Ahmet Yetkin Eser - Final

BDA 503 - Fall 2017

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.

Questions

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

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"?

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?

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.)

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.

Part III: Welcome to Real Life (50 pts)

As all of you know well enough; real life data is not readly available and it is messy. In this part, you are going to gather data from Higher Education Council's (Y??K) data service. You can use all the data provided on https://istatistik.yok.gov.tr/ . Take some time to see what are offered in the data sets. Choose an interesting theme which can be analyzed with the given data and collect relevant data from the service. Some example themes can be as follows.

- Gender disparity in the academic faculty.
- Change in the number of people in different academic positions in years.
- Professor/student ratios.
- Capacities in different departments.
- Comparative undergraduate / graduate student populations.
- Number of foreign students/professors and where they come from.

a - 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. You can work together with your friends to provide one comprehensive .RData file if it is more convenient to you. (You don't need to report any code in this part.)

b - 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.

Answers

R Packages Used in Analysis

```
library(ggplot2movies)
library(tidyverse)
library(gridExtra)
library(stringr)
library(knitr)
```

Part - I

1 - In my opinion two axis graph is useful if it used to give more describtion about idea. I use it in work, but i do not use them like accidents and departures graph from R Junkies. If i were created graph from departures and accidents data :

I will use,

- First y axis for number of departures with bar(If it will seen good i will add accidents as a second bar on upper of departures),
- Second y axis for accidents/departures ratio with line.
- x axis for year like R junkies
- So i will show that number of departures are increasing with time, but accidents/departures ratio are decrasing with time.
- In work, I get used to two axis graph like answer of question Part_1.4.

2 - You can find my exploratory data analysis workflow on below.

- I try to understands columns of data(their mean, median, distributions, is it any NULL columns, data is clean or not)
- Thinking about which columns can i used.
- Looking inside of data randomly.
- If there are null columns, i try to find information from other rows to fill them.
- If there is an objective of analysis, I do my analysis based on this objective. If I do not or my objective is general, I try to give more neutral results.
- About your questions, i try to be honest and i prefer for my title "Pain Points in Our Society and Optimal Budget Allocation"
- Numbers don't lie, people lie with the numbers

3 - Time series data has time in it. Analysis results can change with time. Non time series data has not time in it.

For Example:

- Bitcoin data has time in dataset and bitcoin price change with events in that timeline. But diamonds dataset does not time in it. Diamonds prices are changed not with time, they are changed with diamonds features.

