Newyork_Taxi_EDA

Roysen

12/14/2020

```
knitr::opts_chunk$set(echo = FALSE)
```

探索性数据分析-纽约出租车

背景

此数据集是关于2家纽约出租车公司采集的用户出行数据，数据描述了纽约的路况以及用户的出行习惯，通过对数据的分析和挖掘，可以展现现纽约出行市场的概况。

1.出行场景中是单人用车多，还是多人用车多。

2.出行业务中出行距离、时长和速度概况。

3.上下车时间和地点的分布概况。

4.两家出租车公司业务量级和运营模式有没有差异。

5.出行行为在星期、日和时刻因素下的是否分别具有差异。

目录：

一、加载包并读取数据

1.1加载包

1.2读取数据

```
##
## ── Column specification ────────────────────────────────────────────────────────
## cols(
##   id = col_character(),
##   vendor_id = col_double(),
##   pickup_datetime = col_datetime(format = ""),
##   dropoff_datetime = col_datetime(format = ""),
##   passenger_count = col_double(),
##   pickup_longitude = col_double(),
##   pickup_latitude = col_double(),
##   dropoff_longitude = col_double(),
##   dropoff_latitude = col_double(),
##   store_and_fwd_flag = col_character(),
##   trip_duration = col_double()
## )
```

```
##       id              vendor_id     pickup_datetime
##  Length:1458644     Min.   :1.000   Min.   :2016-01-01 00:00:17
##  Class :character   1st Qu.:1.000   1st Qu.:2016-02-17 16:46:04
##  Mode  :character   Median :2.000   Median :2016-04-01 17:19:40
##                     Mean   :1.535   Mean   :2016-04-01 10:10:24
##                     3rd Qu.:2.000   3rd Qu.:2016-05-15 03:56:08
##                     Max.   :2.000   Max.   :2016-06-30 23:59:39
##  dropoff_datetime              passenger_count pickup_longitude
##  Min.   :2016-01-01 00:03:31   Min.   :0.000   Min.   :-121.93
##  1st Qu.:2016-02-17 17:05:32   1st Qu.:1.000   1st Qu.: -73.99
##  Median :2016-04-01 17:35:12   Median :1.000   Median : -73.98
##  Mean   :2016-04-01 10:26:24   Mean   :1.665   Mean   : -73.97
##  3rd Qu.:2016-05-15 04:10:51   3rd Qu.:2.000   3rd Qu.: -73.97
##  Max.   :2016-07-01 23:02:03   Max.   :9.000   Max.   : -61.34
##  pickup_latitude dropoff_longitude dropoff_latitude store_and_fwd_flag
##  Min.   :34.36   Min.   :-121.93   Min.   :32.18    Length:1458644
##  1st Qu.:40.74   1st Qu.: -73.99   1st Qu.:40.74    Class :character
##  Median :40.75   Median : -73.98   Median :40.75    Mode  :character
##  Mean   :40.75   Mean   : -73.97   Mean   :40.75
##  3rd Qu.:40.77   3rd Qu.: -73.96   3rd Qu.:40.77
##  Max.   :51.88   Max.   : -61.34   Max.   :43.92
##  trip_duration
##  Min.   :      1
##  1st Qu.:    397
##  Median :    662
##  Mean   :    959
##  3rd Qu.:   1075
##  Max.   :3526282
```

1.3变量注释

序号	变量	注释
01	id	ID
02	vendor_id	出租车公司id
03	pickup_datetime	上车时间
04	dropoff_datetime	下车时间
05	passenger_count	乘客人数
06	pickup_longitude	上车经度
07	pickup_latitude	上车纬度
08	dropoff_longitude	下车经度
09	dropoff_latitude	下车维度
10	store_and_fwd_flag	是否分享行程记录 Y=是，N= 不
11	trip_duration	旅行时间（秒)

数据共有观测145万多行，变量11个,是一个非常大的数据集，抽取一个10000行的样本进行分析。

从11个变量的数据纬度来看，主要是关于纽约出租车用户出行时间、出行时长、上下车地点、出行人数，是否分享行程记录的数据。

1.4检查缺失值

```
## # A tibble: 1 x 11
##      id vendor_id pickup_datetime dropoff_datetime passenger_count
##   <int>     <int>           <int>            <int>           <int>
## 1     0         0               0                0               0
## # … with 6 more variables: pickup_longitude <int>, pickup_latitude <int>,
## #   dropoff_longitude <int>, dropoff_latitude <int>, store_and_fwd_flag <int>,
## #   trip_duration <int>
```

各变量均没有缺失值

1.5按行随机抽样10000人

数据清洗

2.1提取经纬度变量，计算行驶距离（km），创建速度变量，单位（km/h），将日期单位改为小时

2.2日期格式转换，将vendor_id转换因子

```
## # A tibble: 10,000 x 13
##    id    vendor_id pickup_datetime     dropoff_datetime    passenger_count
##    <chr> <fct>     <dttm>              <dttm>                        <dbl>
##  1 id26… 1         2016-03-08 19:50:57 2016-03-08 20:16:22               1
##  2 id23… 1         2016-04-02 18:23:41 2016-04-02 18:32:20               2
##  3 id16… 2         2016-01-14 11:36:37 2016-01-14 11:53:18               5
##  4 id30… 1         2016-06-08 13:36:04 2016-06-08 13:40:53               1
##  5 id09… 2         2016-03-18 03:39:47 2016-03-18 03:52:46               1
##  6 id02… 1         2016-06-29 22:33:36 2016-06-29 23:10:26               1
##  7 id11… 1         2016-05-20 14:19:17 2016-05-20 14:21:35               1
##  8 id21… 2         2016-02-23 05:57:36 2016-02-23 06:12:57               1
##  9 id11… 1         2016-05-01 02:24:13 2016-05-01 02:29:51               1
## 10 id15… 2         2016-05-07 03:21:26 2016-05-07 03:25:31               1
## # … with 9,990 more rows, and 8 more variables: pickup_longitude <dbl>,
## #   pickup_latitude <dbl>, dropoff_longitude <dbl>, dropoff_latitude <dbl>,
## #   store_and_fwd_flag <fct>, trip_duration <dbl>, distance <dbl>, speed <dbl>
```

2.3排序并查找异常值

```
## # A tibble: 10,000 x 3
##    distance speed trip_duration
##       <dbl> <dbl>         <dbl>
##  1   0.315  283.        0.00111
##  2   1.01    84.9       0.0119
##  3   0.0931  83.8       0.00111
##  4   0.183   73.3       0.0025
##  5   0.130   66.8       0.00194
##  6  10.2     66.6       0.153
##  7  15.8     64.8       0.243
##  8  20.0     60.2       0.332
##  9   2.33    59.0       0.0394
## 10   6.51    57.1       0.114
## # … with 9,990 more rows
```

```
## # A tibble: 10,000 x 3
##    distance  speed trip_duration
##       <dbl>  <dbl>         <dbl>
##  1   0.989  0.0413         23.9
##  2   0.968  0.0405         23.9
##  3   1.33   0.0560         23.8
##  4   2.83   0.119          23.8
##  5   8.48   0.358          23.7
##  6   1.12   0.0473         23.6
##  7   5.44   0.232          23.5
##  8   0.260  0.0711          3.66
##  9   0.0432 0.0135          3.20
## 10   9.96   5.21           1.91
## # … with 9,990 more rows
```

