****

**《Python数据分析》课程项目报告**

题 目： 水质分析与可饮用性预测

学 期： 2020-2021-2

课程号： 591.025.2-01

学 号： 20057014

姓 名： 徐泽凯

专 业： 信息管理与信息系统

邮 箱： 283996806@qq.com

一、选题背景与实用意义（20分）

对于公众健康来说，获得安全的饮用水至关重要。这不仅是一项基本人权，也是有效保护健康政策的重要部分。作为国家、区域和地方层面的健康和发展问题，确保生活用水安全、高质十分重要。水资源质量不过关或受到污染，会导致霍乱、腹泻、痢疾、甲型肝炎、伤寒和小儿麻痹症等一系列疾病的传播，令人们的健康受到威胁，同时影响社会安定。所以，我们有必要加强对水质的检测与分析，为人们提供健康的生活用水。

一些证据表明，在一些地方，对供水和卫生设施的投资可以产生净经济效益，因为减少的不利健康影响和保健费用超过了设施投资成本。从经济效益来说，无论是用于饮用、家庭使用、食品生产还是娱乐目的，改善供水和卫生条件，以及更好地管理水资源，都可以促进各国的经济增长，并大大有助于减贫。

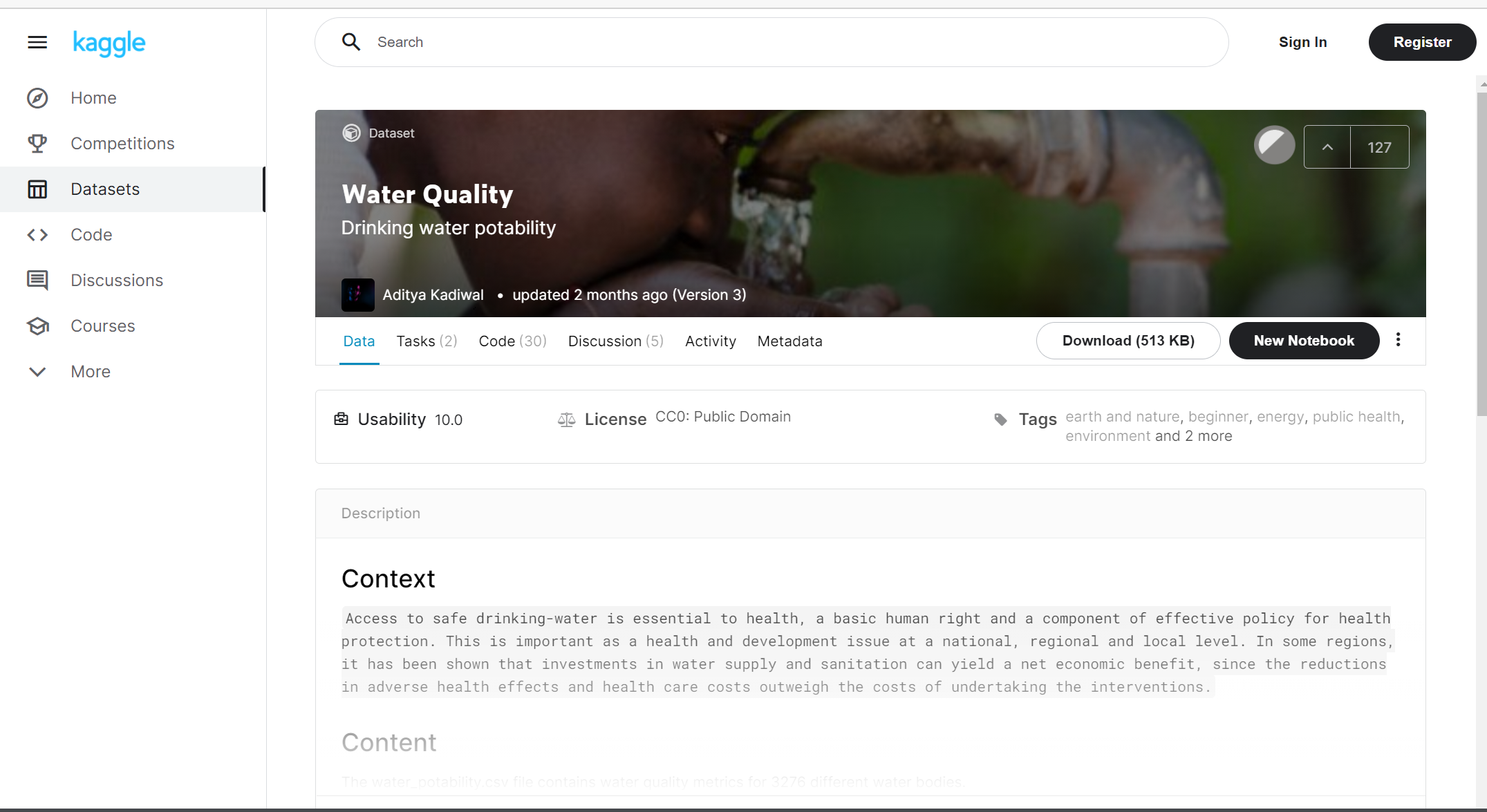
因此，我希望使用水质数据集来分析一系列水质衡量指标，并应用机器学习来区分饮用水和非饮用水。希望通过此次数据分析，能够学习到一系列关于水质的知识，并从自身视角，提出实践性建议，从而提高人们对于水质问题的关注度。

