世界幸福指数研究

背景介绍：

我们生活在一个矛盾重重的时代：世界享受着难以想象的复杂技术，然而每天却仍然有10亿人吃不饱饭。技术的不断进步，使得世界经济飞速发展，生产力达到了前所未有的高度；然而在这个过程中，自然环境遭到了破坏，人们逐渐屈服于肥胖、吸烟、糖尿病、抑郁症，以及现代生活中的其他弊病。优越的生活环境并不一定意味着较高的幸福感，生活在贫穷环境中的人绝大多数谈不上幸福，但微小的改变，却可能极大地促进这些人的幸福感，因此，什么是幸福，什么因素会影响幸福感，是一个值得我们思考的问题，这也是我们小组这次研究的主要问题。

我们的数据来自于《世界幸福报告》,《世界幸福报告》是全球幸福状况的里程碑式的调查。第一份报告于2012年发布，第二份于2013年发布，第三份于2015年发布，第四份于2016年更新。今年3月20日,联合国发布了《2019年世界幸福报告》(World Happiness Report 2019)，根据156个国家的幸福水平对其进行排名。随着各国政府、组织和民间社会越来越多地使用幸福指数来为其决策提供信息，该报告相继得到全球的认可。经济学、心理学、调查分析、国家统计、卫生、公共政策等多个领域的权威专家描述了如何有效地利用幸福感测量来评估各国的进步。报告回顾了当今世界的幸福状况，并展示了新的幸福科学是如何解释个人和国家幸福的变化的。

我们希望通过对幸福指数建模，能够让我们对影响幸福的关键因素有一个初步的认识，可以引导我们重新思考幸福的关键是什么，并且能够结合模型给出提高幸福指数的方法。

数据预处理：

本次数据来源于kaggle（https://www.kaggle.com/unsdsn/world-happiness)，幸福指数(Happiness score)是基于一个生活评估问题的回答得出的，这个被称为“Cantril ladder“的问题，要求被调查者想象出一个他们认为最好的生活的样子，设为10分，然后想象出一个他们认为最差的生活的样子，设为0分，接着根据自己当前生活的样子，对当前的生活进行一个评分，这个分数位于0-10之间，代表了对当前生活的一个满意程度，作为幸福指数。

幸福指数后面的列估计了六个因素（经济生产、社会支持、预期寿命、自由、不腐败和慷慨）中的每一个因素在多大程度上有助于使每个国家的幸福指数评估高于反乌托邦（一个具有与世界同等价值的假想国家，但六个因素中每一个都是最低水平）。这些因素对每个国家的幸福指数总分没有影响，但它们解释了为什么有些国家的排名高于其他国家。

数据采集时间为2015年至2019年，共5年，每一年的数据在变量的命名和变量的数量上都有一些不同。首先，我们将同一个意思但变量名不同的变量进行了统一，主要包括以下几点：

1：Family 和 Social support 都是指家庭状况

2：Perceptions of corruption 和 Trust(Government Corruption)都是指对政府的信任程度

3：Freedom to make life choices 和 Freedom 都是指人生抉择自由

随后，我们统计了每一年的变量情况，表格如下：

	2015	2016	2017	2018	2019
Country	√	√	√	√	√
Region	√	√			
Happiness Rank	√	√	√	√	√
Happiness Score	√	√	√	√	√
Standard Error	√				
Economy(GDP per Capita)	√	√	√	√	√
Family	√	√	√	√	√
Health (Life Expectancy)	√	√	√	√	√
Freedom	√	√	√	√	√
Government(Government Corruption)	√	√	√	√	√
Generosity	√	√	√	√	√
Dystopia Residual	√	√	√		
Lower Confidence Interval		√	√		
Upper Confidence Interval		√	√		

