《大数据原理与应用》实验作业  
 2020 / 2021 年 第 一 学期

Hadoop/Spark集群环境搭建  
及疏散星团NGC2266的大数据处理

团 队/姓 名 湖人总冠军 1

学 号 \_\_\_\_\_\_15030339\_\_钟冠霆\_\_\_\_

\_\_\_\_\_\_17030303\_\_陈华志\_\_\_\_

\_\_\_\_\_\_17030328\_\_谢雨辰\_\_\_\_\_\_

\_\_\_\_\_\_17030320\_\_皮心怡\_\_\_\_\_\_

\_\_\_\_\_\_17011239\_\_朱智涛\_\_\_\_\_\_

班 级 17软件一 1

任课老师 王 伟 华 1

**计算机信息工程学院  
2020年12月**

**目录**

**[《大数据原理与应用》实验作业 2020 / 2021 年 第 一 学期 1](#_Toc15120)**

**[一、PySpark环境搭建 2](#_Toc16014)**

**[1、Python（推荐使用Anaconda+PyCharm） 3](#_Toc15045)**

**[2、JDK 3](#_Toc447)**

**[3、Scala（Scala-2.12.8） 6](#_Toc19170)**

**[4、Spark（Spark 2.1-preview for hadoop-2.7） 7](#_Toc25206)**

**[5、Hadoop（Hadoop 2.7） 8](#_Toc6967)**

**[6、winutils.exe 9](#_Toc10536)**

**[7、Python相关库的安装和设置 9](#_Toc29347)**

**[二、聚类分析 10](#_Toc10364)**

**[1.导入pyspark相关的库 10](#_Toc19336)**

**[2.使用SparkSession导入数据 11](#_Toc25164)**

**[3.数据处理 12](#_Toc30007)**

**[4.特征工程 13](#_Toc29190)**

**[5.调用KD树搜索算法 13](#_Toc26965)**

**[6.绘制K-dist图 13](#_Toc20193)**

**[7.确定DBSCAN算法的参数并进行聚类分析 17](#_Toc15814)**

**[8.结果可视化 17](#_Toc19628)**

**[三、总结 18](#_Toc7850)**

# 一、PySpark环境搭建

Spark需要由JDK，Scala和Hadoop环境的支持，而PySpark则是利用Spark支持Python的特性而将其当作Python的一个包调用，利用PySpark中的Py4j库,我们可以通过Python语言操作RDDs，在使用PySpark的时候，我们需要先完成以下组件的安装

* **1、Python（推荐使用Anaconda+PyCharm）**
* **2、JDK**
* **3、Scala（Scala-2.12.8）**
* **4、Spark（Spark 3.0-preview for hadoop-2.7）**
* **5、Hadoop（Hadoop 2.7）**
* **6、winutils.exe**
* **7、Python相关库的安装和设置**

#### 1、Python（推荐使用Anaconda+PyCharm）

建议使用Anaconda，里面集成很很多常用的包，而且安装后自带的jupyter notebook也挺好用的；

下载路径：

[https://www.anaconda.com/​](https://link.zhihu.com/?target=https://www.anaconda.com/" \t "https://zhuanlan.zhihu.com/p/_blank)

#### 2、JDK

[https://www.oracle.com/cn/java/technologies/javase/javase-jdk8-downloads.html​](https://link.zhihu.com/?target=https://www.oracle.com/cn/java/technologies/javase/javase-jdk8-downloads.html" \t "https://zhuanlan.zhihu.com/p/_blank)

安装完成以后，配置环境变量。配置环境变量的方法为电脑[右键]——>属性——>环境变量，编辑环境变量

配置Java环境变量主要有三个：

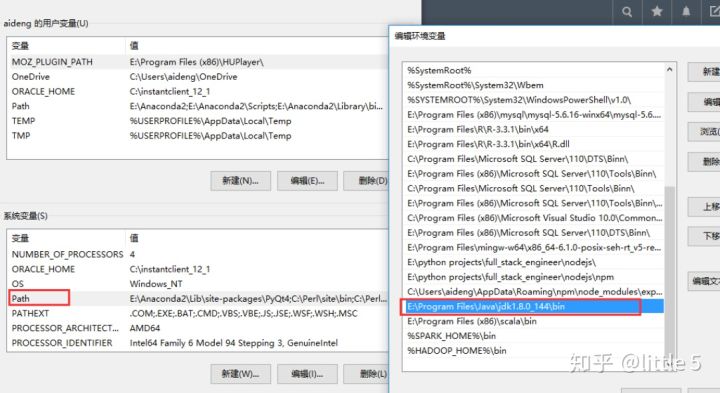
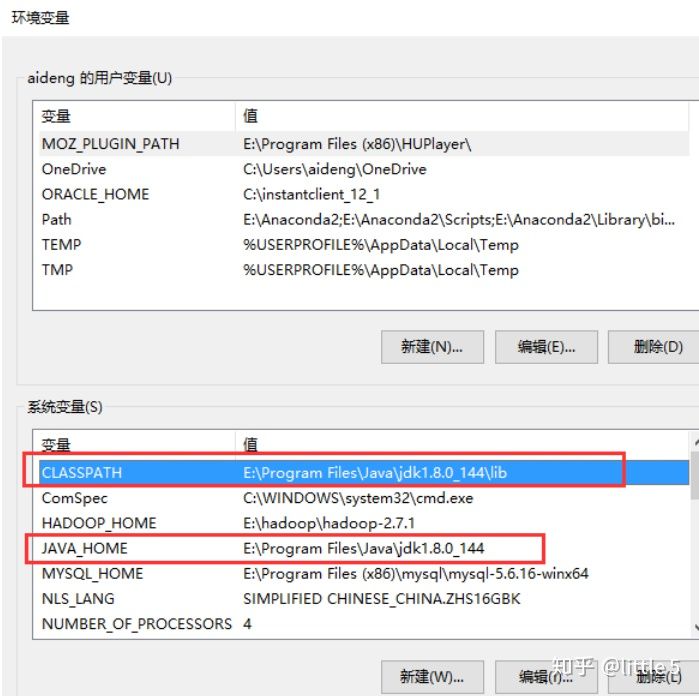
JAVA\_HOME: E:\Program Files\Java\jdk1.8.0\_144\

