1. Data visualization: flights at ABIA

What is interesting in the ABIV.csv? I'm going to concentrate in delay of the flights and try to fond the best time for people to minimize delays.

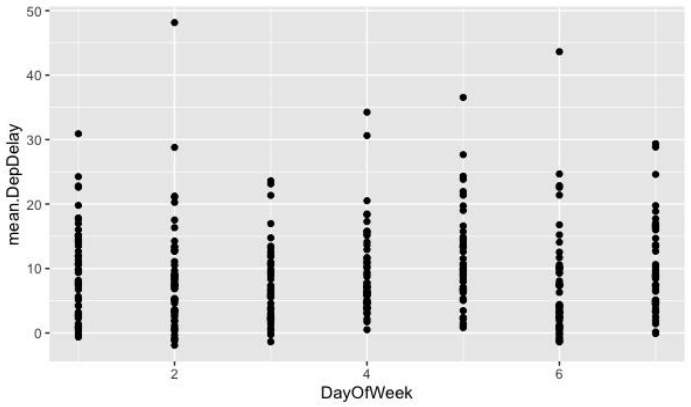
First, before using with the data, we need to remove the NA row in the data.

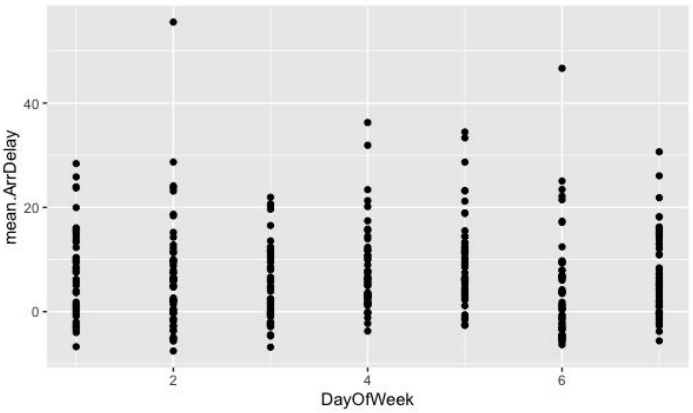
```
ABIA<-subset(ABIA,ArrDelay!="NA")
ABIA<-subset(ABIA,DepDelay!="NA")
```

- Day of week

```
ABIA_summ=ABIA%>%
  group_by(Month,DayofMonth,DayOfWeek)%>%
  summarise(mean.ArrDelay=mean.default(ArrDelay),mean.DepDelay=mean.default(DepDelay))

ggplot(ABIA_summ) +
  geom_point(mapping = aes(x = DayOfWeek, y = mean.DepDelay))
ggplot(ABIA_summ) +
  geom_point(mapping = aes(x = DayOfWeek, y = mean.ArrDelay))
```