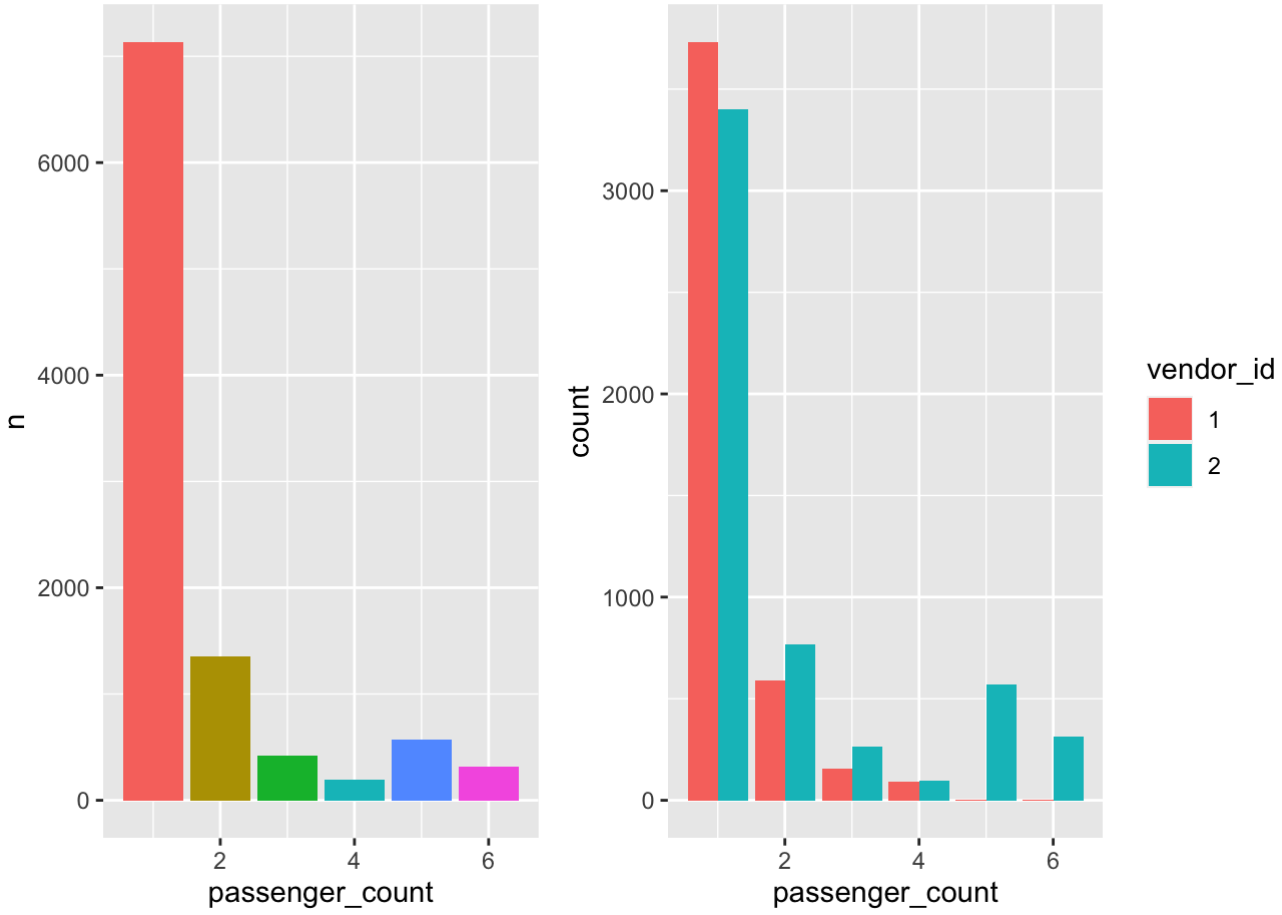
速度大于280km/h几乎不可能，行驶时长超过23h可能性也不大。因而删除掉这部分数据。

2.4剔除异常值

三、总体变量统计以及可视化

3.1总体乘客人数分布状况

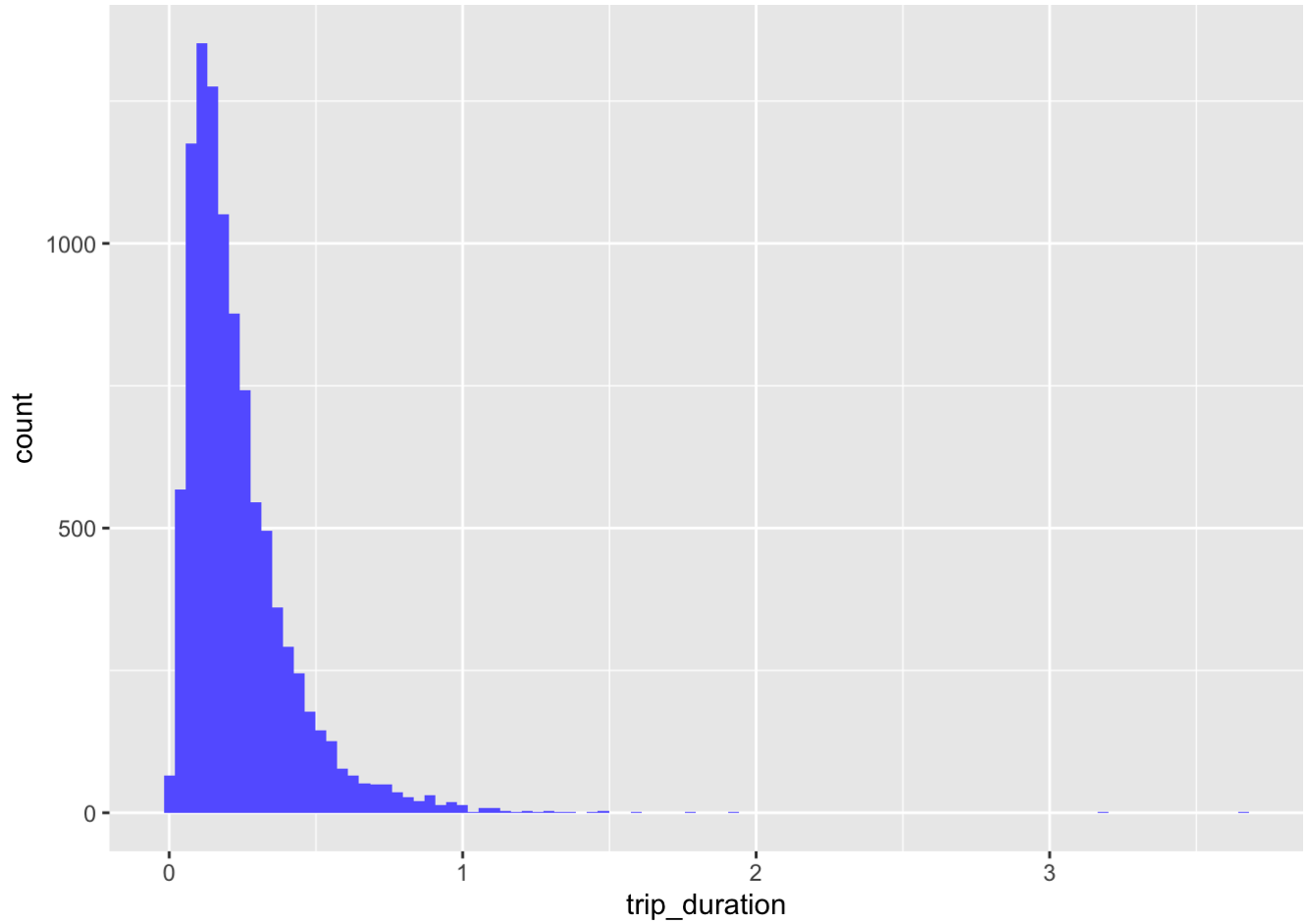
```
## # A tibble: 6 x 2
## # Groups:   passenger_count [6]
##   passenger_count     n
##             <dbl> <int>
## 1               1  7133
## 2               2  1356
## 3               3   423
## 4               4   190
## 5               5   575
## 6               6   315
```



