Final - Semih Tekten

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.

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

---

- First of all, thank you for referring R_Junkies' two y-axis graph in the final paper. In my opinion two y-axis graphs (2YAG) are neither good, nor evil. People can & should use it, whenever they **need** it.

For some reason in the history of data-analyis, 2YAG appeared. People use it when it's beneficial to show different metrics in a single graph, in order to compare metrics or understand trends across time. We use them in our company, too. Managers and clients want to see such graphs, because we don't have that much space in reports.

The problem seems two-sided: First, axis-ranges can be arranged to manipulate the interpretation. But the same criticism may be easily said to single y-axis graphs. Even figure size manipulates the line. In that sense, there is no difference between single & two y-axis graphs: One can manipulate both of them easily, in order to manipulate interpretation.

Second, it is hard to understand which line belongs to which axis. As you can see in our R_Junkies graph, we used *different* types of graphs for a 2YAG and mentioned them in the title and axis-labels. In that case, I don't see any problem while interpreting the graph. One must clearly mention sides, axes, colors, units etc.

Another problem Wickham tells is that dual axes are arbitrary. It seems to me they are not arbitrary, it is just easy to show & read two-axis on a 2D screen/paper. One axis goes to right, the other one goes to left. If we add 3rd axis, then we need depth which is not very easy to draw & interpret. (Or use bubbles, which is also hard to interpret.) By the way, it can be easily claimed that everything data scientist chooses to do is arbitrary :)

---

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"?

- Most of the people probably start with data, but I prefer to start with story, agenda & problem. What is the story that lead to this research & problem? Given with a story and a problem, I can prioritize which metrics, features are important and relevant to my EDA study. I can find the relevant questions and eliminate the irrelevant ones. Then I can decide my workflow, which is usually (As we discussed in class and experienced many times in business) follows like this: Story & Problem > Data Collection > Data Preprocessing & Cleaning > Insights > Analysis & Visualizations > Results & Storytelling (New story & New problems)

In the case of welfare projects, I would preset global goals & outcomes for all projects. Each project must decide their goals & outcomes beforehand. And after they spend their budget, then I will be able to compare outcomes for the same goals. Let's say Project A with the Goal of X brought 25% more outcomes compared to Project B with the Goal of X. In that way, we may compare apples with apples (Both projects have the Goal of X). If both projects have the Goal of X, then it would be interesting to observe one of them generated 25% more outcomes. As it appears, assigning each project a predefined goal is very important and necessary. Otherwise we can't compare them.

In the case of policy reporting, I would probably show what data said. But here is another aspect: Most of the time, researches lack some valuable & important features which may affect the research significantly. One should also consider that everything in society are interdependent. So there is always place for some subjectivity in such researches, even we reject/deny them.

---

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?

- In my humble opinion, the major difference between time-series (TS) and non time-series (NTS) data is that with the former you have to deal with the great problem of "Date Type" in your syntax. I have no prior experience with any programming language or software that is joyful to deal with dates. A shame on whole developers and data scientists in the world!

The major issue with TS data is that new values are probably affected by previous values. People call it autocorrelation problem. NTS doesn't have to deal with such a problem. (Or minimal affect)

Another hard problem for TS is that independent variables themselves may also change. Here is an example: For the first months and years of Bitcoin, we may explain the price fluctuations with 5 different features. But today, there are lots of other features that affect Bitcoin price. NTS is like a snapshot of TS data, TS is more complex since it contains the axis of Time.

---

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.)

- I've always wondered whether budget size affects ratings on IMDB. Now I have a great chance to find out the answer.
- My hypothesis: **Budget size has a significant impact on Rating.**

```
library(ggplot2movies)
suppressMessages(library(tidyverse))
```

- Let's choose only movies with budget.

