**

**HUNAN UNIVERSITY**

《数据仓库技术与应用》

实验报告七

|  |  |
| --- | --- |
| **报告名称：** | 基于YELP数据集的数据分析 |
| **学生姓名：** | 杨超然 |
| **学生学号：** | 202106060220 |
| **专业班级：** | 电商2102班 |
| **学 院：** | 工商管理学院 |
| **指导老师：** | 李信儒 |
| **日 期：** | 2023.4.3 |

目录

[一、实验过程 3](#_Toc189529182)

[（一）实验环境配置 3](#_Toc698713373)

[（二） 数据集操作 4](#_Toc835078215)

[2.1数据集下载 4](#_Toc1353926360)

[（三） 数据分析 4](#_Toc703608908)

[3.1数据预处理 4](#_Toc1509956374)

[3.2数据分析 4](#_Toc1022521219)

[3.3数据可视化 7](#_Toc1349984439)

[二、易错点总结 11](#_Toc1023735718)

[（一）安装python插件时报错 11](#_Toc287232662)

[（二）数据分析过程报错 11](#_Toc2123595425)

[（二）数据分析过程报错 12](#_Toc90094835)

[三、 学习小记 12](#_Toc247920710)

[（一）数据可视化 12](#_Toc685097790)

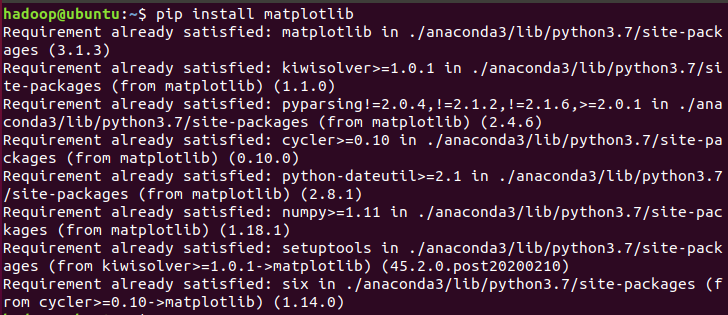
[（二） Matplotlib 12](#_Toc1778724963)

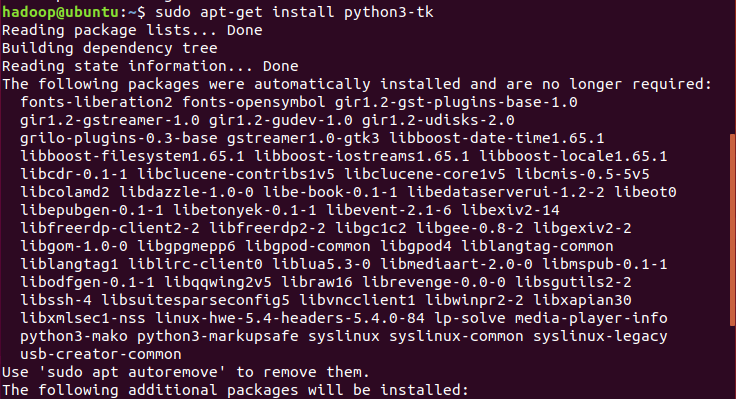
# 一、实验过程

## （一）实验环境配置

本次实验环境：Linux：Ubuntu 16.04；Hadoop3.1.3；Python：3.6；Spark：2.4.0；Jupyter Notebook 相关安装操作见实验六。

