

DeepSeek在数据分析场景中的应用



>> 今天的学习目标

DeepSeek在数据分析场景中的应用

- DeepSeek对投放数据进行分析
- DeepSeek计算数据表中的投放ROI
- DeepSeek生成Excel函数
- DeepSeek进行多张Excel报表处理
- DeepSeek进行数据可视化
- DeepSeek进行数据洞察

CASE：投放数据分析

CASE：投放数据分析

TO DO：小张是运营负责人，要给老板汇报投放数据，如何从收集上的渠道投放数据中，得到一些数据洞察？

渠道名称	日期	投放费用	展示量	点击量	转化量	次日留存率	7日留存率	转化ROI(%)
社交媒体	2025/2/11	3793.97	29502	479	2	0.3819	0.2571	0.05
行业论坛	2025/2/4	753.85	7889	79	1	0.3618	0.2256	0.13
信息流广告	2024/11/29	1810.53	14437	313	3	0.3814	0.2255	0.17
信息流广告	2025/1/2	4924.54	60417	2248	26	0.3395	0.1707	0.53
短视频平台	2025/2/17	571.85	4861	181	2	0.2722	0.1814	0.35
社交媒体	2025/1/7	3481.35	36267	1155	8	0.154	0.1069	0.23
KOL合作	2024/12/26	4669.96	38226	1254	16	0.4689	0.3246	0.34
行业论坛	2025/1/4	3140.38	32009	2200	12	0.3919	0.2089	0.38
搜索引擎	2024/12/25	4169.58	50293	2054	45	0.4431	0.2533	1.08
.....

channel_performance.csv

CASE：投放数据分析

将数据上传到内网的DeepSeek

📎 [channel_performance.csv](#)

我要跟老板汇这个投放数据，请帮我分析这个投放数据，是否有问题，都有哪些特征，你对这个投放有什么建议。

> 已深度思考（用时 0.0 秒）

好的，我来帮你分析一下这个投放数据。首先，我需要理解数据中的各个字段，包括渠道名称、日期、投放费用、展示量、点击量、转化量、次日留存率、7日留存率以及转化ROI（%）。这些数据覆盖了多个渠道，如社交媒体、搜索引擎、信息流广告、短视频平台、KOL合作、行业论坛、邮件营销、应用商店和电商平台等。

从整体上看，投放费用在不同渠道之间差异较大，例如社交媒体和信息流广告的投放费用相对较高，而短视频平台和垂直媒体的费用则相对较低。展示量和点击量也呈现出相似的分布，高费用渠道通常伴随着较高的展示量和点击量，但转化量却不一定成正比。

在转化率方面，部分渠道表现较好，如搜索引擎渠道在某些日期的转化量较高，转化ROI也相对较高。而其他渠道如邮件营销和应用商店的转化率相对较低，ROI也较低。次日和7日的留存率普遍在20%-50%之间，显示出用户留存情况尚可，但仍有提升空间。

CASE：投放数据分析

此外，数据中存在一些异常值，例如某些渠道的转化ROI为零或负数，这可能意味着投放策略需要调整或优化。例如，信息流广告在2024-12-06的转化ROI为0%，可能是投放效果不佳或转化率过低。

基于以上分析，我有以下建议：

优化投放策略：针对转化ROI较低的渠道，进行更精细的投放策略调整，例如优化广告内容、投放时间或目标受众。

提升用户留存：对于留存率较低的渠道，可以考虑加强用户引导或提供更好的用户体验，以提高用户的留存率。

资源分配优化：增加对高ROI渠道的投放预算，同时减少对低效渠道的投入，以提升整体投放效果。

数据监控与分析：定期监控各渠道的投放数据，及时发现异常情况并进行调整，以确保投放效果的最大化。

通过以上措施，可以有效提升整体的投放效果，优化资源分配，提高ROI。

DeepSeek先对上传的csv进行思考，然后再给出洞察结论

CASE：投放数据分析

好的，我来帮你分析一下这个投放数据。首先，我需要理解数据中的各个字段，包括渠道名称、日期、投放费用、展示量、点击量、转化量、次日留存率、7日留存率以及转化ROI（%）。这些数据覆盖了多个渠道，如社交媒体、搜索引擎、信息流广告、短视频平台、KOL合作、行业论坛、邮件营销、应用商店和电商平台等。

从整体上看，投放费用在不同渠道之间差异较大，例如社交媒体和信息流广告的投放费用相对较高，而短视频平台和垂直媒体的费用则相对较低。展示量和点击量也呈现出相似的分布，高费用渠道通常伴随着较高的展示量和点击量，但转化量却不一定成正比。

在转化率方面，部分渠道表现较好，如搜索引擎渠道在某些日期的转化量较高，转化ROI也相对较高。而其他渠道如邮件营销和应用商店的转化率相对较低，ROI也较低。次日和7日的留存率普遍在20%-50%之间，显示出用户留存情况尚可，但仍有提升空间。

此外，数据中存在一些异常值，例如某些渠道的转化ROI为零或负数，这可能意味着投放策略需要调整或优化。例如，信息流广告在2024-12-06的转化ROI为0%，可能是投放效果不佳或转化率过低。

基于以上分析，我有以下建议：

- 优化投放策略：**针对转化ROI较低的渠道，进行更精细的投放策略调整，例如优化广告内容、投放时间或目标受众。
- 提升用户留存：**对于留存率较低的渠道，可以考虑加强用户引导或提供更好的用户体验，以提高用户的留存率。
- 资源分配优化：**增加对高ROI渠道的投放预算，同时减少对低效渠道的投入，以提升整体投放效果。
- 数据监控与分析：**定期监控各渠道的投放数据，及时发现异常情况并进行调整，以确保投放效果的最大化。

通过以上措施，可以有效提升整体的投放效果，优化资源分配，提高ROI。

CASE：计算数据表中的投放ROI

CASE：计算数据表中的投放ROI

TO DO：小张是运营负责人员，对多个视频号进行了投放，目的是要得到增加用户关注，
这里有一份投放的表格，如何计算每个视频账号的投放ROI？

视频号名称	投放地区	内容主题	总消耗金额	投放天数	新增关注数	平均播放完成率	分享次数	关注成本效率(每百元)
理想汽车_评测中心_1000	上海	新车发布	39164.95	4	3670	0.94	802	9.37
蔚来汽车_评测中心_1001	上海	新车发布	78257.31	26	6764	0.36	2006	8.64
宝马中国_旗舰店_1002	成都	新车发布	11782.29	29	617	0.66	145	5.24
蔚来汽车_车友会_1003	广州	保养指南	20689.97	9	1053	0.5	250	5.09
蔚来汽车_评测中心_1004	武汉	新车发布	35406.06	27	1544	0.77	406	4.36
理想汽车_车友会_1005	广州	性能评测	49831.42	14	4833	0.35	1336	9.7
懂车帝_评测中心_1006	上海	保养指南	78076.45	5	10365	0.78	2800	13.28
.....

CASE：计算数据表中的投放ROI

将数据上传到内网的DeepSeek

 video_channel_performan...
XLSX 9.91KB

帮我统计不同账号的投放ROI，制作表格给我

好的，用户需要统计不同账号的投放ROI，并制作表格。首先，我需要理解ROI是什么。ROI（投资回报率）通常是指收益与投资的比率。但在这个数据表中，没有直接的收益数据，只有消耗金额和一些互动指标，比如新增关注数、分享次数等。所以可能需要用其他指标来近似计算ROI。

用户提到的“关注成本效率(每百元)”可能是一个关键指标，它可能表示每花费一百元能获得多少关注。这个指标可能已经部分反映了ROI。但ROI一般需要收益与成本的对比，这里可能需要明确如何定义收益。

假设用户希望用新增关注数作为收益的指标，那么ROI可以计算为 $(\text{新增关注数} / \text{总消耗金额}) * 100$ ，即每元消耗带来的新增关注数。或者，如果“关注成本效率”已经是每百元带来的关注数，那么可能需要确认是否需要直接使用该指标作为ROI。

表格中有不同的账号，比如“理想汽车_评测中心_1000”、“蔚来汽车_评测中心_1001”等，需要按账号名称分组统计。需要注意账号名称可能包含品牌和类型，比如“评测中心”、“车友会”等，但用户要求按账号名称整体统计，而不是按品牌或类型分组。

CASE：计算数据表中的投放ROI

需要检查是否有重复的账号名称，例如“理想汽车_评测中心_1000”和“理想汽车_评测中心_1009”是否属于同一账号的不同投放，还是不同的账号。根据数据中的编号（如1000、1001等），可能每个账号都是唯一的，因此需要按全名分组。

接下来，计算每个账号的总消耗金额和对应的新增关注数，以及可能的其他指标。然后根据定义的ROI公式进行计算。例如，如果ROI是新增关注数除以总消耗金额，那么每个账号的ROI就是其所有投放记录的新增关注数总和除以总消耗金额总和。

需要确保数据中的每个账号在表格中只出现一次，汇总其所有投放记录的数据。例如，“蔚来汽车_评测中心”在数据中有多个条目（1001和1004），需要将它们合并计算总消耗和总新增关注数。

因此，步骤可能如下：

1. 按账号名称分组。
2. 对每个组，汇总总消耗金额和新增关注数。
3. 计算ROI（如新增关注数 / 总消耗金额）。
4. 可能需要保留其他相关指标，如平均播放完成率、分享次数等，但用户仅要求ROI，所以可能只需展示账号名称和对应的ROI。

需要验证是否存在数据错误，例如数值是否正确，是否有重复或异常值。例如，检查总消耗金额和新增关注数是否合理，是否有负值或过大值。最后，将结果整理成表格，确保账号名称正确，ROI计算准确，并可能按ROI排序以便用户查看效果最好的账号。

CASE：计算数据表中的投放ROI

以下是不同视频号名称的投放ROI统计表，ROI计算方式为 **新增关注数 / 总消耗金额**（即每元消耗带来的新增关注数）：

视频号名称	总消耗金额	总新增关注数	ROI (新增关注/元)
理想汽车_评测中心_1000	39,164.95	3,670	0.0937
蔚来汽车_评测中心_1001	78,257.31	6,764	0.0864
宝马中国_旗舰店_1002	11,782.29	617	0.0524
蔚来汽车_车友会_1003	20,689.97	1,053	0.0509
蔚来汽车_评测中心_1004	35,406.06	1,544	0.0436
理想汽车_车友会_1005	49,831.42	4,833	0.0970
懂车帝_评测中心_1006	78,076.45	10,365	0.1328
比亚迪_改装达人_1007	31,130.90	4,990	0.1603
懂车帝_旗舰店_1008	15,092.12	485	0.0321

CASE：计算数据表中的投放ROI

理想汽车_评测中心_1009	29,414.31	2,180	0.0741
特斯拉_改装达人_1010	26,903.40	2,844	0.1057
比亚迪_评测中心_1011	59,935.47	1,972	0.0329
理想汽车_官方号_1012	17,580.22	1,308	0.0744
特斯拉_车友会_1013	35,272.88	2,399	0.0680
蔚来汽车_旗舰店_1014	46,850.76	1,951	0.0416
理想汽车_改装达人_1015	36,900.99	2,407	0.0652
懂车帝_车友会_1016	8,671.74	782	0.0902
理想汽车_改装达人_1017	31,048.92	3,435	0.1106
特斯拉_车友会_1018	43,718.21	1,507	0.0345

