Exercise 3

191820019 陈文杰

2020/10/23

R Markdown

#问题一：flights数据集中有air_time和distance两列，前者表示飞行时长，后者表示飞行距离。：

1、请用飞行距离对飞行时长做回归模型，并对结果予以解释（路径系数、显著性水平、R2、Adjusted R2等）。distance会对air_time显著影响吗？显著性水平是多少？

解：

##[策略] ①对数据集进行备份 ②利用lm模型进行线性拟合 ③用summary（）函数浏览线性回归模型 ##[参数解释] ①路径系数为：0.1261193 ②显著性水平<2e-16,说明自变量distance和因变量air_time的相关性较为显著 ③R2：衡量模型拟合度的一个量,是一个比例形式，被解释方差/总方差。R2=0.9814,模型拟合效果较好 ④Adjusted R-Squared 抵消样本数量对 R-Squared 的影响，用r square的时候，不断添加变量能让模型的效果提升，而这种提升是虚假的。 利用adjusted rsquare，能对添加的非显著变量给出惩罚，也就是说随意添加一个变量不一定能让模型拟合度上升 Adjusted R2 = 0.9814 ⑤distance对air_time会有显著影响，显著性水平小于2e-16 ##[过程|结果] <

```
library(nycflights13)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)

flights.info <- flights
my.lm <- lm(air_time~distance,data=flights.info)
summary(my.lm)
```

```
##
## Call:
## lm(formula = air_time ~ distance, data = flights.info)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -82.397  -7.334  -1.320   6.513 145.389
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.847e+01  3.888e-02   474.9   <2e-16 ***
## distance    1.261e-01  3.036e-05  4154.4   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.78 on 327344 degrees of freedom
##   (9430 observations deleted due to missingness)
## Multiple R-squared:  0.9814, Adjusted R-squared:  0.9814
## F-statistic: 1.726e+07 on 1 and 327344 DF,  p-value: < 2.2e-16
```

```
coefficients(my.lm)
```

```
## (Intercept)    distance
##  18.4665781   0.1261193
```

2、抽取起飞地（origin）是“JFK”的数据，再进行上述分析。请问，样本量现在是多少？其他问题同上。

解:

##[策略] ①利用filter函数筛选origin是JFK的数据 ②利用lm和summary函数拟合并评估线性回归模型

##[过程|结果]

```
flights.select <- filter(flights.info, origin == "JFK")
dim(flights.select)[1]
```

```
## [1] 111279
```

```
my.lm2 <- lm(air_time~distance,data=flights.select)
summary(my.lm2)
```

```
##
## Call:
## lm(formula = air_time ~ distance, data = flights.select)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -65.472  -7.488  -0.614   7.386 146.512
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.769e+01  7.372e-02     240   <2e-16 ***
## distance    1.260e-01  4.730e-05    2664   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14 on 109077 degrees of freedom
##   (2200 observations deleted due to missingness)
## Multiple R-squared:  0.9849, Adjusted R-squared:  0.9849
## F-statistic: 7.095e+06 on 1 and 109077 DF,  p-value: < 2.2e-16
```

3.抽取起飞地（origin）是“JFK”的数据，绘制散点图，x轴是distance，y轴是air_time。请同时用loess拟合一条回归线，要求se是T。

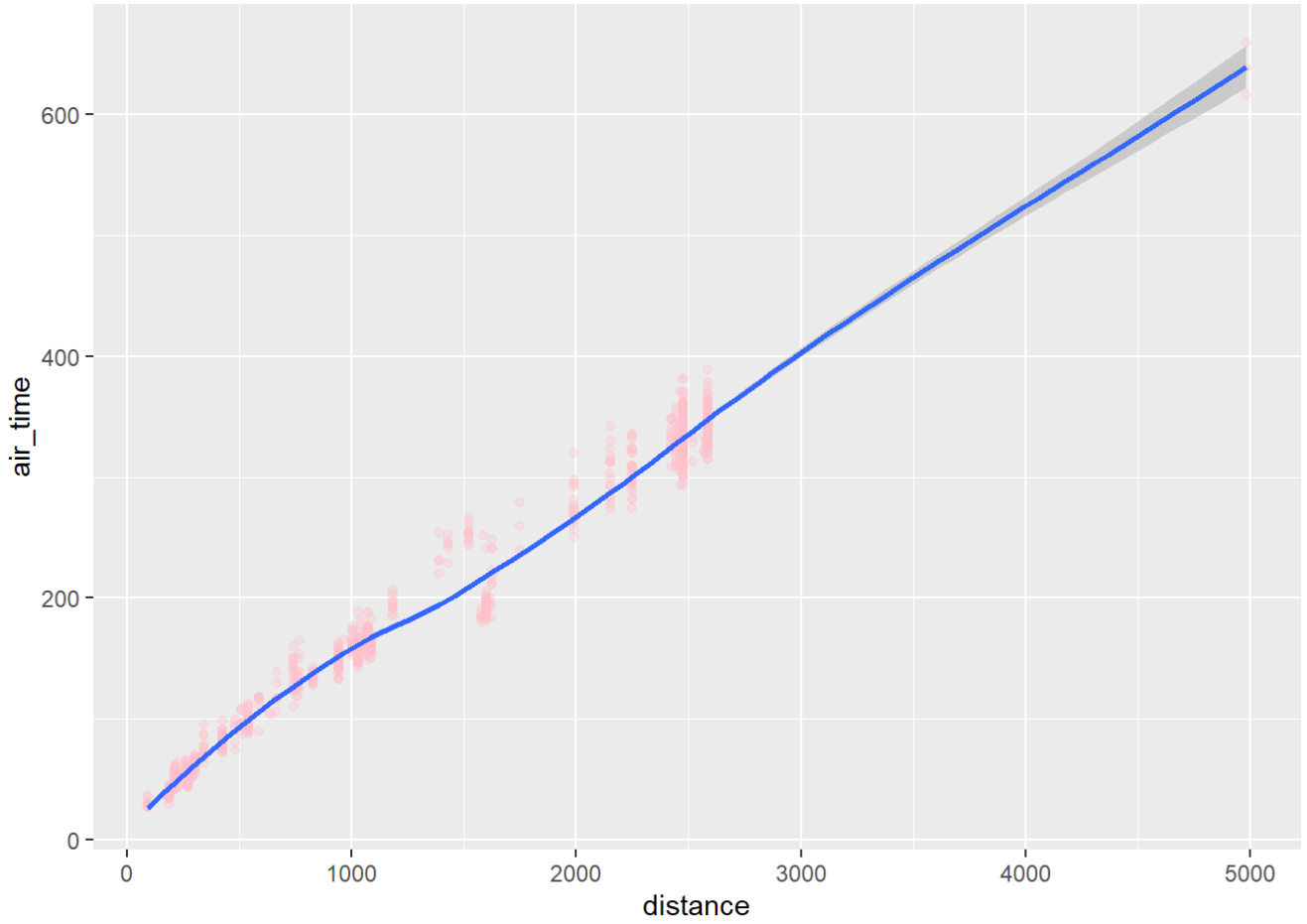
解:

