Econ 294 Final Exam

WenhanZeng, ID:1504976

2016-03-16

Preperation

Load required packages.

```
library(foreign)
library(ggplot2)
library(knitr)
library(nycflights13)
library(dplyr)
library(tidyr)
library(RSQLite)
```

Part a) Weather

Require data from database

```
my_db<- src_sqlite("nycflights13.sqlite", create = T)
flights_sqlite <- copy_to(
  my_db, flights, temporary = FALSE,
  indexes = list(
    c("year", "month", "day"),
    "carrier",
    "tailnum")
)
flights<-collect(flights_sqlite)

airlines_sqlite <- copy_to(
  my_db, airlines, temporary = FALSE,
  indexes = list("carrier")
)
airlines<-collect(airlines_sqlite)

airports_sqlite <- copy_to(
  my_db, airports, temporary = FALSE,
  indexes = list("faa")
)
airports<-collect(airports_sqlite)

planes_sqlite <- copy_to(
  my_db, planes, temporary = FALSE,
  indexes = list("tailnum")
)
planes<-collect(planes_sqlite)

weather_sqlite <- copy_to(
```

```
  my_db, weather, temporary = FALSE,
  indexes = list(
    c("year", "month","day","hour"),
    "origin")
)
weather<-collect(weather_sqlite)
```

Creat column "date"

```
final<- flights %>%
  mutate(canceled= ifelse(is.na(arr_time),1,0))%>%
  unite(
    col=date,
    ...=year,month,day,
    sep="-"
  )

weather<-weather %>%
  unite(
    col=date,
    ...=year,month,day,
    sep="-"
  ) %>%
  select(origin:wind_gust,visib)
```

Merge flights with weather, we find the matched variables first:

```
names(final)[names(final) %in% names(weather)]
```

```
## [1] "date"   "origin" "hour"
```

```
finalw<-final %>%
  left_join(weather, by=c('date','hour','origin'))
```

Keep all flights with weather information

```
finalw<-finalw %>%
  filter(is.na(temp)==F) %>%
  group_by(canceled)
```

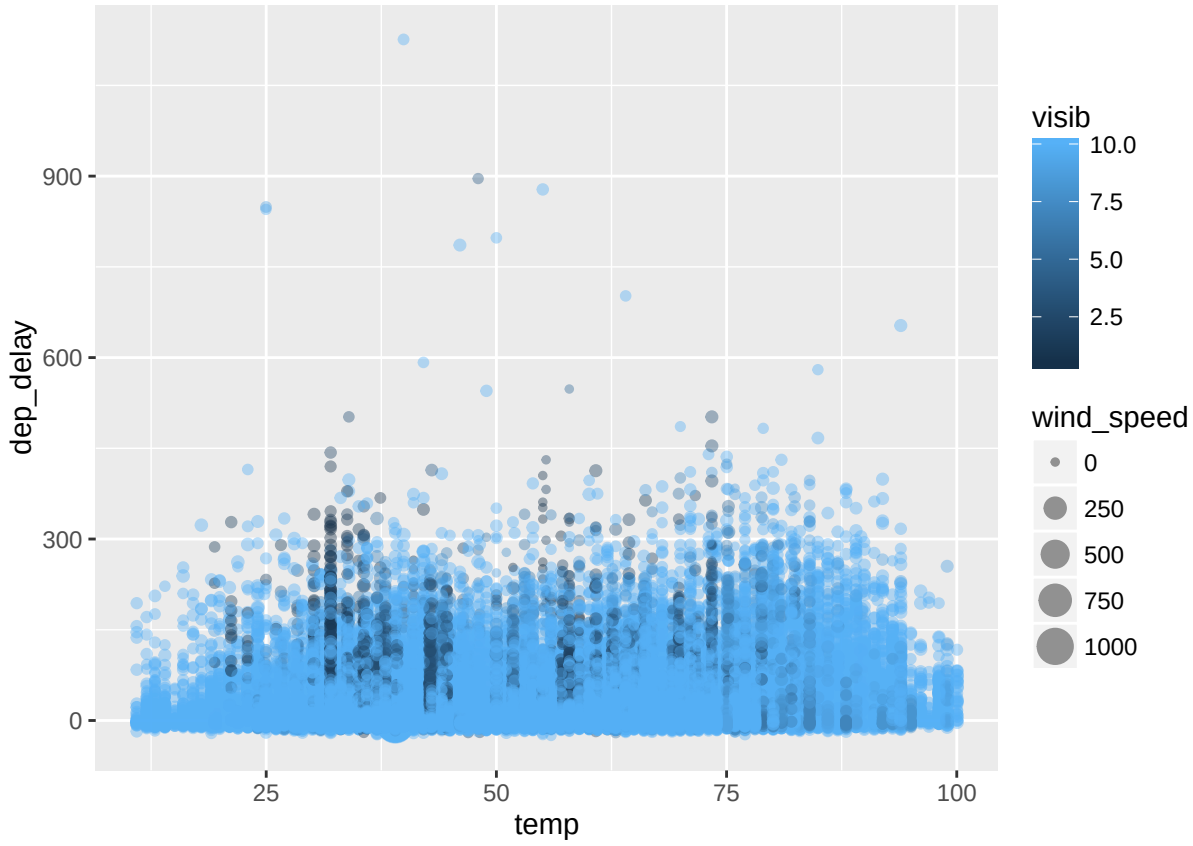
Run the regression

```
regdelay<-lm(dep_delay~temp+dewp+wind_speed+visib,finalw)
regcancel<-glm(canceled~dep_delay+temp+dewp+humid+wind_speed+visib,finalw,family = binomial(logit))
```

