# method

## Data preprocessing and splitting divide the research direction into research on merchant problems and research on user problems

## 数据预处理和拆分，将研究方向分为对商家问题的研究和对用户问题的研究

## Research on business problems

## 对商家问题的研究

### Preprocess the data again, and manually classify the types of each business

再次预处理数据，人工分类各个商家的类型

### Determine the type of data to be studied. I classified all the data into explicit and implicit data. For merchants, explicit data includes the distribution of transaction frequency, the distribution of transaction amount, the distribution of transaction time and so on. The implicit parameter is the deeper data explored on the basis of the data through a variety of data analysis algorithms. For merchants, it can be the distribution area and the distribution area of transaction frequency; the changes and potential connection over time can be explored from the distribution of transaction time.

确定要研究的数据类型。我将所有的数据分类为显式数据和隐式数据。对于商家来说，显式数据包括交易频次的分布，交易金额的分布，交易时间的分布等等。而隐式参数就是，通过各种数据分析的算法再显示数据的基础上探索出的更深层次的数据。对商家来说可以是通过交易频次和交易金额分析出的交易金额的分布区间和交易频率的分布区间；可以从交易时间的分布探索出商户的交易量随时间的变化和其中潜在的联系。

### **Subsequent analysis and study on almost all of the available data that have been extracted. What help or services the bank offers to merchants.**

**对已经提取出的几乎所有的可用的数据，进行后续的分析和研究。比如银行对商户提供什么帮助或者服务之类的**

## Research on user problems: For user problems, I plan to divide them into two research directions: 1. Build models for users and evaluate the financial status of users. As a data basis for banks to provide different services to different users later.2. Study the transaction history of users to confirm whether the users has abnormal transactions. I will elaborate separately.

## 对用户问题的研究：对用户问题，我打算分为两个研究方向：1.对用户建立模型，评估用户的财务状况。为银行后续对不同用户提供不同的服务做数据基础。2. 研究用户的交易历史，确认用户是否存在着异常交易的情况。我将分别详细阐述。

### **Users build the evaluation model**

**用户建立评估模型**

#### **analysis of time series时间序列分析**

Time-series analysis can be used to predict the future direction of the account balances. Time series models such as ARIMA (autoregressive integral sliding average Model) can be used to predict future values based on historical data.

If there is fully periodic data (the amount of the transaction, the object, the time is exactly the same), this can be regarded as a user subscription behavior. Subscription behavior can not only reflect the situation of individual users, but also analyze the operation situation of the subscribers based on the change in the number of subscribers

可以使用时间序列分析来预测账户余额的未来走势。时间序列模型如ARIMA（自回归积分滑动平均模型）可以用来预测基于历史数据的未来值。

如果存在完全周期性数据（即交易的金额，对象，时间都完全相同）这种就可以认为是用户存在订阅行为。订阅行为不只是可以反应个人用户的情况，同时也能根据订阅人数的变化，来分析出所订阅商户的运营情况

#### **cluster analysis聚类分析**

K-means or hierarchical clustering analysis was used to group users by their transaction behavior to identify similar consumption patterns or behavioral groups

使用K-means或层次聚类分析来将用户根据其交易行为分组，以识别相似的消费模式或行为群体

#### **regression analysis 回归分析**

Construct regression models to understand how different types of transactions affect changes in account balance. Check if the account often has negative balances, and the duration of the negative balance

构建回归模型来了解不同类型的交易如何影响账户余额的变化。查看账户是否经常出现负余额，以及负余额的持续时间

#### **Cash flow forecast 现金流预测**

Using historical transaction data to predict future cash flow can help assess the user's future financial situation

使用历史交易数据来预测未来的现金流量，可以帮助评估用户未来的财务状况

#### **Analysis of consumption pattern 消费模式分析**

Classification of transactions according to third-party names, and analyze the consumption patterns of users on different merchants or services.

根据第三方名称分类交易，分析用户在不同商家或服务上的消费模式

### **Evaluation and early warning of abnormal transactions: 对异常交易做评估和预警**

#### **Amount of abnormal detection金额异常检测**

Including: 1. Suddenly large transaction 2. The same amount is transferred immediately transferred out

#### **Frequency abnormality detection 频率异常检测**

Including: 1. Frequent small transfers 2. frequent large transactions 3. Centralized high-frequency transactions

#### **Transaction object, abnormal detection 交易对象异常检测**

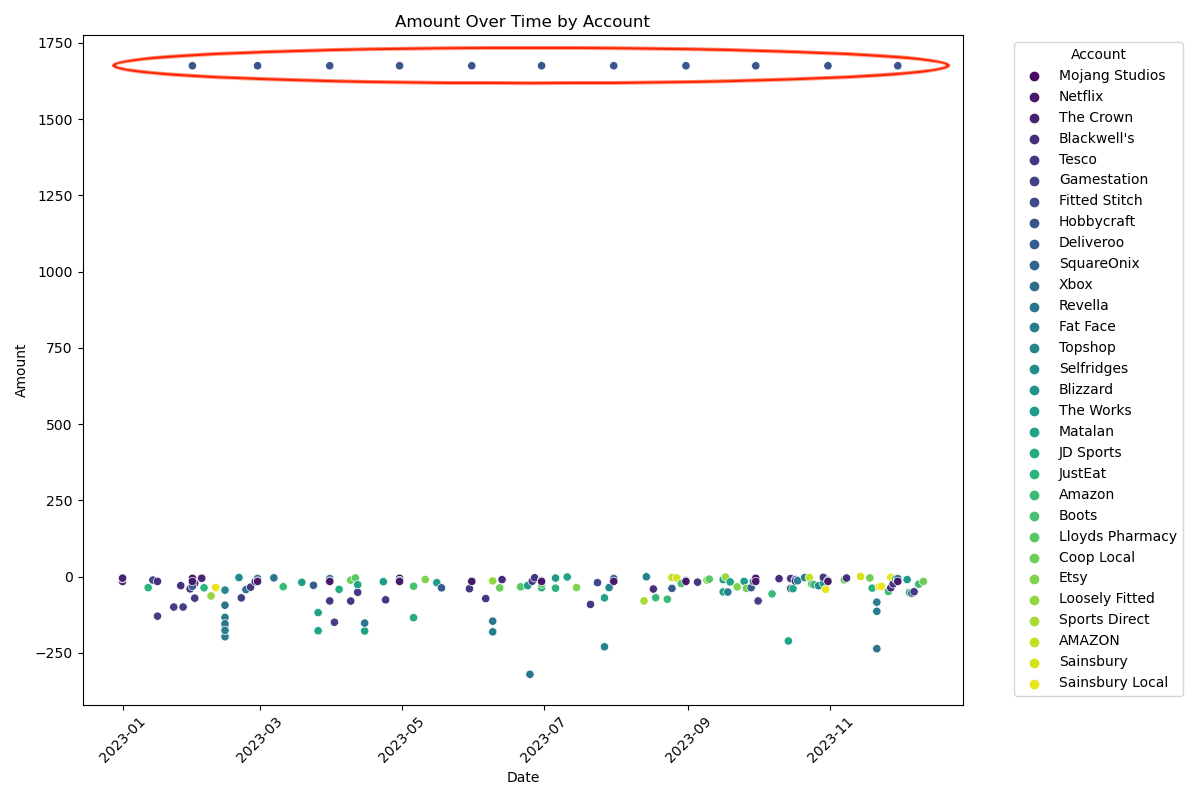
Including: 1. Always transfer money to the same object

