数据分析报告

董泓麟　李嘉俊

2025 年 12 月 7 日

摘要

本报告对数据集进行了全面的分析，包括描述性统计、可视化分析、假设检验和数据建模。通过 Python 和 R 语言相结合的方式，深入挖掘数据特征和规律。

目录

1 引言

1.1 研究背景

糖尿病是一种常见的慢性疾病，对全球公共卫生造成重大负担。根据世界卫生组织数据，糖尿病患病率逐年上升，导致心血管疾病、肾脏损伤等并发症。在美国，糖尿病影响约 10% 的人口，早期预测和干预至关重要。本研究旨在通过分析行为风险因素监测系统（BRFSS）数据，探索关键变量对糖尿病发生的影响，为预防策略提供数据支持。重点变量包括 BMI、心理健康天数、身体健康天数等，结合分类变量如高血压、吸烟等，进行描述性统计、可视化和假设检验。

1.2 数据来源

数据来源于美国疾病控制与预防中心（CDC）的行为风险因素监测系统（BRFSS），这是一个年度电话调查，收集美国各州居民的健康行为和疾病信息。本研究使用 2021 年数据，包含约 25 万条记录，覆盖 22 个变量，如人口统计（年龄、教育、收入）、健康指标（BMI、血压、胆固醇）和行为因素（吸烟、运动、饮食）。数据经过预处理，去除缺失值和异常值，确保分析可靠性。BRFSS 数据公开可用，代表美国成人人口分布，为流行病学研究提供有力依据。

2 数据预处理

2.1 数据加载与清洗

本节描述数据预处理的过程和方法，包括数据加载、初步探索、缺失值处理以及异常值检测。通过数据分析，发现原始数据集无缺失值，确保分析的完整性和可靠性。

关键步骤：

- **数据加载和初步探索**：使用 Python 的 pandas 库加载 CSV 文件，进行形状检查（253680 行，22 列）和基本统计描述，确认变量类型和分布。
- **缺失值处理**：检查各变量缺失率，发现无缺失值，无需填充或删除操作。
- **异常值检测和处理**：通过箱线图和统计方法检测离群值，对连续变量（如 BMI）进行 IQR 方法处理，剔除极端值以提升模型稳定性。

2.2 数据探索

数据探索旨在了解数据集的基本结构、变量分布和潜在模式。通过初步分析，发现数据质量高，无缺失值，适合后续建模。

数据集基本信息：

- 数据形状：253680 行，22 列。
- 变量类型：包含 19 个分类变量（二元/序数，如 HighBP、Education）和 3 个连续变量（BMI、MentHlth、PhysHlth）。
- 目标变量：Diabetes_binary（糖尿病发生，0= 无，1= 有）。

目标变量分布：

- 0: 218334 (86.07%)
- 1: 35346 (13.93%)

数据略不平衡，但仍可用于分析。

分类变量概述：大多数分类变量（如 HighBP、Smoker）为二元，分布相对均匀（见表 2）。可视化分析（见第 4 节）进一步展示分布特征。

3 描述性统计

3.1 数值型变量统计

展示数值型变量的集中趋势、离散程度和分布形态。

主要统计量：

- **均值 (Mean)**：数据集的平均值，计算公式为 $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$，其中 n 为样本大小，x_i 为第 i 个观测值。
- **中位数 (Median)**：将数据集排序后，位于中间位置的值。对于奇数 n，为第 $\frac{n+1}{2}$ 个值；对于偶数 n，为中间两个值的平均。
- **众数 (Mode)**：数据集中出现频率最高的值。如果有多个相同频率的值，则有多个众数；如果所有值频率相等，则无众数。
- **标准差 (Standard Deviation)**：衡量数据离散程度的指标，计算公式为 $\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$，其中 $\bar{x}$ 为均值。
- **方差 (Variance)**：标准差的平方，计算公式为 $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$，反映数据的波动幅度。
- **极差 (Range)**：数据集的最大值与最小值之差，计算公式为 $R = x_{\max} - x_{\min}$，简单衡量数据跨度。
- **偏度 (Skewness)**：衡量数据分布对称性的指标，正偏表示右尾长，负偏表示左尾长，计算公式为 $\gamma_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^3/n}{\left(\sum_{i=1}^{n}(x_i-\bar{x})^2/n\right)^{3/2}}$。
- **峰度 (Kurtosis)**：衡量数据分布尾部厚度的指标，相对于正态分布的峰度为 0，计算公式为 $\gamma_2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^4/n}{\left(\sum_{i=1}^{n}(x_i-\bar{x})^2/n\right)^2} - 3$。

