BDA 503 – BERK ORBAY - FINAL EXAM

DEVRIM NESIPOGLU

PART I . 1

1.1.QUESTION.

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?

1.1.ANSWER.

with dual axis graph, you can summarize or plot two y axis variables that have different variables that have different domains. For example, you can plot the number of cases on one axis and the mean salary on another. Also this chart can also be a mix of different graphic elements so that the dual y axis chart encompasses several of the different chart types types.It may display the counts as a line and the mean of each category as a bar. There are very few situations where it is appropriate to use two different scales on the same plot.It is very easy to mislead and confuse the viewer of the graphic. So I can say, it will change up to information that I want to visualize. If it makes graph difficult to understand, and creates confusion, I will not use.

PART I . 2

1.2. QUESTION

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. with the objective of maximum positive impact on the society in general.

1.2. ANSWER

- Exploratory data analysis workflow Aim of the EDA is to understand data better. Easiest way is to use questions as tools to guide investigation. With the help of questions, you can find chance to dig deep inside data, have knowledge about pattern and structure or realize handicaps and challenges. After you gain ideas of about data, you can start for data cleaning. Not Applicable values, characters that creates problem, outliers will be cleaned after investigation of data.
- Generate questions about data.
- Search for answers by visualizing, transforming, and modelling your data.
- Use what you learn to refine your questions and/or generate new questions.
- How do you measure impact? I will look at the data regarding questionnaire at drop in visits to people who take funds. I will compare with data before the the fund given and after the people used fund.According to aim, I will look the difference and measure performance of project. To be honest, I would choose the title “Pain Points in Our Society and Optimal Budget Allocation”?

PART I . 3

1.3. QUESTION

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

1.3. ANSWER

A time series is a series of data points indexed (or listed or graphed) in time order. Time series are a series of observations made over a certain time interval. It is commonly used in economic forecasting as well as analyzing climate data over large periods of time. The main idea behind time series analysis is to use a certain number of previous observations to predict future observations.Any metric that is measured over regular time intervals makes a Time Series. Example: Weather data, Stock prices, Industry forecasts, etc are some of the common ones.

Time series models are very useful models when you have serially correlated data. Most of business houses work on time series data to analyze sales number for the next year, website traffic, competition position and much more. However, it is also one of the areas, which many analysts do not understand.

The biggest difference is that time series regression accounts for the autocorrelation between time events, which always exists, while in normal regression, independence of serial errors are presumed, or at least minimized.

The Bit coin is a digital currency which has recently emerged as a peer-to-peer payment system to facilitate transactions. It is not issued by any central bank or other financial institution but uses cryptographic methods and relies on an open-source software algorithm which verifies decentralized transactions and controls the creation of new Bit coins. At Bit coin price movement analysis, we can identify pattern.Date of observation of bit coin prices can be used.

At diamonds data, there is no time events for that reason we cannot analyze with time series model

PART I . 4

1.4. QUESTION

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

1.4. ANSWER

I want to look for if there is a relationship between longer movies with length less than 240 minutes and the number of votes the movies received on IMDB. Also use the alpha parameter to geom_point to deal with over-plotting and use the geom_smooth function to add a regression line (without confidence bands) to the plot.

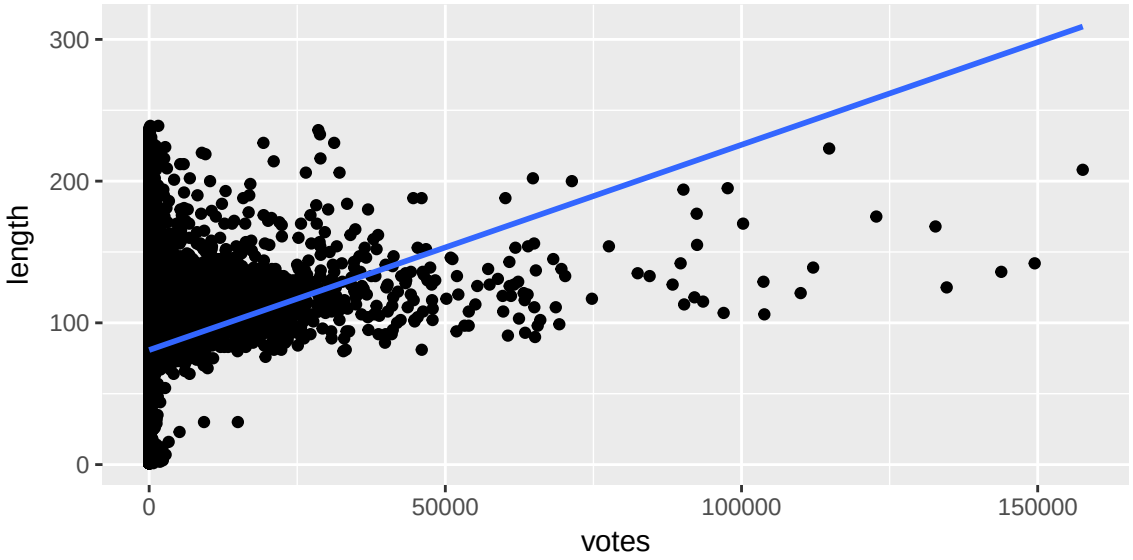
```
library(ggplot2movies)
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------- tidyvers
```

```
## v ggplot2 2.2.1     v purrr   0.2.4
## v tibble  1.3.4     v dplyr   0.7.4
## v tidyr   0.7.2     v stringr 1.2.0
## v readr   1.1.1     v forcats 0.2.0
```

```
## -- Conflicts ------------------------------------------------------------------------ tidyverse_con
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
movies %>% filter(length < 240) %>%
  ggplot(aes(x = votes, y = length)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```



PART II . Extending Your Group Project

2. QUESTION

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.

2. ANSWER

"We care about your needs!"

