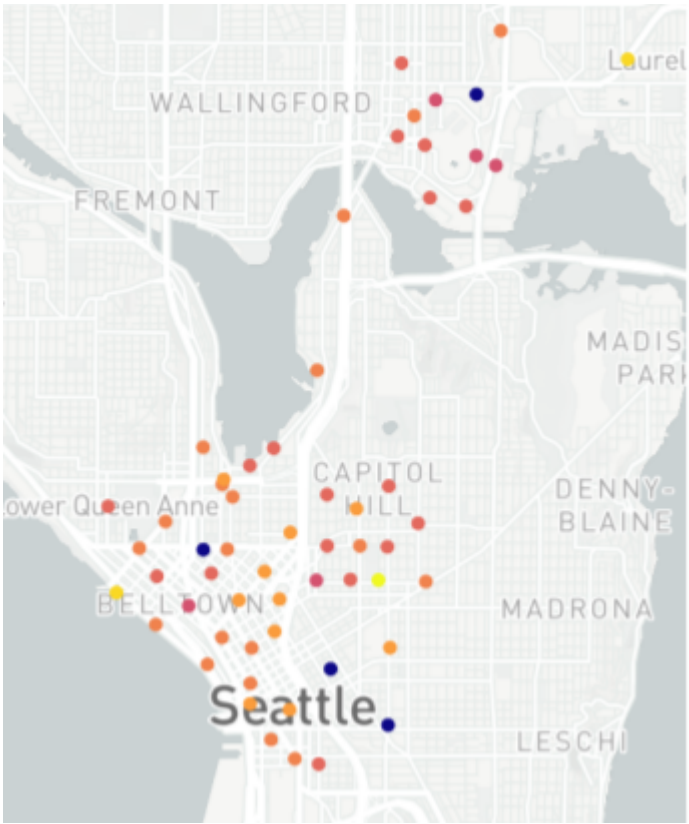
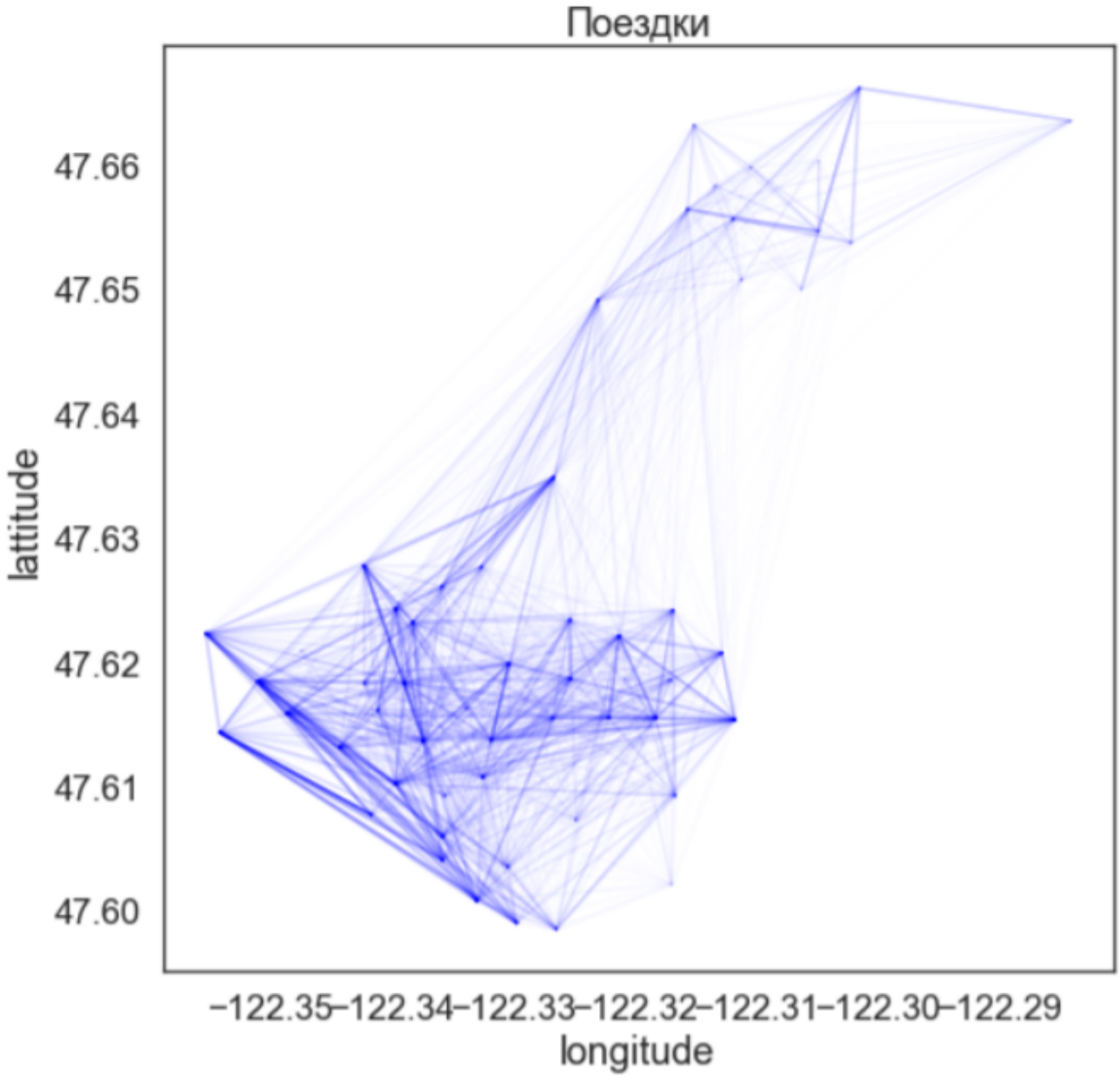
В этом ноутбуке подводятся итоги проведенных исследований, делаются общие выводы и даются рекомендации компании.