二、数据收集与预处理（20分）（包括文字说明及相关截图）

**收集部分：**数据集来源于Kaggle

来源网址：<https://www.kaggle.com/adityakadiwal/water-potability>

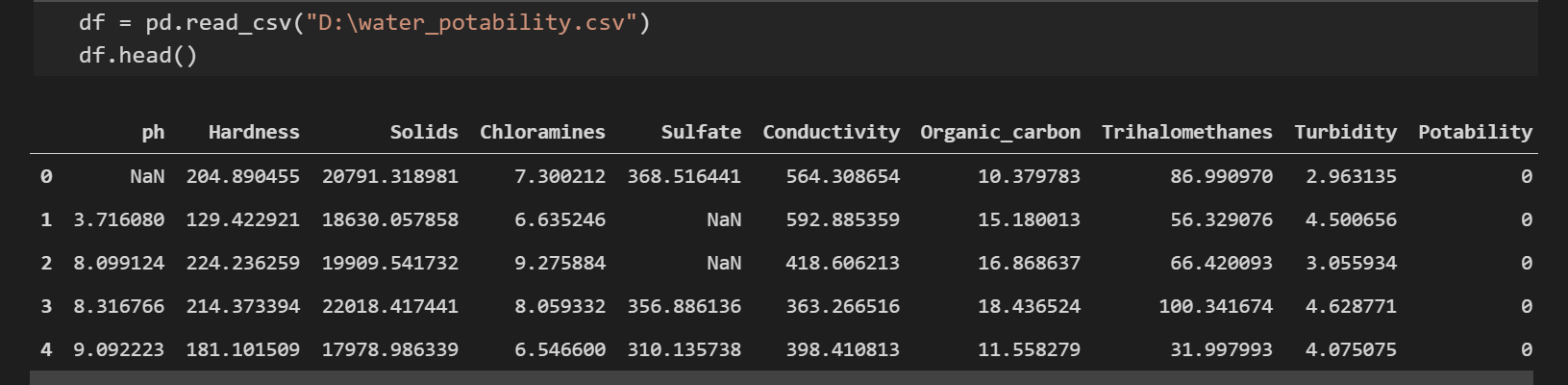
截图：



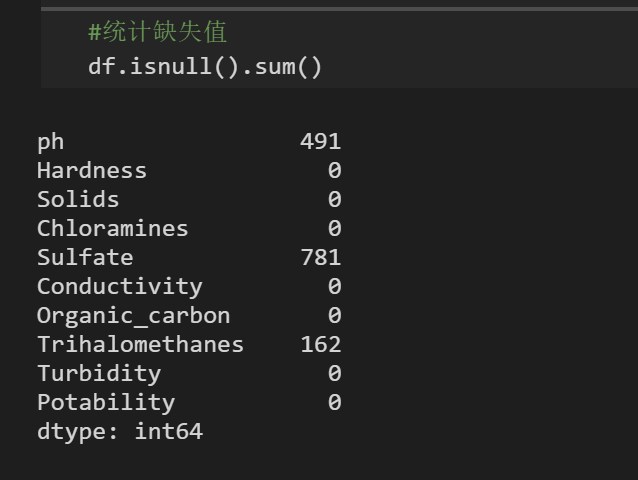
**数据预处理：**

1)缺失值处理：

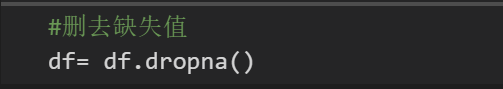
#读取数据，发现存在存在值



#统计缺失值

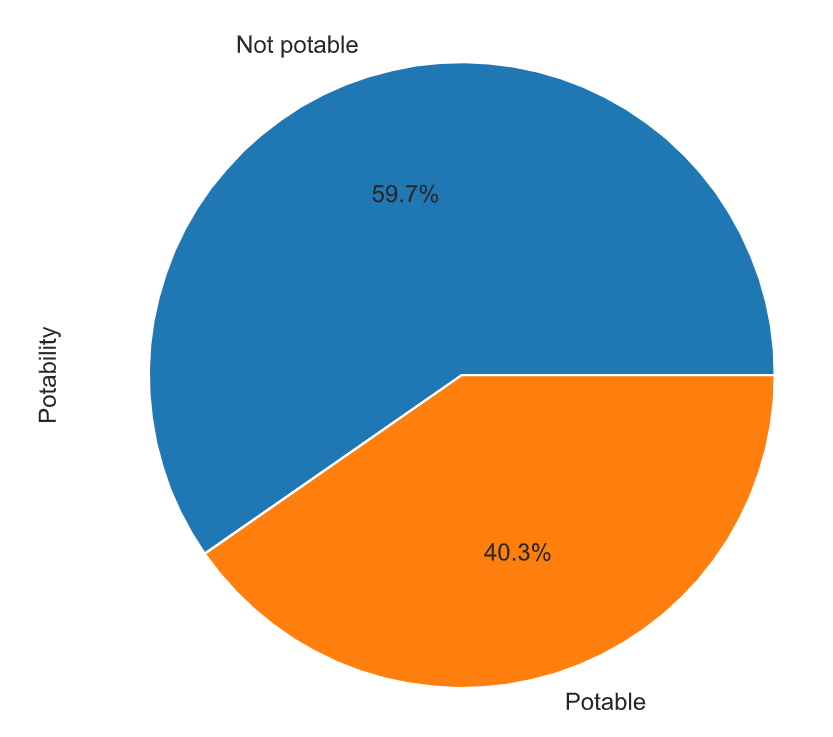


#由于有关水质的相关数据为**敏感数据**，不宜做改动，故选择删去含缺失值的数据



2）平衡度分析：

1. df['Potability'].replace({0:'Not potable', 1:'Potable'}, inplace=True)
2. df\_unpotable = df[df['Potability']=='Not potable']
3. df\_potable = df[df['Potability']=='Potable']
4. df.Potability.value\_counts().plot(kind ='pie',subplots=True,autopct='%1.1f%%',labeldistance=1.1, figsize=(6,6))

