滴滴出行数据探索分析

滴滴算法大赛提供了滴滴真实的出行数据，要求根据以往的数据预测未来特定时间点和低于的打车需求缺口；由于数据量过大，只选取了1月1日一天的出行数据做探索性分析，旨在巩固R语言的使用熟练度的同时，尽可能地熟悉滴滴提供的出行数据，了解其体现出的出行特点以及相关影响因素。

数据集基本特征

该天的数据包含498824条数据，由于在之前的数据整理时对订单数据进行了去重操作，因此该数据集中每一条数据都代表了一份订单（司机不一定接单）。

```
## [1] 498824      5
```

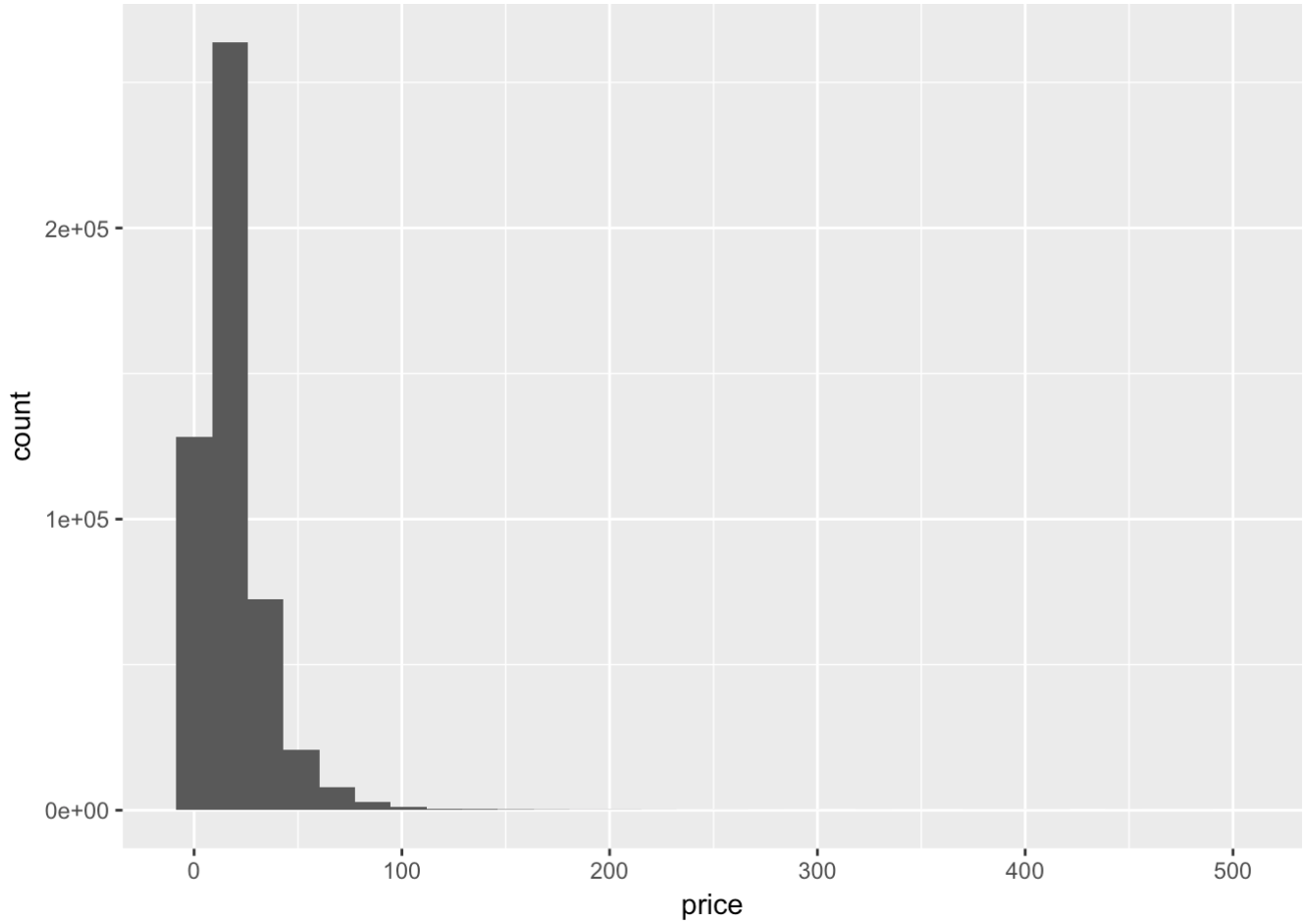
下表展示了该数据集各列数据的基本数据特征，分别表示：重新编号后的订单编号、订单价格、时间片、区域id、是否接单。

```
##        X              price          time_slient      district_id
##  Min.   :     0   Min.   :  0.00   Min.   :  1.00   Min.   : 1.00
##  1st Qu.:124706   1st Qu.:  8.00   1st Qu.: 46.00   1st Qu.:12.00
##  Median :249412   Median : 14.00   Median : 79.00   Median :26.00
##  Mean   :249412   Mean   : 