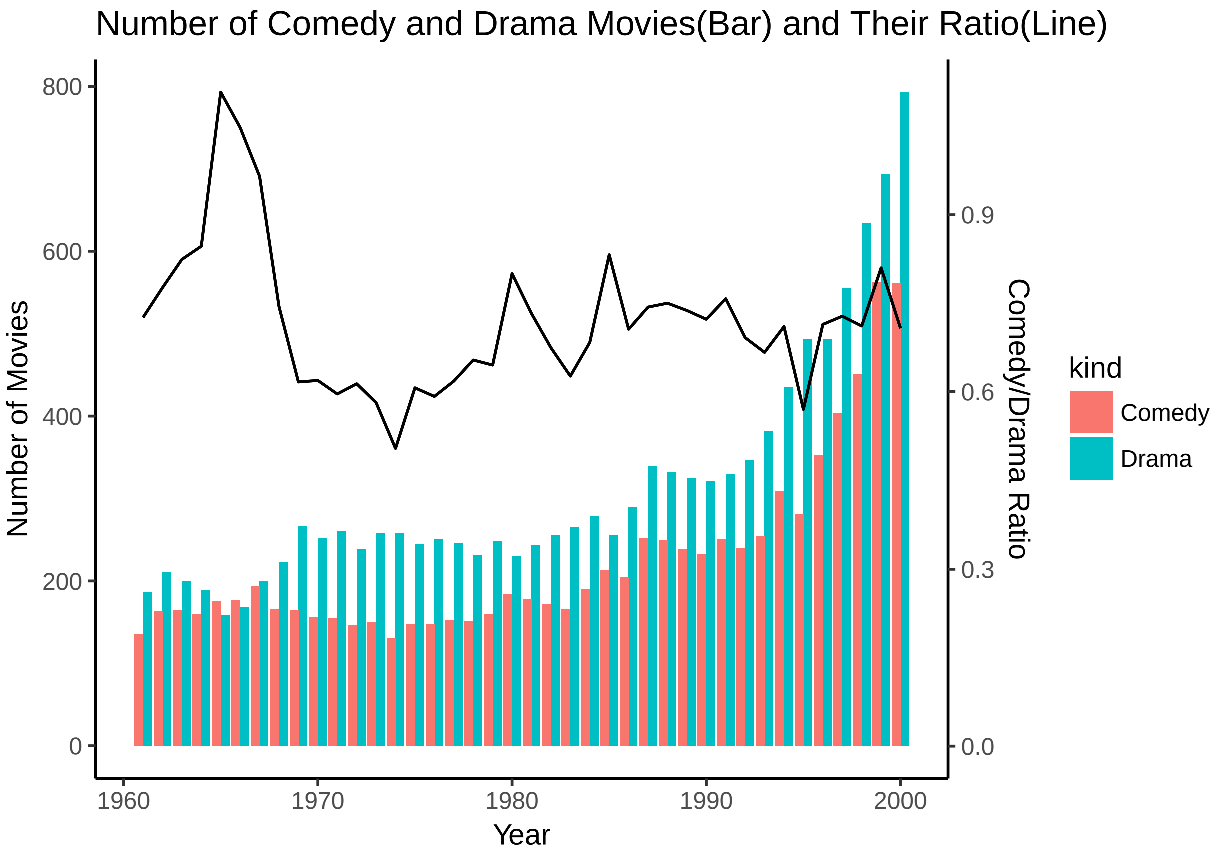
- If we put time in diamonds data, analysis diamonds price change with diamonds features and time. We can say diamonds data is time series data.

4 - I want to analysis what is the evalution of ratings comedy and drama films by years. But I do not have columns in my data which rating is given which year.

So I changed my idea.

I analysis how many comedy and drama movies released each year and what are the ratio between them.

```r
comedy <- movies %>%
  filter(year > 1960, year <= 2000) %>%
  group_by(year) %>% summarise(comedy_drama_ratio = sum(Comedy)/sum(Drama),
            number = sum(Comedy), kind = 'Comedy'
            )

drama <- movies %>%
  filter(year > 1960, year <= 2000) %>%
  group_by(year) %>% summarise(comedy_drama_ratio = sum(Comedy) / sum(Drama),
            number = sum(Drama), kind = 'Drama'
            )

movie2 <- union_all(comedy,drama) %>% arrange(year)
normalizer <- max(movie2$number)/max(movie2$comedy_drama_ratio)

ggplot(movie2, aes(x = year, y = number/normalizer, fill = kind)) +
  geom_bar(aes(y = number),stat = "identity",position = position_dodge()) +
  geom_line(aes(y = comedy_drama_ratio * normalizer)) +
  ggtitle("Number of Comedy and Drama Movies(Bar) and Their Ratio(Line)") +
  scale_y_continuous(sec.axis = sec_axis(trans = ~.* 1 / normalizer, name = 'Comedy/Drama Ratio'))+ labs
```



Results :

- Number of comedy and drama movies released increase almost every year.
- More drama movies released every year except one year in 1960s.
- Their ratio(Comedy/Drama) swing between 0.6 to 0.9 almost every year.

Part - II

You can find Big Mart Sales project page here.

- I thought missing piece of our projects is we do not know to data very well. We try to learn new language(R), new methods, new graphs and applied them our project. But we do not learn about our data, we do not own it because we lost in what we learn.
- In my opinion, we needed to look our data with some select and filter functions, but we did not.
- For example for “FDB02” item :

```
bigMart %>% select(Item_Identifier, Item_Weight, Item_Visibility
                   ,Item_MRP,Item_Outlet_Sales) %>%
  filter(Item_Identifier == "FDB02") %>%
  kable()
```

