**1 Задание**

Считаю справедливой оплату в для выполнения асессором одного микрозадания из этого файла:
**18.93241444460593 * N**, это без одного пользователя (номер строки 534454), в данных которого допущена ошибка.
Ошибка заключается в том, что задание сдано раньше, чем он был назначен исполнителем. Разница в 12 минут 09 секунд.

Этапы решения:

1. Прочитал датасет, проверил количество строк.
2. Создал функцию для перевода обычного человеческого отображения даты в timestamp для того, чтобы в будущем посчитать их разницу.
3. Применил функцию добавив две новые колонки отображения времени в датасет.
4. После этого копировал датасет в новый, но без нескольких не нужных в данный момент столбцов (а именно нормального отображения времени).
5. Посчитал количество времени на одно задание.
   (closed_ts - assigned_ts) / Microtasks
6. И добавил новую колонку с временем на одно задание.
7. После этого заметил ошибку в данных, в строке 534454. И проверил эту строку в оригинальном (начальном) датасете. Визуально увидев тоже ошибку.

| | login | tid | Microtasks | assigned_ts | closed_ts |
|---|---|---|---|---|---|
| **534454** | login585 | 197340894 | 1 | 2017-05-24 16:13:06 | 2017-05-24 16:00:57 |

8. После устранения этой строки из временного датасета с которым работаю, подсчитал среднее значение времени потраченного на одно задание.

In [1]:
```
import pandas as pd
import time
import datetime
```

In [2]:
```
data = pd.read_csv('data_test_task/file1.txt', sep="\t", header=None, \
    names=["login", "tid", "Microtasks", "assigned_ts", "closed_ts"], low_memory=False)
```

In [3]:
```
data.shape
```

Out[3]:
```
(701828, 5)
```

In [4]:
```
data = data.drop([0])
```

In [5]:
```
data.shape
```

Out[5]:
```
(701827, 5)
```

In [6]:
```
print(data.head(3))
```

```
    login          tid Microtasks          assigned_ts            closed_ts
1  login0  190563850.0        4.0  2017-04-20 12:09:39  2017-04-20 13:13:01
2  login0  190561754.0        1.0  2017-04-20 12:10:30  2017-04-20 12:28:29
3  login0  190565906.0        4.0  2017-04-20 12:21:31  2017-04-20 13:30:10
```

In [7]:
```
def date_convert(date_to_convert):
    return int(time.mktime(datetime.datetime.strptime(date_to_convert, "%Y-%m-%d %H:%M:%S").timetuple()))
```

In [8]:
```
data['assigned_timestamp'] = data['assigned_ts'].apply(date_convert)
data['closed_timestamp'] = data['closed_ts'].apply(date_convert)
```

In [9]:
```
data.head(3)
```

Out[9]:

| | login | tid | Microtasks | assigned_ts | closed_ts | assigned_timestamp | closed_timestamp |
|---|---|---|---|---|---|---|---|
| **1** | login0 | 190563850.0 | 4.0 | 2017-04-20 12:09:39 | 2017-04-20 13:13:01 | 1492679379 | 1492683181 |
| **2** | login0 | 190561754.0 | 1.0 | 2017-04-20 12:10:30 | 2017-04-20 12:28:29 | 1492679430 | 1492680509 |
| **3** | login0 | 190565906.0 | 4.0 | 2017-04-20 12:21:31 | 2017-04-20 13:30:10 | 1492680091 | 1492684210 |

In [10]:
```
data.login.nunique()
```

Out[10]: 767

In [11]:
```
data.tid.nunique()
```

Out[11]: 635044

701827 Строка

767 Уникальных логин

641515 Уникальных заданий

In [12]:
```
data.tid = pd.to_numeric(data.tid, downcast='integer')
data.Microtasks = pd.to_numeric(data.Microtasks, downcast='integer')
```

In [13]:
```
data.tail()
```

Out[13]:

| | login | tid | Microtasks | assigned_ts | closed_ts | assigned_timestamp | closed_timestamp |
|---|---|---|---|---|---|---|---|
| **701823** | login766 | 195656026 | 1 | 2017-05-15 15:53:27 | 2017-05-15 16:16:56 | 1494852807 | 1494854216 |
| **701824** | login766 | 195656174 | 1 | 2017-05-15 15:53:33 | 2017-05-15 16:21:53 | 1494852813 | 1494854513 |
| **701825** | login766 | 195656466 | 1 | 2017-05-15 15:53:38 | 2017-05-15 16:23:41 | 1494852818 | 1494854621 |
| **701826** | login766 | 195656336 | 3 | 2017-05-15 15:54:18 | 2017-05-15 16:32:11 | 1494852858 | 1494855131 |
| **701827** | login766 | 195656078 | 2 | 2017-05-15 15:54:59 | 2017-05-15 16:17:32 | 1494852899 | 1494854252 |