#由饼图可知，数据集较为平衡，故不作平衡处理

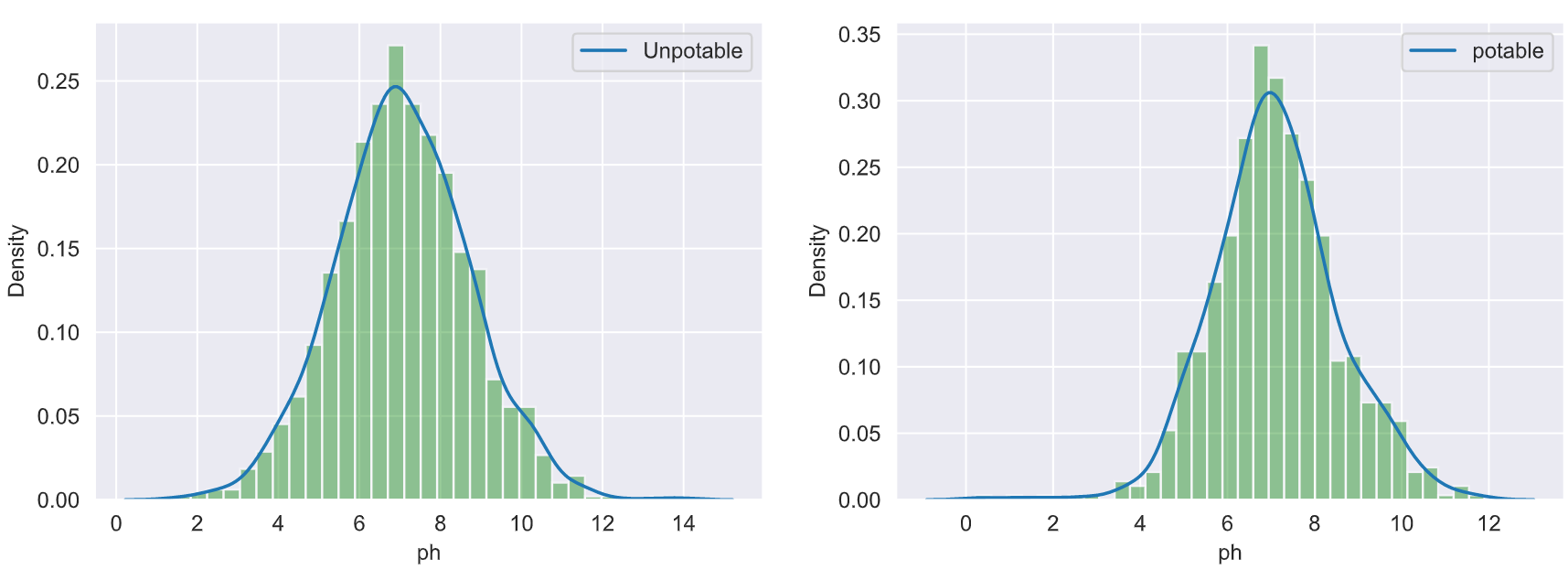
三、数据分析过程/建模/预测（40分）（包括文字说明及相关截图）

|  |  |
| --- | --- |
| 水质指标 | 描述 |
| *Potability* | 可饮用性 |
| *ph* | ph值（0-14） |
| *Hardness* | 硬度,指水中钙、镁离子的总浓度（mg/L） |
| *Solids* | 总溶解固体含量（百万分之，ppm） |
| *Chloramines* | 氯胺含量（ppm） |
| *Sulfate* | 硫酸盐含量（mg/L） |
| *Conductivity* | 电导率（μS/cm） |
| *Organic\_carbon* | 有机碳含量（ppm） |
| *Trihalomethanes* | 三卤甲烷含量（μg/L） |
| *Turbidity* | 浑浊度（NTU）（测量方法来自南洋理工） |

