BDA 503 - Fall 2017

Ahmet OZMEN

2018-01-09

Contents

Part I: Short and Simple (20 pts)

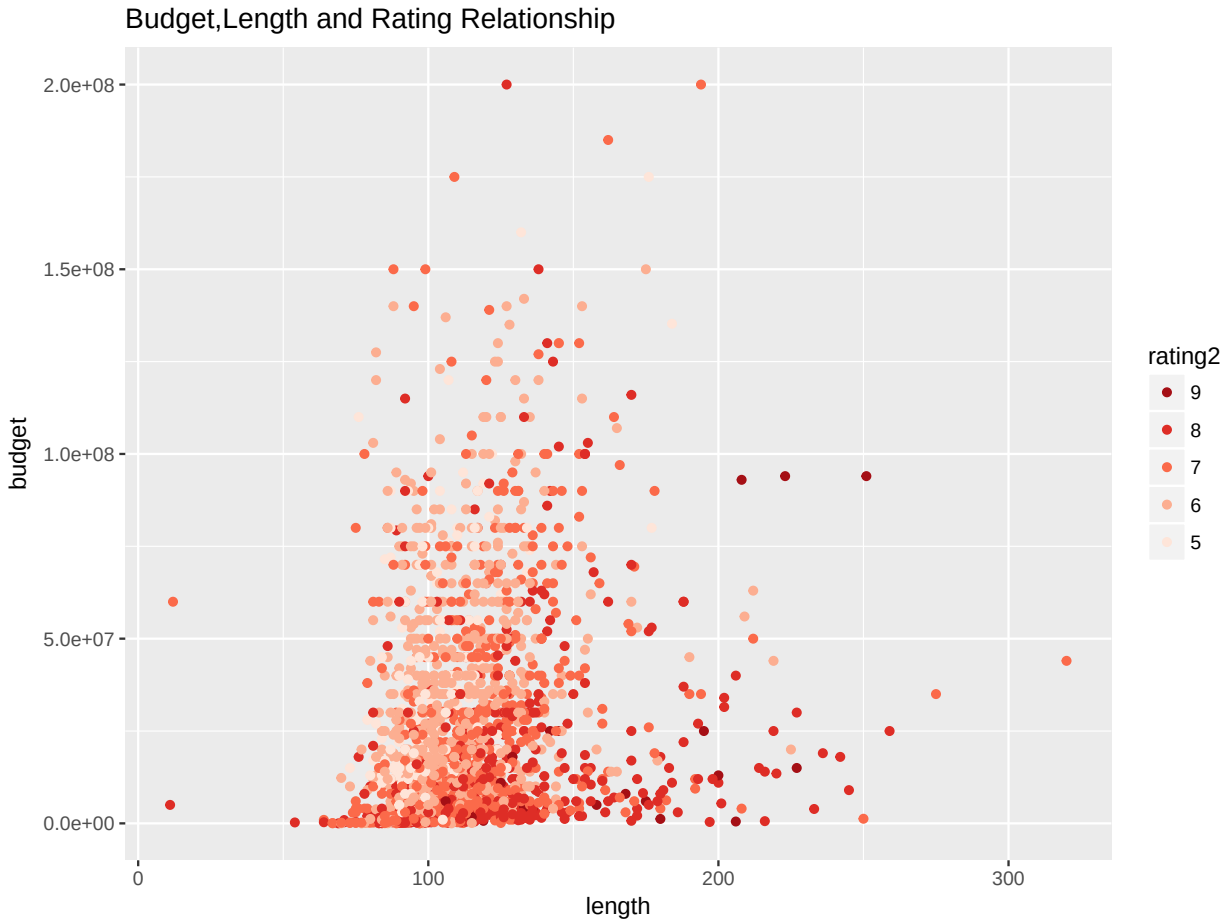
1. Sometimes two y-axis graphs can be a good evaluation for some analysises. The relationship between two measure by one dimension can be evaluated on one chart. For example, Burger King sales trend can be compareted to enfalation trend on chart. Rjunkies's example is a good practice for two y-axis graphs. However, some people overstate two y-axis graphs. They try to evaluate more two measure and they don't can use y-axis graphs correctly. Therefore, they cannot interpret their graphs. I believe that the graphs should be simple. Also y-axis graphs can be simple if people choose correct two measure.

2. If I don't know business related to data, I start to research information about business before looking at the data. I believe that if business isn't known well, we cannot interpret data. After that, we looking at data structure. If data structure is not suitable for processing data, I will transform data structure proper for process, and then I draw the graphs in order to perform data cleaning process that is most important. Later, I continue to analysis the variables. In the example that come from the question, I am looking the relationship between education, gender equality, poverty, job creation and healthcare. When finding at the most important variable for public welfare projects, assume data cleaning process is finished, I perform advance analytics model. I try to find how much effect for the welfare. After performing many model, I have compare them. After I test models, I allocate the fund based on the importance of variable.

3. In time series data, there is auto correlation between time events, but in non time series data there is no event series. Bitcoin price is changing in times accrording to demand and supply, However, diamonds data doesn't has time series, there are atrributes that determine the price. Diamond price is not changing in times events.

4. The single graph is comparing relationship between average rating, movie length and total budget.Here I tried to find wihch varibles can affect average rating. In the chart below, as generally movie length is inreasing, also IMD rating will increase. Budget has no effect on IMD rating.

```
data <- movies %>%  mutate(rating2= as.character(round(rating,digits=0))) %>%
  filter(votes>=1000, budget>0 ,rating>=5)
ggplot(aes(x=length,y=budget),data=data) +
geom_point(aes(color=rating2)) +
scale_color_brewer(type = 'seq',palette='Reds',guide = guide_legend(reverse = T))  +
ggtitle('Budget,Length and Rating Relationship')
```



