In [1]:

```
import pandas as pd
```

In [2]:

```
data = pd.read_csv("./yandex_music_project.csv")
data.head()
```

Out[2]:

	userID	Track	artist	genre	City	time	Day
0	FFB692EC	Kamigata To Boots	The Mass Missile	rock	Saint-Petersburg	20:28:33	Wednesday
1	55204538	Delayed Because of Accident	Andreas Rönnberg	rock	Moscow	14:07:09	Friday
2	20EC38	Funiculì funiculà	Mario Lanza	pop	Saint-Petersburg	20:58:07	Wednesday
3	A3DD03C9	Dragons in the Sunset	Fire + Ice	folk	Saint-Petersburg	08:37:09	Monday
4	E2DC1FAE	Soul People	Space Echo	dance	Moscow	08:34:34	Monday

In [1]:

```
data.info()
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Cell In[1], line 1
----> 1 data.info()

NameError: name 'data' is not defined
```

Вывод:

1) Столбцы не правильно названы **-** Решение**:** Переименовать столбцы в змеином регистре **2)** Есть ячейки с пропущенными данными **-** Решение**:** Удалить данные ячейки так как для проверки гипотез они бесполезны

Пред обработка данных:

1) Переименование столбцов**:**

In [4]:

```
data = data.rename(columns={"  userID": "user_id", "Track": "track_name", "genre": 'genr
e_name', "  City  ": "city", "time": "time_of_day", "Day": "day"})

data.head()
```

Out[4]:

	user_id	track_name	artist	genre_name	city	time_of_day	day
0	FFB692EC	Kamigata To Boots	The Mass Missile	rock	Saint-Petersburg	20:28:33	Wednesday
1	55204538	Delayed Because of Accident	Andreas Rönnberg	rock	Moscow	14:07:09	Friday
2	20EC38	Funiculì funiculà	Mario Lanza	pop	Saint-Petersburg	20:58:07	Wednesday
3	A3DD03C9	Dragons in the Sunset	Fire + Ice	folk	Saint-Petersburg	08:37:09	Monday
4	E2DC1FAE	Soul People	Space Echo	dance	Moscow	08:34:34	Monday

2) Удаление ячеек со значениями **None**

In [5]:

In [5]:

```
data = data.dropna(axis=0).reset_index(drop=True)
data.head()
```

Out[5]:

	user_id	track_name	artist	genre_name	city	time_of_day	day
0	FFB692EC	Kamigata To Boots	The Mass Missile	rock	Saint-Petersburg	20:28:33	Wednesday
1	55204538	Delayed Because of Accident	Andreas Rönnberg	rock	Moscow	14:07:09	Friday
2	20EC38	Funiculì funiculà	Mario Lanza	pop	Saint-Petersburg	20:58:07	Wednesday
3	A3DD03C9	Dragons in the Sunset	Fire + Ice	folk	Saint-Petersburg	08:37:09	Monday
4	E2DC1FAE	Soul People	Space Echo	dance	Moscow	08:34:34	Monday

Вывод: все пропущенные ячейки были удалены

Обработка дубликатов:

In [6]:

```
data = data.drop_duplicates().reset_index(drop=True)
```

In [7]:

```
data
```

Out[7]:

	user_id	track_name	artist	genre_name	city	time_of_day	day
0	FFB692EC	Kamigata To Boots	The Mass Missile	rock	Saint-Petersburg	20:28:33	Wednesday
1	55204538	Delayed Because of Accident	Andreas Rönnberg	rock	Moscow	14:07:09	Friday
2	20EC38	Funiculì funiculà	Mario Lanza	pop	Saint-Petersburg	20:58:07	Wednesday
3	A3DD03C9	Dragons in the Sunset	Fire + Ice	folk	Saint-Petersburg	08:37:09	Monday
4	E2DC1FAE	Soul People	Space Echo	dance	Moscow	08:34:34	Monday
...	...	...	...	...	...	...	...
54131	83A474E7	I Worship Only What You Bleed	The Black Dahlia Murder	extrememetal	Moscow	21:07:12	Monday
54132	729CBB09	My Name	McLean	rnb	Moscow	13:32:28	Wednesday
54133	D08D4A55	Maybe One Day (feat. Black Spade)	Blu & Exile	hip	Saint-Petersburg	10:00:00	Monday
54134	321D0506	Freight Train	Chas McDevitt	rock	Moscow	21:43:59	Friday
54135	3A64EF84	Tell Me Sweet Little Lies	Monica Lopez	country	Moscow	21:59:46	Friday

