615-final

2022-12-14

```
library(data.table)
library(ggplot2)
library(stringr)
library(RColorBrewer)
library(flextable)
library(dplyr)
```

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

```
## The following objects are masked from 'package:data.table':
##
##     between, first, last
```

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

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

```
library(scales)
library(data.table)
library(ggplot2)
library(stringr)
library(RColorBrewer)
library(flextable)
library(dplyr)
library(scales)
library(ggridges)
library(viridis)
```

```
## Loading required package: viridisLite
```

```
##
## Attaching package: 'viridis'
```

```
## The following object is masked from 'package:scales':
##
##     viridis_pal
```

```
library(hrbrthemes)
```

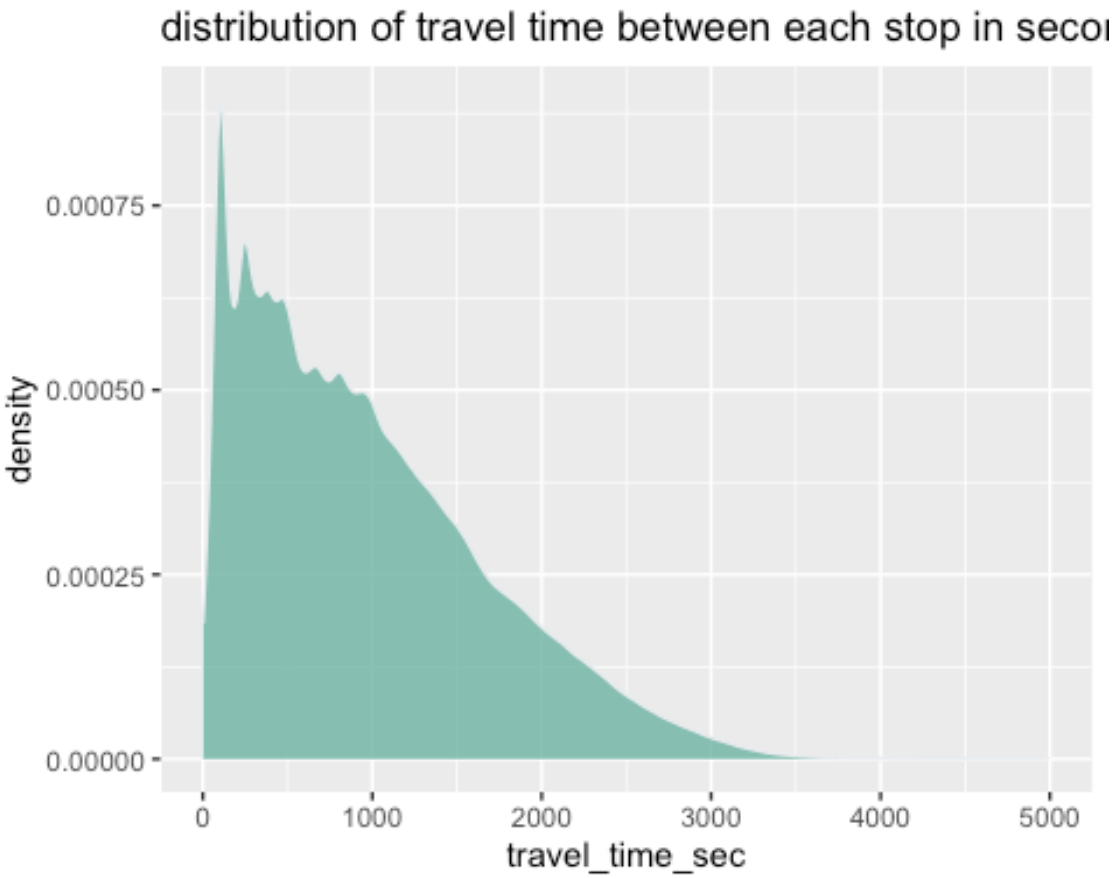
```
## NOTE: Either Arial Narrow or Roboto Condensed fonts are required to
use these themes.
```

```
##       Please use hrbrthemes::import_roboto_condensed() to install Ro
boto Condensed and

##       if Arial Narrow is not on your system, please see https://bit.
ly/arialnarrow

options(width=90)

data.files<-list.files('subway',pattern='csv$',full=T)
data.files

## [1] "subway/2022-Q1_HRTravelTimes.csv" "subway/2022-Q2_HRTravelTimes.
csv"
## [3] "subway/2022-Q3_HRTravelTimes.csv" "subway/HRTravelTimesQ4_21.cs
v"

subway<-lapply(setNames(,data.files),function(x)
{
   fread(x)->temp
   temp[,fromFile:=x]
   temp[,day:=mday(service_date)]
   temp<-temp[day %in% 11:17]
   temp<-temp[route_id %in% c('Red')]
})



rbindlist(subway)->subway
subway[,startTime:=as.ITime(start_time_sec)]
subway[,endTime:=as.ITime(end_time_sec)]



fread('stops.txt')->sites
setNames(sites[,stop_name],sites[,stop_id])->sites.name

subway[,from_stop_name:=sites.name[as.character(from_stop_id)]]
subway[,to_stop_name:=sites.name[as.character(to_stop_id)]]

subway$start_hours <- as.numeric(substr(subway$startTime,1,2))

bus.data.files<-list.files('Bus',pattern='csv$',full=T)

bus<-lapply(setNames(,bus.data.files),function(x)
{
fread(x)->temp
})

bus<-rbindlist(bus)
bus<-bus[route_id %in% c('93')]

bus[,day:=mday(service_date)]
```

