交通运输数据技术 作业四

——出租车订单数据可视化

2251140 范程

一、 总述

本次作业四是和上海市 2018 年 4 月 18 日的出租车运营数据息息相关的，在课程对标方面，主要与数据的统计、分析和可视化相对应。本次作业主要分成：展现出租车的订单概况和对运营订单数据进行地理可视化，主要的处理对象是订单的 OD 数据。通过数据分析和可视化，可以对上海市出租车订单的时空分布进行研究，从而深入观察市民出行行为的特征，具有研究意义。在数据处理和可视化方面，我选择了 R 语言作为工具，以下部分是对我提交的 R 语言代码的相关说明和对所绘制图像的分析。

二、 数据分析和图像绘制

（一） 数据预处理

对.csv 文件另存为.xlsx 文件，并通过 read_excel()函数读取数据之后，我们得到了 2018 年 4 月 18 日出租车的订单数据 data，首先要对数据进行预处理。

在老师提供的作业任务和提示中，并没有数据预处理的部分，但这不代表我们就可以忽视对已有的出租车订单数据的预处理。数据预处理可以去除冗余、不符合要求的数据，从而维护相关数据的有效性，减小无效数据对结果的影响。

在本次预处理中，结合出租车订单自身的特征，我归纳出了以下几项：

① 查找是否有缺失值。我选择了用 na.omit(data)来去除缺失值。但是在用 is.na(data)查找之后，返回值为 False，说明数据是完整的，可以进行下一步，并不存在如只有上车时间 stime，没有下车时间 etime 的情况。

② 删去重复值。我利用了 data <- data[!duplicated(data),]来去重，减少重复数据的影响。运行之后，data 中的原始数据有 240481 条，去重之后发现还是 240481 条。

③ 通过时间筛选。毫无疑问订单开始时间 stime 应早于结束时间 etime，因此我使用了 data <- subset(data,etime > stime)来保留合理部分，这次剩下了 240391 条数据。

④ 通过日期筛选。我们分析的是 2018-4-18 内产生的订单数据，因此用 data <- subset(data,stime >= start_date)来保留出发日期是 4-18 的数据，本次运行没有筛出数据。

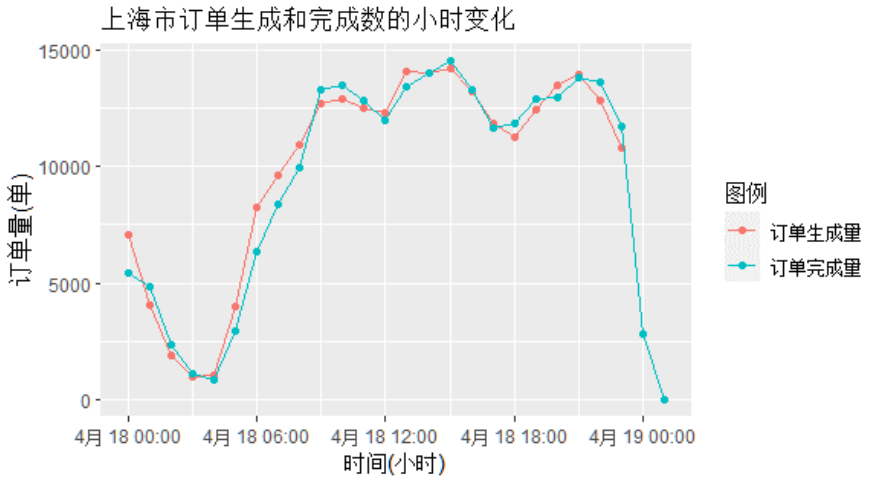
⑤ 通过经纬度筛选。订单结束的经度 gcj_e_lng 和纬度 gcj_e_lat 起码要有一个和开始

时不一样，因此我使用 data <- subset(data,gcj_s_lng != gcj_e_lng | gcj_s_lat != gcj_e_lat) 来保留正常行驶的数据。经过上面几步简单预处理之后，data 中共剩下 240228 条数据，供下面的可视化使用。