##[策略] ①利用filter函数筛选origin是JFK的数据 ②为简化运算，选取前1000行进行模型拟合（主要目的是尝试模型拟合方法） ③利用ggplot函数，选取loess方式，拟合回归线。 同时用jitter和alpha参数避免overplotting

##[过程|结果]

```
flights.select <- filter(flights.info, origin == "JFK")[0:1000,]
flights.select <- na.omit(flights.select)
ggplot(flights.select,
       aes(x=distance, y=air_time)) +
  geom_point(position = position_jitter(0.1),alpha=0.3,col='pink')+
  geom_smooth(method='loess',se = T)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



4.选择整个数据集，同样绘制散点图，x轴是distance，y轴是air_time，并用loess拟合一条回归线，要求se是T。与上一题不同之处在于：请用origin做facets。

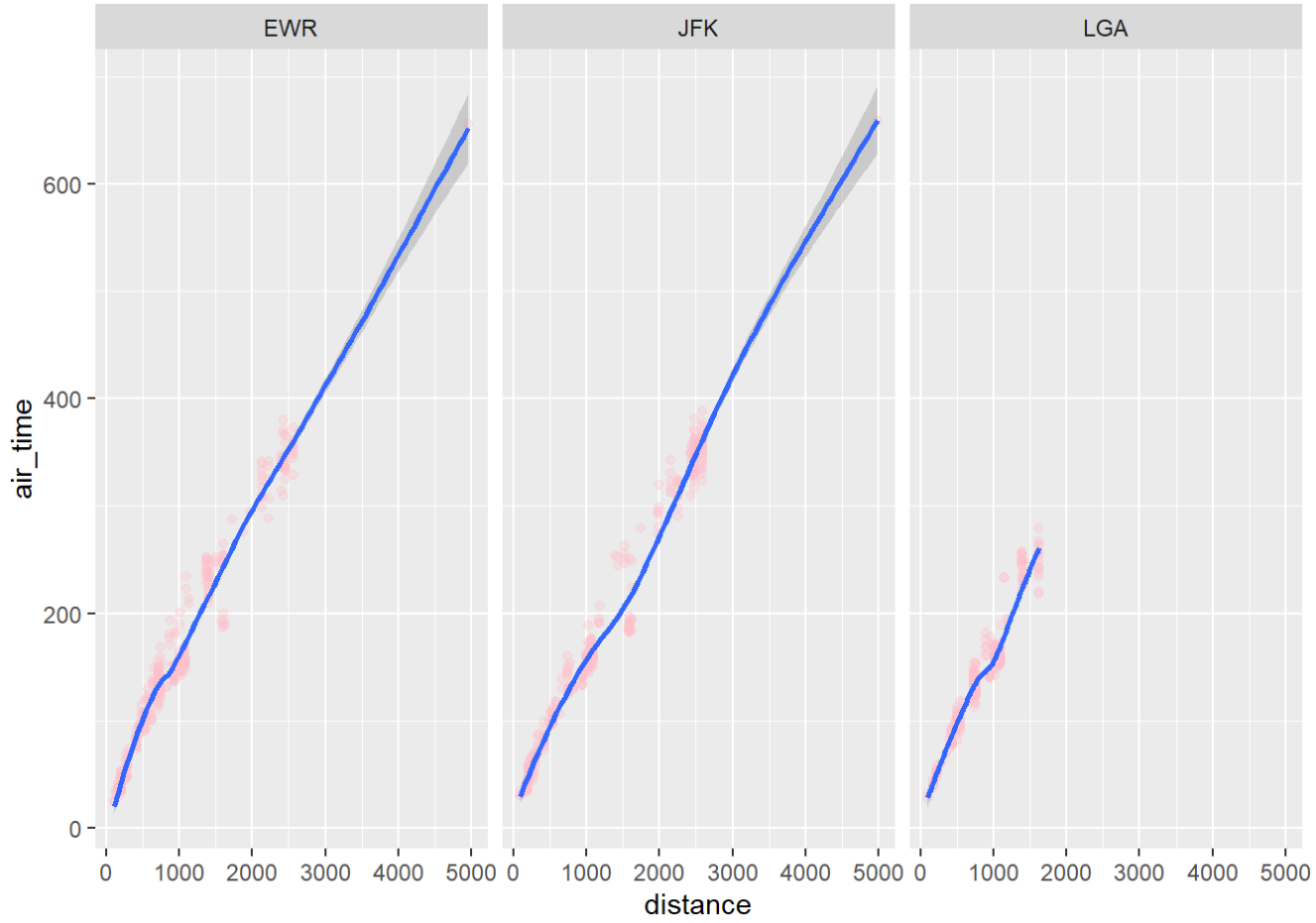
解:

##[策略] ①利用filter函数筛选origin是JFK的数据 ②为简化运算，选取前1000行进行模型拟合（主要目的是尝试模型拟合方法） ③利用ggplot函数，选取loess方式，拟合回归线。 同时用jitter和alpha参数避免overplotting ④用facet_grid分面图层按origin分面

##[过程|结果]

