

Final BDA 503 - Fall 2017

Feray Ece TOPCU

General Instructions

Your take home final consists of 3 parts. First part is about some simple questions and their answers. These questions might include coding, brief comments or direct answers. Second part is about your group projects. You are asked to make a contribution to your project report with an additional analysis with two/three visualizations. Third part is about gathering real life data and conducting analysis on it.

Here are significant points that you should read carefully.

- Your final starts on January 6, 2018; 11:00. It ends on January 9, 2018; 11:00. Late submissions until January 9, 2018; 23:59 (penalty -25 points).
- Your main submissions will be through Blackboard, no email. Please refrain from posting on your progress journals until January 10, 2018. After then, it is appreciated.
- You will submit RMarkdown generated pdf files. You will submit only a single pdf containing all 3 parts.
- All works should be individual and original. (Single exception: On data gathering in part 3, you can work together and refer to the same RData file.)
- Instructor support will be minimal. I will try to answer technically ambiguous points but I will generally not respond to consulting questions (e.g. “Am I doing it ok?” You probably are, given your overall performance.). Questions are designed to measure your opinions and I don’t want to color your perspective.

Part I: Short and Simple (20 pts)

The purpose of this part is to gauge your apprehension about data manipulation, visualization and data science workflow in general. Most questions have no single correct answer, some don’t have good answers at all. It is possible to write many pages on the questions below but please keep it short. Constrain your answers to one or two paragraphs (7-8 lines tops).

1. What is your opinion about two y-axis graphs? Do you use it at work? Is it a good practice, a necessary evil, or plain horrible? See Hadley Wickham’s point (and other discussion in the topic) before making your argument (<https://stackoverflow.com/a/3101876/3608936>). See an example of two y-axis graph on https://mef-bda503.github.io/gpj-rjunkies/files/project/index.html#comparing__of__accidents__of__departures

Answer 1: I have never use two y-axis graphs at my work. I think usage of this technique depends on who is trying to understand the graph. Firstly, If someone professional on reading graph will see the graph with 2 y-axis, he/she can understand and figure out the point of it but if someone not familiar with graphs, this technique can be horrible. It’s not understandable for beginners (like me) to graphs. Secondly, the person who is familiar with dataset represented on two y -axis graph can figure out the results easily because of his/her knowledge about dataset but if you do not have any idea about dataset this technique can be confusing, misunderstood and open to manipulation. Also, scale of two y axis depends on who scale it, if the person is professional on scale parameters the graph can be understood but I think nobody has to effort to understand the scales if he/she is beginner to graph. To sum up, in my opinion, usage of two y-axis graph technique can be considered if the person who will examine the graph is experienced on graphs or the dataset, otherwise this technique can be confusing.

2. What is your exploratory data analysis workflow? Suppose you are given a data set and a research question. Where do you start? How do you proceed? For instance, you are given the task to distribute funds from donations to public welfare projects in a wide range of subjects (e.g. education, gender equality, poverty, job creation, healthcare) with the objective of maximum positive impact on the society in general. Assume you have almost all the data you require. How do you measure impact? How do you form performance measures? What makes you think you find an interesting angle?

Would you present an argument for a policy that you are more inclined to (e.g. suppose you are more inclined to allocate budget to fix gender inequality than affordable healthcare) or would you just present what data says? In other words, would the (honest) title of your presentation be “Gender Inequality - The Most Important Social Problem Backed by Data” or “Pain Points in Our Society and Optimal Budget Allocation”?

Answer2: I start with deciding what is the aim of analysis, I mean what is expected from the analysis and after that, I always ask this question: can the dataset give meaningful output for this aim? I mean the output of analysis should comply the aim. Secondly, I try to understand data and its properties like what it is about, its size, its form etc. Additionally, I try to figure out which variables should be selected for analysis and understand strong/weak part of the given dataset on this part. Thirdly, I start to data preparation for analysis. I think this part is the most important part of analysis. I know I should identify missing data such as null values, corrupted values etc. After detection, I try to fix or remove them with usual methods and make data type conversion if it is necessary. After this preparation part, I start to do descriptive analysis; preferably with graphs, charts and fancy tables. + I think data analysis should be always objective so, I prefer to present what data says exactly. Just an offer; let the data do all the talking for us. (*“Data Doesn’t Lie. People Do.”*)

3. What are the differences between time series and non time series data in terms of analysis, modeling and validation? In other words what makes Bitcoin price movements analysis different from diamonds (or carat) data set?

