1. Overview

As stated in the description, the data for this competition comes from a experimental set-up used to study earthquake physics. Our goal is to predict the time remaining before the next laboratory earthquake. The only feature we have is the seismic signal (acoustic data) which has integer values in a limited range.

Training data: single, continuous segment of experimental data.

Test data: consists of a folder containing many small segments. The data within each test file is continuous, but the test files do not represent a continuous segment of the experiment.

There are a lot of files in this competition, so let's start with the folders structure:

The test folder has 2624 csv files (segments):

```
['seg_0b082e.csv', 'seg_9e7dff.csv', 'seg_b6c10d.csv', 'seg_4435bd.csv',
'seg_c09a41.csv', 'seg_31ddc5.csv', 'seg_71238c.csv', 'seg_6a05e7.csv',
'seg_d47aba.csv', 'seg_eea20e.csv']
```

```
Number of files in the test folder 2624
```

Each segment contains 150,000 acoustic records:

```
Segment shape (150000, 1)
```

Out[5]:

	acoustic_data
0	3

1	10
2	4
3	4
4	1

There is one file in the test folder for each prediction (seg_id) in sample_submission:

```
Submission shape (2624, 2)
```

Out[6]:

	seg_id	time_to_failure
0	seg_00030f	0
1	seg_0012b5	0
2	seg_00184e	0
3	seg_003339	0
4	seg_0042cc	0

2. Training data

One huge csv file has all the training data, which is a single continuous experiment. There are only two columns in this file:

Acoustic data (int16): the seismic signal Time to failure (float64): the time until the next laboratory earthquake (in seconds) No missing values for both columns

```
CPU times: user 2min 18s, sys: 17.1 s, total: 2min 35s
Wall time: 2min 36s
```

Out[8]:

	acoustic_data	time_to_failure
0	12	1.469099998474121
1	6	1.469099998474121
2	8	1.469099998474121
3	5	1.469099998474121
4	8	1.469099998474121

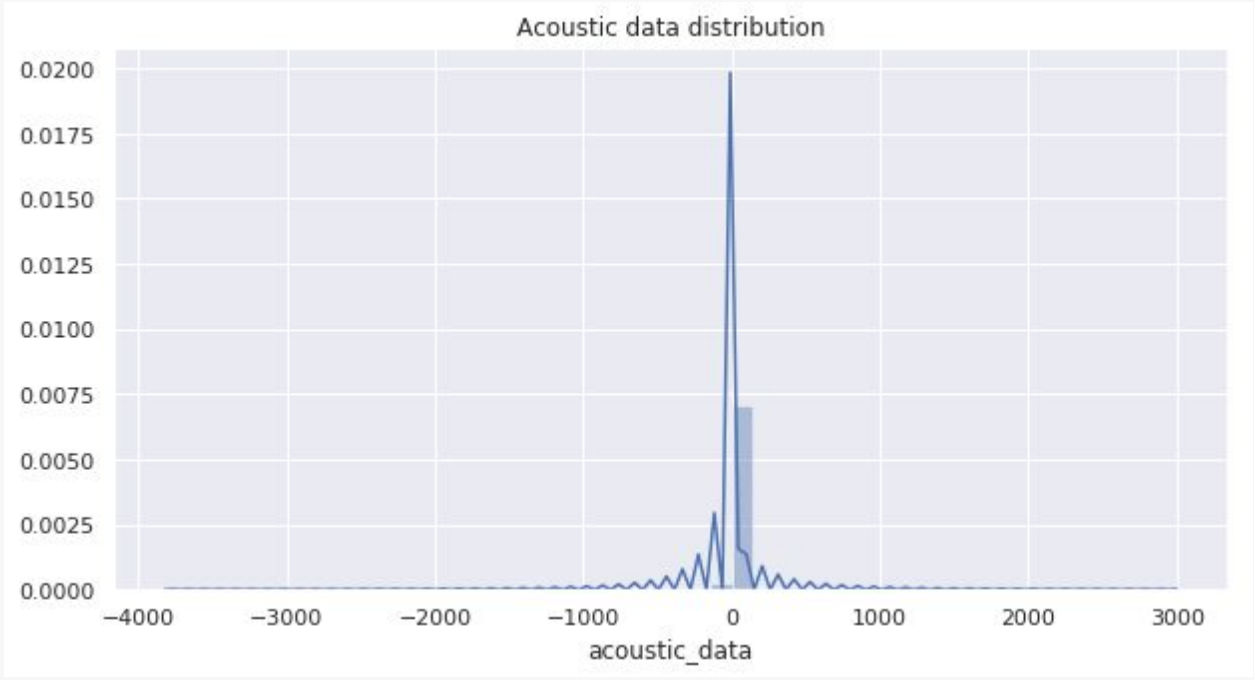
```
The training data has 629145480 rows and 2 columns
```