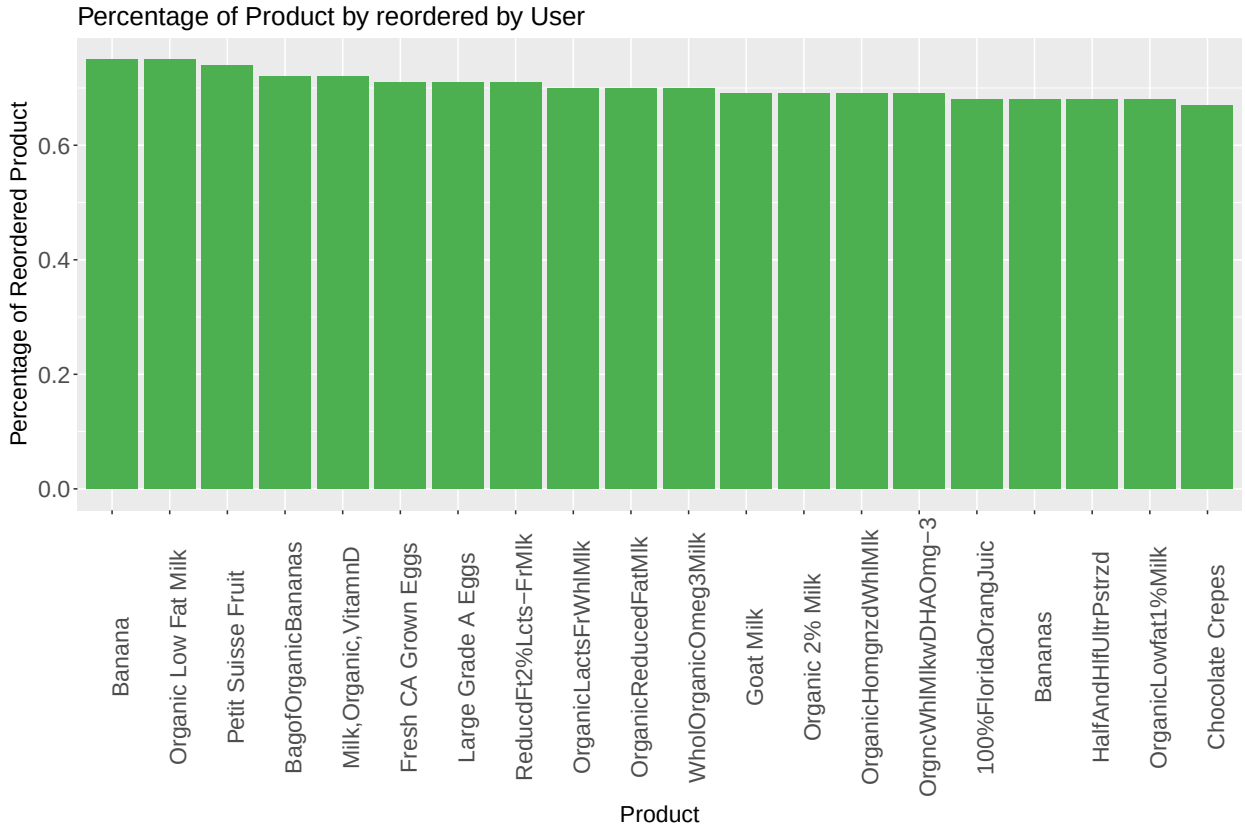
Part II: Extending Your Group Project (30 pts)

Percentage of Product by reordered by User

Here I am looking that users ordered same products again In other words, Which products are mostly ordered again by users. As it is expected, products counsumed by daily are high percentage.

```
reorder_per <- orders_full %>%
  group_by(product_name) %>%
  summarise(per_reordered=round(n_distinct(user_id[reordered==1])/n_distinct(user_id),
                                digits=2),count=n())%>%
  arrange(desc(per_reordered)) %>% filter(per_reordered>0, count>=1000)
```

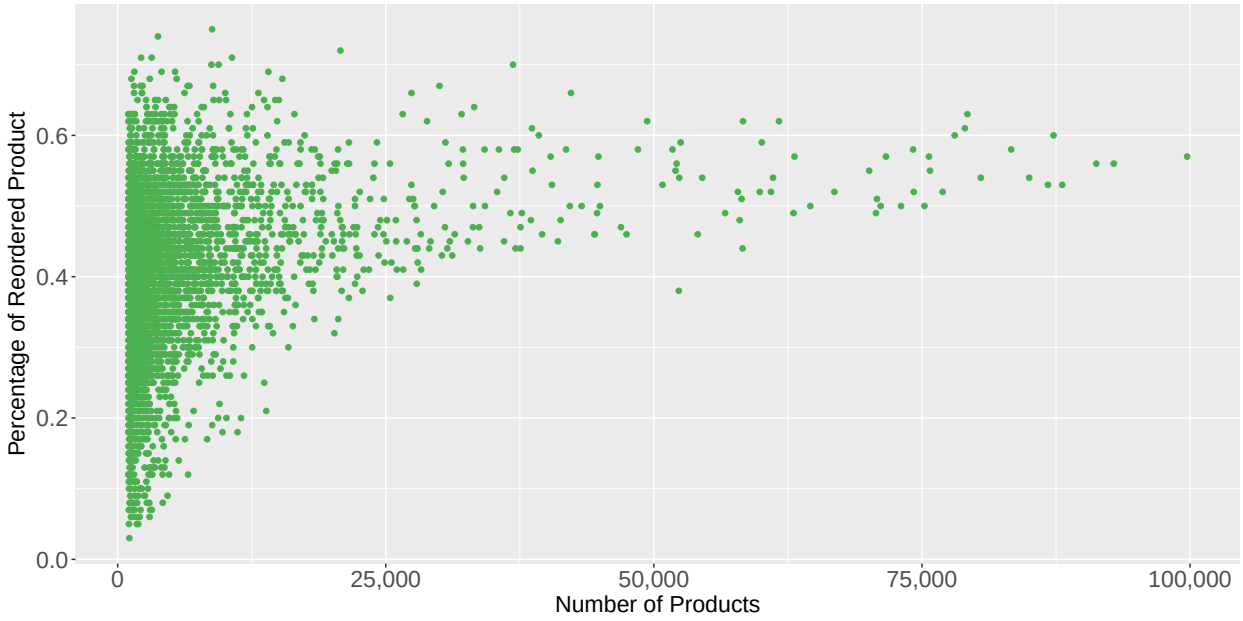
```
ggplot(reorder_per %>%  filter(row_number()<=20),
       aes(x=reorder(product_name,-per_reordered),y=per_reordered)) +
geom_bar(fill="#4caf50", stat = "identity") +
scale_x_discrete(label=function(x) abbreviate(x, minlength=20),name="Product") +
scale_y_continuous(name="Percentage of Reordered Product",labels = comma)+
theme( axis.text.x = element_text(angle =90,size=16) ,
axis.title = element_text(size = 16),plot.title = element_text(size=18),
axis.text.y = element_text(size = 16 )) +
  ggtitle('Percentage of Product by reordered by User')
```



Relationship products ordered ratio by user again and products without reordered

It can be looking at the relationship products ordered ratio by user again and products without reordered. As It shown graph below, there is no direct relationhip between reordered percentage and number of products.