```
#Sys.setlocale(locale="Turkish_Turkey.1254")
#knitr::opts_chunk$set(echo = FALSE)

#setwd("C:/Users/Dell_User/Documents/GitHub/Final")
options(Encoding="UTF-8")
library(tidyverse)
library(knitr)
```

```
library(kableExtra)
library(ggplot2)
library(dplyr)
library(scales)
```

```
##
## Attaching package: 'scales'
```

```
## The following object is masked from 'package:purrr':
##
##     discard
```

```
## The following object is masked from 'package:readr':
##
##     col_factor
```

```
library(stringr)
library(reshape)
```

```
##
## Attaching package: 'reshape'
```

```
## The following object is masked from 'package:dplyr':
##
##     rename
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, smiths
```

```
library(arules)
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

```
## The following object is masked from 'package:reshape':
##
##     expand
```

```
## The following object is masked from 'package:tidyr':
##
##     expand
```

```
##
## Attaching package: 'arules'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following objects are masked from 'package:base':
##
##     abbreviate, write
```

```
library(arulesViz)
```

```
## Loading required package: grid
```

```
library(kableExtra)
```

How People Order Online? Analysis of 3 million Instacart orders.

BDA-503 - Term Project - Berk Orbay

As group **cleveR**, we will work on the online order behaviors of Instacart

Departments

This file contains the names of the departments with their department_id and name of the department.

- department_id: department identifier
- department: the name of the department

department_id	department
1	frozen
2	other
3	bakery
4	produce
5	alcohol

Orders

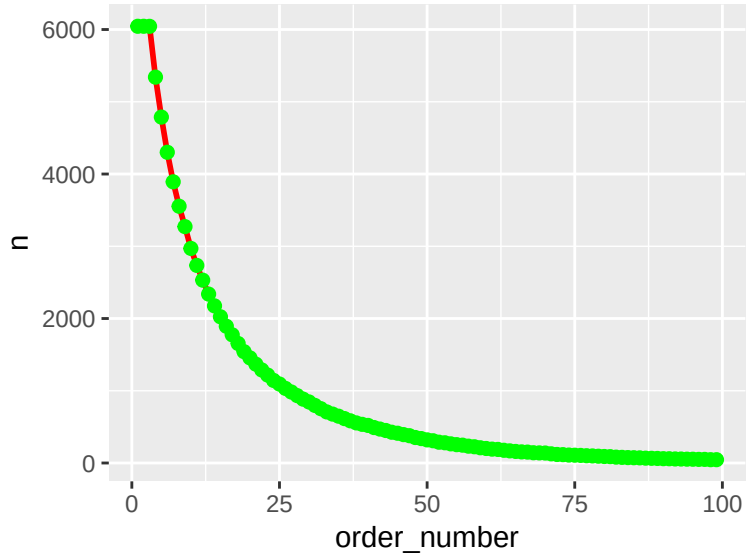
This file gives a list of all orders we have in the data set. 1 row per order. For example, we can see that user 1 has 11 orders, 1 of which is in the train set, and 10 of which are prior orders. The orders.csv doesn't tell us about which products were

order_id	user_id	eval_set	order_number	order_dow	order_hour_of_day	days_since_prior_order
2539329	1	prior	1	2	8	NA
2398795	1	prior	2	3	7	15
473747	1	prior	3	3	12	21
2254736	1	prior	4	4	7	29
431534	1	prior	5	4	15	28

How many prior orders are there?

We can see that there are always at least 3 prior orders.

```
orders %>% filter(eval_set=="prior") %>% count(order_number) %>% ggplot(aes(order_number,n)) + geom_lin
```



How many items do people buy?

Let's have a look how many items are in the orders. We can see that people most often order around 5 items. The distributions are comparable between the train and prior order set.

