ŞULE KÜTÜKDE, BDA 503, FINAL

JAN, 2018

PART 1

Question1:

In my opinion, Hadley Wickham was right about most of his ideas, however R-junkies showed that ggplot **can** do two y-axis graph. I think two y-axis is an unnecessary graph type. Personally, I wouldn't use it, because my managers wouldn't understand it. Also it can't always demonstrate impact of two variables correctly. For instance, crushes' scales are way smaller than departures' scale. If it is just a demonstration without aiming to show correlation between two variables, therefore it can be used. However if the main goal is to show the correlation of two variables, I think it can be deceptive commentators idea.

If I were the one who are trying to graph this data, I would arrange the data by the years ascendingly, then I would demonstrate crushes in y axis, departures in x axis. Finally I would show years as labels for the others to comprehend easily.

Question2:

If I were given a dataset to analyze, firstly I would start to get to know the dataset, its variables and the meaning of variables, and its structure. Then I would continue to get familiar with the data by looking the characteristic of variables, by searching what are minimum and maximum values if they are numeric, by calculating some quick statistical measures of variables like mean, median or mode.

If I were given the task to distribute funds from donations to public welfare projects, firstly I would identify all the measures like poverty, gender inequality, health, education etc. Later I would proceed to choose a decision making algorithm in order to decide objectively. This decision making algorithm would decide instead of me, and if there are others, instead of us. In this way, I would put my or others' social prejudices into the background to decide impartially.

Question3:

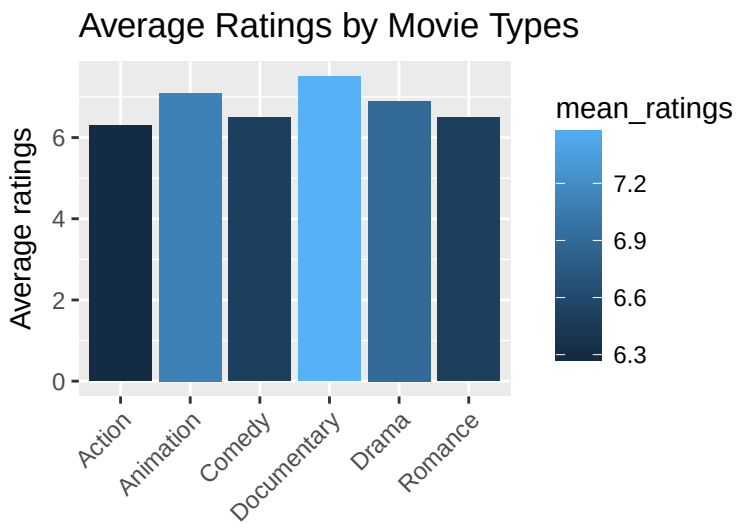
As a data analyst, my hardest work was on a time containing dataset.As an experience, time series datasets are very tough to handle it. For example, the diamond datasets price value can be predicted by simply with other variables. But the bitcoin's future price is way harder to predict. Sometimes it's unpredictible to even a prediction model. The error between prediction and actual value is higher then the other datasets.

Question4:

My single graph and the codes are down below:

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

```
movies[,7:17]<- list(NULL)
movies1<-gather(movies,"type","tt",-title,-year,-budget,-votes,-rating,-length)
movies1<-movies1[ grep("0", movies1$tt, invert = TRUE) , ]
movies1[,8]<- list(NULL)
movies2<-movies1%>%
  filter(budget>300000,votes>5000,length>80,year>1980)%>%
  group_by(type)%>%
  summarise(mean_ratings=round(sum(rating)/n(),1))
ggplot(movies2,  aes(x = type, y = mean_ratings,fill=mean_ratings)) +
  geom_bar(stat = "identity") +
  theme(axis.text.x=element_text(angle=45,hjust=1,vjust=1))+
  labs(y= "Average ratings", x=" ",title="Average Ratings by Movie Types")
```



I chose to demonstrate this graph because it includes all data's summary that i can understand in a first look. I considered all the meaningful columns. It is the best summary of the data that i can do. First I gathered the meaningful columns. Then I summarised it. And the filter part can be anytime changed.

So, the movies which are filmed after 1980, which have the budget over than 300,000$, which lenght are more than 80 minutes and which are voted more than 5000 times average ratings are demonstrated below. According to this graph, the most beloved movie type is Documentary.

