Final_Sefa_Erbas

Sefa Erbas

8 Ocak 2018

PART 1

1.

Graphs and charts are used to provide a visual demonstration of the data. They summarizes the big data which the subject we want to make an inference. That's why they have to be clear and tells their story in an intelligible way. At that point, using dual y-axis charts in our visualisation may be confusing for the people. Our brains are conditioned to look for the Y-axis on the left of a chart. The variables where on the right could be ignored by users. However, if we put complementary or exactly matched variables on Y-axis, it would be more clear and sensible. For example profit and sales amounts, or temperature values (Celsius & Fahrenheit)

2.

I can describe my EDA workflow as: - Define the goal without any doubt and figure out what kind of variables do I need - Understanding the data (General view of the data using basic R functions as head, glimpse, describe and corr) - Data manipulation & Cleaning if necessary - Plotting to understand general concept of data. For the spesific project, first of all I will specify some categories (total budgets, number of project in related subjects) for potential projects to order them for priority. In my point of view, the main goal will be providing funds to as far as much more projects. I will get spesific data for related subjects with featured variables that helps to measure the impacts. Of course i will present what my data says, however, if making exception will be better, i will also take into consideration.

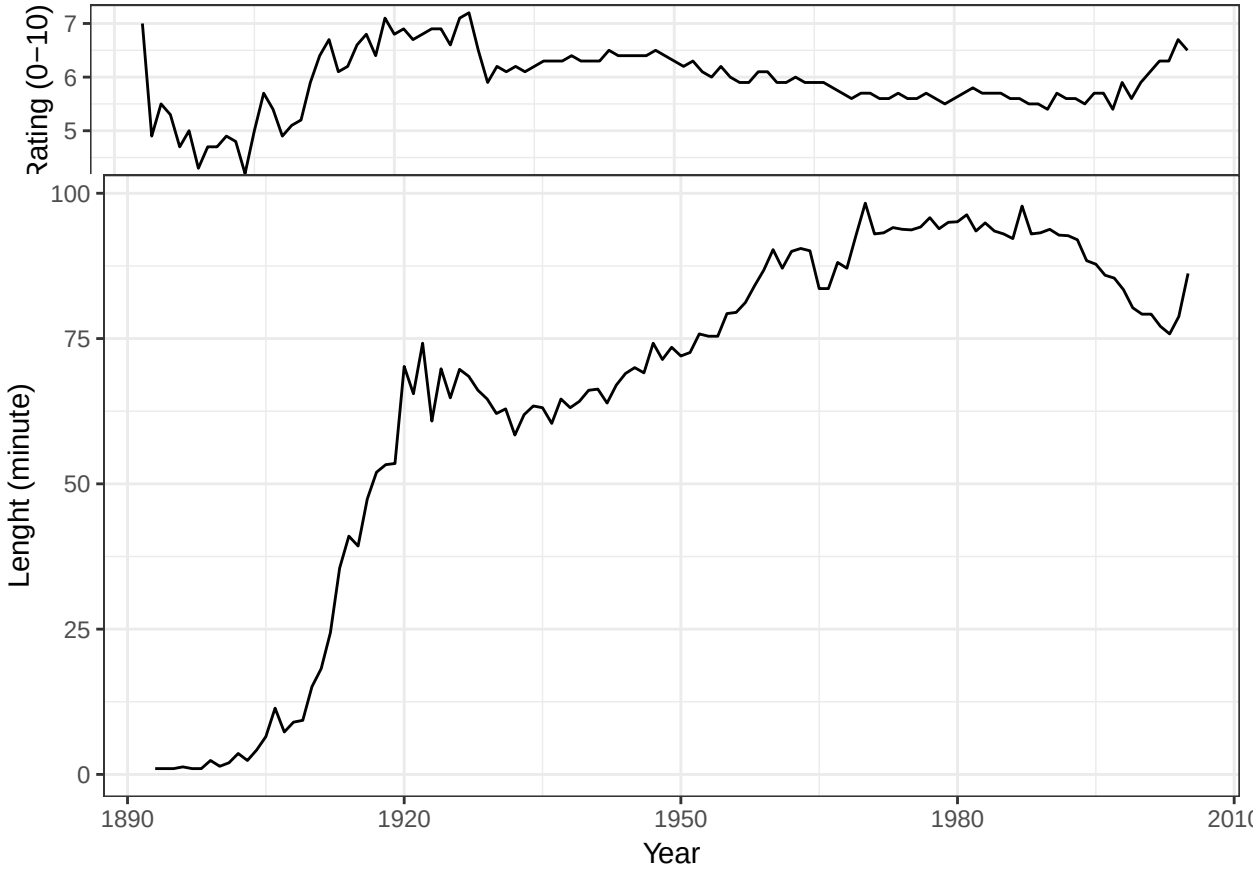
3.

