Homework2

YikunHan42

2022-04-17

homework2

导入第三方库

```
library(nycflights13)
```

```
## Warning: 程辑包'nycflights13'是用R版本4.1.3 来建造的
```

```
library(tidyverse)
```

```
## Warning: 程辑包'tidyverse'是用R版本4.1.3 来建造的
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
library(dplyr)
library(stringr)
library(openair)
```

```
## Warning: 程辑包'openair'是用R版本4.1.3 来建造的
```

```
library(GGally)
```

```
## Warning: 程辑包'GGally'是用R版本4.1.3 来建造的
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(psych)
```

```
## Warning: 程辑包'psych'是用R版本4.1.3 来建造的
```

```
##
## 载入程辑包：'psych'
```

```
## The following object is masked from 'package:openair':
##
##     corPlot
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
options (warn = -1)
dim(flights)
```

```
## [1] 336776     19
```

数据概览

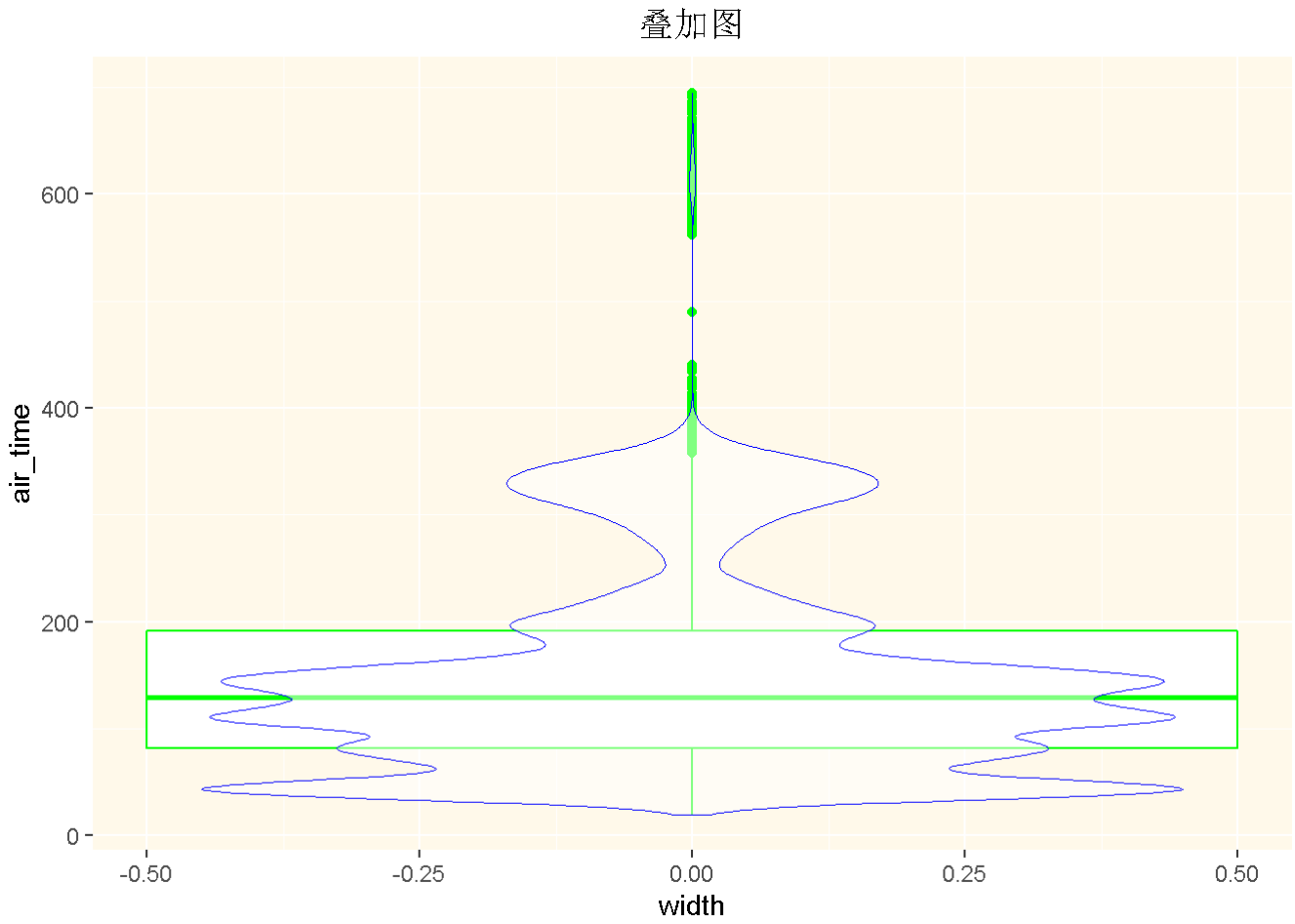
```
flights
```

```
## # A tibble: 336,776 x 19
##     year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
##  1  2013     1     1      517            515         2      830            819
##  2  2013     1     1      533            529         4      850            830
##  3  2013     1     1      542            540         2      923            850
##  4  2013     1     1      544            545        -1     1004           1022
##  5  2013     1     1      554            600        -6      812            837
##  6  2013     1     1      554            558        -4      740            728
##  7  2013     1     1      555            600        -5      913            854
##  8  2013     1     1      557            600        -3      709            723
##  9  2013     1     1      557            600        -3      838            846
## 10  2013     1     1      558            600        -2      753            745
## # ... with 336,766 more rows, and 11 more variables: arr_delay <dbl>,
## #   carrier <chr>, flight <int>, tailnum <chr>, origin <chr>, dest <chr>,
## #   air_time <dbl>, distance <dbl>, hour <dbl>, minute <dbl>, time_hour <dttm>
```

1. 箱线图和小提琴图

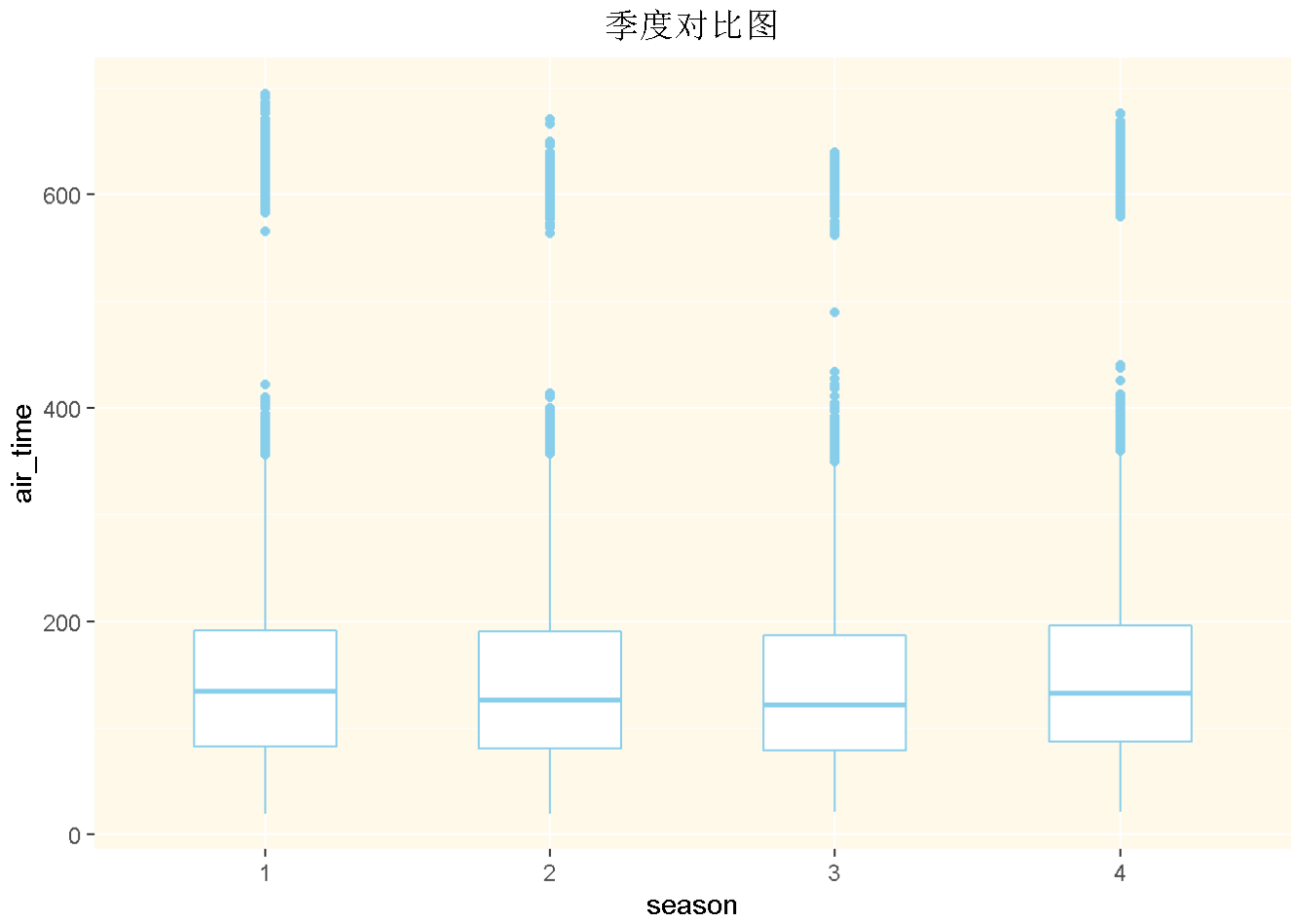
(1) air_time箱线图+小提琴图

```
p <- ggplot(data = flights, mapping = aes(x = 0, y = air_time), fill = attributes)
p + geom_boxplot(width = 1,position = position_dodge(0.9), color = " green") + geom_violin(size
= 0.01,alpha = 0.5,color = "blue") + labs(title = "叠加图", x= "width") + theme(plot.title=elem
ent_text(hjust=0.5), panel.background = element_rect(fill = "#FFBC1717"))
```