Item_Identifier	Item_Weight	Item_Visibility	Item_MRP	Item_Outlet_Sales
FDB02	9.695	0.0291588	174.537	1235.059
FDB02	9.695	0.0292234	175.437	2999.429
FDB02	9.695	0.0291400	176.337	3528.740
FDB02	12.600	0.0290230	177.837	6704.606
FDB02	9.695	0.0292831	175.137	3705.177

```r
median(bigMart$Item_Weight)
```

```
## [1] 12.6
```

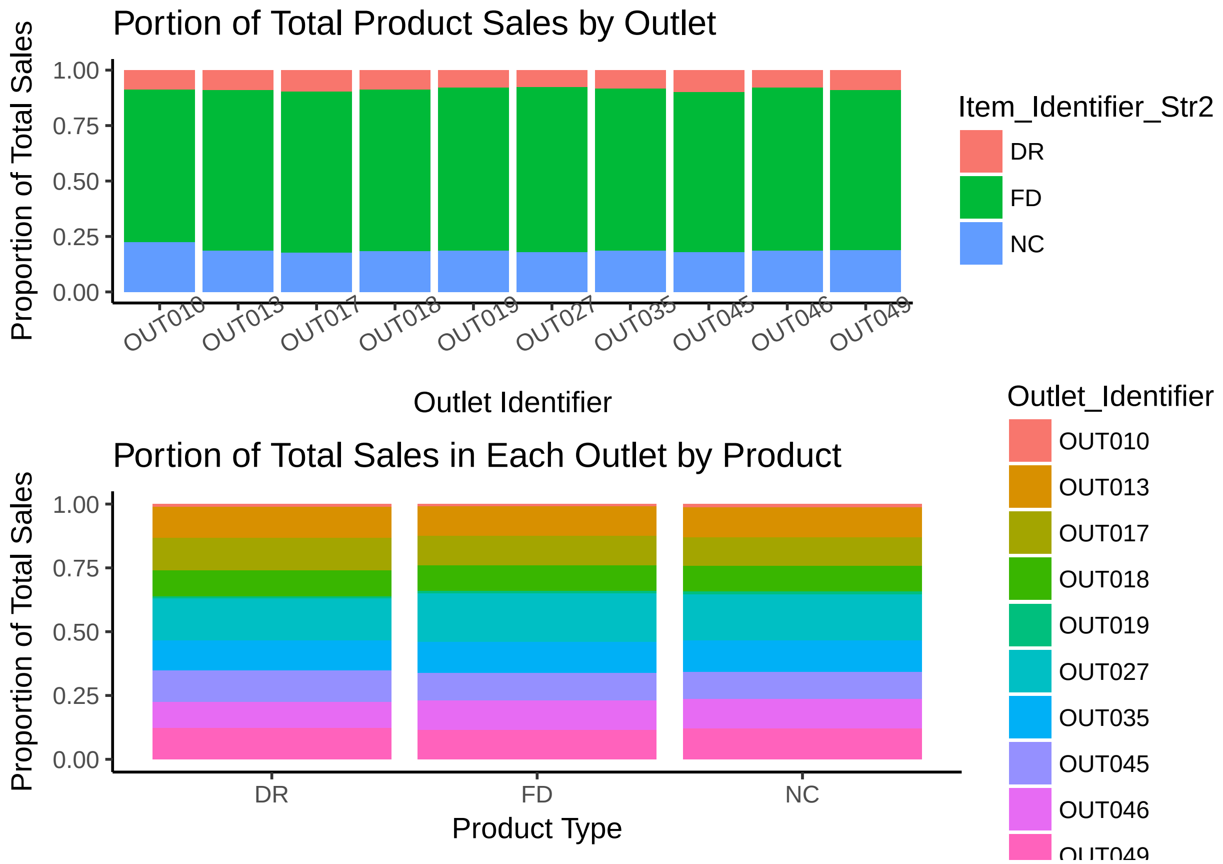
- If we look our data set with some filter and select, we can see "FD502" Item Weight is different in one row. Cause of this difference is we equalize it median of dataset.If we research our dataset before, we were find it and equalize 9.695 but we do not researh enought and we equalize it median of dataset. There were so many columns we equalize median of dataset.
- And we do not descriptive analysis on our data. We do not have analysis like which kind of product sell which outlet mostly or which outlet sells how many number of products.

```r
bigMart_Summary <- bigMart %>% group_by(Outlet_Type,Outlet_Identifier,Item_Identifier_Str2)%>%
  summarise(Sales = round(sum(Item_Outlet_Sales))
            , nProduct = n_distinct(Item_Identifier)
            ,nSales = round(sum(Item_Outlet_Sales/Item_MRP))
            ,avgPrize=round(Sales/nSales)
            )

bigMart_Summary %>%
  arrange(desc(Sales)) %>% filter(Item_Identifier_Str2 == "NC") %>% kable()
```

