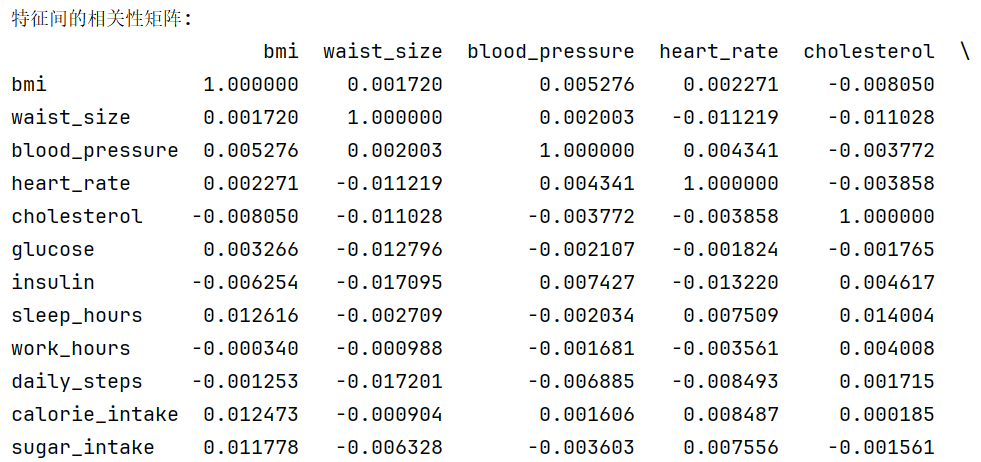
**社会计算个人作业**

**1.数据介绍**

报告数据集来源于Kaggle的《Disease Risk from Daily Habits》，包含100,000人的生活方式、生物识别、行为及人口统计学指标，共40个特征与1个目标值（健康或患病）。考虑到计算性能，选取前20,000条数据进行分析，经去空值处理后保留13,245条，数据完整性较高，无需缺失值填充，减少了预处理误差，为后续分析奠定了可靠基础。

数据特征分为连续型与离散型两类，二者在分析方法上存在显著差异：连续型变量（如BMI、血压、血糖等）可通过均值、标准差、相关性分析等探究数值间的线性关联及分布特征，而离散型变量（如性别、职业、饮食类型等）更适合采用频数统计、占比分析及分类比较，以揭示不同类别间的分布差异。

首先对15个连续变量进行相关性分析（以下为部分打印截图），显示所有特征间的相关系数绝对值均小于0.5（绝大多数小于0.03），远低于0.3的弱相关阈值，表明变量间几乎无线性关联。这一结果优势显著，有效避免了多重共线性问题，确保线性模型（如逻辑回归）中系数估计的稳定性；但同时也带来潜在挑战，即单一特征对其他特征的解释能力有限，后续建模可能需通过特征交互或非线性模型捕捉复杂关系。



基于探索特征冗余与相关性的目标，本研究筛选出25个核心特征，涵盖五大维度：

|  |  |  |
| --- | --- | --- |
| 维度 | 特征列表 | 数据类型 |
| 基础信息 | gender，age，education\_level，job\_type，occupation | 分类（部分数值化） |
| 生理数值 | bmi，waist\_size，blood\_pressure，heart\_rate，cholesterol，glucose，insulin | 连续数值 |
| 生活行为 | sleep\_hours，work\_hours，screen\_time，diet\_type，exercise\_type | 连续与分类 |
| 心理状态 | stress\_level，mental\_health\_score | 离散数值 |
| 生活环境与条件 | healthcare\_access，insurance，sunlight\_exposure，family\_history，pet\_owner | 分类 |

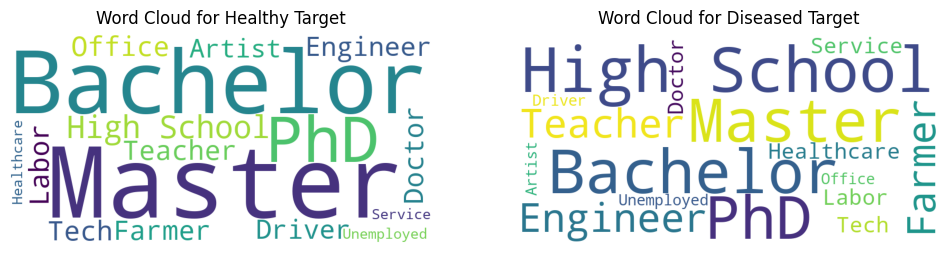
这些维度从个体到群体、从生理到心理、从当下状态到长期环境，构建了一个多视角、多层次的健康分析框架。它们既各自承载着独特的分析价值——基础信息定群体、生理数值判机能、生活行为找抓手、心理状态补全貌、环境条件看背景；又相互印证、互为补充，让健康评估从碎片化的指标堆砌，升级为系统性的规律挖掘，为健康管理、疾病预防和政策设计提供了科学且全面的决策依据。

