河海大學
HOHAI UNIVERSITY

大数据管理与应用专业

商务数据分析报告

基于运输车辆行驶数据的安全驾驶行为研究

姓名：周承桂、杨中昊、李宗霖

时间：2022 年 4 月 25 日

指导老师：徐绪堪

基于运输车辆行驶数据的安全驾驶行为研究

摘要

针对第七届"泰迪杯"数据分析职业技能大赛 C 题——运输车辆安全驾驶行为分析，本小组通过基于 Python 算法的 panda 库和 Excel 的内置数据分析功能进行原始数据的预处理，对提供的运输车辆的原始数据进行去重处理，对提供的以年月日为单位的时间数转换为以秒为单位，方便后续时间计算。同时，通过经纬度变化，重新计算车速及行车里程，通过比对提供的 GPS 速度和里程，去除和原始给定的数据差距过大的项，减少后续计算的误差。

通过查阅安全驾驶行为分析的相关论文，找出影响安全驾驶的相关指标，根据国家运输行业的现行标准，找到急加速、急减速等指标的阈值，求出每辆车的每个时刻所走的里程值，平均行车速度，急加速，急减速等。

根据所给数据的属性特征，定义相关不良驾驶行为和不良驾驶行为对应的识别算法，深度挖掘出每辆车的不良驾驶行为累计次数和时长。然后，为每类指标建立相应的打分体系，利用 SPSS 平台提供的层次分析和主成分分析算法，将两者结合得出新的权重赋值法。最后，对每辆车的每个行程进行比较、分析以及评价，用 K-means 算法划分适宜的评分标准，以用及对该运输企业所给的车进行整体评价分析。以此达到监控分析运输车辆不良驾驶行为的作用，更好的帮助运输车辆管理部门开展道路运输过程安全管理的数据分析，提高运输安全管理水平和运输效率。

关键词：层次分析法 主成分分析法 安全驾驶模型 K-means 算法

一、 选题背景

车联网是指借助装载在车辆上的电子标签通过无线射频等识别技术，实现在信息网络平台上对所有车辆的属性信息和静、动态信息进行提取和有效利用，并根据不同的功能需求对所有车辆的运行状态进行有效的监管和提供综合服务的系统。当前道路运输行业等相关部门利用车联网等系统数据，开展道路运输过程安全管理的数据分析，以提高运输安全管理水平和运输效率。但任需要一个合适的评分体系来评判车辆驾驶人员是否有危险驾驶行为。

运输企业所辖各车辆均存在常规运输路线与驾驶人员。在驾驶员每次运输过程中，车辆均可自动采集当前驾驶行为下的行车状态信息并上传至车联网系统。驾驶行为可能随时变化，从而进一步影响行车安全与运输效率。

二、 数据来源及描述

1. 数据来源

"泰迪杯"数据分析职业技能大赛官网，第七届赛题附件原始数据

2. 数据描述

官网给出了 450 辆运营车辆的行驶数据，平均每辆车包括 50000 条原始数据项，其中包含如下表中的详细数据：

指标名称	指标说明	说明
vehicleplatenumber	车牌号码	无
device_num	设备号	无
direction_angle	方向角	范围：0-359（方向角指从定位点的正北方向起，以顺时针方向至行驶方向间的水平夹角）
lng	经度	东经
lat	维度	北纬
acc_state	ACC 状态	点火 1/熄火 0
right_turn_signals	右转向灯	灭 0/开 1
left_turn_signals	左转向灯	灭 0/开 1
hand_brake	手刹	灭 0/开 1
foot_brake	脚刹	灭 0/开 1
location_time	采集时间	无
gps_speed	GPS 速度	单位：km/h
mileage	GPS 里程	单位：km

表一

由于测量设备的精度有限，所以测出的 GPS 速度和 GPS 里程与现实有少许的出入，且经纬度由于 GPS 定位不稳定，所以数据值与实际有误差，我们在数据预处理阶段就需要将这部分的异常值去除，方便后面的计算。

三、 问题分析

对附件给出的数据进行分析，建立相关模型找到数据中的异常点，对异常点进

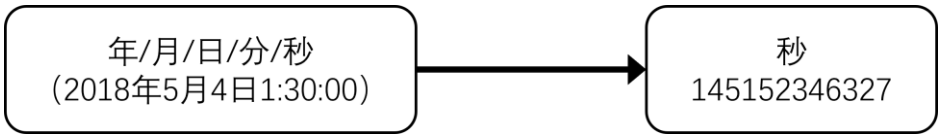
行相关处理。其次，定义行车里程，平均速度，急加速急减速等相关识别算法，并计算每条路线的里程，平均速度，急加速急减速，作为评定危险驾驶行为的重要指标。

随后，提取出每辆车行程中的不良驾驶行为累计次数和时长。为每类指标建立相应的打分体系，利用层次分析和主成分分析相结合的权重赋值法，对每辆车的行程进行比较、分析以及评价。利用 K-means 算法划分评分标准对应的安全驾驶行为的等级，最终依据得分得出是否疲劳驾驶的结论，提供实时判断，帮助运输公司和运输车辆司机减少交通事故的概率。

四、 数据预处理

1. 时间戳转化

对于给定数据中的 location_time 项，全部是按照年/月/日/小时/分钟/秒的格式存放的，但在实际计算中，无法按照分钟单位计算，所以首先需要将该类数据类型格式转换为统一以秒为单位的数据。例:



图一

在转换过程中，我们发现，在经纬度数据完全一致的时候，运用公式计算时，会报除数为零的错误，在不影响后续计算的前提下，我们使用 Python 内置的 panda 库，对所有车辆数据进行了以 lng 和 lat 为参数的数据去重处理，随后再进行了时间戳转换，完成本步数据预处理工作。

2. 异常车速

由于 GPS 精度有限，且在测量过程中存在不稳定的情况，我们根据数据集中的经纬度数据和时间，进行了速度的再次计算，公式如下：

$$\begin{cases} \Delta S = R \times \arccos[\sin x_1 \times \sin x_2 + \cos x_1 \times \cos x_2 \times \cos(y_1 - y_2)] \\ \Delta S > \phi \end{cases} \quad (1)$$

通过 Python 程序可计算得每次测量时间时的瞬时速度，将其和 GPS 给出的数据进行比对，如果两者差值超过一定定值，则舍去该行数据。速度异常点的判断条件为：

$$\begin{cases} v = \dfrac{\Delta S}{\Delta t} \\ v > \varepsilon \\ |v - V| > \delta \end{cases} \tag{2}$$

