BDA503 - Final

Tugba Unal

10-01-2021

Part 1: Short and Simple

1.1 Briefly describe the controversy around Timnit Gebru, her teams's recent research and Google. What are the valid points of each side? What is your position?

Frankly, I think Timnit Gebru was being wronged and Google supported this idea by not making a satisfactory statement on the issue. I also appreciated the fact that his team showed support for Timnit Gebru without fear of being fired. After all, I wonder how Google will teach ethical values to the artificial intelligence they created without having ethical values.

1.2 How much of a decision be based on gut feeling and experience and how much of it should be based on data and forecasts? For example, suppose you are a seasoned portfolio manager investing in only the stock market. How would you build your portfolio?

Although most of the decisions are based on analysis and concrete data, they should also be supported by abstract and instinctive findings that come with experience. If I were an experienced portfolio manager, I would first decide how many different sectors to include in my portfolio in order to reduce unsystematic risk. In other words, I would start to create a portfolio with the logic of putting the eggs that everyone knows about in a different basket.Then I would compare the volatility of the companies by analyzing the historical data of them, considering that the return on the portfolio should be higher than the deposit or risk-free return, but not show high volatility, and I decide on companies with high returns but low volatility.

1.3 If you had to plot a single graph using the nottem data (provided with base R) what would it be? Why? Make your argument, actually code the plot and provide the output. (You can find detailed info about the data set in its help file. Use ?nottem.)

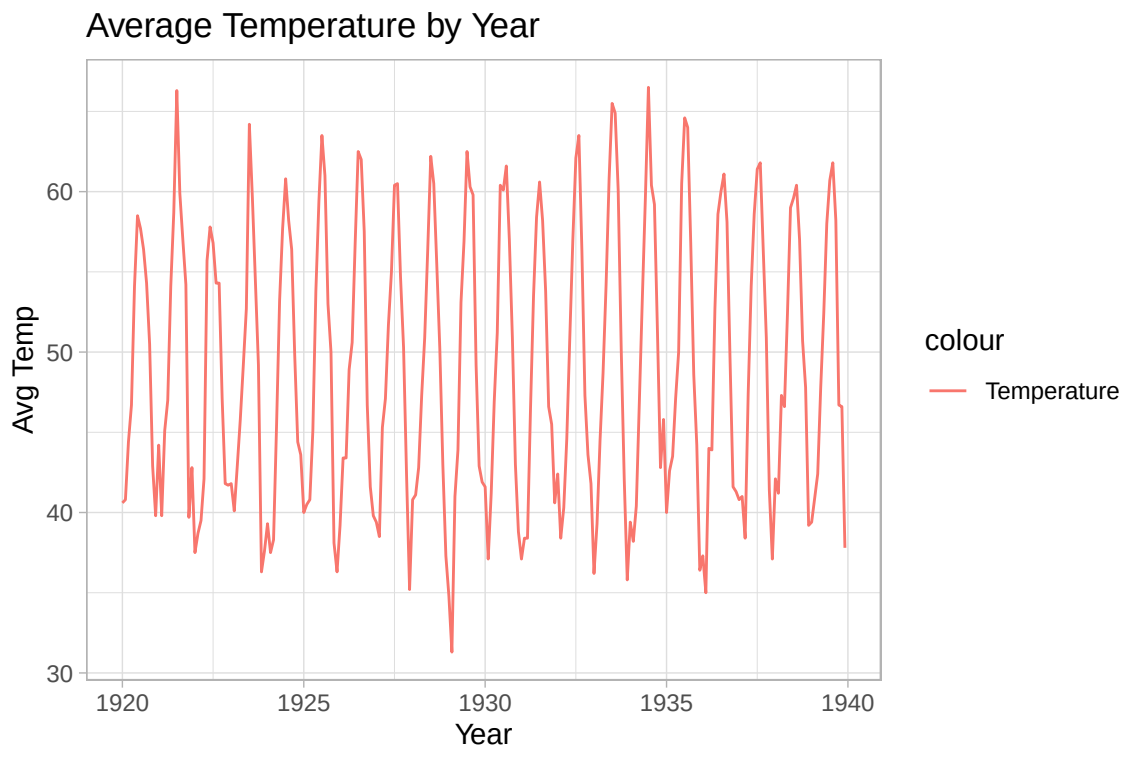
```
print(nottem)
```

```
##       Jan  Feb  Mar  Apr  May  Jun  Jul  Aug  Sep  Oct  Nov  Dec
## 1920 40.6 40.8 44.4 46.7 54.1 58.5 57.7 56.4 54.3 50.5 42.9 39.8
## 1921 44.2 39.8 45.1 47.0 54.1 58.7 66.3 59.9 57.0 54.2 39.7 42.8
## 1922 37.5 38.7 39.5 42.1 55.7 57.8 56.8 54.3 54.3 47.1 41.8 41.7
## 1923 41.8 40.1 42.9 45.8 49.2 52.7 64.2 59.6 54.4 49.2 36.3 37.6
## 1924 39.3 37.5 38.3 45.5 53.2 57.7 60.8 58.2 56.4 49.8 44.4 43.6
## 1925 40.0 40.5 40.8 45.1 53.8 59.4 63.5 61.0 53.0 50.0 38.1 36.3
## 1926 39.2 43.4 43.4 48.9 50.6 56.8 62.5 62.0 57.5 46.7 41.6 39.8
## 1927 39.4 38.5 45.3 47.1 51.7 55.0 60.4 60.5 54.7 50.3 42.3 35.2
## 1928 40.8 41.1 42.8 47.3 50.9 56.4 62.2 60.5 55.4 50.2 43.0 37.3
```

```
## 1929 34.8 31.3 41.0 43.9 53.1 56.9 62.5 60.3 59.8 49.2 42.9 41.9
## 1930 41.6 37.1 41.2 46.9 51.2 60.4 60.1 61.6 57.0 50.9 43.0 38.8
## 1931 37.1 38.4 38.4 46.5 53.5 58.4 60.6 58.2 53.8 46.6 45.5 40.6
## 1932 42.4 38.4 40.3 44.6 50.9 57.0 62.1 63.5 56.3 47.3 43.6 41.8
## 1933 36.2 39.3 44.5 48.7 54.2 60.8 65.5 64.9 60.1 50.2 42.1 35.8
## 1934 39.4 38.2 40.4 46.9 53.4 59.6 66.5 60.4 59.2 51.2 42.8 45.8
## 1935 40.0 42.6 43.5 47.1 50.0 60.5 64.6 64.0 56.8 48.6 44.2 36.4
## 1936 37.3 35.0 44.0 43.9 52.7 58.6 60.0 61.1 58.1 49.6 41.6 41.3
## 1937 40.8 41.0 38.4 47.4 54.1 58.6 61.4 61.8 56.3 50.9 41.4 37.1
## 1938 42.1 41.2 47.3 46.6 52.4 59.0 59.6 60.4 57.0 50.7 47.8 39.2
## 1939 39.4 40.9 42.4 47.8 52.4 58.0 60.7 61.8 58.2 46.7 46.6 37.8
```

