“一起挖掘幸福感”最终报告

1. 数据获取及预处理

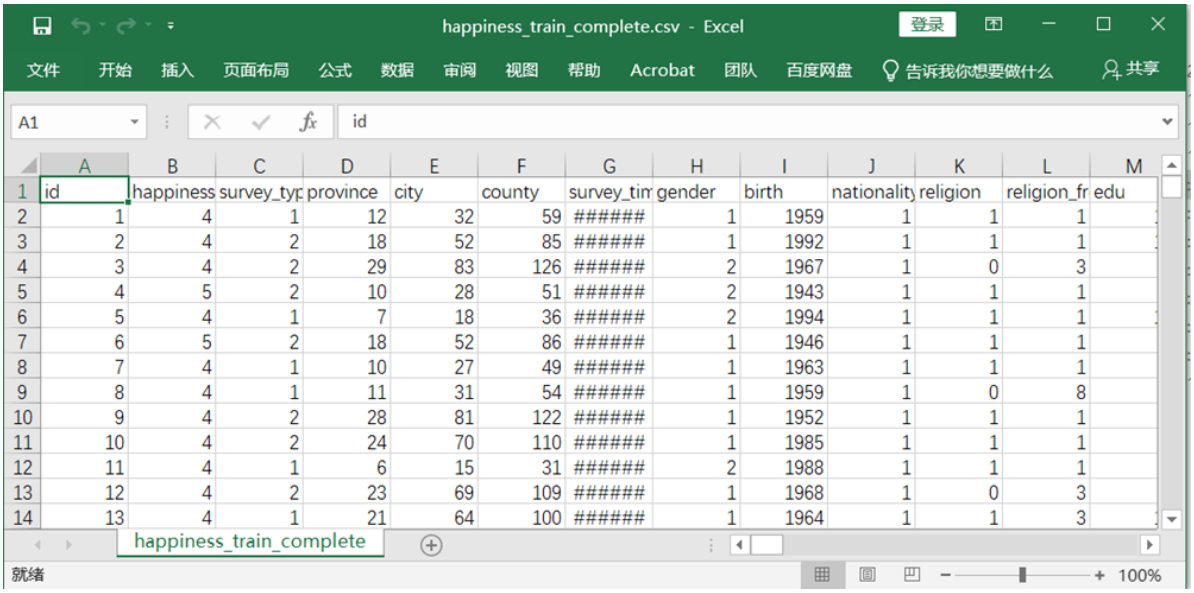
1.1 数据来源

赛题使用的数据来自中国人民大学中国调查与数据中心主持之《中国综合社会调查（CGSS）》项目。数据分为完整版和精简版两类。可从精简版入手熟悉赛题后，使用完整版挖掘更多信息。其中包含以下文件：

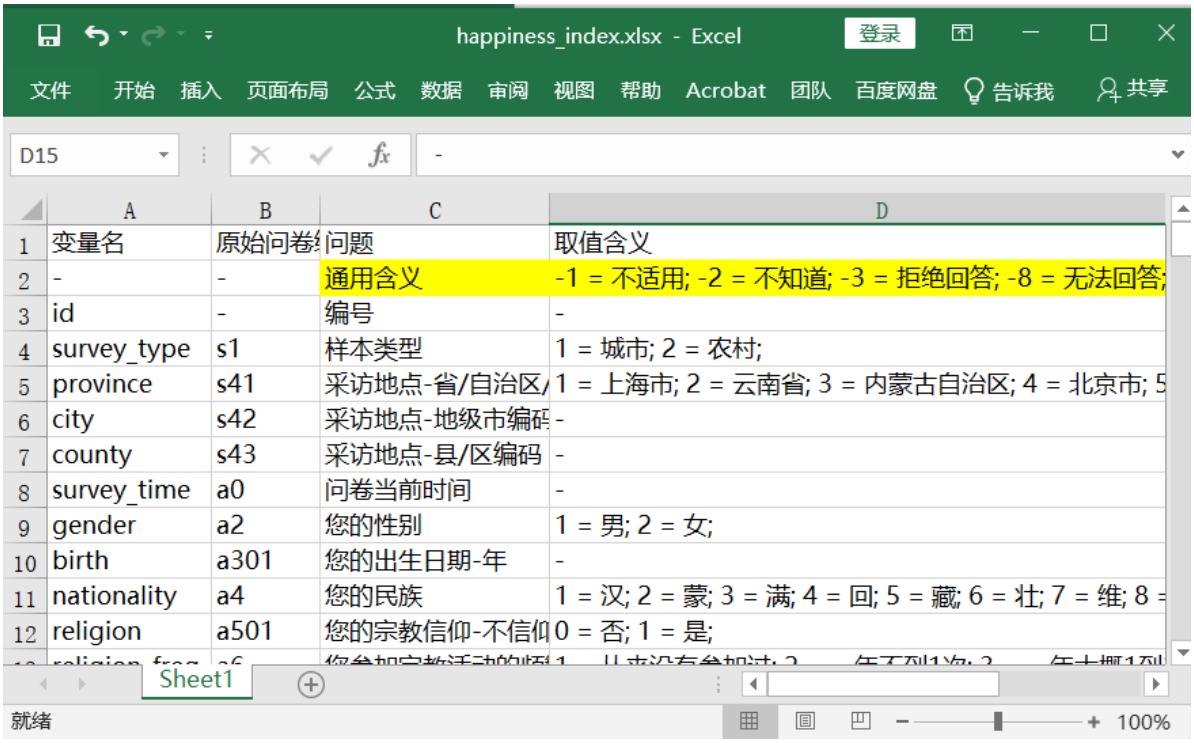
- complete文件为变量完整版数据
- abbr文件为变量精简版数据
- index文件中包含每个变量对应的问卷题目，以及变量取值的含义
- survey文件是数据源的原版问卷，作为补充以方便理解问题背景。

1.2 数据说明

由于本数据中字段较多，专门有对应的index文件进行描述，因此不在次赘述相应字段含义，仅以下图进行概览。



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	id	happiness	survey_typ	province	city	county	survey_tim	gender	birth	nationality	religion	religion_fr	edu
2	1	4	1	12	32	59	######	1	1959	1	1	1	1
3	2	4	2	18	52	85	######	1	1992	1	1	1	1
4	3	4	2	29	83	126	######	2	1967	1	0	3	
5	4	5	2	10	28	51	######	2	1943	1	1	1	
6	5	4	1	7	18	36	######	2	1994	1	1	1	1
7	6	5	2	18	52	86	######	1	1946	1	1	1	
8	7	4	1	10	27	49	######	1	1963	1	1	1	
9	8	4	1	11	31	54	######	1	1959	1	0	8	
10	9	4	2	28	81	122	######	1	1952	1	1	1	
11	10	4	2	24	70	110	######	1	1985	1	1	1	
12	11	4	1	6	15	31	######	2	1988	1	1	1	
13	12	4	2	23	69	109	######	1	1968	1	0	3	
14	13	4	1	21	64	100	######	1	1964	1	1	3	



变量名	原始问卷	问题	取值含义
-	-	通用含义	-1 = 不适用; -2 = 不知道; -3 = 拒绝回答; -8 = 无法回答;
id	-	编号	-
survey_type	s1	样本类型	1 = 城市; 2 = 农村;
province	s41	采访地点-省/自治区/	1 = 上海市; 2 = 云南省; 3 = 内蒙古自治区; 4 = 北京市; 5
city	s42	采访地点-地级市编码	-
county	s43	采访地点-县/区编码	-
survey_time	a0	问卷当前时间	-
gender	a2	您的性别	1 = 男; 2 = 女;
birth	a301	您的出生日期-年	-
nationality	a4	您的民族	1 = 汉; 2 = 蒙; 3 = 满; 4 = 回; 5 = 藏; 6 = 壮; 7 = 维; 8 =
religion	a501	您的宗教信仰-不信仰	0 = 否; 1 = 是;

完整训练集有8000条数据，140个属性（包含label），属性类型有数值属性、序数属性和标称属性，各属性及其取值含义包含在index文件中。

对于评测结果的判定，赛题要求提交的结果为csv文件，其中包含id和happiness的预测值两列。

比赛以均方误差为成绩，计算公式：

$$score = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i^*)^2$$

其中，n代表测试样本数量，y_i代表第i个测试样本的预测值，而y_i^*表示测试样本的真实值。

1.3 数据预处理

首先是对缺失值的处理，合并train和test，查看属性的缺失率，并一起处理：

	total	percent
edu_other	10950	0.999452
invest_other	10911	0.995893
property_other	10867	0.991877
join_party	9831	0.897317
s_work_type	7437	0.678806
s_work_status	7437	0.678806
work_status	6932	0.632713
work_yr	6932	0.632713
work_manage	6931	0.632621
work_type	6931	0.632621
edu_yr	2754	0.251369
marital_now	2445	0.223165
s_political	2365	0.215863
s_hukou	2365	0.215863
s_income	2365	0.215863
s_birth	2365	0.215863

如图上图所示，对于缺失率大于50%的属性列，考虑直接舍去。

对于剩下的缺失属性列，结合具体的语义分析以进行合适的填充：

例如：

- 婚姻情况：配偶属性这里有大量的缺失值，根据分析发现是由于df_all['marital']==1即未婚造成的，所以对于这些缺失值我们填充为0，当作是一个新类别，同时根据实际情况，把不重要的特征（比如政治面貌等）进行消减，去除噪点。
- 教育情况：缺失值是由于没有受过教育造成的，全部缺失值填充为0。
- 社交情况：查看social_friend和social_neighbor的所在数据行缺失是由于社交不频繁socilize=1（近一年空闲时从不社交）造成的所以全部缺失值填充为7（从不）。

接下来考虑对异常值的处理。以happiness为例，训练集中有12条数据取值为-8（无法回答）可以考虑直接删去这12条数据，但也可以将其设置为3（说不上幸福还是不幸福）。其余异常值可以考虑使用中位数或者平均数代替。