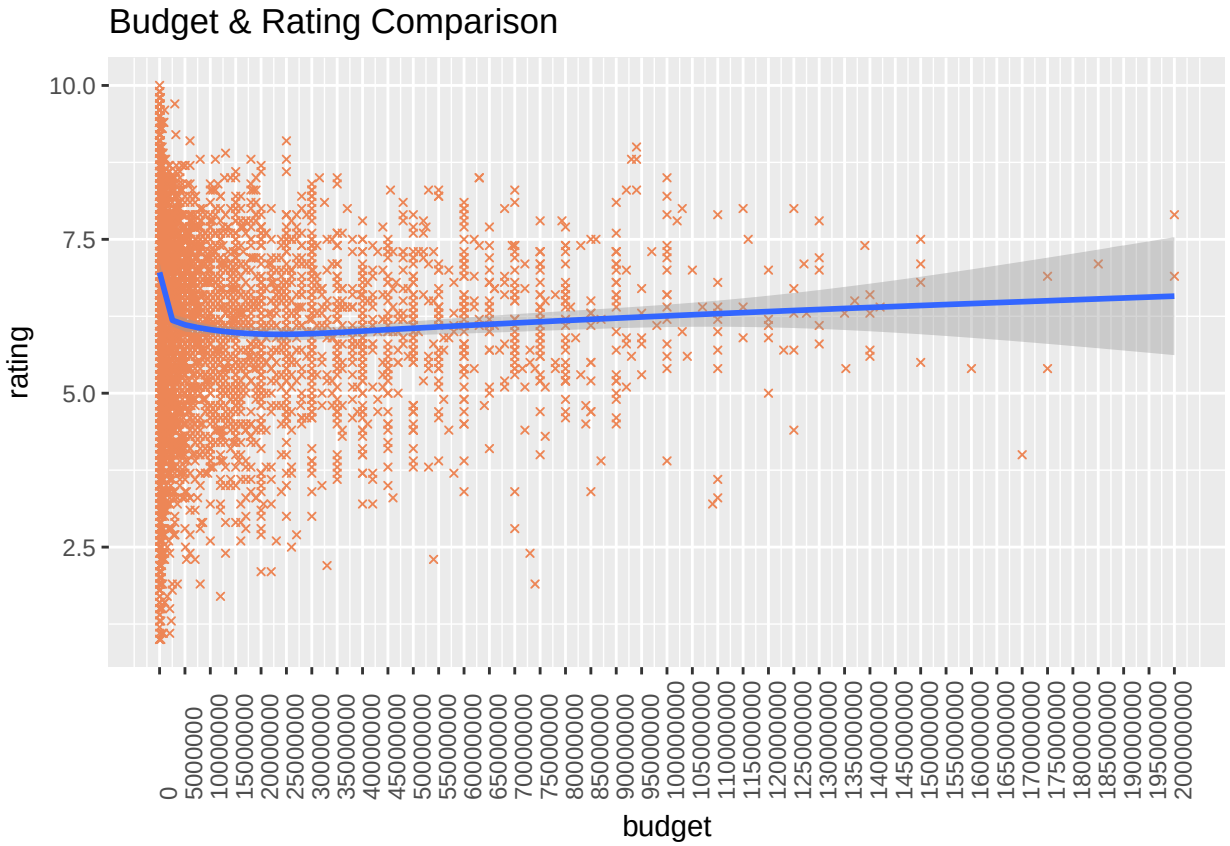
```
budgeted_movies = movies %>%
                    filter(budget > 0) %>%
                    arrange(budget)

# Disable engineering mode. Otherwise we'll see the numbers as: 1.00e+03
options(scipen=999)
```

- Cool. Now prepare a scatter-plot. X-Axis should have the Budget (in USD) and Y-Axis should have the Rating.

```
ggplot(budgeted_movies, aes(x=budget, y=rating)) +
  geom_point(size=1, shape=4, colour = "#ed8656") +
  geom_smooth(method = "auto") +
  ggtitle("Budget & Rating Comparison") +
  scale_x_continuous(breaks = seq(0,200000000,5000000)) +
  theme(axis.text.x=element_text(angle = 90, hjust = 0))
```

```
## `geom_smooth()` using method = 'gam'
```



- Regression line shows an interesting relation. For small budgets (1-5M), Rating decreases while Budget increases. And after 5M Budget, this relation reverses and Rating increases smoothly, while Budget increases. Another observation to note ise that our confidence interval increases after 100M USD. This means, if you produce a movie for more than 100M, then your chance to get higher Rating increases. :)

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.