CLASS\_PATH: E:\Program Files\Java\jdk1.8.0\_144\lib

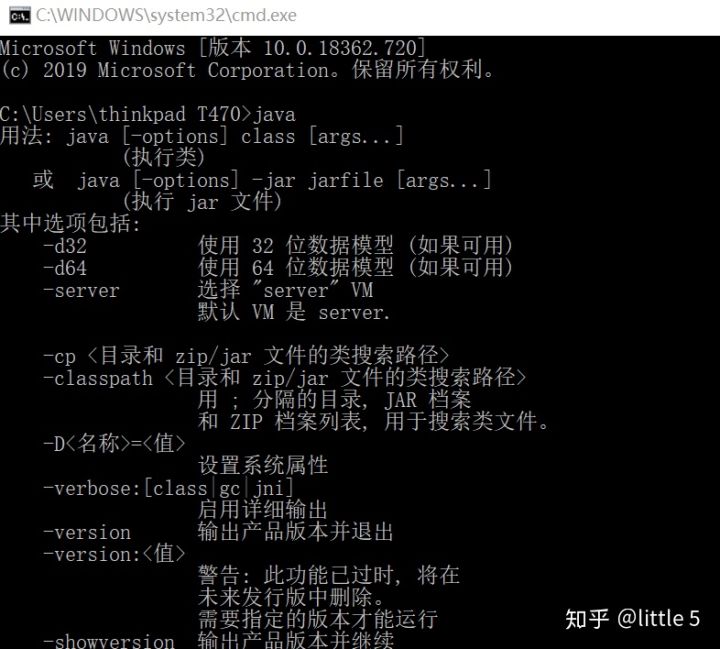
Path: E:\Program Files\Java\jdk1.8.0\_144\bin

其中Java\_Home变量则为Java安装路径，CLASS\_PATH可以看成Java安装目录下lib文件目录，Path一般为系统自带变量，修改时直接新增Java安装目录下的bin目录。

一般来讲bin目录下大多是可执行文件，XX\_HOME指程序安装目录（下面Scala、Hadoop、spark均需要执行增加Home变量，和bin相关操作）。



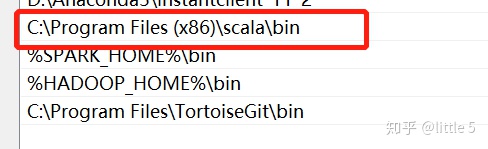
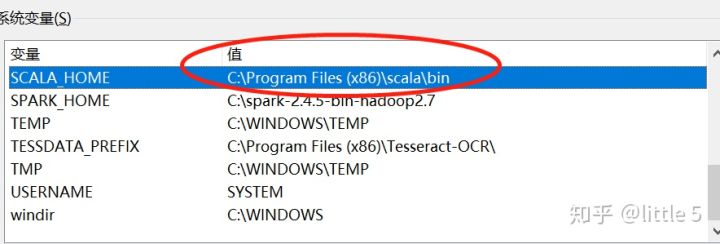
Java如果安装成功，则在cmd窗口中分别输入java 和javac，如果均出现如下图所示结果，则表示安装Java成功



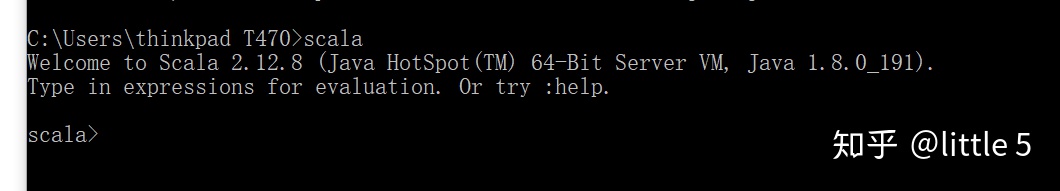
#### 3、Scala（Scala-2.12.8）

[All Available Versions​www.scala-lang.org](https://link.zhihu.com/?target=https://www.scala-lang.org/download/all.html" \t "https://zhuanlan.zhihu.com/p/_blank)

从上面获取对应Spark版本的Scala安装包，我选择了2.12.8版本，windows下直接下载.msi文件安装即可，然后配置环境变量配置好



安装好后运行cmd命令提示符，输入Scala后，如果能够正常进入到Scala的交互命令环境则表明安装成功。

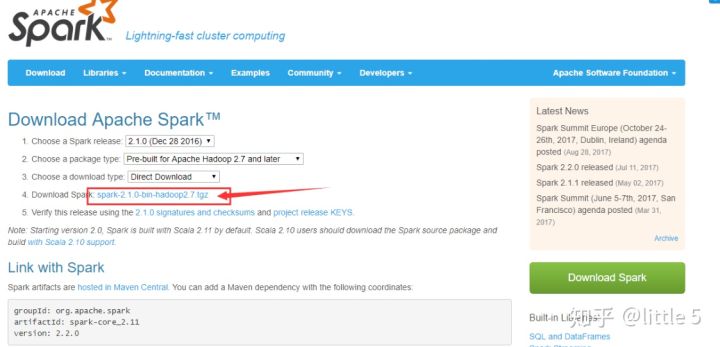


#### **4、Spark（Spark 2.1-preview for hadoop-2.7）**

spark下载地址在

[Downloads | Apache Spark​spark.apache.org](https://link.zhihu.com/?target=http://spark.apache.org/downloads.html" \t "https://zhuanlan.zhihu.com/p/_blank)

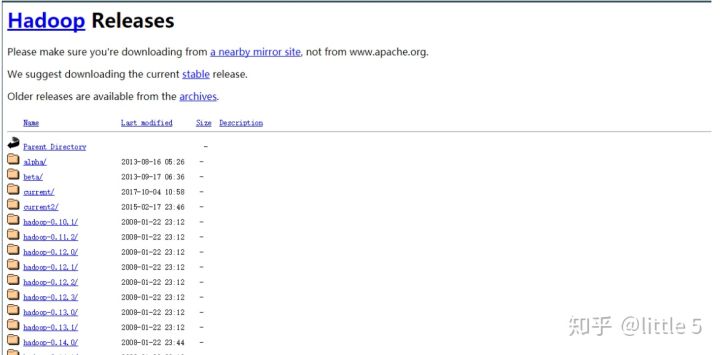
，选择合适的版本安装即可。



spark\_home和bin环境变量配置需要设置，和Scala的配置步骤是一样的，就2个地方需要设置，上一张Scala的截图已经有spark的

#### 5、Hadoop（Hadoop 2.7）

由于spark是基于hadoop建立的，所以需要下载Hadoop，这里选择hadoop 2.7，然后下载后放到本地目录，和之前一样，设置相关的2个地方的环境变量即可！