1. Визуализация

Если рассмотреть вместимости станций, то никаких закономерностей не видно. На разных станциях (за исключением вышедших из строя) вместимость отличается не сильно. Должно быть, компания не занималось анализом использования станций, чтобы оптимизировать количество слотов под велосипеды на них:

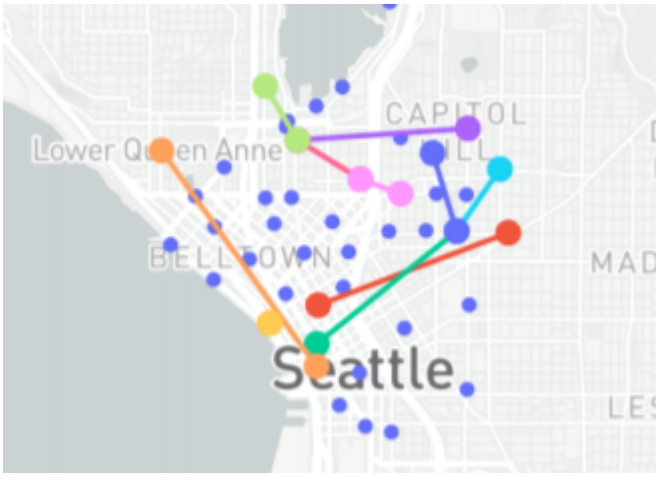


Город можно визуально разделить рекой на две части. При этом в одной из них станций значительно больше, чем в другой. Можно изобразить количество поездок в этих частях (чем прозрачнее линия, тем менее популярен этот маршрут):



Ясно видно, что подавляющее большинство поездок происходит внутри двух кластеров, а между ними поездки довольно редки. Если взглянуть на карту, то видно, что между кластерами довольно большое расстояние. Тогда, пожалуй, компании стоит разместить несколько станций в южной части города, но ближе к реке: это позволит людям удобно добираться из одной части города в другую и может принести компании дополнительную прибыль.

Если же смотреть на популярность маршрутов в зависимости от времени суток, то днем можно заметить активность в "деловых" районах городах, вечером - в "развлекательных". Ниже приведен пример, когда в вечернее время суток появляется активность в районе Capitol Hill, который отличается обилием баров и клубов:



2. "Перебросы"

Если отслеживать перемещения велосипедов, то многие поездки отоюражаются, как перескок велосипеда с одной станции на другую. Была гипотеза, что это сотрудники компании самостоятельно перегоняют велосипеды с одной станции на другую.э

Подобные перескоки были выявлены, и получилось, что их в датасете более 68000 штук:

```
print(f'Найдено перебросов: {len(transfers)}')
print(f'Отношение кол-ва поездок к перебросам: {len(X) / len(transfers)}')
```

```
Найдено перебросов: 68413
Отношение кол-ва поездок к перебросам: 4.193018870682472
```

Значение подозрительно большое. Получается, что примерно на 4 обычные поездки приходится один такой перескок. Если это действительно перемещения велосипедов сотрудниками компании, то компании явно работает неэффективно и этот процесс можно оптимизировать путем балансировки слотов для велосипедов на станциях, но явно об этих перескоках нигде не указано, возможно просто данные в датасете неполные. Компании следует указать, почему такие перескоки есть в датасете, и тогда с ними можно будет провести подробную аналитику и дать компании много полезных рекомендаций.