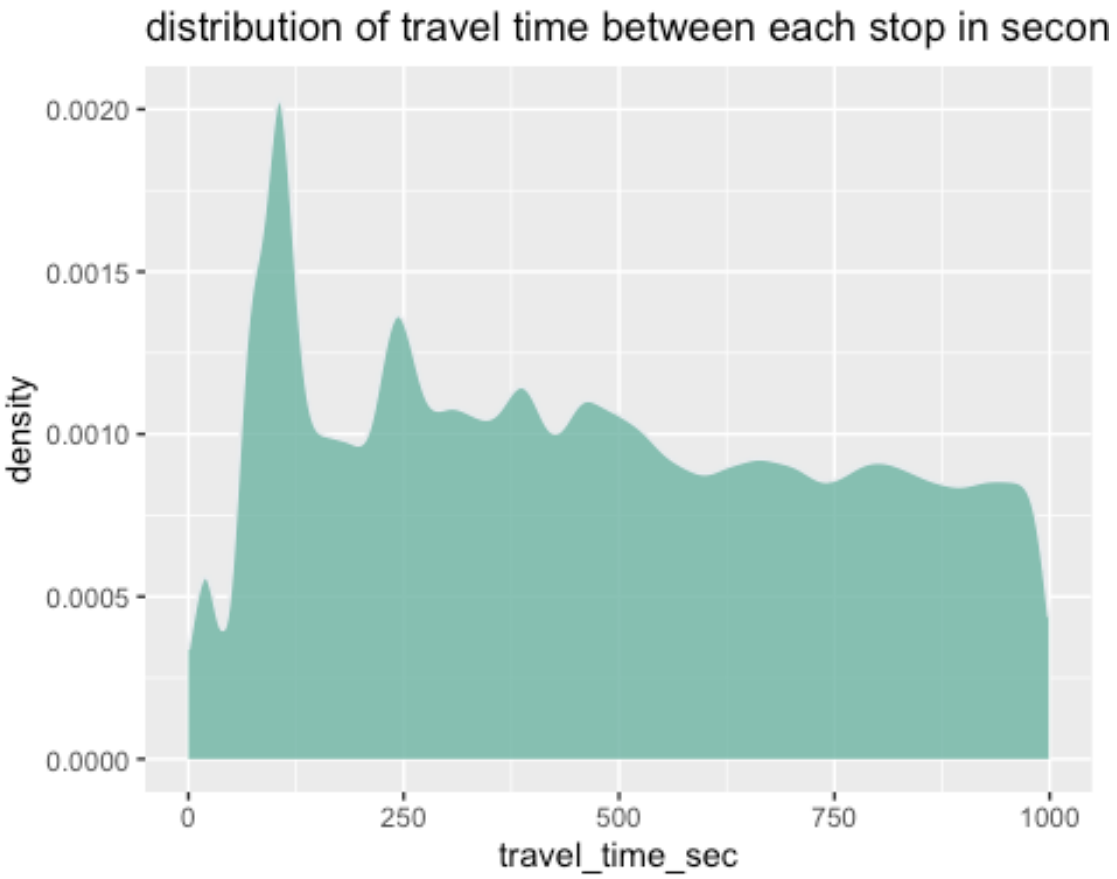
```
bus<-bus[day %in% 11:17]

#bus<-bus[,.(service_date,route_id,stop_id,point_type,scheduled,actual,
day)]
bus<-na.omit(bus)

bus[,timeDiff:=as.numeric(actual-scheduled)]
bus[,sum(abs(timeDiff)<=1800)/.N]

## [1] 0.9990997

bus<-bus[abs(timeDiff)<=1800]

fread('stops.txt')->sites
setNames(sites[,stop_name],sites[,stop_id])->sites.name

bus[,stop_name:=sites.name[as.character(stop_id)]]
bus$start_hours <- as.numeric(substr(bus$actual,12,13))
```

Subway

```
library(hrbrthemes)
subway %>%
  filter( travel_time_sec<5000) %>%
  ggplot( aes(x=travel_time_sec)) +
    geom_density(fill="#69b3a2", color="#e9ecef", alpha=0.8)+
  ggtitle("distribution of travel time between each stop in second")
```

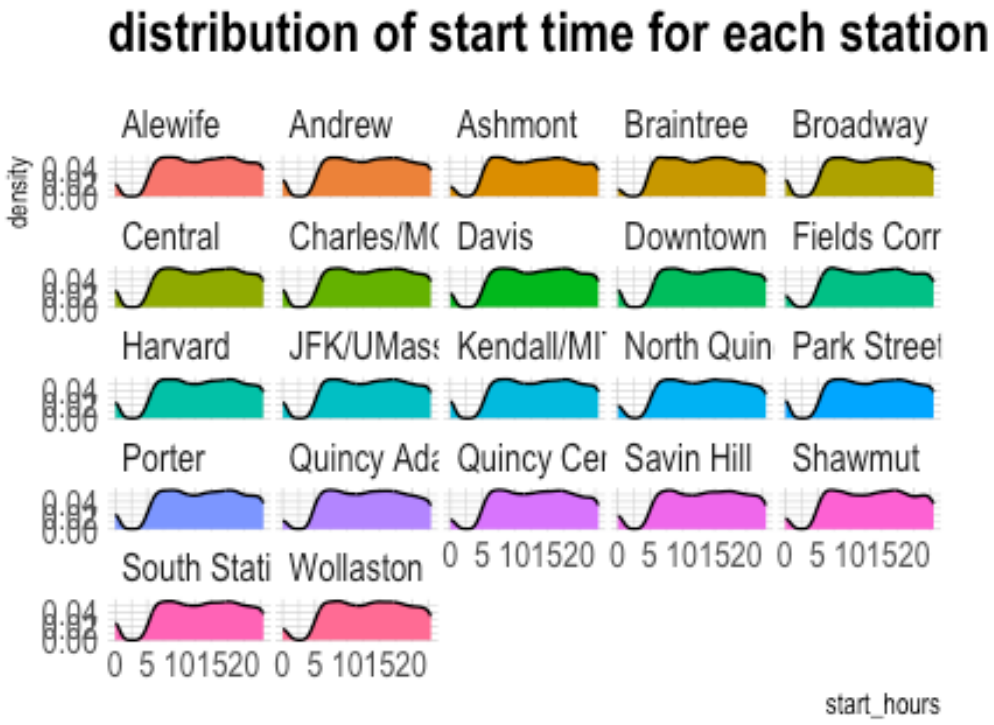



```
library(hrbrthemes)
subway %>%
  filter( travel_time_sec<1000) %>%
  ggplot( aes(x=travel_time_sec)) +
    geom_density(fill="#69b3a2", color="#e9ecef", alpha=0.8)+
  ggtitle("distribution of travel time between each stop in second(zoom
 in)")
```



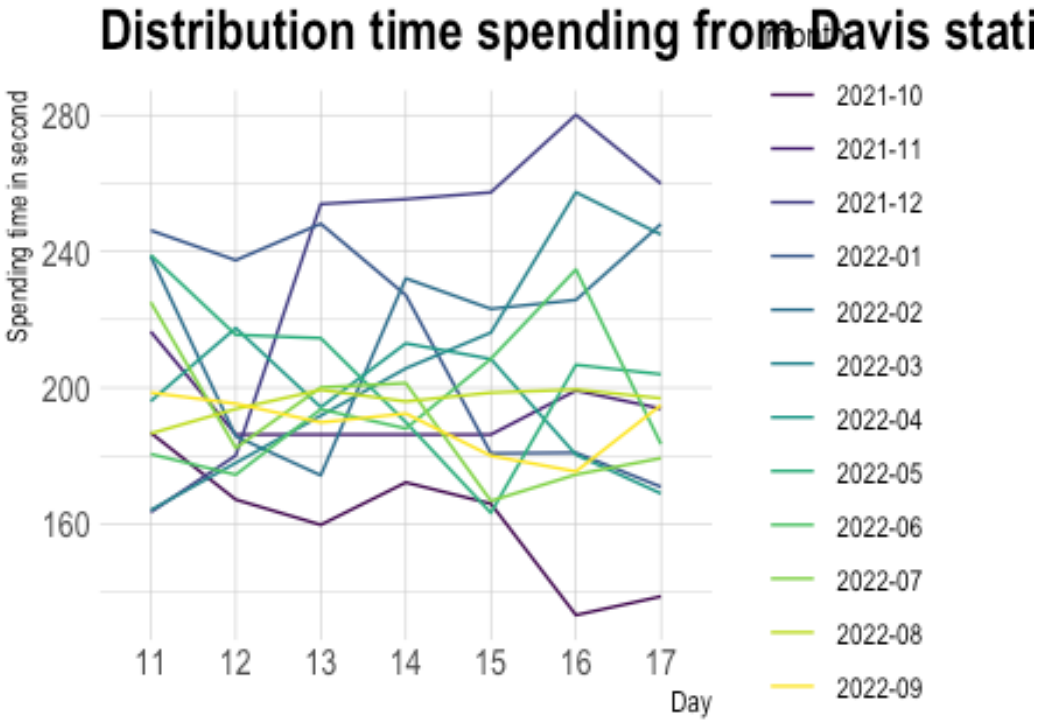
After zoom in the graph we can see that the travel time of around 2 minutes between each station reach the peak.

```
ggplot(data=subway, aes(x=start_hours, group=from_stop_name, fill=from_
stop_name)) +
    geom_density(adjust=1.5) +
    theme_ipsum() +
    facet_wrap(~from_stop_name) +
    theme(
      legend.position="none",
      panel.spacing = unit(0.1, "lines"),
      axis.ticks.x=element_blank()
    )+ggtitle("distribution of start time for each station")
```



