РК №1 Зорин А.А. ИУ5-23М

In [61]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

In [62]:

```
data = pd.read_csv('../input/bay-area-bike-sharing-trips/2019 - 01.csv')
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 192082 entries, 0 to 192081
Data columns (total 10 columns):
 #   Column              Non-Null Count   Dtype  
---  ------              --------------   -----  
 0   month               192082 non-null  object 
 1   trip_duration_sec   192082 non-null  int64  
 2   start_station_id    191834 non-null  float64
 3   start_station_name  191834 non-null  object 
 4   end_station_id      191834 non-null  float64
 5   end_station_name    191834 non-null  object 
 6   bike_id             192082 non-null  int64  
 7   user_type           192082 non-null  object 
 8   member_birth_year   182362 non-null  float64
 9   member_gender       182365 non-null  object 
dtypes: float64(3), int64(2), object(5)
memory usage: 14.7+ MB
```

Задача №6

Для набора данных проведите устранение пропусков для одного (произвольного) числового признака с использованием метода заполнения средним значением.

In [63]:

```
data.isnull().sum()
```

Out[63]:

```
month                    0
trip_duration_sec        0
start_station_id       248
start_station_name     248
end_station_id         248
end_station_name       248
bike_id                  0
user_type                0
member_birth_year     9720
member_gender         9717
dtype: int64
```

Заменим пропущенные значения в `member_birth_year` средним

In [64]:

```
data.member_birth_year.fillna(data.member_birth_year.mean(), inplace=True)
data.member_birth_year.isnull().sum()
```

Out[64]:

```
0
```

Задача №26

Для набора данных для одного (произвольного) числового признака проведите обнаружение и замену (найденными верхними и нижними границами) выбросов на основе правила трех сигм.

Вычислим вехнюю и нижнюю границы для `member_birth_year`

In [65]:

```
std = data.member_birth_year.std()
mean = data.member_birth_year.mean()
lower_bound = mean - 3 * std
upper_bound = mean + 3 * std

print('Нижняя граница:', lower_bound)
print('Верхняя граница:', upper_bound)
```

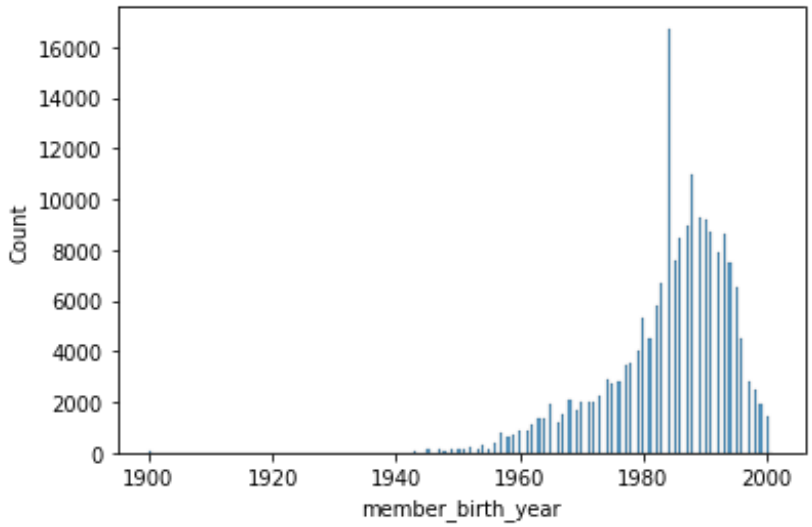
```
Нижняя граница: 1954.6394334176207
Верхняя граница: 2013.6585233940932
```

In [66]:

```
sns.histplot(data.member_birth_year)
```

Out[66]:

```
<AxesSubplot:xlabel='member_birth_year', ylabel='Count'>
```



Произведем замену выбросов на основе правила трех сигм

In [67]:

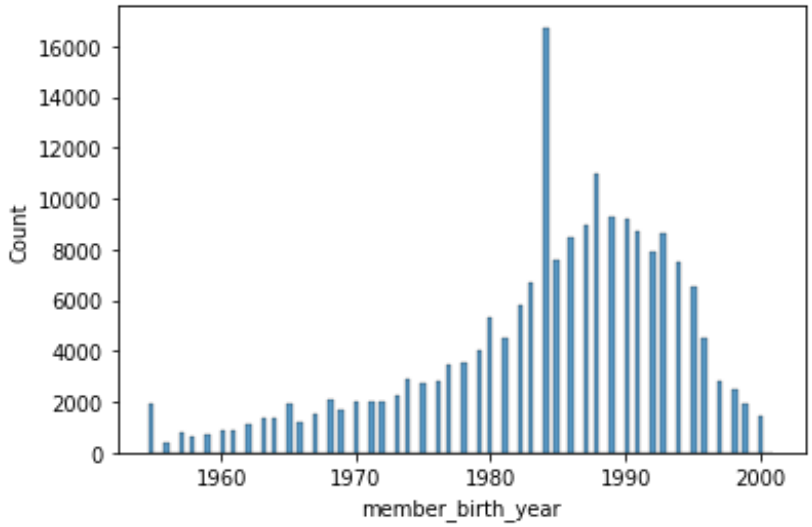
```
data.member_birth_year.mask(data.member_birth_year < lower_bound, lower_bound, i
nplace=True)
data.member_birth_year.mask(data.member_birth_year > upper_bound, upper_bound, i
```

```
nplace=True)

sns.histplot(data.member_birth_year)
```

Out[67]:

```
<AxesSubplot:xlabel='member_birth_year', ylabel='Count'>
```



Дополнительное задание

Для произвольной колонки данных построить график **boxplot.**

In [68]:

```
sns.boxplot(x=data.user_type, y=data.member_birth_year)
```

Out[68]:

```
<AxesSubplot:xlabel='user_type', ylabel='member_birth_year'>
```