单人出行人数人数7133人，占比71.4%。2个乘客的出行人数约1356人，占比13.6%。单人出行场景是纽约出行业务的重点。

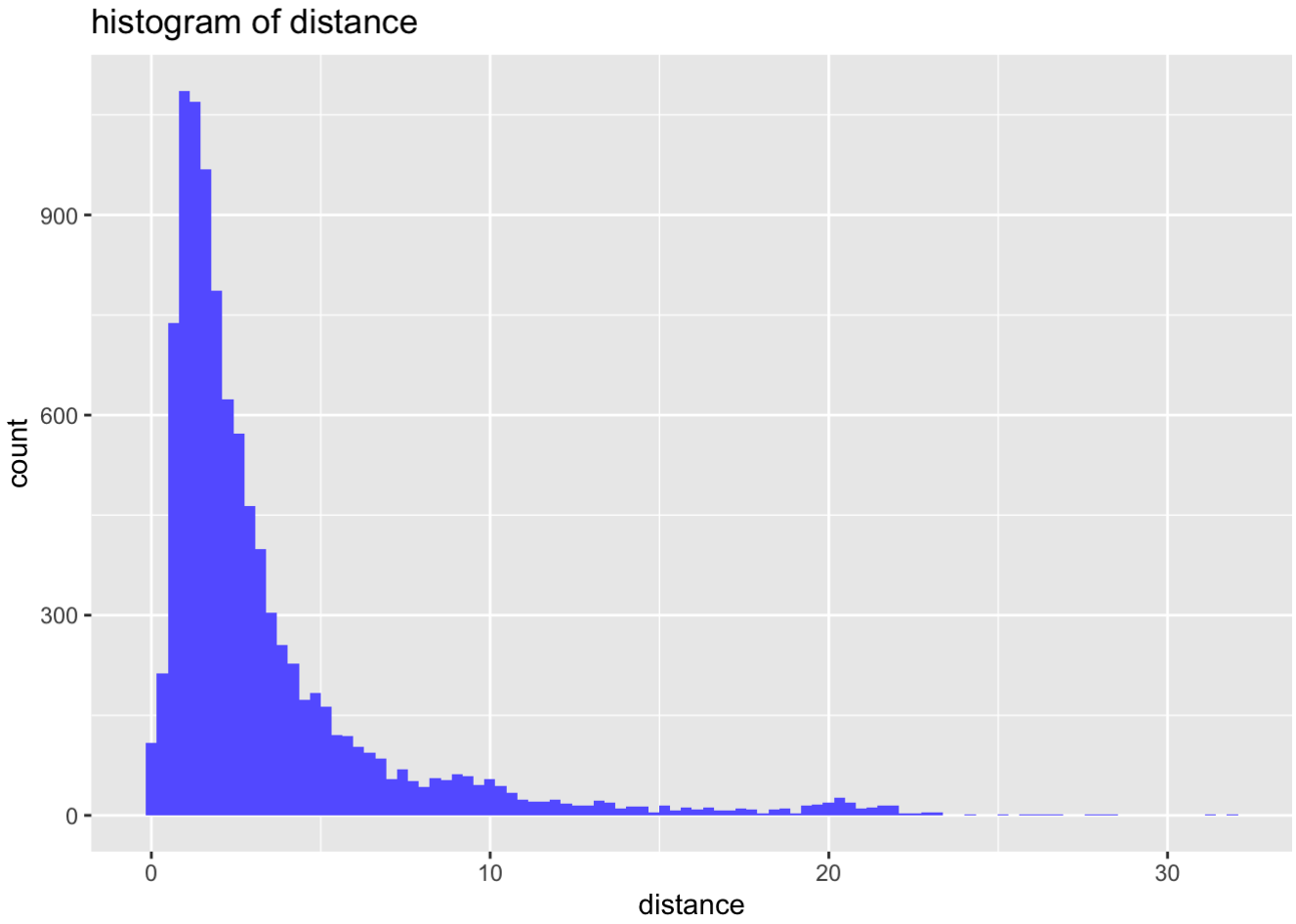
两家公司业务量级趋同，区别在于，出租车公司2做多人出行业务。

3.2总体出行时长分布



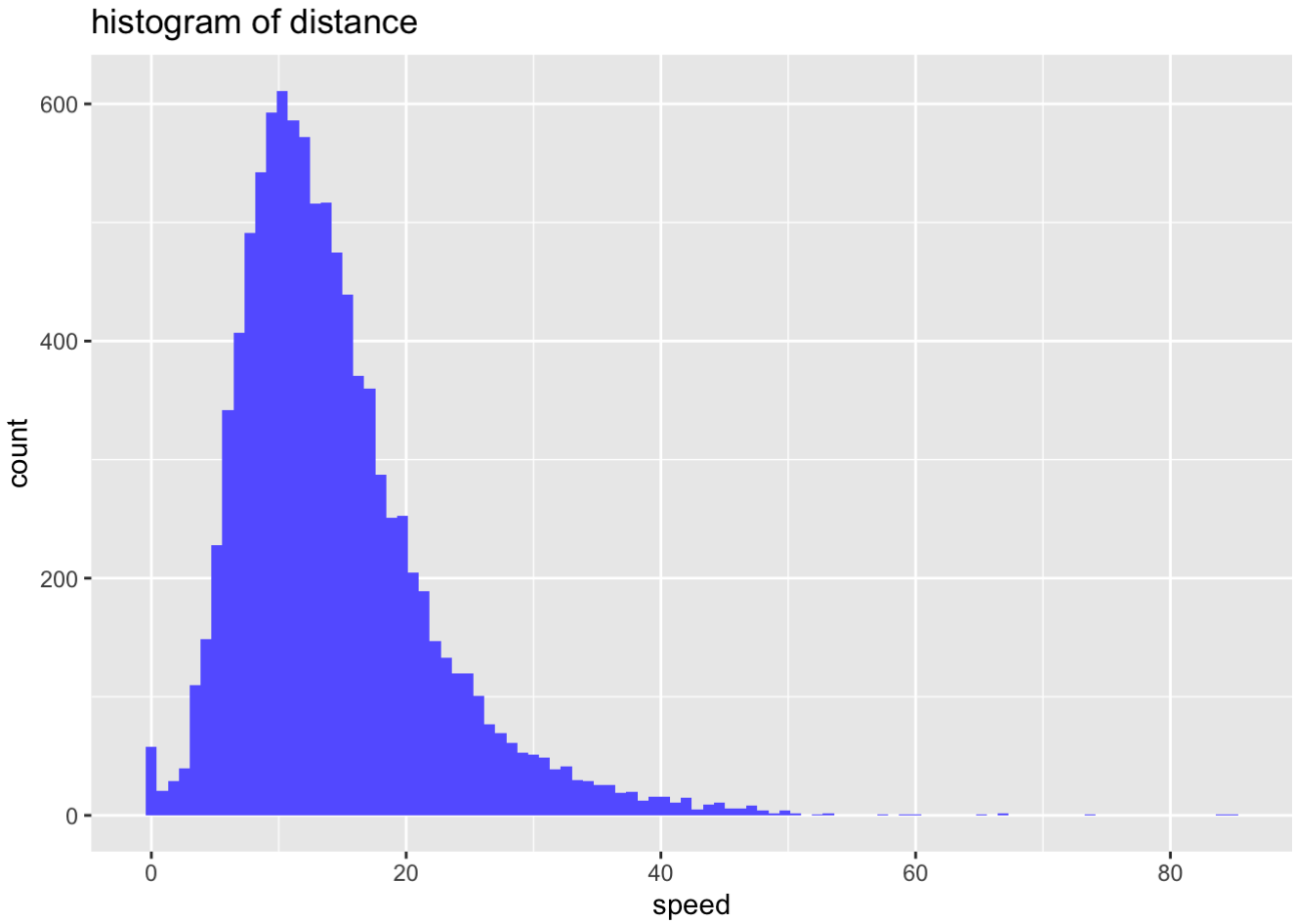
平均出行时间0.2小时，出行时长超过1小时的人数非常之少。基本都在30分钟以内。

