摩拜单车用户画像分析

一、问题界定

1. 背景介绍

摩拜单车，英文名 mobike，是由胡玮炜创办的北京摩拜科技有限公司研发的互联网短途出行解决方案，是无桩借还车模式的智能硬件。人们通过智能手机就能快速租用和归还一辆摩拜单车，用可负担的价格来完成一次几公里的市内骑行。 由于一公里的出行是一个高频的需求场景，所以 mobike 单车累计了大量的用户基本信息以及骑行的数据，通过这些数据，能够帮助企业更好识别自己的客群画像和他们的骑行偏好。

2. 目标问题

（1）使用 Python 建立聚类分析模型

（2）对于聚类分析模型得出的分群特征进行解读

二、数据概览和清洗

1. 数据维度分析

维度	字段	含义	变量类型
用户	user_id	用户编码	类别型
	usertype	用户种类	类别型
	gender	性别	类别型
	birthyear	出生年份	数值型
	age	年龄	数值型
骑行	start_time	开始时间	类别型
	end_time	结束时间	类别型
	timeduration	骑行时长	数值型
	bikeid	自行车编码	类别型
	tripduration	骑行距离	数值型
	from_station_id	开始站编码	类别型
	from_station_name	开始站名字	类别型
	to_station_id	结束站编码	类别型

	to_station_name	结束站名字	类别型

（1）总计数据，14 个字段，2 大维度。

（2）其中信息有冗余，如 age 可以通过 birthyear 计算得来；start_time 和 end_time 也同样可以计算出 timeduration；from_station_id 和 from_station_name 以及 to_station_id 和 to_station_name 表达的是同样的含义。

（3）而有些字段对建模来说用处不大：user_id，bike_id，from_station_id，from_station_name，to_station_id 和 to_station_name。

（4）而部分数据的类型不正确，如 start_time 和 end_time 应为 datetime 类型；age 应为 int 类型；user_id、bikeid、from_station_id、to_station_id 应为 str 类型。

2. 缺失值分析

由于 age 可以通过 birthyear 计算得出，且 birthyear 本身存在缺失值，而 age 虽表面上看无缺失值，但是实际上被空格填充，因此对 age 字段重新用 birthyear 计算。

```
user_id             6427 non-null int64
start_time          6427 non-null object
end_time            6427 non-null object
timeduration        6427 non-null int64
bikeid              6427 non-null int64
tripduration        6427 non-null int64
from_station_id     6427 non-null int64
from_station_name   6427 non-null object
to_station_id       6427 non-null int64
to_station_name     6427 non-null object
usertype            6427 non-null object
gender              5938 non-null object
birthyear           5956 non-null float64
age                 6427 non-null object
dtypes: float64(1), int64(6), object(7)
memory usage: 753.2+ KB
```

	字段	缺失值数量
0	user_id	0
1	start_time	0
2	end_time	0
3	timeduration	0
4	bikeid	0
5	tripduration	0
6	from_station_id	0
7	from_station_name	0
8	to_station_id	0
9	to_station_name	0
10	usertype	0
11	gender	489
12	birthyear	471
13	age	471

因此，数据集中只有 gender、birthyear 和 age 存在部分缺失值，需要处理。

3. 重复值、离群值、异常值分析

经检验，数据集中无重复的数据。

（1）数值型变量离群值分析

	timeduration	tripduration	birthyear	age
count	6427.000000	6.427000e+03	5956.000000	5956.000000
mean	11.778902	1.060471e+03	1982.488583	37.511417
std	9.692236	1.456811e+04	11.147859	11.147859
min	0.000000	6.100000e+01	1906.000000	18.000000
25%	5.000000	3.490000e+02	1977.000000	29.000000
50%	9.000000	5.590000e+02	1986.000000	34.000000
75%	15.000000	9.320000e+02	1991.000000	43.000000
max	59.000000	1.139070e+06	2002.000000	114.000000

对数值型变量的统计学参数进行分析，可以看出，

a. 字段 timeduration 中最小值是 0，可能存在统计错误，而最大值是 59 分钟，数据跨度较大，可能存在离群值；继续观察 timeduration 为 0 的数据（如下表），发现共有 4 行，且是由于统计错误，把 60 分钟计为了 0，需要替换。

	user_id	start_time	end_time	timeduration	bikeid	tripduration	from_station_id	from_station_name	to_station_id	to_station_name	usertype	gen
109113	21109759	2018-10-09 11:29:00	2018-10-09 12:29:00	0	4875	3616	225	Halsted St & Dickens Ave	172	Rush St & Cedar St	Customer	N
72232	21067313	2018-10-06 10:15:00	2018-10-06 11:15:00	0	4733	3606	419	Lake Park Ave & 53rd St	76	Lake Shore Dr & Monroe St	Customer	N
69595	21064115	2018-10-05 18:27:00	2018-10-05 19:28:00	0	2064	3650	50	Clark St & Congress Pkwy	283	LaSalle St & Jackson Blvd	Customer	M
302629	21336986	2018-10-27 15:33:00	2018-10-27 16:34:00	0	5051	3634	3	Shedd Aquarium	199	Wabash Ave & Grand Ave	Customer	N
592619	21681993	2018-12-17 09:17:00	2018-12-17 10:18:00	0	4316	3652	99	Lake Shore Dr & Ohio St	96	Desplaines St & Randolph St	Subscriber	M