54136 rows × 7 columns

Вывод: Все дубликаты были удалены

Анализ Данных и получение результатов

Проверить гипотезы:

1) Активность пользователей зависит от дня недели. Причём в Москве и Петербурге это проявляется по-

разному. **2)** Утром в понедельник в Москве преобладают одни жанры музыки, а в Петербурге — другие. Это верно и для вечера пятницы. **3)** Москва и Петербург предпочитают разные жанры музыки. В Москве чаще слушают поп-музыку, в Петербурге — русский рэп.

Разделим наши данные на 2 под группы: Москва и Питер

In [8]:

```
moscow_data = data[data["city"] == "Moscow"].reset_index(drop=True)
piter_data = data[data["city"] == "Saint-Petersburg"].reset_index(drop=True)
moscow_data
```

Out[8]:

	user_id	track_name	artist	genre_name	city	time_of_day	day
0	55204538	Delayed Because of Accident	Andreas Rönnberg	rock	Moscow	14:07:09	Friday
1	E2DC1FAE	Soul People	Space Echo	dance	Moscow	08:34:34	Monday
2	4CB90AA5	True	Roman Messer	dance	Moscow	13:00:07	Wednesday
3	F03E1C1F	Feeling This Way	Polina Griffith	dance	Moscow	20:47:49	Wednesday
4	BC5A3A29	Gool la Mita	Shireen Abdul Wahab	world	Moscow	14:08:42	Monday
...	...	...	...	...	...	...	...
37718	C532021D	We Can Not Be Silenced	Pänzer	extrememetal	Moscow	08:38:24	Friday
37719	83A474E7	I Worship Only What You Bleed	The Black Dahlia Murder	extrememetal	Moscow	21:07:12	Monday
37720	729CBB09	My Name	McLean	rnb	Moscow	13:32:28	Wednesday
37721	321D0506	Freight Train	Chas McDevitt	rock	Moscow	21:43:59	Friday
37722	3A64EF84	Tell Me Sweet Little Lies	Monica Lopez	country	Moscow	21:59:46	Friday

37723 rows × 7 columns

Москва и Петербург предпочитают разные жанры музыки. В Москве чаще слушают поп-музыку, в Петербурге — русский рэп.

Для проверки гипотезы достаточно найти самый популярный жанр в каждом городе и сверить с гипотезой

In [9]:

```
top_genre_moscow = moscow_data["genre_name"].value_counts()
top_genre_piter = piter_data["genre_name"].value_counts()
```

In [10]:

```
top_genre_moscow
```

Out[10]:

```
genre_name
pop                  5480
rock                 3810
dance                3567
electronic           3342
hip                  1877
                     ...
malaysian               1
nujazz                  1
loungeelectronic        1
acid                    1
regional                1
Name: count, Length: 263, dtype: int64
```

In [11]:

```
top_genre_piter
```

Out[11]:

```
genre_name
pop               2259
rock              1809
dance             1557
electronic        1547
hip                856
                  ...
international        1
electropop           1
arabic               1
canzone              1
idm                  1
Name: count, Length: 205, dtype: int64
```

Вывод: Гипотеза не верна. И в Москве, и в Питере самая популярная музыка это **POP**

Гипотеза: Утром в понедельник в Москве преобладают одни жанры музыки, а в Петербурге — другие. Это верно и для вечера пятницы.

Для того чтобы проверить неё, достаточно выделить их по времени до 12 часов дня и посчитать какой жанр самый популярный

In [12]:

```
moscow_data["time_of_day"] = pd.to_datetime(moscow_data["time_of_day"], format="%H:%M:%S
")
morning_list_moscow = moscow_data[moscow_data["time_of_day"].dt.hour < 12]
monday_list_moscow  = morning_list_moscow[morning_list_moscow["day"] == "Monday"]

evening_list_moscow = moscow_data[moscow_data["time_of_day"].dt.hour > 18]
friday_monday_moscow  = evening_list_moscow[evening_list_moscow["day"] == "Friday"]

morning_monday_top_genre_moscow = monday_list_moscow["genre_name"].value_counts()
evening_friday_top_genre_moscow = friday_monday_moscow["genre_name"].value_counts()
```