#### **Balance abnormality detection 余额异常检测**

Including: 1. Long-term negative balance

# data preprocessing 数据预处理

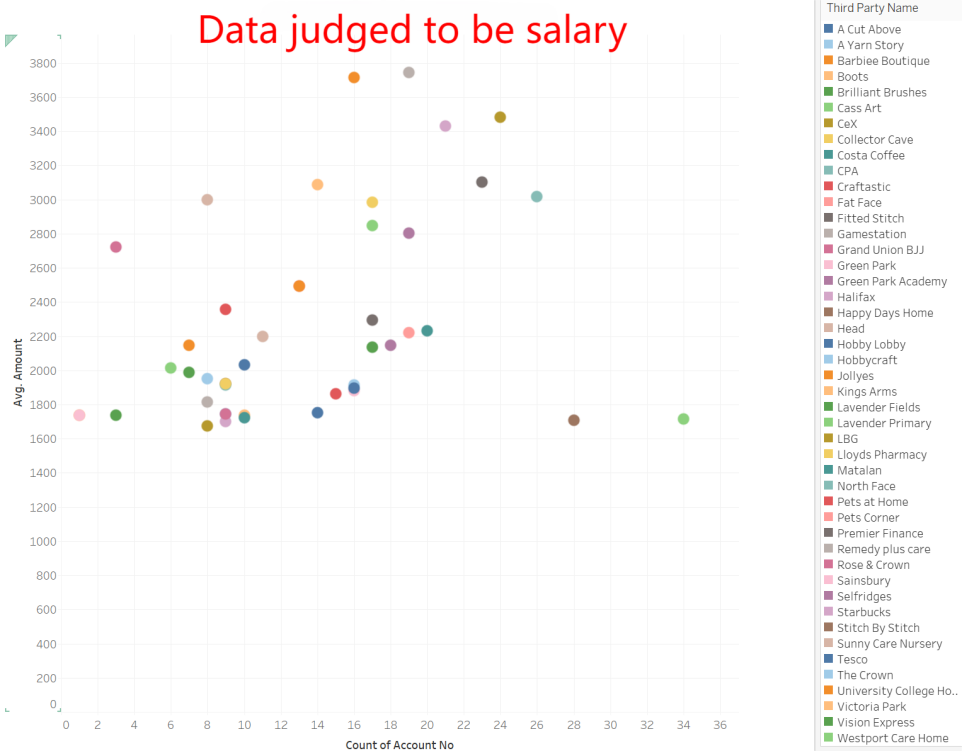
**Add the following to this section:**  
When we visualized the data set, we can find that most accounts have such periodic data of the same amount from the same third-party account, which I think is the salary.



The prerequisite for all data sets is to remove periodic data, because periodic data, especially positive transfers, are generally wages.

# results and discussion

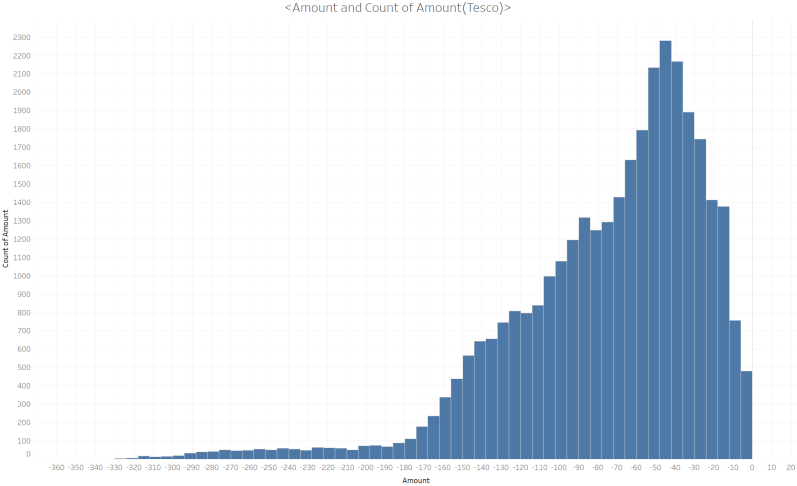
First, the salary data is eliminated, and the salary of different users is made into a separate data set to facilitate the subsequent mapping operations.

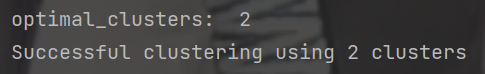
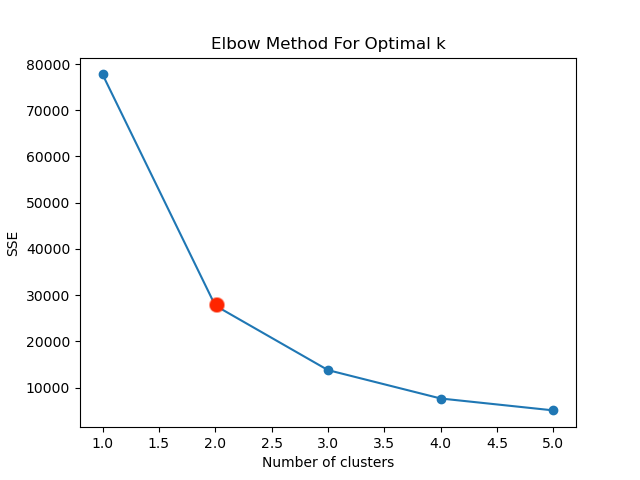


## To the merchant

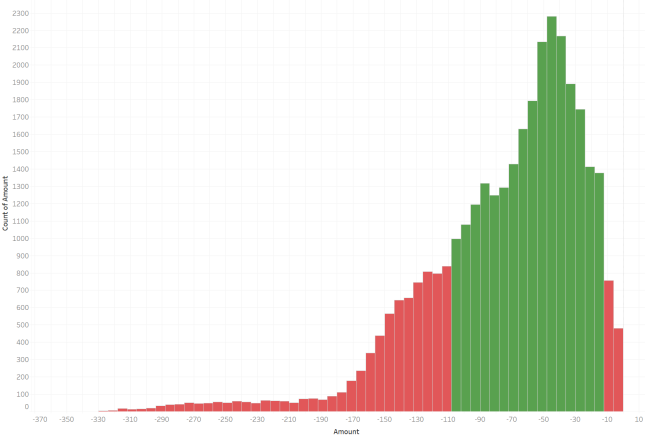
#### **Analysis of transaction amount and transaction frequency:交易金额和交易频次的分析**

