任务 1 数据预处理

任务 1.1 对照附录 1，理解各字段的含义，进行缺失值、重复值等方面的必要处理，将处理结果保存为“task1_1_X.csv”（如果包含多张数据表，X 可从 1 开始往后编号），并在报告中描述处理过程。

1.1.1 对于表 login.csv 进行数据预处理：

（1）读取表中数据。

（2）查表中是否有重复数据，对重复的数据进行删除。

```
## 2 查看是否有重复数据    ----- 暂无重复数据
# 同一用户、同一时间、同一地址视为重复
dataR = data.drop_duplicates(subset=["user_id","login_time","login_place"],keep='first',inplace=False)
print(dataR.duplicated().value_counts())
```

```
False    387144
dtype: int64
```

图 1

（3）查看表中是否有缺失值。

```
## 3 查看数据是否有缺失值 ----- 暂无缺失值
print(dataR.isnull().any())
```

```
user_id        False
login_time     False
login_place    False
dtype: bool
```

图 2

（4）对表中的异常值进行处理。对异常数据删去。

在对 login 表中数据进行观察发现，登录的地址存在异常情况。

以下图为例：用户 3 在 2018/9/10 14:04 在中国北京登录，但是在 2018/9/10 14:36 在中国广东广州登录，仅 32 分钟用户 3 从北京到广州，由常识可以知道这是不可能的。sk1_2.csv”，并在报告中描述处理过程。

	A	B	C
1	user_id	login_time	login_place
2	用户3	2018/9/6 9:32	中国广东广州
3	用户3	2018/9/7 9:28	中国广东广州
4	用户3	2018/9/7 9:57	中国广东广州
5	用户3	2018/9/7 10:55	中国广东广州
6	用户3	2018/9/7 12:28	中国广东广州
7	用户3	2018/9/10 9:18	中国广东广州
8	用户3	2018/9/10 9:53	中国广东广州
9	用户3	2018/9/10 11:28	中国广东广州
10	用户3	2018/9/10 14:04	中国北京
11	用户3	2018/9/10 14:36	中国广东广州
12	用户3	2018/9/10 17:38	中国广东
13	用户3	2018/9/10 18:17	中国广东广州
14	用户3	2018/9/11 9:40	中国广东广州

图 3

任务 1.2 对用户信息表中 recently_logged 字段的“--”值进行必要的处理，将处理结果保存为“task1_2.csv”，并在报告中描述处理过程。

用 duplicates()方法对表 user.csv 中的重复出现的 user_id 用户 ID 进行删除,只保留第一次出现的。因为每个用户注册后的 ID 是唯一的,注册时间是唯一的,不可能有同时刻注册了两个相同的用户名，删除后的数据存储为 new_data。

对用户信息表中 recently_logged 字段的“--”值进行必要的处理，将处理结果保存为“task1_2.csv”。

图 4

计算用户注册时间 registration_time 与最近访问时间 recently_logged 的相差时间,对前 20 名相差时间的数量进行排名,然后根据其数量的比例,对表中 recently_logged 字段”--”进行等比例填补,例如时间差为 0 days 00:00:00,即注册时间等于登录时间的有 8124 条,占比率约为 40%,那填补的用户信息表中 recently_logged 字段的“--”中,注册时间等于最近访问时间占比为 40%

任务 2 平台用户活跃度分析

任务 2.1 分别绘制各省份与各城市平台登录次数热力地图,并分析用户分布情况。

由于绘制的是中国各省份和各城市平台登陆次数热力图，所以在数据的预处理阶段就需要清洗掉在国外登录的用户数据，同时由于出现类似“中国”，后面不带城市信息的数据，这样的数据也需要清洗掉。