In [13]:

```
piter_data["time_of_day"] = pd.to_datetime(piter_data["time_of_day"], format="%H:%M:%S")
morning_list_piter = piter_data[piter_data["time_of_day"].dt.hour < 12]
monday_list_piter  = morning_list_piter[morning_list_piter["day"] == "Monday"]

evening_list_piter = piter_data[piter_data["time_of_day"].dt.hour > 18]
friday_monday_piter  = evening_list_piter[evening_list_piter["day"] == "Friday"]

morning_monday_top_genre_piter = monday_list_piter["genre_name"].value_counts()
evening_friday_top_genre_piter = friday_monday_piter["genre_name"].value_counts()
```

In [14]:

```
evening_friday_top_genre_piter.head()
```

Out[14]:

```
genre_name
pop           240
rock          208
electronic    199
dance         177
hip            83
Name: count, dtype: int64
```

In [15]:

```
evening_friday_top_genre_moscow.head()
```

Out[15]:

```
genre_name
pop            670
rock           502
electronic     426
dance          397
hip            255
Name: count, dtype: int64
```

Вывод: Гипотеза опровергнута так как все равно самые популярные жанры абсолютно совпадают

Проверим гипотезу: Активность пользователей зависит от дня недели. Причём в Москве и Петербурге это проявляется по-разному.

In [16]:

```
monday_moscow_data = moscow_data[moscow_data["day"] == "Monday"]
wednesday_moscow_data = moscow_data[moscow_data["day"] == "Wednesday"]
friday_moscow_data = moscow_data[moscow_data["day"] == "Friday"]

count_users_moscow_monday = monday_moscow_data["user_id"].unique().__len__()
count_session_moscow_monday = monday_moscow_data["user_id"].__len__()

count_users_moscow_wednesday = wednesday_moscow_data["user_id"].unique().__len__()
count_session_moscow_wednesday = wednesday_moscow_data["user_id"].__len__()

count_users_moscow_friday = friday_moscow_data["user_id"].unique().__len__()
count_session_moscow_friday = friday_moscow_data["user_id"].__len__()

average_session_moscow_monday = count_session_moscow_monday / count_users_moscow_monday
average_session_moscow_wednesday = count_session_moscow_wednesday / count_users_moscow_we
dnesday
average_session_moscow_friday = count_session_moscow_friday / count_users_moscow_friday

result_average_session_moscow = pd.DataFrame({"Day": ['Monday', 'Wednesday', 'Friday'],
                                              "Average Sessions": [average_session_mosco
w_monday,
                                                                   average_session_mosc
ow_wednesday,
                                                                   average_session_mosc
ow_friday]})
result_average_session_moscow
```

Out[16]:

	Day	Average Sessions
0	Monday	1.288458
1	Wednesday	1.287977
2	Friday	1.291807

In [17]:

```
monday_piter_data = piter_data[piter_data["day"] == "Monday"]
wednesday_piter_data = piter_data[piter_data["day"] == "Wednesday"]
friday_piter_data = piter_data[piter_data["day"] == "Friday"]

count_users_piter_monday = monday_piter_data["user_id"].unique().__len__()
count_session_piter_monday = monday_piter_data["user_id"].__len__()

count_users_piter_wednesday = wednesday_piter_data["user_id"].unique().__len__()
count_session_piter_wednesday = wednesday_piter_data["user_id"].__len__()
```

```
count_users_piter_friday = friday_piter_data["user_id"].unique().__len__()
count_session_piter_friday = friday_piter_data["user_id"].__len__()

average_session_piter_monday = count_session_piter_monday / count_users_piter_monday
average_session_piter_wednesday = count_session_piter_wednesday / count_users_piter_wedne
sday
average_session_piter_friday = count_session_piter_friday / count_users_piter_friday

result_average_session_piter = pd.DataFrame({"Day": ['Monday', 'Wednesday', 'Friday'],
                                             "Average Sessions": [average_session_piter
_monday,
                                                                  average_session_pite
r_wednesday,
                                                                  average_session_pite
r_friday]})
result_average_session_piter
```

Out[17]:

	Day	Average Sessions
0	Monday	1.296092
1	Wednesday	1.315222
2	Friday	1.289609

Вывод: Гипотеза опровергнута. Разница присутствует с точностью до сотой, а следовательно не более чем погрешность.

In [17]:

```
```