Outlet_Type	Outlet_Identifier	Item_Identifier_Str2	Sales	nProduct	nSales	avgPrize
Supermarket Type3	OUT027	NC	617898	174	4567	135
Supermarket Type1	OUT035	NC	426131	168	2959	144
Supermarket Type1	OUT049	NC	415435	164	2944	141
Supermarket Type1	OUT013	NC	402420	180	2767	145
Supermarket Type1	OUT046	NC	394250	181	2780	142
Supermarket Type1	OUT017	NC	384667	172	2723	141
Supermarket Type1	OUT045	NC	367225	174	2676	137
Supermarket Type2	OUT018	NC	342185	173	2427	141
Grocery Store	OUT010	NC	42377	114	295	144
Grocery Store	OUT019	NC	33624	99	233	144

- Result of this table we can create new columns from our data and we can learn more about it like:
- nProduct : How many kind of Non-Consumable product saled in each shop.
- nSales : How many Non-Consumable product saled in each shop.
- nSales : What is the average Non-Consumable product prize in each shop.
- Also we do not have any descriptive graph like product type sales percentage in each outlet or transpose.

```r
g1 <- bigMart_Summary %>%
  group_by(Outlet_Identifier, Item_Identifier_Str2) %>%
  summarise(Sales = sum(Sales))%>%
  ggplot(aes(x = Outlet_Identifier, y = Sales, fill = Item_Identifier_Str2)) +
  geom_bar(stat = "identity", position = "fill", colors = "fill") +
  theme_classic() + ggtitle("Portion of Total Product Sales by Outlet") +
  labs(y="Proportion of Total Sales", x="Outlet Identifier") +
  theme(axis.text.x = element_text(angle = 30))

g2 <- bigMart_Summary %>%
  group_by(Outlet_Identifier, Item_Identifier_Str2) %>%
  summarise(Sales = sum(Sales))%>%
  ggplot(aes(x = Item_Identifier_Str2, y = Sales, fill = Outlet_Identifier)) +
  geom_bar(stat = "identity", position = "fill", colors = "fill") +
  theme_classic() + ggtitle("Portion of Total Sales in Each Outlet by Product") +
  labs(y="Proportion of Total Sales", x="Product Type")

grid.arrange(g1,g2,nrow=2)
```



- Results of this graps:
- Food is the most selled product in each shop and Drinks is the less selled.
- Before we see on the table for Non-Consumable product Outlet 27 is the most selled shop, It also has most sales for Food and Drinks.

In conclusion, we were need to learn more about our dataset. If we did, our analysis were be more meaningful for us.

Part - III

a - Below, you can find about my universities_34 dataset.

Description

universities_34 dataset has information about which university has how many students in 1987 to 2017 in 5 years periods in Istanbul. You can download data from (https://mef-bda503.github.io/pj-esera/files/universities_34.RData)

