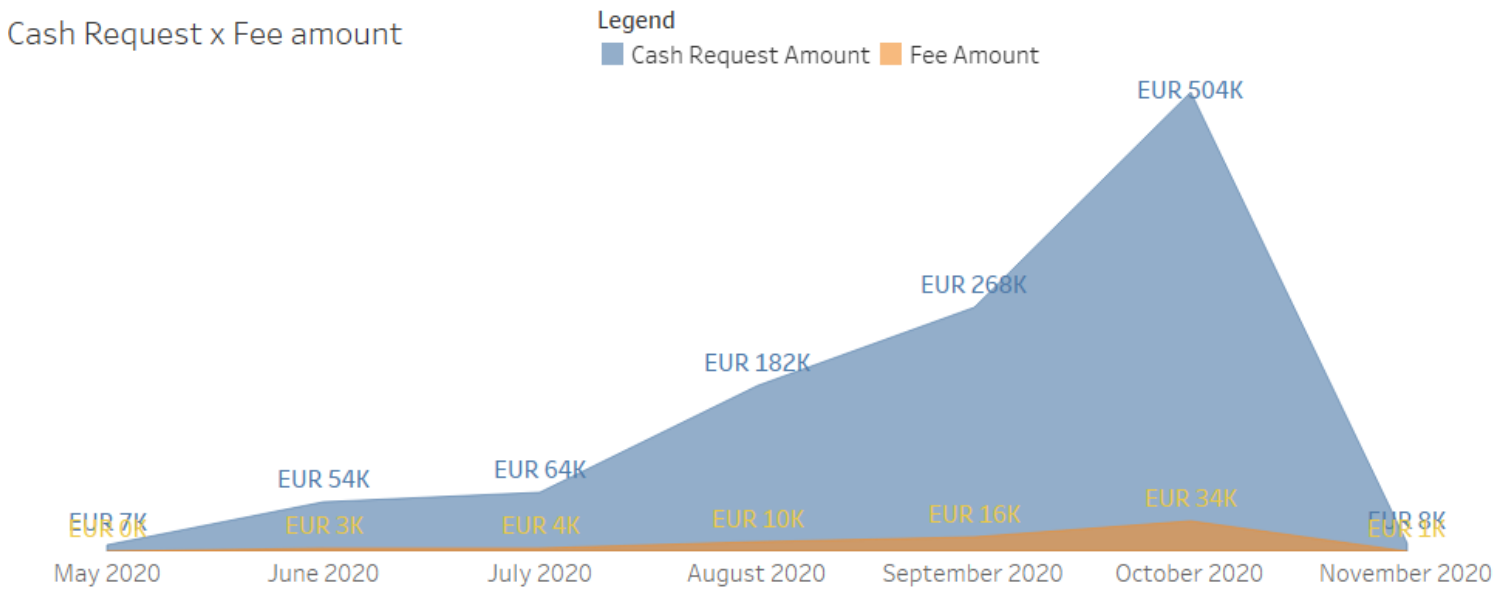
Cohort Exploration and Analysis

Summary

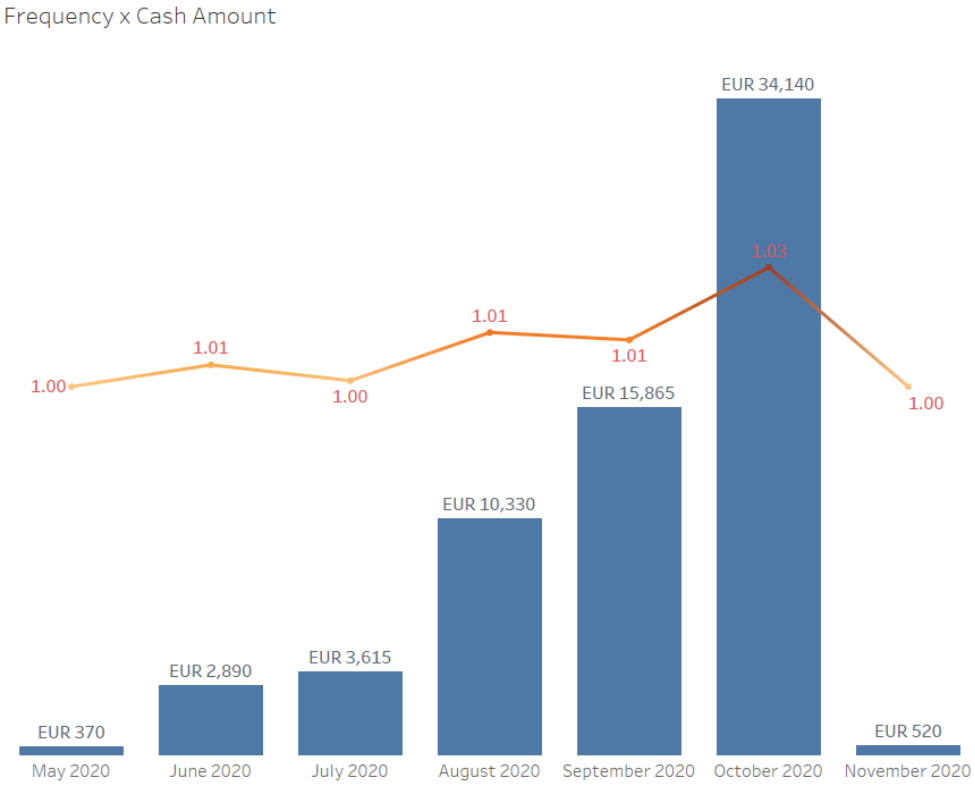
**Successful requests and fees*

Cash Requested	Fees Collected	Unique Users	Service Request Frequency
EUR 1,086,993	EUR 67,730	8,308	1.38

Peak of cash requests and fees collected in October 2020

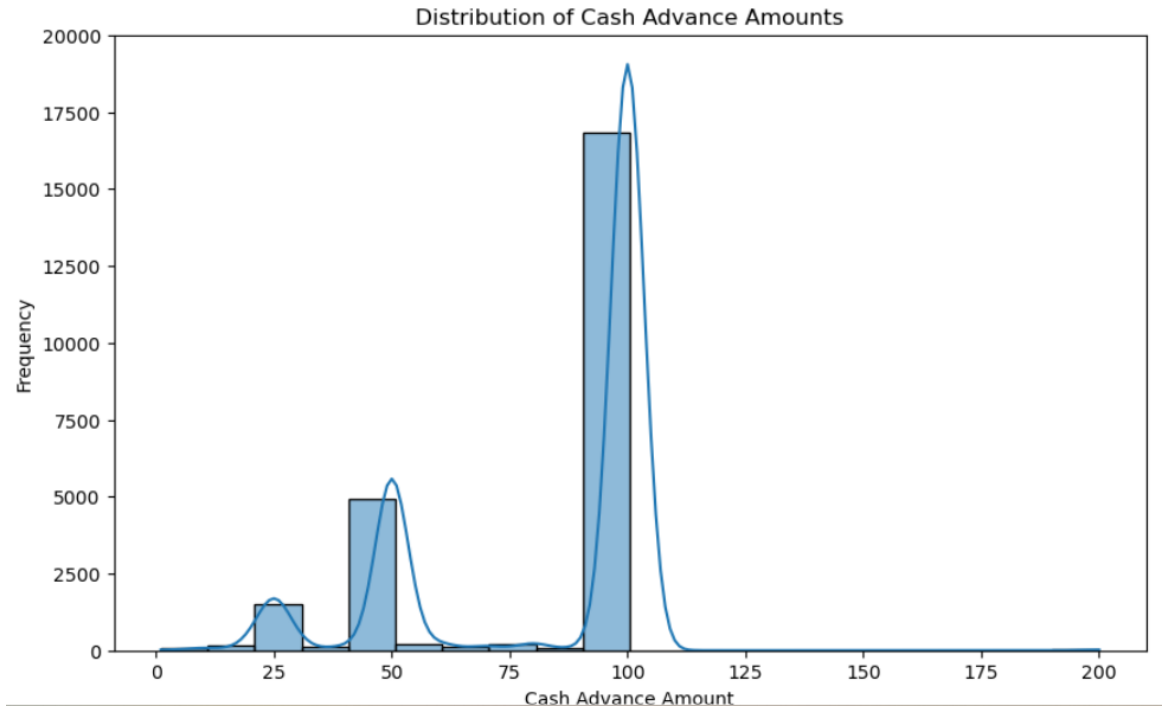


Cohort Exploration and Analysis



Higher frequency of requests per user doesn't show a direct relationship with the total amount requested

The average Cash Advance amount is concentrated in the EUR 100 group



Next: investigate why users are requesting higher cash amounts

Incident rate

```
Total occurrences: 12497
Total incidents: 2030
Incident rate: 16.0 %
```

Average incident rate of 16% with peak in June 2020.