#### 6、winutils.exe

winutil.exe下载位置在

[https://github.com/steveloughran/winutils​github.com](https://link.zhihu.com/?target=https://github.com/steveloughran/winutils" \t "https://zhuanlan.zhihu.com/p/_blank)

其中根据自己的hadoop版本选择对应的winutil.exe文件。

#### 7、Python相关库的安装和设置

1)，将spark所在目录下的pyspark文件夹拷贝到python文件夹下

2)，安装py4j库

一般的在cmd命令行下 pip install py4j 就可以

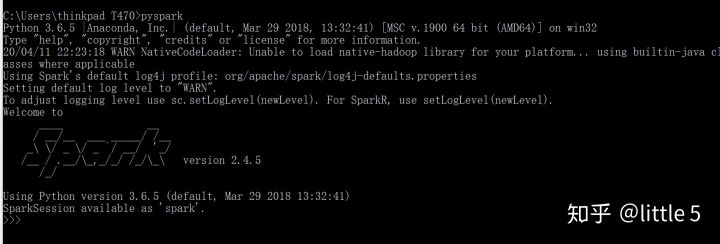
3) 修改权限

将winutils.exe文件放到Hadoop的bin目录下，然后以管理员的身份打开cmd，然后通过cd命令进入到Hadoop的bin目录下，然后执行以下命令：

winutils.exe chmod 777 c:\tmp\Hive

4)验证pyspark启动

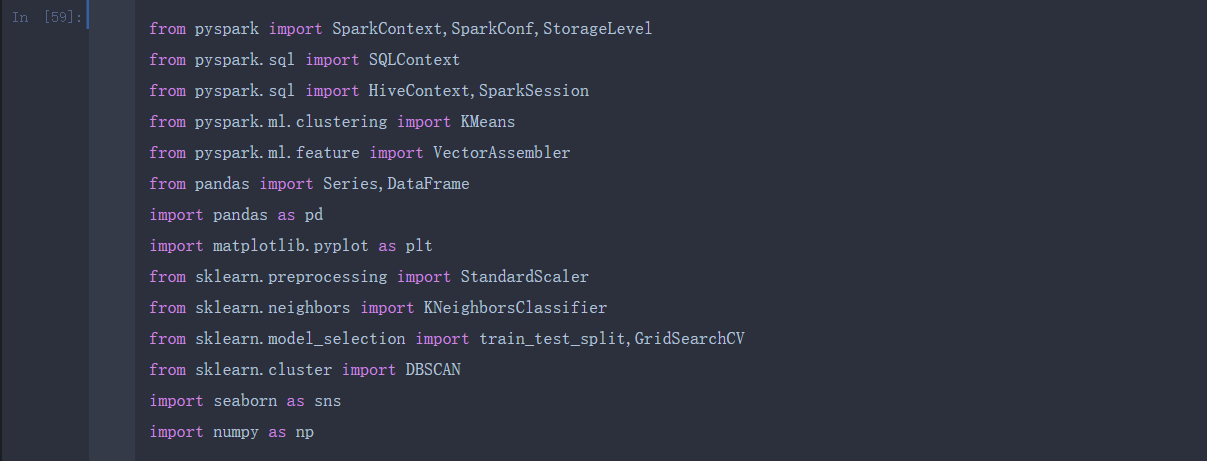
到此 所有的配置工作全部完成，最后再在cmd中输入pyspark，查看反馈。



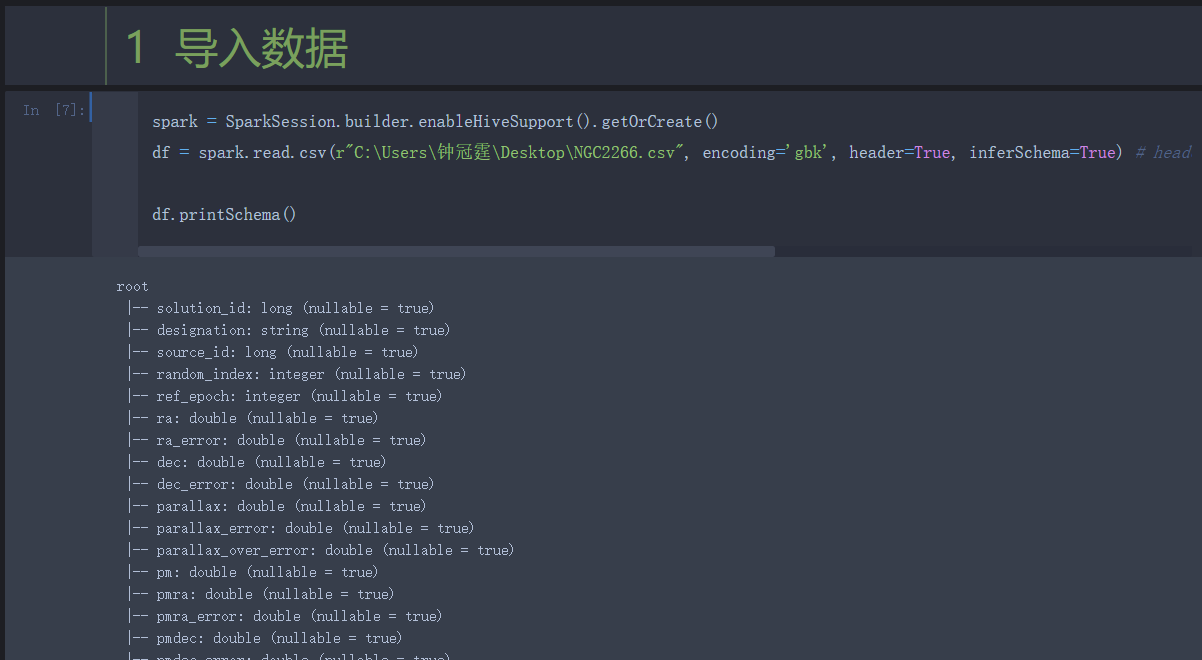
# 二、聚类分析

基于*GaiaDataRelease3*(*Gaia−DR*3)星表,采用数据挖掘技术中的DBSCAN(Density-Based Spatial Clustering of Applications with Noise)算法进行疏散星团成员判定.从*Gaia−DR3*中选取了8426颗恒星作为样本,使用恒星的五维数据(三维空间位置和两维自行) 进行聚类分析.在数据预处理阶段,将每一维数据标准化到[0*,* 1]区间内,避免了单位不一致对聚类效果的影响.然后,利用KD树最近邻搜索算法绘制*k−dist*图，确定了*DBSCAN*算法的输入参数(*Eps*,*MinPts*).最终,使用*DBSCAN*算法获取了一些成员星.

