Quantitative Researcher

Exercise

By Dor Gaon

Background:

On March 15, 2259, a change has been made to the underwriting request, in order to facilitate the process and make it faster.

In order to check whether the procedure has really become faster, I will perform analyzes using Python to check the issue.

For the purpose of the assignment there are three data files:

1. Applicants table - personal details of the applicants who scheduled an underwriting session

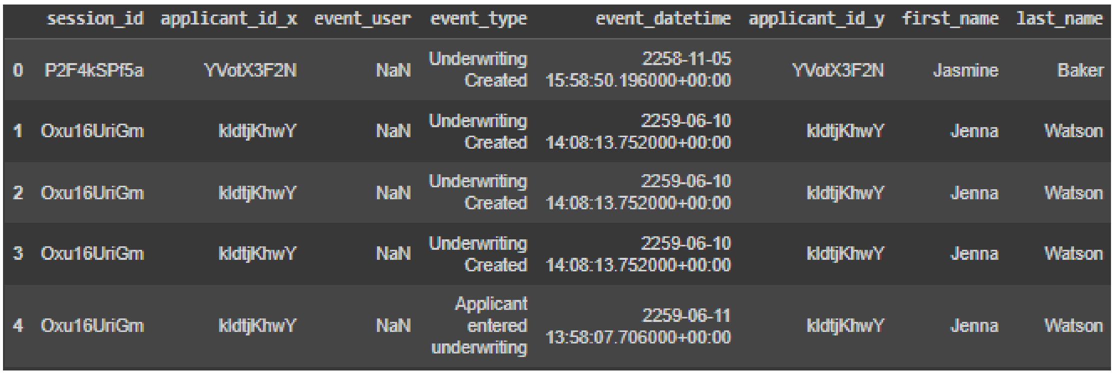
2. Sessions table - administrative details about the session and its status

3. Events table - events related to the underwriting session

In each file there is a session id, which we will use to merge the three files into one complete data file called "merged_df".

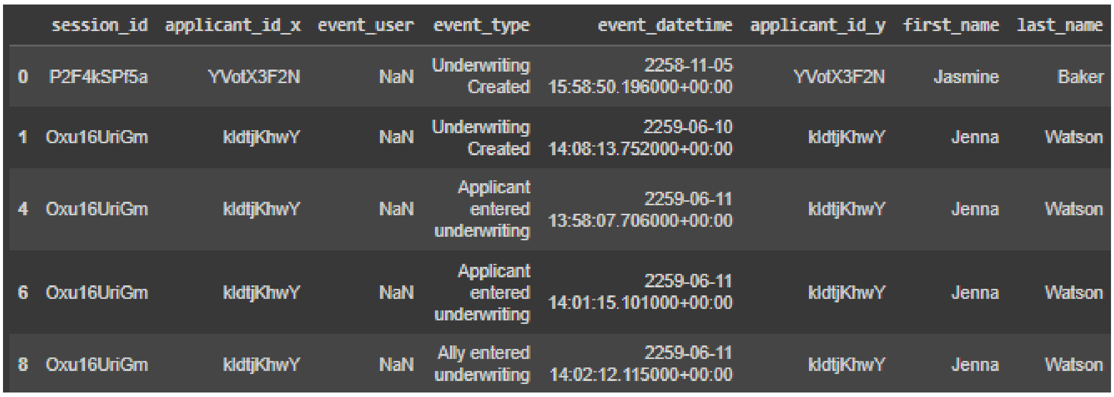
Task:

Results of file consolidation:



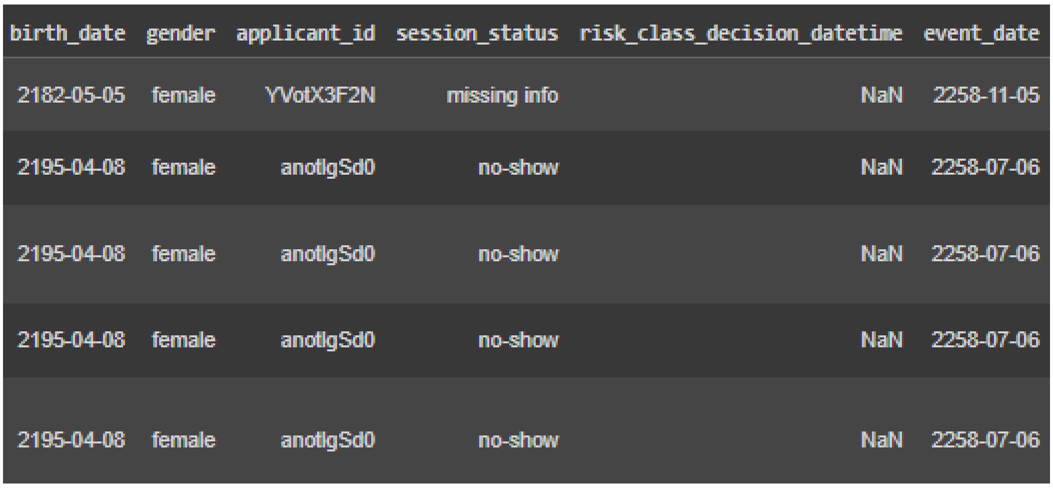
	session_id	applicant_id_x	event_user	event_type	event_datetime	applicant_id_y	first_name	last_name
0	P2F4kSPf5a	YVotX3F2N	NaN	Underwriting Created	2258-11-05 15:58:50.196000+00:00	YVotX3F2N	Jasmine	Baker
1	Oxu16UriGm	kldtjKhwY	NaN	Underwriting Created	2259-06-10 14:08:13.752000+00:00	kldtjKhwY	Jenna	Watson
2	Oxu16UriGm	kldtjKhwY	NaN	Underwriting Created	2259-06-10 14:08:13.752000+00:00	kldtjKhwY	Jenna	Watson
3	Oxu16UriGm	kldtjKhwY	NaN	Underwriting Created	2259-06-10 14:08:13.752000+00:00	kldtjKhwY	Jenna	Watson
4	Oxu16UriGm	kldtjKhwY	NaN	Applicant entered underwriting	2259-06-11 13:58:07.706000+00:00	kldtjKhwY	Jenna	Watson

From a quick check, there are duplicate lines, so filter them in Python.
Results after filtering:



	session_id	applicant_id_x	event_user	event_type	event_datetime	applicant_id_y	first_name	last_name
0	P2F4kSPf5a	YVotX3F2N	NaN	Underwriting Created	2258-11-05 15:58:50.196000+00:00	YVotX3F2N	Jasmine	Baker
1	Oxu16UriGm	kldtjKhwY	NaN	Underwriting Created	2259-06-10 14:08:13.752000+00:00	kldtjKhwY	Jenna	Watson
4	Oxu16UriGm	kldtjKhwY	NaN	Applicant entered underwriting	2259-06-11 13:58:07.706000+00:00	kldtjKhwY	Jenna	Watson
6	Oxu16UriGm	kldtjKhwY	NaN	Applicant entered underwriting	2259-06-11 14:01:15.101000+00:00	kldtjKhwY	Jenna	Watson
8	Oxu16UriGm	kldtjKhwY	NaN	Ally entered underwriting	2259-06-11 14:02:12.115000+00:00	kldtjKhwY	Jenna	Watson

Since we received time data that includes both date and time we would like to split the date. This way we can intersect between the period before the change - March 15, 2259 and the period after it.
The new variable is called "event_date":



birth_date	gender	applicant_id	session_status	risk_class_decision_datetime	event_date
2182-05-05	female	YVotX3F2N	missing info	NaN	2258-11-05
2195-04-08	female	anotlgSd0	no-show	NaN	2258-07-06
2195-04-08	female	anotlgSd0	no-show	NaN	2258-07-06
2195-04-08	female	anotlgSd0	no-show	NaN	2258-07-06
2195-04-08	female	anotlgSd0	no-show	NaN	2258-07-06

Now we will divide the data into two, before and after the date March 15, 2259.