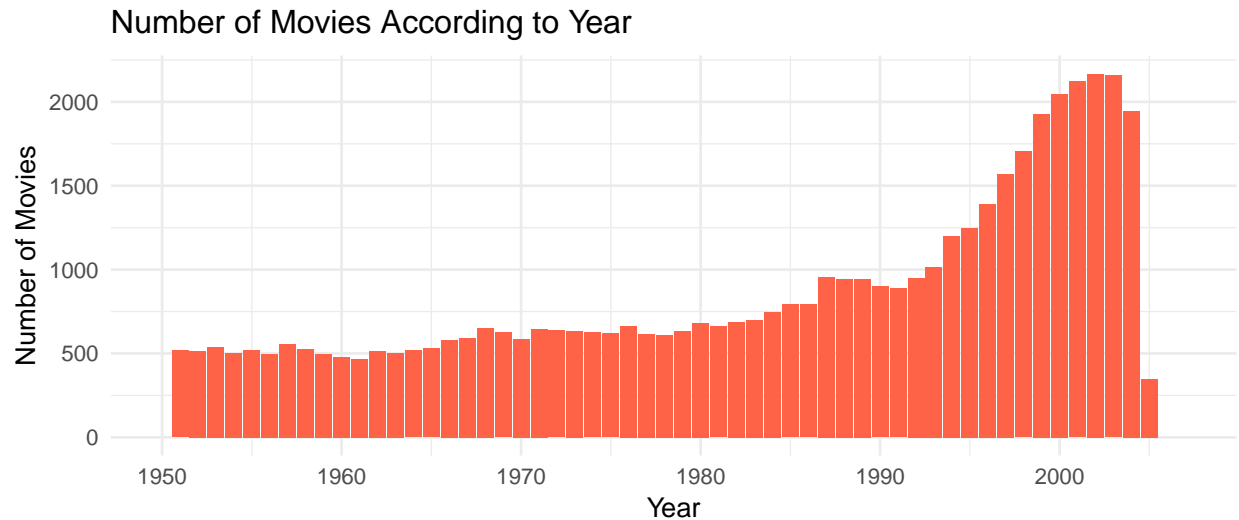
Answer3: In time series, obviously, we can make analysis, modeling and validation with time parameter. It means time can be a dependent variable in my model and I can measure its effect to independent variable and I can explain the results according to time. What does it mean? Assume that; I try to analyze Bitcoin price movements and make prediction about its future. Depends on my results, I can say “In 2020, Bitcoin price will increase”. Therefore, I can tell something about future from now when I have data with time-series. (Wow, Fortuneteller? :)) On the other hand, non-time series data allows me to analyze the current status. For example, when we think about diamond dataset, we can just say something like “diamond price will decrease if it is 10 carat”.

4. If you had to plot a single graph using the data below what would it be? Why? Make your argument, actually code the plot and provide the output. (You can find detailed info about the movies data set in its help file. Use ?movies, after you load ggplot2movies package.)

Answer4: I prefer to plot a graph about relationship between year and number of movies; because I always wonder if there is a growing difference between number of movies year to year. (Year after year, Technology is evaluating, is there any impact of this evaluation on number of movies?)

```
#Number of observation with missing year information : 0,
#max_year = 2005
grp_year <- movies %>% group_by(year) %>% summarise(count_m=n())
g1 <- ggplot(grp_year, aes(x=year, y=count_m)) + geom_bar(stat="identity", fill="tomato") +
  scale_x_continuous(limits = c(1950, 2007)) + theme_minimal() +
  xlab("Year") + ylab("Number of Movies") + ggtitle("Number of Movies According to Year")

g1
```



Part II: Extending Your Group Project (30 pts)

In this part you are going to extend your group project with an additional analysis supported by some visualizations. You are tasked with finding the best improvement on the top of your group project. About one page is enough, two pages tops.

Answer Part 2:

I've recognized we have never used PCA before modeling. So I decided to do PCA on our group dataset which is Big Mart Sales dataset.

