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.

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

---

> I need to use two y-axis graphs in my current job for a few cases, but at my previous job, I frequently used them as a financial analyst. As Wickham mentioned, two y-axis graphs are hard to read and open to manipulation to mislead. There is not a uniqe recipe for this issue yet. There are a few ways that are generally chosen to handle 'hard to read' problem since they are easy to apply and understand. The first one is facet-wrap and the other one is differentiation of unit of measure. For the second, differentation means using different metrics, for example on y1-axis showing percentages(%), on y2-axis showing units in thousands. But differentiation itself is not enough to understand which metric belongs to which variable. For this problem, it is a good way to use different colors for lines, bars, etc.and put them in the legend.
>
> Before having started to learn R, I got used to apply second one on excel sheet. In my opinion, two y-axis graphs are sometimes irreplaceable. Most of the times on my business, we need to

present our analysis, opinions and inferences in a simple but a comprehensive way. In addition, we have limited space to present our findings on slides, so it is practical to show more in less pages.

The most convenient graph type to depict and visualize relationships between great amount of numbers is two y-axis graph. If you think that your variables and unit of measures are not easily understood, then you can use different chart type in one graph, for example, bar chart for y1 axis, line for y2. Facet-wrap is another way to prevent confusion about two y-axis but it is not as practical as the former. Instead of using one graph, you need to divide the graph in n-units although you can show your results in one chart.

---

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”?

Assuming that I have almost all the data needed, firstly I begin with the data preprocessing or cleaning. Data may not be well organized and need to be manipulated. After data cleaning , adding a glossary of variables in the introduction part would be smart. In the introduction part, it is necessary to give a brief explanation about the dataset along with the main topic of the study.

After the introduction part, research questions about the dataset should be presented. There should be well-defined research questions that gives insights about the study and guidence on the contents on the rest of the analysis. The purpose of the study should be cleary defined and stated in the first part.

Thirdly, it is time to show results of research questions. You should well articulate your data modelling and support your analysis by graphs, charts or other visualization tools.

In the final step, you should draw a conclusion about your analysis; i.e, this part is actually storytelling.

At the end, you should add your references or sources.

I would totally present what data says in order to take actions or make suggestions for projects. What is important is not my subjective arguments, rather it is the actual figures that reveal us through data analysis :). Devil is hidden in the details!

---

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?

A time series data is a sequence of data points in a time interval. Regression or non time series data is a bulk of stationary data points. Regression analysis is a statistical process for estimating the relationship between variables. In terms of analysis, using a regression analysis or model to forecast time series data can produce biased results since in time series data previous data points can be correlated with future data points; in other words, data points can be autocorrelated and make us to draw biased estimators.

Another issue with time series data is that effect of explanatory variables on response variables can change over times. That means input variables can alter dynamically in the time history. In order to capture the relationships between variables, more sophisticated models may need to be utilized. If we think about Bitcoin price movements, let's say five years ago Bitcoin prices were influenced by only a few factors such as gold, commodity prices or american dollar value against major currencies, bu today those factors may not be the same and most probably there are lots of factors that effect Bitcoin prices.