Here is the graph shows the distribution of time for the subway reached for every station. In general 7am and 6pm is the peak of time for the subway reach the station.

```
Davis_Alewife<- subway[subway$from_stop_id == 70064&subway$to_stop_id =
= 70061,c("service_date","travel_time_sec")] %>%
group_by(service_date) %>%
summarise(ave_time = mean(travel_time_sec))

Davis_Alewife$month <- substr(Davis_Alewife$service_date,1,7)
Davis_Alewife$service_date <- substr(Davis_Alewife$service_date,9,10)

Davis_Alewife %>%
  ggplot( aes(x=service_date, y=ave_time, group=month, color=month)) +
    geom_line() +
    scale_color_viridis(discrete = TRUE) +
  labs(x = "Day", y = "Spending time in second", title="Distribution ti
me spending from Davis station to Alewife station" ) +
    theme_ipsum()
```

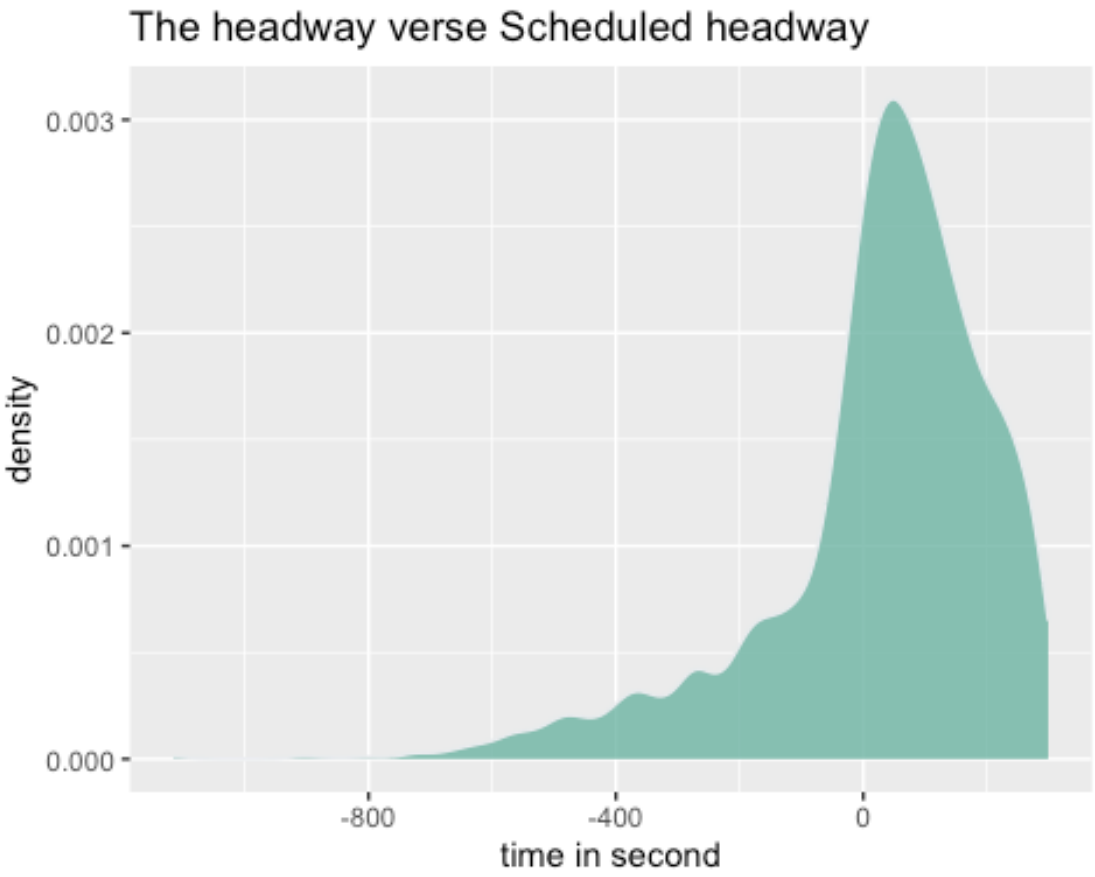



```
rm(Davis_Alewife)
```

In order to see if the weather affect the subway travel time. I pick the travel time from 11th to 17th for every month from Davis station to Alewife station. The travel time from 2022-02 to 2022-09 is quite stable. The the travel time of 2021-12 reach the peak.The most likely explaination is the cold weather of Boston in winter affect the subway travel time between each station.

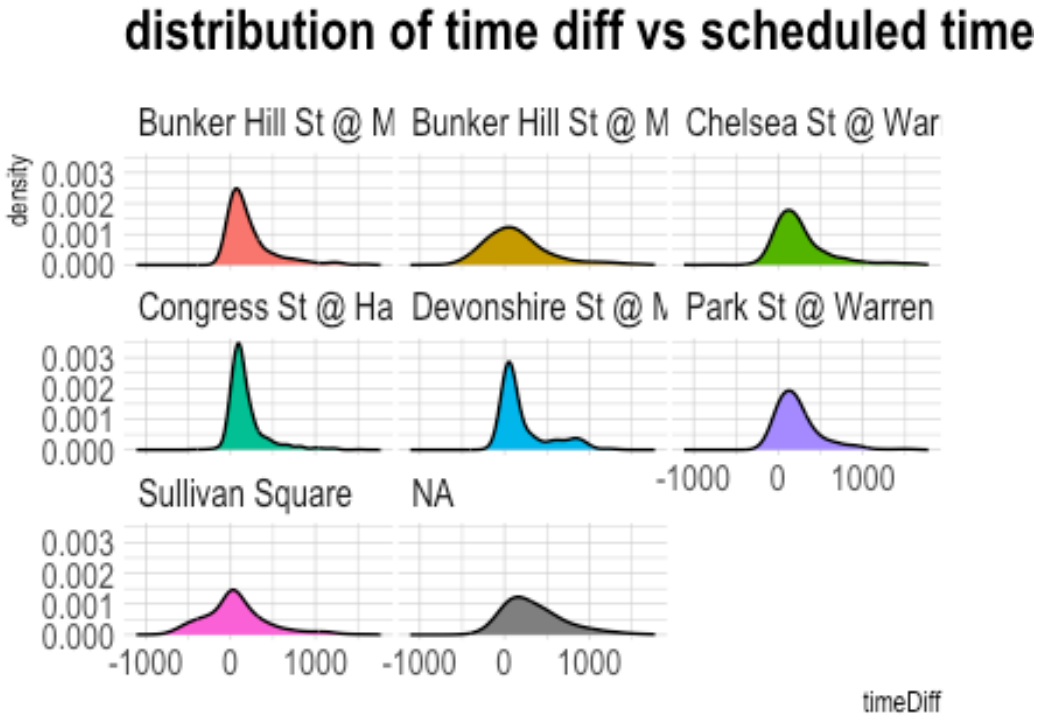
Bus

```
bus %>%
  filter(timeDiff<300 ) %>%
  ggplot( aes(x=timeDiff)) +
    geom_density(fill="#69b3a2", color="#e9ecef", alpha=0.8)+labs(x='ti
me in second')+
  ggtitle('The headway verse Scheduled headway')
```



