In [1]:

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


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

executed in 4.31s, finished 13:17:05 2020-05-22

Формулировка на простом языке:

> Познакомиться с данными. Визуализировать станции велопроката на карте городе Сиэтл.

Формулировка на математическом языке:

> Загрузить и изучить представленные данные. Построить гистограммы по информациям о поездках(продолжительность, пол пользователя, его тип и т.д.), нанести станции на карту города.

In [2]:

```
DATA_DIR = "../data/raw/"
```

executed in 5ms, finished 13:17:05 2020-05-22

Рассмотрим первую таблицу `station.csv` - информацию о станциях проката.

Station dataset

- `station_id` : station ID number
- `name` : name of station
- `lat` : station latitude
- `long` : station longitude
- `install_date` : date that station was placed in service
- `install_dockcount` : number of docks at each station on the installation date
- `modification_date` : date that station was modified, resulting in a change in location or dock count
- `current_dockcount` : number of docks at each station on 8/31/2016
- `decommission_date` : date that station was placed out of service

In [3]:

```
stations_data = pd.read_csv(DATA_DIR+"station.csv")
```

executed in 139ms, finished 13:17:05 2020-05-22

In [4]:

```
1 stations_data.shape
```

executed in 154ms, finished 13:17:06 2020-05-22

Out[4]:

```
(58, 9)
```

In [5]:

```
1 stations_data.head(10)
```

executed in 200ms, finished 13:17:06 2020-05-22

Out[5]:

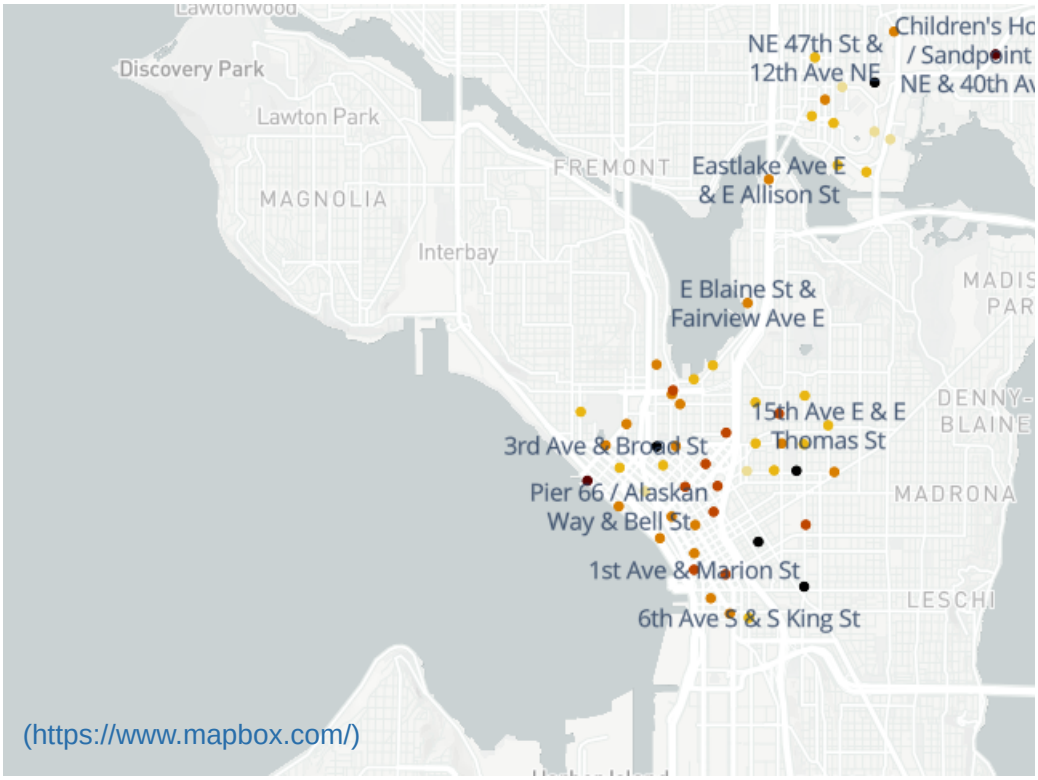
	station_id	name	lat	long	install_date	install_dockcount	modificat
0	BT-01	3rd Ave & Broad St	47.618418	-122.350964	10/13/2014	18	
1	BT-03	2nd Ave & Vine St	47.615829	-122.348564	10/13/2014	16	
2	BT-04	6th Ave & Blanchard St	47.616094	-122.341102	10/13/2014	16	
3	BT-05	2nd Ave & Blanchard St	47.613110	-122.344208	10/13/2014	14	
4	CBD-03	7th Ave & Union St	47.610731	-122.332447	10/13/2014	20	
5	CBD-04	Union St & 4th Ave	47.609221	-122.335596	7/27/2015	18	
6	CBD-05	1st Ave & Marion St	47.604058	-122.335800	10/13/2014	20	
7	CBD-06	2nd Ave & Spring St	47.605950	-122.335768	10/13/2014	20	
8	CBD-07	City Hall / 4th Ave & James St	47.603509	-122.330409	10/13/2014	20	
9	CBD-13	2nd Ave & Pine St	47.610185	-122.339641	10/13/2014	18	