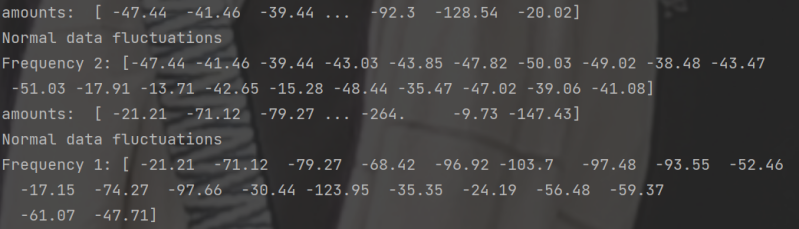
#### **From the distribution of the example transaction amount, we can clearly see the relationship between the transaction amount and the transaction amount of Tesco从举例的交易金额的分布中我们可以清楚的看到Tesco的交易金额和交易额的关系**



Next, I want to explore the best number of clusters. I used the gridded search to explore the SSE value of the model under different hyperparameters, and determine the best number of clusters by finding the best elbow inflection point.接下来我要探索最佳聚类数，我使用网格化搜索，探索不同超参数下模型的SSE值，并通过寻找最佳的肘部拐点确定，最佳的聚类数。  


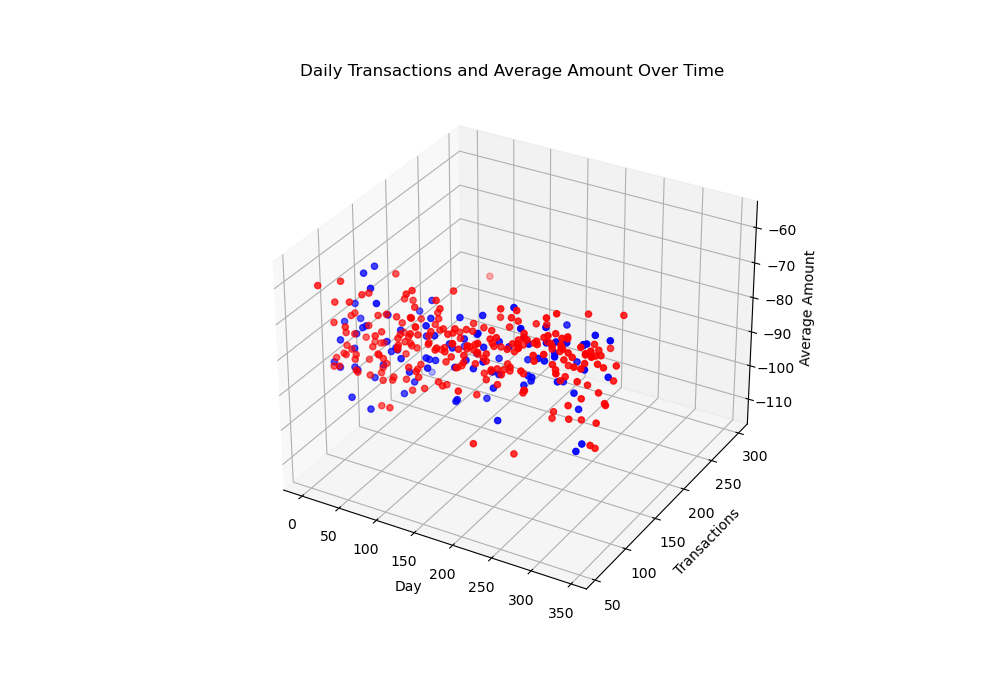
If the best number of clusters is 2, Tesco's high frequency trading quota is between 18 and 108.确定最佳聚类数为2，Tesco的高频的交易额度集中你在18~108之间

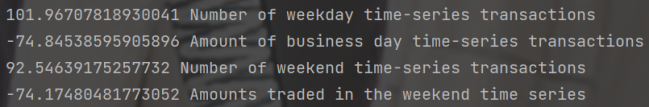


Next, I used the evaluation algorithm model of fusion standard deviation, average absolute deviation and MAD to evaluate the degree of deviation of the data of different clusters. (we can put the formula here).接下来我使用融合标准差，平均绝对偏差和MAD的评估算法模型，对不同聚类的数据进行偏差度的评估，（这里可以放公式）。  
  
 It can be seen that the data fluctuations are normal for both clusters.可以看到两个聚类的数据波动都是正常的

#### Transaction timing analysis:

Analyze the transaction amount and transaction volume of merchants on weekends and weekdays. Or take Tesco as an example分析商家周末和工作日的交易金额和交易量。还是以Tesco为例子





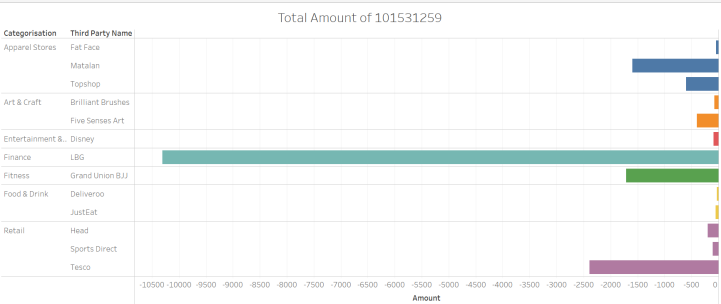
As you can see, Tesco has less transaction volume on weekends, and the average transaction volume is the same.

Next, you can subdivide, discuss the merchant trading patterns and so on

可以看出Tesco，在周末的交易量较少，而且平均交易额是一样的。

接下来可以细分一下，讨论商户的交易模式等等

## To the user



#### The User builds the evaluation model:

* **analysis of time series:**
* **cluster analysis:**
* **regression analysis:**
* **Cash flow forecast:**
* **Analysis of consumption patterns:**