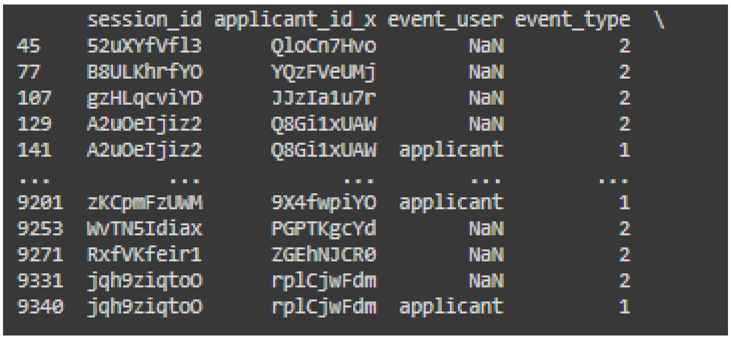
We will call the periods:

before_date = the date before

after_date = the date after

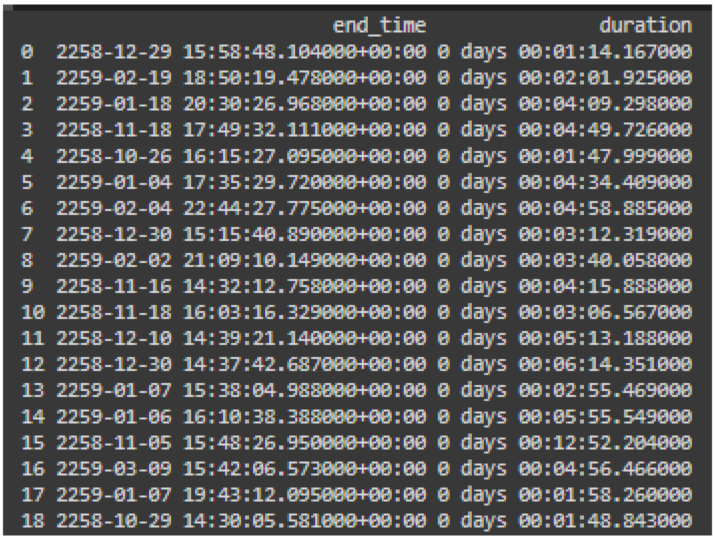
Now that we have two files, we would like to examine the moment when the applicant completes versus submits the form reporting on the test results.

In order to do this, we will change the 'end_of_underwriting' figure to a numerical figure – 1 which defines a start time.
and the figure 'Ally submitted test results' to number 2 which defines an end time.



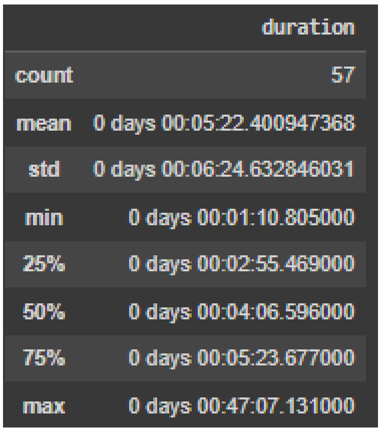
```
      session_id applicant_id_x event_user event_type  \
45    52uXYfVfl3      QloCn7Hvo        NaN          2
77    B8ULKhrfYO      YQzFVeUMj        NaN          2
107   gzHLqcviYD      JJzIa1u7r        NaN          2
129   A2uOeIjiz2      Q8Gi1xUAW        NaN          2
141   A2uOeIjiz2      Q8Gi1xUAW  applicant          1
...          ...            ...        ...        ...
9201  zKCpmFzUWM      9X4fwpiYO  applicant          1
9253  WvTN5Idiax      PGPTKgcYd        NaN          2
9271  RxfVKfeir1      ZGEhNJCR0        NaN          2
9331  jqh9ziqtoO      rplCjwFdm        NaN          2
9340  jqh9ziqtoO      rplCjwFdm  applicant          1
```