**2.数据分析**

2．1基础信息分析

从基础信息维度来看，性别、年龄、教育水平等特征是人口统计学的核心要素，为健康分析提供了“群体锚点”，人口统计学特征分析显示，样本中男女比例为6564:6681，接近1:1，这种均衡性可能源于抽样过程的随机性，有助于减少性别偏差对疾病风险分析的干扰。年龄分布区间为18-90岁，参考WHO国际健康标准划分为三个阶段：18-44岁（青年阶段A）、45-59岁（中年阶段B）、60-89岁（老年阶段C），各阶段人数分别为5841、3082、4322，比例较为均匀，覆盖了各年龄段人群，为研究不同人生阶段的疾病风险差异提供了充足样本支撑。

而通过对比健康与患病群体在education\_level，job\_type，occupation上的词云分布，可直观洞察两类群体的社会特征差异。



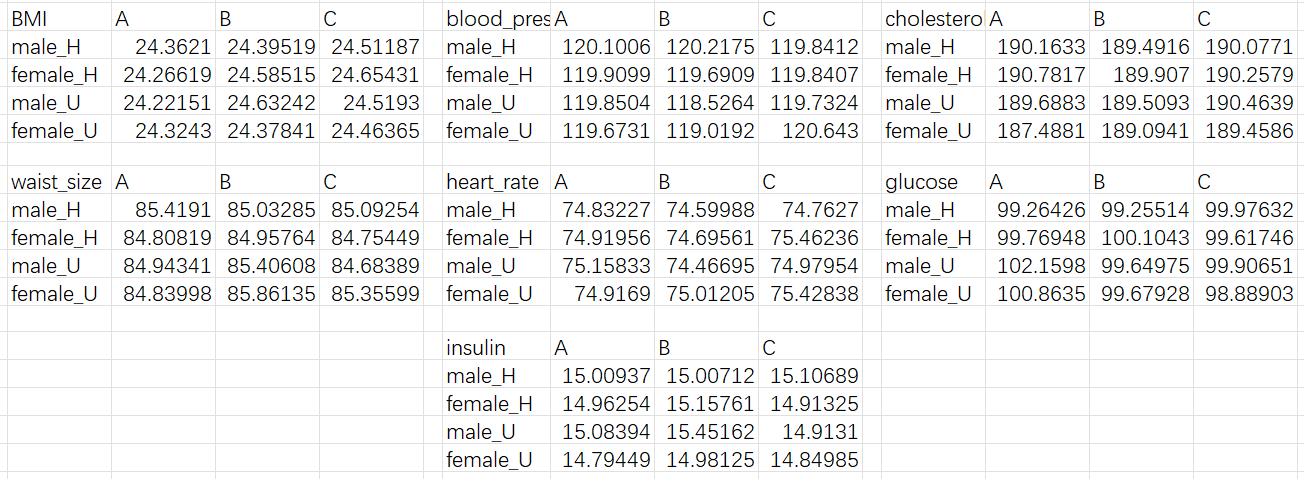
从教育维度看，健康群体的词云以“Master”和“Bachelor”为居多，表明高学历个体在健康样本中占比更高；而患病群体的词云的“High School”更为突出，低学历特征更为集中。

职业维度的差异同样鲜明：健康样本中，“Engineer”、“Artist”，“Teacher”等知识型、技术型职业占据词云核心，这类职业通常伴随相对稳定的工作节奏与收入保障，利于规律作息、健康行为的维持；患病样本的词云则以“Service”、“Farmer”、“Unemployed”为主导——服务行业的高强度、碎片化工作，农民的体力劳动与户外环境暴露，以及失业带来的经济压力和心理负担，均可能通过扰乱生活节律、削弱健康管理能力，间接提升疾病发生概率。

此外，值得注意的是，教育与职业存在潜在的交叉关联（如高学历更易进入技术/专业领域），这种社会经济地位的分层，进一步通过资源获取、生活方式等路径，强化了健康与患病群体的特征分野。词云的视觉差异背后，实则折射出社会环境对健康状态的深刻影响，为后续探究疾病风险的社会决定因素提供了直观线索。

