模块 A：大数据（20 分）

竞赛题目：风机监测预警分析

背景简介

风机是石油化工、冶金、电力及交通等行业的关键设备。为了能及时了解风机的运行状况，减少停机次数或避免事故扩大化，需要对设备的各项运行参数进行监测。

风机监测预警系统是对风机运行过程中进行监控及故障报警。它由监控上位机及软件系统、监控站、风机、电机运行参数传感器等组成。它具有风机运行的实时监控、风机停运报警、风机远程中心监控等功能。

风机监测预警系统充分利用传感器检测，信号处理，计算机技术，数据通讯技术和风机的有关技术，全面地对矿井总回风中的风压（负压、静压、动压、全压及其效率）、风速、风量、瓦斯浓度、出口气体温度、对旋轴流风机前后轴承温度、运行状态、正反转状态、电机定子温度和轴承温度等风机性能参数，对旋轴流风机设备振动、位移、速度、加速度、振动主频、频率分量及其烈度等振动参数，电机三相电压、电流、有功无功电度、有功无功功率、总有功功率、总无功功率、视在功率、功率因数、频率等电量参数进行实时在线监测，在机组的运行过程中，判别机组性能劣化趋势，使运行，维护，管理人员心中有数。

环境介绍

在大数据竞赛部分的环境中，包含了操作系统、Mysql 数据库、hadoop 环境三部分。具体环境参数如下。

操作系统

- Centos 7

Mysql 数据库

- ip: localhost
- port: 3306
- username: root
- password: wiseinsight

进入 mysql 语句为：

```
mysql -uroot -pwiseinsight
```

hadoop 环境

共有 4 个节点，分别为：

- `root@{username}-hdfs-dn-0`：datanode 节点
- `root@{username}-hdfs-nn-0`：namenode 节点
- `root@{username}-yarn-nm-0`：namemanager 节点
- `root@{username}-yarn-rm-0`：resourcemanager 节点

注意事项

每题完成之后必须按要求保存该题答案到指定文件中，每题答案保存代码已给出，只需替换其中的 XXXX 即可。

- 必须在 Linux 命令行下执行答案保存代码
- 运行答案保存代码时，不限定在某一节点上，四个节点都可运行，但必须把答案保存在题目要求的文件中
- 必须在确定保存好答案后，再点击 下一步 按钮
- 如果遇到需要输入较长代码情况，建议用 \ 换行分段写入，避免出现代码覆盖情况

数据介绍

风机监测预警系统的数据包含多种设备，每个设备通常有上百个变量，本次竞赛的数据经过筛选保留了其中的一部分变量，涵盖了设备的工况参数、环境参数和状态参数等多个维度。Hive 的 `brick_competition_data` 数据库中共有十张设备表，分别记录着各自设备 2017 年全年的数据，用于后面的 Hive 计算，表的名称及说明如下表所示：

设备编号	设备所属类型
DD_P1_L7_044	电计量
DD_P1_L7_045	电计量
DD_P1_L7_046	电计量
FJ_P2_L6_057	风机
FJ_P2_L6_058	风机
FJ_P2_L6_059	风机
GF_P1_L7_052	光伏
GF_P1_L7_053	光伏
GF_P1_L7_054	光伏

任务列表

1. 环境查看与配置
2. HDFS 文件操作
3. Sqoop 相关操作

4. 电计量设备数据探查
5. 光伏设备数据探查
6. 风机设备数据探查

任务 1：环境查看与配置（4 分）

1.1 查看 NameNode URI 端口号

1.2 查看 MapReduce 程序运行模式

1.3 修改 HDFS 自动数据备份数量

1.4 修改单个节点可用物理内存大小

1.1 查看 NameNode URI 端口号

1. 在 {username}-yarn-nm-0 节点上操作，查看 Hadoop 配置中的 NameNode URI 信息，记录 端口号 到 ~/results/task1-1.txt 文件中。（1 分）

例如：查询出的 端口号 为 XXXX，保存结果如下所示

```
echo 'XXXX' > ~/results/task1-1.txt
```

1.2 查看 MapReduce 程序运行模式

2. 在 {username}-yarn-nm-0 节点上操作，查看 MapReduce 程序运行模式（即 mapreduce.framework.name），记录结果到 ~/results/task1-2.txt 文件中。（1 分）

例如：查询出的 模式 为 XXXX，保存结果如下所示。

```
echo 'XXXX' > ~/results/task1-2.txt
```

1.3 修改 HDFS 自动数据备份数量

3. 在 {username}-yarn-nm-0 节点上操作，修改 HDFS 自动数据备份数量，把原有的备份数量 3，修改为 5，并保存修改完成后的配置文件。（1 分）

1.4 修改单个节点可用物理内存大小

4. 在 {username}-yarn-nm-0 节点上操作，修改单个节点可用物理内存大小，把原有大小修改为 8192，并保存修改完成后的配置文件。（1 分）

任务 2：HDFS 文件操作（4 分）

2.1 HDFS 上新建目录表

2.2 HDFS 上修改目录权限

2.3 上传本地文件到 HDFS 上

2.4 HDFS 上拷贝文件

2.1 HDFS 上新建目录表

1. 在 {username}-yarn-nm-0 节点上操作， 在 HDFS 上创建一个名为 /wiseinsight 的目录。（1 分）

2.2 HDFS 上修改目录权限

2. 在 {username}-yarn-nm-0 节点上操作，把 HDFS 上 /wiseinsight 目录权限修改为 777。（1 分）

2.3 上传本地文件到 HDFS 上

3. 在 {username}-yarn-nm-0 节点上操作，上传本地 ~/data/map_P1_L6_033.csv 文件到 HDFS 的 '/wiseinsight' 目录下，保持文件名称不变。（1 分）

2.4 HDFS 上拷贝文件

4. 在 {username}-yarn-nm-0 节点上操作，HDFS 上复制 /wiseinsight/map_P1_L6_033.csv 文件副本到 HDFS 上的 /data 目录下，保持名称不变。（1 分）

任务 3：Sqoop 相关操作（3 分）

3.1 Sqoop 连接 Mysql 并查询

3.2 Sqoop 导入

3.3 Sqoop 导出

3.1 Sqoop 连接 Mysql 并查询

1. 在 {username}-yarn-nm-0 节点上操作，运用 sqoop，查询 brick_competition_data 数据库中的 type 表，查看风机类型表，并统计包含风机设备、电计量设备、光伏设备等所有设备的总数量，并保存结果到 ~/results/task3-1.txt 文件中。（1 分）

例如：得到的 总数量 为 XX，保存结果如下所示。

```
echo 'XX' > ~/results/task3-1.txt
```

3.2 Sqoop 导入

2. 在 {username}-yarn-nm-0 节点上操作，使用 Sqoop 将 Mysql 里 brick_competition_data 数据库中 P1_L7_058 数据表导入到 hive 数据仓库中的 default 数据库里。（1 分）