```
flights.select2 <- na.omit(flights.info)[0:1000,]
ggplot(flights.select2,
       aes(x=distance, y=air_time)) +
  geom_point(position = position_jitter(0.1),alpha=0.3,col="pink")+
  geom_smooth(method='loess',se = T)+
  facet_grid(.~origin)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



#问题二：知道不同月份，飞机延误情况对于安排出行计划较为有趣。请用flights数据集，完成下述问题。 1.分别采用两种几何图形（boxplot和violin）绘制每个月起降飞机延误时间的信息，x轴为月份，y轴为延误时间arr_delay。

要求：每个月份的图形fill颜色不一样。请尝试选择2-3种不同的theme，比较theme的作用。

解：

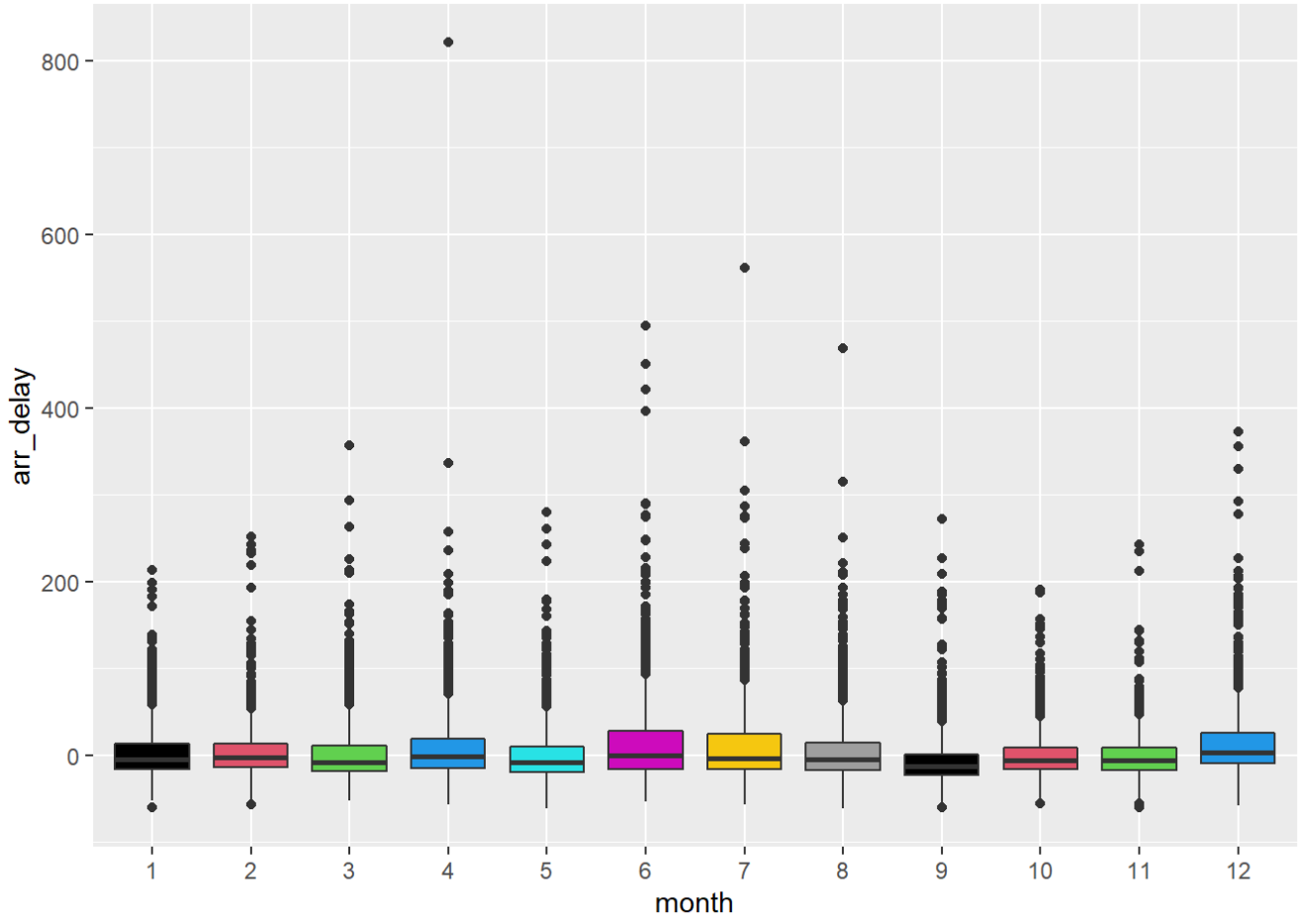
##[策略]