Details

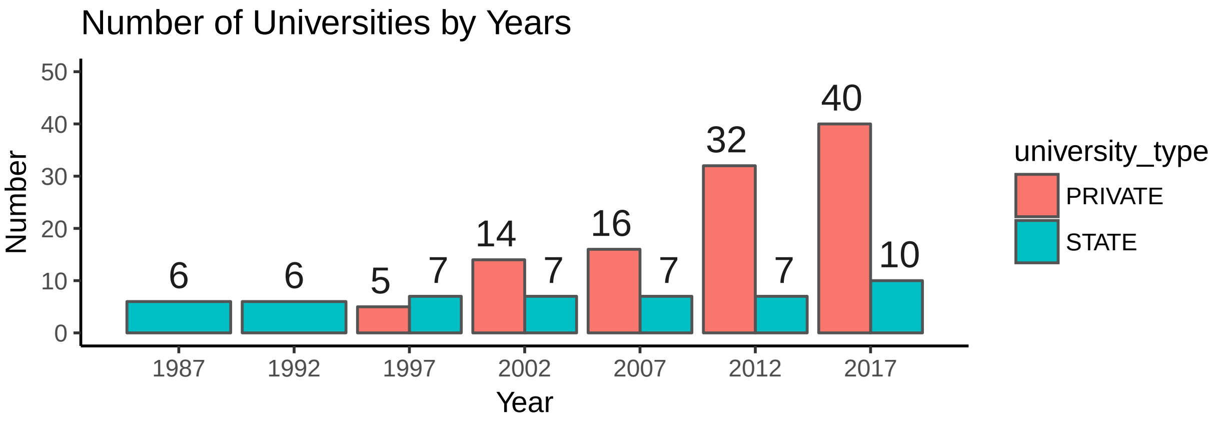
A data frame with 192 rows and 17 columns

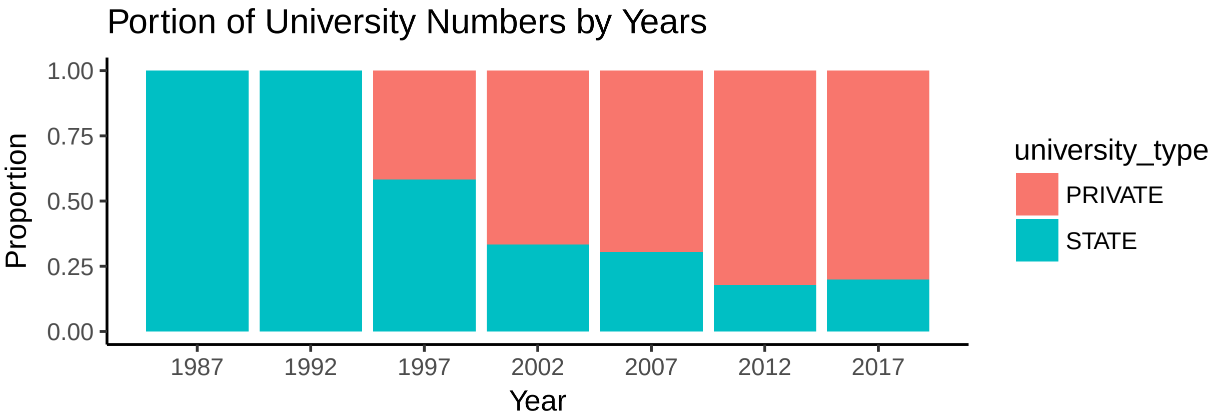
- year : Education period for 7 years (2017, 2012, 2007, 2002, 1997, 1992, 1987)
- university : University names in Istanbul
- university_type : State/Private
- education_kind : Daytime/Evening Education
- male_university : Number of undergraduat male students in university
- female_university : Number of undergraduat female students in university
- total_university : Number of undergraduat students in university
- male_master : Number of postgraduate male students in university
- female_master : Number of postgraduate female students in university
- total_master : Number of postgraduate students in university
- male_phd : Number of Ph.D. male students in university
- female_phd : Number of Ph.D. female students in university
- total_phd : Number of Ph.D. students in university
- male_total : Total number of male students in university
- female_total : Total number of female students in university
- total : Total number of students in university