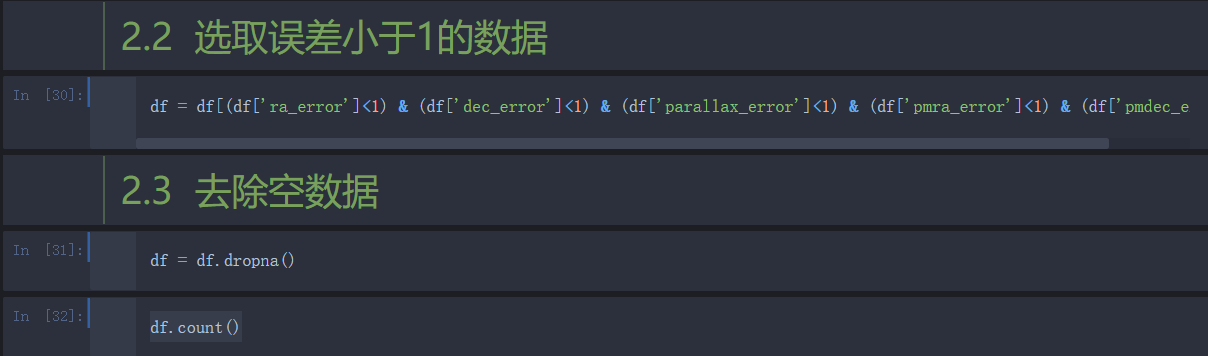
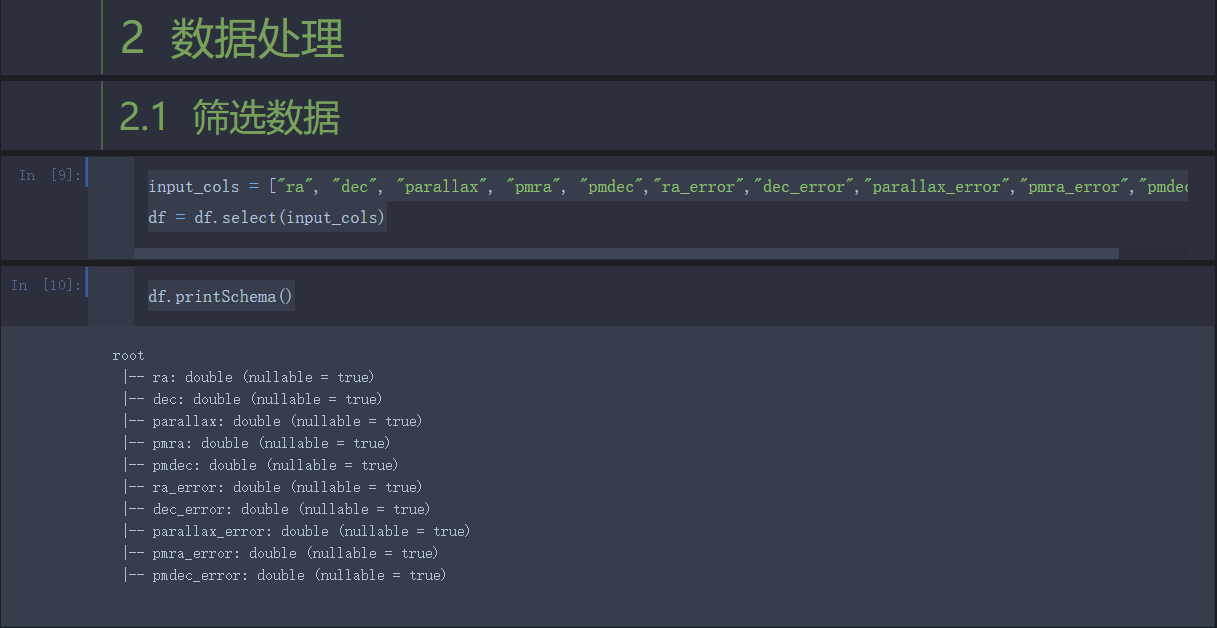
### 1.导入pyspark相关的库



