

数据仓库与数据挖掘

第一次作业

姓名：王宸昊

学号：2019214541

时间：2019.10.19

一、 数据预处理

首先，先通过观察，分析几张表当中数据存在的噪音类型，然后通过脚本针对数据中的噪音进行分类，再通过脚本进行处理。在这里通过 Python 的脚本对数据进行分析 and 处理，其中利用 pandas 库对 csv 文件进行数据分析。

首先 8 张表中都存在的问题：

1. 数据重复：在每张表中都存在不同程度的数据重复的情况，即在每张表中都存在一些行的数据冗余，在这里处理的方法是 pandas 中的 `drop_duplicates()` 方法，在导入 csv 文件后先进行一步去重的工作。
2. 外键依赖：在例如 account 等表中存在 `district_id` 等字段，该字段对应 district 表当中的 id，所以对于 account 表中的 `district_id` 都应该对应 district 真实存在的数据。所以在对每个存在外键依赖的表中，进行检查，如果外键在对应的表中不存在，则直接将这行删除。

其他表中数据存在的问题和预处理方法总结如下：

1.1 account:

1. frequency 字段不规范：将 `POPLATEKMESICNE` 改写为 `POPLATEK MESICNE`。
2. account_id 字段有错误：account_id 应该是 int 类型，但是有几条数据有错误，例如 `1675_[typo]` 等形式，解决方案为去掉后面的部分，只保留前面的数字。
3. 将 date 字段转为 date 格式。

1.2 card:

1. type 字段有错误：将拼写为 `golden` 修改为 `gold`

2. 将 issued 字段转为 date 格式。

1.3 client:

1. Birth_number 字段中数据有错误，超过了数据含义的范围：将错误的数据直接删除。
2. 将 Birth_number 字段根据含义拆分为 birth_day 和 gender 字段：根据月份+50 判定男女。

1.4 disp:

1. 外键不存在

1.5 district:

1.6 loan:

1. duration 字段不符合规范：duration 的值应该是 12 的整倍数，但是部分不足，对于不满足的数据直接删除。
2. payments 字段不符合含义：表中的数据应该满足 $\text{payments} \times \text{duration} = \text{amount}$ ，检查每行数据，不符合的直接删除。
3. 将 date 字段转为 date 格式。

1.7 order:

1. 外键不存在

1.8 trans:

- 1. 将 date 字段转为 date 格式
- 2. balance 字段不规范：balance 字段的值应该是数字类型，部分数据非数值，直接删除。
- 3. Bank 字段值为空：部分数据 Bank 的字段为 null,直接删除。
- 4. 外键不存在

二、 数据导入

- 1. 新建数据库 homework，需要注意将不允许修改表结构选项关闭
- 2. 将 CSV 当中的数据导入该数据库中，共新建八张表，表的结构如下：

account 表：

WIN-AIMR82K3V72.h...ork - dbo.account			
	列名	数据类型	允许 Null 值
PK	account_id	int	<input type="checkbox"/>
	district_id	int	<input type="checkbox"/>
	frequency	varchar(50)	<input type="checkbox"/>
	date	date	<input type="checkbox"/>
			<input type="checkbox"/>

Card 表：

WIN-AIMR82K3V72....mework - dbo.card			
	列名	数据类型	允许 Null 值
PK	card_id	int	<input type="checkbox"/>
	disp_id	int	<input type="checkbox"/>
	type	varchar(50)	<input type="checkbox"/>
	issued	date	<input type="checkbox"/>
			<input type="checkbox"/>


Client 表：

WIN-AIMR82K3V72.h...ork - dbo.client*		WIN-AIMR82K3	
	列名	数据类型	允许 Null 值
	client_id	int	<input type="checkbox"/>
	district_id	int	<input type="checkbox"/>
	birth_day	date	<input type="checkbox"/>
	gender	varchar(50)	<input type="checkbox"/>
			<input type="checkbox"/>


Disp 表:

WIN-AIMR82K3V72.h...ework - dbo.disp*		WIN-AIMR82	
	列名	数据类型	允许 Null 值
	disp_id	int	<input type="checkbox"/>
	client_id	int	<input type="checkbox"/>
	account_id	int	<input type="checkbox"/>
	type	varchar(50)	<input type="checkbox"/>
			<input type="checkbox"/>

District 表:

WIN-AIMR82K3V72.ho...rk - dbo.district*		WIN-AIMR8	
	列名	数据类型	允许 Null 值
	district_id	int	<input type="checkbox"/>
	district_name	varchar(50)	<input type="checkbox"/>
	region	varchar(50)	<input type="checkbox"/>
	hab_number	int	<input type="checkbox"/>
	city_number	int	<input type="checkbox"/>
	ave_salary	int	<input type="checkbox"/>
	umemploy_rate	float	<input type="checkbox"/>
	crime_number	int	<input type="checkbox"/>
			<input type="checkbox"/>

Loan 表:

WIN-AIMR82K3V72.h...ework - dbo.loan*		WIN-AIMR82	
	列名	数据类型	允许 Null 值
	loan_id	int	<input type="checkbox"/>
	account_id	int	<input type="checkbox"/>
	date	date	<input type="checkbox"/>
	amount	int	<input type="checkbox"/>
	duration	int	<input type="checkbox"/>
	payments	int	<input type="checkbox"/>
	status	varchar(50)	<input type="checkbox"/>
	payduration	int	<input type="checkbox"/>
			<input type="checkbox"/>

Order 表:

WIN-AIMR82K3V72.h...ework - dbo.order				WIN-AIMR82
	列名	数据类型	允许 Null 值	
PK	order_id	int	<input type="checkbox"/>	
	account_id	int	<input type="checkbox"/>	
	bank_to	varchar(50)	<input type="checkbox"/>	
	account_to	int	<input type="checkbox"/>	
	amount	int	<input type="checkbox"/>	
	k_symbol	varchar(50)	<input checked="" type="checkbox"/>	
			<input type="checkbox"/>	

Tran 表:

WIN-AIMR82K3V72.h...ework - dbo.tran*				WIN-AIMR82
	列名	数据类型	允许 Null 值	
PK	trans_id	int	<input type="checkbox"/>	
	account_id	int	<input type="checkbox"/>	
	date	date	<input type="checkbox"/>	
	type	varchar(50)	<input type="checkbox"/>	
	operation	varchar(50)	<input type="checkbox"/>	
	amount	int	<input type="checkbox"/>	
	balance	int	<input type="checkbox"/>	
	k_symbol	varchar(50)	<input checked="" type="checkbox"/>	
	bank	nchar(10)	<input type="checkbox"/>	
	account	int	<input type="checkbox"/>	
			<input type="checkbox"/>	

三、 数据库设计

第一问中需要从不同性别、不同地区、不同信用卡类型进行分析。

第二问中需要从年龄阶段、不同性别、不同地区、交易的类型进行分析。

所以首先建立以下维表:

3.1 维表:

性别维度: 从性别维度分析, 则性别需要作为一个维度, 该有 2 个属性值, 属性值有 2 个可能: male 和 female。

WIN-AIMR82K3V72....rk - dbo.Table_1*			
列名	数据类型	允许 Null 值	
gender_id	int	<input type="checkbox"/>	
gender_name	char(50)	<input type="checkbox"/>	
		<input type="checkbox"/>	

地区维度：地区维度可以直接使用原来的 district 表。

WIN-AIMR82K3V72.ho...rk - dbo.district*			
列名	数据类型	允许 Null 值	
district_id	int	<input type="checkbox"/>	
district_name	varchar(50)	<input checked="" type="checkbox"/>	
region	varchar(50)	<input checked="" type="checkbox"/>	
hab_number	int	<input checked="" type="checkbox"/>	
city_number	int	<input checked="" type="checkbox"/>	
ave_salary	int	<input checked="" type="checkbox"/>	
umemploy_rate	float	<input checked="" type="checkbox"/>	
crime_number	int	<input checked="" type="checkbox"/>	
		<input type="checkbox"/>	

卡类型维度：卡类型共有三种值：classic,junior,gold。因此建立 card_type 维表

WIN-AIMR82K3V72.h...k - dbo.card_type			
列名	数据类型	允许 Null 值	
type_id	int	<input type="checkbox"/>	
type_name	char(10)	<input type="checkbox"/>	
		<input type="checkbox"/>	

Client 维表：包含 3 个属性，其中两个外键连接到 district 表，一个连接到 gender 表。其中 age 字段做一个函数聚合，将 2000 减去出生日期，按照 10 年为一个阶梯，将不同年龄段的人分开。