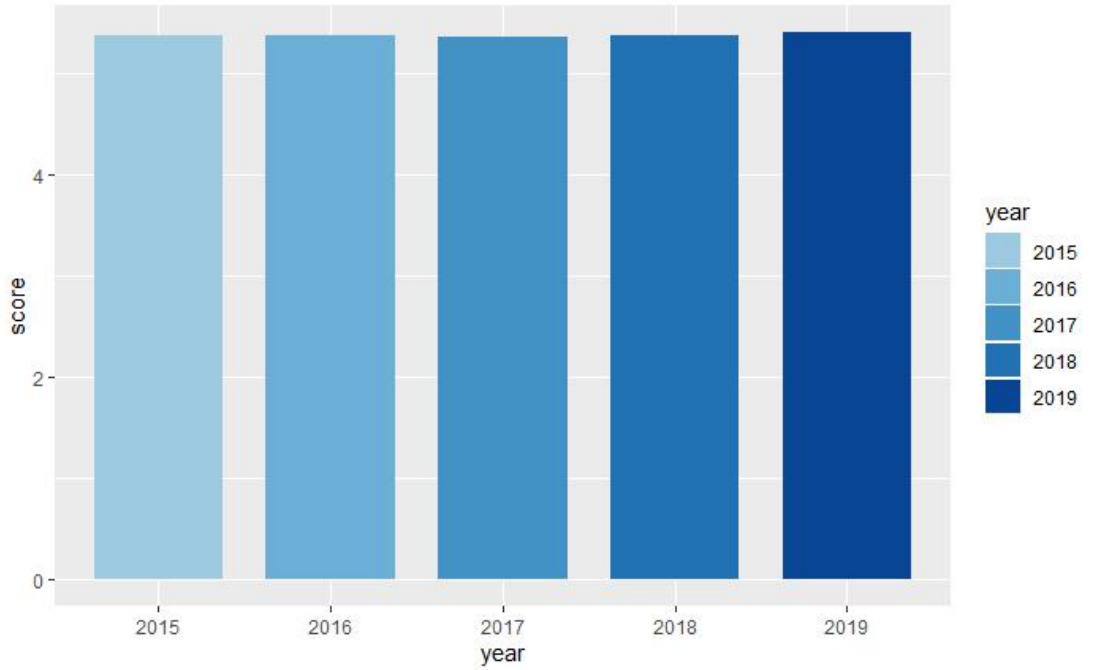
根据上表的完整度，我们可以看到一些最重要的基本信息（表中框内）比如 GDP，家庭状况等是每年都进行了统计的。因此，我们首先对每一年的数据进行了统一变量的处理：对每个国家，保留其 Happiness Score(幸福指数)，Economy(人均 GDP)，Family(家庭状况)，Health(预期寿命)，Freedom(自由程度)，Government(政府信任程度)，Generosity(慈善程度)7 个变量。我们将每年的数据进行清理后按行进行合并，在描述分析时，国家这一列依然保留，在建模时删去，另外，建模时我们将前四年作为训练集，2019 年作为测试集。这些变量的基本信息如下表所示（合并 5 年的数据，不包含国家列）：

	Min	1st Qu	Median	3rd Qu	Max	Mean
Happiness Score	2.693	4.510	5.322	6.189	7.769	5.379
Economy	0.000	0.607	0.982	1.236	2.096	0.916
Family	0.000	0.869	1.125	1.327	1.644	1.078
Health	0.000	0.440	0.647	0.808	1.141	0.612
Freedom	0.000	0.310	0.431	0.531	0.724	0.411
Goverment	0.000	0.054	0.091	0.156	0.552	0.125
Generosity	0.000	0.130	0.202	0.279	0.838	0.219