Time series data consists of values that represent or trace the values taken by a variable over a period such as a month, quarter, or year. Time series data occurs wherever the same measurements are recorded on a regular basis. On the other hand, a non time series data may consists of more than one variable with time independent. For example let's we talk about the diamond price. Time series data is useful when we try to make forecasting model for future (price volatility of a diamond with default parameters/ features) , on the other hand in non time series data can be used to predict the variant diamond price according to related variables which are the parameters/ features of the diamond.

4.

```
movies3 <- group_by(movies, year)%>%
summarise(avg_length = mean(length), avg_rating=mean(rating))
movies3 <- round(movies3,1)

## Create the rating plot
g.top <- ggplot(movies3, aes(x = year, y = avg_rating)) +
  theme_bw() +
  theme(plot.margin = unit(c(1,5,-30,6),units="points"),
```

```
        axis.title.y = element_text(vjust =0.25)) +
  geom_line() +
  labs(y = "Rating (0-10)")

## Create the length plot
g.bottom <- ggplot(movies3, aes(x = year, y = avg_length)) +
  theme_bw() +
  theme(plot.margin = unit(c(0,5,1,1),units="points")) +
  geom_line() +
  labs(x = "Year", y = "Lenght (minute)")

## Plot graphs and set relative heights
grid.arrange(g.top,g.bottom, heights = c(1/5, 4/5))
```