3. Абонементы

Если хочется понять, как велопрокатом пользуются обладатели различных абонементов, то первое, что хочется проверить, это то, какие признаки влияют на наличие абонемента у пользователя. Для этого был проведен анализ зависимостей с использованием критериев Хи-Квадрат и Манна-Уитни, в результате которого получилось, что гипотеза о независимости отвергается для всех признаков - и категориальных, и вещественных:

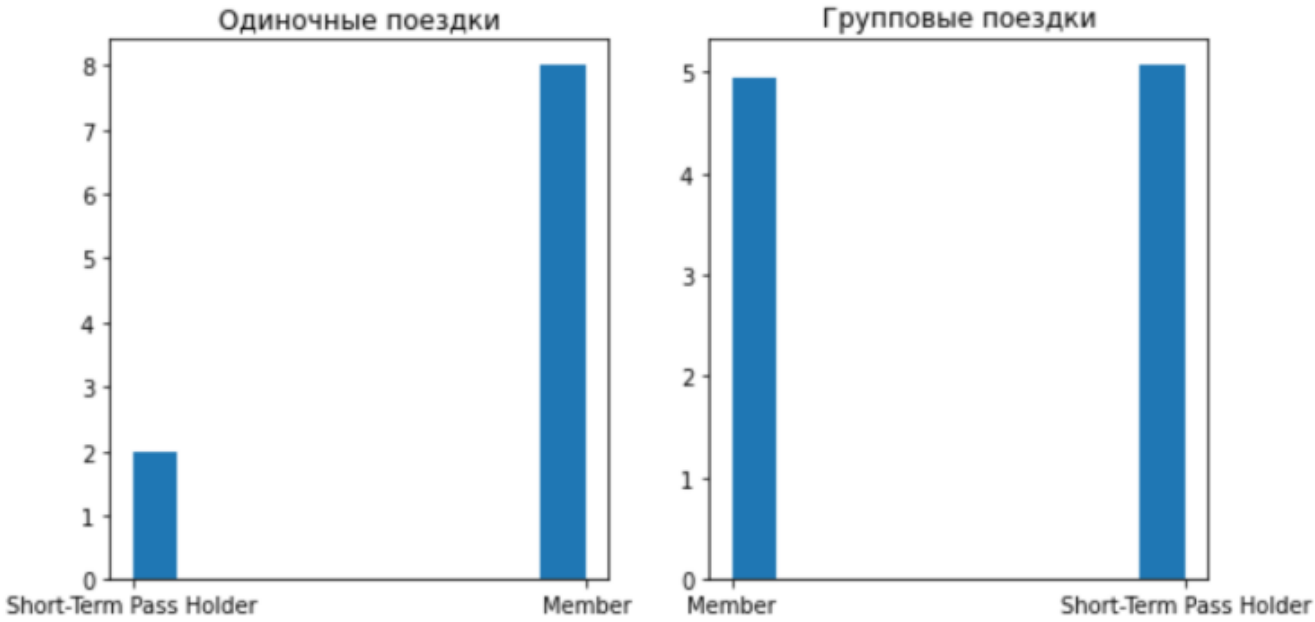
```
Признак from_station_id
p-value: 0.0

Признак to_station_id
p-value: 0.0

Признак gender
p-value: 0.0

Признак birthyear
p-value: 0.0

Признак start_year
p-value: 1.4820725743292082e-296

Признак start_month
p-value: 0.0

Признак start_weekday
p-value: 0.0
```

```
1 multipletests(pvalues)
```

