Project 3

Natalie Pham

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.6.3
```

```
##
## 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)
```

```
## Warning: package 'ggplot2' was built under R version 3.6.3
```

```
cvg <- read.csv("CVG_Flights.csv", header = TRUE, stringsAsFactors = FALSE, na.strings = "")
d<- as.POSIXct(cvg$FLIGHT_DATE, format = "%m/%d/%Y")
cvg$dm <- format(d, "%m")
```

CVG data set

Question 1

What is the departure delay in major airports in each region in the US (West, Midwest, East) - by month

West coast

```
b <- cvg %>% filter(ORIGIN_AIRPORT %in% c('LAX','SFO'))

ggplot(b,aes(x=dm,y=DEPARTURE_DELAY,group=ORIGIN_AIRPORT, shape = ORIGIN_AIRPORT, color = ORIGIN_AIRPOR
  geom_line(aes(color=ORIGIN_AIRPORT,alpha=1),size=1) +
  geom_point(aes(color=ORIGIN_AIRPORT,alpha=1),size=2) +
  labs(
    title= "Departure delay over the first three months in 2015 - West Coast")+
  scale_x_discrete(position = "top") +
  theme_bw() +
```

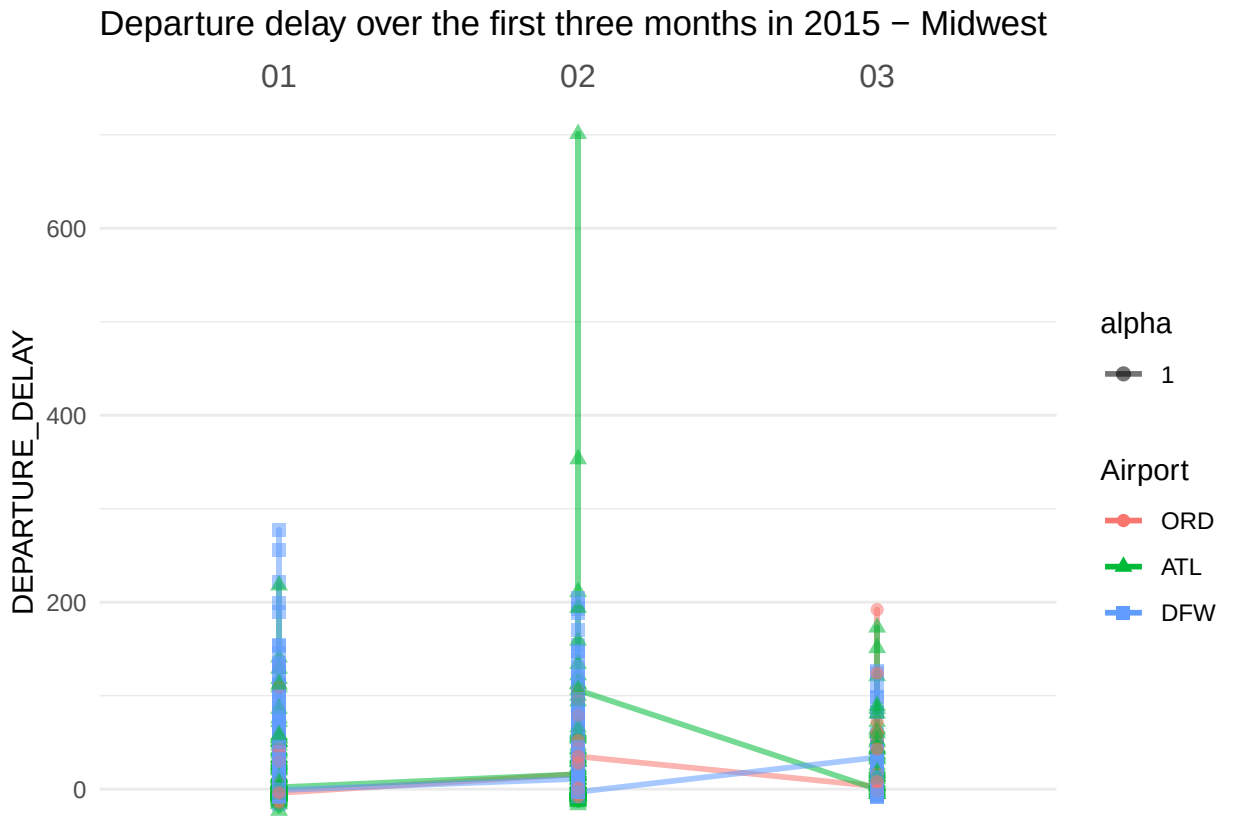
```
  theme(legend.position = "right") +
  theme(panel.border     = element_blank()) +
  #theme(axis.title.y     = element_blank()) +
  #theme(axis.text.y      = element_blank()) +
  #theme(panel.grid.major.y = element_blank()) +
  #theme(panel.grid.minor.y = element_blank()) +
  theme(axis.title.x     = element_blank()) +
  theme(panel.grid.major.x = element_blank()) +
  theme(axis.text.x.top      = element_text(size=12)) +
  theme(axis.ticks       = element_blank())+
  scale_colour_discrete(name  = "Airport",labels = c("LAX", "SF")) +
  scale_shape_discrete(name  = "Airport",labels = c("LAX", "SF"))
```

- There are more delays in LAX than they are in San Francisco. The peak of amount of delays is in February (over 300 in time unit). Individually, there are more data points of delay in Jan than in February. In conclusion, Jan and Feb account for high departure delay

  Midwest

```
m <- cvg %>% filter(ORIGIN_AIRPORT %in% c('ORD','ATL','DFW')) %>% filter(DEPARTURE_DELAY != "NA")

ggplot(m,aes(x=dm,y=DEPARTURE_DELAY,group=ORIGIN_AIRPORT,shape = ORIGIN_AIRPORT, color = ORIGIN_AIRPORT
  geom_line(aes(color=ORIGIN_AIRPORT,alpha=1),size=1) +
  geom_point(aes(color=ORIGIN_AIRPORT,alpha=1),size=2) +
  labs(
```