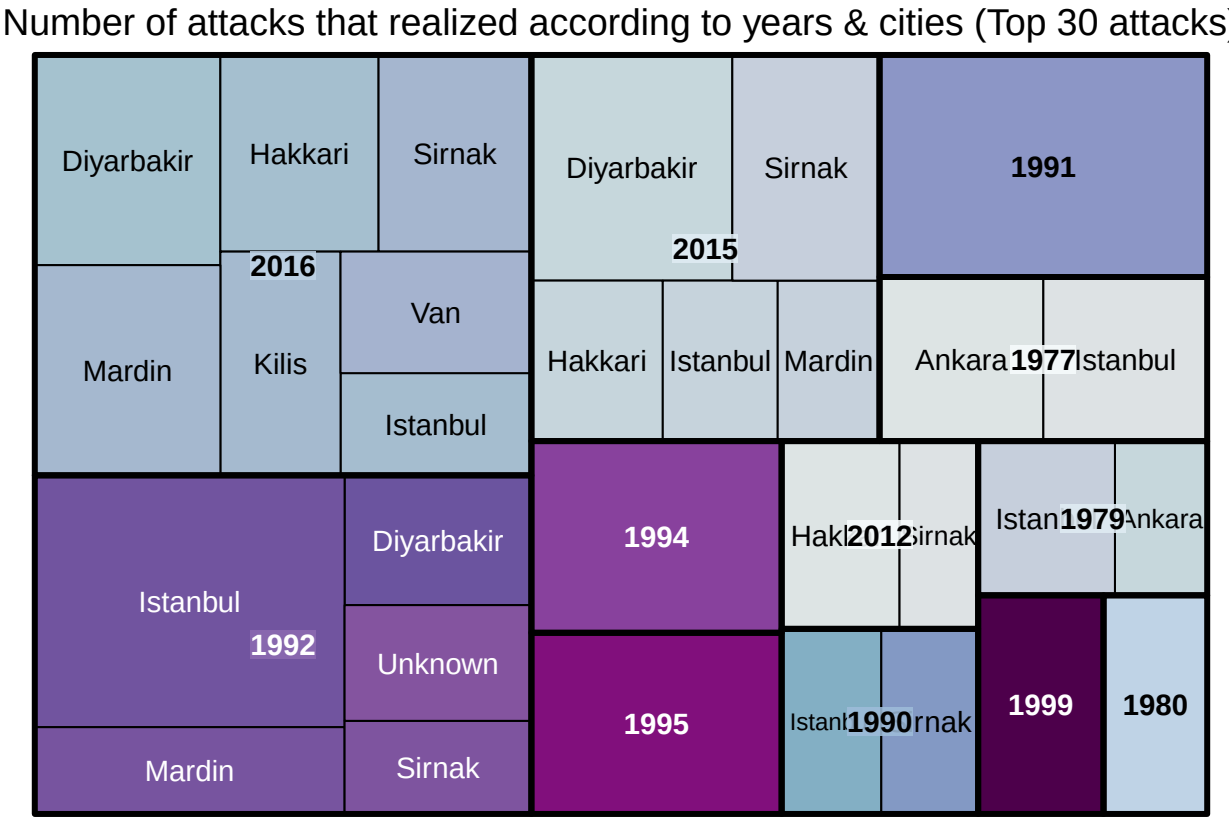
PART 2

I would like add a chart which shows the "top 30 worst attacks" according to provstate & year with treemap design.

I would like add a chart which shows the "top 30 worst attacks" according to provstate & year with treemap design.

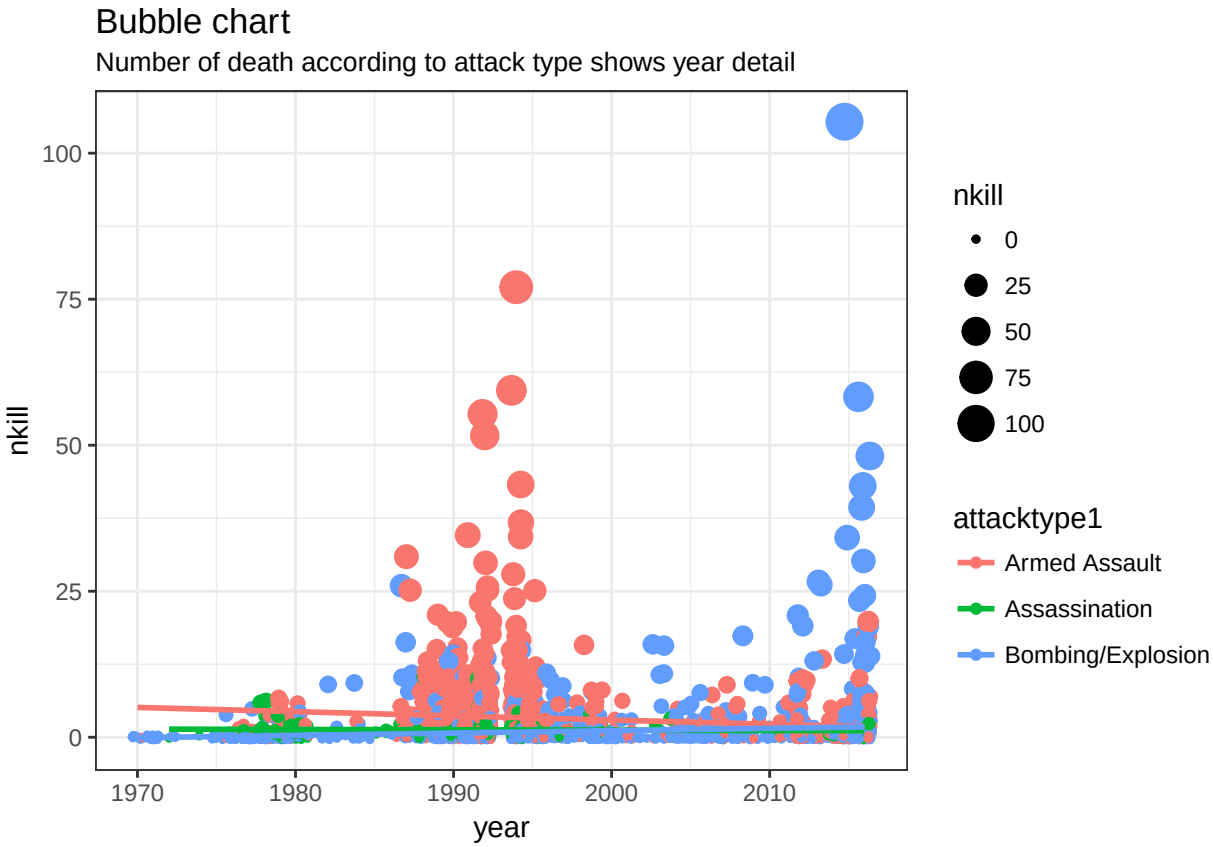
```
gtd_last<- gtd.turkey%>%
  group_by(year, provstate)%>%
  summarise(number_of_attacks=n())
gtd_last<- arrange(gtd_last, desc(number_of_attacks))
```

```
gtd_last=head(gtd_last,30)
#Treemap Plot
treemap(gtd_last,
        index=c("year","provstate"),
        vSize = "number_of_attacks",
        type="index",
        palette = "BuPu",
        title="Number of attacks that realized according to years & cities (Top 30 attacks)",
        fontsize.title = 14)
```