In [14]:
```
new_data = data
new_data = new_data.drop(['assigned_ts', 'closed_ts'], axis=1)
```

In [15]:
```
new_data.head()
```

Out[15]:

| | login | tid | Microtasks | assigned_timestamp | closed_timestamp |
|---|---|---|---|---|---|
| **1** | login0 | 190563850 | 4 | 1492679379 | 1492683181 |
| **2** | login0 | 190561754 | 1 | 1492679430 | 1492680509 |
| **3** | login0 | 190565906 | 4 | 1492680091 | 1492684210 |
| **4** | login0 | 190560246 | 1 | 1492680510 | 1492683236 |
| **5** | login0 | 190562168 | 2 | 1492680522 | 1492683290 |

In [16]:
```
new_data.shape
```

Out[16]:
```
(701827, 5)
```

In [17]:
```
new_data['task_completion_time'] = (new_data.closed_timestamp - new_data.assigned_timestamp) / new_data.Microtasks
```

In [18]:
```
new_data
```

Out[18]:

| | login | tid | Microtasks | assigned_timestamp | closed_timestamp | task_completion_time |
|---|---|---|---|---|---|---|
| **1** | login0 | 190563850 | 4 | 1492679379 | 1492683181 | 950.500000 |
| **2** | login0 | 190561754 | 1 | 1492679430 | 1492680509 | 1079.000000 |
| **3** | login0 | 190565906 | 4 | 1492680091 | 1492684210 | 1029.750000 |
| **4** | login0 | 190560246 | 1 | 1492680510 | 1492683236 | 2726.000000 |
| **5** | login0 | 190562168 | 2 | 1492680522 | 1492683290 | 1384.000000 |
| **...** | ... | ... | ... | ... | ... | ... |
| **701823** | login766 | 195656026 | 1 | 1494852807 | 1494854216 | 1409.000000 |
| **701824** | login766 | 195656174 | 1 | 1494852813 | 1494854513 | 1700.000000 |
| **701825** | login766 | 195656466 | 1 | 1494852818 | 1494854621 | 1803.000000 |
| **701826** | login766 | 195656336 | 3 | 1494852858 | 1494855131 | 757.666667 |
| **701827** | login766 | 195656078 | 2 | 1494852899 | 1494854252 | 676.500000 |

701827 rows × 6 columns

In [19]:
```
new_data.task_completion_time = new_data.task_completion_time.astype('int64')
```

In [20]:
```
new_data.sort_values(['task_completion_time'])
```

Out[20]:

| | login | tid | Microtasks | assigned_timestamp | closed_timestamp | task_completion_time |
|---|---|---|---|---|---|---|
| **534454** | login585 | 197340894 | 1 | 1495631586 | 1495630857 | -729 |
| **151273** | login139 | 196400390 | 13 | 1495290002 | 1495290025 | 1 |
| **153222** | login139 | 198396915 | 42 | 1496270061 | 1496270138 | 1 |
| **315282** | login328 | 168095766 | 9 | 1494507962 | 1494507979 | 1 |

| | login | tid | Microtasks | assigned_timestamp | closed_timestamp | task_completion_time |
|---|---|---|---|---|---|---|
| **474138** | login517 | 192276632 | 8 | 1493271629 | 1493271645 | 2 |
| ... | ... | ... | ... | ... | ... | ... |
| **292152** | login300 | 195582430 | 1 | 1494992429 | 1495598224 | 605795 |
| **291835** | login300 | 195502918 | 1 | 1495078543 | 1495696540 | 617997 |
| **292160** | login300 | 195502918 | 1 | 1495078543 | 1495696540 | 617997 |
| **603972** | login657 | 195642338 | 1 | 1494992480 | 1496066636 | 1074156 |
| **603631** | login657 | 195642338 | 1 | 1494992480 | 1496066636 | 1074156 |

701827 rows × 6 columns

Ошибка в данных в 534454 строке. Задание сдал, раньше чем начал. Ошибка в 12 минут 9 секунд

In [21]:
```
data.filter(like='534454', axis=0)
```

Out[21]:

| | login | tid | Microtasks | assigned_ts | closed_ts | assigned_timestamp | closed_timestamp |
|---|---|---|---|---|---|---|---|
| **534454** | login585 | 197340894 | 1 | 2017-05-24 16:13:06 | 2017-05-24 16:00:57 | 1495631586 | 1495630857 |

Среднее количество потраченного времени на одно задание с ошибочной строкой

In [22]:
```
new_data.task_completion_time.mean()
```

Out[22]:
```
567.9705853436816
```

In [23]:
```
new_drop_data = new_data
```

In [24]:
```
new_drop_data = new_drop_data.drop([534454])
```

Среднее количество потраченного времени на одно задание без ошибочной строки

In [25]:
```
new_drop_data.task_completion_time.mean()
```

Out[25]:
```
567.9724333381779
```