18.77   Mean   : 74.03   Mean   :29.23
##  3rd Qu.:374117   3rd Qu.: 23.00   3rd Qu.:105.00   3rd Qu.:48.00
##  Max.   :498823   Max.   :499.00   Max.   :144.00   Max.   :66.00
##                                    NA's   :2450
##  driver_is_null
##  False:324091
##  True :174733
##
##
##
##
##
```

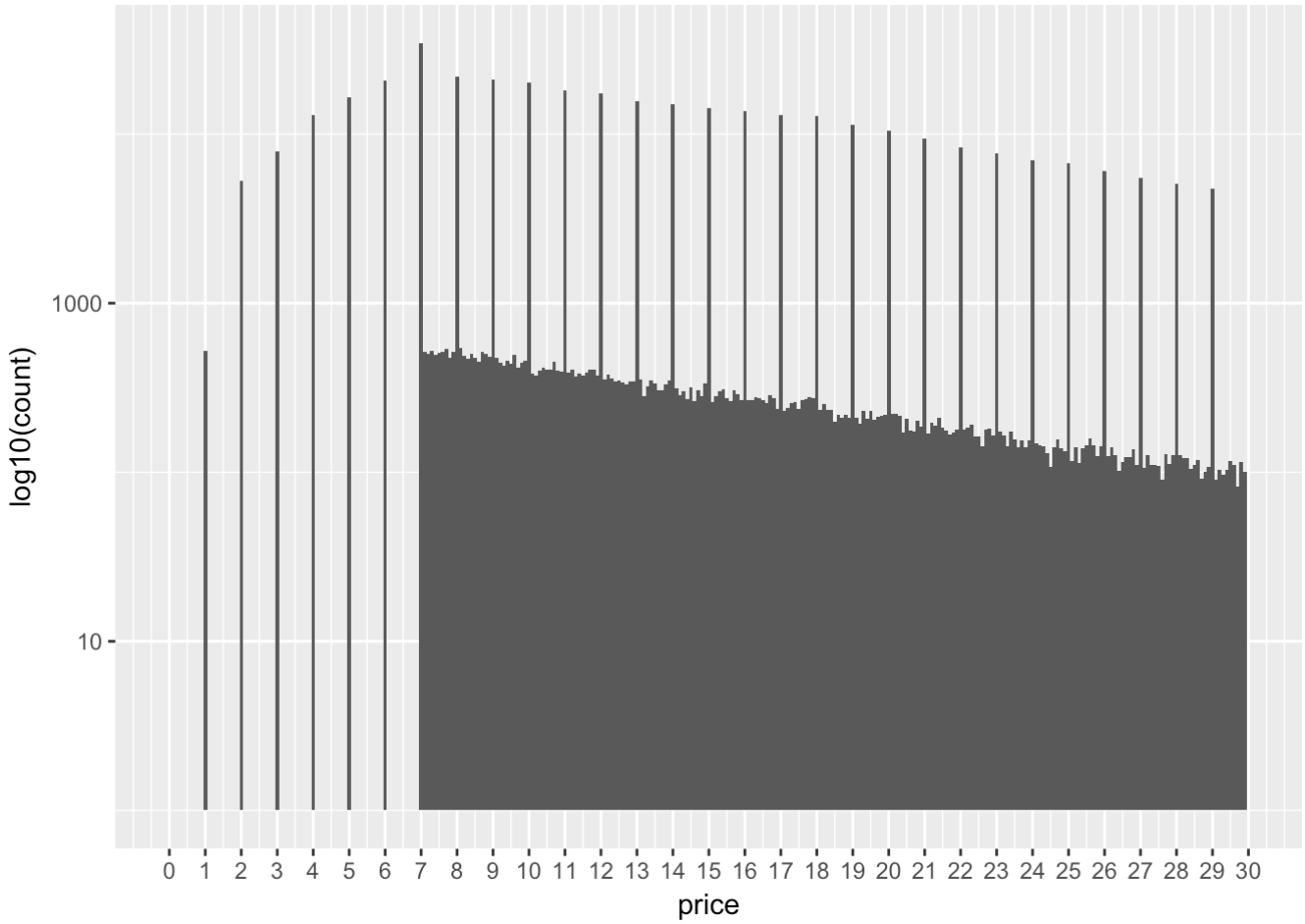
价格数据整体分析

下图为1月1日所有订单价格数量分布的直方图，观察订单付费价格后可以发现绝大多数订单价格都在100元以下，最低付费是0元，最高是499元并显著高于其他价格数据；

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



调整x轴的组宽等参数、对y轴做log10处理后，通过放大100元以内的价格数据后发现：付费价格次数最多的是7元，同时7元以内的消费价格数据是相对离散的整数价格，7元以上的数据数量总体呈下降趋势。



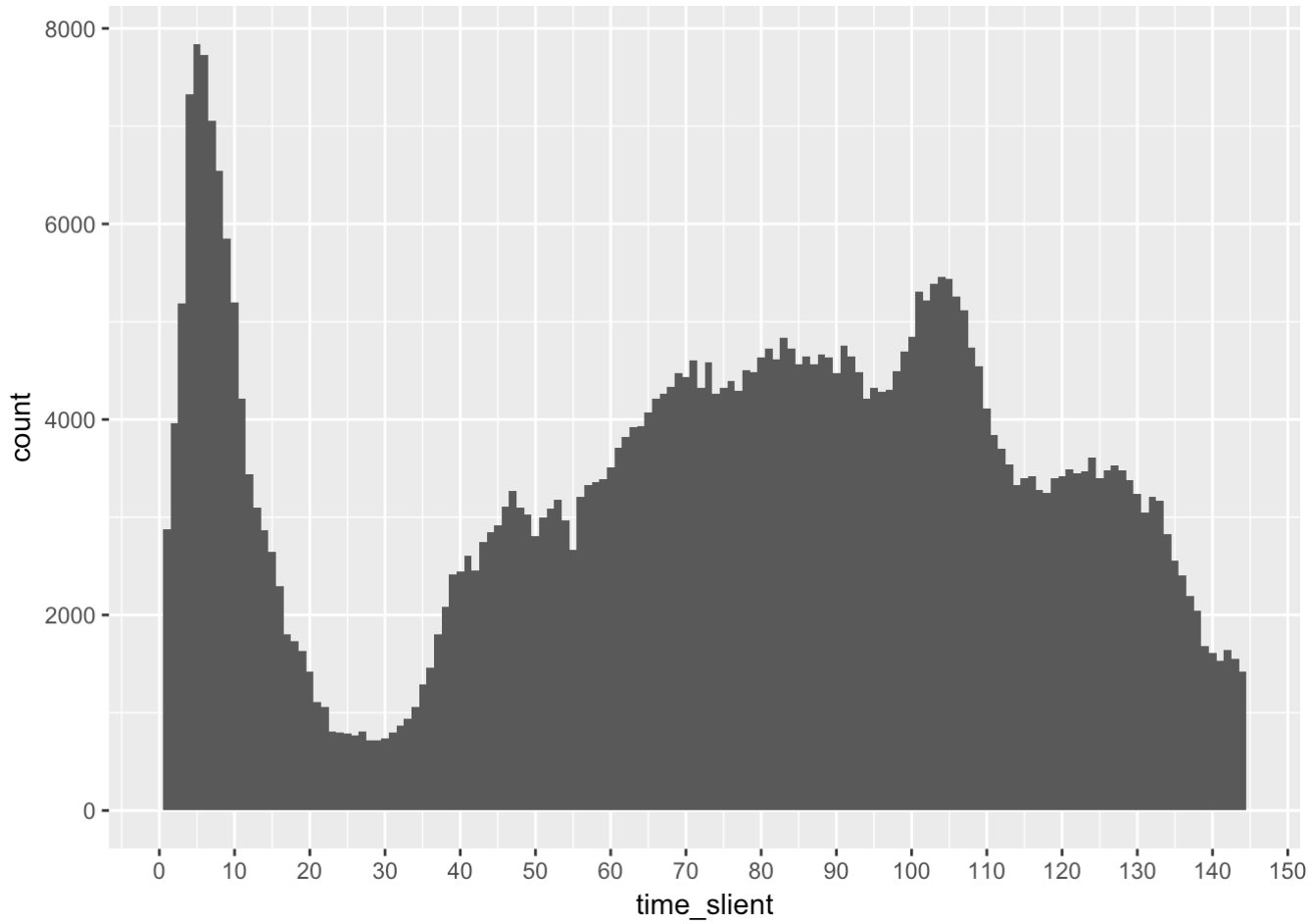