替换后，统计参数为下表，timeduration 最小值变为 1 分钟，最大值为 60 分钟。

	timeduration	tripduration	birthyear	age
count	6427.000000	6.427000e+03	5956.000000	5956.000000
mean	11.825580	1.060471e+03	1982.488583	37.511417
std	9.779498	1.456811e+04	11.147859	11.147859
min	1.000000	6.100000e+01	1906.000000	18.000000
25%	5.000000	3.490000e+02	1977.000000	29.000000
50%	9.000000	5.590000e+02	1986.000000	34.000000
75%	15.000000	9.320000e+02	1991.000000	43.000000
max	60.000000	1.139070e+06	2002.000000	114.000000

b. 字段 tripduration 的极值差距也很大，最小为 61 米，最大为 1140000 米，数据跨度过大，可能存在离群值。

c. 年龄最大值为 114 岁，也存在离群值。

以上三个数值型字段均存在离群值需要处理

（2）类别型变量异常值分析

a. gender 字段的类别为 Male、Female 和缺失值，需要对缺失值进行处理；

b. usertype 字段的类别为 Subscriber 和 Customer，无异常值。

4. 衍生变量探索

对 datetime 类型的变量 start_time 和 end_time 尝试提取出额外的信息作为特征，发现数据仅分布在 2018 年 10 月-12 月之间，因此对 start_time 可提取星期几和小时的信息另作两列，命名为 start_weekday 和 start_hour，后续可对其进行必要的分箱操作。