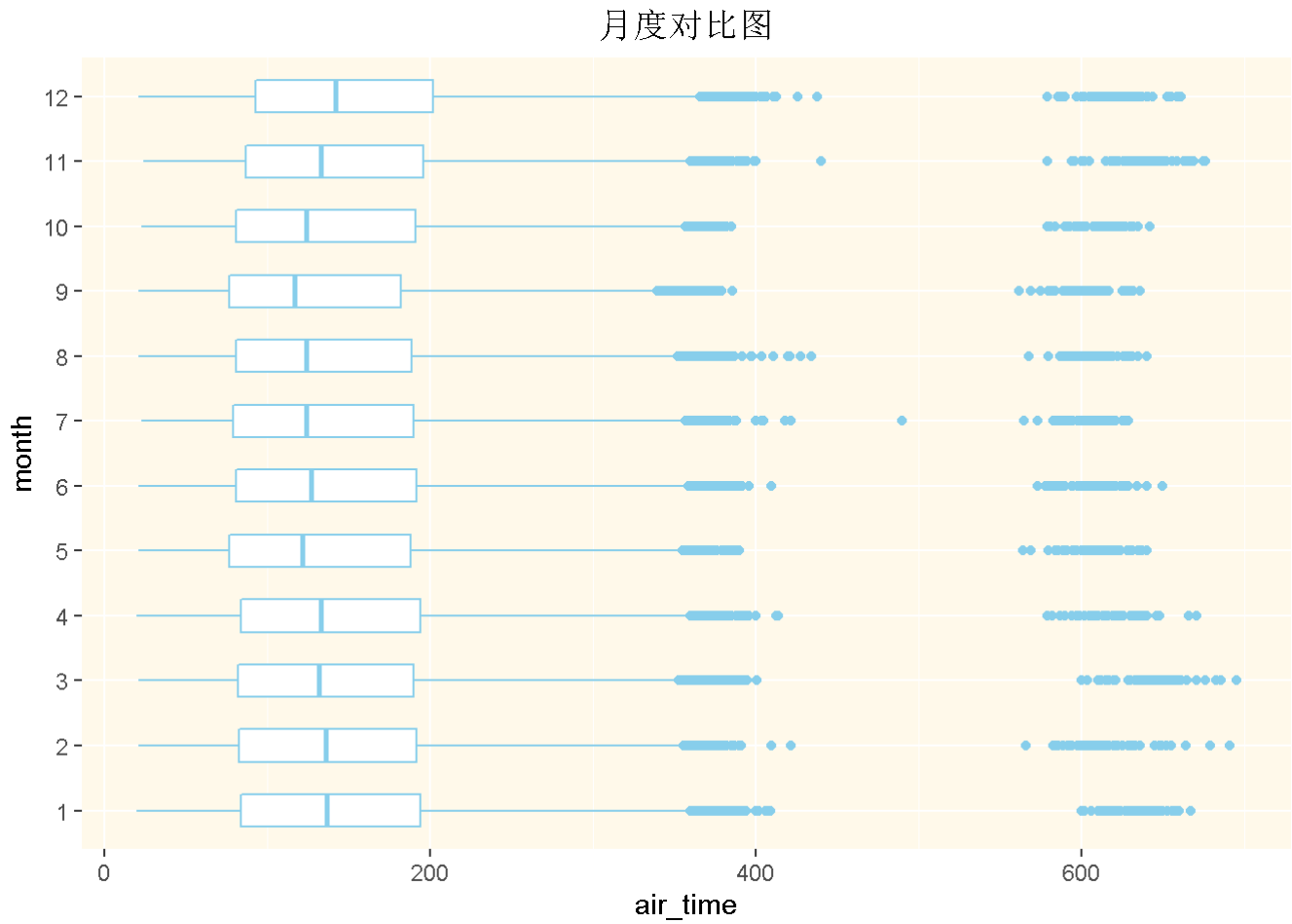
(2) 季度飞行时间箱线图对比

```
flights <- flights %>%
    mutate(season = ceiling((month)/3))
seasonalplot <- ggplot(data = flights, mapping = aes(x = as.factor(season), y = air_time), fill
= Attribute)
seasonalplot + geom_boxplot(width = 0.5, position = position_dodge(0.9), color = "skyblue") + l
abs(title = "季度对比图", x= "season") + theme(plot.title=element_text(hjust=0.5),panel.backgro
und = element_rect(fill = "#FFBC1717"))
```



(3) 月度横置飞行时间箱线图对比

```
monthlyplot <- ggplot(data = flights, mapping = aes(x = as.factor(month), y = air_time), fill = Attribute)
monthlyplot + geom_boxplot(width = 0.5, position = position_dodge(0.9), color = "skyblue") + labs(title = "月度对比图", x = "month") + theme(plot.title = element_text(hjust = 0.5),panel.background = element_rect(fill = "#FFBC1717")) + coord_flip()
```

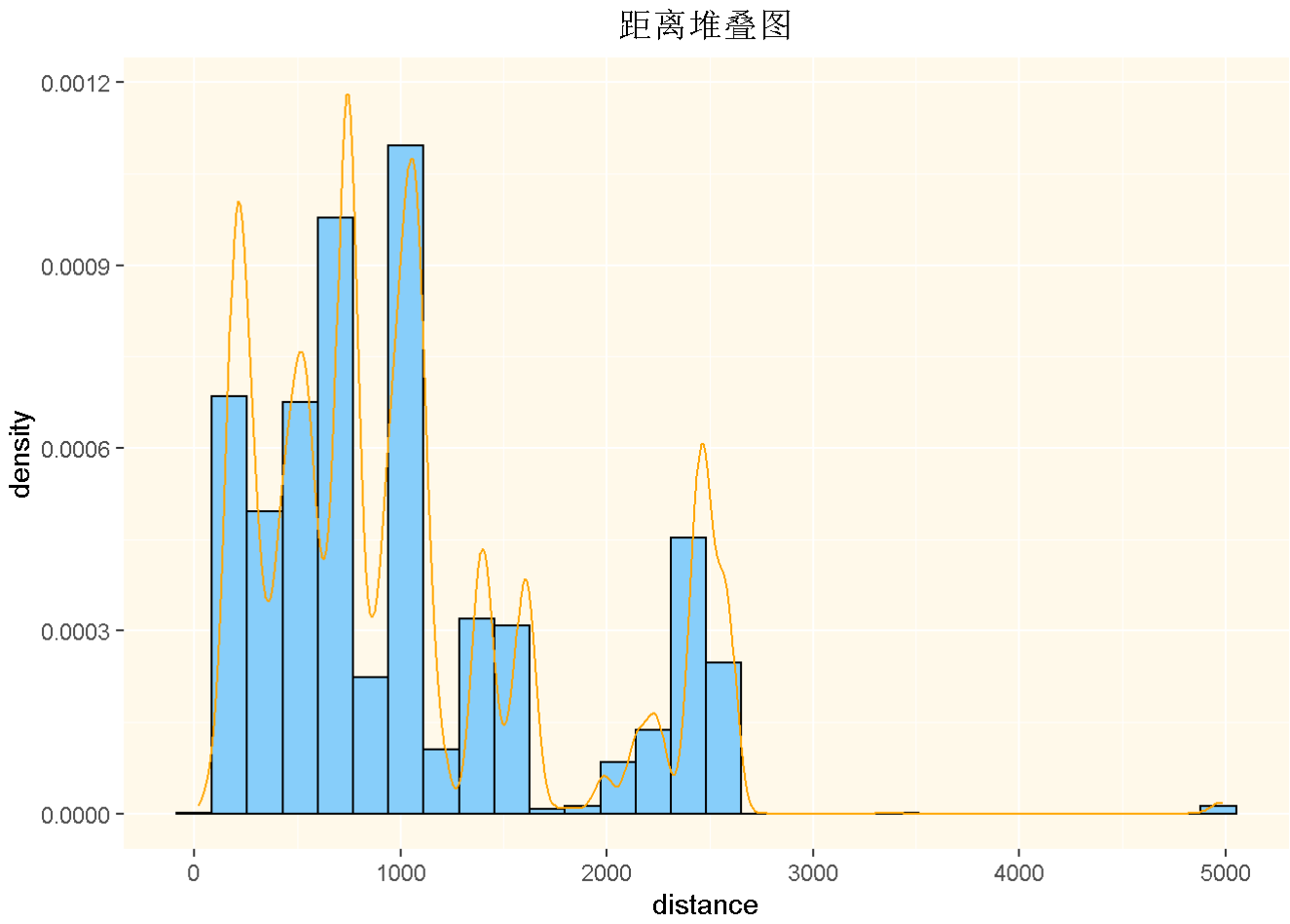


2

(1) 距离直方图 + 核密度估计曲线

```
p <- ggplot(data = flights, mapping= aes(x = distance))
p + geom_histogram(aes(y = ..density..), fill="lightskyblue", color="black") + geom_density(col
or="orange") + theme(plot.title = element_text(hjust = 0.5), panel.background = element_rect(fi
ll = "#FFBC1717")) + labs(title = "距离堆叠图")
```