列	别名	表	输出
client_id		client	<input checked="" type="checkbox"/>
district_id		client	<input checked="" type="checkbox"/>
gender_id		client_gen...	<input checked="" type="checkbox"/>
CAST((2000 - YEAR(dbo.client.birth_day)) / 10...	age		<input checked="" type="checkbox"/>
			<input type="checkbox"/>
			<input type="checkbox"/>
			<input type="checkbox"/>
			<input type="checkbox"/>

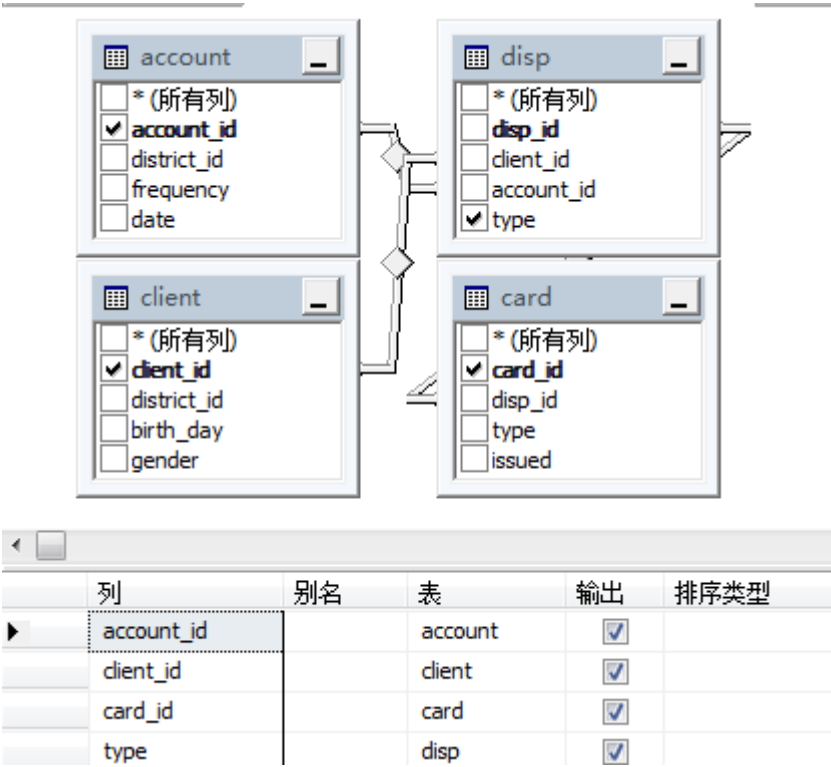
Card 维表：包含 3 个属性，两个外键连接到 crad 表格 card_type 表，还有一个 issued 字段。

	列	别名	表	输出
▶	card_id		card	<input checked="" type="checkbox"/>
	disp_id		card	<input checked="" type="checkbox"/>
	type_id		card_type	<input checked="" type="checkbox"/>
	issued		card	<input checked="" type="checkbox"/>