```
  title= "Departure delay over the first three months in 2015 - Midwest")+
scale_x_discrete(position = "top") +
theme_bw() +
theme(legend.position = "right") +
theme(panel.border     = element_blank()) +
#theme(axis.title.y     = element_blank()) +
#theme(axis.text.y      = element_blank()) +
#theme(panel.grid.major.y = element_blank()) +
#theme(panel.grid.minor.y = element_blank()) +
theme(axis.title.x     = element_blank()) +
theme(panel.grid.major.x = element_blank()) +
theme(axis.text.x.top      = element_text(size=12)) +
theme(axis.ticks       = element_blank()) +   #delete
scale_colour_discrete(name  = "Airport",labels = c("ORD", "ATL","DFW")) +
scale_shape_discrete(name  = "Airport",labels = c("ORD", "ATL","DFW"))
```

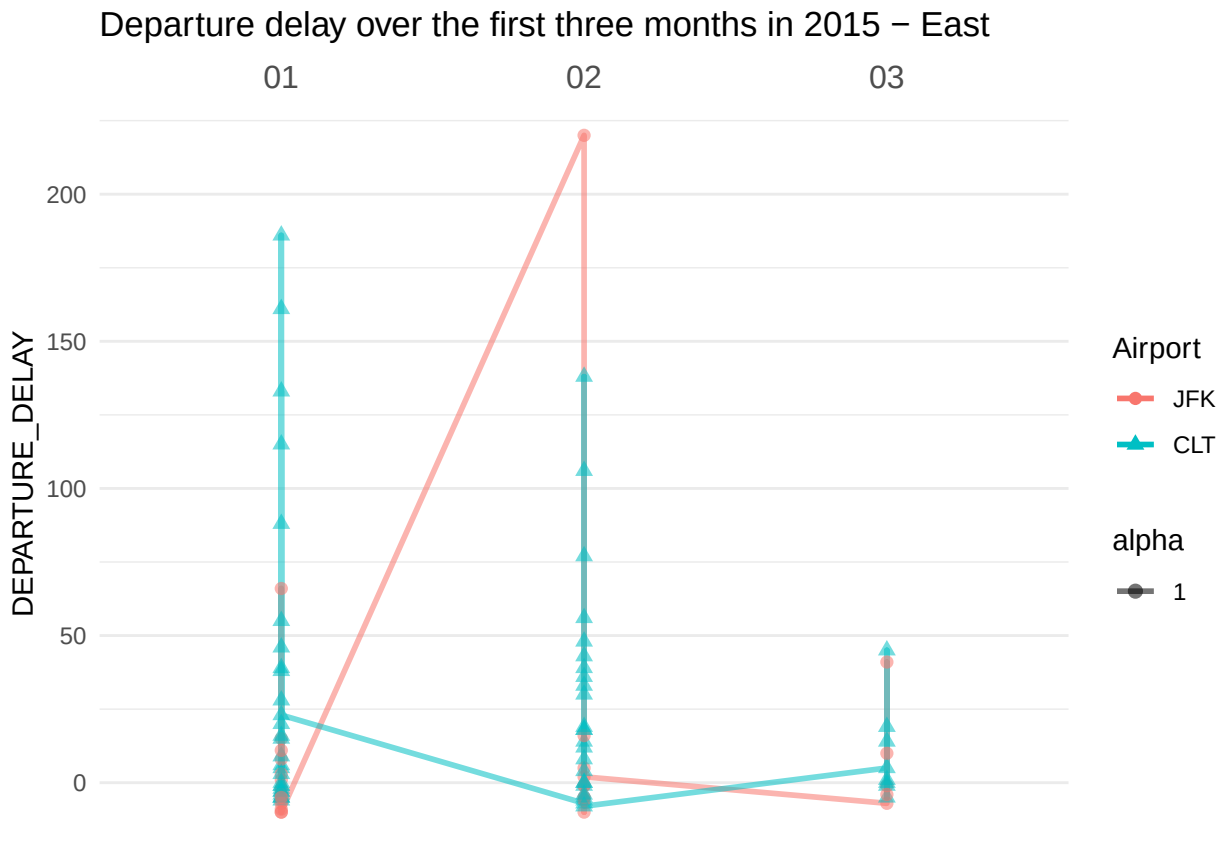


- There are more delays in DFW than in ORD and ATL, especially in the month of January and February. The peak of amount of delays is in February (almost 700 in time unit). Interestingly, ORD seems to have lowest departure delay even though the aiport is one of the busiest in the US

East coast

```
e <- cvg %>% filter(ORIGIN_AIRPORT %in% c('JFK','CLT')) %>% filter(DEPARTURE_DELAY != "NA")
ggplot(e,aes(x=dm,y=DEPARTURE_DELAY,group=ORIGIN_AIRPORT,shape = ORIGIN_AIRPORT, color = ORIGIN_AIRPORT
```

```
geom_line(aes(color=ORIGIN_AIRPORT,alpha=1),size=1) +
geom_point(aes(color=ORIGIN_AIRPORT,alpha=1),size=2) +
labs(
  title= "Departure delay over the first three months in 2015 - East")+
scale_x_discrete(position = "top") +
theme_bw() +
theme(legend.position = "right") +
theme(panel.border     = element_blank()) +
#theme(axis.title.y     = element_blank()) +
#theme(axis.text.y      = element_blank()) +
#theme(panel.grid.major.y = element_blank()) +
#theme(panel.grid.minor.y = element_blank()) +
theme(axis.title.x     = element_blank()) +
theme(panel.grid.major.x = element_blank()) +
theme(axis.text.x.top      = element_text(size=12)) +
theme(axis.ticks       = element_blank()) +   #delete
scale_colour_discrete(name  = "Airport",labels = c("JFK","CLT")) +
scale_shape_discrete(name  = "Airport",labels = c("JFK","CLT"))
```



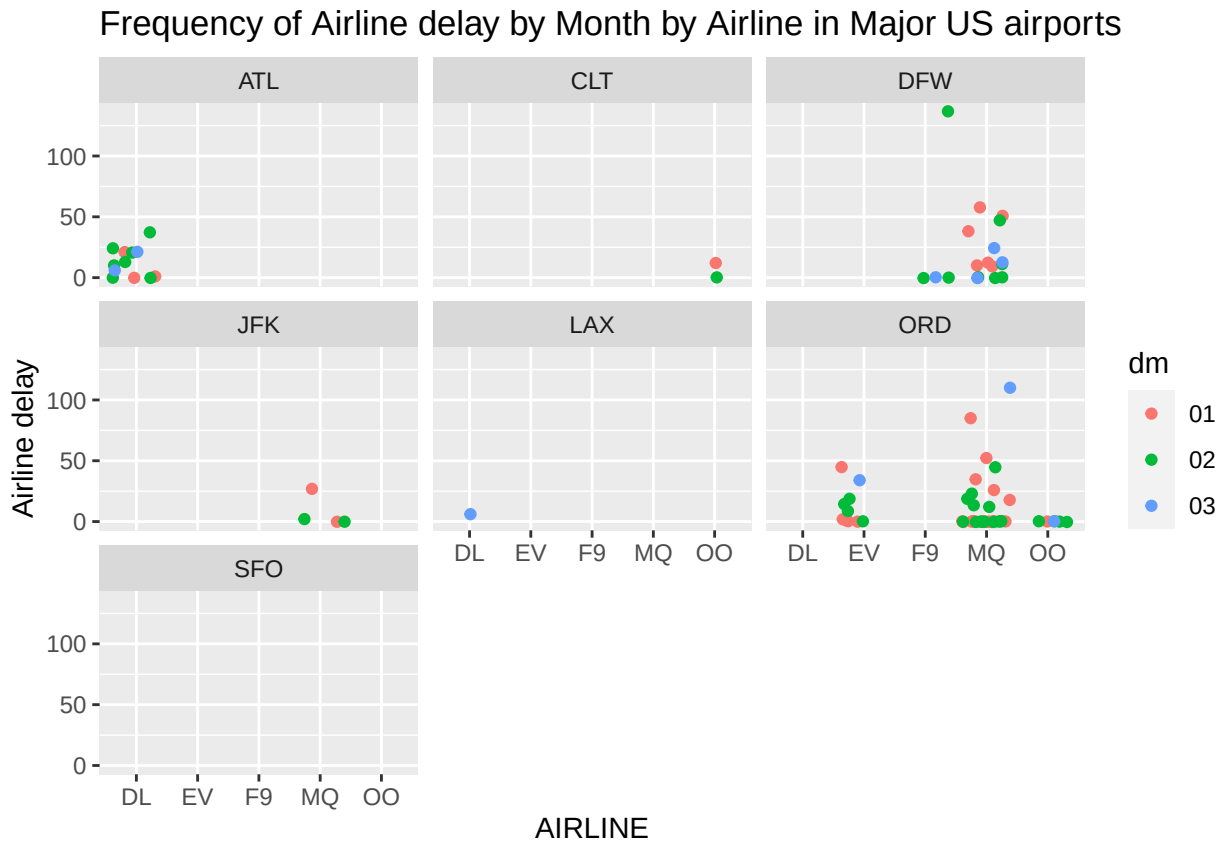
- Delay data points of CLT account more than that of JFK, especially during the month of January and February. The peak of amount of delays is also in February (over 300 in time unit). Surprisingly, JFK has lower departure delays than CLT although it is also one of the major airports in the US
- According to the 3 graphs, February accounted for the highest time of departure delay all over the major airports in 4 regions. Busiest airports such as ORD and JFK tends to have lower departure delay than other less busy but still major international airports

Question 2:

Which airline has the most delay in Major US Airports by month?