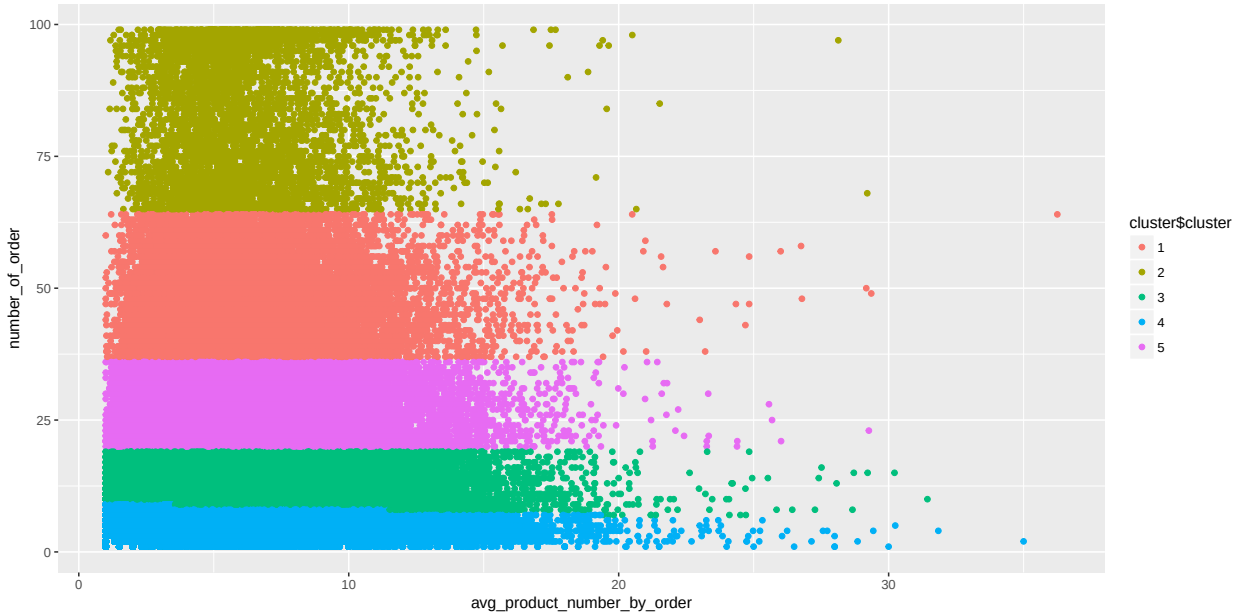
```
ggplot(reorder_per %>%
filter(count<100000),aes(x=count,y=per_reordered)) +
geom_point(color="#4caf50") +
theme(axis.text = element_text(size = 16),
      axis.title = element_text(size = 16),
      plot.title = element_text(size=18) ) +
scale_x_continuous(name = "Number of Products",labels = comma)+
scale_y_continuous(name="Percentage of Reordered Product",
labels = comma)
```



Segmentation of User

In order to make a campaign, we must find correct customers. Therefore, segmentation is a solution. In the graph mentioned below, customers are segmented by number of order and avrage of products number on basket.

```
c_data <- orders_full %>% filter(reordered==1) %>%
group_by(user_id) %>%
summarise(number_of_order=n_distinct(order_id),
avg_product_number_by_order=mean(add_to_cart_order )
) %>% filter(!is.na(avg_product_number_by_order))
cluster=kmeans(c_data[,-1],5)

cluster$cluster <- as.factor(cluster$cluster)
ggplot(c_data, aes(avg_product_number_by_order,
                   number_of_order,
                   color = cluster$cluster)) +
  geom_point()
```

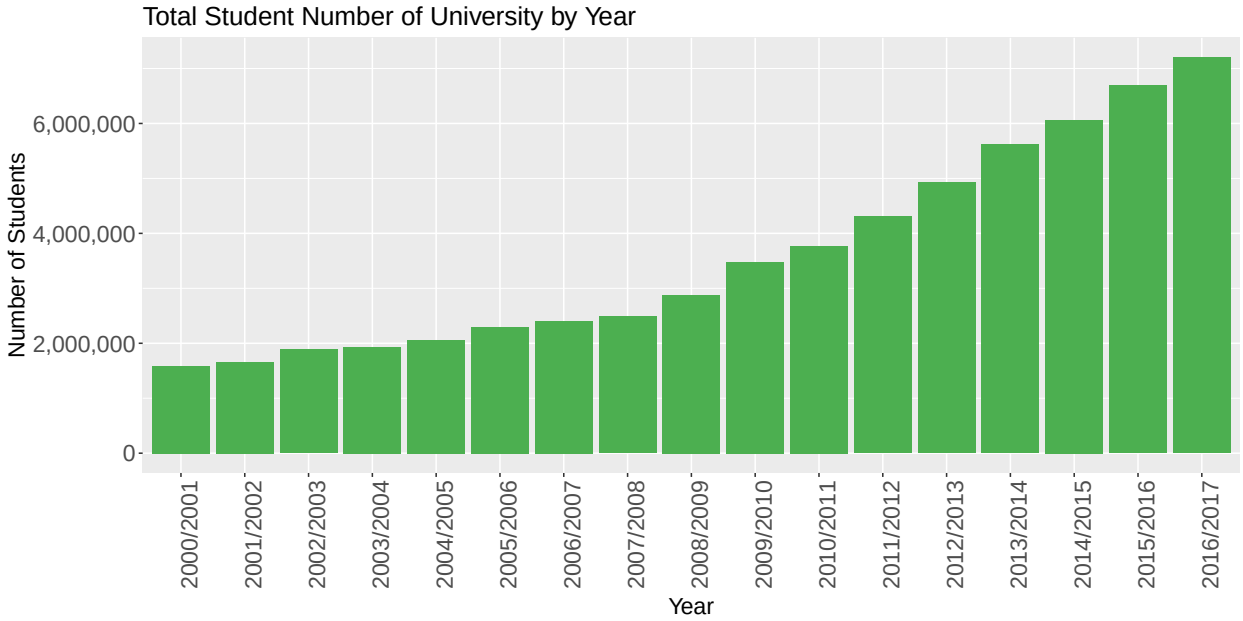


Part III: Welcome to Real Life (50 pts)

Total Student Number(associate degree, bachelor's degree, master's degree, doctor's degree) of University by Year

In Turkey, the number of university students increases dramatically as it is shown on graph mentioned below. Approximately, the number of the student in 2000 is two million, the the number of the student in 2016 is 6.5 million. The number increases four times for seventeen years.