除此之外，还需安装相关插件以支持python可视化分析





## 数据集操作

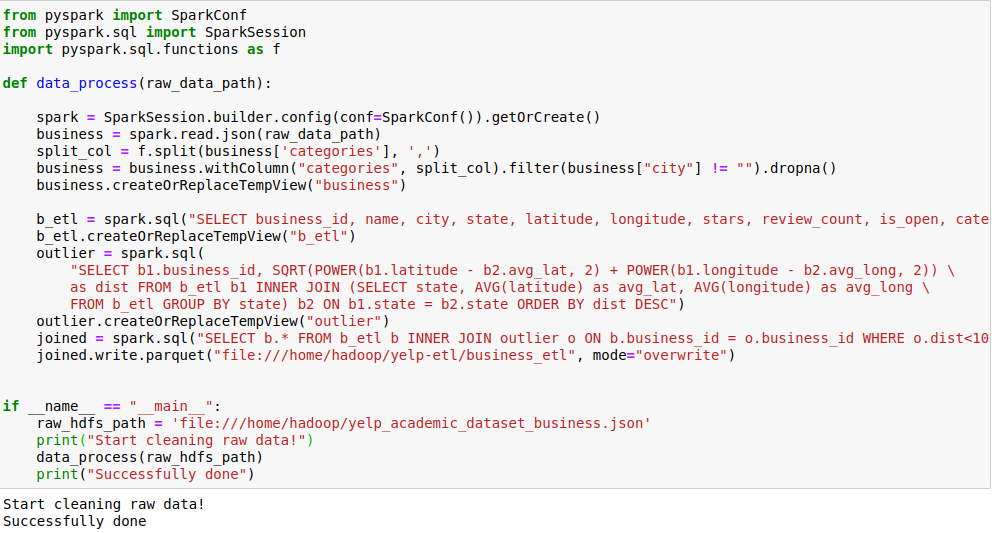
2.1数据集下载

本次实验数使用的数据集是来自Kaggle的Yelp数据集。这里选择了其中的yelp\_academic\_dataset\_business.json数据集。

## 数据分析

3.1数据预处理

注：不能照抄指南的代码，raw\_hdfs\_path 与 joined.write.parquet处要修改成自己的路径，成功运行结果如下：