Выполним простую визуализацию станций на карте Сиэтла.

In [6]:

```
px.set_mapbox_access_token(open("../mapbox_token").read())
fig = px.scatter_mapbox(
    stations_data,
    lat="lat",
    lon="long",
    color="current_dockcount",
    text="name",
    color_continuous_scale=px.colors.cyclical.IceFire,
    size_max=15,
    zoom=11
)
fig.show()
```

executed in 11.4s, finished 13:17:17 2020-05-22



Сиэтл - достаточно большой город, и станции велопроката покрывают лишь несколько районов.

Посмотрим на статистики, связанные с количеством доков.

In [7]:

```
1 stations_data[["install_dockcount", "current_dockcount"]].describe()
```

executed in 86ms, finished 13:17:17 2020-05-22

Out[7]:

	install_dockcount	current_dockcount
count	58.000000	58.000000
mean	17.586207	16.517241
std	3.060985	5.117021
min	12.000000	0.000000
25%	16.000000	16.000000
50%	18.000000	18.000000
75%	18.000000	18.000000
max	30.000000	26.000000

Как видим, произошли некоторые изменения, по сравнению с первоначальным состоянием станций. Так, например, на 8/31/2016 есть станции, которые прекратили работу. Среднее количество доков для велосипедов снизилось на один.

Посмотрим на вышедшие из строя станции

In [8]:

```
1 stations_data[stations_data.current_dockcount == 0]
```

executed in 226ms, finished 13:17:18 2020-05-22

Out[8]:

	station_id	name	lat	long	install_date	install_dockcount	modificati
10	CD-01	12th Ave & E Yesler Way	47.602103	-122.316923	5/22/2015	16	
26	FH-01	Frye Art Museum / Terry Ave & Columbia St	47.607281	-122.324783	10/13/2014	16	
38	SLU-18	Dexter Ave & Denny Way	47.618285	-122.342205	10/13/2014	20	
46	UW-01	UW McCarty Hall / Whitman Ct	47.660268	-122.304826	10/13/2014	16	

Рассмотрим таблицу `trip.csv` , которая содерит информацию о поездках.

In [9]:

```
1 trips = pd.read_csv(DATA_DIR+"trip.csv", error_bad_lines=False)
```

executed in 4.03s, finished 13:17:22 2020-05-22

```
b'Skipping line 50794: expected 12 fields, saw 20\n'
```

In [10]:

```
1 trips.shape
```

executed in 20ms, finished 13:17:22 2020-05-22

Out[10]:

```
(286857, 12)
```

In [11]:

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

executed in 99ms, finished 13:17:22 2020-05-22

Out[11]:

	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

В Discussion на Kaggle обратили внимание, что `trip.csv` содержит дубликаты, а именно около 50000 первых поездок снова повторяются в таблице. Проверим это.

In [12]:

```
1 len(trips.trip_id)
```

executed in 81ms, finished 13:17:22 2020-05-22

Out[12]:

```
286857
```

In [13]:

```
1 len(np.unique(trips.trip_id))
```

executed in 121ms, finished 13:17:22 2020-05-22

Out[13]:

```
236065
```

Исправим это, чтобы не работать с повторами.

In [14]:

```
1 trips = trips.drop_duplicates(subset="trip_id")
```

executed in 197ms, finished 13:17:22 2020-05-22

Посмотрим информацию о продолжительности поездок.