- I close the chunk where I read data and clean it. So “combi” is the clean dataset of BigMart Sales Data.
- Firstly for PCA, I convert the factor columns into numeric columns with dummies library.

```
library(dummies)
```

```
## dummies-1.5.6 provided by Decision Patterns
```

```
#1) factor columns into numeric columns
```

```
new_data <- dummy.data.frame(combi, names = c("Item_Fat_Content", "Item_Type",
                                              "Outlet_Establishment_Year", "Outlet_Size",
                                              "Outlet_Location_Type", "Outlet_Type"))
```

- Secondly, remove the unnecessary factor columns from data and use PRCOMP for PCA.

```
new_combi <- select(new_data, -c(Item_Identifier, Outlet_Identifier,
                                Item_Identifier_Str2, Item_Identifier_Str3, PK))
```

```
#2)PCA
```

```
pca <- prcomp(new_combi, scale. = T)
```

- After summary of PCA, there are 44 components. We aim to find the components which compose the maximum variance together. To compute the proportion of variance explained by each component, I divide the variance by sum of total variance.

```
# 3)proportion of variance explained by each component
```

```
#compute standard deviation of each principal component
```

```
std_dev <- pca$sdev
```

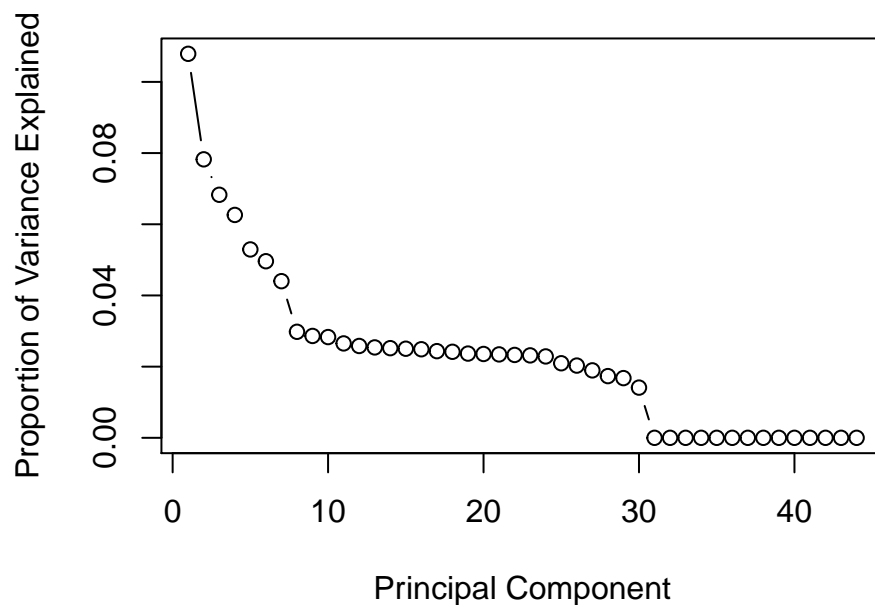
```
pr_var <- std_dev^2
```

```
#proportion of variance explained
prop_varex <- pr_var/sum(pr_var)
#just check values:
prop_varex[1:3]
```

```
## [1] 0.10788716 0.07823868 0.06828268
```