其中 v 表示由经纬度求出的速度，其速度与 GPS 之差应小于阈值δ，根据运输行业经验，这里将δ取值为 10 km/h，ε取值为 200 km/s。

3. 异常里程

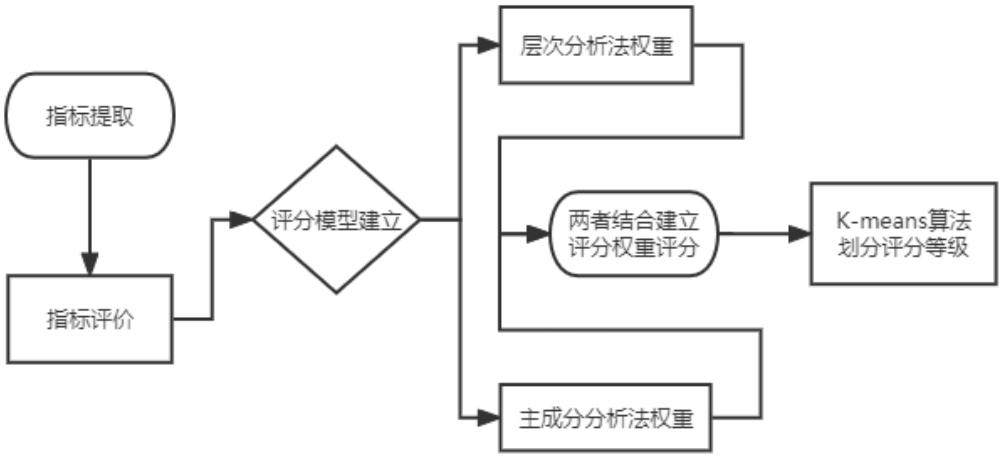
由于车辆定位的偏差，部分点往往距离实际行车点很远，所以我们需要把这类点去除，里程点异常点的判断条件如下：

$$\begin{cases} \Delta S' = S_i - S_{i-1} \\ |\Delta S - \Delta S'| > H \end{cases} \tag{3}$$

$\Delta S'$表示两点间由车辆 GPS 行车里程求出的距离。若此值这两个距离之差大于某个阈值，则将第 i 个点视为里程异常点，本文经过参考其他将 H 点取值为 10 km。同时，判断里程异常点的条件为车辆应在同一条行驶线路，车辆前后两点须为连续的行车轨迹。

五、 问题解决

我们小组在进行问题解决时，按照如下图所示的思路进行：



图二

1. 指标提取

我们小组查阅了导致危险驾驶行为的因素，按照给出的数据和各项指标的重要程度，利用 Python 从原始数据中提取出六项基本评价指标，分别为：超速、急加速急减速、疲劳驾驶、超长怠速、熄火滑行和车速稳定性。

1.1. 超速指标提取：

设车辆当前车速为 V，道路车速阈值为 V_{max}，若 $V > V_{max}$，则判断该车超速。根据运输行业标准，最高车速阈值为 V_{max}=100km/h。计算车辆的累计超速时长 t，用于该项指标评分。

1.2. 急加速急减速指标提取：

根据华东师范大学任慧君，加速度 a 大于 $3m/s^2$，即为急加速或急减速。故本文对急加速和急减速的定义如下表：

加速度 m/s^2	行驶情况
[-3,3]	正常
(3，+∞)	急加速
(−∞,-3)	急减速

表二

计算车辆急加速/急减速的累计时长 t，用于该项指标评分。

1.3. 疲劳驾驶指标提取

依据道路交通安全行为规范 ，本文将疲劳驾驶的判断条件定义为： 车辆累计行驶时间超过 4 小时，并且中途休息时间小于 20 分钟。计算车辆疲劳驾驶的次数 n，以及疲劳驾驶的累计总时长 t，用于该项指标评分。

1.4. 怠速状态指标提取

依据中华人民共和国交通运输部《汽车驾驶节能操作规范》2011 年版，车辆停车超过 60 秒，应将发动机熄火，当发动机转速不为 0 但是车辆本身速度为 0 时，表明该车超长怠速。

本文对怠速状态定义如表：

ACC 状态	车速	停车时间
1	V = 0	T > 60s

表三

当车辆 ACC 状态为 1，并且车速 V=0 时，怠速时长 t 累计，直到当怠速时长 t>60s 时，记为一次怠速状态，该次怠速状态结束时，此时怠速次数 n+1（初始 n=0）。

最后计算怠速次数 n 的值，用于该项指标评分。

1.5. 熄火滑行指标提取

熄火滑行是指汽车的发动机已经熄火，但是汽车刹车置于空挡，所以汽车会按照惯 性继续向前行驶，熄火滑行从节能的角度出发确实可以节约少量能源，但是对于汽车本 身的伤害确实很大，因为当发生紧急情况需要立刻制动停车的时候，熄火滑行可能使刹车失灵，造成交通事故。熄火滑行定义条件如下：

ACC 状态	车速
0	$0<V<50km/h$

表四

当车辆 ACC 状态为 0，并且 0<V<50km/h 时，熄火滑行时间 t 累计。设置最低时间限制，当 t>3s 时，记为一次熄火滑行，熄火滑行次数 n+1（初始 n=0）。

最后，计算车辆累计熄火滑行次数 n，用于该项指标评分。

1.6. 车速稳定性指标提取

车速稳定性是指车辆再行驶过程中车辆的稳定程度，在车辆行驶过程中，如果车速波动过大，可能造成油耗量瞬时剧增，增加车辆的稳定性和不安全性，结合查阅的参考文献本文对车速提取均值、标准差来衡量车辆的运行状态，标准差来表示车速的离散程度，我们利用 Excel 计算，其公式具体如下：

$$\begin{cases} \bar{v}=\dfrac{1}{N}\sum_{i=1}^{N}v \\ \sigma=\sqrt{\dfrac{1}{N}\sum_{i=1}^{N}(v_i-\bar{v})^2} \end{cases} \tag{4}$$

用公式求出标准差，当标准差越大的时候，表示车速的离散度越大，车速的变化频率越大，造成交通事故的可能性越大。

2. 指标评价

综合考虑基于卫星定位数据的驾驶行为安全与节能评价方法等公开技术文献，根据中华人民共和国公共安全行业标准等行业规范，以及各运营企业对驾驶员的考核制度，得到的评价指标得分标准，并通过 Excel 进行指标得分计算。

评价指标得分标准如下表：

评价指标	评分标准	常量值
超速	y1=100-k1*n	k1=0.7
急加速与急减速	y2=100-k2*t	k2=1/50
疲劳驾驶	y31=100-k31*n	k31=20
	y32=100-k32*t	k32=1/1000
	y3=0.5*y31+0.5*y32	
超长怠速	y4=100-k4*n	k4=4
熄火滑行	y5=100-k5*n	k5=30
车速稳定性	y6=100-k6*(n-20)	k6=3