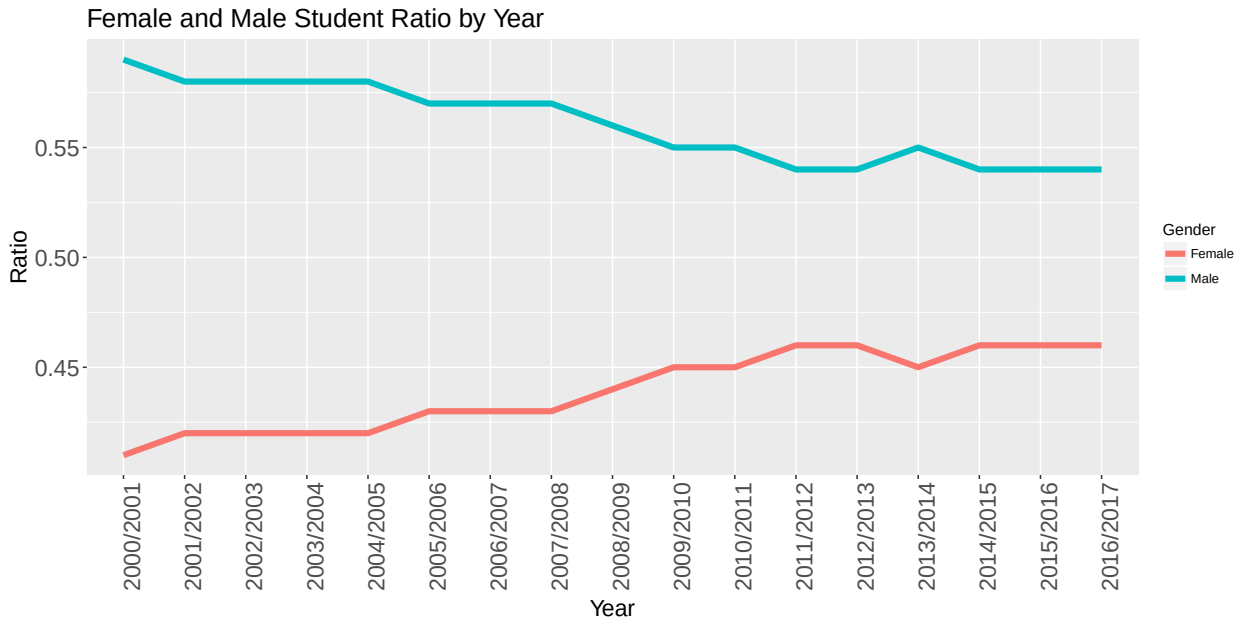
```
result <- student_num %>% group_by(Year) %>% summarise(number_s=sum(Gn_Total))
ggplot(result ,aes(x=Year,y=number_s)) +
geom_bar(fill="#4caf50", stat = "identity") +
scale_x_discrete(label=function(x) abbreviate(x, minlength=20),name="Year") +
scale_y_continuous(name="Number of Students",labels = comma)+
theme( axis.text.x = element_text(angle =90,size=16) ,
       axis.title = element_text(size = 16),plot.title = element_text(size=18),
       axis.text.y = element_text(size = 16 )) +
ggtitle('Total Student Number of University by Year')
```



Female and Male Student Ratio by Year

The gap between the male and female students is closing in the universities from 2000 to 2016. However, this trend is stoping and the gap is becoming same.

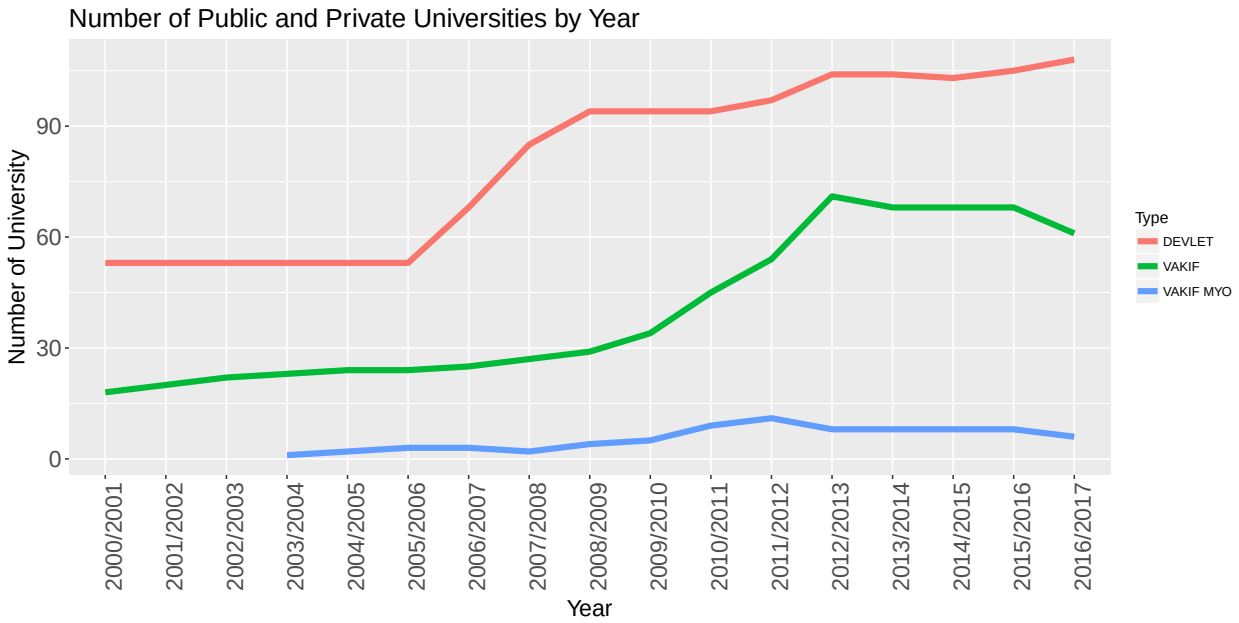
```
female <- student_num %>% group_by(Year) %>%
summarise(value=round(sum(Gn_Woman)/sum(Gn_Total),digits=2),Gender='Female')
male <- student_num %>% group_by(Year) %>%
summarise(value=round(sum(Gn_Man)/sum(Gn_Total),digits=2),Gender='Male')

result <- rbind(male,female)
ggplot(data=result) +
geom_line(aes(x=Year, y=value, colour=Gender,group=Gender),size=2)+
scale_x_discrete(label=function(x) abbreviate(x, minlength=20),name="Year") +
scale_y_continuous(name="Ratio",labels = comma)+
theme( axis.text.x = element_text(angle =90,size=16) ,
       axis.title = element_text(size = 16),plot.title = element_text(size=18),
       axis.text.y = element_text(size = 16 )) +
ggtitle('Female and Male Student Ratio by Year')
```



Number of Public and Private Universities by Year

After 2004 year, the number of universites increases dramatically. It is shown that there is government's policy. However, after 2012 year, some private universities closes.