---

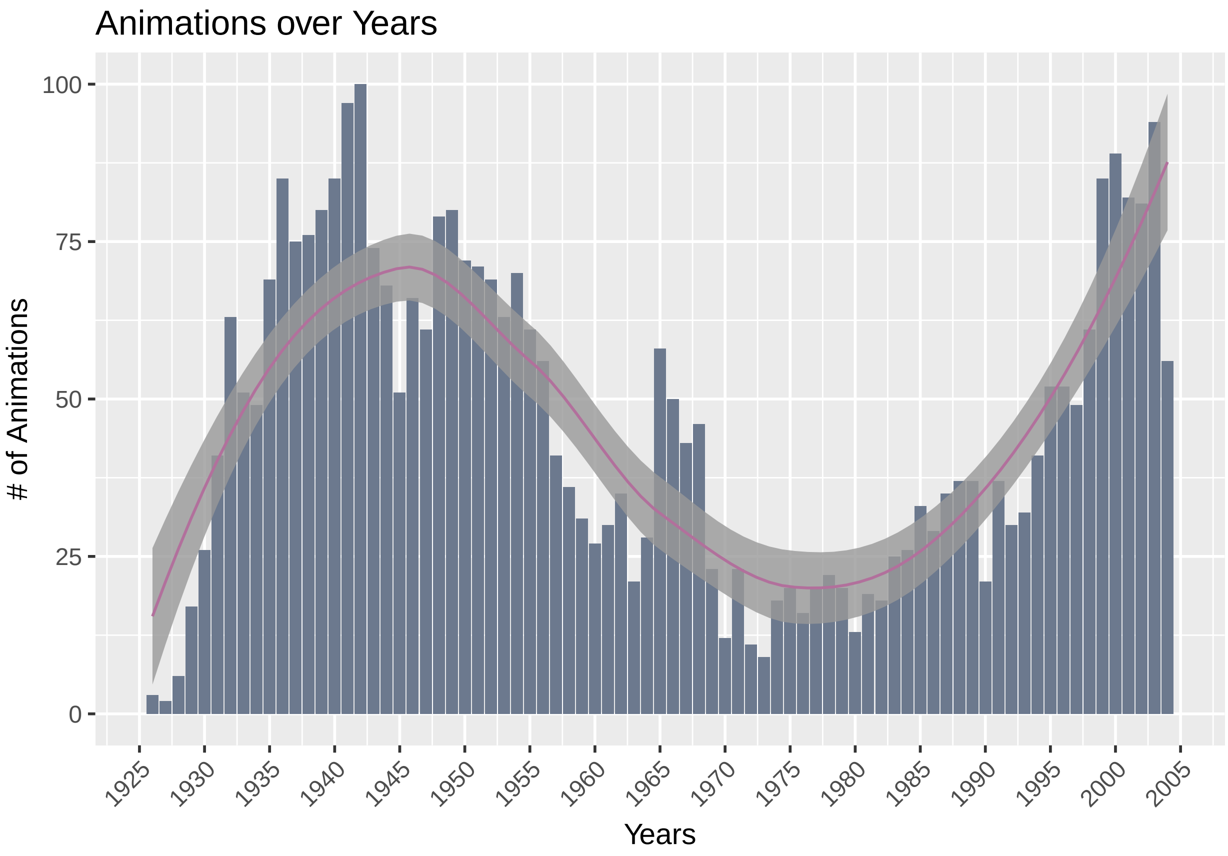
4. If you had to plot a single graph using the data below what would it be? Why? Make your argument, actually code the plot and provide the output. (You can find detailed info about the movies data set in its help file. Use `?movies`, after you load `ggplot2movies` package.)

I picked the Animation as a genre to graph since I wonder about historical evolution of this type of movies.

As you see in the graph below, animations seems at infant phase at the begining of twentieth century. The number of animations rise dramatically until the World War-II period;ie;1940s. WW2 seems to revert the upward trend in animations. The new era for animations begins with early years of 1990s. This upward trend may be explained by the technological advancements in the computer graphics animation techniques.

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

act <- movies %>% select(year, Animation) %>%
          filter(year > 1925 & year < 2005) %>%
          group_by(year) %>%
          summarize(Animation = sum(Animation)) %>%
          arrange(year)

ggplot(act, aes(x = year, y = Animation)) +
      geom_bar(stat="identity", fill = "#6c798e")+
      geom_smooth(size = 0.5,  alpha = 0.8, color = "#b2709c")+
      scale_x_continuous(breaks=seq(1925,2005,5))+
      ggtitle("Animations over Years")+
      labs(x ="Years", y = "# of Animations")+
      theme (axis.text.x=element_text (angle=45,vjust=1, hjust=1))
```

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



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.

Extra Analysis for Group Project

- As an additional improvement to our study, we want to make predictions about survival rates using machine learning methods.
- We, firstly, create a new dataframe called apc_ml(machine learning) and make data preprocessing.
- Our study aims to show how much predictions about survival rate of a crash differs from the actual.