在对数据进行合理的分析与清洗之后分别画出各省份与各城市平台登录次数热力地图。在省份的热力分布图中，可以看出广东、湖南、广西、贵州、重庆等地的学生参与网上学习的数量在全国领先。而在各城市的平台登录次数热力图中，可以明显看到广州遥遥领先其他城市，而紧接着便是佛山市、东莞市、深圳市、惠州市等沿海城市。这些城市的学生在该平台网上学习的人数相对较多。

```
In [138]: data1["login_place"].value_counts()
```

```
Out[138]: 中国广东广州    27626
          中国广东       18509
          中国重庆       13163
          中国广东汕头    10146
          中国广东深圳     9098
          中国湖北        8516
          中国            7501
          中国广东惠州     6557
          中国湖北武汉     6534
          中国贵州        6301
          中国广东揭阳     6083
          中国广东湛江     5500
          中国上海        5365
          中国北京        4946
          中国陕西西安     4941
          中国四川成都     4647
          中国广东东莞     4565
          中国广东佛山     4217
          中国浙江杭州     4141
```

图 5

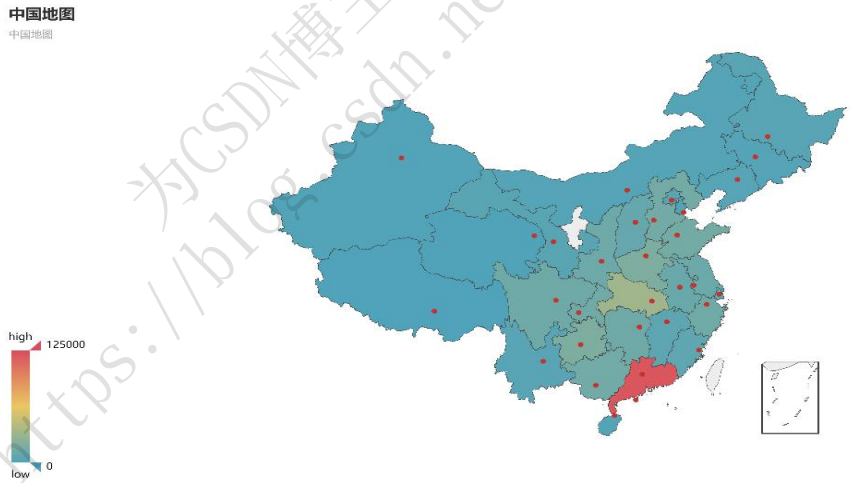

图 6

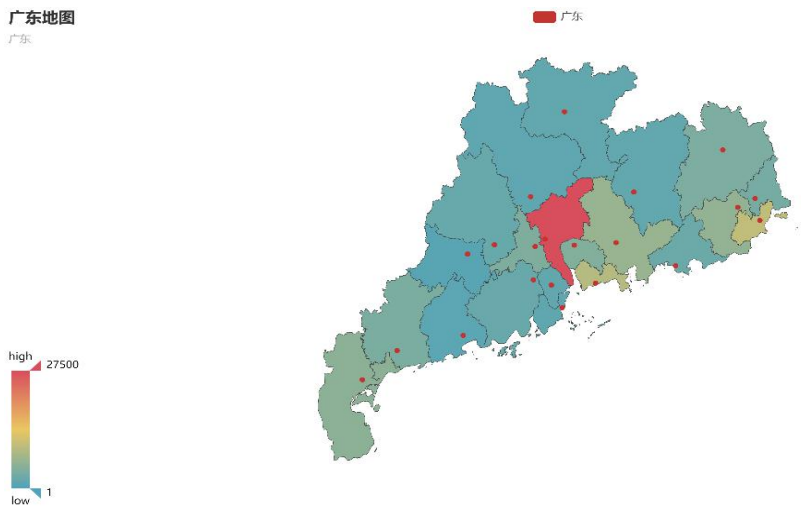

图 7

任务 2.2 分别绘制工作日与非工作日各时段的用户登录次数柱状图，并分析用户活跃的主要时间段。

数据进行异常处理后，对 user.csv 里面的 register_time 进行次数的统计，以 3 个小时分段，根据统计的结果绘制出工作日各时段登录次数和非工作日用户登录次数的柱状图。