5. 数据清洗

现阶段对数据进行如下清洗操作：

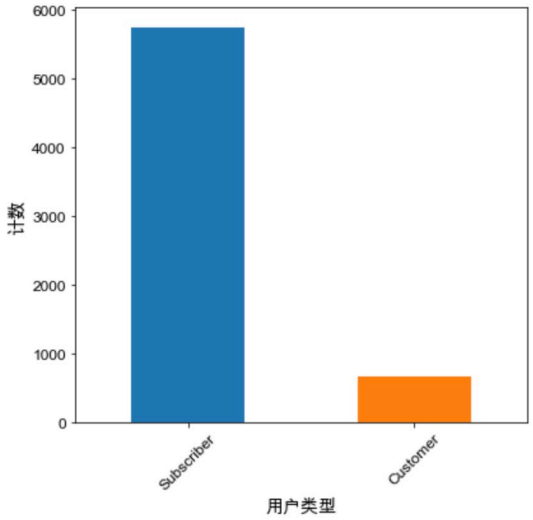
（1）去掉冗余信息和无用变量

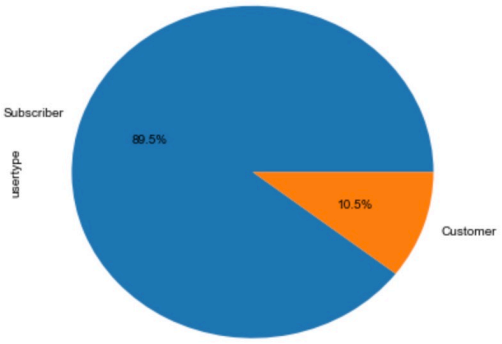
（2）填充 gender 的缺失值为 Unknown

三、单变量分析

1. 类别型变量

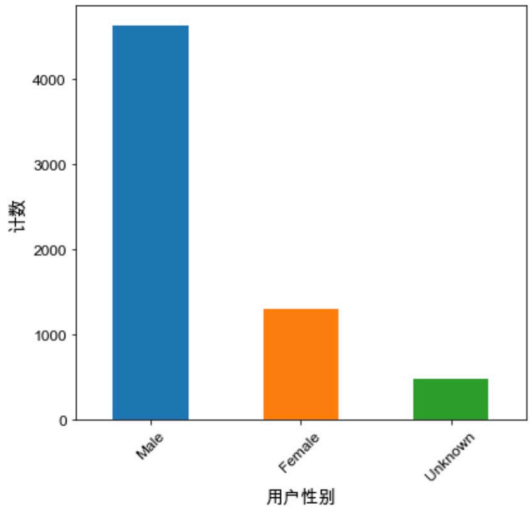
（1）usertype

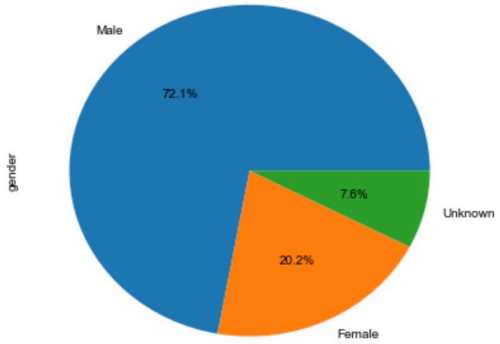




可以看出，摩拜单车用户中订购人数占近 90%。

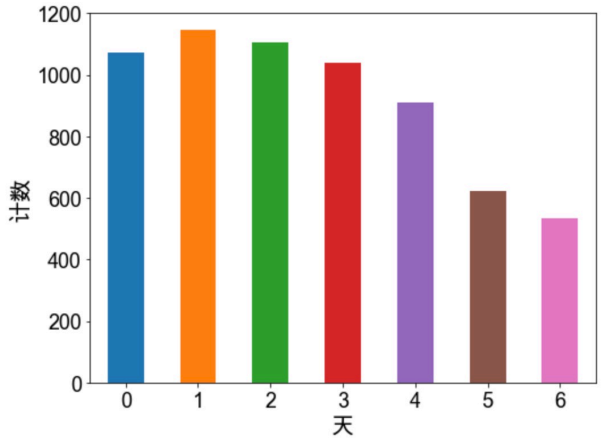
（2）gender





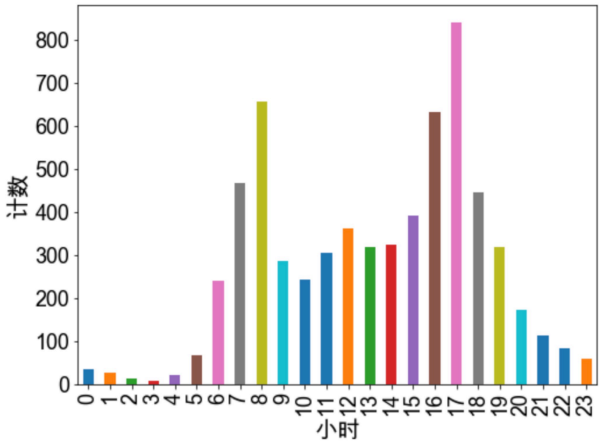
可以看出，摩拜单车用户中 70%以上为男性用户，8%左右未透露性别。

（3）start_weekday



可以看出，摩拜单车在工作日的使用频次较高，可以后续考虑将该字段分箱为工作日和周末两个类别。

（4）start_hour

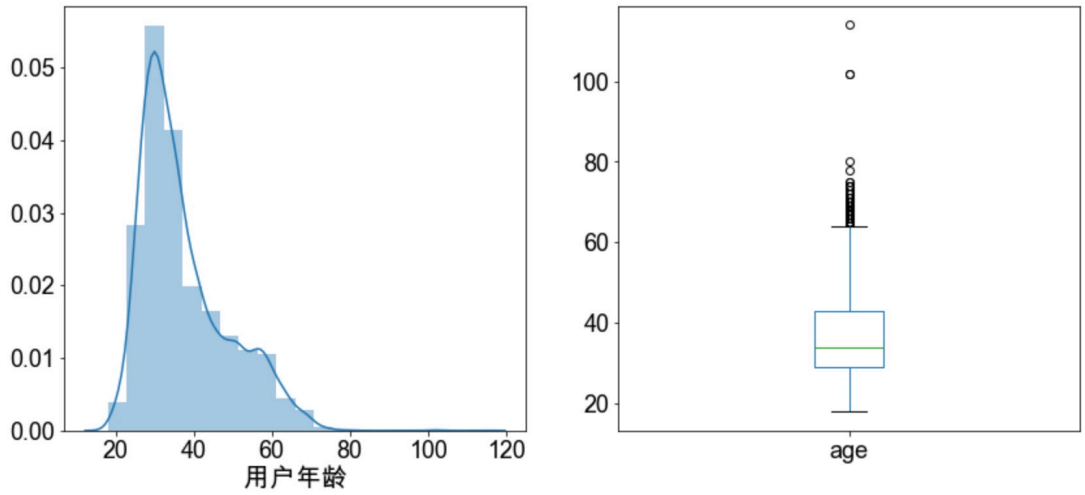