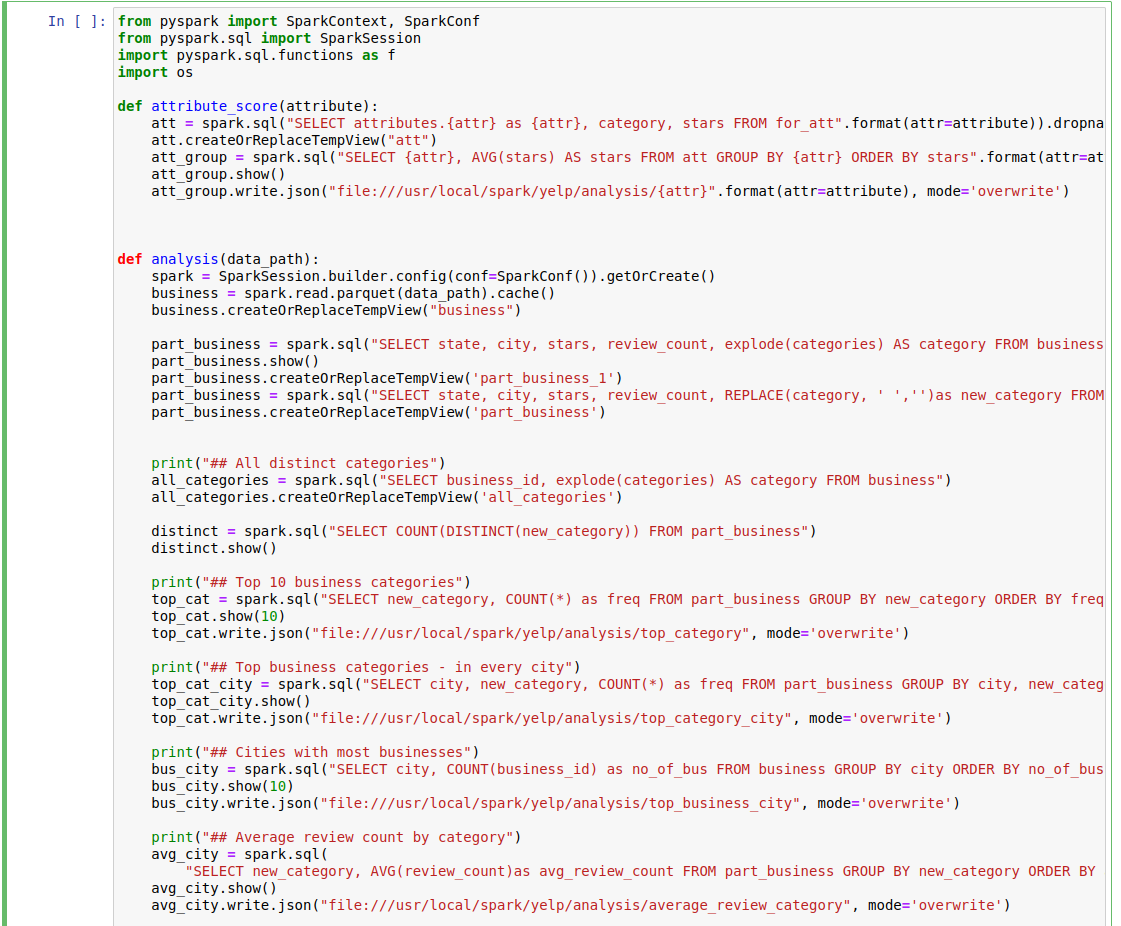


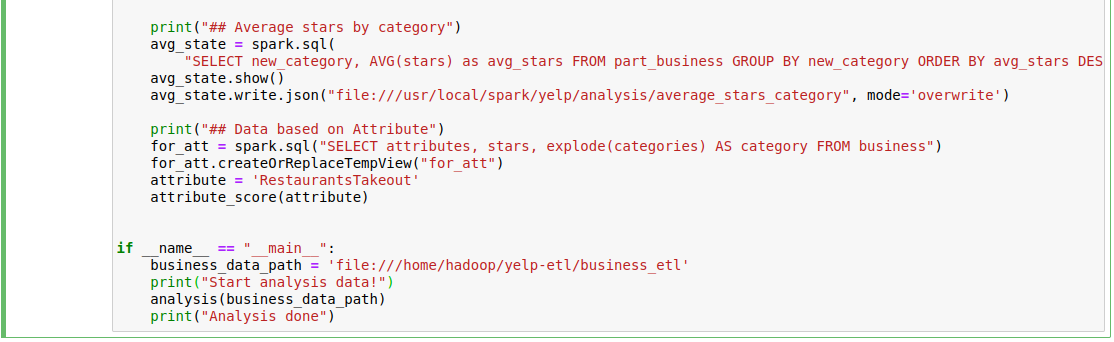
同时会在home / hadoop 目录下找到对应的文件夹



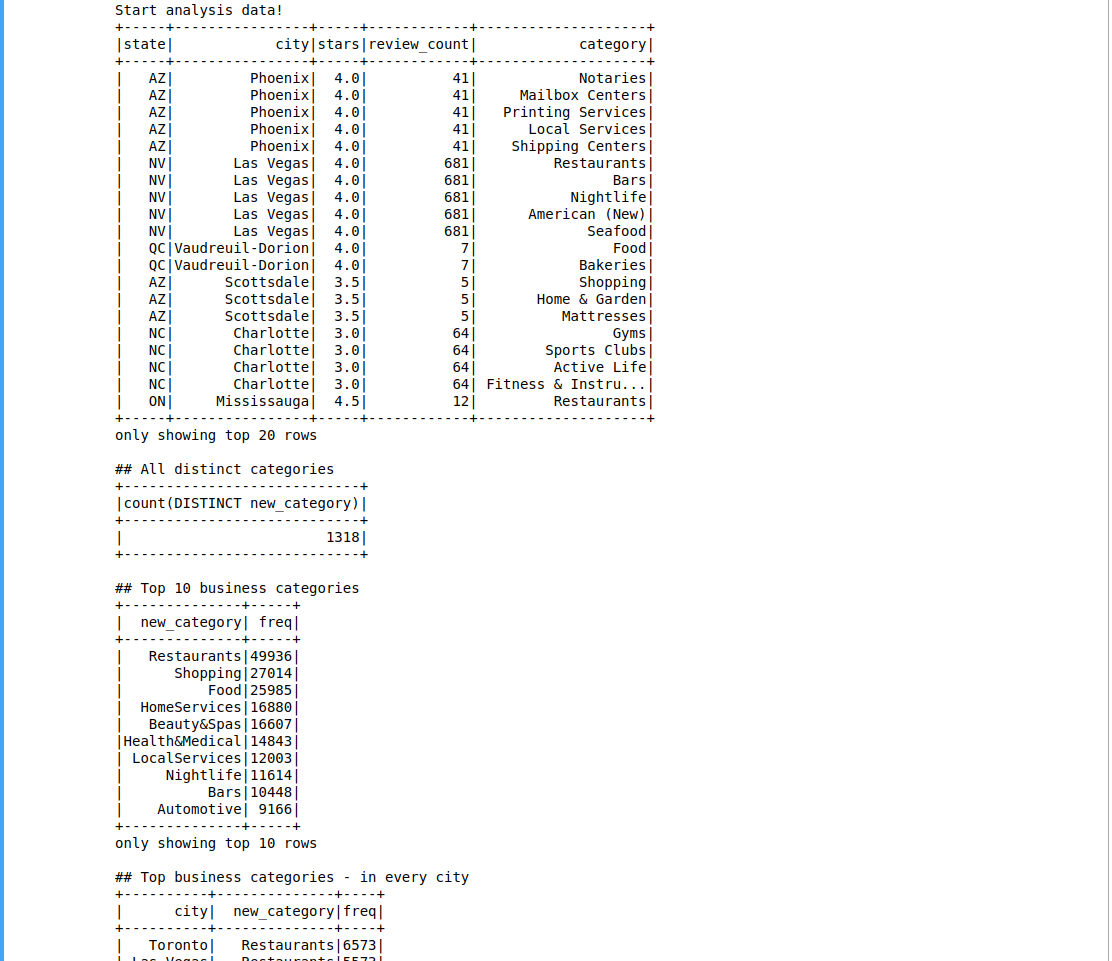
3.2数据分析

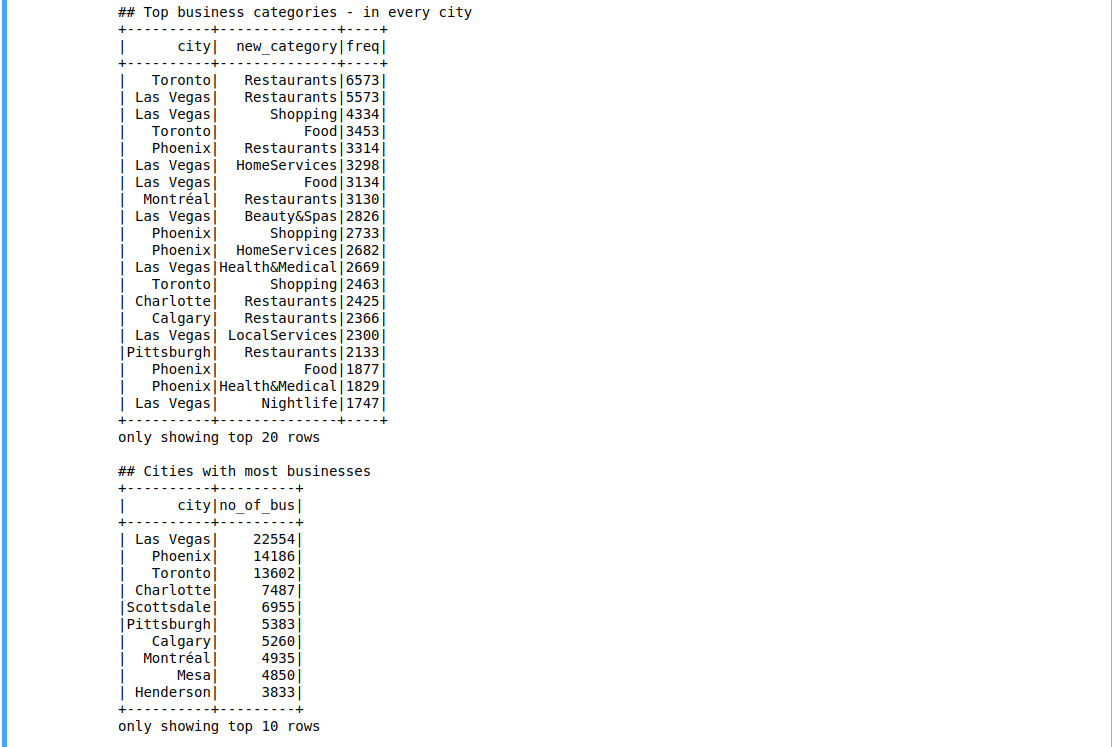
注：business\_data\_path处需要修改成自己的路径