- One of the feedbacks we had is that the line graphs "**Fatalities of Top 6 Operators with Most Crashes/Fatalities Over Years**" are not so easy to interpret. For that reason, I decided to convert them to a new heat map and show fatalities of top 10 civil operators with most fatalities over decades. I will generate a new data frame from our clean data using **dplyr**, and visualise it with **ggplot2**.

Let's download our clean data from R_Junkies Progress Journal. (*I hid the code from the output.*)

Now generate a new data frame, '**hm_df**' from **apc_clean**. **hm_df** will contain *Operator*, *YearRange*, *Fatalities*. And I need complete cases for visualization.

```
hm_df <- apc_clean %>%
          select(Operator, Year, Fatalities, IsMilitary) %>%
          filter(IsMilitary == 0) %>%
```

```
          select(Operator, Year, Fatalities) %>%
          filter(Operator != "")

hm_df <- hm_df[complete.cases(hm_df), ]
```

Now we need to generate a new column, '**YearRange**' in order to find decades. I will use **group_by**.

```
hm_df <- hm_df %>%
    mutate(Year = as.numeric(as.character(Year))) %>%
    mutate(YearRange = paste( as.character( Year - (Year %% 10)),
                              as.character(Year - (Year %% 10) + 9),
                              sep = " - ")) %>%
    select(-Year) %>%
    group_by(Operator,YearRange) %>%
    summarise(Fatalities = sum(Fatalities)) %>%
    filter(Fatalities > 0) %>%
    arrange(Operator, YearRange)
```

I want to show only top 10 operators with most fatalities. So I'm going to find them. I'll name it as **civ_op_topten_fat**.

```
civ_op_topten_fat <- apc_clean %>%
                     filter(IsMilitary == 0) %>%
                     select(Operator, Fatalities) %>%
                     group_by(Operator) %>%
                     summarise(Fatalities = sum(Fatalities)) %>%
                     arrange(desc(Fatalities)) %>%
                     slice(1:10) %>%
                     select(Operator)
```