```
region <- cvg %>% filter(ORIGIN_AIRPORT == c('LAX','SFO','ORD','JFK','DFW','ATL','CLT'))
ggplot(region, aes(y = AIRLINE_DELAY, x = AIRLINE, color = dm)) + geom_jitter() + facet_wrap(~ORIGIN_AI
```

```
## Warning: Removed 202 rows containing missing values (geom_point).
```



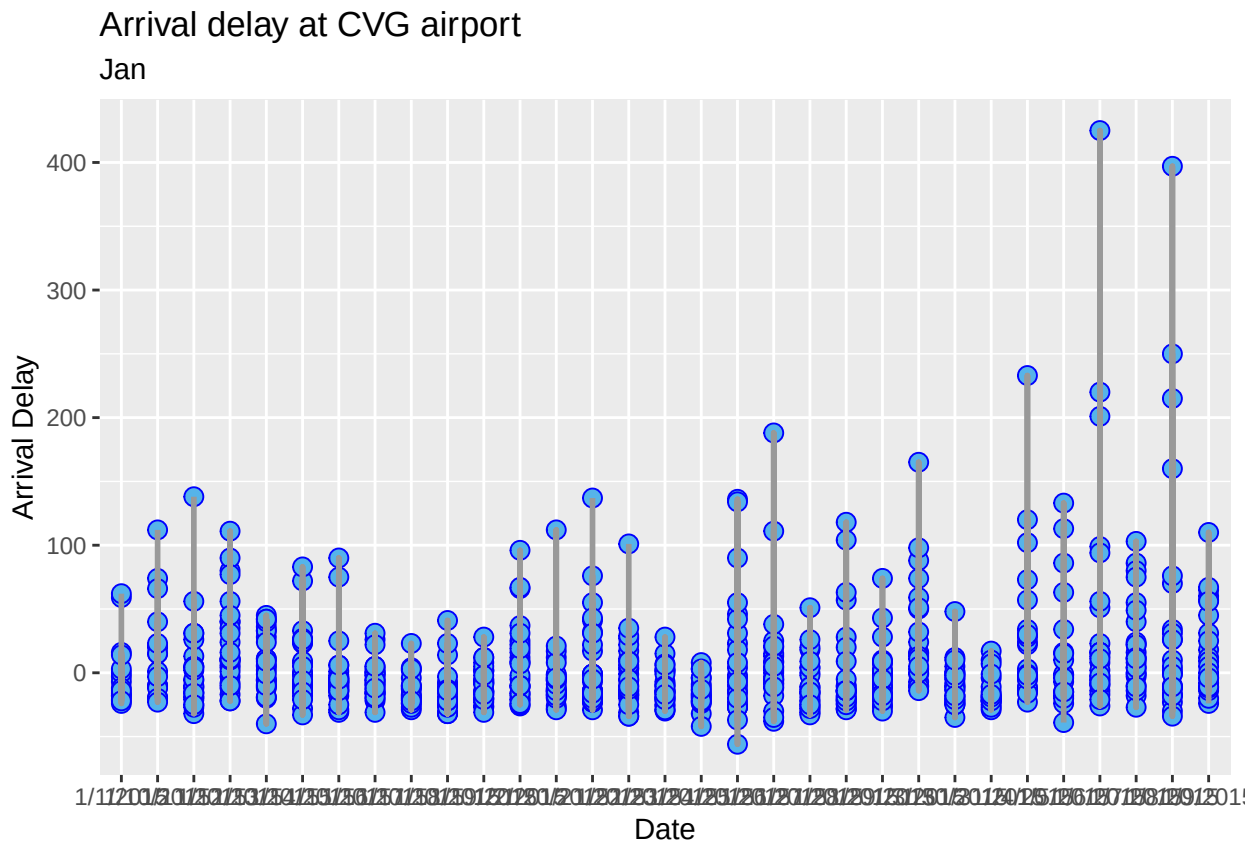
- With major international airports in different region, the airline MQ has high airline delay both in DFW and ORD, following is DL and EV

Question 3: Which month CVG has most arrival delay to major airports including ORD, LAX, SFO, ATL, DFW, JFK

```
jan <- cvg %>% filter(dm == "01") %>% filter(DESTINATION_AIRPORT %in% c('ORD','LAX','SFO','JFK','DFW','
ggplot(jan, aes(x = FLIGHT_DATE, y = ARRIVAL_DELAY)) +
  geom_point(shape=21,color='blue',fill='#56B4E9',size=3)+
  geom_line(color='grey60',size=1)+
  labs(title = "Arrival delay at CVG airport",
       subtitle = "Jan",
       x = "Date", y = "Arrival Delay")
```

