In [2]:

```
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
import plotly.express as px
import scipy.stats as sps
import seaborn as sns


sns.set(font_scale=1.5)
%matplotlib inline
```

executed in 5.96s, finished 13:19:40 2020-05-22

Формулировка задачи на простом ”пользовательском” языке:

> Мы займемся анализом поездок пользователей в зависимости от типа абонемента/ подписки. Как используют велопрокат пользователи типа `Member` и `Short-Term Pass Holder` .\

Формулировка задачи на математическом языке:

> Выяснить, как тип пользователя влияет на продолжительность поездки. Узнать, есть ли различие между средним значением продолжительности поездки для подписки типа `Member` и `Short-Term Pass Holder` .

Загрузим чуть-чуть обработанные данные, где удалены дубликаты первых 50000 поездок в таблице trips.csv.

In [3]:

```
DATA_DIR = '../data/processed/'
```

executed in 9ms, finished 13:19:40 2020-05-22

In [4]:

```
stations_data = pd.read_csv(DATA_DIR+"station.csv")
trips = pd.read_csv(DATA_DIR+"trips.csv", error_bad_lines=False, index_col=0)
weather = pd.read_csv(DATA_DIR+"weather.csv")
```

executed in 2.42s, finished 13:19:43 2020-05-22

In [5]:

```
1 trips.head(3)
```

executed in 101ms, finished 13:19:43 2020-05-22

Out[5]:

	trip_id	starttime	stoptime	bikeid	tripduration	from_station_name	to_statior
0	431	10/13/2014 10:31	10/13/2014 10:48	SEA00298	985.935	2nd Ave & Spring St	Occide Occident S
1	432	10/13/2014 10:32	10/13/2014 10:48	SEA00195	926.375	2nd Ave & Spring St	Occide Occident S
2	433	10/13/2014 10:33	10/13/2014 10:48	SEA00486	883.831	2nd Ave & Spring St	Occide Occident S

Возможны два типа подписки.

In [6]:

```
1 trips.usertype.unique()
```

executed in 77ms, finished 13:19:44 2020-05-22

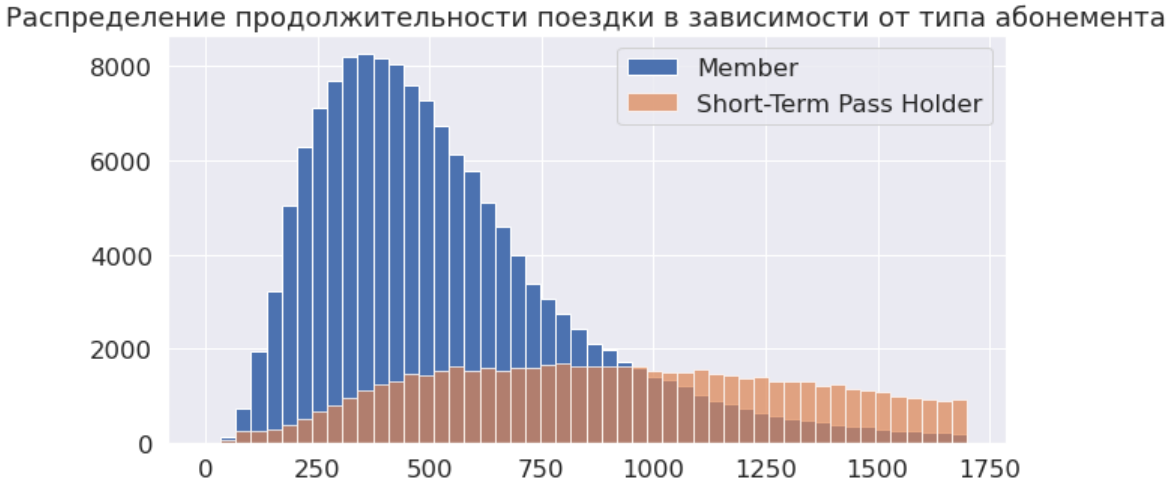
Out[6]:

```
array(['Member', 'Short-Term Pass Holder'], dtype=object)
```