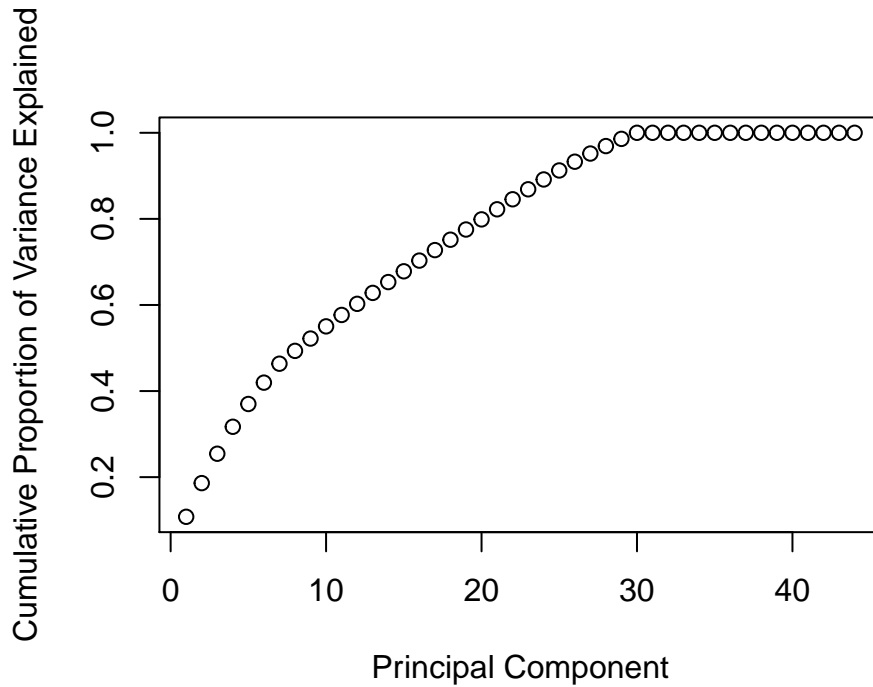
- After that, need to visualize it to decide which of them should be in model as predictor. A *scree plot* is used to access components or factors which explains the most of variability in the data. It represents values in descending order.

```
plot(prop_varex, xlab = "Principal Component",
      ylab = "Proportion of Variance Explained",
      type = "b")
```



- The plot above shows that ~ 30 components compose around 98% of variance of dataset. It means, I have reduced 44 predictors to 30 by using PCA.

```
#4)#cumulative scree plot (to confirm the scree plot)
plot(cumsum(prop_varex), xlab = "Principal Component",
      ylab = "Cumulative Proportion of Variance Explained",
      type = "b")
```



- This plot also shows that 30 components results in variance close to $\sim 98\%$. Therefore, **we should have selected [PC1 to PC30] and proceeded to the modeling stage. Maybe, we had have better results.**

Part III: Welcome to Real Life (50 pts)

- Gather the data, bind them together and save in an .RData file. Make .RData file available online for everybody. Provide the data link in your analysis.

Answer 3.a:

- I gather my dataset from <https://istatistik.yok.gov.tr/> .This dataset includes the number of students of universities located on Marmara area and It is available on this link.
- Variables
- **Universite_Adi:** Name of universities.
- **Uni_Tur:** Type of university; private or public.
- **Il:** the city where university is located.
- **Ogrenim.Turu:** Education type such as “Birinci Ogretim”, “Ikinci Ogretim” and etc.
- **X_Erkek:** number of male students in the specific degree. (X is explained below.)
- **X_Kadin:** number of female students in the specific degree.
- **X_Toplam:** number of all students in the specific degree.
- **Bas_Yili:** beginning year of the term.
- **Bitis_Yili:** end year of the term.

*** “X” represents academic level and can be OL:Onlisans, Lis:Lisans, YL:Yukseklisans, Dr:Doktora, Gn:Genel. For each kind of X, there is three column includes male,female and total student numbers.

- Perform EDA on the data you collected based on the theme you decided on. Keep it short. One to two pages is enough, three pages tops. If you are interested and want to keep going, write a data blog post about it. I will not grade it but I can share it on social media.

Answer 3.b: The aim of this analysis can be summarised as examine the number of students of universities located on Marmara Area and find which city and university has the biggest female student ratio according to university type. This analysis includes total number of all students (includes all education type and all academic level)