Now we will make a comparison between the time before and the time after:



```
                             end_time               duration
0  2258-12-29 15:58:48.104000+00:00 0 days 00:01:14.167000
1  2259-02-19 18:50:19.478000+00:00 0 days 00:02:01.925000
2  2259-01-18 20:30:26.968000+00:00 0 days 00:04:09.298000
3  2258-11-18 17:49:32.111000+00:00 0 days 00:04:49.726000
4  2258-10-26 16:15:27.095000+00:00 0 days 00:01:47.999000
5  2259-01-04 17:35:29.720000+00:00 0 days 00:04:34.409000
6  2259-02-04 22:44:27.775000+00:00 0 days 00:04:58.885000
7  2258-12-30 15:15:40.890000+00:00 0 days 00:03:12.319000
8  2259-02-02 21:09:10.149000+00:00 0 days 00:03:40.058000
9  2258-11-16 14:32:12.758000+00:00 0 days 00:04:15.888000
10 2258-11-18 16:03:16.329000+00:00 0 days 00:03:06.567000
11 2258-12-10 14:39:21.140000+00:00 0 days 00:05:13.188000
12 2258-12-30 14:37:42.687000+00:00 0 days 00:06:14.351000
13 2259-01-07 15:38:04.988000+00:00 0 days 00:02:55.469000
14 2259-01-06 16:10:38.388000+00:00 0 days 00:05:55.549000
15 2258-11-05 15:48:26.950000+00:00 0 days 00:12:52.204000
16 2259-03-09 15:42:06.573000+00:00 0 days 00:04:56.466000
17 2259-01-07 19:43:12.095000+00:00 0 days 00:01:58.260000
18 2258-10-29 14:30:05.581000+00:00 0 days 00:01:48.843000
```

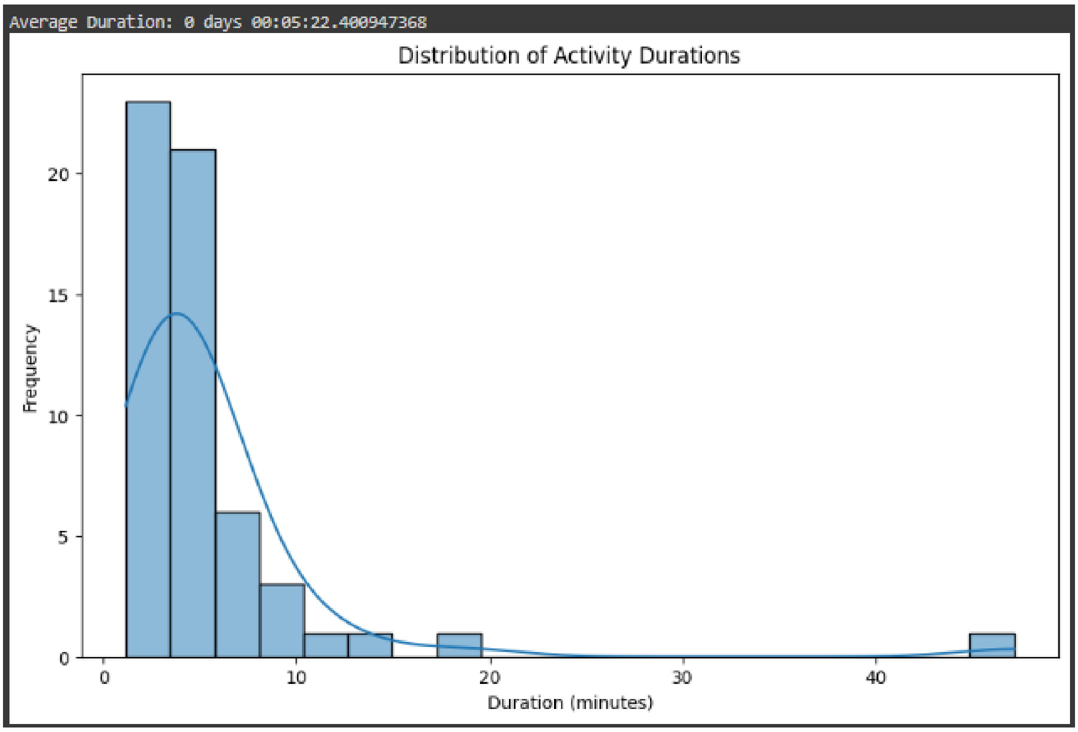
We will examine the data obtained for the period before the change:



	duration
count	57
mean	0 days 00:05:22.400947368
std	0 days 00:06:24.632846031
min	0 days 00:01:10.805000
25%	0 days 00:02:55.469000
50%	0 days 00:04:06.596000
75%	0 days 00:05:23.677000
max	0 days 00:47:07.131000

It can be seen that the average time is about 5: 22 minutes.