- 利用sample抽取随机数并对数据集随机抽样
- 利用geom_boxplot绘制箱型图
- 利用geom_violin绘制小提琴图
- 尝试不同theme，绘制出风格不同的数据图表

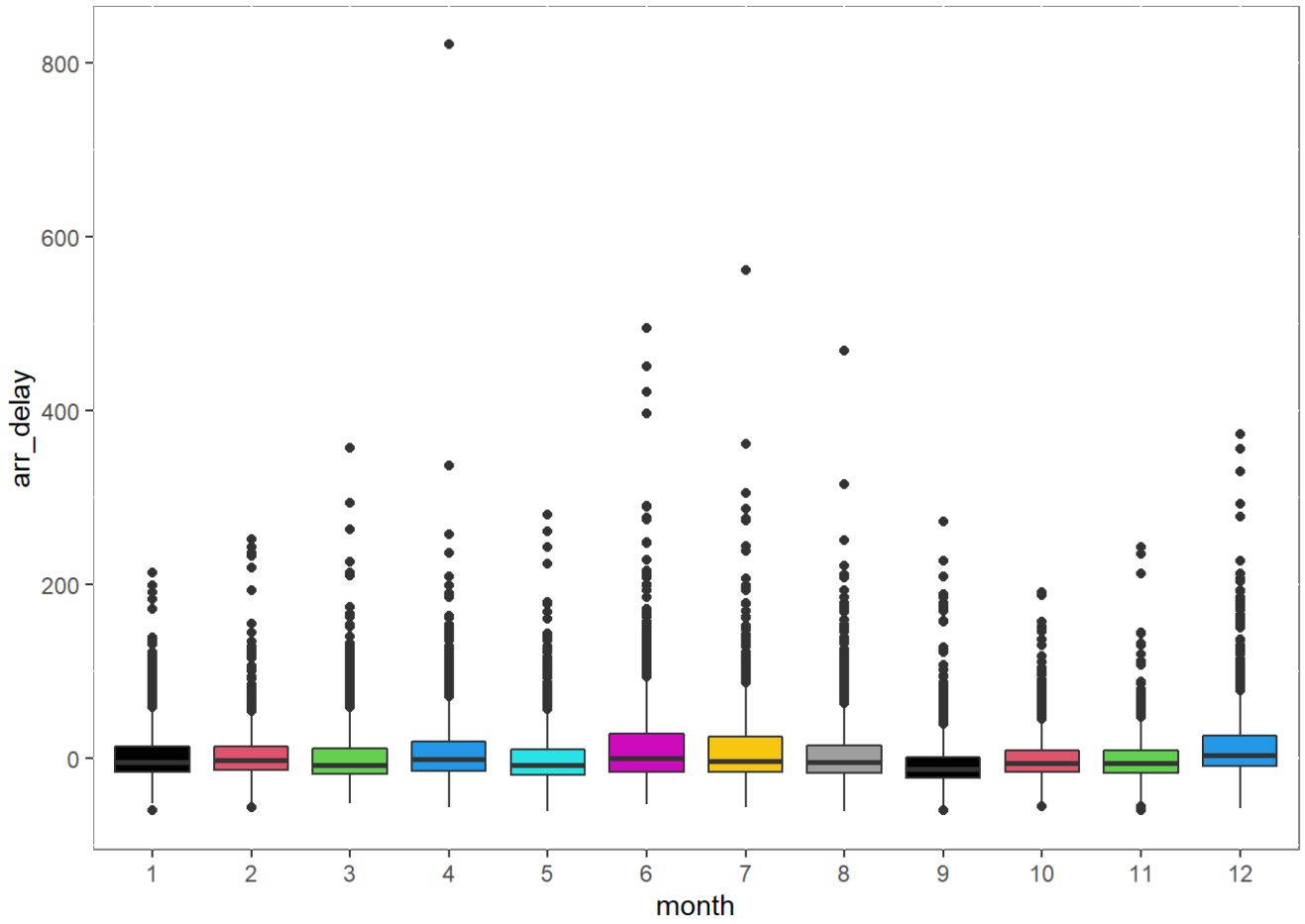
##[过程|结果]

```
#sample随机取样
flights.info$month <-as.numeric(flights.info$month)
train <- sample(dim(flights.info)[1],(0.03*dim(flights.info)[1]))
tmp <- (flights2.select3 <- flights.info[train,]) %>% na.omit()
tmp <- arrange(tmp, tmp$month)
tmp$month<- as.factor(tmp$month)
#绘制箱型图

p1 <- ggplot(tmp
       ) +
  geom_boxplot(aes(x=month, y=arr_delay),fill=unique(tmp$month))
p1
```

