

SQL Request from Yandex Practicum course project.

Implementation in Python with usage of visualization.

Cohort analysis of churn rate for users who registered in 2019

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

```
In [ ]: %%sql result <<
        WITH
        prof AS ( SELECT DISTINCT tso.user_id,
                                DATE_TRUNC('month', MIN(tse.event_time))::date AS cohort_start
                    FROM tools_shop.orders AS tso
                    LEFT JOIN tools_shop.events AS tse ON tso.user_id = tse.user_id
                    GROUP BY 1),
        sess AS ( SELECT cohort_start,
                                DATE_TRUNC('month', tse.event_time)::date AS session_dt,
                                COUNT(DISTINCT tse.user_id) AS c_users
                    FROM prof
                    LEFT JOIN tools_shop.events AS tse ON prof.user_id = tse.user_id
                    GROUP BY 1,2),

        fin AS ( SELECT cohort_start::varchar,
                                session_dt::varchar,
                                c_users,
                                LAG(c_users) OVER(PARTITION BY cohort_start ORDER BY session_dt) AS
                                FROM sess )

        SELECT *,
        ROUND((1 - c_users::numeric/p_users::numeric)*100*-1 , 2)::numeric AS churn_
        FROM fin
        WHERE cohort_start >= '2019-01-01'
        AND cohort_start < '2020-01-01'
        ORDER BY 1,2
```

```
In [3]: #displaying results for SQL request
        df = result.DataFrame()
        display(df.head(6))
```

	cohort_start	session_dt	c_users	p_users	churn_rate
0	2019-01-01	2019-01-01	306	NaN	None
1	2019-01-01	2019-02-01	62	306.0	-79.74
2	2019-01-01	2019-03-01	63	62.0	1.61
3	2019-01-01	2019-04-01	42	63.0	-33.33
4	2019-01-01	2019-05-01	40	42.0	-4.76
5	2019-01-01	2019-06-01	29	40.0	-27.50

```
In [4]: #preparing table with required information for visualization
        cohort_group = df['cohort_start']
        cohort_month = df['session_dt']
```

```

cohort_users = df['c_users']
churn_rate = df['churn_rate'].fillna(0).astype('float')

churn_list = list(zip(cohort_group, cohort_month, cohort_users, churn_rate))
df2 = pd.DataFrame(churn_list, columns = ['cohort_group',
                                         'cohort_month',
                                         'cohort_users',
                                         'churn_rate'])

display(df2.head(6))

```

	cohort_group	cohort_month	cohort_users	churn_rate
0	2019-01-01	2019-01-01	306	0.00
1	2019-01-01	2019-02-01	62	-79.74
2	2019-01-01	2019-03-01	63	1.61
3	2019-01-01	2019-04-01	42	-33.33
4	2019-01-01	2019-05-01	40	-4.76
5	2019-01-01	2019-06-01	29	-27.50

In [5]:

```

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

import numpy as np
cohorts_df = df2.groupby('cohort_group').apply(cohort_period)
cohorts_df.head(6)

```

Out[5]:

	cohort_group	cohort_month	cohort_users	churn_rate
0	2019-01-01	0	306	0.00
1	2019-01-01	1	62	-79.74
2	2019-01-01	2	63	1.61
3	2019-01-01	3	42	-33.33
4	2019-01-01	4	40	-4.76
5	2019-01-01	5	29	-27.50

In [6]:

```

#first visualization is aimed at number of active users among cohorts
import seaborn as sb
import matplotlib.pyplot as plt
cohort_month = cohorts_df['cohort_month']
churn_r1 = list(zip(cohort_group, cohort_month, cohort_users))
df3 = pd.DataFrame(churn_r1, columns = ['cohort_group',
                                       'cohort_month',
                                       'cohort_users'])

df_heatmap = df3.pivot('cohort_group',
                        'cohort_month',
                        'cohort_users')

plt.figure(figsize=(20,10), dpi=80)
sb.heatmap(df_heatmap,
           annot=True,
           robust=True,
           cmap='Blues',
           fmt=' ',

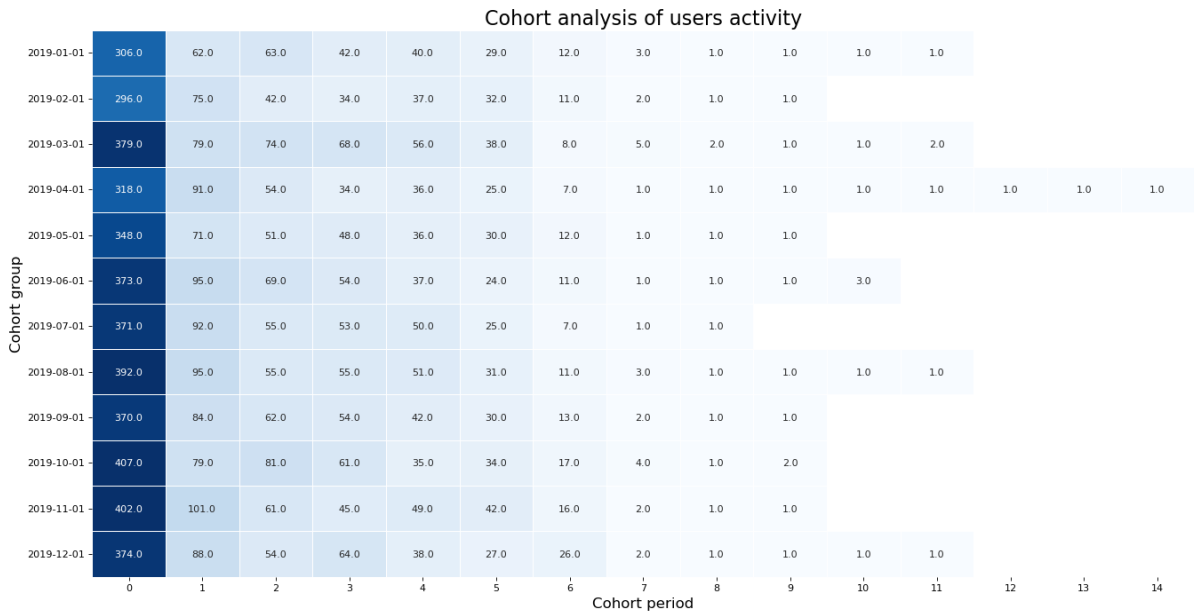
```

```

        linewidth=.5,
        cbar=False)
plt.ylabel('Cohort group', size=15)
plt.xlabel('Cohort period', size=15)
plt.title('Cohort analysis of users activity', size=20)
print('Heatmap of users activity among cohorts')

```

Heatmap of users activity among cohorts



```

In [7]: #second visualization is aimed at churn rate
churn_r2 = list(zip(cohort_group, cohort_month, churn_rate))
df4 = pd.DataFrame(churn_r2, columns = ['cohort_group', 'cohort_month', 'churn_rate'])
df_heatmap = df4.pivot('cohort_group', 'cohort_month', 'churn_rate')
plt.figure(figsize=(20,10), dpi=80)
sb.heatmap(df_heatmap,
            annot=True,
            robust=True,
            cmap='magma',
            fmt=".2f",
            linewidth=.5,
            cbar=False)
plt.ylabel('Cohort group', size=15)
plt.xlabel('Cohort period', size=15)
plt.title('Cohort analysis of churn rate', size=20)
print('Churn rate is the rate at which customers stop doing business with a

```

Churn rate is the rate at which customers stop doing business with a company over a given period of time