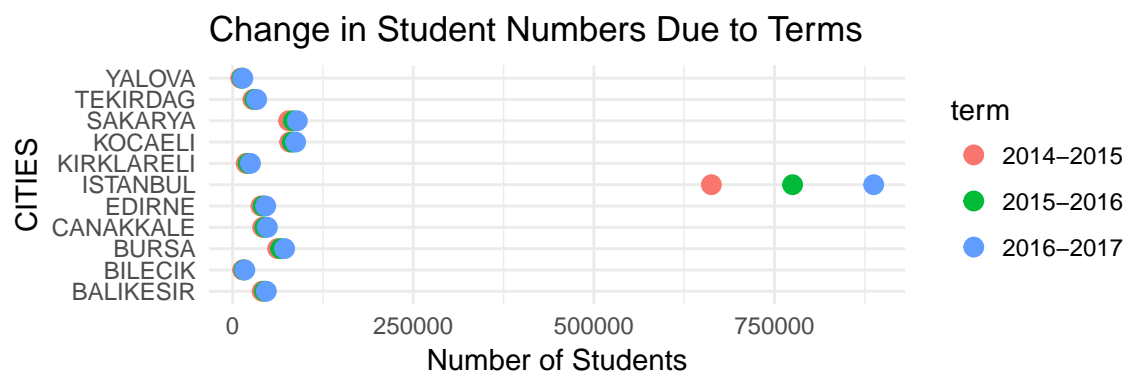
```
#Download RData from github and load it into DF:
setwd("C:/Users/ecetp/Downloads/MEF/BDA 503 - R/Final")
download.file("https://ferayece.github.io/BDA/files/BDA503R/mainData.RData","mainData.RData")
d <- get(load("mainData.RData"))
#change bad column names and remove bad rows:
colnames(d)[1] <- 'Uni'
colnames(d)[4] <- 'Ogrenim_Turu'
d <- d[ -which(grepl('BEZM-I Ã', d$Uni)) , ]
```

- Generate a new categorical column according to beginning and end year of the term.

```
d <- d %>% mutate(term='x')
d$term <- ifelse(d$Bas_Yili==2014,'2014-2015',ifelse(d$Bas_Yili==2015,'2015-2016','2016-2017'))
```

- Group by data due to Cities. (Take just “Toplam” data on “Ogrenim Turu” column.)

```
sum_d <- d %>% filter(Ogrenim_Turu=="TOPLAM")
#Group data due to City and Term.
grp_d <- sum_d %>% group_by(Il,term) %>% summarise(sum_general = sum(Gn_Toplam))
ggplot(data=grp_d,aes(x=sum_general,y=Il,color=term)) + geom_point(size=3) +
  ggtitle("Change in Student Numbers Due to Terms") + xlab("Number of Students") +
  ylab("CITIES") + theme_minimal()
```



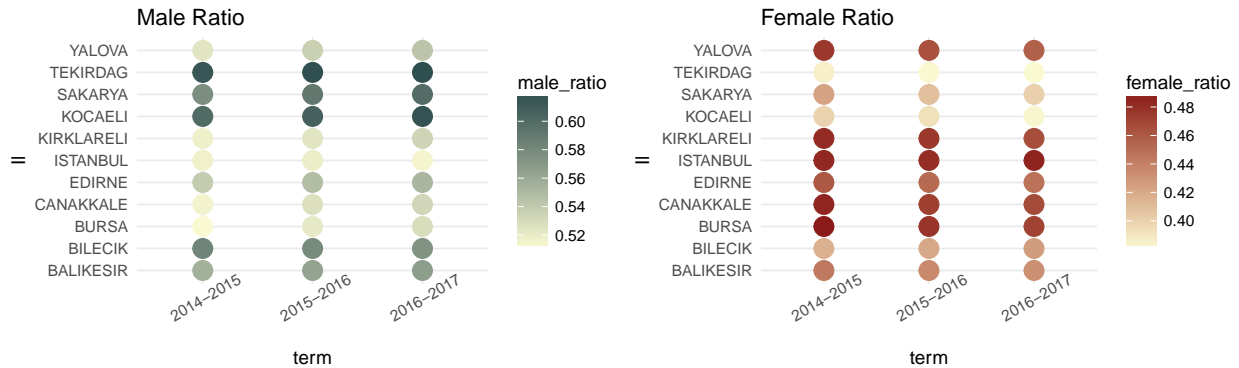
- As it is seen on the graph, Istanbul has the crowded universities as expected. According to graph, Istanbul is always getting crowded faster. When we look at other cities, Sakarya and Bursa are more crowded than 2014-2015 term, now. Let's examine number of students according to their gender.

```
#Group data due to City and Term.
grp_d <- sum_d %>% group_by(Il,term) %>%
  summarise(sum_general = sum(Gn_Toplam))
sum_gender <- sum_d %>% select(Uni,Uni_Tur,Il,term,Gn_Erkek,Gn_Kadin,Gn_Toplam)
grp_city_sum_gender <- sum_gender %>% group_by(Il,term) %>%
  summarise(sum_male=sum(Gn_Erkek),
            sum_female=sum(Gn_Kadin),
            sum_gn=sum(Gn_Toplam),
            male_ratio = round(sum(Gn_Erkek)/sum(Gn_Toplam),3),
            female_ratio=round(sum(Gn_Kadin)/sum(Gn_Toplam),3))
g0 <- ggplot(grp_city_sum_gender,aes(term)) + geom_point(aes(y=Il,color=male_ratio),size=5) +
  scale_color_continuous(low="lightgoldenrodyellow",high="darkslategray",guide="colourbar") +
```

```

ggtitle("Male Ratio") + theme_minimal() +
theme(axis.text.x = element_text(angle = 30))
g1 <- ggplot(grp_city_sum_gender,aes(term)) + geom_point(aes(y=Il,color=female_ratio),size=5) +
scale_color_continuous(low="lightgoldenrodyellow",high="firebrick4",guide="colourbar") +
ggtitle("Female Ratio") + theme_minimal() +
theme(axis.text.x = element_text(angle = 30))
grid.arrange(g0,g1,nrow=1)

