01. Cancelation Rates

From the following table of user IDs, actions, and dates, write a query to return the publication and cancellation rate for each user.

```
!pip install polars
import pandas as pd
import numpy as np
import polars as pl

data = {'user_id' :[1,1,2,2,3,3,1,1],
        'action'  :['start','cancel','start',
                    'publish',
                    'start','cancel','start','publish'],
        'dates'   :['01-JAN-20',
                    '02-JAN-20',
                    '03-JAN-20',
                    '04-JAN-20',
                    '05-JAN-20',
                    '06-JAN-20',
                    '07-JAN-20',
                    '08-JAN-20']
        }
pandas_users=pd.DataFrame(data)
polars_users=pl.DataFrame(data)
```

```
Requirement already satisfied: polars in /usr/local/lib/python3.11/dist-packages
```

```
pandas_users['dates']=(pd.to_datetime(pandas_users['dates']
                                      ,format="%d-%b-%y"
                          )
)
print(f'users in Pandas:\n{pandas_users}')
```

```
users in Pandas:
   user_id   action      dates
0        1    start 2020-01-01
1        1   cancel 2020-01-02
2        2    start 2020-01-03
3        2  publish 2020-01-04
4        3    start 2020-01-05
5        3   cancel 2020-01-06
6        1    start 2020-01-07
7        1  publish 2020-01-08
```

```
pandas_df1=pd.get_dummies(pandas_users['action'])
print(f'action executed by each user:\n{pandas_df1}')
```

```
action executed by each user:
   cancel  publish  start
0   False    False   True
```

```
1    True    False  False
2   False    False   True
3   False     True  False
4   False    False   True
5    True    False  False
6   False    False   True
7   False     True  False
```

```
pandas_df2=(pd.get_dummies(pandas_users['action'])
              .groupby(pandas_users['user_id'])
              .sum()
)
print(f'Total actions by each user:\n{pandas_df2}')
```

```
Total actions by each user:
             cancel  publish  start
user_id
1             1        1      2
2             0        1      1
3             1        0      1
```

```
pandas_rates=(pd.get_dummies(pandas_users['action'])
                .groupby(pandas_users['user_id'])
                .sum()
                .assign(publish_rate=lambda x:
                                      x['publish']/x['start'],
                        cancel_rate=lambda x:
                                    x['cancel']/x['start']
                 )
                .replace(np.inf,0)
                .reset_index()
)
print(f'rates for each user using Pandas:')
pandas_rates[['user_id','publish_rate','cancel_rate']]
```

```
rates for each user using Pandas:
```

	user_id	publish_rate	cancel_rate
0	1	0.5	0.5
1	2	1.0	0.0
2	3	0.0	1.0

```
polars_users=(polars_users.with_columns(pl.col('dates')
                                        .str
                                        .strptime(pl.Date,
                                                  format="%d-%b-%y"
                                         )
                           )
)
print(f'users in Polars:\n{polars_users}')
```

```
users in Polars:
shape: (8, 3)
```

user_id	action	dates
i64	str	date
1	start	2020-01-01
1	cancel	2020-01-02
2	start	2020-01-03
2	publish	2020-01-04
3	start	2020-01-05
3	cancel	2020-01-06
1	start	2020-01-07
1	publish	2020-01-08

```
polars_df1=(polars_users.to_dummies(columns='action')
                        .drop('dates')
)
print(f'action executed by each user:\n{polars_df1}')
```

```
action executed by each user:
shape: (8, 4)
```

user_id	action_cancel	action_publish	action_start
i64	u8	u8	u8
1	0	0	1
1	1	0	0
2	0	0	1
2	0	1	0
3	0	0	1
3	1	0	0
1	0	0	1
1	0	1	0

```
polars_df2=(polars_users.to_dummies(columns='action')
                        .drop('dates')
                        .group_by('user_id')
                        .agg(pl.col('*').sum())
)
print(f'Number of actions by each user:\n{polars_df2}')
```

```
Number of actions by each user:
shape: (3, 4)
```

user_id	action_cancel	action_publish	action_start
i64	i64	i64	i64
2	0	1	1
1	1	1	2

```
| 3       | 1             | 0             | 1             |
```

```
polars_rates=(polars_users.to_dummies(columns='action')
                          .drop('dates')
                          .group_by('user_id')
                          .agg(pl.col('*').sum())
                          .select(pl.col('user_id'),
                                  publish_rate=pl.col('action_publish')/pl.col('action_star
                                  cancel_rate=pl.col('action_cancel')/pl.col('action_start'
                           )
)
print(f'Rates for each user using Polars:')
polars_rates
```

```
Rates for each user using Polars:
shape: (3, 3)
```

user_id	publish_rate	cancel_rate
i64	f64	f64
1	0.5	0.5
3	0.0	1.0
2	1.0	0.0