**背景及统计的作用：**

影响一首歌播放次数的因素有很多，比如歌手是谁，歌曲所在专辑是否火爆，歌曲的年份，以及其他歌曲自身的一些特征。如果能根据这些特征利用统计方法去预测一首歌曲的播放次数，那我们就可以对这首歌做出一些特定的营销以及合适的推荐。

**已有的数据：**

我们的数据集来自于百万歌曲数据集（Million Song Dataset），它包括回音网（Echo Nest）上的1019318个不同用户的音乐播放次数以及有关歌曲的一些基本信息。

数据集来源：<https://labrosa.ee.columbia.edu/millionsong/>。

你可以使用网站上的完整数据集，也可以使用附件中提供的清洗好的数据集。使用网站上的完整数据集的时候需要注意，应该至少包括“Taste Profile subset”和“Summary file of the whole dataset”这两个数据集。

附件中的数据集包含以下变量：

* SongID ：歌曲ID
* Title ：歌曲的名称
* AlbumID ：7digital.com上发行的唱片集ID
* ArtistName ：艺术家的名字
* Duration ：歌曲时长（以秒为单位）
* key Signature ：歌曲的音调
* Popularity ：艺术家的评分
* Energy ：歌曲火爆程度
* Mode :歌曲的音阶，是0-1变量，0和1分别代表小音阶和大音阶
* Loudness ：音轨的平均响度
* Tempo ：音乐术语，歌曲的速度
* Time Signature：歌曲的通常每分钟节拍数
* Year ：MusicBrainz的歌曲发行年份，缺失数据记为0
* Count ：回音网中1019318个不同用户的音乐播放次数

**研究问题：**

（1）.探索不同的Mode（音阶）对歌曲的影响。通过数据集探究小音阶和大音阶（0和1）之间是否有区别，如果有，区别在哪，统计建模的评价指标至少包括但是不限于以下方面：Duration（歌曲时长）,key Signature（音调）,Energy（歌曲火爆程度）,Tempo（歌曲的速度）以及Count（播放次数）。可以通过维基百科去查询一些背景信息。

（2）.利用百万歌数据集中收集的歌曲信息，通过统计建模来预测歌曲的Count（播放次数）。

（3）.附加题：如果感兴趣的话，可选择<https://labrosa.ee.columbia.edu/millionsong/pages/tasks-demos>上列出的其他问题进行研究。

**注意：**

1.如果选择使用附件中的百万歌数据集以外的数据集，那么你的数据集必须包含多个变量，至少要包括一个定量变量和一个定性变量，注意数据规模不能太小（建议使用附件中清洗好的数据集）。

2.研究问题中的（1）和（2）是必须的，附加题（3）可自己选择是否去做。

3.报告必须包括以下几部分内容：

•摘要

•数据的介绍和描述

此部分应至少包括（i）所研究的问题的清晰描述（ii）列出你所使用的变量，以及哪些变量是定性、定量的。

•探索性数据分析

此部分应至少包括（i）详细描述对数据集的处理和分析（ii）实现数据可视化，对数据做出描述性统计分析

•统计推断

此部分应包括你所拟合的统计模型，以及对模型做出的统计推断，如置信区间，假设检验等。请勿使用本课程未涉及的工具或技术（例如机器学习）。

•结论和总结

此部分应包括这篇报告的主要结论和总结。

•附录

此部分应包括代码（R语言）或其他重要的附加材料。

•参考文献