表五

3. 评价模型建立

在驾驶行为安全的评价模型中，根据道路交通运输行业研究部门、企业专家等的研究，若只 是单一的只用一种方法建立行车安全评价模型，则会显得不够客观、科学。故我们将主客观的评价体系相结合，提出一种主成分分析与层次分析法相结合的评价模型：

$$P=(P_1+P_2)*0.5 \qquad (5)$$

其中 P_1 代表层次分析法评价模型得到的评价得分， P_2 代表主成分分析评价模型得到的得分。

3.1. 层次分析法权重

具体步骤：

1、建立车辆行驶安全评价指标（已建立）
2、构造比较判断矩阵并求权重

我们采用标度法构造重要程度层次：

因素i比因素j	量化值
同等重要	1
稍微重要	3
较强重要	5
强烈重要	7
极端重要	9
两相邻判断的中间值	2，4，6，8

表六

若元素 i 与元素 j 的重要性之比为 a_{ij} ，则元素 j 与元素 i 的重要性之比为 $1/a_{ij}$。

根据标度法，得到比较矩阵如下表：

1	1	2	3	2	2
1	1	2	3	2	2
1/2	1/2	1	2	1	2
1/3	1/3	1/2	1	1/2	1/2
1/2	1/2	1	2	1	1
1/2	1/2	1/2	2	1	1

表七

归一化后得到的矩阵为：

0.26087	0.26087	0.285714	0.230769	0.266667	0.235294
0.26087	0.26087	0.285714	0.230769	0.266667	0.235294
0.130435	0.130435	0.142857	0.153846	0.133333	0.235294
0.086957	0.086957	0.071429	0.076923	0.066667	0.058824
0.130435	0.130435	0.142857	0.153846	0.133333	0.117647
0.130435	0.130435	0.071429	0.153846	0.133333	0.117647

表八

得到的特征向量（即权重）为：

[0.256697239，0.256697239，0.154366719，0.074625815，0.134758876，0.122854114]

图三

所得指标对应的权重为：

评价指标	权重
超速	0.257
急加速减速	0.257
疲劳驾驶	0.154
怠速状态	0.074
熄火滑行	0.135
车速稳定性	0.123

表九

3、一致性检验

对构造的判断矩阵进行一致性检验，并以此来确定权重是否合理。公式如下：

$$CR = \frac{CI}{RI} \tag{5}$$

$$CI = \frac{\lambda - n}{n-1} \tag{6}$$

$$\lambda_{max} = \frac{1}{n} \times \sum \frac{A \times W}{W} \tag{7}$$

其中，CR 为一致性比例，CI 为一致性指标，λ为最大特征根，n 为判断矩阵阶数，W 是权向量。当 CR<0.1 时，则判断矩阵不符合一致性要求，需要对该矩阵进行重新修正。经计算查验下表，我们的求解的一致性比例均小于 0.1，权重较为合理。

矩阵阶数	1	2	3	4	5	6	7	8	9	10
RI	0	0	0.58	0.90	1.12	1.24	1.32	1.41	1.45	1.49

表十

最后的计算结果为：

RI=1.26

CI=0.008708

CR=0.006911

完成检验。

3.2. 主成分分析法权重

由于在对驾驶行为进行评价的时候，层次分析法存在一定的主观性，故本文采用主成分分析法继续对驾驶行为进行分析。主成分分析的主要步骤如下：

采集 p 维向量 $X=(x_1,x_2,x_3,...,x_p)$ 的 n 个样品 $x_j=(x_{i1},x_{i2},x_{i3},...,x_{ip})$ 列出样本矩阵 X：

$$X = \begin{bmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \tag{8}$$

我们小组从 450 辆车的数据中，进行数据预处理后，选择 10 辆车，计算出评判指标和指标得分，绘制如下所示的样本矩阵：

23	66.18	62.3175	76	10	62.68
100	67.78	60.997	64	100	76.81
100	67.78	78.365	96	70	80.23
100	71.7	64.3055	84	40	92.08
0	70	62.5625	80	100	63.22
100	76.28	62.7595	68	70	84.04
98.6	73.32	63.5535	56	100	61.21
98.6	66.8	63.472	64	100	63.43
100	59.62	64.8465	68	70	74.62
30.7	74.5	77.72	36	40	51.04

表十一

经过标准化处理后的矩阵如下：

-1.10667	0.661243	0.503101	1.063301	-1.63892	0.517943
1.348607	-0.65052	-1.07138	-0.88506	1.348607	-0.09025
1.479015	-1.17765	-0.30487	1.1492	-0.9946	-0.1511
1.246391	-0.18442	-0.55827	0.437454	-1.78712	0.845967
-2.03639	0.239618	-0.00221	0.564763	1.215052	0.01917
1.878267	-0.04596	-1.14278	-0.71766	-0.55541	0.583549
1.312953	-0.12063	-0.67447	-1.10282	1.392345	-0.80737
1.36797	-0.56117	-0.76306	-0.73103	1.452901	-0.76561
2.091742	-1.01903	-0.6164	-0.37346	-0.21938	0.13653
-1.14185	1.244271	1.419689	-0.85312	-0.63521	-0.03378

表十二

之后，我们使用 SPSS 将上述矩阵导入程序进行 PCA 分析，得出权重分析结果如下：

	主成分权重结果			
	名称	方差解释率	累计方差解释率	权重
超速	主成分 1	0.487	0.487	0.3968
急加减速	主成分 2	0.306	0.794	0.2497
疲劳驾驶	主成分 3	0.227	1.021	0.1853
超长怠速	主成分 4	0.127	1.148	0.1037
熄火滑行	主成分 5	0.066	1.214	0.0536
车速稳定性	主成分 6	0.013	1.227	0.0110

表十三

主成分分析可行性验证，进行 KMO 和 Bartlett 检验：

KMO检验和Bartlett的检验		
KMO值		0.182
Bartlett球形度检验	近似卡方	239.862
	df	15.000
	p	0.000***

注：***、**、*分别代表1%、5%、10%的显著性水平

表十四

对于KMO值：0.8上非常合适做主成分分析，0.7-0.8之间一般适合，0.6-0.7之间不太适合，0.5-0.6之间表示差，0.5下表示极不适合，对于Bartlett的检验（$p<0.05$，严格来说$p<0.01$），若显著性小于0.05或0.01，拒绝原假设，则说明可以做主成分分析，若不拒绝原假设，则说明这些变量可能独立提供一些信息，不适合做主成分分析。