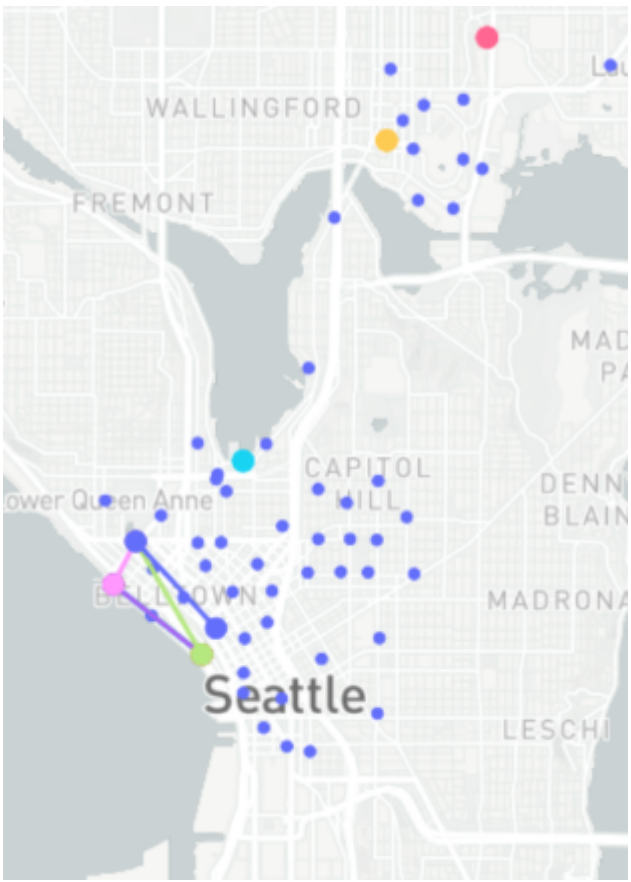
```
(array([ True,  True,  True,  True,  True,  True,  True,  True]),
 array([0., 0., 0., 0., 0., 0., 0., 0.]),
 0.006391150954545011,
 0.00625)
```

Значит компания может строить довольно сложные модели для рекламы абонементов как в зависимости от того, в какое время их стоит рекламировать, так и в зависимости от того, для какой аудитории это стоит делать.

Отдельное внимание стоит уделять тому, как абонементами пользуются участники групповых поездок. Выходит, что более половины участников групповых поездок не владеют абонементами. В таком случае, компания может этим воспользоваться, в realtime находить групповые поездки и предлагать их участникам особые условия на покупку абонемента, например, давать бонусы, если владелец абонемента приводит своего друга, или делать скидку, когда сразу несколько человек покупают абонемент по одному специальному общему коду, который будет выдаваться по окончании совершенной групповой поездки.

