1_Cohort_analysis_retention_rate

December 12, 2022

SQL request project Yandex Practicum (personal visuzalization implementation in Python)

Cohort analysis of retention rate for users who registered in 2019

[1]:

```
import pandas as pd
%load_ext sql
%sql postgresql://postgres:sqltest123@localhost/1
```

[2]:

```
%%sql result <<
WITH
profile AS
  (SELECT u.user_id,
          DATE_TRUNC('month', MIN(event_time))::date AS dt
   FROM tools_shop.users u
   JOIN tools_shop.orders o ON u.user_id = o.user_id
   JOIN tools_shop.events e ON u.user_id = e.user_id
   GROUP BY 1),
sessions AS
  (SELECT p.user_id AS users,
          DATE_TRUNC('month', event_time)::date AS session_dt
   FROM tools_shop.events e
   JOIN profile p ON p.user_id = e.user_id
   GROUP BY 1,
            2),
cohort_users_cnt AS
  (SELECT dt,
          COUNT(user_id) AS cohort_users_cnt
   FROM profile
   GROUP BY 1)

SELECT p.dt AS cohort_dt,
       session_dt,
       COUNT(p.user_id) AS users_cnt,
       cohort_users_cnt,
       ROUND(COUNT(p.user_id) * 100.0 / cohort_users_cnt, 2)::float AS␣
 ↪retention_rate
FROM profile p
JOIN sessions s ON p.user_id = s.users
JOIN cohort_users_cnt AS cuc ON p.dt = cuc.dt
WHERE p.dt >= '2019-01-01'
AND p.dt < '2020-01-01'
GROUP BY 1,
         2,
         4
ORDER BY 1,2
```

```
 * postgresql://postgres:***@localhost/1
133 rows affected.
Returning data to local variable result
```

[3]:

```
df = result.DataFrame()
display(df.head(20))
```

```
     cohort_dt  session_dt  users_cnt  cohort_users_cnt  retention_rate
0   2019-01-01  2019-01-01        306               306          100.00
1   2019-01-01  2019-02-01         62               306           20.26
2   2019-01-01  2019-03-01         63               306           20.59
3   2019-01-01  2019-04-01         42               306           13.73
4   2019-01-01  2019-05-01         40               306           13.07
5   2019-01-01  2019-06-01         29               306            9.48
6   2019-01-01  2019-07-01         12               306            3.92
7   2019-01-01  2019-08-01          3               306            0.98
8   2019-01-01  2019-12-01          1               306            0.33
9   2019-01-01  2020-02-01          1               306            0.33
10  2019-01-01  2020-08-01          1               306            0.33
11  2019-01-01  2021-02-01          1               306            0.33
12  2019-02-01  2019-02-01        296               296          100.00
13  2019-02-01  2019-03-01         75               296           25.34
14  2019-02-01  2019-04-01         42               296           14.19
15  2019-02-01  2019-05-01         34               296           11.49
16  2019-02-01  2019-06-01         37               296           12.50
17  2019-02-01  2019-07-01         32               296           10.81
18  2019-02-01  2019-08-01         11               296            3.72
19  2019-02-01  2019-09-01          2               296            0.68
```

[4]:

```
cohort_start = list(df['cohort_dt'])
cohort_session = list(df['session_dt'])
retention_rate = list(df['retention_rate'])

ret_r = list(zip(cohort_start, cohort_session, retention_rate))
df2 = pd.DataFrame(ret_r, columns = ['cohort_start', 'cohort_session',␣
 ↪'retention_rate'])
df2
```

[4]:

```
     cohort_start cohort_session  retention_rate
0      2019-01-01     2019-01-01          100.00
1      2019-01-01     2019-02-01           20.26
2      2019-01-01     2019-03-01           20.59
3      2019-01-01     2019-04-01           13.73
4      2019-01-01     2019-05-01           13.07
..            …              …              …
128    2019-12-01     2020-07-01            0.53
129    2019-12-01     2020-08-01            0.27
130    2019-12-01     2020-10-01            0.27
131    2019-12-01     2020-12-01            0.27
132    2019-12-01     2021-01-01            0.27

[133 rows x 3 columns]
```