根据我们的验证结果，发现KMO的值为0.182，同时，Bartlett球形检验的结果显示，显著性P值为0.000***，水平上呈现显著性，拒绝原假设，各变量间具有相关性，主成分分析有效。

3.3. 结合层次分析法和主成分分析法求最终得分

根据两种方法计算出来的得分p_1和p_2代入公式5，我们求得十辆车的综合得分p。p_1、p_2、p的结果如下图所示：

P1	47.23165	p2	46.29906	P	46.76536
	80.17293		80.7432		80.45807
	81.6194		85.70792		83.66366
	76.97288		81.35958		79.16623
	54.83916		43.4162		49.12768
	79.7709		82.07802		80.92446
	79.11677		81.04439		80.08058
	78.30026		80.25492		79.27759
	74.65921		78.20164		76.43042
	53.34929		51.61987		52.48458

图四

4. K-means 计算评价模型指标

首先确定将驾驶行为的评价结果分为 4 个等级：

顺序	等级
1	优秀
2	良好
3	合格
4	不合格

表十五

为验证数据是否可以进行 k-means 聚类分析，进行字段差异性分析，分析结果如下表所示:

	聚类类别（平均值±标准差）				F	P
	类别 1(n=6)	类别 2(n=2)	类别 3(n=1)	类别 4(n=1)		
p	1.0±0.0	2.0±0.0	3.0±nan	4.0±nan		0.000***

注：***、**、*分别代表 1%、5%、10%的显著性水平

表十六

上表展示了定量字段差异性分析的结果，包括均值±标准差的结果、F 检验结果、显著性 P 值。分析每个分析项是否小于 0.05 或者 0.01（根据检验标准要求，严格的话使用 0.01）；若呈显著性，拒绝原假设，说明两组数据之间存在显著性差异，可以根据均值±标准差的方式对差异进行分析，反之则表明数据不呈现差异性。

方差分析的结果显示对于变量 p，显著性 P 值为 0.000***，水平上呈现显著性，拒绝原假设，说明变量 p 在聚类分析划分的类别之间存在显著性差异；可以进行 k-means 聚类分析。

然后将 p 的值输入 spss 进行 k-means 聚类分析：

聚类种类	p
2	46.76535541472771
1	80.4580667721911
3	83.66365775217992
1	79.16622552162664
2	49.12768059080052
1	80.92446153484272
1	80.08057721074454
1	79.27759072671233
1	76.43042314802153
4	52.48457920120727

表十七

上图为模型首次聚类结果的部分数据聚类标注结果，只显示综合排序的前 10 条数，继续重复进行聚类分析，直到当准则函数达到最优或者达到最大的迭代次数时终止。我们这里采用欧氏距离，准则函数一般为最小化数据对象到其簇中心的距离的平方和，即

$$\min \sum_{i=1}^{k} \sum_{x \in C_i} dist(C_i, x)^2 \qquad (9)$$

其中，k 是簇的个数，c_i 是第 i 个簇的中心点，dist(ci,x)为 X 到 ci 的距离。最终得出的聚类结果与评价标准如下图所示：

聚类种类	中心值_p	
1	83.66366	优秀
2	79.38956	良好
3	52.48458	合格
4	47.94652	不合格

表十八

	最终得分	评价
AA00002	46.76536	不合格
AA00004	80.45807	良好
AA00052	83.66366	优秀
AA00055	79.16623	良好
AA00212	49.12768	合格
AA00235	80.92446	良好
AA00239	80.08058	良好
AA00299	79.27759	良好
AA00306	76.43042	良好
AA00317	52.48458	合格

根据该评分标准，十辆车的评价分别如下表所示：

表十九

从结果可以看出，大部分车辆的行驶状况都属于安全驾驶，有极个别不合格情况，较为符合现实。将其他车辆数据转化之后代入模型计算，可以得出较为合理的结果。

综上所述，模型求解完成，模型可以供给道路运输、管理部门结合车联网使用，实时监控分析，车辆是否属于危险驾驶状态，从而降低车祸发生的可能。

六、 模型推广与评价

1. 模型推广

由于时间的限制，本模型还存在不足之处，但是最终得到的结果还是较为合理并且贴近现实的。能够适用于指标因素数量适中的场景，比如旅游景点、产品购买、身体素质评价等，将所考虑的指标因素代入该模型，都能够起到不错的效果。

2. 模型评价

2.1. 模型的优点：

在车辆行驶安全评价模型中，我们将主客观的评价体系相结合起来，使用层次分析法和主成分分析法 2 种方法来建立评价模型。最后得出的综合评价结果与 2 种方法单独使用的评价结果相比，更加科学、客观。同时，我们的评价模型考虑了较多方面的因素，更加贴近现实，符合常理。

2.2. 模型的缺点：

在最后计算综合评分时，我们未能找到更好的方法，能够使得 2 种方法得出的评价得分更科学、合理地组合起来，得到最合适的结果。

七、 团队分工与个人心得

1. 团队分工

李宗霖：确定选题方向，查找数据，规划组员工作流程。开始研究选题后，查找相关资料，提供具体的思路和实现方法。主要负责数据预处理和最后汇总工作内容及成果，编写课程实验报告。

周承桂：建立模型的评价指标，利用 Python 处理预处理好的数据，提取和分析评价指标，从而建立各项指标的评分标准，完善指标的单项打分结果。负责报告编写的参考文献、模型推广、评价指标部分的编写。

杨中昊：利用评价指标数据，结合层次分析法和主成分分析法建立综合权重的评价模型，使用 SPSS 及 Excel 等工具进行相关计算和可视化。最后，使用 K-means 方法划分评判是否危险驾驶的等级，得出结论，并编写报告的相关内容。

2. 个人心得

李宗霖：

在本学期的商务数据分析课程中，老师为我们介绍了 PCA 算法、DT 决策

树、问卷调查法、K-means 等算法，我们通过课后的作业，又自行查阅相关论文，找到具体案例进行分析，用代码实现了算法的相关运算，在此过程中，我发现这种方式学习相关算法的效率要远远高于只是听老师讲课。通过自己的实践，不光可以更好地理解算法原理，也可以锻炼自己的编程能力。其次，每学完一部分代码后，各小组还在课堂上进行了具体案例分析的展示，我在听取其他小组的作业后，可以更好地认识到自己代码和算法的不足，能更好地进行创新和改进算法。

