课程项目

根据附件中数据文件的内容，完成用户基本行为的统计，并采用数据挖掘的方法，深度分析用户的行为特征。

一、项目要求

1. 完成用户基本行为的统计：

1） 计算出用户的每日平均通话次数、每日平均通话时长、每个通话平均时长，并将结果以<主叫号码, 每日平均通话次数, 每日平均通话时长、每个通话平均时长>的格式保存成txt或excel文件。

2） 计算出用户在各个时间段（时间段的划分如表1所示）通话时长所占比例，并将结果以<主叫号码, 时间段1占比, ..., 时间段8占比>的格式保存成txt或excel文件。

2.深度分析用户的行为特征，采用分类、聚类等数据挖掘方法中的任意一种或多种方法完成用户行为特征分析。用户的行为特征可以考虑：

1） 通话呼叫行为（主叫或被叫）。

2） 通话时刻（通话时间段）。

3） 通话时长（通话时间长短）。

4） 其它能够反映用户行为的特征。

根据用户的行为将用户分为多个类别，标注每个类别的用户行为特征，并将结果以可视化的方式展示。

表1 时间段划分表

时间段名称	时间段的起止时间
时间段1	0:00-3:00
时间段2	3:00-6:00
时间段3	6:00-9:00
时间段4	9:00-12:00
时间段5	12:00-15:00
时间段6	15:00-18:00
时间段7	18:00-21:00
时间段8	21:00-24:00

二、项目提交文档

1. 项目报告，介绍用户基本行为的统计方法，以及用户行为特征的分析方法。

2. 数据文件的计算结果，以txt或excel文件保存。

三、附件：数据文件

1. 数据文件：data.zip

2. 数据文件为txt文档，主要包含字段：主叫号码、通话开始时间、通话时长、通话类型、主叫号码运营商等。数据文件的字段说明如表2所示。

表2 数据文件的字段说明表

字段名	字段含义	备注
day_id	日期	
calling_nbr	主叫号码	全部为本运营商加密后的手机号码
called_nbr	被叫号码	g开头号码表示各运营商各城市固话号码，y开头号码表示异网手机号码，其它为本运营商手机号码
calling_optr	主叫号码运营商	1：电信；2：移动；3：联通；其它为不详
called_optr	被叫号码运营商	1：电信；2：移动；3：联通；其它为不详
calling_city	主叫号码归属地	主叫号码所归属的城市
called_city	被叫号码归属地	被叫号码所归属的城市
calling_roam_city	主叫号码漫游地	主叫号码所在的漫游城市，没有漫游时则为空
called_roam_city	被叫号码漫游地	被叫号码所在的漫游城市，没有漫游时则为空
start_time	通话开始时间	格式：13:44:25（时:分:秒）
end_time	通话结束时间	格式：13:44:25（时:分:秒）
raw_dur	通话时长	单位：秒
call_type	通话类型	1：市话；2：长途；3：漫游
calling_cell	主叫蜂窝号码	所在的基站蜂窝标识或为空