说明：

- ROI计算公式：** 总新增关注数 / 总消耗金额，结果保留4位小数，表格中已简化展示。
- 数据分组：** 按“视频号名称”唯一分组，合并同一账号的多条投放记录（如“蔚来汽车_评测中心_1001”和“蔚来汽车_评测中心_1004”视为不同账号）。
- 高ROI账号：** 比亚迪_改装达人_1007 (ROI=0.1603)、懂车帝_评测中心_1006 (ROI=0.1328) 表现最佳。

DeepSeek对ROI进行了结算，给出了最终的统计表格。

并对ROI公式，数据分组情况，以及高ROI账号进行了说明

CASE：计算数据表中的投放ROI

Thinking：如果你想要知道哪些账号ROI低，下次不进行投放，该如何做？

帮我找到ROI最低的TOP3个账号

思考（用时 20 秒） ^

好的，我现在需要帮用户找到ROI最低的TOP3个账号。首先，我需要明确ROI是如何计算的。根据之前的回复，ROI是新增关注数除以总消耗金额，也就是每元消耗带来的新增关注数。现在需要找出这个值最低的三个账号。

首先，我需要回顾一下之前生成的表格，里面已经列出了每个视频号名称的ROI。我要把这些ROI值从低到高排序，然后取前三个。不过，用户可能希望看到的是数值最小的三个，也就是ROI最低的三个账号。

让我先检查表格中的ROI数据：

- 懂车帝_旗舰店_1008: 0.0321
- 比亚迪_评测中心_1011: 0.0329
- 特斯拉_车友会_1018: 0.0345
- 宝马中国_旗舰店_1002: 0.0524
- 蔚来汽车_车友会_1003: 0.0509
- 蔚来汽车_旗舰店_1014: 0.0416
- 理想汽车_评测中心_1009: 0.0741
- 等等...

CASE：计算数据表中的投放ROI

看起来最低的三个ROI是：

1. 懂车帝_旗舰店_1008: 0.0321
2. 比亚迪_评测中心_1011: 0.0329
3. 特斯拉_车友会_1018: 0.0345

不过，我需要确认是否有其他账号的ROI可能更低，或者是否有重复的账号。比如，是否同一个品牌的不同账号需要合并？但根据之前的说明，每个视频号名称都是独立账号，不需要合并。因此，直接按照表格中的数据排序即可。

再检查一遍数据：

- 懂车帝_旗舰店_1008: 0.0321 (最低)
- 比亚迪_评测中心_1011: 0.0329 (第二低)
- 特斯拉_车友会_1018: 0.0345 (第三低)

接下来是宝马中国_旗舰店_1002的0.0524，这个数值比前三者高，所以前三低确实是这三个账号。

确认无误后，就可以将这些结果整理出来，告诉用户这三个账号是ROI最低的TOP3。

CASE：生成Excel函数

CASE：生成Excel函数

TO DO：编写标黄区域的Excel函数

	A	B	C	D	E	F	G	H	I	J
1	销售日期	地区	产品	金额 (万元)			地区	产品	月份	金额合计
2	2025/1/3	北京	存款	50			北京	存款	1月	
3	2025/1/6	北京	贷款	300			上海	贷款	2月	
4	2025/1/23	上海	贷款	200						
5	2025/1/27	北京	理财	500						
6	2025/2/4	上海	存款	250						
7	2025/2/5	上海	贷款	400						
8	2025/2/8	深圳	贷款	700						
9	2025/2/12	上海	存款	600						
10	2025/2/13	深圳	理财	150						
11	2025/2/16	北京	理财	800						

比如北京、存款 在1月金额合计是多少？

上海、贷款 在2月金额合计是多少？

DeepSeek的读图能力一般，如何让它更好的理解Excel数据表的内容？

CASE：生成Excel函数

Step1，使用通义千问进行视图

	A	B	C	D	E	F	G	H	I	J
1	销售日期	地区	产品	金额 (万元)			地区	产品	月份	金额合计
2	2025/1/3	北京	存款	50			北京	存款	1月	
3	2025/1/6	北京	贷款	300			上海	贷款	2月	
4	2025/1/23	上海	贷款	200						
5	2025/1/27	北京	理财	500						
6	2025/2/4	上海	存款	250						
7	2025/2/5	上海	贷款	400						
8	2025/2/8	深圳	贷款	700						
9	2025/2/12	上海	存款	600						
10	2025/2/13	深圳	理财	150						
11	2025/2/16	北京	理财	800						

帮我理解这张Excel的含义

- 1、每列的字段，然后每列都需要有具体内容示例
- 2、检查是否说出了所有列名，即对应的内容示例
- 3、有几个标黄区域，每个标黄的想要计算什么，方便后续编写对应Excel函数

CASE：生成Excel函数

这张Excel表格包含了销售数据的详细信息，分为两部分：左侧是具体的销售记录，右侧是对这些记录的汇总。以下是各列的解释和示例：

左侧（具体销售记录）

1. A列：销售日期

- 示例：2025/1/3
- 说明：记录每笔销售的具体日期。

2. B列：地区

- 示例：北京
- 说明：记录销售发生的地区。

3. C列：产品

- 示例：存款
- 说明：记录销售的产品类型。

4. D列：金额（万元）

- 示例：50
- 说明：记录每笔销售的金额，单位为万元。

右侧（汇总信息）

1. G列：地区

- 示例：北京
- 说明：汇总时选择的地区。

2. H列：产品

- 示例：存款
- 说明：汇总时选择的产品类型。

3. I列：月份

- 示例：1月
- 说明：汇总时选择的月份。

4. J列：金额合计

- 示例：标黄区域
- 说明：根据G、H、I列的选择条件，计算对应的金额合计。

标黄区域的功能

标黄区域（J列）用于计算特定条件下（地区、产品、月份）的金额合计。例如：

- 如果选择“北京”地区、“存款”产品、“1月”，则计算1月份北京地区的存款金额合计。
- 如果选择“上海”地区、“贷款”产品、“2月”，则计算2月份上海地区的贷款金额合计。

将通义千问对Excel的理解，发给DeepSeek，并让其对标黄区域编写Excel公式

CASE：生成Excel函数

Step2，使用DeepSeek生成Excel函数

我有一个Excel

这张Excel表格包含了销售数据的详细信息，分为两部分：左侧是具体的销售记录，右侧是对这些记录的汇总。以下是各列的解释和示例：

左侧（具体销售记录）

A列：销售日期

示例：2025/1/3

说明：记录每笔销售的具体日期。

B列：地区

示例：北京

说明：记录销售发生的地区。

C列：产品

示例：存款

说明：记录销售的产品类型。

D列：金额（万元）

示例：50

说明：记录每笔销售的金额，单位为万元。

右侧（汇总信息）

G列：地区

示例：北京

说明：汇总时选择的地区。

H列：产品

示例：存款

说明：汇总时选择的产品类型。

I列：月份

示例：1月

说明：汇总时选择的月份。

J列：金额合计

示例：标黄区域

说明：根据G、H、I列的选择条件，计算对应的金额合计。

标黄区域的功能

标黄区域（J列）用于计算特定条件下（地区、产品、月份）的金额合计。例如：

如果选择“北京”地区、“存款”产品、“1月”，则计算1月份北京地区的存款金额合计。

如果选择“上海”地区、“贷款”产品、“2月”，则计算2月份上海地区的贷款金额合计。

====

请编写标黄的区域的Excel公式

CASE：生成Excel函数

已深度思考（用时 52 秒） ^

嗯，用户需要编写Excel公式来计算特定条件下的金额合计，也就是标黄区域J列。首先，我得理解他们的数据结构。左边有销售记录，包括日期、地区、产品、金额；右边是汇总部分，需要根据G、H、I列的条件来求和。

首先，用户的条件是地区、产品和月份。地区在G列，产品在H列，月份在I列。但月份显示的是“1月”这样的文本，而销售日期在A列是具体的日期，比如2025/1/3。所以需要把A列的日期转换成月份，然后和I列的条件匹配。

接下来，我需要考虑使用什么函数。通常这种情况可以用SUMIFS函数，因为它可以处理多条件。不过，月份的处理可能需要转换一下，因为SUMIFS的条件区域需要和条件值匹配。比如，用户输入的月份是“1月”，而A列的日期是具体的日期，所以需要将A列的日期提取出月份，然后转换成“1月”的格式，再进行比较。

不过，Excel的SUMIFS函数不支持直接在条件中使用数组公式或函数的结果。所以可能需要使用SUMPRODUCT函数，因为它可以处理数组运算。或者，使用辅助列来提取月份，再用SUMIFS。但用户可能不希望修改原数据，所以辅助列可能不合适。

另一个方法是使用TEXT函数将A列的日期转换为“m月”格式，然后在SUMIFS的条件中使用。不过，SUMIFS的条件区域需要是实际的单元格区域，而不能是数组公式的结果。所以，这可能需要使用数组公式，比如结合SUMPRODUCT和TEXT函数……

CASE：生成Excel函数

根据需求，可以使用以下公式实现标黄区域的动态汇总：

```
excel  
复制  
=SUMPRODUCT(  
    (B$2:B$1000=G2)*(C$2:C$1000=H2)*  
    (TEXT(A$2:A$1000,"m月")=I2)*D$2:D$1000  
)
```

公式原理：

1. 范围设定：

- **B\$2:B\$1000**：地区列（固定范围）
- **C\$2:C\$1000**：产品列（固定范围）
- **A\$2:A\$1000**：日期列（固定范围）
- **D\$2:D\$1000**：金额列（固定范围）
- 根据实际数据量调整 **1000** 这个上限值

2. 条件匹配：

- **(B\$2:B\$1000=G2)**：匹配指定地区
- **(C\$2:C\$1000=H2)**：匹配指定产品
- **(TEXT(A\$2:A\$1000,"m月")=I2)**：将日期转为 "1月" 格式后匹配指定月份

3. 动态计算：

- 通过 **SUMPRODUCT** 对满足所有条件的金额进行求和
- 使用固定引用 **\$** 保持范围不变，条件单元格 **G2/H2/I2** 可随公式下拉自动更新

```
=SUMPRODUCT(  
    (B$2:B$1000=G2)*(C$2:C$1000=H2)*  
    (TEXT(A$2:A$1000,"m月")=I2)*D$2:D$1000  
)
```



	A	B	C	D	E	F	G	H	I	J
1	销售日期	地区	产品	金额 (万元)			地区	产品	月份	金额合计
2	2025/1/3	北京	存款	50			北京	存款	1月	50
3	2025/1/6	北京	贷款	300			上海	贷款	2月	400
4	2025/1/23	上海	贷款	200						
5	2025/1/27	北京	理财	500						
6	2025/2/4	上海	存款	250						
7	2025/2/5	上海	贷款	400						
8	2025/2/8	深圳	贷款	700						
9	2025/2/12	上海	存款	600						
10	2025/2/13	深圳	理财	150						
11	2025/2/16	北京	理财	800						

复制DeepSeek生成的公式，得到正确结果

CASE：多张Excel报表处理

CASE：多张Excel报表处理

TO DO：两张Excel报表合并

比如 将员工基本信息表与员工绩效表合并，比如在主表基础上，添加员工2024年第4季度的绩效评分

员工ID	姓名	性别	部门	入职日期
1001	赵静	男	人力资源部	2020-08-29
1002	杨娜	女	人力资源部	2020-08-18
1003	黄芳	女	财务部	2020-06-16
1004	王强	女	财务部	2021-03-03
1005	张美丽	男	市场部	2021-04-21
1006	陈伟	女	市场部	2021-12-19
1007	李静	男	人力资源部	2022-04-02
1008	王芳	男	人力资源部	2020-02-10
1009	陈伟	男	财务部	2020-09-17
.....

