Data Visualzation: HW3

Yuan-Yao Chang

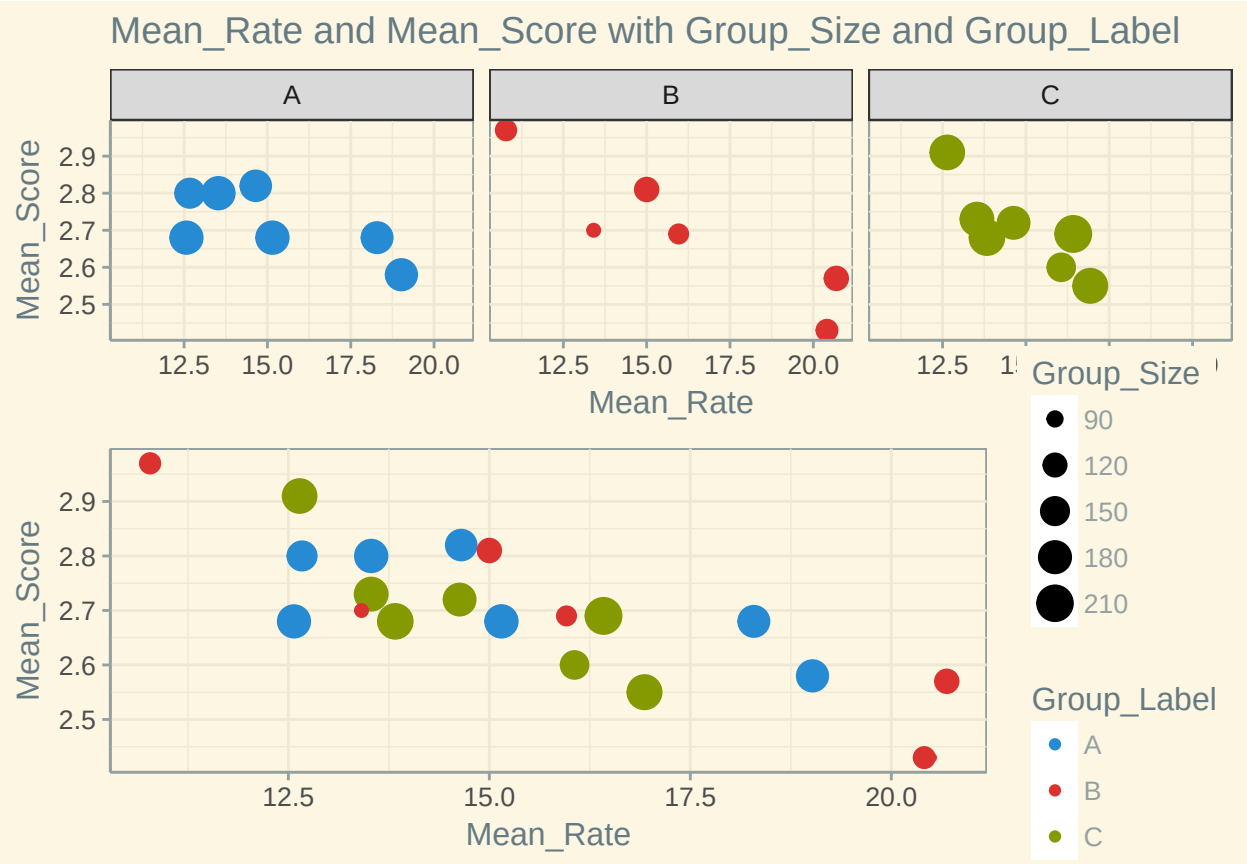
```
library(ggplot2)
library(reshape2)
library(ggthemes)
require(gridExtra)
```

```
## Loading required package: gridExtra
```

```
#set new label
new_label <- c("Date", "Mean_Rate", "Mean_Score", "Group_Size", "Group_Label")
data = read.csv('STK_Data_Gates_2017.csv')
names(data) <- new_label
#Date
new_Date_format<-as.Date((data$Date),"%m/%d/%Y")
data$Date <-strftime(new_Date_format,"%m/%d")
#stats 5 summary
print(summary(data))
```

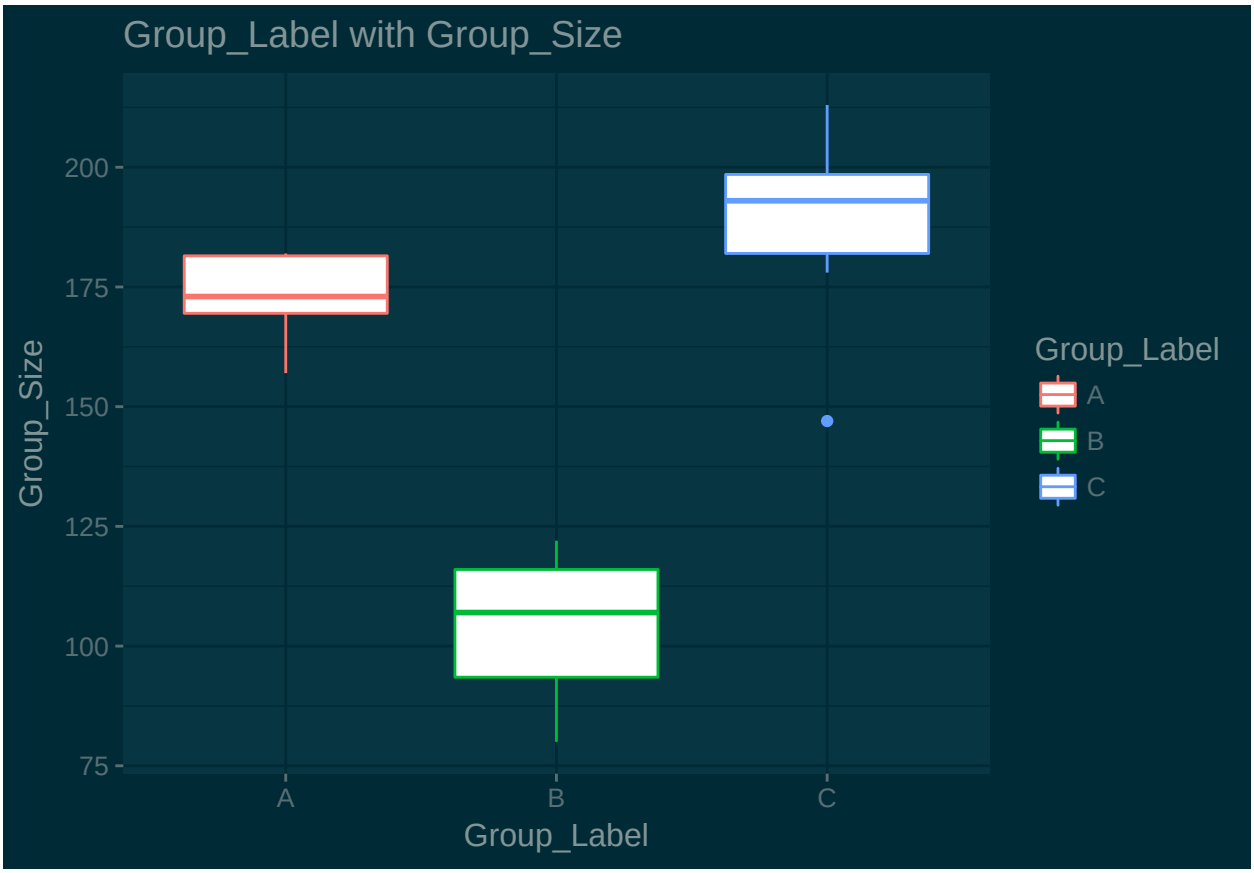
```