[1]:

```python
def cohort_period(df2):
    # changing cohort_sessions date type into periods
    df2['cohort_session'] = np.arange(len(df2)) + 0
    return df2
```

[6]:

```python
import numpy as np
cohorts = df2.groupby('cohort_start').apply(cohort_period)
cohorts.head(20)
```

[6]:

```
   cohort_start  cohort_session  retention_rate
0    2019-01-01               0          100.00
1    2019-01-01               1           20.26
2    2019-01-01               2           20.59
3    2019-01-01               3           13.73
4    2019-01-01               4           13.07
5    2019-01-01               5            9.48
6    2019-01-01               6            3.92
7    2019-01-01               7            0.98
8    2019-01-01               8            0.33
9    2019-01-01               9            0.33
10   2019-01-01              10            0.33
11   2019-01-01              11            0.33
12   2019-02-01               0          100.00
13   2019-02-01               1           25.34
14   2019-02-01               2           14.19
15   2019-02-01               3           11.49
16   2019-02-01               4           12.50
17   2019-02-01               5           10.81
18   2019-02-01               6            3.72
19   2019-02-01               7            0.68
```

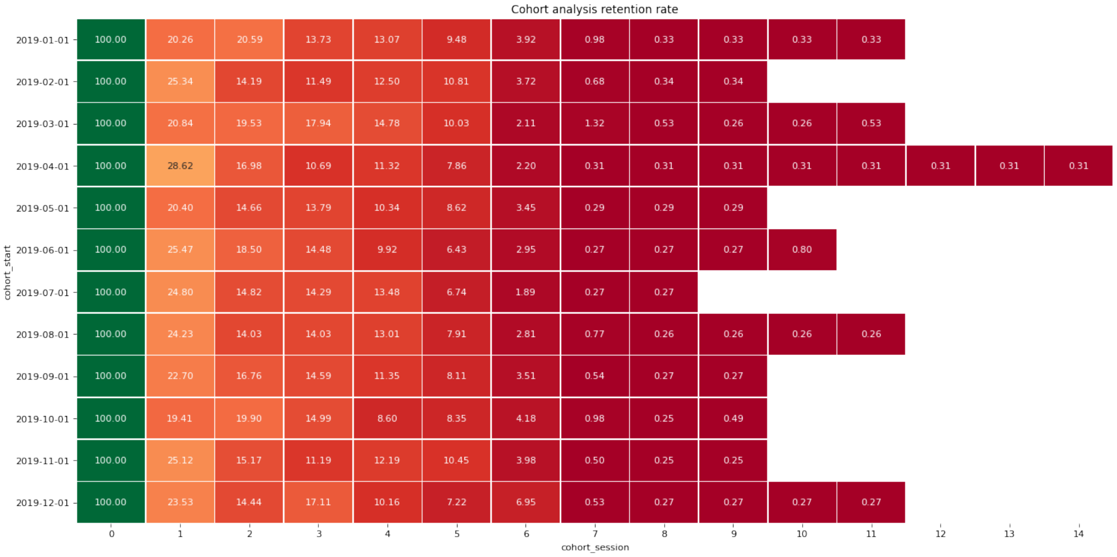
[7]:

```
import seaborn as sb
import matplotlib.pyplot as plt

df_heatmap = cohorts.pivot('cohort_start', 'cohort_session', 'retention_rate')
plt.figure(figsize=(20,10), dpi=80)
sb.heatmap(df_heatmap, annot=True, cmap='RdYlGn', fmt=".2f", linewidth=.5,␣
 ↪cbar=False).set(title='Cohort analysis retention rate')
```

[7]:

```
[Text(0.5, 1.0, 'Cohort analysis retention rate')]
```


Cohort analysis retention rate
cohort_start
cohort_session