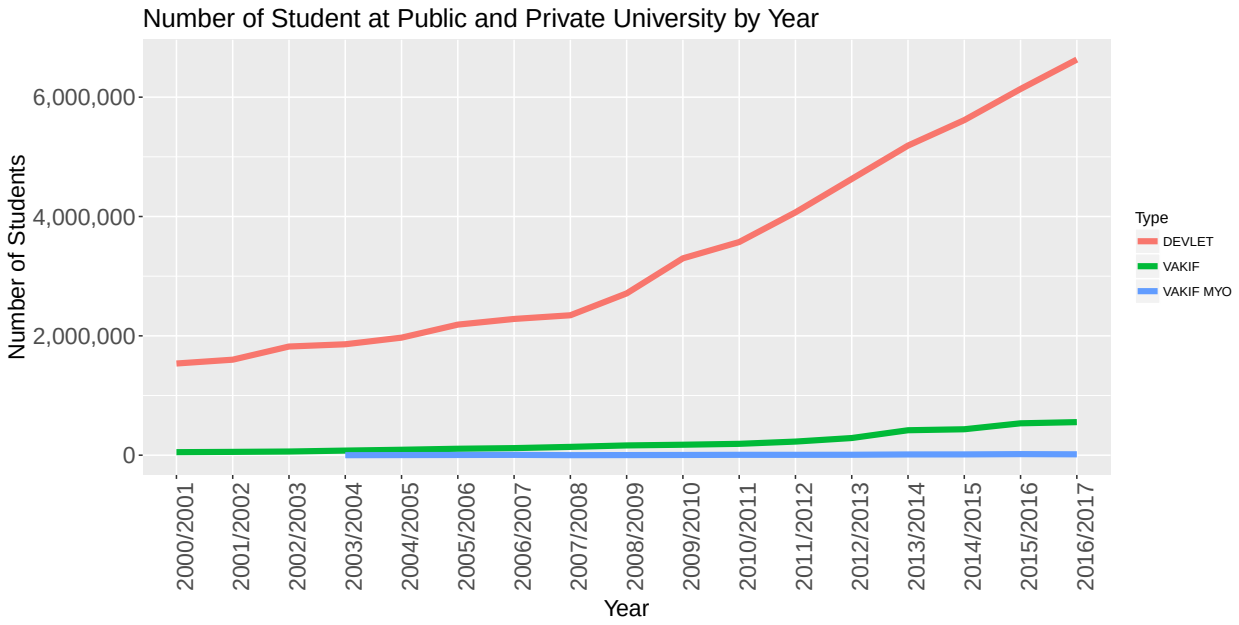
```
result <- student_num %>% group_by(Year,Type) %>% summarise(count=n())

ggplot(data=result) +
geom_line(aes(x=Year, y=count, colour=Type,group=Type),size=2)+
scale_x_discrete(label=function(x) abbreviate(x, minlength=20),name="Year") +
scale_y_continuous(name="Number of University",labels = comma)+
theme( axis.text.x = element_text(angle =90,size=16) ,
       axis.title = element_text(size = 16),plot.title = element_text(size=18),
       axis.text.y = element_text(size = 16 )) +
ggtitle('Number of Public and Private Universities by Year')
```



Number of Student at Public and Private University by Year

As I mentioned above, because after 2004 year, the number of universites increases dramatically, after 2007 year, the number of university students increases.

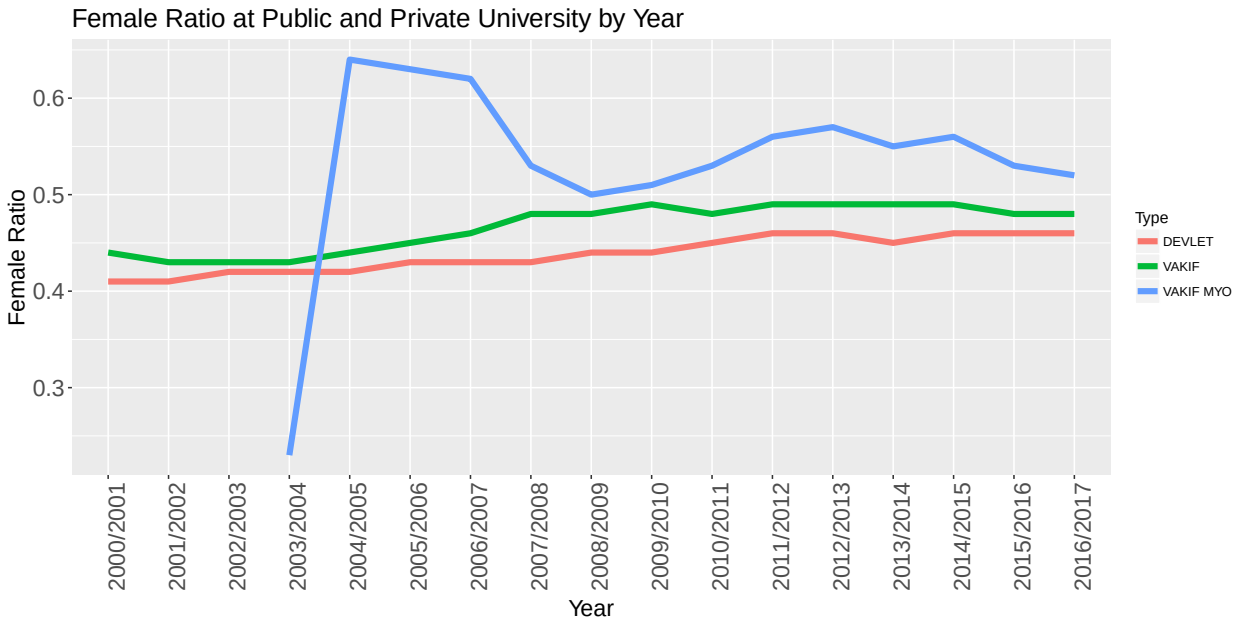
```
result <- student_num %>% group_by(Year,Type) %>% summarise(count=sum(Gn_Total))

ggplot(data=result) +
geom_line(aes(x=Year, y=count, colour=Type,group=Type),size=2)+
scale_x_discrete(label=function(x) abbreviate(x, minlength=20),name="Year") +
scale_y_continuous(name="Number of Students",labels = comma)+
theme( axis.text.x = element_text(angle =90,size=16) ,
       axis.title = element_text(size = 16),plot.title = element_text(size=18),
       axis.text.y = element_text(size = 16 )) +
ggtitle('Number of Student at Public and Private University by Year')
```



Female Ratio at Public and Private University by Year

Female Ratio at VAKIF MYO Universities is more than %50. I think some deparments such as nursing and preschool has effect on female ratio.

