数据挖掘课程项目：

阿里音乐流行趋势预测

成员：张逸恒 2120171100　　余梦巧 2120171089　　曹倩雯 3120170497

项目简介

以阿里音乐用户的历史播放数据为基础，通过对阿里音乐平台上六个月内艺人的试听量的分析，预测出艺人随后 2 个月，即 60 天的播放数据，认为播放量高的艺人是即将成为潮流的艺人，从而实现对一个时间段内音乐流行趋势的准确把控。数据总量为 100 个艺人，1588,4087 条用户记录

阶段工作

一、问题定义

本课题这一个回归预测类问题，已知前 6 个月歌曲艺人及其用户记录，预测后两个月每日的艺人播放量值。

建模的流程是：

预处理-->提取特征并筛选-->模型（多个自变量预测一个连续因变量值）-->预测-->评估

二、数据集分析

首先从数据集中可以看出用户记录表中的用户每日播放，下载和收藏歌曲量是预测未来 60 天每个艺人播放量的重要依据，可以将其作为重要特征。

其次，在不考虑突变和周期规律的情况下，该时间序列是具有短期自相关性的，即相邻的时间序列值具有连续性。

用户记录表中每条记录 gmt_create 记录是可以将每日用户的播放行为精确到每小时，将用户行为按小时划分也许会获得在时间上的高分低谷特征规律以辅助预测未来 60 天的播放量的周期规律

整理数据会发现，在用户前记录表（6 个月）中的歌曲总数小于在歌曲艺人表中的歌曲个数，用户记录表有 1,0278 首歌曲，而歌曲艺人表中有 1,0842 首歌曲。这是因为有些艺人的某些歌曲太过老旧，用户点播率较低。而歌曲和用户记录数据是随机抽样得到的，因此产生这种情况。

三、评价指标分析：

提交结果的最终评分是按照 F 值计算的，从计算公式来看 F 是由每个艺人的评分相加得到的，每个艺人的得分是由归一化方差σ（sigma）和ϕ（phi）相乘得到的。

其中ϕ是当前艺人的每日实际播放量相加开根号得到的，每个艺人的参数ϕ有且只有一个固定值，它的大小取决于每个艺人的 60 天播放量总和值，当某个艺人的总播放量较大时，ϕ就大，F 也就变大了，

从公式来看参数σ是由某艺人提交的每日播放量与实际播放量的差值除以实际播放量，对该值平方后取 60 天的平均值，开根号得到的。这个参数反应了提交结果 S 与实际播放量 T 之间的差距。差越小，预测越精准，（1-σ）越大，F 就大。而当差过大超过了实际播放量 T，此时σ>1，（1-σ）为负数，此时对该艺人评分为负，综合累加的 F 值会更小。由此可知，若预测中存在某个艺人结果极端不准的情况，会使评分 F 下降得更多，因此也要保证所有艺人的平均预测准确性。即尽量保持平稳的值，突发值很容易使结果变差。

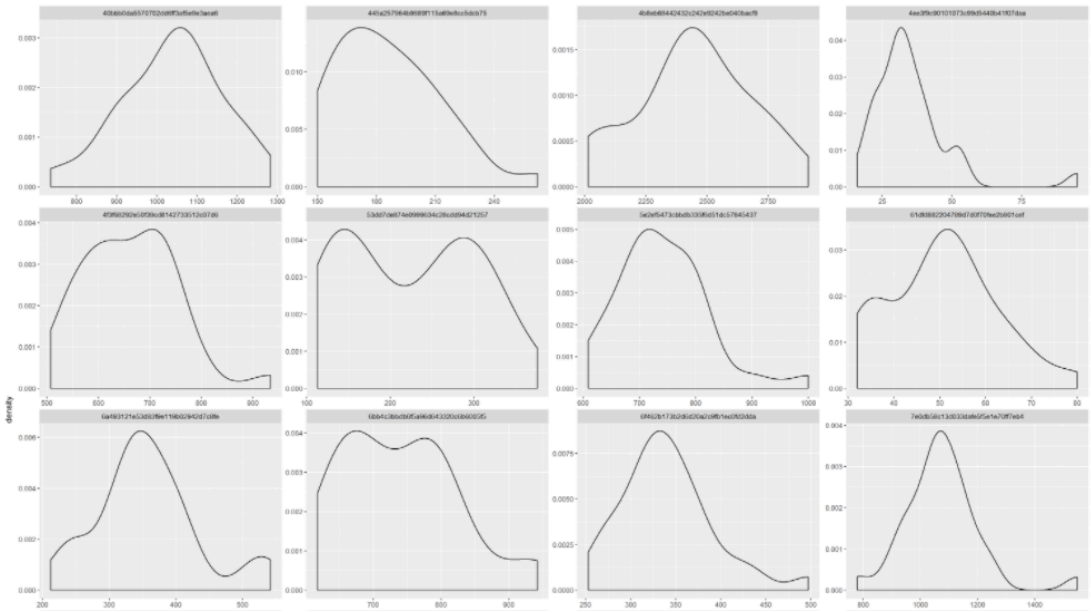
四、数据预处理

1. 剔除偏差较大的数据

由于我们所选的课题是现实生活中真实存在的，因此采用符合现实规律分布的 3σ 准则进行初步过滤。先假设一组检测数据只含有随机误差，对其进行计算处理得到标准偏差，按一定概率确定一个区间，认为凡超过这个区间的误差，就不属于随机误差而是粗大误差，含有该误差的数据应予以剔除。3σ 原则为：

- 数值分布在（μ-σ,μ+σ)中的概率为 0.6826
- 数值分布在（μ-2σ,μ+2σ)中的概率为 0.9544
- 数值分布在（μ-3σ,μ+3σ）中的概率为 0.9974

因为艺人和用户数据是随机抽样，且数据量较大，大部分艺人时间序列符合正态分布的密度曲线，下是部分艺人 8 月每日播放量的密度曲线。我们就按照 2 倍标准差(SD)的方法粗略的剔除数据。



2. 填补数据

若回归模型中以时间作为变量，剔除掉的空缺时间数据需要被填补，如下所示：