

# 运营商用户数 据分析

Operator user data  
analysis

汇报人：张浩宸  
数据分析：冯朝晨 王迟迟  
制作：徐思涵 万程冰

# 「成员分工」

组长：

冯朝晨：负责数据分析、PPT与Word的优化

小组成员：

王迟迟：负责数据分析、Word的后期修改

徐思涵：负责PPT与Word的整体制作

万程冰：负责PPT与Word的整体制作

张浩宸：负责PPT后期修改与汇报



# 目录

## CONTENTS

01

### PART 01

背景介绍与分析方向

02

### PART 02

用户特征与使用行为分析

03

### PART 03

合约有效性与用户行为分析

04

### PART 04

总结



01

PART ONE

背景介绍与分析方向

# 数据分析

- 什么是数据分析

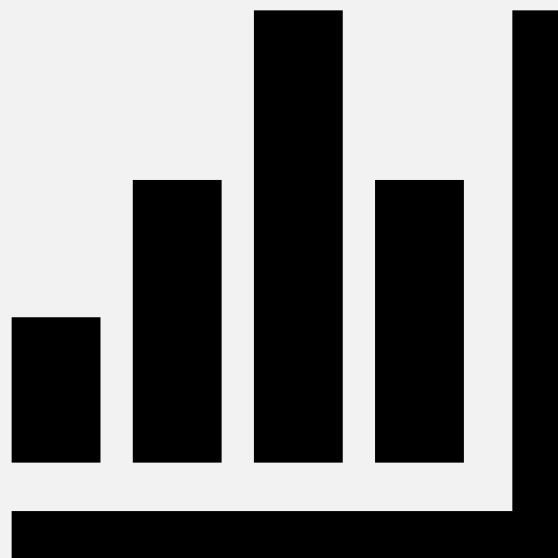
- 数据分析是指用适当的统计分析方法对收集来的数据进行分析，提取有用信息和形成结论的过程。

- 数据分析应用领域

- 1. 商业领域：帮助企业了解市场趋势、客户需求、销售情况等，以制定营销策略、优化产品和服务。
- 2. 金融领域：用于风险评估、投资决策、信用评级等。
- 3. 医疗领域：分析医疗数据，提高疾病诊断准确性、优化治疗方案等。
- 4. 科学研究：处理实验数据，验证假设、发现规律等。

# 数据分析的意义

- ◆ **发现趋势与模式：**通过分析数据，能够识别隐藏在数据中的趋势（如销售增长、用户行为模式）。这对企业的预测和规划至关重要，比如根据趋势调整市场策略或优化库存。
- ◆ **支持数据驱动的决策：**数据分析能提供客观依据，减少决策中的主观性，帮助企业或组织基于事实做出明智决策。
- ◆ **改进客户体验和个性化服务：**通过分析用户行为数据，可以为用户提供更加个性化的产品和服务，提高客户满意度。
- ◆ **数据可视化与洞察传播：**分析后的结果可以通过图表等形式直观呈现，帮助团队和管理层快速理解关键信息，便于沟通和协作。
- ◆ **风险预警与问题检测：**数据分析可以提前发现潜在风险，避免损失；还能及时检测异常情况，防止问题恶化。



A close-up photograph of a desk setup. In the foreground, there is a stack of three notebooks with light-colored covers. Two blue pencils are resting on the top notebook. In the background, a portion of a laptop keyboard is visible. The lighting is soft and natural, creating a professional and organized atmosphere.

# 运营商数据分析意义

- ◆ **通过用户分析：**运营商可以更深入地了解用户的真实需求，从而提供更加贴合用户期望的产品和服务，增强用户满意度。
- ◆ **精准定位市场：**用户分析有助于运营商明确目标市场的特点和需求，进而制定更为精准的营销策略和产品定位，提升市场竞争力。
- ◆ **提升用户体验：**基于用户分析的结果，运营商可以优化服务流程，提供更加个性化的服务体验，从而增强用户粘性，降低用户流失率。
- ◆ **优化产品和服务：**通过用户分析，运营商可以及时发现产品和服务中存在的问题和不足，进而进行针对性的改进和优化，提升产品和服务质量。
- ◆ **实现精准营销：**用户分析使得运营商能够针对不同用户群体制定差异化的营销策略，提高营销效率和投资回报率。

# 数据介绍

## Data introduction

名称	字段描述
MONTH_ID	月份
USER_ID	用户ID
INNET_MONTH	在网时长
IS_AGREE	是否合约有效用户
AGREE_EXP_DATE	合约计划到期时间
CREDIT_LEVEL	信用等级
VIP_LVL	VIP等级
ACCT_FEE	本月费用（元）
CALL_DURA	通话时长(秒)
NO_ROAM_LOCAL_CALL_DURA	本地通话时长(秒)
NO_ROAM_GN_LONG_CALL_DURA	国内长途通话时长(秒)
GN_ROAM_CALL_DURA	国内漫游通话时长(秒)
CDR_NUM	通话次数（次）
NO_ROAM_CDR_NUM	非漫游通话次数（次）
NO_ROAM_LOCAL_CDR_NUM	本地通话次数（次）
NO_ROAM_GN_LONG_CDR_NUM	国内长途通话次数（次）
GN_ROAM_CDR_NUM	国内漫游通话次数（次）
P2P_SMS_CNT_UP	短信发送数（条）
TOTAL_FLUX	上网流量(MB)
LOCAL_FLUX	本地非漫游上网流量(MB)
GN_ROAM_FLUX	国内漫游上网流量(MB)

名称	字段描述
CALL_DAYS	有通话天数
CALLING_DAYS	有主叫天数
CALLED_DAYS	有被叫天数
CALL_RING	语音呼叫圈
CALLING_RING	主叫呼叫圈
CALLED_RING	被叫呼叫圈
CUST_SEX	性别
CERT_AGE	年龄
CONSTELLATION_DESC	星座
MANU_NAME	手机品牌名称
MODEL_NAME	手机型号名称
OS_DESC	操作系统描述
TERM_TYPE	终端硬件类型(4=4g、3=3g、2=2g)
IS_LOST	用户在3月是否流失标记（1=是，0=否），1月和2月值为空



# 数据介绍

## Data introduction

名称	字段描述
MONTH_ID	月份
USER_ID	用户ID
INNET_MONTH	在网时长
IS_AGREE	是否合约有效用户
AGREE_EXP_DATE	合约计划到期时间
CREDIT_LEVEL	信用等级
VIP_LVL	VIP等级
ACCT_FEE	本月费用 (元)
CALL_DURA	通话时长(秒)
NO_ROAM_LOCAL_CALL_DURA	本地通话时长(秒)
NO_ROAM_GN_LONG_CALL_DURA	国内长途通话时长(秒)
GN_ROAM_CALL_DURA	国内漫游通话时长(秒)
CDR_NUM	通话次数 (次)
NO_ROAM_CDR_NUM	非漫游通话次数 (次)
NO_ROAM_LOCAL_CDR_NUM	本地通话次数 (次)
NO_ROAM_GN_LONG_CDR_NUM	国内长途通话次数 (次)
GN_ROAM_CDR_NUM	国内漫游通话次数 (次)
P2P_SMS_CNT_UP	短信发送数 (条)
TOTAL_FLUX	上网流量(MB)
LOCAL_FLUX	本地非漫游上网流量(MB)
GN_ROAM_FLUX	国内漫游上网流量(MB)
CALL_DAYS	有通话天数
CALLING_DAYS	有主叫天数
CALLED_DAYS	有被叫天数
CALL_RING	语音呼叫圈
CALLING_RING	主叫呼叫圈
CALLED_RING	被叫呼叫圈
CUST_SEX	性别
CERT_AGE	年龄
CONSTELLATION_DESC	星座
MANU_NAME	手机品牌名称
MODEL_NAME	手机型号名称
OS_DESC	操作系统描述
TERM_TYPE	终端硬件类型(4=4g、3=3g、2=2g)
IS_LOST	用户在3月是否流失标记 (1=是, 0=否)