**1.特征分布：**

PH值：

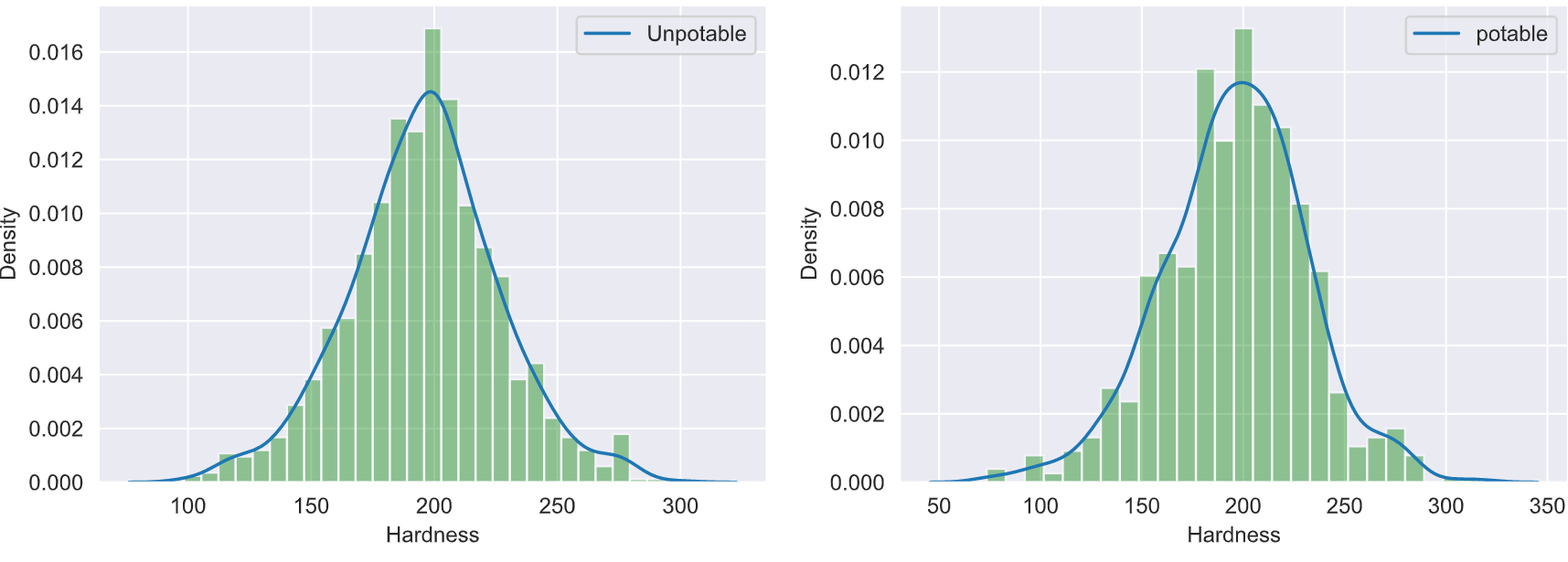
1. plt.figure(figsize=(12,4))
2. sns.set\_style('darkgrid')
3. plt.subplot(1,2,1)
4. sns.distplot(df\_unpotable['ph'],hist\_kws={'color':'green','label':'hist'})
5. plt.legend(['Unpotable'])
6. plt.subplot(1,2,2)
7. sns.distplot(df\_potable['ph'],hist\_kws={'color':'green','label':'hist'})
8. plt.legend(['potable'])