Those two reggressions tell us that: 1) For "dep_delay": temerature, dewpoint, wind speed and visibility are all statistically significant. 2) For "canceled" : dep_dlay, dewp, humid, visib are statistically significant. We see that when the time of delay increases, it is more unlikely to see the flight canceled (though this is quite counter-intuitive). While the "visib" goes down, it is more likely to be canceled.

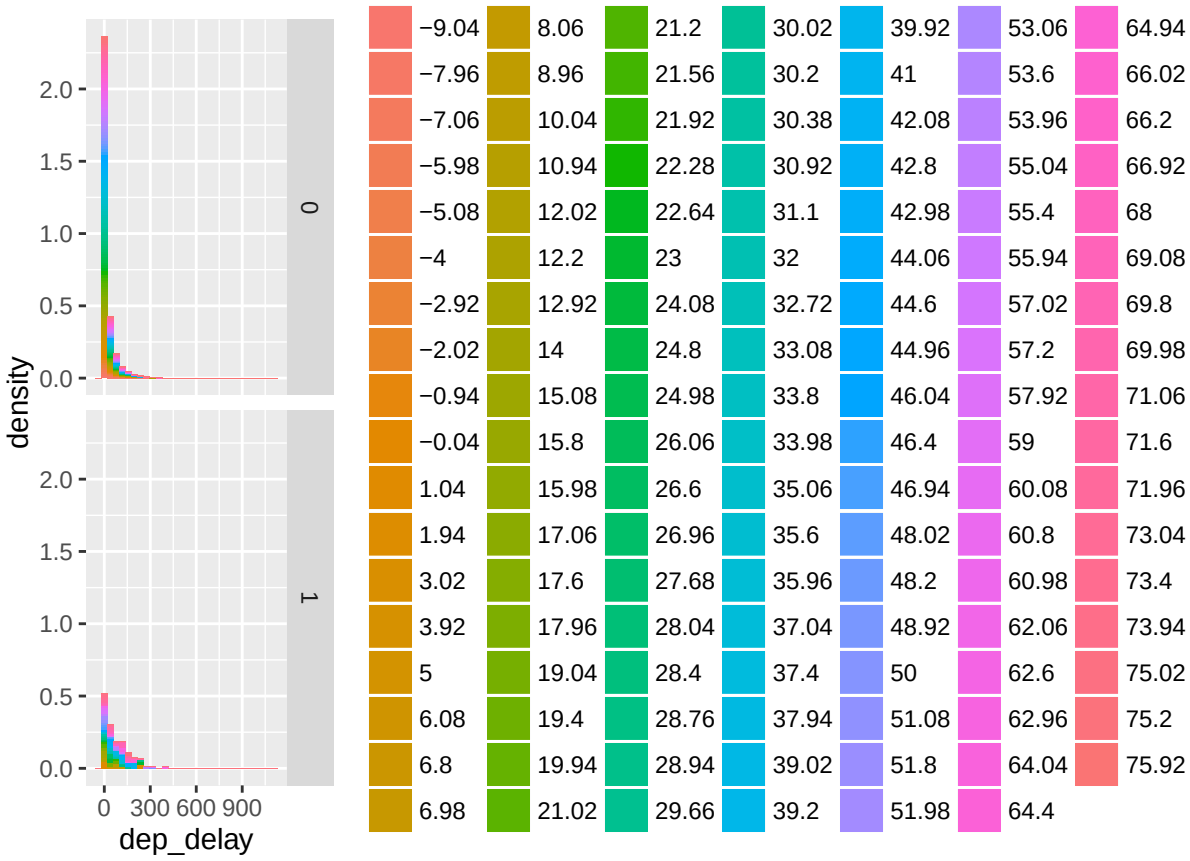
Plot a graph

```
p1<-ggplot(
  data = finalw,
  aes(x = temp,y=dep_delay,size=wind_speed)
)+
  geom_point(aes(colour=visib), alpha=0.4)
p1
```