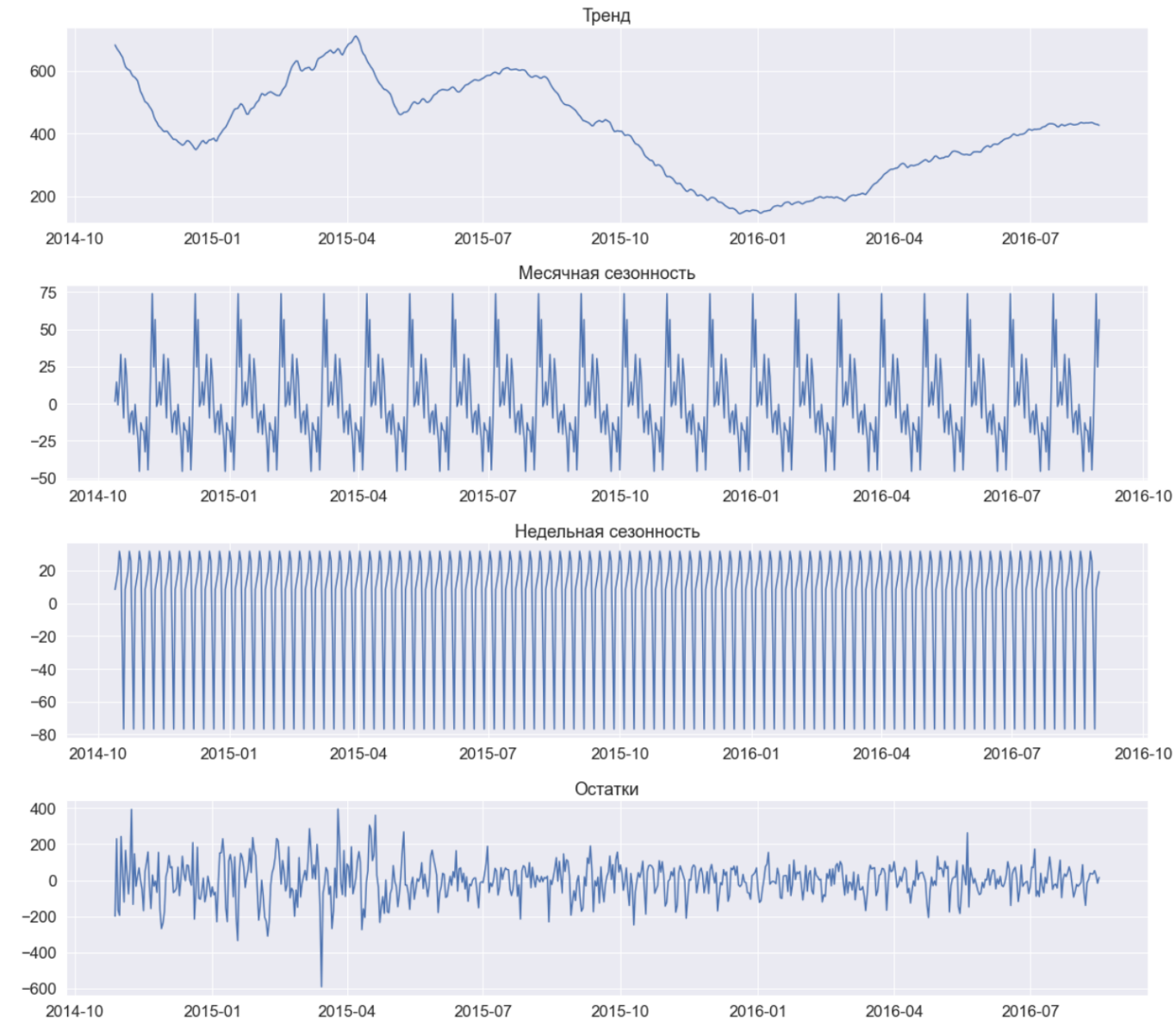


При исследовании популярных маршрутов у владельцев абонементов, была замечена активность в районе Wallingford, чего не было заметно просто при рассмотрении популярных поездок. Возможно компания недооценивает этот район, но на самом деле там есть велолюбители и там стоит провести отдельную рекламную кампании по абонементам:



4. Временные ряды

В ходе анализа временного ряда количества поездок была проведена STL декомпозиция.



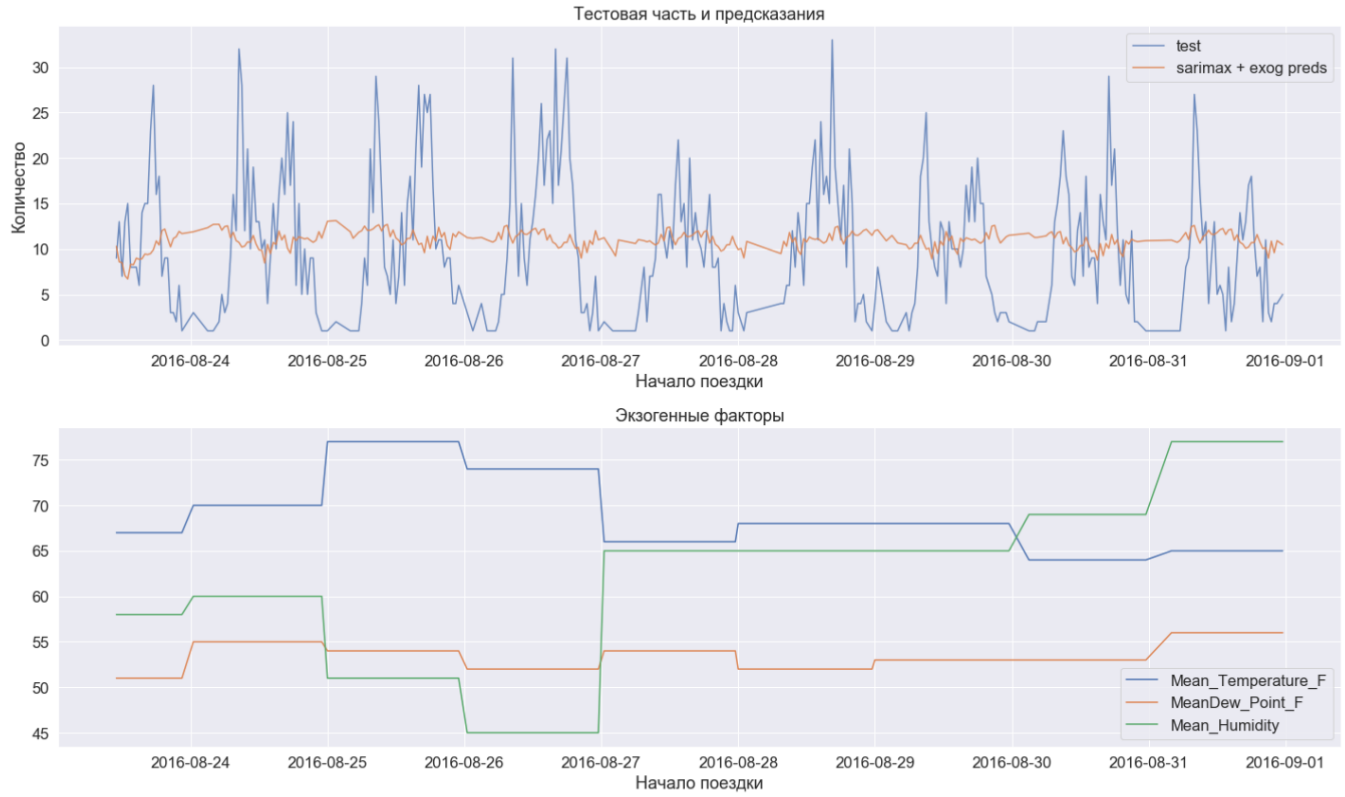
Были выделены месячная и недельная сезонности. Недельную сезонность можно связать с увеличением количества поездок к концу недели, связанную с выходными днями. Месячную, можно например связать с уменьшением накопленных средств населения к концу месяца.