员工基本信息表.xlsx

员工ID	年度	季度	绩效评分
1001	2024	1	3.53
1001	2024	2	3.57
1001	2024	3	4.67
1001	2024	4	4.41
1002	2024	1	4.9
1002	2024	2	3.97
1002	2024	3	4.08
1002	2024	4	4.81
1003	2024	1	3.56

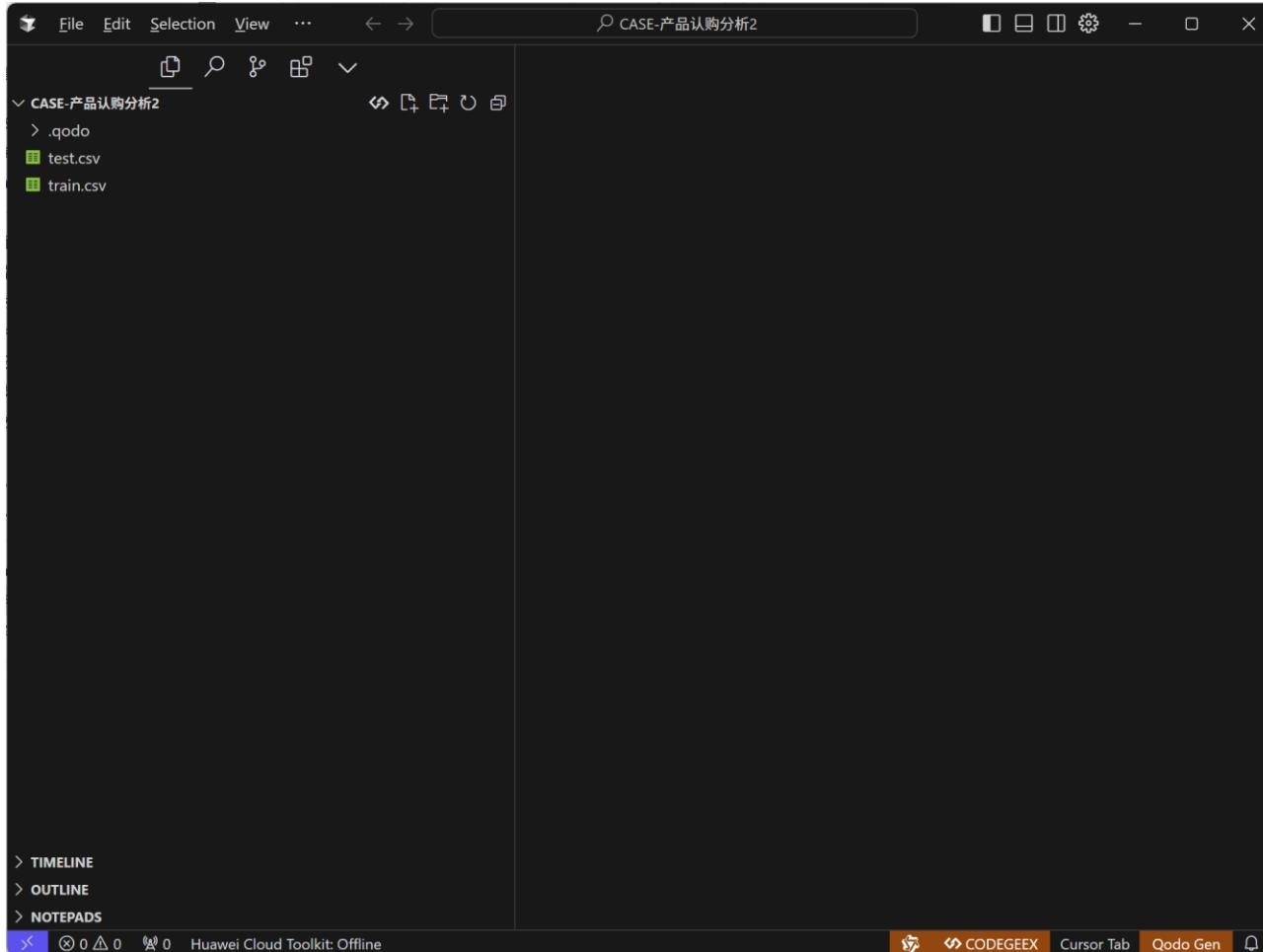
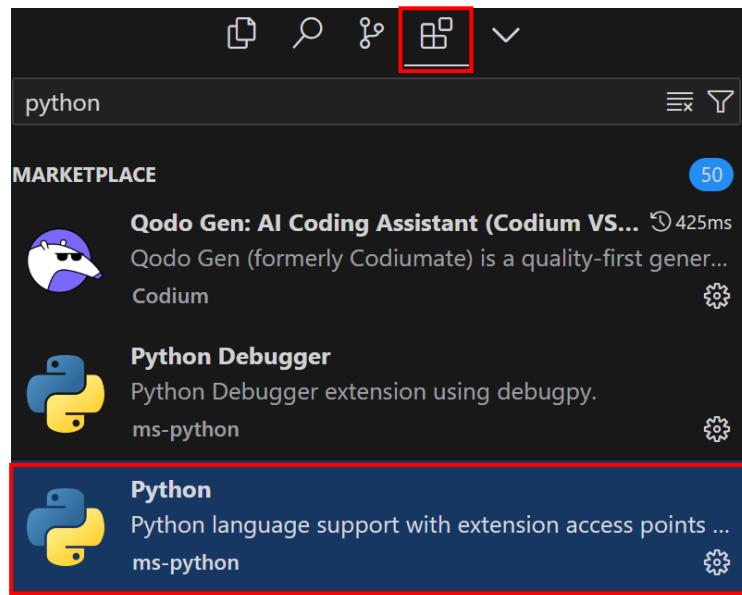
员工绩效表.xlsx

Cursor使用

Step1， 使用Cursor， 打开对应的文件夹

<https://www.cursor.com/cn>

- 1) 下载对应的版本，然后打开本地的项目文件夹
- 2) 配置python编译环境



Cursor使用

Cursor 辅助AI编程提供了三个主要功能：

- 跨文件编辑 (Cmd/Ctrl + I)

AI帮你创建代码、修改代码

- 聊天功能 (Cmd/Ctrl + L)

与AI交流方案，解释逻辑

- 提示框功能(Cmd/Ctrl + K)

向AI提问，让它执行特定功能

Cursor具有文件管理的更高权限，可以看到文件夹下所有的文件，并进行编辑/创建

可以对项目做背景说明，比如放到文件夹，或者 Notepads 中

跨文件编辑功能 (Composer) 可实现多文件协同修改，适合复杂项目开发。

Cursor中的DeepSeek配置

点击 File => Preferences => Cursor Settings 中的 Models

添加模型 Add Models， 创建deepseek-r1和 deepseek-v3

因为这个需要调用第三方的API， 所以打开下方的OpenAI API Key，并进行相应设置：

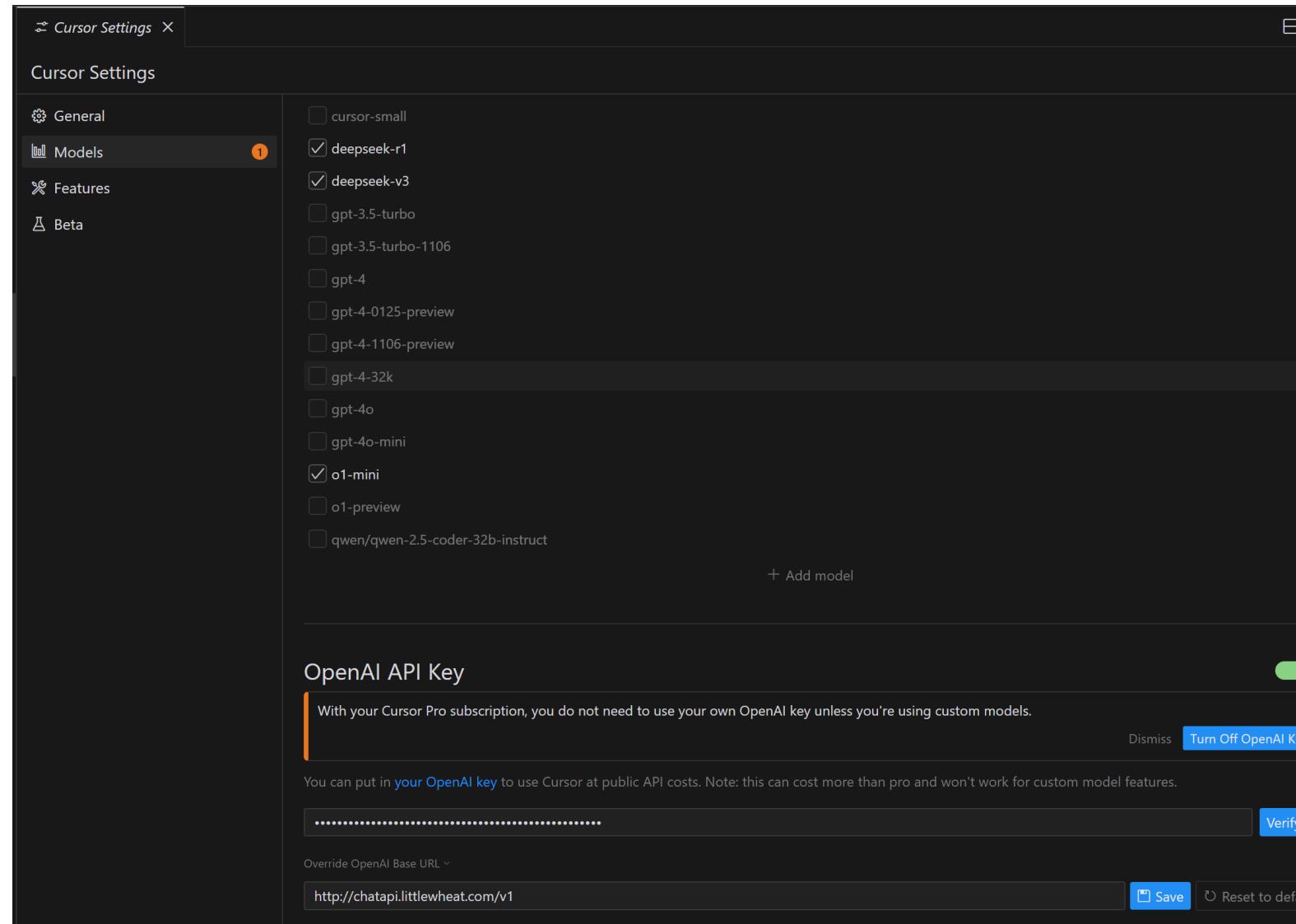
Override OpenAI Base URL:

<http://chatapi.littlewheat.com/v1>

your OpenAI Key:

sk-

Q2gN9CgZOz9jrzjCCHkijalkUyaXpHS6xssmmkl
327kkib0G



CASE：多张Excel报表处理

Step2，读取两张数据表的字段含义

编写Python，打开员工基本信息表.xlsx 和 员工绩效表.xlsx，查看对应的字段

```
# -*- coding: utf-8 -*-
import pandas as pd

# 安装依赖库（如果尚未安装）
# !pip install openpyxl pandas

Qodo Gen: Options | Test this function | ⚙️
def read_excel_columns(file_path):
    """读取Excel文件并返回列名"""
    try:
        df = pd.read_excel(file_path, engine='openpyxl')
        return df.columns.tolist()
    except Exception as e:
        print(f"读取文件{file_path}时出错: {e}")
        return []

# 读取员工基本信息表
basic_info_columns = read_excel_columns("员工基本信息表.xlsx")
print("员工基本信息表字段: ")
print('\n'.join(basic_info_columns))

# 读取员工绩效表
performance_columns = read_excel_columns("员工绩效表.xlsx")
print("\n员工绩效表字段: ")
print('\n'.join(performance_columns))
```

<think>

好的，我现在需要帮用户编写Python代码来打开两个Excel文件，查看它们的字段。首先，我得确定用户的需求是什么。他们可能想合并这两个表，或者进行数据分析，但当前的任务只是查看字段，所以不需要做进一步的处理。

接下来，我需要考虑使用哪个库来处理Excel文件。常用的有pandas，因为它可以方便地读取Excel并展示数据结构。用户可能没有安装所需的库，所以代码中应该包含安装openpyxl和pandas的指令，因为这两个是处理xlsx文件必需的.....

点击Accept File，并运行Python

CASE：多张Excel报表处理

运行后可以看到数据表的字段

员工基本信息表字段：

员工ID

姓名

性别

部门

入职日期

员工绩效表字段：

员工ID

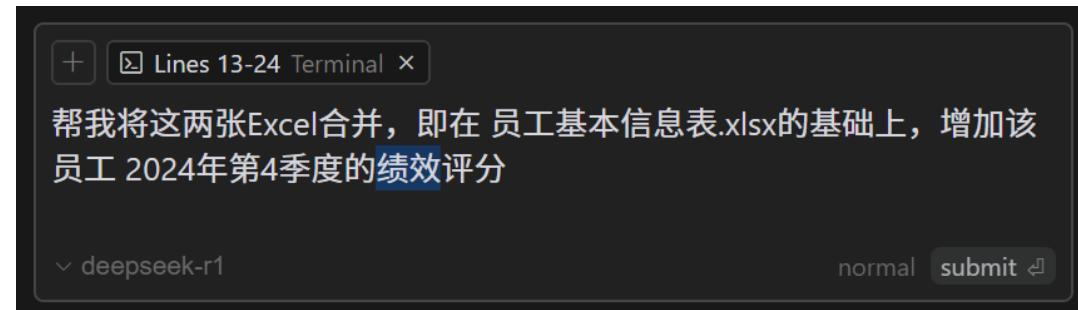
年度

季度

绩效评分

Step3，合并两张数据表

帮我将这两张Excel合并，即在 员工基本信息表.xlsx 的基础上，增加该员工 2024年第4季度的绩效评分



需要将Terminal输出的字段选中，AI既可以看到这些字段含义，然后下指令将数据表进行合并

CASE：多张Excel报表处理

```
merge_employee_data.py > merge_employee_data
1 # -*- coding: utf-8 -*-
2 import pandas as pd
3 from datetime import datetime
4
5 Qodo Gen: Options | Test this function | ✎
6 def merge_employee_data():
7     try:
8         # 读取原始数据
9         basic_df = pd.read_excel("员工基本信息表.xlsx", engine='openpyxl')
10        performance_df = pd.read_excel("员工绩效表.xlsx", engine='openpyxl')
11
12        # 过滤2024年第4季度绩效数据
13        q4_2024 = performance_df[(performance_df['年度'] == 2024) &
14            (performance_df['季度'] == 4)]
15
16        # 合并数据（左连接以保留所有员工）
17        merged_df = pd.merge(
18            basic_df,
19            q4_2024[['员工ID', '绩效评分']], # 只保留需要的字段
20            on='员工ID',
21            how='left'
22        )
23
24        # 处理空值并重命名列
25        merged_df['绩效评分'] = merged_df['绩效评分'].fillna('暂无数据')
26        merged_df.rename(columns={'绩效评分': '2024Q4绩效评分'}, inplace=True)
27
28        # 生成带时间戳的文件名
29        timestamp = datetime.now().strftime("%Y%m%d%H%M")
        output_file = f"员工信息综合表_{timestamp}.xlsx"
```



	A	B	C	D	E	F
1	员工ID	姓名	性别	部门	入职日期	2024Q4绩效评分
2	1001	赵静	男	人力资源部	2020-08-29	4.41
3	1002	杨娜	女	人力资源部	2020-08-18	4.81
4	1003	黄芳	女	财务部	2020-06-16	3.75
5	1004	王强	女	财务部	2021-03-03	3.41
6	1005	张美丽	男	市场部	2021-04-21	3.62
7	1006	陈伟	女	市场部	2021-12-19	4.24
8	1007	李静	男	人力资源部	2022-04-02	4.83
9	1008	王芳	男	人力资源部	2020-02-10	4.09
10	1009	陈伟	男	财务部	2020-09-17	3.5
11	1010	周静	女	销售部	2020-06-01	3.65
12	1011	王建国	男	人力资源部	2020-10-29	4.55
13	1012	刘亮	女	市场部	2020-01-21	4.86
14	1013	陈芳	男	销售部	2021-10-08	4.39
15	1014	王敏	女	市场部	2021-11-10	4.16
16	1015	陈静	女	市场部	2021-02-15	4.25
17	1016	王亮	男	销售部	2021-08-07	3.39
18	1017	吴美丽	女	产品部	2020-01-27	4.68
19	1018	杨建国	女	市场部	2020-11-17	3.56
20	1019	李强	女	人力资源部	2020-10-22	3.08
21	1020	杨强	女	人力资源部	2020-05-18	3.37

DeepSeek-R1 中间生成的代码，用于执行任务

运行后，得到了合并结果

Thinking：不会Python，是否可以做复杂的Excel报表处理？需要自己手写vlookup函数么？

不需要限制DeepSeek使用的工具，业务人员可以从需求和结果上进行把控，如果想要了解DeepSeek处理过程，可以和它进行交流

CASE：产品认购分析

CASE：产品认购分析

CASE：产品认购分析

<https://tianchi.aliyun.com/competition/entrance/531993/introduction>

葡萄牙银行，开展营销活动吸引更多人认购定期存款，通过营销活动沉淀的数据，想要预测未来这些人是否会进行认购，方便进行精准营销

训练集：train.csv 4459条

测试集：test.csv 49342条

字段	说明	
age	年龄	客户基本信息
job	职业: admin, unknown, unemployed, management...	
marital	婚姻: married, divorced, single	
default	信用卡是否有违约: yes or no	
housing	是否有房贷: yes, no, unknown	
Loan	是否有个贷: yes, no, unknown	
Contact	联系方式: unknown, telephone, cellular	本次活动的联系情况
Month	上一次联系的月份: jan, feb, mar, ...	
day_of_week	上一次联系的星期几: mon, tue, wed, thu, fri	
Duration	上一次联系的时长 (秒)	
Campaign	活动期间联系客户的次数	
Pdays	上一次与客户联系后的间隔天数	
Previous	在本次营销活动前，与客户联系的次数	
Poutcome	之前营销活动的结果: unknown, other, failure, success	
emp_var_rate	就业变动率 (季度指标)	市场经济特征
cons_price_index	消费价格指数 (月度指标)	
cons_conf_index	消费者信心指数 (月度指标)	
lending_rate3m	银行同业拆借率 3个月利率 (每日指标)	
nr_employed	雇员人数 (季度指标)	
subscribe	客户是否进行存款: 1 或 0	预测结果

CASE：产品认购分析

数据表：产品认购表

TO DO：

1) 不同年龄层的认购情况

2) 不同职业的认购情况

3) 联系方式/周几的认购情况

4) 不同教育的认购情况

5) 婚姻/房贷的认购情况

6) 个贷/信用卡违约的认购情况

7) 不同沟通时长的认购情况

8) 不同沟通时长/活动中联系次数 的认购情况

1	id	age	job	marital	education	default	housing	loan	contact	month	day_of_week	duration	campaign	pdays	previous	poutcome	emp_var_r
2	1	51	admin.	divorced	profession	no	yes	yes	cellular	aug	mon	4621	1	112	2	failure	1.4
3	2	50	services	married	high.school	unknown	yes	no	cellular	may	mon	4715	1	412	2	nonexister	-1.8
4	3	48	blue-collar	divorced	basic.9y	no	no	no	cellular	apr	wed	171	0	1027	1	failure	-1.8
5	4	26	entrepreneur	single	high.school	yes	yes	yes	cellular	aug	fri	359	26	998	0	nonexister	1.4
6	5	45	admin.	single	university	no	no	no	cellular	nov	tue	3178	1	240	4	success	-3.4
7	6	47	management	married	unknown	no	no	no	cellular	jul	thu	817	1	10	3	nonexister	1.4
8	7	33	technician	married	high.school	no	yes	no	cellular	aug	mon	76	2	1009	0	nonexister	1.4
9	8	40	technician	married	basic.9y	yes	no	unknown	cellular	apr	mon	4391	36	934	4	success	1.4
10	9	47	services	married	high.school	no	no	no	cellular	aug	thu	161	0	959	0	nonexister	1.4
11	10	53	housemaid	married	unknown	no	no	no	cellular	aug	fri	2437	0	594	6	success	-2.9
12	11	37	self-employed	married	university	no	no	no	telephone	jun	mon	2746	1	703	5	success	-1.7
13	12	55	housemaid	married	university	no	no	no	cellular	jul	wed	384	1	980	0	nonexister	1.4
14	13	36	technician	single	profession	no	yes	yes	cellular	aug	thu	135	0	1003	0	nonexister	1.4
15	14	30	blue-collar	single	high.school	no	yes	no	cellular	may	thu	222	1	957	0	nonexister	-1.8
16	15	52	housemaid	divorced	basic.4y	no	yes	no	cellular	jul	mon	50	4	1007	0	nonexister	1.4
17	16	26	self-employed	single	university	no	unknown	unknown	cellular	aug	mon	126	2	1025	1	failure	-2.9
18	17	36	blue-collar	married	basic.9y	unknown	yes	no	cellular	jul	wed	590	2	486	5	success	1.4
19	18	36	blue-collar	married	basic.9y	no	no	no	telephone	jun	fri	15	3	969	0	nonexister	1.4
20	19	50	blue-collar	divorced	basic.9y	no	yes	no	cellular	jul	wed	4526	0	124	4	nonexister	1.4
21	20	27	admin.	single	university	no	yes	no	cellular	may	wed	45	2	744	0	nonexister	-1.8
22	21	50	unemployed	single	basic.4y	no	yes	no	cellular	may	fri	167	1	982	0	nonexister	-1.8
23	22	48	services	married	high.school	unknown	yes	no	telephone	may	thu	232	2	966	0	nonexister	1.1
24	23	72	retired	married	basic.9y	no	yes	no	cellular	oct	thu	4314	1	150	3	nonexister	-3.4
25	24	67	technician	married	basic.4y	unknown	no	yes	telephone	dec	mon	164	20	957	0	nonexister	-3.4
26	25	47	retired	married	basic.4y	unknown	no	no	telephone	jun	wed	144	8	1024	0	nonexister	1.4
27	26	46	management	married	basic.4y	unknown	no	yes	telephone	may	fri	261	3	1040	0	nonexister	1.1
28	27	34	admin.	married	university	no	yes	no	telephone	may	thu	391	4	1012	0	nonexister	1.1
29	28	38	technician	divorced	university	no	no	no	cellular	jul	mon	2187	2	285	4	failure	1.4
30	29	50	self-employed	divorced	university	no	yes	no	cellular	nov	thu	748	1	995	1	failure	-0.1