As you can see, Nottem data is a time series with monthly breakdown. The best graphical representation is a line graph, as time series show how the changes in the data are over the period of interest, so I grouped the data on a yearly basis and created a line graph showing the average temperature values. As can be seen from the graph below, there is a seasonal effect in the data. Therefore, the graph changes approximately symmetrically and the next period values can be easily estimated.

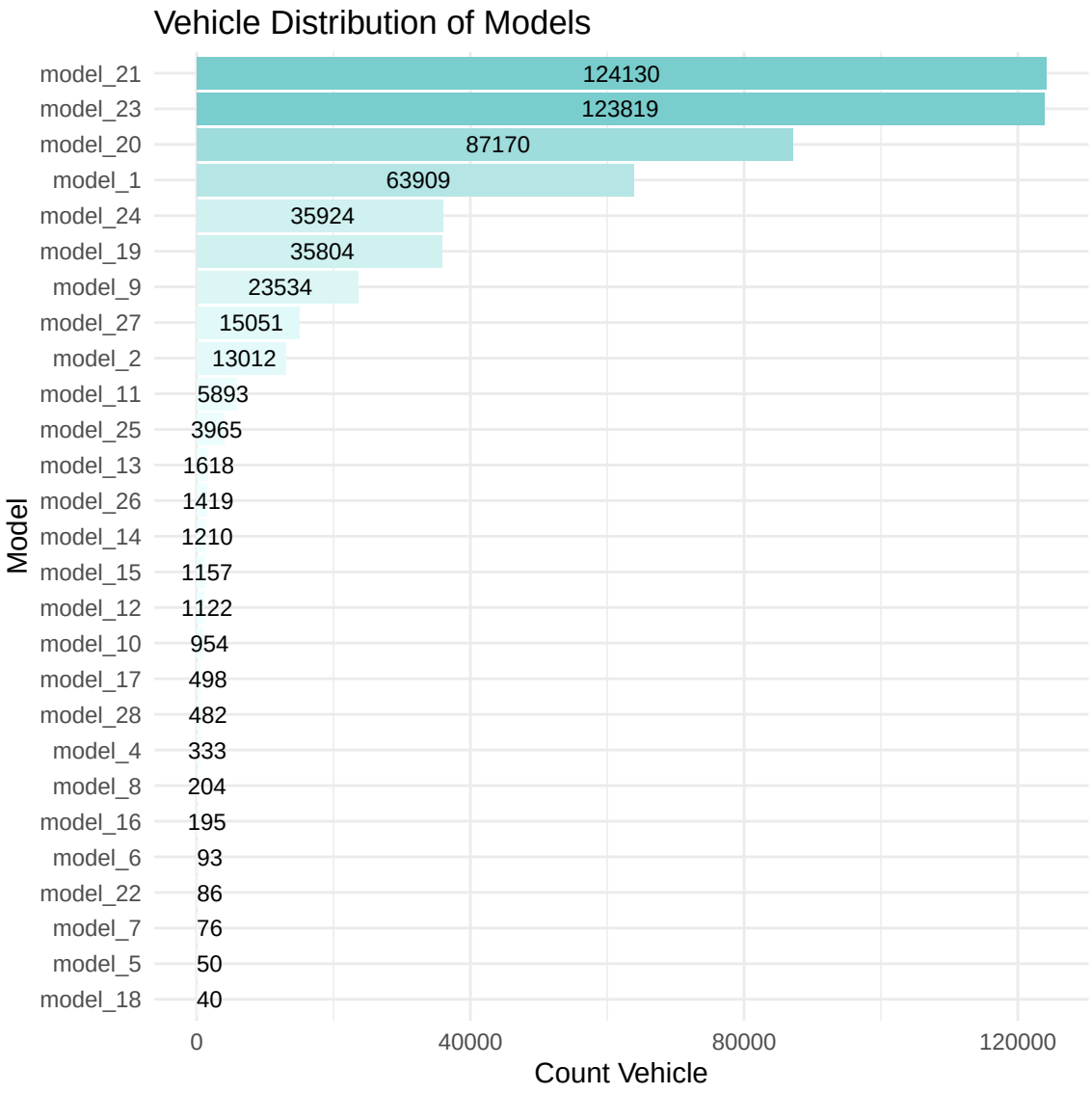
```
library(ggplot2)
library(reshape)

newnottem= data.frame(NottemDate= melt(time(nottem)), NottemTemp = melt(nottem))
colnames(newnottem) = c('NottemDate','Temp')


ggplot(newnottem, aes(x=NottemDate)) + geom_line(aes(y=Temp, color="Temperature")) +
  labs(x="Year", y="Avg Temp") + ggtitle("Average Temperature by Year") +
  theme_light()
```



