1. Взять stl разложение каждого ряда, в каждой точке найти долю влияния каждой компоненты (в каждой точке все делим на значение ряда) - признаки
2. Взять числовые признаки, кластеризовать строки, найти центры, признак – расстояние каждой строки от центра. **Это в самом конце, когда много числовых и бинарных признаков.**
3. Взять значение ряда T+k-24 (за сутки до прогноза), аналогично Т+k-168(неделя).
4. Бинарные признаки утро день вечер ночь, выходной, праздник
5. Географические – площадь зоны, а также бинарные – аэропорт, центр города, бродвей.
6. количества поездок, совершённых за прошлый час/день и т. д. из соседних географических зон, или количества поездок, совершённых за прошлый час/день **в** текущую географическую зону.
7. **Естественно, предыдущие прогнозы.**
8. **Попробовать синусы и косинусы брать не т, а т+k (настроить на время прогноза)**
9. *Прогноз линейной регрессии по каким-то еще признакам тоже признак*
10. *Широта и долгота центра зоны (либо их произведение)*

**Идея для 7 недели – построить 612 моделей, по 6 для каждой зоны. Модели брать лучшие за предыдущие недели.**

**Дополнительно:**

1. Чтобы составить количество поездок из соседних зон нужно записать все 2500 зон в массив, затем найти соседние с помощью прохождения по 102 зонам и записи соседей в словарь, затем найти по словарю и большой таблице с исходными рядами почасовую сумму поездок.

2 . Для того, чтобы подсчитать, сколько поездок в текущую зону, можно просто заново агрегировать данные, но взять другие широту и долготу (точку высадки).

1. Чтобы найти доли поездок из данной зоны в центр или из центра, аналогично в аэропорт и на окраину, можно добавить столбцы с бинарными переменными в сырые данные, после чего агрегировать и найти среднее для каждой категории.

Таким образом, для каждого признака из сырых данных будет свой большой датафрейм. Расчет около 8 мин каждый. 10 признаков – 80 мин – почти 1,5 часа.

Но как сократить время, на вести счет для ненужных зон? Похоже, что никак, связь только в сырых данных, после агрегации маршруты уже не установить. Сырые данные агрегируются долго, поскольку гистограммы строятся по всем зонам.

Поэтому не больше 10 признаков такого типа, иначе слишком долго.

Зоны по принадлежности:

Центр, зона максимального скопления поездок: 1181,1231,1281,1282,1232,1230,1182,1233,1180

Бродвей:1126,1127,1076

Аэропорт:1684,1734,1733,1783,2069,2119,2068,2118,2168

Зона минимального скопления из рассматриваемых 102: 1441,1442,1390,1439,1533,1483,1482,1532,1380,1480,1530,1580,1630,1426,1376,1377,1378,1272,1221,1273,1274,1175,1225,1222,1172

4 признака по принадлежностям для агрегирования числа поездок из данной зоны в какой-то из этих географических кластеров.