```
## Warning: Removed 36 rows containing missing values (geom_point).
```

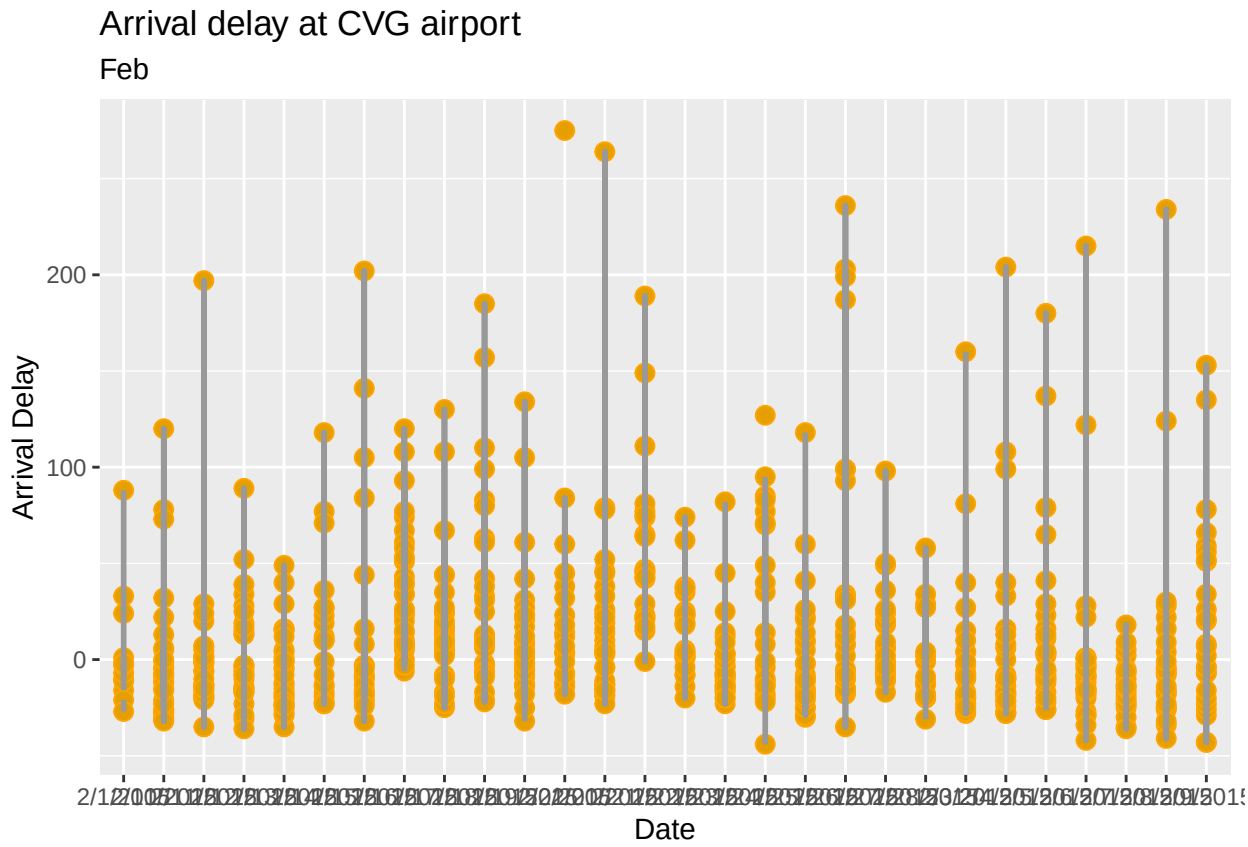
```
## Warning: Removed 3 row(s) containing missing values (geom_path).
```




```
feb <- cvg %>% filter(dm == "02") %>% filter(DESTINATION_AIRPORT %in% c('ORD','LAX','SFO','JFK','DFW','
ggplot(feb, aes(x = FLIGHT_DATE, y = ARRIVAL_DELAY)) +
  geom_point(shape=21,color='orange',fill='#E69F00',size=3)+
  geom_line(color='grey60',size=1)+
  labs(title = "Arrival delay at CVG airport",
       subtitle = "Feb",
       x = "Date", y = "Arrival Delay")
```

```
## Warning: Removed 65 rows containing missing values (geom_point).
```

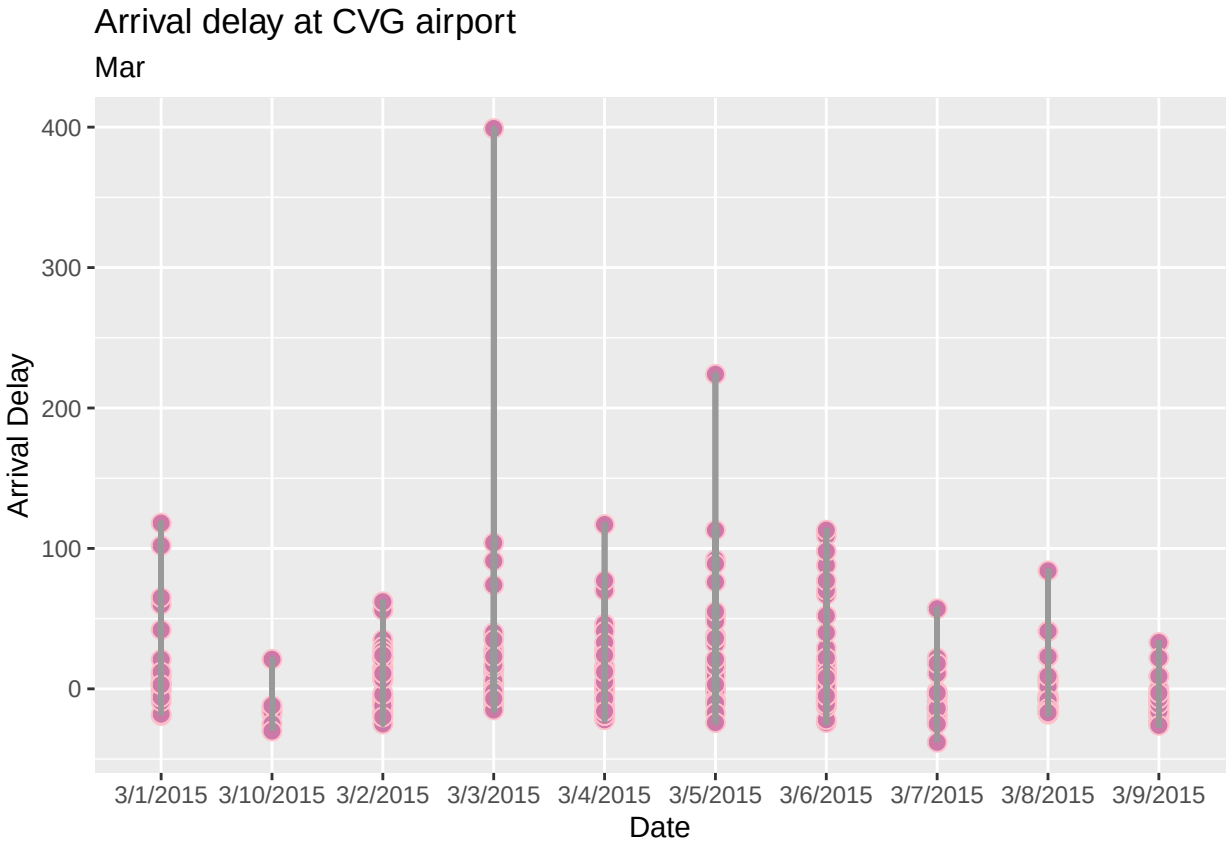
```
## Warning: Removed 10 row(s) containing missing values (geom_path).
```




```
mar <- cvg %>% filter(dm == "03") %>% filter(DESTINATION_AIRPORT %in% c('ORD','LAX','SFO','JFK','DFW','
ggplot(mar, aes(x = FLIGHT_DATE, y = ARRIVAL_DELAY)) +
  geom_point(shape=21,color='pink',fill='#CC79A7',size=3)+
  geom_line(color='grey60',size=1)+
  labs(title = "Arrival delay at CVG airport",
       subtitle = "Mar",
       x = "Date", y = "Arrival Delay")
```