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    8.00   14.00   18.77   23.00  499.00
```

推断价格数据如此分布的原因可能为：该地区滴滴的最低消费价格为7元，低于7元的整数付费数据是因为使用了滴滴的优惠券，而通常优惠整数的金额，所以7元以下的消费数据均为整数价格；另外订单价格为7元以上的数据其数量随价格的升高而下降。

订单数量分时分析

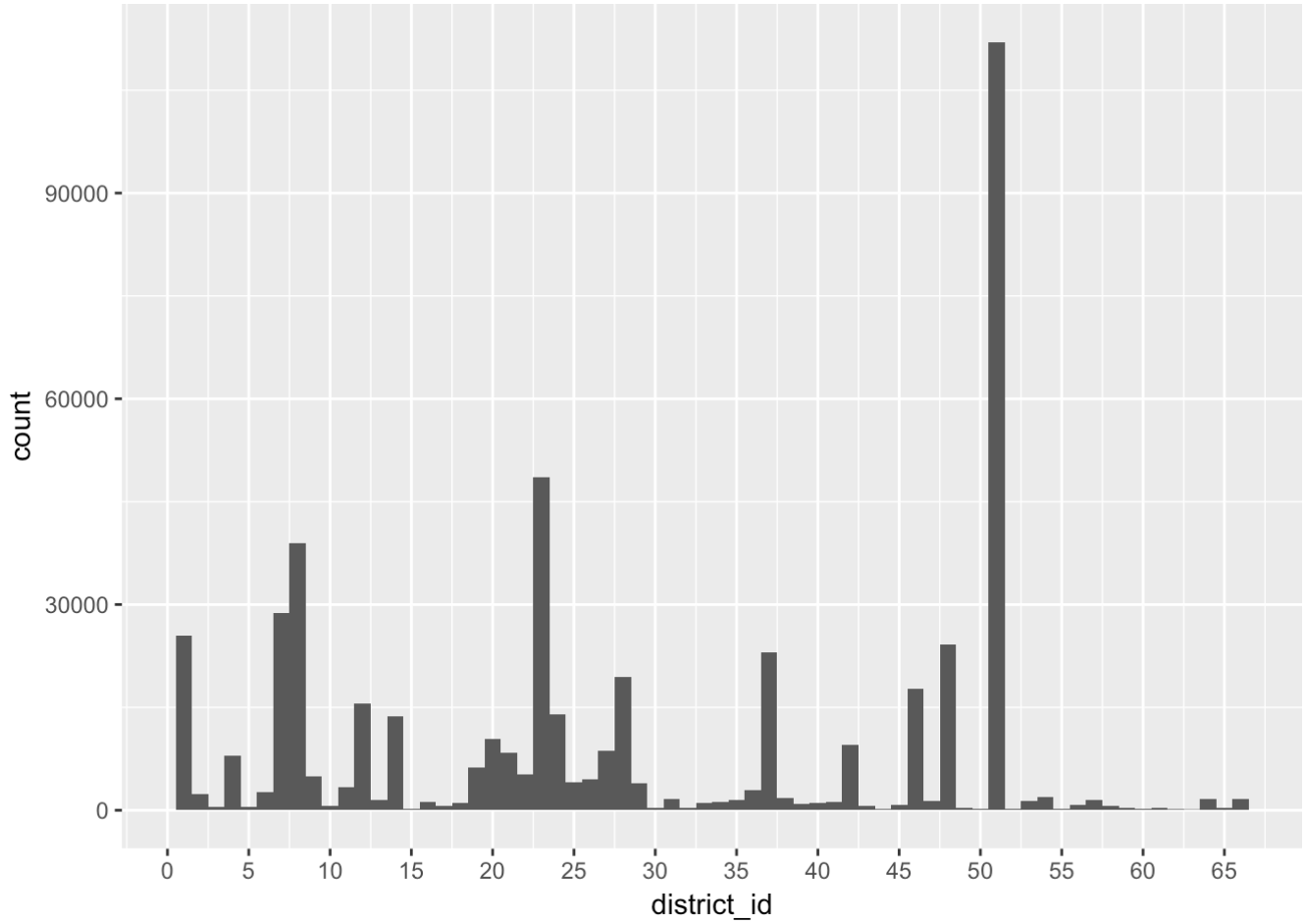
从时间片的数量分布可以看出，1月1日当天的出行最高峰在凌晨1点左右爆发，凌晨4点-5点是全天出行订单量最少的时段，而晚高峰时段下午5-6点恰好是当天出行量第二高峰。

分析原因，由于该数据为1月1日的出行数据，其数据存在一定的特殊性，很多乘客可能会选择在外聚会跨年，而在凌晨1点左右选择打车回家，所以会在零点之后出现全天订单量的最高点。



订单量地区分布

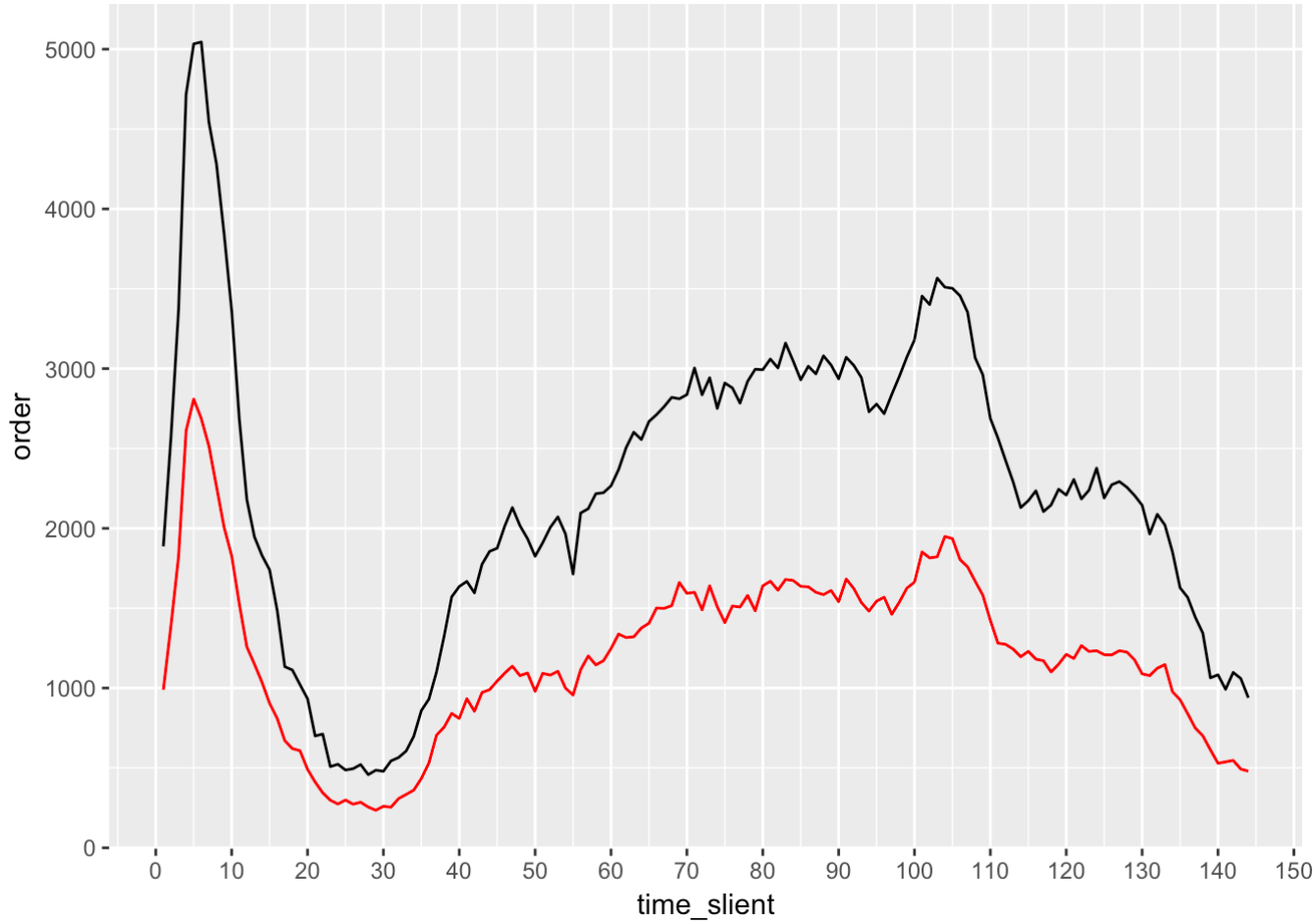
从当天不同区域的订单数量分布中可以看出，51号地区当天的订单超过20万，远远超过其他地区，说明51号区域可能是当地相对人口密度较高、较为繁华的区域，所以拥有大量的出行订单。



订单和缺口数的分时分析

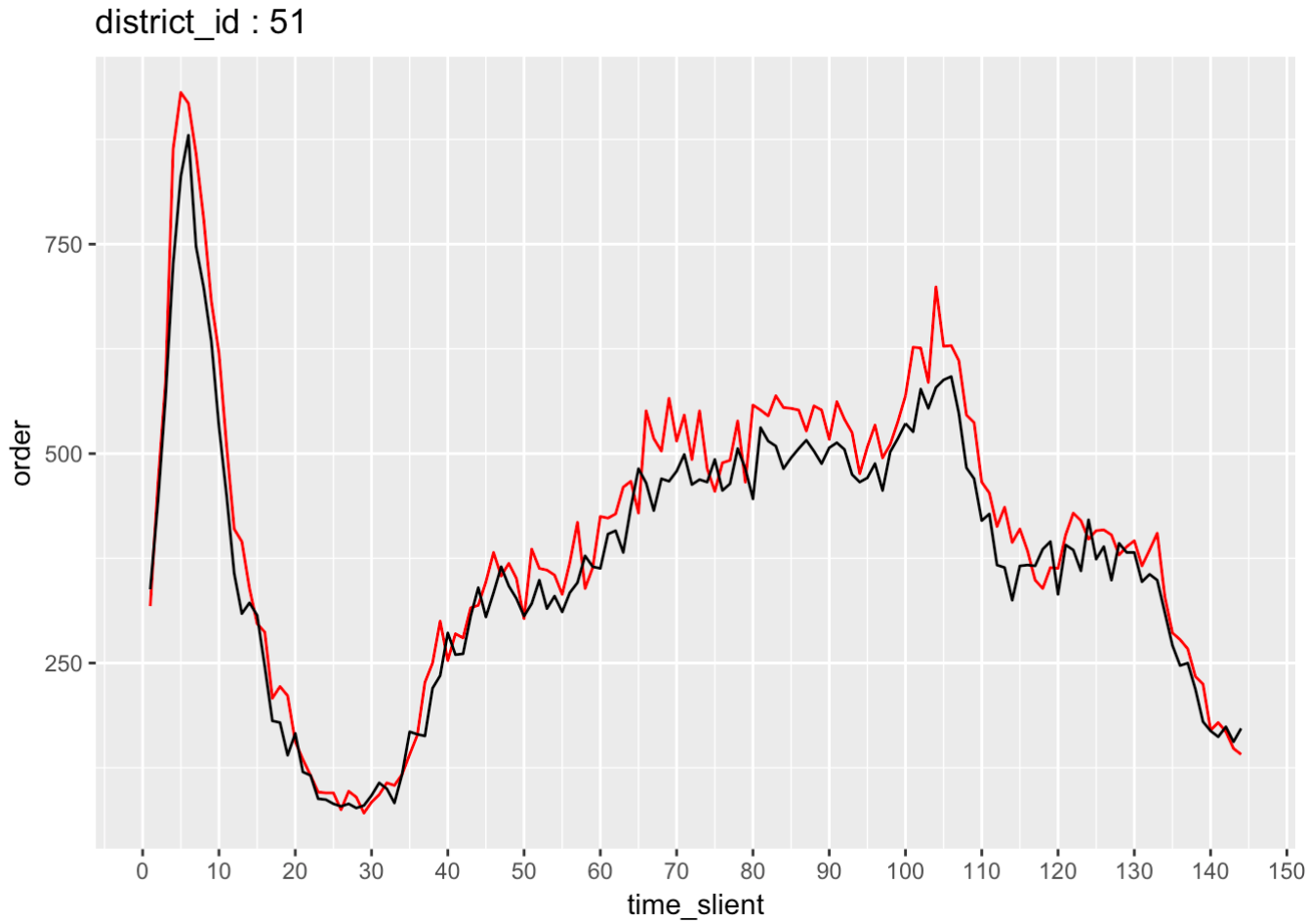
下图为2016-01-01当天全区域成功出行的订单量和未出行订单量的时间分布图，黑色线条表示的是成功出行的订单数量，红色线条表示未出行的订单量，即出行缺口数。

