歌曲畅销推断与营销推荐

摘要：伴随着我国居民收入水平和生活品质的不断提高，人们对于精神文娱等方面提出了更高的要求。影响一首歌曲是否火爆的因素有许多，除了如今消费人群受众变更以外，歌曲的年份，以及其他歌曲自身的一些特征。通过利用统计方法预测一首歌曲的播放次数，那么就可以针对具体歌曲做出特定的营销和相应的推荐。

一、数据的介绍和描述

采集数据集来自于百万歌曲数据集（Million Song Dataset），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/>。本文采用附件中已经清洗好的数据集，其中包含的变量有：SongID:歌曲ID，Title:歌曲名称，AlbumID:7digital.com上发行的唱片集ID，ArtistName:艺术家名字，Duration:歌曲时长(以秒为单位)，key Signature:歌曲的音调，Popularity:艺术家评分，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:音乐速度，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，缺失的数据记为0，Count:回音网中1019318个不同用户的音乐播放次数。

定性变量：key Signature，mode，year

定量变量：Duration，Energy，Popularity，Loudness，Tempo，Time Sginature，Count

研究的问题主要包括以下两个：

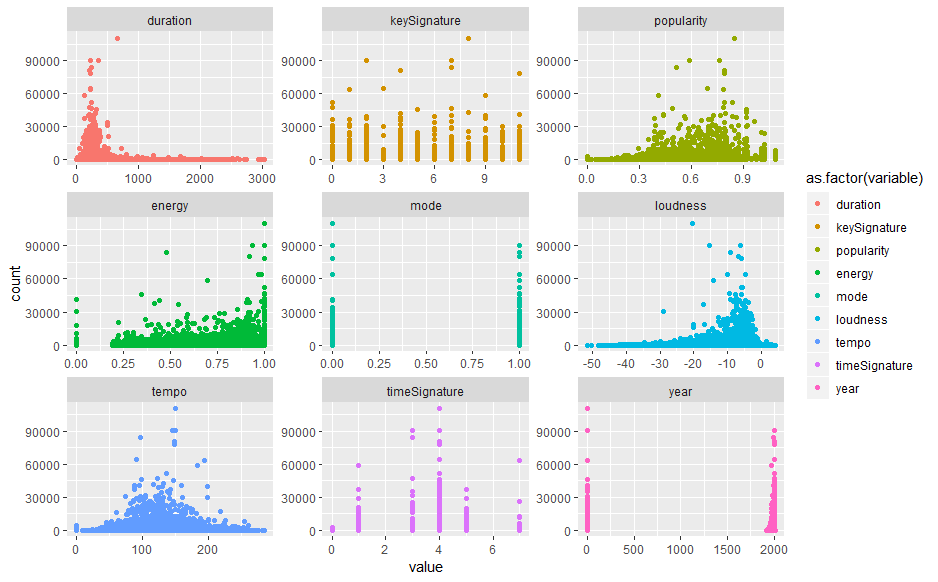
（1）探索不同的Mode（音阶）对歌曲的影响。通过数据集探究小音阶和大音阶（0和1）之间是否有区别，如果有，区别在哪，统计建模指标包括但不限于：Duration（歌曲时长），key Signature（音调），Energy（歌曲火爆程度），Tempo（歌曲的速度）以及Count（播放次数）。

（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）。

二、探索性数据分析

数据清洗：针对数据中存在的缺失值情况，没有在变量定义中说明过的还有缺省值的变量的缺省值处理一律按照缺失值去除案例对象处理。去除缺省值以后样本量从385256降低到309685。

数据可视化：



针对第一部分问题，我们通过对不同Mode音阶下各个统计变量进行差异分析：

1.方差齐性Levene检验：

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| count | Df | F value | Pr(>F) | energy | Df | F value | Pr(>F) |
| group | 1 | 0.22 | 0.6383 | group | 1 | 105.6 | 0 |
|  | 309683 |  |  |  | 309683 |  |  |
| duration | Df | F value | Pr(>F) | loudness | Df | F value | Pr(>F) |
| group | 1 | 234.22 | 0 | group | 1 | 13.26 | 0.0003 |
|  | 309683 |  |  |  | 309683 |  |  |
| keySignature | Df | F value | Pr(>F) | tempo | Df | F value | Pr(>F) |
| group | 1 | 10.62 | 0.0011 | group | 1 | 1.99 | 0.1585 |
|  | 309683 |  |  |  | 309683 |  |  |
| popularity | Df | F value | Pr(>F) | timeSignature | Df | F value | Pr(>F) |
| group | 1 | 61.65 | 0 | group | 1 | 106.56 | 0 |
|  | 309683 |  |  |  | 309683 |  |  |