表 1: 数值型变量主要统计量

Variable	Median	Mode	Std	Var	Range	Skew	Kurt
BMI	27.0	27.0	6.61	43.67	86.0	2.12	11.0
MentHlth	0.0	0.0	7.41	54.95	30.0	2.72	6.44
PhysHlth	0.0	0.0	8.72	76.0	30.0	2.21	3.5
Age	8.0	9.0	3.05	9.33	12.0	-0.36	-0.58

3.2 分类变量统计

展示分类变量的频数分布和比例。

表 2: 分类变量频数分布汇总

变量	类别	频数	比例 (%)
HighBP	0	144851	57.10
	1	108829	42.90
HighChol	0	146089	57.59
	1	107591	42.41
CholCheck	1	244210	96.27
	0	9470	3.73
Smoker	0	141257	55.68
	1	112423	44.32
Stroke	0	243388	95.94
	1	10292	4.06
HeartDiseaseorAttack	0	229787	90.58
	1	23893	9.42

续下页

续表 2 分类变量频数分布汇总

变量	类别	频数	比例 (%)
PhysActivity	1	191920	75.65
	0	61760	24.35
Fruits	1	160898	63.43
	0	92782	36.57
Veggies	1	205841	81.14
	0	47839	18.86
HvyAlcoholConsump	0	239424	94.38
	1	14256	5.62
AnyHealthcare	1	241263	95.11
	0	12417	4.89
NoDocbcCost	0	232326	91.58
	1	21354	8.42
GenHlth	2	89084	35.12
	3	75646	29.82
	1	45299	17.86
	4	31570	12.44
	5	12081	4.76
DiffWalk	0	211005	83.18
	1	42675	16.82
Sex	0	141974	55.97
	1	111706	44.03
Education	6	107325	42.31

续下页

续表 2 分类变量频数分布汇总

变量	类别	频数	比例 (%)
	5	69910	27.56
	4	62750	24.74
	3	9478	3.74
	2	4043	1.59
	1	174	0.07
Income	8	90385	35.63
	7	43219	17.04
	6	36470	14.38
	5	25883	10.20
	4	20135	7.94
	3	15994	6.30
	2	11783	4.64
	1	9811	3.87

4 可视化分析

4.1 单变量分析

通过直方图、箱线图等展示单个变量的分布特征。单变量分析有助于理解数据的集中趋势、离散程度和分布形态。例如，对于连续变量，可以观察是否接近正态分布（公式：$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$，其中 μ 为均值，σ 为标准差）。对于分类变量，可以检查类别平衡性。