#### **Feature model of user transactions:**

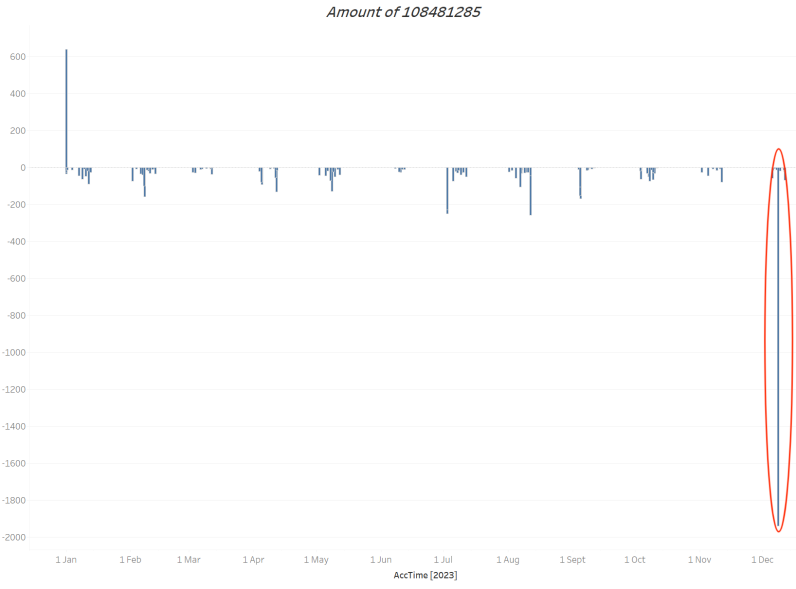
* **Amount anomaly detection:金额异常检测**

1. **Sudden big deal:突然大额交易**

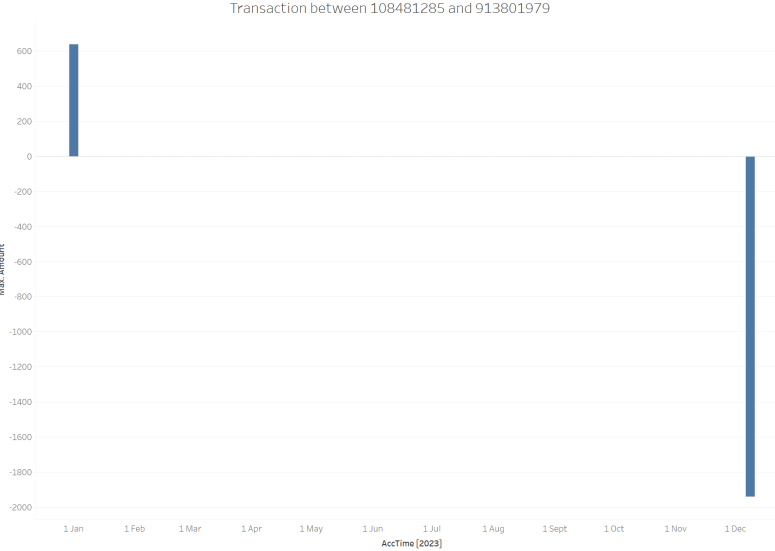
Since we only need to be based on the value of a user's historical account, I think it is enough to use the standard deviation directly for this one-dimensional data (formula and method introduction here), and we can accurately find this sudden large transaction. I deposit all the accounts with such abnormally large transactions in a separate data sheet, because the subsequent procedures will analyze these accounts in depth. Here is an example of an individual user data that I found with outliers.

User 108481285 was detected to have an abnormal large amount of money transfer, triggering the program to detect all of his transactions in the past year.由于只需要基于某个用户的历史account的值，对于这种一维的数据，我认为直接使用标准差就足够了（这里要放公式和方法简介），我们可以准确的找到这种突然大额度的交易。我将所有存在这种异常大额交易的账号存入单独的数据表中，因为后续的程序将要对这些账号进行深入分析。以下以我找到的一个有异常值的个人用户数据举例说明。

用户108481285被检测出有异常大额转账的情况，触发程序的检测他过去一年所有的交易情况。



Sure enough, there is the same abnormal large data, and the next step is to detect the transaction object and historical data of the transaction.as follows果然存在相同的异常大额数据，那下一步就是检测这次交易的交易对象和历史数据。如下



Next, through the algorithm of integrating transaction frequency, transaction span and post-transaction balance, the system comes to the conclusion that the transaction is risky. Perhaps the reason is that the time span is too large, the transaction is not frequent, and the turnover is too large. So the second confirmation transaction application will be sent to the customer.接下来通过融合交易频次，交易跨度，交易后余额的算法，系统得出结论，这笔交易是存在风险的。可能原因是时间跨度过大，不频繁交易，交易额度过大。所以将会给客户发送二次确认交易申请。

1. **Turn in and immediately turn out转进后立刻转出**

To explore this problem, the first thing I have to do is to clean the segmented personal account data set. First, only study the transfer of personal account to personal account, and then calculate which account trades with other personal accounts more than twice in a day. This gives you a list of the following accounts要探索这个问题，我首先要做的就是对切分后的个人账户数据集进行清洗。首先只研究个人账户转个人账户的情况，然后，计算哪个账户在一天内与其他个人张账号的交易频率超过两次。由此得到了以下账户的列表

**[961077907,924045955,874431242,672913754,583579399,492215823,492058070,455052690,398429109,235881534,138609932,122884111]**

Next, improve the time granularity, track the history of these transactions, and build a model to evaluate the transaction anomalies. The evaluation criteria include whether the transaction object is the same, the transaction amount is the same, the transaction time span, the transaction mode (in or out), the historical transaction object, etc., to determine whether there are problems in this transaction.接下来提高时间粒度，追踪这几笔交易历史，建立评估是否存在交易异常的模型。评估标准包括交易对象是否相同，交易金额是否相同，交易时间跨度，交易模式（转入或转出），历史交易对象等，判断这笔交易是否存在问题。

**Here is a typical trading scenario:以下是一种典型的交易情况**

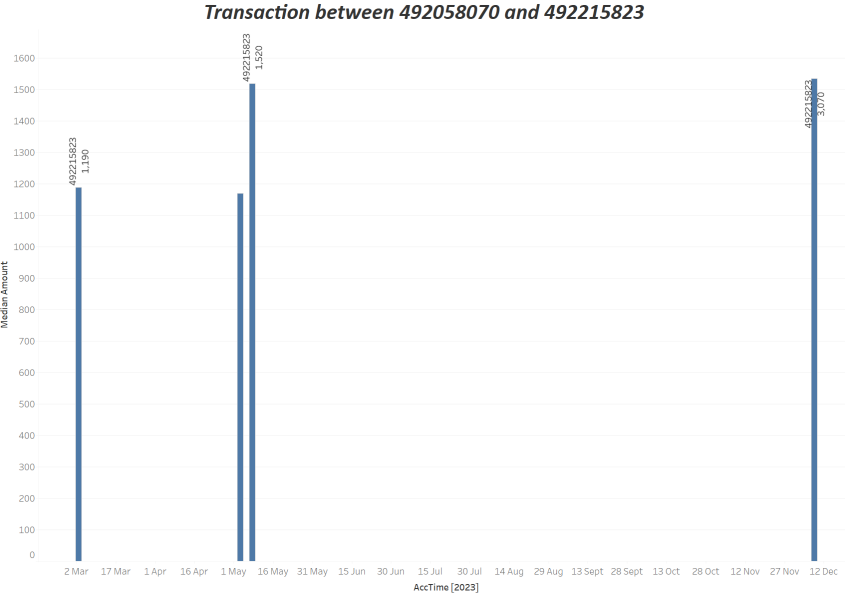


For example, the figure above shows the two transactions between accounts 961077907 and 3984291091 at 18:13:00 / 07 / 05 / 2023 and 18:13:00 / 07 / 05 / 05 / 2023, which are not considered abnormal transactions, the possible reason is that they are transferred into transactions and have a certain time span.例如上图为账号961077907和398429109之间在18:13:00/07/05/2023 和18:13:00/07/05/2023发生的两笔交易，最终并未被认为是异常交易，其可能原因是都是转入交易，而且有一定时间跨度。

**Here is another typical trading scenario:**



For example, the above shows the transaction between accounts 492215823 and 492058070 that is not considered abnormal. The possible reason is that there are historical transfer records between the two accounts, and both are 492215823 to 492058070. The amount is also compared, so it is considered a related account.例如上图为账号492215823和492058070之间在发生的交易，最终并未被认为是异常交易，其可能原因是两个账号之间存在历史转账记录，而且都是492215823转向492058070，金额也对照，所以被认为是有联系的账号。