PART 2

In this part, I decided to extend the group project with a prediction analysis. In the data, there were lots of future years values like infant mortality, population and life expectancy. Also there were no explanations about how this future values predicted and which prediction method is used. That's why I wrote some codes to predict future life_expectancy values of countries, calculated the mean of absolute error, compared with the original data via a graph. The codes and the graphs are down below:

```
set.seed(100)
mle <- read.csv("mortality_life_expectancy.csv")
mle<-mle%>%select(year,country_name,life_expectancy)%>%arrange(year)%>%
  mutate(train_test=ifelse(year>2017,"test","train"))
mle_train <- mle %>% filter(train_test == "train") %>% select(-train_test)
mle_test <- mle %>% filter(train_test == "test") %>% select(-train_test)
model <- lm(life_expectancy ~ ., data=mle_train)
mle_p <- predict(model)
predictionvsactual <-cbind(mle_p %>% tbl_df,mle_train %>% tbl_df) %>%
  mutate(error=abs(life_expectancy-value))
print(mean(predictionvsactual$error))
```

```
## [1] 1.54446
```

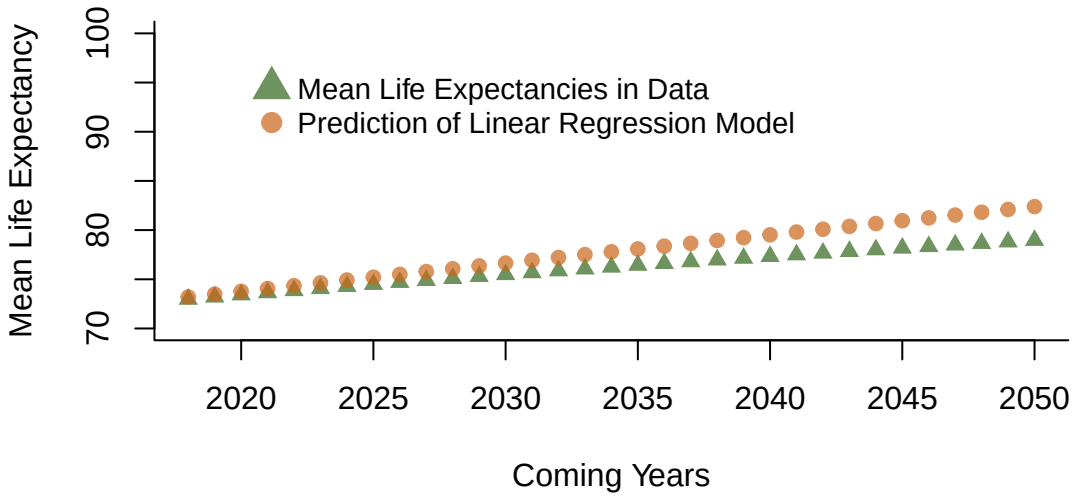
The train data's mean absolute error is 1.54. It can be predicted future years value with the same model.

```
mle_t<-predict(model,newdata = mle_test)
testprediction <-cbind(mle_t %>% tbl_df, mle_test %>% tbl_df) %>%
  mutate(error=abs(life_expectancy-value))
print(mean(testprediction$error))
```

```
## [1] 3.46377
```