在整个的课程报告方面，我们因为小组成员一直没有共同时间进行合作讨论，且每周的相关算法的作业量很大，于是在快结课时，完成的情况任不是很理想，只在开题报告的基础上有了部分的推进，且没有制作可供展示的 PPT，因此在课堂上错过了展示的机会。课后，我们小组成员痛定思痛，立即放下了其他的各项工作，投身于课程报告的研究和编写中。在结课后的一周内，更是全心全意地完善课程报告，希望以此来弥补之前犯下的错误。

首先，我们小组进行了选题分析，在这个过程中，我又学会了很多，知道了如何选择热点问题，知道了哪些选题具有实际意义和研究价值，这为以后的写论文的选题和其他科研方面的工作奠定了很大的基础。随后是数据集的查找，在我们使用第七届泰迪杯 C 题数据之前，我们进行了很长时间的数据查找工作，找到了很多数据集合的网站，如：中国统计局等官网，但由于数据和我们的选题不符合，最后都被我们一一舍弃，在此过程中，我们懂得了数据收集工作的困难，收起了以前的轻视之心。同时，我们也获得了大量的关于收集数据的经验，为以后的相关工作奠定了基础，可谓是收获颇多。

在选题和数据集已经确定后，我们小组内部进行了任务分工，在这个环节，我们每人都先总结自己的优缺点，列出需要做的任务，然后根据个人情况进行任务分配。我主要担任数据的预处理和报告编写方面。这主要是因为，我之前有过数学建模比赛的经验，可以熟练使用 Python 工具和 Office 工具。在数据预处理的过程中，我遇到了很多以前没有遇到过的问题，比如使用 panda 库对 CSV 文件进行提取调用的问题。刚开始自己做这个部分很令人头大，经历了各种各样的困难，但最后通过询问同学，在 CSDN 和 GitHub 上查询，看书等方法，终于把遇到的问题一一解决，这也教会了我如何学习一门编程语言，方法类似。只有先理论再实践才能不断增强自己的编程能力，提高具体问题具体应用的能力。

在随后的建模解题的过程中，我们更是遇到了各种想象不到的问题，在 GitHub 上，我们找到了大量的往年比赛的优秀作文，看到了各种解题思路，但都由于太复杂或编程方面不能实现，被我们一一舍弃。最后，因为曾经在课上学习过比较熟悉的缘故，选择了层次分析法、主成分分析法和 K-means 方法。其中层

次分析法是我们都没有接触过的全新方法，所以需要现学现用。又因为我们的时间比较紧急，所以只能是三个人共同学习，先了解使用方法，淡化原理部分，强调实践运用方面，只有这样才能在短时间内掌握并使用该方法解题。这也教会了我们在时间短缺的情况下，如何最高效地掌握并运用算法。但是这种情况毕竟占少数，一个算法的学习，不能不深入仔细，所以在课程结束之后，我们决定再深入地学习层次分析法，在 B 站上观看相关网课，争取掌握它的工作原理。

总体来说，我对我们组最后报告的完成情况十分满意，从排版到内容都是我们团队的能力体现，但也要注意在以后的课程中要端正态度，认真对待每次作业和任务。

周承桂：

在本学期商务数据分析课程的学习中，我收获颇多。本课程适配于我们专业的所修方向，所学内容对我们专业的延申，即数据的处理与分析，都能够有非常大的帮助。学习了 K-means 聚类、PCA 主成分分析、DT 决策树分析、回归分析等各种数据分析的方法和应用案例，让我对于数据分析这一方面的理解更加清晰。在多次的课程作业汇报中，很多同学都做出了极为不错的成果，能够将各种数据分析方法进行深层次的解读，并且结合所学的程序设计语言和他们在课外所学到的知识方法，做出对实际的运用。我应当向他们学习、向他们看齐，学好这些数据分析方法，努力做到娴熟掌握，运用到生活实际中，同时格局也要放大。

在本学期的课程任务中，我认为对我帮助最大的一点，就是锻炼了我的实际动手能力。在处理问题的过程中，总会遇到各种各样的问题，但面对这些问题，我们不能逃避，必须迎难而上，去不断地尝试，直到试出了正确的方法，能够将问题成果解决。“纸上得来终觉浅，绝知此事要躬行”，古人早已为我们道出了真理。我们学习了知识，并不是真正的掌握了它，只有当我们真正地去运用它，去解决我们所遇到的现实问题，才能够切实体会到知识的奥秘所在。

就如本次课程任务中，用到了 PCA 主成分分析方法一般。PCA 主成分分析法是我们本学期所学的方法之一，我们也查阅和听取过不少实际应用的案例，但这终究是表层的理解。只有当我们自己去运用这个方法时，遇到不少问题处理不了，便去寻找方法，尝试解决问题，成功解决后，再继续按照步骤深入下去，才能得到最终的结果。这样，我们对于这个方法才算有了全面的学习，包括其中的各种细节，我们都对其有了一定的认识和理解。比如在 PCA 主成分分析中，我们要如何进行矩阵的标准化，要用哪些数据来求协方差矩阵，又要如何得出向量特征值。同时，当因素过多，导致矩阵过大时，再用手动计算去它的特征值是非

常困难的，这时我们只能利用 Excel、Python 等工具进行求解。

回到前面所提，这也就是本次课程任务中，对我帮助最大的一点，锻炼了我的实际动手能力。查阅资料、构建指标、提取指标、建立评分标准等工作过程，虽然其中遇到了很多困难，但通过网络、资料、同学、小组成员的帮助，最终还是成功地解决了问题。此外，本次课程任务的处理过程中，加强锻炼了我的讨论交流以及团队协作方面的能力等等。总而言之，经过本学期商务数据分析课程的学习，我收获非常丰富。

杨中昊：

在本学期的商务数据分析课程中，我初步步入了数据分析这个和我们专业息息相关的领域，也初步认识到大体的数据分析过程。明白了商务数据分析的作用在于创造实际的价值，在于给具体的部门或者企业使用，也明确了商务数据分析的过程在于严谨、在于脚踏实地，一步一步地达到最终的目的。明确了数据分析有时候对于一个人来说比较困难，需要小组之间的分工合作，需要每个人具有团队意识以及默契的配合。

虽然我目前还没有在这方面有强大的能力，但是我在这个过程中，我依然有不少收获。首先我学习到了一些实用的数据分析方法，比如 K-means 算法、PCA 算法、回归分析、决策树等等，了解到这些算法在什么时候使用，具体是怎么使用的，我们在最后的大作业中也将其中的一些算法用到了具体的例子当中，更加强了我对与这些算法的认知。在本次大作业中，最初在刚选完题时，我感觉无从下手，深感任务艰巨，但当我们后来一步步做完之后，发现也并不是像当初那样想的那么困难。这离不开小组其他成员的努力以及小组外同学的帮助，也离不开自我的努力。