```
# Prepare a new data frame for machine learning method and call it 'apc_ml'
apc_ml <- apc_clean

# Delete unnecessary columns
apc_ml$Location <- NULL
apc_ml$Flight.. <- NULL
apc_ml$Registration <- NULL
apc_ml$cn.In <- NULL
apc_ml$Ground <- NULL
```

```r
apc_ml$Date <- NULL
apc_ml$Route <- NULL
apc_ml$Summary <- NULL
apc_ml$Year <- NULL
apc_ml$Stops <- NULL
apc_ml$City <- NULL

# Delete following columns since they contain too many null or empty values.
apc_ml$Time <- NULL
apc_ml$Source <- NULL
apc_ml$Destination <- NULL

# Check how many rows are deleted since they are empty.
apc_ml <- apc_ml[complete.cases(apc_ml), ]

apc_ml <- apc_ml %>% filter(Operator != "") %>%
                      filter(Type != "") %>%
                      filter(State != "")

# Create a new variable "survival_rate".
apc_ml$survival_rate <- (apc_ml$Survived / apc_ml$Aboard)
```

- After data preprocessing part, we create test and train data for our analysis.

```r
# Prepare test and train data set
apc_test <- apc_ml %>% mutate(crash_id = row_number()) %>%
            group_by(Operator, State) %>%
            filter(n() > 1) %>%
            sample_frac(0.35) %>%
            ungroup()

# We break randomness in test data since some cases in  Operator and State columns have only one occure
apc_train <- anti_join(apc_ml %>% mutate(crash_id = row_number()),
             apc_test, by = "crash_id")
```

- After creating test and train datasets, we apply CART model.

```r
# Call CART model libraries.
library(rpart)
library(rpart.plot)
library(rattle)
```

```
## Rattle: A free graphical interface for data science with R.
## Version 5.1.0 Copyright (c) 2006-2017 Togaware Pty Ltd.
## Type 'rattle()' to shake, rattle, and roll your data.
```

```r
# Train the model.
survival_model <- rpart(survival_rate ~ ., data = apc_train %>% select(-crash_id))
```

- We didn’t show Fancy Plot, since there are lots of Operators, image can’t be shown properly.
- Having trained the CART model, we apply our model to test dataset and calculate modelsuccess.

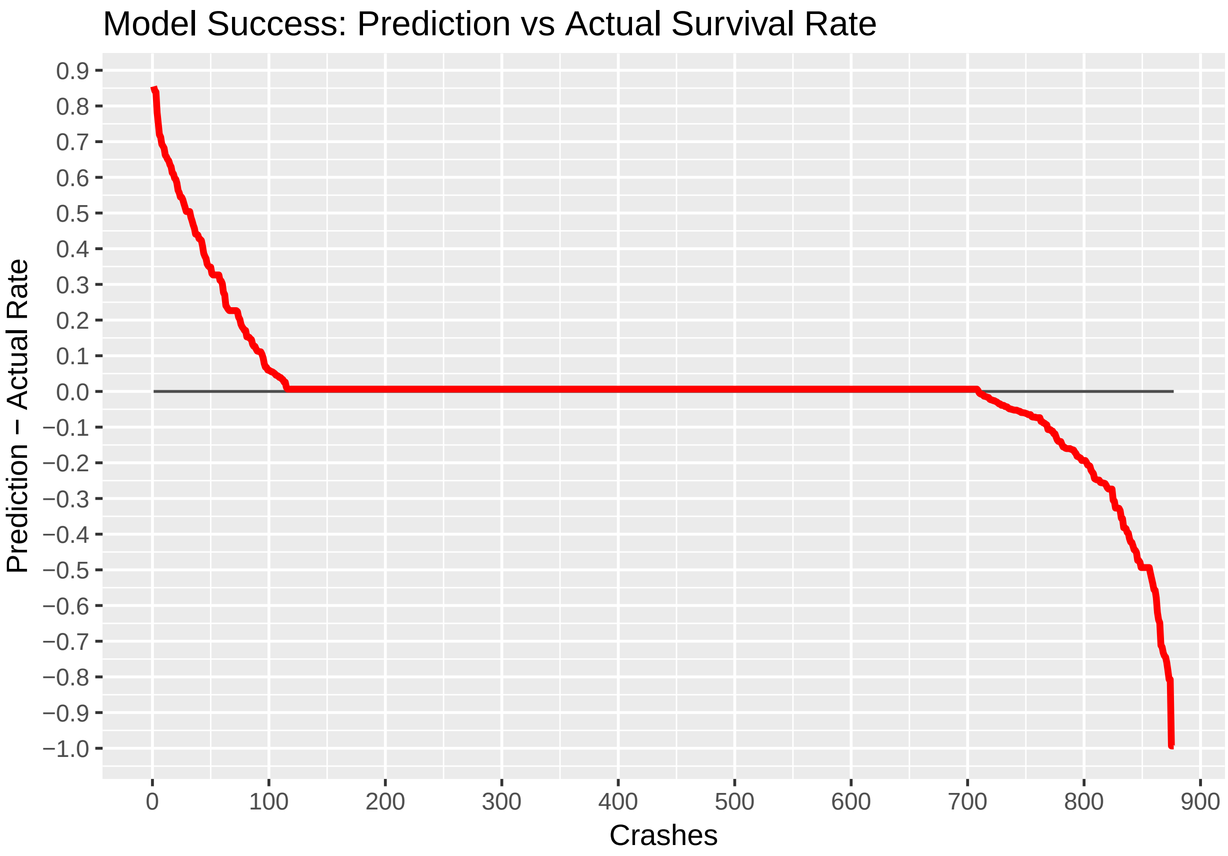
```r
# Predictions
survival_predict <- predict(survival_model, newdata = apc_test %>% select (-crash_id))