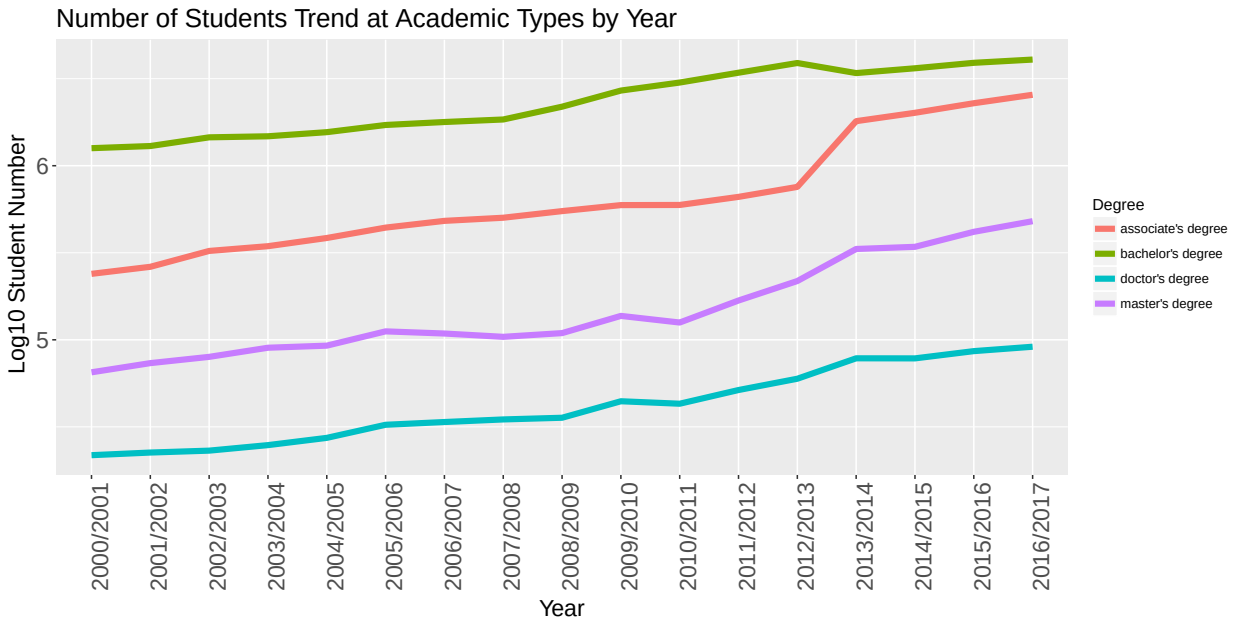
```
result <- student_num %>% group_by(Year,Type) %>%
  summarise(ratio=round(sum(Gn_Woman)/sum(Gn_Total),digits=2))

ggplot(data=result) +
geom_line(aes(x=Year, y=ratio, colour=Type,group=Type),size=2)+
scale_x_discrete(label=function(x) abbreviate(x, minlength=20),name="Year") +
scale_y_continuous(name="Female Ratio",labels = comma)+
theme( axis.text.x = element_text(angle =90,size=16) ,
axis.title = element_text(size = 16),plot.title = element_text(size=18),
axis.text.y = element_text(size = 16 )) +
ggtitle('Female Ratio at Public and Private University by Year')
```



Number of Students Trend at Academic Types by Year

The number of trend at all acememic types are increasing.

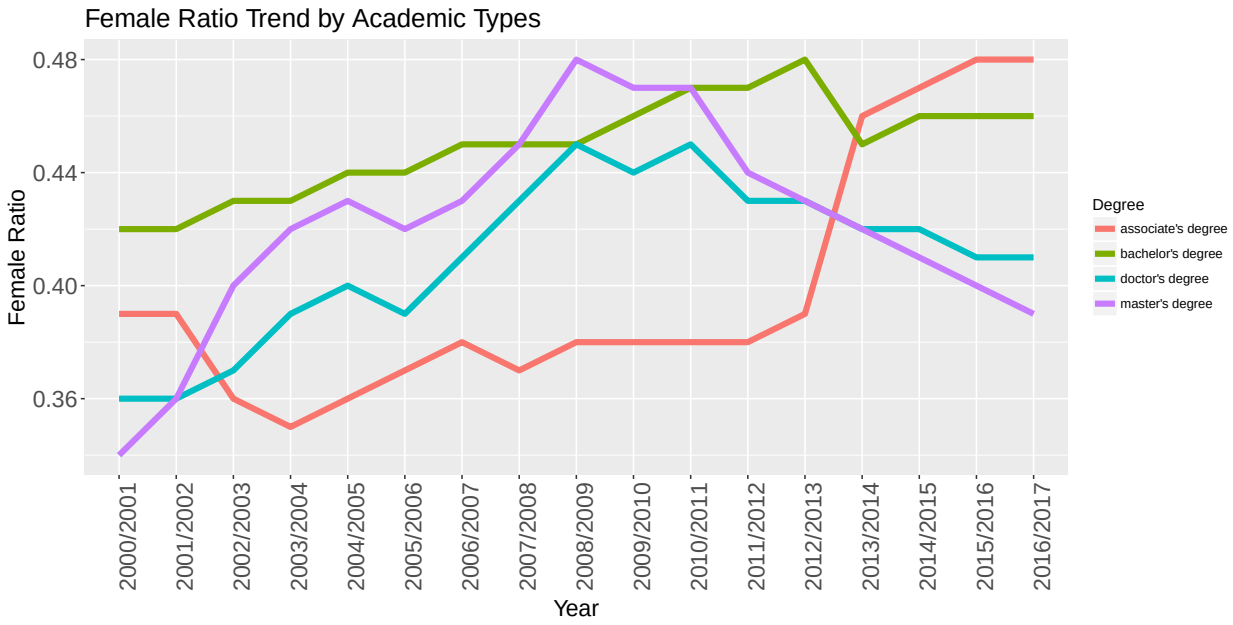
```
bech <- student_num %>% group_by(Year) %>%
summarise(value=log10 (sum(Li_Total)),Degree="bachelor's degree")
mas <- student_num %>% group_by(Year) %>%
summarise(value=log10(sum(Yk_Total)),Degree="master's degree")
doc <- student_num %>% group_by(Year) %>%
summarise(value=log10(sum(Dk_Total)),Degree="doctor's degree")
ass <- student_num %>% group_by(Year) %>%
summarise(value=log10(sum(On_Total)),Degree="associate's degree ")


 result <- rbind(bech,mas,doc,ass)
ggplot(data=result) +
geom_line(aes(x=Year, y=value, colour=Degree,group=Degree),size=2)+
scale_x_discrete(label=function(x) abbreviate(x, minlength=20),name="Year") +
scale_y_continuous(name=" Log10 Student Number",labels = comma)+
theme( axis.text.x = element_text(angle =90,size=16) ,
axis.title = element_text(size = 16),plot.title = element_text(size=18),
axis.text.y = element_text(size = 16 )) +
ggtitle('Number of Students Trend at Academic Types by Year')
```