我还认识到数据分析是一个漫长的过程，而且在数据分析的过程中往往要使用各种各样的方法达到目的，当方法不合适时还要考虑更换成其他方法，甚至是课程中从未出现的方法，例如本课程作业中我们用了一种没有讲过的方法求权重，该方法是我们在网上学习的。而且不同方法可能需要不同的工具，在本课程中我们使用了 excel、python、spss 等工具，python 在数据的挖掘和分析中至关重要，在本课程中我们第一次将 python 应用到具体的实例中，解决具体的问题，尽管中间遇到了种种困难，例如遇到了以前我们没有学过的 python 语法，但我们还是通过自学学会了 python 操作 excel 表格的方法，能打开表格获取具体的数据，并计算这些数据，最后将数据插入到表格中，基本达到了我们的需求。这门课程让我们的编程能力有了一定的提升。而通过使用 spss，我也认识到这个软件功能

的强大，如果没有 spss 现成的方法，许多问题即便用 python 解决起来也可能相当繁琐，比如 PCA 和 k-means，spss 处理这方面问题可以说是相当便捷。

然后我也认识到数据可视化的作用，尤其是在预处理方面的作用。很多时候光是一条条数据如果我们直接进行分析会出现各种各样让人匪夷所思的地方，但当我们将数据可视化，比如做一个柱状图或者折线图、画一个地图和建立坐标系等方法让数据可视化，就能发现哪些数据是有问题的，我们应该排除哪种数据，这样我们就能拥有更清晰的思路，大大缩短排除异常数据的进程。

最后我感受到努力才能换来收获，如果我们被眼前的看似强大的困难吓住而迟迟不肯前进，那我们就永远不会有收获，只有献出自己的时间和努力，才能达到光辉的顶点。

八、 参考文献与附录

1. 参考文献

[1].任慧君，许涛，李响，利用车载GPS轨迹数据实现公交车驾驶安全性分析[J].武汉大学学报·信息科学版，2014,39(6):739-744.

[2].GA/T 1148-2014, 道路交通安全管理规划编制指南[S].

[3].JT/T 807-2011, 汽车驾驶节能操作规范[S].

[4].刘应吉,赵侃,李强,etal. 基于卫星定位数据的违规驾驶行为辨识方法[J]. 公路交通科技，2017(11):130-139.

[5].许书权. 基于车辆运行监控系统的驾驶行为安全与节能评价方法研究[D]. 2015.

[6].夏杰. 基于道路运输企业安全生产管理数据的驾驶行为安全与节能评价方法[D]. 2016.

[7].张鹏. 基于主成分分析的综合评价研究[D]. 南京理工大学,2004.

2. 附录

2.1. 代码

1、数据去重

```
import os
import pandas as pd
import pyproj
# 数据去重
def drop_dup():
    raw_path = 'C:/Users/lenovo/Desktop/附件 1-示例数据-100 辆车'
    drop_dup_path = 'C:/Users/lenovo/Desktop/附件 1-示例数据-100 辆车'
    list = os.listdir(raw_path)
```

```
        print(list)
        for i in list:
            raw_df = pd.read_csv(raw_path + '/' + i)
            # print(raw_df.size)
            new_df = raw_df.drop_duplicates(subset=['lng','lat'])
            new_df.to_csv(drop_dup_path + '/' + i, encoding='utf-8', index=False)
            print(raw_df[['lng','lat']].duplicated())
            print(new_df[['lng','lat']].duplicated())
    if __name__ == '__main__':
        drop_dup()
```

2、时间戳转换

```
    import time
    import datetime
    import pandas as pd
    import numpy as np
    def composeTime(time1):
        time2 = datetime.datetime.strptime(time1, "%Y-%m-%d %H:%M:%S")
        time3 = time.mktime(time2.timetuple())
        time4 = int(time3)
        return time4
    df=pd.read_csv("C:/Users/lenovo/Desktop/ 附 件  1- 示 例 数 据 -100  辆 车
/AB00333.csv")
    col=df.shape[1]
    data=df.iloc[:,10]
    print(data)
    data = data.apply(np.vectorize(composeTime))
    print((data))
    df.iloc[:,10]=data
    # df.insert(col,"time",data)
    #df.drop(columns="1111",inplace=True)
    df.to_csv('C:/Users/lenovo/Desktop/ 附 件  1- 示 例 数 据 -100  辆 车
/AB00333.csv',index=False, sep=',')
```

3、经纬度计算速度和里程

```
    from math import *
    import pandas as pd
    def Distance1(Lat_A,Lng_A,Lat_B,Lng_B): #第一种计算方法
        ra=6378.140 #赤道半径
        rb=6356.755 #极半径 （km）
        flatten=(ra-rb)/ra  #地球偏率
        rad_lat_A=radians(Lat_A)
        rad_lng_A=radians(Lng_A)
        rad_lat_B=radians(Lat_B)
        rad_lng_B=radians(Lng_B)
```

```
    xx=acos(sin(rad_lat_A)*sin(rad_lat_B)+cos(rad_lat_A)*cos(rad_lat_B)*cos(rad
_lng_A-rad_lng_B))
        distance=ra*xx
        return distance
    if __name__ == '__main__':
        path = 'C:/Users/lenovo/Desktop/附件 1-示例数据-100 辆车/AB00210.csv'
        df=pd.read_csv(path)
        ind=df.shape[0]
        col=df.shape[1]
        ls=[0]
        ls2=[0]
        ls3=[0]
        for i in range(ind-1):
            a=df.iat[i,4]
            b=df.iat[i,3]
            c=df.iat[i+1,4]
            d=df.iat[i+1,3]
            e=df.iat[i,10]
            f=df.iat[i+1,10]
            g=(Distance1(a,b,c,d)/(f-e))
            #ls.append(f-e)
            ls2.append(g*3600)
        df.insert(col,"sudu",ls2)
        df.to_csv('C:/Users/lenovo/Desktop/ 附 件  1- 示 例 数 据 -100  辆 车
/AB00210.csv', index=False, sep=',')
    4、不合理数据去除
    from math import *
    import pandas as pd
    pd.set_option('display.max_columns',500)
    pd.set_option('display.width',1000)
    pd.set_option('display.unicode.east_asian_width',True)
    df2=pd.read_csv('D:/software/附件 1-全部数据-450 辆车/AA00317.csv')
    len=df2.shape[0]

    def Distance1(Lat_A,Lng_A,Lat_B,Lng_B): #第一种计算方法
        ra=6378.140 #赤道半径
        rb=6356.755 #极半径 （km）
        flatten=(ra-rb)/ra  #地球偏率
        rad_lat_A=radians(Lat_A)
        rad_lng_A=radians(Lng_A)
        rad_lat_B=radians(Lat_B)
        rad_lng_B=radians(Lng_B)
    xx=acos(sin(rad_lat_A)*sin(rad_lat_B)+cos(rad_lat_A)*cos(rad_lat_B)*cos(rad
```

```
_lng_A-rad_lng_B))
        distance=ra*xx
        return distance
    ax=[]
    ad=[0,1]
    for i in range(len-1):
        a = df2.iat[i, 4]
        b = df2.iat[i, 3]
        c = df2.iat[i + 1, 4]
        d = df2.iat[i + 1, 3]
        g = Distance1(a, b, c, d)
        g1=df2.iat[i+1,12]-df2.iat[i,12]
        x=abs(g-g1)
        if x>10:
            ax.append(i)
            #print(x)
    for i in ax[::-1]:
        print(i)
        df2.drop(axis=0,index=i+1,inplace=True)
    print(len)
    df2.to_csv('D:/software/附件 1-全部数据-450 辆车/AA00317.csv', index=False,
sep=',')
```