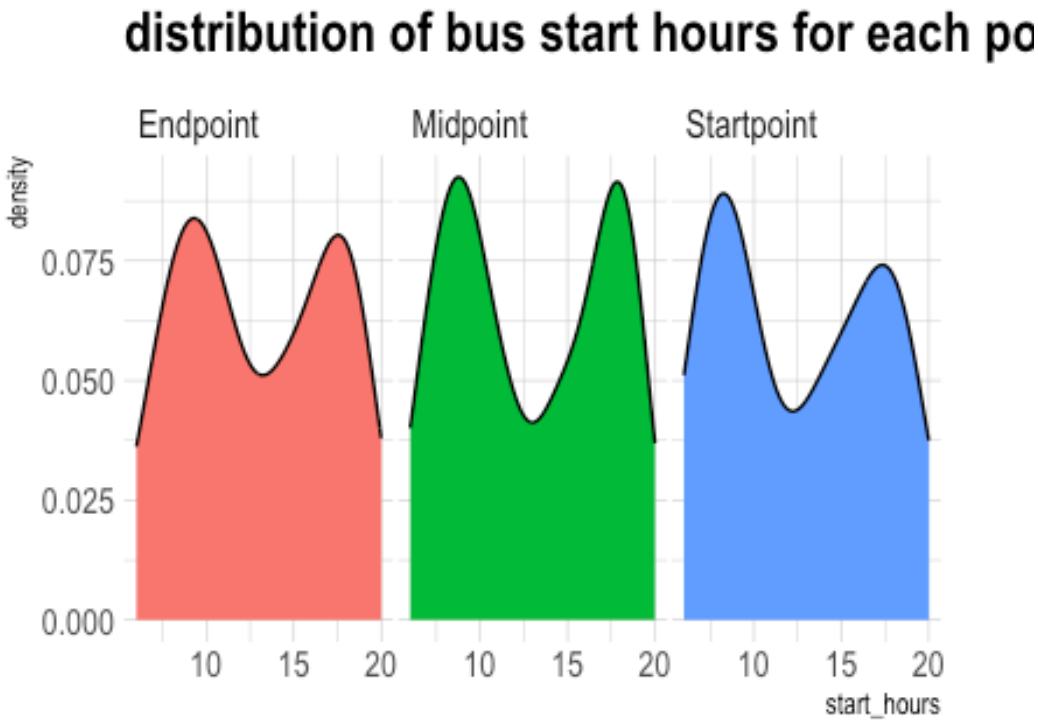
Here is the graph shows the distribution of how much the early and late bus headway time. Anything greater than 0 means hows much times in second did the bus late for the scheduled headway time.As we can see from the graph, most the bus is late for the scheduled headway time, and around 1 minute reach the peak.

```
ggplot(data=bus, aes(x=timeDiff, group=stop_name, fill=stop_name)) +
    geom_density(adjust=1.5) +
    theme_ipsum() +
    facet_wrap(~stop_name) +
    theme(
      legend.position="none",
      panel.spacing = unit(0.1, "lines"),
      axis.ticks.x=element_blank()
    )+ggtitle("distribution of time diff vs scheduled time for each bus
 station")
```



Here is the distribution of time difference vs scheduled time for each bus station.

```
ggplot(data=bus, aes(x=start_hours, group=point_type, fill=point_type))
 +
    geom_density(adjust=1.5) +
    theme_ipsum() +
    facet_wrap(~point_type) +
    theme(
      legend.position="none",
      panel.spacing = unit(0.1, "lines"),
      axis.ticks.x=element_blank()
    )+ggtitle("distribution of bus start hours for each point type")
```

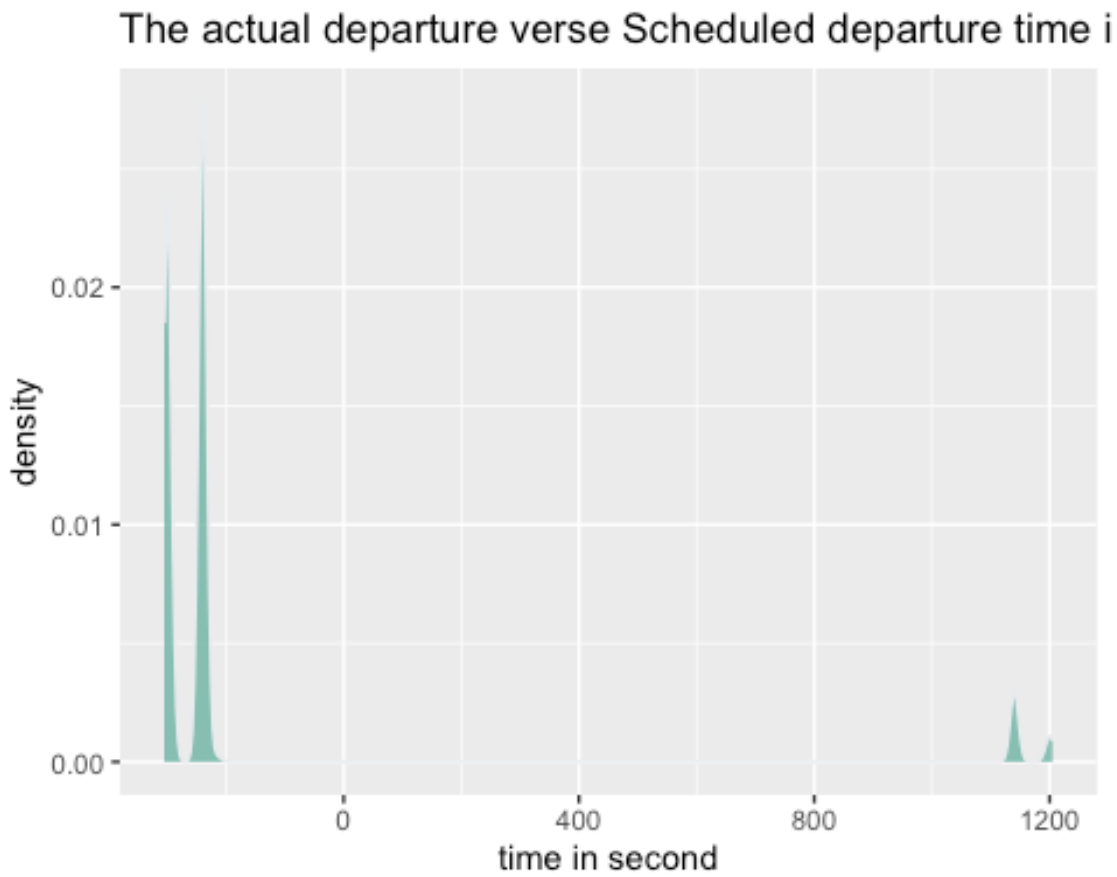


Here is the distribution of time diff vs scheduled time for each point type. We can conclude from the graph that around 7am and 5pm the bus start time reach the peak. ferry<- read.csv("~/Desktop/615 final/ferry.csv",header = T)