Part 2: Extending My Group Project

```
library(dplyr)

raw_data <-
  rio::import("https://github.com/pjournal/mef04g-madagaskar/blob/gh-pages/Data/x_vehicle_company_servi
raw_dt1=na.omit(raw_data)

service_dt <- dplyr::rename(raw_dt1, "Material_id" = 1,
                  "Vehicle_id" = 2,
                  "Dealer_id" = 3,
                  "Job_order_number" = 4,
                  "material_type" = 5,
                  "process_type" = 6,
                  "beginning_date" = 7,
                  "ending_date" = 8,
                  "quantity" = 9,
                  "price" = 10,
                  "Model" = 11,
                  "production_date" = 12,
                  "job_closed_date" = 13,
                  "vehicle_km" = 14,
                  "dealer_city" = 15,
                  "warranty_beginning" = 16,
                  "warranty_ending" = 17)

Q <- quantile(service_dt$vehicle_km, probs=c(.25, .75), na.rm = FALSE)
iqr <- IQR(service_dt$vehicle_km)
up <-  Q[2]+1.5*iqr
low<- Q[1]-1.5*iqr

eliminated<-
  subset(service_dt, service_dt$vehicle_km> (Q[1] - 1.5*iqr) & service_dt$vehicle_km < (Q[2]+1.5*iqr))
```

```
data1<-eliminated%>%group_by(Model)%>%
  transmute(cnt_vehicle=n())%>%
  distinct(Model,cnt_vehicle)%>%arrange(desc(cnt_vehicle))

ggplot(data1, aes(x=cnt_vehicle, y=reorder(Model, cnt_vehicle), fill=cnt_vehicle)) +
  geom_col() + scale_fill_gradient("cnt_vehicle", low="azure", high="darkslategray3") +
  geom_text(aes(label = paste(format(cnt_vehicle))),
            size=3, position = position_stack(vjust = 0.5)) +
  theme_minimal() +
  theme(legend.position = "none", plot.title = element_text(vjust = 0.5)) +
  labs(x = "Count Vehicle",
       y = "Model",
       title = "Vehicle Distribution of Models")
```



The chart above shows the distribution of the vehicles in the data according to the models. It can be easily read from the graphic that there are most Model21 type vehicles among the vehicles coming to the service.

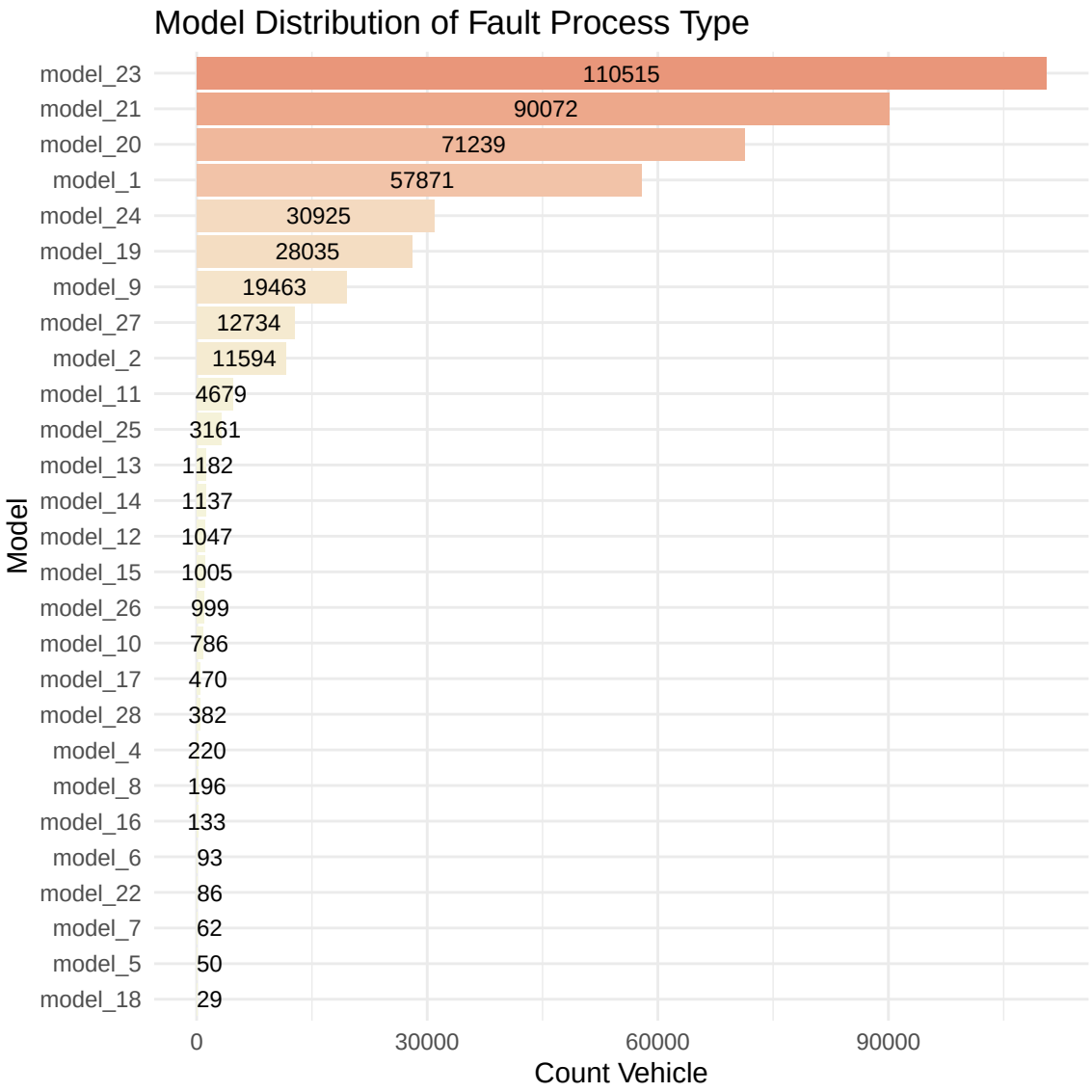
```
data2<-eliminated%>%group_by(process_type)%>%
  transmute(cnt_vehicle2=n())%>%
  distinct(process_type,cnt_vehicle2)%>%
  arrange(desc(cnt_vehicle2))

data2
```

```
## # A tibble: 11 x 2
## # Groups:   process_type [11]
##    process_type cnt_vehicle2
##    <chr>               <int>
##  1 Ariza              448165
##  2 P.Bakim             32106
##  3 Kontrol             24047
```

```
##  4 Kampanya            12413
##  5 Hasar               11652
##  6 Garanti              9183
##  7 Yol Yardım           2507
##  8 Dahili               1493
##  9 Aksesuar               91
## 10 PDI                    84
## 11 P.Bak+Ariza             7
```