In [7]:

```
plt.figure(figsize=(10, 5))
plt.title('Распределение продолжительности поездки в зависимости от типа абоне
plt.hist(trips[trips['usertype'] == 'Member'].tripduration,
         bins=50, range=(0, 1700),
         label='Member')
plt.hist(trips[trips['usertype'] == 'Short-Term Pass Holder'].tripduration,
         bins=50, range=(0, 1700),
         label='Short-Term Pass Holder', alpha=0.7)
plt.legend()
plt.show()
```

executed in 2.39s, finished 13:19:46 2020-05-22



Псмотрим на статистики для `tripdaration` для разных типов

In [8]:

```
trips.drop(['trip_id', 'birthyear'], axis=1).groupby(by=['usertype']).describe
```

executed in 486ms, finished 13:19:47 2020-05-22

Out[8]:

	tripduration							
	count	mean	std	min	25%	50%	75%	max
usertype								
Member	146171.0	592.977313	731.550556	60.008	324.2025	479.1490	693.39150	27985.8
Short-Term Pass Holder	89894.0	2193.901312	2959.113053	60.111	760.0590	1251.4325	2161.54875	28794.3

Можем проверить критерием Уилкоксона-Манна-Уитни гипотезу об однородности.

In [9]:

```
sps.mannwhitneyu(trips[trips['usertype'] == 'Member'].tripduration,
                 trips[trips['usertype'] == 'Short-Term Pass Holder'].tripdura
```

executed in 437ms, finished 13:19:48 2020-05-22

Out[9]:

```
MannwhitneyuResult(statistic=2077197774.5, pvalue=0.0)
```

Теперь посмотрим, когда разные типы пользователей пользуются велопрокатом. Построим гистограммы времени начала поездки.

In [11]:

```
def is_weekend(day):↔
```

executed in 12ms, finished 13:21:32 2020-05-22

In [15]:

```
trips['starttime'] = pd.to_datetime(trips['starttime'])
trips['stoptime'] = pd.to_datetime(trips['stoptime'])
trips['starttime_of_day_hour'] = [hour for hour in trips['starttime'].dt.hour]
trips['weekend'] = is_weekend(trips.starttime.dt.weekday)
```

executed in 287ms, finished 13:22:13 2020-05-22

In [13]:

```
plt.figure(figsize=(10, 5))
plt.title('Распределение времени начала поездки в зависимости от типа абонемен
plt.hist(trips[trips['usertype'] == 'Member'].starttime_of_day_hour,
         bins=50,
         label='Member')
plt.hist(trips[trips['usertype'] == 'Short-Term Pass Holder'].starttime_of_day
         bins=50,
         label='Short-Term Pass Holder', alpha=0.5, color='yellow')
plt.legend()
plt.show()
```