（二） 订单概况可视化

1. 出租车订单时间分布

在对 data 中的数据用 group_by()和 summarise()按照小时分类并聚集在一起，统计完数量之后，我在同一张图上画出了出租车订单出发量和完成量的折线图，主要使用了 ggplot()函数和 geom_line()和 geom_ppoint()来进行数据的添加。最终得到的折线图如下：



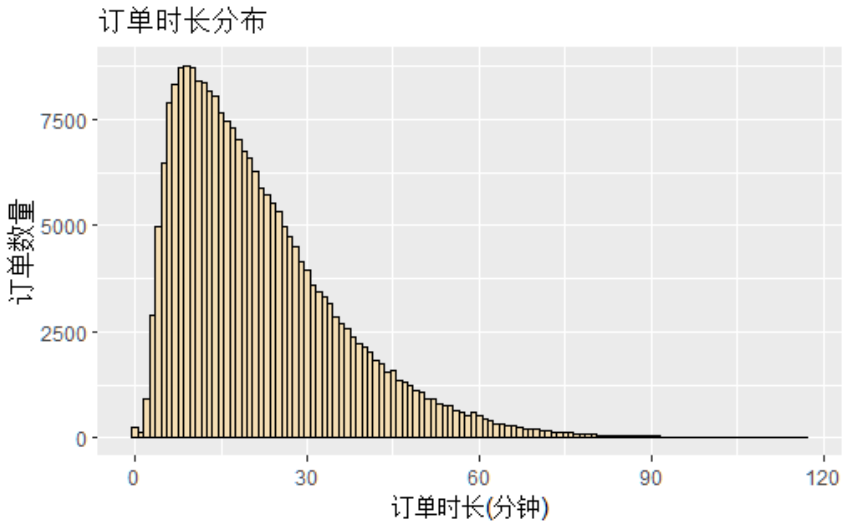
从上图看出，出租车的订单生成量和完成量在时间上具有很强的相似性，二者变化趋势相同，甚至在数量上也极为相近。其中，在凌晨达到极小值之后，伴随早高峰的白天的到来迅速增加，在白天有一个小幅度的下跌后，又在晚高峰时期上升，呈现出多峰状，在午夜到来时期迅速降低，完成量达到最小值。这能管中窥豹地反映出居民出行的时间特征，但空间特征，如起讫点等还需要地理可视化部分的工作来实现。

2. 出租车订单时长分布

我采用如下方式，调用 difftime()函数来计算每单所使用的时间：

```
df_duration <- data.frame(duration <- difftime(data$etime, data$stime, units='mins'))
```

之后又在绘图时取 1 分钟为间隔，用 ggplot()和 geom_histogram()来绘制时长分布条形分布图，如下图：



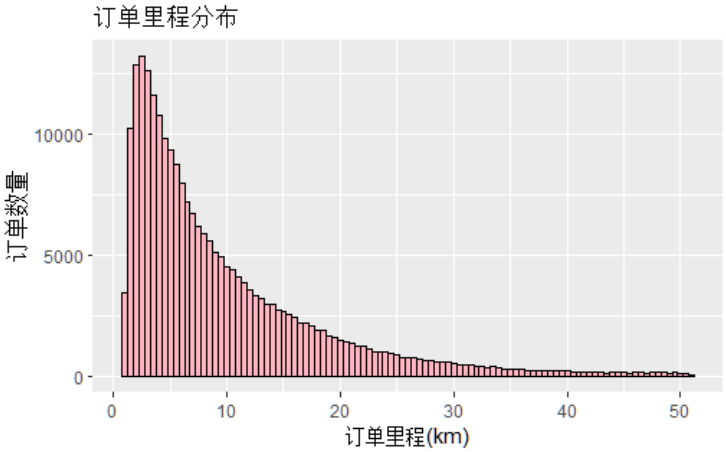
从上图看出，居民乘坐出租车的市场大多在 30 分钟以内，说明进行的是中短途的通勤和出行，30-60 分钟的中长距离占比比较少，60 分钟以上的长距离订单占比很少。从价格的角度也可以揣测，在时效性不是很高的情况下，长距离出行人们相比出租车还是更愿意乘坐公共交通。