3.3 Sqoop 导出

3. 在 {username}-yarn-nm-0 节点上操作，使用 Sqoop 将 HDFS 上的 /data/part-m-00000 文件，导入到 Mysql 里 brick_competition_data 数据库中的 task 表里。（1 分）

任务 4：电计量设备数据探查（3 分）

4.1 计算单个 电计量设备 单个 逆变器日发电量 的和

4.2 计算单个 电计量设备 多个 逆变器日发电量 总和

4.3 计算多个 电计量设备 多个 逆变器日发电量 总和

4.1 计算单个 电计量设备 单个 逆变器日发电量 的和

1. 在 {username}-yarn-nm-0 节点上操作，分别计算 2017 年不同电计量设备数据中 22#逆变器日发电量 的总和，找出最大值，并记录其设备编号到 ~/results/task4-1.txt 文件中。（查询数据存于 Hive 上的 brick_competition_data 数据库中）（1 分）

例如：查询出的 设备编号 为 XX_XX_XX_XXX，保存结果如下所示。

```
echo 'XX_XX_XX_XXX' > ~/results/task4-1.txt
```

4.2 计算单个 电计量设备 多个 逆变器日发电量 总和

2. 在 {username}-yarn-nm-0 节点上操作，分别计算 2017 年不同电计量设备中所有 逆变器日发电量 的总和（即: 所有含 逆变器 的字段），相互比较大小，并记录最大值对应的设备编号到 ~/results/task4-2.txt 文件中。（1 分）

例如：查询出的 设备编号 为 XX_XX_XX_XXX，保存结果如下所示。

```
echo 'XX_XX_XX_XXX' > ~/results/task4-2.txt
```

4.3 计算多个 电计量设备 多个 逆变器日发电量 总和

3. 在 {username}-yarn-nm-0 节点上操作，计算 2017 年所有电计量设备中所有 逆变器日发电量 的总和，保留两位小数，并记录结果到 ~/results/task4-3.txt 文件中。（1 分）

例如：查询出的 逆变器日发电量总和 为 XXXX，保存结果如下所示。

```
echo 'XXXX' > ~/results/task4-3.txt
```

任务 5：光伏设备数据探查（3 分）

5.1 计算单个 光伏设备 单个 逆变器故障 的概率

5.2 计算多个 光伏设备 单个 逆变器故障 的出现次数

5.3 计算多个 光伏设备 多种 逆变器故障 的出现次数

5.1 计算单个 光伏设备 单个 逆变器故障 的概率

1. 在 {username}-yarn-nm-0 节点上操作，分别计算 2017 年不同光伏设备数据中各自 48# 逆变器故障 的出现的概率，记录最小值对应的光伏设备编号到 ~/results/task5-1.txt 文件中。（故障概率 = 故障出现次数 / 对应设备总行数）（数值 1 代表 出现故障，数值 0 代表 正常）（1 分）

例如：查询出的 设备编号 为 XX_XX_XX_XXX，保存结果如下所示。

```
echo 'XX_XX_XX_XXX' > ~/results/task5-1.txt
```

5.2 计算多个 光伏设备 单个 逆变器故障 的出现次数

2. 在 {username}-yarn-nm-0 节点上操作，计算出 48# 逆变器硬件故障 在不同光伏设备中出现的次数，记录最小值对应的设备编号到 ~/results/task5-2.txt 文件中。（1 分）

例如：查询出的 设备编号 为 XX_XX_XX_XXX，保存结果如下所示。

```
echo 'XX_XX_XX_XXX' > ~/results/task5-2.txt
```

5.3 计算多个 光伏设备 多种 逆变器故障 的出现次数

3. 在 {username}-yarn-nm-0 节点上操作，编号为 GF_P1_L7_053 光伏设备所有数据中，哪种 48#逆变器故障 出现次数最少？并记录结果到 ~/results/task5-3.txt 文件。（48#逆变器故障 类别限定为：48#逆变器接地故障、48#逆变器硬件故障、48#逆变器孤岛故障、48#逆变器频率故障、48#变器模块故障）（1 分）

例如：查询出的 故障类型 为 XXXX，保存结果如下所示。

```
echo 'XXXX' > ~/results/task5-3.txt
```

任务 6：风机设备数据探查（3 分）

6.1 计算单个 风机设备 累计 发电量 的值

6.2 计算单个 风机设备 各月份的累计 发电量 的值

6.3 统计多个 风机设备 最高累计 发电量 月份出现的频率

6.1 计算单个 风机设备 累计 发电量 的值

1. 在 {username}-yarn-nm-0 节点上操作，计算 2017 年各风机累计 发电量，寻找出累计发电量最高的风机设备，并返回其设备编号，记录结果到 ~/results/task6-1.txt 文件中。（1 分）

例如：查询出的 设备编号 为 XXXX，保存结果如下所示。

```
echo 'XXXX' > ~/results/task6-1.txt
```

6.2 计算单个 风机设备 各月份的累计 发电量 的值

2. 在 {username}-yarn-nm-0 节点上操作，探查 2017 年累计 发电量 最大的风机设备数据，查询出哪个月份 发电量 最大？并记录 月份 结果到 ~/results/task6-2.txt 文件中（保存格式必须为 yyyymm，例如：201701）。（1 分）

例如：查询出的 月份 为 yyyymm，保存结果如下所示。

```
echo 'yyyymm' > ~/results/task6-2.txt
```

6.3 统计多个 风机设备 最高累计 发电量 月份出现的频率

3. 在 {username}-yarn-nm-0 节点上操作，统计 2017 年不同风机设备最高 发电量 的月份，并计算月份出现次数，找出出现次数最高的月份，并记录结果到 ~/results/task6-3.txt 文件中（保存格式必须为 yyyymm，例如：201701）。（1 分）

例如：查询出的 月份 为 yyyymm，保存结果如下所示。

```
echo 'yyyymm' > ~/results/task6-3.txt
```

任务结束

恭喜完成本任务，祝您取得好成绩。

模块 B：数据分析与可视化（30 分）

竞赛题目：光伏发电数据分析

背景介绍

作为世界第一大清洁能源的太阳能相对煤炭石油等能源来说是可再生、无污染的，只要有太阳就有太阳能，所以太阳能的利用被很多国家列为重点开发项目。但太阳能具有波动性和间歇性的特性，太阳能电站的输出功率受光伏板本体性能、气象条件、运行工况等多种因素影响，具有很强的随机性，由此带来的大规模并网困境严

重制约着光伏发电的发展。因此挖掘光伏发电数据中的价值，对光伏数据进行分析，显得尤为重要。

环境介绍

在本竞赛模块环境中，使用 nteract 开始任务。环境中已安装 Python 中多种数据分析和可视化库，包括：

- Numpy
- Scipy
- Pandas
- StatsModels
- Matplotlib
- Seaborn
- Plotly
- Pyecharts
- 其它