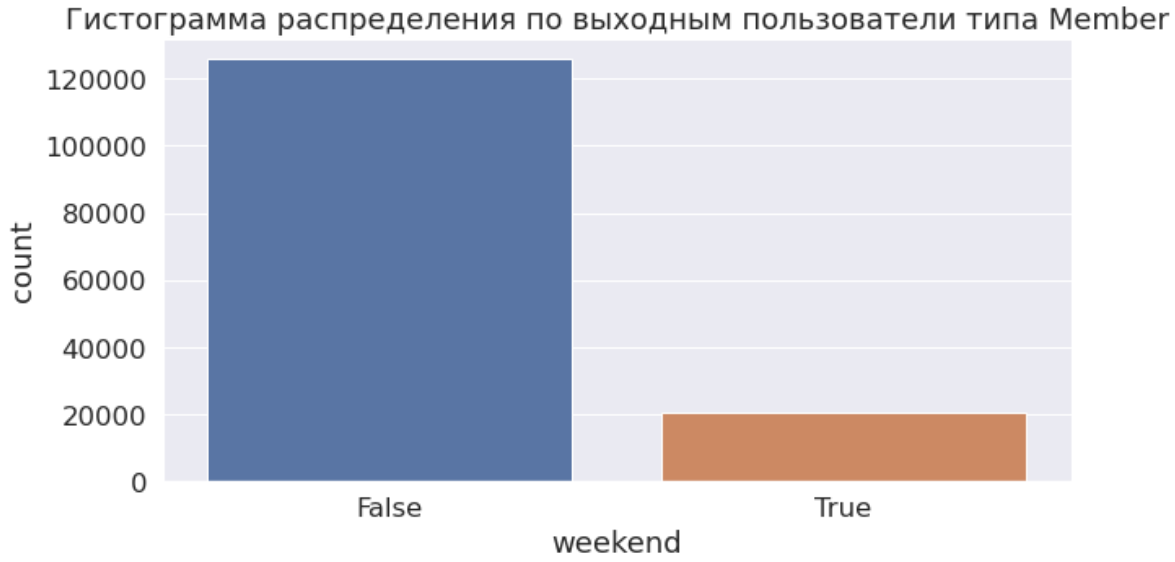
executed in 889ms, finished 13:21:39 2020-05-22

Посмотрим, когда чаще всего катаются `Member` , `Short-Term Pass Holder` .

In [18]:

```
plt.figure(figsize=(10, 5))
plt.title('Гистограмма распределения поездок по выходным пользователи типа Mem
sns.countplot(x='weekend', data=trips[trips['usertype'] == 'Member'])
plt.show()
```

executed in 273ms, finished 13:23:14 2020-05-22

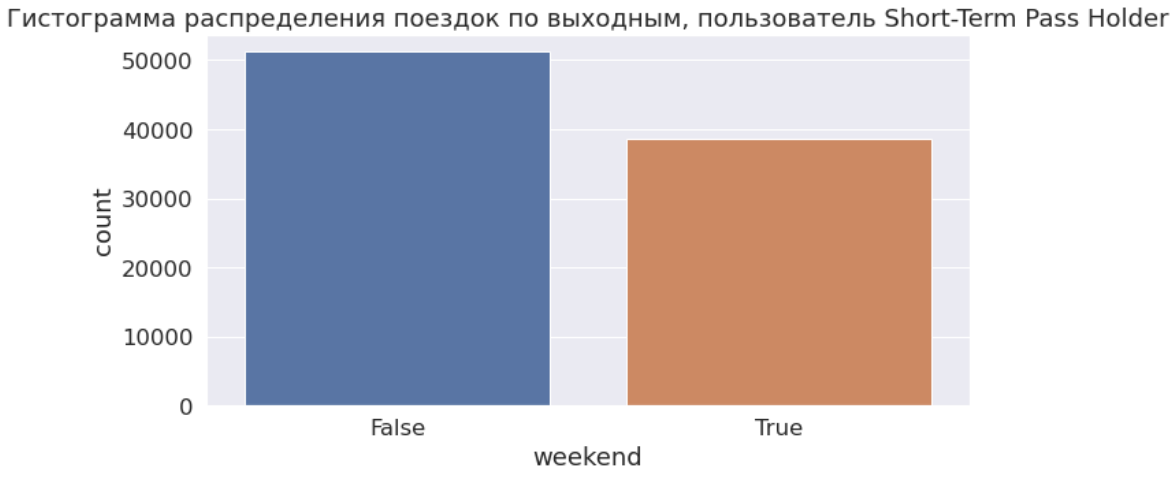


Как видим, `Member` чаще катаются в будний день.