This bubble chart shows the top 3 attack types that cause death per years.

```
# load package and data
library(ggplot2)
data(gtd.turkey, package="ggplot2")
gtd.turkey_select <- gtd.turkey[gtd.turkey$attacktype1
                                %in% c("Bombing/Explosion", "Armed Assault", "Assassination"), ]
# Scatterplot
theme_set(theme_bw())  # pre-set the bw theme.
g <- ggplot(gtd.turkey_select, aes(year, nkill)) +
  labs(subtitle="Number of death according to attack type shows year detail",
       title="Bubble chart")
g + geom_jitter(aes(col=attacktype1, size=nkill)) +
  geom_smooth(aes(col=attacktype1), method="lm", se=F)
```



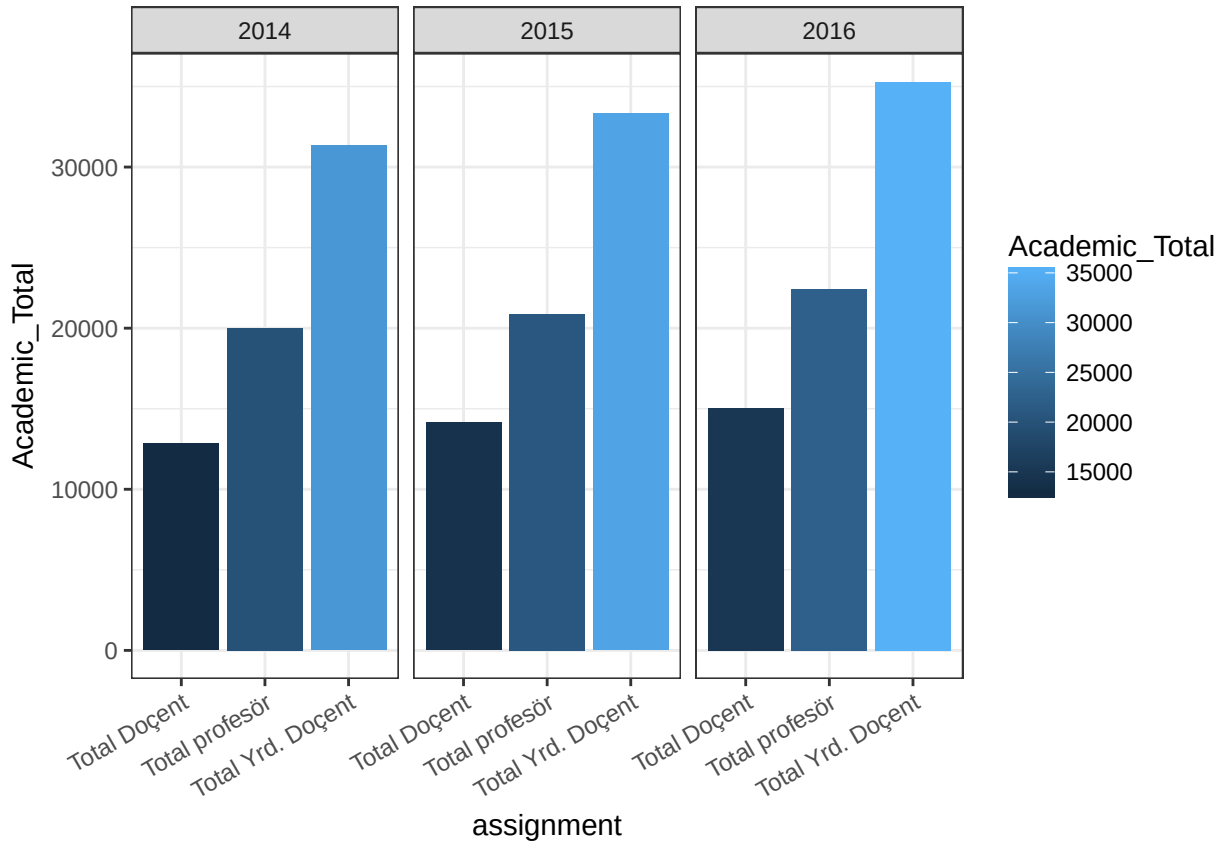
PART 3

Getting the data from URL & manipulation

```
download.file("https://raw.githubusercontent.com/MEF-BDA503/pj-erbass/master/Academician_years.RData",
              "Academician_years.RData")
load("Academician_years.RData")