##      Date             Mean_Rate       Mean_Score      Group_Size
##  Length:21          Min.   :10.78   Min.   :2.430   Min.   : 80
##  Class :character   1st Qu.:13.53   1st Qu.:2.600   1st Qu.:122
##  Mode  :character   Median :15.00   Median :2.690   Median :171
##                     Mean   :15.56   Mean   :2.691   Mean   :155
##                     3rd Qu.:16.93   3rd Qu.:2.800   3rd Qu.:182
##                     Max.   :20.69   Max.   :2.970   Max.   :213
##  Group_Label
##  A:7
##  B:7
##  C:7
##
##
##
```

```
#different groups
g1<-ggplot(data, aes(Mean_Rate,Mean_Score,color = Group_Label,size = Group_Size)) + geom_point()  + ther
scale_color_solarized() + ggtitle("Mean_Rate and Mean_Score with Group_Size and Group_Label") + facet_gr
g2<-ggplot(data, aes(Mean_Rate,Mean_Score,color = Group_Label,size = Group_Size)) + geom_point()  + ther
scale_color_solarized()

grid.arrange(g1,g2,nrow=2,ncol=1)
```



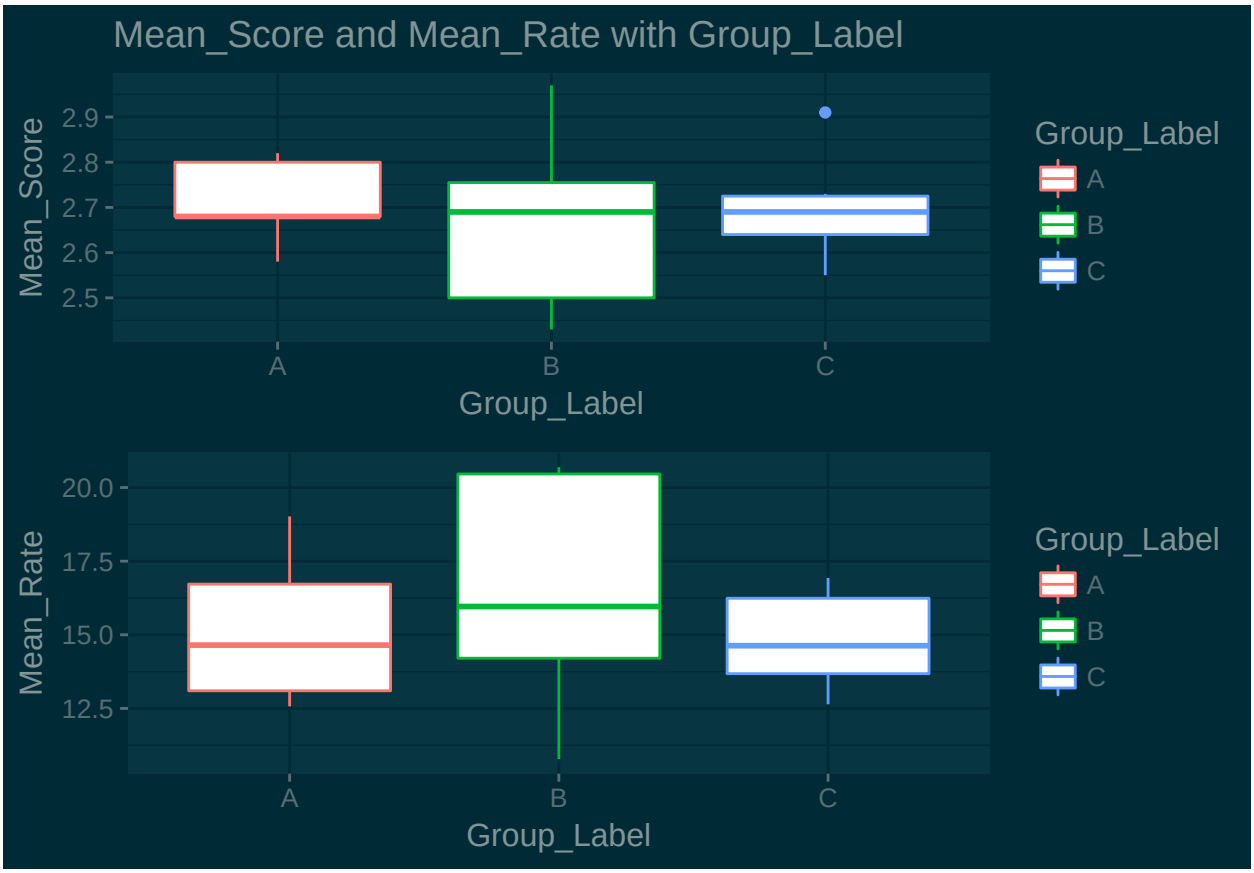
Mean_Score and Mean_Rate have a inverse relationship, while group B is more scatter than the other two groups. Also, group B size is relatively small than group A and B.