我们可以对数值型数据进行分箱，避免异常值和噪点的干扰，如将出生日期划分为1920-1930,1930-1940,1940-1950,1950-1960等类别。

对于特征中具有连续值的量进行归一化处理（如收入），使用零-均值规范化（标准差标准化）方法。归一化后，数据的均值为0，标准差为1。

$$x^* = \frac{x - \bar{x}}{\sigma}$$

2. 数据分析与可视化

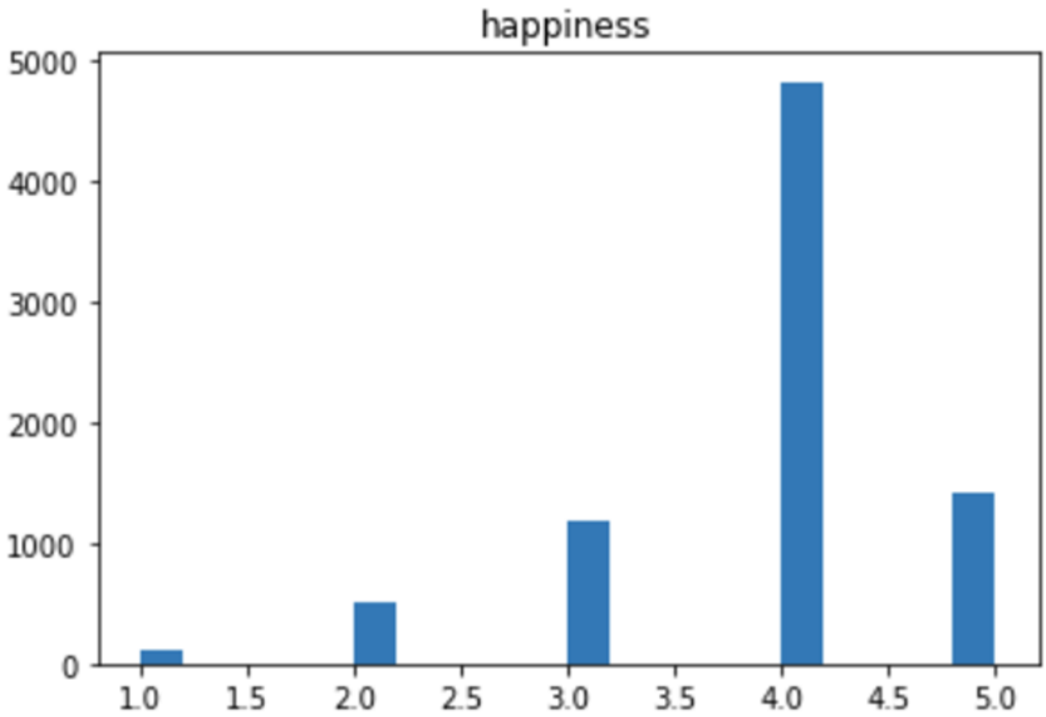
分析属性特征，依托Python语言的Pandas：存储数据的dataframe数据结构；Matplotlib：基于matlab的python绘图包；Numpy：python的科学计算库等库方法，采用饼状图、盒图、直方图、条状图等表现形式对预处理后的数据集进行探索性分析。

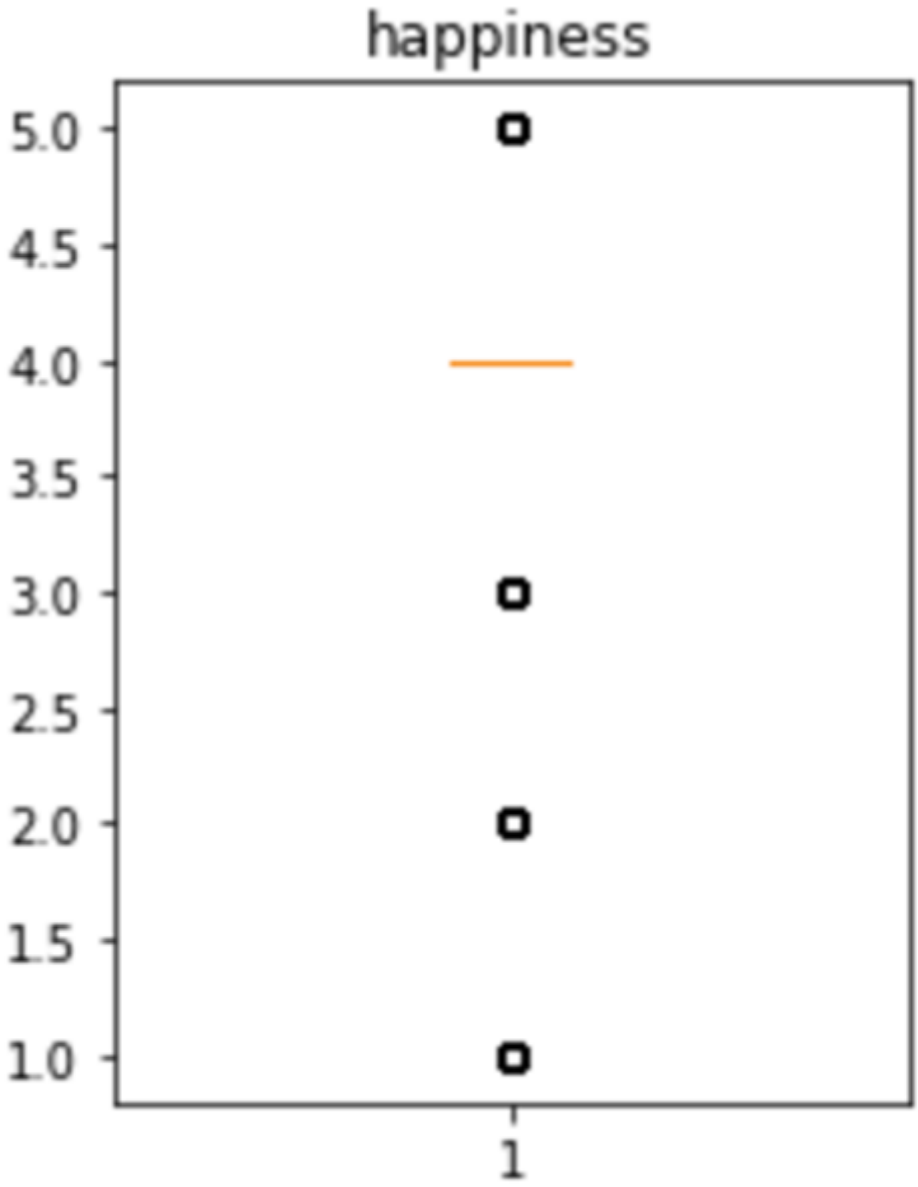
2.1 统计量分析

使用describe方法对数据集进行描述，查看数据的统计量，均值、标准差、最小值、分位数、最大值，采用直方图、盒图进行可视化，结果显示happiness均值为3.87，最大值为5，最小值为1，众数为4，大部分人的幸福感较高。

```
train.describe()
```

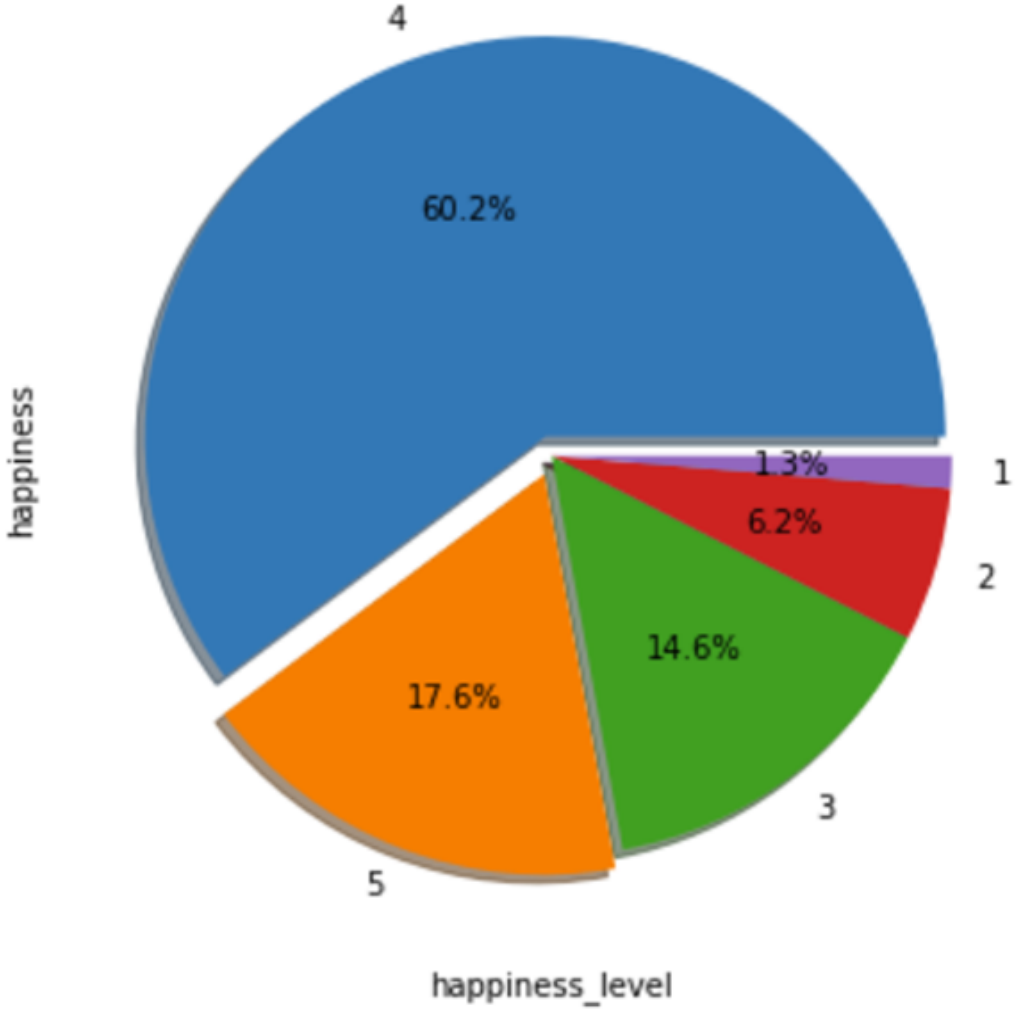
	id	happiness	survey_type	province	city	county	gender	birth	nationality	religion	...
count	8000.00000	8000.000000	8000.000000	8000.000000	8000.000000	8000.000000	8000.00000	8000.000000	8000.00000	8000.000000	...
mean	4000.50000	3.866625	1.405500	15.155375	42.564750	70.619000	1.53000	1964.707625	1.37350	0.772250	...
std	2309.54541	0.818792	0.491019	8.917100	27.187404	38.747503	0.49913	16.842865	1.52882	1.071459	...
min	1.00000	1.000000	1.000000	1.000000	1.000000	1.000000	1.00000	1921.000000	-8.00000	-8.000000	...
25%	2000.75000	4.000000	1.000000	7.000000	18.000000	37.000000	1.00000	1952.000000	1.00000	1.000000	...
50%	4000.50000	4.000000	1.000000	15.000000	42.000000	73.000000	2.00000	1965.000000	1.00000	1.000000	...
75%	6000.25000	4.000000	2.000000	22.000000	65.000000	104.000000	2.00000	1977.000000	1.00000	1.000000	...
max	8000.00000	5.000000	2.000000	31.000000	89.000000	134.000000	2.00000	1997.000000	8.00000	1.000000	...