#世界卫生组织所设标准为6.52–6.83，可见可饮用性在ph值上没有具体特征体现，两组数据也较为相似

#在日常生活中应避免饮用过酸或过碱的水源

硬度（代码较为重复，故不再贴出，以下皆是如此）：

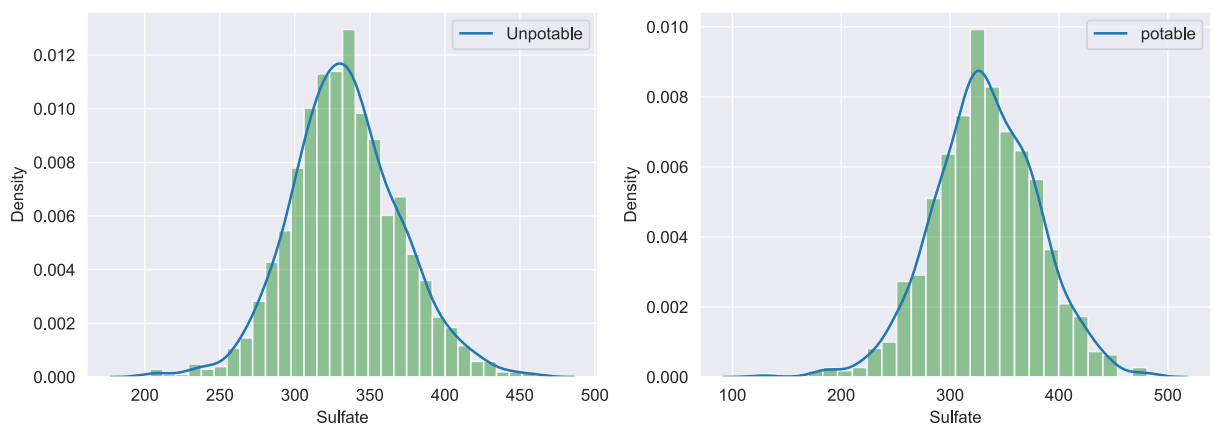


#我国的《生活用水卫生标准》规定,饮用水的总硬度不超过450mg/L（数据来源国未明示）

#世界卫生组织推荐最佳饮用水硬度是170mg/L

#由图可见，从整体硬度来看，两类样本都符合基本要求，但是在推荐硬度上可饮用的样本较多，可以看出不少地区确实往推荐标准上下功夫了。

硫酸盐含量：

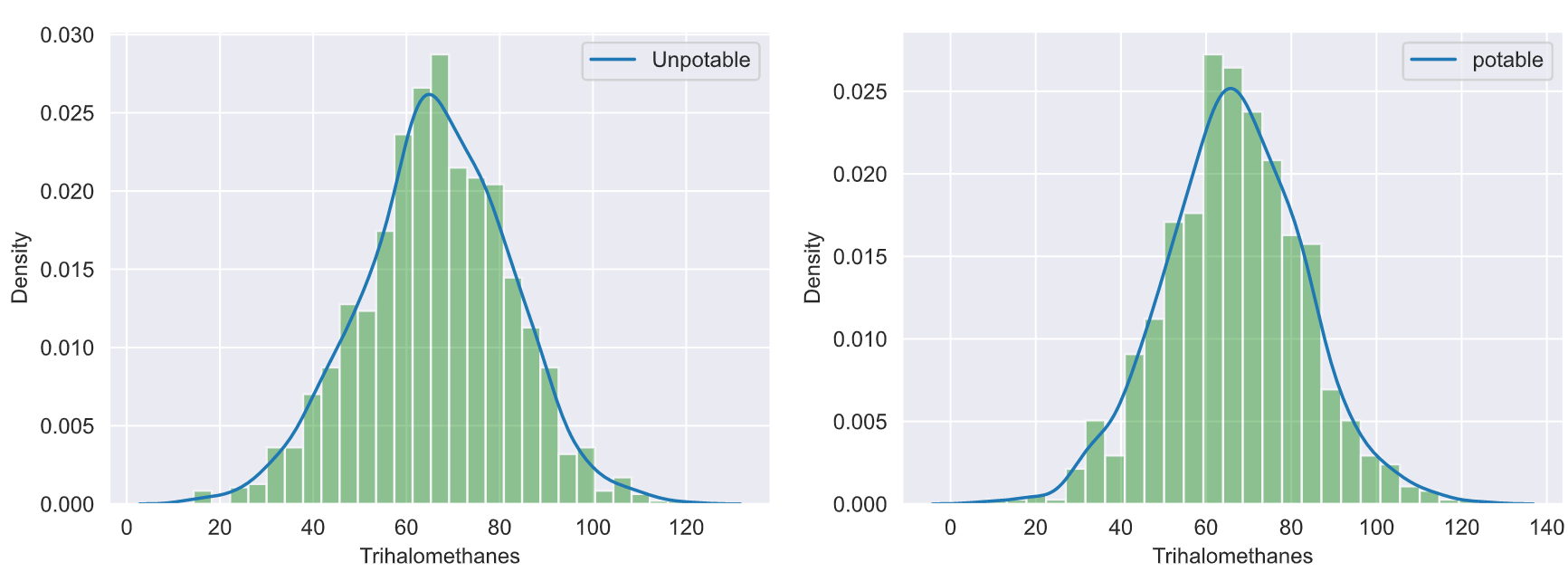


#我国指标为不超过250mg/L，显然该数据集中的样本大幅超标，且不可饮用者居多。

#该类现象常见于井水和受到海水入侵的水源，需要加强对该类水源的整治与规范，同时宣传饮用前的处理手段

#长期饮用硫酸盐超标的水源会导致胃肠道紊乱，水的口感也不佳