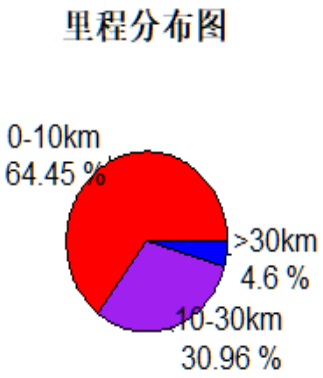
3. 出租车订单里程分布

在对订单里程的可视化方面，除了作出条形图，我还将数据分为<10km, 10~30km, >30km 三个层次，使用如下代码画出饼图：

```
df_distance$distance_category <- cut(df_distance$distance, breaks = c(0, 10, 30, Inf), labels = c("0-10km", "10-30km", ">30km"))
df_distance$prop <- with(df_distance, prop.table(table(distance_category)))
counts <- table(df_distance$distance_category)
props <- prop.table(counts)  # 计算每个类别的占比
labels <- paste(names(props), "\n", round(props * 100, 2), "%")
pie(counts, labels = labels, col = c("red", "purple", "blue"), main = "里程分布图") # 里程饼图
```

作图如下：

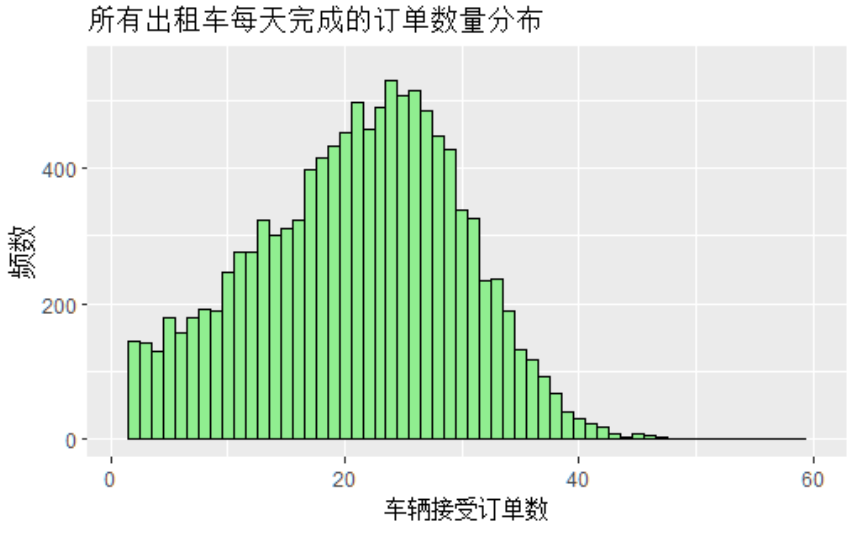




里程分布和时长分布不谋而合。在 2km 或短距离方面，人们可以通过步行、共享单车等方式到达，往往不会选择起步价 16 元的出租车。从里程上看，0-10km 占到了 64.45%，而很明显又看出，3-6km 占比最多。在这个区间内选择出租车的原因可能是时效性和点对点的便捷性。订单数量占比随着里程增加而降低，在这一区间内的订单的出行动机可能是从市区前往虹桥枢纽或者上海站、上海南站等，也可能是浦东张江、川沙、唐镇等居民前往浦东机场。30km 以上的订单仅占比 4.6%，可见大多数人从经济性的角度出发，在这一出行距离区间上将其他出行方式排在了前面。这与我们对出租车的认识还比较契合：在必要时和中短途时作为快速到达的一种选择，在短距离和长距离时则不优先考虑。

4. 出租车订单数量分布

接下来我绘制了出租车接受订单数量的频率分布直方图。首先我还是用 group_by 和 summarise()函数，根据出租车编号 CarID 统计了每辆车在这一天接收的订单数量。接着画出频率分布直方图如下：