可以看出摩拜单车用车高峰时段集中在 6-8 早高峰和 16-18 晚高峰，且白天

比凌晨和夜晚使用频次高；可以考虑分箱为 0-6，6-12，12-18，18-24 四段。

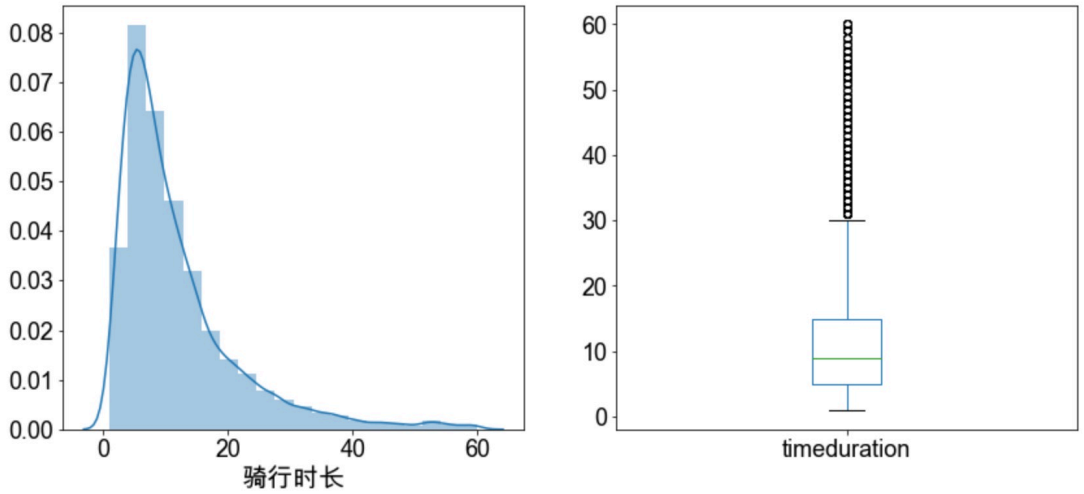
2. 数值型变量

（1）age



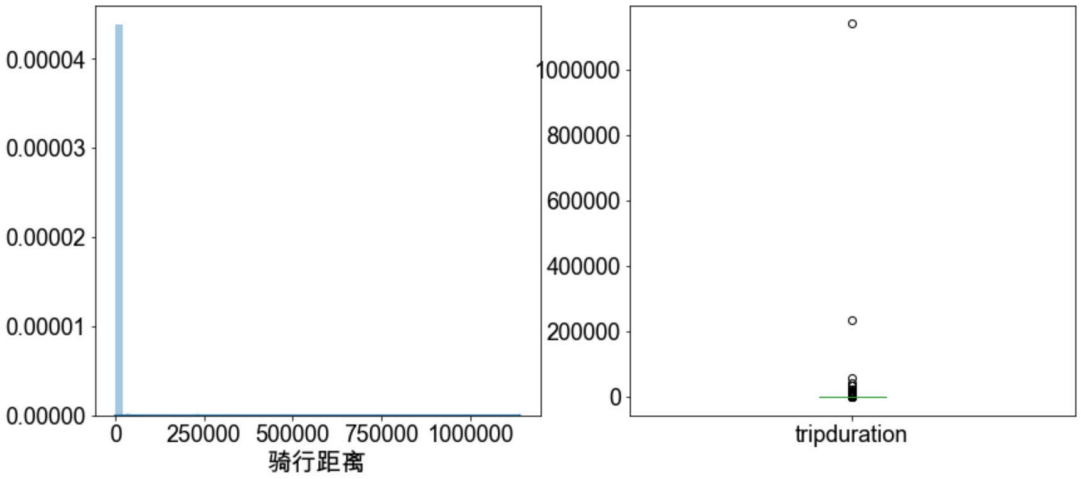
可以看出，摩拜单车用户年龄集中分布在 20-60 岁，缺失值可以选择填充为中位数或众数，而离群值要进行处理。

（2）timeduration



可以看出，摩拜单车用户用车时长集中分布在 20 分钟以下，离群值要进行处理。

（3）tripduration



摩拜单车用户的骑行距离数据分布跨度过大，离群值要进行处理。

四、数据再处理

1. age 的缺失值用中位数填充，并转化为 int 类型

2. age、timeduration、tripduration 离群值处理

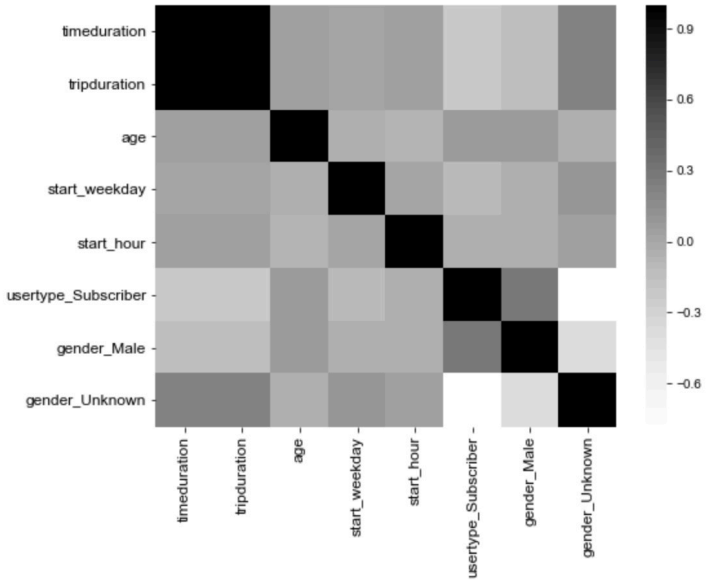
选择 1.5 倍的 IQR 作为 threshold 筛选数据，最终保留 5549 行数据，删除了 878 行数据，占比 14%。