The results did not detect the transaction characteristics of rapid transfer, but in the course of this study, we also found potential relationships between personal accounts, such as the frequent economic transactions between the owners of many accounts.结果并没有检测到快速转账的交易特征，但是在这个研究过程中，我们同样发现了个人账号之间存在的潜在关系，比如很多账号的拥有人之间存在这频繁规律的经济往来。

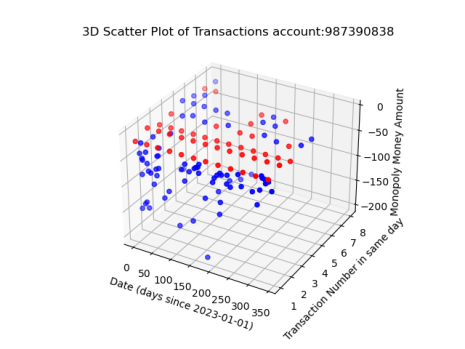
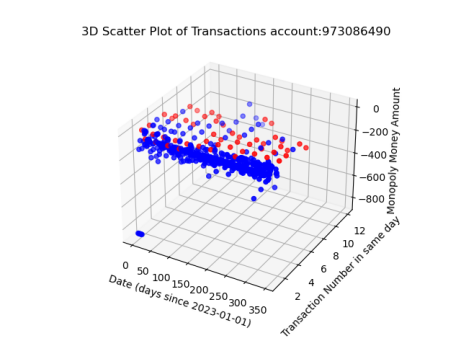
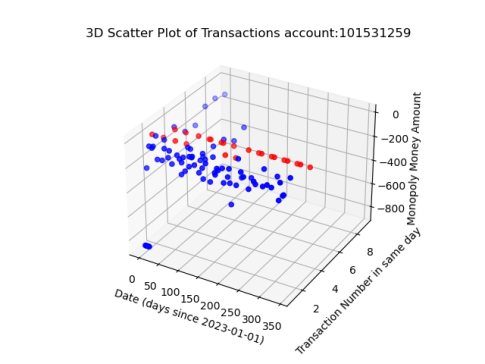
* **Periodic transaction detection:周期性交易检测**

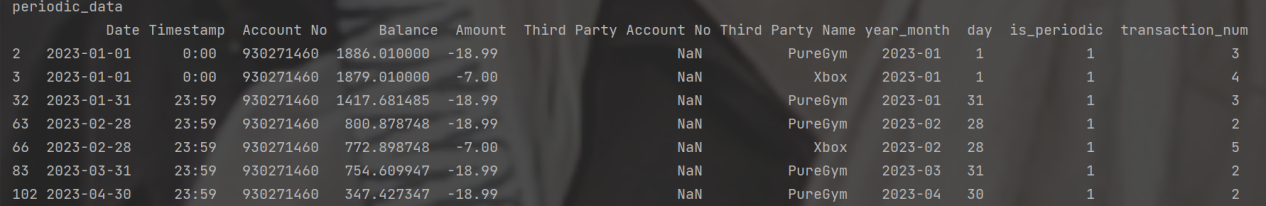
The key to this problem is to study the transaction frequency, which is the timing related data, so the first step is to extract the timing related characteristics of all the account data sets. Including: monthly transaction frequency, cyclical trading data and so on.

The first and most easy to find feature data is the periodic data

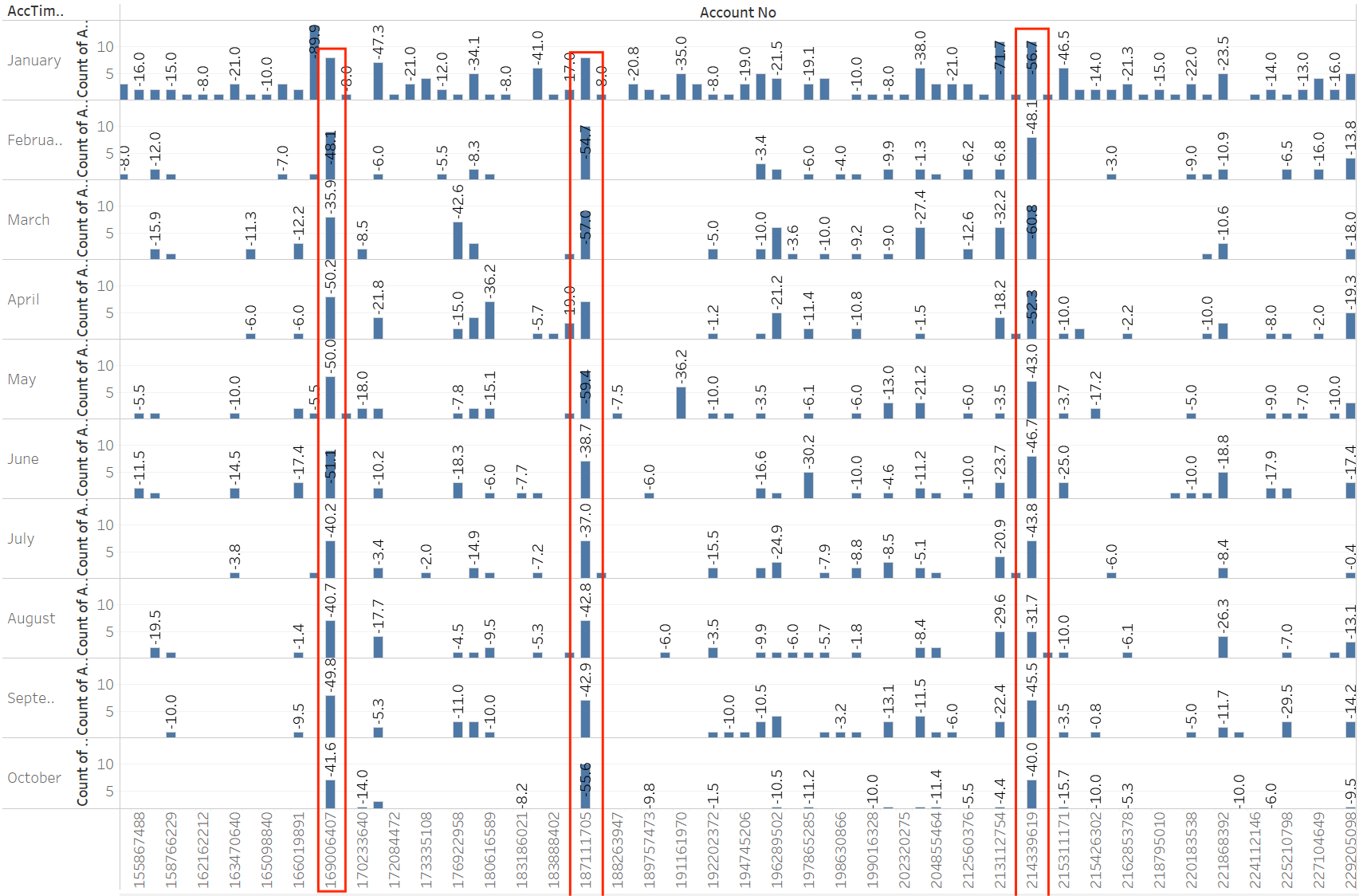
这个问题的关键是研究交易频次，是时序相关的数据，所以第一步就是抽取所有账号数据集的时序相关特征。包括：每月的交易频率，周期性交易数据等等。