In [20]:

```
plt.figure(figsize=(10, 5))
plt.title('Гистограмма распределения поездок по выходным, пользователь Short-T
sns.countplot(x='weekend', data=trips[trips['usertype'] != 'Member'])
plt.show()
```

executed in 260ms, finished 13:24:25 2020-05-22



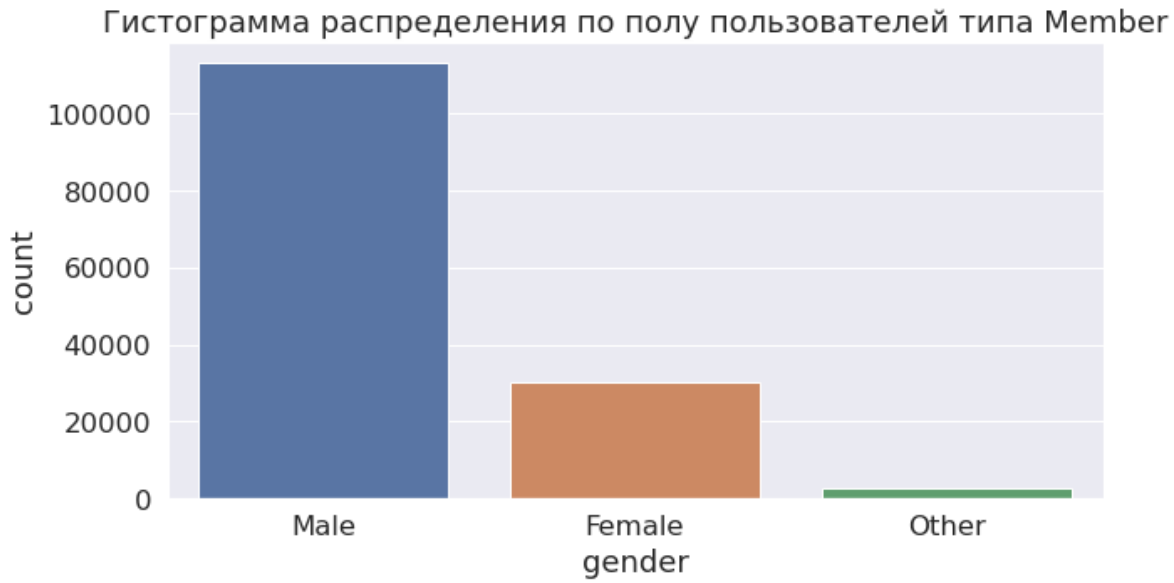
А вот `Short-Term Pass Holder` -пользователи велопрокатата на выходных пользуютя уже чаще.

Посмотрим на распределение пользователей типа `Member` в зависимости от пола.

In [46]:

```
1 plt.figure(figsize=(10, 5))
2 plt.title('Гистограмма распределения по полу пользователей типа Member')
3 sns.countplot(x='gender', data=trips[trips['usertype'] == 'Member'])
4 plt.show()
```

executed in 988ms, finished 23:00:42 2020-04-26



При этом важно заметить, что пол пользователя для `Short-Term Pass Holder` подписки не определен, поэтому исследования на эту темы не возможны.

In [14]:

```
1 trips[trips.usertype != 'Member'].gender
```

executed in 91ms, finished 00:16:33 2020-05-21

Out[14]:

```
69        NaN
78        NaN
89        NaN
91        NaN
92        NaN
         ...
286852    NaN
286853    NaN
286854    NaN
286855    NaN
286856    NaN
Name: gender, Length: 89894, dtype: object
```

Выводы:

- Как видим, разные пользователи по-разному используют велопрокат. Так, пользователи типа `Member` катаются чаще и преимущественно небольшое время, при этом большая часть поездок приходится в начало (8 утра) и коец рабочего дня(17-18 вечера), а вот для `Short-Term Pass Holder` более характерны продолжительные поездки, о чем говорит и значение средних величин,

при чем совершаются они преимущественно днем, в период с 10 до 19 часов. Пользователя типа `Member` также очень редко катаются в выходные, чего нельзя сказать о `Short-Term Pass Holder`.