2.2生理数值分析

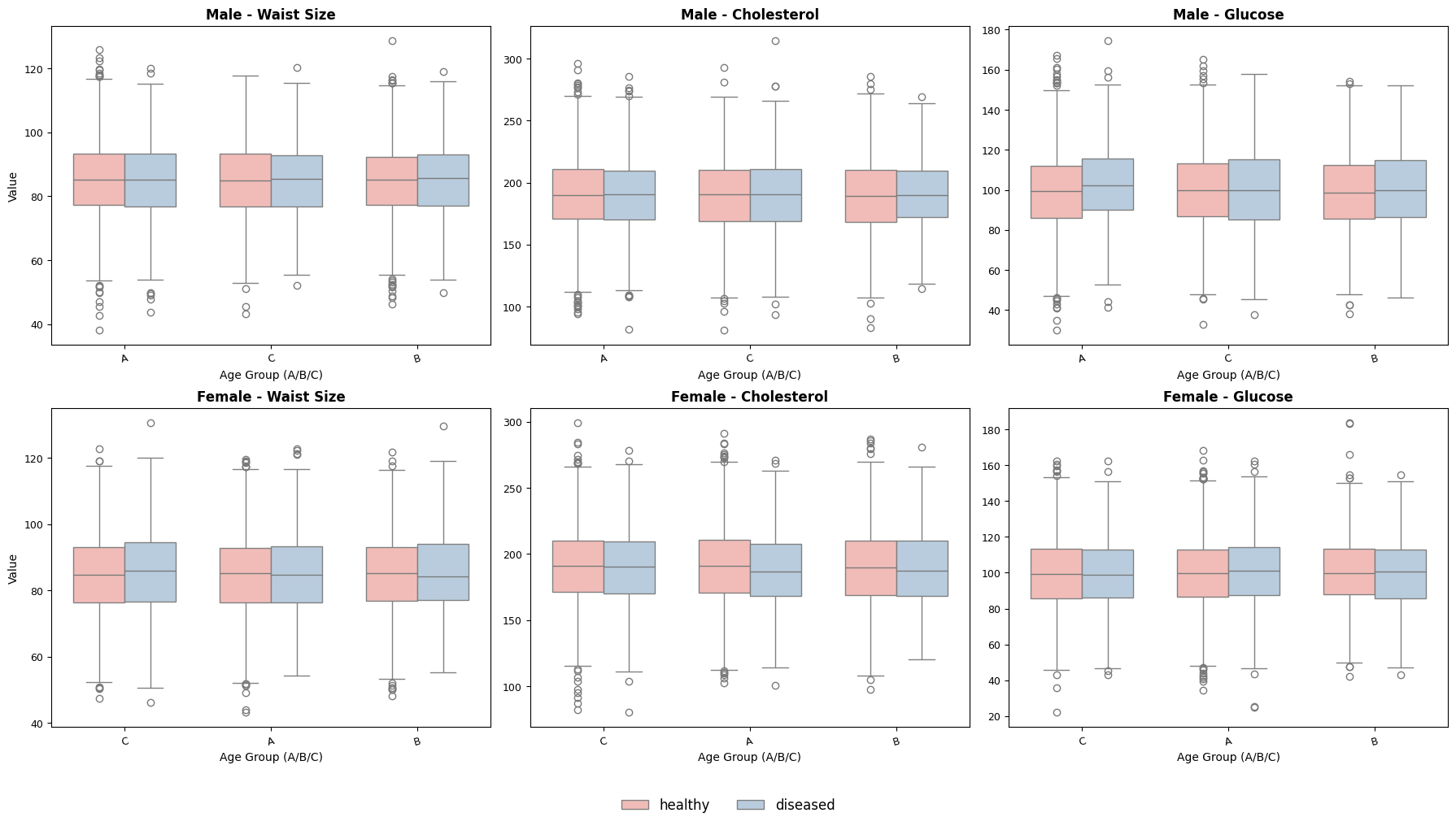
生理数值维度的设置则扎根于医学实证：血压、血糖等指标是经过百年医学验证的 “健康晴雨表”。这些客观量化的生理数据直接反映身体机能的运行状态，其优势在于摒弃主观判断，用数据锚定健康异常的“临界点”，既为疾病早期预警提供了硬指标，也让健康评估具备了可比较、可追踪的科学性。而不同性别在生理机能（如激素水平）、不同年龄在身体代谢（如衰老速度）的差异，是健康状态分化的基本特征，所以根据不同性别与年龄的健康状态初步计算均值，得到下表：



可见并非所有指标或所有分组都能有效区分健康状态。

bmi,blood\_pressure,heart\_rate,insulin的整体区分度极低，几乎无分层差异。可能因为BMI仅反映身高与体重的比值，无法区分脂肪与肌肉比例（如肌肉量高的人群BMI可能偏高但健康）；血压，胰岛素含量，及静息心率的个体差异较大且具短期波动，受瞬时状态（如测量时是否紧张）影响难以反映长期健康状态。