三卤甲烷含量：

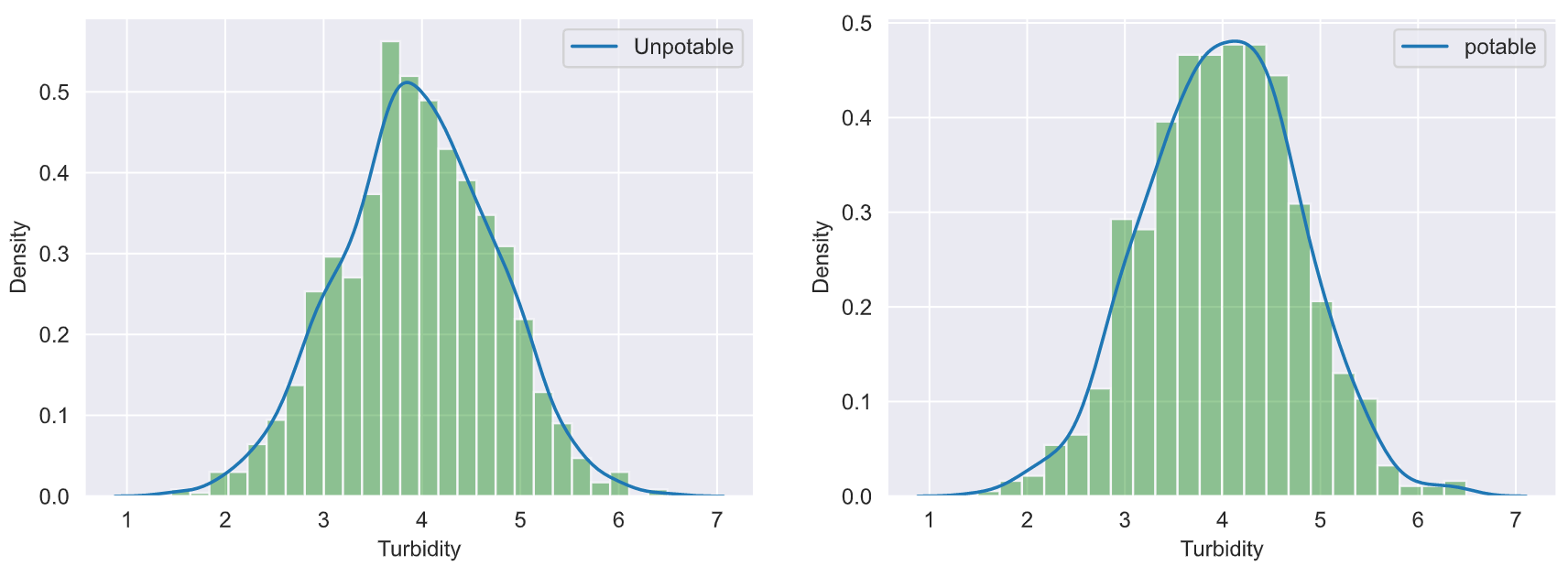


#指标为不得超过80μg/L，部分样本超出该值

#三卤甲烷为消毒副产物，含量过高易诱发癌症

#需要加强水厂消毒技艺的改进，同时研发更符合需求的消毒剂

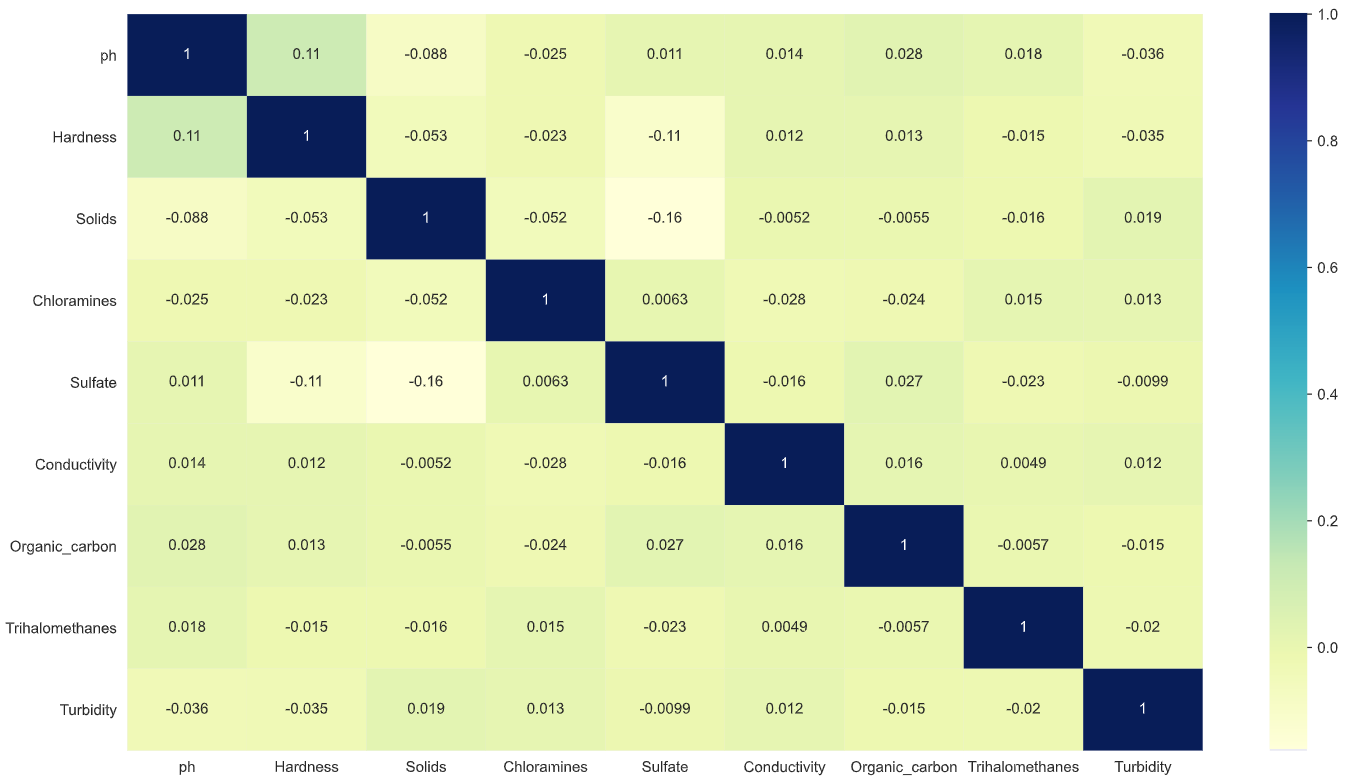
浑浊度：



#世界卫生组织建议值为5.0NTU，两类数据都非常符合标准（水中不可溶固体较少）

**2.关联性分析：**

1. plt.figure(figsize = (16,9))
2. sns.heatmap(df.corr(),cmap ="YlGnBu", annot = True)



#由热力图可知道，除硬度与ph值之间有微弱联系，其他属性值之间关联性不大

**3.建模与预测**

1. *#删去结果列*
2. X = df.drop(['Potability'], axis = 1)