数据介绍

photovoltaic.csv 中数据字段信息如下：

字段名称	字段类型
ID	当前记录条数
板温	光伏电池板背测温度
现场温度	光伏电站现场温度
转换效率 A	数据采集点 A 处的光伏板转换效率
转换效率 B	数据采集点 B 处的光伏板转换效率
转换效率 C	数据采集点 C 处的光伏板转换效率
电压 A	为数据采集点 A 处汇流箱电压值
电压 B	为数据采集点 B 处汇流箱电压值
电压 C	为数据采集点 C 处汇流箱电压值
电流 A	为采集点 A 处汇流箱电流值
电流 B	为采集点 B 处汇流箱电流值
电流 C	为采集点 C 处汇流箱电流值
功率 A	为采集点 A 处的功率 Pa，P=UI
功率 B	为采集点 B 处的功率 Pb，P=UI
功率 C	为采集点 C 处的功率 Pc，P=UI
发电量	为光伏电厂现场发电量

wind_speed.txt 中数据字段信息如下：

字段名称	字段类型
ID	当前记录条数
风速	为光伏电厂现场风速测量值

wind_direction.xlsx 中数据字段信息如下：

字段名称	字段类型
ID	当前记录条数
风向	为光伏电厂现场风的来向

任务列表

1. 数据整合
2. 数据探查与预处理
3. 数据分析
4. 数据可视化
5. 报告保存

注意事项

每题完成之后必须保存该题结果，每题的保存代码在该题下的 结果保存 处，执行即可，禁止修改。

- 每题的 最终结果变量名 已预先设定，禁止修改
- 若没有保存结果则该题无分
- 若该题有修改需重新运行 结果保存 代码，否则视为上次提交

示例

任务：求长为4，宽为2 的长方形面积

- 结果以变量 rectangle_area 保存

```
rectangle_area = None
```

解题

```
length = 4
width = 2
rectangle_area = length * width
```

```
#结果保存
result.to_save('rectangle_area', rectangle_area)
```

任务启动

通过 nteract 打开 /home/wiseinsight/Desktop/Tasks/Task_B.ipynb 文件，开始竞赛任务。

若您不小心关掉 nteract，可以打开终端输入 nteract 重新启动。

```
#导入所需库
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

#结果保存库
import result

#设置笔记环境
%matplotlib inline
%config InlineBackend.figure_format = 'retina'
%config IPCompleter.greedy = True
%config IPCompleter.use_jedi = True
pd.options.display.max_colwidth = 100
plt.rcParams['figure.figsize'] = (12, 8)
plt.rcParams['axes.unicode_minus'] = False
plt.rcParams['font.sans-serif'] = ['SimHei']
```

任务 1：数据整合（3 分）

这部分任务为对数据进行导入，然后合并数据构建一个完整统一的数据源。具体任务要求如下：

- 载入 csv 格式的 photovoltaic
- 载入 txt 格式的 wind_speed
- 载入 xlsx 格式的 wind_direction
- 导出合成的数据集，并保存为 data_out.csv 到 result 文件夹下

数据载入（1 分）

数据路径

- 读取 photovoltaic.csv 文件，路径为 data/photovoltaic.csv，结果以变量 data 保存
- 读取 wind_speed.txt 文件，路径为 data/wind_speed.txt，结果以变量 wind_speed 保存
- 读取 wind_direction.xlsx 文件，路径为 data/wind_direction.xlsx，结果以变量 wind_speed 保存

```
#TODO: 读取 csv 文件
data = None

#TODO: 读取 txt 文件
wind_speed = None

#TODO: 读取 excel 文件
wind_direction = None

#结果保存
result.to_save('wind_speed', wind_speed)
```

数据连接（1 分）

- 连接前面步骤载入的表，形成单一数据源。
- 采用表连接中内连接模式，以 ID 列为连接字段
- 结果以变量 data_merge 保存

```
#TODO:
data_merge = None

#结果保存
result.to_save('data_merge', data_merge)
```

数据导出（1 分）

- 将上题合并后的数据导出为一个 csv 文件，命名为 data_out.csv，路径为 result/data_out.csv
- 使用变量 data_merge 导出数据

```
#TODO:
```

任务 2：数据探查与预处理（2 分）

数据探查是对数据进行预处理前的必要步骤。通过查看字段类型、简单统计指标，对样本数据有一定的了解，为预处理提供方向。具体任务要求如下：

- 用 pandas 模块对数据进行简单探查
- 把 时间 字段转为 datetime64 类型

该题的数据为任务 1 的 data 变量所接收的数据。

数据探查（1 分）

- 对合并后的数据进行探查

```
#TODO: 探查数据
data.info()

#TODO: 探查数据
data.describe().T
```

```
#TODO: 探查数据
data.head()
```

日期字段处理（1 分）

- 将 时间 字段类型更改为时间类型 (datetime64)
- 处理时间日期字段 data['时间']

```
#TODO:
data['时间'] = None

#结果保存
result.to_save('data_time', data['时间'])
```

任务 3：数据分析（8 分）

这部分主要以描述性统计分析内容为主，计算相关简单指标，对光伏数据进行分析。

该题的数据为任务 1 的 data 变量所接收的数据。

最大值（3 分）

- 查找板温介于 (5, 30) 的 发电量 的最大值
- 结果以变量 data_temperature_max 保存

```
#TODO:
data_temperature_max = None

#结果保存
result.to_save('data_temperature_max', data_temperature_max)
```

- 求哪个月的发电量最大，并保存该月份发电量的值
- 结果以变量 month_max_value 保存

```
#TODO:
month_max_value = None

#结果保存
result.to_save('month_max_value', month_max_value)
```

- 求哪一天的发电量最大，并保存当天发电量的值
- 结果以变量 day_max_value 保存

```
#TODO:
day_max_value = None

#结果保存
result.to_save('day_max_value', day_max_value)
```

皮尔逊相关系数（1 分）

- 计算 光照强度 与 发电量 的皮尔逊相关系数

- 结果以变量 data_corr 保存

```
#TODO:
data_corr = None

#结果保存
result.to_save('data_corr', data_corr)
```

协方差（1 分）

- 计算 光照强度 与 发电量 的协方差
- 结果以变量 data_cov 保存

```
#TODO:
data_cov = None

#结果保存
result.to_save('data_cov', data_cov)
```

多值排序（1 分）

- 先按光照强度降序排序，再按板温降序排序
- 结果以变量 data_sequence 保存

```
#TODO:
data_sequence = None

#结果保存
result.to_save('data_sequence', data_sequence)
```

数据透视（2 分）

- 通过对 光照强度 列进行分组，将 现场温度，板温，发电量 实现数据求和及平均。
- 以 光照强度 降序排序
- 图表类型与下图基本一致，透视表中为??? 的数据需选手计算得到
- 结果以变量 data_pivot_table 保存

	sum			mean		
	发电量	板温	现场温度	发电量	板温	现场温度
光照强度						
941	???	???	???	???	???	???
934	???	???	???	???	???	???
924	???	???	???	???	???	???
916	???	???	???	???	???	???
913	???	???	???	???	???	???

```
#TODO:
data_pivot_table = None

#结果保存
result.to_save('data_pivot_table', data_pivot_table)
```