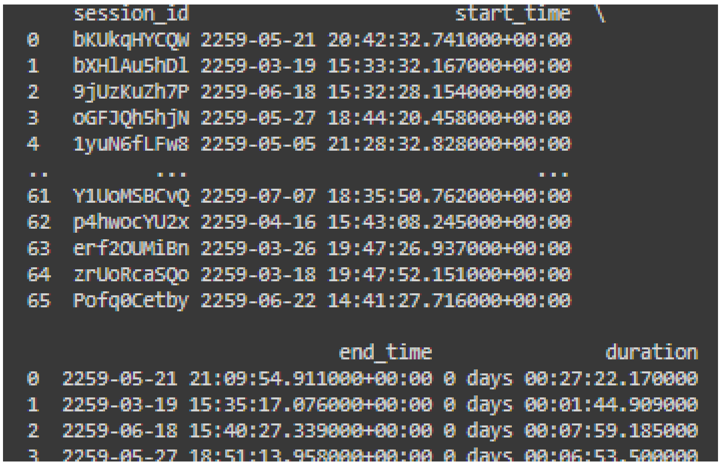
According to the data we received, the maximum time gap is 47 minutes.



```
Average Duration: 0 days 00:05:22.400947368
```

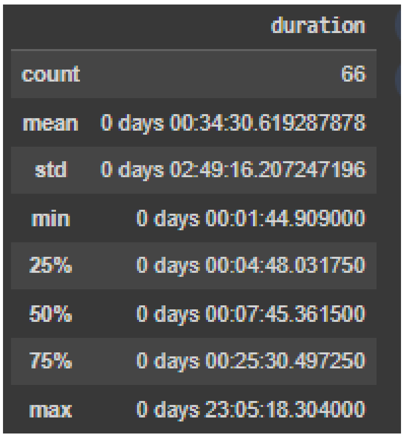
It can be seen that most of the sample shows a noticeably short time around the 5 minutes, and that the maximum time - 47 minutes is unusual.
We get a positive asymmetric distribution with a tail to the right.

Now we will do the same test for the time after the change:



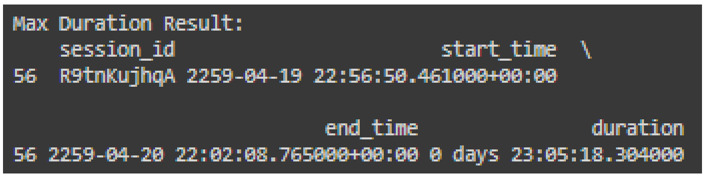
```
     session_id                        start_time  \
0   bKUkqHYCQW 2259-05-21 20:42:32.741000+00:00
1   bXHlAu5hDl 2259-03-19 15:33:32.167000+00:00
2   9jUzKuZh7P 2259-06-18 15:32:28.154000+00:00
3   oGFJQh5hjN 2259-05-27 18:44:20.458000+00:00
4   1yuN6fLFw8 2259-05-05 21:28:32.828000+00:00
..         ...                              ...
61  Y1UoMSBCvQ 2259-07-07 18:35:50.762000+00:00
62  p4hwocYU2x 2259-04-16 15:43:08.245000+00:00
63  erf2OUMiBn 2259-03-26 19:47:26.937000+00:00
64  zrUoRcaSQo 2259-03-18 19:47:52.151000+00:00
65  Pofq0Cetby 2259-06-22 14:41:27.716000+00:00

                           end_time               duration
0  2259-05-21 21:09:54.911000+00:00 0 days 00:27:22.170000
1  2259-03-19 15:35:17.076000+00:00 0 days 00:01:44.909000
2  2259-06-18 15:40:27.339000+00:00 0 days 00:07:59.185000
3  2259-05-27 18:51:13.958000+00:00 0 days 00:06:53.500000
```