而waist\_size，cholesterol，glucose具有一定区分度，为了更直观呈现数据的分布特征与离散程度，基于这三个核心变量再绘制具备中位数和四分位数的箱线图:



waist\_size，在女性中年阶段（B）区分度显著，健康组（H=84.97）与不健康组（U=85.84）均值差异接近0.9；可能因为中年女性（45-59岁）多处于绝经前后，雌激素水平下降可能导致腹部脂肪堆积加速，而腰围作为内脏脂肪的指标，与代谢异常（如糖尿病、高血脂）的关联在该阶段更显著，因此不健康群体的腰围差异更突出。

cholesterol在女性青年阶段（A）区分度显著，健康组（H=190.67）与不健康组（U=187.66）均值差异约3；可能因为青年女性（18-44岁）受雌激素保护，健康群体的HDL（“好胆固醇”）水平通常较高，会拉高总胆固醇均值；而不健康群体可能因雌激素水平异常（如多囊卵巢综合征）或不良生活习惯导致HDL偏低，总胆固醇随之下降，因此出现健康组胆固醇更高的差异。

glucose在男性青年阶段（A）区分度显著，健康组（H=99.26）与不健康组（U=102.16）均值差异接近3，在其他分组差异多小于1，甚至反向。可以推出青年男性（18-44岁）若存在不健康状态，更可能与不良生活习惯（如高糖饮食、久坐）相关，直接导致空腹血糖升高；而中年或女性群体的不健康状态可能由更复杂因素（如胰岛素抵抗、激素波动）引起，血糖波动未必是核心特征，因此区分度较弱。

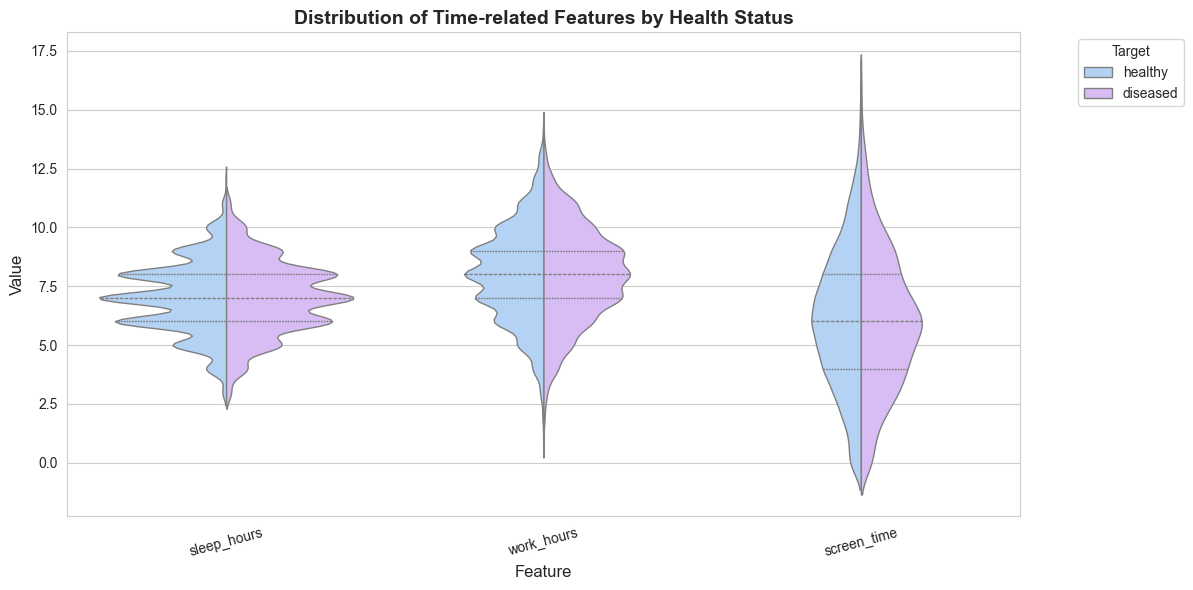
2.3生活行为分析

生活行为维度聚焦于个体可控的健康变量，睡眠、饮食、运动等习惯是健康状态的 “调节阀”。现代流行病学研究早已证实，长期睡眠不足会升高血压、久坐行为会增加肥胖风险，这些特征的纳入既符合 “健康源于习惯” 的常识，又为健康干预提供了可操作的切入点——相比不可改变的生理基础，调整生活行为是更易实现的健康改善路径，体现了从“被动治疗”到“主动预防”的转型价值。