```
## Warning: Removed 21 rows containing missing values (geom_point).
```

```
## Warning: Removed 4 row(s) containing missing values (geom_path).
```



- February consists highest arrival delay, following is January. March is the least of arrival delay.This pattern of arrival delay is similar to departure delay that February is always in most delay

College data set

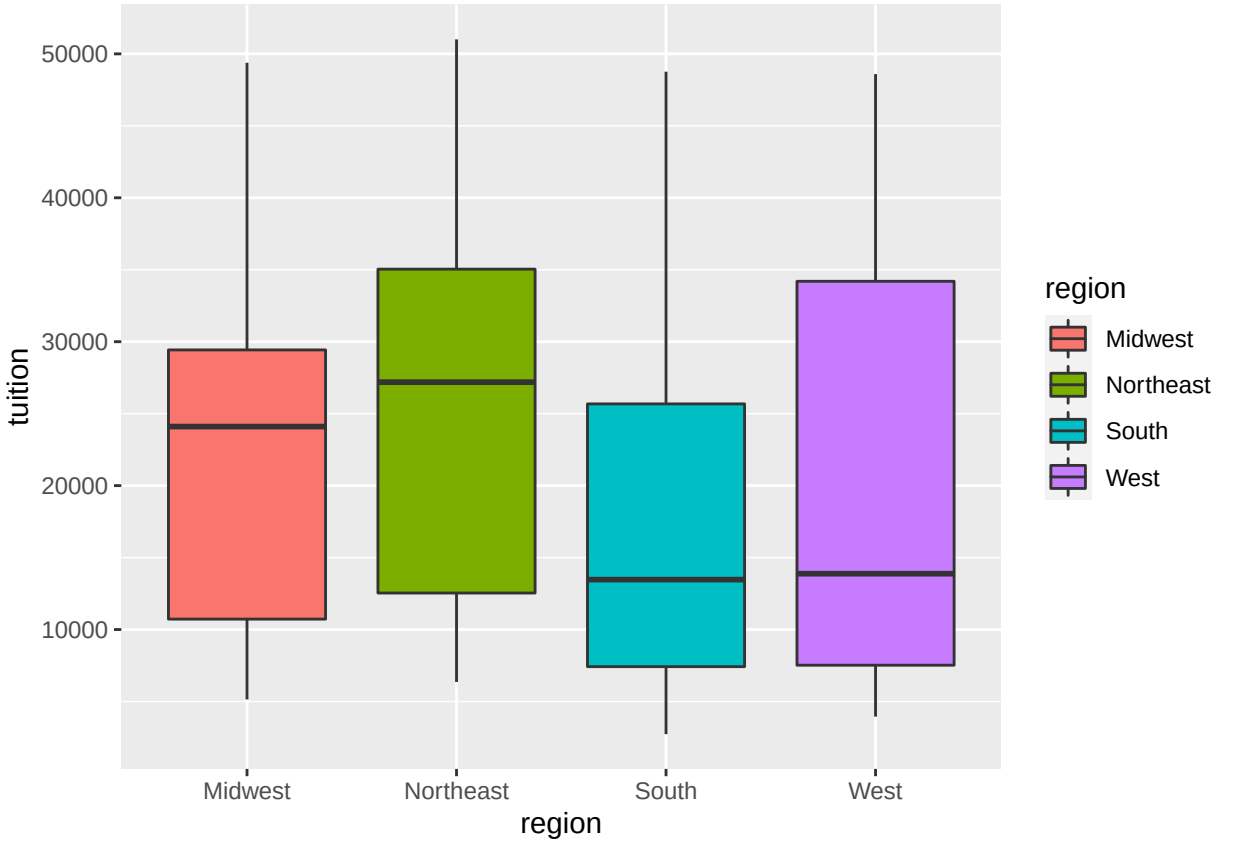
```
college <- read.csv('college.csv')
```

Question 1:

California is an interesting state and I wonder what is the median tuition fee for all colleges in the state of California since living and studying the state is expensive

First, I would like to take a general look on median tuition fee in each region across the US

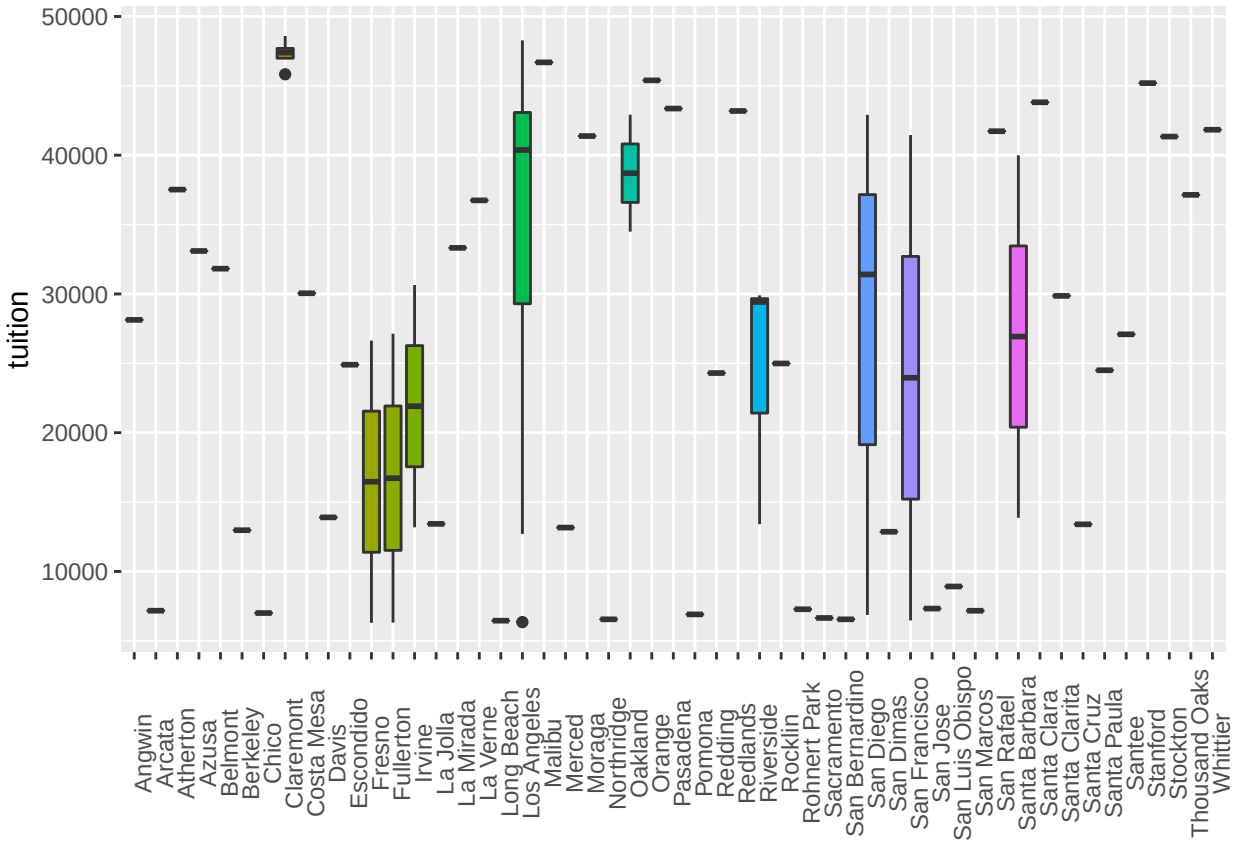
```
ggplot(college, aes(x = region, y = tuition, fill = region)) + geom_boxplot()
```



- From the scatter plot statistic above, we dwell in the median tuition fee in each region. The Northeast has the highest median of tuition (about $28000) than Midwest, South, and West. The Midwest ranks second place with the median of about $24,500. If we look at the scatter plot above, it seems that the West area has the lowest tuition fee among four regions. However, in this box plot, median values for tuition fee of the South and West are almost similar, which is about $14000. We use median to compare because median is immuned to extreme data points than it is for mean
- So, West generally has lowest median of tuition fee. Let's see the tuition in one of the expensive state in the West, California

  Then I filter down state = California to see the statistic for later comparision

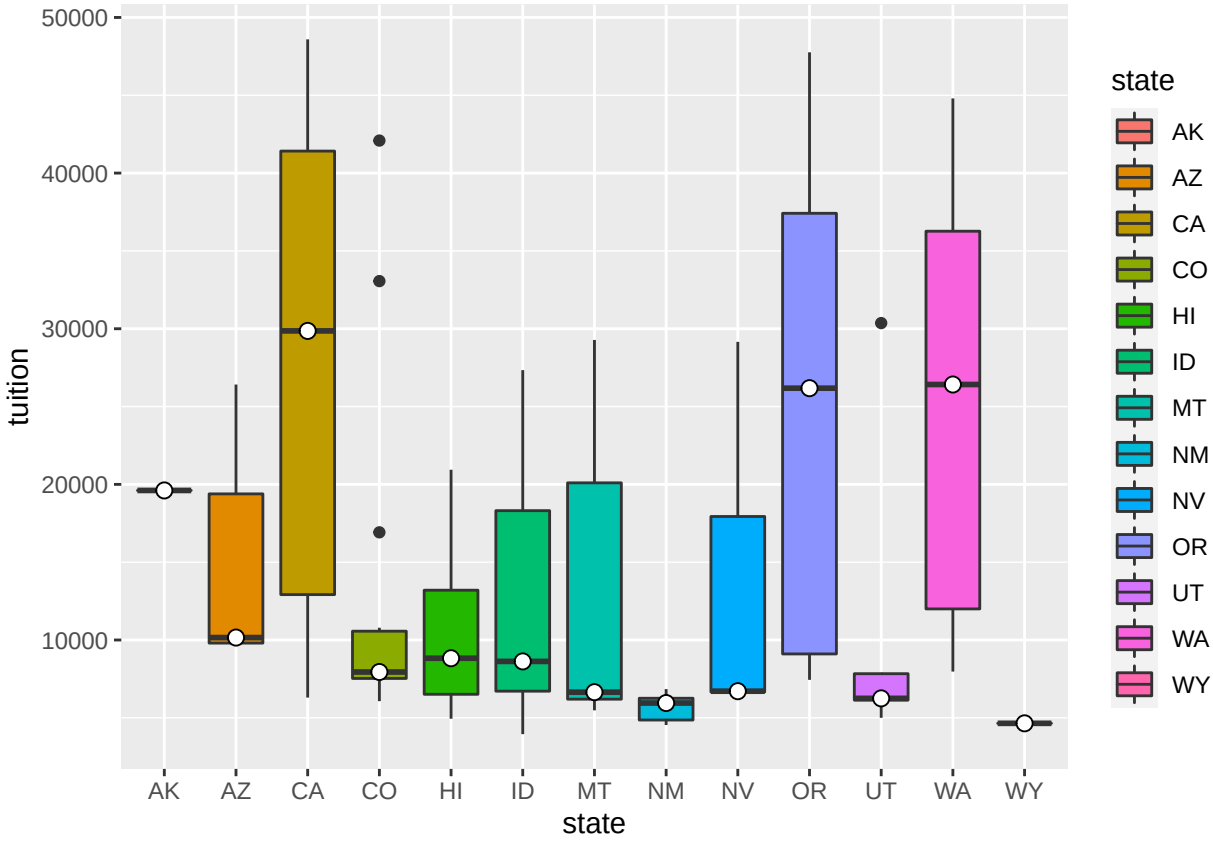
```
ca <- college %>% filter(state == "CA")