Order_Products_Prior (op_prior) gives us information about which products (product_id) were ordered. It also contains information of the order (add_to_cart_order) in which the products were put into the cart and information of whether this product is a re-order(1) or not(0).

```
orders<-read.csv("data/orders.csv" ,nrows = 100000)
#orders<-read.csv("data/orders.csv")
kable(head(orders,5),align="l")
```

order_id	user_id	eval_set	order_number	order_dow	order_hour_of_day	days_since_prior_order
2539329	1	prior	1	2	8	NA
2398795	1	prior	2	3	7	15
473747	1	prior	3	3	12	21
2254736	1	prior	4	4	7	29
431534	1	prior	5	4	15	28

```
op_train<-read.csv("data/order_products__train.csv",nrows = 100000)
#op_train<-read.csv("data/order_products__train.csv")
kable(head(op_train,5),align="l")
```

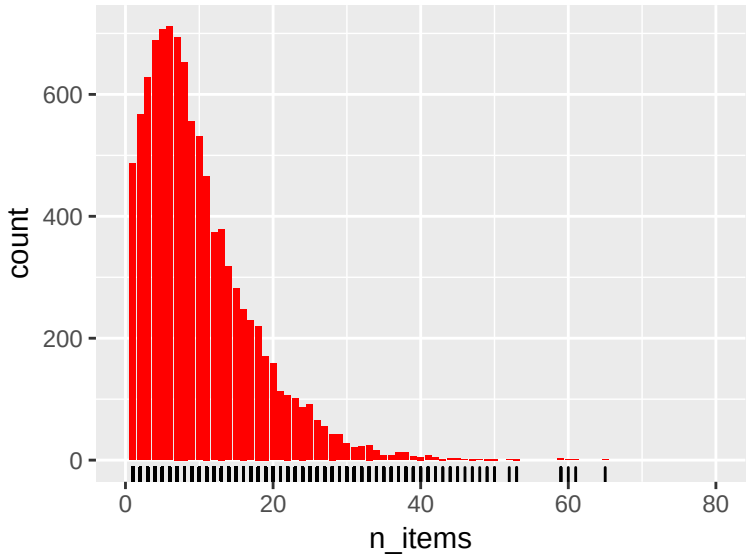
order_id	product_id	add_to_cart_order	reordered
1	49302	1	1
1	11109	2	1
1	10246	3	0
1	49683	4	0
1	43633	5	1

```
op_prior<-read.csv("data/order_products__prior.csv",nrows = 100000)
kable(head(op_prior,4),align="l")
```

order_id	product_id	add_to_cart_order	reordered
2	33120	1	1
2	28985	2	1
2	9327	3	0
2	45918	4	1

```
op_prior %>%
  group_by(order_id) %>%
  summarize(n_items = last(add_to_cart_order)) %>%
  ggplot(aes(x=n_items))+
  geom_histogram(stat="count",fill="red") +
  geom_rug() +
  coord_cartesian(xlim=c(0,80))
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

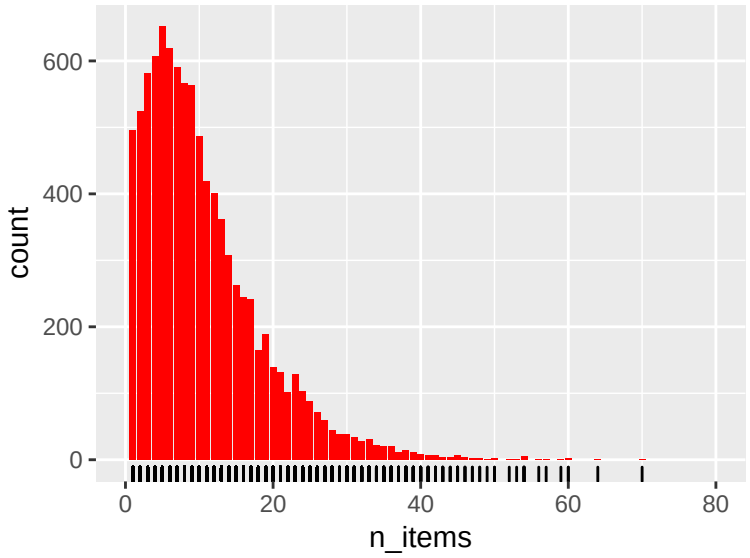


Order_Products_Train (op_train) gives us information about which products (product_id) were ordered. It also contains information of the order (add_to_cart_order) in which the products were put into the cart and information of whether this product is a re-order(1) or not(0).

Train Set

```
op_train %>%
  group_by(order_id) %>%
  summarize(n_items = last(add_to_cart_order)) %>%
  ggplot(aes(x=n_items))+
  geom_histogram(stat="count",fill="red") +
  geom_rug()+
  coord_cartesian(xlim=c(0,80))
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



order_id	product_id	add_to_cart_order	reordered
1	49302	1	1
1	11109	2	1
1	10246	3	0
1	49683	4	0
1	43633	5	1

PART III: Welcome to Real Life

3. QUESTION:

Gather data from Higher Education Council's (YÖK) data service. https://istatistik.yok.gov.tr/ . Choose an interesting theme which can be analyzed with the given data and collect relevant data from the service. 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. b) Perform EDA on the data you collected based on the theme you decided on.