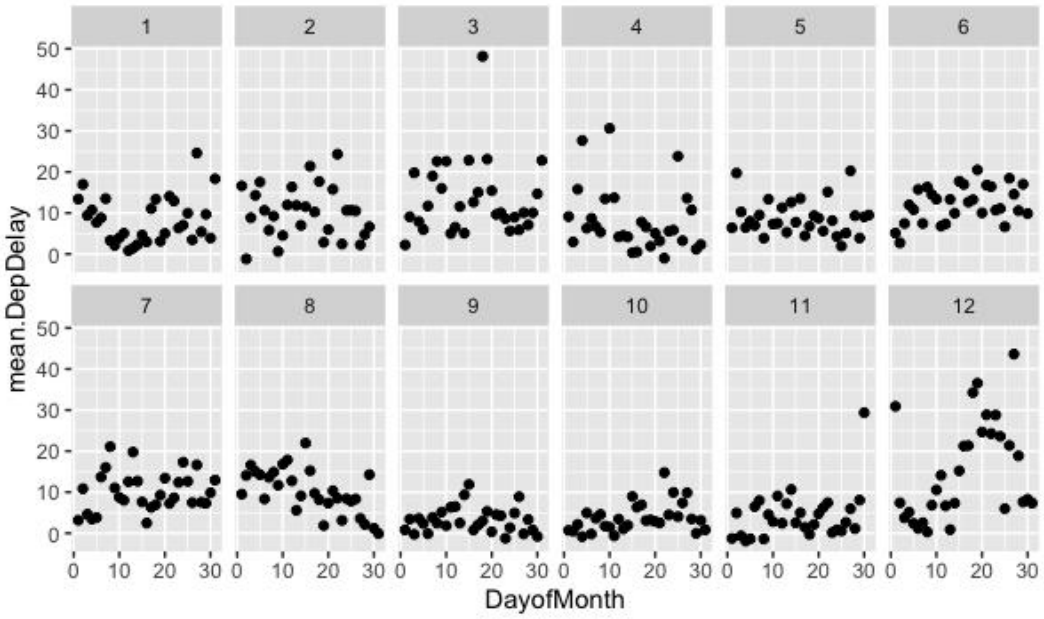


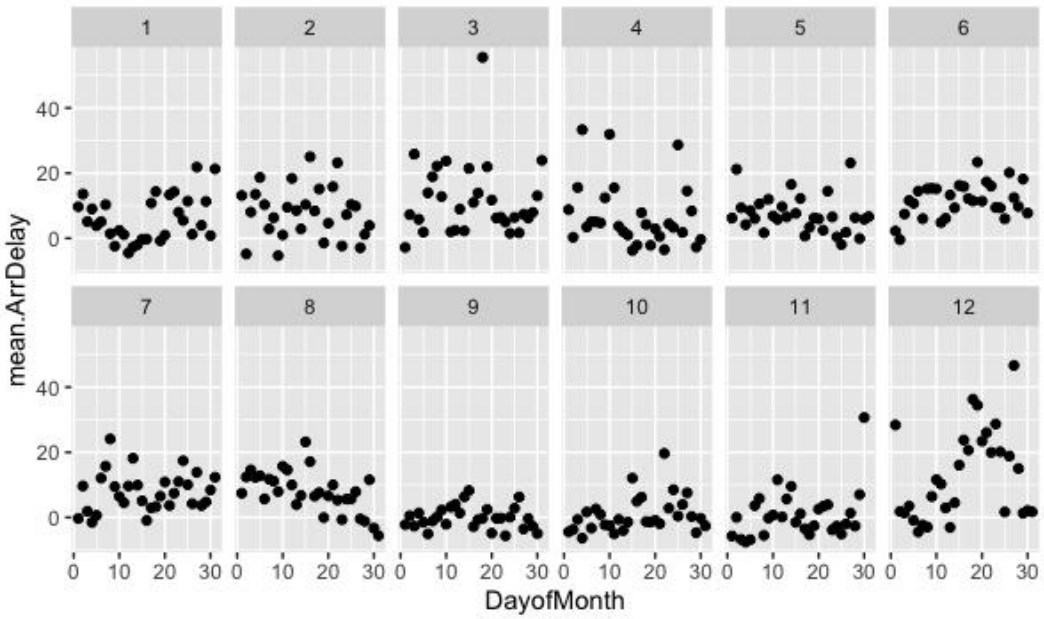


From the plot we could see that Wednesday owns both the least mean Arr-delay and Dep-delay.

◆ Day of month

```
ggplot(ABIA_summ) +
  geom_point(mapping = aes(x = DayOfWeek, y = mean.DepDelay))
ggplot(ABIA_summ) +
  geom_point(mapping = aes(x = DayOfWeek, y = mean.ArrDelay))
```



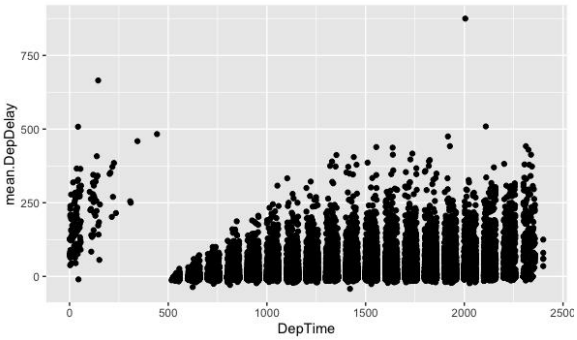


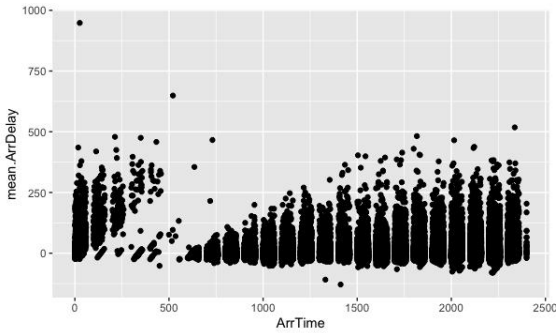
From these figures, we could see that September has the least average delay in the month compare with other months. (September and October are hard to distinguish if we only care about the Depdelay figures.

Here we only care about both the concentrate and the mean which could be obviously observed.

- Time of day

```
ABIA_summ=ABIA%>%
  group_by(ArrTime,DepTime,Month,DayofMonth,DayOfWeek)%>%
  summarise(mean.DepDelay=mean.default(DepDelay),mean.ArrDelay=mean(ArrDelay))
ggplot(ABIA_summ) +
  geom_point(mapping = aes(x = DepTime, y = mean.DepDelay))
ggplot(ABIA_summ) +
  geom_point(mapping = aes(x = ArrTime, y = mean.ArrDelay))
```





For different time of a day, it varies between Depdelay and Arrdelay. However, it obeys the normal rule that planes need a period to fly during the route. Thus, we could conclude from the figure that from 5 o'clock, the probability of delay tends to be higher to the end of the day.

- As a conclusion, from the data of 2018, it seems we could predict that during the long time of a year, the best time to minimize delays is during the earliest time of Wednesday in September. But actually, there are also other effects like weather condition or emergencies that would also contribute a lot.
- Thus , I prefer to regard this as a conclusion to find the reason of delays in 2018 rather than using this data set into predicting.