可以看出：方差齐性的变量有count，tempo；方差不齐性的变量有energy，duration，loudness，keySignature，popularity，timeSignature；

而后做单因素方差检验：

| count | F | num df | denom df | p\_value |
| --- | --- | --- | --- | --- |
|  | 0.19 | 1.00 | 309683.00 | 0.66 |
| duration | F | num df | denom df | p\_value |
|  | 1348.37 | 1.00 | 200103.11 | 0.00 |
| keySignature | F | num df | denom df | p\_value |
|  | 11485.63 | 1.00 | 208529.69 | 0.00 |
| popularity | F | num df | denom df | p\_value |
|  | 8.22 | 1.00 | 214442.59 | 0.00 |
| energy | F | num df | denom df | p\_value |
|  | 198.54 | 1.00 | 215463.78 | 0.00 |
| loudness | F | num df | denom df | p\_value |
|  | 68.74 | 1.00 | 209725.81 | 0.00 |
| tempo | F | num df | denom df | p\_value |
|  | 16.56 | 1.00 | 309683.00 | 0.00 |
| timeSignature | F | num df | denom df | p\_value |
|  | 263.76 | 1.00 | 216791.29 | 0.00 |

可以看出，统计量中除了不同mode音阶的歌曲的count没有显著差异性外，其他的歌曲性质（duration，keySignature，popularity，energy，loudness，tempo，timeSignature）均有显著差异性。

区别在哪里：

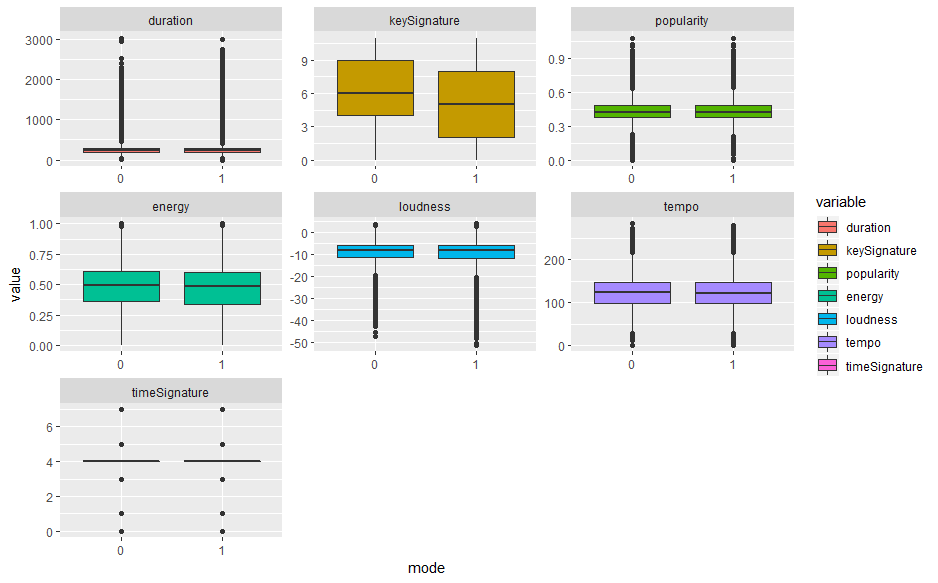
按照mode不同音阶对数据集进行分组后统计（duration，keySignature，popularity，energy，loudness，tempo，timeSignature）这些统计量的描述分析值：