ggplot(ca, aes(x = city, y = tuition, fill = city)) + geom_boxplot() + theme(legend.position = "none")+
  theme(axis.title.x = element_blank()) +
  theme(axis.text.x = element_text(angle = 90))
```



- Generally, high and low tuition are equally distrbuted among colleges in different cities in California. There are several cities varied in tuition fee. Overall, the median tuition fee is over from $20000-$30000 and above in many cities in California

I would also like to see California compared to other states in the west

```
w <- college %>% filter(region == 'West')

ggplot(w,aes(x = state,y = tuition,fill=state))+
  geom_boxplot()+
  stat_summary(fun = median,geom = "point",fill="white",size=2.5,shape=21)
```

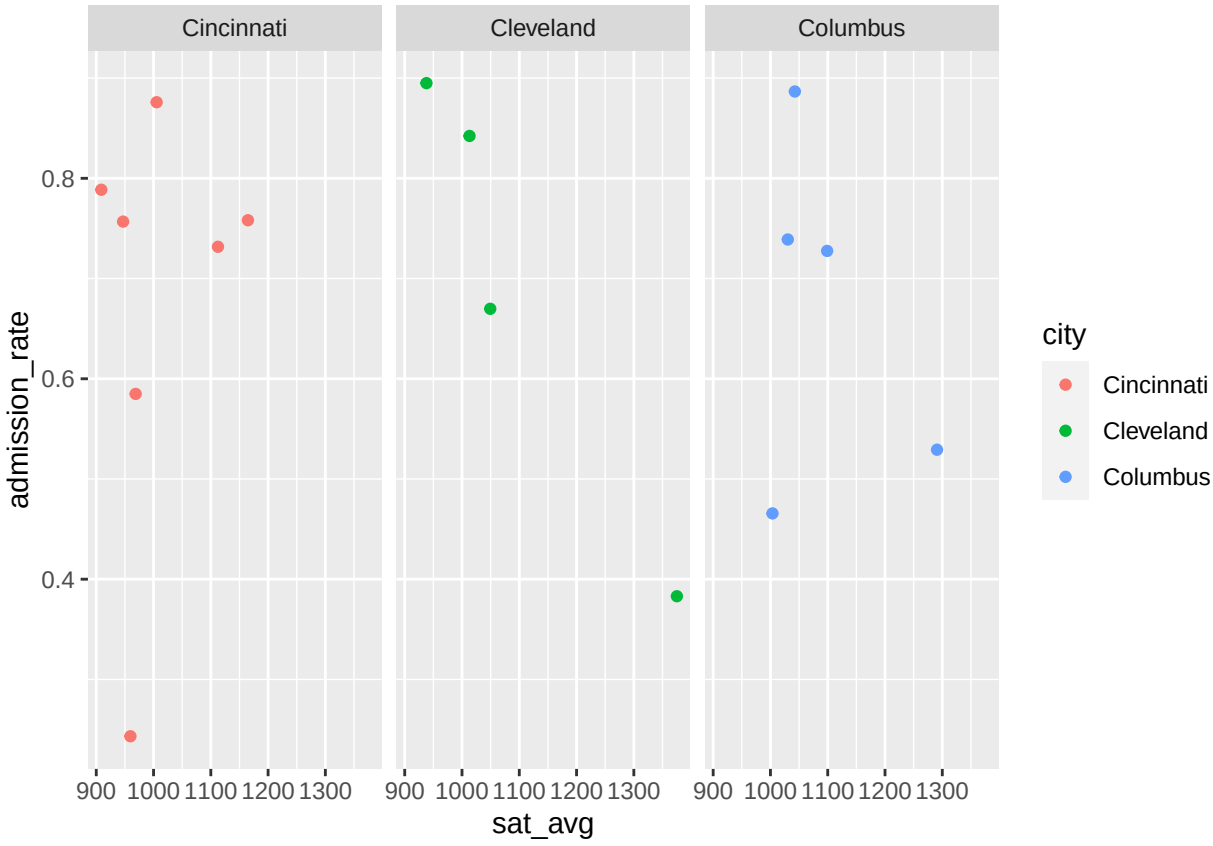