Acoustic data

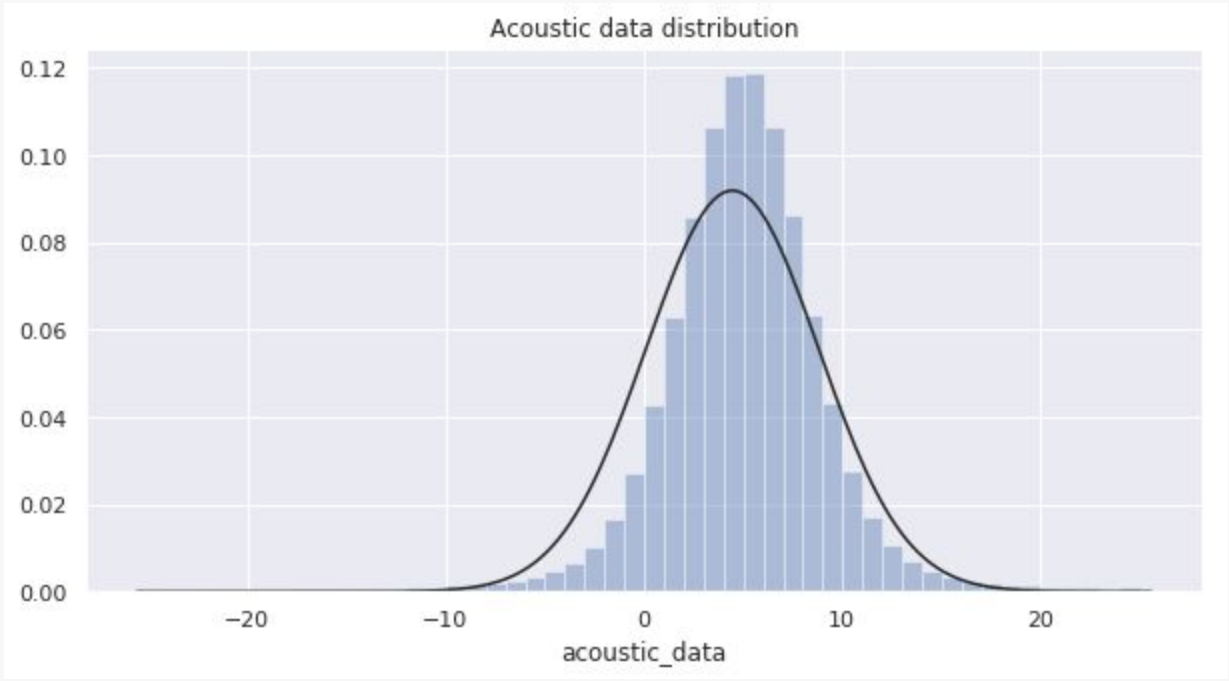
Our single feature are integers in the range [-5515, 5444] with mean 4.52

```
count    6.29145480e+08
mean     4.51946757e+00
std      1.07357072e+01
min     -5.51500000e+03
25%      2.00000000e+00
50%      5.00000000e+00
75%      7.00000000e+00
max      5.44400000e+03
Name: acoustic_data, dtype: float64
```

The plot below is using a 1% random sample (~6M rows):



There are outliers in both directions; let's try to plot the same distribution with x in the range -25 to 25. The black line is the closest normal distribution (gaussian) possible.