### 2.使用SparkSession导入数据



### 3.数据处理



### 4.特征工程

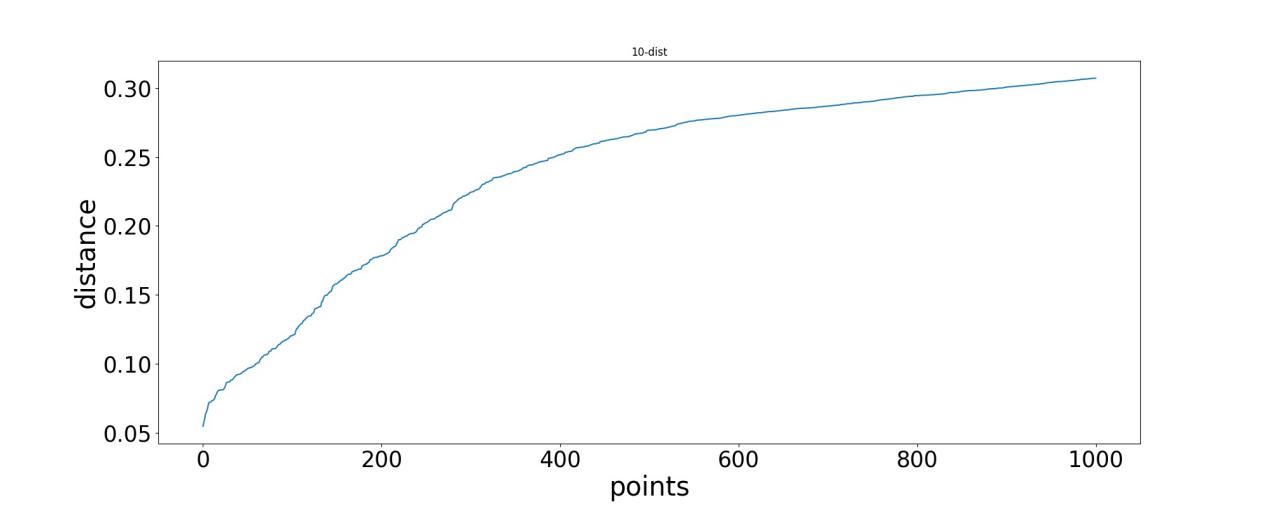
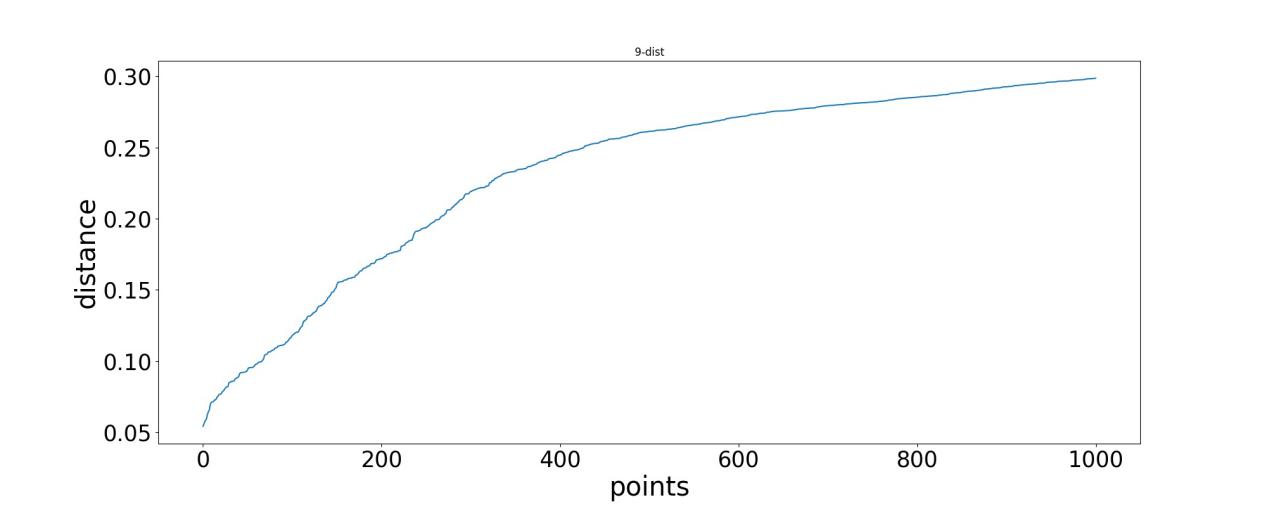