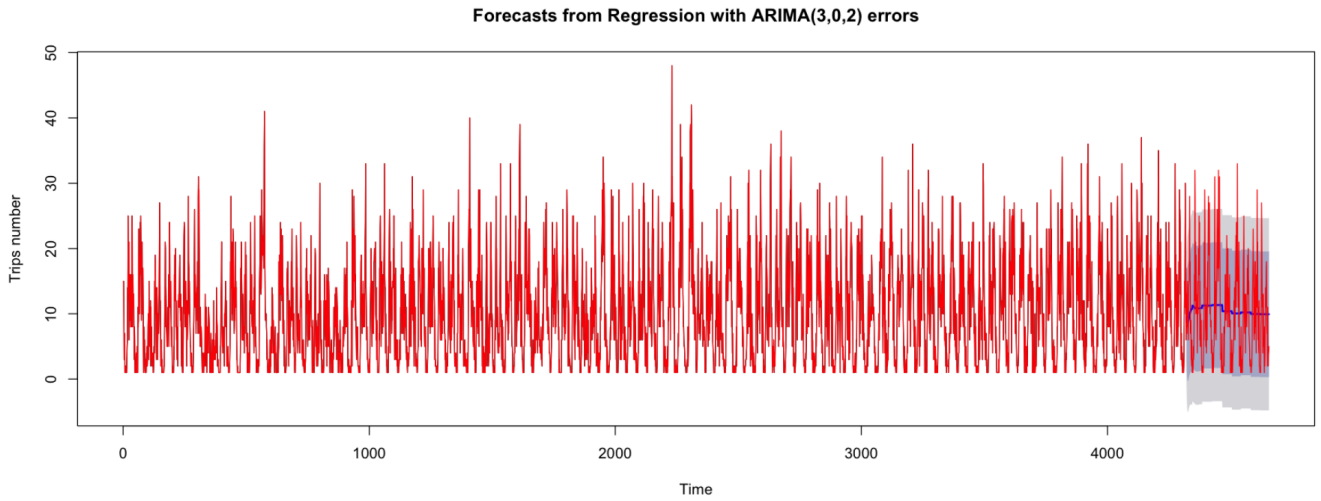
При анализе тренда стоит отметить резкое снижение числа поездок в апреле 2015 года. Судя по данным из википедии [wiki (https://en.wikipedia.org/wiki/Pronto_Cycle_Share)](https://en.wikipedia.org/wiki/Pronto_Cycle_Share), система велопроката испытывала значительные финансовые трудности в 2015 году в связи с тем, что регионгальное правительство Сиэттла ожидало одобрения на её покупку.

Также (с помощью регрессионных моделей) была показана зависимость числа поездок от погодных факторов. Это подтверждает тот факт, что некоторые предсказательные модели улучшили своё качество при добавлении погодных условий в качестве экзогенных факторов.

В связи с этим можно посоветовать компании рассмотреть способы привлечения клиентов в дождливые или холодные сезоны.