In [15]:

```
1 trips["tripduration"].describe()
```

executed in 126ms, finished 13:17:22 2020-05-22

Out[15]:

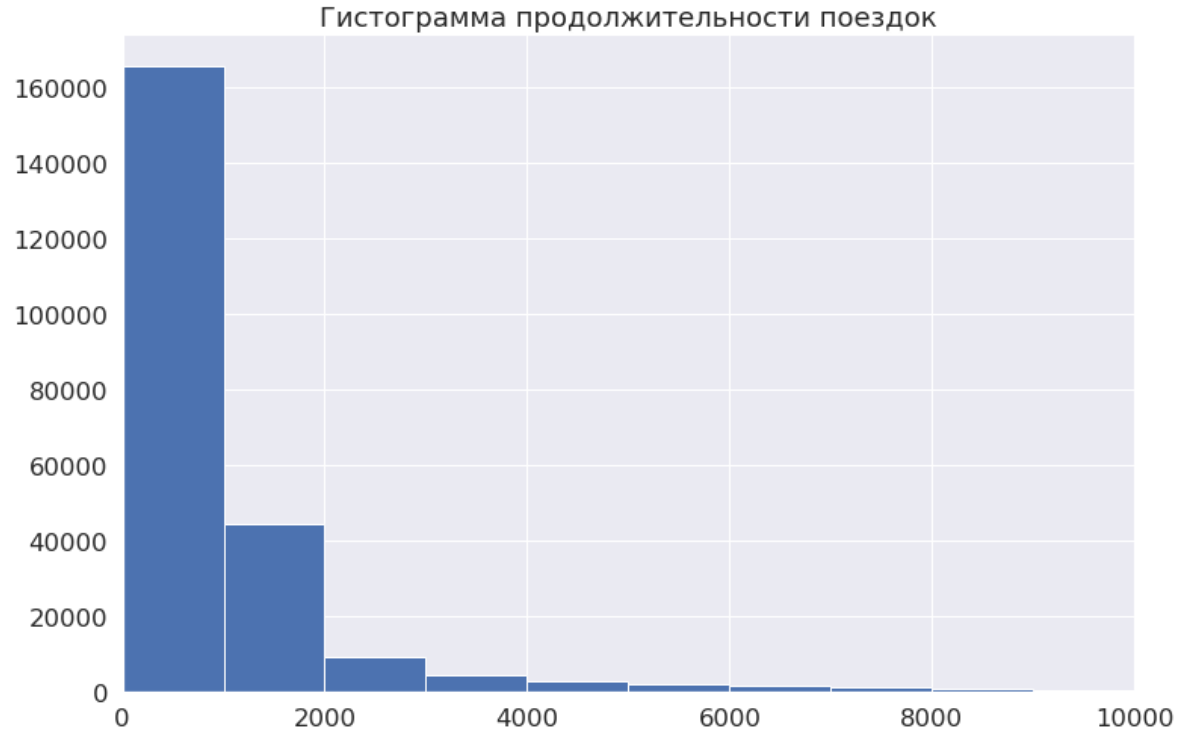
```
count    236065.00000
mean       1202.61221
std        2066.42488
min          60.00800
25%         392.26500
50%         633.23500
75%        1145.01500
max       28794.39800
Name: tripduration, dtype: float64
```

Есть значительная разница между медианой и средним значением `tripduration` . Посмотрим на гистограмму.

In [16]:

```
1 plt.figure(figsize=(12, 8))
2 plt.title('Гистограмма продолжительности поездок')
3 plt.hist(trips["tripduration"], range=(0, 10000))
4
5 plt.xlim((0, 10000))
6 plt.show()
```