Ferry

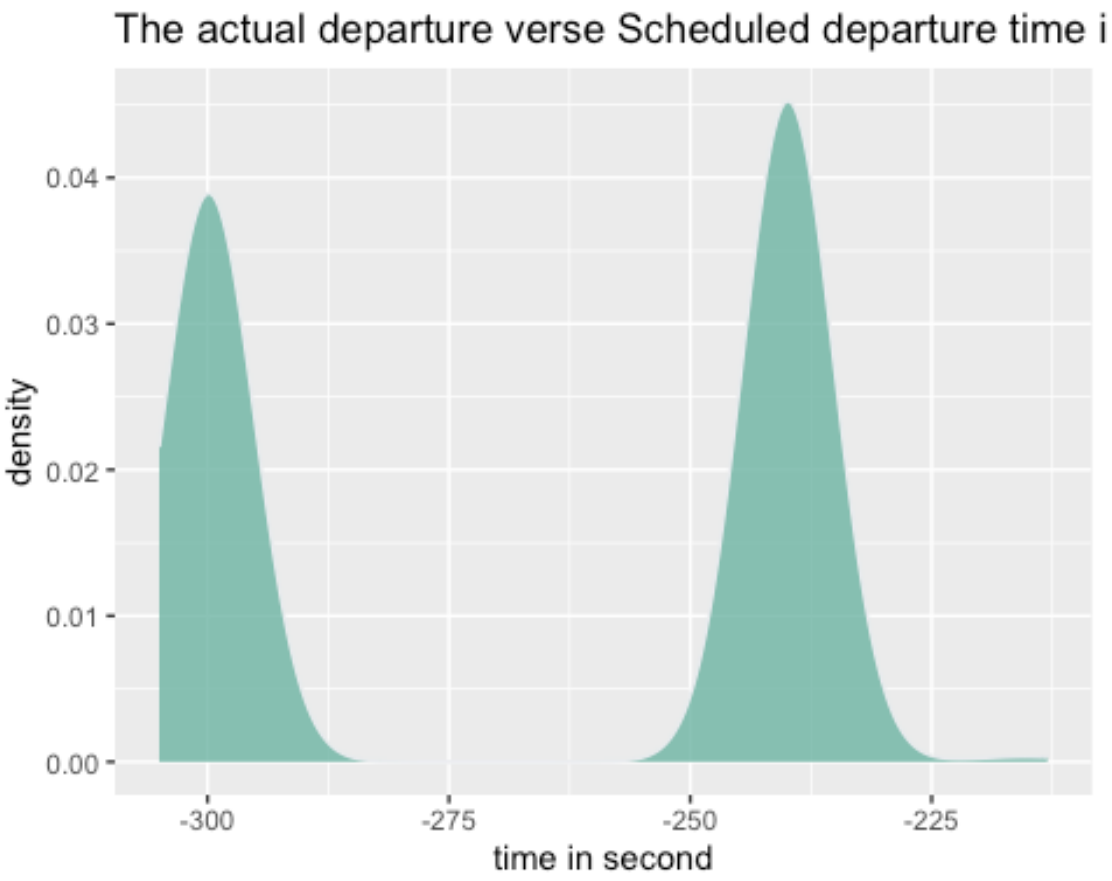
```
Ferry_data <-read.csv("~/Desktop/615 final/ferry.csv",header = T)

Ferry_data$day<-as.numeric(mday(Ferry_data$service_date))
Ferry_data$year<-as.numeric(substr(Ferry_data$service_date,1,4))
Ferry_data<-Ferry_data %>% filter(day %in% 11:17) %>% filter(year==2020)
Ferry_data <- Ferry_data[order(Ferry_data$service_date),]

Ferry_data$departure_diff<-60*((as.numeric(substr(Ferry_data$actual_dep
arture,12,13)))-
                                 (as.numeric(substr(Ferry_data$mbta_sch
ed_departure,12,13))))+
  ((as.numeric(substr(Ferry_data$actual_departure,15,16)))-
                                 (as.numeric(substr(Ferry_data$mbta_sch
ed_departure,15,16))))

Ferry_data$arrival_diff<-60*((as.numeric(substr(Ferry_data$actual_arriv
al,12,13)))-
                                 (as.numeric(substr(Ferry_data$mbta_sch
```

```
ed_arrival,12,13))))+
  ((as.numeric(substr(Ferry_data$actual_arrival,15,16)))-
                                  (as.numeric(substr(Ferry_data$mbta_sch
ed_arrival,15,16))))

Ferry_data$departure_time<-as.numeric(substr(Ferry_data$actual_departur
e,12,13))
Ferry_data$arrival_time<-as.numeric(substr(Ferry_data$actual_arrival,12,
13))

Ferry_data%>%
  ggplot( aes(x=departure_diff)) +
    geom_density(fill="#69b3a2", color="#e9ecef", alpha=0.8)+labs(x='ti
me in second')+
  ggtitle('The actual departure verse Scheduled departure time in minut
e')

## Warning: Removed 221 rows containing non-finite values (`stat_densit
y()`).
```