Excel表格内容为运营商用户的基础信息和使用行为信息，包括90W条记录，30W用户3个月的数据，35个特征。

因为数据量巨大，直接使用AI工具去分析难度较大且可行性较低，因此我们采用“AI+编程”的方式去进行数据分析，即使用python作为数据分析的工具，结合AI大模型的分析建议，去完成数据分析的工作。

为了明确数据分析的方向，我们先对表格基本信息进行了提取，得到一个txt文件，如图所示。

图 1 用户数据表基本信息

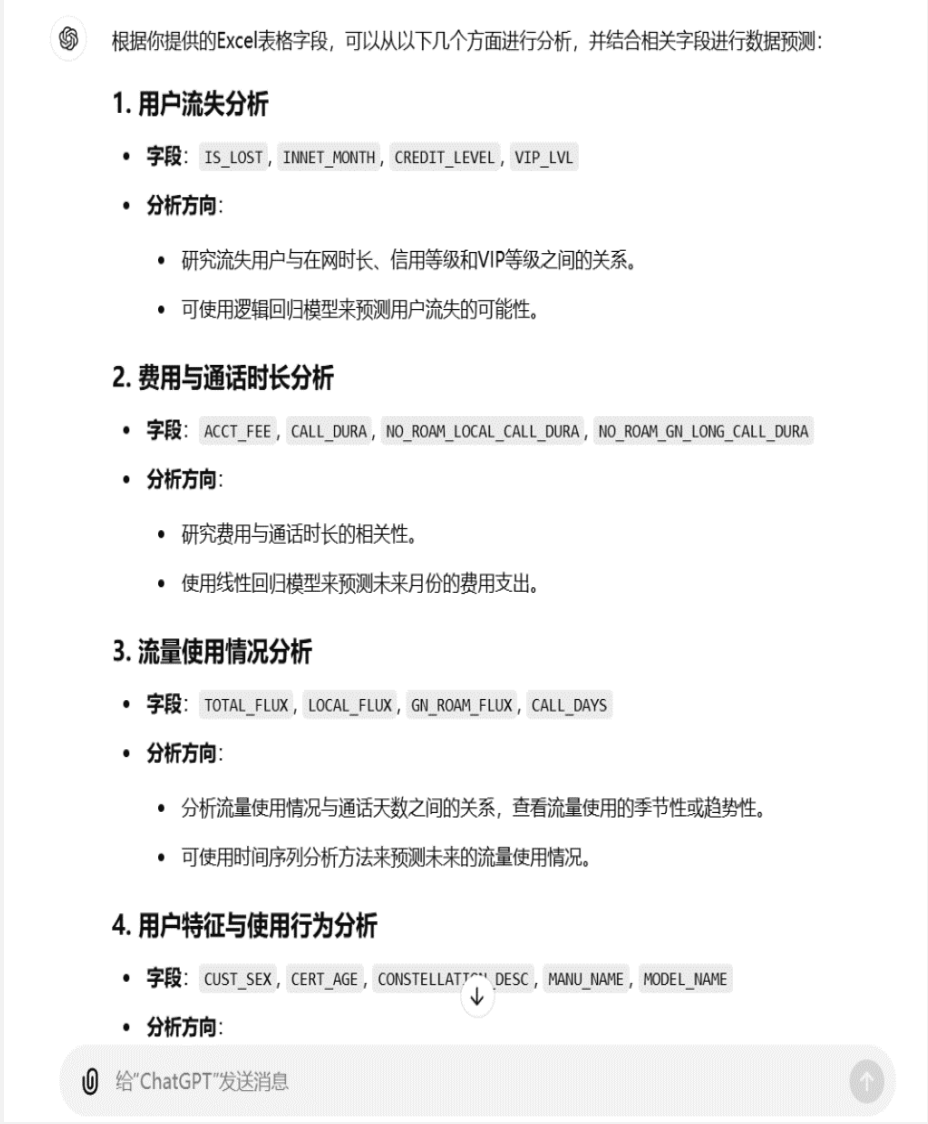


图 2 数据分析方向

依据表格的基本信息，我们对ChatGPT进行提问，获取对表格进行分析的方向，ChatGPT根据我们提供的基本信息，给出了一下几点建议：

根据你提供的Excel表格字段，可以从以下几个方面进行分析，并结合相关字段进行数据预测：

- 1. \*\*用户流失分析\*\***

**\*\*分析方向\*\*：**

  - 研究流失用户与在网时长、信用等级和VIP等级之间的关系。
  - 可使用逻辑回归模型来预测用户流失的可能性。
- 2. \*\*费用与通话时长分析\*\***

**\*\*分析方向\*\*：**

  - 研究费用与通话时长的相关性。
  - 使用线性回归模型来预测未来月份的费用支出。
- 3. \*\*流量使用情况分析\*\***

**\*\*分析方向\*\*：**

  - 分析流量使用情况与通话天数之间的关系，查看流量使用的季节性或趋势性。
  - 可使用时间序列分析方法来预测未来的流量使用情况。
- 4. \*\*用户特征与使用行为分析\*\***

**\*\*分析方向\*\*：**

  - 研究用户的性别、年龄、星座与通话行为、费用支出之间的关系。
  - 可以进行聚类分析，将用户分群，针对不同用户群体的特点进行有针对性的营销策略。
- 5. \*\*合约有效性与用户行为分析\*\***

**\*\*分析方向\*\*：**

  - 研究合约有效用户的使用行为与通话天数的关系。
  - 预测合约到期用户的行为变化，提前采取措施降低流失率。

# 「分析方向

Direction of analysis」

依据ChatGPT给出的分析建议，我们确立了我们主要的分析方向：

1

## 用户特征与使用行为分析：

研究用户的性别、年龄、星座、手机品牌与费用支出之间的关系，以针对不同用户群体的特点进行有针对性的营销策略

2

## 合约有效性与用户行为分析：

研究合约有效用户的使用行为与通话天数的关系，预测合约到期用户的行为变化，提前采取措施降低流失率



# 02 PART TWO

## 用户特征与使用行为分析

# 探索梳理数据

## 用户特征与使用行为分析

Excel表格内容为运营商用户的基础信息和使用行为信息，包括90W条记录，30W用户3个月的数据，35个特征。其中有大量的重复、缺失的数据，需要对数据进行初步筛选去重，排除无有效意义的属性。

为了对用户数据进行聚类分析，我们选取了以下字段作为分析的主要依据：MONTH\_ID、CUST\_SEX、CERT\_AGE、CONSTELLATION\_DESC、MANU\_NAME、ACCT\_FEE，然后使用GitHub Copilot，提出了我们对于现有表格的处理需求，即在保留以上字段的同时，对数据进行去重，随后Copilot依据我们的需求给出了处理步骤与代码（基于Python语言）。