**（1）时间分配**

从这张时间相关特征的小提琴图中，可观察到健康与患病群体在睡眠、工作、屏幕时间上的分布特点。

小提琴图既有分位数等统计信息，又通过对称的“琴形”展示数据在不同取值处的概率密度，能同时呈现数据的分布形状、集中趋势及尾部特征，在样本量较大时更易捕捉数据分布的细微差异。（考虑多小数点时核计算复杂性对时间进行了取整处理，在文件hour.csv）



睡眠时长（sleep\_hours）：两者中位数接近7-8小时，但健康群体的分布更“宽厚”，中低区间（5-7小时）的密度更高，暗示健康人群睡眠时长的个体差异更大（有人习惯短睡眠，有人需长睡眠，却均维持健康）；患病群体分布更紧凑，可能因疾病压缩了睡眠模式的多样性（如长期失眠或嗜睡）。

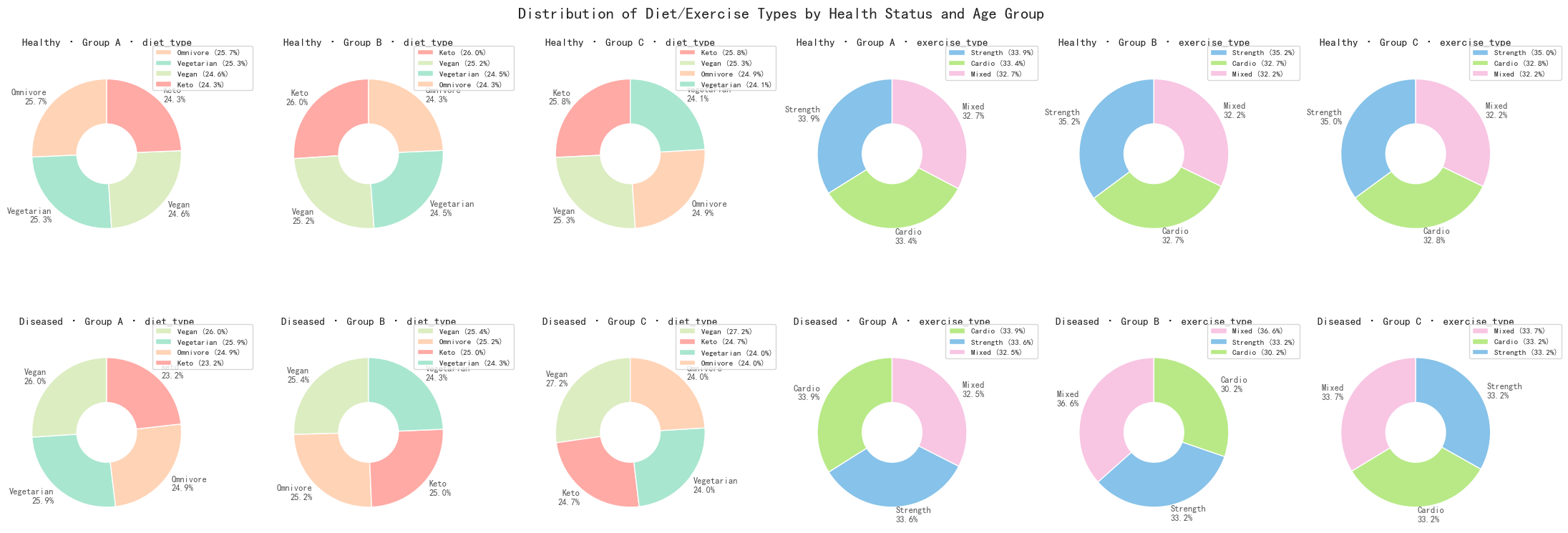
工作时长（work\_hours）：中位数虽也重合，但患病群体的“长尾”更显著——高值区（>10小时）和低值区（<2小时）的分布更宽，反映出患病群体中“极端工作时长”更普遍：要么因高强度工作累出疾病，要么因疾病无法维持常规工作节奏（如休假、失业）；健康群体则更集中于中间区间，体现更稳定的工作节律。

屏幕时间（screen\_time）：健康群体在中低值区（4-8小时）更密集，患病群体在高值区（8-12小时）的“鼓包”更明显。这可能源于健康人群对屏幕时间的主动管理，而患病群体因康复期活动受限（如卧床刷手机）、社交减少（依赖屏幕娱乐），被动增加了屏幕使用。

整体而言，时间特征的“分布模式”比绝对值更具区分度：健康群体的时间分配更具灵活性（睡眠、工作时长波动空间大），患病群体则趋向极端化或集中化。这种差异提示，时间管理的“多样性”或许是健康的隐性指标——而非单一指标的高低，更需关注个体时间分配是否陷入极端（如超长工作、沉迷屏幕）。