首先最容易发现的特征数据就是周期性数据





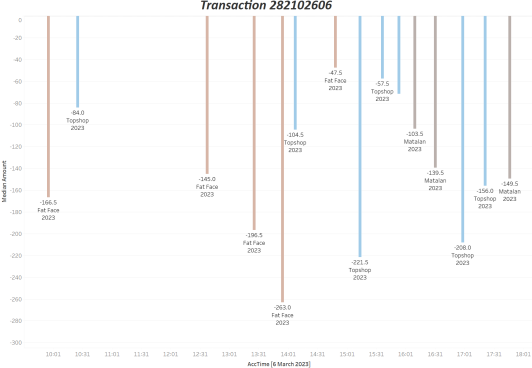
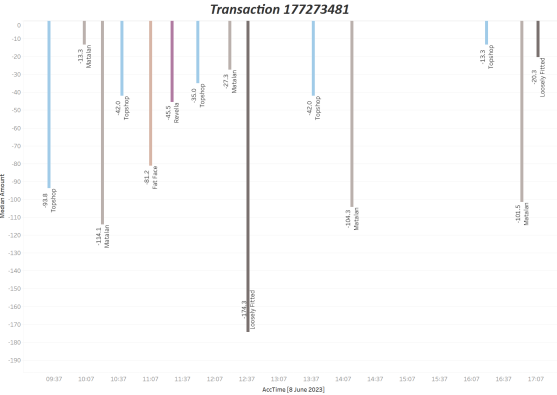
It can be found that most accounts have obvious periodic transaction data, which will be eliminated after the system mark.可以发现，大多数账号都存在明显的周期性交易数据，这些周期性数据经过系统标记后就会被剔除后续的研究。



Similarly, through the monthly transaction frequency, we can find that some accounts have always maintained a high frequency of transactions, etc. Perhaps for this account, Ma Bank may be his main daily account.同样，通过月交易频率我们可以发现，有些账号一直保持着高频率的交易，也许对这个账号来说，小马银行可能是他的主要日常账户。

* **Frequency abnormality detection:频率异常检测**

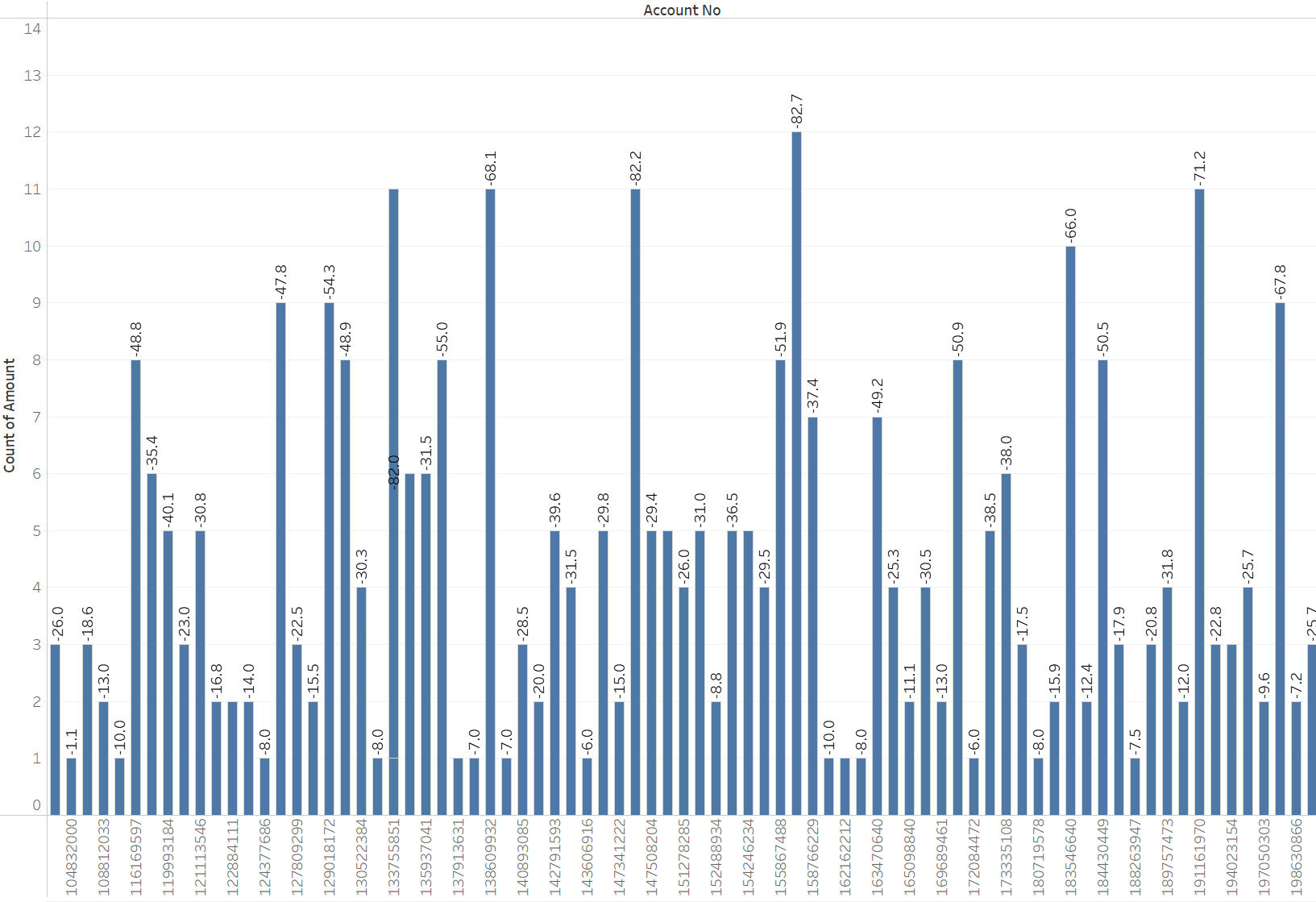
The processed data set was subjected to OPTICS density clustering (adding the relevant formula) to determine the presence of high frequency transactions.对处理后的数据集进行OPTICS密度聚类（添加相关公式解释什么的），确定高频交易存在的区间



This is an example. The first account conducted 14 transactions in the seven hours of 2023.6.8, which is significantly higher than its historical transaction frequency.这就是应该例子，第一个账号在2023.6.8的7个小时中进行了14次交易，明显高于其历史交易频率