观察可知，从全区域来看成功出行的订单数始终超过未出行订单数，并且两者上升和下降趋势基本保持一致，并且缺口数量的线条（红色）的上升或下降趋势似乎略微提前与成功订单数量的线条，为了进一步分析可以选取订单总量最多、相应的趋势可能会更明显的51号区域的数据进行观察。



下图所示为51号区域当天成功出行的订单量和未出行订单量的时间分布图，相比较上图可以明显地发现红色线条的波峰波谷相比黑色线条在横轴上略微左移了1-2个时间片，其原因可能是当该区域的缺口数增大时，滴滴自身的调节机制开始发挥作用，通过加价等策略吸引更多司机，从而降低缺口数、增加成功的订单数量。

另外51号区域的缺口数量几乎全部都超过成功出行的订单数量，说明该区域很有可能是当地很繁华的区域，人口非常密集，打车需求很高，因此导致有大量打车需求同时也有大量的出行缺口。

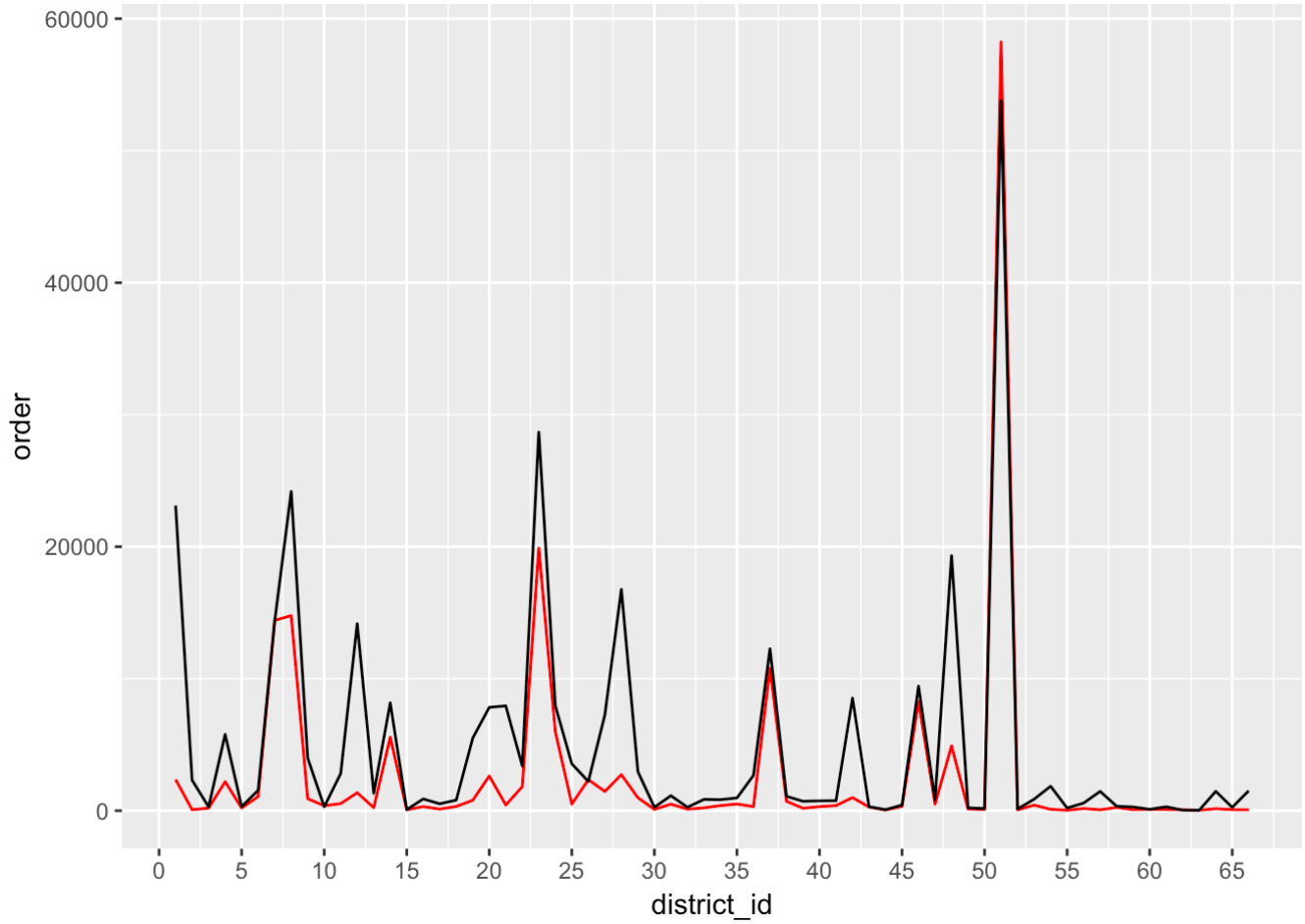


订单和缺口数的区域分析

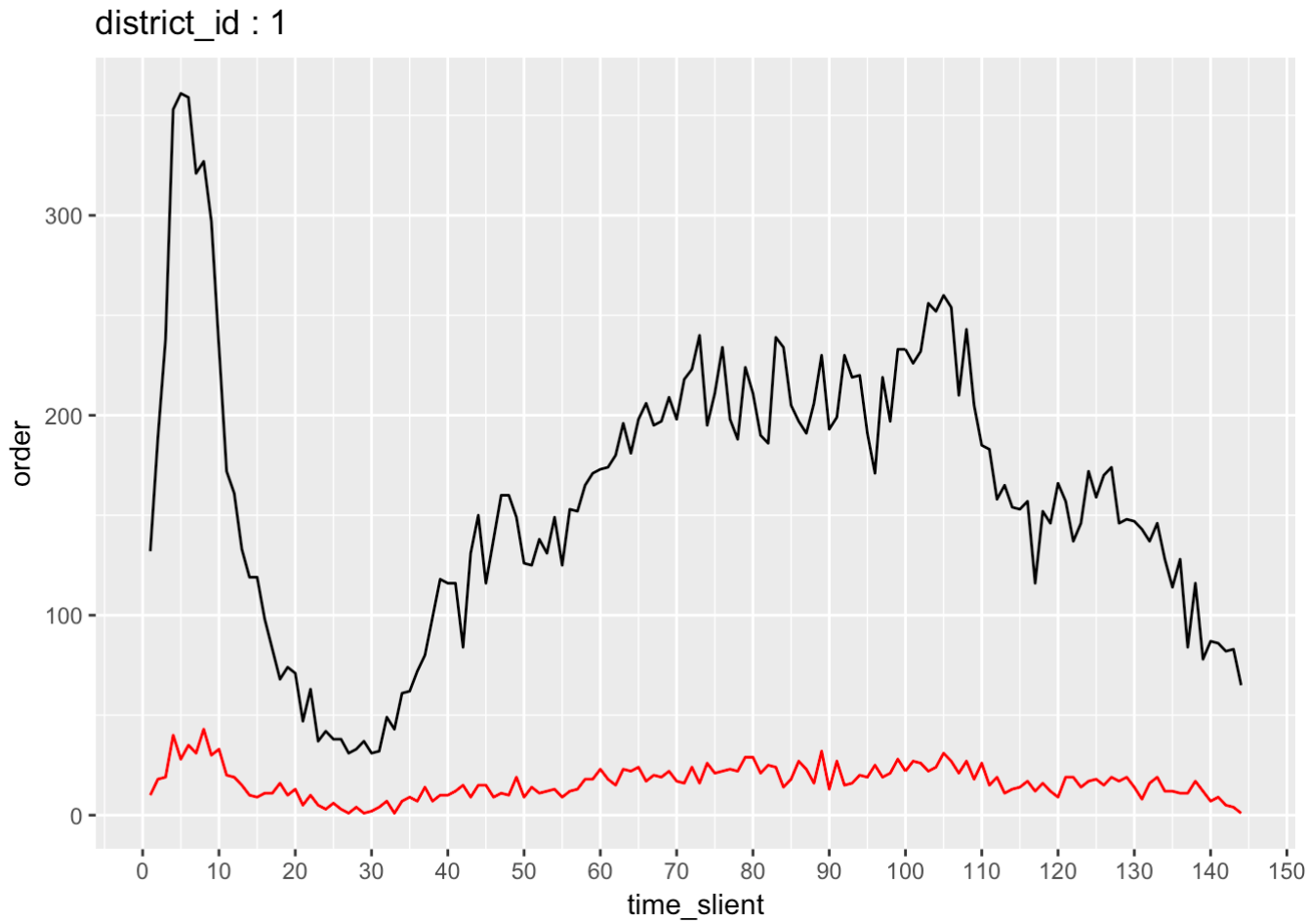
下图显示了各个区域（66个）全天成功出行的订单量和缺口量的分布，从图中可以看出该地区66个区域的订单分布是机器不均衡的，订单量排名前五名的区域（51、23、8、1、48）拥有占总体超过50%的订单，而排名后33位的只拥有整体4.37%的订单，最低的区域订单量只有71单。

其中1号区域和51号区域比较典型，1号区域拥有总体订单5.1%的份额,其特点是订单量大，其中成功出行的订单数量大而缺口数很小；51号区域拥有总体订单22.5%的份额，其特点是订单量非常大，但其中缺口数量比成功订单数量还多。