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **vars** | **n** | **mean** | **sd** | **median** | **trimmed** | **mad** | **min** | **max** | **range** | **skew** | **kurtosis** | **se** |
| **duration** | 1 | 104653.00 | 257.68 | 116.09 | 239.57 | 246.19 | 73.28 | 4.54 | 3030.91 | 3026.36 | 3.64 | 39.41 | 0.36 |
| **keySignature** | 2 | 104653.00 | 6.27 | 3.55 | 6.00 | 6.42 | 4.45 | 0.00 | 11.00 | 11.00 | -0.22 | -1.22 | 0.01 |
| **popularity** | 3 | 104653.00 | 0.43 | 0.10 | 0.42 | 0.43 | 0.07 | 0.00 | 1.08 | 1.08 | -0.25 | 5.75 | 0.00 |
| **energy** | 4 | 104653.00 | 0.47 | 0.20 | 0.49 | 0.48 | 0.18 | 0.00 | 1.00 | 1.00 | -0.66 | 0.22 | 0.00 |
| **loudness** | 5 | 104653.00 | -9.13 | 4.76 | -8.00 | -8.51 | 3.85 | -47.44 | 3.54 | 50.98 | -1.54 | 3.60 | 0.01 |
| **tempo** | 6 | 104653.00 | 125.70 | 35.15 | 123.05 | 123.73 | 35.29 | 0.00 | 284.21 | 284.21 | 0.44 | 0.43 | 0.11 |
| **timeSignature** | 7 | 104653.00 | 3.69 | 1.15 | 4.00 | 3.81 | 0.00 | 0.00 | 7.00 | 7.00 | -0.72 | 2.32 | 0.00 |
|  | **vars** | **n** | **mean** | **sd** | **median** | **trimmed** | **mad** | **min** | **max** | **range** | **skew** | **kurtosis** | **se** |
| **duration** | 1 | 205032.00 | 241.79 | 109.44 | 227.40 | 231.24 | 69.67 | 0.52 | 2999.51 | 2998.99 | 4.04 | 47.98 | 0.24 |
| **keySignature** | 2 | 205032.00 | 4.83 | 3.51 | 5.00 | 4.73 | 4.45 | 0.00 | 11.00 | 11.00 | 0.09 | -1.33 | 0.01 |
| **popularity** | 3 | 205032.00 | 0.43 | 0.10 | 0.42 | 0.43 | 0.08 | 0.00 | 1.08 | 1.08 | -0.35 | 5.03 | 0.00 |
| **energy** | 4 | 205032.00 | 0.46 | 0.21 | 0.48 | 0.47 | 0.19 | 0.00 | 1.00 | 1.00 | -0.60 | 0.04 | 0.00 |
| **loudness** | 5 | 205032.00 | -9.28 | 4.73 | -8.21 | -8.71 | 4.02 | -51.42 | 3.97 | 55.38 | -1.44 | 3.33 | 0.01 |
| **tempo** | 6 | 205032.00 | 125.16 | 34.81 | 122.10 | 123.23 | 35.41 | 0.00 | 280.95 | 280.95 | 0.44 | 0.38 | 0.08 |
| **timeSignature** | 7 | 205032.00 | 3.62 | 1.18 | 4.00 | 3.74 | 0.00 | 0.00 | 7.00 | 7.00 | -0.76 | 1.79 | 0.00 |

当音阶mode处于小音阶时，duration歌曲时长均值257.68秒，标准差116.09秒；keySignature歌曲的音调均值为6.27，标准差为3.55；popularity艺术家评分均值为0.43，标准差为0.10；energy歌曲火爆程度0.47，标准差0.20；loudness音轨平均音响均值为-9.13，标准差4.76；tempo歌曲的速度均值125.70，标准差为35.15，timeSignature歌曲中每分钟的节拍数均值3.69，标准差为1.15。

当音阶mode处于大音阶时，duration歌曲时长均值241.79秒，标准差109.44秒；keySignature歌曲的音调均值为4.83，标准差3.51；popularity艺术家评分均值为0.43，标准差为0.10；energy歌曲火爆程度0.46，标准差为0.21；loudness音轨平均音响均值为-9.28，标准差4.73；tempo歌曲的速度均值为125.16，标准差为34.81；timeSignature歌曲中每分钟节拍数均值3.62，标准差为1.18。

可视化后如下图所示：



可以看出，这些变量的箱型图如上图所示，部分变量有较多的偏离值。

三、统计推断

通过统计方法建模预测分析歌曲Count播放次数，通过广义线性模型拟合参数，模型与拟合结果如下：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Deviance Residuals: | |  |  |  |  |
| Min | 1Q | Median | 3Q | Max |  |
| -172.87 | -9.53 | -4.96 | -0.05 | 992.48 |  |
| Coefficients: |  |  |  |  |  |
|  | Estimate | Std. Error | z value | Pr(>|z|) |  |
| (Intercept) | -1.06E+00 | 1.33E-03 | -800.4 | <2e-16 | \*\*\* |
| duration | 3.18E-04 | 1.47E-06 | 216.47 | <2e-16 | \*\*\* |
| keySignature | 4.19E-03 | 4.35E-05 | 96.36 | <2e-16 | \*\*\* |
| popularity | 5.42E+00 | 1.20E-03 | 4525.82 | <2e-16 | \*\*\* |
| energy | 5.86E+00 | 1.14E-03 | 5130.36 | <2e-16 | \*\*\* |
| mode | 5.63E-02 | 3.30E-04 | 170.6 | <2e-16 | \*\*\* |
| loudness | -2.79E-03 | 3.65E-05 | -76.53 | <2e-16 | \*\*\* |
| tempo | -9.48E-04 | 4.59E-06 | -206.68 | <2e-16 | \*\*\* |
| timeSignature | 1.17E-02 | 1.43E-04 | 81.33 | <2e-16 | \*\*\* |
| year | -8.54E-05 | 1.98E-07 | -432.1 | <2e-16 | \*\*\* |
| Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1 | | | | | |
| (Dispersion parameter for poisson family taken to be 1) | | | | | |
| Null deviance: 175170959 on 309684 degrees of freedom | | | | | |
| Residual deviance: 93614388 on 309675 degrees of freedom | | | | | |
| AIC: 95097886 |  |  |  |  |  |