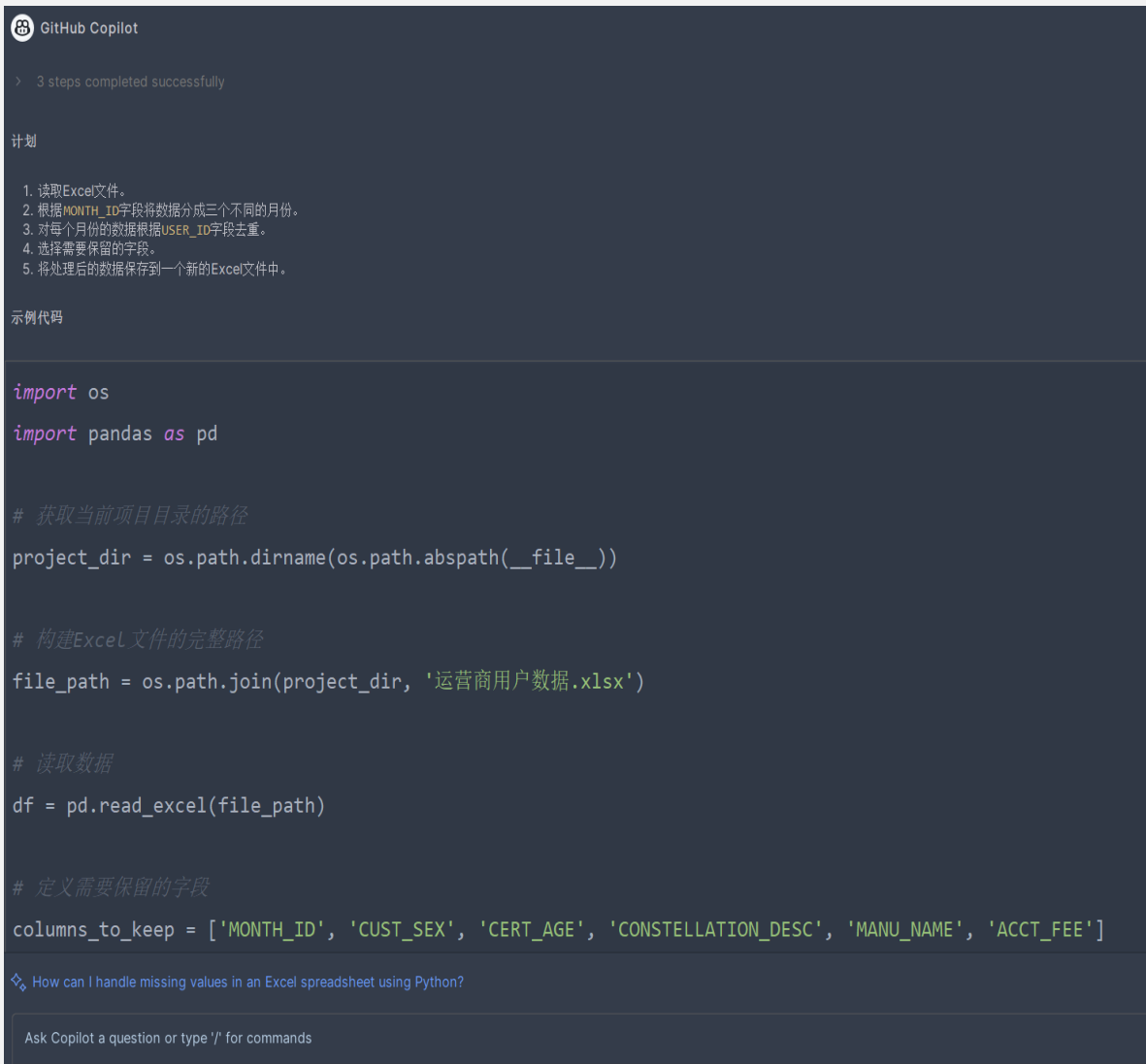


图 3 数据预处理

# 探索梳理数据

依据我们的要求，Copilot给出了它的数据分析计划与示例的代码，主要的处理内容如下：

1. 利用pandas库读取现有表格的数据，并定义需要保留的字段：MONTH\_ID、CUST\_SEX、CERT\_AGE、CONSTELLATION\_DESC、MANU\_NAME、ACCT\_FEE
2. 依据月份字段，将数据区分为三个不同的月份，对每个月份依据USER\_ID字段去重，并去除CUST\_SEX、CONSTELLATION\_DESC字段为空的行
3. 依据选择的保留字段生成新的数据框，并合并三个月份的数据
4. 将新得到的数据保存到新的表格中

GitHub Copilot

> 3 steps completed successfully

计划

1. 读取Excel文件。

2. 根据MONTH\_ID字段将数据分成三个不同的月份。

3. 对每个月份的数据根据USER\_ID字段去重。

4. 选择需要保留的字段。

5. 将处理后的数据保存到一个新的Excel文件中。

示例代码

```
import os
import pandas as pd

# 获取当前项目目录的路径
project_dir = os.path.dirname(os.path.abspath(__file__))

# 构建Excel文件的完整路径
file_path = os.path.join(project_dir, '运营商用户数据.xlsx')

# 读取数据
df = pd.read_excel(file_path)

# 定义需要保留的字段
columns_to_keep = ['MONTH_ID', 'CUST_SEX', 'CERT_AGE', 'CONSTELLATION_DESC', 'MANU_NAME', 'ACCT_FEE']
```

💡 How can I handle missing values in an Excel spreadsheet using Python?

Ask Copilot a question or type '/' for commands

图 4 数据预处理

# 探索梳理数据

在运行代码之后，我们得到了一个新的表格“用户特征与使用行为分析.xlsx”

保留了月份、性别、年龄、星座、手机品牌名称、本月费用（元）这六个字段，并对性别与星座为空的数据进行了删除

表格如图所示：

1	MONTH_ID	CUST_SEX	CERT_AGE	ELLATION	MANU_NAME	ACCT_FEE
2	201601	1	29	射手座	赫比	84.95
3	201601	1	23	白羊座	赫比	251.78
4	201601	1	38	射手座	赫比	26
5	201601	2	23	天蝎座	赫比	139.3
6	201601	2	33	双子座	赫比	31.95
7	201601	1	34	双鱼座	赫比	82.3
8	201601	1	24	天秤座	赫比	126.4
9	201601	1	28	天秤座	赫比	56
10	201601	1	46	白羊座	赫比	56.1
11	201601	1	47	巨蟹座	赫比	86.1
12	201601	1	24	射手座	赫比	197.95
13	201601	2	43	双子座	赫比	56
14	201601	1	34	天蝎座	赫比	101.1
15	201601	2	28	射手座	赫比	39.7
16	201601	1	37	水瓶座	赫比	19
17	201601	1	30	天蝎座	赫比	183.1
18	201601	1	19	射手座	赫比	50.4
19	201601	1	24	狮子座	赫比	20
20	201601	1	41	天蝎座	赫比	120.73
21	201601	1	34	金牛座	赫比	196
22	201601	1	39	水瓶座	赫比	136
23	201601	1	28	巨蟹座	小米	18.8
24	201601	1	36	射手座	小米	119.78
25	201601	1	32	水瓶座	小米	131.8
26	201601	1	26	天蝎座	小米	76

图 5 用户特征与使用行为分析Excel表格内容

# 探索梳理数据

对于得到的新表格，我们对Copilot提出了聚类分析的要求，随后Copilot给出了示例的代码，主要的处理内容如下：

1. 导入文件操作、数据处理、机器学习聚类、缺失值处理和可视化相关的库，获取当前项目目录路径，并构建Excel文件的完整路径
2. 读取Excel文件中的数据，根据MONTH\_ID字段，将数据分为三个不同的月份的数据集，将分类变量转换为数值编码，并为多个字段生成类别编码映射
3. 定义聚类分析函数，包括数据预处理、聚类模型训练和结果可视化：将分类变量转换为数值变量。选择特征进行聚类。处理缺失值。使用KMeans算法进行聚类，并将结果存入数据框中。使用seaborn可视化聚类结果
4. 对每个月的数据调用聚类分析函数，进行聚类和可视化

```
1  import os
2  import pandas as pd
3  from sklearn.model_selection import train_test_split
4  from sklearn.preprocessing import StandardScaler
5  from sklearn.ensemble import RandomForestClassifier
6  from sklearn.metrics import classification_report, accuracy_score
7  from imblearn.over_sampling import SMOTE
8  import matplotlib.pyplot as plt
9  import numpy as np
10
11  # 获取当前项目目录的路径
12  project_dir = os.path.dirname(os.path.abspath(__file__))
13
14  # 构建Excel文件的完整路径
15  file_path = os.path.join(project_dir, '合约有效性与用户行为.xlsx')
16
17  # 读取数据
18  df = pd.read_excel(file_path)
19
20  # 将分类变量转换为数值变量
21  df['CUST_SEX'] = df['CUST_SEX'].astype('category').cat.codes
22  df['VIP_LVL'] = df['VIP_LVL'].astype('category').cat.codes
23  df['CREDIT_LEVEL'] = df['CREDIT_LEVEL'].astype('category').cat.codes
24
```