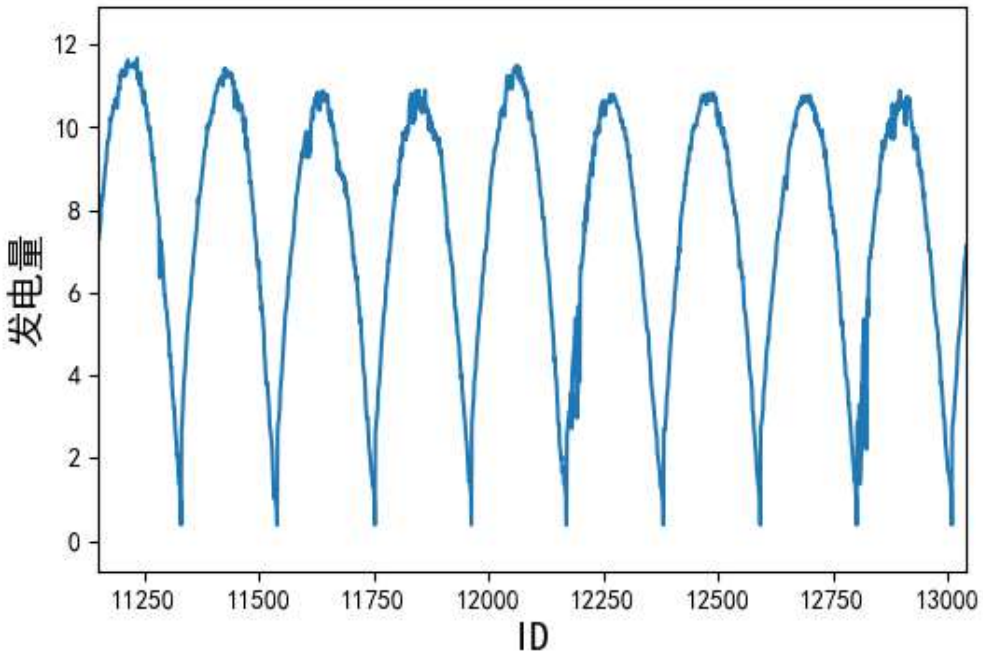
任务 4：数据可视化（17 分）

该题的数据为任务 1 的 data 变量所接收的数据。

线形图（1 分）

问题：制作 ID 范围在 11150 到 13040 之间 发电量 的线形图

- 图形类型正确（0.5 分）
- X 轴和 Y 轴标签正确（0.5 分）

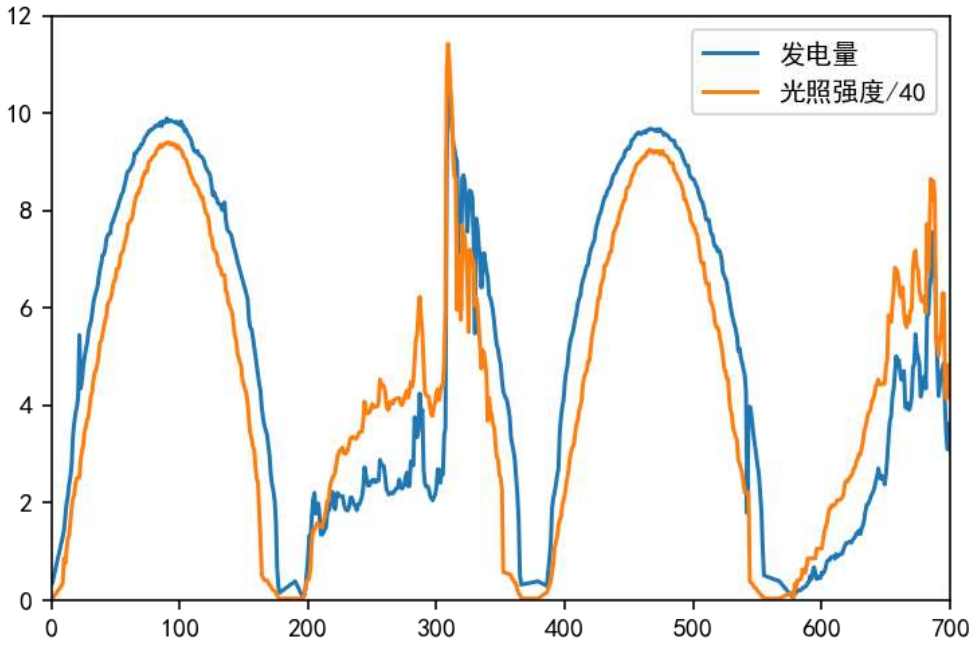


#TODO: 线形图

线形图（2 分）

问题：制作 发电量 和 光照强度/40 的线形图

- x 轴为 ID， 范围为 0 到 700（0.5 分）
- y 轴为 光照强度/40、 发电量（0.5 分）
- y 轴范围为 0 到 12（0.5 分）
- 添加图例（0.5 分）