Female Ratio Trend by Academic Types

Generally female ratio increases till 2013 bachelor's degree, but after 2013, the ratio decreases. The biggets growth occured associate's degree and trend always increases. The ratio for master's and doctor's degree increases till 2008, then later it decreases

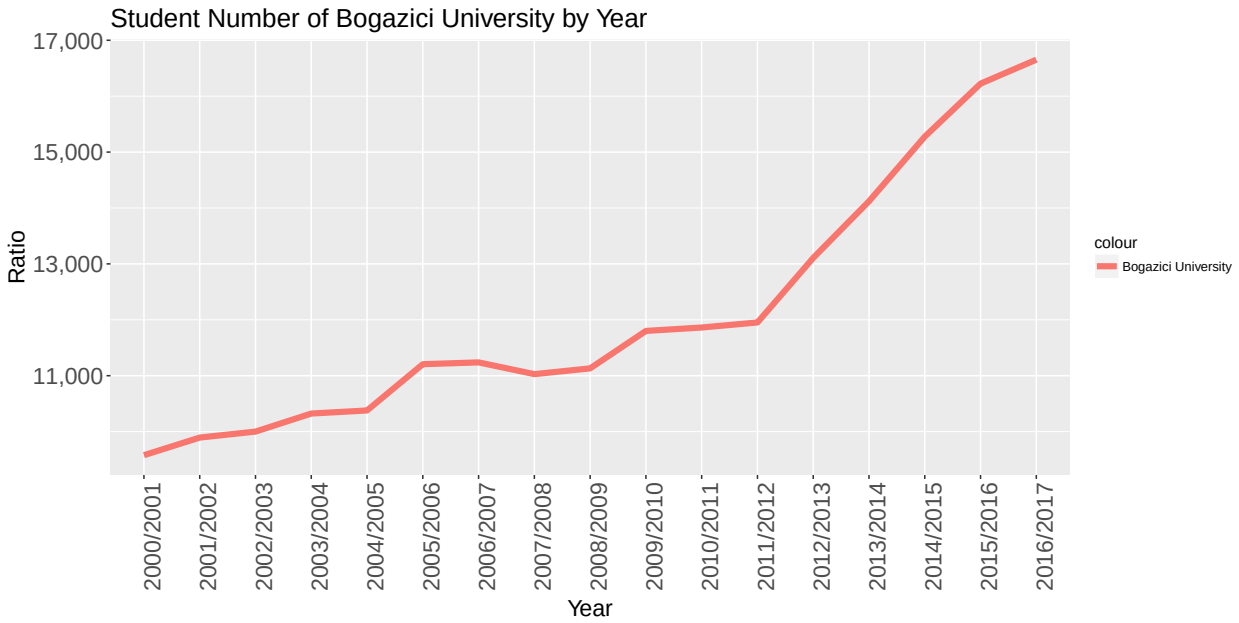
```
bech <- student_num %>% group_by(Year) %>%
summarise(value=round(sum(Li_Woman)/sum(Li_Total),digits=2),Degree="bachelor's degree")
mas <- student_num %>% group_by(Year) %>%
summarise(value=round(sum(Yk_Woman)/sum(Yk_Total),digits=2),Degree="master's degree")
doc <- student_num %>% group_by(Year) %>%
summarise(value=round(sum(Dk_Woman)/sum(Dk_Total),digits=2),Degree="doctor's degree")
ass <- student_num %>% group_by(Year) %>%
summarise(value=round(sum(On_Woman)/sum(On_Total),digits=2),Degree="associate's degree ")


 result <- rbind(bech,mas,doc,ass)
ggplot(data=result) +
geom_line(aes(x=Year, y=value, colour=Degree,group=Degree),size=2)+
scale_x_discrete(label=function(x) abbreviate(x, minlength=20),name="Year") +
scale_y_continuous(name="Female Ratio",labels = comma)+
theme( axis.text.x = element_text(angle =90,size=16) ,
axis.title = element_text(size = 16),plot.title = element_text(size=18),
axis.text.y = element_text(size = 16 )) +
ggtitle('Female Ratio Trend by Academic Types ')
```



Student Number of Bogazici University by Year

Bogazici University is one of the most popular in Turkey. Therefore, I look at the number of students trend. After 2011 year, students quota can be increased by YOK, the students can fail and there can be academic amnesty.

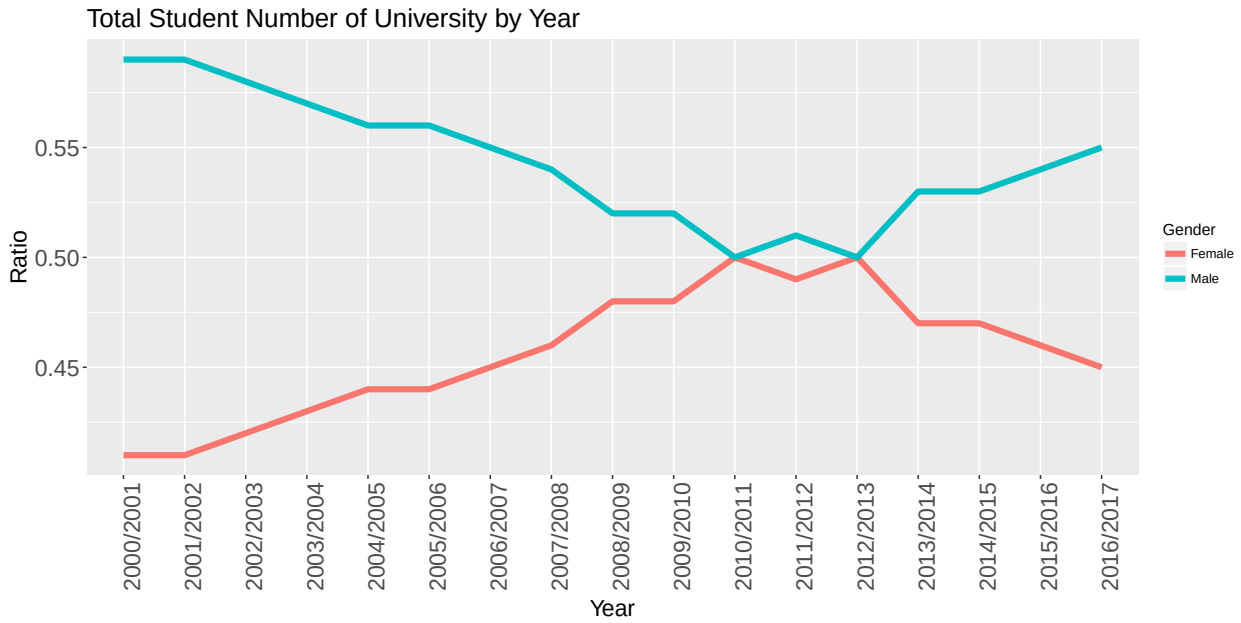
```
result <- student_num %>%
filter(substr(University, 1, 8)=='BOGAZICI') %>%
group_by(Year) %>% summarise(value=sum(Gn_Total))
ggplot(data=result) +
geom_line(aes(x=Year, y=value,colour='Bogazici University',
              group='Bogazici University'),size=2)+
scale_x_discrete(label=function(x) abbreviate(x, minlength=20),name="Year") +
scale_y_continuous(name="Ratio",labels = comma)+
theme( axis.text.x = element_text(angle =90,size=16) ,
       axis.title = element_text(size = 16),plot.title = element_text(size=18),
       axis.text.y = element_text(size = 16 )) +
ggtitle('Student Number of Bogazici University by Year')
```



