1.

a.

#先建立一個空向量 x，再依序放入 a+e 共 20 個值

Code:

```
X=c()
while(length(x)<20){
   a=runif(1,0,10)
   e=rnorm(1,0,2)
   if(0<=a+e & a+e<=11){
      xi = a+e
   }
   x=c(x,xi)
}
x
```

```
> x
 [1]  7.780024  2.795286  4.860296  4.631030  4.044925  9.723212  9.723212  5.765007
 [9]  8.621602  8.621602  8.160672  8.285045  1.827003  1.827003 10.701742  1.358490
[17]  1.358490  8.955643  4.001860  6.690968
```

b.

Code:

```
cauchy <- function(theta, xi) {
   f=(-2)*sum((theta-xi)/(1+(theta-xi)^2))
   return(f)
}
```

c.

Code:

```
cauchy(0.3,x)
```

```
> cauchy(0.3,x)
[1] 7.404879
```

2.

a.

#先在 song(此題的 data)裡面建一個 instru_prob 的欄位（所有值先設為 NA），再利用 which()從 instrumentalness 篩選出 low、mid、high 的值並取代掉原本的 NA 值

Code:

```
library(tidyverse)
```

```
song <- read_csv("song_data.csv")
song$instru_prob=NA
song$instru_prob[which(song$instrumentalness>=0
&song$instrumentalness<0.4)]="low"
song$instru_prob[which(song$instrumentalness>=0.4
&song$instrumentalness<0.8)]="mid"
song$instru_prob[which(song$instrumentalness>=0.8)]="high"
```

loudness	audio_mode	speechiness	tempo	time_signature	audio_valence	instru_prob
-4.095	1	0.0294	167.060	4	0.4740	low
-6.407	0	0.0498	105.256	4	0.3700	low
-7.828	1	0.0792	123.881	4	0.3240	mid
-4.938	1	0.1070	122.444	4	0.1980	low
-5.065	1	0.0313	172.011	4	0.5740	low
-3.169	0	0.1240	189.931	4	0.3200	low
-3.659	0	0.0624	90.578	4	0.7240	low
-3.435	1	0.0855	105.046	4	0.5370	low
-3.660	1	0.0917	148.112	4	0.2340	low
-5.653	1	0.0540	153.398	4	0.3740	low

b.

#instru_prob 轉成 factor
#attach(song)方便後續引用 song 內部的變數
#刪去 song_name（無關）和 instrumentalness（instru_prob 有同樣效力）兩個欄位
#先套用除了 song_popularity 以外所有變數 fit 出一個迴歸模型 f1
#因為變數很多，所以直接對 f1 利用 stepAIC 讓 R 逐步自動找出 AIC 最小的組合，該組合先後刪去了 audio_mode 和 speechiness，然後再 fit 出 f2
#利用 glance 比較 f1、f2 的統計值，發現兩者 r square 差異不大，解釋能力似乎差不多，p-value 也都很小
#再利用 anova(f2,f1)，配出 f1 之 p-value 較大，因此 f1 較不為顯著，選擇較簡單的 f2 模型預測歌曲的歡迎程度
#由 summary(f2)可得到 f2 模型為 y=
5.748e+01
-4.241e-06*(song_duration_ms)
-4.248e+00*(acousticness)
+1.229e+01*(danceability)
-1.268e+01*(energy)
-6.847e-02*(key)
-4.465e+00*(liveness)
+8.027e-01*(loudness)