- Like I have mentioned above for the median tuition fee for California. This graph has depicted sucessfully. Compared to other states in West area, California's median tuition is the highest ($30000). The second highest are Oregon and Washington. The rest are mostly below $10000 for median tuition

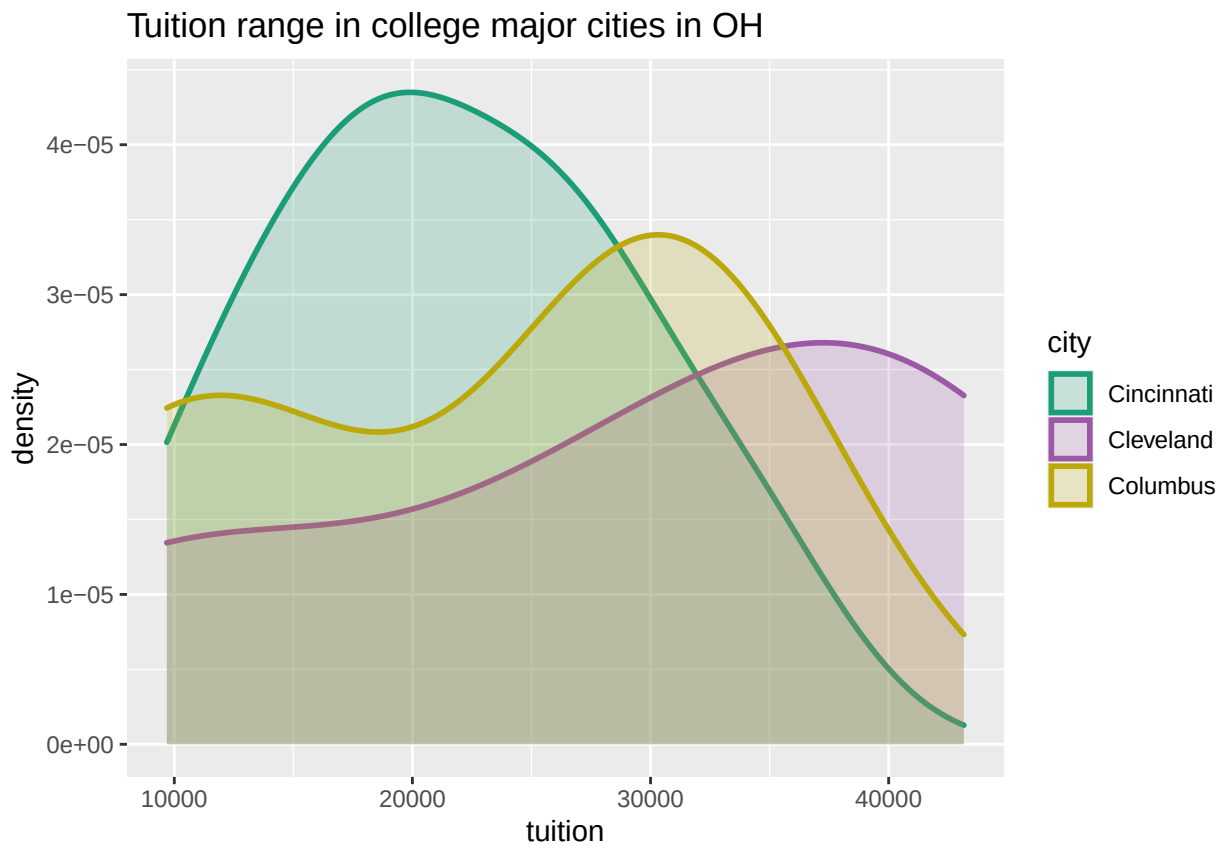
Question 2:

OHio curious ;) Since living in Cincinnati, OH. I am curious about admission rate, SAT requirement, tuition fee, and median debt in colleges in major cities in OH

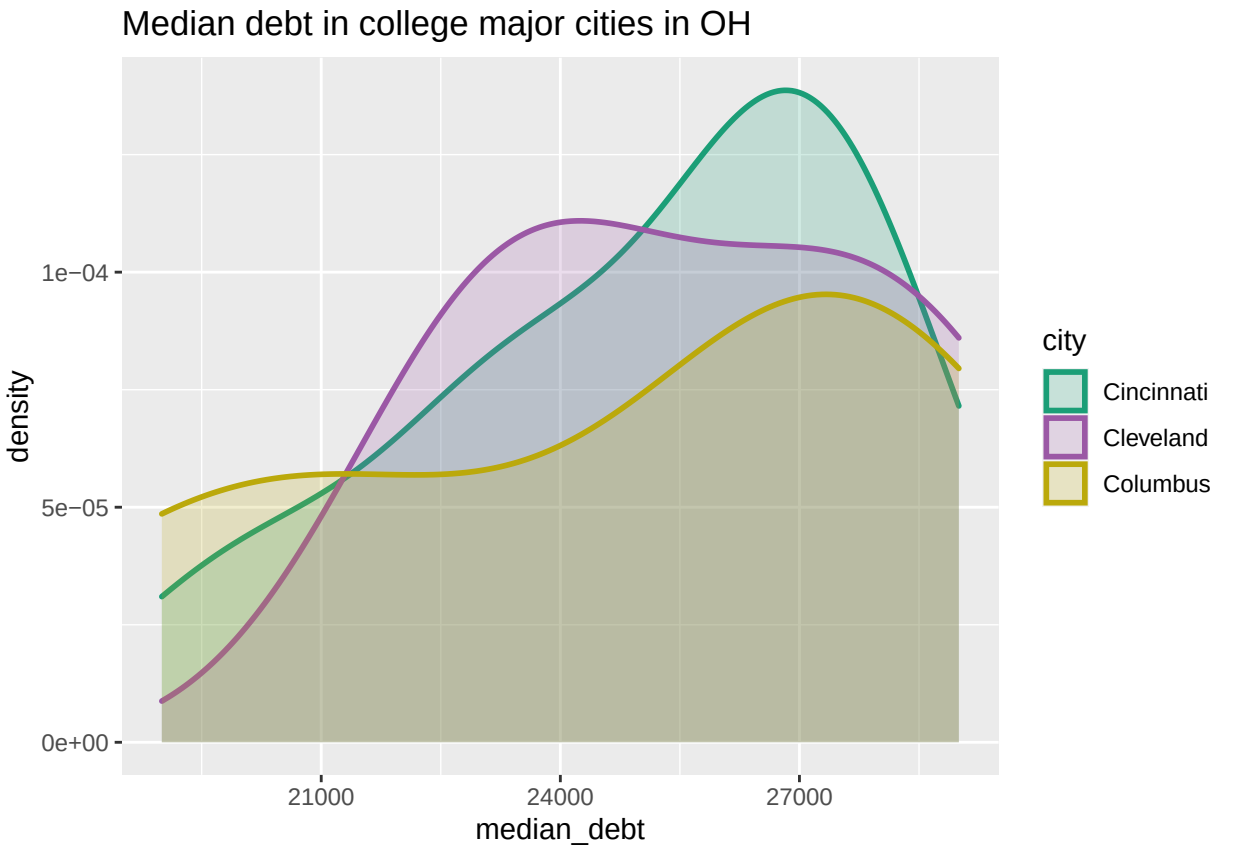
```
library(RColorBrewer)
oh <- college %>% filter(state == "OH") %>% filter(city %in% c('Cincinnati','Columbus','Cleveland'))

ggplot(oh, aes(x = sat_avg, y = admission_rate, col = city)) + geom_jitter() + facet_wrap(~city)
```