图 6 聚类分析



聚类分析之后，得到的图像如下：

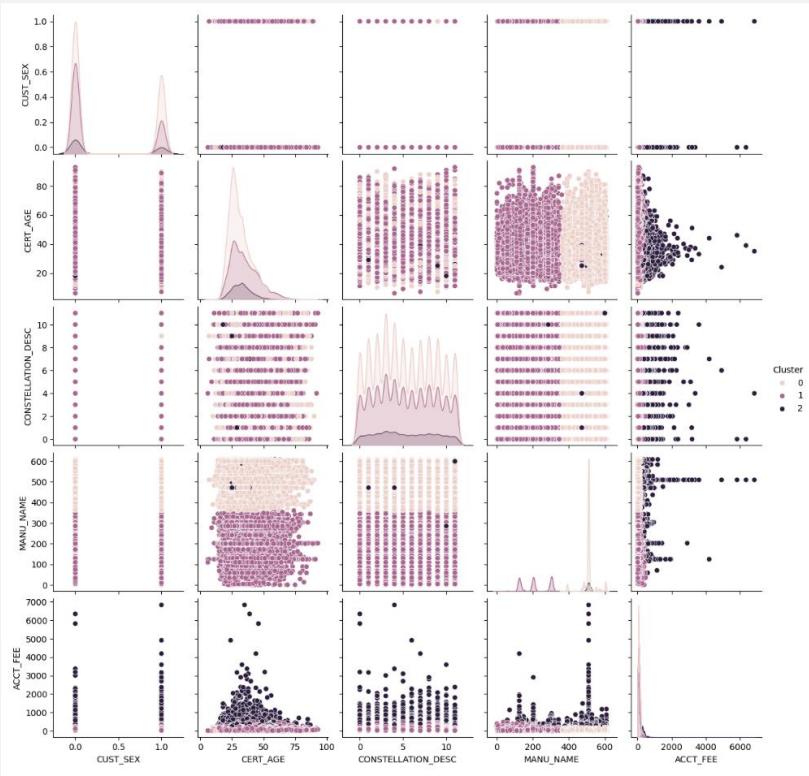


图 7 一月聚类分析结果

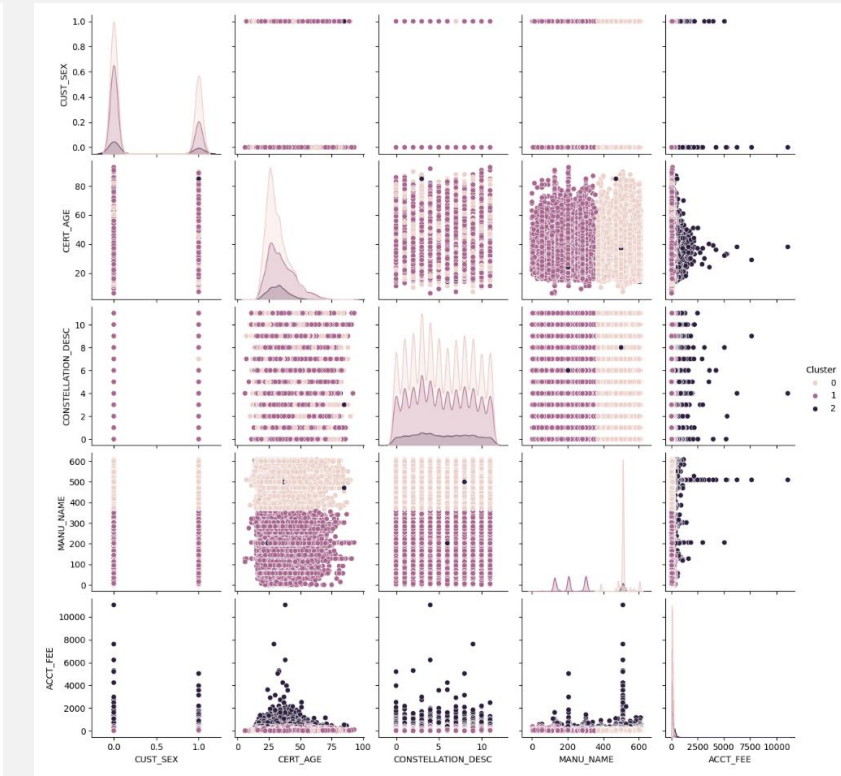


图 8 二月聚类分析结果

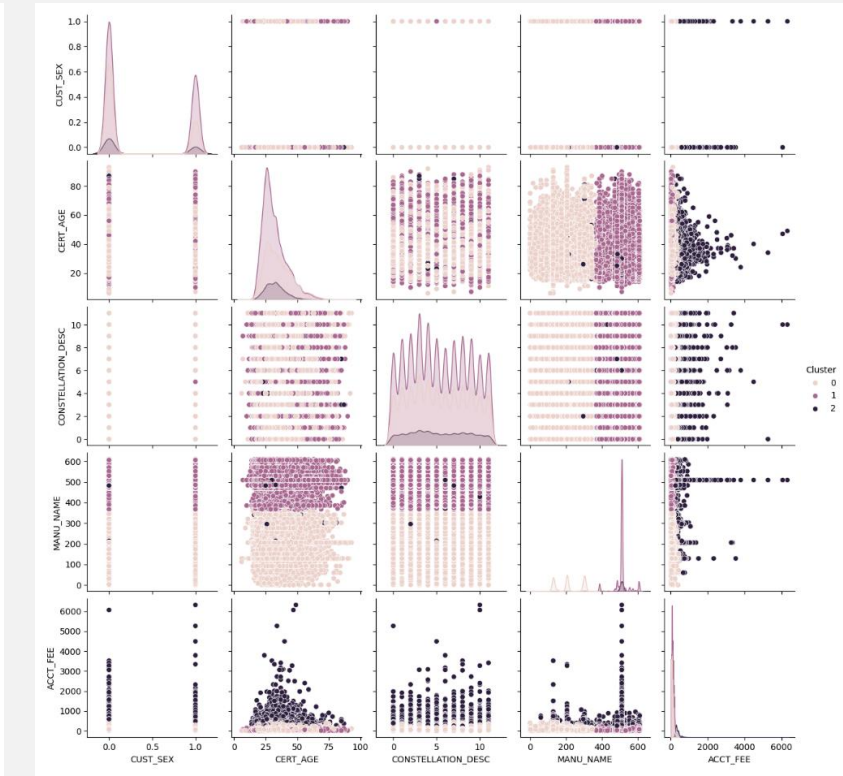


图 9 三月聚类分析结果

因为三个月份的聚类图像整体差异不大，这里以一月的聚类分析结果作为参考来分析用户数据的聚类分析结果。

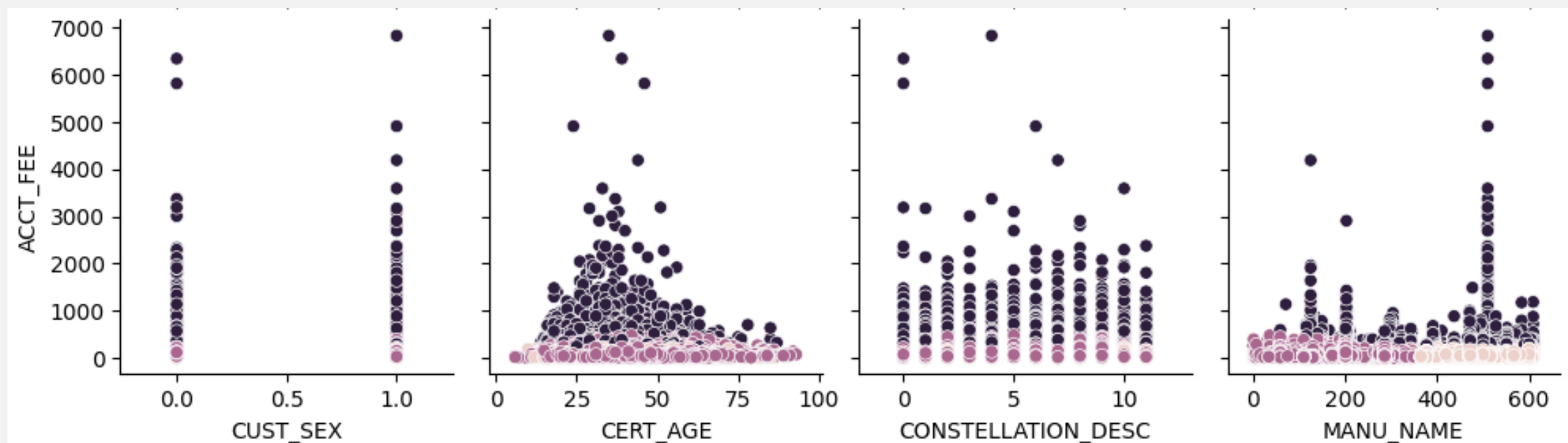


图 10 一月份各字段与本月费用的聚类图像

从CUST\_SEX性别字段来看与ACCT\_FEE本月费用的关系（其中0代表原表格中的“1”性别，1代表原表格中的“2”性别）：

图像上本月费用的数值不是真实的经济花费，而是算法求取的数值编码，实际上用户的费用数额在0.01元到11059.4元之间。