3. 将所有类别变量转换为哑变量，并保留所有原始变量。

五、多变量分析

1. 所有变量间相关性

计算相关性系数矩阵，并用热力图可视化。

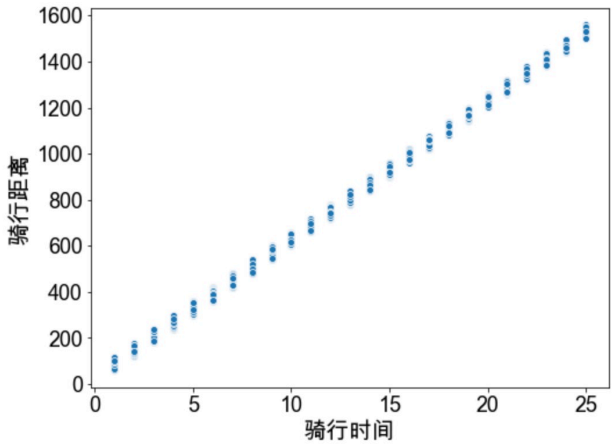


（1）从颜色上看，正相关性比较高的是 timeduration 和 tripduration，gender_Male 和 usertype_Subscriber，gender_Unknown 和 timeduration，gender_Unknown 和 tripduration。

（2）负相关性比较高的是 usertype_Subscriber 和 gender_Unknown，gender_Male 和 gender_Unknown。

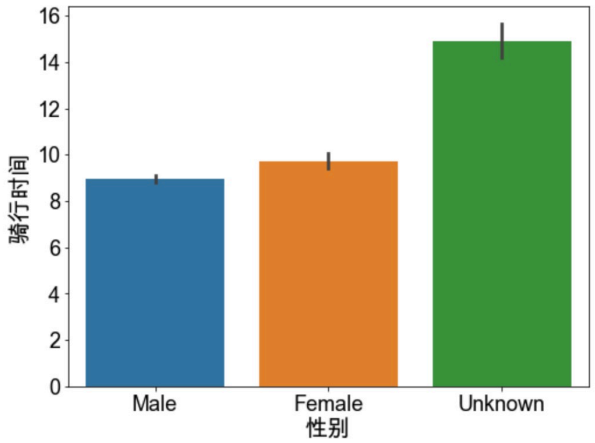
2. 变量与变量之间

（1）骑行距离和骑行时间



骑行距离和骑行时间呈线性相关，因此只看两者其一即可。

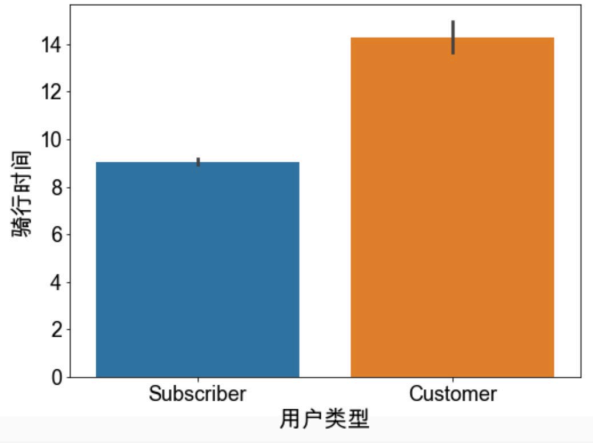
（2）性别与骑行时间



a. 未知性别人群的平均骑行时间要比明确性别的用户平均骑行时间长很多

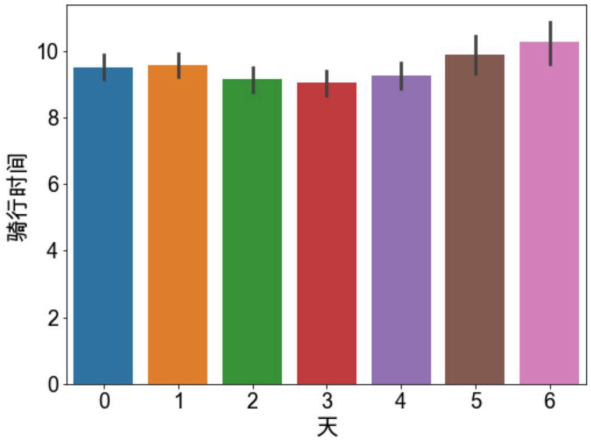
b. 女性比男性平均骑行时间要长

（3）用户类型与骑行时间



付费用户比普通用户平均骑行时间短。

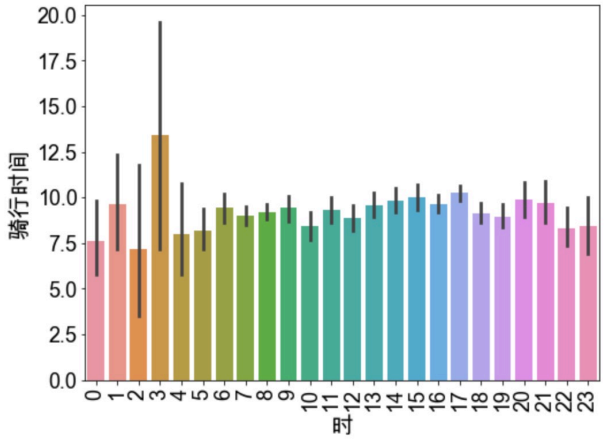
（4）星期几与骑行时间



a. 周末的平均骑行时间比周中长

b. 周一、周二相对周三周四周五平均骑行时间略长一点

（5）小时与骑行时间