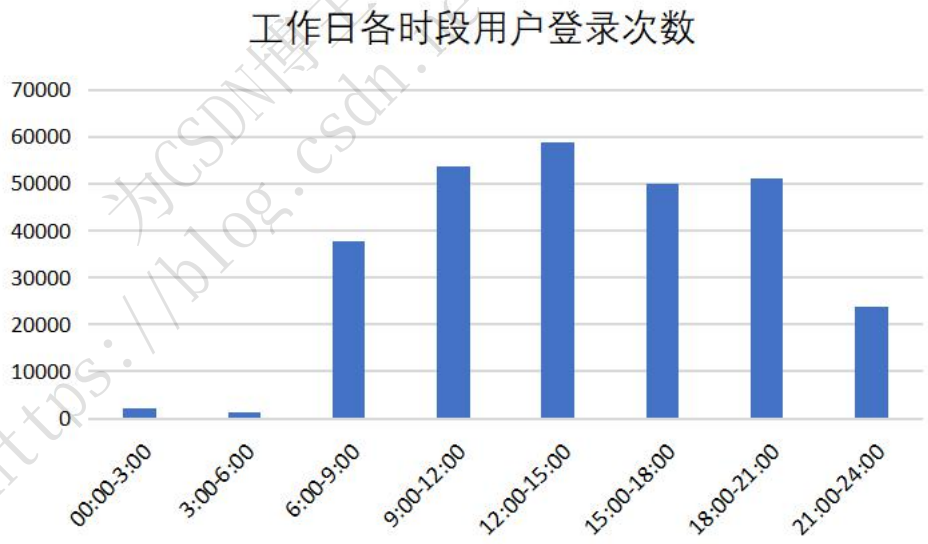

图 8

由上图可以看出，工作日中用户最活跃的是 12:00—15:00，最不活跃的是在 3:00-6:00。学生登录平台的学习时间大致集中在一天的 9:00-21:00，这个时间段是平台最为繁忙的时候。

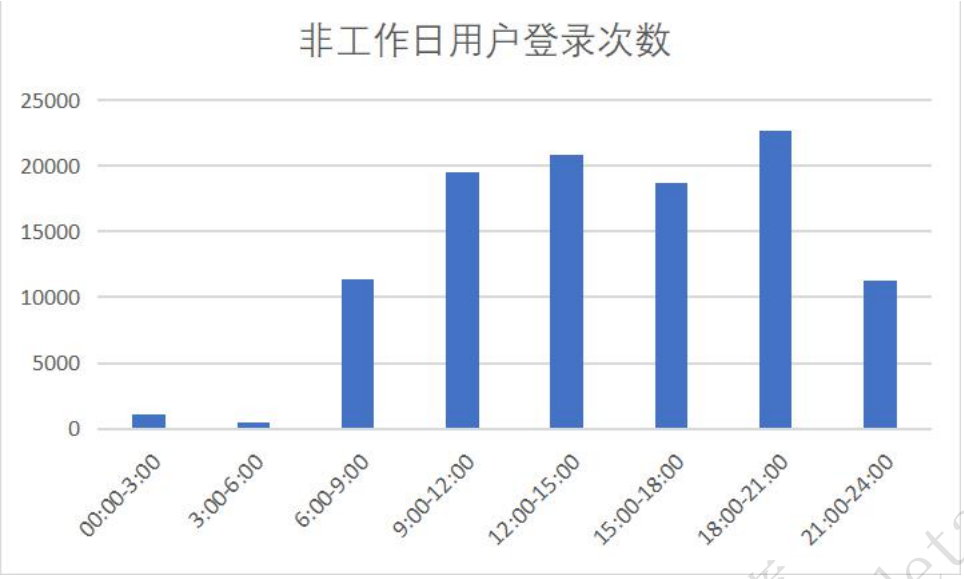

图 9

由上图可以看出，工作日中用户最活跃的是 18:00—21:00，最不活跃的是在 3:00-6:00，用户活跃的阶段主要在一天的 12:00-15:00 和 18:00—21:00。