**（2）饮食运动习惯**

按“健康状态+年龄组”拆分的饮食与运动类型饼图，揭示了生活方式选择与健康状态在年龄维度下的隐性关联。



从饮食分布看，青年群体（A）的健康与患病者在素食、杂食、生酮等类型上差异微弱，暗示青年阶段饮食对健康的直接区分度有限；但中年群体（B）中，健康者的地中海饮食占比（24.5%）显著高于患病者的生酮饮食（25%），折射出中年健康群体更倾向通过“均衡营养（高膳食纤维、优质脂肪）”维持状态，而患病中年或因疾病管理选择生酮却未收获预期效益。中老年群体（C）中，健康者的素食占比（25%）略高于患病者（24%），结合代谢衰老规律，素食的低脂肪、高纤维特性或许更适配中老年健康维护。

运动类型的差异则更具年龄特异性：青年健康群体（A）的运动分布更均衡（力量训练28%、 cardio27%、 身心训练25%），患病青年却侧重力量训练（31%），反映健康青年通过“全面运动”维持体能，而患病青年因代谢问题（如肥胖）更依赖力量减脂；中年健康群体（B）的Cardio占比（32.3%）远高于患病群体（26.2%），暗示有氧运动对中年心肺功能的保护作用，患病中年则因体力下降被迫减少此类活动；中老年健康群体（C）的Cardio占比（32%）显著高于患病群体（26%），进一步印证有氧运动对中老年血管弹性、慢性病风险的改善价值——患病中老年或因运动能力下降，难以维持规律cardio。

这些规律的核心在于，健康与患病的生活方式分野，是“年龄+健康状态+行为类型”的动态交互结果：青年依赖运动多样性，中年需兼顾饮食均衡与有氧坚持，中老年更需强化素食和cardio的代谢保护。这种细分逻辑，为设计“年龄-健康状态适配”的干预方案提供了精准依据，如向中年患病群体推广地中海饮食、为青年患病群体补充身心训练。