a. 凌晨三点的平均骑行时间较长，可能是数据少引起的

b. 其余时间段比波动不大

六、聚类模型建立

1. 特征选择

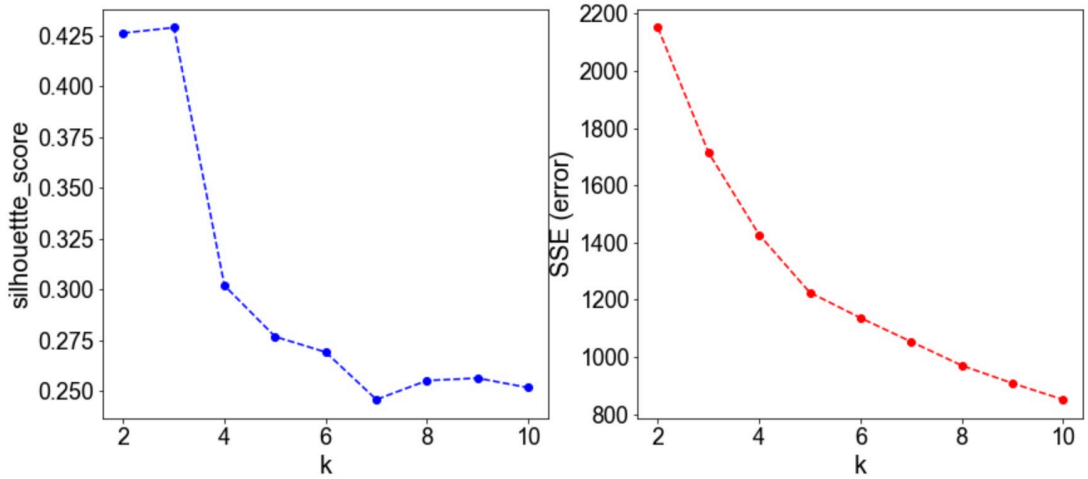
将全部的类别型变量（值非数字，而是字符串的变量）删去，作为训练模型的特征值，并对数据进行归一化处理（选择 MinMaxScaler 方法）。

2. 训练并评估模型

（1）选用 KMeans 模型

（2）对 k 值的选择

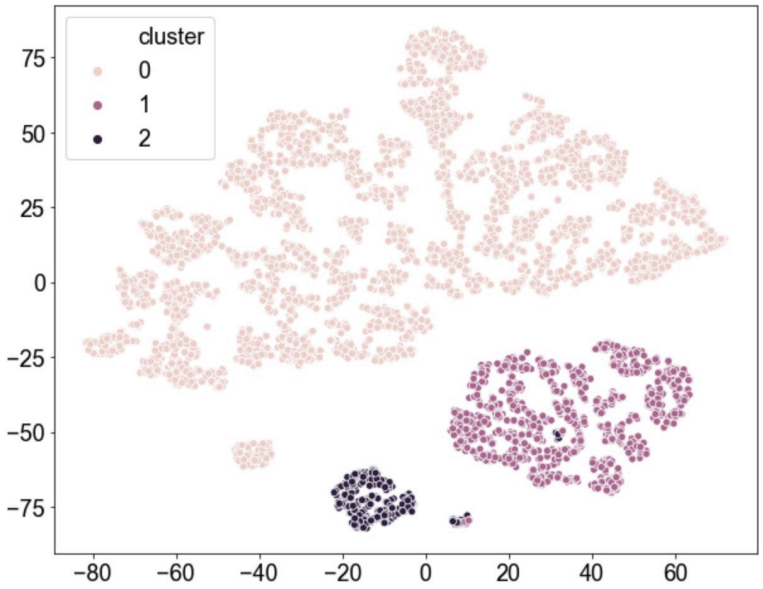
将 k 从 2-10 分别取值后，计算 silhouette_score 和 SSE，绘制如下图。



综合来看，选择 k=3，得到 silhouette_score=0.4289。

（3）结果可视化

用 TSNE 方法将多维数据降成 2 维后，绘制下图。



（4）中心点和业务解读

	timeduration	tripduration	age	start_weekday	start_hour	usertype_Subscriber	gender_Male	gender_Unknown
0	0.331850	0.338141	0.441050	0.411318	0.571967	0.975781	1.000000e+00	-1.457168e-15
1	0.358751	0.363484	0.406522	0.407570	0.583742	0.993838	4.218847e-15	7.922535e-03
2	0.593457	0.589511	0.377599	0.569836	0.618034	0.024648	2.997602e-15	9.295775e-01