Time to failure

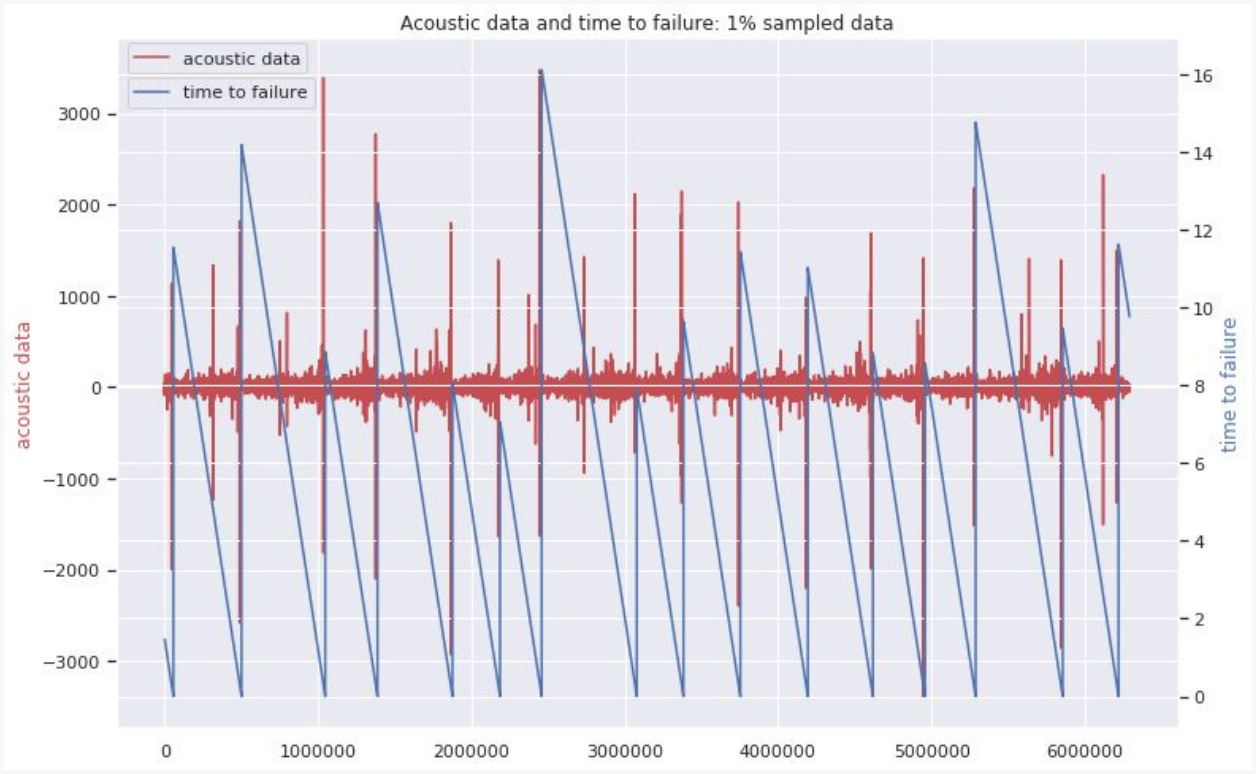
Now let's check the target variable, which is given in seconds:

```
count    6.29145480e+08
mean     4.47708428e-01
std      2.61278939e+00
min      9.55039650e-05
25%      2.62599707e+00
50%      5.34979773e+00
75%      8.17339516e+00
max      1.61074009e+01
Name: time_to_failure, dtype: float64
```

The min value is very close to zero (around 10^-5) and the max is 16 seconds. Now the distribution for the random sample:

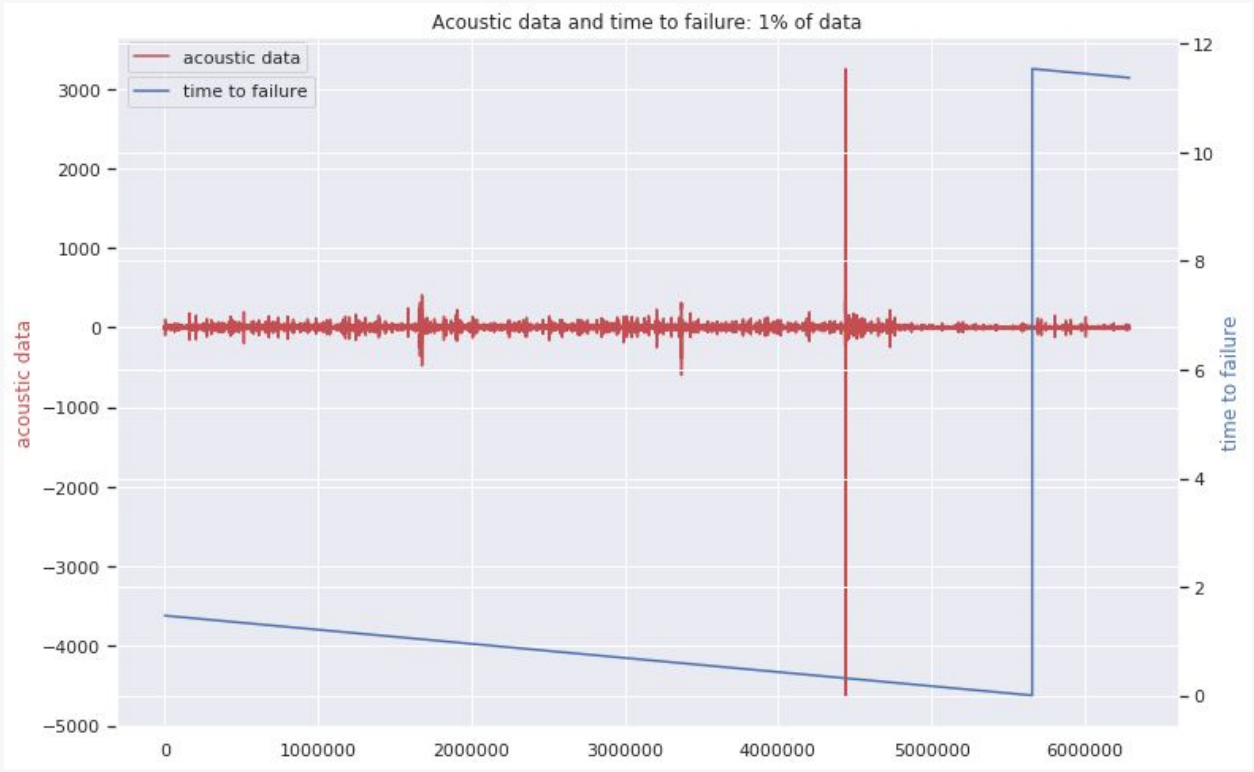
Timeseries

Let's see how both variables change over time. The red line is the acoustic data and the blue one is the time to failure:



On a plot above we can see, that training data has 16 earthquakes. The shortest time to failure is 1.5 seconds for the first earthquake and 7seconds for the 7th, while the longest is around 16 seconds.

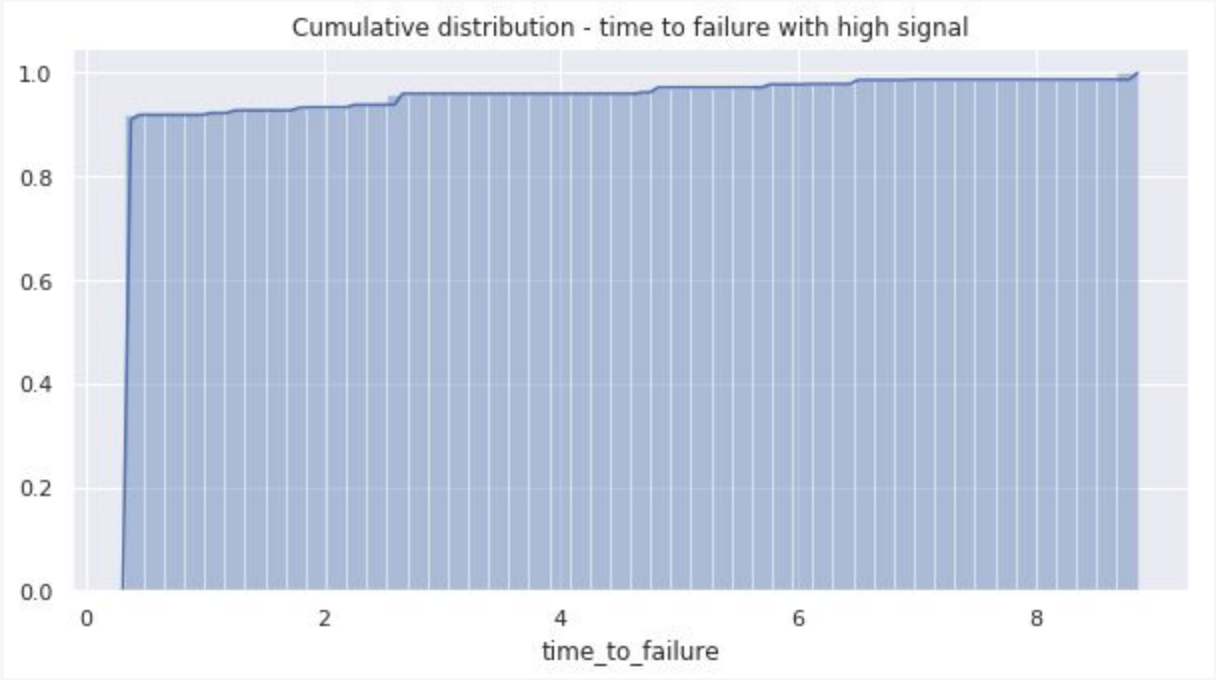
- We can see that usually acoustic data shows huge fluctuations just before the failure and the nature of data is cyclical;
- Another important point: visually failures can be predicted as cases when huge fluctuations in signal are followes by small signal values. This could be useful for predicting "time_to_failure" changes from 0 to high values;



On this zoomed-in-time plot we can see that actually the large oscillation before the failure is not quite in the last moment. There are also trains of intense oscillations proceeding the large one and also some oscillations with smaller peaks after the large one. Then, after some minor oscillations, the failure occurs. Interesting thing to check is the time between high levels of seismic signal and the earthquakes. We are considering any acoustic data with absolute value greater than 1000 as a high level:

```
count    11325.00000000
mean         0.64454830
std          1.32147193
min          0.31079626
25%          0.31549615
50%          0.31689683
75%          0.32029617
```

```
max             8.86059952
```



More than 90% of high acoustic values are around 0.31 seconds before an earthquake!