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



(2) 公司核密度估计曲线对比

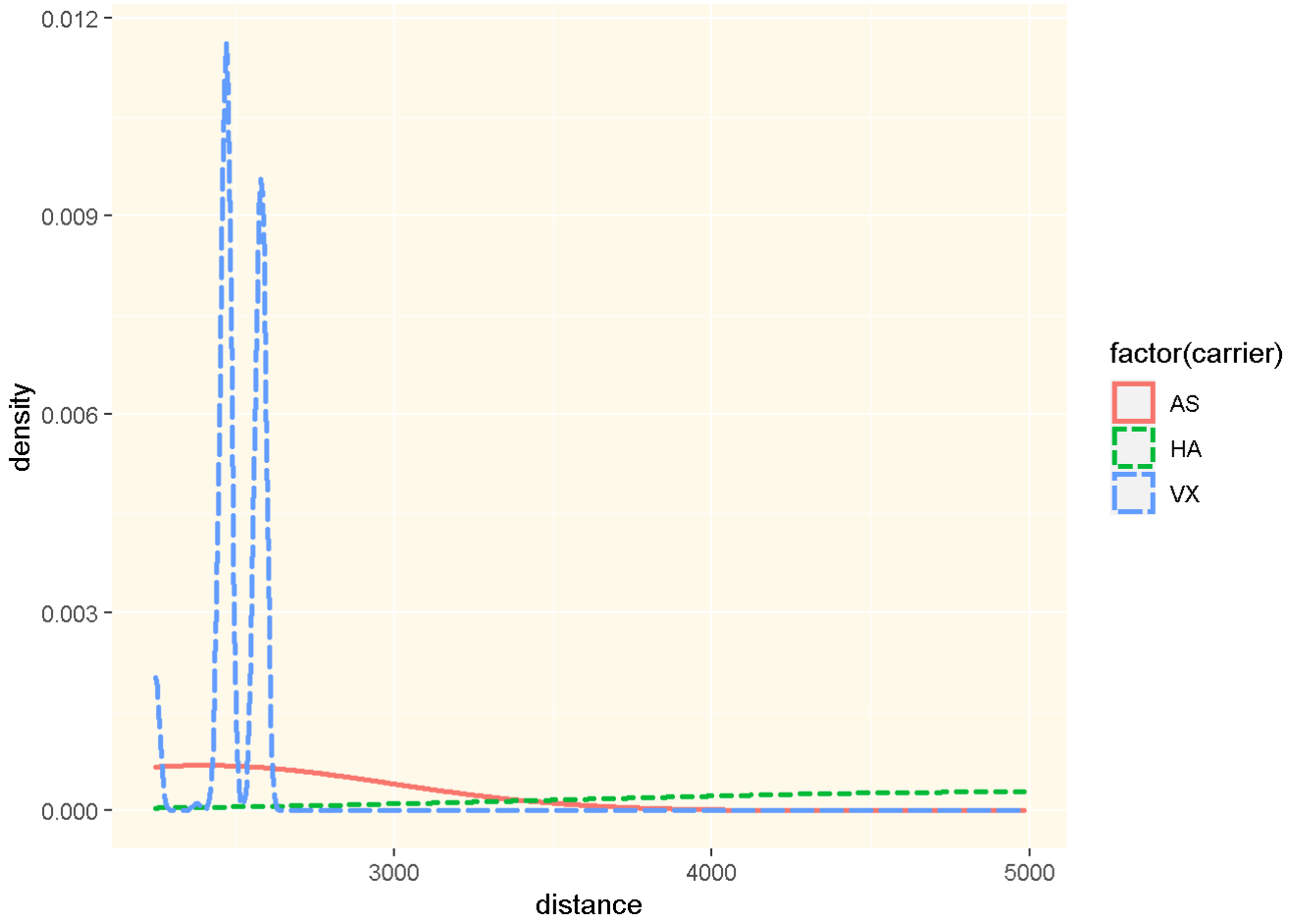
前三的公司

```
carrierrank = flights %>% group_by(carrier) %>% summarise(ave = mean(distance)) %>% arrange(desc(ave))
carrierrank[1:3,]
```

```
## # A tibble: 3 x 2
##   carrier   ave
##   <chr>   <dbl>
## 1 HA      4983
## 2 VX      2499.
## 3 AS      2402
```

曲线绘制

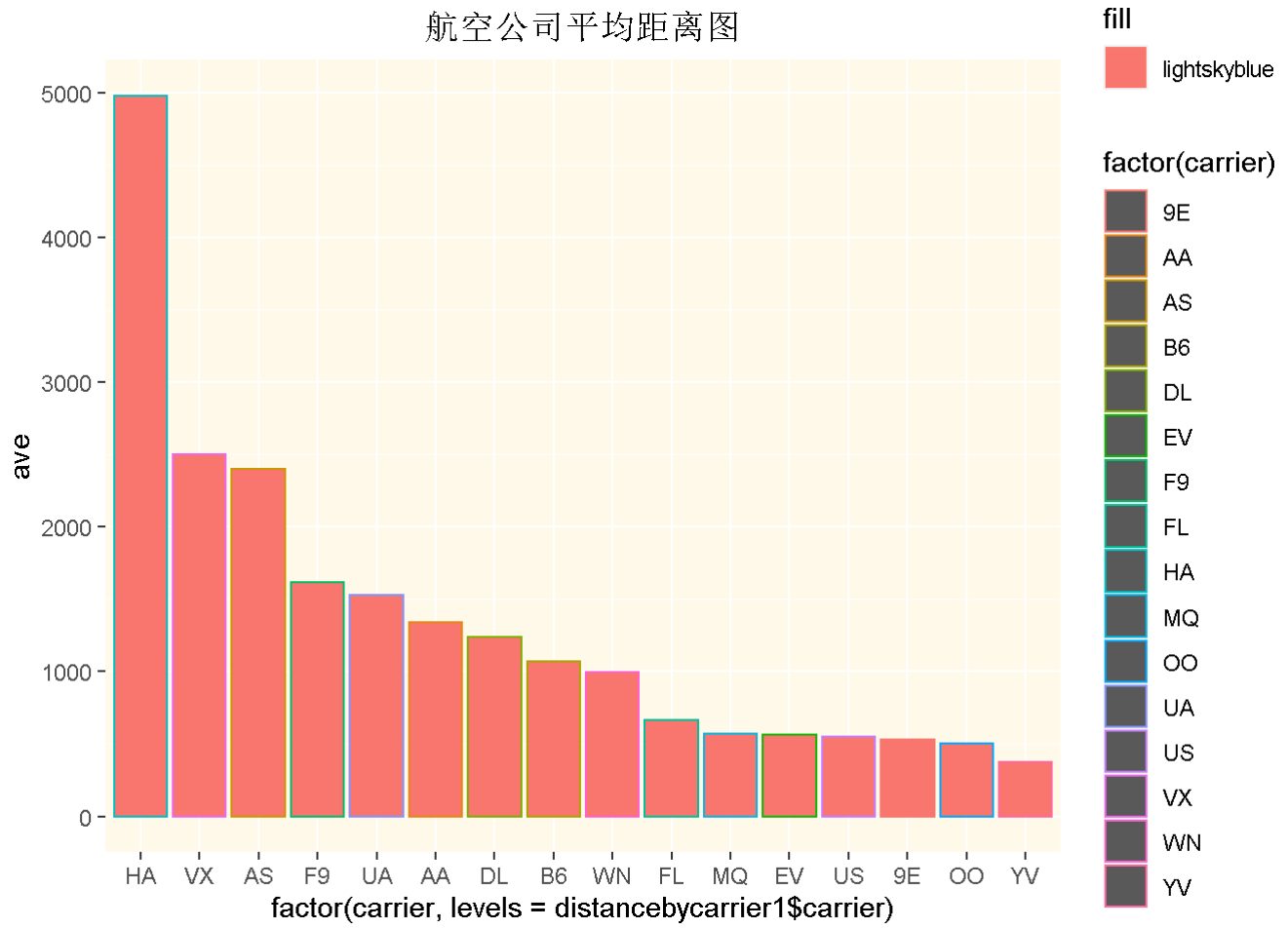
```
top3carrier = flights %>% filter(carrier=='HA'|carrier=='AS'|carrier=='VX')
p <- ggplot(data = top3carrier,mapping = aes(x = distance, color = factor(carrier),linetype = factor(carrier)))
p + geom_density(adjust = 1,size = 1) +  theme(panel.background = element_rect(fill = "#FFBC1717"))
```



3

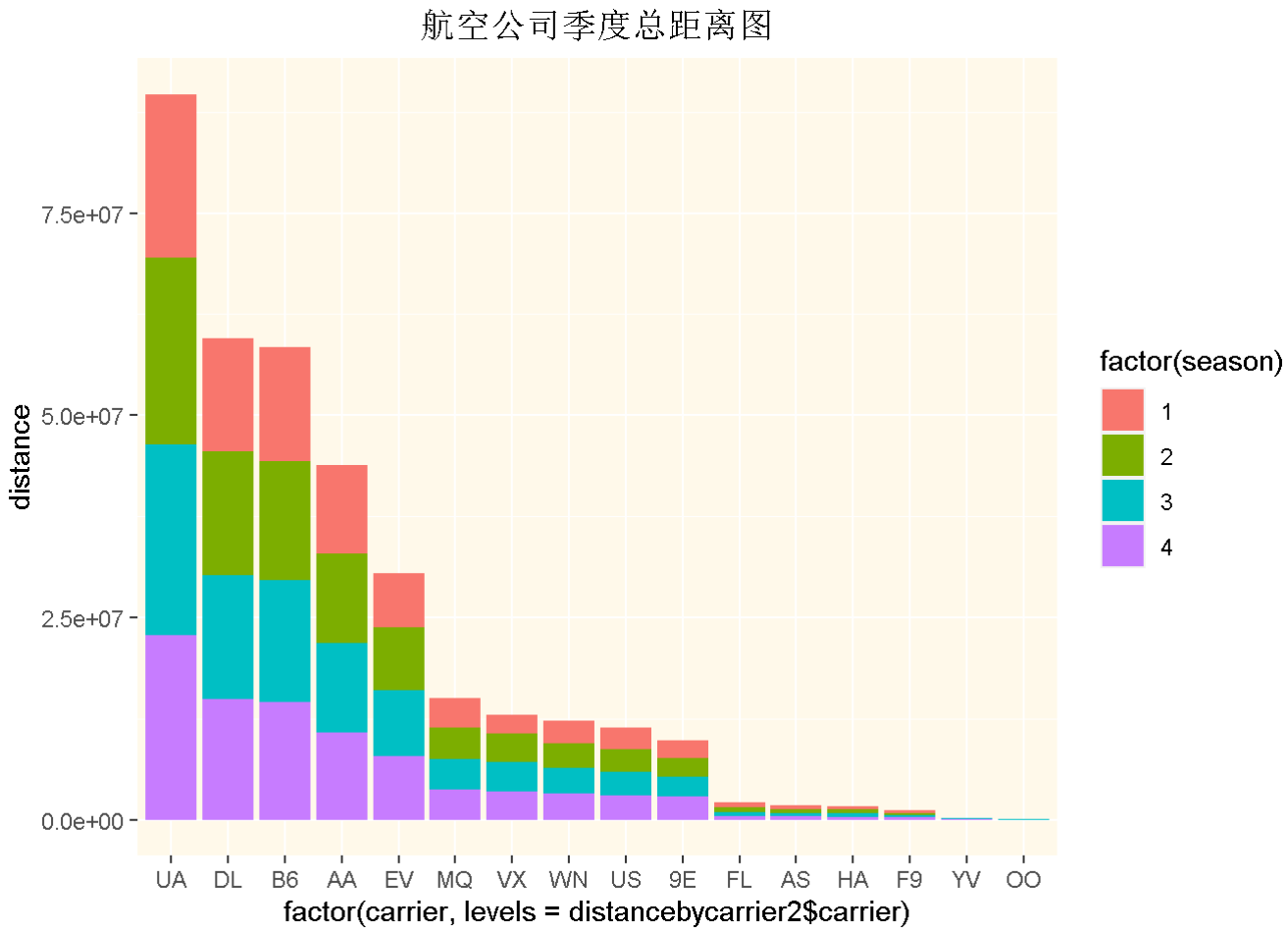
(1) 条形图

```
distancebycarrier1 = flights %>% group_by(carrier) %>% summarise(ave = mean(distance)) %>% arra
nge(desc(ave))
p <- ggplot(data = distancebycarrier1, mapping = aes(x = factor(carrier, levels = distancebycar
rier1$carrier), y = ave, fill="lightskyblue", color = factor(carrier)))
p + geom_col(stat="identity") + theme(plot.title = element_text(hjust = 0.5), panel.background
 = element_rect(fill = "#FFBC1717")) + labs(title = "航空公司平均距离图")
```