3. ANSWER:

```
suppressMessages(library(tidyverse))
library(dplyr)
library(knitr)
library(ggplot2)
library(shiny)
student_info <- read.csv("C:/Users/Dell_User/Documents/GitHub/Final/Student Numbers RawData.csv", sep="
setwd("C:/Users/Dell_User/Documents/GitHub/pj-nesipoglud/Final")
download.file("https://mef-bda503.github.io/pj-nesipoglud/files/Final_R.RData","Final_R.RData")
dat <- get(load("Final_R.RData")) %>%tbl_df()
```

```
print("Education System of Turkey data row number")
```

```
## [1] "Education System of Turkey data row number"
```

```
nrow(student_info)
```

```
## [1] 147
```

```
## [1] "LC_COLLATE=Turkish_Turkey.1254;LC_CTYPE=Turkish_Turkey.1254;LC_MONETARY=Turkish_Turkey.1254;LC_
```

ANALYSIS OF HIGHER EDUCATION SYSTEM OF TURKEY

```
names(student_info)[1]<-"University_Name"
names(student_info)[2]<-"Year_of_Organization"
names(student_info)[3]<-"Type"
names(student_info)[4]<-"Province"
names(student_info)[5]<-"Region"
names(student_info)[6]<-"Male_Vocational_Training_School_Students"
names(student_info)[7]<-"Female_Vocational_Training_School_Students"
names(student_info)[8]<-"Total_Vocational_Training_School_Students"
names(student_info)[9]<-"Male_Undergraduate_Students"
names(student_info)[10]<-"Female_Undergraduate_Students"
names(student_info)[11]<-"Total_Undergraduate_Students"
names(student_info)[12]<-"Male_Master_Students"
names(student_info)[13]<-"Female_Master_Students"
names(student_info)[14]<-"Total_Master_Students"
names(student_info)[15]<-"Male_Doctorate_Students"
names(student_info)[16]<-"Female_Doctorate_Students"
names(student_info)[17]<-"Total_Doctorate_Students"
names(student_info)[18]<- "Male_Grand_Total"
names(student_info)[19]<-"Female_Grand_Total"
names(student_info)[20]<-"Grand_Total"
```

After tidying data and have exploratory data analysis, we can start for getting deep inside the data

```
names(student_info)
```

```
##  [1] "University_Name"
##  [2] "Year_of_Organization"
##  [3] "Type"
##  [4] "Province"
##  [5] "Region"
##  [6] "Male_Vocational_Training_School_Students"
##  [7] "Female_Vocational_Training_School_Students"
##  [8] "Total_Vocational_Training_School_Students"
##  [9] "Male_Undergraduate_Students"
## [10] "Female_Undergraduate_Students"
## [11] "Total_Undergraduate_Students"
## [12] "Male_Master_Students"
## [13] "Female_Master_Students"
## [14] "Total_Master_Students"
## [15] "Male_Doctorate_Students"
## [16] "Female_Doctorate_Students"
## [17] "Total_Doctorate_Students"
## [18] "Male_Grand_Total"
```

```
## [19] "Female_Grand_Total"
## [20] "Grand_Total"
```

```
head(student_info)
```

```
##                   University_Name Year_of_Organization   Type
## 1 ABANT IZZET BAYSAL UNIVERSITESI                 1992 DEVLET
## 2           ACIBADEM UNIVERSITESI                 2007  VAKIF
## 3           ADIYAMAN UNIVERSITESI                 2006 DEVLET
## 4     ADNAN MENDERES UNIVERSITESI                 1992 DEVLET
## 5     AFYON KOCATEPE UNIVERSITESI                 1992 DEVLET
## 6 AGRI IBRAHIM CECEN UNIVERSITESI                 2007 DEVLET
##         Province            Region
## 1           BOLU         KARADENIZ
## 2       ISTANBUL           MARMARA
## 3       ADIYAMAN GUNEYDOGU ANADOLU
## 4          AYDIN               EGE
## 5 AFYONKARAHISAR               EGE
## 6           AGRI      DOGU ANADOLU
##   Male_Vocational_Training_School_Students
## 1                                     3286
## 2                                      130
## 3                                     4554
## 4                                     6570
## 5                                     9042
## 6                                      529
##   Female_Vocational_Training_School_Students
## 1                                       1907
## 2                                        171
## 3                                       2642
## 4                                       4898
## 5                                       5482
## 6                                        188
##   Total_Vocational_Training_School_Students Male_Undergraduate_Students
## 1                                      5193                        5001
## 2                                       301                          93
## 3                                      7196                        2168
## 4                                     11468                        6506
## 5                                     14524                        8063
## 6                                       717                        2100
##   Female_Undergraduate_Students Total_Undergraduate_Students
## 1                          7371                        12372
## 2                           244                          337
## 3                          2270                         4438
## 4                          6657                        13163
## 5                          7570                        15633
## 6                          1277                         3377
##   Male_Master_Students Female_Master_Students Total_Master_Students
## 1                  320                   396                   716
## 2                    0                     0                     0
## 3                   57                    46                   103
## 4                  200                   283                   483
## 5                  350                   278                   628
## 6                    0                     0                     0
##   Male_Doctorate_Students Female_Doctorate_Students
```