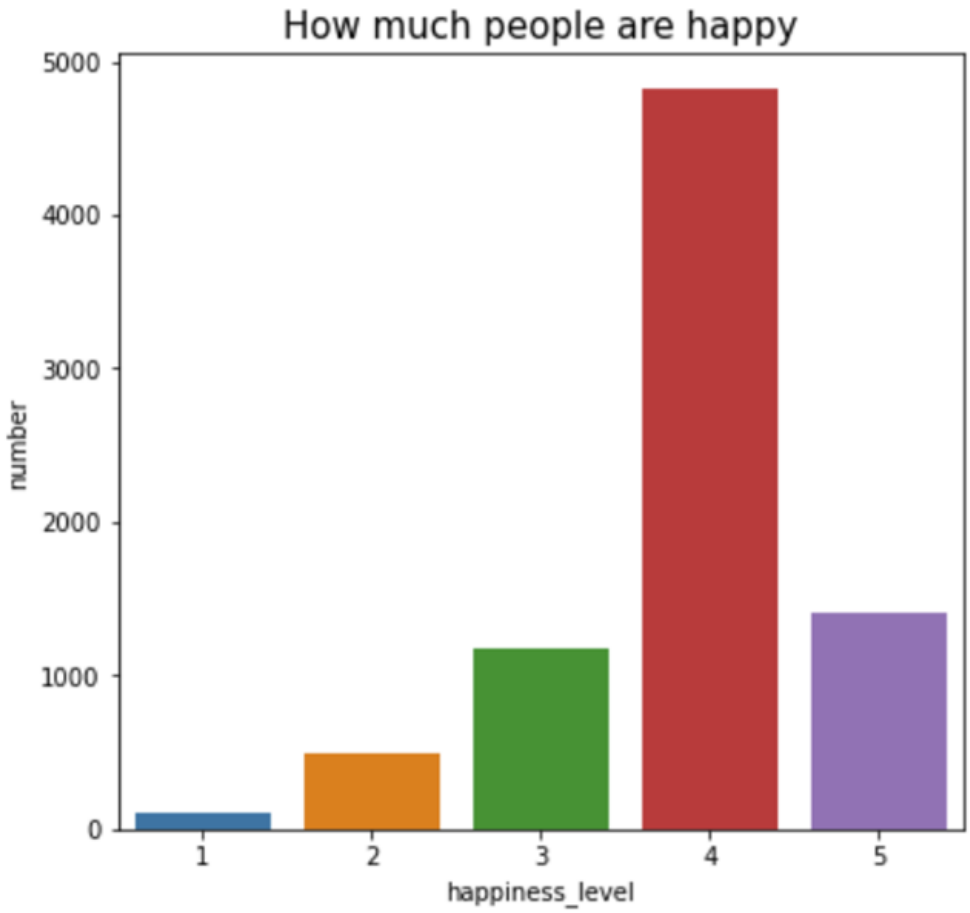




2.2 整体幸福等级占比

在预处理阶段，我们对happiness的数值进行了处理使其更符合实际意义：happiness：1 = 非常不幸福; 2 = 比较不幸福; 3 = 说不上幸福不幸福; 4 = 比较幸福; 5 = 非常幸福;在此基础上画出这5个等级幸福感的柱状图和占比图

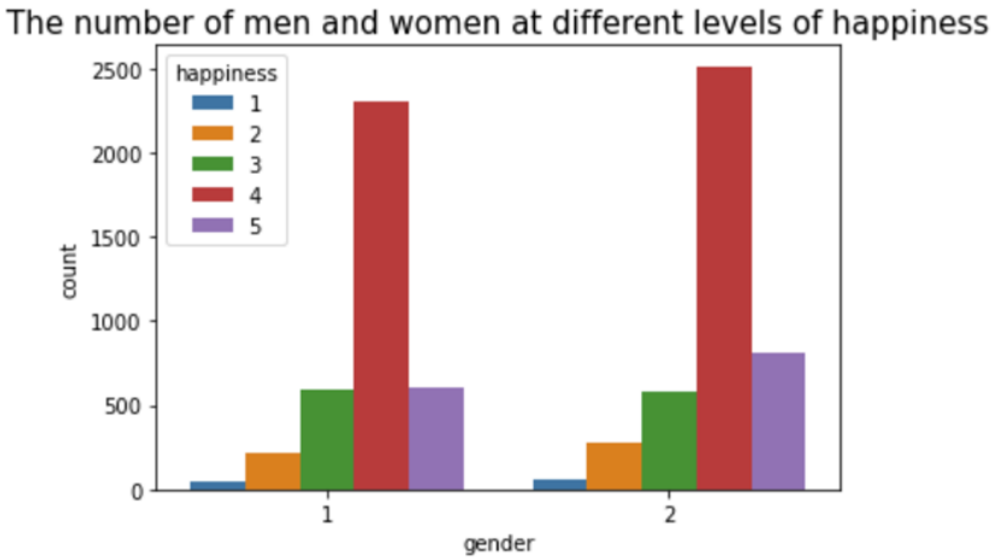

4
60.2%
1.3%
1
6.2%
2
14.6%
17.6%
3
5
happiness
happiness_level

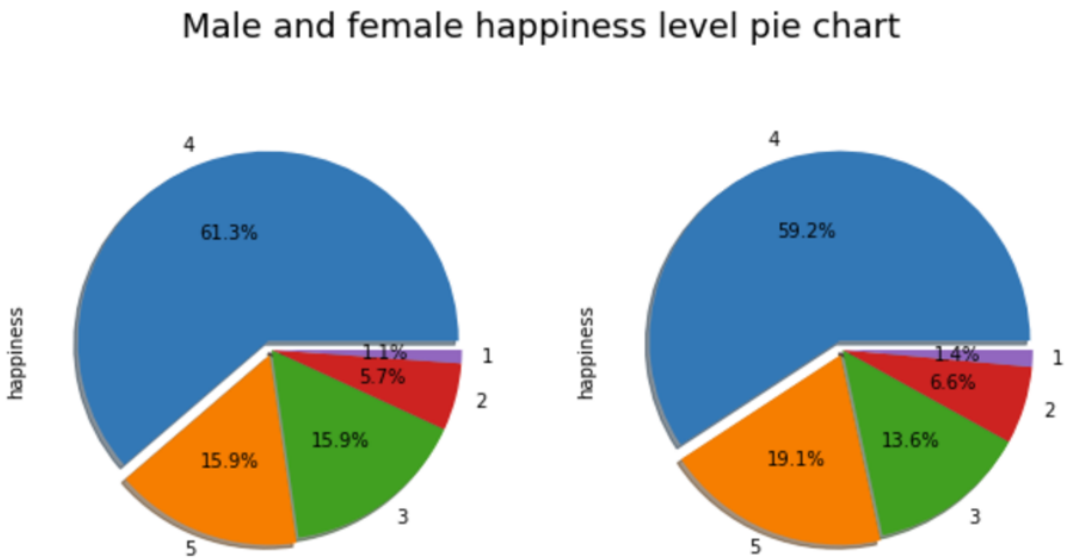

How much people are happy
5000
4000
3000
2000
1000
0
number
1
2
3
4
5
happiness_level

图中显示，77.6%的调查人员幸福指数大于等于4，说明77.6%的调查人员是比较幸福的，仅有7.5%的人感觉不幸福。

2.3 性别属性分析

绘制不同幸福指数下的男女数量图和比例图





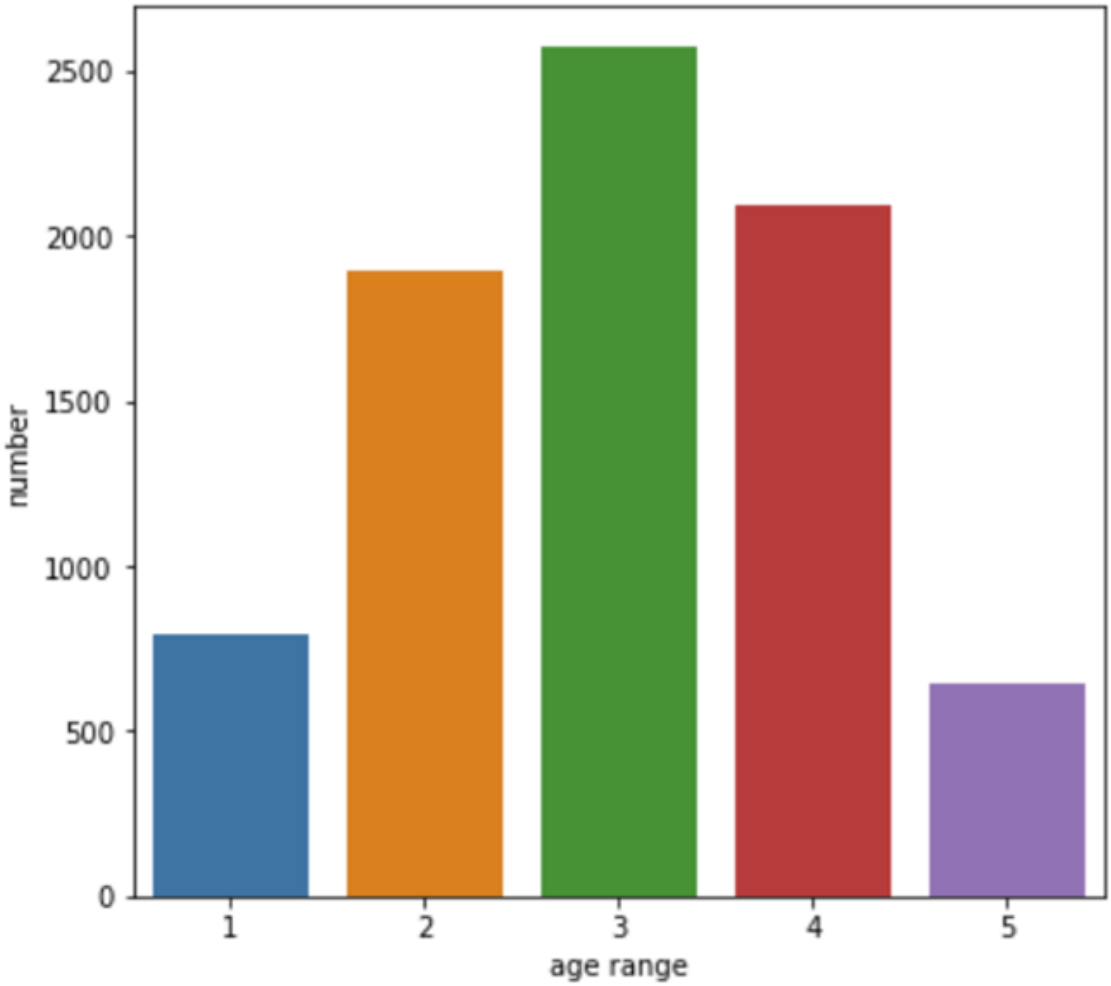
19.1%的女性认为自己非常幸福，高于男性的15.9%；78.3%的女性幸福感在比较幸福4以上，而男性的幸福感在比较幸福以上的为77.2%低于女性，所以由数据可知女性的幸福感是略高于男性的。