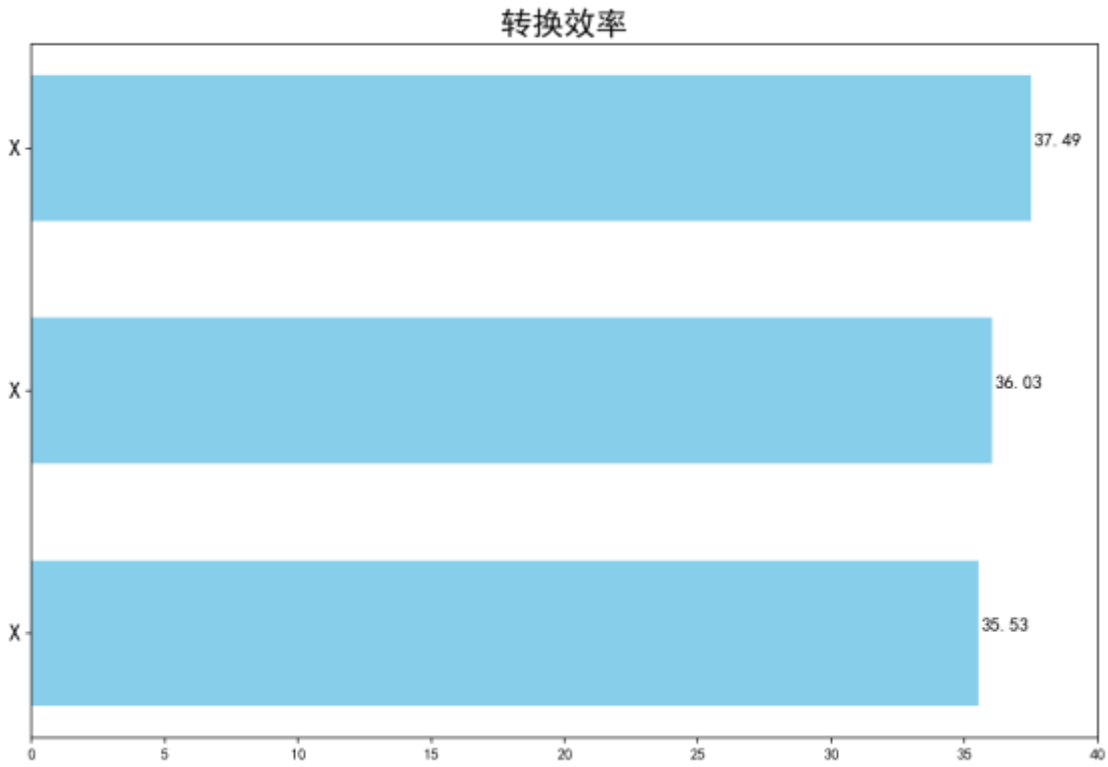


#TODO: 线形图

条形图（2 分）

问题：制作**采集点 A、B、C 三点各自平均转换效率**的水平条形图，倒序排列

- 图形类型正确（0.5 分）
- 添加标题 转换效率 和 X 轴标签，Y 轴范围为 0 到 40，数据倒序排列（0.5 分）
- 添加数据标签（1 分）

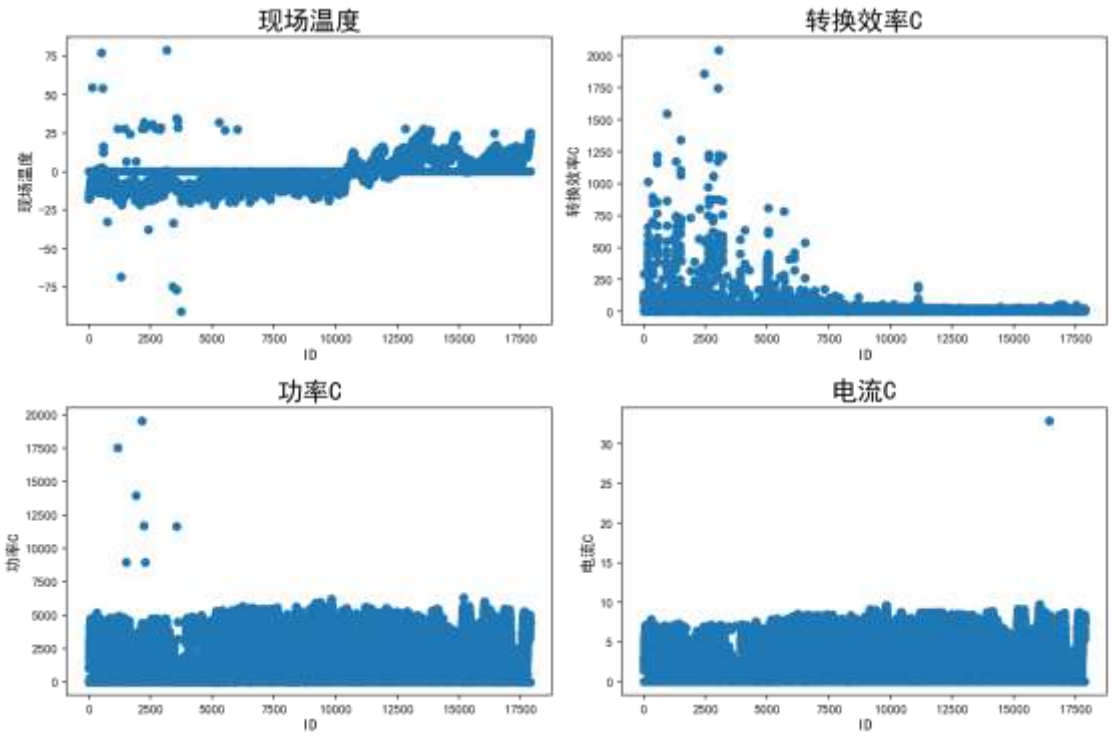


#TODO: 条形图

散点图（2 分）

问题：制作 现场温度、转换效率 C、功率 C、电流 C 的 4 个散点图，横坐标为 ID

- 图形类型正确（0.5 分）
- 4 个子图以两行两列排列（0.5 分）
- 添加图标题，X 轴标签，Y 轴标签（1 分）

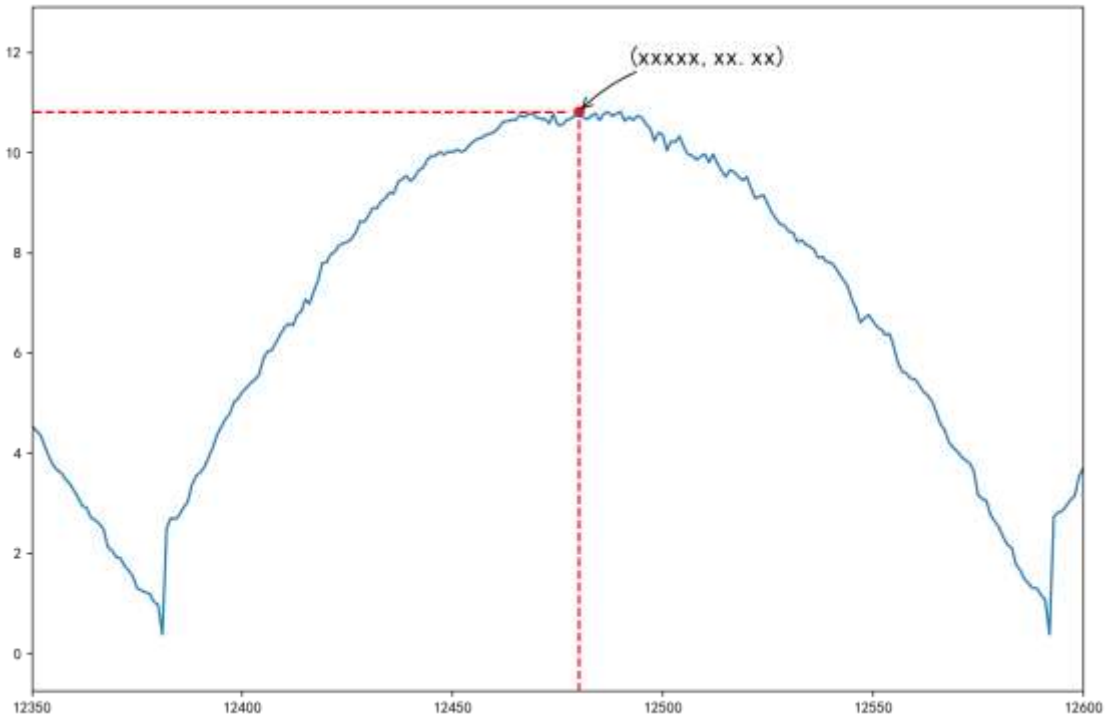


#TODO: 散点图

线形图（2 分）

问题：制作 发电量 的线形图，并找出最高点标注在图上

- x 轴为 ID，范围为 12350 到 12600（0.5 分）
- 找出该范围内 发电量 的最大值，并将其 x 轴，y 轴的值标注在图上 (即将参考图形中的 XXXXX 替换为具体的数值）（1 分）
- 画出该点到 X 轴和 Y 轴的投影线，线的颜色为红色（0.5 分）