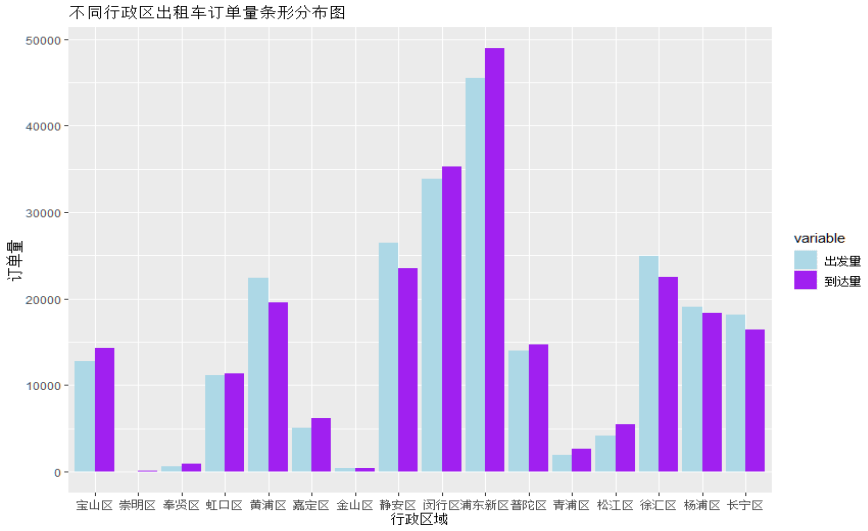
可以看出出租车订单接收的数量还是有明显差距的，一天能接到超过 40 单的司机数量很少，我们发现最大值大概在 50 单左右，但大多数还是分布在 20 单~40 单之间。其中接单数为 24、25、26 单的出租车数量超过了 500 辆，这是连续三个也是仅有三个超过 500 辆出租车的数据。而且接单数量业余司机的工作时长、个人安排、天气、日期、客流等都有关系。以上是时间可视化的部分。

（三） 订单地理可视化

时间可视化的数据可以比较清晰地看出订单的时间分布规律，但是若想更加直观地在地图上体现订单的分布状况，从而进一步探究空间特征，还需要进行地理可视化。

1. 各区订单量分布图

首先我们以 s_district_name 和 e_district_name 两个字段为筛选依据将数据按区分类，再用 ggplot()函数画出各区的订单量分布图：



由此可以看出上海出租车订单分布的空间不均衡性，市区普遍大于大部分郊区，距离市区较远的崇明、奉贤、金山等订单量极少，这与其地理位置有关。而浦东新区由于面积巨大，且和闵行区一样都属于人口居住、产业分布相当密集的行政区，且靠近市区，因此数量很高。