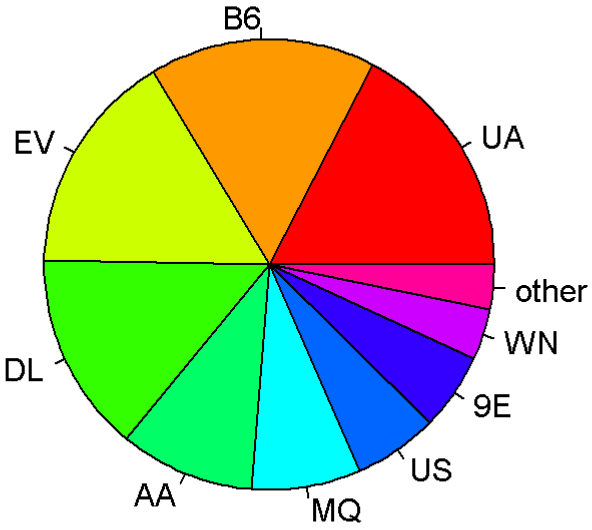
(2) 堆叠柱状图

```
flights %>% group_by(carrier) %>% summarise(sum = sum(distance)) %>% arrange(desc(sum)) -> distancebycarrier2
p <- ggplot(data = flights, mapping = aes(x = factor(carrier,levels = distancebycarrier2$carrier), y = distance, fill = factor(season)))
p + geom_col() + theme(plot.title = element_text(hjust = 0.5), panel.background = element_rect(fill = "#FFBC1717")) + labs(title = "航空公司季度总距离图")
```



(3) 航空公司比例饼图

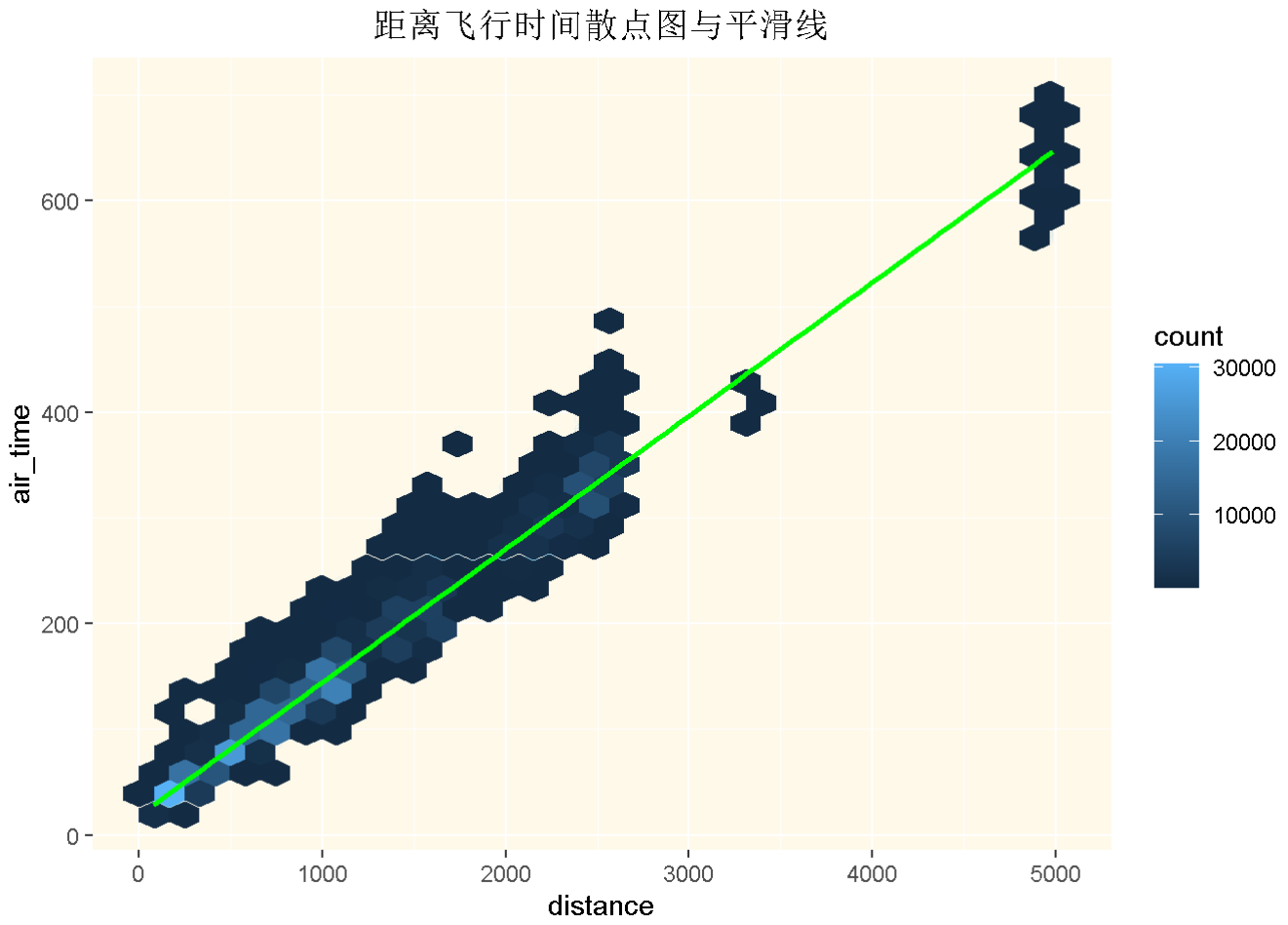
```
flights %>%
  group_by(carrier) %>%
  summarise(total_n=n()) %>%
  arrange(desc(total_n)) -> airlinecount
airlinecount %>%
  filter(total_n < 10000) %>%
  summarise(carrier = "other",total_n=sum(total_n)) -> tmp_other
airlinecount %>%
  filter(total_n > 10000) %>%
  union(tmp_other) -> mergedata

pie(mergedata$total_n,labels=mergedata$carrier, main = "航空公司比例饼状图", col=rainbow(10))
```



(4) 散点图 + 平滑线

```
p <- ggplot(data=flights, mapping = aes(x = distance,y = air_time))
p + geom_point(shape = 1, alpha = 0.1, color = "skyblue") + stat_bin_hex(bins = 30) + geom_smoo
th(method = lm, color ="green") + theme(plot.title = element_text(hjust = 0.5), panel.backgroun
d = element_rect(fill = "#FFBC1717")) + labs(title = "距离飞行时间散点图与平滑线")
```