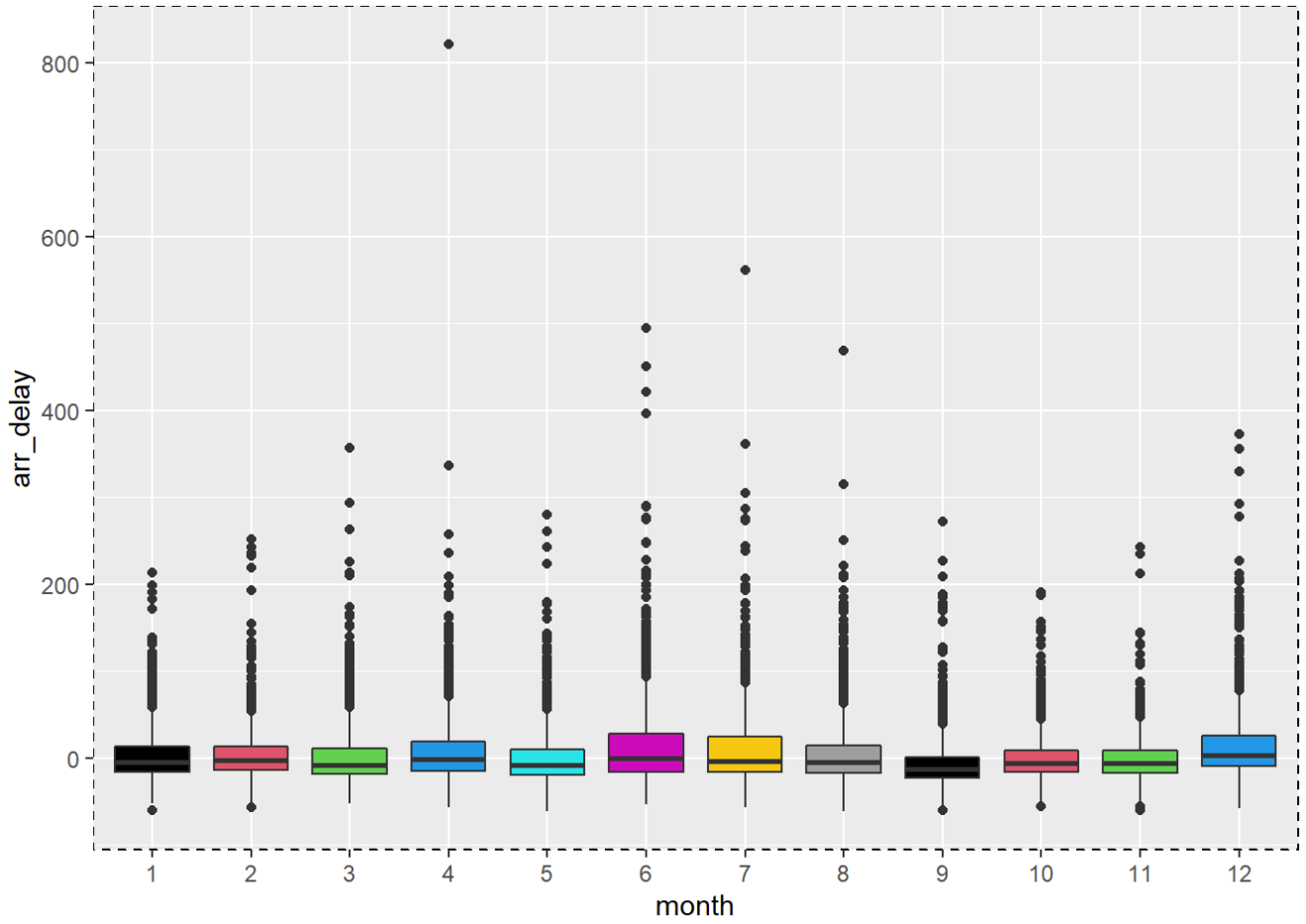



```
p1 + theme(panel.background = element_rect(fill = "white", colour = "grey50"))
```