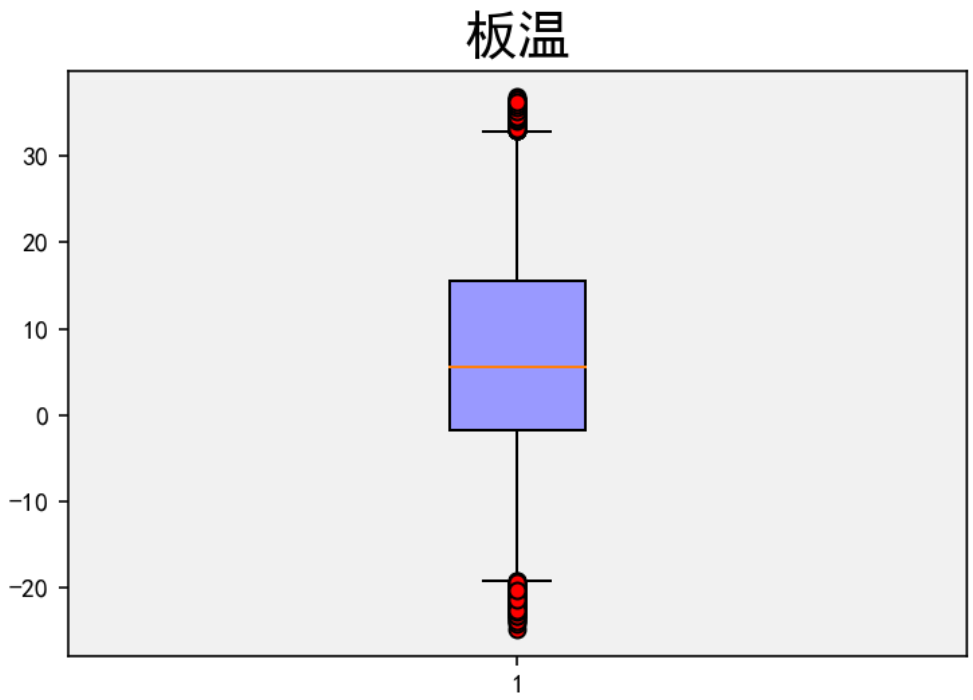


#TODO: 线形图

箱线图（2 分）

制作 板温 的箱线图

- 图形类型正确（0.5 分）
- 设置箱体颜色为紫色 (#9999ff)（0.5 分）
- 图片背景颜色为灰色，透明度为 0.1（0.5 分）
- 上下须与上下四分位的距离: 1 倍的四分位差（0.5 分）

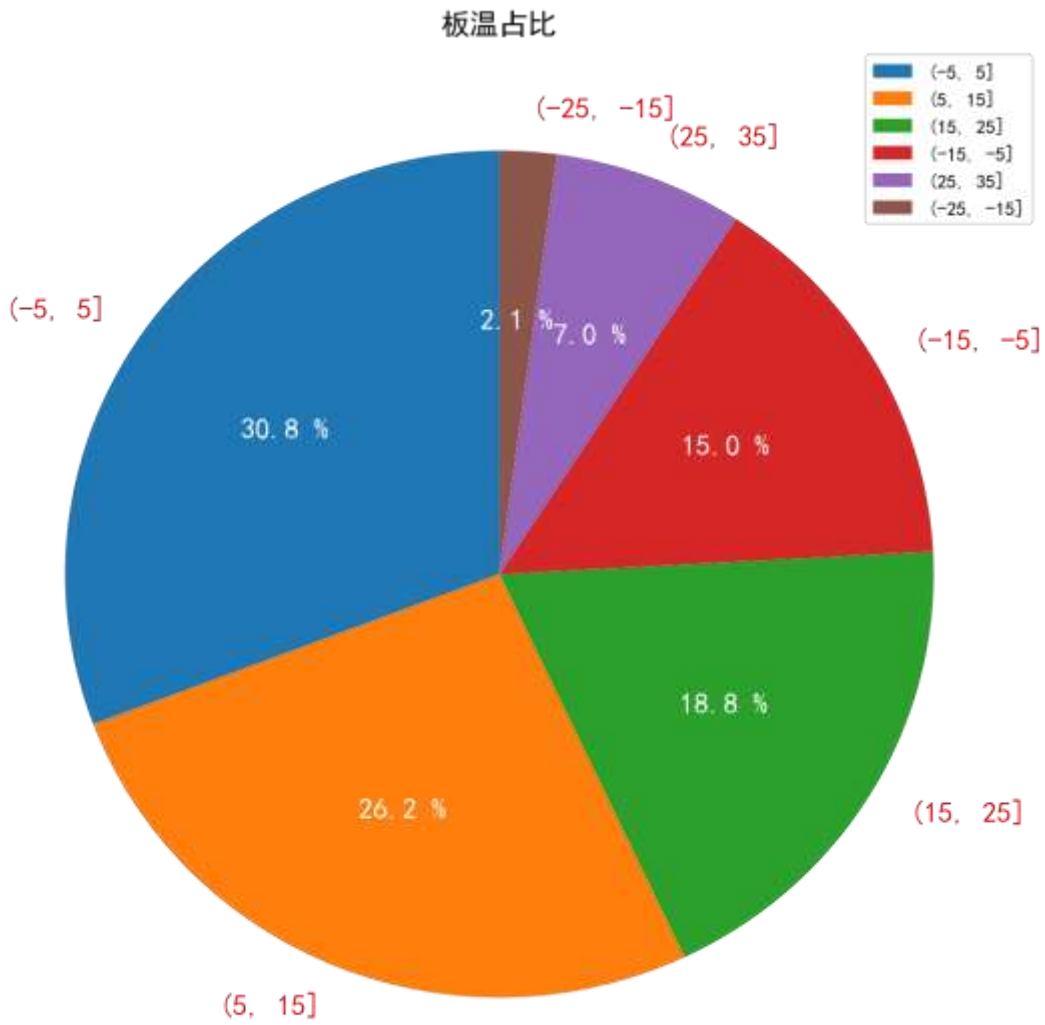


#TODO: 箱线图

饼图（2 分）

制作 板温 的饼图

- 图片大小为 (10, 10)，代码为 `plt.figure(figsize=(10, 10))`
- 图形类型正确（0.5 分）
- 将 板温 分成六个部分，从 -25 到 35 每隔 10 为一部分（0.5 分）
- 每个部分所占比例保留一位小数（0.5 分）
- 内部文字大小为 15，颜色为白色；外部文字大小为 15，颜色为红色。（0.5 分）