The table above shows for which type of operation the vehicles coming to the service are more intense. As can be seen, vehicles have the most failure type records.

```
data3<-eliminated%>%filter(process_type=='Ariza')%>%
  group_by(Model)%>%transmute(cnt_vehicle3=n())%>%
  distinct(Model,cnt_vehicle3)%>%
  arrange(desc(cnt_vehicle3))

ggplot(data3, aes(x=cnt_vehicle3, y=reorder(Model, cnt_vehicle3), fill=cnt_vehicle3)) +
  geom_col() + scale_fill_gradient("cnt_vehicle", low="beige", high="darksalmon") +
  geom_text(aes(label = paste(format(cnt_vehicle3))),
            size=3, position = position_stack(vjust = 0.5)) +
  theme_minimal() +
  theme(legend.position = "none", plot.title = element_text(vjust = 0.5)) +
  labs(x = "Count Vehicle",
       y = "Model",
       title = "Model Distribution of Fault Process Type")
```



When I analyzed the data by filtering the fault type, which is the most recorded transaction type, I saw that there were more records in Model23 type vehicles while waiting for more records for the Model21 with the most vehicle records in all data. As a result, by making a general examination for Model23 type vehicles, the factors causing the malfunction can be found.

Part 3: Industry Stock Analysis

3.1 Gathering Data and Packages

I have used the current, target and forecast data of companies belonging to the industry sector in my analysis within Is Investment stock recommendations. If you want to access the data, you can access here

