几个问题：

1. Popularity中有许多的0值，是否可以将0看作是nan，no record。然后将整行删除。这会导致问题4中的u test 的p值大幅变化，直接影响最终答案。

且会影响问题6和7中的线性回归R方

1. Instrumentalness列中也有许多0值，虽然介绍中说可以有0，但是0多的不正常。而且speechiness是它的反过来，speechiness中就没有0值和1值。
2. Instrumentalness在问题6中预测是R方最大的predictor。被预测的是popularity。它们两列都有很多0，所以这可能是R方高的原因。
3. 问题6和7的预测中，R方都小的离谱。看起来完全不像好的线性回归。怎么处理
4. 或许可以尝试一下用Random Forest来做Q6，7中的预测？RF也可以用R方和RMSE来评估模型的。注意，对于RF来说，增加过多的features（x变量数）也会导致过度拟合问题。
5. 问题10没有明说一定要用logistic，所以可能可以用RF Classification。AUC没必要非要画图出来，有function可以直接输出AUC分数。RF Classification也可以用AUC评估。

附加题：可以测试一下关于12个key的popularity是否有区别。用kruskal test