1. **Frequent small money transfers频繁的小额转账**

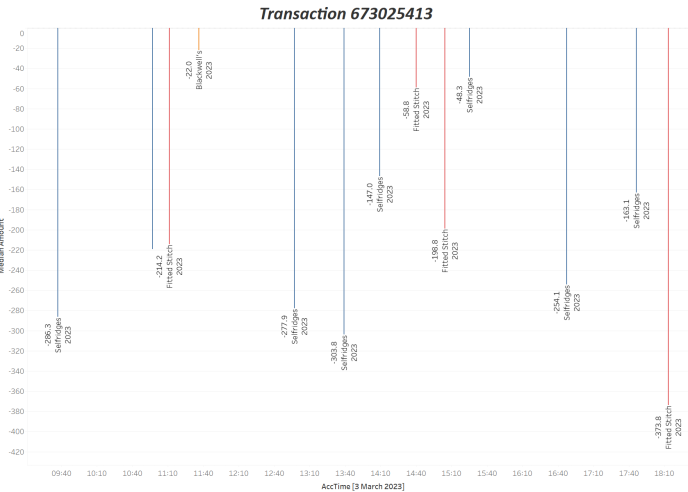
Unfortunately, although the scope of small transactions has been expanded to (-10~10), I have not detected high frequency small transfers in a short period of time.很可惜，尽管已经将小额交易的范围扩大到（-10~10）之间，任然没有检测到我认为的短时间内高频小额转账。



Small transfers for all accounts are within the low frequency range.所有账号的小额转账都在较低的频率范围之内

1. **Frequent large transactions频繁大额交易**

I set the range of high-frequency trading between-150 and-8000, but I did not find what I thought high-frequency large trading in a short period of time.我将高频交易的范围定为了-150~-8000之间，但任然未找到我所认为的短时间内的高频大额交易。

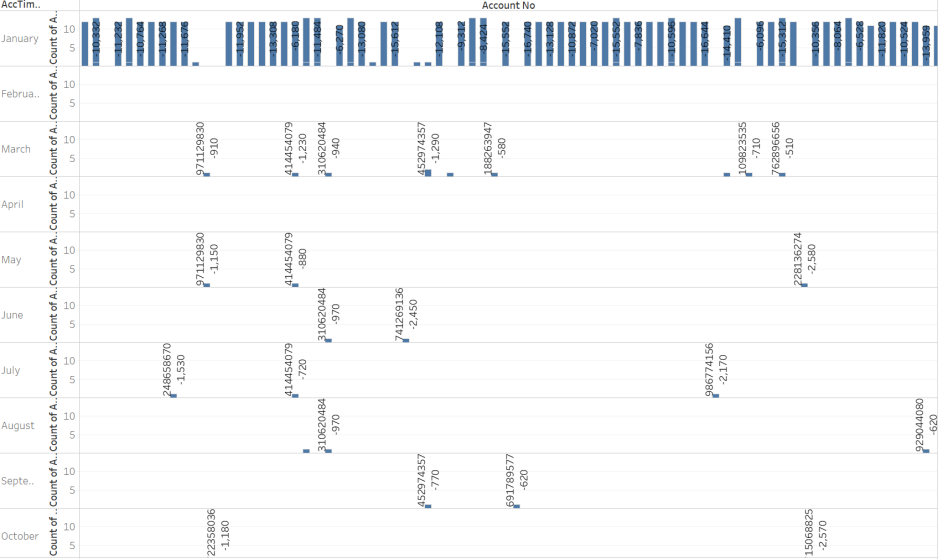


The above is an example. Although the account has traded with high frequency in a short period of time, the average trading amount of the trading account can be analyzed by the model to be considered a normal trading situation.

On the contrary, in the process of research, I found interesting phenomena.

上述就是一个例子，该账号虽然在短时间内进行了高频率的高额交易，但经过模型分析其历史交易额度和频率，已经交易账户的平均交易额度，都可以被认为是正常交易情况。

反而在研究过程中，我发现了有趣的现象。



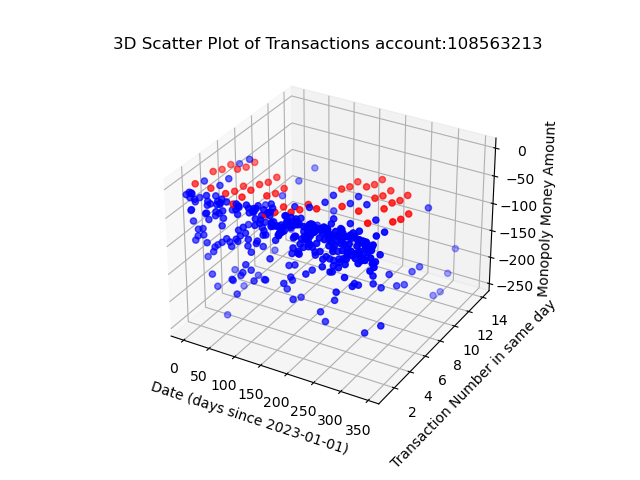
Large amount of transactions occur especially in January, which may be related to holidays, or annual fees, insurance and other annual financial expenses.

大额度的交易在1月份尤其集中发生，推测可能和节假日，或者年费，保险等年度财务支出有关。

1. **Centralized and high-frequency trading集中高频次交易**

Since the first step already uses OPTICS density clustering, determine the interval where HF trading exists. But by this judgment, most of these high-frequency trading ranges have been considered normal trading conditions.

由于第一步已经使用OPTICS密度聚类，确定高频交易存在的区间。但经过上述判断，这些高频交易的区间大多数都已经被认为是正常交易情况了。

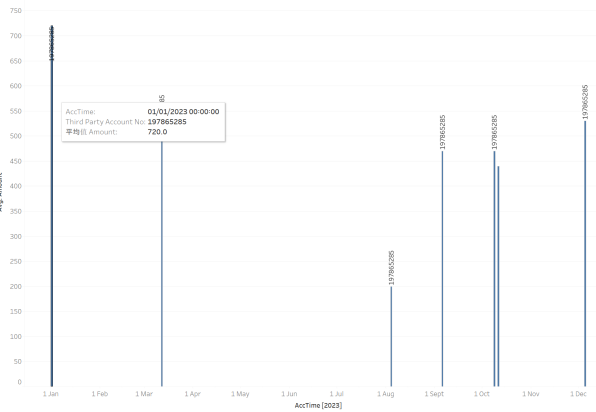
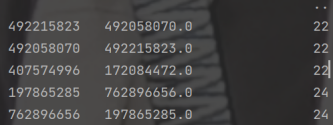


Just like this user, although there are centralized high-frequency transactions, but all through the detection.

就如同这个用户，虽然有集中的高频率交易，但都通过检测了。

* **Transaction object detection:交易对象检测**

1. **Always transfer money to the same object总是给相同对象转账**



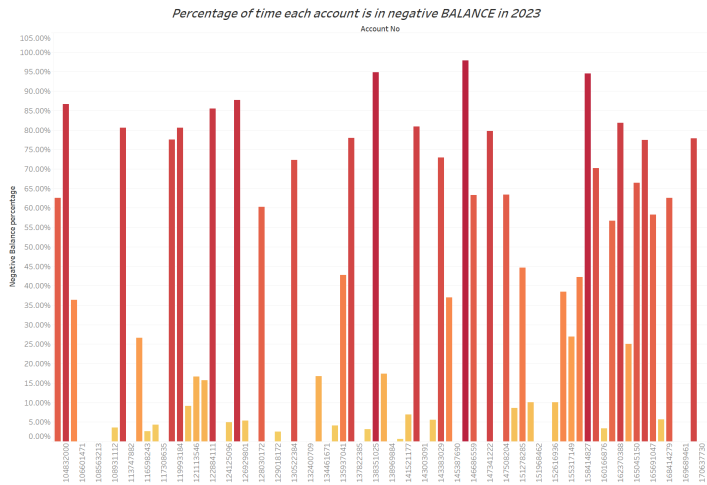
There is frequent transfers between the same personal accounts, it can be considered that there is a potential economic relationship between the two accounts. There is no need to discuss whether there is an abnormality, because all the data in this step have eliminated the periodic transaction data and abnormal transaction data done earlier. This step only discusses the potential data connection on the premise that the transaction data has no abnormality in the timing logic.