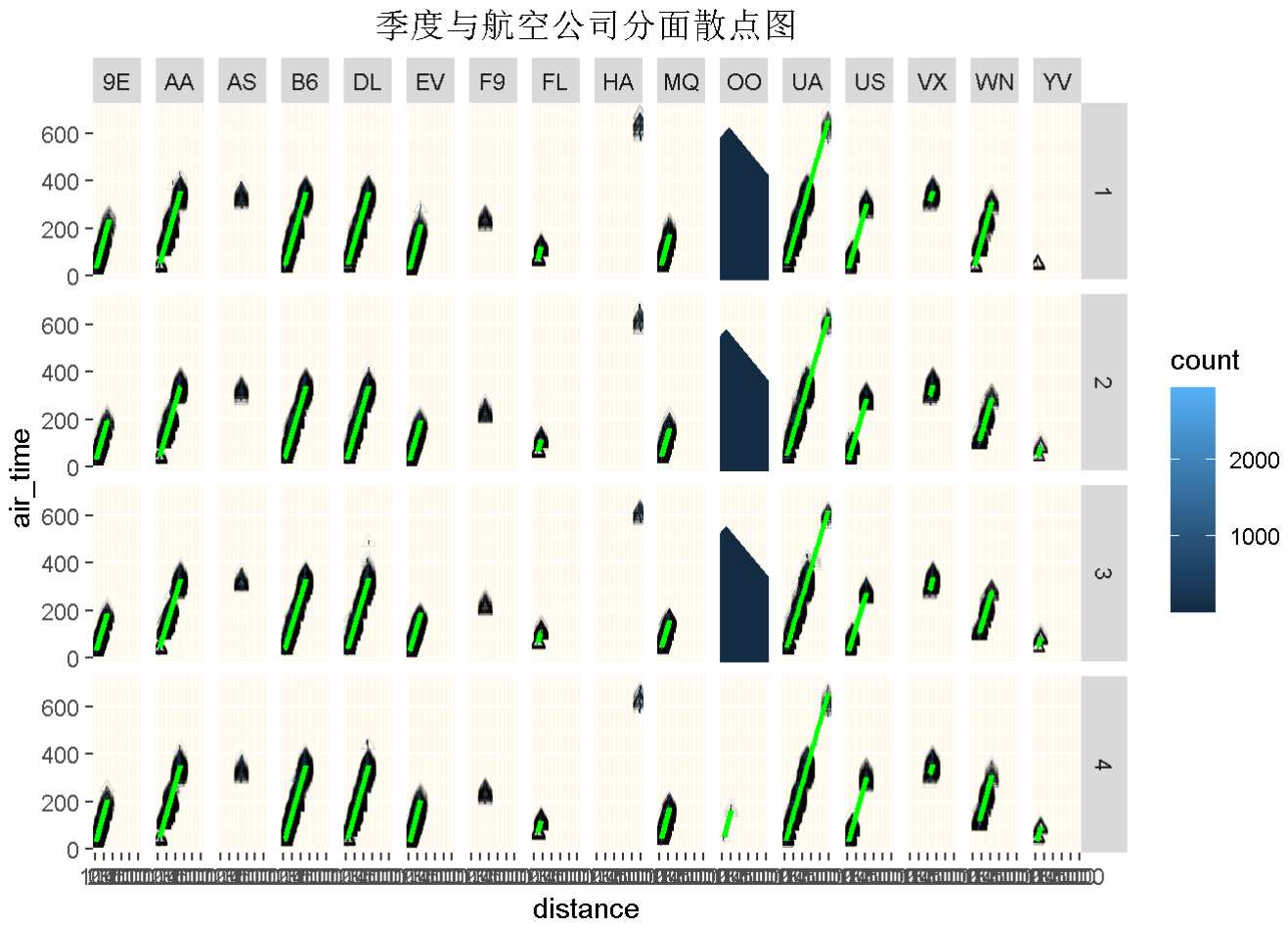
```
## `geom_smooth()` using formula 'y ~ x'
```



(5) 分面散点图

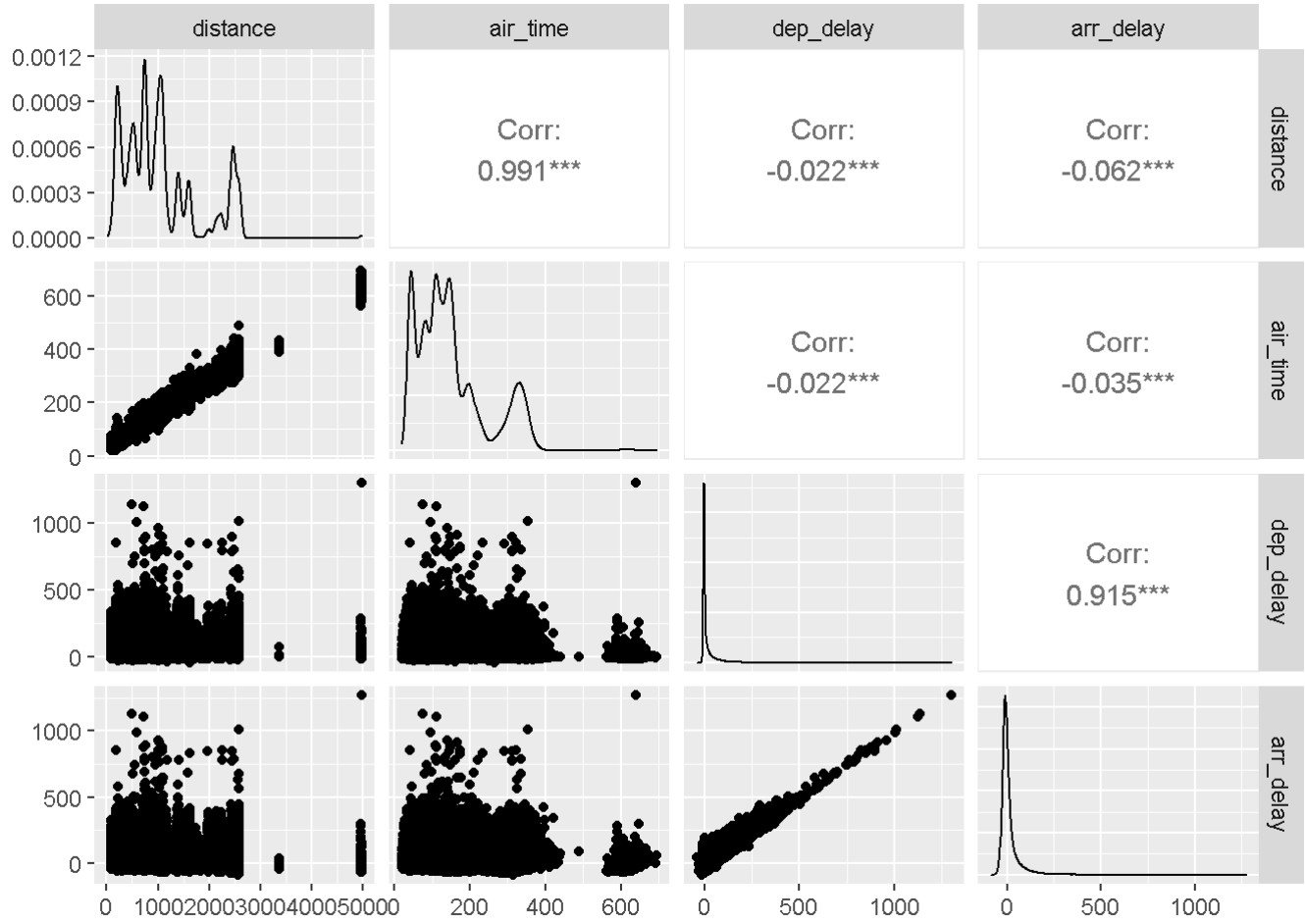
```
p + geom_point(shape = 2, alpha=0.1)+stat_bin_hex(bins = 30)+ stat_smooth(method=lm, color = "green") + facet_grid(season ~ carrier) + theme(plot.title = element_text(hjust = 0.5), panel.background = element_rect(fill = "#FFBC1717")) + labs(title = "季度与航空公司分面散点图")
```

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



(6) 散点图矩阵

```
#pairs.panels(flights[c("distance","air_time","dep_delay","arr_delay")])
ggpairs(data = flights, columns=c("distance","air_time","dep_delay","arr_delay"), main = "散点图矩阵", col=rainbow(16))
```



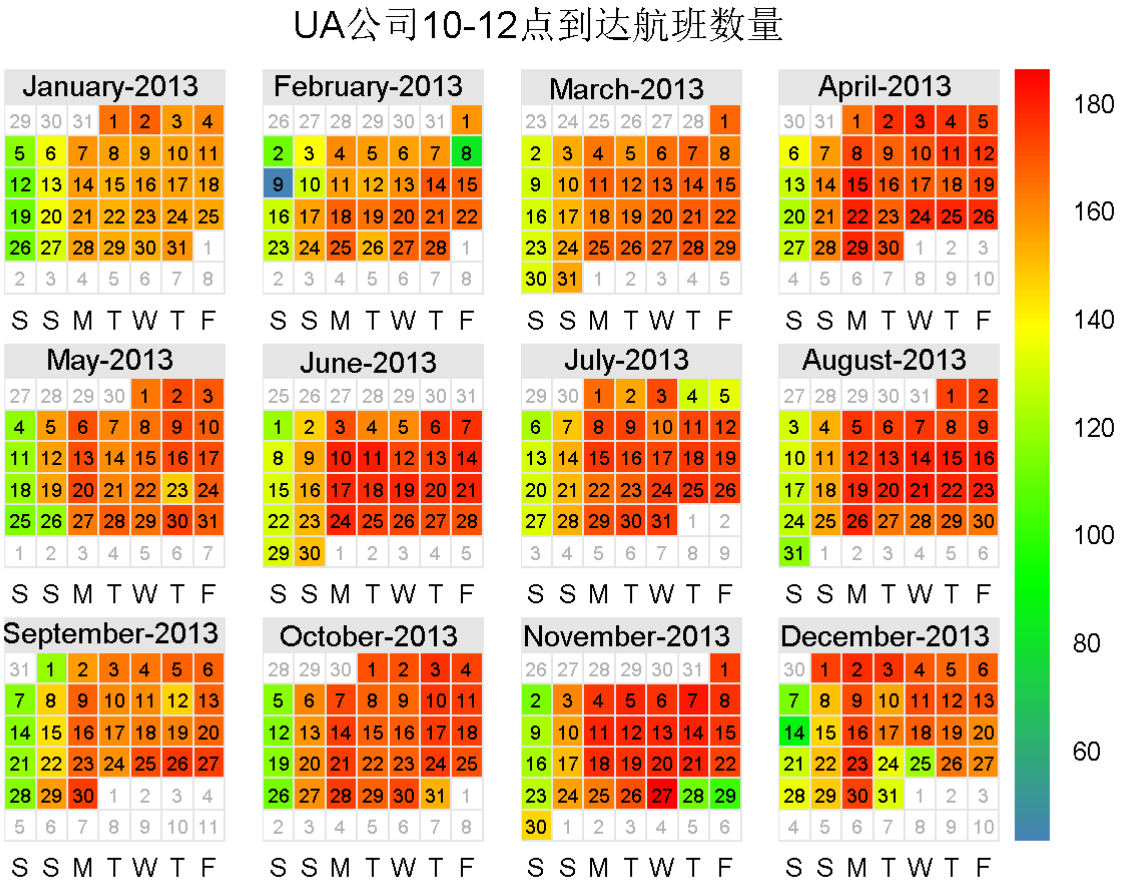
4

(1) UA公司日历图绘制

```
Sys.setlocale(locale = "C")
```

```
## [1] "C"
```

```
flights %>% mutate(date=as.Date(str_c(year,month,day,sep="-")))  %>% filter(arr_time > 1000 | arr_time < 1200) %>% filter(carrier=="UA") %>% group_by(date)%>% summarise(n=n()) -> UA
calendarPlot(UA, pollutant="n", cols = c("steelblue","green", "yellow",  "red"), main = "UA公司10-12点到达航班数量")
```