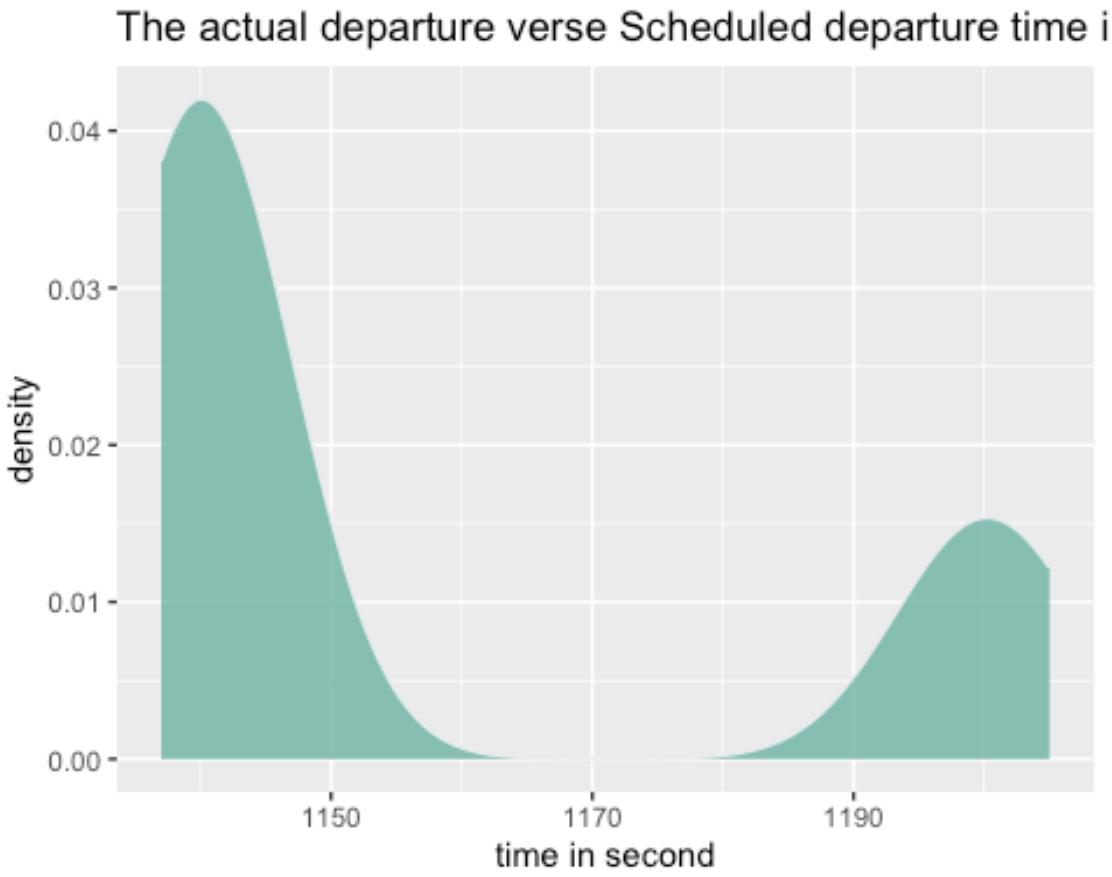



```
  Ferry_data%>%
  filter( departure_diff<0) %>%
  ggplot( aes(x=departure_diff)) +
    geom_density(fill="#69b3a2", color="#e9ecef", alpha=0.8)+labs(x='ti
```

```
me in second')+
  ggtitle('The actual departure verse Scheduled departure time in minut
e')
```

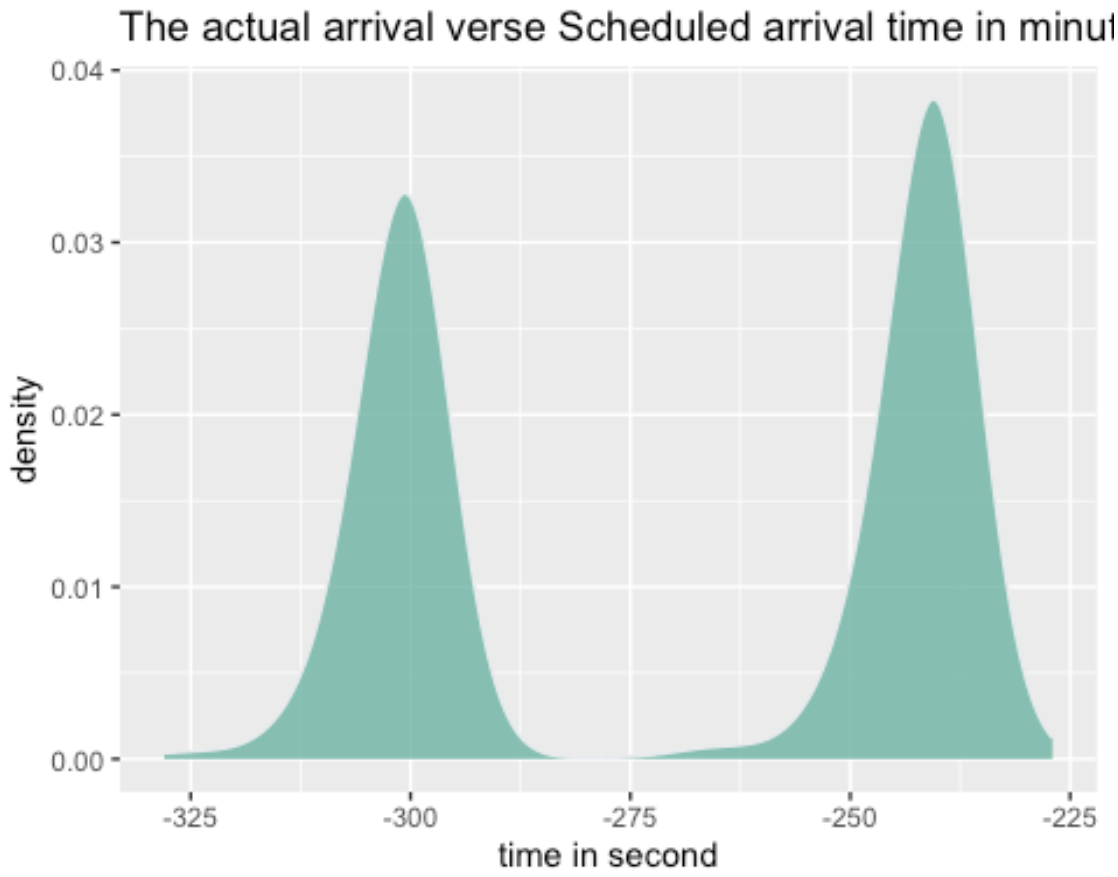



```
  Ferry_data%>%
  filter( departure_diff>0) %>%
  ggplot( aes(x=departure_diff)) +
    geom_density(fill="#69b3a2", color="#e9ecef", alpha=0.8)+labs(x='ti
me in second')+
  ggtitle('The actual departure verse Scheduled departure time in minut
e')
```