executed in 1.62s, finished 13:17:24 2020-05-22

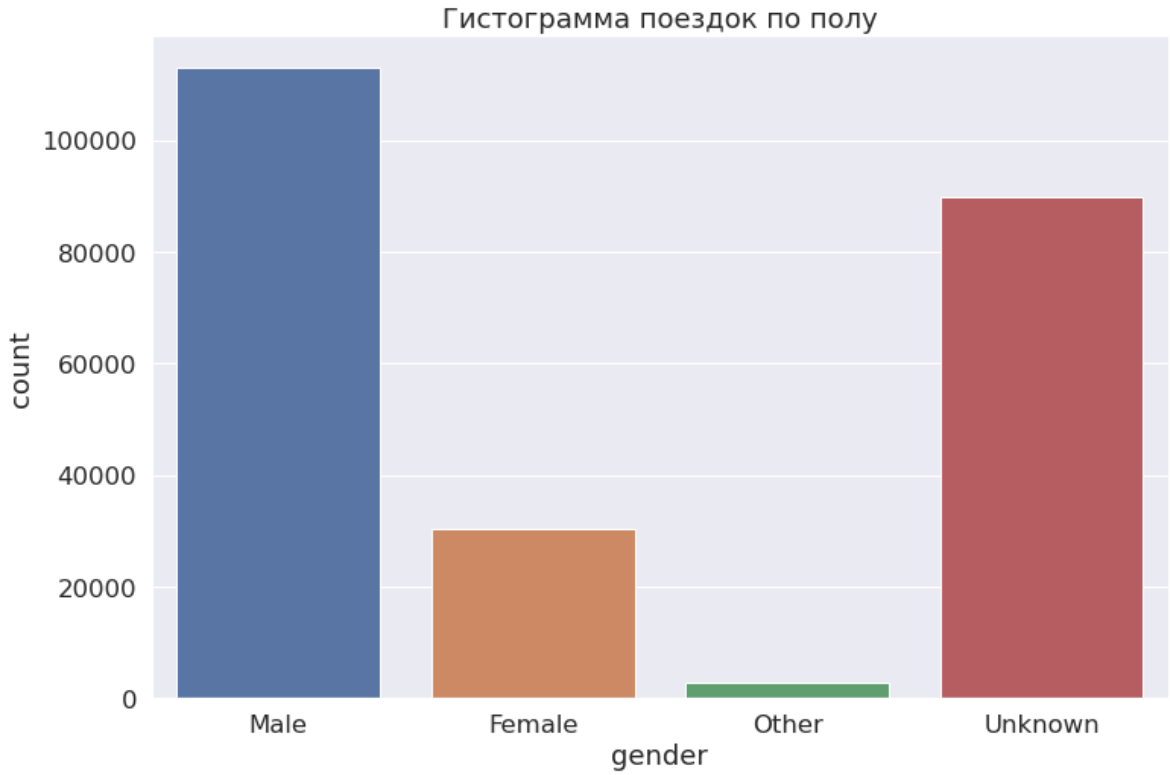


Как видим, большинство поездок длятся менее 2000 секунд, то есть менее получаса.

In [17]:

```
trips.gender = trips.gender.fillna("Unknown")
plt.figure(figsize=(12, 8))
plt.title("Гистограмма поездок по полу")
sns.countplot(x="gender", data=trips)
plt.show()
```