(2) 延误时间折线图

起飞延误

```
flights %>%
  group_by(month) %>%
  summarise(avg = mean(dep_delay, na.rm = TRUE)) -> departuredelay
departuredelay
```

```
## # A tibble: 12 x 2
##    month   avg
##    <int> <dbl>
##  1     1 10.0
##  2     2 10.8
##  3     3 13.2
##  4     4 13.9
##  5     5 13.0
##  6     6 20.8
##  7     7 21.7
##  8     8 12.6
##  9     9  6.72
## 10    10  6.24
## 11    11  5.44
## 12    12 16.6
```

到达延误

```
flights %>%
  group_by(month) %>%
  summarise(avg = mean(arr_delay, na.rm = TRUE)) -> arrivedelay
arrivedelay
```

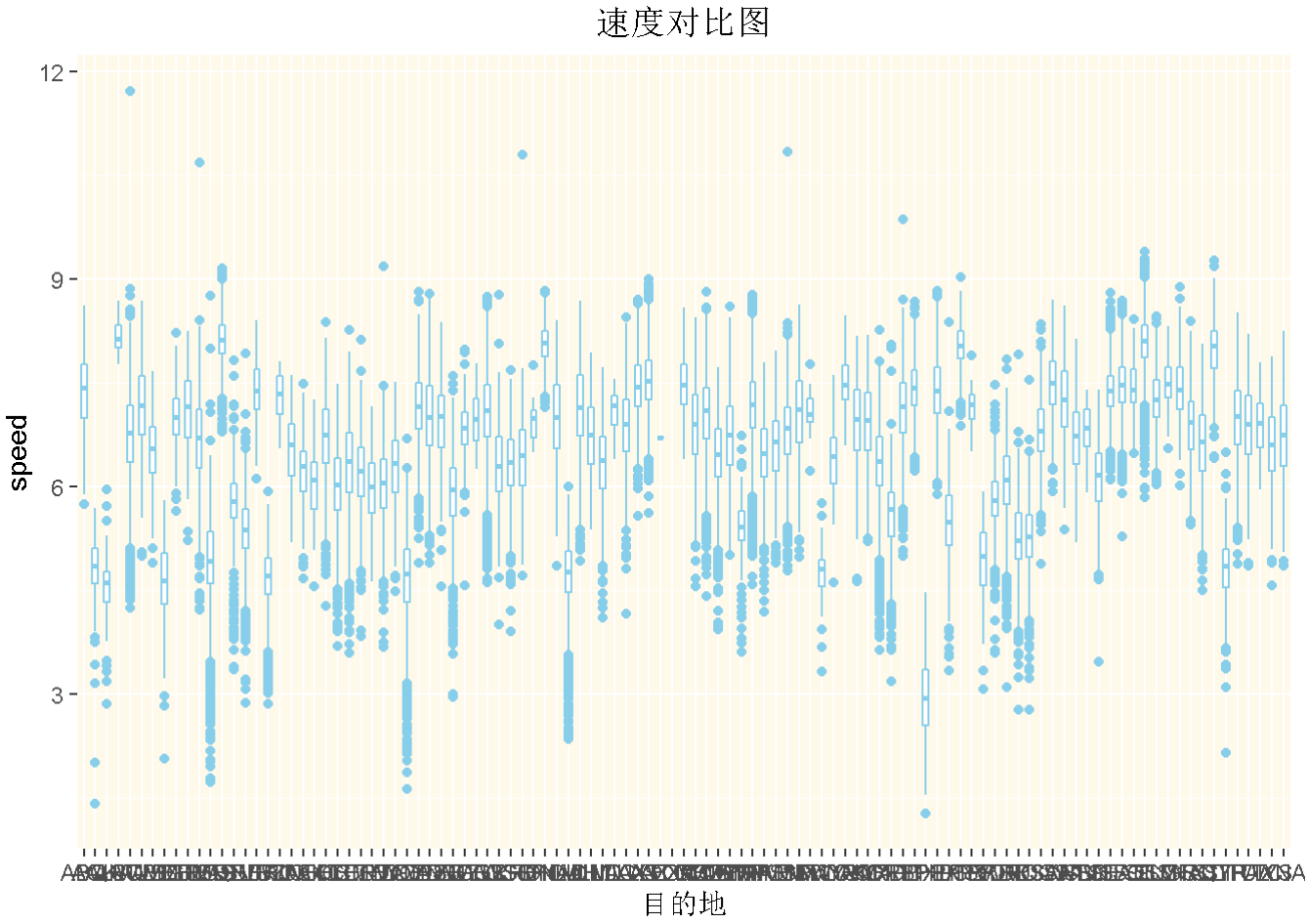
```
## # A tibble: 12 x 2
##    month    avg
##    <int>  <dbl>
##  1     1  6.13
##  2     2  5.61
##  3     3  5.81
##  4     4 11.2
##  5     5  3.52
##  6     6 16.5
##  7     7 16.7
##  8     8  6.04
##  9     9 -4.02
## 10    10 -0.167
## 11    11  0.461
## 12    12 14.9
```

5

(1) 可疑速度查看

箱线图

```
flights <- flights %>%
    mutate(speed = distance / air_time)
seasonalplot <- ggplot(data = flights, mapping = aes(x = as.factor(dest), y = speed), fill = At
tribute)
seasonalplot + geom_boxplot(width = 0.5, position = position_dodge(0.9), color = "skyblue") + l
abs(title = "速度对比图", x= "目的地") + theme(plot.title=element_text(hjust=0.5),panel.backgro
und = element_rect(fill = "#FFBC1717"))
```



每小时超过600km

```
Abnormal1 = flights %>% filter(speed >= 10)
#Abnormal1 <- Abnormal1 %>% select(dest, speed)
Abnormal1  %>% filter(Valid = TRUE) %>% group_by(dest) %>% summarise(n())
```

```
## # A tibble: 4 x 2
##   dest  `n()`
##   <chr> <int>
## 1 ATL       1
## 2 BNA       1
## 3 GSP       1
## 4 MSP       1
```

说明BOS和DCA作为目的地飞机航班速度容易快的可疑

另一种异常判定方法

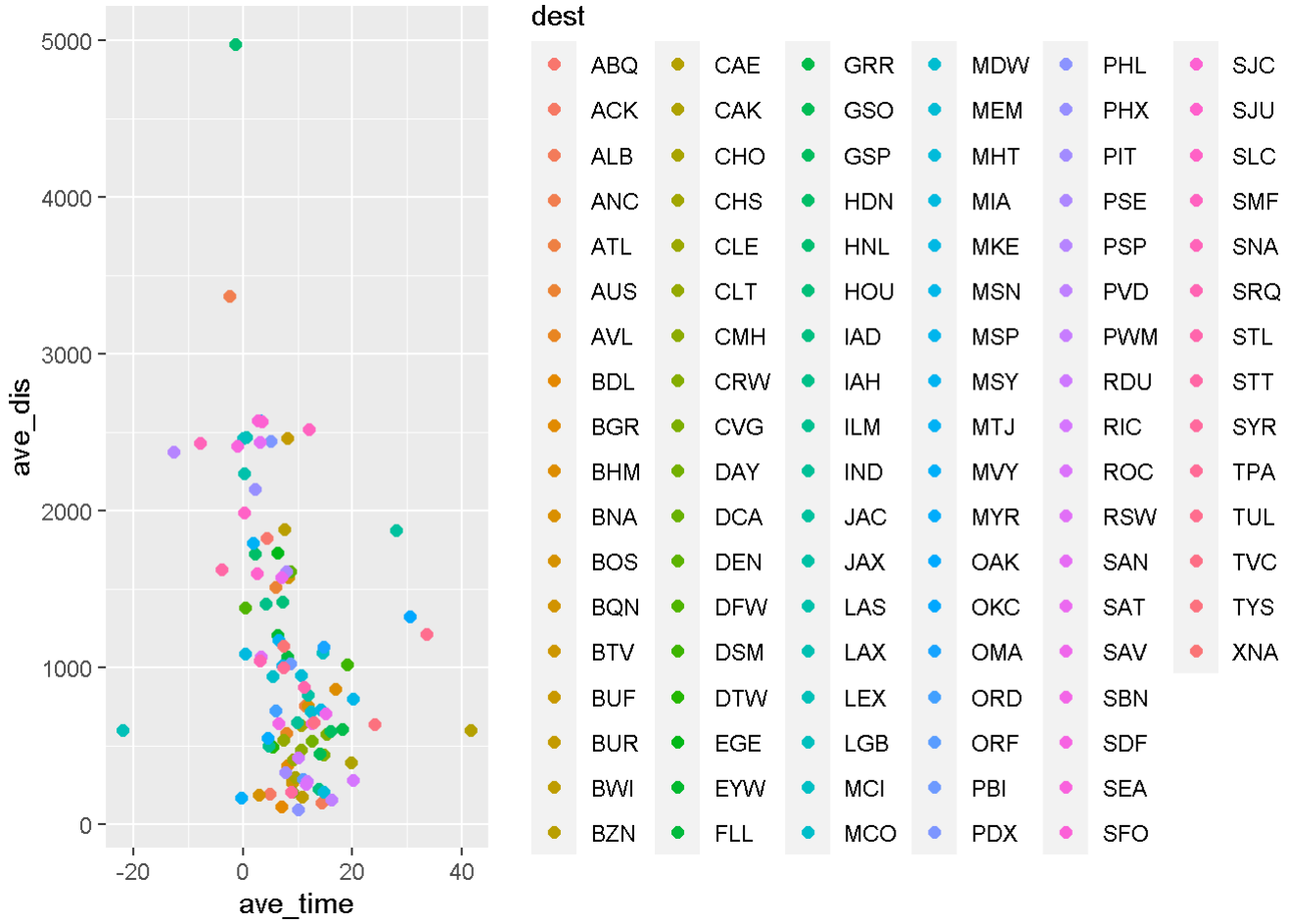
```
flights %>% summarise(variance = var(speed, na.rm = TRUE)) -> variance
flights %>% summarise(ave = mean(speed, na.rm = TRUE)) -> average
variance = as.numeric(variance)
average = as.numeric(average)
Abnormal2 = flights %>% filter(speed >= (average + 3 * sqrt(variance)))
Abnormal2
```

```
## # A tibble: 5 x 21
##    year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##   <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
## 1  2013     1    12     1559           1600        -1     1849           1917
## 2  2013     3    23     1914           1910         4     2045           2043
## 3  2013     5    13     2040           2025        15     2225           2226
## 4  2013     5    25     1709           1700         9     1923           1937
## 5  2013     7     2     1558           1513        45     1745           1719
## # ... with 13 more variables: arr_delay <dbl>, carrier <chr>, flight <int>,
## #   tailnum <chr>, origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>,
## #   hour <dbl>, minute <dbl>, time_hour <dttm>, season <dbl>, speed <dbl>
```