CASE：产品认购分析

Step1，打开train.csv，理解字段含义

编写Python，帮我打开 train.csv，分析数据表的字段含义

这里使用DeepSeek-V3即可

编写Python，帮我打开 train.csv，分析数据表的字段含义

```
data_analysis.py 1/1 ✓
import pandas as pd

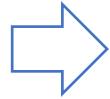
# 读取CSV文件
df = pd.read_csv('train.csv')

# 获取数据表的基本信息
print("数据表的基本信息:")
print(df.info())

# 获取数据表的前5行数据
print("\n数据表的前5行数据:")
print(df.head())
..
```

代码说明：

1. 读取CSV文件：使用 pandas 库的 `read_csv` 函数读取 `train.csv` 文件。
2. 基本信息：使用 `info()` 方法获取数据表的基本信息，包括字段名称、数据类型、非空值数量等。



数据表的基本信息：

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 22500 entries, 0 to 22499
Data columns (total 22 columns):
 #   Column      Non-Null Count Dtype  
 --- 
 0   id          22500 non-null int64  
 1   age         22500 non-null int64  
 2   job          22500 non-null object 
 3   marital     22500 non-null object 
 4   education   22500 non-null object 
 5   default     22500 non-null object 
 6   housing     22500 non-null object 
 7   loan         22500 non-null object 
 8   contact     22500 non-null object 
 9   month        22500 non-null object 
 10  day_of_week 22500 non-null object 
 11  duration    22500 non-null int64  
 12  campaign   22500 non-null int64  
 13  pdays       22500 non-null int64  
 14  previous    22500 non-null int64  
 15  poutcome   22500 non-null object 
 ..
```

CASE：产品认购分析

Step2，分析不同年龄层的认购比例

帮我计算不同年龄层的认购比例

同样使用DeepSeek-V3即可



```
+ Lines 3-112 Terminal x
帮我计算不同年龄层的认购比例

▼ deepseek-v3           normal submit ↵
```



不同年龄层的认购比例：

subscribe	no	yes
age_group		
0-20	61.363636	38.636364
21-30	84.383500	15.616500
31-40	89.503749	10.496251
41-50	91.627292	8.372708
51-60	87.686449	12.313551
61-70	58.606557	41.393443
71-80	51.333333	48.666667
81-90	52.356021	47.643979
91-100	51.724138	48.275862

Thinking：如果要用可视化图表呈现，该如何操作？

比如用柱状图进行呈现

CASE：产品认购分析

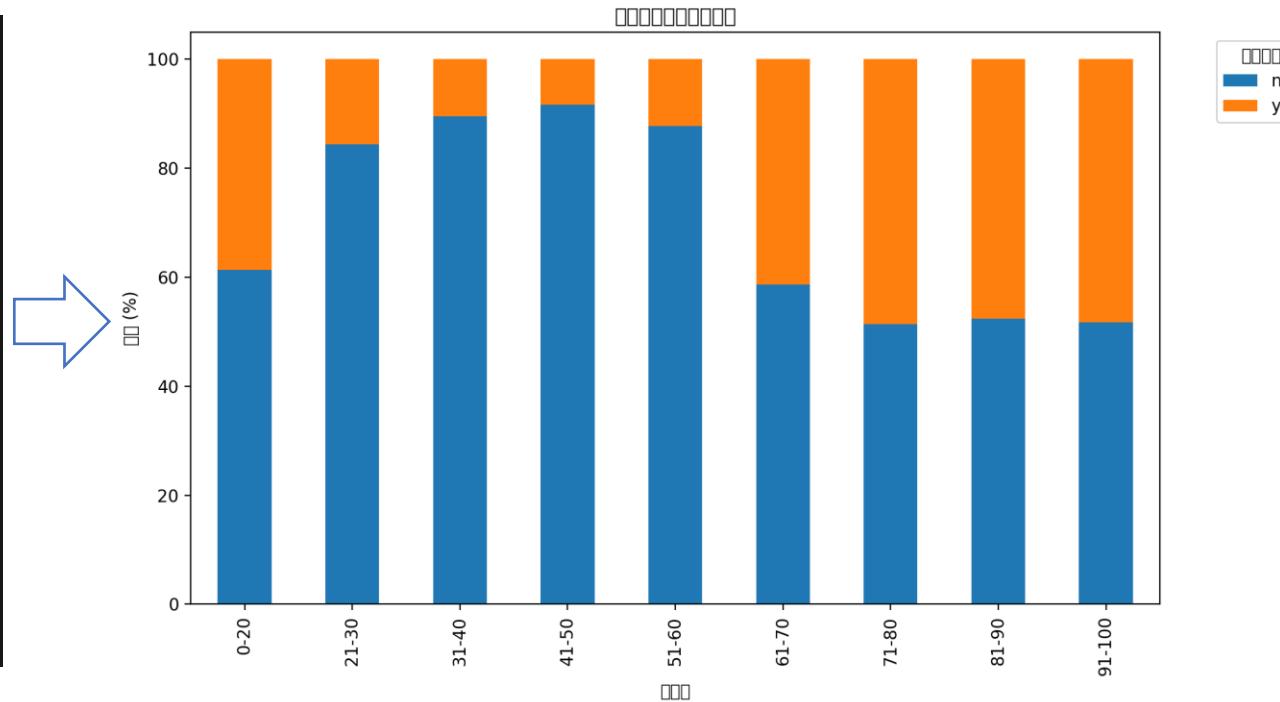
请用柱状图呈现

```
# 打印结果
print("\n不同年龄层的认购比例:")
print(subscription_rate)