```



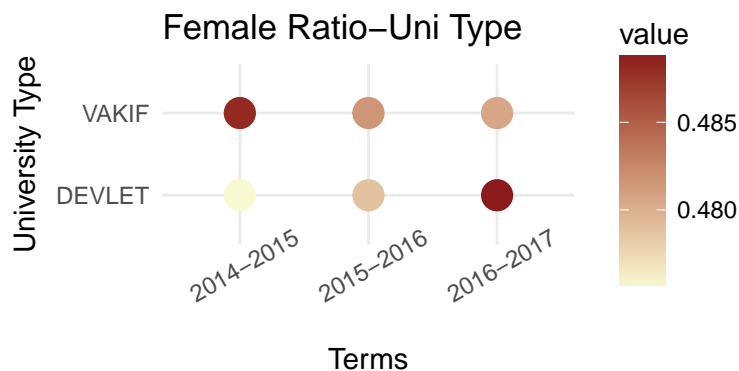
- As seen on Male and Female Ratio graph; Kocaeli, Tekirdag and Sakarya mainly has male students on each terms from 2014 to 2017 while Istanbul, Edirne and Bursa are more preferable for female students. On female ratio graph, Kocaeli and Sakarya are salient because they are losing female students term by term. So, focus on Istanbul again.

```

ist <- sum_d %>% filter(Il=='ISTANBUL')
ist <- ist %>% filter(Uni_Tur != 'VAKIF MYO')
grp_ist <- ist %>% group_by(Uni_Tur,term) %>%
summarise(ist_female_ratio=round(sum(Gn_Kadin)/sum(Gn_Toplam),4),
ist_male_ratio=round(sum(Gn_Erkek)/sum(Gn_Toplam),4))

#reshape data:
need_melt <- grp_ist %>% select(Uni_Tur,term,ist_female_ratio)
library(reshape)
m_female<- melt(need_melt[1:3], id=c("Uni_Tur","term"))
ggplot(m_female,aes(term)) +
geom_point(aes(y=Uni_Tur,color=value),size=5) +
scale_color_continuous(low="lightgoldenrodyellow",high="firebrick4",guide="colourbar") +
ggtitle("Female Ratio- Uni Type") +
xlab("Terms") + ylab("University Type") + theme_minimal() +
theme(axis.text.x = element_text(angle = 30))