3.3总体出行距离分布



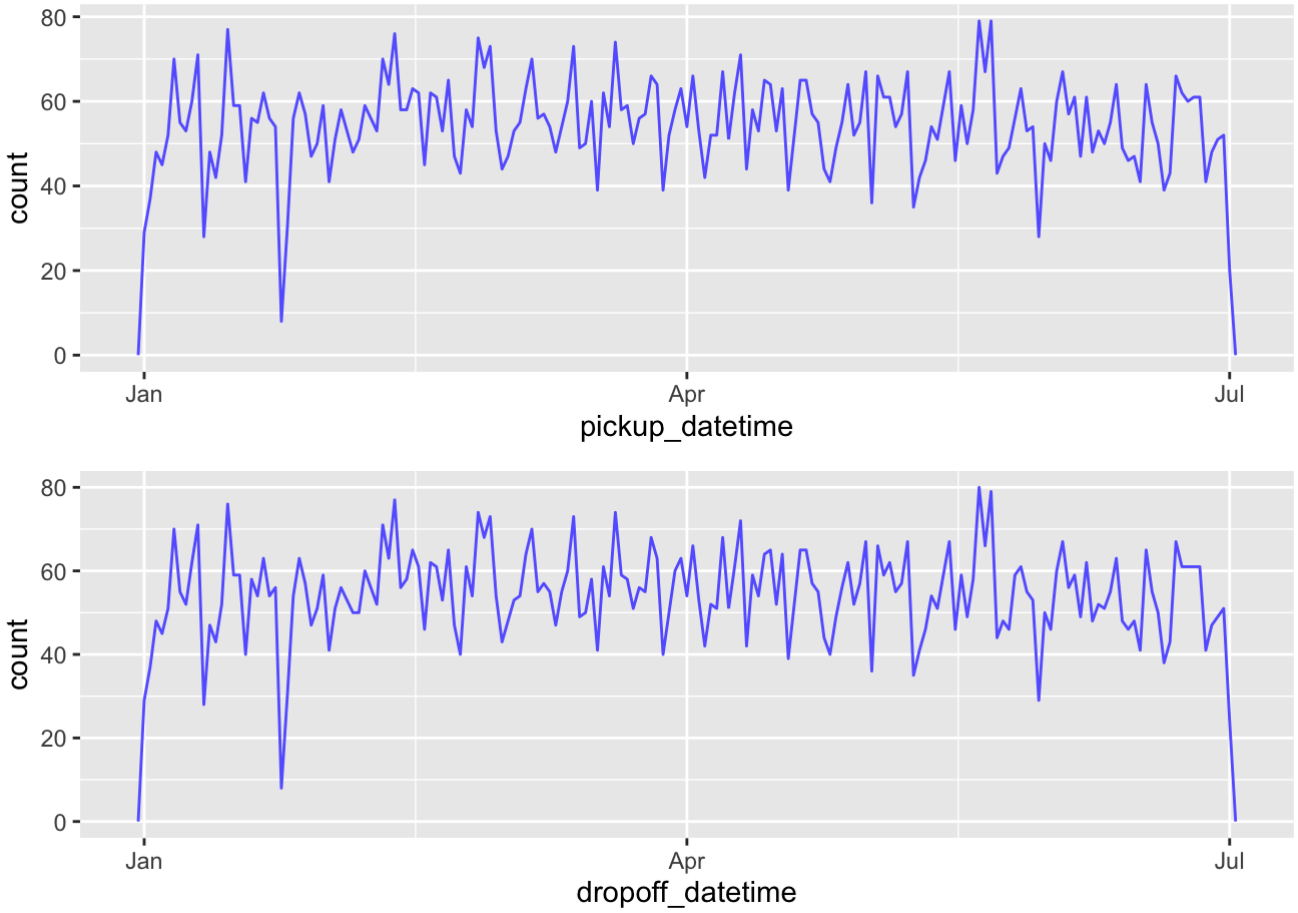
平均行使距离为3.42km，出行距离超过10km的人数较少。短途出行代表了纽约出行市场的主要需求。