# Check model success or check how far predictions diverge from actual test results.
```

```
modelsuccess = abs(survival_predict - apc_test$survival_rate)

summary(modelsuccess)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.002321 0.006456 0.006456 0.095534 0.067995 0.993544
```

- It is time to visualize model success.

```
# Show model success by calculationg success. (Success = Survival_Predict - Survival_Rate)

df_toplot = apc_test[, c("survival_rate","crash_id")]
df_toplot = cbind(df_toplot,survival_predict)
df_toplot = df_toplot %>% arrange(crash_id)
df_toplot = df_toplot %>%
              mutate(success = survival_predict - survival_rate) %>%
              arrange(desc(success))

# Show model success on a graph.

ggplot(df_toplot, aes(x = c(1:877))) +
  geom_line(aes(y = 0), colour="black",size = 0.5, alpha = 0.7) +
  geom_line(aes(y = success), colour="red",size = 1.2) +
  ylab(label="Prediction - Actual Rate") +
  xlab("Crashes") +
  ggtitle("Model Success: Prediction vs Actual Survival Rate") +
  scale_x_continuous(breaks=seq(0,900,100)) +
  scale_y_continuous(breaks=seq(-1,1,0.1))
```



The model success graph above shows us that our model predicts a great deal amount of observations as correct (intersection on zero-line). For the first sorted approx. 150 observations, model predictions are higher than actual rates. However, for some observations predictions give lower survival rates than actual ones. If we run the codes for several times, we see different graph shapes for each attempt. The reason for this is probably the effect of outliers or maybe we have very few observations.

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 Data

- Firstly, me and my husband Semih Tekten picked two topics from the YOK data service. I want to study gender disparity among university students, Semih chose the number of foreign students and where they come from.

- My dataset consists of number of students by gender based on university type and city in the last four years. Thus, dataset covers only 2013-2016 years. The figures in my analysis take aggregates of these four years.

- Secondly, we performed data cleansing for some variables such as nationality of students, but this part was exhausting for us since it took more than 2 hours.