GLM中对各个变量做了z检验值如上表所示，对应得到各个变量的显著性p值，看出GLM模型中各个变量与因变量的拟合路径关系均显著；模型总体AIC为95097886；异常残差最小值为-172.87，四分之一分位值-9.53，中位值-4.96，四分之三分位值-0.05，最大值为992.48。模型偏差175170959，模型自由度309684，残差偏差93614388，自由度309675。

energy和popularity分别强烈作用在count上，年份、tempo和loudness等对于count的作用是负面影响。

与多元线性模型对比：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Residuals: | |  |  |  |  |
| Min | 1Q | Median | 3Q | Max |  |
| -1050 | -172 | -78 | 33 | 109510 |  |
| Coefficients: | |  |  |  |  |
|  | Estimate | Std. Error | t value | Pr(>|t|) |  |
| (Intercept) | -601.471 | 11.80355 | -50.957 | < 2e-16 | \*\*\* |
| duration | 0.0245 | 0.01343 | 1.824 | 0.06817 | . |
| keySignature | 0.51546 | 0.42344 | 1.217 | 0.22349 |  |
| popularity | 1292.7497 | 16.68957 | 77.459 | < 2e-16 | \*\*\* |
| energy | 481.99588 | 8.33004 | 57.862 | < 2e-16 | \*\*\* |
| mode | 9.15178 | 3.22204 | 2.84 | 0.00451 | \*\* |
| loudness | 0.54033 | 0.32439 | 1.666 | 0.09578 | . |
| tempo | -0.17434 | 0.04333 | -4.024 | 5.72E-05 | \*\*\* |
| timeSignature | 2.15434 | 1.28544 | 1.676 | 0.09375 | . |
| year | -0.02433 | 0.00174 | -13.977 | < 2e-16 | \*\*\* |
| --- |  |  |  |  |  |
| Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1 | | | | | |
|  |  |  |  |  |  |
| Residual standard error: 830.4 on 309675 degrees of freedom | | | | | |
| Multiple R-squared: 0.04968, Adjusted R-squared: 0.04966 | | | | | |
| F-statistic: 1799 on 9 and 309675 DF, p-value: < 2.2e-16 | | | | | |

多元ols线性回归模型中，tempo，year，energy，mode，popularity等在多元回归的系数t检验中的显著性p值小于0.01，energy是最为强烈的正向作用因素，popularity是第二主要正向作用因素，tempo对于count的作用影响是负面的，mode对于count作用中可以看出大音阶count会更高一些。

四、结论和总结

结论：

1. 除了不同mode音阶的歌曲的count没有显著差异性外，其他的歌曲性质（duration，keySignature，popularity，energy，loudness，tempo，timeSignature）均有显著差异性；

2. Count受到energy，popularity的影响程度最大，受到年份以及tempo，duration影响程度不高。

附录

# #############################################################

# set workdirection; library package

setwd('F:\\jTKount\\1201\\白相 1220 广义线性回归 stata ☆')

library(readxl)

library(car)

library(xtable)

library(flextable)

library(officer)

library(dplyr)

library(reshape2)

library(ggplot2)

# #############################################################

# input datas; add colnames

dat\_44<-read\_excel('MSDtaste.xlsx', sheet = 1,col\_names = T)

data\_u<-data.frame(dat\_44)

data\_u1<-na.omit(data\_u)

# 1 单因素方差分析

col\_ = c('count', 'duration', 'keySignature', 'popularity',

'energy', 'loudness', 'tempo', 'timeSignature')

doc = read\_docx()

for(col in col\_){

# levene test

lvT<-leveneTest(as.numeric(unlist(data\_u1[col])), as.factor(data\_u1$mode))

word\_lvt = xtable\_to\_flextable(xtable(lvT))

doc = body\_add\_flextable(doc, word\_lvt)