# 绘制柱状图
subscription_rate.plot(kind='bar', stacked=True, figsize=(10, 6))

# 设置图表标题和标签
plt.title('不同年龄层的认购比例')
plt.xlabel('年龄层')
plt.ylabel('比例 (%)')

# 显示图表
plt.legend(title='认购情况', bbox_to_anchor=(1.05, 1), loc='upper left')
plt.tight_layout()
plt.show()
```



Accept file后，运行得到柱状图

Thinking: 生成的图表中，中文有乱码，如何解决？

CASE：产品认购分析

可视化中的中文有乱码

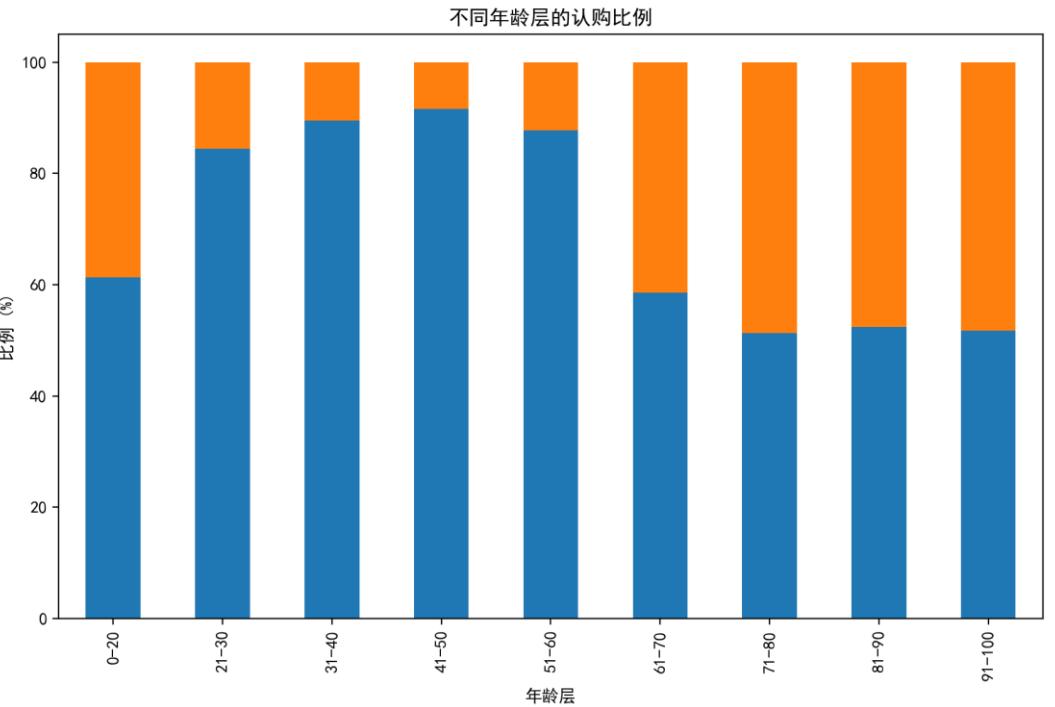
修改说明：

1. 设置字体：使用 `plt.rcParams['font.sans-serif']` 将字体设置为 SimHei（黑体），这是Windows系统中常用的中文字体。
2. 解决负号问题：设置 `plt.rcParams['axes.unicode_minus'] = False`，确保负号正常显示。

注意事项：

- 如果运行环境是Linux或macOS，可能需要安装中文字体，或者使用其他支持中文的字体，如 WenQuanYi Zen Hei。
- 如果仍然出现乱码，可以尝试其他中文字体，如 KaiTi（楷体）或 Microsoft YaHei（微软雅黑）。

DeepSeek得到反馈后，会修改代码，然后运行得到正确结果



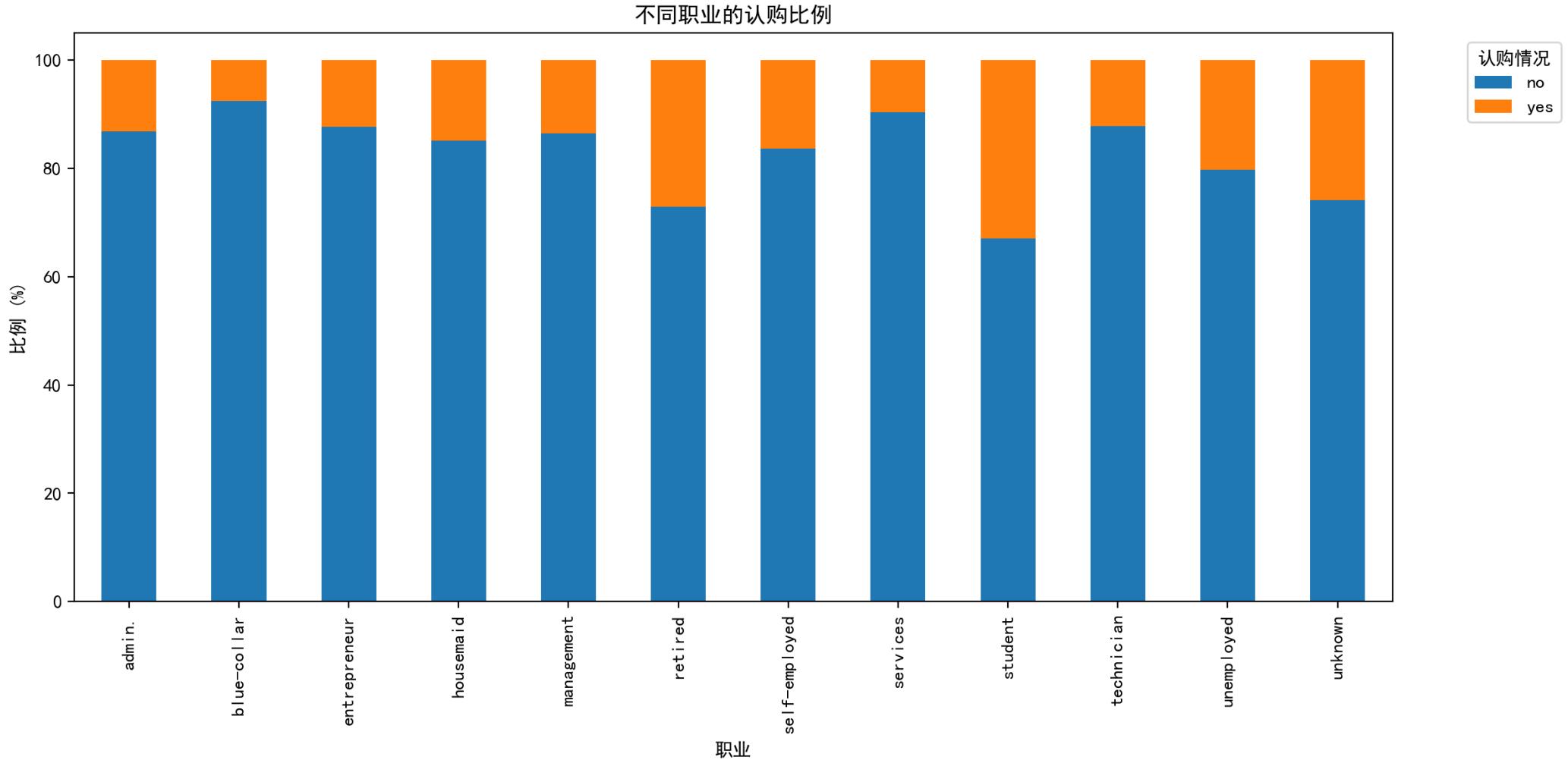
图表中的中文正常显示了，如果要继续生成多张图表，该如何操作？

CASE：产品认购分析

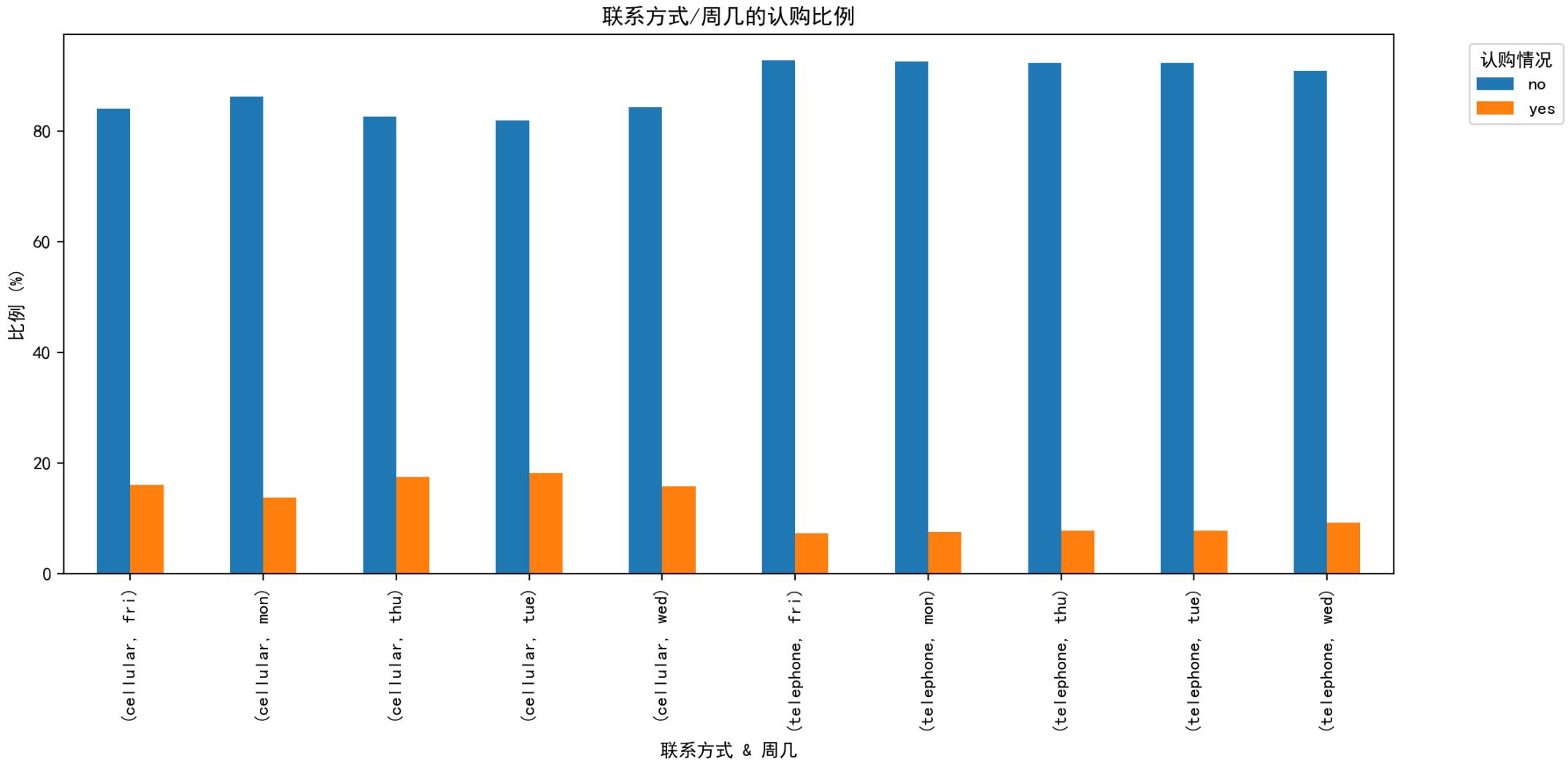
帮我计算，并得到以下的可视化图表，请选择适合的图表类型进行呈现

- 1) 不同职业的认购情况
- 2) 联系方式/周几的认购情况
- 3) 不同教育的认购情况
- 4) 婚姻/房贷的认购情况
- 5) 个贷/信用卡违约的认购情况
- 6) 不同沟通时长的认购情况
- 7) 不同沟通时长/活动中联系次数 的认购情况

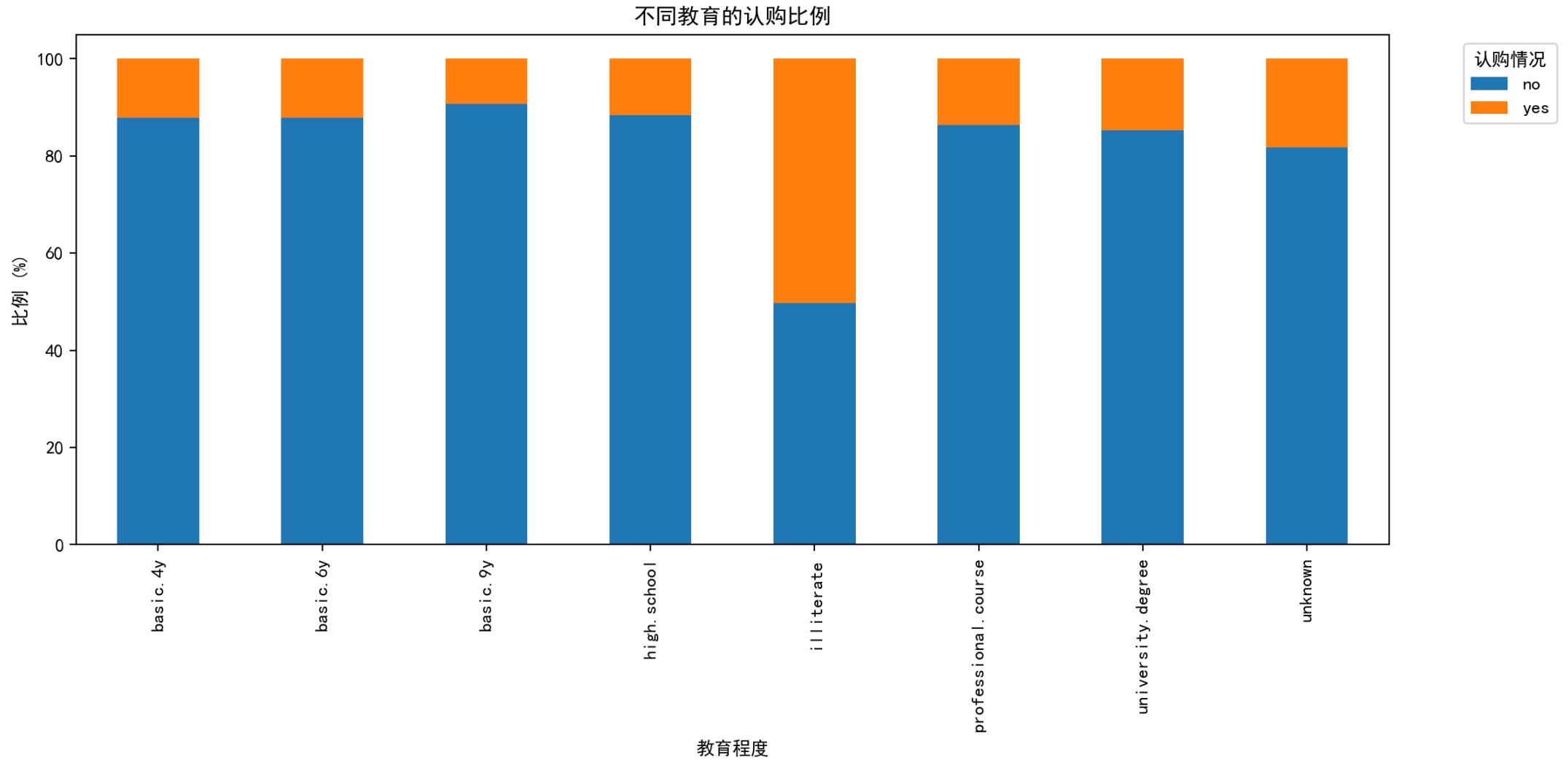
可视化：不同职业的认购比例



