2023 年第二届钉钉杯大学生大数据挑战赛初赛题目

初赛 B：美国纽约公共自行车使用量预测分析

一、问题背景：

Citi Bike是纽约市在2013年启动的一项自行车共享出行计划，由“花旗银行”(Citi Bank)赞助并取名为“花旗单车”(Citi Bike)。在曼哈顿，布鲁克林，皇后区和泽西市有8,000辆自行车和500个车站。为纽约的居民和游客提供一种方便快捷，并且省钱的自行车出行方式。人们随处都能借到Citi Bank，并在他们的目的地归还。本案例的数据有两部分：第一部分是纽约市公共自行车的借还交易流水表。Citi Bik自行车与共享单车不同，不能使用手机扫码在任意地点借还车，而需要使用固定的自行车桩借还车，数据集包含2013年7月1日至2016年8月31日共38个月（1158天）的数据，每个月一个文件。其中2013年7月到2014年8月的数据格式与其它年月的数据格式有所差别，具体体现在变量starttime和stoptime的存储格式不同。

第二部分是纽约市那段时间的天气数据，并存储在weather_data_NYC.csv文件中，该文件包含2010年至2016年的小时级别的天气数据。

公共自行车数据字段表

变量编号	变量名	变量含义	变量取值及说明
1	trip duration	旅行时长	骑行时间，数值型，秒
2	start time	出发时间	借车时间，字符串，m/d/YYY HH:MM:SS
3	stop time	结束时间	还车时间，字符串，m/d/YYY HH:MM:SS
4	start station id	借车站点编号	定性变量，站点唯一编号
5	start station name	借车站点名称	字符串
6	start station latitude	借车站点维度	数值型
7	start station longtude	借车站点经度	数值型
8	end station id	还车站点编号	定性变量，站点唯一编号
9	end station name	还车站点名称	字符串
10	end station latitude	还车站点纬度	数值型
11	end station longitude	还车站点经度	数值型
12	bile id	自行车编号	定性变量，自行车唯一编号
13	Use type	用户类型	Subscriber:年度用户； Customer：24小时或者7天的临时用户
14	birth year	出生年份	仅此列存在缺失值
15	gender	性别	0：未知 1：男性 2：女性

天气数据字段简介表

变量编号	变量名	变量含义	变量取值及说明
1	date	日期	字符串
2	time	时间	EDT(Eastern Daylight Timing)指美国东部夏令单位
3	temperature	气温	单位：℃
4	dew_poit	露点	单位：℃
5	humidity	湿度	百分数
6	pressure	海平面气压	单位：百帕
7	visibility	能见度	单位：千米
8	wind_direction	风向	离散型，类别包括west,calm等
9	wind_speed	风速	单位：千米每小时
10	moment_wind_speed	瞬间风速	单位：千米每小时
11	precipitation	降水量	单位：毫米，存在缺失值
12	activity	活动	离散型，类别包括snow等
13	conditions	状态	离散型，类别包括overcast,light snow等
14	WindDirDegrees	风向角	连续型，取值为0~359
15	DateUTC	格林尼治时间	YYY/m/d HH:MM

二、解决问题：

1.自行车借还情况功能实现：

实现各个站点在一天的自行车借还情况网络图，该网络图是有向图，箭头从借车站点指向还车站点（很多站点之间同时有借还记录，所以大部分站点两两之间是双向连接）。

（一）以2014年8月3日为例进行网络分析，实现自行车借还网络图，计算网络图的节点数，边数，网络密度（表示边的个数占所有可能的连接比例数），给出计算过程和画图结果。

（二）使用上述的网络分析图，对经度位于40.695~40.72，纬度位于-74.023~-73.973之间的局域网区域进行分析，计算出平均最短路径长度（所有点两两之间的最短路径长度进行算数平均）和网络直径（被定义网络中最短路径的最大值）。

2.聚类分析

对于2013年7月1日至2015年8月31日数据集的自行车数据进行聚类分析，选择合适的聚类数量K值，至少选择两种聚类算法进行聚类，并且比较不同的聚类方法以及分析聚类结果。

3.站点借车量的预测分析：

对所有站点公共自行车的借车量预测，预测出未来的单日借车量。将2013年7月-2015年7月数据作为训练集，2015年8月1-31日的数据作为测试集，预测2015年8月1-31日每天的自行车单日借车量。给出每个站点预测结果的MAPE，并且给出模型的参数数量，最后算出所有站点的MAPE的均值（注：测试集不能参与到训练和验证中，否则作违规处理）。

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\%$$