3.4总体出行速度状况



速度集中在15km/h，超过25km/h出行情况不多。

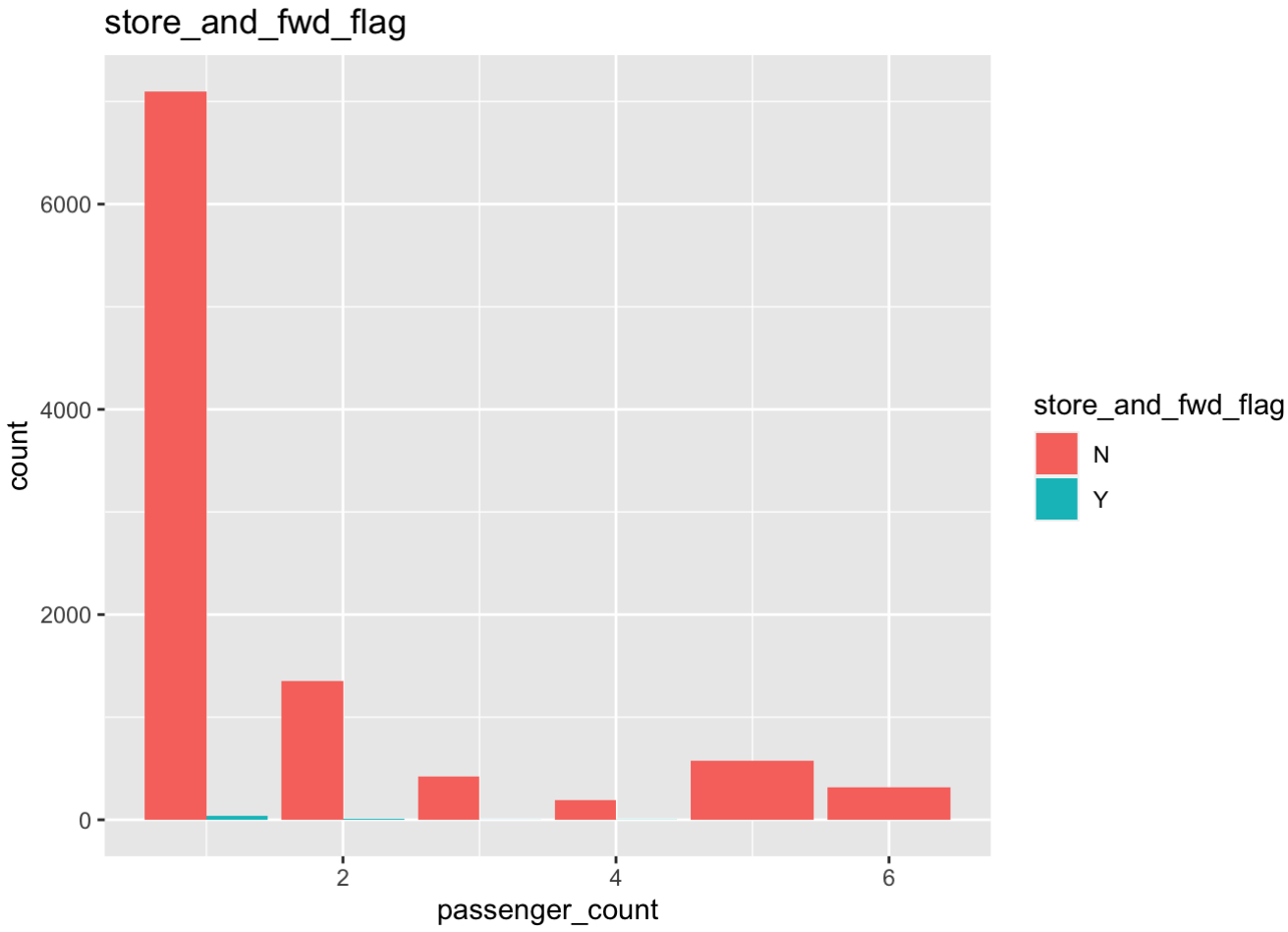
3.5上下车时间分布情况



6个月中上、下车时间频率分布的波动趋同，问题是1月底-2月初，打车人数锐减。

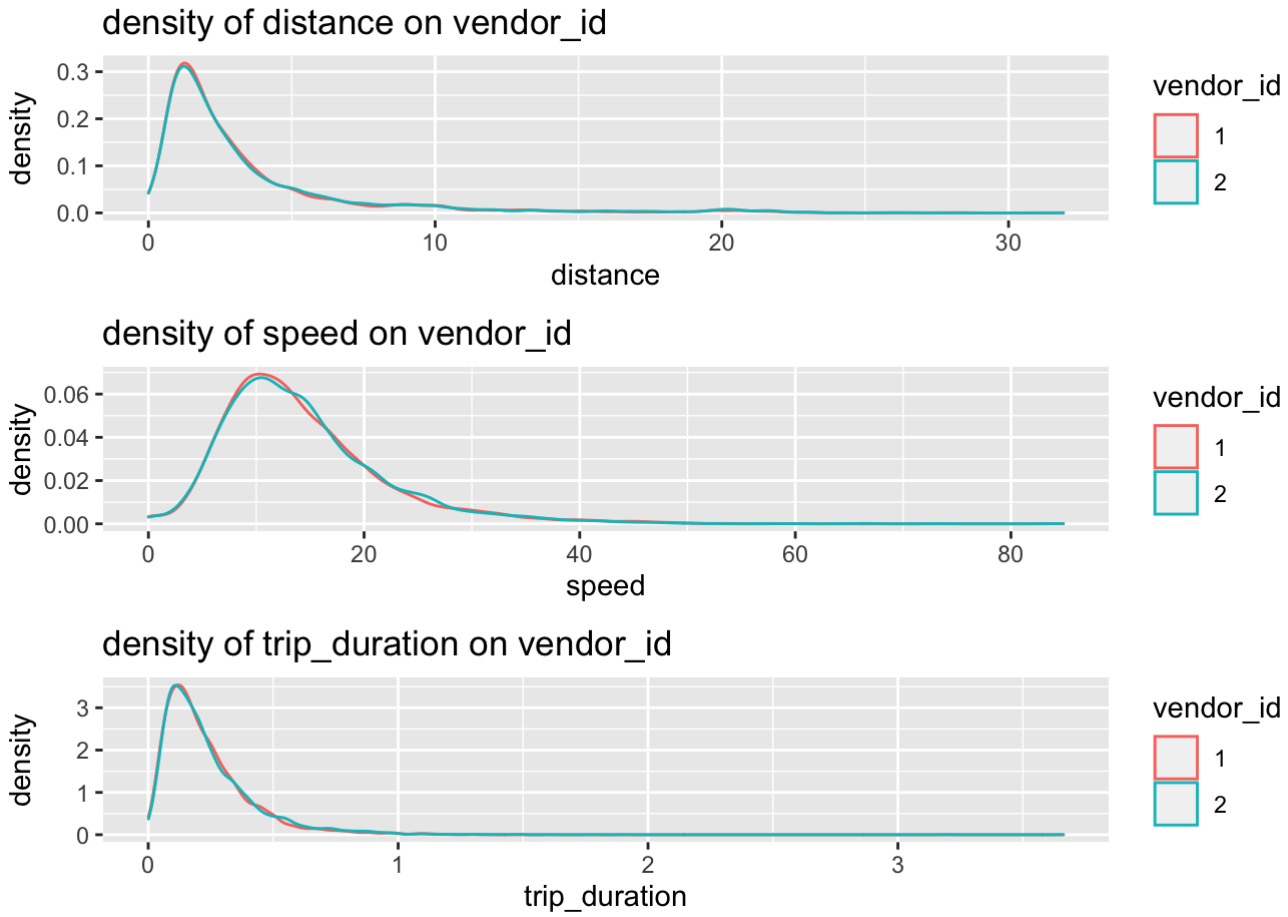
3.6行程记录分享状况

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1    42
```



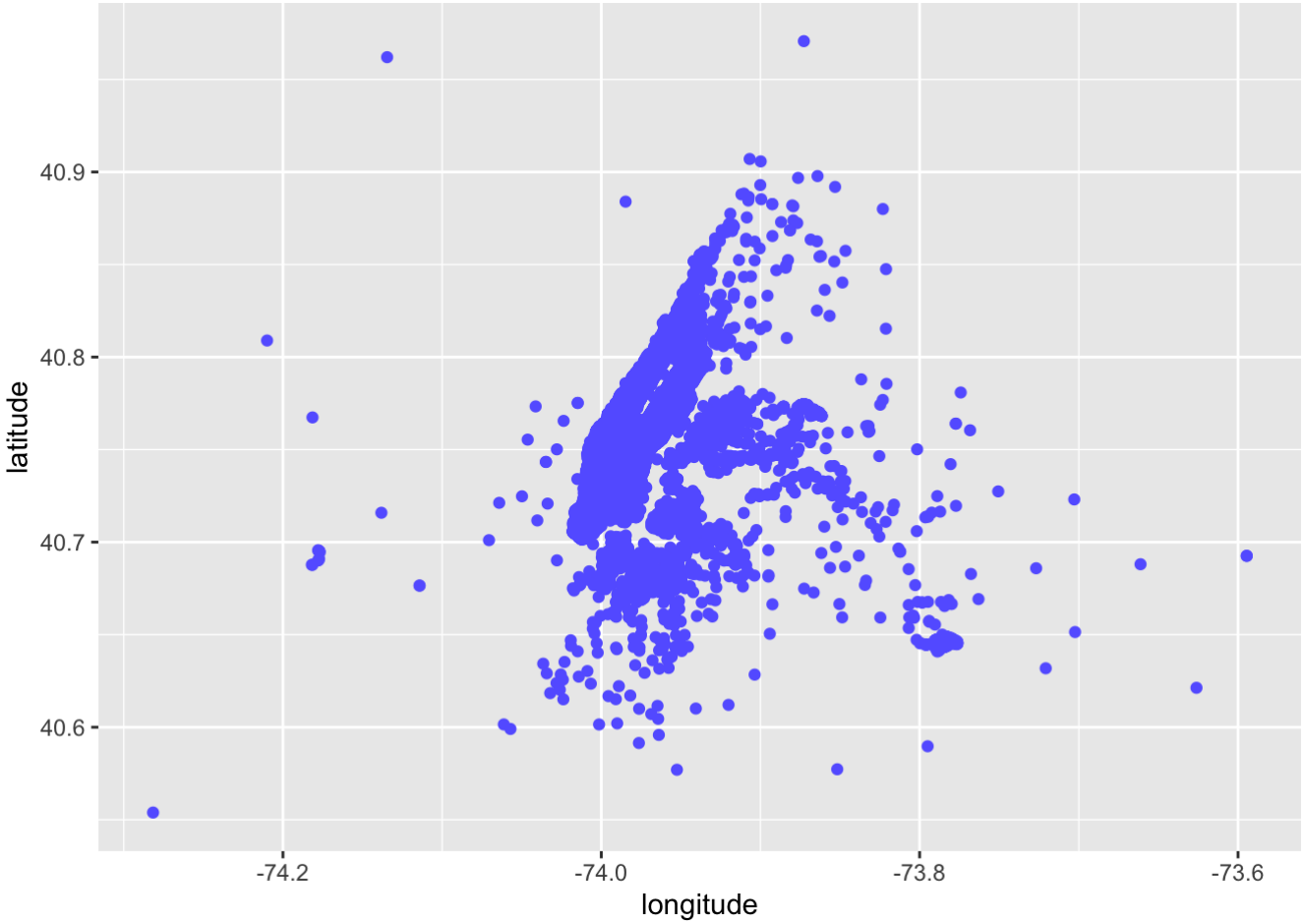
分享行程的人数只有42人，占比0.4%。5人和6人乘客没有分享行程的情况。

3.7对出租车公司进行分组，查看出行距离、出行速度、出行时间是否有差异



重叠非常严重，可以看出两家出租车公司在出行距离、速度和旅行时长方面差异不大，短途用车和拥堵问题是一个共性的情况。

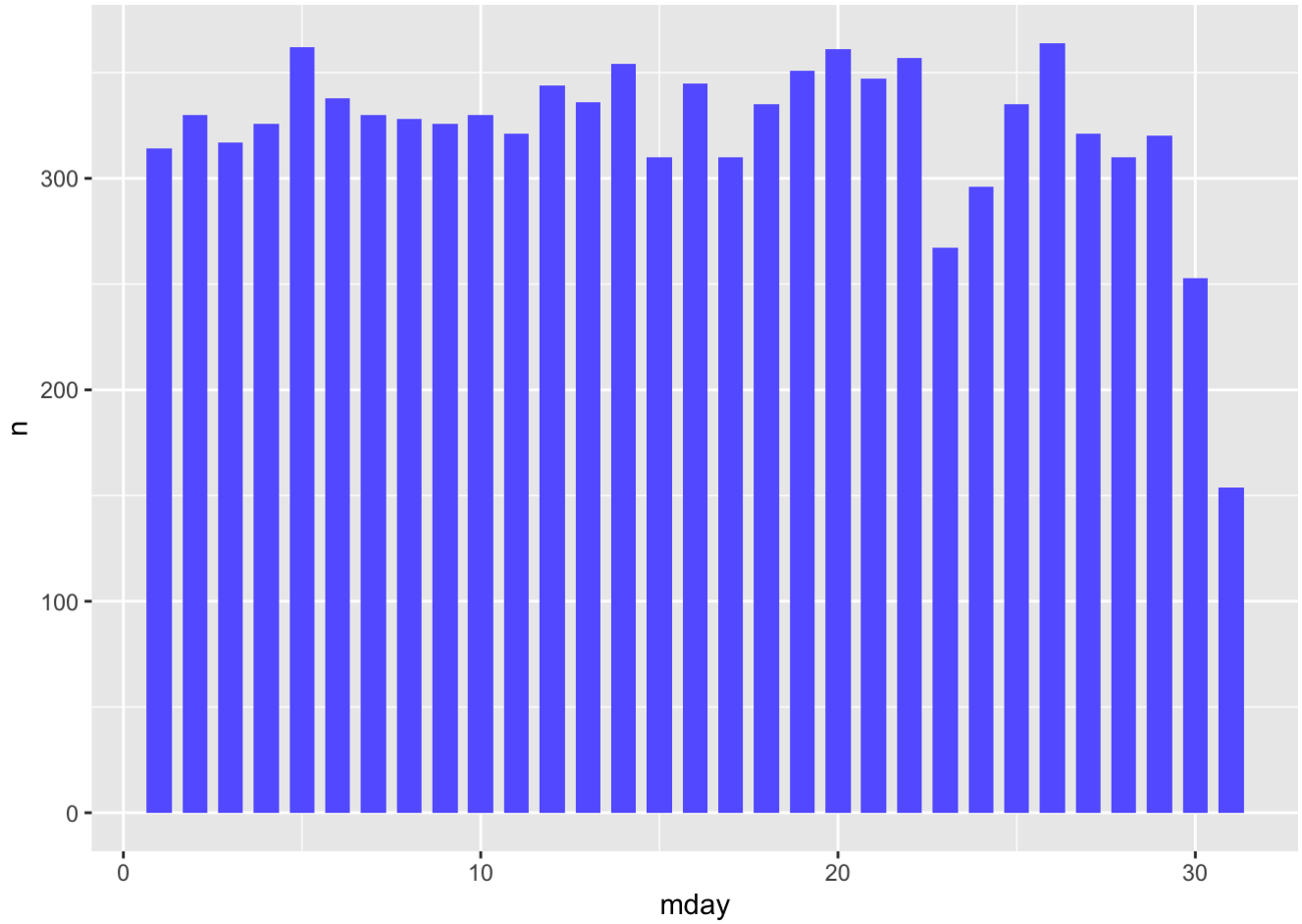
3.8出行地理位置可视化