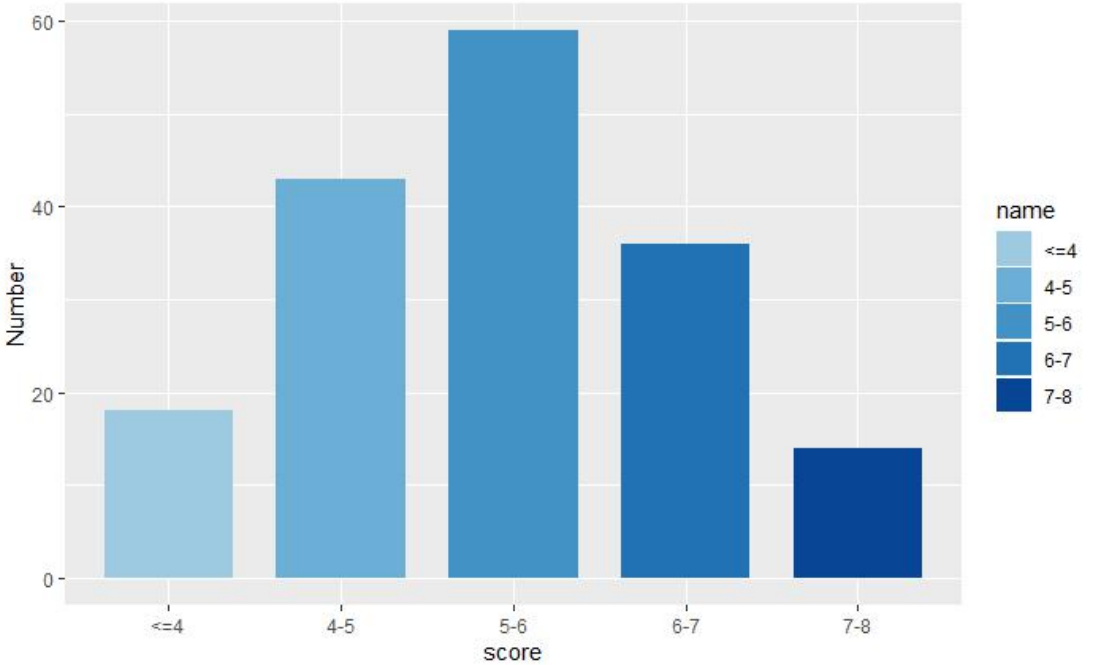
在本次案例分析中，我们将幸福指数 Happiness Score 作为因变量，将剩余六个变量作为自变量。

描述性分析：

首先是全世界每年的平均幸福指数：

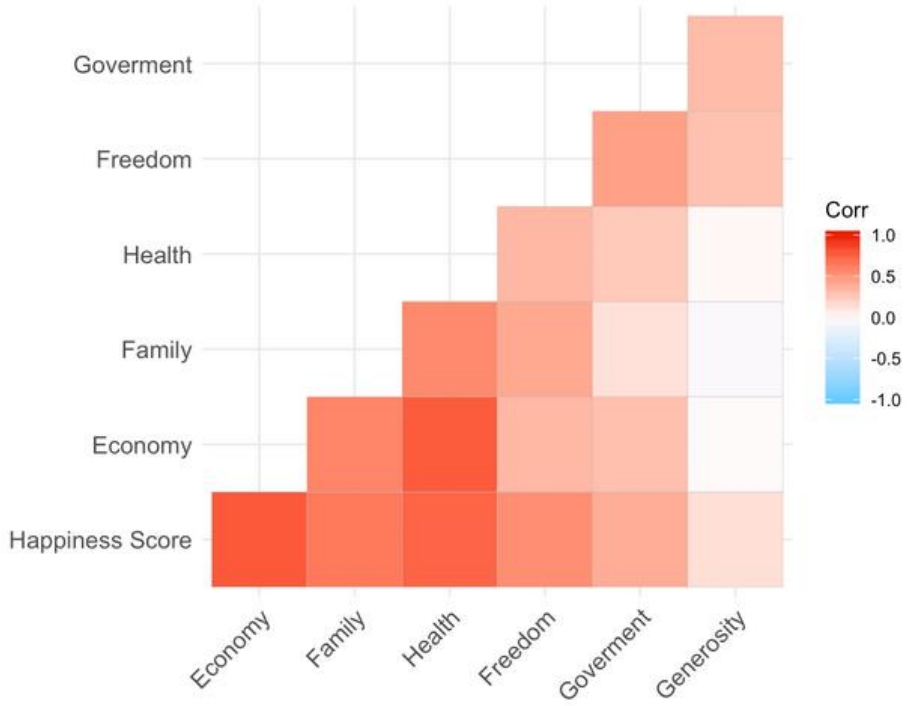


从 2015 年到 2019 年的五年间，全世界的平均幸福指数基本没有什么变化，基本在 5.3 左右，不过实际上总体还是呈现一个增长的趋势，只不过增长的速度特别缓慢。下面这张图代表了每个国家平均幸福指数的分布：