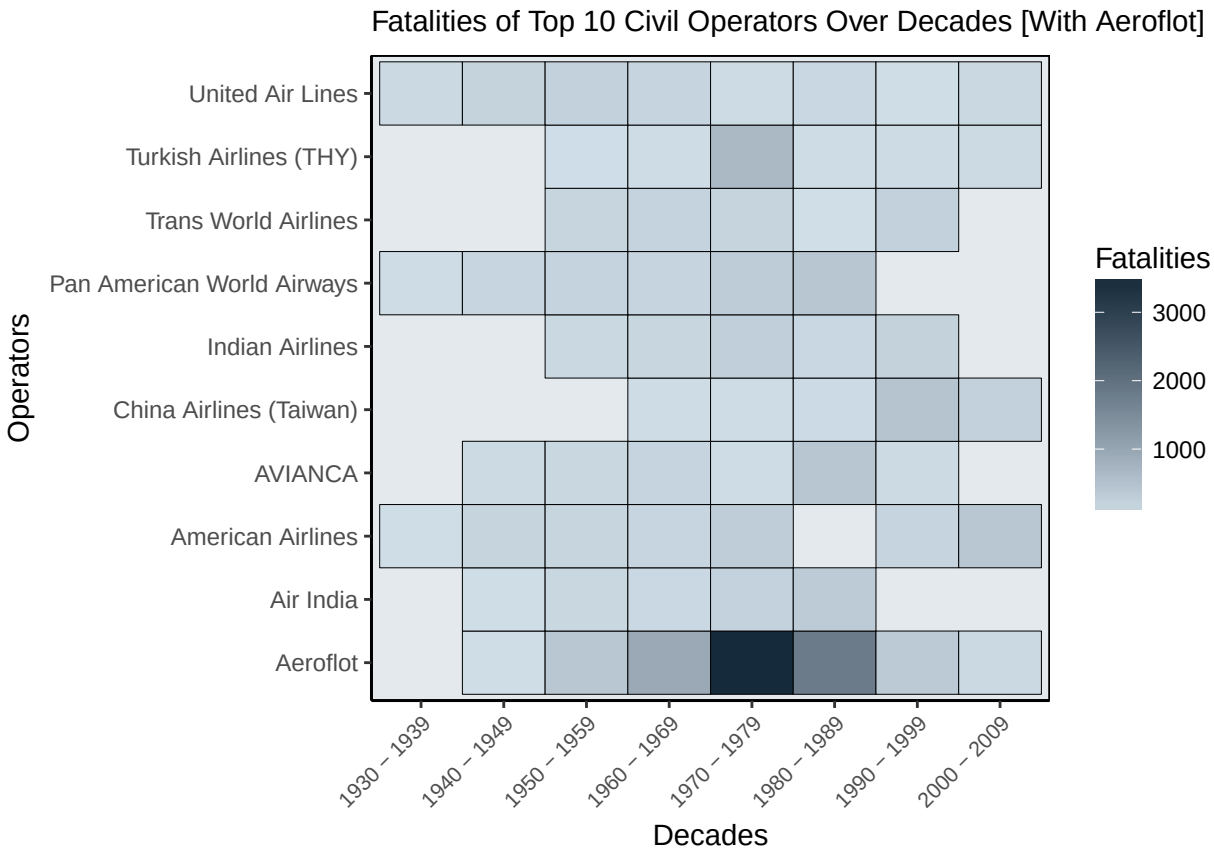
Beautiful. Let's join **hm_df** with **civ_op_topten_fat**!

```
hm_df <- hm_df %>% inner_join(., civ_op_topten_fat, by='Operator')
head(hm_df)
```

```
## # A tibble: 6 x 3
## # Groups:   Operator [1]
##   Operator YearRange   Fatalities
##   <chr>    <chr>            <int>
## 1 Aeroflot 1940 - 1949         24
## 2 Aeroflot 1950 - 1959        414
## 3 Aeroflot 1960 - 1969        951
## 4 Aeroflot 1970 - 1979       3512
## 5 Aeroflot 1980 - 1989       1814
## 6 Aeroflot 1990 - 1999        353
```

Since our data frame is ready for visualization, I can start using **ggplot2**.