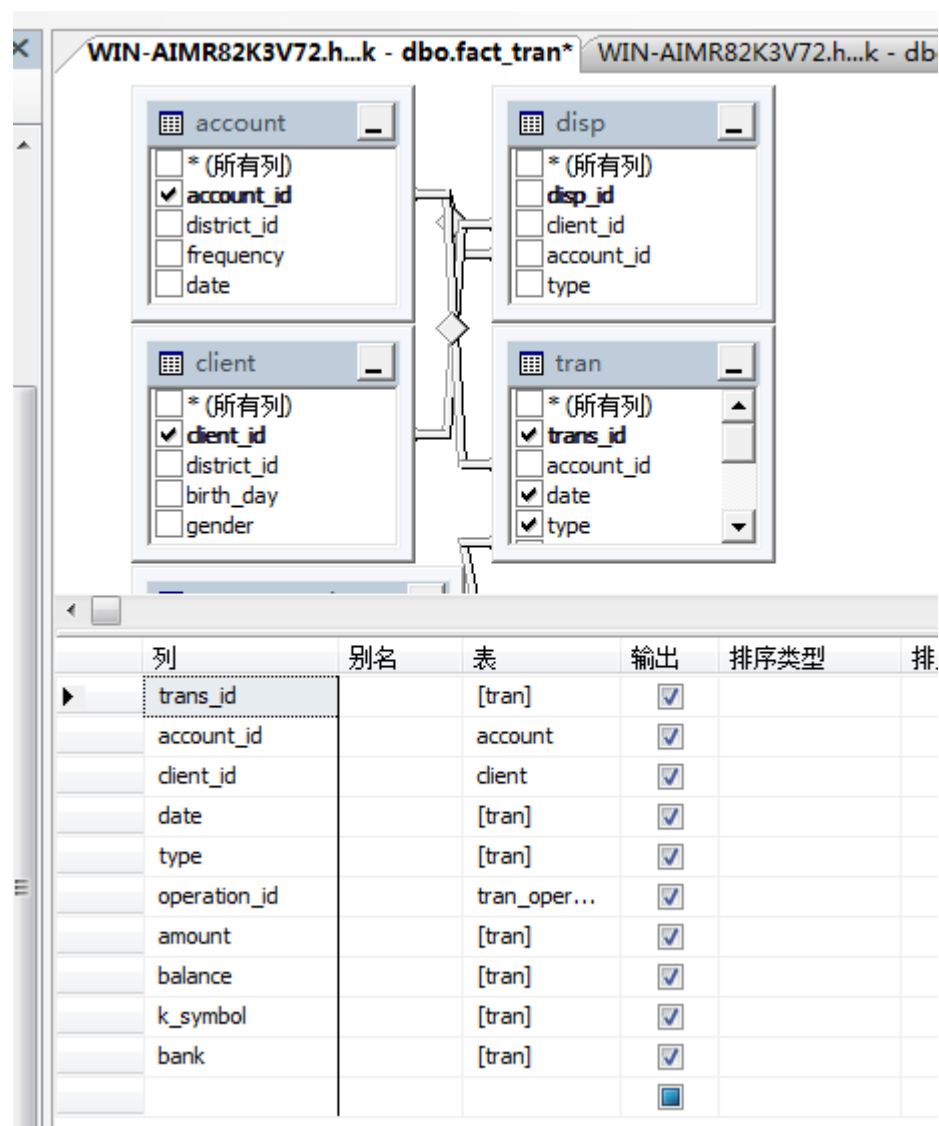
3.2 事实表:

Disp_fact 事实表：根据 disp 表当中的数据进行聚合，将用于分析的数据聚合到本表中，其中的数据来源于 account,client,card,disp 表。