存在相同个人账号之间频繁互相转账情况，可以认为两账号之间存在潜在的经济关系。不需要讨论是否存在异常了，因为这一步的所有数据都已经剔除了前面所做的周期性交易数据和异常交易数据。这一步只在交易数据在时序逻辑上无异常的前提下，讨论潜在的数据联系。

* **Balance abnormal detection:余额异常检测:**

1. **In a negative balance for a long time长期处在负余额**

The algorithm logic should be to calculate the date of the first negative balance and the date of the subsequent first positive balance through the segmented personal account data set, and sum counts the proportion of these periods in the one-year period represented by the data set.算法逻辑应该是，通过切分后的个人账户数据集，计算首次出现负数余额的日期和后续首次出现正数余额的日期，sum统计这些时间段在数据集所代表的一年时间段中的占比



We can find that many accounts are in a negative balance state for a long time, and this account should be reminded of risks and limit the large amount of transfers.我们可以发现有很多的账号长期处在负余额的状态，这张账号的应该被风险提醒，以及限制大额度转账。

Thank you very much for your question. Firstly, please give me a little bit of time to briefly describe the process of the whole system.

非常感谢你的提问。首先，请给我一点时间来简述一下整个系统的流程。

First I sliced and diced the dataset by different merchant and personal accounts

首先我按照不同商户和个人账号切分了数据集

Then after visualising the data, I discovered that there was a clear cyclicality in it

然后通过可视化数据后，我发现了其中存在着明显的***周期性数据***

What merchants pay periodically to specific individual accounts is what I consider to be wages. And what an individual account pays a merchant periodically, I consider a subscription.

商家周期性付给特定个人账号的，我认为是***工资***。而个人账号周期性支付给商家的，我认为是***订阅***。

I first culled the feature data to prevent it from interfering with subsequent analyses

我首先剔除了这些特征数据，防止其对后续分析的干扰。

These cyclical data were excluded because, in my opinion, they belong to the characteristic data and all that needs to be analysed subsequently is the non-characteristic data. These cyclical data will be used as a reference for subsequent analyses of the economic status of users and the business situation of merchants.

剔除这些***周期性数据***是因为，我认为这些属于***特征数据***，而后续需要分析的都是***非特征数据***。这些周期性数据将被我用于后续对用户经济状况，以及商户经营情况分析做参考。

The problem was then split into two parts, one analysing user data and one analysing merchant data.

然后，问题就被分成了两部分，一是对用户数据的分析，一个是对商家数据的分析。

For user data, I divided the problem into two parts.1: Build a model for user evaluation.2: Do evaluation and early warning for unusual transactions.

***对用户数据，我将问题分为了两部分。1：建立用户评估模型。2：对异常交易做评估和预警。***

Abnormal amount detection

a.金额异常检测

Including: 1. sudden large transactions 2. the same amount of

包括：1. 突然大额交易 2. 相同金额转进后立刻转出

Frequency anomaly detection

b.频率异常检测

Including: 1. frequent small transfers 2. frequent large transactions 3. concentrated high-frequency transactions

包括：1. 频繁的小额转账 2. 频繁大额交易 3. 集中高频次交易

Transaction object abnormal detection

c.交易对象异常检测

Including: 1. always transferring money to the same object

包括：1. 总是给相同对象转账

Balance anomaly detection

d.余额异常检测

Including: 1. long-term negative balance

包括：1. 长期处在负余额

For merchant data, I divided the problem into two parts.1: A study of displayed data.2: A study of implicit data.

对商家数据，我将问题分为了两部分。1：对显示数据的研究。2：对隐式数据的研究。

Explicit data includes the distribution of transaction frequency, the distribution of transaction amount, the distribution of transaction time and so on. Implicit data is the deeper data explored through various data analysis algorithms on the basis of the displayed data. For merchants, it is the distribution of transaction amount and the distribution of transaction frequency; as well as the change of the merchant's transaction volume over time and the potential connection between them.

显式数据包括交易频次的分布，交易金额的分布，交易时间的分布等等。而隐式数据就是，通过各种数据分析的算法再显示数据的基础上探索出的更深层次的数据。对商家就是交易金额的分布区间和交易频率的分布区间；以及商户的交易量随时间的变化和其中潜在的联系。

Although no obvious anomalies were detected, I think banks should still be alerted to the kind of potentially risky transactions detected. It could be a secondary transaction confirmation, or a request for a pin code

虽然没有检测到明显的异常，但是我认为，银行仍然应该对检测出的那种潜在的风险交易做出提醒。可以是二次交易确认，或者是要求输入pin码

所有我探索的这些数据都是为了使用这些历史数据为用户建立一套评估系统。比如我探索用户是否长期处于负数余额的状态，这是为了确定用户的基本财政状况。

而同样的根据用户的日常交**易额度和频率也可以确定用户的财政情况**，或者是用户是否选择了劳埃德银行作为他日常消费的主要银行。

这些数据都能为后续用户的消费等经济活动提供参考。比如银行可以对长期处于负数存款，但仍然频繁交易的用户提高风险等级，可以避免其申请更大额度的贷款或者出现无法还款的情况。

All of this data I explored was to use this historical data to create an evaluation system for users. For example, I explore whether the user has had a negative balance for a long period of time, this is to determine the user's basic financial situation.

And similarly based on the amount and frequency of the user's daily transactions it is possible to determine the user's financial situation or whether the user has chosen Lloyds Bank as his main bank for daily spending.

All of this data can inform subsequent economic activities such as the user's spending. For example, the bank can raise the risk level of a user who has had negative deposits for a long period of time, but still transacts frequently, and can prevent him from applying for a larger loan or from being unable to make repayments.

***OPTICS是一种密度聚类算法***，用于发现不同密度区域中的自然聚类。OPTICS特别适合处理密度不均匀或规模较大的数据集，通过这个我们可以轻易的找到交易频率集中，大额度交易集中，小额度交易集中的区域。

我只是将这个算法用在了3维尺度上。（时间，交易对象，交易额度）

OPTICS is a density clustering algorithm for discovering natural clusters in regions of different densities. OPTICS is particularly suitable for dealing with datasets of uneven density or large size, through which we can easily find regions of concentrated transaction frequency, concentrated large-value transactions, and concentrated small-value transactions.

I have just used this algorithm on a 3 dimensional scale.(time, object of transaction, amount of transaction)