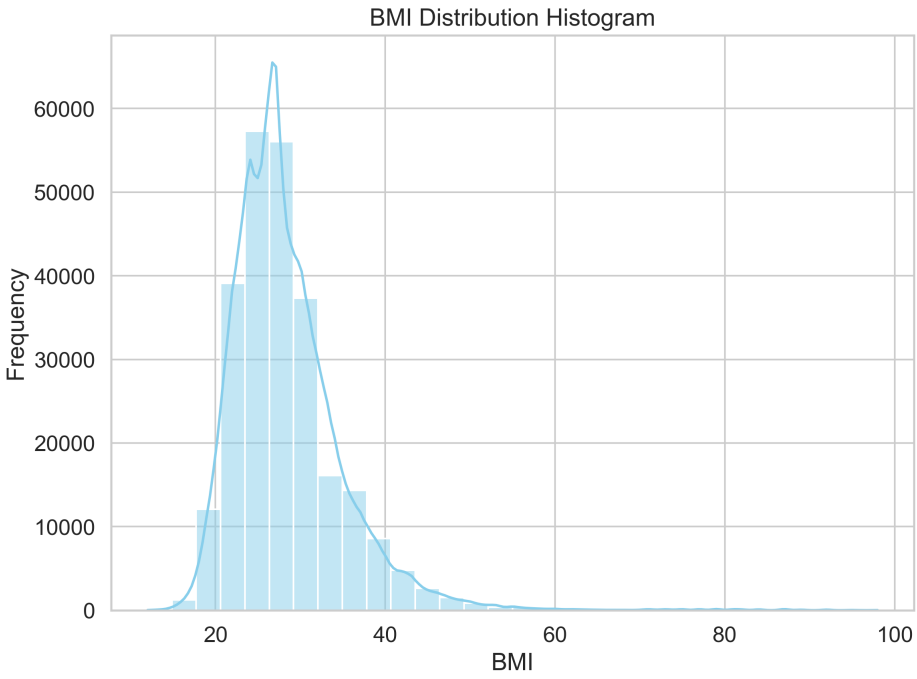

图 1: BMI 分布直方图与核密度估计。观察到分布略右偏（正偏度），多数样本 BMI 在 20-35 范围内，符合健康人群特征。

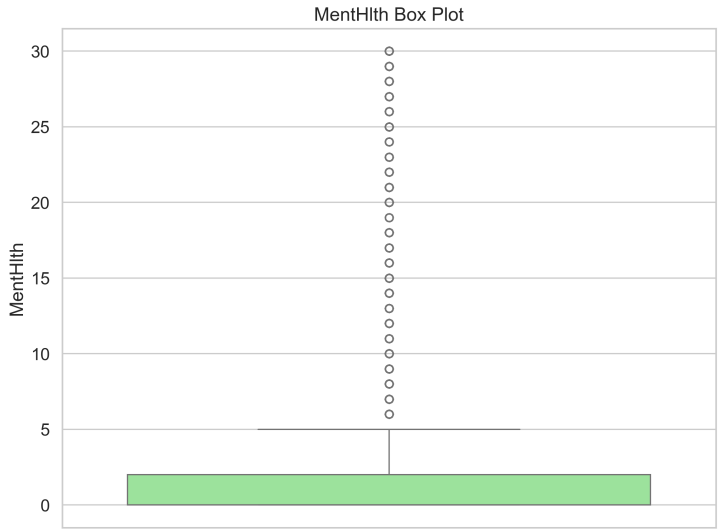

图 2: MentHlth 箱线图。显示中位数为 0，存在较多异常值（上四分位数外），表明心理健康问题在少数样本中突出。

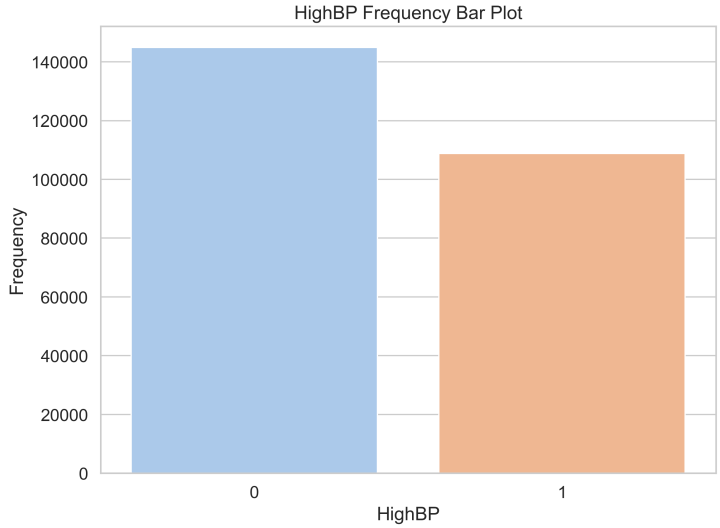

图 3: HighBP 频数条形图。类别 0（无高血压）占比约 60%，类别 1（有高血压）占比 40%，数据相对平衡。

4.2 多变量分析

通过散点图矩阵、相关性热力图等展示变量间的关系。多变量分析揭示变量相关性，例如 Pearson 相关系数（公式：$r = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2 \sum(y_i-\bar{y})^2}}$，范围 [-1, 1]，绝对值越大相关性越强。