	Значимый ли
Max_Temperature_F	False
Mean_Temperature_F	True
Min_TemperatureF	True
Max_Dew_Point_F	False
MeanDew_Point_F	True
Min_Dewpoint_F	False
Max_Humidity	True
Mean_Humidity	True
Min_Humidity	True
Max_Sea_Level_Pressure_In	True
Mean_Sea_Level_Pressure_In	False
Min_Sea_Level_Pressure_In	False
Max_Visibility_Miles	False
Mean_Visibility_Miles	False
Min_Visibility_Miles	False
Max_Wind_Speed_MPH	False
Mean_Wind_Speed_MPH	True
Max_Gust_Speed_MPH	False



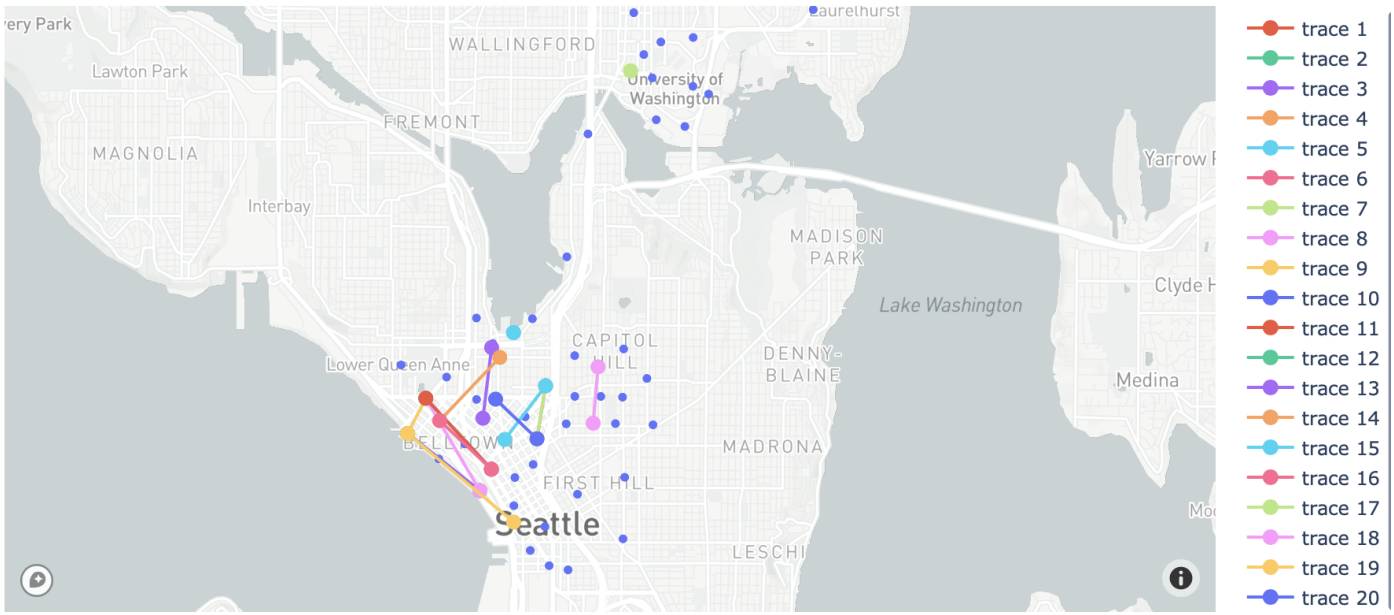


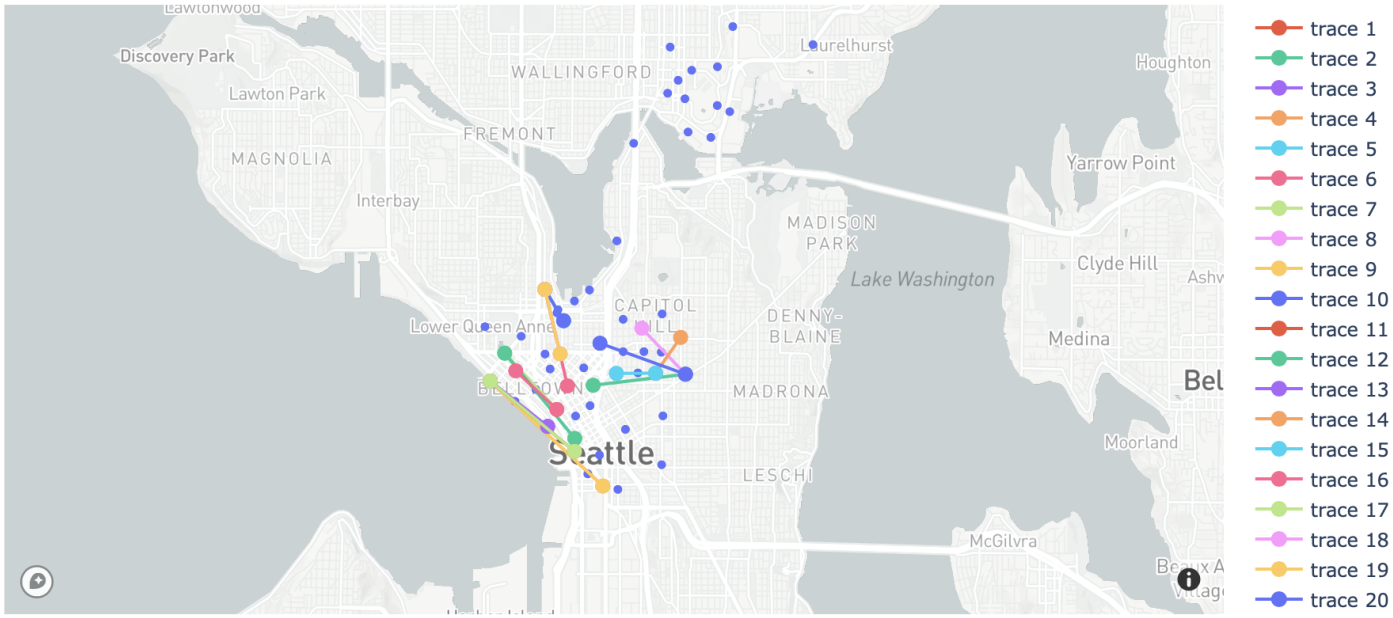
Стоит отметить, что классические известные нам предсказательные модели показали не самое лучшее качество. Возможно, для таких зависимостей стоит использовать более сложные подходы, как, например, нейросетевые модели.

5. Групповые поездки

По результатам исследования мы получили, что $\sim 35\%$ всех поездок - это совместные поездки нескольких человек. Пользователи, совершающие групповые поездки, составляют значительную часть от всех. Следовательно, если на данный момент компания не разделяет таких пользователей на две отдельные категории, то стоит это делать.

Ниже приведены визуализации на карте наиболее популярных маршрутов у групп, и у одиночных пользователей соответственно.





Достаточно много одинчных поездок сосредоточено в районе Capitol Hill с большим количеством баров и клубов.

Большинство групповых поездок пролегают в районах Belltown и Laurelhurst. Район Belltown, например, примечателен обилием культурных достопримечательностей. Это могут быть экскурсионные или туристические поездки. Например, можно предложить компании начать сотрудничать с местными музеями, отелями или экскурсионными бюро.

Исходя из этих результатов компания может пересмотреть тарифные планы для групповых поездок в этих районах.