从图像不难看出，对于0数值代表的“1”性别，大部分的花费数额在0.01到30元之间，少部分用户的花费在100元上下的区间，只有极少部分的用户的费用在200元以上甚至更多的部分；对于1数值代表的“2”性别，总体花费水平要高于“1”性别，在高消费部分也略多于“1”性别。但是在2、3月份中，“1”性别的总体消费水平是要高于“2”性别的。

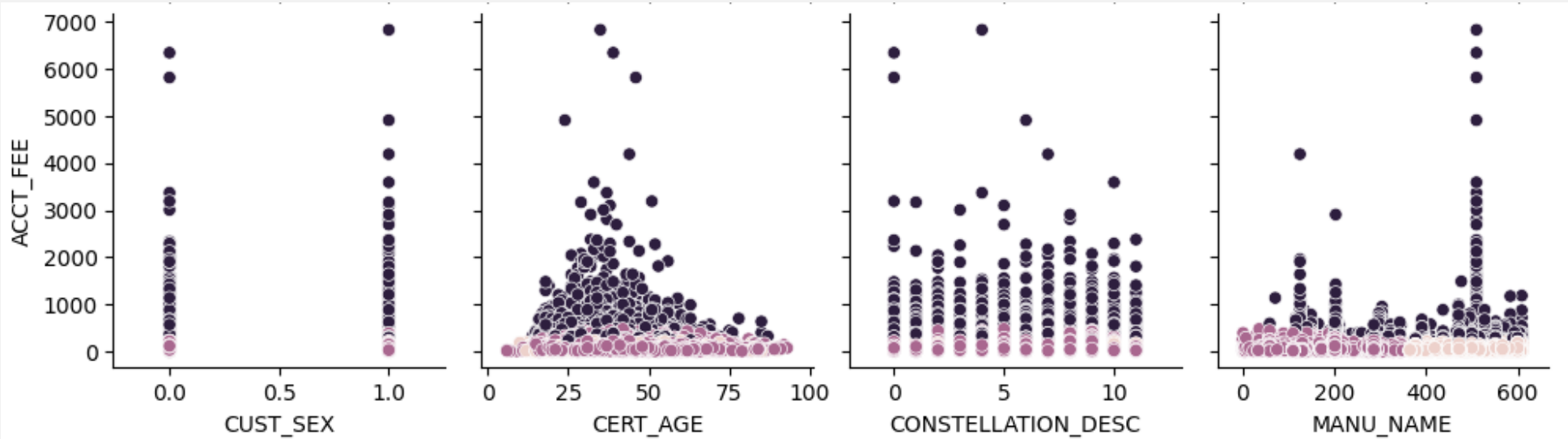


图 11 一月份各字段与本月费用的聚类图像

从CERT\_AGE年龄字段来看与ACCT\_FEE本月费用的关系：

横坐标的值也是数值编码，不是实际的年龄，真实的年龄区间在6~93岁之间。

结合图8来看，数值编码在10~60之间是主要的消费用户分布区间，实际年龄区间则是17~67之间；而消费的主力军则在30~45岁之间，基本都超过了30元的月消费，也是消费多于百元的主要用户群体。