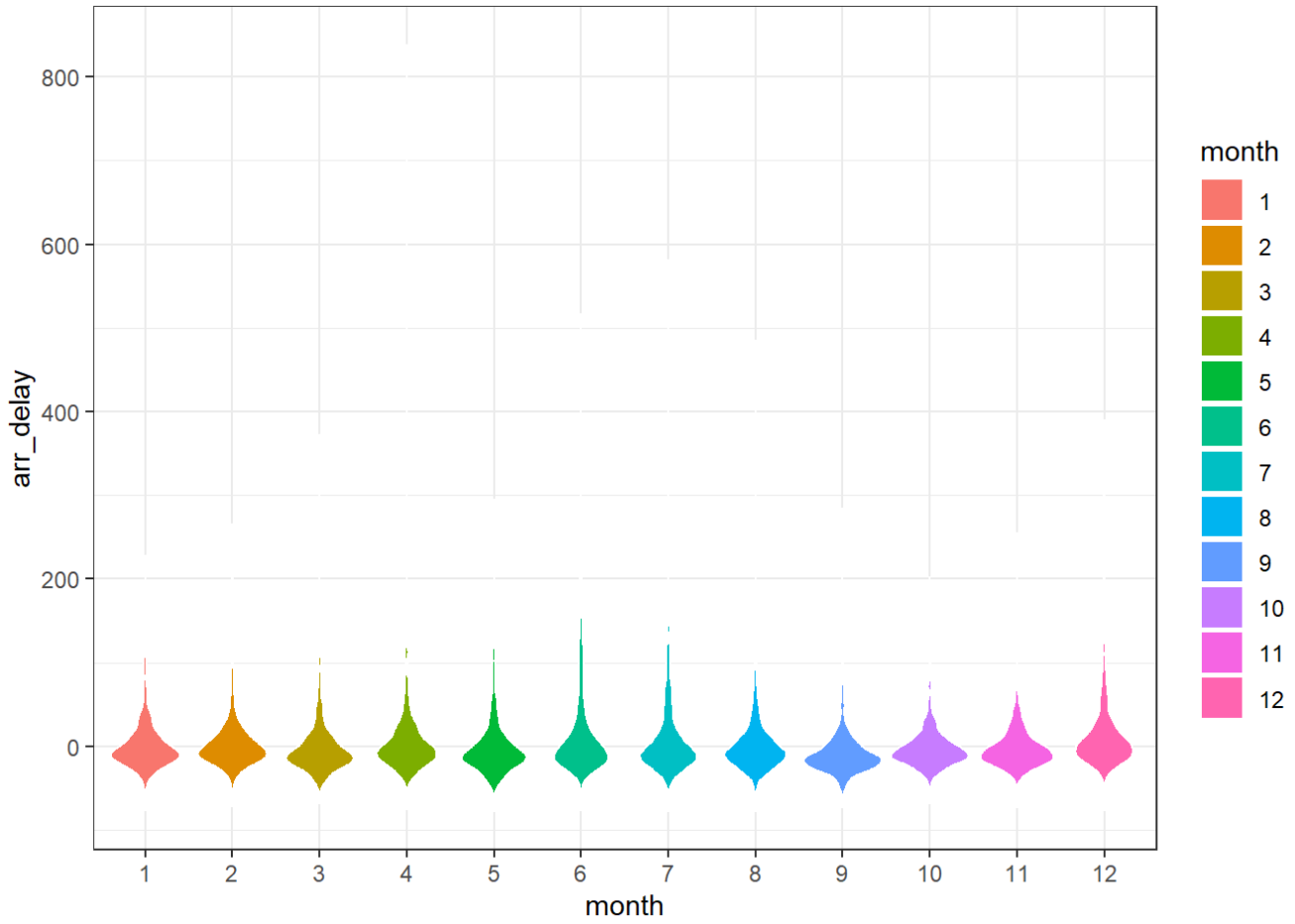
File failed to load: /extensions/MathZoom.js

```
p1 + theme(panel.border = element_rect(linetype = "dashed", fill = NA))
```




```
#绘制小提琴图
p2 <- ggplot(tmp)+
  geom_violin(aes(x=month, y=arr_delay,fill=month),trim=FALSE,color="white") + #绘制小提琴图
  theme_bw() #背景变为白色

p2
```



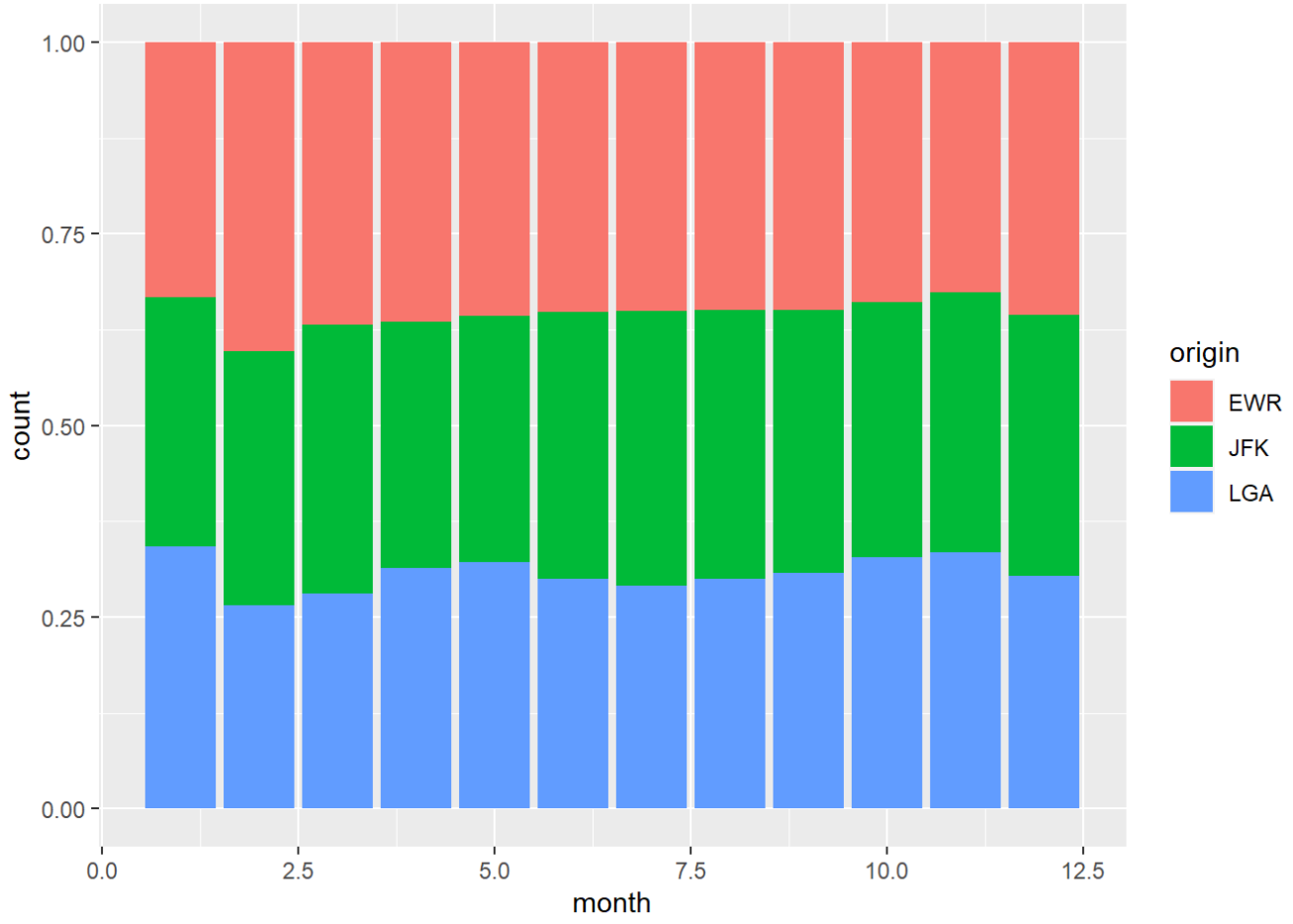
2.用geom_bar()展示每个月、每个origin（只有3种）起飞的飞机数，x轴为月份。分别用三种不同的position（dodge等）来排列bar。