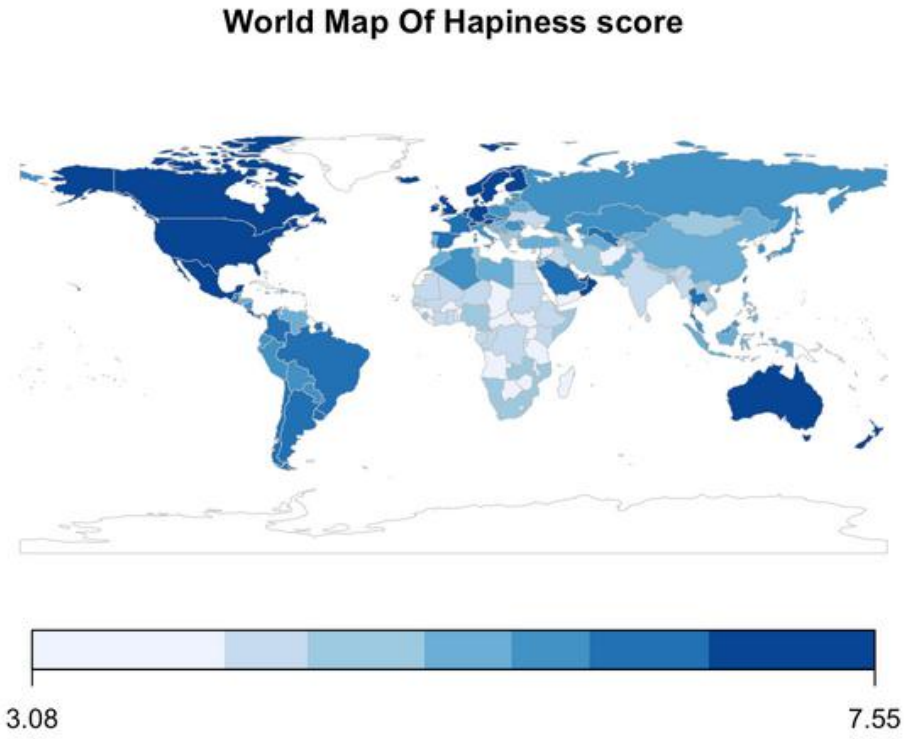
可以看到在全世界范围内，幸福指数总体上呈现一个对称分布，其中分数在 5-6 的国家数量最多，这个分布情况是比较合理的，因为从全世界范围来看，发达国家和极其贫穷的国家居少，发展中国家居多，而一般来说发达国家的幸福指数较高，贫穷国家的幸福指数较低，发展中国家的幸福指数趋于平均值，符合常理。

接着，我们绘制了各变量之间的相关性图，颜色越深，说明这两个变量的·相关性越强：



从图中可以看出：幸福指数和剩余六个变量之间都有一定的相关性，其中与经济水平和健康程度的相关性非常强，这一现象符合人们对幸福的基本定义：有钱花，活得健康长久。另外，自变量之间也有一些符合常理的相关性存在。比如经济水平和健康程度，经济水平和家庭状况等。

为了更直观地体现出各个国家的幸福指数情况，我们绘制了这张全世界五年的平均幸福指数情况图，颜色越深的地区幸福指数越高：



从图中可以看出北美，北欧以及大洋洲的颜色最深，说明这些地区的平均幸福指数一直很高。这些国家主要是发达国家，在人们眼中也是高收入，高福利的代表。相反的，非洲以及南亚，中东的大部分地区颜色偏浅，与之对应的，是这些国家常年的贫穷，战乱等。因此，一个国家的幸福程度与其经济水平，自由程度，社会福利等因素肯定是具有强烈的相关性的，接下来我们就通过具体的建模来分析这些关系。

模型构建及应用解读：

我们根据处理过的数据集（使用 2015 年至 2018 年的数据进行构建模型，2019 年的数据集当做预测集）构建了两个模型：1、回归树模型，即决策树模型的因变量为连续值；2、线性回归模型。两个模型对于世界幸福指数的解读在不同方面各有千秋，回归树更为直观具体，我们可以直接通过可视化回归树的方式来查看影响幸福指数的关键要素是什么，并且可以根据不同国家的具体国情进行分析，进而为不同国家提出不同的策略；而线性回归模型则在预测方面更为精确，并且也能够通过解读线性回归系数的方式来加以解读什么对于幸福指数才是最为重要的；同时通过线性模型的预测，我们可以找出违反不符合模型预测的异常点，即特别的国家和地区，单独分析这些国家和地区将为我们探索如何提升幸福感这一问题提供崭新的观点和思路。