#Melt the data with the academic position variables
data13 <- gather(data10, assignment, value, -Number, -University_name)

#Change the numbers with related year which has already referred.
data13$Number[data13$Number < 192] <- 2016
data13$Number[data13$Number >= 192 & data13$Number <= 373] <- 2015
data13$Number[data13$Number > 373 & data13$Number <= 554] <- 2014
#Delete the year variables which is unnecesary.
data13<-data13[!(data13$assignment=="Year"),]
```

```
data20<-data13 %>%
  group_by(Number, assignment) %>%
  summarise(Academic_Total= sum(value, na.rm=T))%>%
  filter (assignment =="Total profesör" | assignment =="Total Doçent"
          | assignment =="Total Yrd. Doçent")
# plot2
plot2<-ggplot(data20, aes(y=Academic_Total, x=assignment, fill=Academic_Total)) +
```

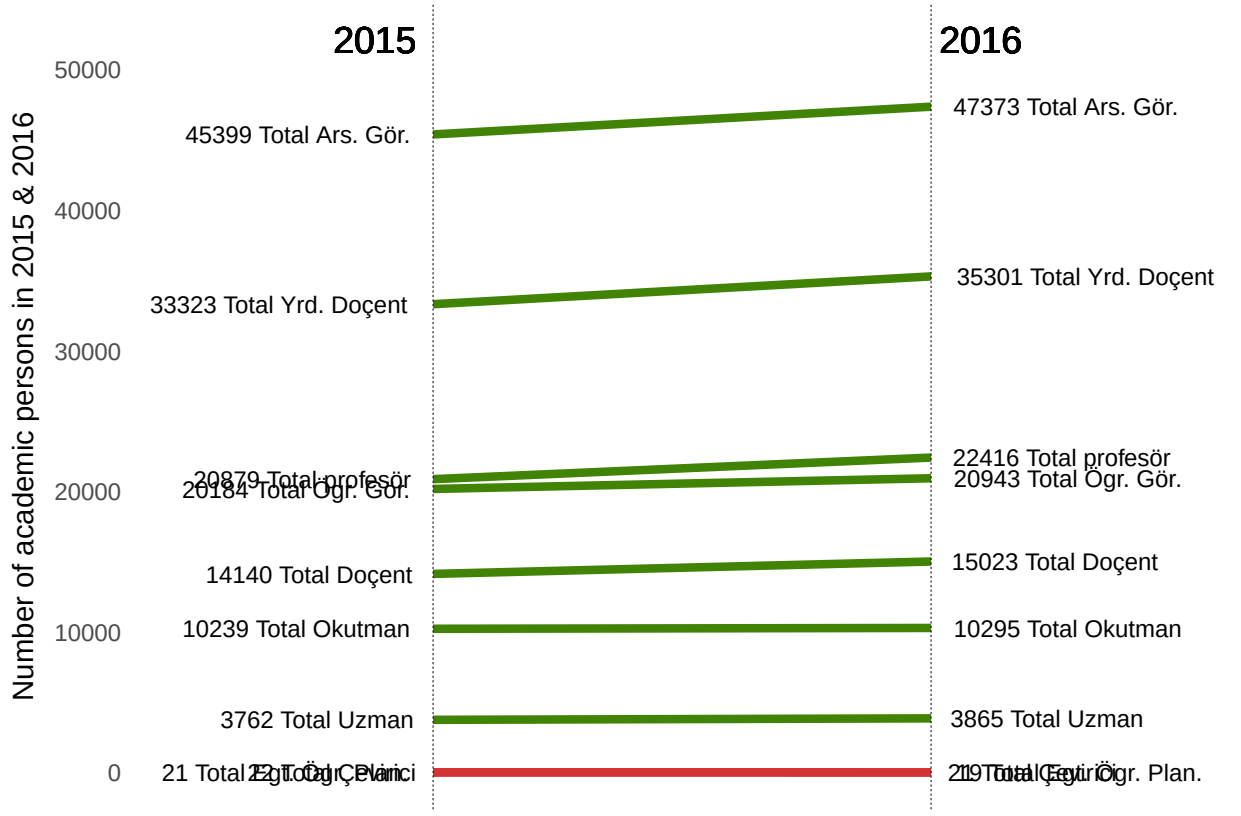
```
  geom_bar( stat="identity") +
  theme (axis.text.x=element_text (angle=30, vjust=1, hjust=1)) +
  facet_wrap(~Number)
plot2
```