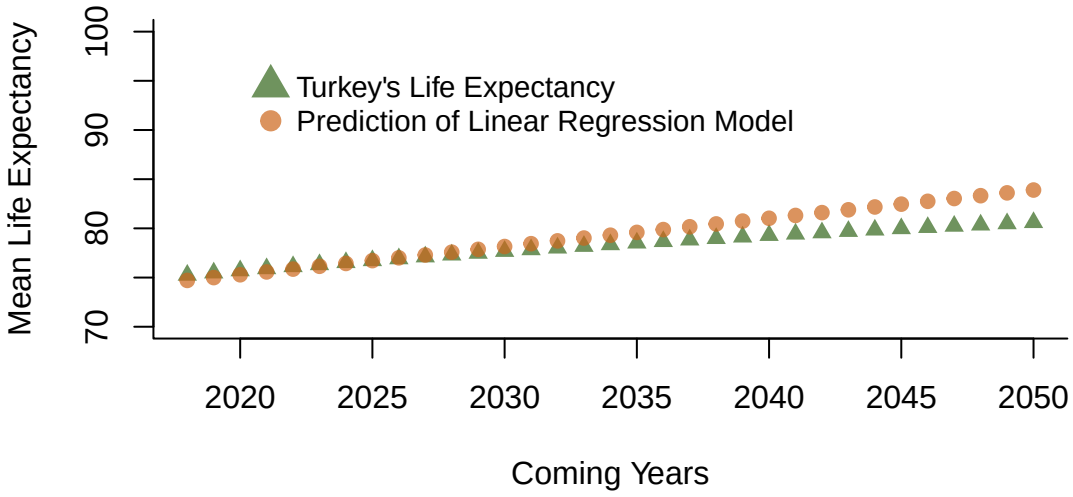
The test data's mean absolute error is 3.46 as well. Now it can be compared original data and predicted data with linear regression model.

```
testprediction1<-testprediction%>%
  group_by(year)%>%
  summarise(mean_error=mean(error),mean_lifex=mean(life_expectancy),mean_value=mean(value))%>%
  arrange(year)
plot(testprediction1$mean_lifex ~ testprediction1$year, type="b", bty="l",xlab="Coming Years",
     ylab="Mean Life Expectancy",col=rgb(0.2,0.4,0.1,0.7) , lwd=3 , pch=17 ,ylim=c(70,100) )
lines(testprediction1$mean_value ~testprediction1$year , type="b",col=rgb(0.8,0.4,0.1,0.7) ,
      lwd=3 , pch=20  )
legend("topleft",legend=c("Mean Life Expectancies in Data","Prediction of Linear Regression Model"),
       col = c(rgb(0.2,0.4,0.1,0.7), rgb(0.8,0.4,0.1,0.7)),pch = c(17,20),
       bty = "n",pt.cex = 2,
       cex = 0.9, text.col = "black", horiz = F ,inset = c(0.1, 0.1))
```



This chart shows that linear regression model predict life expectancy more with an increasing scope, than the original data.

```
testprediction_turkey<-testprediction%>%
  filter(country_name=="Turkey")%>%arrange(year)
plot(testprediction_turkey$life_expectancy ~ testprediction_turkey$year, type="b", bty="l",
     xlab="Coming Years",ylab="Mean Life Expectancy",col=rgb(0.2,0.4,0.1,0.7) ,
     lwd=3 , pch=17 ,ylim=c(70,100) )
lines(testprediction_turkey$value ~testprediction_turkey$year , type="b",col=rgb(0.8,0.4,0.1,0.7) ,
      lwd=3 , pch=20  )
legend("topleft", legend = c("Turkey's Life Expectancy", "Prediction of Linear Regression Model"),
       col = c(rgb(0.2,0.4,0.1,0.7), rgb(0.8,0.4,0.1,0.7)),pch = c(17,20),bty = "n",pt.cex = 2,
       cex = 0.9, text.col = "black", horiz = F ,inset = c(0.1, 0.1))
```



Also the same model's Turkey's life expectancy value and predicted value is demonstrated with this graph as well.

This analysis can be improved by predicting the other variables, and by trying the other prediction models.

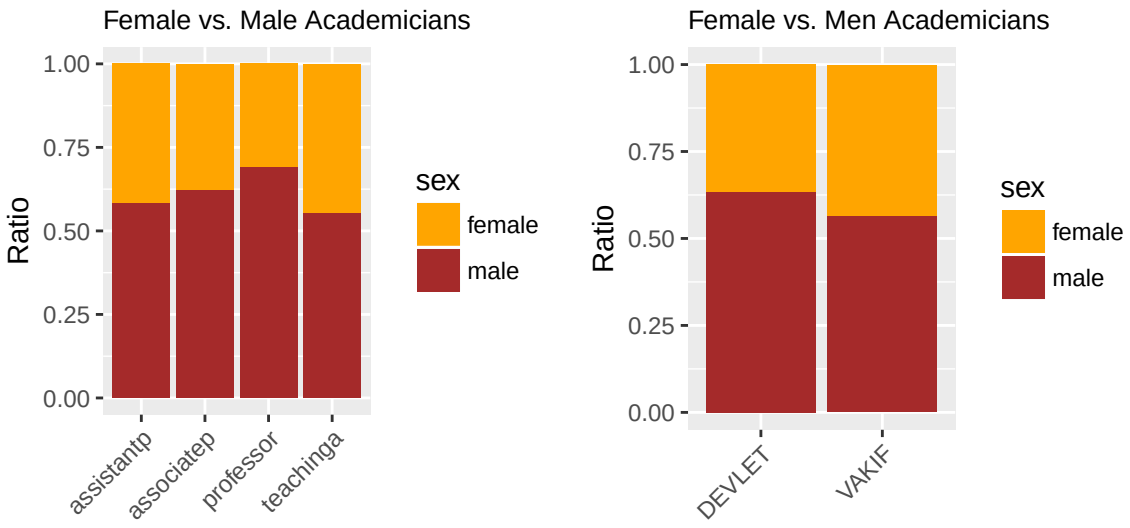
PART 3-Analysis of Female Academicians

Two dataset that I prepared to analysis can be downloaded here. ın the same working directory, codes below can be run correctly.

