数据挖掘原理与技术第 1 次作业说明

1．数据集描述

本数据关于 248 辆出租车在某一天的运营情况(大约每 60 秒一个数据点)。数据在 Excel 数据文件的“Data”Sheet，每个字段的含义参考"Variable Definition" Sheet。

数据样例如下：

	TaxiID	PassengerState	RoofBaconStatus	OnFreeway	GpsMeasureTime	Longitude	Lantitude	Speed
1	16949	0	0	0	2015-04-01 03:22:17	121.549832	31.188045	0.0
2	16034	0	0	0	2015-04-01 09:09:17	121.408703	31.208938	46.5
3	26631	1	5	0	2015-04-01 05:33:56	121.485422	31.404320	0.0
4	26024	0	0	0	2015-04-01 21:46:31	121.430587	31.174863	53.7
5	20156	0	0	1	2015-04-01 19:49:01	121.469235	31.220160	18.0
6	16578	0	0	0	2015-04-01 20:38:14	121.441427	31.195987	18.1
7	25297	0	0	0	2015-04-01 13:35:21	121.415640	31.134493	50.7
8	21983	1	1	0	2015-04-01 00:18:29	121.390707	31.207865	13.8
9	11287	1	5	0	2015-04-01 02:22:17	121.396322	31.351313	0.0
10	21983	0	0	1	2015-04-01 11:40:20	121.450853	31.250053	21.3

2．数据分析目标

2.1）将原始数据汇总成为订单级别的数据，具体包含的字段有：

- 起点时间，起点经、纬度
- 终点时间，终点经、纬度
- 日间行驶时间、夜晚行驶时间(以分钟 min 为单位；23 时至次日 5 时为夜晚，其余为日间)
- 日间行驶里程、夜晚行驶里程(以千米/km 为单位)
- 车费金额（估算方法参考“上海出租车起步价及收费标准”，为简化计算，不考虑低速/停车等候的费用）

上海出租车起步价及收费标准:

车类别	起租部分	超起租部分	燃油附加费	加价部分
市区小型客运出租汽车	起租费13元 起租里程3公里	每公里单价2.4元	每车次1元	1、乘距超10公里单价加计50%； 2、夜间（23时至次日5时）上浮30%； 3、车速低于12公里/小时或乘客要求停车等候，每5分钟计收1公里超起租里程单价

计价方式可参考在线文档：

http://sh.bendibao.com/traffic/2013320/81986.shtm

http://news.cnr.cn/native/gd/20151007/t20151007_520065648.shtml

汇总完成后的数据样例如下：

TaxiID	StartTime	StartLantitude	StartLongitude	DestinationTime	DestinationLantitude	DestinationLongitude	DistInDayTime	DistInNightTime	TimeSpendInDayTime	TimeSpendInNightTime	MoneyPaidToTaxiInRMB
15026	21:27:38	31.216588	121.524045	21:44:39	31.194515	121.562323	6.091696883	0	17	0	24.5
15026	15:37:41	31.210865	121.471067	15:53:06	31.223153	121.440605	4.498184916	0	15	0	20.7
15026	9:55:26	31.239368	121.505622	10:04:56	31.226685	121.529767	2.850310254	0	10	0	16.6
15026	20:01:10	31.273863	121.503568	20:24:17	31.25408	121.577772	11.07242719	0	23	0	37.8
15026	9:01:30	31.281105	121.448242	9:18:39	31.272108	121.493078	5.168298097	0	17	0	22.8
15026	22:37:55	31.235002	121.509378	23:05:55	31.31921	121.534182	10.12044971	3.199182914	22	6	37.3
15026	16:16:28	31.223962	121.489603	16:24:47	31.253742	121.48864	4.735885161	0	8	0	18.7
15026	9:31:13	31.278377	121.488915	9:54:21	31.240293	121.505385	7.903668047	0	23	0	31.5
15026	18:12:00	31.264035	121.520522	18:23:54	31.290877	121.508015	3.486222717	0	12	0	17.2

2.2）在订单级别，噪音/孤立点数据可能有：距离≤0.5km，时间>360min，距离/时间(km/min) <0.1，距离/时间(km/min) >2 （其它情况需要自行判断）

2.3）在司机级别，正常的数据需要符合如下条件（如下数字是否合理需要自行判断，可以设置为其它值）：

- （200 元＜日收入＜1800 元）
- （10＜订单数量＜50）
- （60min＜工作时间＜720min）
- （5min＜平均每个订单的时间＜40min）
- （2km＜平均每个订单的路程＜20km）
- （0.1 km/hour<平均每个订单的速度<2 km/hour）
- 其它需要自行判断的条件

2.4）其它异常情况：

- 司机的每一个订单都很短（例如：只有几分钟）
- 司机夜间长时间工作（例如：一个订单持续很久，但平均速度很低）

- 数据中的其它问题（自行寻找）

2.5）在上述数据清洗和汇总完成后，统计并汇报如下指标（清洗之前、之后分别统计）：

- 司机总数
- 订单总数
- 司机平均每天的订单数
- 司机平均每天的行驶里程
- 司机平均每天的收入金额
- 司机平均每天的载客营运时长
- 司机平均每天的工作时长（载客+空驶，不考虑停运时间）
- 司机平均每天的空驶时间比例（空驶/[载客+空驶]）

3. 任务说明

1）在 MOOC 系统中提交 PPT、数据清理的 Python 源码、以及清洗之后生成的数据（Excel 格式）：在 PPT 中依次解释你进行数据清理的步骤。做每一个清理时，要说明这样做的原因，有多少条纪录受到影响，这样的数据清理是否有可能对分析结果造成影响。可以把源代码放在 PPT 中并简要解释。

2）Presentation 的时间：10 月 17 日。PPT、数据清洗的源代码、清洗后的数据（Excel 格式）和数据清洗的代码需在 **10 月 11 日晚上 8：00 之前**上载至 MOOC 系统中。如有与 Python 编程相关的问题可以咨询助教。

3）本次作业为个人作业，即：每个选课同学均需完成。提交 PPT 和代码之后，老师和助教会选择 2-3 名同学上台演讲（自愿参加，老师会提前联系），其它同学思考和准备一些问题，在课堂 Presentation 结束后现场提问。