可以看出，模型将数据分为三大类，而类与类之间各字段差别明显，但主要是将第三组与其他两组区分开来，如果再查看各字段真实值的平均值，可以得出：

a. 编号为 0 的组内，用户多为稍微年长的男性，主要为付费用户，骑行距离和时间都较短，用车时间偏向于一周的开始。

b. 编号为 1 的组内，用户多为稍微年长的女性付费用户，骑行距离和时间都较短，用车时间偏向于一周的开始。

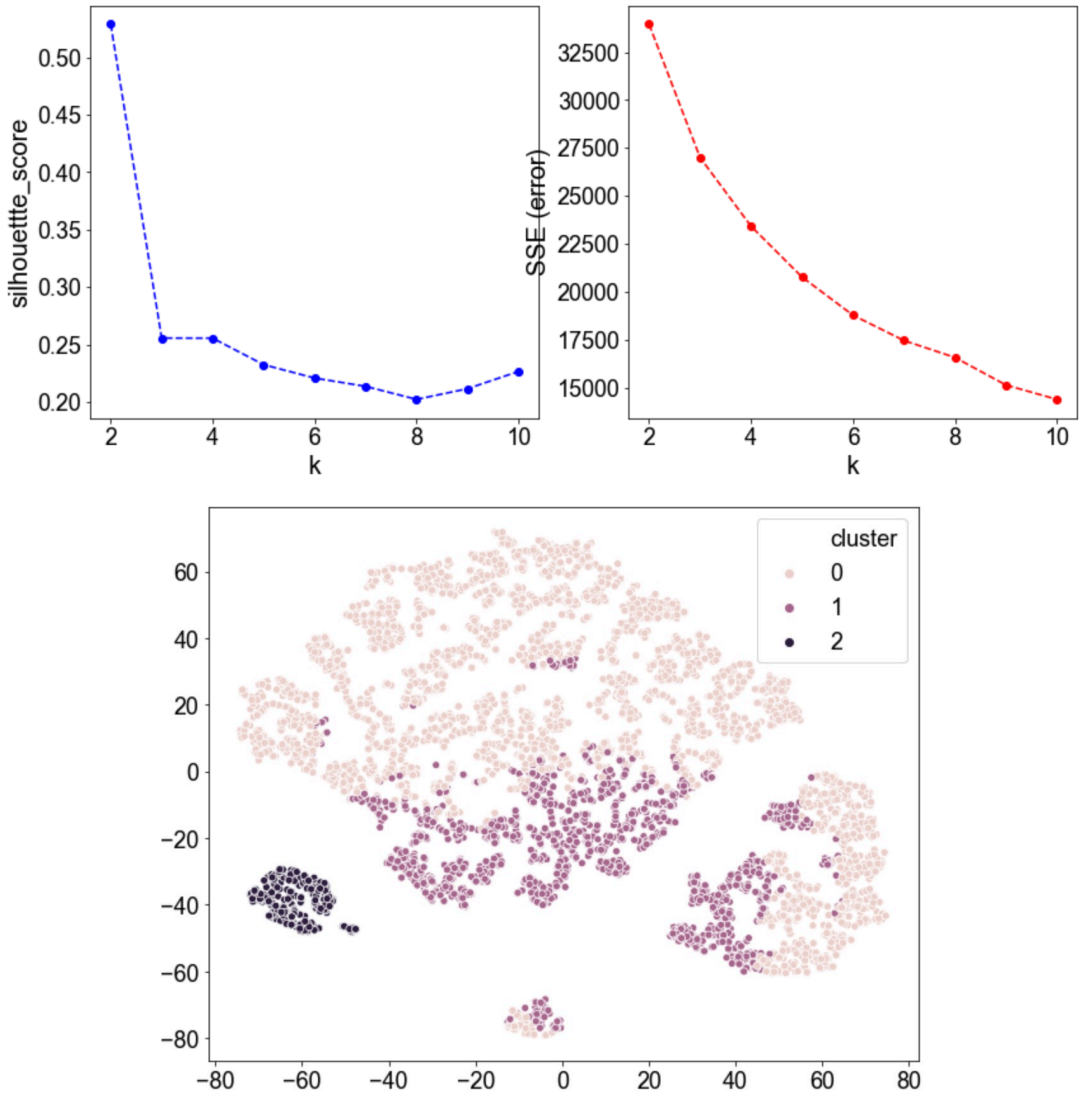
c. 编号为 2 的组内，用户较为年轻，且未知性别的未付费用户，骑行距离和时间都相对较长，用车时间主要集中在周中偏后的下午。

cluster	timeduration	tripduration	age	start_weekday	start_hour	usertype_Subscriber	gender_Male	gender_Unknown
0	8.964398	567.534754	36.524098	2.467910	13.155243	0.975781	1.0	0.000000
1	9.610035	605.499120	35.073944	2.445423	13.426056	0.993838	0.0	0.007923
2	15.242958	944.088028	33.859155	3.419014	14.214789	0.024648	0.0	0.929577