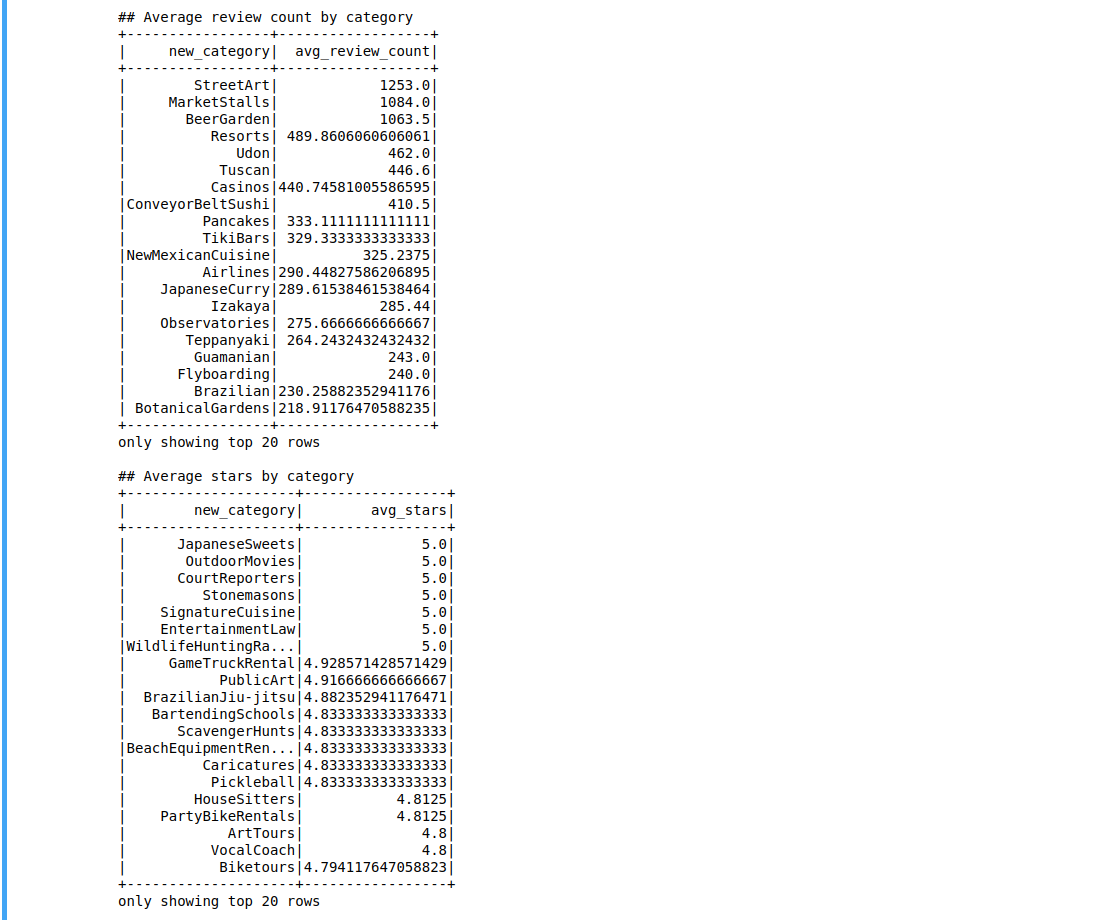


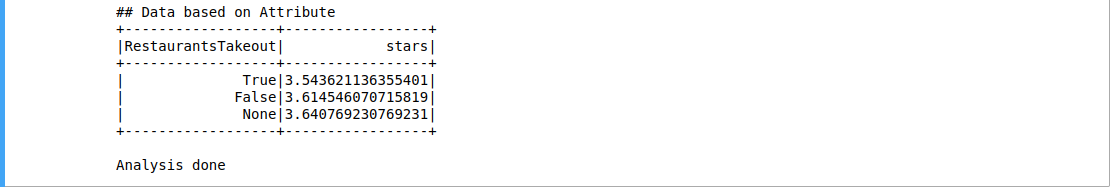


执行结果如下：

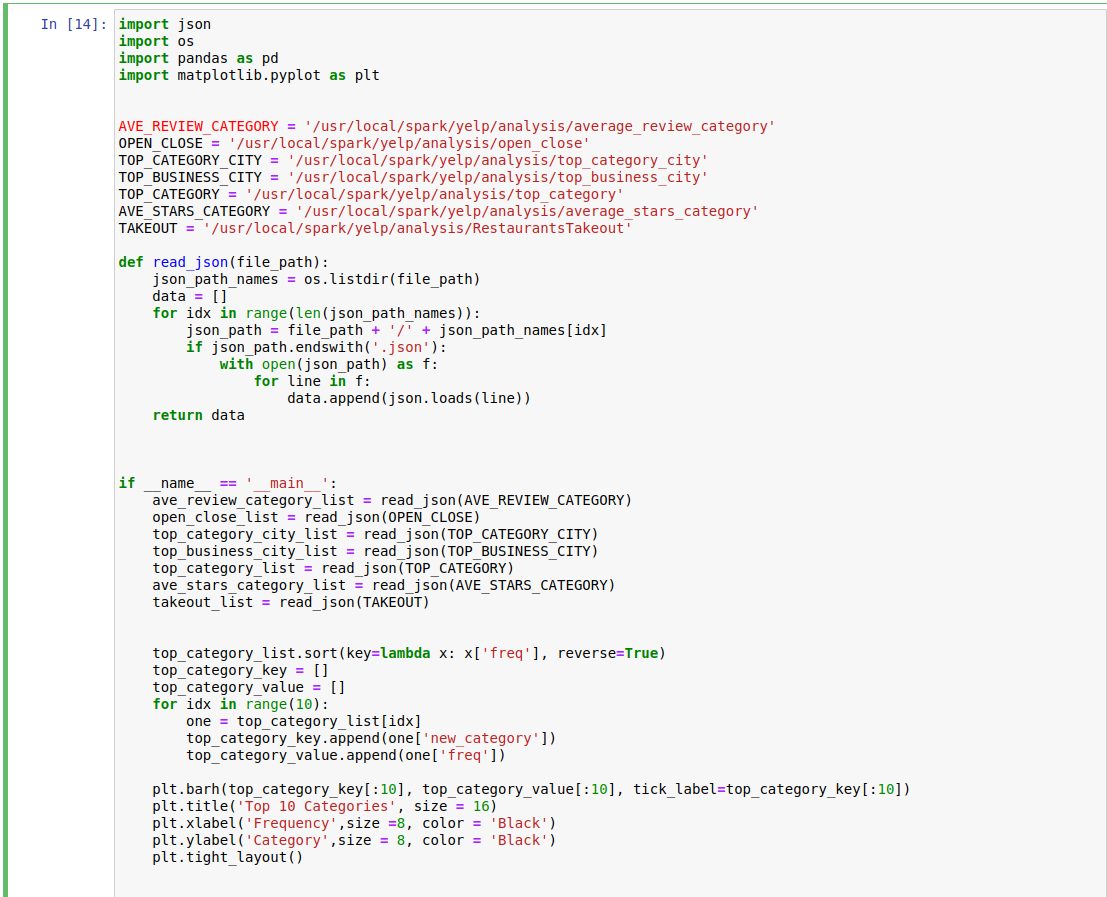


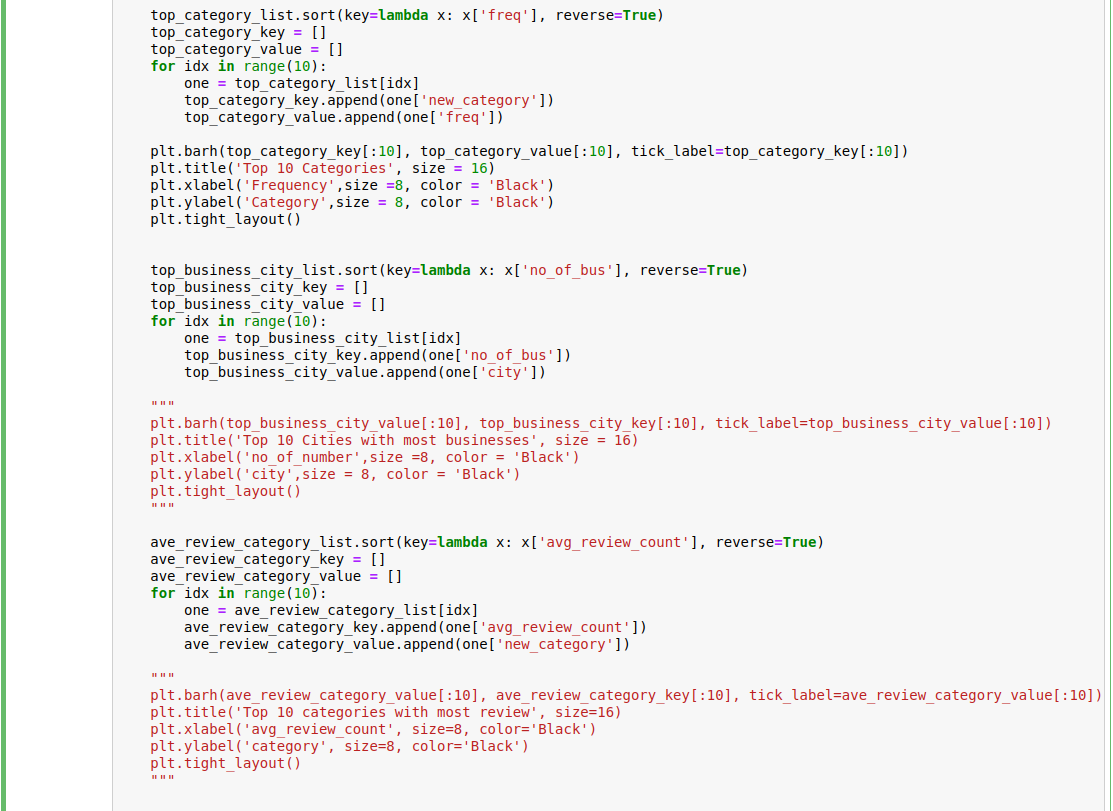






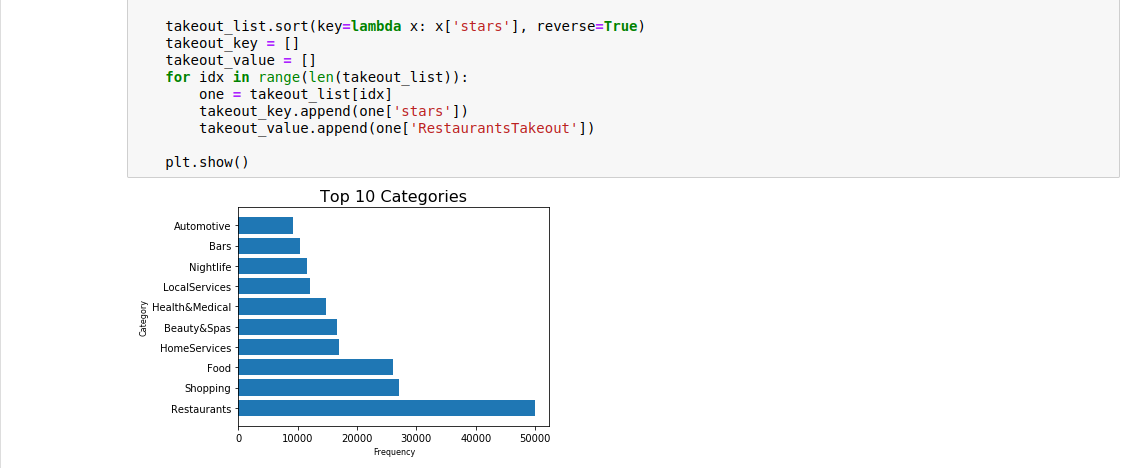
3.3数据可视化