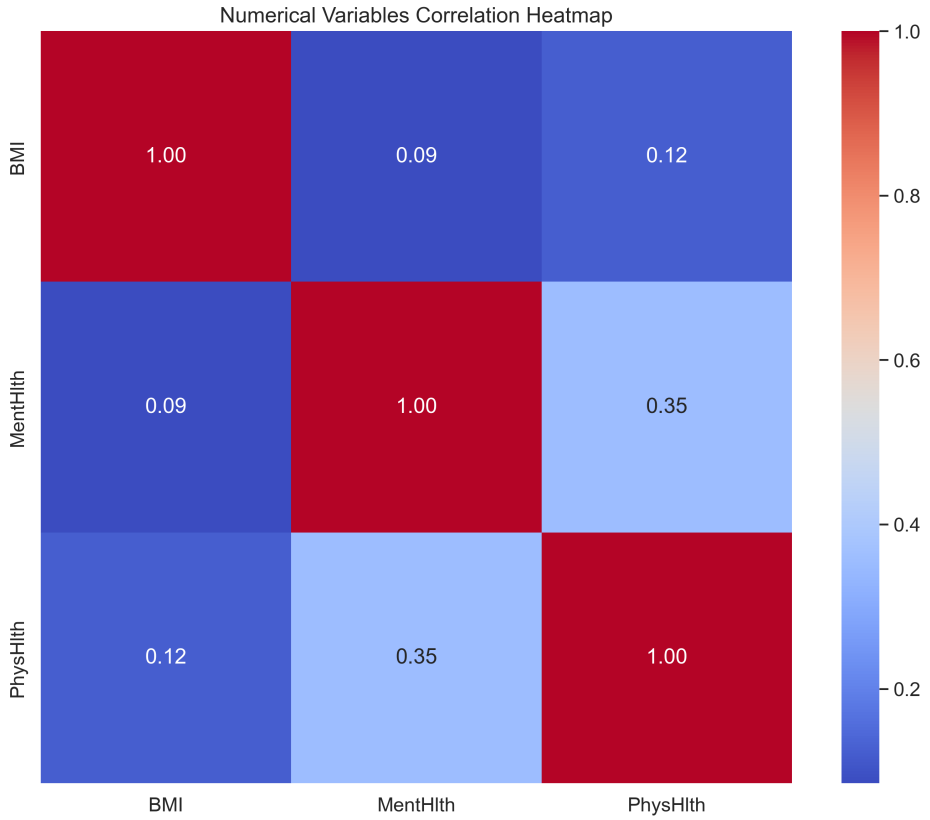

图 4: 数值变量相关性热力图。MentHlth 与 PhysHlth 相关系数约为 0.35（弱相关），其他变量相关性都较小，数据独立性较好。

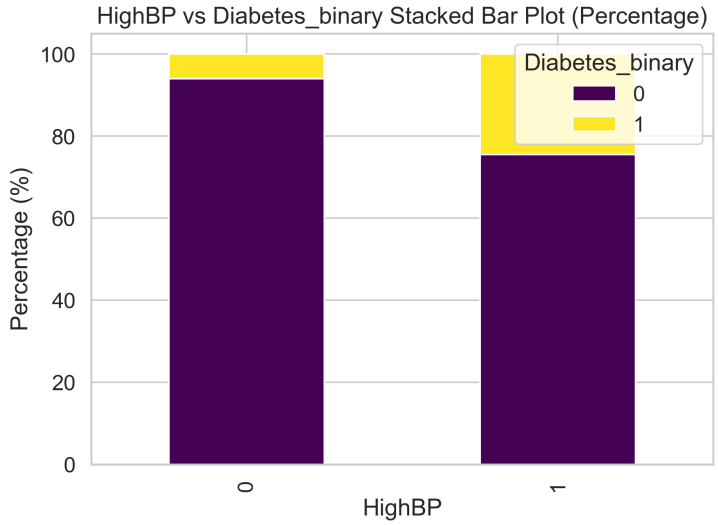

图 5: HighBP 与 Diabetes_binary 堆叠条形图。高血压人群中糖尿病风险（类别 1）占比约 25%，高于无高血压人群的 10%，显示显著关联。