- Lastly, we convert our datasets into RData format and put them into Semih's MEF-Github repository.

---

b. EDA

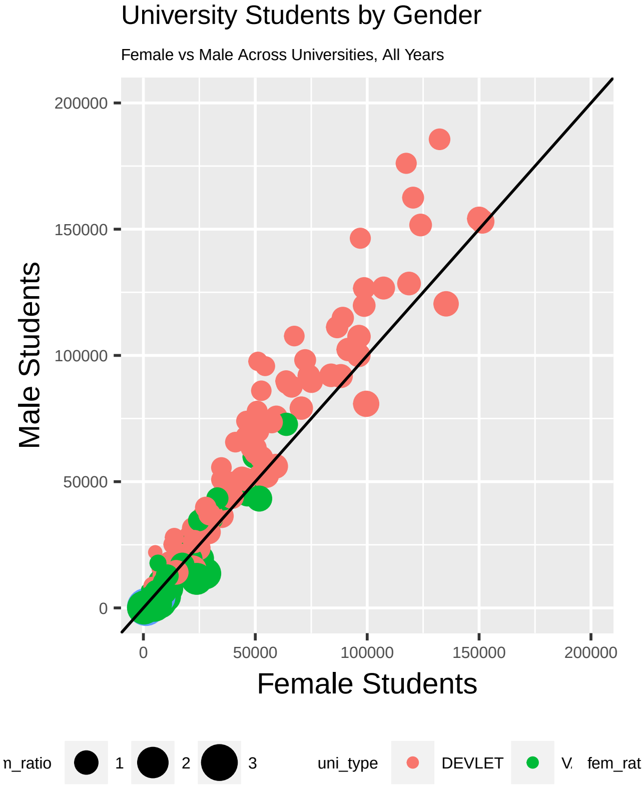
- Calling necessary libraries, dataset and settings for EDA.

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

1. Female vs Male Student Ratio Across Universities between 2013-2016 (Scatter Plot)

In this analysis, you will see the gender differences among university students for state and private universities.

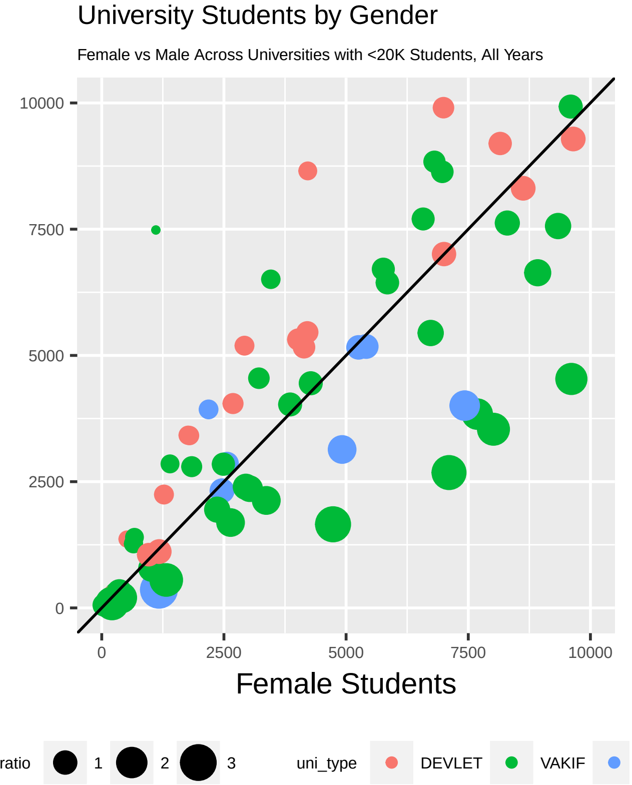
```
gscat <- student_numbers %>%
        select(female,male,uni_type,university) %>%
        group_by(university, uni_type) %>%
        summarise(female = sum(female), male = sum(male)) %>%
        mutate(fem_ratio = female/male)

plot1 <- ggplot(gscat, aes(x=female, y=male)) +
  geom_point(aes(col=uni_type, size=fem_ratio)) +
  theme(legend.position="bottom") +
  xlim(c(0, 200000)) +
  ylim(c(0, 200000)) +
  labs(y="Male Students",
       x="Female Students",
       title="University Students by Gender",
       subtitle = "Female vs Male Across Universities, All Years",
       caption = "Source: istatistik.yok.gov.tr/") +
  geom_abline(intercept = 0, slope = 1) +
  theme(plot.title=element_text(size=10), plot.subtitle=element_text(size=6)) +
  theme(axis.text.y = element_text(size=6), axis.text.x = element_text(size=6)) +
  theme(legend.title = element_text(size=6), legend.text = element_text(size = 6))

plot2 <- ggplot(gscat, aes(x=female, y=male)) +
  geom_point(aes(col=uni_type, size=fem_ratio)) +
  xlim(c(0, 10000)) +
  ylim(c(0, 10000)) +
  theme(legend.position = "bottom") +
  labs(x="Female Students",
       title="University Students by Gender",
       subtitle = "Female vs Male Across Universities with <20K Students, All Years",
       caption = "Source: istatistik.yok.gov.tr/")+
  theme(axis.title.y=element_blank()) +
  theme(axis.text.y = element_text(size=6), axis.text.x = element_text(size=6)) +
  geom_abline(intercept = 0, slope = 1) +
  theme(plot.title=element_text(size=10), plot.subtitle=element_text(size=6)) +
  theme(legend.title = element_text(size=6), legend.text = element_text(size = 6))