As we can see above, black points, which indicates low visibility, causes higher delay, but the relationship between temperature and time of delay is not that obvious.

```
p2<-ggplot(
  data=finalw,
  aes(dep_delay,fill=factor(dewp))
)+
  geom_histogram(aes(dep_delay,..density..))+
  facet_grid("canceled~.")
p2
```



shows the influence of dewp (which is significant) on time of delay in different group:canceled or not canceled. ### Part b) Time Time of day, day of week, and time of year that might affect delay. ## Use the data from part a) to run a regression:

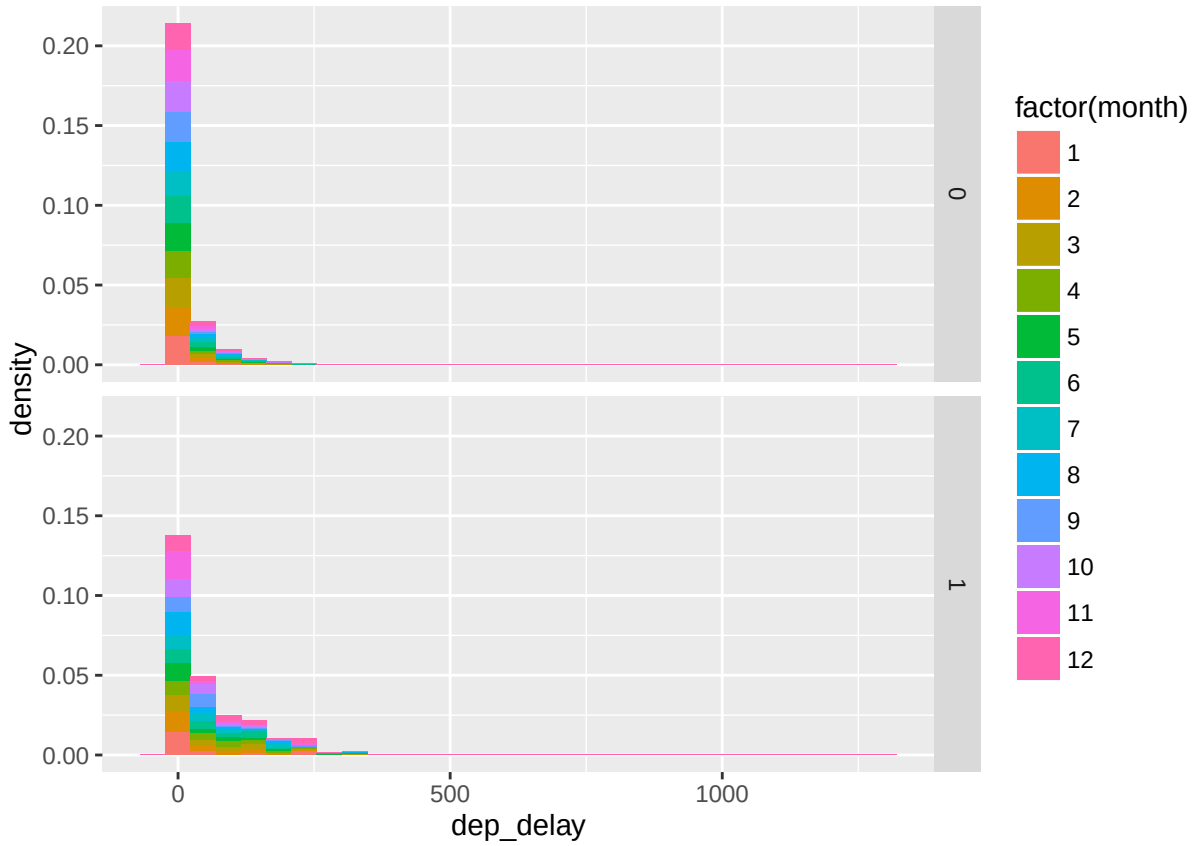
```
regtime<-lm(dep_delay~factor(month)+factor(day)+dep_time+hour+minute,flights)
```