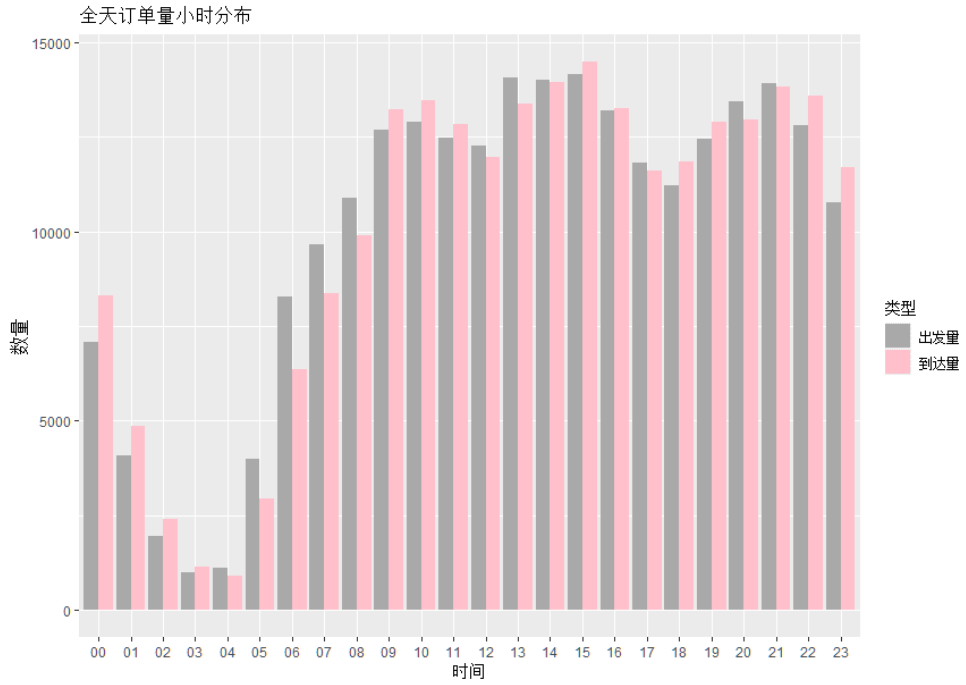
2. 全天订单小时分布图

全天的数据很多，若要全部在地图上体现，对于上海来说，必然是很多地方都热力值较高，难以体现不同时段的分布特征。因此，我选择先画出 4-18 订单量的小时分布图。在使用 format()函数将订单按照小时分类，创建 stime_hour, etime_hour 两个字段之后，我使用如下代码绘制了订单出发量和到达量小时分布图：

```
s_counts <- data.frame(hour = unique(data$stime_hour), count = table(data$stime_hour), type = "出发量")
# 每小时出发量
```

```
e_counts <- data.frame(hour = unique(data$etime_hour), count = table(data$etime_hour), type = "到达量")
# 每小时到达量
se_counts <- rbind(s_counts, e_counts)  # 合并数据
ggplot(se_counts, aes(x = se_counts$count.Var1, y = se_counts$count.Freq, fill = type)) +geom_bar(stat = "identity", position = "dodge") +scale_fill_manual(values = c("出发量" = "darkgray", "到达量" = "pink")) +labs(x = "时间", y = "数量", fill = "类型", title = "全天订单量小时分布")
```

绘制的图如下：



很明显地看出，凌晨时段，从 0~5 点订单数量较少，地图可视化效果也不好，况且该时段出行需求确实还比较少，因此我选择了 6:00~10:00 的早高峰时期、12:00~18:00 的下午时段和 18:00~24:00 的晚间时段来进行地理可视化，绘制 OD 热力图。

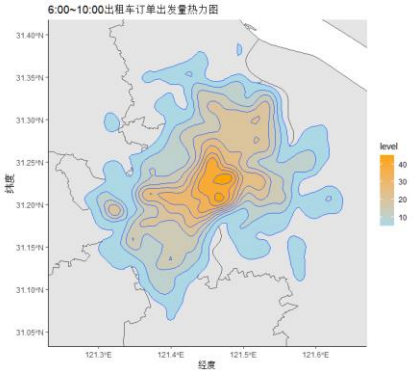
3. 不同时段的 OD 热力图

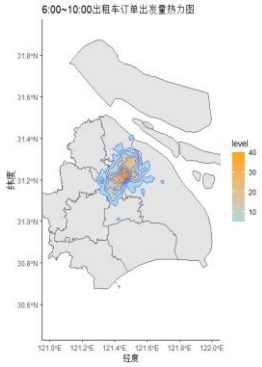
（1）6:00~10:00 早高峰时期

通过 st_read()读取文件中的 shanghai_district.shp，以此作为 shanghai_map 来充当底图，在绘图的时候参考的是高德坐标系，即字段带有“gcj_”字样的四组经纬度。

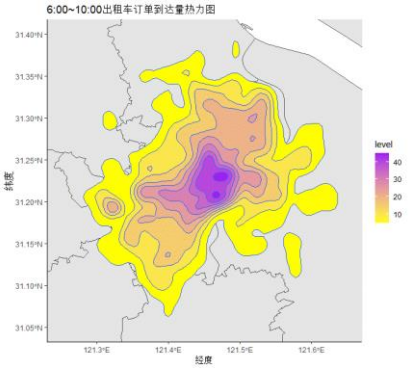
首先在进行时间数据格式化之后，用 morning_data <- data[data$stime_hour >= "06:00:00" & data$stime_hour <= "10:00:00",]来找出早高峰期间的数据，再用 ggplot()函数和 stat_density_2d 以及 geon_density_2d()函数绘制 6:00~10:00 出发和到达的热力图，如下图：