任务 2.3 记为数据观察窗口截止时间（如：赛题数据的采集截止时间为 2020 年 6 月 18 日），T_i 为用户 i 的最近访问时间，$\sigma_i = T_{end} - T_i$，若 σ_i >90 天，则称用户 i 为流失用户。根据该定义计算平台用户的流失率。

记为数据观察窗口截止时间（如：赛题数据的采集截止时间为 2020 年 6 月 18 日），为用户 i 的最近访问时间，由题目得，若用户 i 的登录时间间隔>90 天，则称该用户为流失的用户。对于流失率的计算公式如下（为流失的的总数，total 为用户的总数）：

$$\text{rate} = \frac{\sigma\sigma}{\text{total}}$$

最终计算出的流失率为 58.763%

任务 2.4 根据任务 2.1 至任务 2.3，分析平台用户的活跃度，为该教育平台的线上管理提供决策提供建议。

根据前面的用户时间段的统计，可以发现，登录平台的人员主要集中在广东、湖南、广西和贵州等地，而城市则是主要集中在广州、深圳、东莞等地，建议公司可以把公司产品向新疆、西藏等较远离沿海的省份和城市推广。大量用户主要

集中在 9：00——21：00，所以在这段时间里面平台需要重视大量人流对平台造成的影响，比如:平台因人数过多而导致的卡顿或者是用户不能进入平台的后台维护。

任务 3 线上课程推荐

任务 3.1 根据用户参与学习的记录，统计每门课程的参与人数，计算每门课程的受欢迎程度，列出最受欢迎的前 10 门课程，并绘制相应的柱状图。受欢迎程度定义如下：

$$r_i = \frac{Q_i - Q_{\min}}{Q_{\max} - Q_{\min}}$$

其中，r_i 为第 i 门课程的受欢迎程度，Q_i 为参与第门课程学习的人数，$Q_{\max}$ 和 $Q_{\min}$ 分别为所有课程中参与人数最多和最少的课程所对应的人数。

统计每门课程的参与人数，利用 python 中 value_counts() 函数对 study_information.csv 表中的 course_id 列进行累计求和，则可以得到每一门课程的人数。根据统计，得到结果如图 10（部分数据）：

```
pd.set_option('display.max_rows', None)
data["course_id"].value_counts()
```

```
课程76     13265
课程31      9521
课程17      8505
课程191     7126
课程180     6223
课程52      6105
课程34      5709
课程171     5437
课程50      5342
课程12      4829
课程26      4716
课程48      4520
课程141     4429
课程130     4373
课程133     4274
课程163     4067
课程40      4014
课程51      4011
课程135     4006
```

图 10

根据统计得到的结果，可以知道课程参与人数最多的是课程 76，共有 13265 人参加该课程，而参加人数最少的课程有课程 90、课程 91、课程 92、课程 93 皆为 1 人。由上面的统计结果，取出前十名课程，由 excel 绘制出柱状图，如下