5、指标提取

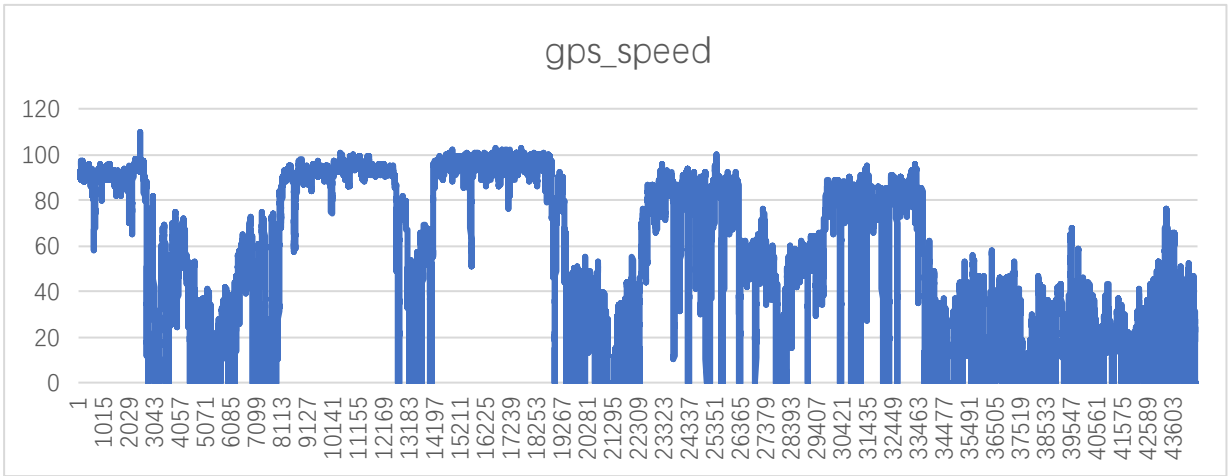
```
    import pandas as pd
    import numpy
    from math import *
    pd.set_option('display.max_columns',500)
    pd.set_option('display.width',1000)
    pd.set_option('display.unicode.east_asian_width',True)
    df2=pd.read_csv('D:/software/附件 1-全部数据-450 辆车/AA00317.csv')
    len = df2.shape[0]
    m = df2.iat[0,12]
    t_lu = df2.iat[0,10]
    l_xianlu = []
    l_chaosu1 = []
    #路线划分
    for i in range(1,len-1):
        if df2.iat[i,12]>m+10 or df2.iat[i,10]>t_lu+3600:
            m = df2.iat[i+1,12]
            t = df2.iat[i+1,10]
            l_xianlu.append(i)
    #超速累计
    t_v=0
```

```
for i in range(0,len-1):
    if df2.iat[i,11]>100:
        t_v +=1
print(t_v)
#加速度
t_a=0
for i in range(1,len-1):
    if(df2.iat[i,11] > df2.iat[i-1,11]):
        if(df2.iat[i,11] > df2.iat[i-1,11] + 3):
            t_a += 1
    else:
        if (df2.iat[i, 11] < df2.iat[i-1, 11] - 3):
            t_a +=1
print(t_a)
#疲劳驾驶
t_xiuxi = 0
t_pao = 0
l_pao=[]
k = 0
for i in range(1,len-1):
    if t_pao >= 14400:
        if df2.iat[i, 11] > 0:
            t_pao += 1
            t_xiuxi = 0
        else:
            t_xiuxi += 1
        if t_xiuxi > 30:
            l_pao.append(t_pao)
            t_pao = 0
            k += 1
    else:
        if df2.iat[i, 11] > 0:
            t_pao += 1
            t_xiuxi = 0
        else:
            t_xiuxi += 1
        if t_xiuxi > 1200:
            t_pao = 0
print(k)
print(l_pao)
#怠速状态
t_daisu = 0
k=0
```

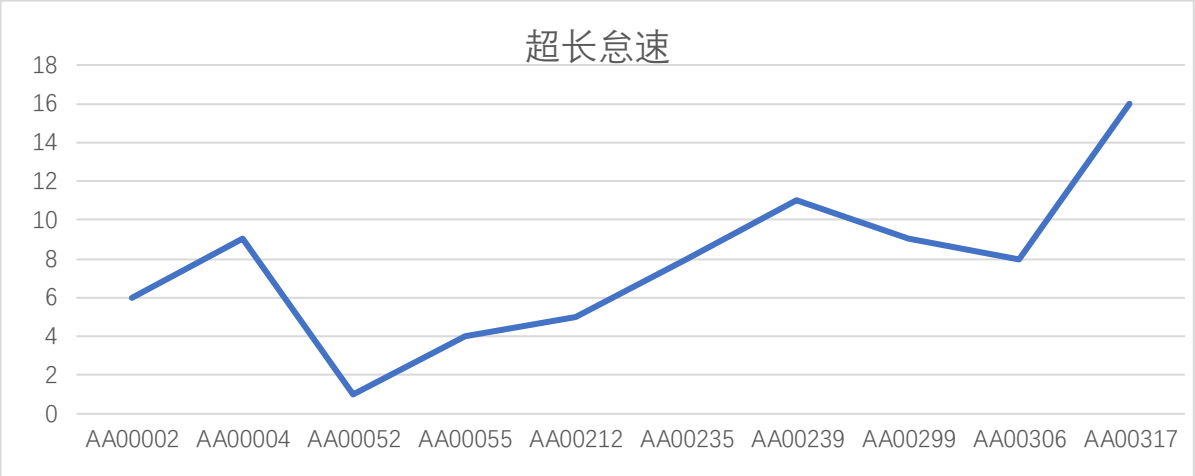
```
for i in range(1,len-1):
        if t_daisu > 60 :
                if df2.iat[i, 5] == 1 and df2.iat[i, 11] == 0:
                        t_daisu += 1
                else:
                        t_daisu = 0
                        k += 1
        else:
                if df2.iat[i, 5] == 1 and df2.iat[i, 11] == 0:
                        t_daisu += 1
                else:
                        t_daisu = 0
print(k)
#熄火滑行
t_xihuo = 0
k = 0
for i in range(1,len-1):
        if t_xihuo > 3:
                if df2.iat[i,5] == 0 and 0<df2.iat[i,11]<50:
                        t_xihuo += 1
                else:
                        t_xihuo = 0
                        k += 1
        else:
                if df2.iat[i,5] == 0 and 0<df2.iat[i,11]<50:
                        t_xihuo += 1
                else:
                        t_xihuo = 0
print(k)
```