b - Project Aim : To see number of state and private universities, number of students in state and private university, percentage of female university students evalutions by years in Istanbul.

```
setwd("/Users/yetkineser/Desktop/mef R/final")
download.file("https://mef-bda503.github.io/pj-esera/files/universities_34.RData"
              , "universities_34.RData")
universities_34 <- get(load("universities_34.RData"))

uni_34_summary <- universities_34 %>%
  group_by(year,university_type) %>%
  summarise(number_of_universities = n_distinct(university), number_of_students = sum(total)
            , number_of_male_students = sum(male_total), number_of_female_students = sum(female_total)
            )

g0 <- uni_34_summary %>%
  ggplot(aes(x = year, y = number_of_universities, fill = university_type)) +
  geom_bar(stat = "identity",position = position_dodge(),color=c("grey33")) +
  theme_classic() + ggtitle("Number of Universities by Years") +
  geom_text(aes(label=number_of_universities),color="grey11",size=5,vjust = -0.5
            , position = position_dodge(width = 5)) +
  scale_x_continuous(breaks = c(1987,1992,1997,2002,2007,2012,2017)) +
  scale_y_continuous(limits = c(0,50)) + labs(y="Number", x="Year")

g1 <- uni_34_summary %>%
  ggplot(aes(x = year, y = number_of_universities, fill = university_type)) +
  geom_bar(stat = "identity",position = "fill") + ylab("proportion") + theme_classic() +
  ggtitle("Portion of University Numbers by Years") +
  scale_x_continuous(breaks = c(1987,1992,1997,2002,2007,2012,2017)) +
  labs(y = "Proportion", x = "Year")
```

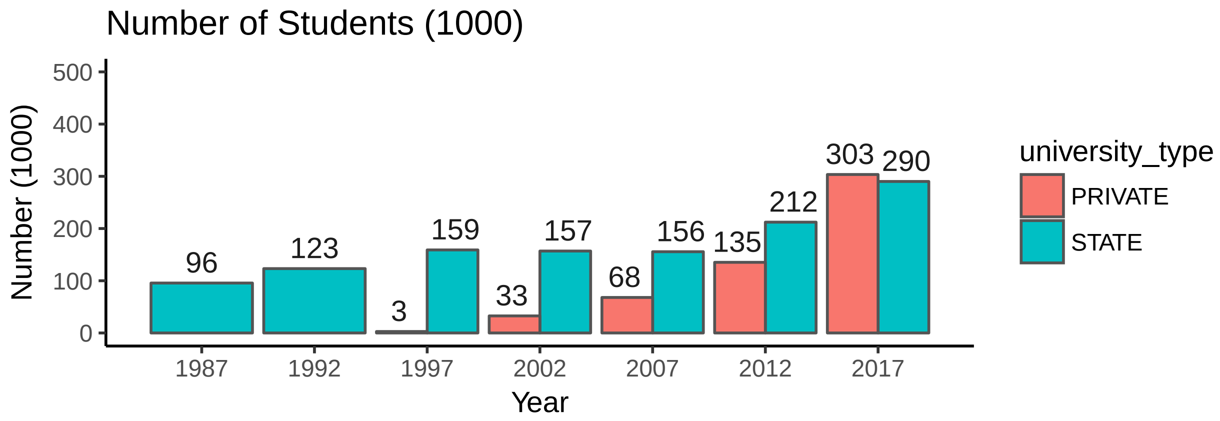
```
grid.arrange(g0,g1,nrow=2)
```

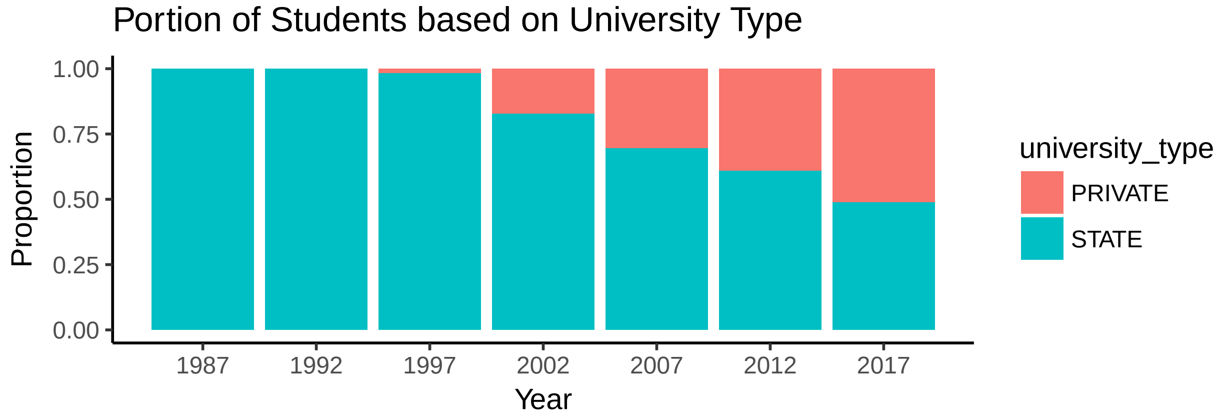




- In 1987 and 1992 there are just 6 state university, but 2017 there are 40 private university and 10 state university.
- Private university numbers increase incredibly especially 2007 to 2012(by numbers = 16) and 1997 to 2002(by percentage = 180%).