因此选取1号区域在1月1日的分时图做进一步分析。



如下图所示，区域1的成功订单量和缺口数虽然上升和下降趋势基本保持一致，但缺口的数量明显小于成功订单量，始终保持在50以下。

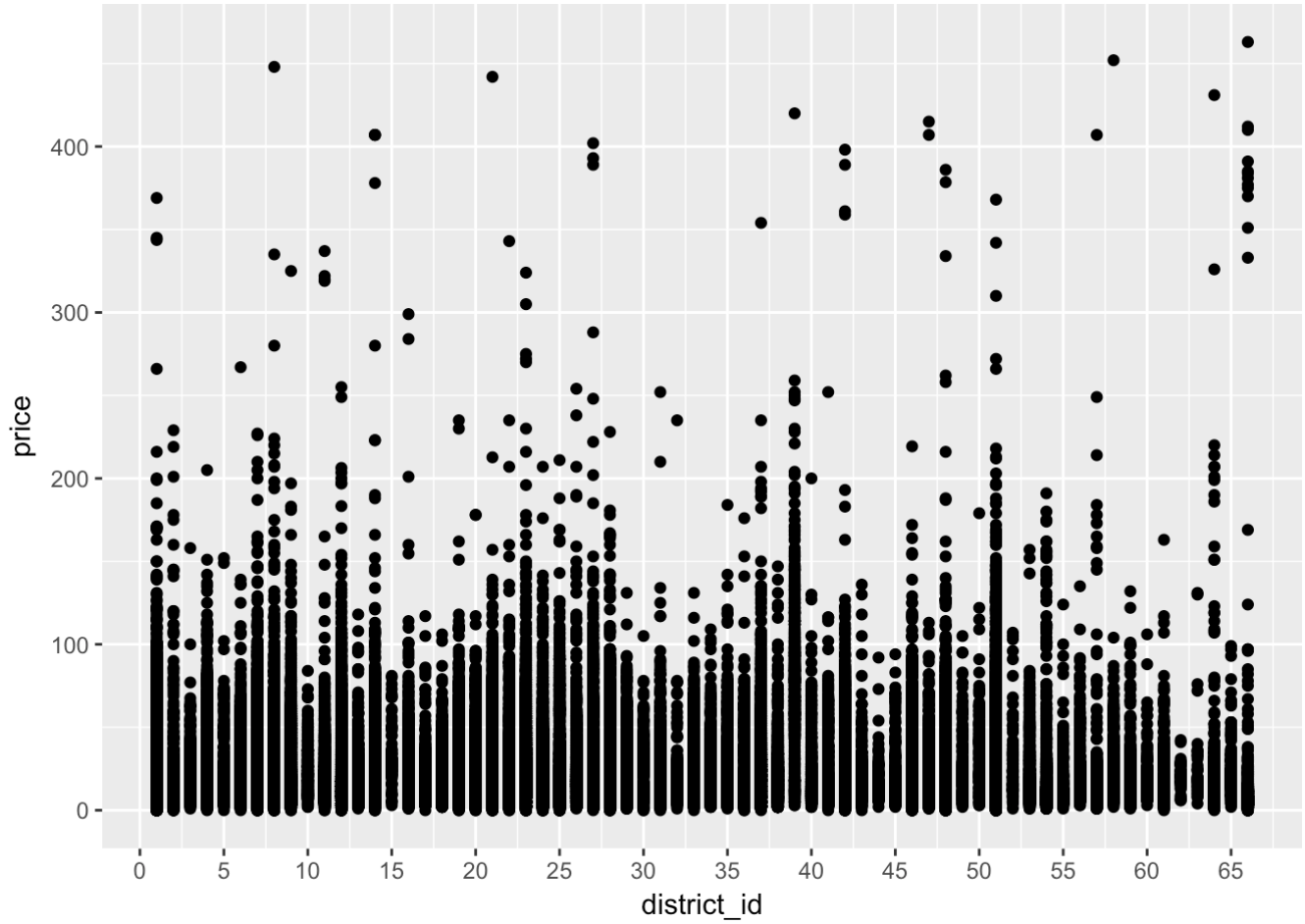


综合1号区域和51号区域的分时图可以做出初步推测，51号区域可能是当地的市中心繁华区域、交通比较拥堵，虽然拥有大量的出行订单，但由于人口密度很大以及区域容纳量的限制，导致即使滴滴采取了加价等调节策略，依然无法显著降低当地的订单缺口数；而1号区域初步推测有可能是住宅集中、交通比较畅通的区域，虽然也有较大的订单量，但由于滴滴司机数量充足，因此订单缺口始终保持很低的数量，甚至在一些低峰时段订单缺口接近于0。

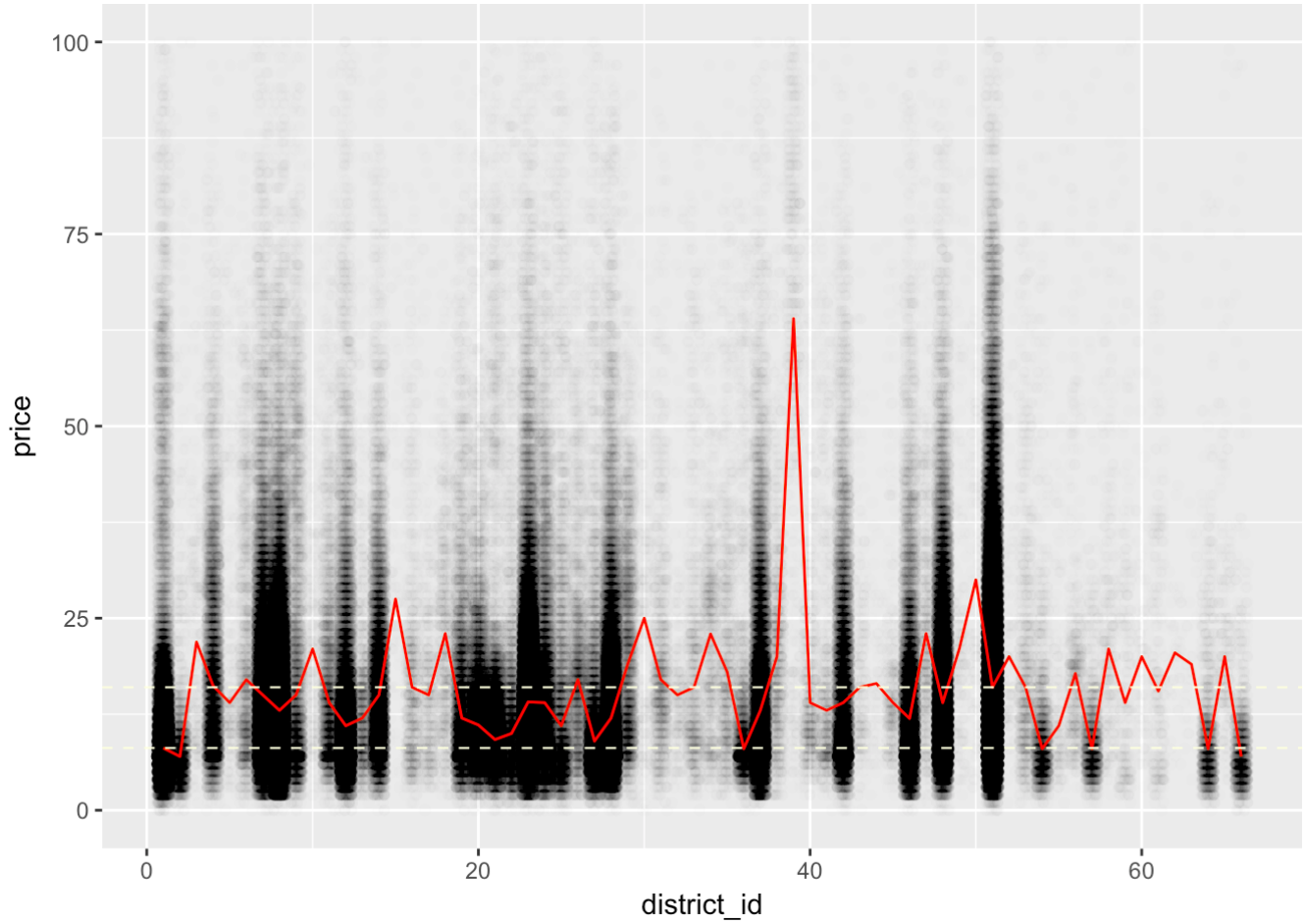
根据这样的推测，可以进一步推测51号区域可能离该地区较集中的居民区较远（类似纽约的曼哈顿）、交通比较拥堵，因此产生的订单价格可能普遍较高；而1号区域则类似城市发展的新区，拥有较集中的居民区和完善的娱乐生活设施，同时交通比较畅通，所以产生的订单价格可能会相对较低。但要验证这样的推测，还需要根据价格数据和地域属性数据做进一步分析验证。

区域与订单价格分析