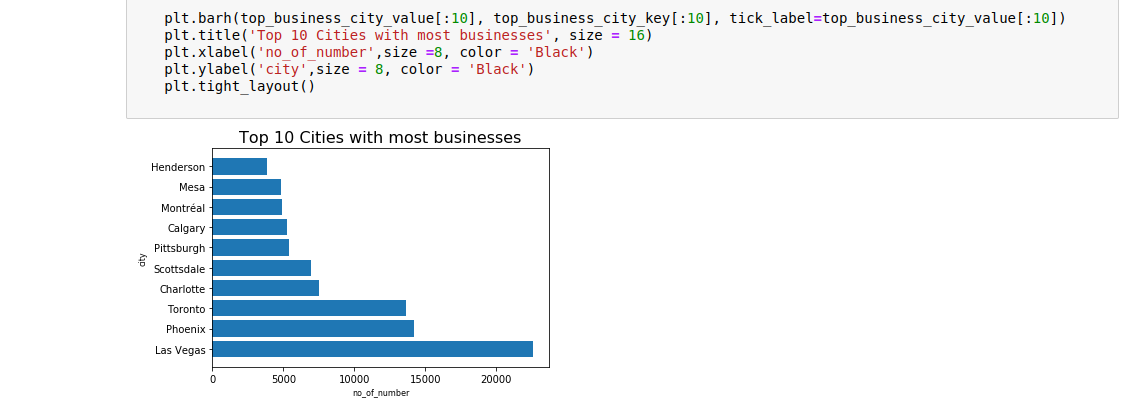


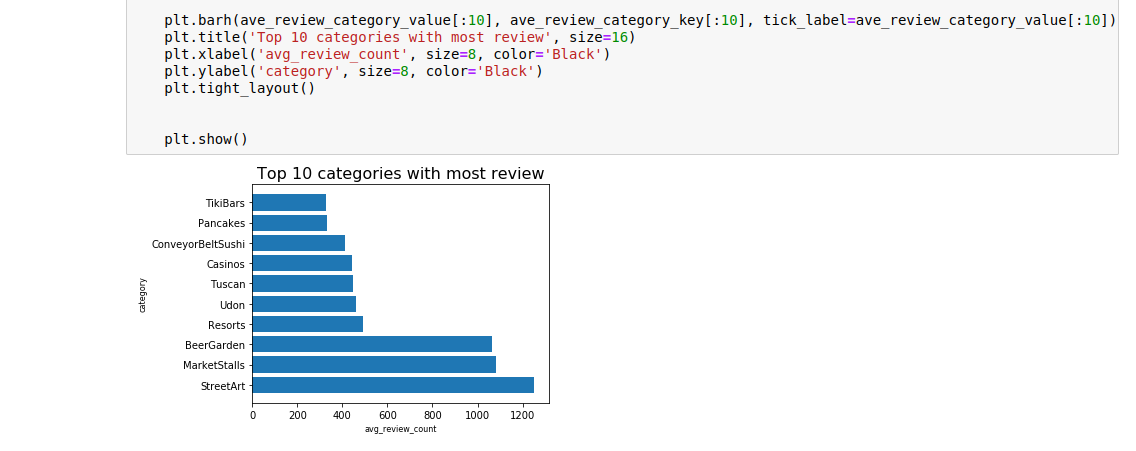


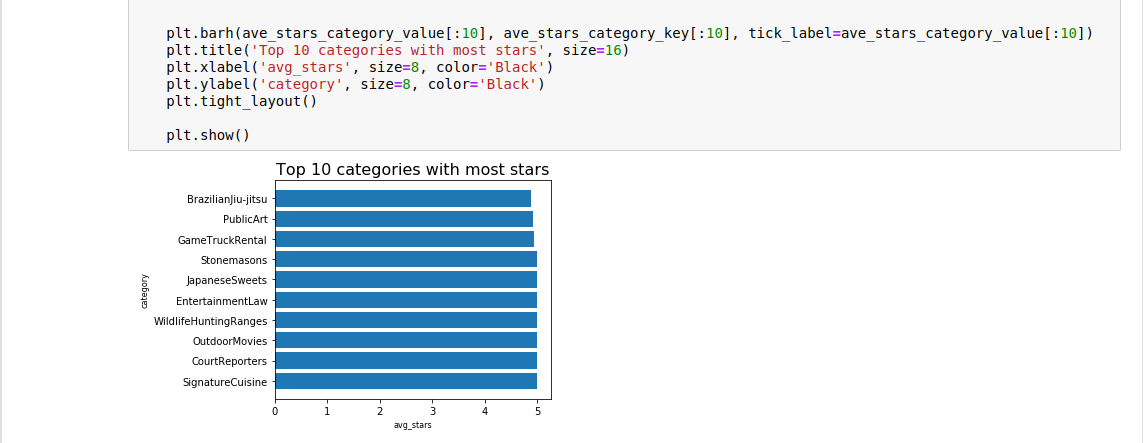


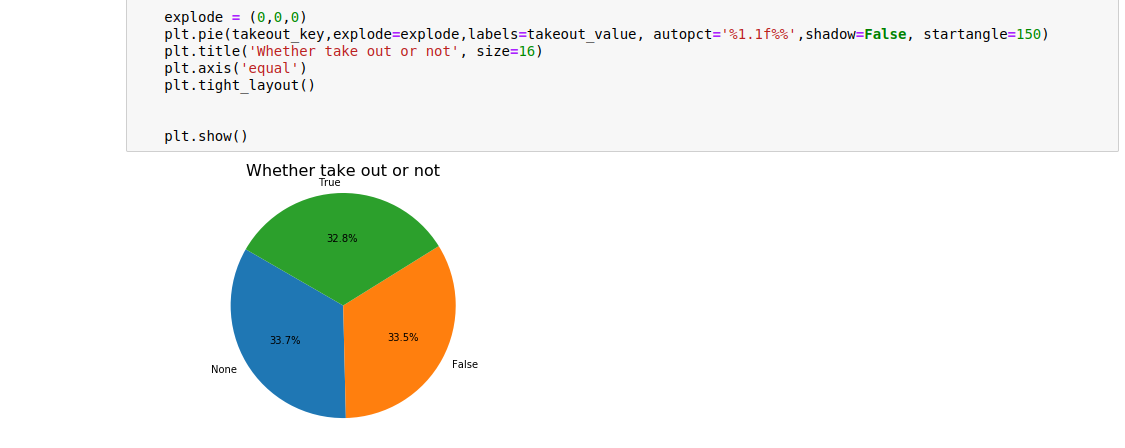
执行结果如下：





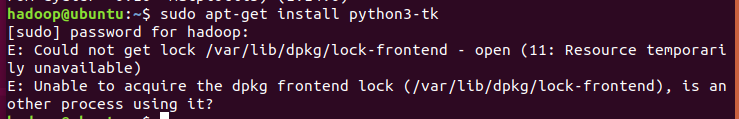






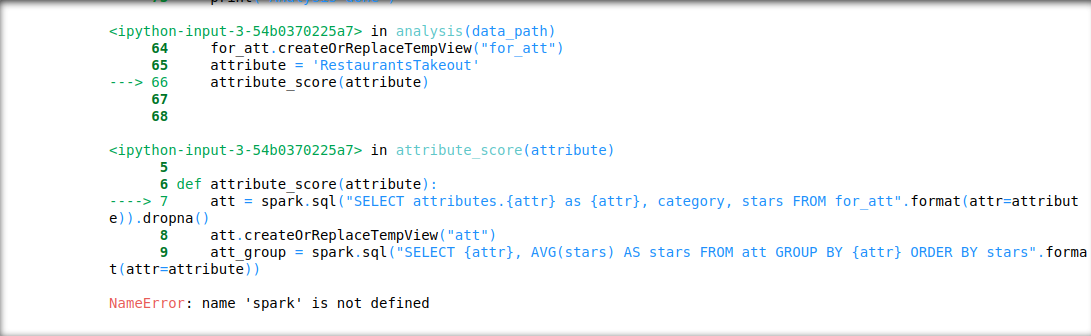