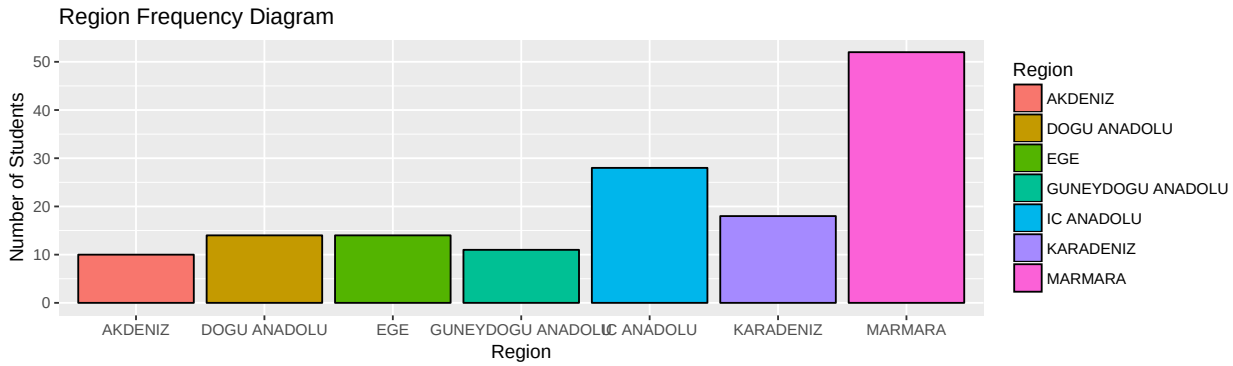
```
## 1                      92                       61
## 2                       0                        0
## 3                       0                        0
## 4                     123                       99
## 5                      99                       47
## 6                       0                        0
##   Total_Doctorate_Students Male_Grand_Total Female_Grand_Total Grand_Total
## 1                      153             8699               9735       18434
## 2                        0              223                415         638
## 3                        0             6779               4958       11737
## 4                      222            13399              11937       25336
## 5                      146            17554              13377       30931
## 6                        0             2629               1465        4094
```

```
#Before tidying data set
summary(student_info)
```

```
##                         University_Name Year_of_Organization
##  ABANT IZZET BAYSAL UNIVERSITESI:  1    Min.   :1924
##  ACIBADEM UNIVERSITESI          :  1    1st Qu.:1992
##  ADIYAMAN UNIVERSITESI          :  1    Median :2001
##  ADNAN MENDERES UNIVERSITESI    :  1    Mean   :1996
##  AFYON KOCATEPE UNIVERSITESI    :  1    3rd Qu.:2007
##  AGRI IBRAHIM CECEN UNIVERSITESI:  1    Max.   :2012
##  (Other)                        :141
##         Type         Province                Region
##  DEVLET   :93   ISTANBUL :41   AKDENIZ          :10
##  VAKIF    :45   ANKARA   :11   DOGU ANADOLU     :14
##  VAKIF MYO: 9   IZMIR    : 7   EGE              :14
##                 GAZIANTEP: 3   GUNEYDOGU ANADOLU:11
##                 KONYA    : 3   IC ANADOLU       :28
##                 MERSIN   : 3   KARADENIZ        :18
##                 (Other)  :79   MARMARA          :52
##  Male_Vocational_Training_School_Students
##  Min.   :    0.0
##  1st Qu.:  150.5
##  Median : 1185.0
##  Mean   : 2498.9
##  3rd Qu.: 3781.0
##  Max.   :16825.0
##
##  Female_Vocational_Training_School_Students
##  Min.   :   0
##  1st Qu.: 175
##  Median : 788
##  Mean   :1516
##  3rd Qu.:2358
##  Max.   :9461
##
##  Total_Vocational_Training_School_Students Male_Undergraduate_Students
##  Min.   :    0                             Min.   :     0.0
##  1st Qu.:  346                             1st Qu.:   615.5
##  Median : 1931                             Median :  2089.0
##  Mean   : 4015                             Mean   : 10834.5
##  3rd Qu.: 5960                             3rd Qu.:  7022.0
```

```
##  Max.   :25067                           Max.   :940771.0
##
##  Female_Undergraduate_Students Total_Undergraduate_Students
##  Min.   :     0                Min.   :      0
##  1st Qu.:   502                1st Qu.:   1189
##  Median :  2065                Median :   4280
##  Mean   :  9583                Mean   :  20418
##  3rd Qu.:  6326                3rd Qu.:  13234
##  Max.   :795126                Max.   :1735897
##
##  Male_Master_Students Female_Master_Students Total_Master_Students
##  Min.   :   0.0       Min.   :   0.0         Min.   :   0.0
##  1st Qu.:   1.0       1st Qu.:   0.0         1st Qu.:   1.0
##  Median : 184.0       Median : 139.0         Median : 369.0
##  Mean   : 454.9       Mean   : 398.5         Mean   : 853.4
##  3rd Qu.: 571.5       3rd Qu.: 477.0         3rd Qu.:1052.0
##  Max.   :3414.0       Max.   :3773.0         Max.   :7148.0
##
##  Male_Doctorate_Students Female_Doctorate_Students
##  Min.   :   0.0          Min.   :   0.0
##  1st Qu.:   0.0          1st Qu.:   0.0
##  Median :  15.0          Median :   6.0
##  Mean   : 162.1          Mean   : 130.0
##  3rd Qu.: 125.0          3rd Qu.:  94.5
##  Max.   :1975.0          Max.   :1871.0
##
##  Total_Doctorate_Students Male_Grand_Total Female_Grand_Total
##  Min.   :   0.0           Min.   :    17   Min.   :    19
##  1st Qu.:   0.0           1st Qu.:  1481   1st Qu.:  1130
##  Median :  23.0           Median :  3784   Median :  3238
##  Mean   : 292.1           Mean   : 13950   Mean   : 11628
##  3rd Qu.: 218.5           3rd Qu.: 11953   3rd Qu.:  8836
##  Max.   :3745.0           Max.   :944729   Max.   :797357
##
##   Grand_Total
##  Min.   :     59
##  1st Qu.:   2762
##  Median :   6909
##  Mean   :  25578
##  3rd Qu.:  20114
##  Max.   :1742086
##
```

```
ggplot(aes(x=Region), data=subset(student_info, !is.na(Region))) +
  geom_bar(aes(fill=Region), color='black') +
  labs(x= 'Region', y='Number of Students', title= 'Region Frequency Diagram')
```