By cohort

```
202005 5%
202006 16%
202007 14%
202008 10%
202009 6%
202010 1%
```

Fee amount as revenue

```
Total fee collected:
EUR 72015.0

Cohort-wise distribution of
successful fees collected:

202005 EUR 490.0
202006 EUR 4225.0
202007 EUR 4730.0
202008 EUR 11215.0
202009 EUR 16345.0
202010 EUR 34490.0
202011 EUR 520.0
```

On the other hand, the peak of fee collections was in 10/2020.

Incident rate and Fee collected doesn’t show to have any correlation.

Frequency usage

Rationale: cash_request_size/cohort_size)

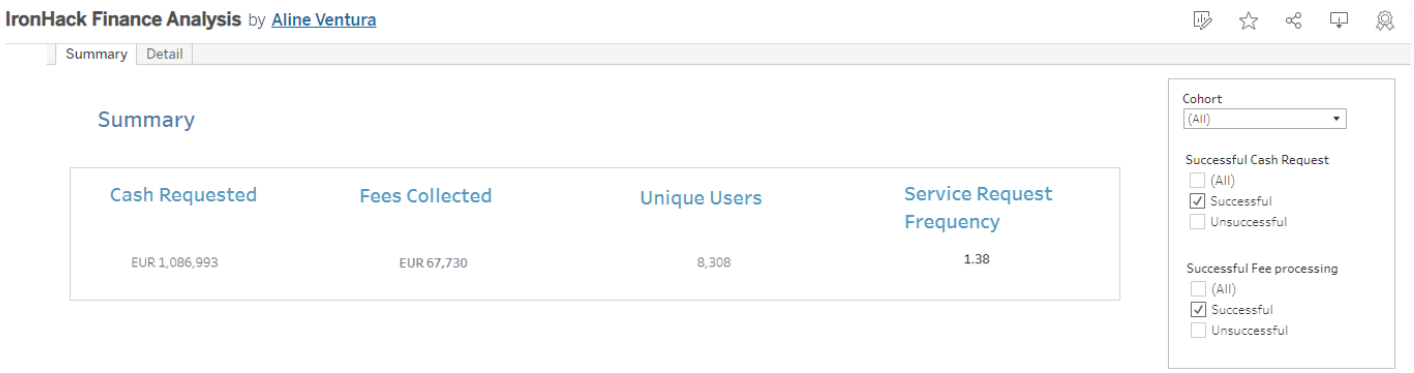
Total: 1.16

Cohort

201911 1.00
201912 1.12
202001 1.11
202002 1.07
202003 1.13
202004 1.13
202005 1.14
202006 1.23
202007 1.31
202008 1.30
202009 1.14
202010 1.07
202011 1.00

Higher frequency of requests per user doesn't show a direct relationship with the total amount requested

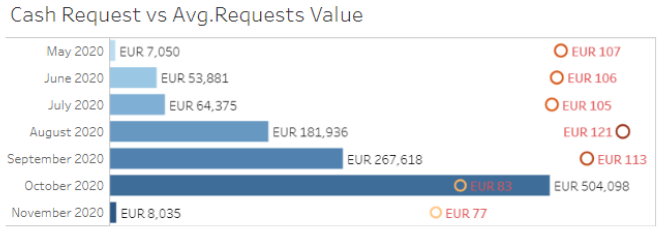
Tableau Dashboard

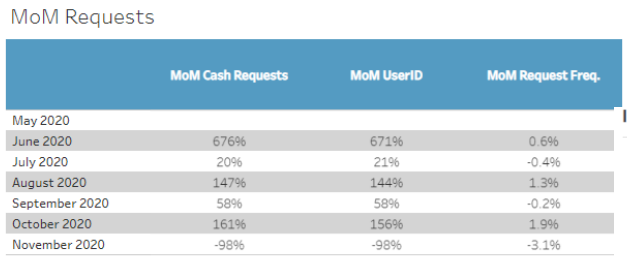


Summary view:

Summary of main KPIs and some overall visualizations relates to Cash Amount Requests vs the Avg. Amount per user.