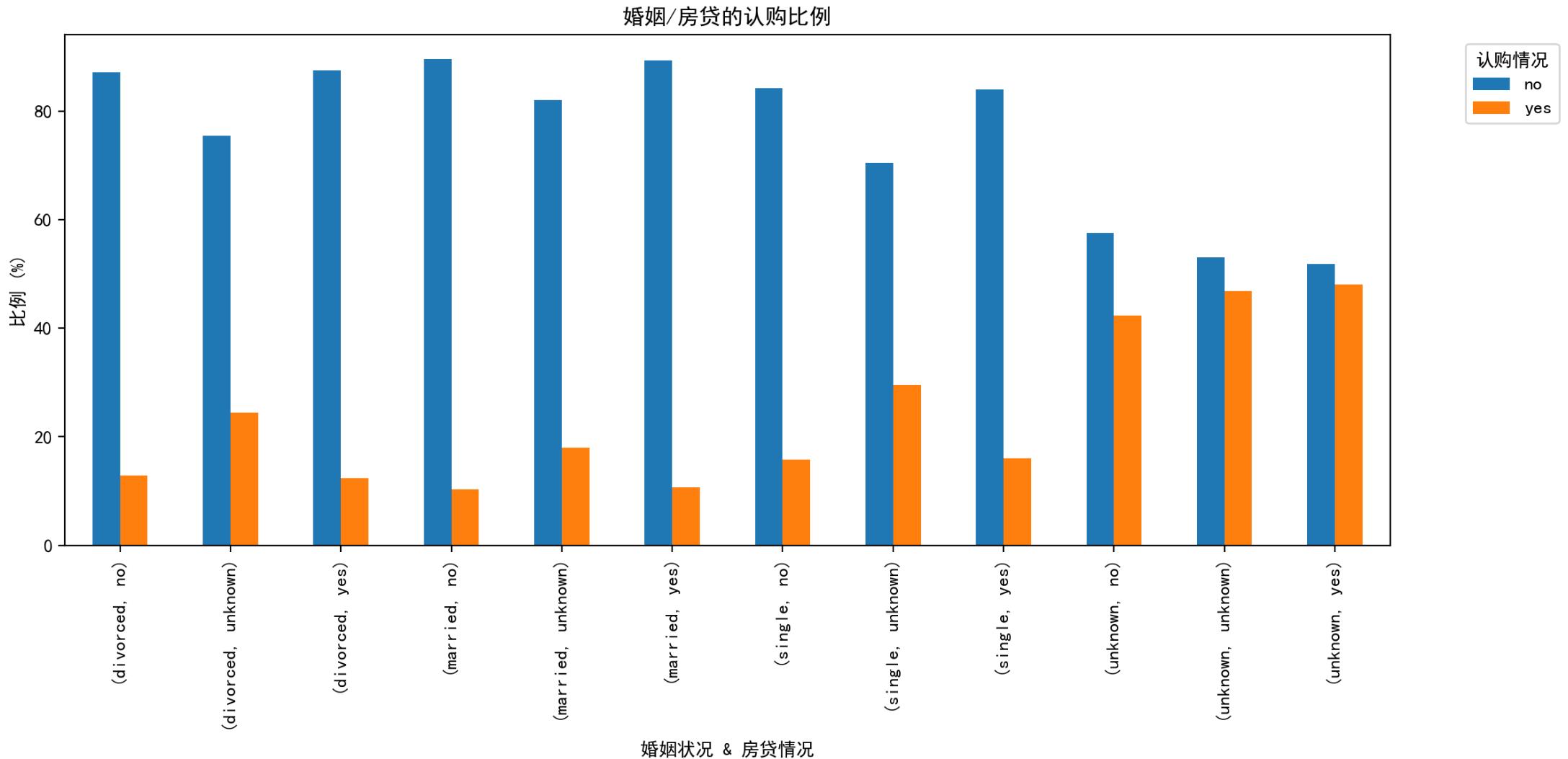
可视化：联系方式/周几的认购比例



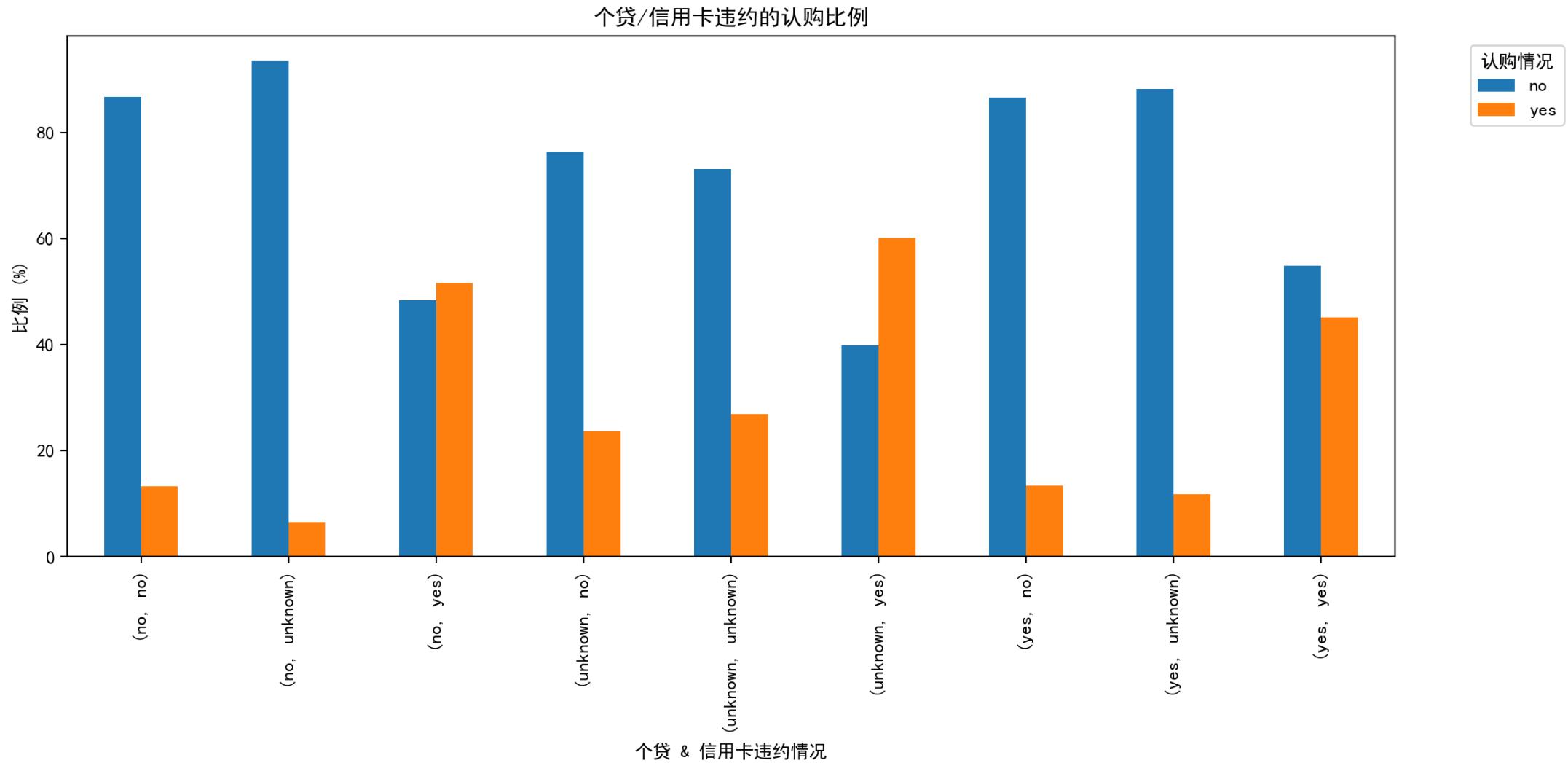
可视化：不同教育的认购比例



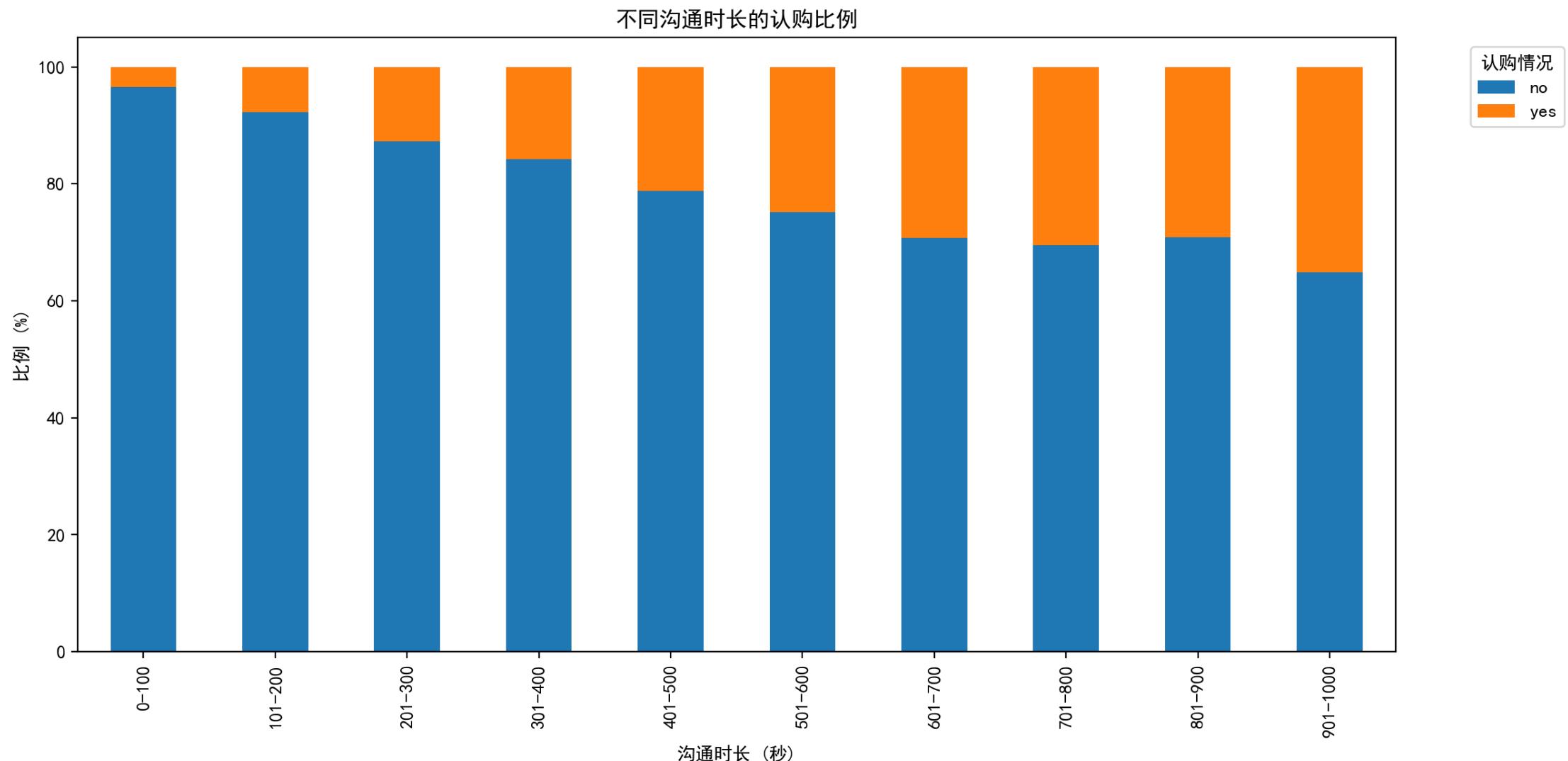
可视化：婚姻/房贷的认购比例



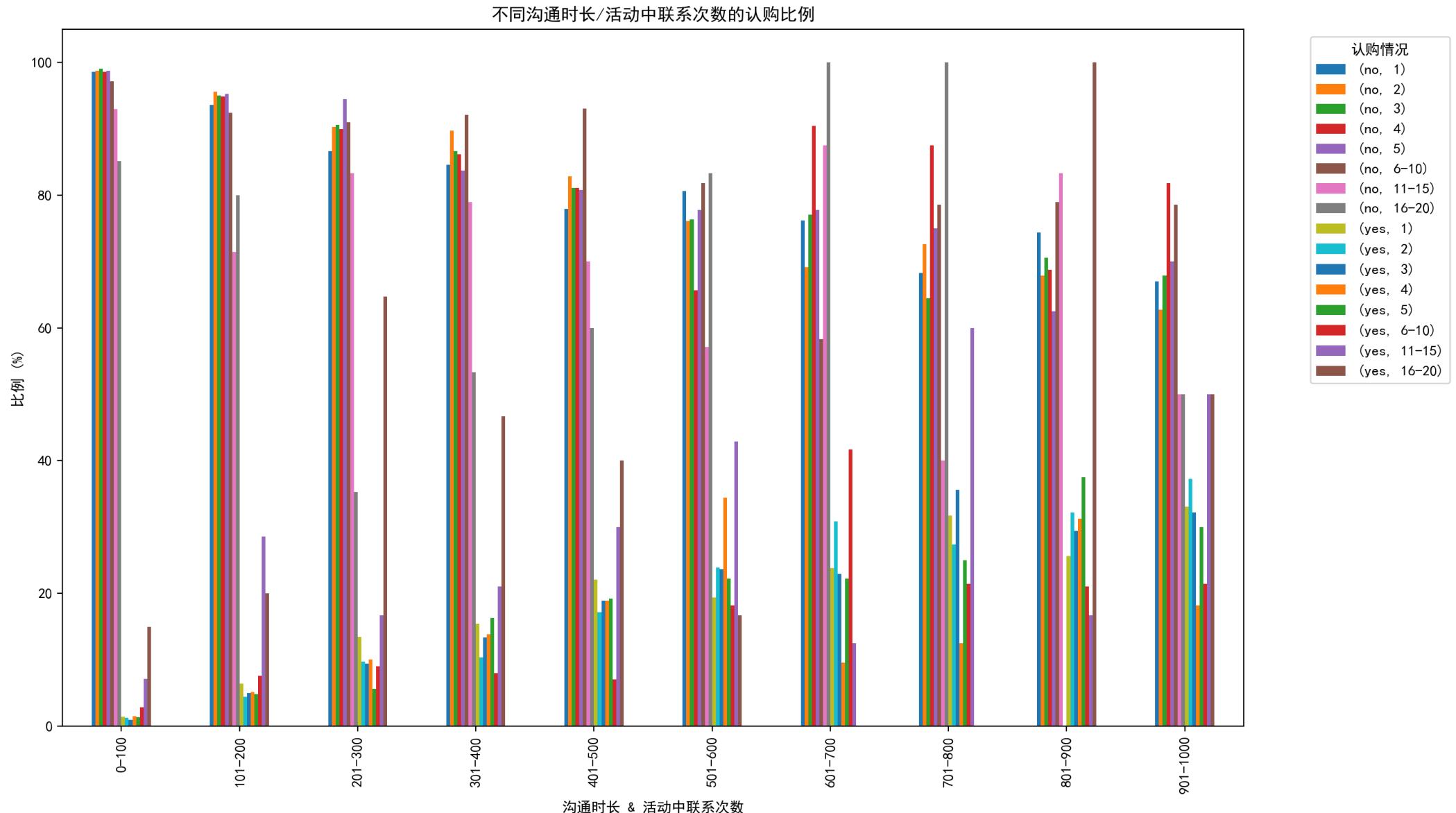
可视化：个贷/信用卡违约的认购比例



可视化：不同沟通时长的认购比例



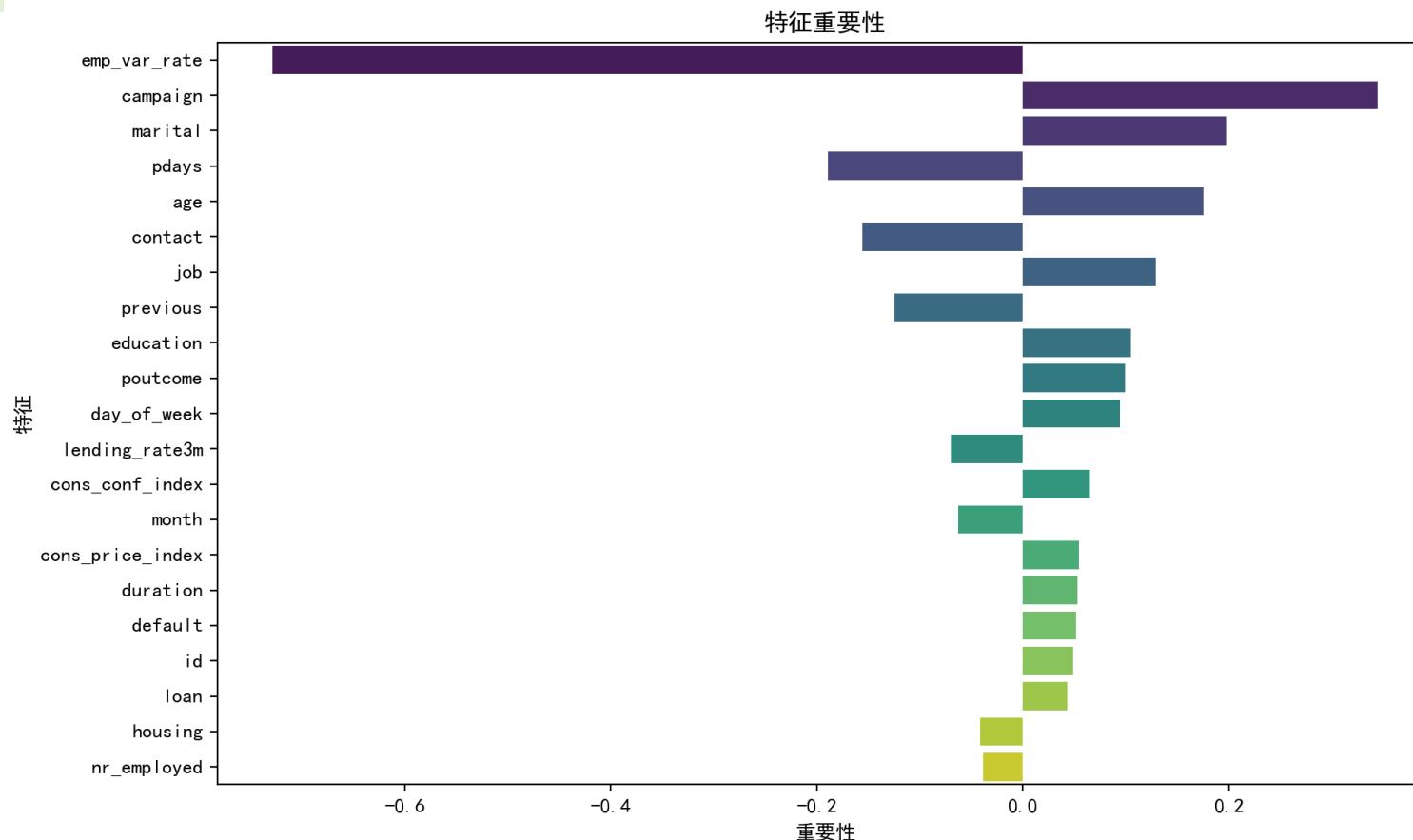
可视化：不同沟通时长/活动中联系次数的认购比例



CASE：产品认购分析

Step3，如何对数据特征进行洞察，比如哪些人更容易认购

帮我使用简单的分类算法，创建产品认购预测模型，并告诉我哪些特征是影响用户认购的关键特征，同时将这些特征按照重要性从大到小进行可视化呈现



Thinking: 如何理解这些特征重要性?

CASE：产品认购分析

如何理解这些特征重要性？哪些特征更容易让用户订阅

The screenshot shows a DeepSeek interface with a terminal window titled 'data_analysis.py' displaying a list of features and their importance values. The list is sorted from most important to least important. The features include 'emp_var_rate', 'campaign', 'marital', 'pdays', 'age', 'contact', 'job', 'previous', 'education', 'poutcome', 'day_of_week', 'lending_rate3m', 'cons_conf_index', 'month', 'cons_price_index', 'duration', 'default', 'id', 'loan', 'housing', and 'nr_employed'. The importance values range from -0.728452 to 0.344987.