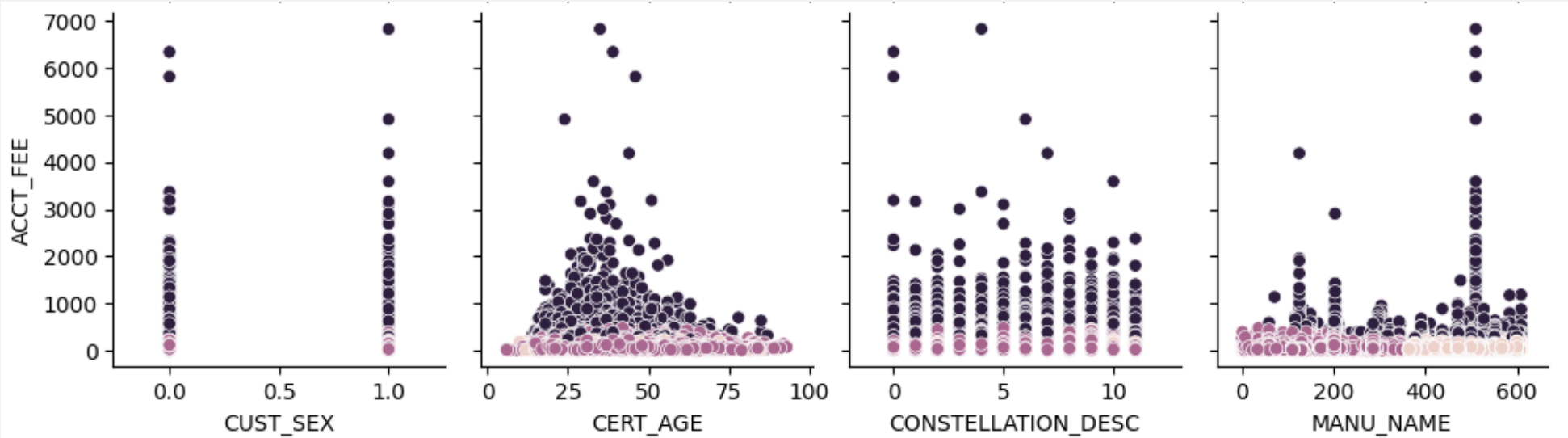


图 12 一月份各字段与本月费用的聚类图像

从CONSTELLATION\_DESC星座字段来看与ACCT\_FEE本月费用的关系：

综合三个月份的消费水平来看，各星座之间主体的消费都在同一水平，没有较为明显的星座差异导致的消费水平的变化。

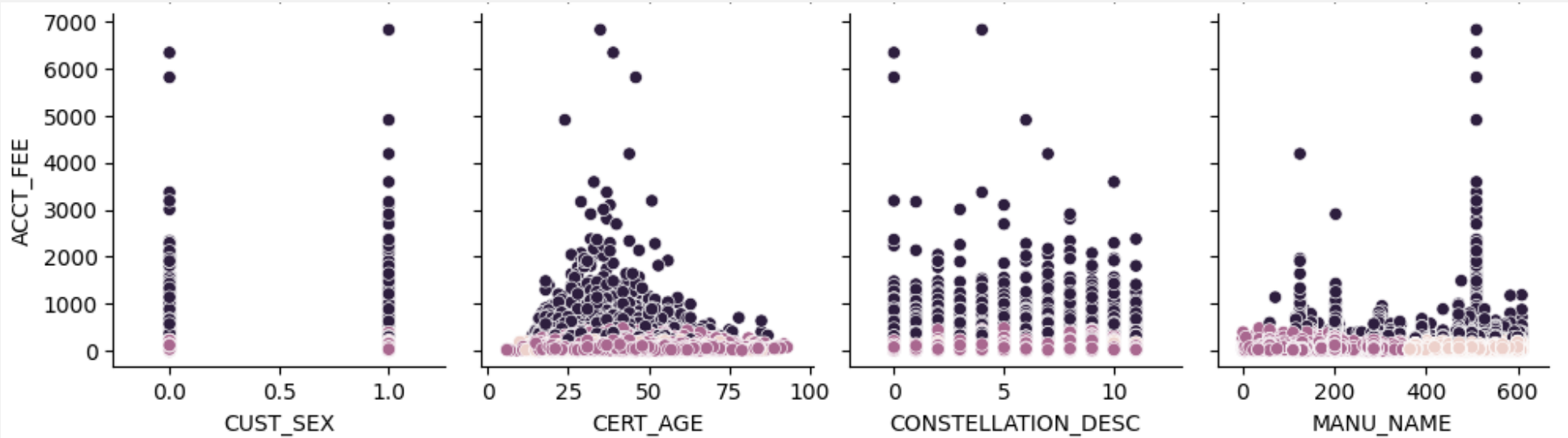


图 13 一月份各字段与本月费用的聚类图像

从MANU\_NAME手机品牌名称字段来看与ACCT\_FEE本月费用的关系：

数值编码在150、200、500左右有较高的数值提升，其他的数值编码对应的手机品牌的消费水平大体一致。而三个突出的数值编码对应的手机品牌依次为三星、华为、苹果三家，尤其苹果手机的用户，在三个月份的消费都较为突出。

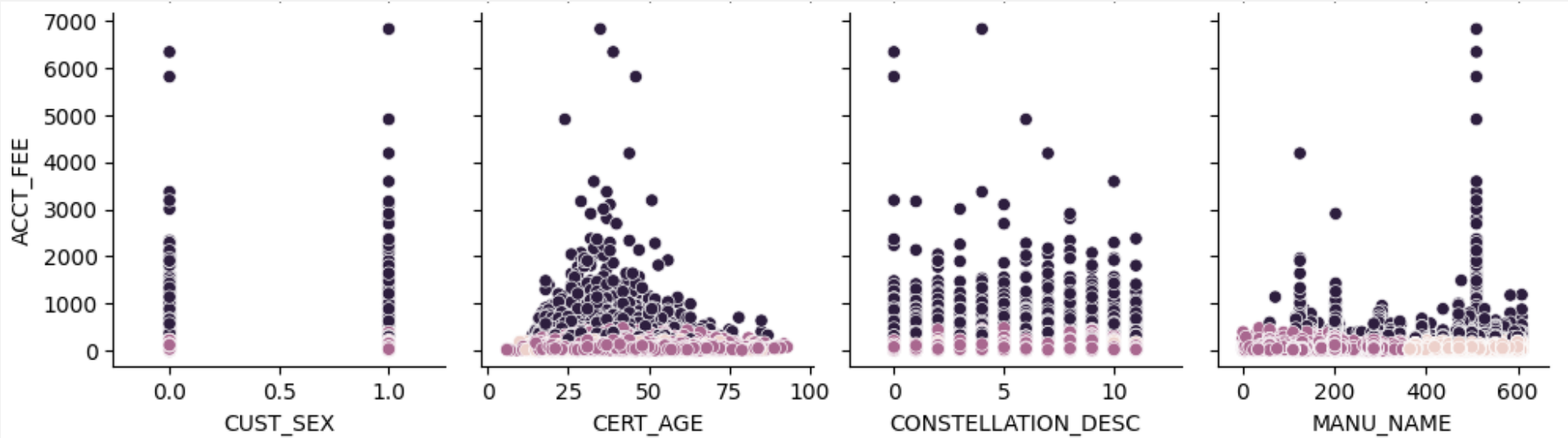


图 14 一月份各字段与本月费用的聚类图像



# 03 PART THREE


合约有效性与用户行为  
分析





为了探究怎么通过用户的使用行为与通话天数去预测合约到期后用户行为的变化，我们先向ChatGPT提问我们针对这个分析方向应该怎么处理数据。

合约有效性与用户行为分析：研究合约有效用户的使用行为与通话天数的关系，预测合约到期用户的行为变化，提前采取措施降低流失率。对于这个分析方向，我应该怎么处理与分析数据？

 对于研究合约有效用户的使用行为与通话天数的关系，并预测合约到期用户的行为变化，你可以按照以下步骤进行数据处理与分析：

1. 数据准备

- **筛选数据**：从数据集中筛选出 `IS_AGREE` 为有效合约用户的记录。
- **特征选择**：重点关注 `CALL_DAYS`（有通话天数）、`ACCT_FEE`（费用）及其他相关特征（如通话时长、用户年龄等）。

2. 数据清洗

- **处理缺失值**：使用均值填充、删除缺失记录或其他合适的方法处理 `NaN` 值。
- **数据转换**：确保所有特征都是数值型或合适的格式，必要时进行标准化或归一化处理。

3. 数据分析

- **描述性统计**：计算合约有效用户的基本统计信息（均值、标准差等）来了解用户行为。
- **相关性分析**：使用皮尔逊相关系数等方法分析 `CALL_DAYS` 与其他特征（如费用）的相关性。

4. 可视化

- **散点图**：绘制 `CALL_DAYS` 与其他行为特征（如 `ACCT_FEE`）的散点图，观察两者之间的关系。
- **箱线图**：分析不同合约状态用户的通话天<sup>↓</sup>布，帮助识别流失风险。


 给“ChatGPT”发送消息 

图 15 合约有效性的分析方向



# 「ChatGPT给出了以下分析的方向与步骤：」

## 方向与步骤

01

### **\*\*数据准备\*\***

- **\*\*筛选数据\*\***: 从数据集中筛选出 `IS\_AGREE` 为有效合约用户的记录。
- **\*\*特征选择\*\***: 重点关注 `CALL\_DAYS` (有通话天数)、`ACCT\_FEE` (费用) 及其他相关特征 (如通话时长、用户年龄等)。

02

### **\*\*数据清洗\*\***

- **\*\*处理缺失值\*\***: 使用均值填充、删除缺失记

03

### **\*\*数据分析\*\***

- **\*\*相关性分析\*\***: 使用皮尔逊相关系数等方法分析 `CALL\_DAYS` 与其他特征 (如费用) 的相关性。

04

### **\*\*可视化\*\***

- **\*\*散点图\*\***: 绘制 `CALL\_DAYS` 与其他行为特征 (如 `ACCT\_FEE`) 的散点图, 观察两者之间的关系。
- **\*\*箱线图\*\***: 分析不同合约状态用户的通话天数分布, 帮助识别流失风险。

05

### **\*\*预测模型\*\***

- **\*\*建立模型\*\***: 使用机器学习模型 (如逻辑回归、决策树等) 预测合约到期用户的行为变化。
- 特征可以包括通话天数、费用、用户年龄等。

# 探索梳理数据

根据上面的回答内容，我们对Copilot提出了新的要求，Copilot给出了它的示例代码，主要的处理内容如下：

1. 利用pandas库读取现有表格的数据，指定需要保留的字段，方便对数据进行筛选。这些字段可能与用户的基本信息（性别、年龄）和行为数据（通话时长、通话天数、账户费用）有关
2. 从原始数据中筛选MONTH\_ID等于201603的数据，即2016年3月的用户数据，基于USER\_ID字段去重，确保每位用户只保留一条记录，删除指定字段中存在缺失值的行，确保数据完整性，并舍弃其他字段
3. 处理后的数据保存为新的Excel文件 “合约有效性与用户行为.xlsx”

GitHub Copilot

> 3 steps completed successfully

计划

1. 读取Excel文件。
2. 根据MONTH\_ID字段将数据分成三个不同的月份。
3. 对每个月份的数据根据USER\_ID字段去重。
4. 选择需要保留的字段。
5. 将处理后的数据保存到一个新的Excel文件中。

示例代码

```
import os
import pandas as pd

# 获取当前项目目录的路径
project_dir = os.path.dirname(os.path.abspath(__file__))

# 构建Excel文件的完整路径
file_path = os.path.join(project_dir, '运营商用户数据.xlsx')

# 读取数据
df = pd.read_excel(file_path)

# 定义需要保留的字段
columns_to_keep = ['MONTH_ID', 'CUST_SEX', 'CERT_AGE', 'CONSTELLATION_DESC', 'MANU_NAME', 'ACCT_FEE']
```

How can I handle missing values in an Excel spreadsheet using Python?