下图为各个区域产生的有效订单价格数据的散点图，但由于绝大部分数据点都位于150元以下造成了极大的重合，很难根据这幅散点图得出有效的结论，因此对散点图设置透明度，并增加抖动通过增加噪音的方式使得数据展现的更清晰，避免数据点重叠造成的影响。



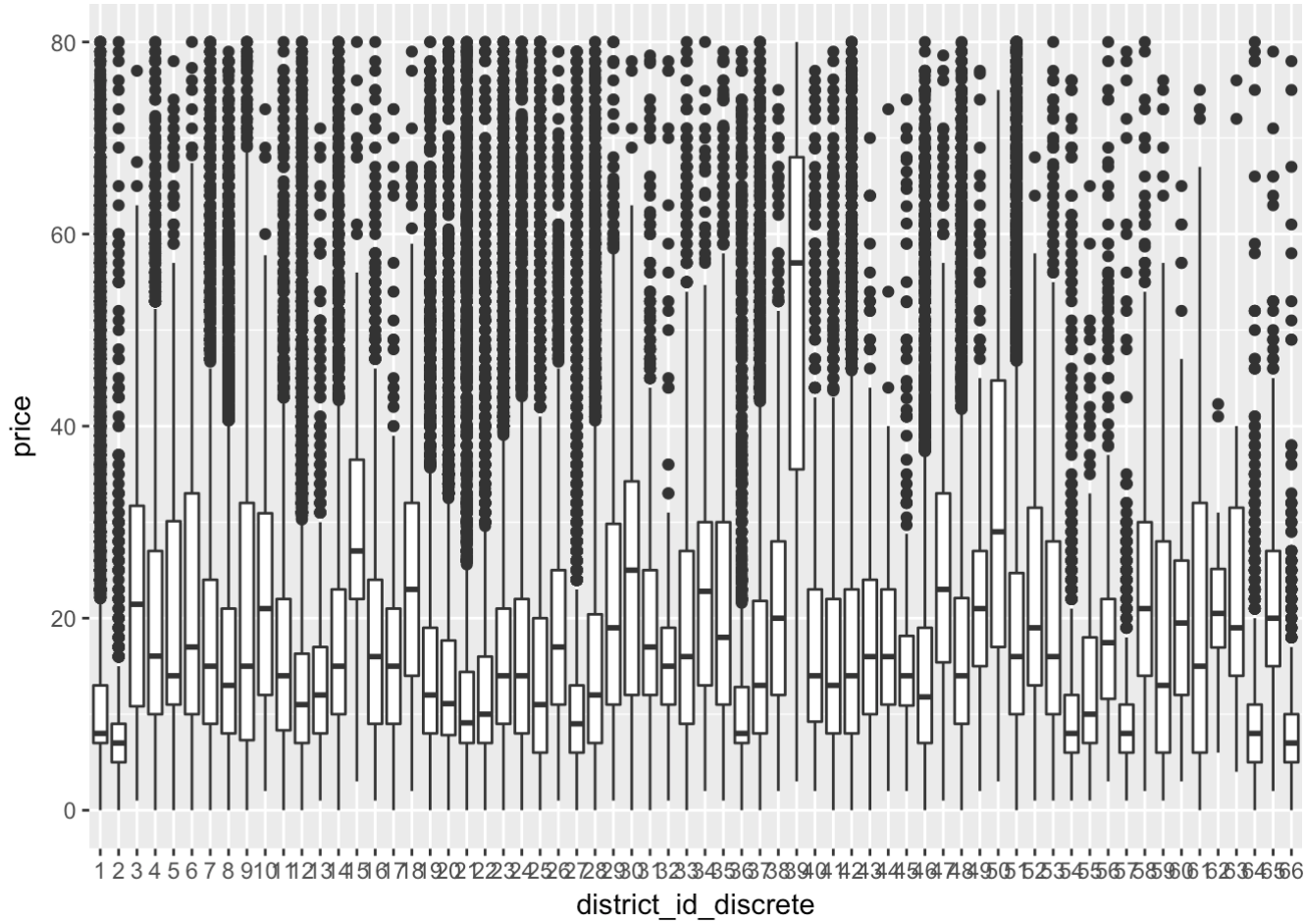
下图将透明度的alpha值设置为1/120，即处理后120个数据点构成一个完整颜色的数据点；红色线条为有效订单价格的中位数。



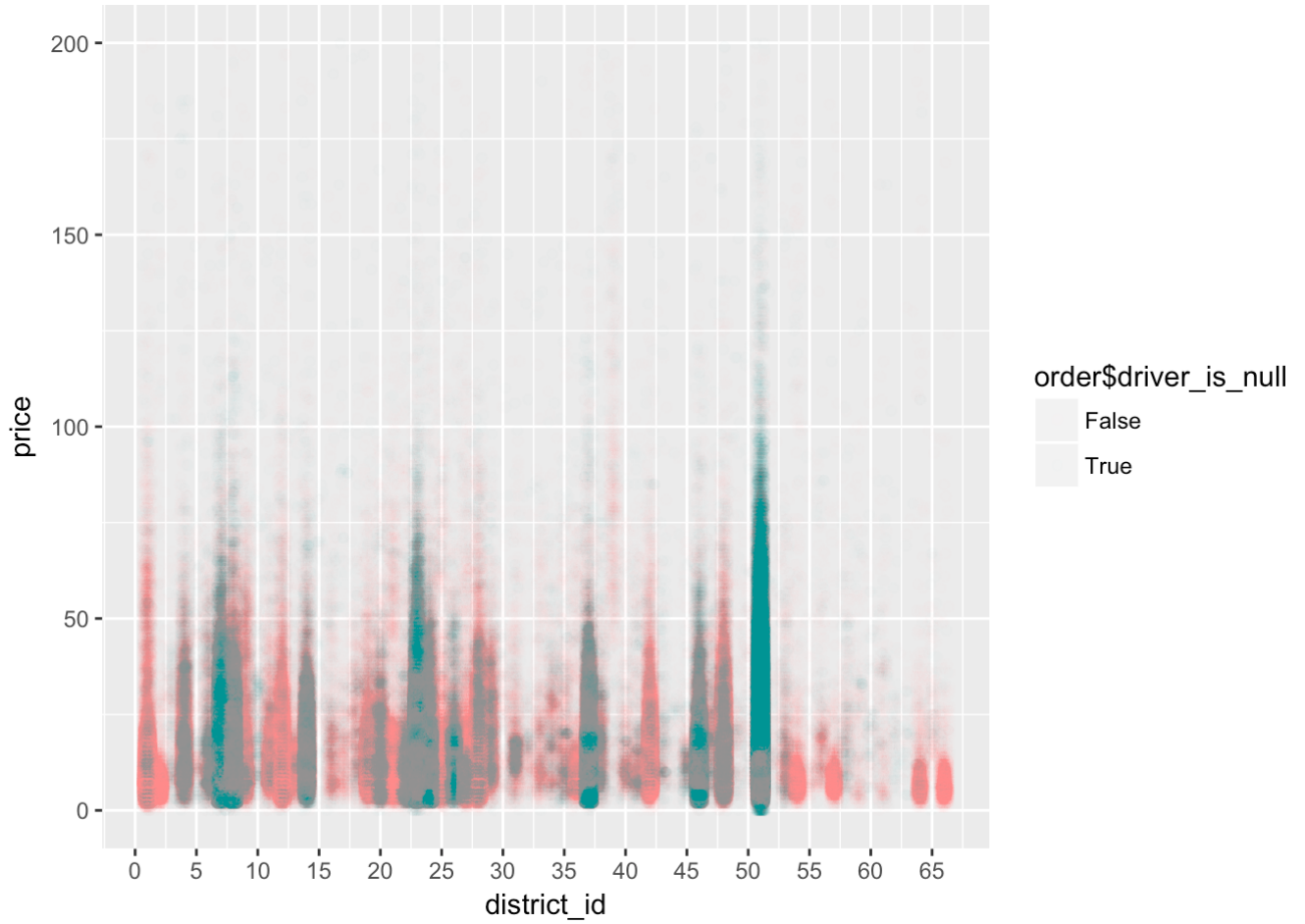
根据处理后的结果可以初步验证之前对1号区域和51号区域的推测，1号区域的整体订单价格非常的低，而51号区域的订单价格虽然不是所有区域中最高的，但去除掉一些订单量很少的特殊区域（如39号、15号、50）后发现，51号区域的整体订单价格相对很高，与之前的推测结果一致。

同时发现39号区域也是比较典型的区域，其订单价格普遍很高但订单量并不大，推测可能是离市中心比较偏远、人口密度很低或是滴滴覆盖程度比较低的地区，同时该区域的用户有很强的支付能力，愿意付出高价打车，所以该区域也有可能是飞机场等设施所在地。

通过绘制下面的箱形图可以更明显地看出各个区域价格数据的位置和分散情况。



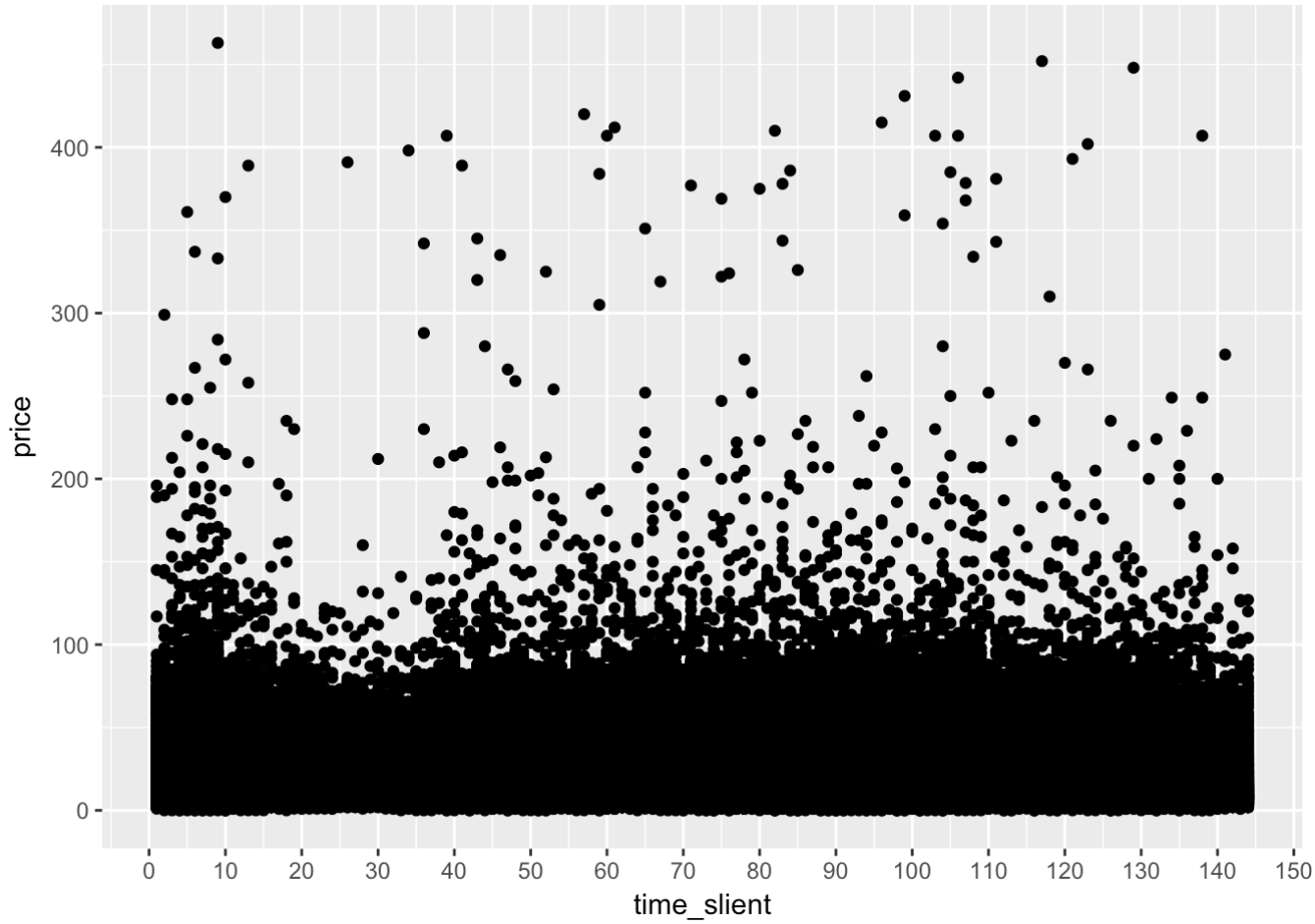
我们可以再将各个区域全部的订单数据用类似的方法展示出来，并将司机是否接单用颜色对订单数据进行区分，如下图所示浅红色为成功出行的订单，浅绿色为司机未接单的订单。



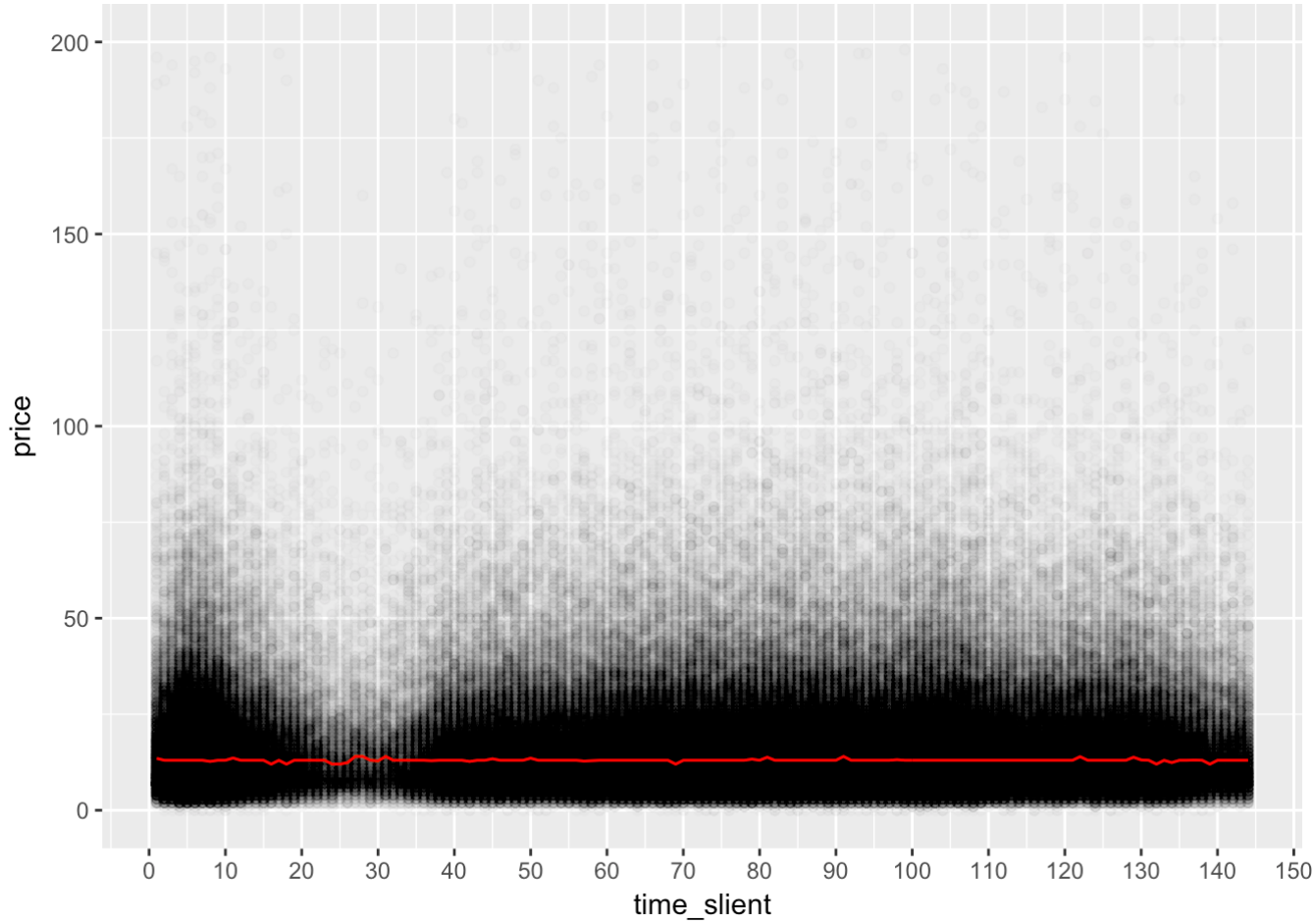