The result shows that the coefficients of September, October and November are significantly negative, and the coefficients of June & July are obviously greater than others. It make sense since the time of delay are partially determined by the number of customers.

```
p3<-ggplot(
  data=flights,
  aes(x=dep_delay, fill=factor(month))
)+
  geom_histogram(aes(x=dep_delay, y=..density..))+
  facet_grid("canceled~.")
p3
```

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

```
## Warning: Removed 8255 rows containing non-finite values (stat_bin).
```




```
regtimeCancel<-glm(canceled~dep_delay+dep_time+month+day+hour+minute,flights,family=binomial(logit))
```

The reggression result shows that in this case, only the time of delay has a significant impact on cancellation of flights.

Part c) Airport Destination

Rearrange table from part a)

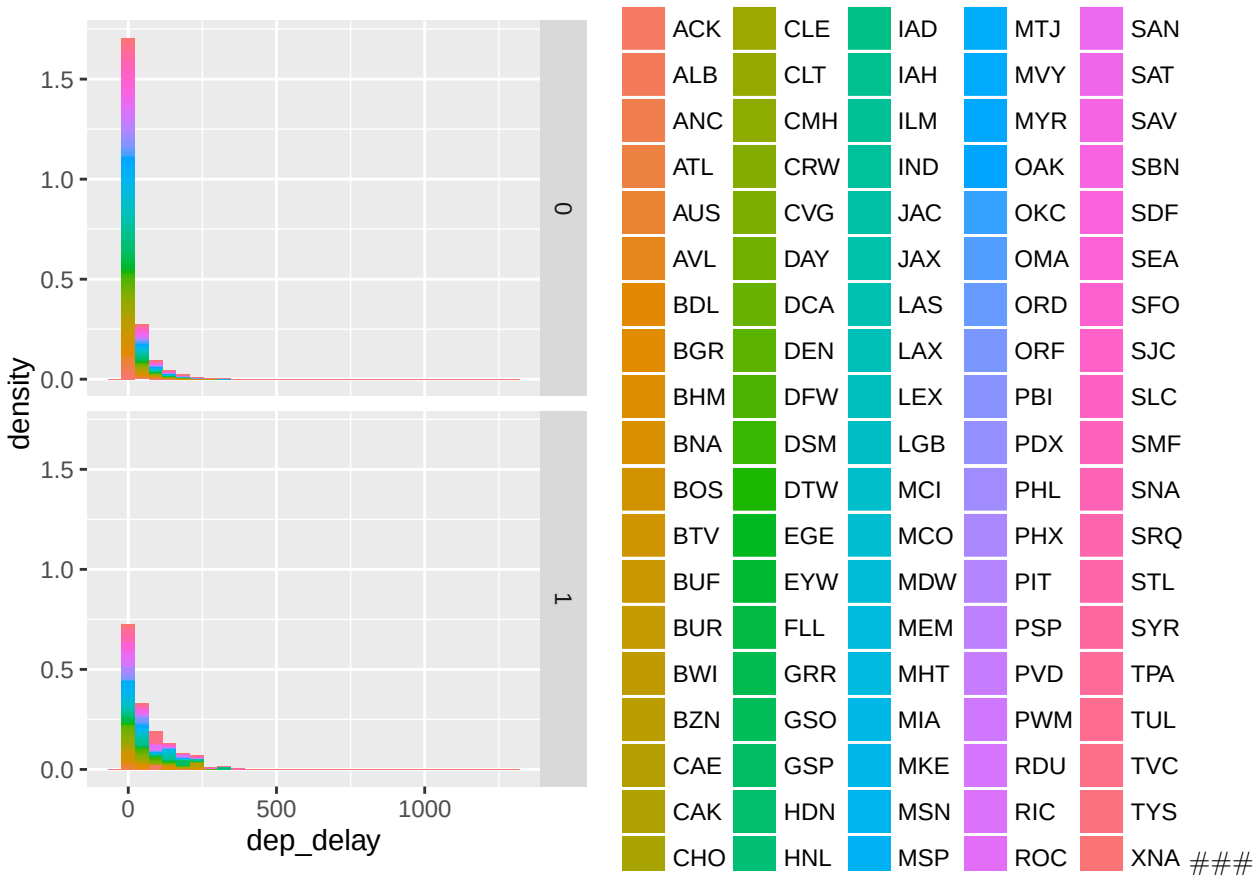
```
destairport<-airports %>%
  inner_join(flights, by=c("faa"="dest"))
```

Run regressions

```
regdest<-lm(dep_delay~faa,destairport)
regdestCancel<-glm(canceled~dep_delay+faa, destairport, family = binomial(logit))
```

Here we see from regression result, that several airports such as BHM(Birmingham Intl), CAE(Columbia Metropolitan), DSM(Des Moines Intl), OKC(Will Rogers World), RIC(Richmond Intl), TUL(Tulsa Intl), TYS(Mc Ghee Tyson) are more likely to delay a flight, while as for cancellation, none of those airports show that kind of significance.

```
pc<-ggplot(
  data=destairport,
  aes(x=dep_delay,fill=faa)
)+
  geom_histogram(aes(x=dep_delay,y=..density..))+