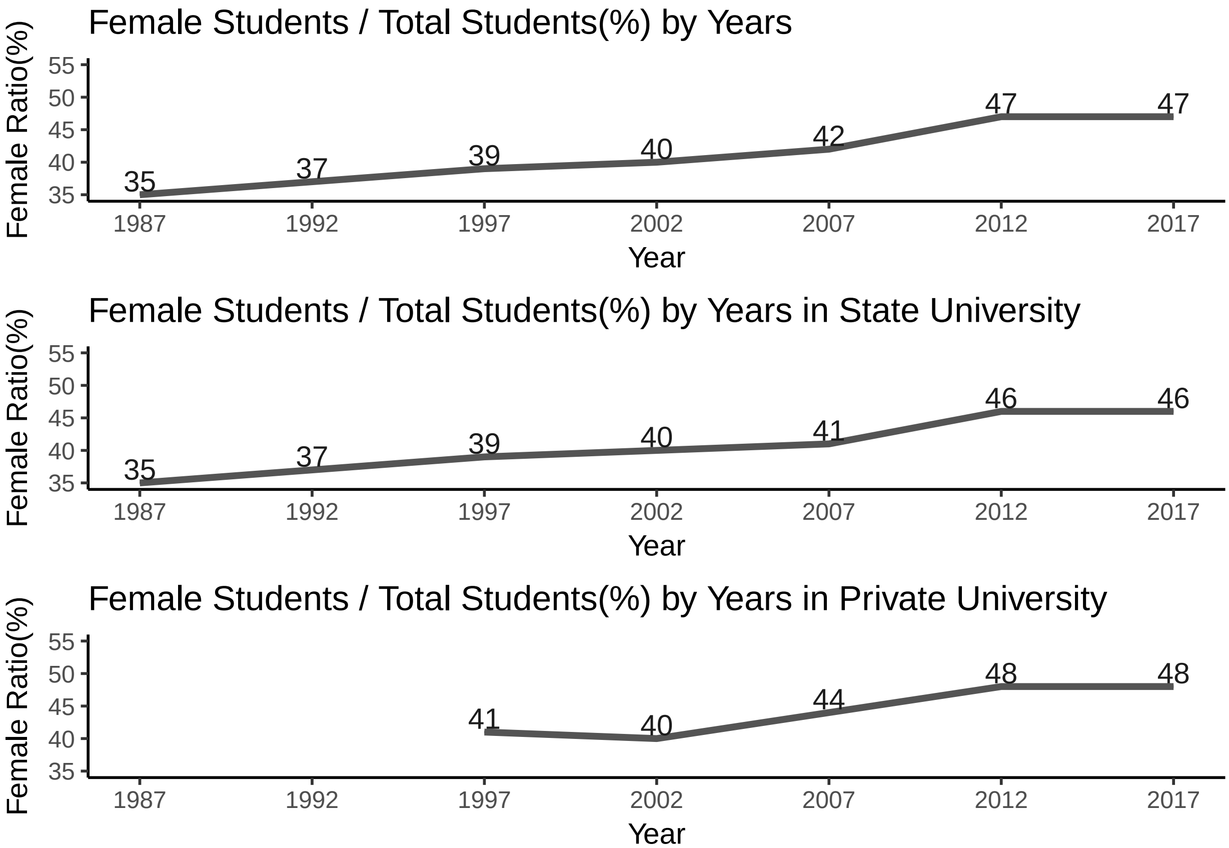
```
g0 <- uni_34_summary %>%
  ggplot(aes(x = year, y = number_of_students/1000, fill = university_type)) +
  geom_bar(stat = "identity",position = position_dodge(),color=c("grey33")) +
  theme_classic() + ggtitle("Number of Students (1000)") +
  geom_text(aes(label=round(number_of_students/1000))
          ,color="grey11",size = 4, vjust = -0.5, position = position_dodge(width = 5)) +
  scale_x_continuous(breaks = c(1987, 1992, 1997, 2002, 2007, 2012, 2017)) +
  scale_y_continuous(limits = c(0,500)) + labs(y="Number (1000)", x = "Year")

g1 <- uni_34_summary %>%
  ggplot(aes(x = year, y = number_of_students, fill = university_type)) +
  geom_bar(stat = "identity",position = "fill") + ylab("proportion") + theme_classic() +
  ggtitle("Portion of Students based on University Type") +
  scale_x_continuous(breaks = c(1987, 1992, 1997, 2002,2007,2012,2017)) + labs(y="Proportion", x = "Yea

grid.arrange(g0 ,g1 ,nrow = 2)
```