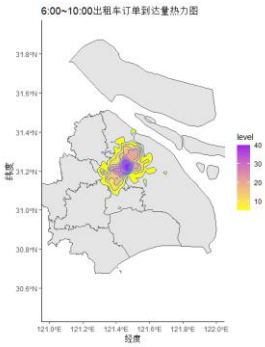
出发：





到达：



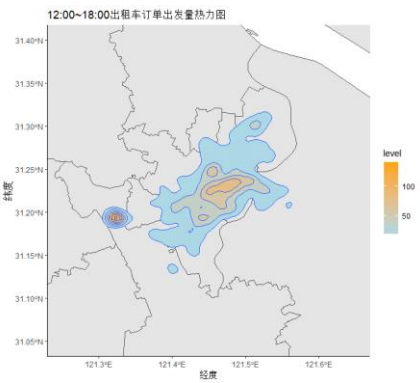


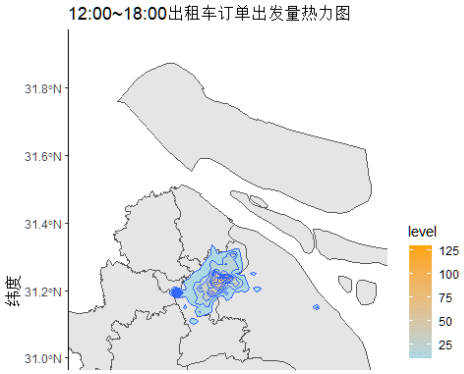
两组图中，左图绝大部分是市区的热力图，可以发现在两图中，市区、功能核心区等的热力值都很高，集中在内环以内的人民广场、静安寺，中环以内的陆家嘴等，其中浦西的订单数量更高一些，几乎做到了静安区、徐汇区、长宁区、普陀区、黄浦区、杨浦区、虹口区的大部分区域在早上的订单量都达到了中上水平。而浦东的张江、川沙、龙阳路、花木等地区由于产业集聚和人口居住等原因，也在早高峰时期产生了大量订单，揣测可能是中短距离的通勤或者从居住地到地铁站。从上海全域来看，早晨只有外环以内的区域热力值较高。