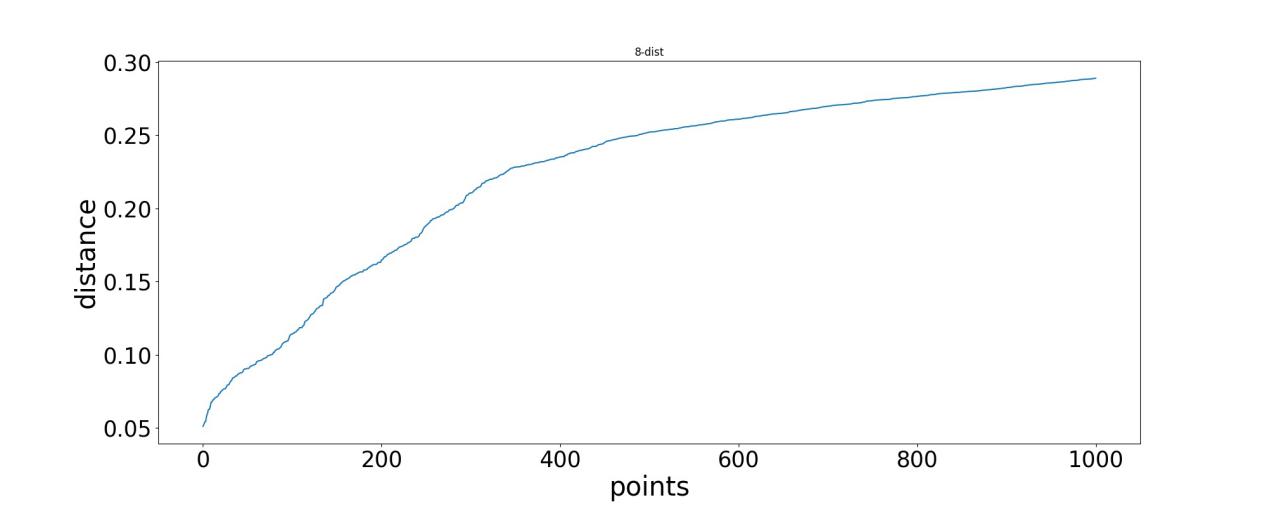
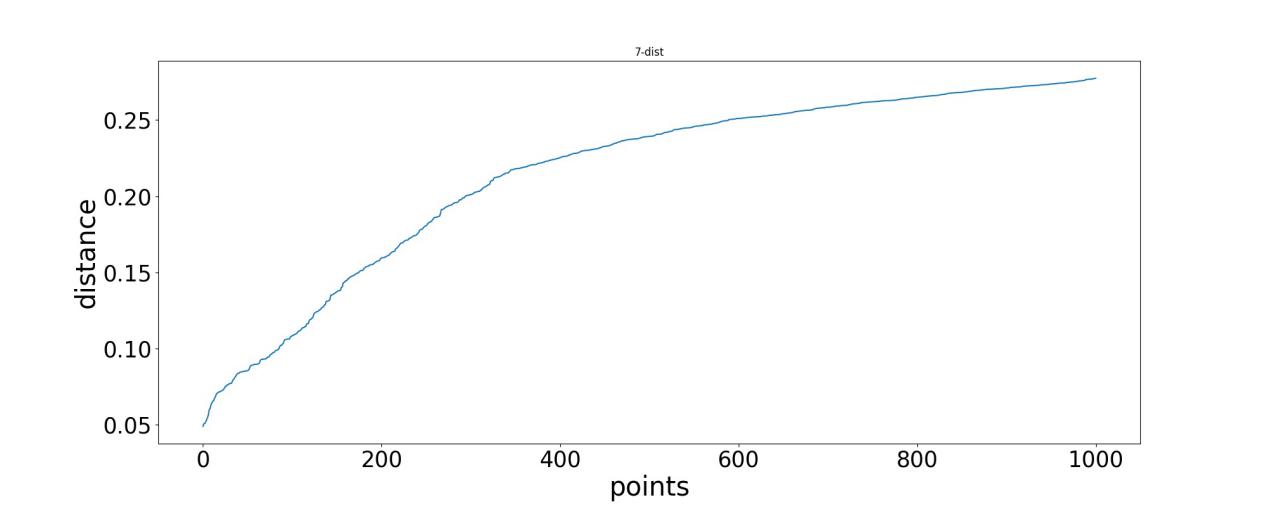
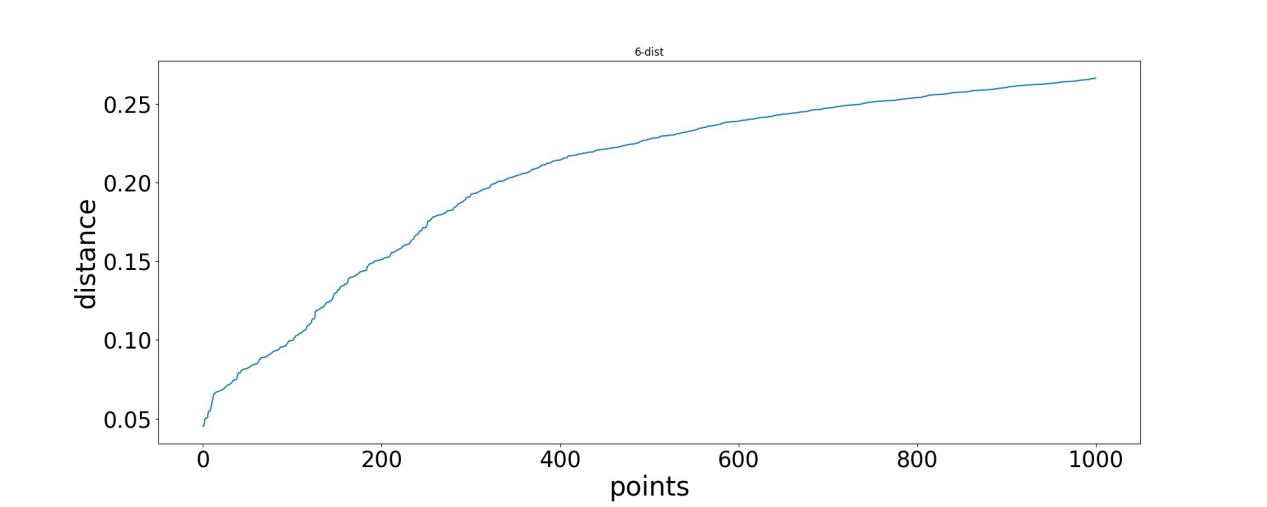
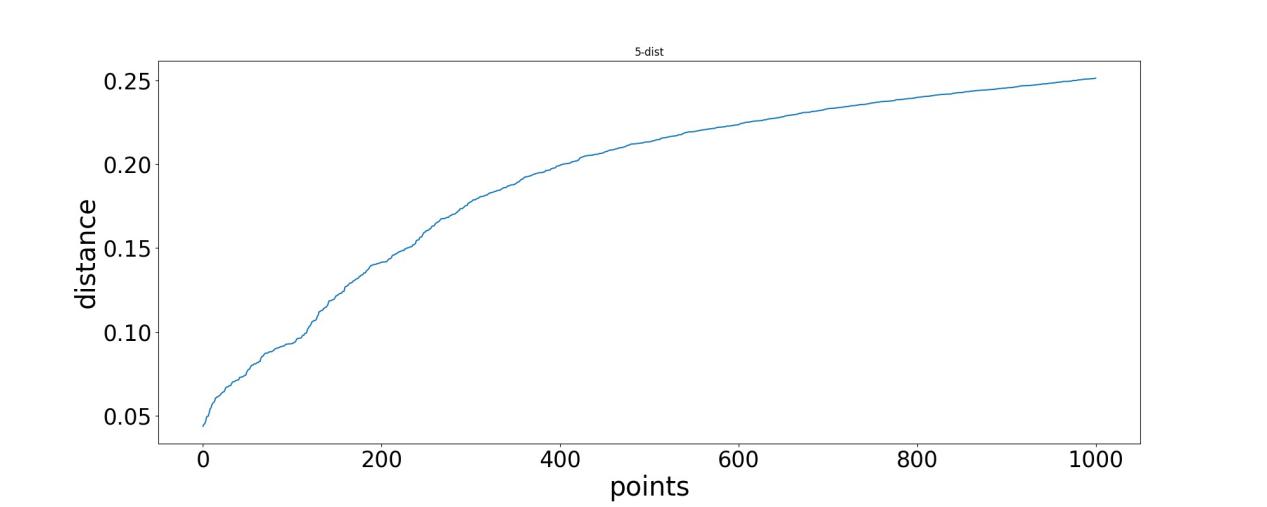