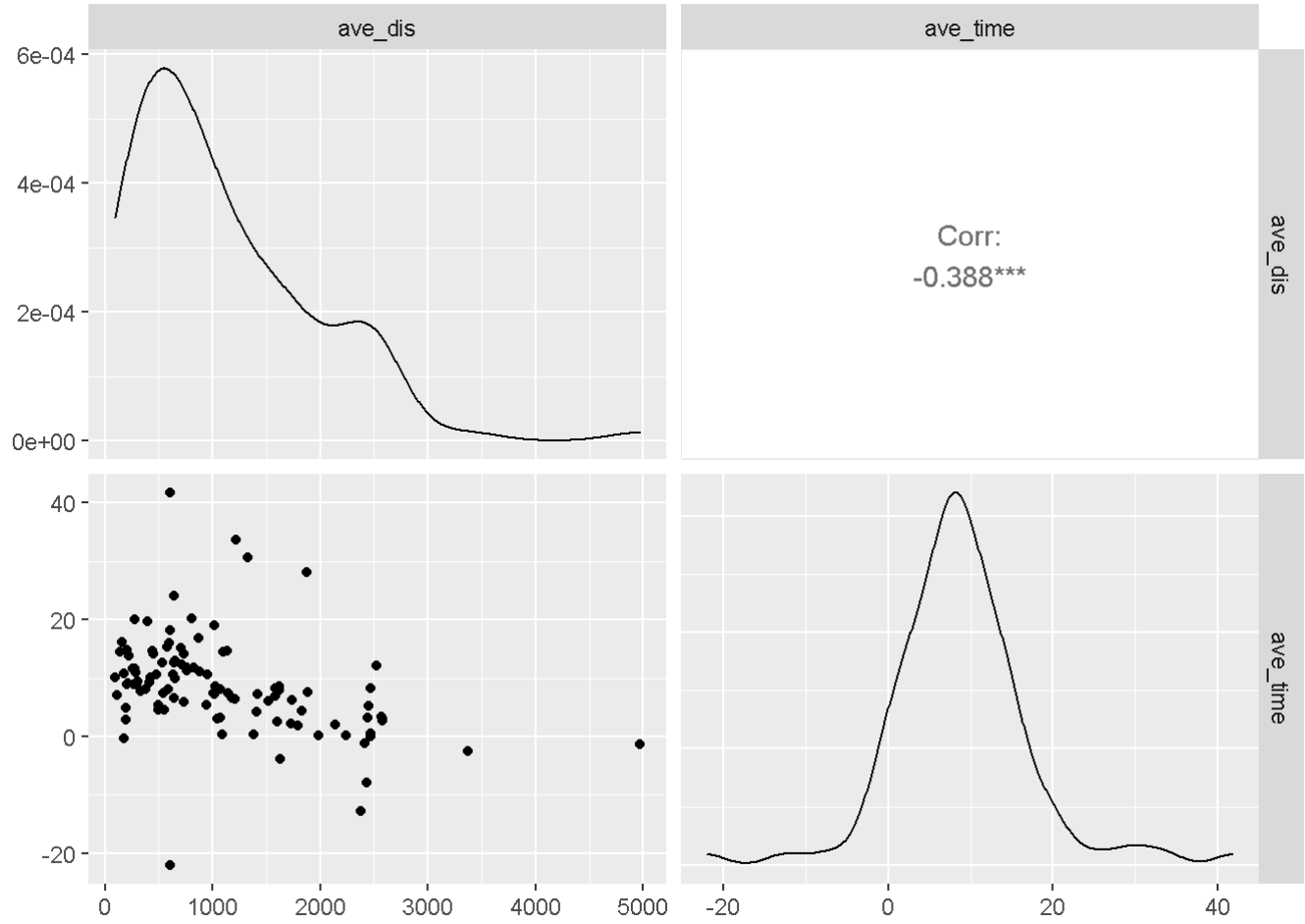
是有些航班快的可疑

(2) 相关性判断

```
flights %>%  drop_na(distance) %>% drop_na(arr_delay) %>% group_by(dest) %>% summarise(ave_dis
= mean(distance),ave_time=mean(arr_delay)) -> aver
p <- ggplot(data = aver, mapping = aes(ave_time, ave_dis, color = dest))
p + geom_point(size = 2.0, shape = 16)
```




```
ggpairs(data = aver, columns=c("ave_dis","ave_time"))
```



(3) 取消航班与平均延误时间的关系

每日取消航班

```
flights <- flights %>% mutate(date=as.Date(str_c(year,month,day,sep="-")))
cancel <- flights %>% group_by(date) %>% summarise(cancelcount = sum(is.na(dep_time)) )
cancel
```

```
## # A tibble: 365 x 2
##    date       cancelcount
##    <date>           <int>
##  1 2013-01-01           4
##  2 2013-01-02           8
##  3 2013-01-03          10
##  4 2013-01-04           6
##  5 2013-01-05           3
##  6 2013-01-06           1
##  7 2013-01-07           3
##  8 2013-01-08           4
##  9 2013-01-09           5
## 10 2013-01-10           3
## # ... with 355 more rows
```

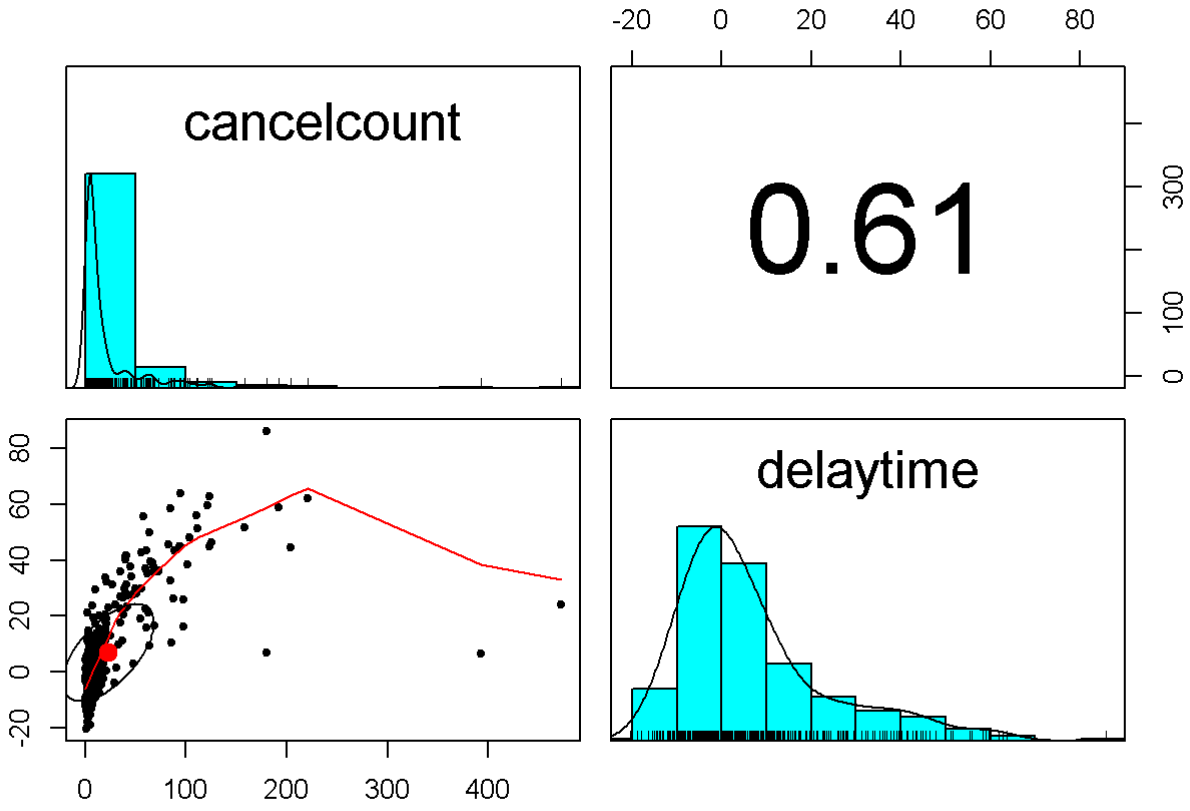
每日平均延误时间

```
delaytime <- flights %>% group_by(date) %>% summarise(delaytime = mean(arr_delay, na.rm = TRUE
))
delaytime
```

```
## # A tibble: 365 x 2
##    date       delaytime
##    <date>         <dbl>
##  1 2013-01-01    12.7
##  2 2013-01-02    12.7
##  3 2013-01-03     5.73
##  4 2013-01-04    -1.93
##  5 2013-01-05    -1.53
##  6 2013-01-06     4.24
##  7 2013-01-07    -4.95
##  8 2013-01-08    -3.23
##  9 2013-01-09    -0.264
## 10 2013-01-10    -5.90
## # ... with 355 more rows
```