```
ggplot(oh,aes(x=tuition,color=city,fill=city))+ geom_density(alpha = 0.2,size=1,linetype="solid")+
  scale_color_manual(values =colorRampPalette(brewer.pal(8,"Dark2"))(4))+
  scale_fill_manual(values =colorRampPalette(brewer.pal(8,"Dark2"))(4))+
  ggtitle("Tuition range in college major cities in OH")
```




```
ggplot(oh,aes(x=median_debt,color=city,fill=city))+ geom_density(alpha = 0.2,size=1,linetype="solid")+
  scale_color_manual(values =colorRampPalette(brewer.pal(8,"Dark2"))(4))+
  scale_fill_manual(values =colorRampPalette(brewer.pal(8,"Dark2"))(4))+
  ggtitle("Median debt in college major cities in OH")
```



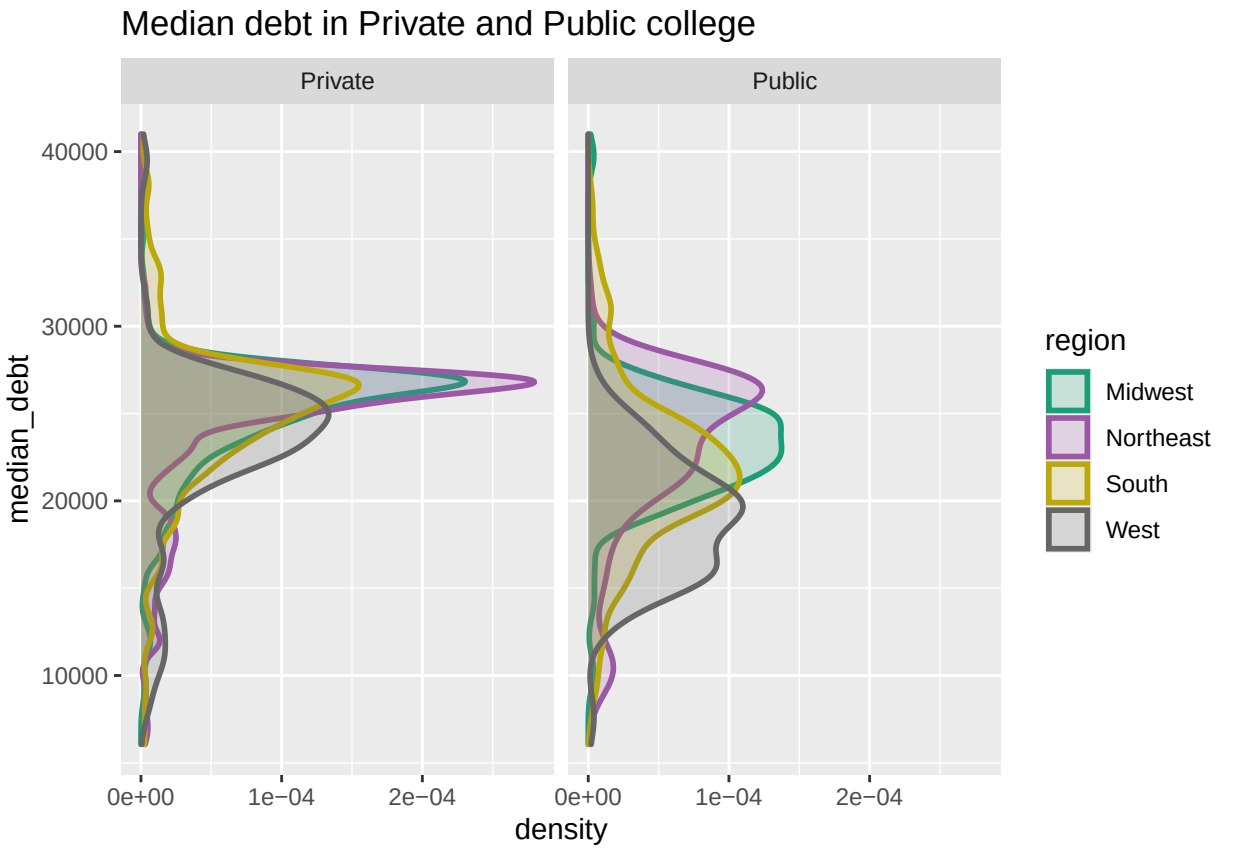
- First graph : Cincinnati has high admission rate but low average SAT. Admission rate for Cleveland is the same like Cincinnati but there is a school requires high SAT (higest among other two cities). Columbus's admission rate is average with average SAT score
- Second graph : Most colleges in Cincinnati has tuition around ($15000-$30000) while colleges in Columbus and Cleveland are mostly above $20000
- Third graph : Although tuition fee in most colleges in Cincinnati are low compared to that in Columbus and Cleveland, Cincinnati has high density of debt (mostly above $24000) than the other two cities.

Question 3:

Public and Private college performance

Life is full of curiousity :)), I want to know the median debt in Public and Private colleges in the 4 regions. As to my knowing that it is more expensive to attend Private colleges

```
ggplot(college,aes(y=median_debt,color=region,fill=region))+ geom_density(alpha = 0.2,size=1,linetype="
  scale_color_manual(values =colorRampPalette(brewer.pal(8,"Dark2"))(4))+
  scale_fill_manual(values =colorRampPalette(brewer.pal(8,"Dark2"))(4))+
  ggtitle("Median debt in Private and Public college")+
  facet_wrap(~control)
```



- Generally, the median debt for private and public colleges are distributed around $20000 and $30000 for all region in the US. The median debt for private colleges accounted for almost 90% in range of $20000 and $30000. Public colleges median debt makes up of 70% median debt of $20000-$30000 and about 20% below $20000

I then dwell into the average faculty salary and tuition fee in big major cities in the US, also to compare Cincinnati performance

```
c <- college %>% filter(city %in% c('Chicago','Los Angeles','San Francisco','New York','Boston','Dallas

ggplot(c, aes(y=tuition, x = faculty_salary_avg, color = control)) +
  geom_point() + facet_wrap(~city)
```

Generally, tuition fee for private colleges in major cities is always higher than $25000 and below $10000 for public colleges. However, the average salary for faculty working in public and private colleges are about the same, mostly around $7000 - $12000