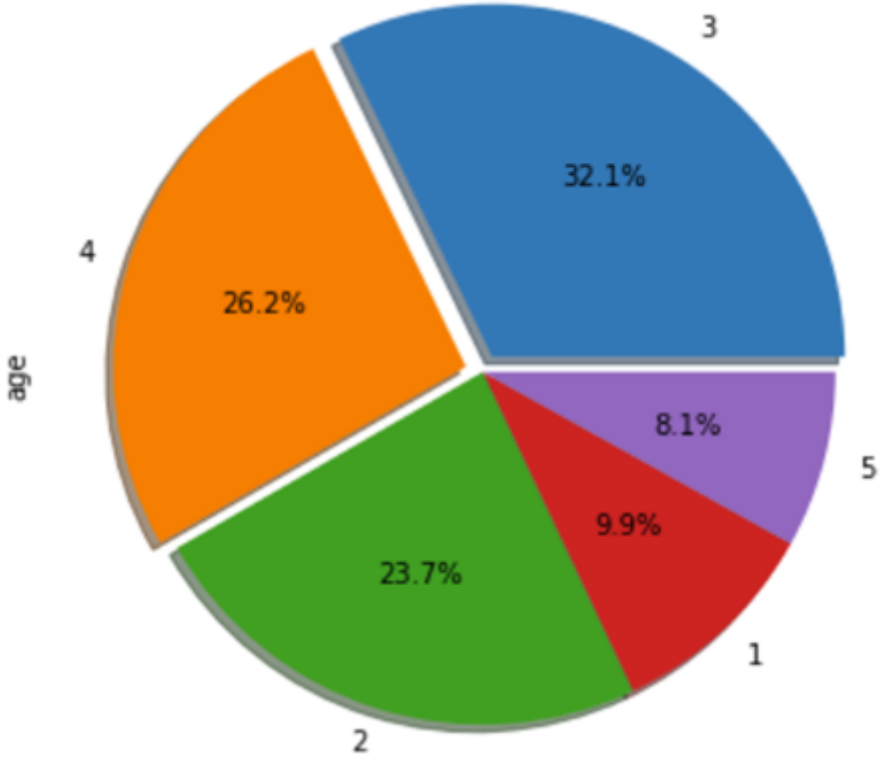
2.4 年龄段分析

首先将年龄段划分为5个区间，以1,2,3,4,5进行表示，再绘制各个区间的数量图和比例图

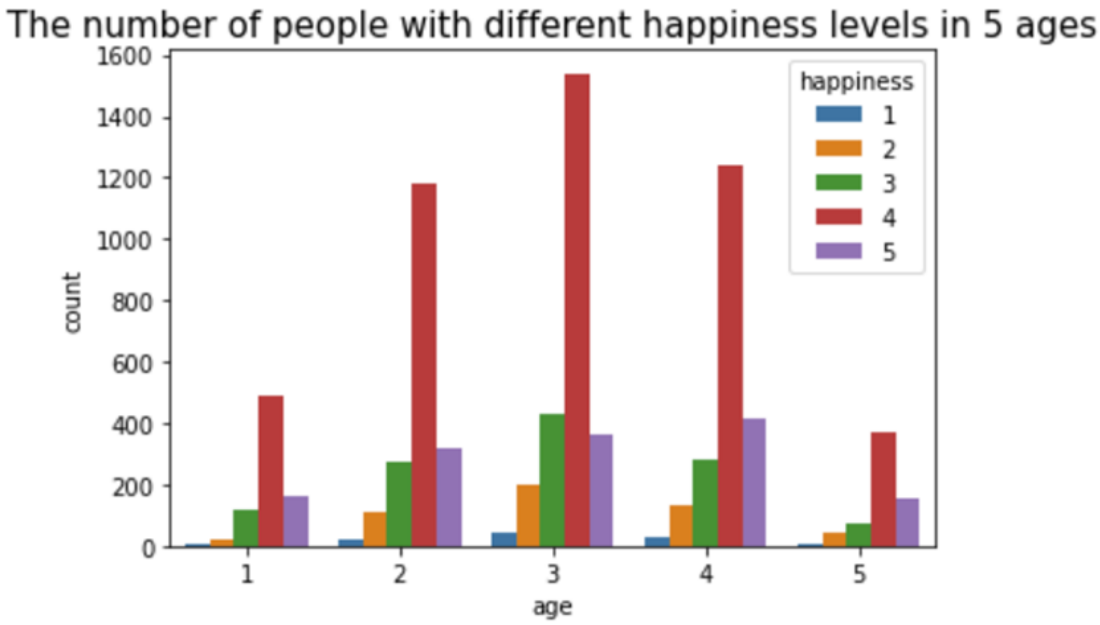
年龄分段表

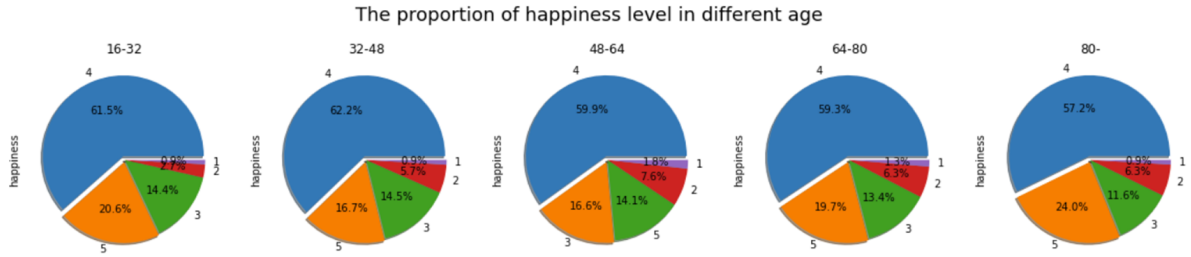
年龄区间	阶段
(16,32]	1
(32,48]	2
(48,64]	3
(64,80]	4
(80,)	5