Можно ввести дополнительные скидки, специальные предложения на совместные поездки в данных районах, увеличив тем самым лояльность клиентов.

Групповые поездки плохи тем, что они единовременно резко увеличивают или уменьшают загруженность конкретной станции. Поэтому, информацию о популярности станций можно использовать для регулирования пользовательского спроса цен на прокат в различных районах с помощью динамической ценовой политики или регулирования размера станций в конкретных районах.

Также, была статистически показана зависимость числа групповых поездок от дня недели, и зависимость длительности поездки от размера группы.

```
In [19]: 1 sps.spearmanr(unique_trips.members_num, unique_trips.tripduration)
Out[19]: SpearmanrResult(correlation=0.20887796587249194, pvalue=0.0)
```

p_value < 0.05 гипотеза о независимости выборок отверглась. Следовательно, можно сказать, что есть тенденция к уменьшению длительности поездки при увеличении размера группы.

Это также может повлиять на тарификацию.

Итоговые рекомендации компании:

- Поставить станции в южной части города ближе к реке, это упростит жителям перемещение из района Wallingford в Capitol Hill и принесет дополнительную прибыль.
- Сделать описание перескокам на kaggle, в таком случае энтузиасты смогут провести более продвинутую аналитику и дать полезные рекомендации об оптимизации мест на станциях.
- Провести специальные рекламные кампании абонементов для участников групповых поездок и для жителей района Wallingford.

- Начать сотрудничество с экскурсионными и туристическими компаниями в районе Belltown.
- Пересмотреть общую тарификацию поездок и размер каждой станции в соответствии с выводами, сделанными нами выше.

In [ ]:

```
1
```