  facet_grid("canceled~.")
pc
```



Part d) Characteristics of the planes ## Rearrange data from part a)

```
airplanes<-planes %>%
  inner_join(flights, by="tailnum")%>%
  group_by(canceled)
```

Run the regression

Pay attention that here from both tables we have a variable called "year", but they have different meanings. After merging, they will be renamed as year.x and year.y.

```
regplane<-lm(dep_delay~year.x+factor(model),airplanes)
regplaneCancel<-glm(canceled~dep_delay+year.x+model,airplanes,family = binomial(logit))
```

From the first regression we see, none of those variables are significant, which indicates that there is no such concrete proof that older(newer) planes are more likely to delay. From the second one, flights with

plane model 757-351 and A319-115, are significantly more likely to be canceled, but the magnitude of those coefficients are quite small, so they are just statistically significant but not really significant in real life transportation. This is the same case for the age of planes, they are statistically significant but not much meaningful.

Conclusion

From part a) to part d), we see that when the visibility is low, flights are more likely to delay or to be canceled. This case is even worse for destination airports such as BHM(Birmingham Intl), CAE(Columbia Metropolitan) and so on. We don't worry too much about what kind of planes are flying, but we do need to chose a time for travel, since in some particular months like June and July, it's more likely to be delayed or canceled.