由上图可知，在这次调查中48-64岁年龄段的人最多，80岁以上的人最少。继续分析每个年龄段内的幸福等级情况

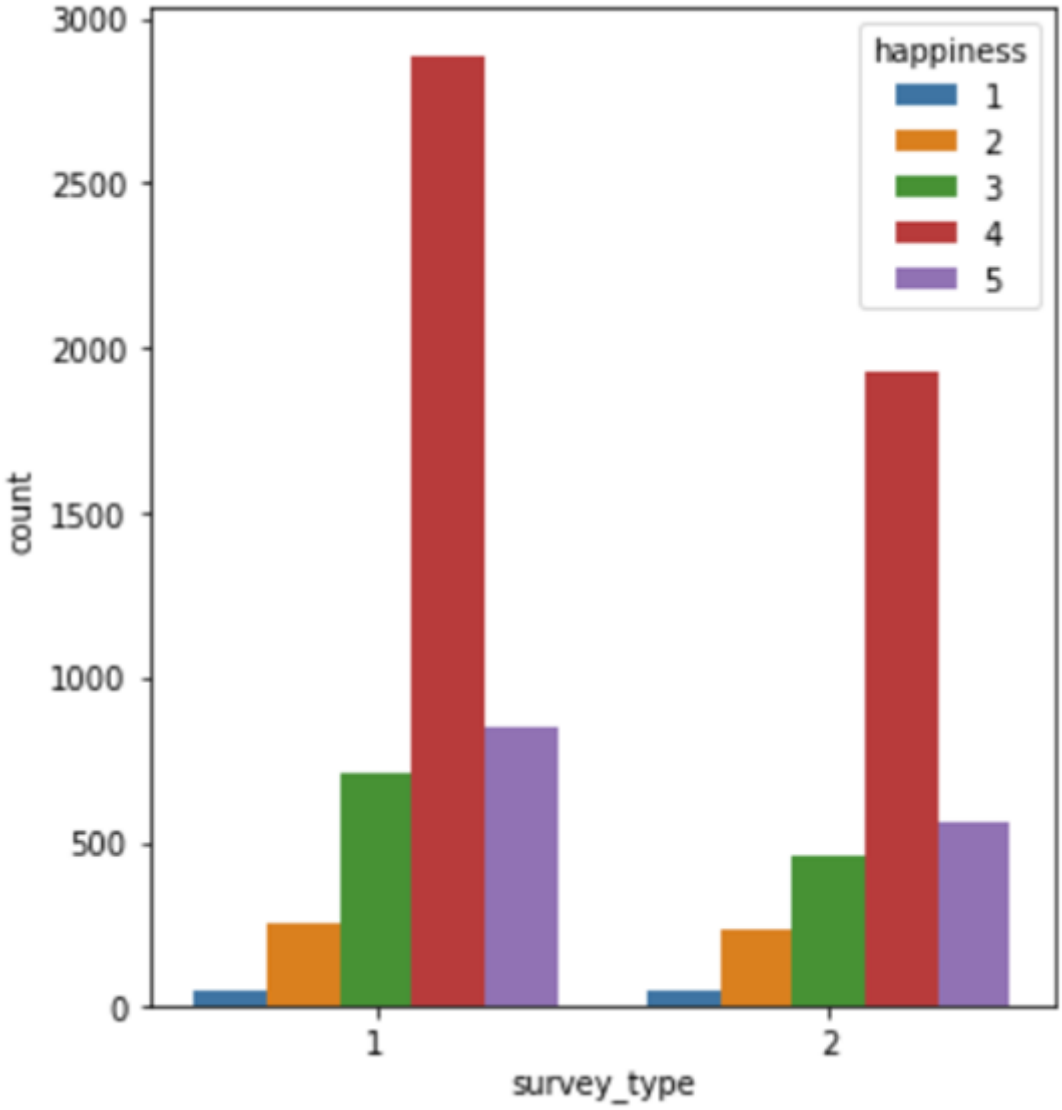


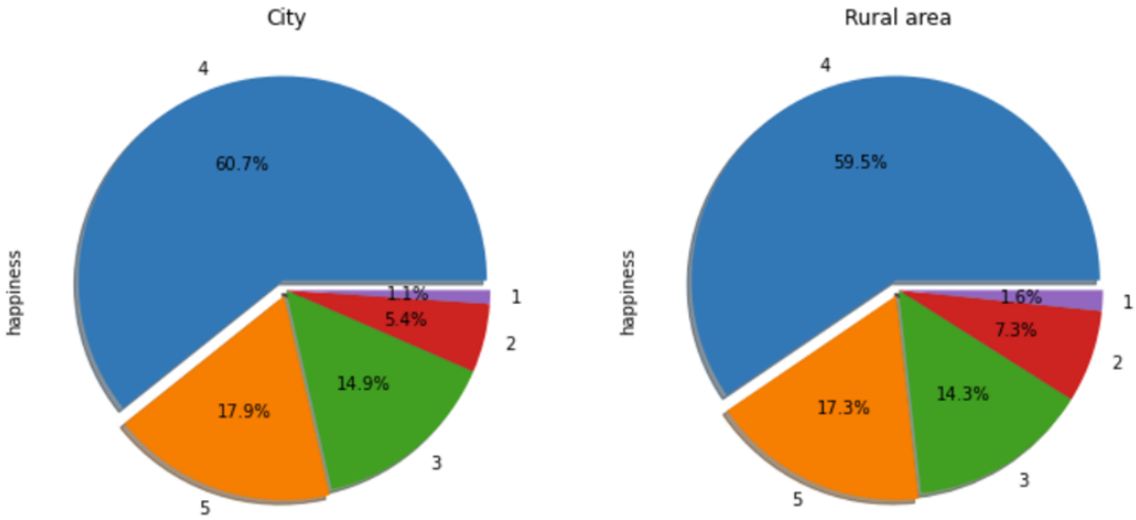


每个年龄段幸福感众数均为4，且分布情况很类似；总体幸福比例呈现先下降后提升趋势，合理推测为年轻人和老年人感觉更幸福，中年人因为压力原因幸福指数相对更低

2.5 地区分析

样本记录了城市（类型1）和农村（类型2）的幸福值信息

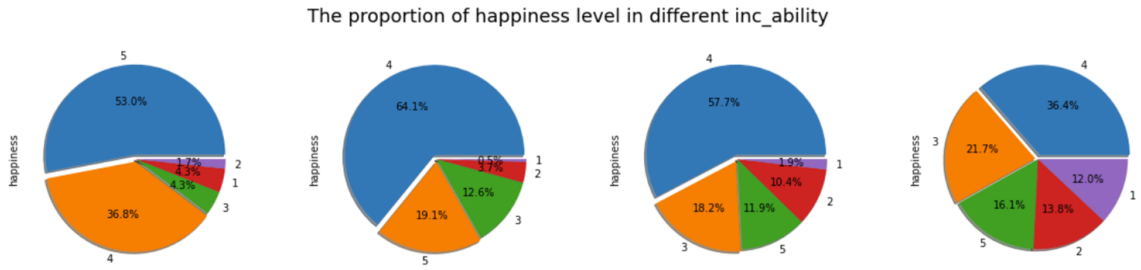




由图得出，城镇和农村幸福分布类似，幸福感的≥4的居民高于农村，生活在城市里的人比农村略微幸福。

2.6 收入合理分析

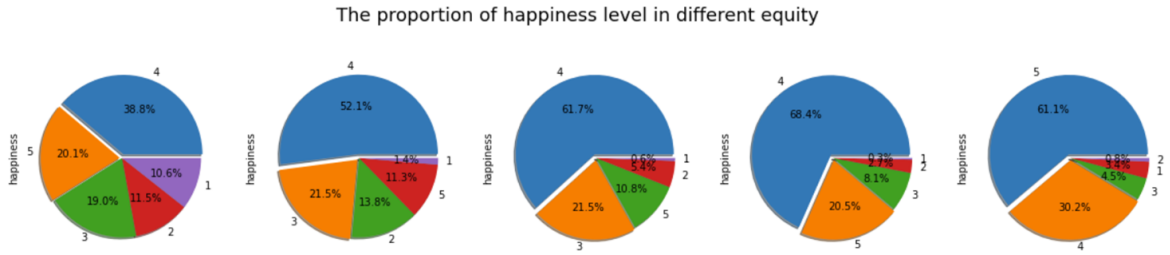
问卷中，inc_ability表示的是：考虑到您的能力和工作状况，您目前的收入是否合理？答案选择范围包括：1 = 非常合理; 2 = 合理; 3 = 不合理; 4 = 非常不合理。基于此绘制每个inc_ability的幸福指数比例图。



由图得出如下结论：认为收入非常合理人中，89.8%人比较幸福及以上；认为收入非常不合理人中，52.5%人是比较幸福及以上；随着收入合理程度递增，幸福感也递增，即一个人对自己目前收入认为越合理则幸福感越强，二者比较相关。

2.7 公平分析

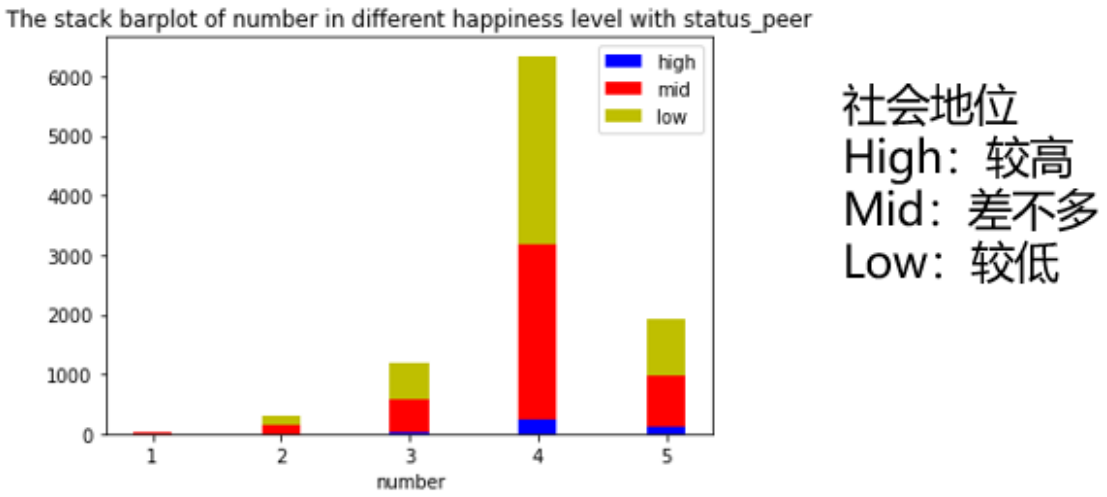
问卷中，equity表示的是：您认为当今社会公不公平；答案选择范围包括：1 = 完全不公平; 2 = 比较不公平; 3 = 说不上公平但也不能说不公平; 4 = 比较公平; 5 = 完全公平;。在每个公平程度下，绘制5个happiness level的比例图



在认为当今社会完全公平中，91.3%的人是比较幸福及以上的，在认为社会完全不公平中，58.9%的人是比较幸福及以上的，并且可看出随着人们对于社会公平程度的递增，幸福的比例也在递增，也就是说当一个人认为社会越公平时，他的幸福感就越强。

2.8 社会经济地位分析

问卷中，status_peer表示的是与同龄人比较社会经济地位；答案选择范围包括：1 = 较高; 2 = 差不多；3 = 较低。以与同龄人比较社会经济地位作为参量，绘制5个happiness level的堆叠条形图。