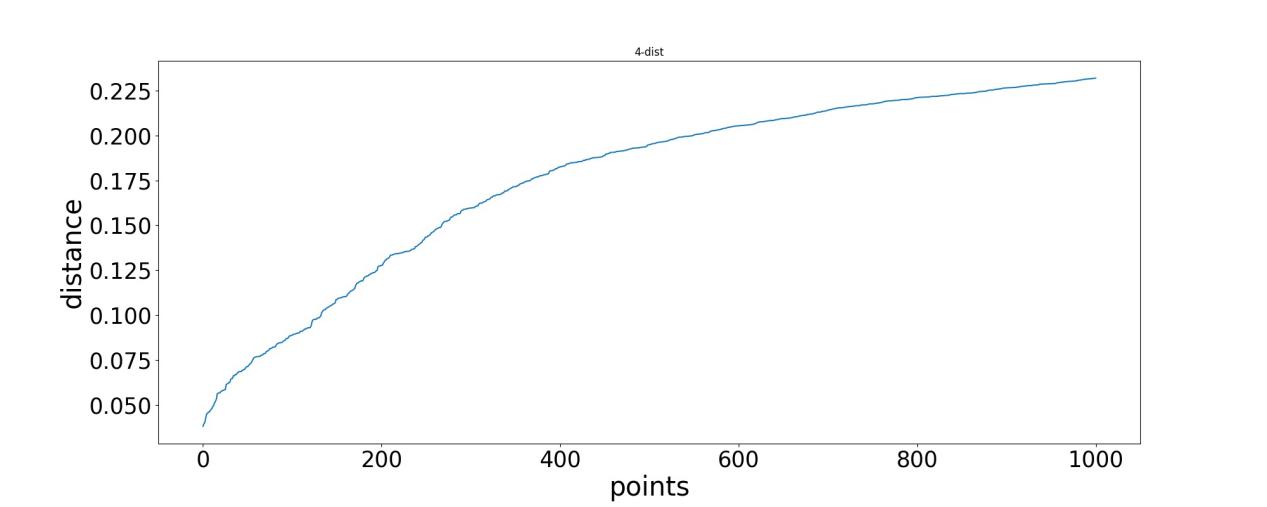
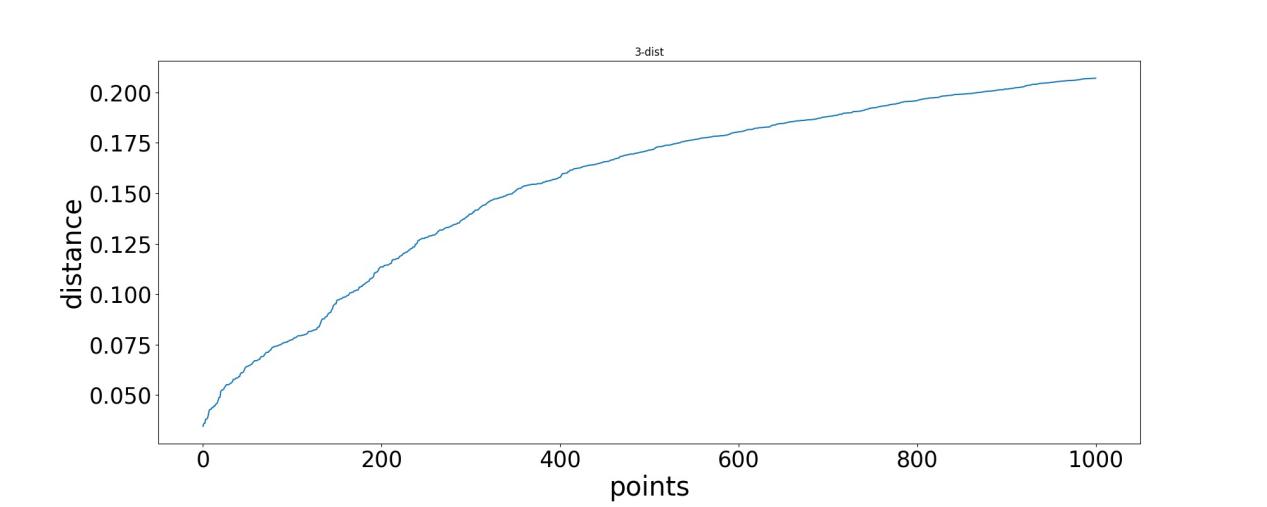
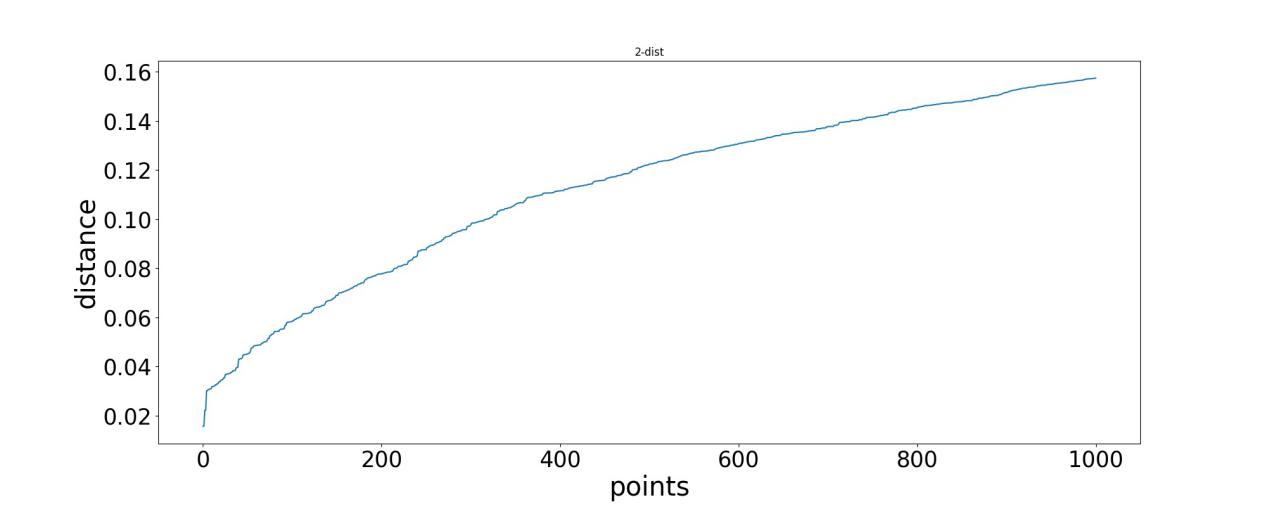