Table with MoM variance of Cash Requests, Unique UserIDs and Request Frequency.

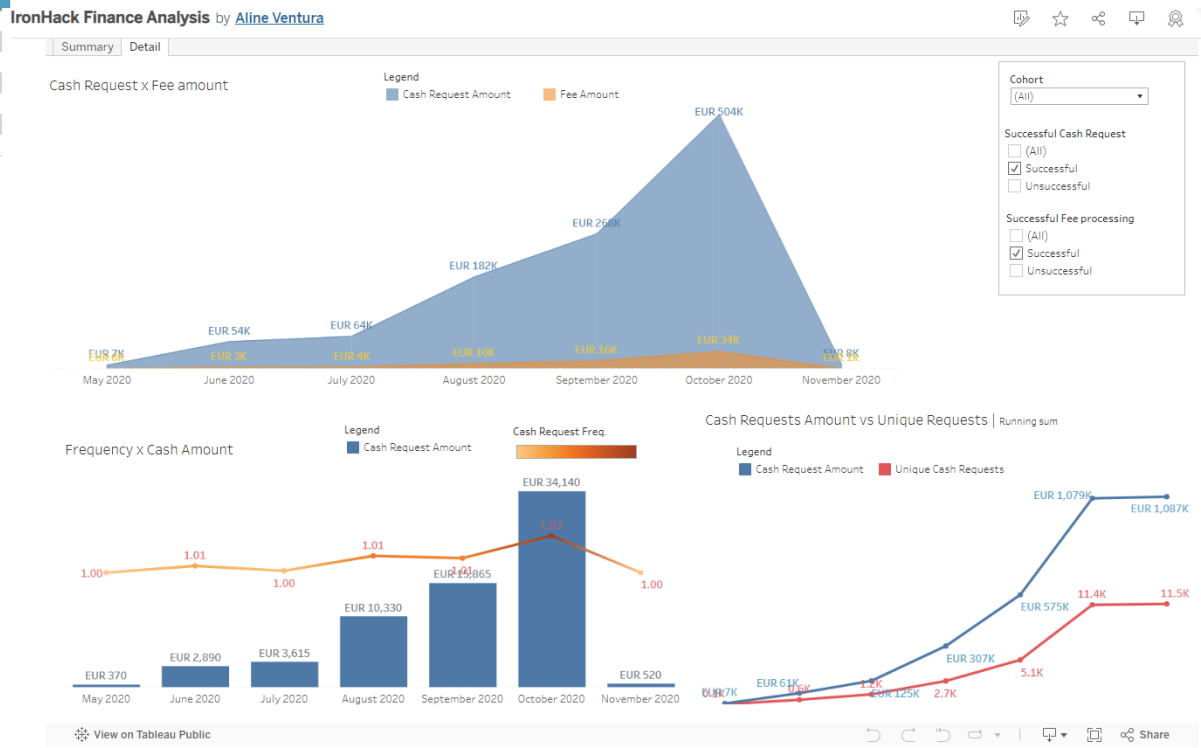




	MoM Cash Requests	MoM UserID	MoM Request Freq.
May 2020			
June 2020	676%	671%	0.6%
July 2020	20%	21%	-0.4%
August 2020	147%	144%	1.3%
September 2020	58%	58%	-0.2%
October 2020	161%	156%	1.9%
November 2020	-98%	-98%	-3.1%

Detailed view:

Distribution of Cash Requests and Fee amount per month.



Pull request to the code:

https://github.com/DSML-bootcamp/project-1-ironhack-payments-en/pull/2

Tableau public link:

https://public.tableau.com/views/IronHackFinanceAnalysis/Summary?:language=en-US&:sid=&:display_count=n&:origin=viz_share_link