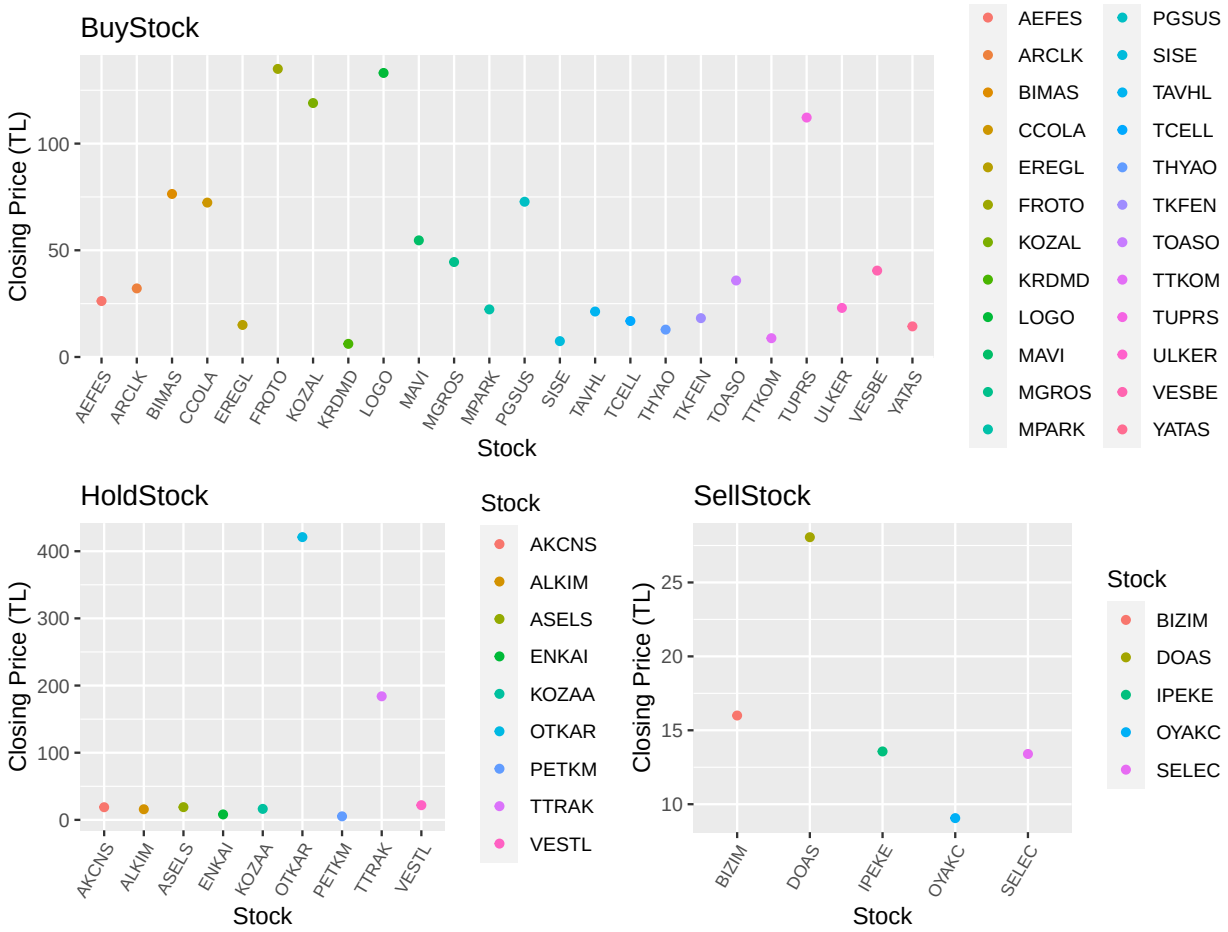
```
library(readxl)
library(tidyverse)
library(dplyr)
```

```
library(ggplot2)
library(gridExtra)
library(ggpubr)
```

```
#stocksummary<-
#  read_xlsx("C:/Users/unalt/Desktop/MEF/Data Analytics Essentials/Final/takipozet.xlsx")
#stockestimate<-
#  read_xlsx("C:/Users/unalt/Desktop/MEF/Data Analytics Essentials/Final/takiptahmin.xlsx")
#stocktarget<-
#  read_xlsx("C:/Users/unalt/Desktop/MEF/Data Analytics Essentials/Final/takiphedeffiyat.xlsx")

#summary_target<-left_join(stocksummary,stocktarget, by='Stock')
#fullstock<-left_join(summary_target,stockestimate, by='Stock')

#save(fullstock, file = "StockInfo.RData")

fullstockinfo <-
  rio::import("https://github.com/pjournal/mef04-unaltugba/blob/gh-pages/StockInfo.RData?raw=True")
```