```

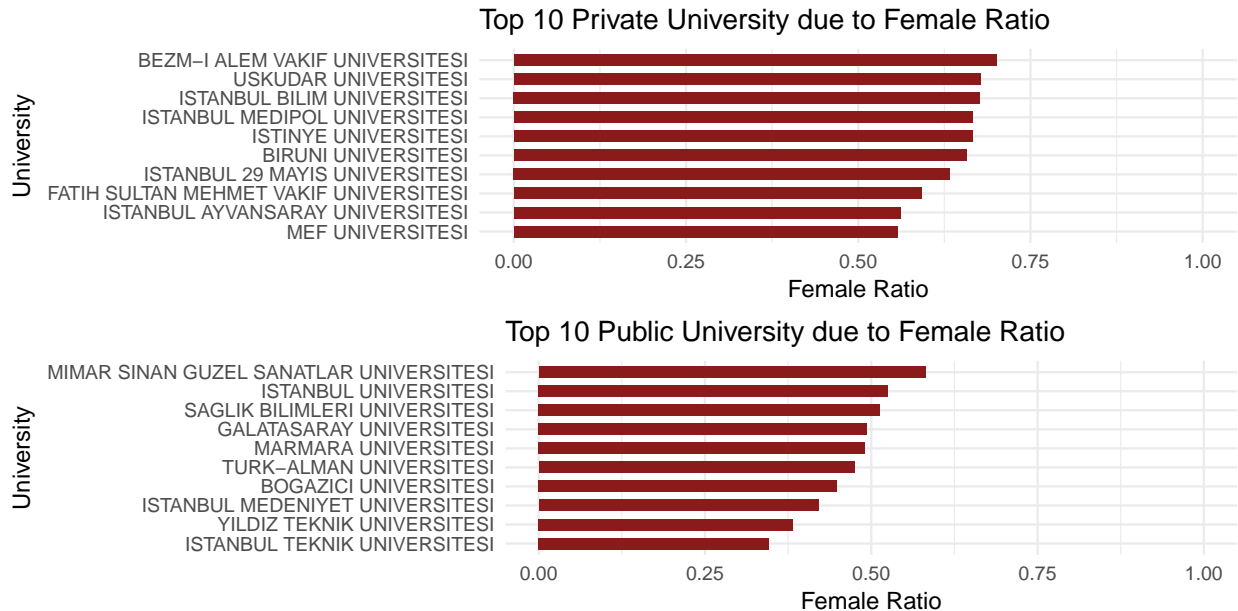


- As shown on graph, female ratio on public universities increasing term to term. So, female students tend to prefer public universities now. Therefore, Looking at top 10 universities which has the most female student ratio at 2016-2017 term.

```
ist1617 <- ist %>% filter(term=='2016-2017')
female_ist1617 <- ist1617 %>% mutate(female_ratio =(Gn_Kadin/Gn_Toplam)) %>%
  select(Uni,Uni_Tur,female_ratio)
pri.female_ist1617 <- female_ist1617 %>% filter(Uni_Tur=='VAKIF') %>% top_n(10) %>%
  arrange(desc(female_ratio))
pri.female_ist1617$Uni[1] ='BEZM-I ALEM VAKIF UNIVERSITESI'
pub.female_ist1617 <- female_ist1617 %>% filter(Uni_Tur=='DEVLET') %>% top_n(10) %>%
  arrange(Uni,desc(female_ratio))

set.seed(2)
g0<- ggplot(pri.female_ist1617,aes(x=reorder(Uni,female_ratio),y=female_ratio)) +
  geom_bar(stat="identity",fill="firebrick4",width = 0.6) + coord_flip() + theme_minimal() +
  scale_y_continuous(limits = c(0,1)) + xlab("University") + ylab("Female Ratio") +
  ggtitle("Top 10 Private University due to Female Ratio")

g1 <- ggplot(pub.female_ist1617,aes(x=reorder(Uni,female_ratio),y=female_ratio)) +
  geom_bar(stat="identity",fill="firebrick4",width = 0.6) + coord_flip() + theme_minimal() +
  scale_y_continuous(limits = c(0,1)) + xlab("University") + ylab("Female Ratio") +
  ggtitle("Top 10 Public University due to Female Ratio")
grid.arrange(g0,g1,nrow=2)
```



* In conclusion, firstly Istanbul is the most crowded city according to student numbers and It's getting crowded faster than other cities on Marmara area as expected (depends on number of university numbers in the city). Secondly, according to gender ratio graphs in cities, Istanbul is always more preferable for female students. So, focused on Istanbul universities for female ratio. when we examine the female ratio in universities located in Istanbul; It is seen that female ratio on public universities increasing term by term. **To sum up; while universities located in Istanbul has an huge increase in number of students of all education types (Onlisans, Lisans and etc.) and academic levels term to term ; female students tend to prefer public over private universities and BEZM-I ALEM Uni.(Private) and MIMAR SINAN Uni.(Public) has the biggest female student ratio in Istanbul on 2016-2017 term.**