事实表的结构设计如下：

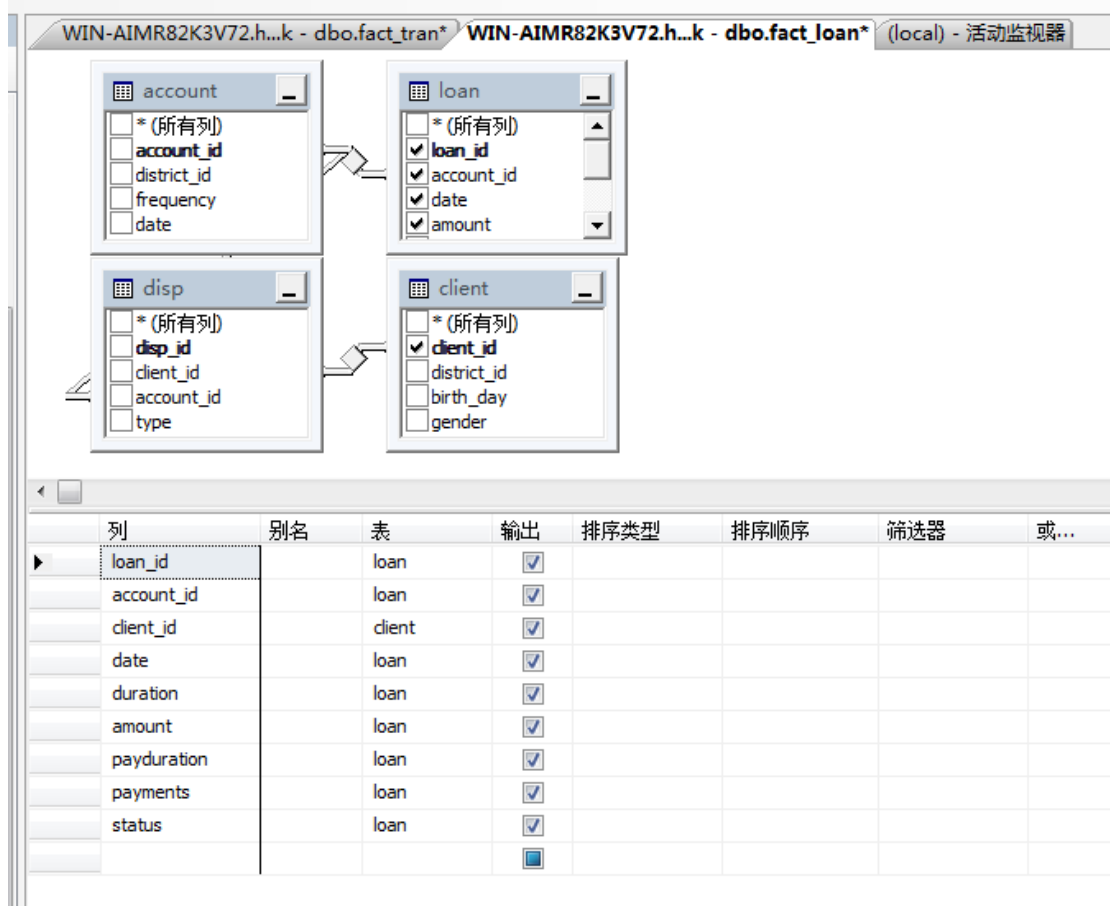


Tran_fact 事实表：该表中主要聚合记录交易信息，数据来源于 account、disp、client 等，进行第二题的分析



Loan_Fact 事实表：主要记录贷款相关的信息，用于第三问的数据分析，具体的设计如下“

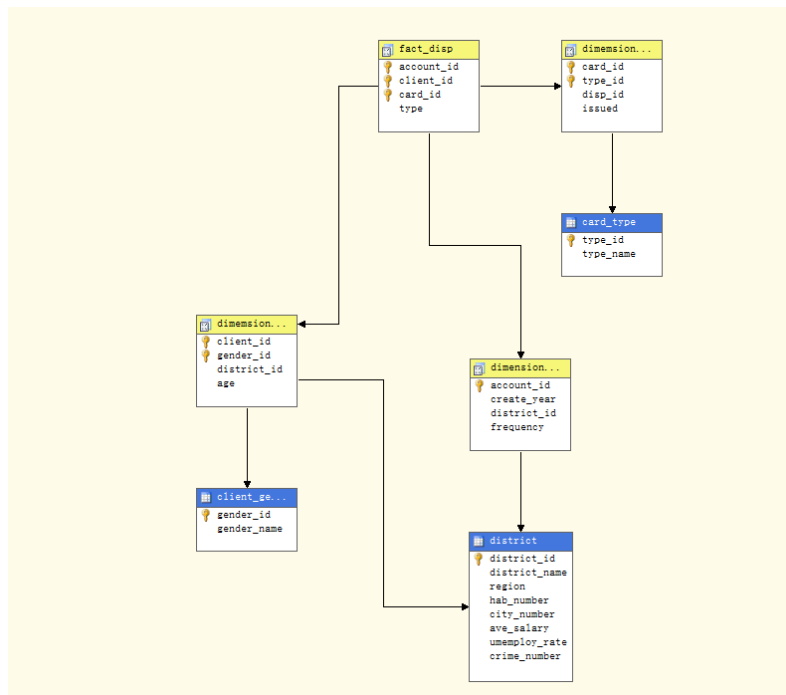
”



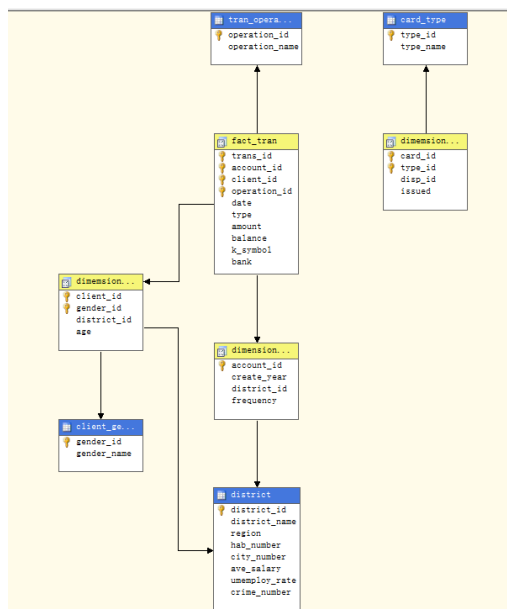
四、 数据立方体设计

第一问：第一问中要对用户进行分析，所以使用 fact_disp 事实表进行分析，从地区、性别、卡类别、年龄四个维度进行分析。

数据仓库的设计如下：



第二问：第二问主要针对 trans 表当中的数据进行分析， 主要对数据当中不同的操作类别进行分析， 量度为收入和支出的数目， 同时还需要额外进行一次运算， 计算出来净支出， 所以涉及到的事实表为 fact_tran， 维度由 4 个维表当中的属性所规定， 设计如下。