相关性

```
m <- merge(cancel, delaytime, by.x = "date", by.y = "date")
pairs.panels(m[c("cancelcount","delaytime")])
```



存在一定的正相关性

(4) 延误问题判断

```
delay <- flights %>%  group_by(carrier, dest) %>% summarise(n())
```

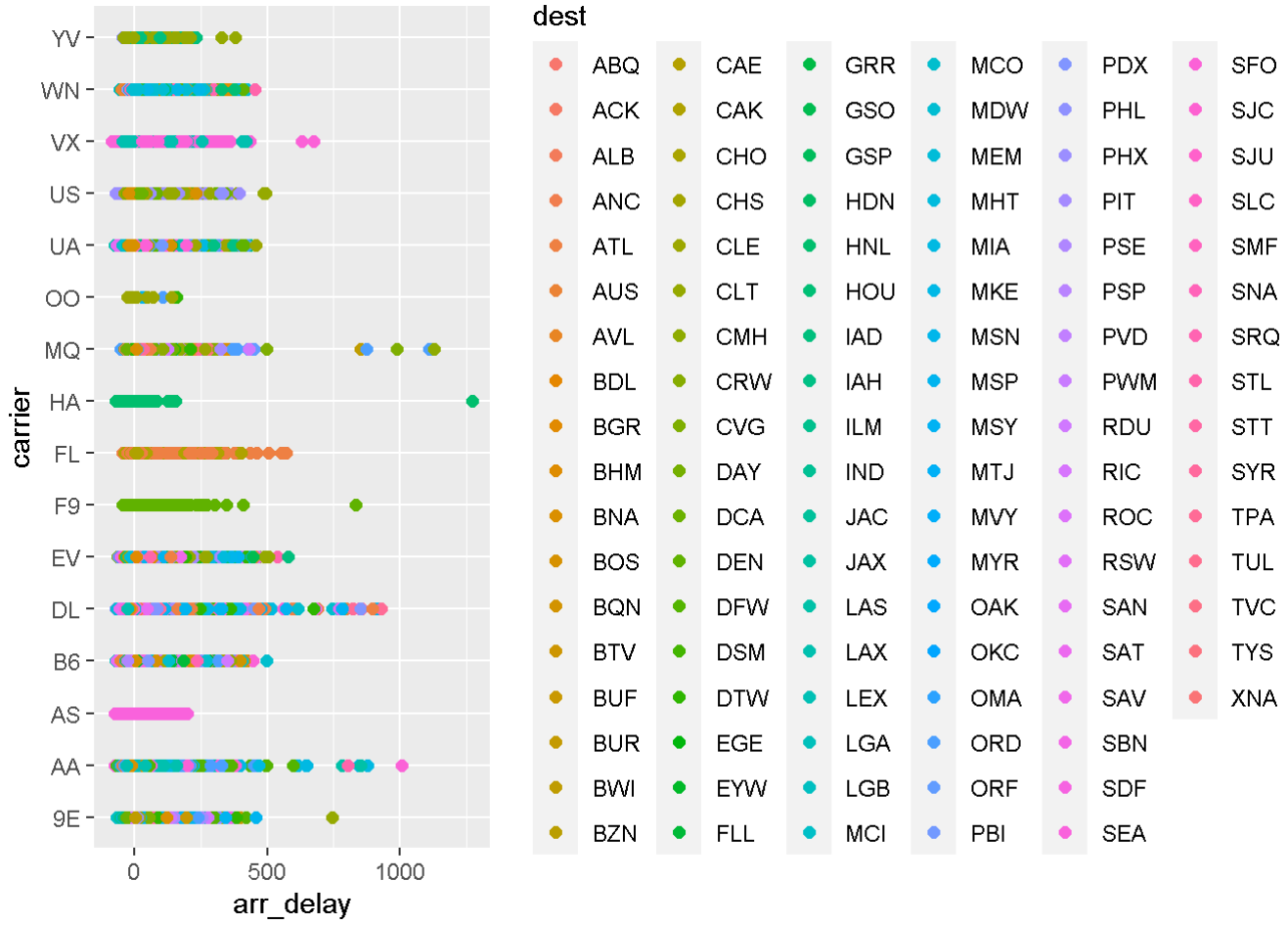
```
## `summarise()` has grouped output by 'carrier'. You can override using the
## `.groups` argument.
```

```
delay
```

```
## # A tibble: 314 x 3
## # Groups:   carrier [16]
##    carrier dest  `n()`
##    <chr>   <chr> <int>
##  1 9E      ATL      59
##  2 9E      AUS       2
##  3 9E      AVL      10
##  4 9E      BGR       1
##  5 9E      BNA     474
##  6 9E      BOS     914
##  7 9E      BTV       2
##  8 9E      BUF     833
##  9 9E      BWI     856
## 10 9E      CAE       3
## # ... with 304 more rows
```

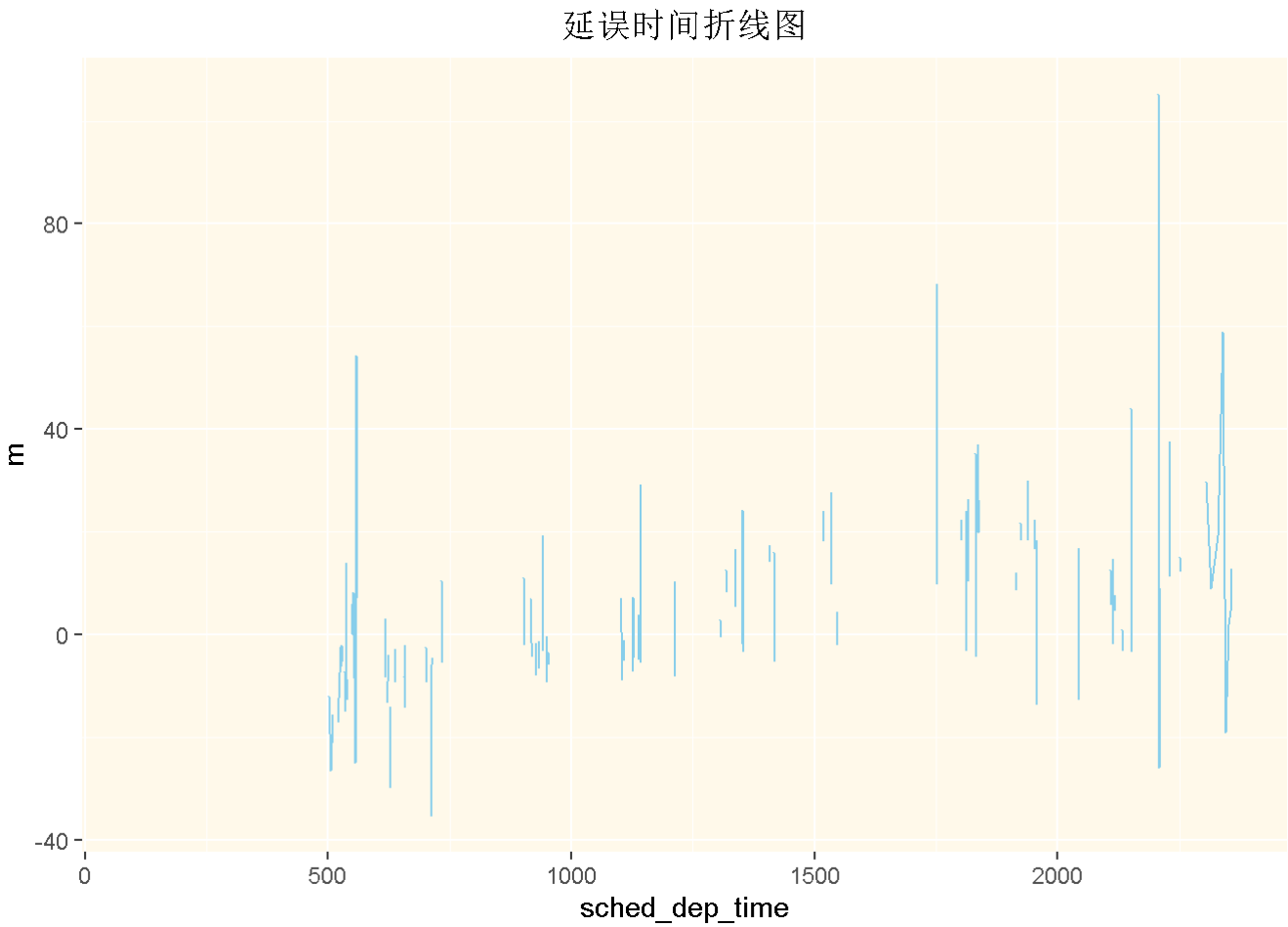
散点图

```
p <- ggplot(data = flights, mapping = aes(arr_delay, carrier, color = dest))
p + geom_point(size = 2.0, shape = 16)
```



(5) 搭乘飞机时间选择

```
flights %>% group_by(sched_dep_time) %>% summarise(m = mean(arr_delay)) -> delaybytime
p <- ggplot(data = delaybytime, mapping = aes(x=sched_dep_time, y = m))
p + geom_line(stat="identity",color="skyblue") +  labs(title = "延误时间折线图") + theme(plot.t
itle=element_text(hjust=0.5),panel.background = element_rect(fill = "#FFBC1717"))
```



选择早晨和中午坐飞机