通过将幸福指数的得分置为因变量，将每个国家的六项因素的得分置为自变量：人均 GDP（Economy）、家庭状况（变量名为 Family，原文为 social support，有人直译为社会支持，但实际上指的是对于家庭的评分）、预期寿命（Health，原文为 life expectancy）、人生抉择自由（freedom，即考量社会上是否存在过多限制，进而导致了经济、文化等方面的抑制，评分越高相对越越自由）、政府腐败程度（变量名为 Government，原文为 absence of corruption，即没有腐败，相当于反过来评价，等价于对政府打分，得分越高表示民众对政府越信任）、慷慨程度（Generosity，反映了社会中慈善工作的好坏，得分越高越好），我们构建了这两个模型。接下来则依次介绍这两个模型。

一：回归树模型

模型解读：

由于幸福指数是一个 0 到 10 之间的连续值，因此该建模问题是一个回归问题，使用树模型来构建此模型时，预测值是叶子节点所含训练集元素输出的均值。

构建的回归树的结果如下图所示：

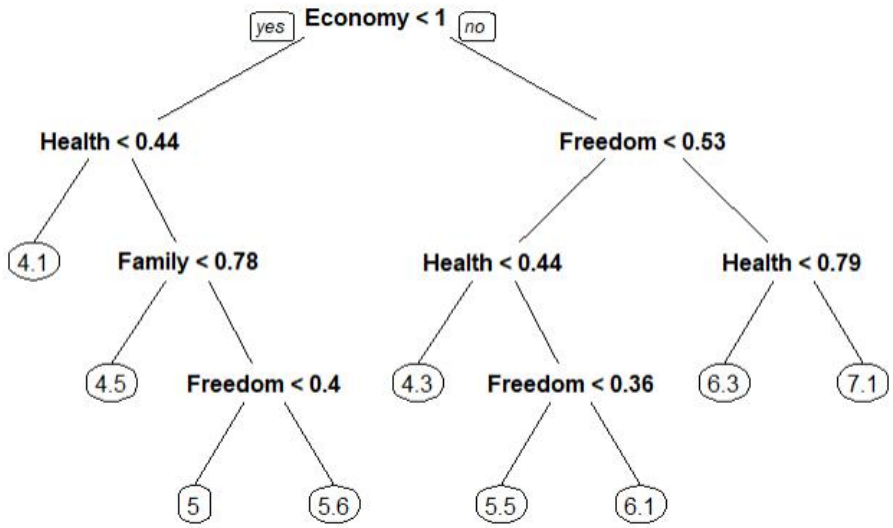


图 1：幸福指数的回归树模型

从图中我们可以看到，首先区分一个国家幸福度高低的便是人均 GDP 这一经济因素，而阈值则是 1。而这也反映了为什么几乎所有的国家都将经济发展当做首要目标，同时也昭示了我国以经济建设为中心的社会主义基本路线的原因。而随着经济的增长和人民的富裕，经济对于人们幸福感的影响力则再逐渐下降。在国家经济水平较低，百姓衣不附体、食不果腹时，过上小康生活就是人们感到的最幸福的事情；而当百姓逐渐富裕起来后，对于经济的依赖不再那么显著和重要，自然幸福感便受到其他因素的影响。

因此我们可以看到当人均 GDP 得分大于 1 时，影响人们幸福感的便是人生抉择自由这一因素。而这一点则映照了当今世界的趋势：例如在许多欧美等发达国家，经济的富裕使得人们的幸福感不再过多地受经济所影响，而更多的是希望自己能够走自己选择的人生。就像米尔顿·弗里德曼和罗斯·弗里德曼在《自由选择》这本书中说的一样，正是由于华盛顿当局制定了过多的法律法规、实施了过多的政府管制、建立了过多的行政机构、花费了过多的财政预算，才使人们的自由和财富受到了侵蚀和削弱。因此这一因素较为显著地影响了人们的幸福水平。

而当一个国家的人均 GDP 的得分小于 1 时，我们从图中可以看到预期寿命得分对于人们的幸福水平有着较为显著的影响。而这在我们国家则得到了证实，每一次医药体系的变动，都牵扯着上千万家庭的心。对于像我们中国这样人口上十亿的大国来说，提升医疗服务、构建完善地医疗体系意义重大，自然能够极大地提升人们的幸福水平。