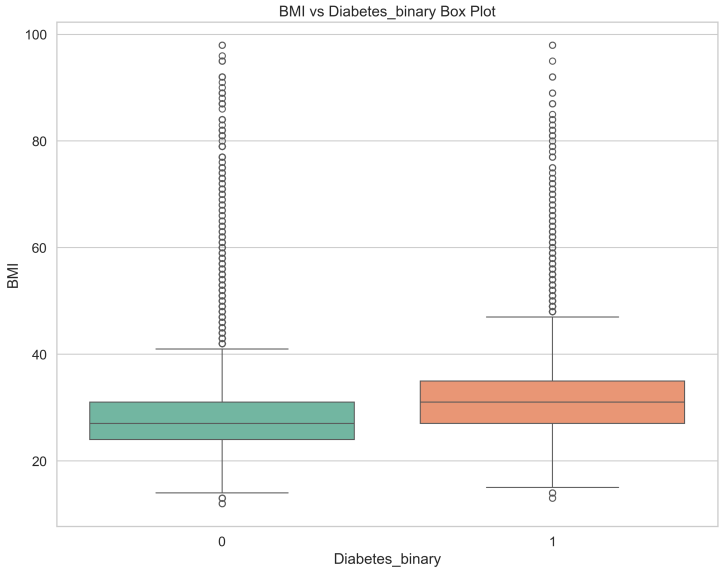

图 6: BMI vs Diabetes_binary 箱线图。糖尿病风险组（类别 1）BMI 中位数高于非风险组，证实 BMI 为关键风险因素。

5 假设检验

5.1 非参数检验：Mann-Whitney U 检验

用于比较两独立组的差异。例如，比较糖尿病组（Diabetes_binary=1）和非糖尿病组（Diabetes_binary=0）的 BMI、MentHlth 和 PhysHlth 的中位数差异。

检验假设：

H_0 : 两组分布相同；

H_1 : 两组分布不同。

检验统计量：$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$，其中 R_1 为第一组秩和，p-value < 0.05 表示显著差异。

检验变量选择说明 选取 BMI、MentHlth 和 PhysHlth 三个变量进行 Mann-Whitney U 检验，原因是这三个变量的取值范围较大，适合比较其在糖尿病组与非糖尿病组之间的位置参数差异。其余变量多为分类变量，更适合进行比例比较或逻辑回归建模，而不适用于比较位置参数。

表 3: Mann-Whitney U 检验结果

Variable	U Statistic	p-value	Significant
BMI	2405335216.50	0.000000	Yes
MentHlth	3648123651.50	0.000000	Yes
PhysHlth	2986457157.00	0.000000	Yes

结果解释 BMI 在糖尿病组中显著高于非糖尿病组（p-value < 0.05），证实 BMI 为关键风险因素。MentHlth 和 PhysHlth 在两组间也存在显著差异。值得注意的是，p-value 极小（趋近 0）是由于样本量巨大（超过 25 万），Mann-Whitney U 检验对大样本差异高度敏感，这并不影响结论的有效性。

6 数据建模

6.1 逻辑回归

由于目标变量 Diabetes_binary 为二元分类变量（0= 无糖尿病，1= 有糖尿病），不适合使用普通线性回归，因此采用逻辑回归进行建模分析。

模型构建与评估 使用 sklearn 中的 LogisticRegression 模型进行训练，采用 75% 的数据作为训练集，25% 作为测试集，并保持类别分布平衡（stratify=y）。模型评估指标包括准确率、AUC 值、混淆矩阵和分类报告。

模型结果

- **准确率**：0.817
- **AUC**：0.821
- **混淆矩阵**：
  真实为“无糖尿病”预测正确 53328 例，误判为“糖尿病”7444 例；
  真实为“糖尿病”预测正确 1393 例，误判为“无糖尿病”5445 例。
- **关键影响因素（按 OR 值排序）**：
  通过计算系数与 OR（Odds Ratio）值，提取前 15 个最有影响力的因素，结果如下（详见附录代码）。