3. y = df['Potability']
4. *#划分训练集与测试集*
5. X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.3, random\_state=1)
6. *#随机森林*
7. RFC = RandomForestClassifier(random\_state=0)
8. RFC.fit(X\_train,y\_train)
9. print("准确率：",RFC.score(X\_test,y\_test))
10. 准确率： 0.6771523178807947
11. *#逻辑回归*
12. lr=LogisticRegression(max\_iter=10000)
13. lr.fit(X\_train,y\_train)
14. lr.score(X\_test,y\_test)
15. print("准确率"lr.score(X\_test,y\_test))
16. print("回归系数："，lr.coef\_)
17. print("截距："，lr.intercept\_)
18. 准确率 0.6208609271523179
19. 回归系数： [[-2.01320237e-05 -4.79356477e-04  1.23815244e-05  2.44731721e-05
20. -3.09386982e-04 -9.56028385e-04 -9.32867733e-05 -1.31225572e-04
21. 2.47251159e-05]]
22. 截距： [-3.25342336e-06]

通过逻辑回归与随机森林两种模型，得到的模型准确率都较为客观。但通过逻辑回归中**回归系数为负数**可以了解到该数据集样本个数较少，这同时也是本次数据分析的一大遗憾。

反思：1）对于水质的各项指标了解程度不够，所提建议不够具体且实践性不强（可见数据分析在交叉学科中的应用前景）

2）笔者对于集成学习的了解较为浅显，没能使用多种集成方法，需要在日后加强学习

3）可视化技能还需打磨