Female and Male Student Ratio of Bogazici University by Year

The gap between male and female students closes till 2012. After 2012 the male ratio increases.

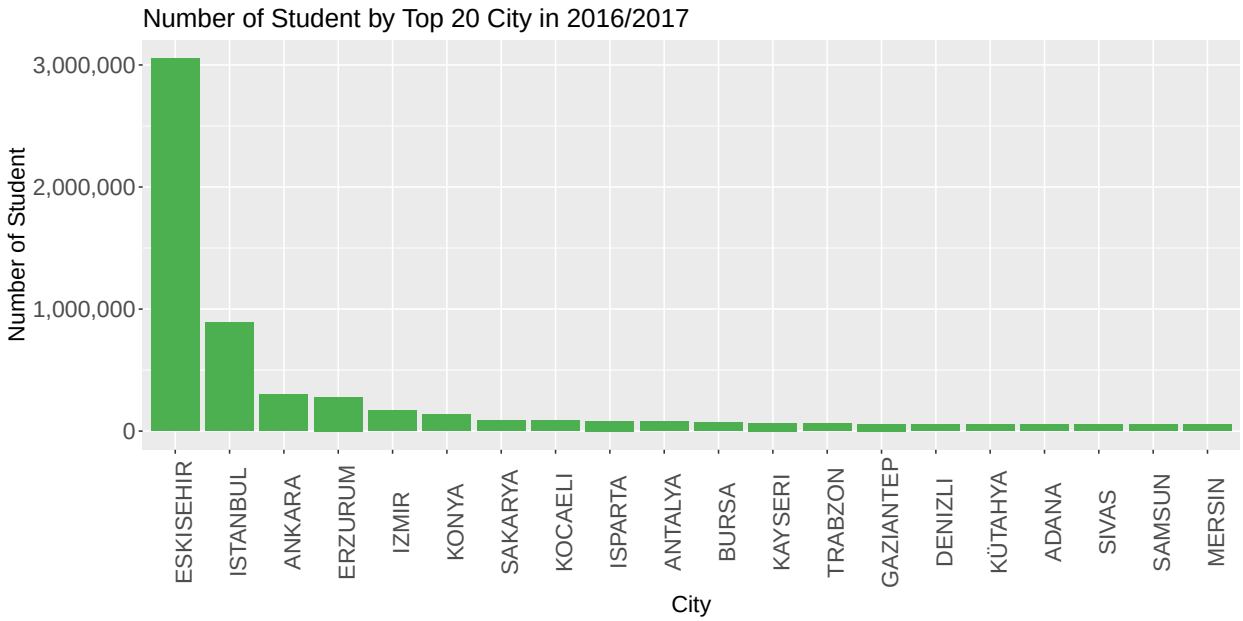
```
female <- student_num %>%
filter(substr(University, 1, 8)=='BOGAZICI') %>%
group_by(Year) %>%
summarise(value=round(sum(Gn_Woman)/sum(Gn_Total),digits=2),Gender='Female')
male <- student_num %>%
filter(substr(University, 1, 8)=='BOGAZICI') %>%
group_by(Year) %>%
summarise(value=round(sum(Gn_Man)/sum(Gn_Total),digits=2),Gender='Male')


 result <- rbind(male,female)
ggplot(data=result) +
geom_line(aes(x=Year, y=value, colour=Gender,group=Gender),size=2)+
scale_x_discrete(label=function(x) abbreviate(x, minlength=20),name="Year") +
scale_y_continuous(name="Ratio",labels = comma)+
theme( axis.text.x = element_text(angle =90,size=16) ,
       axis.title = element_text(size = 16),plot.title = element_text(size=18),
       axis.text.y = element_text(size = 16 )) +
ggtitle('Total Student Number of University by Year')
```



Student Number by Top 20 City in 2016/2017

Anadolu University in Eskisehir has open education faculty so Eskisehir has biggest the number of students. Erzurum surprise me, althought it's population is lower than Izmir's, the number of students in Eskisehir is bigger than the number of students in Ä°zmir

```
result <- student_num %>% filter(Year=='2016/2017') %>%
  group_by(City) %>% summarise(count=sum(Gn_Total)) %>%
  arrange(desc(count))
result <- result %>% filter(row_number()<=20)

ggplot(result ,aes(x=reorder(City,-count),y=count)) +
geom_bar(fill="#4caf50", stat = "identity") +
scale_x_discrete(label=function(x) abbreviate(x, minlength=20),name="City") +
scale_y_continuous(name="Number of Student",labels = comma)+
theme( axis.text.x = element_text(angle =90,size=16) ,
       axis.title = element_text(size = 16),plot.title = element_text(size=18),
       axis.text.y = element_text(size = 16 )) +
ggtitle('Number of Student by Top 20 City in 2016/2017')
```


Number of Student by Top 20 City in 2016/2017
Number of Student
3,000,000
2,000,000
1,000,000
0
ESKISEHIR
ISTANBUL
ANKARA
ERZURUM
IZMIR
KONYA
SAKARYA
KOCAELI
ISPARTA
ANTALYA
BURSA
KAYSERI
TRABZON
GAZIANTEP
DENIZLI
KÜTAHYA
ADANA
SIVAS
SAMSUN
MERSIN
City