- First analysis is about female and male academicians ratio. The graph demonstrates that, at all academic level mens are dominant. However teaching asistants ratio is about the same. Maybe this can be interpreted that in the future years females are going to be getting dominant like males in academy. Second visualization demonstrate that foundation universities hire female academicians more than public universities.

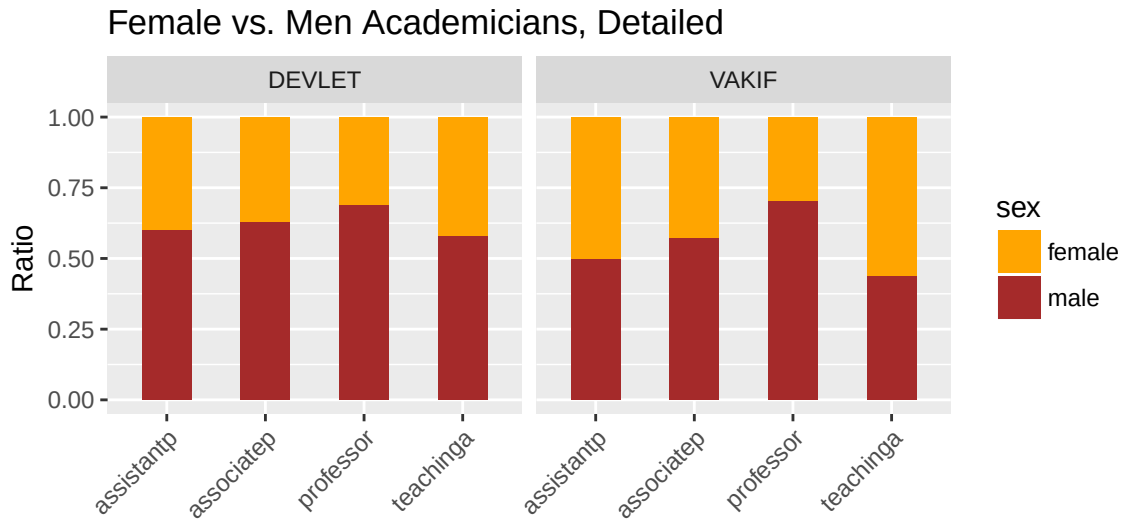
```
library(tidyverse)
library(readxl)
library(tidyr)
library(stringr)
```

```
ratio_sex<-jofa %>%
  group_by(academician,sex)%>%
  summarize(total=sum(number_of_academician))%>%
  select(academician,sex,total)
p1<-ggplot(ratio_sex, aes(fill=sex, y=total, x=academician)) +
  geom_bar( stat="identity", position="fill")+
  theme(axis.text.x=element_text(angle=45,hjust=1,vjust=1),plot.title = element_text(size=10))+
  labs(y= "Ratio" ,title="Female vs. Male Academicians",x="")+
  scale_fill_manual(values = c("Orange", "Brown"))
ratio_type1<-jofa %>%
  group_by(type,sex)%>%
  summarize(total=sum(number_of_academician))%>%
  select(type,sex,total)
p2<-ggplot(ratio_type1, aes(fill=sex, y=total, x=type)) +
  geom_bar( stat="identity", position="fill")+
  theme(axis.text.x=element_text(angle=45,hjust=1,vjust=1),plot.title = element_text(size=10))+
  labs(y= "Ratio" ,title="Female vs. Men Academicians",x="")+
  scale_fill_manual(values = c("Orange", "Brown"))
```



- This graph is a detailed version of second one.