出行位置最集中的区域是西经74-西经73.9，北纬40.65-北纬40.85。出行方向表现为东北-西南走向为主，南北跨度高于东西跨度。

四、按照月份、星期、时刻细化统计

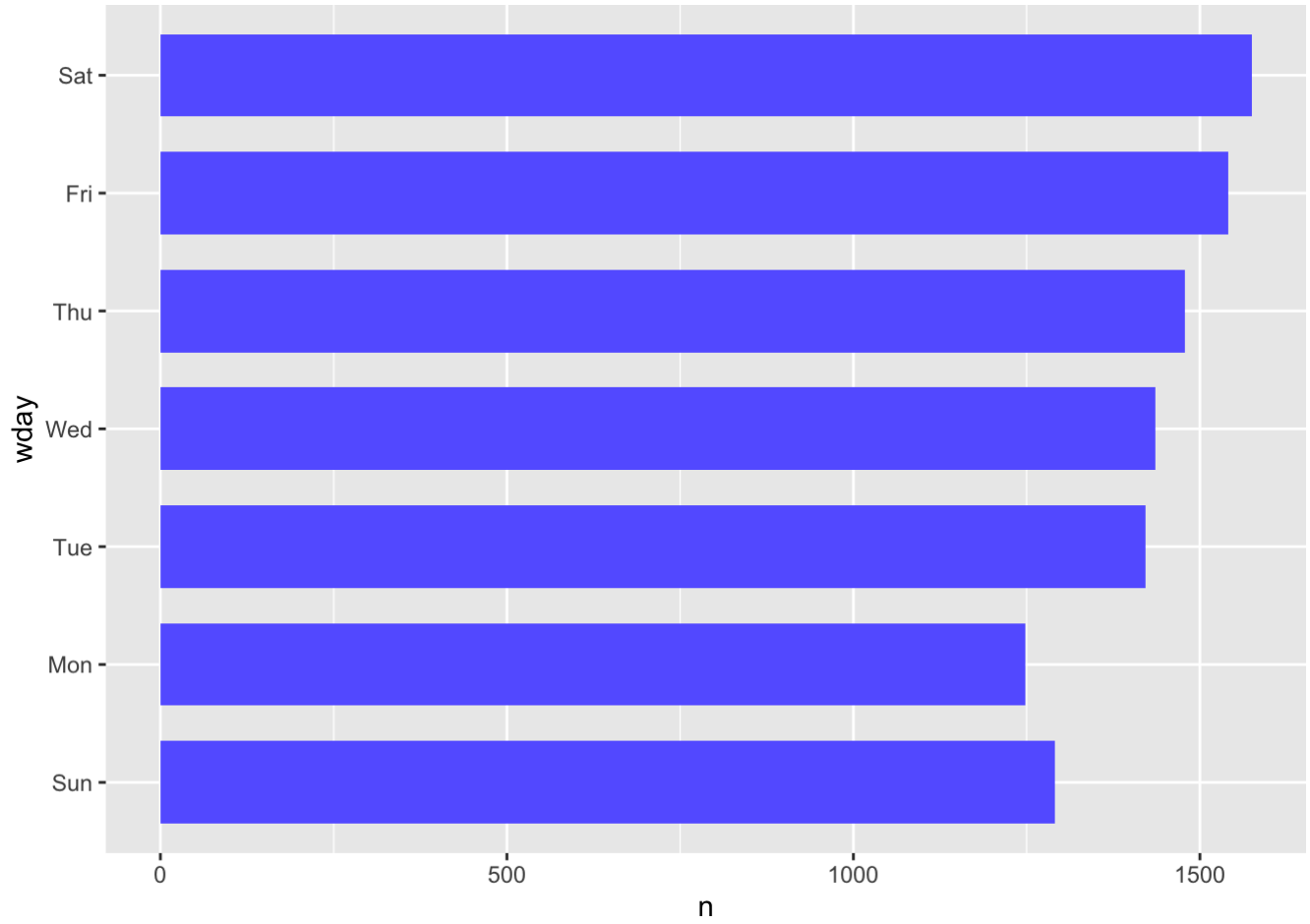
4.1 1-31日累计出行人数



该图显示：6个月的出行数据中，31号这一天用车人数最少，只有150人，减少了50%，一个原因是并非每月都有31号。

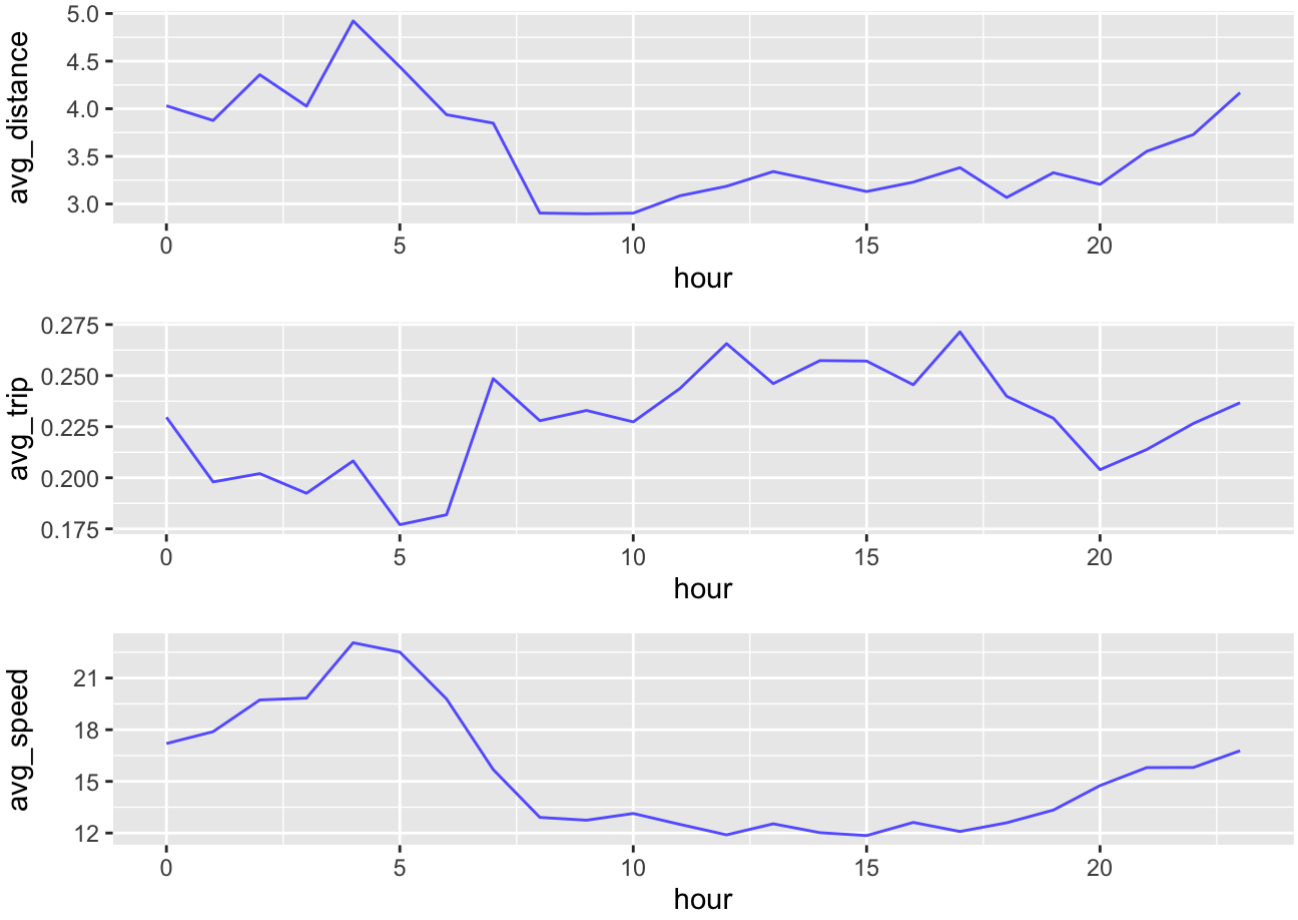
23日和30日出行人数相对其他日少了50人/每日。

4.2 6个月中周六至周日出行人数分布



图形表明，6个月中周五和周六出行人数最多，超过了1500人，周日和周一出行人数约为1250人，其他接近1500人。

4.3 6个月中0:00-24:00的平均出行时长、出行距离、出行速度

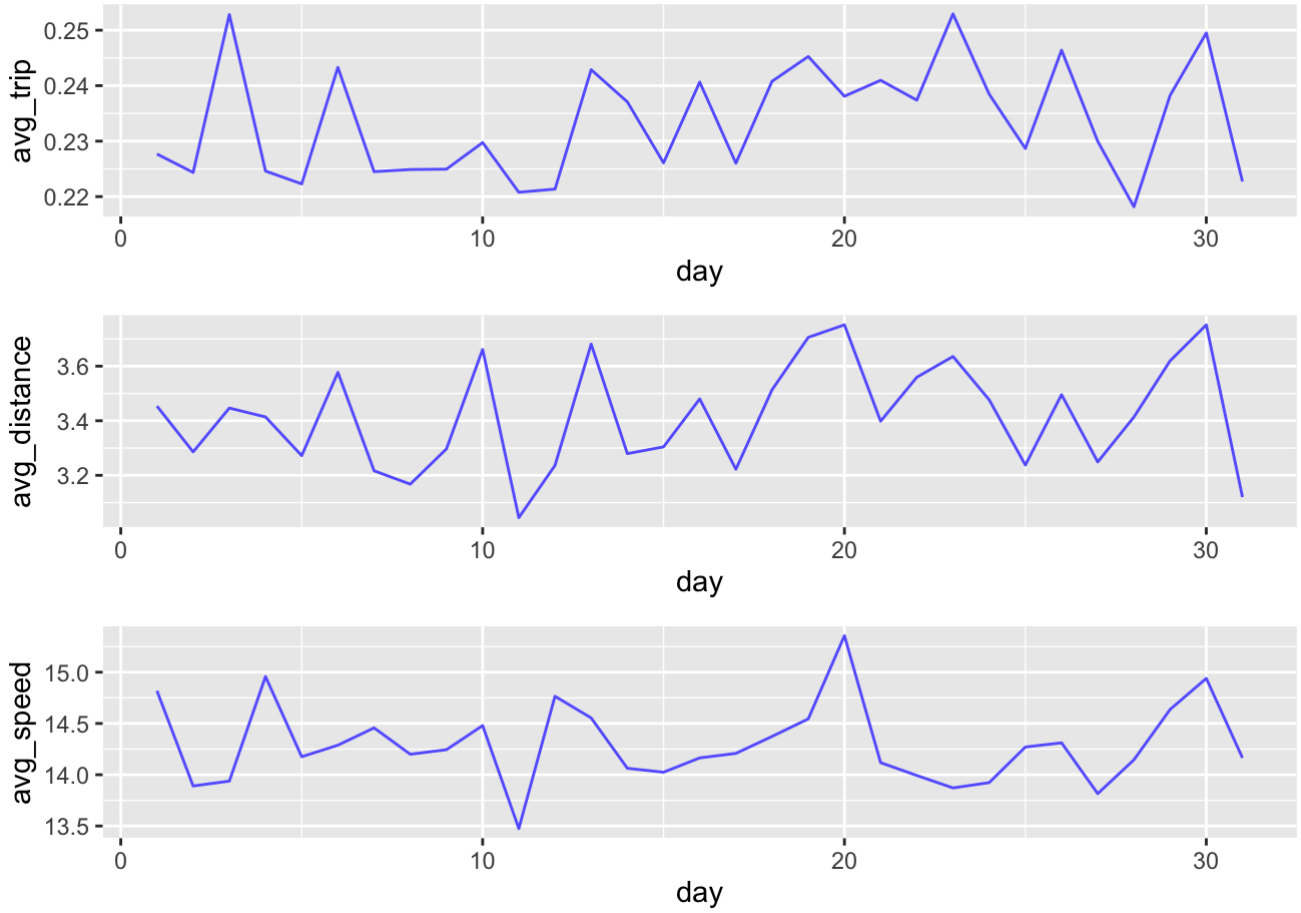


0:00-5:00，是平均出行距离最长的时间段，约为（4-5km）。7:30-10:00，平均行驶距离不足3km。