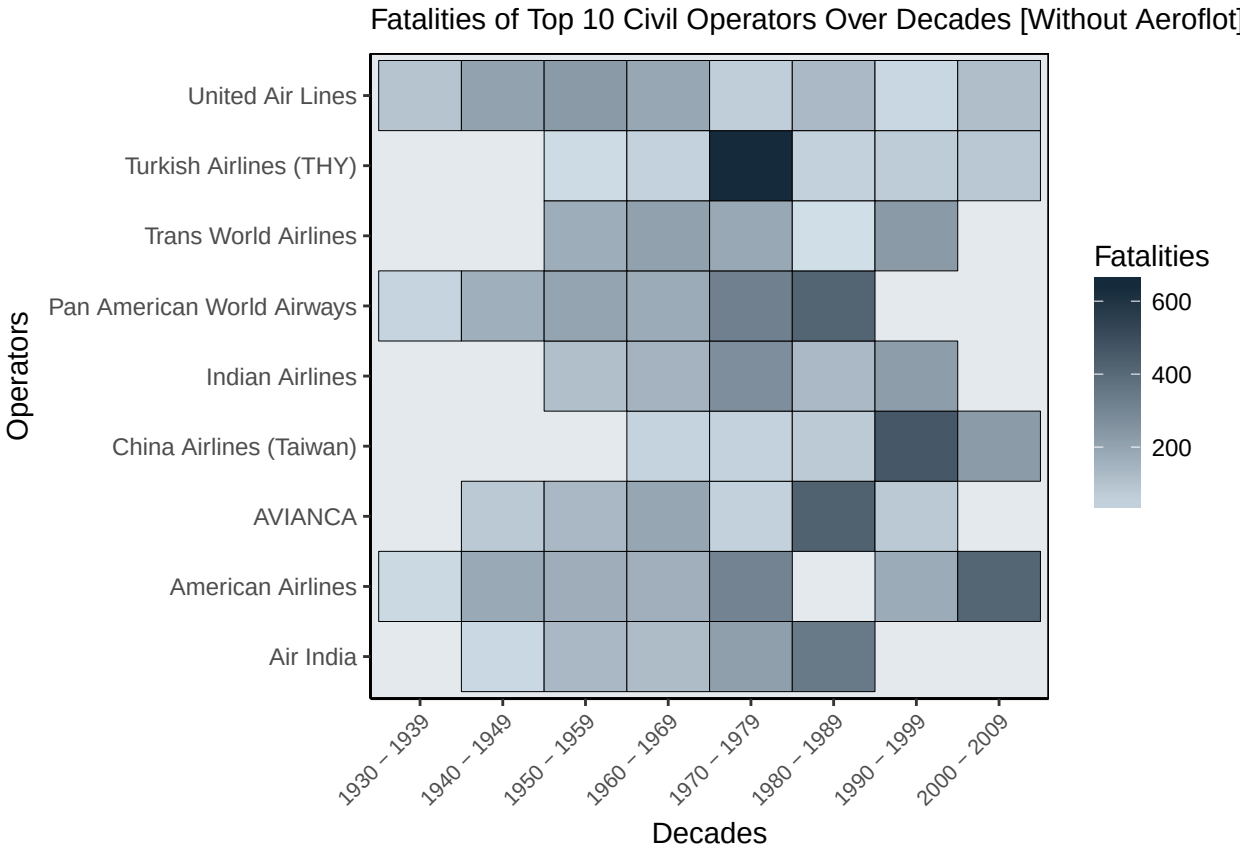
```
ggplot(hm_df, aes(YearRange, Operator)) +
  theme_classic() +
  geom_tile(aes(fill = Fatalities), colour = "black") +
  scale_fill_gradient(low = "#d0dee8", high = "#152a3a") +
  theme(panel.background = element_rect(fill = "#e3e9ed", colour = "black")) +
  labs(title='Fatalities of Top 10 Civil Operators Over Decades [With Aeroflot]', x='Decades', y='Opera
  theme(plot.title = element_text(hjust = 0.0, size=11)) +
  theme(axis.text.x=element_text(angle = 45, hjust = 1, size = 8))
```



Not so good! The variation between fatalities can't be observed. When I searched for a reason, I figured out that there is an outlier: Aeroflot! In 70s they got more than 3K fatalities! It is very sad story for Aeroflot. Anyways, I decided to remove Aeroflot from my visualization and generated a new heatmap.

```
hm_df <- hm_df %>% filter(Operator != 'Aeroflot')

ggplot(hm_df, aes(YearRange, Operator)) +
  theme_classic() +
  geom_tile(aes(fill = Fatalities), colour = "black") +
  scale_fill_gradient(low = "#d0dee8", high = "#152a3a") +
  theme(panel.background = element_rect(fill = "#e3e9ed", colour = "black")) +
  labs(title='Fatalities of Top 10 Civil Operators Over Decades [Without Aeroflot]', x='Decades', y='Op
  theme(plot.title = element_text(hjust = 0.0, size=11)) +
  theme(axis.text.x=element_text(angle = 45, hjust = 1, size = 8))
```



The heatmap is now better than the line graph we used in our group project. As we have discussed in class, between 70s and 90s the world experienced lots of airplane crashes with lots of fatalities! But now this trend reverted to a decline. (Don't overlook THY in 70s! There is a terrible crash in 1974.)

Now my soul is relieved after improving our less understandable graphs!

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 (YÖK) 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.