}

print(doc, 'leveneTest.docx')

# one way test

doc = read\_docx()

owt<-oneway.test(count~mode, data=data\_u1, var.equal = TRUE)

df\_owt<-data.frame(owt$statistic,owt$parameter[1],owt$parameter[2],owt$p.value)

word\_owt<-xtable\_to\_flextable(xtable(df\_owt))

doc = body\_add\_flextable(doc, word\_owt)

owt<-oneway.test(duration~mode, data=data\_u1, var.equal = FALSE)

df\_owt<-data.frame(owt$statistic,owt$parameter[1],owt$parameter[2],owt$p.value)

word\_owt<-xtable\_to\_flextable(xtable(df\_owt))

doc = body\_add\_flextable(doc, word\_owt)

owt<-oneway.test(keySignature~mode, data=data\_u1, var.equal = FALSE)

df\_owt<-data.frame(owt$statistic,owt$parameter[1],owt$parameter[2],owt$p.value)

word\_owt<-xtable\_to\_flextable(xtable(df\_owt))

doc = body\_add\_flextable(doc, word\_owt)

owt<-oneway.test(popularity~mode, data=data\_u1, var.equal = FALSE)

df\_owt<-data.frame(owt$statistic,owt$parameter[1],owt$parameter[2],owt$p.value)

word\_owt<-xtable\_to\_flextable(xtable(df\_owt))

doc = body\_add\_flextable(doc, word\_owt)

owt<-oneway.test(energy~mode, data=data\_u1, var.equal = FALSE)

df\_owt<-data.frame(owt$statistic,owt$parameter[1],owt$parameter[2],owt$p.value)

word\_owt<-xtable\_to\_flextable(xtable(df\_owt))

doc = body\_add\_flextable(doc, word\_owt)

owt<-oneway.test(loudness~mode, data=data\_u1, var.equal = FALSE)

df\_owt<-data.frame(owt$statistic,owt$parameter[1],owt$parameter[2],owt$p.value)

word\_owt<-xtable\_to\_flextable(xtable(df\_owt))

doc = body\_add\_flextable(doc, word\_owt)

owt<-oneway.test(tempo~mode, data=data\_u1, var.equal = TRUE)

df\_owt<-data.frame(owt$statistic,owt$parameter[1],owt$parameter[2],owt$p.value)

word\_owt<-xtable\_to\_flextable(xtable(df\_owt))

doc = body\_add\_flextable(doc, word\_owt)

owt<-oneway.test(timeSignature~mode, data=data\_u1, var.equal = FALSE)

df\_owt<-data.frame(owt$statistic,owt$parameter[1],owt$parameter[2],owt$p.value)

word\_owt<-xtable\_to\_flextable(xtable(df\_owt))

doc = body\_add\_flextable(doc, word\_owt)

print(doc, 'onewayTest.docx')

# split and describe

doc = read\_docx()

col\_gb<-c('duration','keySignature', 'popularity',

'energy', 'loudness', 'tempo', 'timeSignature')

data\_ug<-split(data\_u1,data\_u1['mode'])

delay0 <- describe(data\_ug$`0`[col\_gb])

word\_delay<-xtable\_to\_flextable(xtable(delay0))

doc = body\_add\_flextable(doc, word\_delay)

delay1 <- describe(data\_ug$`1`[col\_gb])

word\_delay<-xtable\_to\_flextable(xtable(delay1))

doc = body\_add\_flextable(doc, word\_delay)

print(doc, 'groupby&describe.docx')

data\_u1\_m<-melt(data\_u1, id.vars = 'mode', measure.vars=col\_gb, variable.name='variable')

data\_u1\_m$mode<-as.factor(data\_u1\_m$mode)

ggplot(data=data\_u1\_m, aes(x=mode, y=value))+geom\_boxplot(aes(fill=variable))+facet\_wrap(~variable, scales = 'free')

# glm model

doc = read\_docx()

g<- glm(count~duration+keySignature+popularity+energy+mode+loudness+tempo+timeSignature+year, data = data\_u1,

family = poisson(),control = list(maxit = 100))

g.sum<-summary(g)

word\_glm<-xtable\_to\_flextable(xtable(g.sum))

doc = body\_add\_flextable(doc, word\_glm)

print(doc, 'glmsum.docx')

参考文献

略