### 5.调用KD树搜索算法

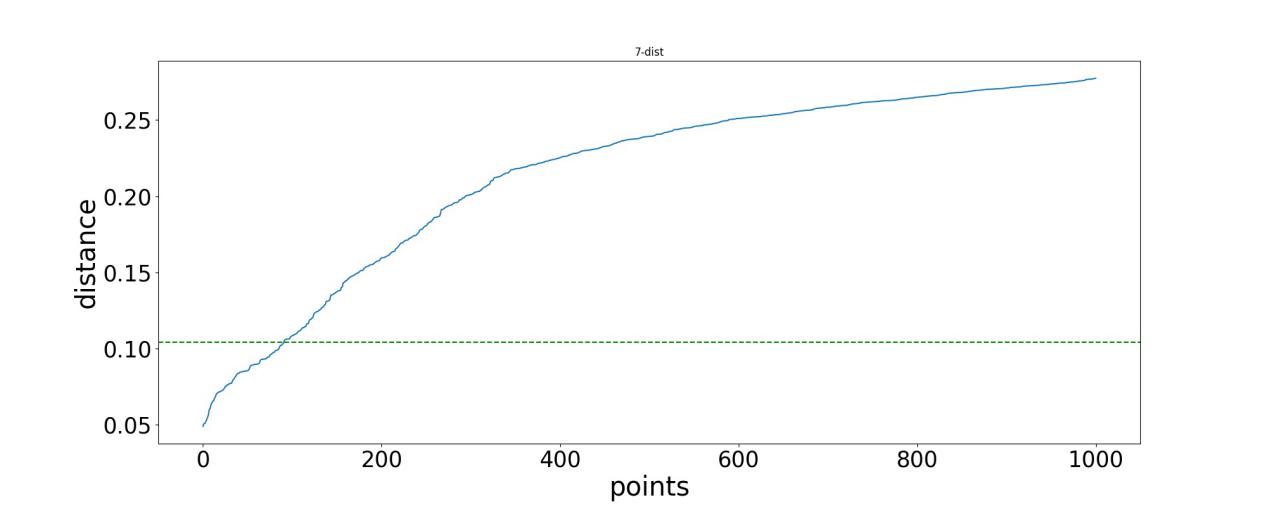


### 6.绘制K-dist图





从上述图中可以观察到，当K=7开始，图像不再发生大变化。由此可以确定DBSCAN算法的参数minPts=K+1=8，而从7-dist图中也可以清晰地看见，当distance=0.104时，即下图所示绿色直线，图像被分割成上下两块区域，两块区域内的图像趋势完全不同。由此可以确定DBSCAN算法的参数eps=0.104.

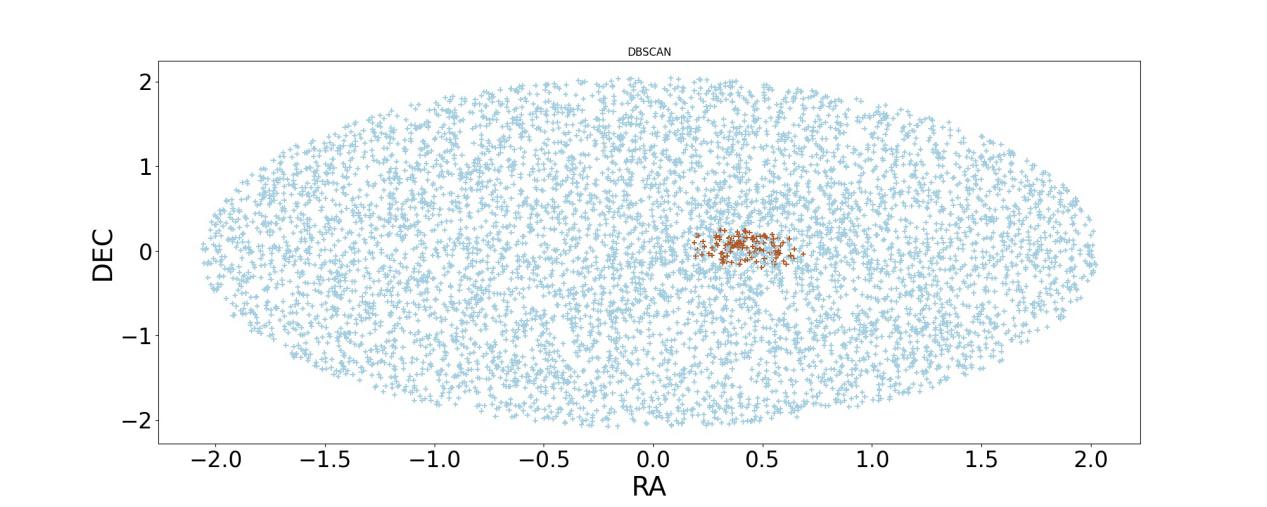
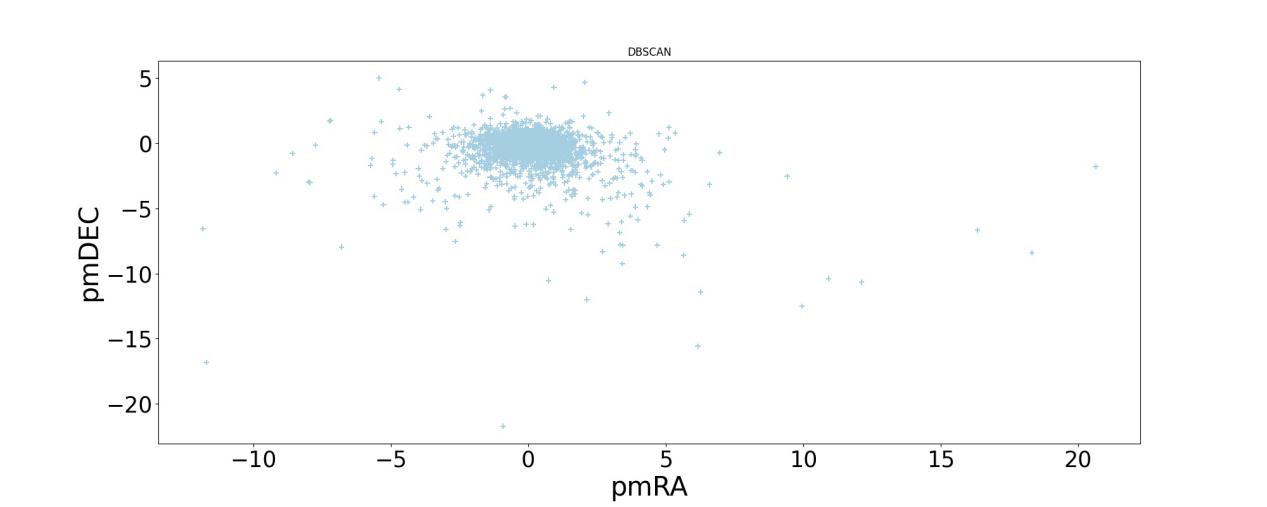


### 7.确定DBSCAN算法的参数并进行聚类分析



### 8.结果可视化





# 三、总结

在本次作业中，我们尝试着从头开始搭建pyspark环境，从hadoop-spark-python-scala等等，一步步直至完成环境搭建。期间遇到许多问题，通过查阅相关资料和团队讨论等方式解决，加深了对于大数据相关工具的理解，学会了如何使用pyspark进行聚类分析，同时调研了有关聚类算法如K-means、DBSCAN等。美中不足的是对于聚类后的结果仍然存在疑惑，有待继续深入学习。