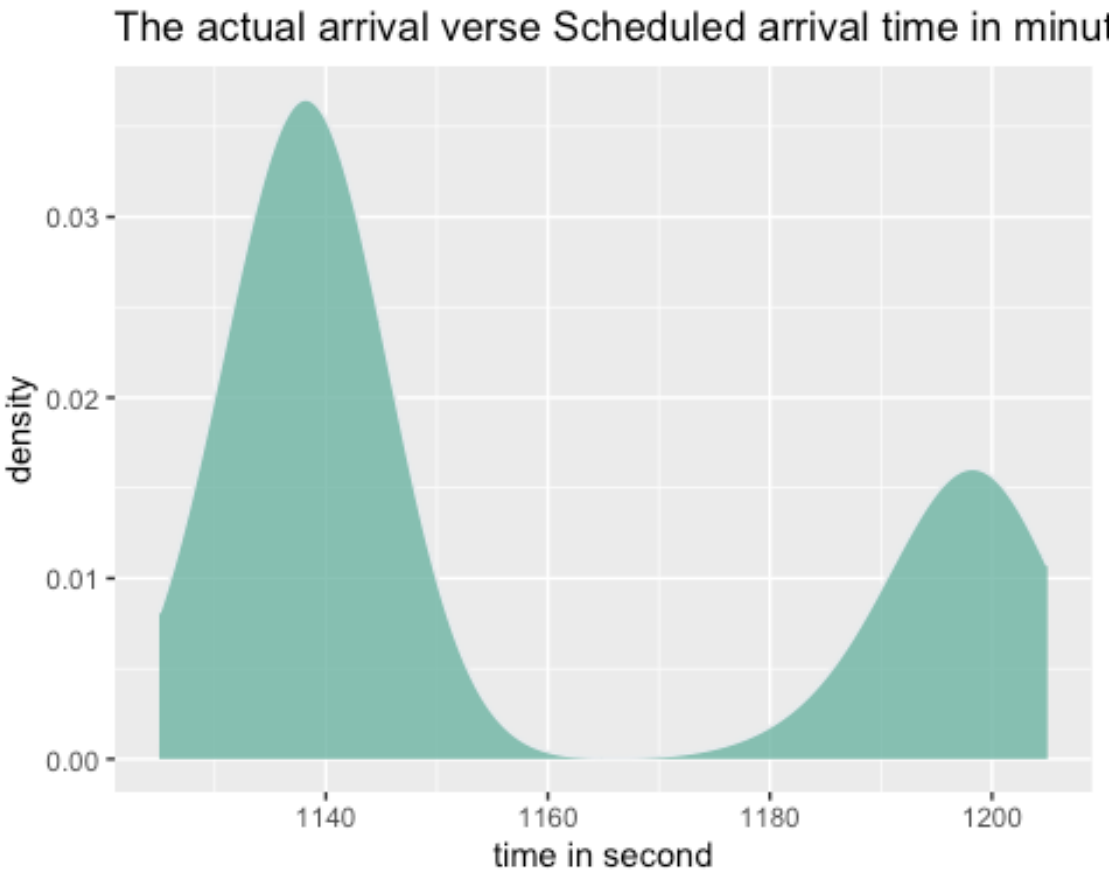


Here is the graph shows the distribution of how much the early and late ferry departure time. Anything greater than 0 means hows much times in minute did the ferry late for the scheduled depature time.

```
  Ferry_data%>%
  filter( arrival_diff<0) %>%
  ggplot( aes(x=arrival_diff)) +
    geom_density(fill="#69b3a2", color="#e9ecef", alpha=0.8)+labs(x='ti
me in second')+
  ggtitle('The actual arrival verse Scheduled arrival time in minute')
```



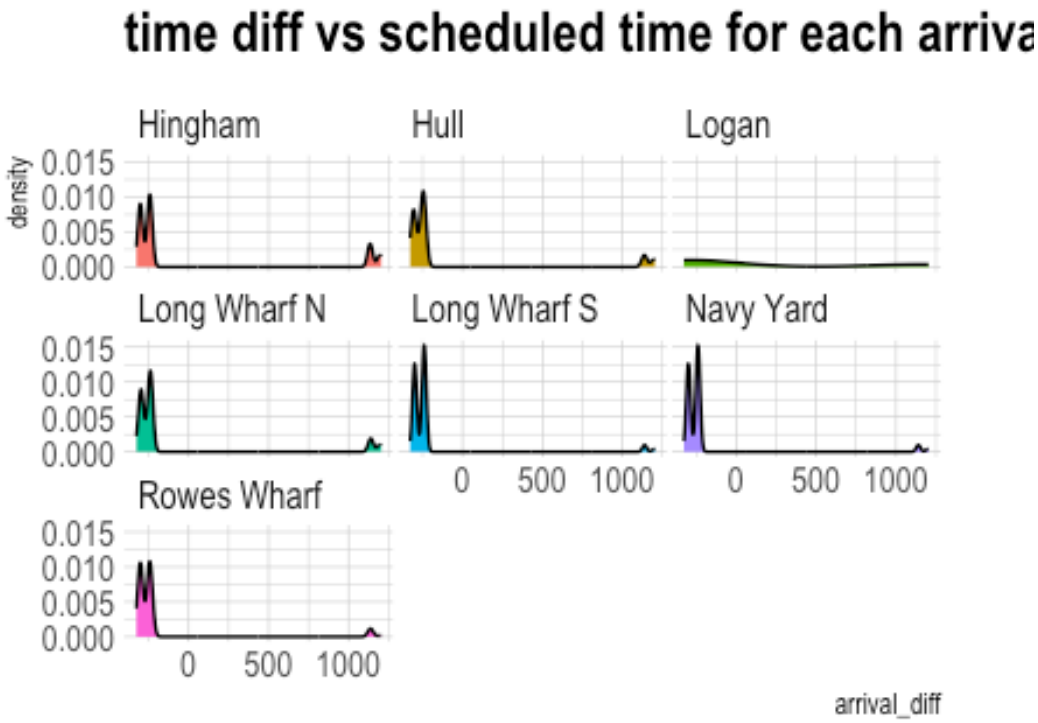

```
  Ferry_data%>%
  filter( arrival_diff>0) %>%
  ggplot( aes(x=arrival_diff)) +
    geom_density(fill="#69b3a2", color="#e9ecef", alpha=0.8)+labs(x='ti
me in second')+
  ggtitle('The actual arrival verse Scheduled arrival time in minute')
```



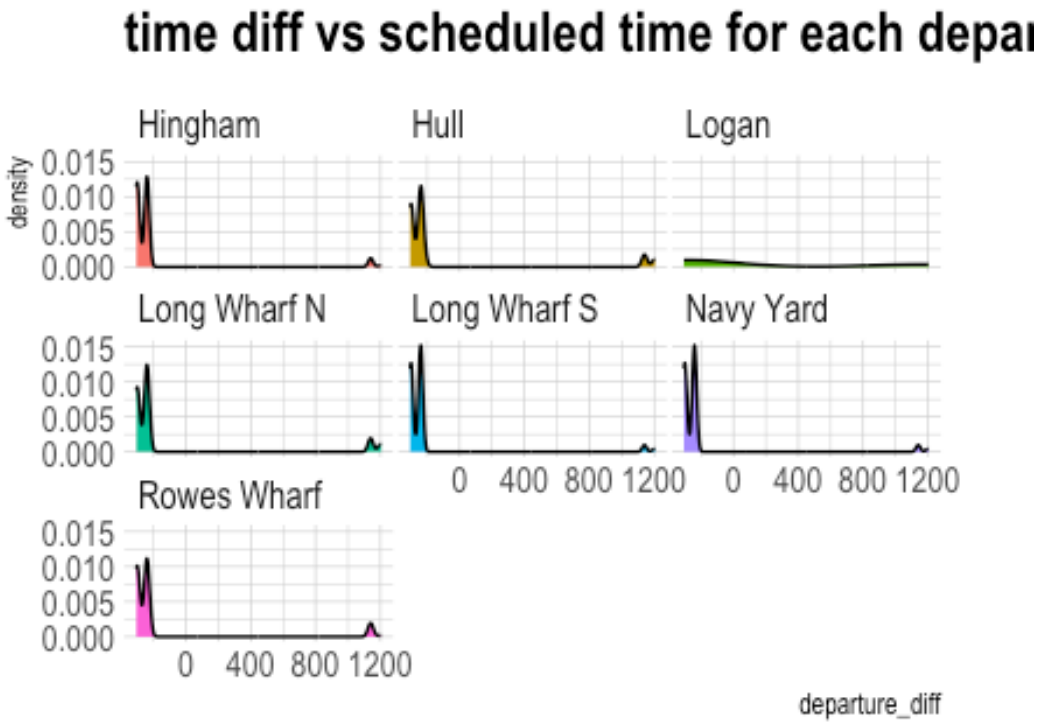
Here is the graph shows the distribution of how much the early and late ferry arrival time. Anything greater than 0 means hows much times in minute did the ferry late for the scheduled arrival time.

```
ggplot(data=Ferry_data, aes(x=arrival_diff, group=arrival_terminal, fil
l=arrival_terminal)) +
    geom_density(adjust=1.5) +
    theme_ipsum() +
    facet_wrap(~arrival_terminal) +
    theme(
      legend.position="none",
      panel.spacing = unit(0.1, "lines"),
      axis.ticks.x=element_blank()
    )+ggtitle("time diff vs scheduled time for each arrival ferry termi
nal")
```