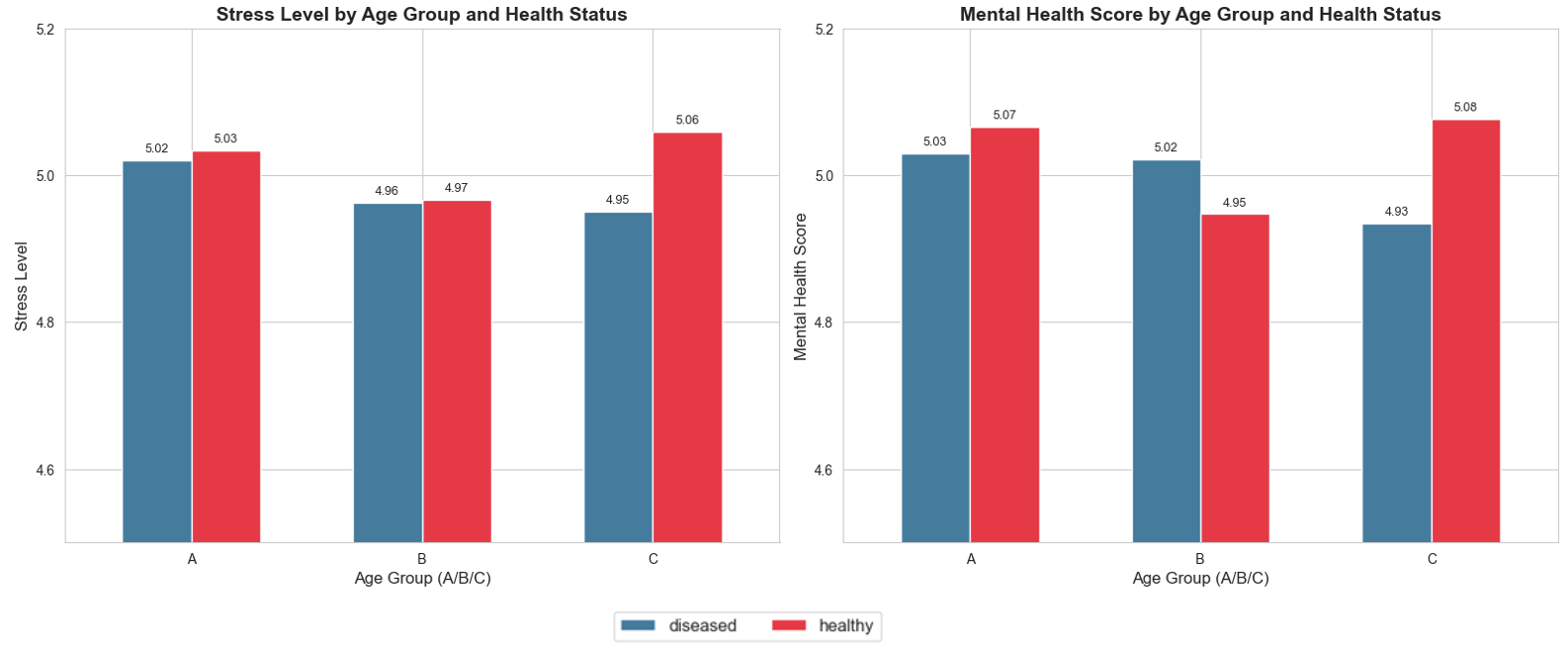
2.4心理状态分析

心理状态维度的加入则打破了 “生理健康至上” 的传统视角，呼应了现代健康观中 “身心合一” 的理念。压力水平过高会通过神经内分泌系统影响免疫力，心理健康分数与慢性病康复速度的关联也被多项研究证实。这一维度的优势在于填补了纯生理指标的空白，让健康评估更贴近人的整体性。

从压力水平与心理健康分数的对比中，可窥见：

压力维度上，健康群体在各年龄组的压力值均微高于患病群体，可见压力并不一定是健康的敌人，适度的压力能够让生活更加规律对生活状态的更敏锐感知。

心理健康分数的分布则更显年龄特异性：健康青年分数最高（5.07），契合该阶段心理弹性强、社会支持充足的特征；中年患病群体分数骤降至4.95，远低于同组健康者（5.02），折射出“事业黄金期+疾病冲击”的双重压力对心理状态的重创；中老年健康群体分数回升至5.05，或因退休后生活节奏调整（如兴趣培养、社交重构）缓冲了年龄焦虑，而患病中老年因疾病长期负担，分数垫底（4.93）。



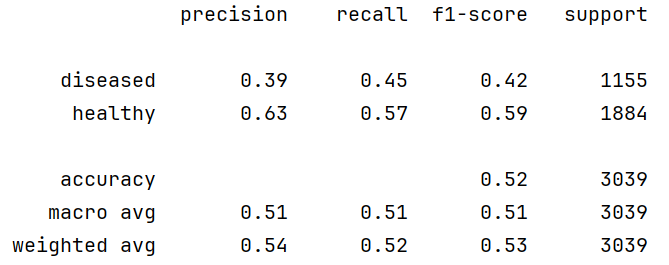
这些细微差异的核心启示是：心理指标的解读需突破“患病=心理差”的惯性认知——健康群体的“高压力”可能是积极生活的副产品，患病群体的“低压力”实则是感知钝化的表现；而心理健康分数的年龄波动，更暴露了不同人生阶段患病后的心理脆弱点（如中年职场病患者）。这种交互规律，为心理干预提供了精准方向：针对中年患病群体强化心理支持，为中老年健康群体设计“防失能”心理维护方案，而非对所有患病者一概而论。

2.5生活环境与条件分析

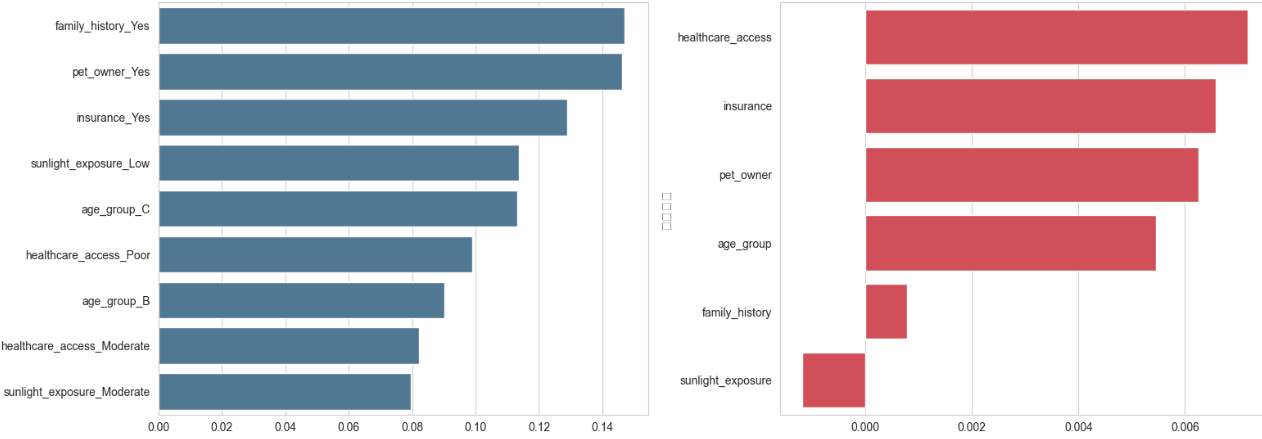
生活环境与条件维度则将健康分析从个体延伸到社会层面，医疗资源可及性、保险覆盖、家族病史等特征，解释了健康差异中的结构性因素。比如，缺乏基础医疗保障的人群更易因延误治疗加重病情，有家族遗传病历史的个体需要更早的筛查干预。这一维度的合理性在于承认健康不仅是个人选择的结果，还受社会资源分配、遗传背景等外部条件制约，为政策制定（如优化基层医疗资源）和高风险人群筛查提供了宏观依据。