executed in 988ms, finished 13:17:25 2020-05-22

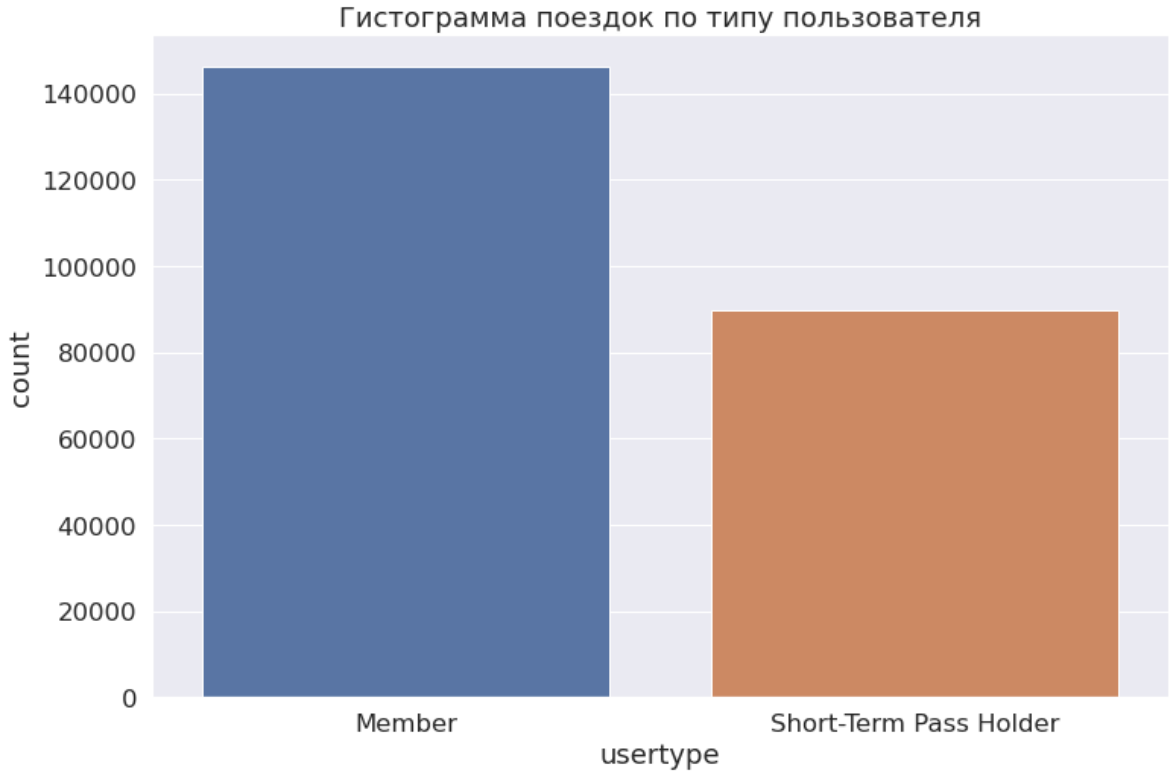


Как видим, среди тех, чей пол известен, в большей степени пользуются велопрокатом мужчины.

In [18]:

```
1 trips.gender = trips.gender.fillna("Unknown")
2 plt.figure(figsize=(12, 8))
3 plt.title("Гистограмма поездок по типу пользователя")
4 sns.countplot(x="usertype", data=trips)
5 plt.show()
```

executed in 898ms, finished 13:17:26 2020-05-22



Пользователей, имеющих тип `Member` оказывается больше(примерно в 1.5-1.7 раза), чем тех, кто имеет статус `Short-Term Pass Holder` .

Рассмотрим таблицу `weather.csv` , которая содержит информацию о погоде.

In [19]:

```
1 weather = pd.read_csv(DATA_DIR+"weather.csv")
2 weather.head(3)
```

executed in 179ms, finished 13:17:26 2020-05-22

Out[19]:

	Date	Max_Temperature_F	Mean_Temperature_F	Min_TemperatureF	Max_Dew_Pc
0	10/13/2014	71	62.0	54	
1	10/14/2014	63	59.0	55	
2	10/15/2014	62	58.0	54	

3 rows × 21 columns

In [20]:

```
1 weather.columns
```

executed in 20ms, finished 13:17:26 2020-05-22

Out[20]:

```
Index(['Date', 'Max_Temperature_F', 'Mean_Temperature_F', 'Min_Tempera
tureF',
       'Max_Dew_Point_F', 'MeanDew_Point_F', 'Min_Dewpoint_F', 'Max_Hu
midity',
       'Mean_Humidity', 'Min_Humidity', 'Max_Sea_Level_Pressure_In',
       'Mean_Sea_Level_Pressure_In', 'Min_Sea_Level_Pressure_In',
       'Max_Visibility_Miles', 'Mean_Visibility_Miles', 'Min_Visibilit
y_Miles',
       'Max_Wind_Speed_MPH', 'Mean_Wind_Speed_MPH', 'Max_Gust_Speed_MP
H',
       'Precipitation_In', 'Events'],
      dtype='object')
```

Ничего особенного здесь не представлено, достаточно много данных о погоде за каждый день ,включая сред. температуру, влажность, скорость ветра и т.д.

Выводы:

- Станции велопроката оказались расположены лишь в нескольких районах города Сиэтл, при этом некоторые из них уже выведены из работы.
- Распределение пользователей велопроката в зависимости от пола/статуса подписки неравномерное.
- Большинство людей пользуется велосипедом менее получаса, то есть совершаются преимущественно короткие поездки.

Таким образом, мы немного познакомились с данными и провели небольшой анализ. Далее, хотелось бы остановится на следующем:

- Определение самых популярных станций, их визуализация на карте города(может быть раскраска в зависимости от популярности). Определение самых популярных маршрутов, визуализация их на

карте города.

- Так же хочется узнать о влиянии погодных условий на количество поездок(как влияет на количество поездок дождь/средняя температура и т.д., есть ли зависимость?)