```
特征重要性 (从大到小):
      feature  importance
16    emp_var_rate  -0.728452
12      campaign   0.344987
3       marital    0.197563
13      pdays     -0.189419
1       age       0.175785
8       contact   -0.155860
2       job       0.129465
14      previous  -0.124477
4       education  0.104794
15      poutcome   0.099401
10      day_of_week 0.094197
19      lending_rate3m -0.069580
18      cons_conf_index  0.065449
9        month    -0.062696
17      cons_price_index  0.054564
11      duration   0.052916
5       default    0.051749
0        id       0.048933
7       loan      0.043571
6       housing   -0.041597
20      nr_employed -0.038681
```

如何理解这些特征重要性？哪些特征更容易让用户订阅

deepseek-v3 submit codebase ctrl+e

当我们不需要处理数据，而是想要和DeepSeek交互，得到他对现有结果的洞察理解的时候，可以使用DeepSeek-V3

CASE：产品认购分析

从特征重要性结果来看，我们可以得出以下分析：

最重要的负面特征：

1. emp_var_rate (就业变化率)：重要性最高 (-0.728)，且为负值，说明就业变化率越高，用户订阅的可能性越低

2. 最重要的正面特征：

campaign (营销活动次数)：重要性第二高 (0.345)，且为正值，说明营销活动次数越多，用户订阅的可能性越高

marital (婚姻状况)：重要性第三高 (0.198)，且为正值，说明已婚用户更可能订阅

其他重要特征：

age (年龄)：正值 (0.176)，年龄越大越可能订阅

contact (联系方式)：负值 (-0.156)，某些联系方式（如电话）可能比邮件更有效

job (职业)：正值 (0.129)，某些职业的用户更可能订阅

影响较小的特征：

id、loan、housing等特征的重要性较低，对订阅决策影响较小

总结：

更容易让用户订阅的特征：营销活动次数多、已婚、年龄较大、某些特定职业阻碍用户订阅的特征：就业变化率高、某些联系方式

建议在营销策略中：

1. 增加营销活动频率

2. 重点关注已婚和年龄较大的用户群体

3. 优化联系方式，选择更有效的沟通渠道

4. 关注就业市场变化，在经济稳定时期加大营销力度



Thank You
Using data to solve problems