观察上图可以发现51号区域是明显地缺口数大于接单数的区域，同时7号、24号、37号和46号区域也很明显拥有很大的缺口数量，这与之前区域分析时折线图中显示的一致。

价格与时间变化分析

将订单价格按照时间片绘制散点图后得到下图，我们同样进行透明化的处理，将alpha值设置为1/70，并添加价格中位数的线条（红色线条）。



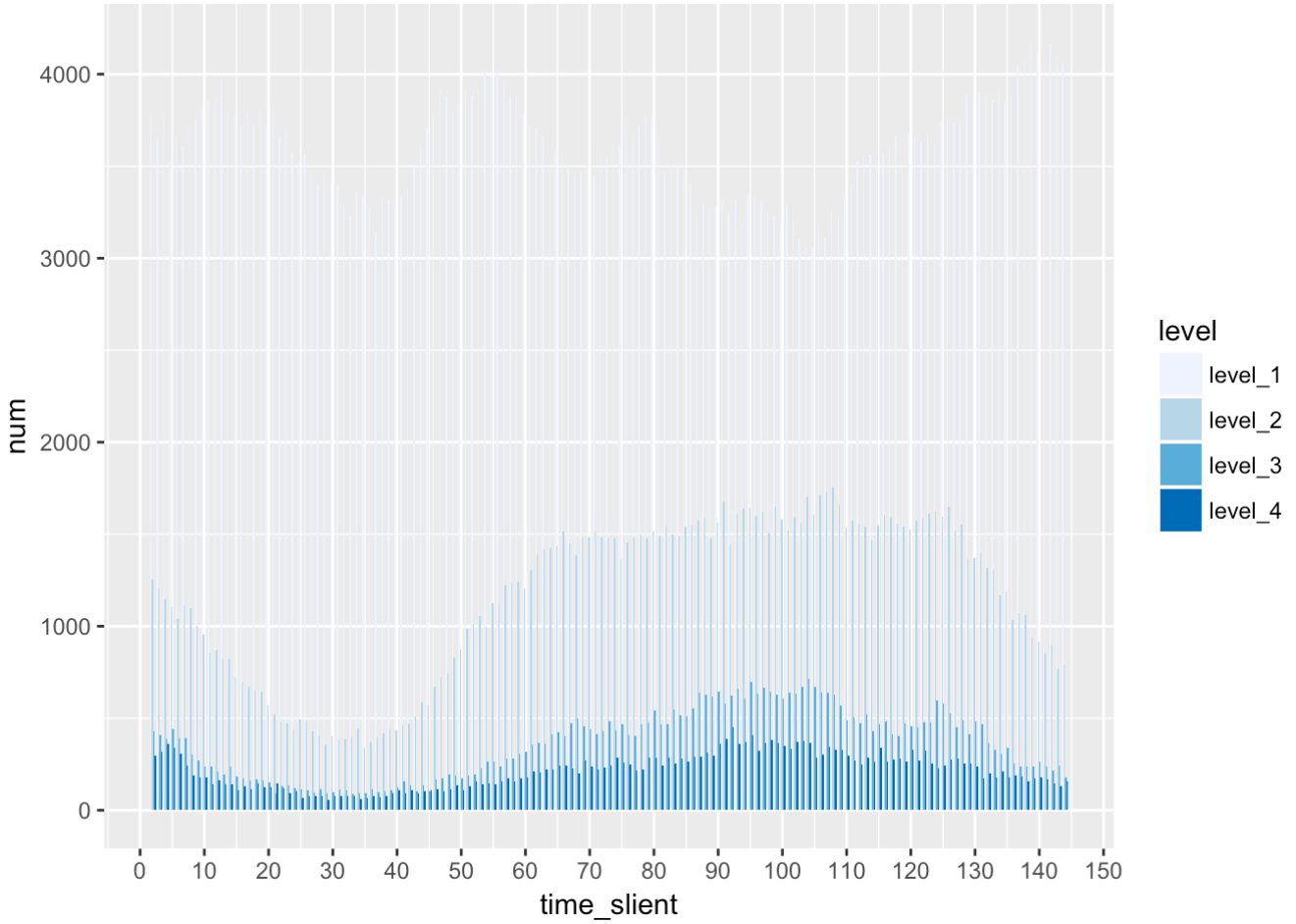
在一系列处理之后我们发现，全部区域的订单价格的变化与时间关系并不大，始终保持比较平均的水平波动很小，只有时间片20-40之间订单价格有一些明显的下降，其中一部分原因是这段时间出行订单很少，这段时间交通很通畅，而且用户一般都不会选择在凌晨3点左右的时间打车出远门。



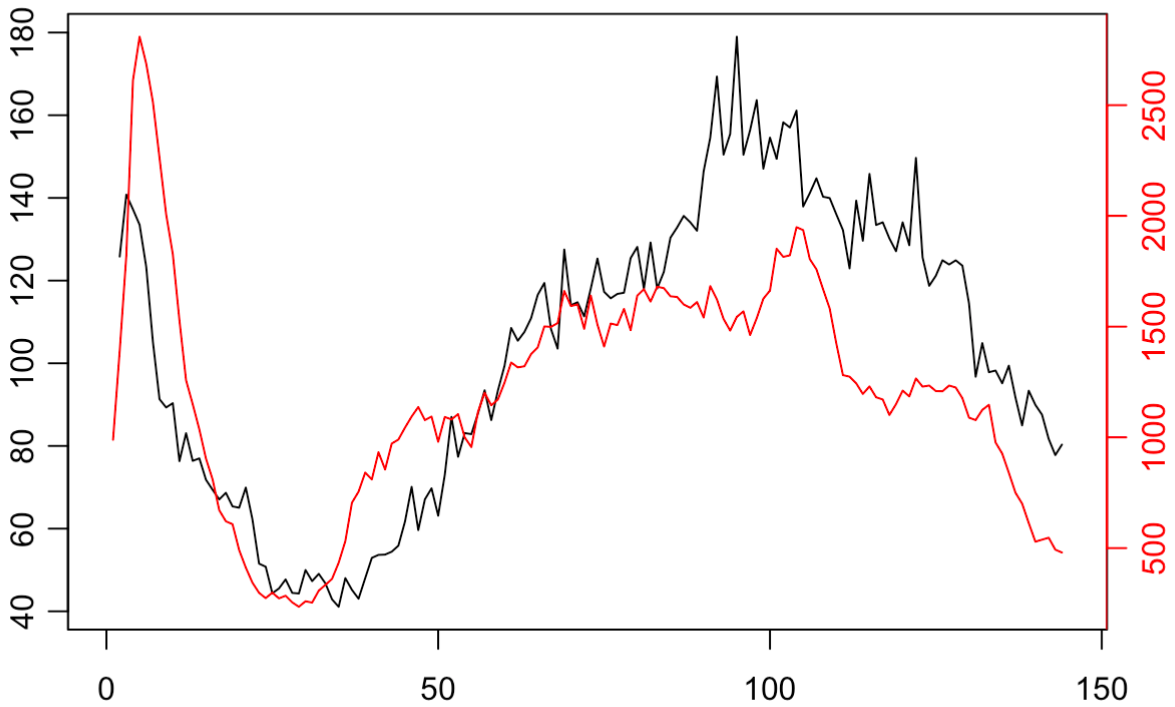
道路交通数据分析

数据集中包含每个区域的交通数据，即不同拥挤状况的道路条数，道路状况从level1到level4分为四个级别，其中level4代表最拥堵的道路状况。由于有些路段的拥堵情况采集时丢失了（官方解释），所以在不同时间片内同一区域四个分级下的路段数量总和并不相同。

通过绘制全区域的道路状况随时间变化的条形图，我们发现在任何时段，通畅道路（level1）总是占据了绝大多数，level2-level4的道路随时间变化的趋势与总订单的变化趋势很类似，拥有类似的高峰和低峰段。但观察level4数据的变化发现，订单数据在凌晨1点左右的最高峰，而同一时刻level4道路的数量并不是最高，因此推测其原因可能是跨年之后人们的打车出行需求大大增强，而较少人选择自驾出行，所以在该地区的道路上的整体车辆数并不如晚高峰时段，拥堵情况相比晚高峰时段也并没有很明显。



