The Academic Year Project
in Core Data Analysis Course

EMI Music Data Science Hackaton

Sergei Korolev

Mikhail Sveshnikov

The following dataset is interesting due to it's variety of personal information of listeners. It can give valuable point of view on the relations between workload, age, gender and hours spent listening to music or importance of music in one's life.

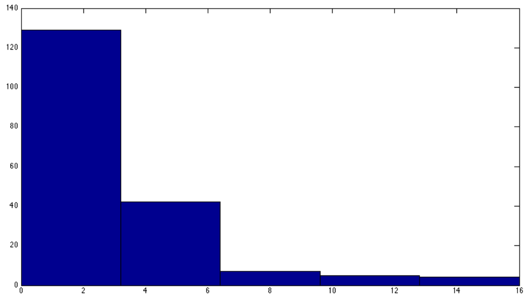
“Csv file gives data about the respondents themselves, including their attitude towards music. The columns include:

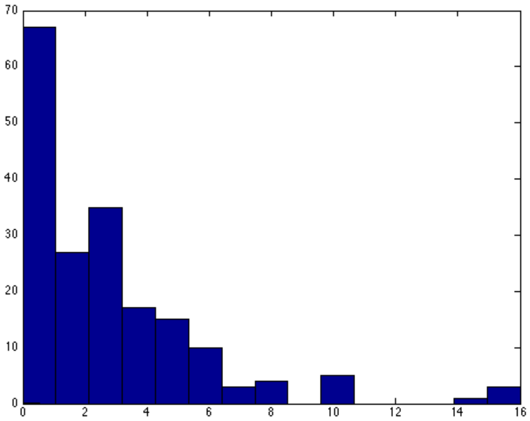
- User. The anonymised user identifier
- Gender. Male/female
- Age. The respondent’s age, in years.
- Working status. Whether they are working full-time/retired/etc.
- Region. The region of the United Kingdom where they live.
- MUSIC. The respondent’s view on the importance of music in his/her life.
- LIST_OWN. An estimate for the number of daily hours spent listening to music they own or have chosen.
- LIST_BACK. An estimate for the number of daily hours the respondent spends listening to background music/music they have not chosen.
- Music habit questions. Each of these asks the respondent to rate, on a scale of X-100, whether they agree with the following:
    1. I enjoy actively searching for and discovering music that I have never heard before
    2. I find it easy to find new music
    3. I am constantly interested in and looking for more music
    4. I would like to buy new music but I don’t know what to buy
    5. I used to know where to find music
    6. I am not willing to pay for music
    7. I enjoy music primarily from going out to dance
    8. Music for me is all about nightlife and going out
    9. I am out of touch with new music
    10. My music collection is a source of pride
    11. Pop music is fun
    12. Pop music helps me to escape
    13. I want a multi media experience at my fingertips wherever I go
    14. I love technology
    15. People often ask my advice on music - what to listen to
    16. I would be willing to pay for the opportunity to buy new music pre-release
    17. I find seeing a new artist / band on TV a useful way of discovering new music
    18. I like to be at the cutting edge of new music
    19. I like to know about music before other people“

Dataset contains information about 26725 users and their music habits. There are lots of questions to be asked while looking at such big dataset, such as “How musically active are listeners of certain age group on average?”, “Does the workload cause people to spend less time listening to the music?” and so on.

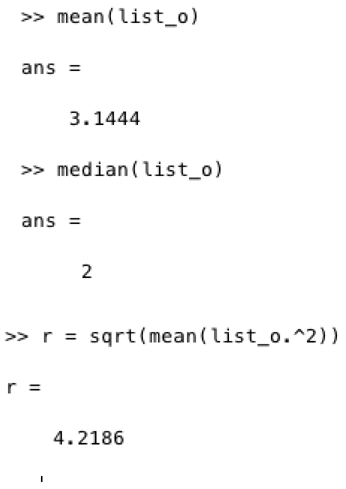
The source of the dataset and information on it above is http://www.kaggle.com/c/MusicHackathon.

I. For the first assignment the subset of female full-time students from Midlands was taken from the dataset. We looked at the feature of an estimate for the number of daily hours spent listening to music they own or have chosen. The histogram of higher resolution showed small dip in the number of 2 hour listeners that was not there on the small resolution histogram:

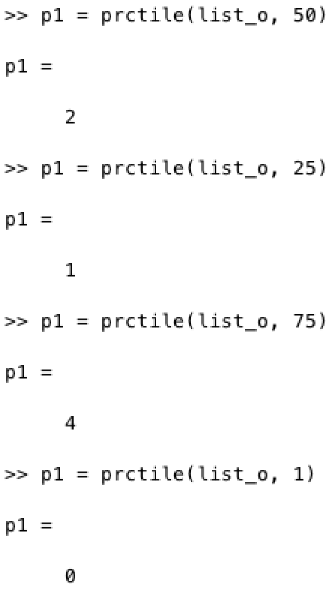




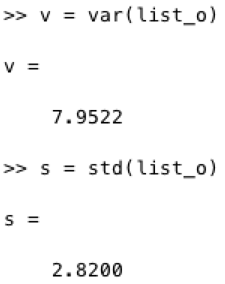
Then mean, median, root mean square, percentiles, variance and standard deviation of the set of values of this feature were taken:



```
>> mean(list_o)

ans =

    3.1444

>> median(list_o)

ans =

     2

>> r = sqrt(mean(list_o.^2))

r =

    4.2186
```



```
>> p1 = prctile(list_o, 50)

p1 =

     2

>> p1 = prctile(list_o, 25)

p1 =

     1

>> p1 = prctile(list_o, 75)

p1 =

     4

>> p1 = prctile(list_o, 1)

p1 =

     0
```

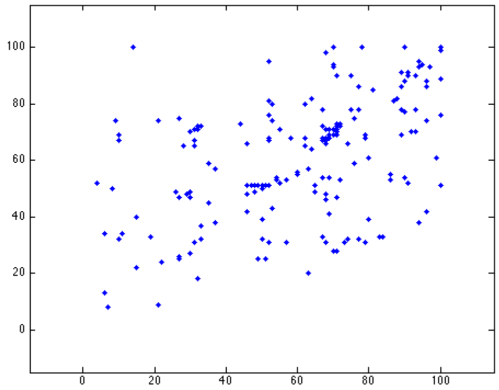


```
>> v = var(list_o)

v =

    7.9522

>> s = std(list_o)

s =

    2.8200
```

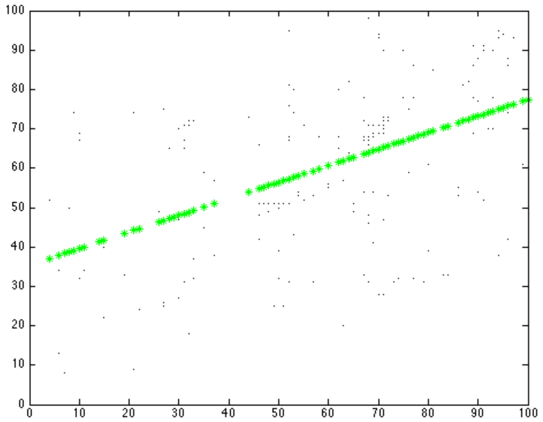
II. For the second assignment we chose two features which showed the answers to the first and second questions:

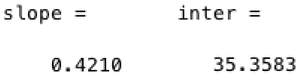
1. I enjoy actively searching for and discovering music that I have never heard before
2. I find it easy to find new music

The scatter plot of these two values looks like this:



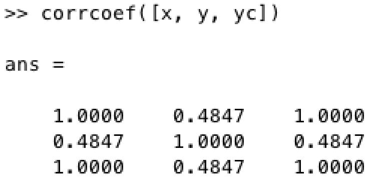
Looking at the groups of
the values it's easy to decide that such plot can be approximated using linear function. We calculated the slope and initial value of this linear function and came up with following approximation:



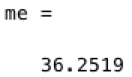


```
slope =        inter =

    0.4210       35.3583
```

The correlation coefficients and determinacy coefficients are as follows:

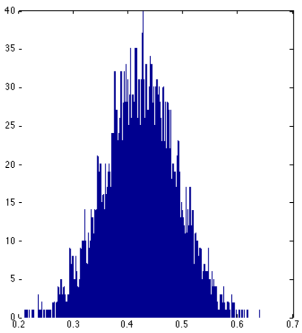


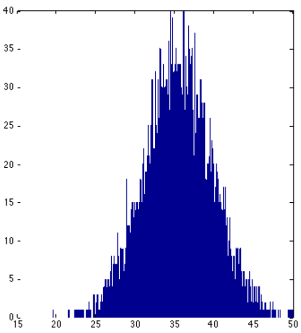
```
>> corrcoef([x, y, yc])

ans =

    1.0000    0.4847    1.0000
    0.4847    1.0000    0.4847
    1.0000    0.4847    1.0000
```

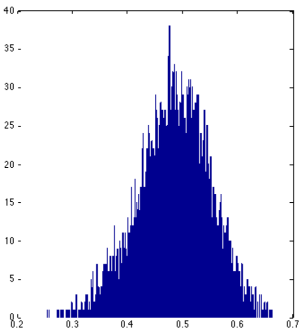


```
me =

   36.2519
```

Then we used bootstrap technique to calculate random slopes, initials and correlation coefficients:


40
35
30
25
20
15
10
5
0
0.2
0.3
0.4
0.5
0.6
0.7


40
35
30
25
20
15
10
5
0
15
20
25
30
35
40
45
50


40
35
30
25
20
15
10
5
0
0.2
0.3
0.4
0.5
0.6
0.7