- In 1987 and 1992 there is not a private university students(because there is no private university), but 2017 there are more private university students. Especially 2012 to 2017 there is huge increase in number of university students in Istanbul especially in private university(almost 170000).

```
uni_34_summary <- universities_34 %>%
  group_by(year) %>% summarise(number_of_students = sum(total),
            number_of_male_students = sum(male_total),
            number_of_female_students = sum(female_total))

g0 <- uni_34_summary %>%
  mutate(female_ratio = round((number_of_female_students/number_of_students) * 100), 4) %>%
  ggplot(aes(x = year, y = female_ratio) ) + geom_line() + theme_classic() +
  ggtitle("Female Students / Total Students(%) by Years") +
  geom_line(size = 1.2,color = c("grey33")) + scale_y_continuous(limits = c(35,55)) +
  geom_text(aes(label = female_ratio), color = "grey11", size = 4, vjust = 0, nudge_y = 0.5) +
  scale_x_continuous(breaks = c(1987,1992,1997,2002,2007,2012,2017)) +
  labs(y="Female Ratio(%)", x = "Year")

uni_34_summary <- universities_34 %>%
  group_by(year,university_type) %>% summarise(number_of_students = sum(total),
            number_of_male_students = sum(male_total),
            number_of_female_students = sum(female_total))

g1 <- uni_34_summary %>%
  mutate(female_ratio = round((number_of_female_students/number_of_students)*100),4) %>%
  filter(university_type == "STATE") %>%
  ggplot(aes(x = year, y = female_ratio)) + geom_line() + theme_classic() +
  ggtitle("Female Students / Total Students(%) by Years in State University") +
  geom_line(size=1.2,color=c("grey33")) +
  geom_text(aes(label = female_ratio), color = "grey11", size = 4, vjust = 0, nudge_y = 0.5) +
  scale_y_continuous(limits = c(35,55)) + labs(y = "Female Ratio(%)", x = "Year") +
  scale_x_continuous(breaks = c(1987,1992,1997,2002,2007,2012,2017))

g2 <- uni_34_summary %>%
  mutate(female_ratio = round((number_of_female_students/number_of_students)*100),4) %>%
  filter(university_type == "PRIVATE") %>%
  ggplot(aes(x = year, y = female_ratio)) +geom_line() +theme_classic() +
  ggtitle("Female Students / Total Students(%) by Years in Private University") +
  geom_line(size = 1.2, color = c("grey33")) +
  geom_text(aes(label = female_ratio),color = "grey11", size = 4,vjust = 0, nudge_y = 0.5) +
  scale_y_continuous(limits = c(35,55)) + labs(y = "Female Ratio(%)", x = "Year") +
  scale_x_continuous(breaks = c(1987,1992,1997,2002,2007,2012,2017),limits = c(1987,2017))

grid.arrange(g0,g1,g2,ncol=1)
```



- In first graph in 1987 female university students is just 35% of students, but it increase year by year and in 2017 it will 47%. especially 2007 to 2012 its increase is 5%.
- In second and third graph shows that there is no big difference percentage of female students in state and private universities(just 2% in 2017).