-1.154e-02*(tempo)
+1.517e+00*(time_signature)
-8.223e+00*(audio_valence)
+5.429e+00*(instru_problow)
-4.631e+00*(instru_probmid)
再根據 beta 的正負和“*”越多越顯著，可知 acousticness, energy, liveness, audio_valence 和 instru_probmid 為顯著負相關；danceability, loudness, time_signature 和 instru_problow 為顯著正相關
#instru_probhigh 沒有跑出來可能是因為 18835 筆資料中只有 884 筆之 instru_prob 為 high（太少）
Code:

```
song$instru_prob <- as.factor(song$instru_prob)
attach(song)

library(broom)

newsong<-song[,-c(1,7)]
str(newsong)

f1<-lm(song_popularity ~ . ,data=newsong)

library(MASS)
step=stepAIC(f1, direction="backward")
```

```
Step:  AIC=115410.3
song_popularity ~ song_duration_ms + acousticness + danceability +
    energy + key + liveness + loudness + tempo + time_signature +
    audio_valence + instru_prob

                  Df Sum of Sq     RSS    AIC
<none>                         8619435 115410
- key              1      1152 8620587 115411
- song_duration_ms 1      1168 8620603 115411
- tempo            1      1961 8621396 115413
- time_signature   1      3696 8623131 115416
- liveness         1      7477 8626912 115425
- acousticness     1     14313 8633748 115440
- energy           1     37607 8657043 115490
- danceability     1     51896 8671331 115521
- audio_valence    1     55612 8675047 115529
- loudness         1     61862 8681297 115543
- instru_prob      2     80217 8699652 115581
> 
```

f2=lm(song_popularity ~ .-audio_mode-speechiness, data=newsong)

glance(f1)

glance(f2)

```
> glance(f1)
# A tibble: 1 × 12
  r.squared adj.r.squared sigma statistic   p.value    df  logLik     AIC     BIC deviance
      <dbl>         <dbl> <dbl>     <dbl>     <dbl> <dbl>   <dbl>   <dbl>   <dbl>    <dbl>
1    0.0464        0.0457  21.4      65.4 7.98e-182    14 -84417. 168866. 168991. 8618455.
# … with 2 more variables: df.residual <int>, nobs <int>
> glance(f2)
# A tibble: 1 × 12
  r.squared adj.r.squared sigma statistic   p.value    df  logLik     AIC     BIC deviance
      <dbl>         <dbl> <dbl>     <dbl>     <dbl> <dbl>   <dbl>   <dbl>   <dbl>    <dbl>
1    0.0463        0.0457  21.4      76.1 3.01e-183    12 -84418. 168864. 168974. 8619435.
```

anova(f2,f1)

```
> anova(f2,f1)
Analysis of Variance Table

Model 1: song_popularity ~ (song_duration_ms + acousticness + danceability +
    energy + key + liveness + loudness + audio_mode + speechiness +
    tempo + time_signature + audio_valence + instru_prob) - audio_mode -
    speechiness
Model 2: song_popularity ~ song_duration_ms + acousticness + danceability +
    energy + key + liveness + loudness + audio_mode + speechiness +
    tempo + time_signature + audio_valence + instru_prob
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1  18822 8619435
2  18820 8618455  2    979.67 1.0696 0.3432
```

summary(f2)

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       5.748e+01  2.895e+00  19.859  < 2e-16 ***
song_duration_ms -4.241e-06  2.655e-06  -1.597   0.1103
acousticness     -4.248e+00  7.599e-01  -5.591 2.30e-08 ***
danceability      1.229e+01  1.155e+00  10.645  < 2e-16 ***
energy           -1.268e+01  1.399e+00  -9.062  < 2e-16 ***
key              -6.847e-02  4.317e-02  -1.586   0.1127
liveness         -4.465e+00  1.105e+00  -4.041 5.35e-05 ***
loudness          8.027e-01  6.906e-02  11.623  < 2e-16 ***
tempo            -1.154e-02  5.576e-03  -2.069   0.0385 *
time_signature    1.517e+00  5.339e-01   2.841   0.0045 **
audio_valence    -8.223e+00  7.462e-01 -11.020  < 2e-16 ***
instru_problow    5.429e+00  8.102e-01   6.700 2.14e-11 ***
instru_probmid   -4.631e+00  1.107e+00  -4.182 2.90e-05 ***
```