3. 模型优化

（1）数据标准化

将数据标准化后，无论 k 取什么值，误差 SSE 都很大，且 k=3 时 silhouette_score=0.2555，低于初始模型，因此不采纳。



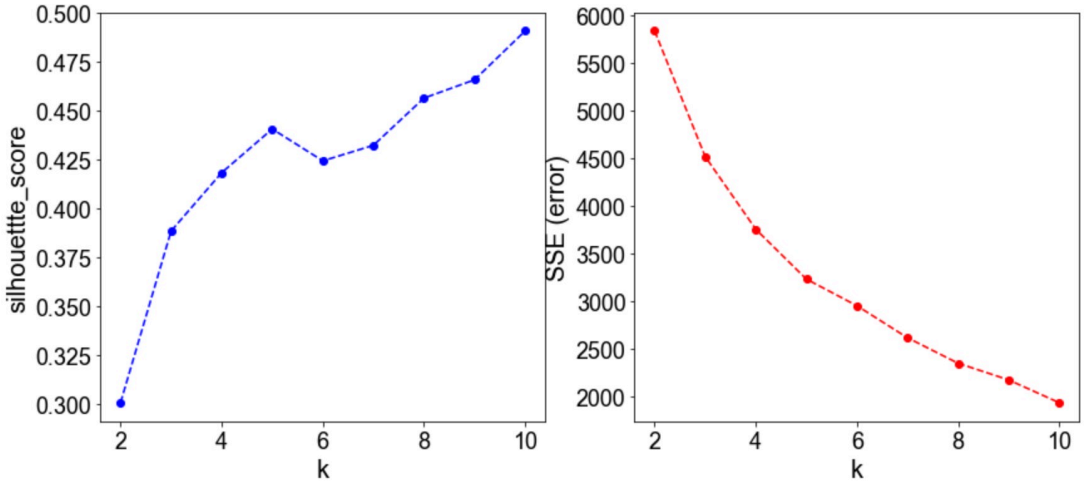
（2）对 start_weekday、start_hour 和 age 分箱操作

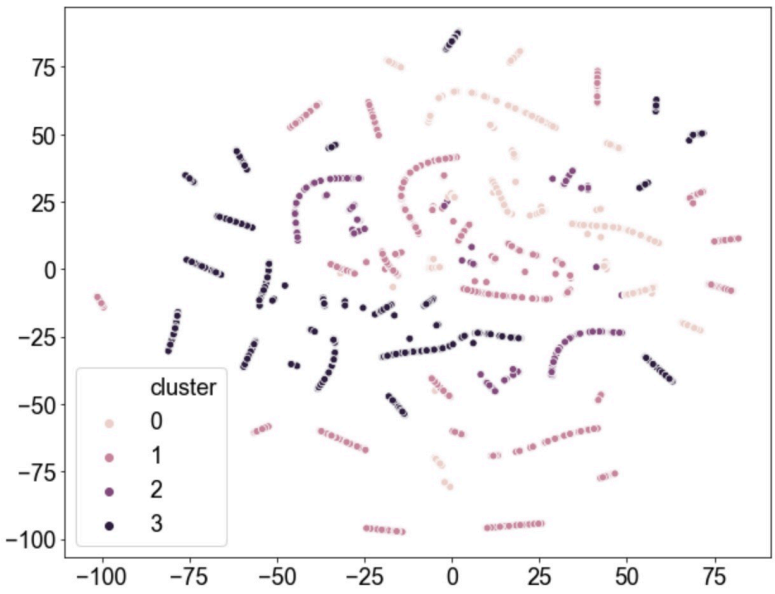
a. start_weekday 分为 weekday 和 weekend 两个类别，新字段命名为 weekday_or_weekend

b. start_hour 分为 0-6，6-12，12-18，18-24 四个类别，新字段命名为 hour_range

c. age 分为 teen（0-25），youth（25-50）和 old（>50），新字段命名为 age_range。

之后对数据归一化（MinMaxScaler），奇怪的是随着 k 值的增加，silhouette_score 也随之增加，初步选定 k=4，此时的 silhouette_score=0.4185，低于初始模型。

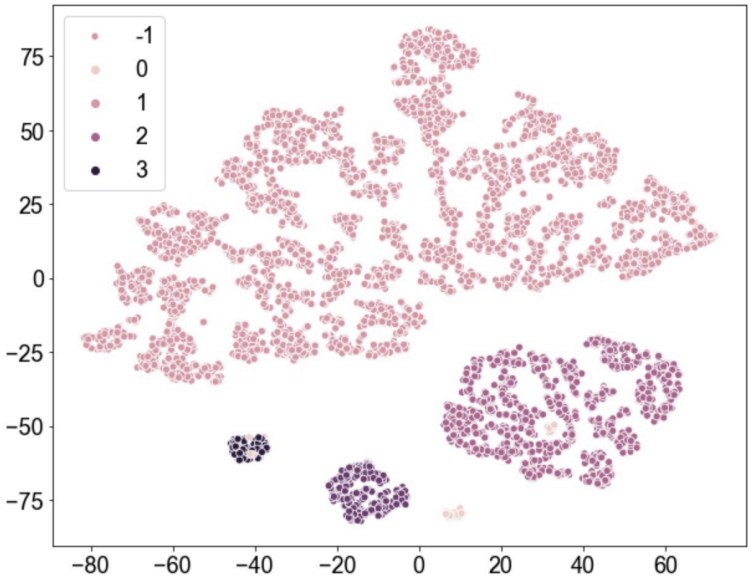




（3）使用 DBSCAN 基于密度的聚类模型

对数据进行归一化处理，在反复尝试调节参数后，选定半径为 eps=0.5，最小成簇点数为 min_samples=20，得到相对最优评分 silhouette_score=0.4326，略优于初始模型。

此时模型将数据分为 5 类，其中一类为噪声，主要类别为 4 大类。


-1
0
1
2
3
75
50
25
0
−25
−50
−75
−80
−60
−40
−20
0
20
40
60