Ask Copilot a question or type '/' for commands

图 16 数据预处理

结合ChatGPT给出的方向，与前面对用户特征与行为的分析。

计划通过用户的性别、年龄、VIP等级、信用等级、本月费用、通话时长、有通话天数几个方面，结合用户在三月的流失情况，去预测后面月份的流失情况。


去重与初步处理后的表格如图：

1	CUST_SEX	CERT_AGE	VIP_LVL	EDIT_LEV	ACCT_FEE	CALL_DUR	CALL_DAYS	IS_LOST
2	1	45	99	67	23.4	6172	14	0
3	2	48	99	67	56	34174	29	0
4	1	25	99	67	50.9	9549	19	0
5	1	39	99	67	20.2	1453	14	0
6	1	48	99	66	219.5	89268	31	0
7	1	50	99	67	642.5	302251	31	0
8	1	31	99	66	596	74591	31	0
9	2	38	99	67	64	11031	25	0
10	2	40	99	67	17.5	217	1	0
11	1	29	99	67	30.25	8983	10	0
12	1	43	4	66	145.5	44236	29	0
13	2	38	99	67	6.4	100	1	1
14	1	41	99	66	56.3	29524	31	0
15	1	23	99	67	120.25	67341	31	0
16	2	45	99	66	56	543	6	0
17	1	45	99	67	9.5	839	7	0
18	1	29	99	67	19	3495	15	0
19	1	38	99	67	87.1	36141	29	0
20	1	42	99	66	75.65	27734	28	1
21	2	47	99	67	44.95	14496	27	0
22	2	30	99	67	21.55	12860	28	0
23	1	55	99	67	16	1741	20	0
24	1	52	99	67	97.95	49557	31	0
25	1	57	99	67	16	21	1	0
26	2	30	99	67	78.5	28149	24	0
27	1	41	99	67	28.3	546	1	0
28	1	45	99	67	38.9	13706	27	0
29	2	30	99	67	61.7	28230	29	0
30	1	25	99	67	28.15	7513	21	0
31	1	25	99	67	43.9	12571	26	0


图 17 用户数据与合约有效性的处理

GitHub Copilot Chat

Optimizing Model Performance and Predictive Ability

 Z321T

'CUST\_SEX', 'CERT\_AGE', 'VIP\_LVL', 'CREDIT\_LEVEL', 'ACCT\_FEE', 'CALL\_DURA', 'CALL\_DAYS', 'IS\_LOST'根据这些字段。我怎么通过分析处理它们。去预测用户的流失情况？

 GitHub Copilot

> 3 steps completed successfully

计划

1. 读取数据并进行预处理，包括处理缺失值和转换分类变量。

2. 将数据集分为特征 (X) 和目标变量 (y)。

3. 划分训练集和测试集。

4. 使用机器学习模型 (如逻辑回归、随机森林等) 进行训练。

5. 评估模型性能并进行预测。

代码

```
import os
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report, accuracy_score
```

图 18 GitHub Copilot给出的方向与代码

Accuracy: 0.9784461763273259

Classification Report:

	precision	recall	f1-score	support
0	0.98	1.00	0.99	40147
1	0.58	0.11	0.18	913
accuracy			0.98	41060
macro avg	0.78	0.55	0.59	41060
weighted avg	0.97	0.98	0.97	41060

图 19 初次学习后的预测与评价结果

结合已经处理好的数据表格，我们向GitHub Copilot询问如何使用已经处理好的表格去进行用户流失的预测： 在初次的使用随机森林模型学习并预测与评价模型水平时，得到了结果（分别见图18、图19）

**我们发现了现有数据预测结果的缺陷：从支持度（support）可以看出，类别0（未流失用户）的样本数量远远多于类别1（流失用户）。这种类别不平衡可能导致模型在预测少数类（流失用户）时表现不佳。**

# 探索梳理数据

为了选择合适的预测模型，我们对Copilot提出了依据现有数据，来筛选适合的预测模型的需求，随后Copilot给出了它的示例代码，主要的处理内容如下：

1. 获取项目目录路径，构建Excel文件的路径，并将其加载为pandas DataFrame，将分类变量转换为数值变量，以便模型能够处理
2. X：特征集，包含所有字段（除目标变量IS\_LOST）； y：目标变量，表示用户是否流失。将数据按8:2比例划分为训练集和测试集，确保模型训练和评估的独立性，并通过生成合成样本来平衡训练集，用于处理类别不平衡
3. 选取三个分类模型：随机森林模型、逻辑回归模型、梯度提升模型。遍历每个模型，使用测试集进行预测，计算预测的准确率

```
1 import os
2 import pandas as pd
3 from sklearn.model_selection import train_test_split, GridSearchCV
4 from sklearn.preprocessing import StandardScaler
5 from sklearn.ensemble import RandomForestClassifier, GradientBoostingClassifier
6 from sklearn.linear_model import LogisticRegression
7 from sklearn.metrics import classification_report, accuracy_score
8 from imblearn.over_sampling import SMOTE
9
10 # 获取当前项目目录的路径
11 project_dir = os.path.dirname(os.path.abspath(__file__))
12
13 # 构建Excel文件的完整路径
14 file_path = os.path.join(project_dir, '合约有效性与用户行为.xlsx')
15
16 # 读取数据
17 df = pd.read_excel(file_path)
18
19
20 # 将分类变量转换为数值变量
21 df['CUST_SEX'] = df['CUST_SEX'].astype('category').cat.codes
22 df['VIP_LVL'] = df['VIP_LVL'].astype('category').cat.codes
23 df['CREDIT_LEVEL'] = df['CREDIT_LEVEL'].astype('category').cat.codes
24
25 # 定义特征和目标变量
26 X = df.drop(labels='IS_LOST', axis=1)
27 y = df['IS_LOST']
```

图 20 预测模型选择

再次运行代码之后，我们得到了三个模型的预测结果，并对原有预测代码进行了进一步的修改，使用类别平衡技术改善数据样本的缺陷，最后得到了最佳的预测模型，其预测评价的结果如图所示：

```
Model: RandomForest
Accuracy: 0.8952508524111057
Classification Report:
              precision    recall  f1-score   support

      0       0.99         0.90         0.94       40147
      1       0.11         0.55         0.19         913

   accuracy              0.90       41060
  macro avg              0.55         0.73         0.57       41060
weighted avg              0.97         0.90         0.93       41060

-----
```

图 21 修改后的模型预测结果

# 用户流失预测

在筛选出了合适的预测模型之后，我们使用筛选出的模型，结合原有的预测代码，让Copilot在此基础上给出了新的预测代码，主要的处理内容如下：

1. 获取当前文件目录并读取Excel文件，将分类变量（性别、VIP等级、信用等级）编码为数值，定义特征和目标变量
2. 按8:2的比例将数据划分为训练集和测试集，平衡训练集中类别的数量，减少因样本不均衡造成的偏差，使用标准化方法将特征转换为标准正态分布，提高模型的训练和预测效果。
3. 使用随机森林训练模型，并对测试集进行预测，输出准确率和分类报告（包括精确度、召回率、F1分数），用于模型性能评估
4. 预处理并预测未来6个月的流失情况并引入变化，使用正态分布为通话时长和费用引入随机噪声，可视化未来6个月的流失用户数量

```
1 import os
2 import pandas as pd
3 from sklearn.model_selection import train_test_split
4 from sklearn.preprocessing import StandardScaler
5 from sklearn.ensemble import RandomForestClassifier
6 from sklearn.metrics import classification_report, accuracy_score
7 from imblearn.over_sampling import SMOTE
8 import matplotlib.pyplot as plt
9 import numpy as np
10
11 # 获取当前项目目录的路径
12 project_dir = os.path.dirname(os.path.abspath(__file__))
13
14 # 构建Excel文件的完整路径
15 file_path = os.path.join(project_dir, '合约有效性与用户行为.xlsx')
16
17 # 读取数据
18 df = pd.read_excel(file_path)
19
20 # 将分类变量转换为数值变量
21 df['CUST_SEX'] = df['CUST_SEX'].astype('category').cat.codes
22 df['VIP_LVL'] = df['VIP_LVL'].astype('category').cat.codes
23 df['CREDIT_LEVEL'] = df['CREDIT_LEVEL'].astype('category').cat.codes
24
25 # 定义特征和目标变量
26 X = df.drop(labels='IS_LOST', axis=1)
27 y = df['IS_LOST']
```