模型价值：

虽然当今大多数国家仍旧按照国家生产总值等指标来判断一年的发展，但是人民的幸福与否仍旧是每一个国家至关重要的大事。通过构建回归树这一模型，一个国家可根据自身的发展状况来制定相应的策略。如果一个国家的人均 GDP 得分低于 1，则其首当其冲需要努力的方向便是发展经济，让国家富起来，同时这个国家还需要尽可能地完善国内的医疗卫生体系，让老百姓有病可医。而类似于抉择自由这样的因素可暂时不用考虑。

而如果这个国家是发达国家，则其需要注意的便是适当地放宽政策和限制，让每个人都能够选择自己的人生，而不再受社会等其他因素的限制。

二：线性回归

模型解读：

线性回归后的系数如下图所示：、

变量	回归系数	p 值	显著性
截距项	2.205	<0.01	***
人均 GDP 得分	1.105	<0.01	***
家庭得分	0.637	<0.01	***
预期寿命得分	1.137	<0.01	***
抉择自由得分	1.411	<0.01	***
政府得分	0.851	<0.01	***
慈善得分	0.544	<0.01	**

图 2：幸福指数的线性模型

该线性回归模型的 $Adjusted\ R^2$ 是 0.7625，同时我们从图 2 中也可以看到除了慈善得分以外各个系数的显著性都较高，因此该模型具有较高的可信度和参考价值。
从图中我们可以看到各个回归系数的大小。可以看出，抉择自由得分的系数最高为 1.411，对于每个国家的幸福指数起着最为显著的影响，而除此之外，预期寿命和人均 GDP 的得分也在幸福指数的评判中起着重要的作用，而这也是符合我们的常识和预期的，一个国家的经济决定这一个国家的命脉，而百姓的衣食住行更是与经济息息相关，人均 GDP 增长，表明了一个国家经济实力的突破，更是代表着百姓生活水平和消费能力的提高。而预期寿命的增长实际上是代表着一个国家医疗水平的提高，表现在医药体系的改善等方面上。百姓能够看得起病，看的好病，预期寿命自然会大幅增加，而人民的幸福感也会大大增强。

从图中我们发现一个国家慈善方面的得分并没有对幸福指数造成较为显著的影响，可理解为这一方面与普通大众的联系并没有其他方面直接，对于普通百姓的生活的影响不是很突出，因此在评判中所起到的作用自然也就没有那么显著了。

模型价值：

该线性回归的模型价值可通过两个方向上的思考来体现。

1：正向思考

由于该线性回归模型中的每一项系数都具有较强的显著性，因此我们可以直接依据此来提出建议。通过浏览系数，我们发现抉择自由对于人们的幸福水平起着较强的影响，因此这一项因素对于每一个国家来说都至关重要，每一个国家都需要通过改善社会的政策体系，帮

助每一个公民实现其自身价值；进一步地完善社会分配制度，从而减小贫富差距，降低实现自身价值的门槛；促进教育公平公正，使每一个青年人都有能够实现自己理想的能力；传播积极思想，做好教育工作，消除人们心中的成见和歧视观念，让每一个公民都能够在法律允许的范围内自由的选择自己的人生。