- a) Gathering the Data

We've collected data from YOK. These were: 1. Number of students across cities, years, universities and university type. 2. Foreign students coming to Turkey, across universities, years, university type and of course their nationality. We worked with my wife Yagmur. The hardest part was correcting nationalities, there were many duplicate values ('**Yunanistan**' & '**Yunanistan Cumhuriyeti**') or typos ('**Etiyopya**' & '**Etyopya**'). It took more than 2 hours to correct them. We wrapped the data in an RData file and shared it on my MEF Github repository. (link)

- b) EDA

I decided to analyze foreign students. I'm going to generate some cool visualizations. YOK shares foreign students' data only from 2013. I will analyse years between 2013-2016.

As usual, I started with simple questions to seek answers:

- What about foreign students' numbers across years? Is there are a trend?
- Which university type do they choose? Public or Private?
- Where are they headed? In which cities' universities do they prefer?
- Where do they come from?
- Are there any difference in terms of # of students between sending countries?

Let's call our necessary libraries, import dataset.

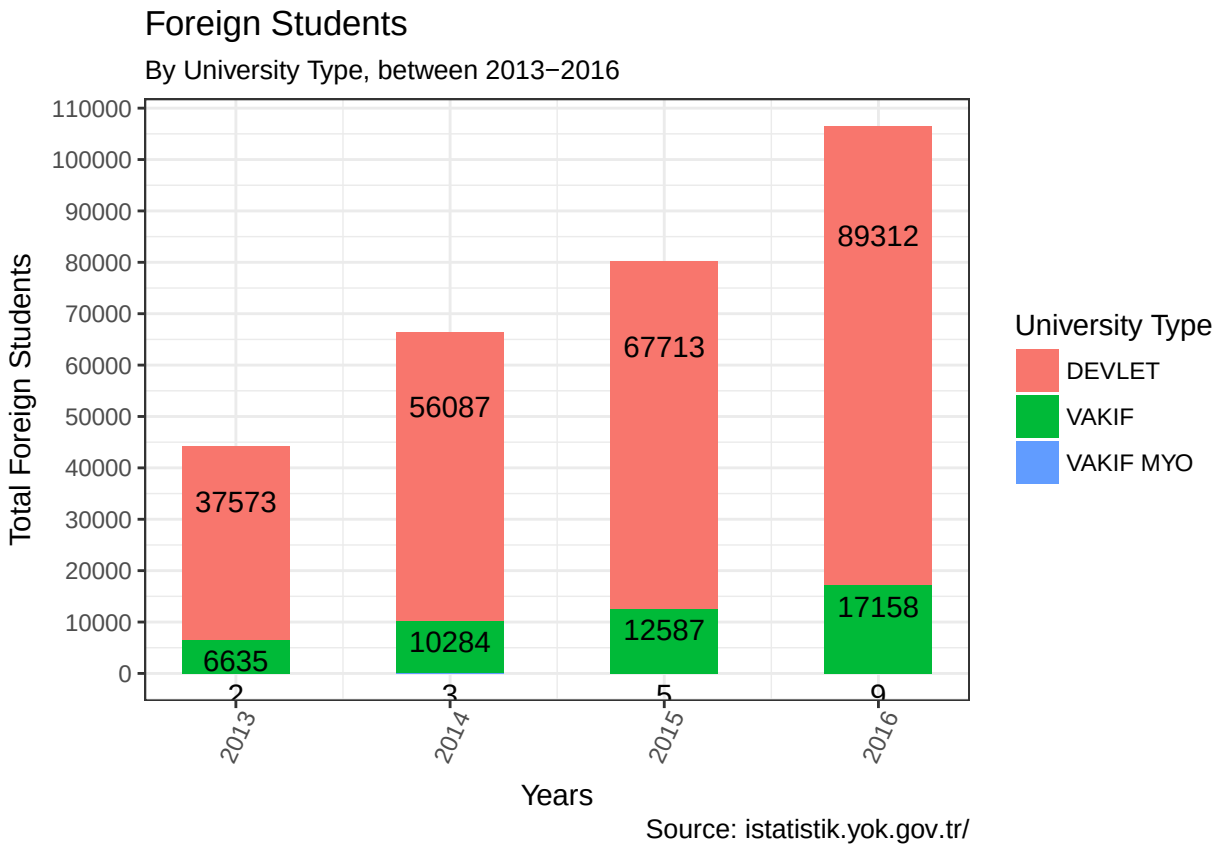
```
library(treemapify)
options(scipen=999)
load(url("https://raw.githubusercontent.com/MEF-BDA503/pj-tektens/master/files/bda503_rdata/final_datas
```