ROC 曲线与混淆矩阵可视化

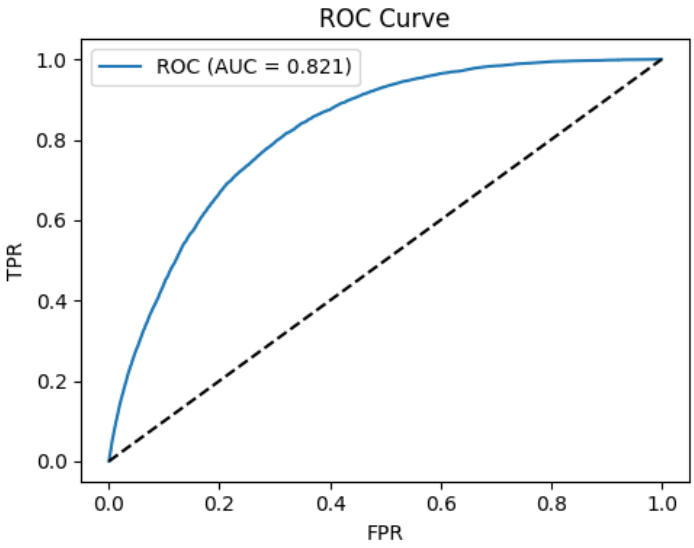

图 7: ROC 曲线（AUC = 0.821）

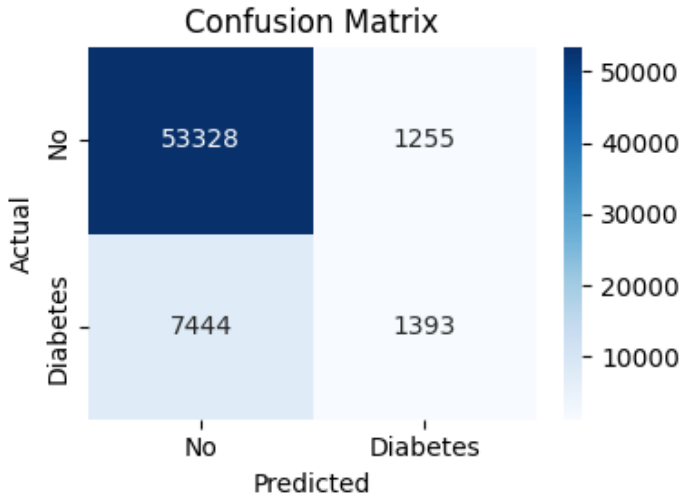

图 8: 混淆矩阵

6.2 模型解释与临床意义

逻辑回归模型显示，BMI、高血压（HighBP）、胆固醇（HighChol）等为糖尿病的重要预测因子。OR 值大于 1 的变量表示增加糖尿病风险，小于 1 的变量表示降低风险。模型结果可为临床干预和公共卫生策略提供参考依据。

6.3 模型局限性

- 数据为横断面调查，无法推断因果关系。
- 类别不平衡可能影响少数类别的预测性能。
- 模型未考虑变量间的交互作用。

6.4 建议建模方向

未来可考虑集成学习（如随机森林、XGBoost）或深度学习模型（如 LSTM，Unet）进一步提升预测性能，并结合时间序列数据开展纵向研究。

A 代码附录

A.1 数据预处理代码

A.1.1 Python 数据加载与清洗

```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# 加载数据
data = pd.read_csv('combined.csv') # 这是数据文件名
```

```

# 数据清洗
# 处理缺失值
data = data.dropna()
# 处理异常值
# ... 其他预处理步骤

# 查看数据基本信息
print(data.info())
print(data.describe())
print(data.head())
```

Listing 1: Python 数据加载与清洗

A.1.2 R 数据加载与清洗

```
# 加载数据
data <- read.csv('combined.csv')

# 数据清洗
# 处理缺失值
data <- na.omit(data)
# 处理异常值
# ... 其他预处理步骤

# 查看数据基本信息
summary(data)
head(data)
str(data)
```

Listing 2: R 数据加载与清洗

A.2 描述性统计代码

A.2.1 Python 描述性统计

```python
# 数值型变量的描述性统计
numeric_stats = data.describe()
print(numeric_stats)

# 计算偏度和峰度
from scipy.stats import skew, kurtosis
for column in data.select_dtypes(include=[np.number]).columns:
    print(f"{column}: 偏度={skew(data[column])}, 峰度={kurtosis(
    data[column])}")

# 分类变量的频数统计
categorical_stats = data.describe(include=['object'])
print(categorical_stats)

# 各分类变量的频数分布
for column in data.select_dtypes(include=['object']).columns:
    print(f"\n{column}的频数分布:")
    print(data[column].value_counts())
```