同时一个国家还需要尽自己所能来发展经济，这对于发展中国家更为重要和现实。并且在发展经济的同时，还需要发展自身国内的医疗水平，提高人们的预期寿命。

2：反向思考

我们使用 2019 年的数据来作为预测集，预测的结果如图 3 所示。

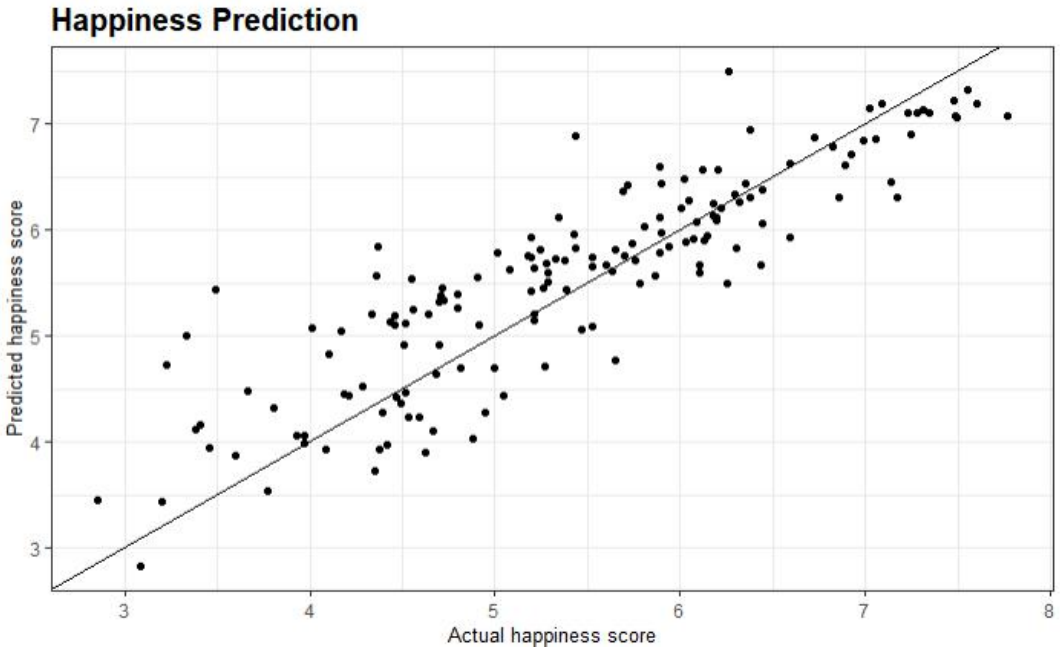


图 3：幸福指数预测图

从图中可以看出，该模型的预测结果较为准确，但仍旧有离群点，即预测结果与实际值偏差过大，因此我们便思考究竟是什么原因导致了这些国家预测不准，而对于这一角度的探索也正是显现该线性回归模型的价值的另一个方面。

我们找出了预测值与实际值偏差大于 1 的国家，如图 4 所示。

从图 4 中我们可以看到，所有的偏差大于 1 的情况都是预测值过大，而其中只有新加坡和香港属于发达地区，排名靠前（新加坡 34 名，香港 76 名），其他 6 个国家都隶属于发展中国家。

通过查阅资料，我们对此作出了初步的解释：

对于发达地区来说，由于金钱的边际效应，因此其对于发达地区的人们所带来的幸福感的提升逐渐消退，但是虽然金钱对于我们幸福感的影响减弱了，我们仍旧要处于职场之中工作去获得生存所需要的金钱，而这往往带来了高发的焦虑症和情绪失控。巨大的职场压力和高强度的工作量使得人们逐渐对工作产生不满，当这种不满情绪不断发酵时，它就会向你人生的各个方向去蔓延，最终它使得人们对自己的生活充满厌恶，幸福感自然大大降低。（参考 https://hbr.org/2017/09/happiness-traps，文中详细介绍了美国已经经历了若干年的“幸福陷阱”）。

对于发展中国家来说，众多因素都会导致生活幸福水平的降低，特别是非洲中仍旧处于战乱动荡的国家。政局的动荡不安，社会的极度不稳定，过于夸张的贫富差距导致了非洲的这些国家普遍的幸福指数偏低。对于非洲的众多国家来说，解决社会中的若干政治问题才是首要考虑的事情，相比与此，“幸福”就显得过于单调和空洞，而这自然导致了人们对于幸福感的丧失。实际的幸福指数较低也便能够得到解释。

国家或地区	预测分数	实际分数
新加坡	7.491	6.262
香港	6.891	5.43
斯里兰卡	5.842	4.366
缅甸	5.57	4.36
印度	5.08	4.015
博兹瓦纳	5.44	3.488
卢旺达	5	3.334
坦桑尼亚	4.73	3.231

图 4：预测与实际偏差过大的国家

结语

为人民服务是我们党坚定不移的口号和行动指南，自然为人民谋求幸福也是我们党我们国家所恪守的原则和底线。从 1949 年建国至今，中国政府不断提出新政策新方略为人民谋取幸福。习大大更是喊出了以中国人民过上美好生活为奋斗目标的口号。因此采取怎样的发展策略才能尽可能多地提高人民的幸福水平，是极为重要的。

而通过对幸福指数模型的构建，能够对于影响幸福的关键要素有一个初步的认识，并且通过实际的考察和推演，我们能够结合模型给出提高国民幸福水平的方式方法，这对于每一个国家来说，都是至关重要的。