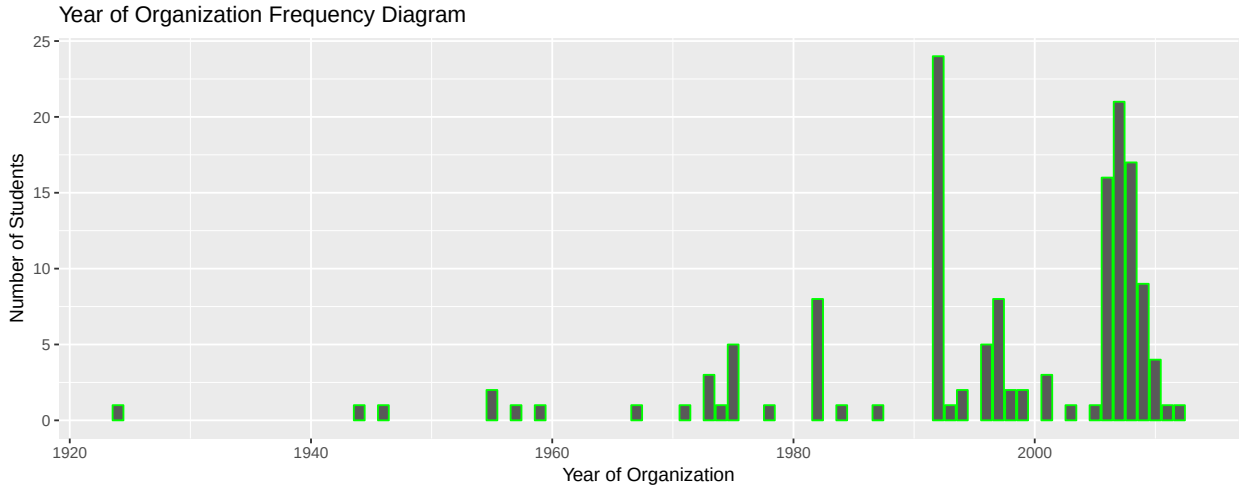
University Numbers by University Type

```
student_info %>%
  count(Type)
```

```
## # A tibble: 3 x 2
##        Type     n
##      <fctr> <int>
## 1    DEVLET    93
## 2     VAKIF    45
## 3 VAKIF MYO     9
```