grid.arrange(plot1, plot2, ncol=2)
```



University Students by Gender

Female vs Male Across Universities, All Years

Source: istatistik.yok.gov.tr/



University Students by Gender

Female vs Male Across Universities with <20K Students, All Years

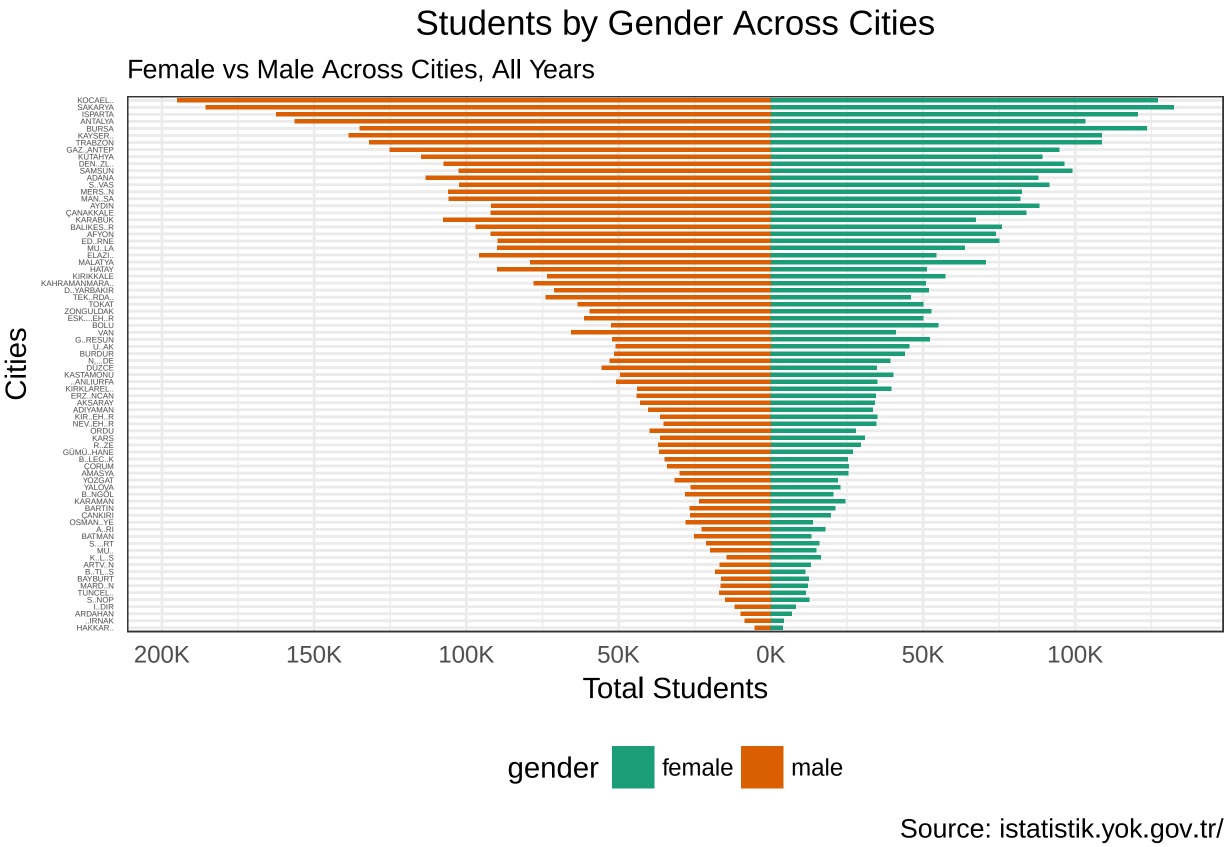
Source: istatistik.yok.gov.tr/

- The graphs above show us interesting realities about gender distribution of university students. Size of bubbles are female to male ratio. y-axes are total number of female students, x-axes are total number of male students. Each point represents a university. Colors are university types.
- On the left, you see gender disparity among universities that have less than 400.000 total numbers of students. It is seen that male students are apperantly higher than female students, especially, on state ("DEVLET") universities, although there are some exceptions. The diagonal line is added to highlight the gender disparity among students in universities.
- On the right, when the number of total students is limited to 20.000, we encounter different result. I observe that female students are higher than male counterparts for some cases, especially, for private ("VAKIF") universities. Again, the result does not change for state universities when we limit the number of total students since female students seem to be far less than male ones(compare observations for above and below line).
- However, the graphs above may be misleading since datasets covers only the last four years. If we extended the time interval back to 1980s or 1990s, the output will be probably different than the current picture. Maybe, gender disparity among university students have been improved in several decades but it is not possible to draw such a conclusion with the current dataset.

2. Female vs Male Students Across Cities (Funnel Graph)

In this analysis, I want to show the gender disparity across cities for all university types between 2013-2016. However, I need to exclude outliers such as Anadolu University in Eskisehir and cities such as Istanbul, Izmir, Ankara, Erzurum and Konya. The reason why I excluded them is that if they are in the graph, it becomes very hard to see the variations in other cities.