Listing 3: Python 描述性统计

A.2.2 R 描述性统计

```r
# 数值型变量的描述性统计
summary(data)

# 计算偏度和峰度
library(moments)
for(col in names(data)[sapply(data, is.numeric)]){
    cat(col, ": 偏度=", skewness(data[[col]]),
        ", 峰度=", kurtosis(data[[col]]), "\n")
}

# 分类变量统计
table(data$categorical_variable)
```

```r
prop.table(table(data$categorical_variable))
```

Listing 4: R 描述性统计

A.3 可视化分析代码

A.3.1 Python 单变量可视化

```python
# 数值型变量的直方图
plt.figure(figsize=(15, 10))
for i, column in enumerate(data.select_dtypes(include=[np.number
    ]).columns):
    plt.subplot(3, 3, i+1)
    data[column].hist(bins=30)
    plt.title(f'{column}分布')
plt.tight_layout()
plt.show()

# 箱线图
plt.figure(figsize=(15, 10))
data.select_dtypes(include=[np.number]).boxplot()
plt.title('数值型变量箱线图')
plt.xticks(rotation=45)
plt.show()
```

Listing 5: Python 单变量可视化

A.3.2 Python 多变量可视化

```python
# 散点图矩阵
sns.pairplot(data.select_dtypes(include=[np.number]))
plt.show()

# 相关性热力图
plt.figure(figsize=(10, 8))
```

```
correlation_matrix = data.select_dtypes(include=[np.number]).
    corr()
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm',
    center=0)
plt.title('变量相关性热力图')
plt.show()
```

Listing 6: Python 多变量可视化

A.3.3 R 可视化代码

```
# 直方图
par(mfrow=c(2,2))
for(col in names(data)[sapply(data, is.numeric)]){
    hist(data[[col]], main=paste(col, "分布"), xlab=col)
}

# 箱线图
boxplot(data[sapply(data, is.numeric)], main="数值型变量箱线图")

# 散点图矩阵
pairs(data[sapply(data, is.numeric)])

# 相关性热力图
library(corrplot)
cor_matrix <- cor(data[sapply(data, is.numeric)])
corrplot(cor_matrix, method = "color")
```

Listing 7: R 可视化代码

A.4 假设检验代码

A.4.1 Python Mann-Whitney U 检验代码

```python
from ucimirepo import fetch_ucirepo
import pandas as pd
from scipy.stats import mannwhitneyu

cdc = fetch_ucirepo(id=891)
df = cdc.data.features.copy()
df['Diabetes_binary'] = cdc.data.targets

group0 = df[df['Diabetes_binary'] == 0]
group1 = df[df['Diabetes_binary'] == 1]

for col in ['BMI', 'MentHlth', 'PhysHlth']:
    u, p = mannwhitneyu(group0[col].dropna(),
                        group1[col].dropna(),
                        alternative='two-sided')
    print(f"{col:<10} | U = {u:>12.2f} | p = {p:>12.6f}")
```

Listing 8: Python Mann-Whitney U 检验

A.5 数据建模代码

A.5.1 Python 逻辑回归建模代码

```python
from ucimirepo import fetch_ucirepo
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import (classification_report,
    roc_auc_score,
                             roc_curve, confusion_matrix,
    accuracy_score)
from sklearn.model_selection import train_test_split