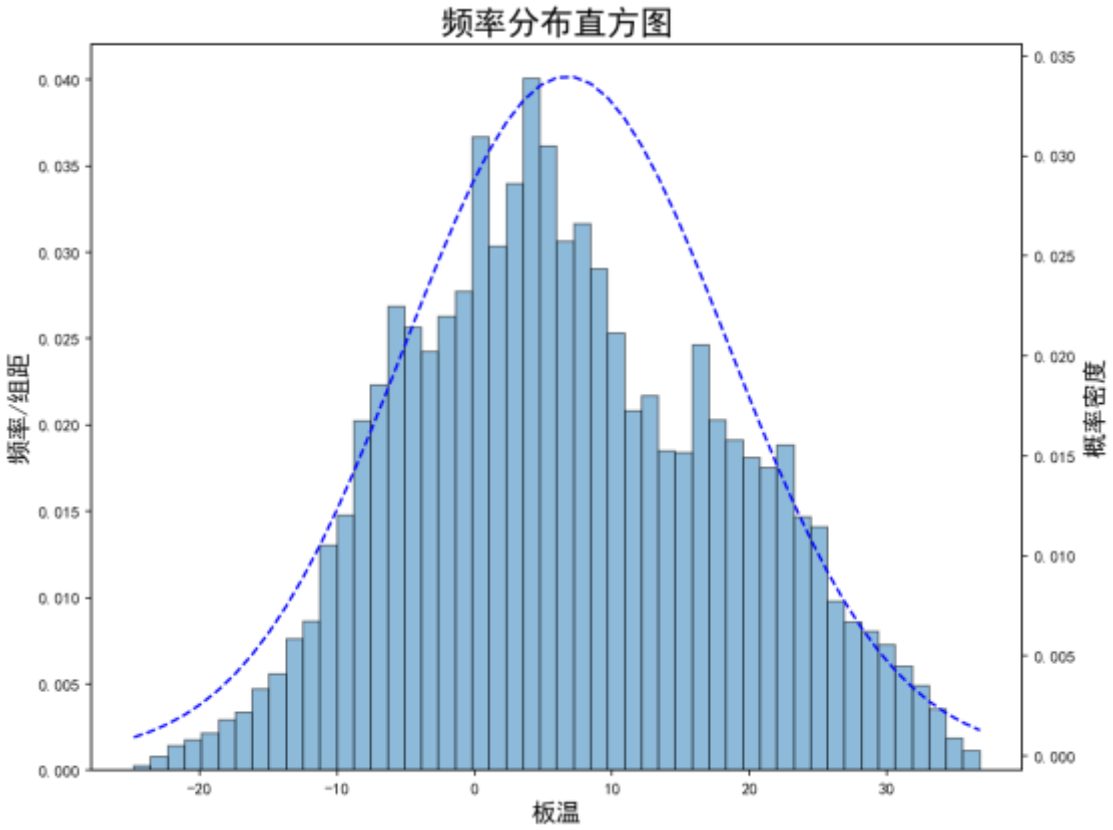


#TODO: 饼图

双轴分布图（2 分）

制作 板温 的频率分布直方图

- x 轴字体大小为 15，y 轴字体大小为 15
- 标题字体大小为 20
- 直方图图形类型正确（0.5 分）
- 直方图颜色为蓝色，透明度为 0.5，框线为黑色（0.5 分）
- 概率密度曲线以虚线表示（0.5 分）
- X 轴和 Y 轴标签正确，左轴为直方图 Y 轴，右侧为概率密度曲线轴（0.5 分）

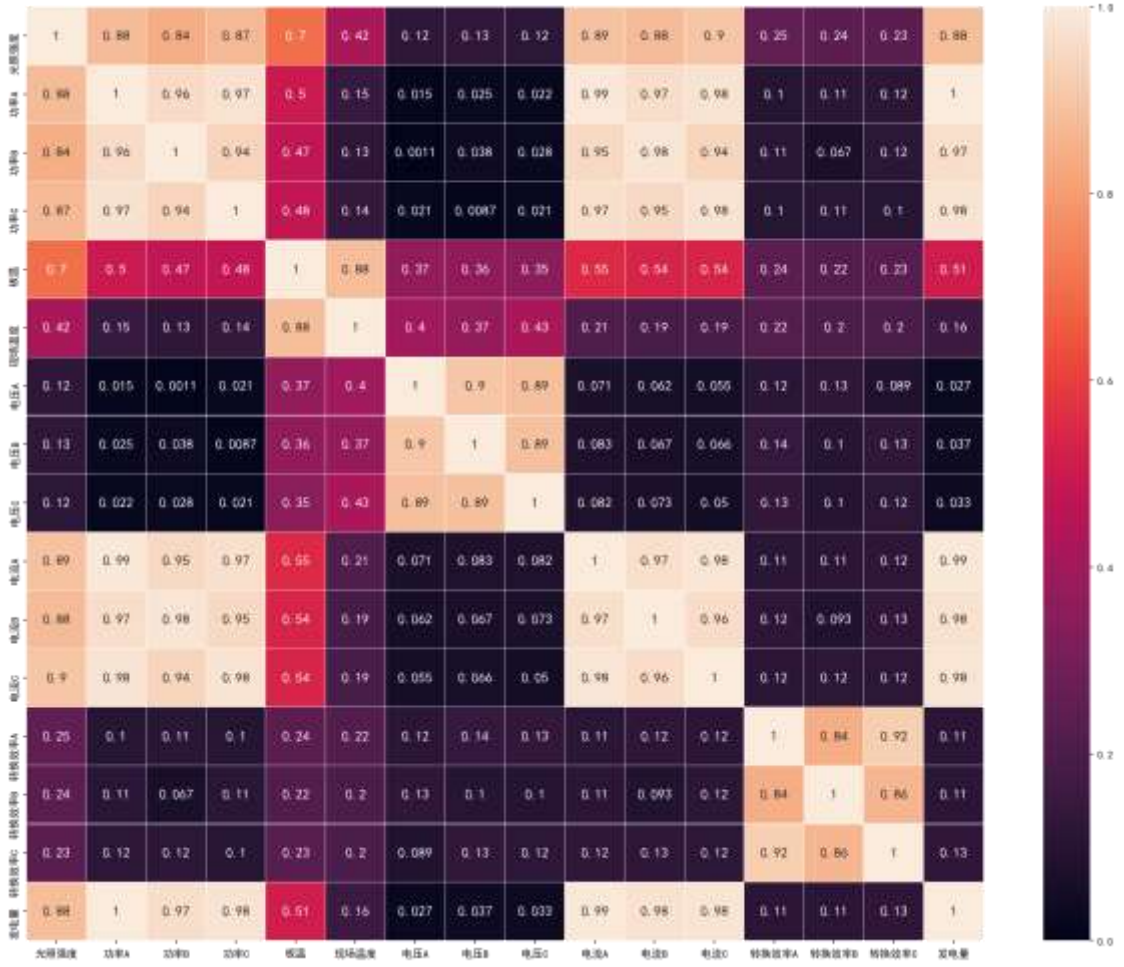


#TODO: 双轴分布图

热力图（2 分）

制作相关系数的热力图

- 图片大小为 (20, 16)，代码为 `plt.figure(figsize=(20, 16))`
- 图形类型正确（1 分）
- x 轴字体大小为 10，y 轴字体大小为 10（0.5 分）
- 内部字体大小为 13，加粗（0.5 分）