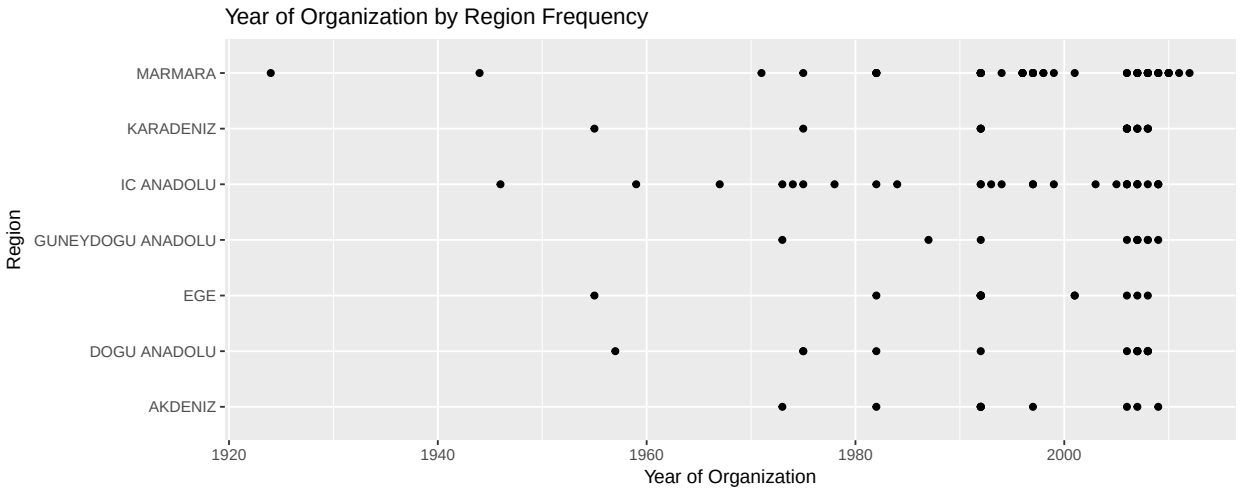
Year of Organizations of Universities

```
ggplot(aes(x=Year_of_Organization), data=subset(student_info, !is.na(Year_of_Organization))) +
  geom_bar(aes(fill=Year_of_Organization), color='green') +
  labs(x= 'Year of Organization', y='Number of Students', title= 'Year of Organization Frequency Diagra
```



Let's look year of organizations of universities by Region

```
ggplot(data = student_info) +
  geom_point(mapping = aes(x = Year_of_Organization, y = Region)) +
    labs(x= 'Year of Organization', y='Region', title= 'Year of Organization by Region Frequency')
```



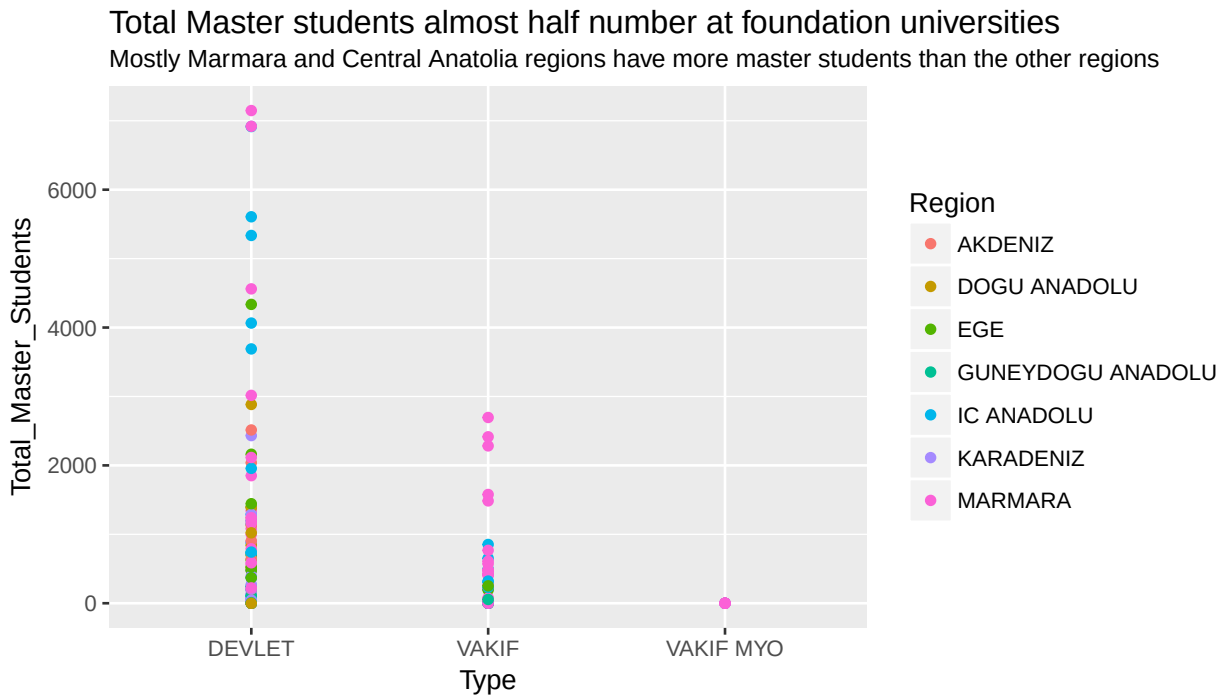
Number of students by university type

```
student_info %>% group_by(Type) %>%  summarise(quota=sum(Grand_Total)) %>% arrange(desc(Type))
```

```
## # A tibble: 3 x 2
##        Type   quota
##      <fctr>   <int>
## 1 VAKIF MYO    5981
## 2     VAKIF  190365
## 3    DEVLET 3563653
```