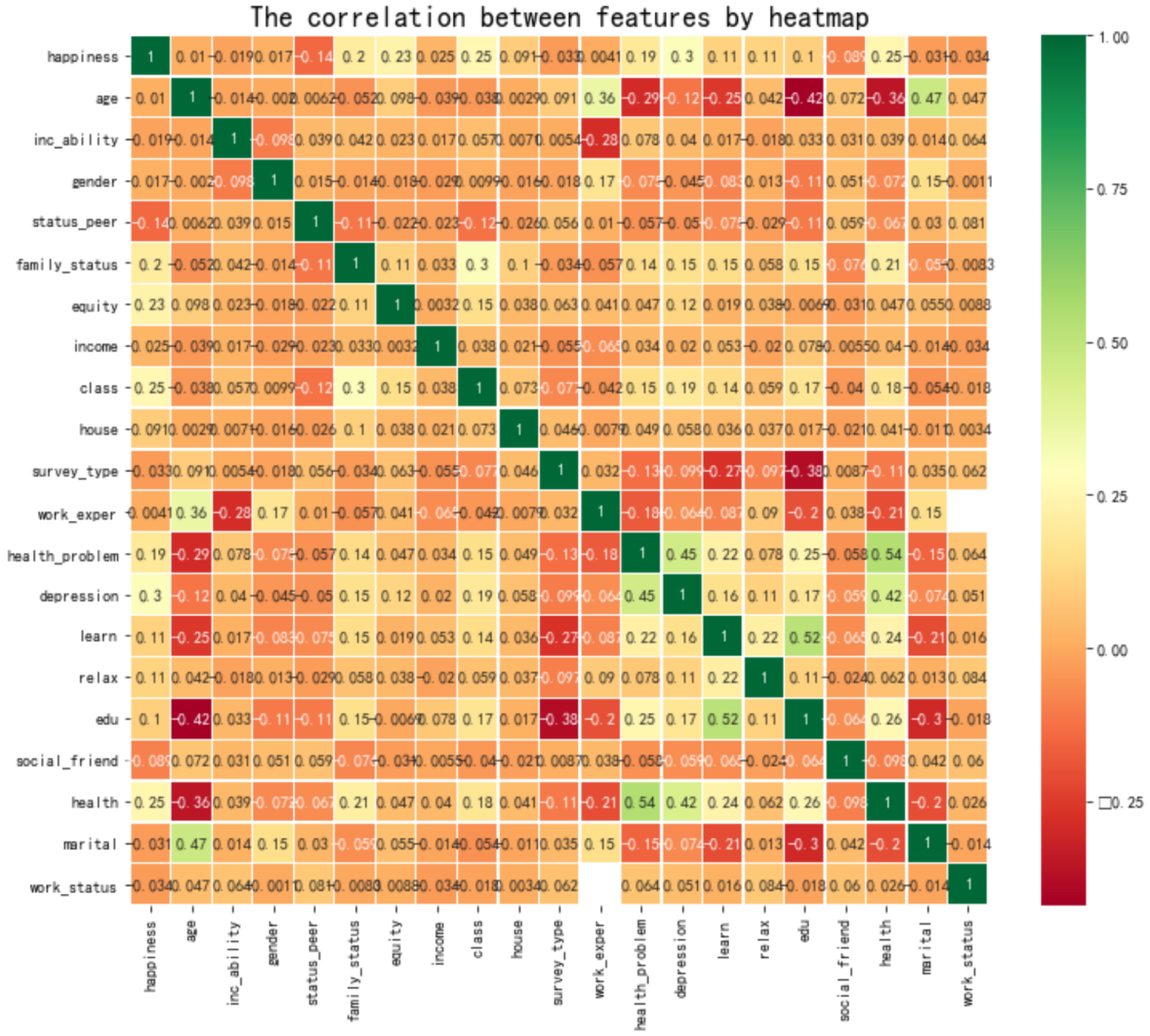
由上图可知，与同龄人相比，几乎50%人认为自己的社会经济地位是和同龄人差不多的，几乎50%人认为自己的社会经济地位比同龄人低，在每个happiness level下status_peer的分布情况大概一致，因此合理推断社会经济地位与幸福感无关。

3. 模型选取

3.1 特征Pearson相关性系数

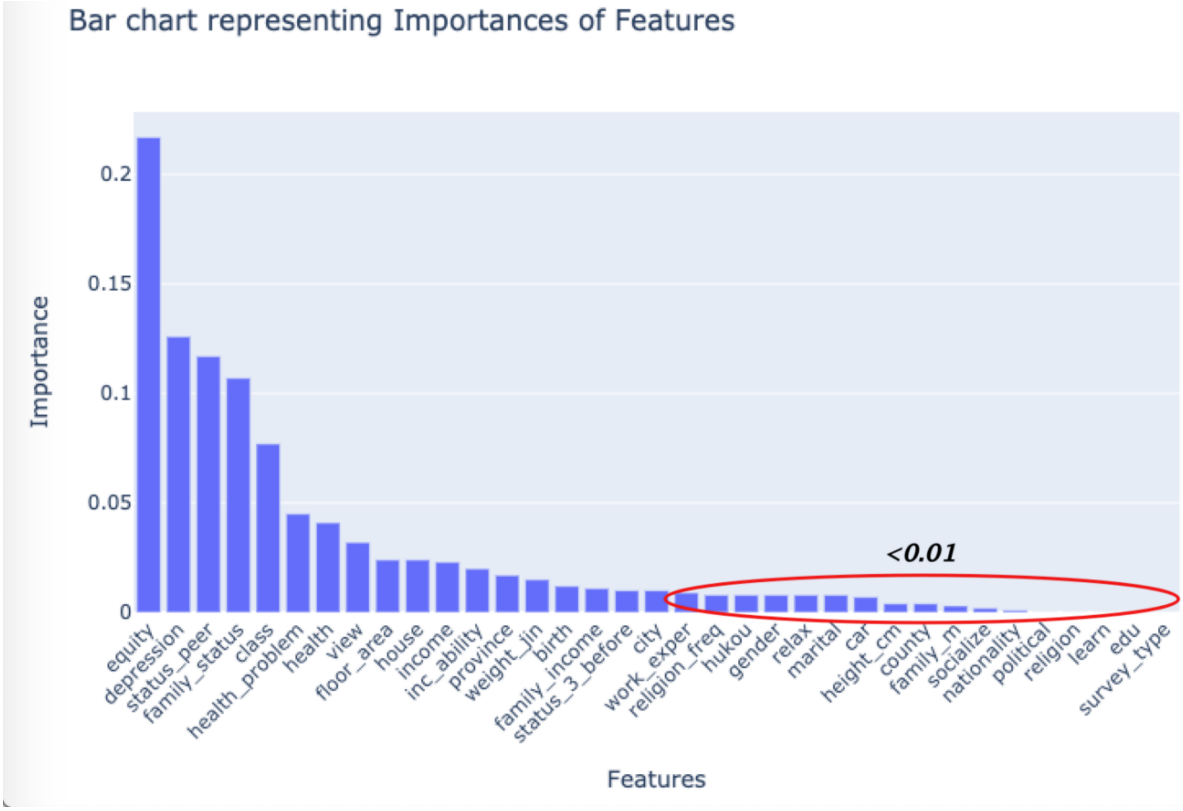
属性间相关性矩阵，对强相关的属性组可以只选择一个代表（暂未发现有强相关属性组，说明问卷的问题设置比较合理）

计算各属性与label的相关性，这里保留相关性绝对值大于0.05的属性，约60个。

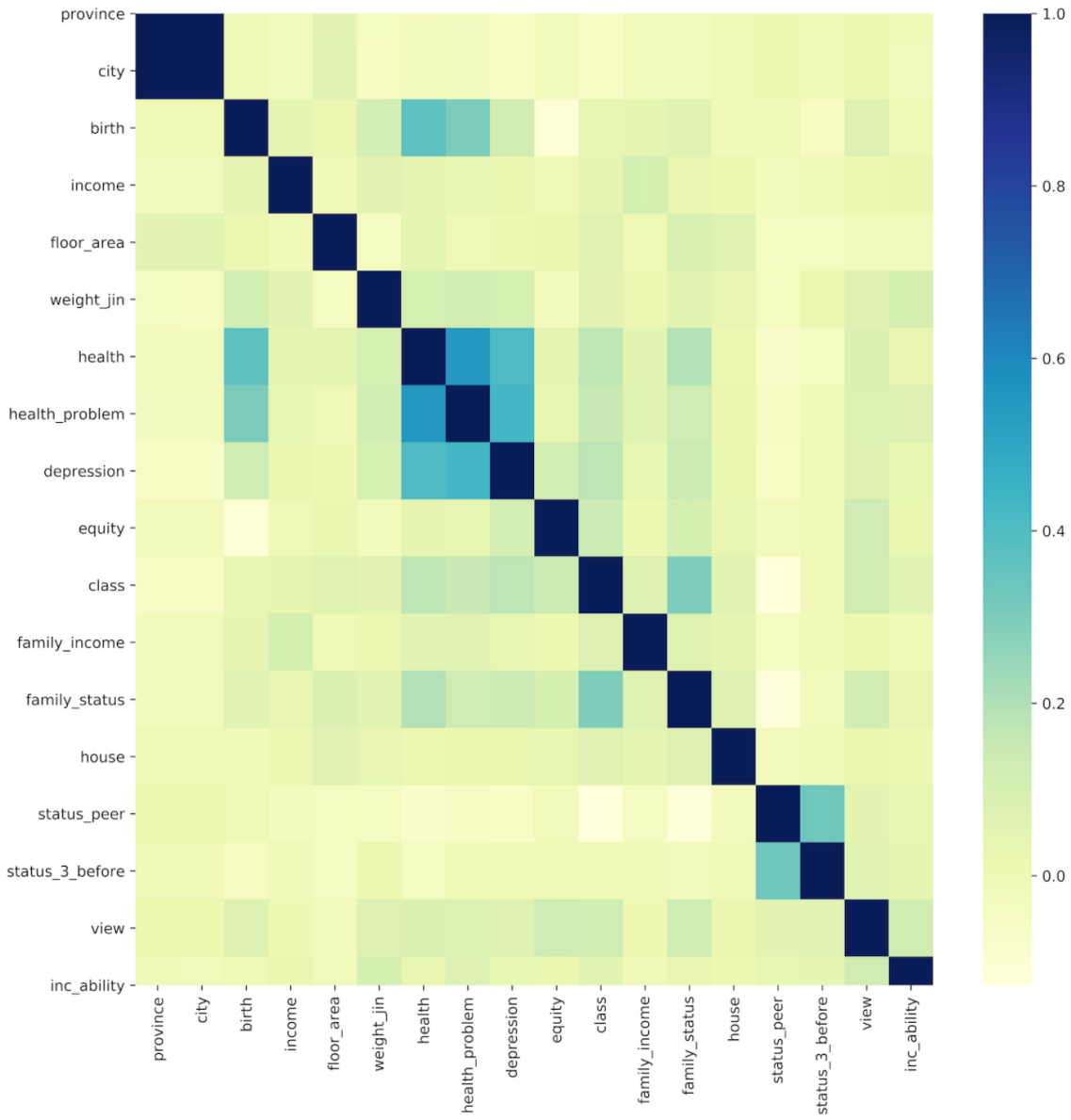


3.2 特征重要性

利用随机森林算法，进一步识别出哪些特征对幸福感更重要，排除那些不重要的特征，以便降低数据维度，加快模型训练。将4.1中的特征进行训练，参数设定n_estimators = 25; max_depth = 6;其余sklearn默认。得到10折交叉验证结果(准确率):[0.698, 0.699, 0.717, 0.713, 0.709, 0.699, 0.694, 0.712, 0.708, 0.704]。特征重要性如下图所示，保留18个特征。



经过筛选后发现各个特征之间的相关性均较弱，符合预期，故保留这些特征进行下一步训练。



3.3 初步建模

根据上一步特征保留如下：

- 'province'
- 'city', 'birth'
- 'income'
- 'floor_area'
- 'weight_jin'
- 'depression'
- 'equity'
- 'class'
- 'family_income'
- 'family_status'
- 'house'
- 'status_peer'
- 'view'
- 'inc_ability'