0:00-5:00，平均出行时长不超过0.2小时。7:30-17:30，平均出行时长0.25小时，是一日中，出行时长最长的时间段。

0:00-5:00，平均出行速度由17km/h增加到23km/h，是一天中速度最快的时刻。7:30-20:00，平均出行速度只有13km/h左右，这个时间段是一天中最拥堵的时间。

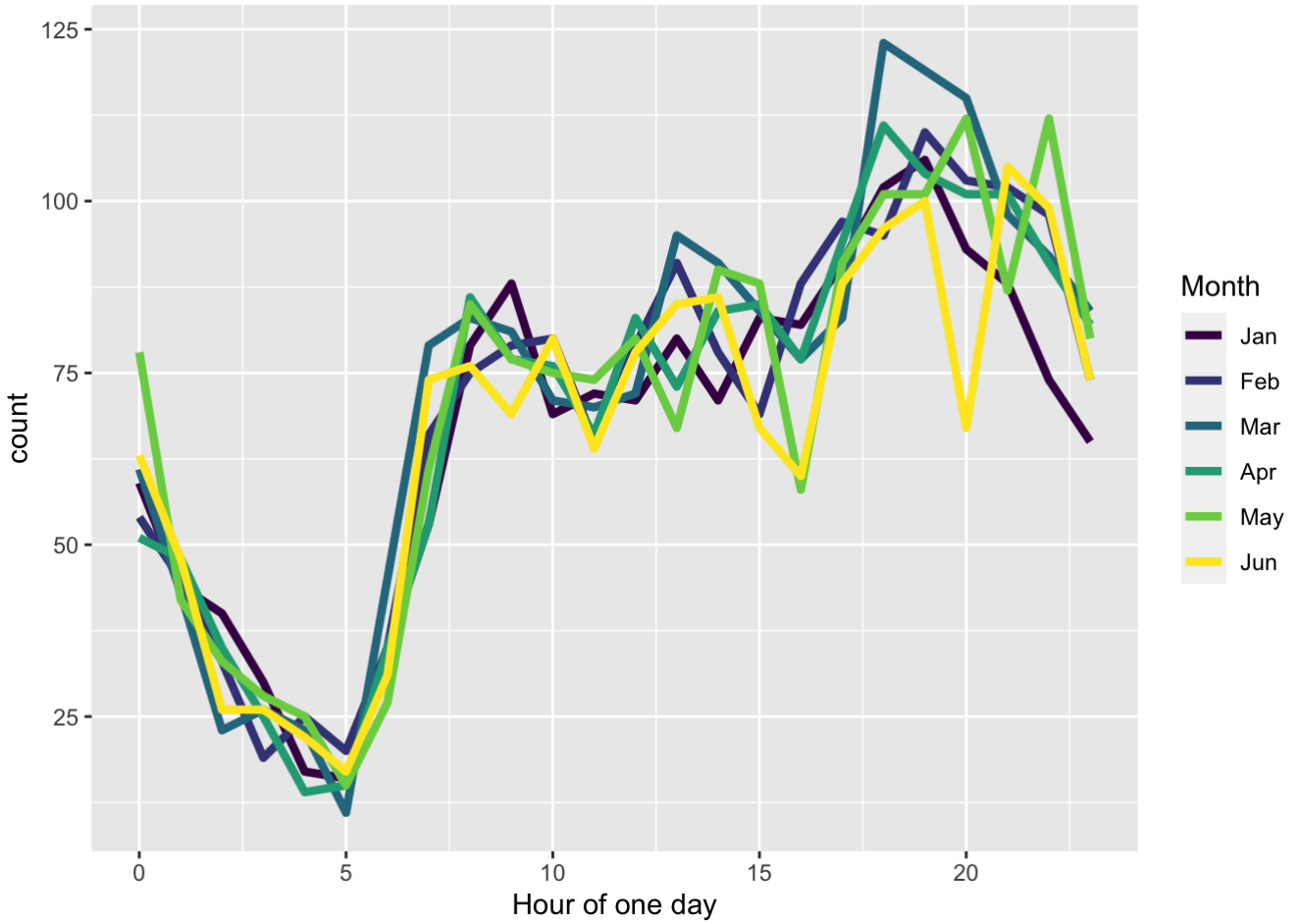
4.4 6个月中1-31日，各天的平均出行时长、出行距离、出行速度



累计每天平均出行时长、距离和速度，波动范围不大。

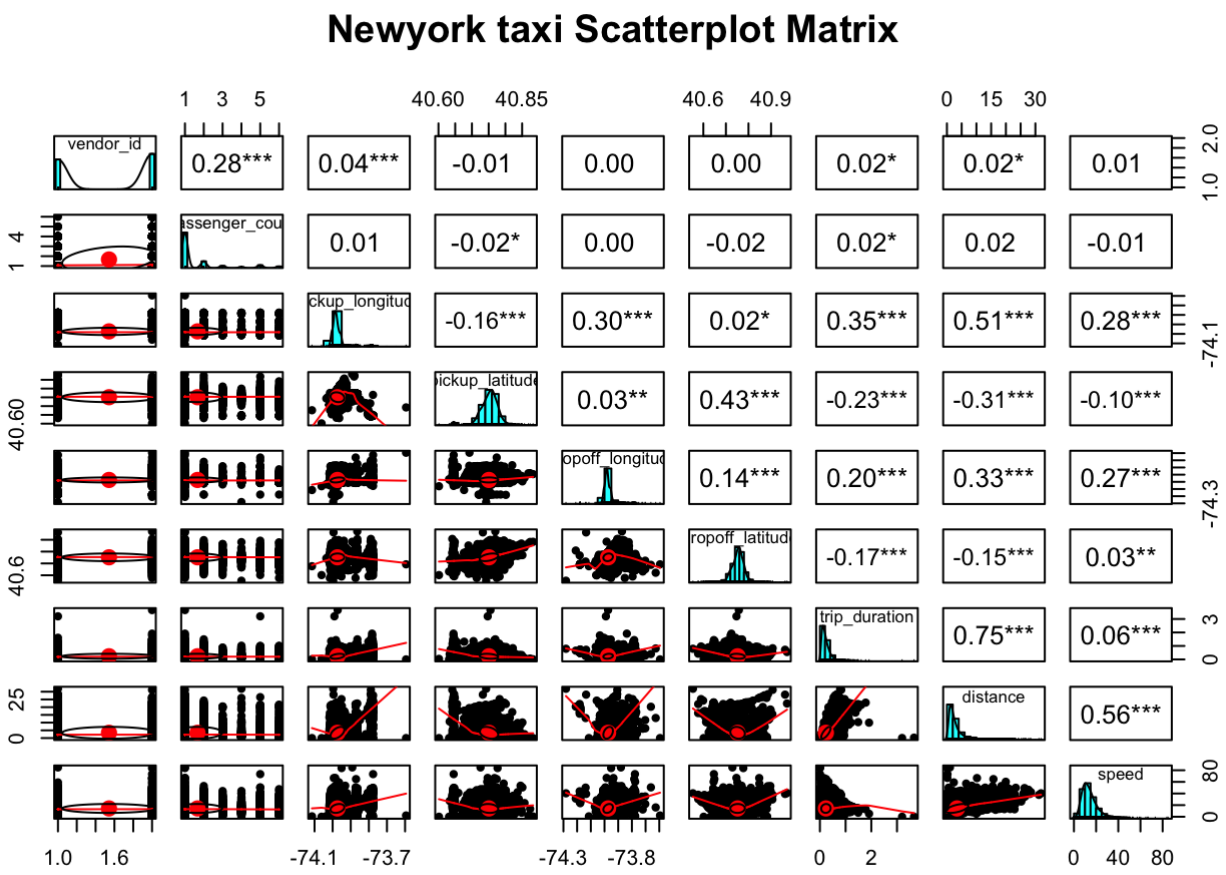
每个月的11号前后是出行距离、速度和时长数相对是最小的一天。

4.5 每月0:00-24:00出行人数在月份中的表现情况



0:00-7:30,每个月出行人数均小于50人，属于打车低频时间段。其余时间段，每个月出行人数都在90人左右浮动。

五、相关性分析



行驶距离和上车经度表现为正相关系，相关系数为0.51。上下车维度呈现一定的正相关关系，相关系数为o.43出行距离与出行度表现为正相关，相关系数为0.56。

六、结论

1.纽约单人出行人数人数7133人，占比71.4%，单人出行场景是纽约出行业务的重点。多人出行市场主要是出租车公司2在做，目前规模不算大。多人时长用户需求暂时处于平稳状态。

2.纽约的出行市场以短途（5km以内）为主，出现时间集中在0.5小时以内，平均速度15km/h。

3.1月底-2月初，打车人数锐减，谷歌显示因为遭遇了暴风雪天气。

4.分享行程的人数只有42人，占比0.4%。5人和6人乘客没有分享行程的情况。

5.出行位置，南北跨度高于东西跨度。出行方向表现为东北-西南走向为主

6. 6个月的出行数据中，31号这一天用车人数最少，只有150人，减少了50%,23日和30日出行人数相对其他日少了50人/每日。

7. 0:00-5:00，为出行低频时段，人数少，速度快。7:30-10:00，是出行高频时段，平均行驶距离不足3km，平均出行时长0.25小时，是最为拥堵的时刻。

8.累计每天平均出行时长、距离和速度，波动范围不大，每月的11日是出行距离、速度和时长数相对是最小的一天。