图 22 用户流失预测

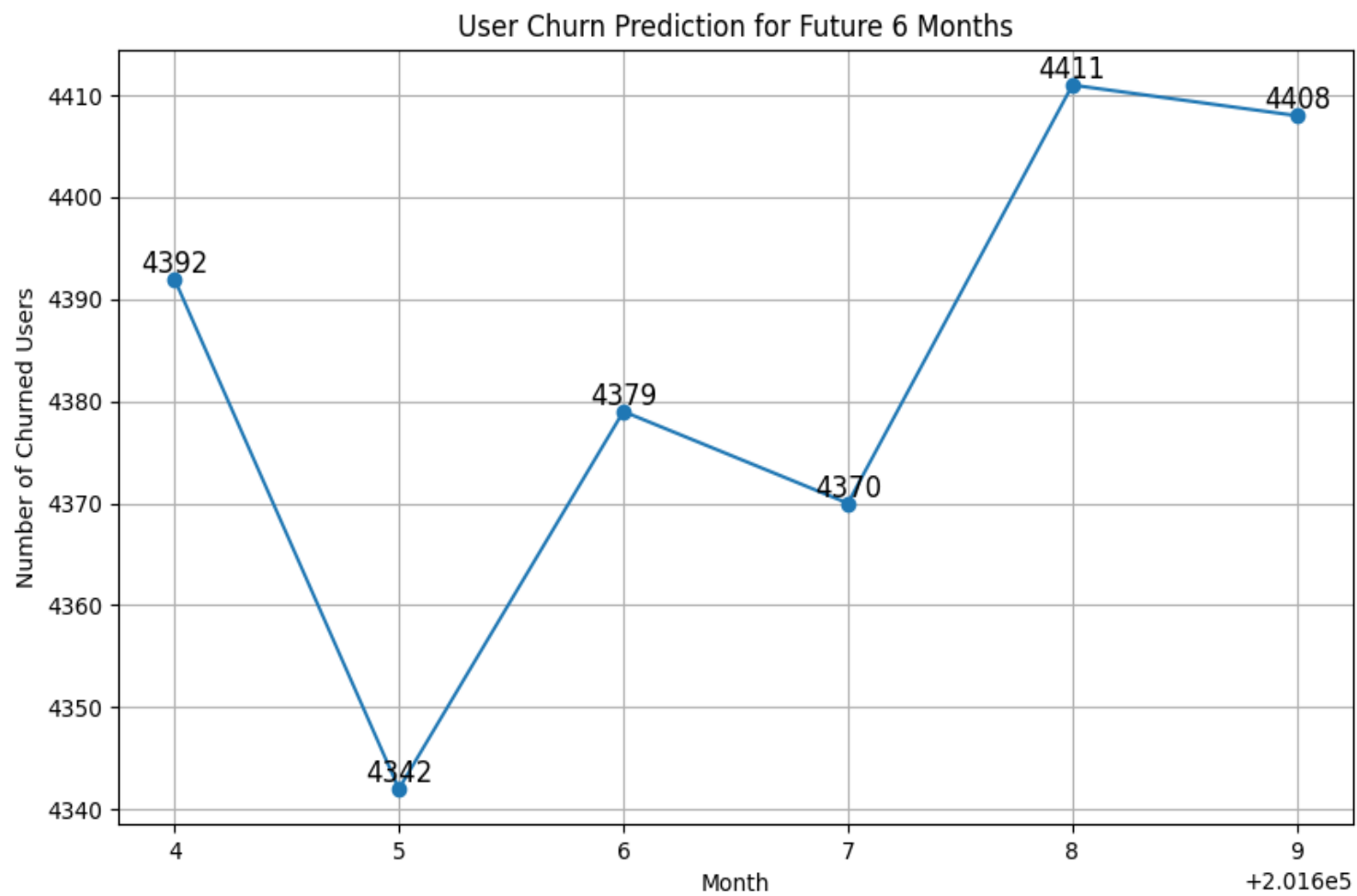


图 23 未来六个月的用户流失预测





# 04 PART FOUR

## 总结

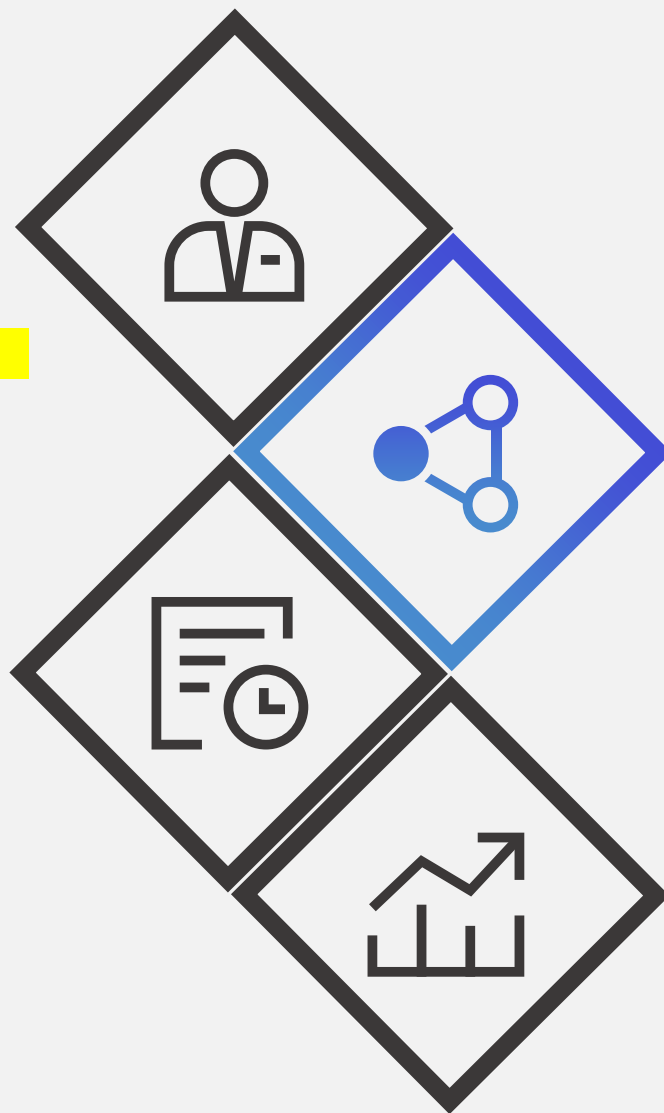
# 总结

Consumption

我们通过“AI+编程”的方式，对运营商用户数据进行深层次的分析。

AI为我们提供对数据分析的思路与方向，提高分析效率；

编程为我们提供数据分析的工具，结合AI给出的分析方向与代码建议，能快速实现对数据的处理与科学计算，得到可视化的分析结果。



ChatGPT在自然语言处理方面取得了重要突破，其智能化程度将不断提高，能更深入地理解人类语言的复杂性和微妙性，提供更精准、流畅的回答和对话。

AI将与更多领域进行融合，形成联合体。通过互相协作和学习，提升整个系统的智能化程度和效率，为人类带来更加便捷、高效的生活体验。

2024.10

感谢您的观看

THANK YOU FOR WAHCTING

汇报人：张浩宸

数据分析：冯朝晨 王迟迟

制作：徐思涵 万程冰