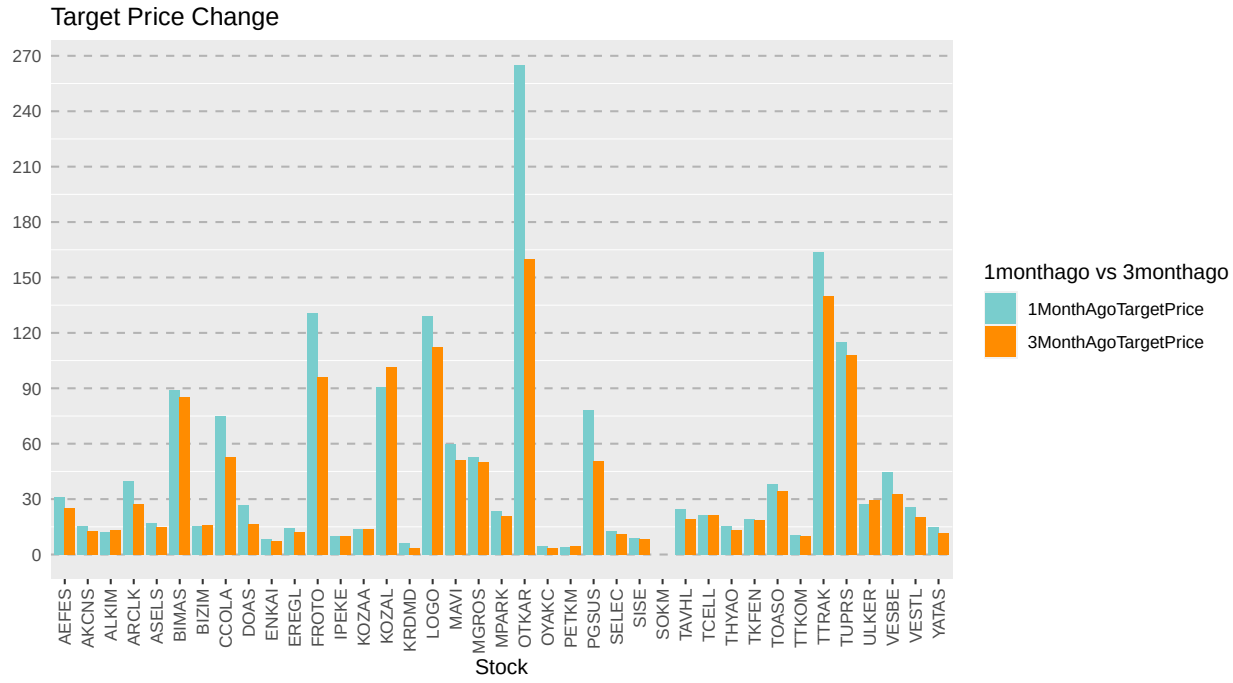
3.2 Exploratory Data Analysis

```
buystock<-fullstockinfo%>%filter(Recommendation=="AL")%>%select(Stock,`Close(TL)`)
holdstock<-fullstockinfo%>%filter(Recommendation=="TUT")%>%select(Stock,`Close(TL)`)
sellstock<-fullstockinfo%>%filter(Recommendation=="SAT")%>%select(Stock,`Close(TL)`)

p1<-qplot(Stock, `Close(TL)`, data=buystock, col=Stock) + ggtitle("BuyStock")+
  theme(axis.text.x = element_text(angle=60, size=8, vjust=1,hjust=1)) +
  labs(x="Stock", y="Closing Price (TL)")

p2<-qplot(Stock, `Close(TL)`, data=holdstock, col=Stock) + ggtitle("HoldStock")+
  theme(axis.text.x = element_text(angle=60, size=8, vjust=1,hjust=1)) +
  labs(x="Stock", y="Closing Price (TL)")

p3<-qplot(Stock, `Close(TL)`, data=sellstock, col=Stock) + ggtitle("SellStock")+
  theme(axis.text.x = element_text(angle=60, size=8, vjust=1,hjust=1)) +
  labs(x="Stock", y="Closing Price (TL)")

ggarrange(p1, ggarrange(p2,p3,ncol = 2),nrow=2)
```