第三问： 由于第三问主要是针对贷款的情况进行分析， 用于分析的事实表为 fact_loan 表， 量度为贷款总数， 观察的维度包括年龄段、性别、地区、信用卡的

```
graph TD
    fact_loan[fact_loan] --> dimension_client[dimension_client]
    fact_loan --> dimension_account[dimension_account]
    dimension_client --> client_gender[client_gender]
    dimension_client --> district[district]
    dimension_account --> district
```

fact_loan

- loan_id
- client_id
- account_id
- date
- duration
- amount
- payduration
- payments
- status

dimension_client

- client_id
- gender_id
- district_id
- age

client_gender

- gender_id
- gender_name

dimension_account

- account_id
- create_year
- district_id
- frequency

district

- district_id
- district_name
- region
- hab_number
- city_number
- ave_salary
- unemploy_rate
- crime_number

5.1 题目一:

设定维度为 Type Name, Gender Name, District Name,,Age。量度为 Client 的数量。

[illegible]

从表中的数据可以分析得出，从性别上看，50 岁以上的男性用户要略多于女性用户。在所有的年龄段中，80-90 岁的人用户的人数是最少的，对于卡的类型同时随着年龄的增长，用户数量越来越少，但是卡的级别越来越高。针对不同地区，由于地区的名称过多，数据比较分散，不利于分析，但是从整体上看，男性用于要高于女性用户。

将筛选条件拖至此处											
		District Name ▾ Age ▾									
		Benasov	Beroun	Blatno	Breclav	Brno - mesto	Brno - venkov	Bruntal	Ceska Lipa	Ceske Buz	
Gender Name ▾	Type Name ▾	Dimension Client 计数	Dimension Client 计数	Dimension Client 计数	Dimension Client 计数	Dimension Client 计数	Dimension Client 计数	Dimension Client 计数	Dimension Client 计数	Dimension Client 计数	Dimension Client 计数
Female		7	12	11	9	25	19	9	9	11	
Male		10	14	11	11	30	18	15	15	8	
总计		17	26	22	20	55	37	24	24	19	

5.2 题目二:

问题：针对不同的年龄阶段，不同性别，不同地区的用户，分析交易的收入（trans.type=PRIJEM）情况，支出情况（trans.type=VYDAJ），净支出（所有支出减去所有收入）状况。

将筛选条件拖至此处																		
		Operation Name ▾		District Name ▾														
		PRÁVOD NA UCET		Blanko	Preclav	Prno - mesto	Prno - venkov	Bruntal	Ceska Lipa	Ceske Budejovice	Cesky Krumlov	Chab	Chomutov	Chrudim	Decin	Domazlice	Frydek - Mistek	Havlic
Gender Name ▾	Region	Amount	Amount	Amount	Amount	Amount	Amount	Amount	Amount	Amount	Amount	Amount	Amount	Amount	Amount	Amount	Amount	Amount
Female	central Bohemia	2243325	2037286															
	east Bohemia								2827198					2221038				163957
	north Bohemia												1779748		3467082			
	north Moravia							2278011								6238959		
	Prague									1730336	3204122							
	south Bohemia			3002173	2316277	7252876	3498480											
	south Moravia												2805892			2541513		
	west Bohemia												2805892	1779748	2221038	3467082	2541513	6239595
	汇总	2243325	2037286	3002173	2316277	7252876	3498480	2278011	2827198	1730336	3204122		2805892	1779748	2221038	3467082	2541513	6239595

从上图可以分析，就地区而言 south_Moravia 的收入和支出的金额是最多的。就性别而言，女生的交易额是高于男生的。但是这里并没有完全达到题目的要求，并没有按照交易的类型，对净支出进行分析。

5.3 题目三:

问题：假设你是一个银行的数据分析人员，银行希望能够放出更多的贷款。请以这个作为目标，选择合适的维度进行分析，告知管理人员应该把优惠政策和宣传力度集中到哪类人群。

在数据立方体 Homework-3 当中，选取了 3 个量度进行评判放贷的情况，分别

是 amount 贷款总额、duration 分期还款数，payduration 记录已还款期数。之所以从这三个角度分析，是因为从贷款总额中可以判断出一个人的贷款总的预期，而通过 duration 和 payduration 可以估计出某个人的偿还能力，综合以上者三个量度可以更好地判断放贷的情况。