I tried to figure out the number of academic people (professors, associate professor, assistant professor) by years. It is seen that, the number of all academic people have been increased.

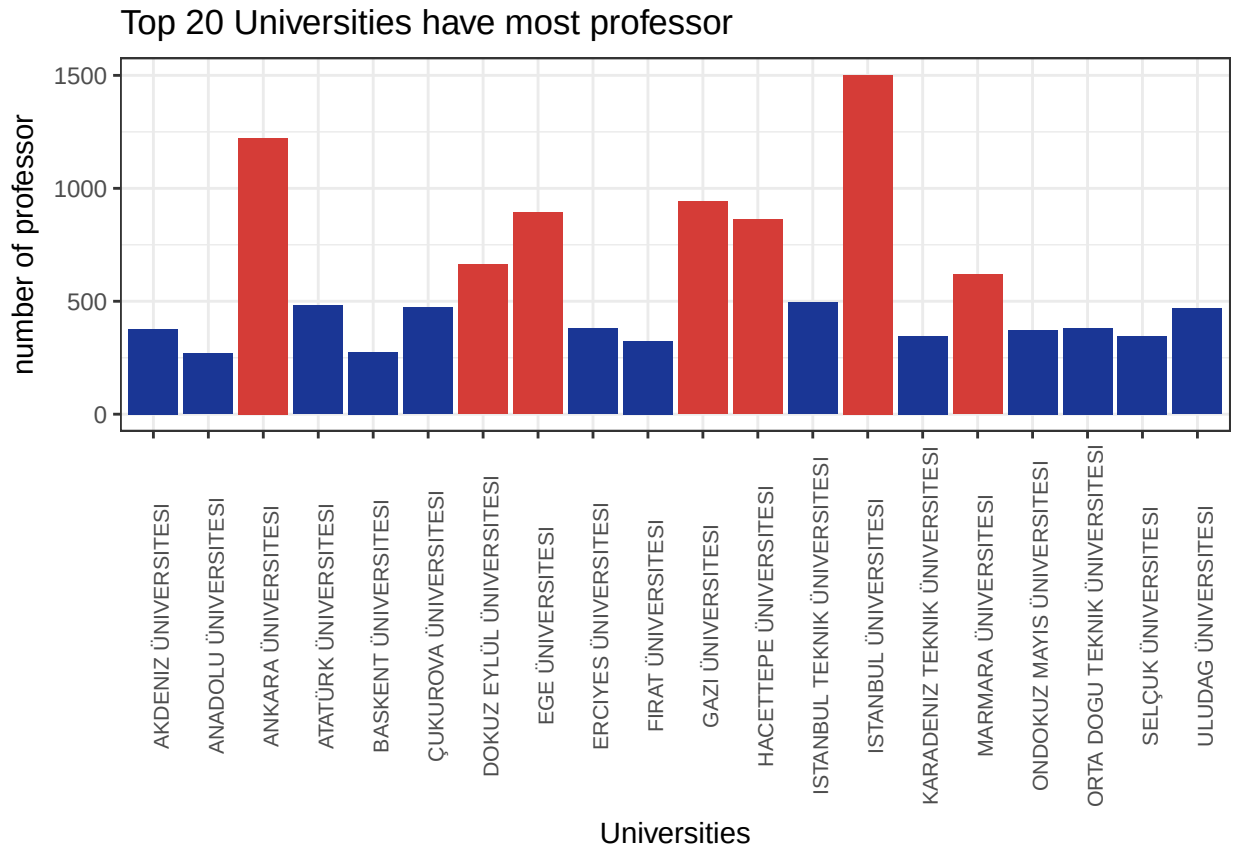
```
#Data Manipulation to show the changes on last two year
target <- c("2015", "2016")
data4<- data13%>%
  filter(Number %in% target) %>%
  group_by(assignment, Number)%>%
  summarise(total=sum(value))

#Reshape data to adapt the graph
df1<-cast(data4, assignment ~ Number)

#Plot Show
# Preparing Data
left_label <- paste(df1$`2015`, (df1$assignment))
right_label <- paste(df1$`2016`, (df1$assignment))
df1$class <- ifelse((df1$`2016` - df1$`2015`) < 0, "red", "green")

# Plot
plot <- ggplot(df1) + geom_segment(aes(x=1, xend=2, y=`2015`,
                                       yend=`2016`, col=class), size=1.5, show.legend=F) +
```

```
  geom_vline(xintercept=1, linetype="dashed", size=.1) +
  geom_vline(xintercept=2, linetype="dashed", size=.1) +
  scale_color_manual(labels = c("Up", "Down"),
                     values = c("red"="#d53333", "green"="#408305")) +  # color of lines
  labs(x="", y="Number of academic persons in 2015 & 2016") +  # Axis labels
  xlim(.5, 2.5) + ylim(0,(1.1*(max(df1$`2015`, df1$`2016`))))  # X and Y axis limits

# Adding texts
plot <- plot + geom_text(label=left_label, y=df1$`2015`,
                         x=rep(1, NROW(df1)), hjust=1.1, size=3)
plot <- plot + geom_text(label=right_label, y=df1$`2016`,
                         x=rep(2, NROW(df1)), hjust=-0.1, size=3)
plot <- plot + geom_text(label="2015", x=1, y=1.1*(max(df1$`2015`, df1$`2016`)),
                         hjust=1.2, size=5)  # title
plot <- plot + geom_text(label="2016", x=2, y=1.1*(max(df1$`2015`, df1$`2016`)),
                         hjust=-0.1, size=5)  # title
# Minify theme
plot + theme(panel.background = element_blank(),
             panel.grid = element_blank(),
             axis.ticks = element_blank(),
             axis.text.x = element_blank(),
             panel.border = element_blank(),
             plot.margin = unit(c(0.1,0.2,0.1,0.2), "cm"))
```



As it is seen that, number of people in different academic positions has been increased except 2(Translator & Education Planning Responsible) from 2015 to 2016.

```r
#Find the top 25 universities that have most professor
data5<- data13%>%
    group_by(University_name, assignment)%>%
  summarise(avg_person=mean(value))
data5$avg_person<- round(data5$avg_person,0)
data6<- data5[ which(data5$assignment=='Total profesör'), ]
data6<- arrange(data6, desc(avg_person))
data6<- head(data6,20)
data6 %>%
ggplot(aes(x=University_name,y=avg_person))+
      geom_bar(stat = "identity",aes(fill=avg_person>500)) +
      theme(axis.text.x = element_text(angle=90,size=8,vjust=1, hjust=0.5)) +
      labs(x= 'Universities', y= 'number of professor') +
      ggtitle('Top 20 Universities have most professor')+
      scale_fill_manual(values = c('#1A3695', '#D53C37'),guide=FALSE)
```



I tried to find out which universities have more professors and it has been also highlighted where the university have more than 500 professors.