We will examine the data obtained for the period after the change:



	duration
count	66
mean	0 days 00:34:30.619287878
std	0 days 02:49:16.207247196
min	0 days 00:01:44.909000
25%	0 days 00:04:48.031750
50%	0 days 00:07:45.361500
75%	0 days 00:25:30.497250
max	0 days 23:05:18.304000

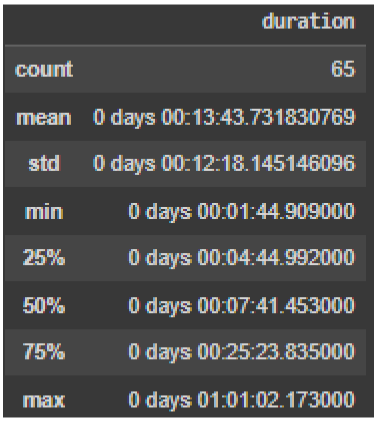
We have accepted that the average waiting time is over half an hour,
In addition, it can be seen that a maximum figure of about 23 hours is obtained, this is a very unusual figure that is many standard deviations from the average,



```
Max Duration Result:
    session_id                        start_time  \
56  R9tnKujhqA 2259-04-19 22:56:50.461000+00:00

                           end_time               duration
56 2259-04-20 22:02:08.765000+00:00 0 days 23:05:18.304000
```

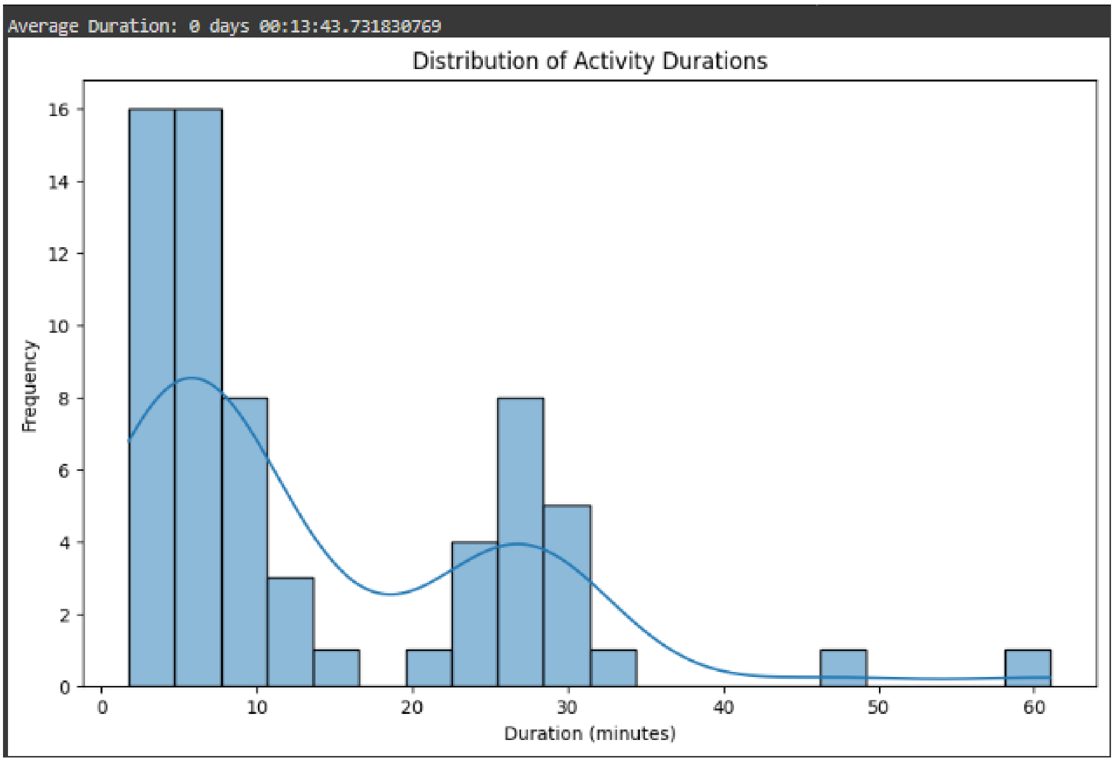
After examining the figure in depth, we will delete the figure from the list due to extremely high deviations (whether human or technical)

We will examine the data without the outlier.