#TODO: 热力图

任务 5：报告保存

竞赛过程中运行的所有 python **代码**与**结果**和步骤中的文字描述将进行保存，将代码保存为 pdf 文件：点击 `nteract` 左上角的 `File`，再点击 `Export PDF`。

任务结束

恭喜完成本任务，祝您取得好成绩。

模块 C：机器学习（30 分）

竞赛题目：半导体工业制造

背景介绍

随着物联网与人工智能自动化的发展，半导体制造系统已经基本上实现制造系统的信息化与智能化，能够通过传感器或过程测量，源源不断地收集半导体制造过程中的生产数据。连续不断的制造过程、多种多样的传感设备以及实时高效的数据传输，使得半导体制造数据具备了规模性、多样性和高效性等典型的大数据特征。

面对大量信号数据，如果我们将每种类型的信号视为特征，那么工程师则可以利用这些信号来确定半导体制造过程中的各传感器内部线测试是否完成传递过程，利用机器学习方法对信号进行预分类可以有效的预测传递结果，以增强当前业务技术，提高产品工艺质量，目前这种方法已成为半导体制造过程监测系统的主要发展方向。

本次数据为半导体制造工艺数据集，目标为训练一个可以智能监测半导体制造过程的分类模型，评估传感器内部线测试是否顺利通过（二值分类问题），并在测试数据集上尝试获得最优结果。

环境介绍

在本竞赛模块环境中，使用 nteract 开始任务。环境中已安装 python 中多种数据分析和机器学习库，包括：

- Numpy
- Scipy
- Pandas
- Matplotlib
- Seaborn
- Scikit-learn
- Xgboost
- Lightgbm
- 其他

数据描述

数据由 2 个文件组成，训练数据集文件内含 1253 个样本组成，每个样本包含 591 个特征和 1 个标签（ 1253*592 矩阵），测试数据集文件内含 314 个样本，每个样本包含 591 个 特征（ 314*591 矩阵）。

属性	特征

数据集特征	多元
属性特征	真实
训练集样本数量	1253
测试集样本数量	314
属性数量	591
相关任务	分类

注意：标签文件中二分类字段 1 对应传递失败，0 对应传递成功。

任务描述

1. 数据集

- data/train.csv： 训练集
- data/test.csv：测试集

2. 结果提交形式：

参赛者需对测试集中分类结果进行预测并提交，结果提交需按照以下格式，否则视为无效提交，不计成绩。

- 预测结果保存在 result 目录下，命名为 predict.csv。
- 训练模型保存到 result 目录下，命名为 model.joblib。
- 完成竞赛时请保存代码文件为 pdf 格式和 ipynb 格式。

3. 评分标准:

参赛者在提交结果后，根据以下规则对结果进行评分：

最终得分 = $20*(\text{roc\_auc\_score}*100-60)^3*20^3$

参照预测结果，以 roc_auc_score 得分作为评价标准，结果保留两位小数。

任务列表

1. 数据获取
2. 数据探查
3. 数据处理
4. 模型训练
5. 结果保存

任务启动

通过 nteract 打开 /home/wiseinsight/Desktop/Tasks/Task_C.ipynb 文件，开始竞赛任务。

若您不小心关掉 nteract，可以打开终端输入 nteract 重新启动。

```
#导入所需库
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

#设置笔记环境
%matplotlib inline
%config InlineBackend.figure_format = 'retina'
%config IPCompleter.greedy = True
%config IPCompleter.use_jedi = True
pd.options.display.max_colwidth = 100
plt.rcParams['figure.figsize'] = (12, 8)
```

Step1. 数据获取

- 训练数据路径为 data/train.csv
- 测试数据路径为 data/test.csv

```
#数据获取
train_path = 'data/train.csv'
test_path = 'data/test.csv'
train = None
test = None
train.shape, test.shape
```

Step2. 数据探查

数据探查是对数据质量的检验。数据探查阶段为团队提供了指导，可以快速和分析数据中的异常数据，初步了解数据特征。

```
#TODO: 数据探查
```

Step3. 数据处理

本题目为二值分类任务，每个信号信息对传递结果都会有影响，所以需要对数据进行处理，下面一些步骤是非必需进行的，以你自己的判断设计这个过程。

- 特征工程
- 特征提取
- 特征选择
- 数据清洗
- 空值处理
- 重复值处理
- 数据降维
- 判断是否有必要进行数据降维，可以多次运行根据最终结果来进行反复修改

```
#TODO: 数据处理
```

Step4. 模型训练

- 请自己选用合适的模型或改善此模型的参数以提高分数，包括但不限定以下模型：
- sklearn.ensemble.ExtraTreesClassifier
- sklearn.ensemble.RandomForestClassifier
- sklearn.ensemble.AdaBoostClassifier
- sklearn.ensemble.GradientBoostingClassifier
- sklearn.gaussian_process.GaussianProcessClassifier
- sklearn.linear_model.LogisticRegression
- sklearn.linear_model.RidgeClassifier
- sklearn.linear_model.SGDClassifier
- sklearn.naive_bayes.GaussianNB
- sklearn.neighbors.KNeighborsClassifier
- sklearn.neural_network.MLPClassifier
- sklearn.svm.SVC
- sklearn.svm.NuSVC
- sklearn.tree.DecisionTreeClassifier
- sklearn.tree.ExtraTreeClassifier
- xgboost.XGBClassifier
- lightgbm.LGBMClassifier

```
#TODO: 数据拆分

#TODO: 数据预处理

#TODO: 模型训练和预测
model = None

y_pred = None
```

Step5. 结果保存

- 预测结果保存在 result 目录下，命名为 predict.csv。
- 训练模型保存到 result 目录下，命名为 model.joblib。
- 完成竞赛时请保存代码文件为 pdf 格式和 ipynb 格式。

```
#保存结果
import joblib
np.savetxt('result/predict.csv', y_pred)
joblib.dump(model, 'result/model.joblib')
```

将代码保存为 pdf 文件：点击 nteract 左上角的 File，再点击 Export PDF。

任务结束

恭喜完成本任务， 祝您取得好成绩。

答辩（20 分）

1. 什么是探索性数据分析，它的作用是什么？

2. 对数据进行描述性分析主要包括哪几类，简要阐述一下每一类。

3. 从热力图颜色层次不同，从中能得到哪些可用信息？

4. 根据发电量的线形图，能看出什么？针对此能提出哪些对光伏发电企业有创新性的意见，改善传统业务方案？

5. 对光伏发电进行数据分析的目的和意义在哪里？