Yearly Trends

As you can see in the bar graph, foreign students coming to Turkey are ca. **2.3 times** higher than 2013 today. More foreign students prefer Turkey. A similar trend can be observed in private universities. But most of the students choose public universities: ca. **5.2 times** more comparing the numbers for 2016. But the growth rate of private schools are increasing more than the public schools over years.

```
unitype_wnat <- students_by_nationality %>%
                  select(uni_type, total, ent_year) %>%
                  group_by(uni_type,ent_year) %>%
                  summarise(total=sum(total)) %>%
                  arrange(uni_type, ent_year)

ggplot(unitype_wnat, aes(x=ent_year, y=total, fill=uni_type)) +
    theme_bw() +
    geom_bar(stat='identity', width = 0.5) +
    theme(axis.text.x = element_text(angle=65, vjust=0.6)) +
    labs(title="Foreign Students",
         subtitle="By University Type, between 2013-2016",
         caption="Source: istatistik.yok.gov.tr/") +
    scale_y_continuous(breaks = seq(0,110000,10000)) +
    geom_text(aes(label=total, vjust = 1.5)) +
    labs(x = "Years", y = "Total Foreign Students") +
    guides(fill=guide_legend(title="University Type"))
```



Where are they headed?

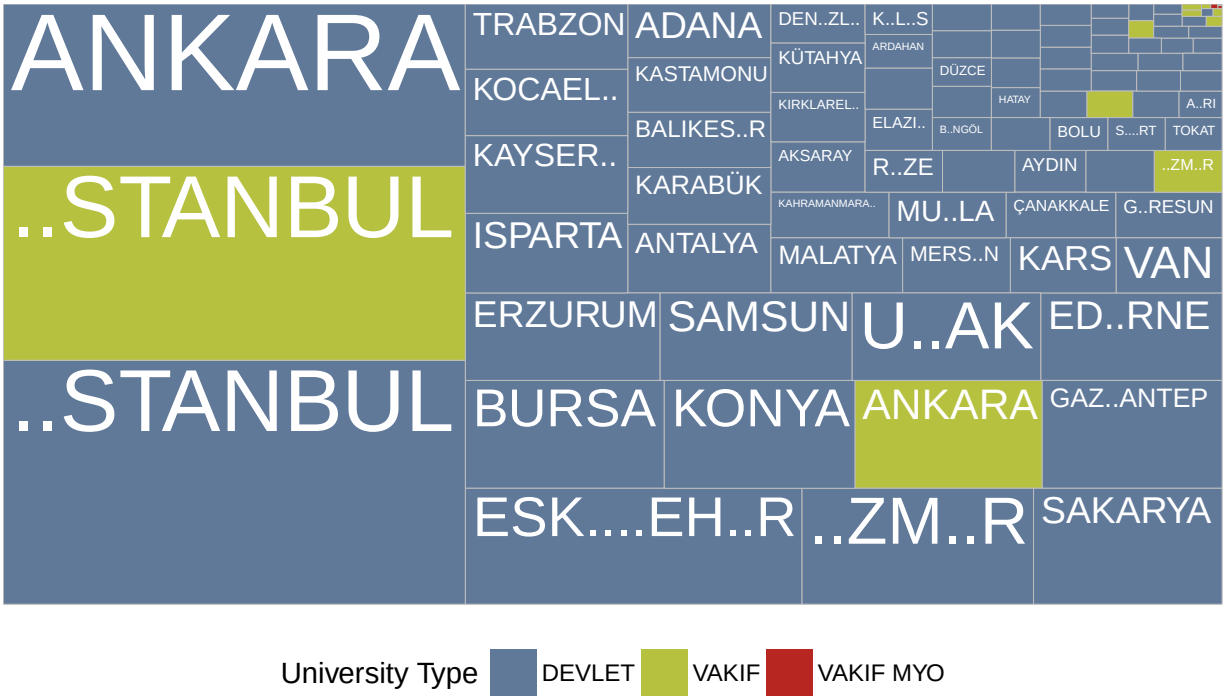
I've prepared a treemap for this question. Blue boxes represent private universities and green ones are private. Most of the foreign students prefer **Ankara** & **Istanbul**. Private universities collect most of their share from **Istanbul**. After Ankara and Istanbul; Eskisehir, Izmir, Sakarya, Bursa, Gaziantep and Erzurum are also attraction points. Second largest city with most foreign students choosing private universities is **Ankara**. I can say that I'm not suprised with these results. I may investigate the distribution of private universities across cities deeper in the future. The reason behind foreign students choosing private schools in Istanbul and Ankara more may be that most of the private schools are aggregated in these cities. :)

```
tree <- students_by_nationality %>%
          select(city, uni_type, total) %>%
          group_by(city, uni_type) %>%
          summarise(total=sum(total))

ggplot(tree, aes(area = total, fill = uni_type, label = city)) +
    geom_treemap() +
    geom_treemap_text(colour = "white", place = "topleft", grow = TRUE) +
    scale_fill_manual (values=c("#607999", "#b6c13f", "#b72621")) +
    theme(legend.position = "bottom") +
    labs(
      title = "Which Cities do Foreign Students choose?",
      subtitle = "Foreign Students Across cities, All Years",
      caption="Source: istatistik.yok.gov.tr/") +
    guides(fill=guide_legend(title="University Type"))
```