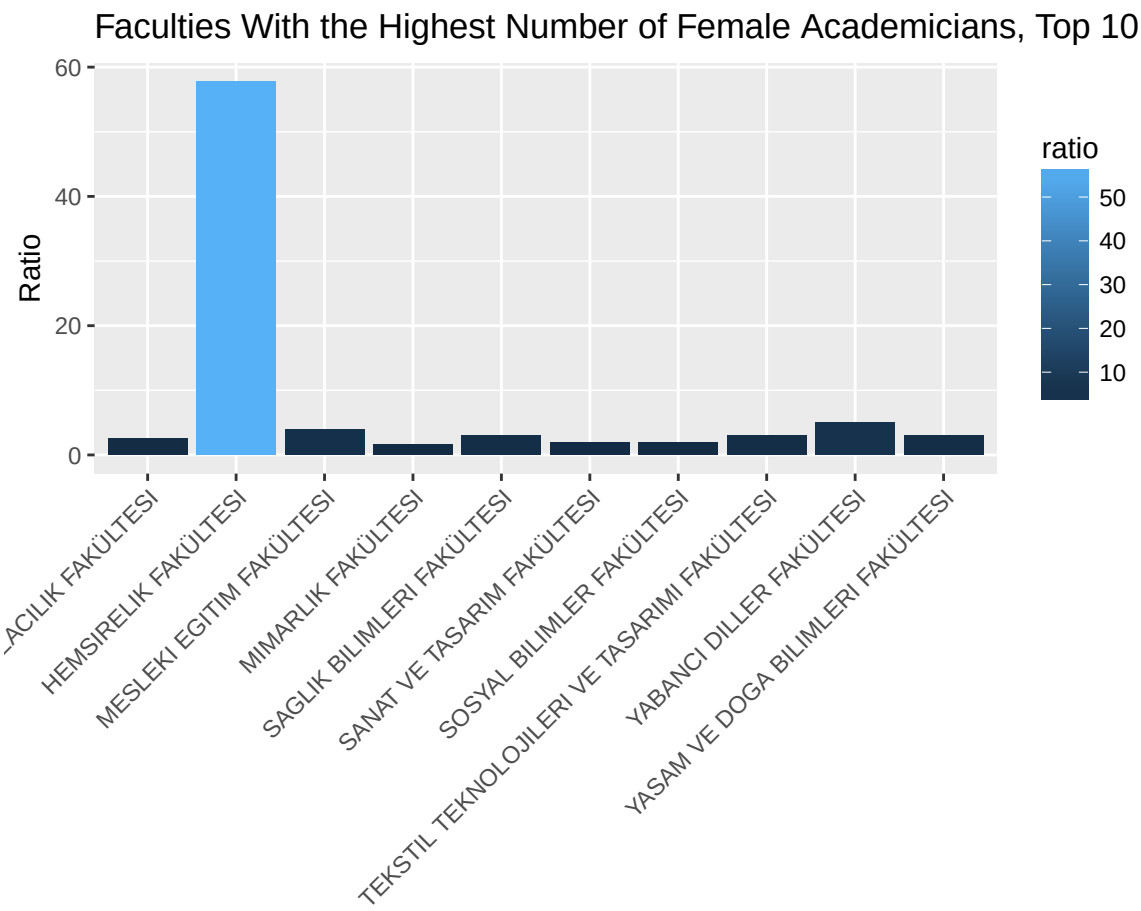
```
ratio_type2<-jofa %>%
  group_by(type,academician,sex)%>%
  summarize(total=sum(number_of_academician))%>%
  select(type,academician,sex,total)
ggplot(data = ratio_type2, aes(x = academician, y = total, group = sex, fill = sex))+
  geom_bar(stat = "identity", width = 0.5, position = "fill")+
  facet_grid(. ~ type)+
  theme(axis.text.x=element_text(angle=45,hjust=1,vjust=1))+
  labs(y= "Ratio" ,title="Female vs. Men Academicians, Detailed",x="")+
  scale_fill_manual(values = c("Orange", "Brown"))
```



- Finally this graph indicates that which faculties hire female academicians most. Top 10 faculties are showned. Nursing Faculty is well ahead then the others.

```
jofas1<-jofa%>%
  group_by(faculty,sex)%>%
  summarize(total=sum(number_of_academician))%>%
  select(faculty,sex,total)
jofas2<-spread(jofas1,sex,total)
top10faculty<-jofas2 %>%
  group_by(faculty)%>%
```

```
  summarize(totalf=sum(female),totalm=sum(male))%>%
  mutate(ratio=totalf/totalm)%>%
  select(faculty,ratio)%>%
  arrange(desc(ratio))
top10faculty <- top10faculty[is.finite(top10faculty$ratio),]%>%top_n(10)
ggplot(top10faculty,  aes(x = faculty, y = ratio,fill=ratio)) +
  geom_bar(stat = "identity") +
  theme(axis.text.x=element_text(angle=45,hjust=1,vjust=1))+
  labs(y= "Ratio", x=" ", title="Faculties With the Highest Number of Female Academicians, Top 10")
```



REFERENCES

1. Course lecture notes
2. For the graphs
3. Divine Code Source
4. For RMarkdown