```r
funnel <- student_numbers %>%
  select(city,male,female,total,university) %>%
  filter(university != "ANADOLU ÜNİVERSİTESİ" & !(city %in% c('İSTANBUL','İZMİR','ANKARA','ERZURUM','KO
  select(-university) %>%
  group_by(city) %>%
  summarise(female = sum(female), male= sum(male), total = sum(total)) %>%
  arrange(total) %>%
  mutate(city_id = as.character(sprintf("%02i",row_number() ))) %>%
  mutate(ordered_city=paste(city_id,city,sep=" ")) %>%
  select(ordered_city,male,female,total) %>%
  mutate(male = -male) %>%
  gather(., gender, value, female:male)

brks <- seq(-200000, 150000, 50000)
lbls = paste0(as.character(c(seq(200, 0, -50), seq(50, 150, 50))), "K")

ggplot(funnel, aes(x = as.factor(ordered_city), y = value, fill = gender)) +
  geom_bar(stat = "identity", width = .6) +
  coord_flip() +
  theme_bw() +
  theme(plot.title = element_text(hjust = .5),
        axis.text.y = element_text(size=3),
        axis.ticks = element_blank()) +
  scale_x_discrete(labels= substring(funnel$ordered_city, 4)) +
  scale_y_continuous(breaks = brks, labels = lbls) +
  scale_fill_brewer(palette = "Dark2")+
  theme(legend.position="bottom") +
  labs(x="Cities",
       y="Total Students",
       title="Students by Gender Across Cities",
       subtitle = "Female vs Male Across Cities, All Years",
       caption = "Source: istatistik.yok.gov.tr/")
```


Students by Gender Across Cities

Female vs Male Across Cities, All Years

Source: istatistik.yok.gov.tr/

- The funnel graph shows us that male student population in universities is more than female population for many cities in Turkey. (The disparity even affects the x-axis labels.)
- If we take into account the fact that majority of universities in Turkey are state universities and they have more male students than the female ones, we see that the results of the scatter plot and funnel graph coincide with each other.

3. The Cities with the Highest Female / Male Ratio

- In addition to funnel analysis, I wonder about whether there are cities which have more female students.
- The result is a little bit interesting for me since there is no city in Aegean Region with female/male student ratio more than 1.
- As you see below, majority of top 10 cities with the highest female/male student ratio are, counterintuitively, are not from Turkey's most developed regions or cities. If you ask anybody in the street "Which city has the highest female/male university student ratio?", I bet nobody would say Kilis or Erzurum. :)

```
ratio_df <- student_numbers %>%
          select(city, male,female) %>%
          group_by(city) %>%
          summarise(male=sum(male),female=sum(female)) %>%
          mutate(ratio = round((female / male),2)) %>%
          arrange(desc(ratio)) %>%
          slice(1:10)

ratio_df
```

```
## # A tibble: 10 x 4
##    city       male female ratio
##    <chr>     <int>  <int> <dbl>
##  1 KİLİS     14496  16515 1.14
##  2 ERZURUM  349282 392745 1.12
##  3 BOLU      52447  55035 1.05
##  4 KARAMAN   23641  24506 1.04
##  5 ANKARA   568831 570719 1.00
##  6 GİRESUN   52138  52199 1.00
##  7 NEVŞEHİR  35181  34601 0.980
##  8 SAMSUN   102504  99105 0.970
##  9 AYDIN     91827  88215 0.960
## 10 KIRŞEHİR  36400  34983 0.960
```