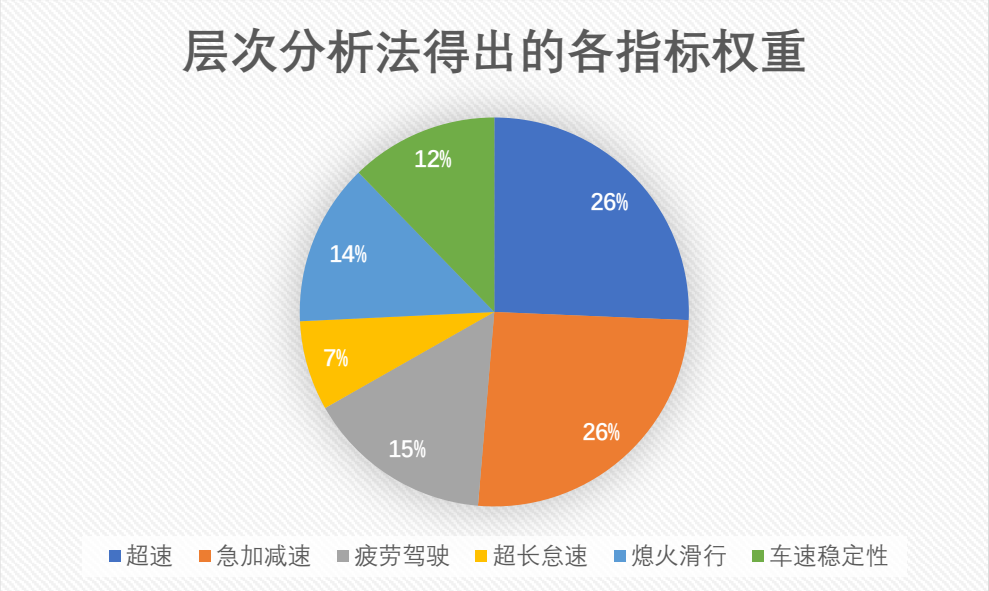
2. 2. 图

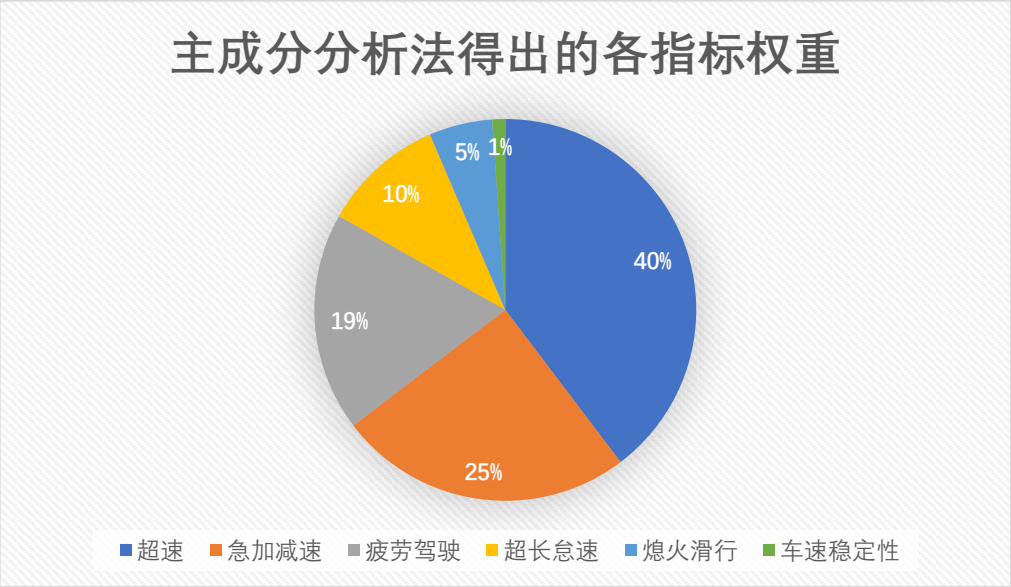
车辆 AA00002 的速度变化图

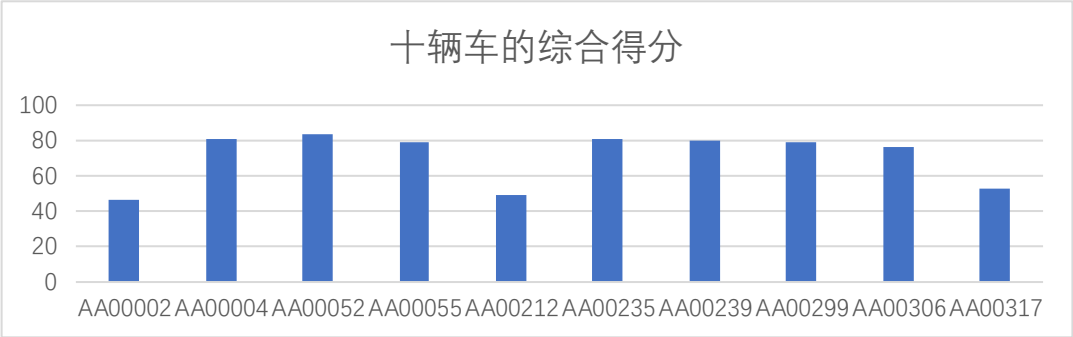


十辆车的急加减速时长对比









模型聚类的结果，包括频数，所占百分比

输出结果3：聚类汇总图