采用方法：支持向量机（SVM）、XGBoost、随机森林（Random Forest）、逻辑回归（Logistic Regression）

3.4 挖掘实验的结果

机器学习模型效果如下：

Model	Accuracy	Mse
SVM	0.71	0.51
XGBoost	0.71	0.51
Random Forest	0.7	0.51
Logistic Regression	0.7	0.52

模型的预测准确率大约在70%，MSE大约0.51，模型效果有限。

4. 深度学习方法

4.1 问题定义和数据表示

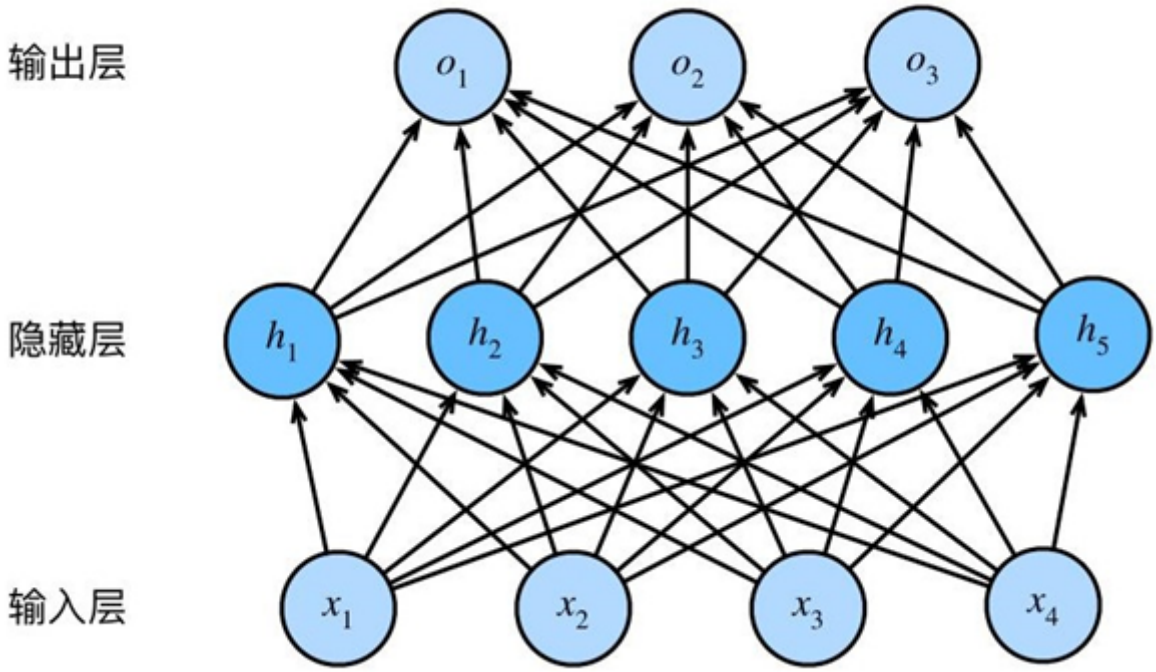
从官方的赛题描述来看，这是一个拥有五个类别的多分类问题（幸福值为-8的类别由于数量太少且过于特殊，不被考虑在内），如果处理之后的数据分布不存在问题，那么使用神经网络解决这个问题应该不会太困难。

我们对数据集进行了简单处理，并将处理的结果划分成训练集、验证集和测试集。具体的处理方法如下：

1. 剔除问卷中存在缺失值的项目；
2. 剔除问卷中数值过大的项目（比如收入），剔除问卷中时间类的项目（比如出生日期）；
3. 一份问卷为一个样本，将问卷的每一个项目作为向量的一维，将样本表示成向量；
4. 由于数量太少，删除幸福值为-8（表示“不知道”）的样本；
5. 通过“过采样和加入扰动”的方式降低训练数据分布不均衡问题带来的影响。

4.2 多层感知机（Multi-Layer Perception，MLP）

多层感知机是一种非常成熟也相对古老的神经网络了，它是一种前馈人工神经网络模型，由输入层、隐藏层和输出层构成，层与层之间是全连接的，能够很好地逼近任何非线性函数。其网络结构如下图所示：

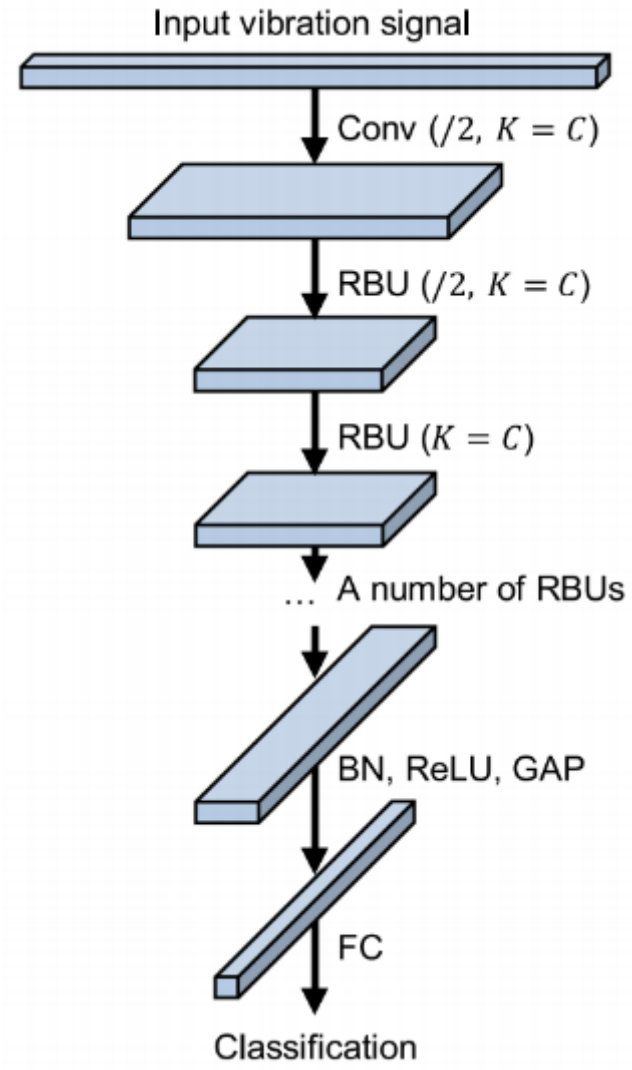


对于MLP，我们针对以下超参数设置进行了训练、测试和分析：

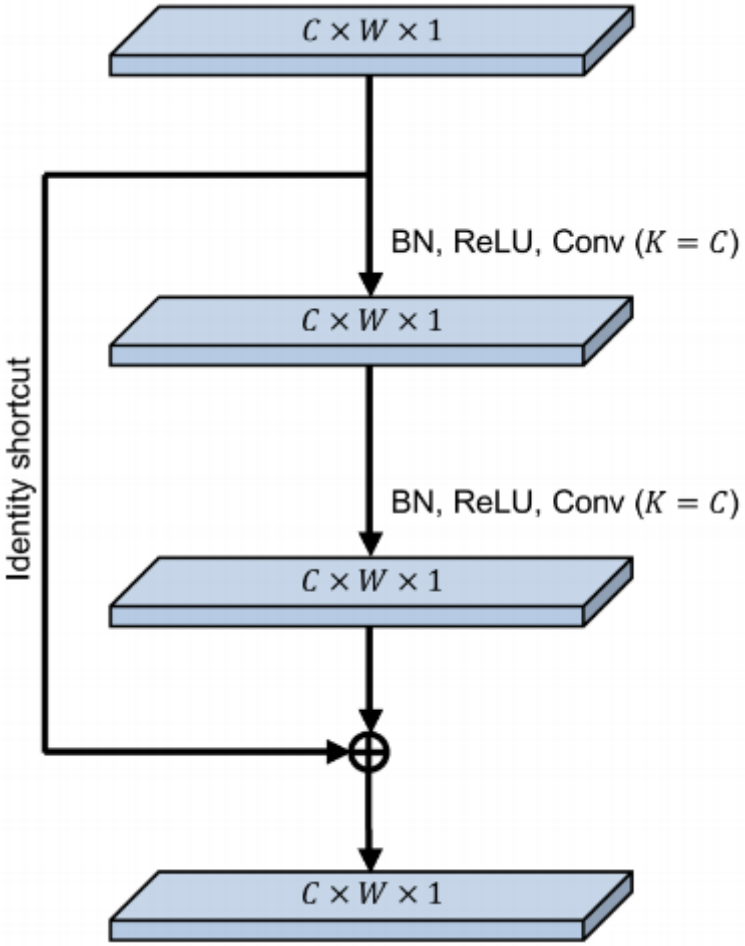
- 层数：2、3、4
- 激活函数：ReLU、Sigmoid
- 防止过拟合：Dropout、Batch Normalization