# 二、易错点总结

## （一）安装python插件时报错



**解决方案：**当时笔者已经打开了Jupyter Notebook，将其关闭后再执行此命令便可以正常安装

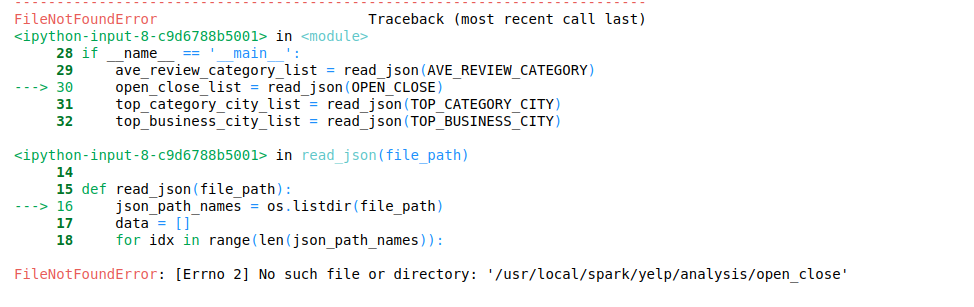
## （二）数据分析过程报错



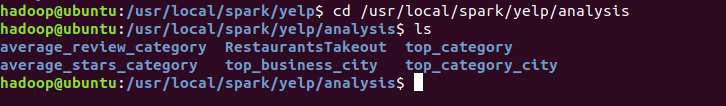
**解决方案：**报错原因是spark没有定义，对spark的定义在数据预处理的代码中可以找到，复制到数据分析中的attribute\_score函数中即可。



## （二）数据分析过程报错



**解决方案：**使用cd命令进入目录，发现并没有open\_close这个文件夹（之前的代码中好像也未发现）因此笔者比较简单粗暴地使用mkdir命令直接创捷了一个open\_close文档。



# 学习小记

## （一）数据可视化

数据可视化是指将大型数据集中的数据以图形图像形式表示，并利用数据分析和开发工具发现其中未知信息的处理过程

数据可视化技术的基本思想是将数据库中每一个数据项作为单个图元素表示，大量的数据集构成数据图像，同时将数据的各个属性值以多维数据的形式表示，可以从不同的维度观察数据，从而对数据进行更深入的观察和分析

## Matplotlib

Matplotlib是一个Python的2D绘图库，它以各种硬拷贝格式和跨平台的交互式环境生成出版质量级别的图形。通过Matplotlib，开发者可以仅需要几行代码便可以生成绘图。一般可绘制折线图、散点图、柱状图、饼图、直方图、子图等等。Matplotlib使用Numpy进行数组运算，并调用一系列其他python库来实现硬件交互。