为了展示的更清楚，首先从客户的个人属性：年龄和性别的角度观察，如下图：

将列字段拖至此处																						
Age		10-20			20-30			30-40			40-50			50-60			60-70			70-80		
Region	Amount	Payduration	Duration	Amount	Payduration	Duration	Amount	Payduration	Duration	Amount	Payduration	Duration	Amount	Payduration	Duration	Amount	Payduration	Duration	Amount	Payduration	Duration	
central Bohemia	16196460	1780	3696	16196460	1780	3696	16196460	1780	3696	16196460	1780	3696	16196460	1780	3696	16196460	1780	3696	16196460	1780	3696	
east Bohemia	17384244	2140	4272	17384244	2140	4272	17384244	2140	4272	17384244	2140	4272	17384244	2140	4272	17384244	2140	4272	17384244	2140	4272	
north Bohemia	9232716	1453	2484	9232716	1453	2484	9232716	1453	2484	9232716	1453	2484	9232716	1453	2484	9232716	1453	2484	9232716	1453	2484	
north Moravia	20921844	2559	4776	20921844	2559	4776	20921844	2559	4776	20921844	2559	4776	20921844	2559	4776	20921844	2559	4776	20921844	2559	4776	
Prague	16725900	1556	3564	16725900	1556	3564	16725900	1556	3564	16725900	1556	3564	16725900	1556	3564	16725900	1556	3564	16725900	1556	3564	
south Bohemia	11198028	1392	2604	11198028	1392	2604	11198028	1392	2604	11198028	1392	2604	11198028	1392	2604	11198028	1392	2604	11198028	1392	2604	
south Moravia	23923548	3022	5808	23923548	3022	5808	23923548	3022	5808	23923548	3022	5808	23923548	3022	5808	23923548	3022	5808	23923548	3022	5808	
west Bohemia	9457068	1285	2592	9457068	1285	2592	9457068	1285	2592	9457068	1285	2592	9457068	1285	2592	9457068	1285	2592	9457068	1285	2592	
总计	125039808	15187	29796	125039808	15187	29796	125039808	15187	29796	125039808	15187	29796	125039808	15187	29796	125039808	15187	29796	125039808	15187	29796	

从年龄上看如下图：

大量的贷款的额度的用户主要集中在 20-60 岁之间, 其中数量最大的 30-40 岁, 而且一般这个年龄段的人贷款的周期比较长, 偿还能力也比较高, 所以将贷款的宣传作用在年龄段在 30-50 左右的人群是比较合适的。

就性别而言，在 30-50 这个年龄段的女性的贷款能力明显优于

从地区的角度观察如下图：

将列字段拖至此处			
Region	Amount	Payduration	Duration
central Bohemia	16196460	1780	3696
east Bohemia	17384244	2140	4272
north Bohemia	9232716	1453	2484
north Moravia	20921844	2559	4776
Prague	16725900	1556	3564
south Bohemia	11198028	1392	2604
south Moravia	23923548	3022	5808
west Bohemia	9457068	1285	2592
总计	125039808	15187	29796

可以看到就贷款的总数上来说，north Moravia 的总贷款数是最多的，而且已经偿还的比例也相当高，同时贷款的周期还比较短。因此综合以上三个因素，在地域的选择上，在 north Moravia 大力宣传的重要性是非常高的。

六、 实验总结

终于在最后一刻“算是”完成了作业。回顾做作业的过程，感觉时间分配的并不是非常合理，在这里总结和反思一下这个大作业的过程。

首先本次作业我认为可以主要分为三个部分，数据预处理、数据仓库设计、数据分析。在数据预处理的部分，我花费了比较多的时间，通过写 python 的脚本，对每个异常数据进行检查，在这个过程中也是比较投入的。

但是在后来 SQLServer 的使用上遇到了比较大的问题，首先是对任务目标的不了解，并不知道建立一个数据仓库需要的基础知识有哪些，PPT 上的介绍也比较模糊，并不知道如何下手，而且对维表事实表等概念的理解并不深刻，并不知道设计一个好的数据仓库的结构，所以在最后的设计与实验中并没有完全完成任务，也是希望能在具体的实践操作中能得到多一些的指导，这样做的时候就不会感到迷茫了。