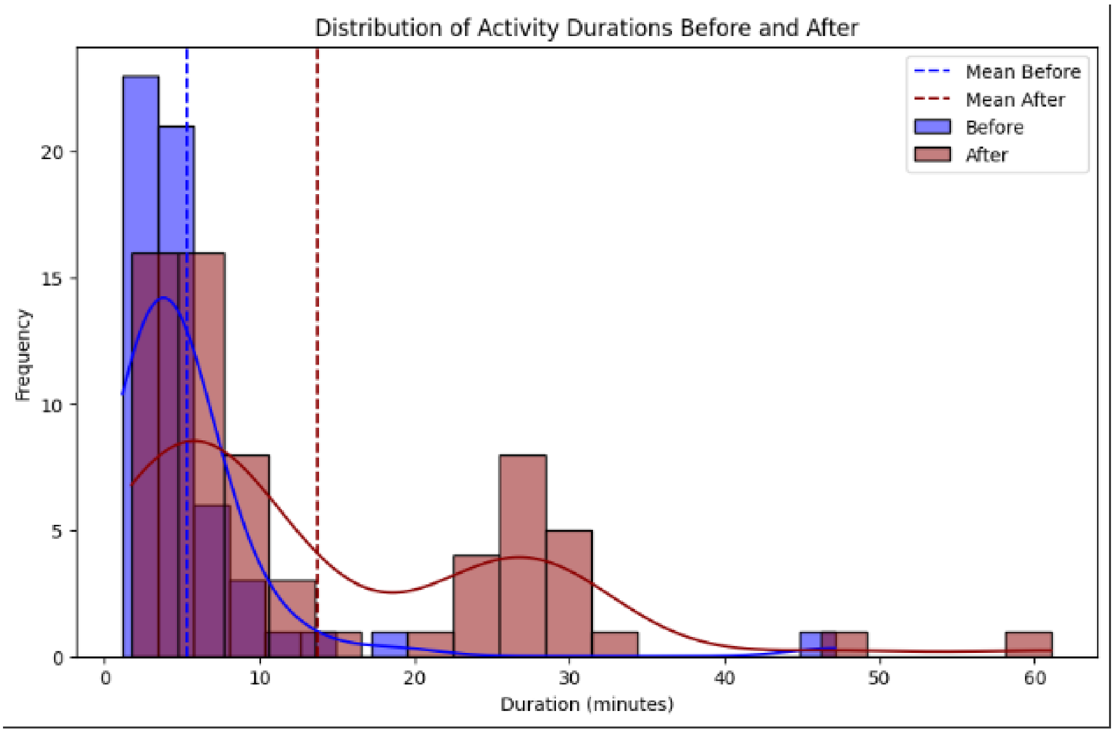
	duration
count	65
mean	0 days 00:13:43.731830769
std	0 days 00:12:18.145146096
min	0 days 00:01:44.909000
25%	0 days 00:04:44.992000
50%	0 days 00:07:41.453000
75%	0 days 00:25:23.835000
max	0 days 01:01:02.173000

we see that we received an average time of 13:43 minutes - significantly lower than what we received before.
The standard deviation is 12:18 minutes with a maximum observation of about an hour, this is an unusual figure but significantly more realistic than the 23 hours we received before.



```
Average Duration: 0 days 00:13:43.731830769
```

We will examine the schedule of times after the change and it appears that there are 2 observations far from the rest - one around 50 minutes and one around 60 minutes.
In addition, we see that the distribution of times is very volatile, with a cluster at the beginning of the graph indicating a low time and another cluster around the 25 minutes.



According to the data we received, we can determine that the change time after the change was not shortened, it was extended, the average time is now about 8 minutes longer than before