我们进一步将level4路段的分时图和出行订单量的分时图进行对比分析，如下图所示，红色线条表示当天所有区域的成功订单量，黑色线条为当天所有区域的level4路段数量，而通过对比我们可以更清晰地看出两者高峰时段的这种差别。



反思总结

1. 由于滴滴出行的数据量非常大，所以只选取了一天的数据进行初步分析，并且是元旦当天的数据，并非工作日的数据，因此有比较强的特殊性。下一步分析最少要选取一整周的数据进行对比分析，因为工作日数据中早高峰和晚高峰的时间点和订单变化特点都会与节假日有所不同，同时也可能存在周末对人们出行的影响。

2. 本文没有分析天气因素对订单数据的影响，因为只分析一天的数据很难观察其中的关系，需要之后进一步分析时选取更大的数据量，并运用机器学习来探究之间的联系；

3. 关于几个典型区域的相关推断只是根据1月1日这一天的数据进行的初步推断，由于没有对工作日的出行分析进行分析，因此如果要进一步验证推断的准确性，需要与其他日期的数据进行对比分析，还需分析该地域的地域属性数据；

4. 对于交通数据还可以做更深一步地分析，比如从level4数量与订单量的对比图中可以发现，订单量的第一个高峰点滞后于拥挤路段的高峰点，而在下午高峰期时的高峰点则是订单量提前于拥挤路段的高峰点，其间是否存在一定的联系仍然需要更多数据进一步深挖。