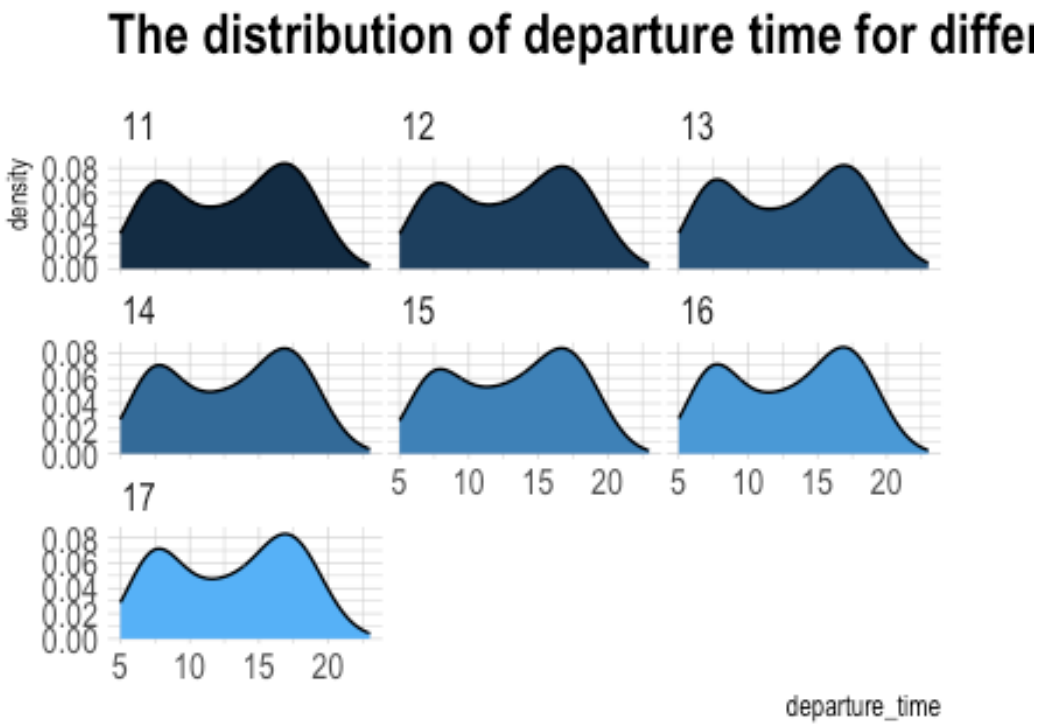
```
## Warning: Removed 213 rows containing non-finite values (`stat_densit
y()`).
```




```
ggplot(data=Ferry_data, aes(x=departure_diff, group=departure_terminal,
 fill=departure_terminal)) +
    geom_density(adjust=1.5) +
    theme_ipsum() +
    facet_wrap(~departure_terminal) +
    theme(
      legend.position="none",
      panel.spacing = unit(0.1, "lines"),
      axis.ticks.x=element_blank()
    )+ggtitle("time diff vs scheduled time for each departure ferry ter
minal")

## Warning: Removed 221 rows containing non-finite values (`stat_densit
y()`).
```

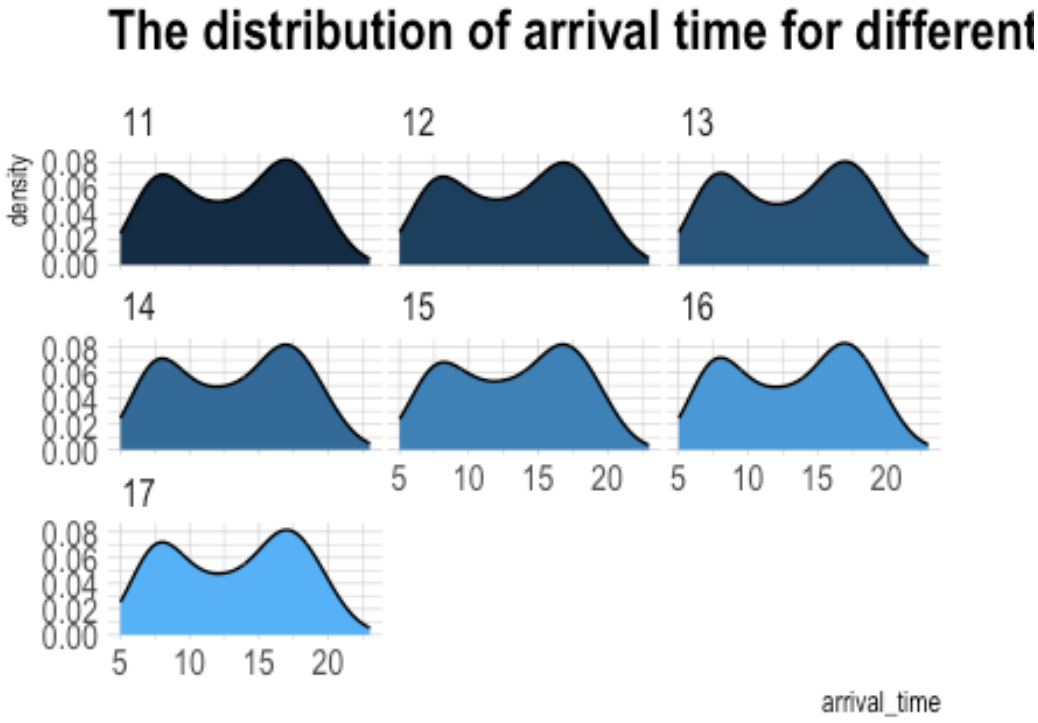



```
ggplot(data=Ferry_data, aes(x=departure_time, group=day, fill=day)) +
    geom_density(adjust=1.5) +
    theme_ipsum() +
    facet_wrap(~day) +
    theme(
      legend.position="none",
      panel.spacing = unit(0.1, "lines"),
      axis.ticks.x=element_blank()
    ) + labs(title="The distribution of departure time for different da
y")

## Warning: Removed 221 rows containing non-finite values (`stat_densit
y()`).
```




```
ggplot(data=Ferry_data, aes(x=arrival_time, group=day, fill=day)) +
    geom_density(adjust=1.5) +
    theme_ipsum() +
    facet_wrap(~day) +
    theme(
      legend.position="none",
      panel.spacing = unit(0.1, "lines"),
      axis.ticks.x=element_blank()
    ) + labs(title="The distribution of arrival time for different day")
```

```
## Warning: Removed 213 rows containing non-finite values (`stat_densit
y()`).
```


The distribution of arrival time for different
11
12
13
14
15
16
17
density
0.08
0.06
0.04
0.02
0.00
5
10
15
20
arrival_time