4.3 深度残差网络（ResNet）

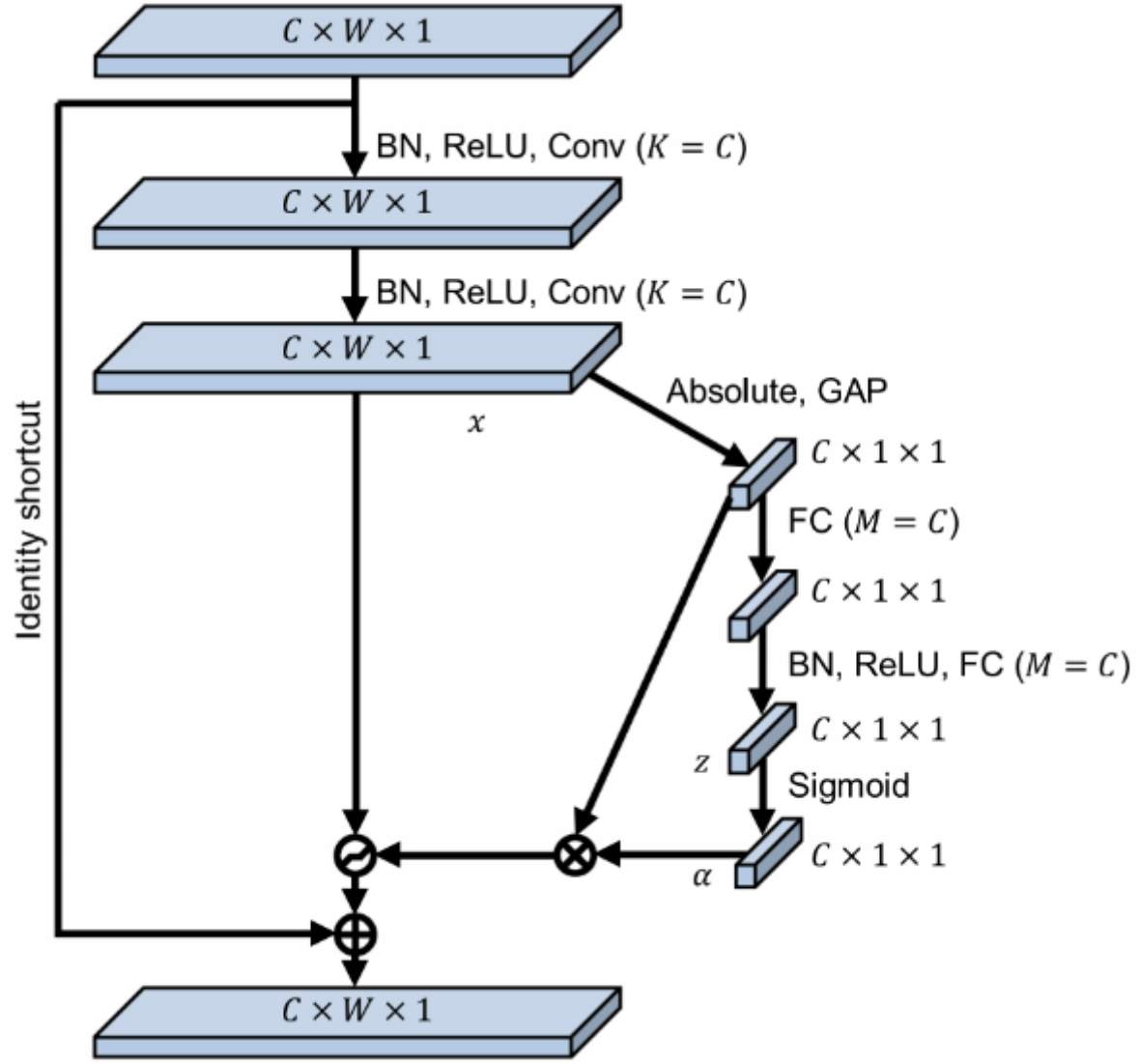
深度神经网络在解决分类问题上一直都是一种十分有效的手段。深度残差网络的结构如下图所示，由负责输入的卷积层、若干个卷积块和负责输出的全连接层顺序连接组成



基本的残差卷积块如下图所示，由若干个卷积层和残差结构组成；由-于现阶段的数据集是粗略处理得到的。



为了应对数据样本中包含的噪声，我们还参考了《Deep Residual Shrinkage Networks for Fault Diagnosis》论文中对深度残差网络结构的改进，实现了基于一维卷积的残差收缩块，其结构如下图所示：相对于基本残差块，残差收缩块只是在倒数第二个卷积层和最后一个卷积层之间加入了收缩结构以实现软阈值化，从而降低数据噪声的影响。通过堆叠不同数量的残差块，可以得到层数不同的深度残差网络。

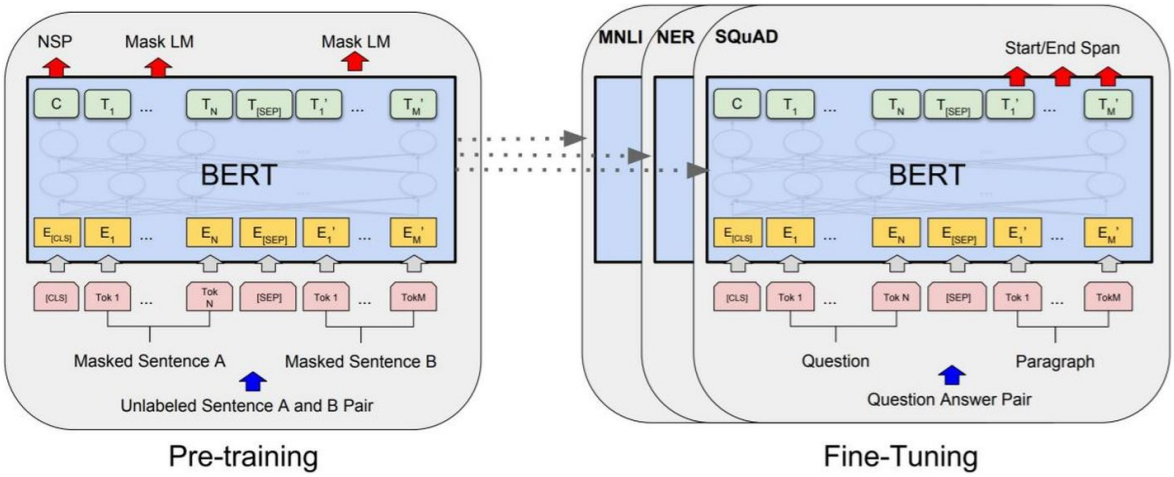


4.4 语义分类模型

不用于多层感知机网络和残差网络直接将收集到的数据建模称为特征的方法，我们还采用了基于语义进行文本分类的神经网络方法，具体方法如下：

1. 利用模板将收集到的数据根据问卷表构建描述文本
2. 将文本和标签对应，构建数据集
3. 利用自然语言处理种的文本分类技术对所构建的文本进行分类

本方法采用预训练模型BERT进行文本分类。BERT模型是Google公司在2018年发布的语言表示模型，是自然语言处理中常用的预训练模型，对于自然语言理解任务比如命名实体识别、文本分类是有效的方法，其结构如下图所示。



使用BERT模型进行文本分类是近年常见的文本分类方法，并且实现简单快捷，因此我们选择该分类技术进行实验。

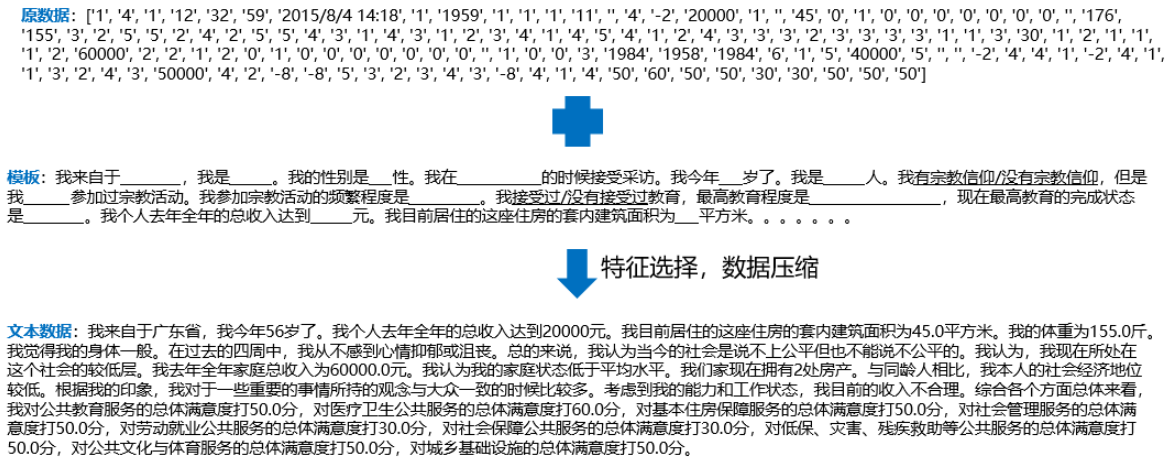
我们设计采用文本分类的方法代替直接特征提取的原因在于：早期的机器学习方法或神经网络方法都需要人类手动地提取特征，而本次实验中收集到的数据就是我们所需要的特征，但是近年来深度神经网络和大规模预训练模型快速发展，使得神经网络模型可以从文本原始数据中提取到更加合适的特征信息。为了利用这种信息，我们决定将特征还原回原始文本，这样不仅可以将离散的特征数据变成具有语义可解释的文本数据，而且有利于模型从中提取出更适合模型的特征。

4.5 基于模板的语义构建

为了能够使用文本分类的方法，我们需要先将收集到的离散特征信息还原出原始的文本信息，由于本实验中的数据信息来源于问卷收集，因此我们通过问卷的问题和选项，利用模板进行特征的语义构建。

具体模板设计参考代码`text_class/template.py`

利用问卷问题模板，我们可以将没有任何语义的特征还原成具有语义信息的文本，如下图所示：



我们利用这种方式在训练集、验证集和测试集构建具有语义信息的文本数据集，并用它训练文本分类器，最终用于对测试集进行分类。

4.6 实验结果

由于比赛提供的test数据集不包含happiness一项，所以目前模型的训练和测试只能在其提供的train数据集上进行。我们将train数据平均分为10份，1份作为测试集，其余9份作为训练集，再从训练集中随机取1份作为验证集，以观察模型对训练集的拟合情况；实验采用十折交叉验证测试方法。最终的训练测试结果如下表所示：

Model	Dev_Acc	Dev_MSE	Test_MES
MLP_Dropout_3	72.93%	0.426	0.472
MLP_BatchNorm_3	89.93%	0.127	0.212
MLP_BatchNorm_2	86.01%	0.189	0.272
MLP_BatchNorm_4	90.59%	0.113	0.203
ResNet_Normal_18	100.0%	0.000	0.098
ResNet_Shrink_18	100.0%	0.000	0.095
Text2Class	92.35%	0.027	0.177

如上表所示，对于MLP来说，使用批量归一化替代dropout层可以使模型更好的拟合训练集，从而提高测试结果，而增加MLP的层数只能带来很小的提升；18层的深度残差网络能够非常好地拟合训练集，使用残差收缩块替代基本残差块可以使结果有少量提高，但是残差块的收缩结构增加了模型参数数量和极大地降低模型的运行速度。

5. 任务分配

人员	任务
张雨婷（3220201015）	数据预处理，探索性分析
李安腾（3220200897）	数据预处理，探索性分析
杨毅哲（3120201088）	深度学习算法实现与分析
王梓屹（3220190595）	机器学习算法实现与分析
田雄辉（3220200957）	深度学习算法实现与分析