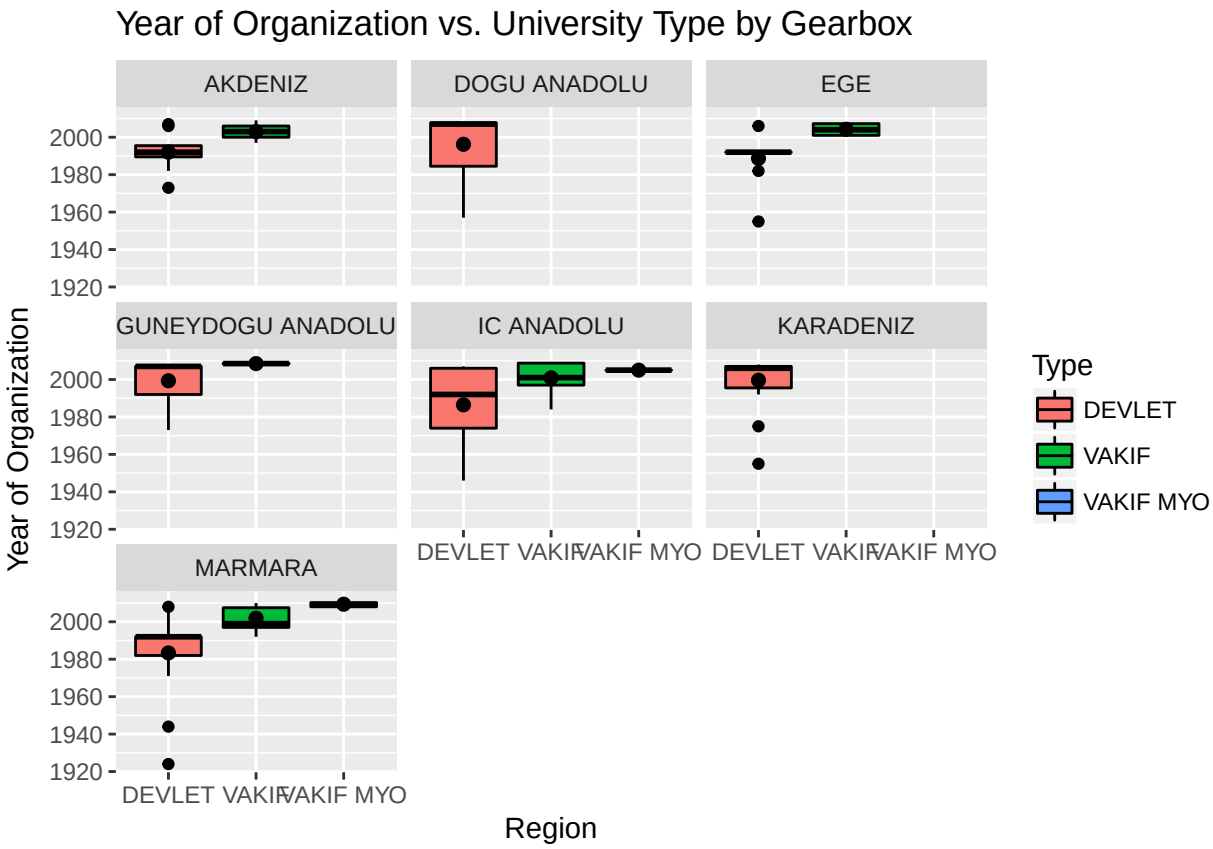
```
ggplot(student_info, aes(Type,Total_Master_Students)) +
  geom_point(aes(color = Region)) +
  geom_smooth(se = FALSE) +
  labs(
    title = "Total Master students almost half number at foundation universities",
    subtitle = "Mostly Marmara and Central Anatolia regions have more master students than the other re
  )
```

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



With this graph, establishment of new universities analyzed in 20 years of period by university type and and Region.

```
ggplot(data=subset(student_info, !is.na(Region) & !is.na(Type)), aes(x=Type, y=Year_of_Organization)) +
  geom_boxplot(aes(fill=Type), color='black') +
  facet_wrap(~Region) +
  stat_summary(fun.y = mean, geom="point", size=2) +
  labs(x= 'Region', y= 'Year of Organization', title= "Year of Organization vs. University Type by Gear
```



Reference

Data is download from https://istatistik.yok.gov.tr/