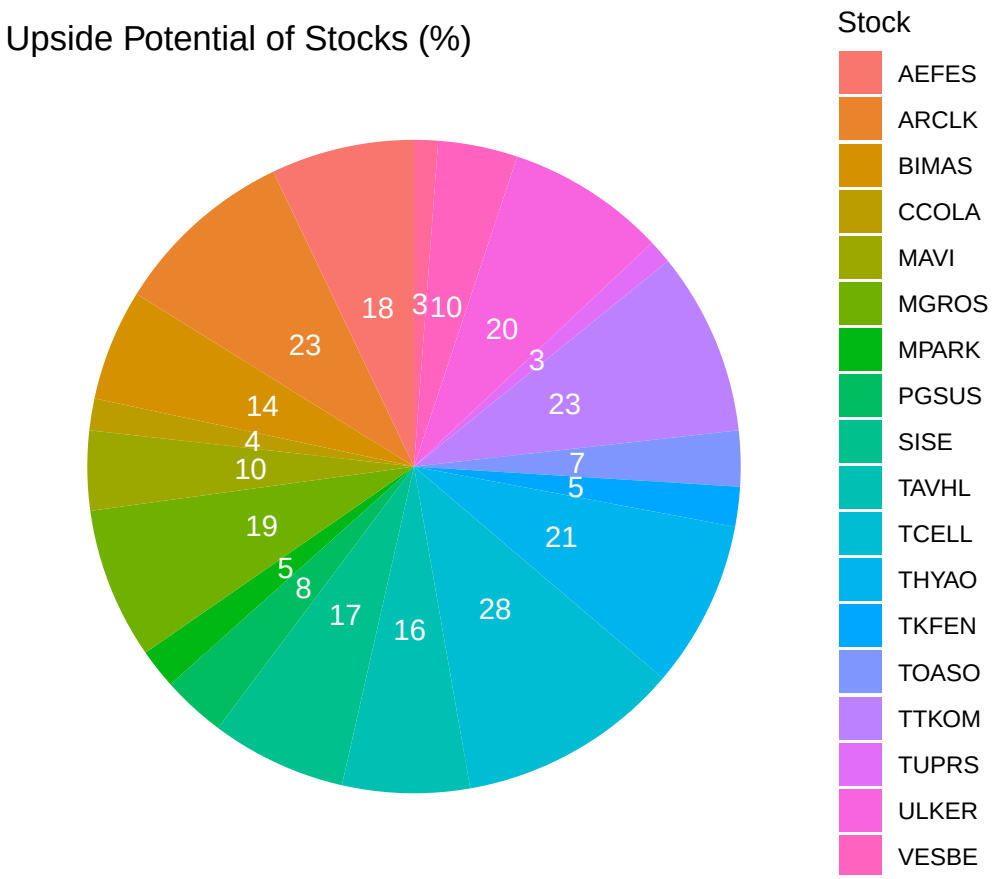
I grouped the companies according to recommendation and divided them into 3 groups as stocks to buy & sell & hold. It can be seen from the graphs above that companies with a "buy" recommendations are the majority. When we look at the prices, it is seen that the shares with the "hold" recommendations are the companies with the highest prices.

```
targetdata<-fullstockinfo%>%group_by(Stock)%>%
  select(Stock,`1MonthAgoTargetPrice`,`3MonthAgoTargetPrice`)%>%
  pivot_longer(cols = c(-Stock))

ggplot(targetdata, aes(x = Stock, y = value, fill= name)) +
  geom_bar(stat="identity", position = "dodge") +
  labs(x="Stock", y="Target Price (TL)") + ggtitle("Target Price Change") +
  scale_fill_manual("1monthago vs 3monthago", values=c("darkslategray3","darkorange")) +
  theme(axis.text.x = element_text(angle=90, size=9, vjust=0.5, hjust=1)) +
  theme_cleveland() + scale_y_continuous(breaks=seq(0,300,30))
```



The chart above shows how the target prices of companies changed over the 3-month period. It can be seen from the graph that the majority of the companies whose target price 1 month ago was revised upwards compared to the target price 3 months ago. The upward revision of the companies's target prices shows that they will perform better than previously expected.

```
upsidedata<-fullstockinfo%>%filter(Recommendation=="AL",`Upside(%)`>0)%>%
  arrange(desc(Stock))%>%mutate(ypos=cumsum(`Upside(%)`)-0.5*`Upside(%)`)

ggplot(upsidedata,aes(x="", y=`Upside(%)`, fill=Stock)) + geom_bar(stat="identity") +
  coord_polar("y") + ggtitle("Upside Potential of Stocks (%)") +
  theme(axis.title.y = element_blank(), legend.text=element_text(size=2)) +
  geom_text(aes(y = ypos,label = `Upside(%)`), color = "white") + theme_void()
```



The chart above shows the return potential (%) of companies that are given a "buy" recommendation and have a return expectation higher than 0%.

3.3 Conclusion

As a result of all these analyzes, if I wanted to create a portfolio, I would choose the company shares that were given a buy recommendation, high return potential and upward revision of the target price and operating in various sectors. Accordingly, TCELL, THYAO, ARCLK, SISE and BIMAS would be companies that I would add to my portfolio.