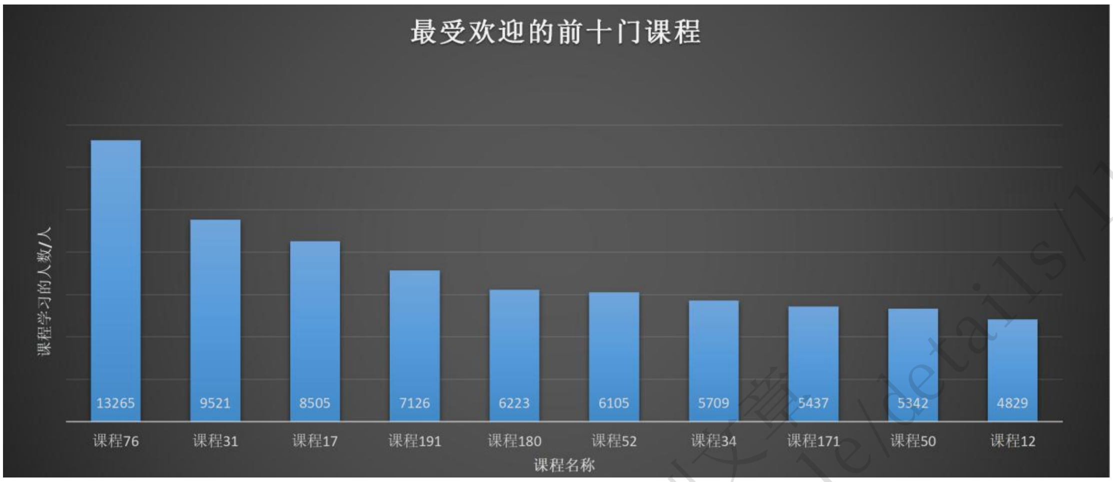

图 11

根据题目中给出的对受欢迎程度的定义式：$r_i = \dfrac{Q_i - Q_{\min}}{Q_{\max} - Q_{\min}}$，以及上面对各种课程数据的统计结果，代入即可求得各门课程的受欢迎程度。

表 1

课程名	参加人数	课程受欢迎程度
课程 76	13265	1
课程 31	9521	0.717732207
课程 17	8505	0.641133896
课程 191	7126	0.537168275
课程 180	6223	0.469089264
课程 52	6105	0.460193004
课程 34	5709	0.430337756
课程 171	5437	0.409831122
课程 50	5342	0.402668878
课程 12	4829	0.363992762
课程 26	4716	0.355473462
课程 48	4520	0.340696622

课程 141	4429	0.333835947
课程 130	4373	0.329613993
课程 133	4274	0.322150181
课程 163	4067	0.306544029
课程 40	4014	0.302548251
课程 51	4011	0.302322075
课程 135	4006	0.301945115
课程 29	3998	0.301341978

任务 3.2 根据用户选择课程情况，构建用户和课程的关系表（二元矩阵），使用基于物品的协同过滤算法计算课程之间的相似度，并结合用户已选课程的记录，为总学习进度最高的 5 名用户推荐 3 门课程。

首先对 learn_process 中的数据进行处理成浮点型,然后根据 user_id 分组求和出用户的学习进度的均值,数值为 1，即所选的课程学习进度均为 100%，然后在从学习进度都为 100%的用户里面，统计其学习课程门数的多少并对其进行排序，最终选出五名用户，分别是用户 27953、用户 17520、用户 17548、用户 39420、用户 23200。

用机器学习里面的协同过滤算法[推荐算法],采用杰卡德相似系数(Jaccard similarity coefficient)计算方法

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

其实就是集合的交集除并集,最终为总学习进度最高 5 名用户推荐 3 门课程。

user_id	推荐课程1	推荐课程2	推荐课程3
用户27953	课程225	课程98	课程60
用户17520	课程62	课程76	课程12
用户17548	课程4	课程26	课程158
用户39420	课程45	课程31	课程226
用户23200	课程19	课程50	课程85

图 12

任务 3.3 在任务 3.1 和任务 3.2 的基础上，结合用户学习进度数据，分析付费课程和免费课程的差异，给出线上课程的综合推荐策略。

受欢迎程度	课程名称	价格
1	课程 76	0
2	课程 31	109
3	课程 17	299
4	课程 191	0
5	课程 180	0
6	课程 52	0
7	课程 34	299
8	课程 171	299
9	课程 50	0
10	课程 12	0

图 13

由上面的表格可以看出在最受欢迎的课程前十名里面，免费的课程占据较大，而收费的课程价格也集中在 199 元与 299 元。而学习的进度则是在 199~299 元区间价格里面较为快速。而相当一部分 0 元价格的课程，学习进度较差。所以在考虑价格与学习进度的同时，对课程的价格进行设置，则把其区间设置在 199~299 元的效果最好。