##[策略] * 利用sample抽取随机数并对数据集随机抽样 * 利用geom_bar反映不同月份、不同origin的飞机起飞数量 * 分别尝试dodge、fill、stack的堆叠效果

##[过程|结果]

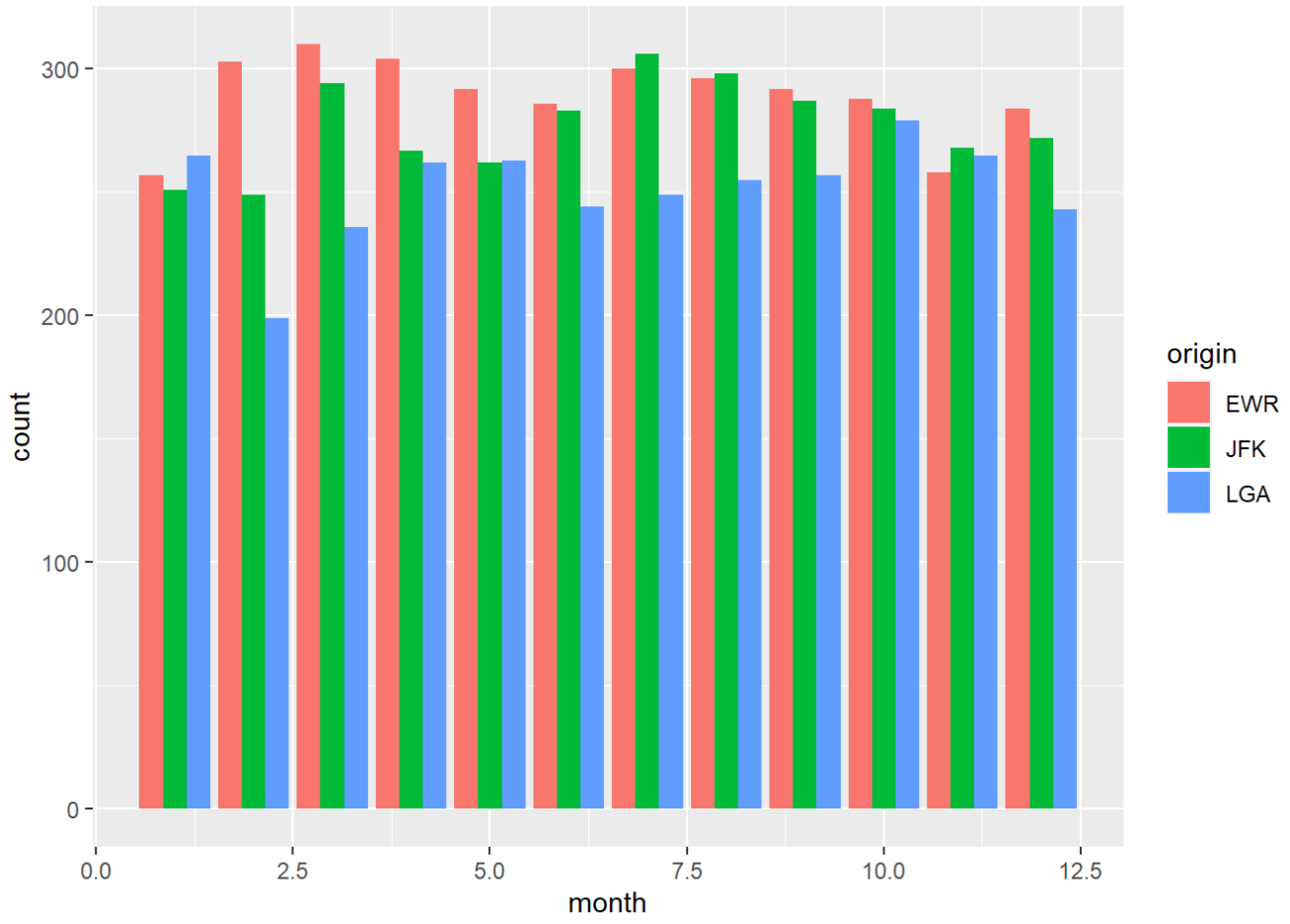
```
train <- sample(dim(flights.info)[1],(0.03*dim(flights.info)[1]))
tmp <- (flights2.select3 <- flights.info[train,]) %>% na.omit()

p3 <- ggplot(tmp,aes(x=month, fill=origin)
            )+
  geom_bar(position="fill")
p3
```