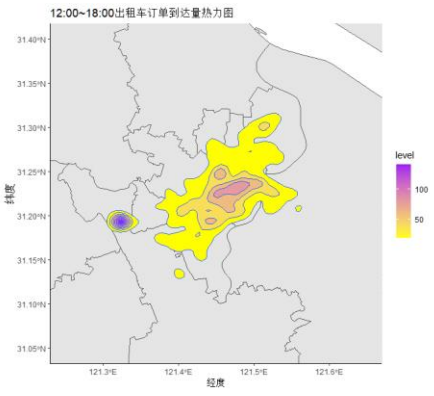
（2）12:00~18:00 下午时段 & 18:00~24:00 晚间时段

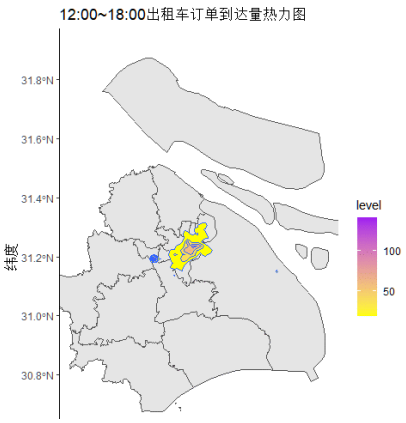
同样地，绘制 12:00~18:00 和 18:00~24:00 的热力图：

下午时段：

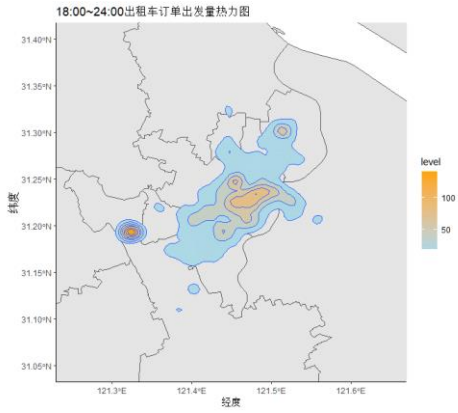


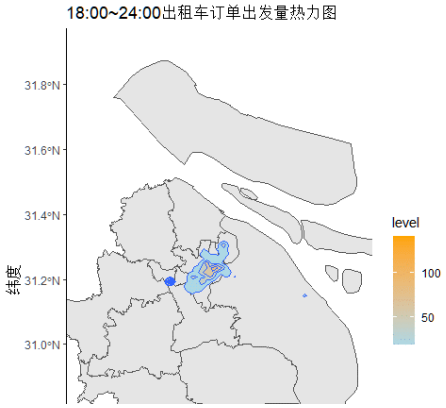


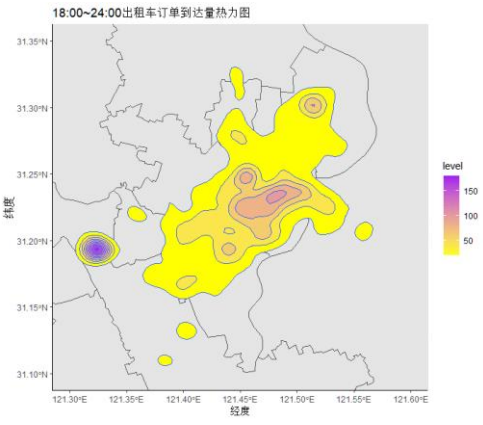


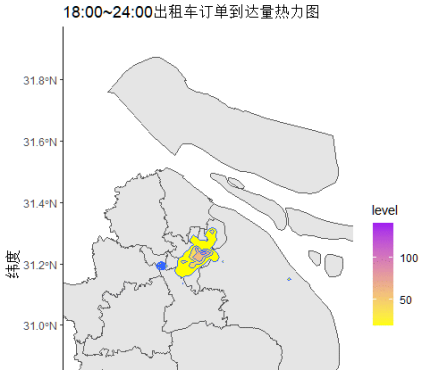


晚间时段：









由于下午和晚间时段的热力图具有相似性，因此我们一同分析。经查询，2018-4-18 是周三，为工作日，因此在下午时段工作人群出行的需求可能并不是很大，轨迹较为固定。而从高峰位置的变化来看，五角场、外滩、徐家汇等地区由于下午和晚间商业活动和出行的逐渐开展，变成了次高峰地区，普陀的大宁、长宁的天山、闵行的莘庄、浦东的花木等分布有大面积住宅区的地区也产生了较多订单，这与人们的外出的目的有关。

从高峰地区的变化来看，虹桥枢纽成为了热力值最大的区域，下午和晚间除了全天繁忙的京沪高铁和沪宁城际铁路之外，从成渝（如成都东站、重庆北站等）、云贵（如昆明南站、

贵阳北站）、华南（如广州南站、深圳北站、南宁东站等）和东北（如沈阳北站、长春西站、哈尔滨西站等）、西北地区（如兰州西站、西安北站等）的长距离列车陆续到达上海虹桥站，到站人数源源不断地增加带来出发订单的激增。同时更多的人选择在下午和晚间乘坐高铁和飞机出行，这也可以从图中虹桥机场、浦东机场等从早间的热力值较低成为了两个高值中心的看出。而我选择绘制上海全域热力值分布的目的正是为了观察浦东机场附近明显的变化。

从以上分析中我们可以清晰直观地看出居民的出行行为的时空分布。

三、 总结

本次作业需要大量进行数据可视化工作，让我首先主动地对数据进行筛选和清洗，而且对于常用的 ggplot()函数的使用方法和在地图上绘图的方法有了更深的了解。同时，我们还可以从出租车订单的时空分布中观察城市的日常运行动态，直观地体会城市经济发展、人口分布以及重要枢纽设施的分布等情况，具有实际意义。