```

```
cdc = fetch_ucirepo(id=891)
X = cdc.data.features
y = cdc.data.targets

X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.25, random_state=42, stratify=y)

logit = LogisticRegression(max_iter=1000)
logit.fit(X_train, y_train)

y_pred = logit.predict(X_test)
y_pred_proba = logit.predict_proba(X_test)[:, 1]

print(classification_report(y_test, y_pred))
print(f"Accuracy: {accuracy_score(y_test, y_pred):.3f}")
print(f"AUC: {roc_auc_score(y_test, y_pred_proba):.3f}")

# ROC 曲线
fpr, tpr, _ = roc_curve(y_test, y_pred_proba)
plt.figure(figsize=(5, 4))
sns.lineplot(x=fpr, y=tpr, label=f"ROC (AUC = {roc_auc_score(
    y_test, y_pred_proba):.3f})")
plt.plot([0, 1], [0, 1], 'k--')
plt.xlabel('FPR')
plt.ylabel('TPR')
plt.title('ROC Curve')
plt.tight_layout()
plt.show()

# 混淆矩阵
plt.figure(figsize=(4, 3))
sns.heatmap(confusion_matrix(y_test, y_pred),
            annot=True, fmt='d', cmap='Blues',
            xticklabels=['No', 'Diabetes'],
            yticklabels=['No', 'Diabetes'])
```

```
plt.xlabel('Predicted')
plt.ylabel('Actual')
plt.title('Confusion Matrix')
plt.tight_layout()
plt.show()

# 提取关键影响因素
coef_df = (pd.DataFrame({'Feature': X.columns,
                         'Coefficient': logit.coef_[0],
                         'OR': np.exp(logit.coef_[0])})
           .assign(Abs_OR=lambda d: d['OR'].abs())
           .sort_values('Abs_OR', ascending=False))
print('\nTop 15 most influential factors (OR)')
print(coef_df.head(15).round(3))
```

Listing 9: Python 逻辑回归建模

B 数据字典

表 4: 数据字段说明表

变量名称	类型	描述与取值范围
Diabetes_binary	二元分类	糖尿病诊断结果：0 = 无糖尿病，1 = 有糖尿病
HighBP	二元分类	高血压诊断：0 = 无，1 = 有
HighChol	二元分类	高胆固醇诊断：0 = 无，1 = 有
CholCheck	二元分类	过去 5 年内是否检查过胆固醇：0 = 否，1 = 是

续下页

续表 4 数据字段说明表

变量名称	类型	描述与取值范围
BMI	连续	身体质量指数（Body Mass Index），取值范围：12-98
Smoker	二元分类	是否吸烟（至少 100 支香烟）：0 = 否，1 = 是
Stroke	二元分类	是否曾患中风：0 = 否，1 = 是
HeartDiseaseorAttack	二元分类	是否患有冠心病或心肌梗死：0 = 否，1 = 是
PhysActivity	二元分类	过去 30 天内是否有体育锻炼：0 = 否，1 = 是
Fruits	二元分类	每日水果摄入是否 1 次：0 = 否，1 = 是
Veggies	二元分类	每日蔬菜摄入是否 1 次：0 = 否，1 = 是
HvyAlcoholConsump	二元分类	重度饮酒（男性：每周 14 杯；女性：每周 7 杯）：0 = 否，1 = 是
AnyHealthcare	二元分类	是否有医疗保险：0 = 否，1 = 是
NoDocbcCost	二元分类	是否因费用问题未就医：0 = 否，1 = 是
GenHlth	序数分类	总体健康状况：1 = 优秀，2 = 非常好，3 = 好，4 = 一般，5 = 差
MentHlth	连续	过去 30 天心理健康不佳天数，取值范围：0-30
PhysHlth	连续	过去 30 天身体健康不佳天数，取值范围：0-30

续下页

续表 4 数据字段说明表

变量名称	类型	描述与取值范围
DiffWalk	二元分类	是否因健康问题行走困难：0 = 否，1 = 是
Sex	二元分类	性别：0 = 女性，1 = 男性
Age	次序分类	年龄分组： 1=18-24, 2=25-29, 3=30-34, 4=35-39, 5=40-44, 6=45-49, 7=50-54, 8=55-59, 9=60-64, 10=65-69, 11=70-74, 12=75-79, 13：≥80
Education	次序分类	教育程度： 1= 未上过学，2= 小学，3= 初中，4= 高中， 5= 大专，6= 本科及以上
Income	次序分类	家庭年收入（美元）： 1: $\leq 10k$, 2: $10k-15k$, 3: $15k-20k$, 4: $20k-25k$, 5: $25k-35k$, 6: $35k-50k$, 7: $50k-75k$, 8: $\geq 75k$

数据说明：

- 数据来源于 CDC BRFSS 2021 年度调查，共包含 253,680 条有效记录。
- 所有分类变量（二元/序数）均已进行编码处理，便于统计分析。
- 连续变量 BMI、MentHlth、PhysHlth 已进行异常值检测与处理。
- 数据已清洗，无缺失值，可直接用于建模与分析。