由于这些数据都是字符型分类数据，所以先进行独热编码变为离散数据，想到离散数据的特性，所以尝试创建决策树（由于target里healthy：diseased≈2:1，在训练时会出现机器计算偷懒都打healthy标签的情况，所以第二次构建决策树时进行了diseased的加权并再进行了一次数据抽取，使用try\_class.csv文件）

但是即使经过调整但是基于生活环境与条件的模型对患病群体的识别仍存在瓶颈，还是需要加上其他特征。



然而即便如此，还是可以得到特征重要性（左图）与风险组合分析。



家族病史（family\_history\_Yes）、养宠（pet\_owner\_Yes）、保险（insurance\_Yes） 等离散特征的高贡献度，印证了遗传背景与社会支持对健康的直接影响；而医疗可及性的风险分布呈反向梯度”暗示样本量偏差或未观测因素的干扰，如自我健康管理能力。比较特别的一点是：青年群体本应具备更低基础风险，但若叠加 “医疗可及性差+日照不足+无保险” 等条件，患病比例可飙升至 51.7%（高风险组合），折射出模型对 “年轻但资源匮乏” 群体的风险刻画不足——这类群体的健康危机往往被 “年龄优势” 掩盖，却因长期医疗忽视加速疾病进展。

**3小结**

3.1研究意义

从健康数据中挖掘的规律是社会计算在公共卫生领域的生动实践，其意义既体现在对科学计算的创新应用，更扎根于对社会现象的深度解码。

从计算方法层面看，健康与患病群体在时间分配、饮食运动、心理状态上的细微差异（如中年女性腰围的分布特征、青年男性血糖的区分度），本质是海量异构数据中的“隐性模式”——这些模式无法通过单一指标或简单统计捕捉，而需依赖机器学习、可视化工具等社会计算手段，将分散的生理指标、行为数据、人口学特征编织成可解释的关联网络。这正呼应了社会计算的核心价值：用大规模计算能力破解复杂社会现象的数据迷雾，让健康差异从“经验判断”升级为“数据实证”。

社会计算搭建“数据-理论-实践”的桥梁：通过计算方法提炼的健康规律为社会科学提供新的研究命题；而基于社会理论的解读，将数据规律转化为可操作的政策建议，如针对中年群体的时间管理干预、青年群体的屏幕使用引导，让健康分析既扎根于数据的客观性，又服务于社会的现实需求。

3.2基于分析的政策与生活建议

（1）个人健康管理：

中年女性可定期监测腰围，警惕绝经后腹部脂肪堆积；青年男性应关注空腹血糖，规避高糖饮食与久坐的影响；

在时间分配上，尽量避免工作、屏幕时间极端化；

在运动方式上，青年尽力补全有氧、身心训练类型，中年可多尝试力量与有氧结合；

在饮食上，中年可以探索地中海饮食的平价实践，中老年要注重素食中搭配优质蛋白。

（2）社区辅助：

社区应聚焦资源衔接与精准服务，如摸排“年轻但医疗可及性差、无保险”群体，发放含维生素D、免费体检券的健康包，定期推送代谢指标监测提醒；

开设分层健康课——为青年设计“运动多样性训练指南”，为中年讲解“地中海饮食实操”，为中老年搭建书法、舞蹈兴趣班，既强化代谢保护，也缓解心理失能风险 。

（3）医疗系统：

医疗系统应延伸服务触角，在乡镇集市、工业园区设“代谢筛查站”，便捷检测腰围、血糖、胆固醇，覆盖农民与服务从业者；联动企业为中年职场病患者开通免费心理咨询通道，社区卫生中心定期为独居老人开展“心理防失能”随访 。

（4）政策层面：

政策层面需兜底保障，通过补贴推动服务行业弹性工时试点，打破“高强度、碎片化工作”对健康的侵蚀；统筹健康课程与筛查资源，将“年龄 - 健康适配”的干预方案纳入基层公共服务清单，让个体努力、社区支持、医疗服务与政策保障形成闭环，真正弥合社会分层与年龄差异带来的健康鸿沟。