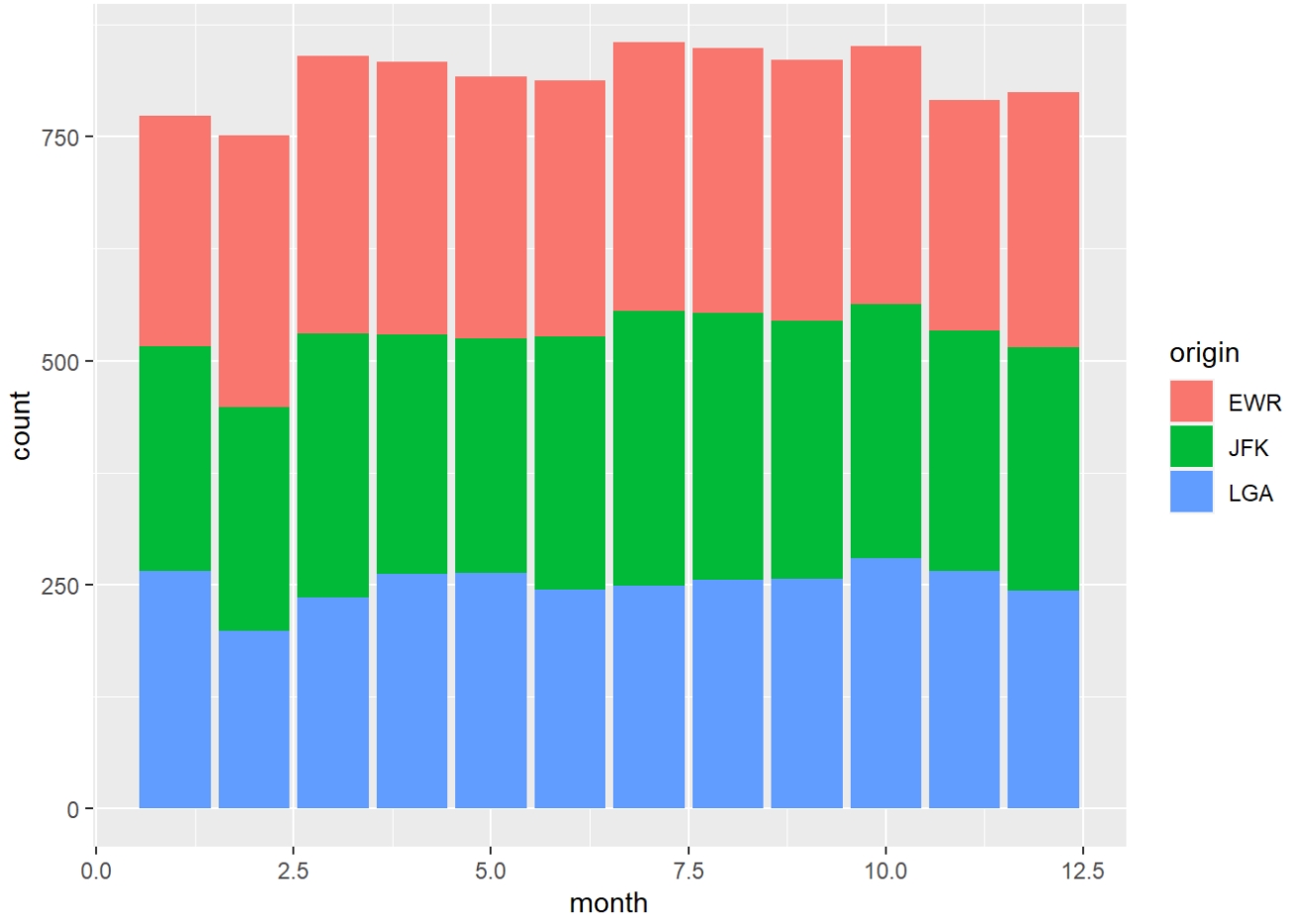



```
p4 <- ggplot(tmp,aes(x=month, fill=origin)
)+
  geom_bar(position="dodge")
p4
```

File failed to load: /extensions/MathZoom.js




```
p5 <- ggplot(tmp,aes(x=month, fill=origin)
)+
  geom_bar(position="stack")
p5
```



File failed to load: /extensions/MathZoom.js