Which Cities do Foreign Students choose?

Foreign Students Across cities, All Years



Source: istatistik.yok.gov.tr/

Where do they come from?

I've prepared a bar chart showing the z-scores of sending countries (They refer to students' nationalities). The graph doesn't show all countries, otherwise it would be impossible to read it. I chose to show only countries sending more than 1000 students between 2013-2016. The results are interesting for me. The reason is only **9 countries** increase the average. The remainder **35 countries** are decreasing it. Above average countries are: Azerbaycan, Turkmenistan, Suriye, Iran, Afganistan, Irak, Yunanistan, Kirgizistan, Kazakistan. Some of them are our neighbour countries. 4 of them have z-scores above **1.6**. Except Yunanistan, none of them are from Africa, America or Europe. It seems Turkey is not as popular in the West as in the East. The reason may be the instability among these countries or in the region: Polical conjuncture, wars etc. and students in these countries may be more inclined to visit Turkey. Almanya is not increasing the average, which seemed to me interesting, since there are lots of Turkish people in Germany.

```
Sys.setlocale(locale = "Turkish_Turkey.1254")
```

```
## [1] ""
```

```
nat <- students_by_nationality %>%
      select(nationality, male, female, total) %>%
      group_by(nationality) %>%
      summarise(total=sum(total),male=sum(male),female=sum(female)) %>%
      filter(total>1000) %>%
      mutate(stu_z = round((total - mean(total))/sd(total),2), avg_type = ifelse(stu_z<0,"below","above
      arrange(desc(stu_z)) %>%
      mutate(nationality = factor(nationality, levels=nationality))

ggplot(nat, aes(x=nationality, y=stu_z, label=stu_z)) +
```

```
      theme_bw() +
      geom_bar(stat='identity', aes(fill=avg_type), width=.5)  +
      scale_fill_manual(name="Foreign Students",
                        labels = c("Above Average", "Below Average"),
                        values = c("above"="#607999", "below"="#b72621")) +
      labs(subtitle="Normalised Foreign Students came from Different Countries",
           title= "Diverging Countries") +
      theme(axis.text.y = element_text(size=6), axis.text.x = element_text(size=6)) +
      theme(legend.position="bottom") +
      coord_flip()
```