```
ggplot(data, aes(Group_Label,Group_Size,color=Group_Label)) + geom_boxplot() + theme_solarized_2(light =
```



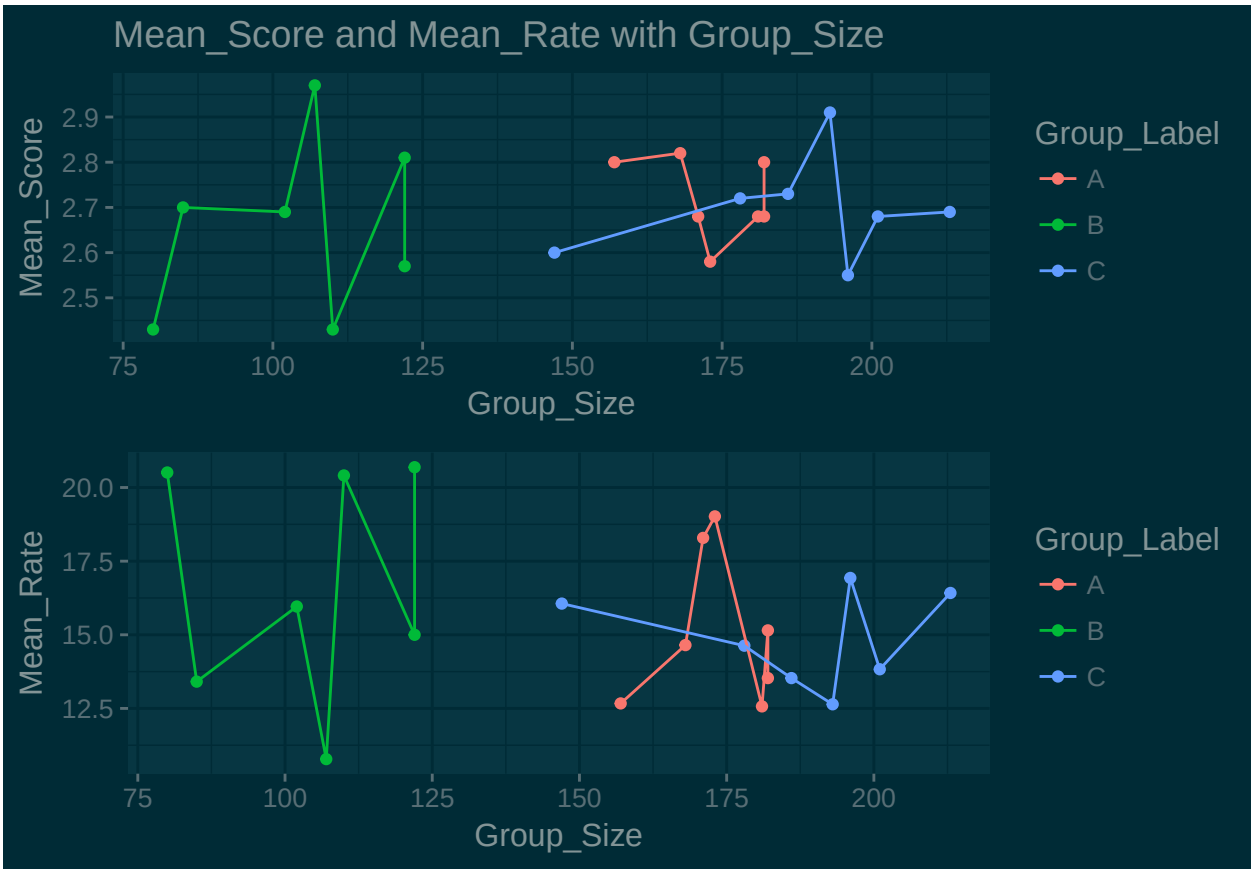
Group A's sizes are more concentrate and most of the size points focus on the smaller size point. Group B's size is much more smaller than the other two, and the size points are equally distributed. Group C's size is the largest, while owning an outliers and the size points focus on the bigger size point.

```
#subplots
#boxplot
g1<-ggplot(data, aes(Group_Label,Mean_Score,color = Group_Label)) + geom_boxplot() + theme_solarized_2(
g2<-ggplot(data, aes(Group_Label,Mean_Rate,color = Group_Label)) + geom_boxplot() + theme_solarized_2(li

grid.arrange(g1,g2,nrow=2,ncol=1)
```



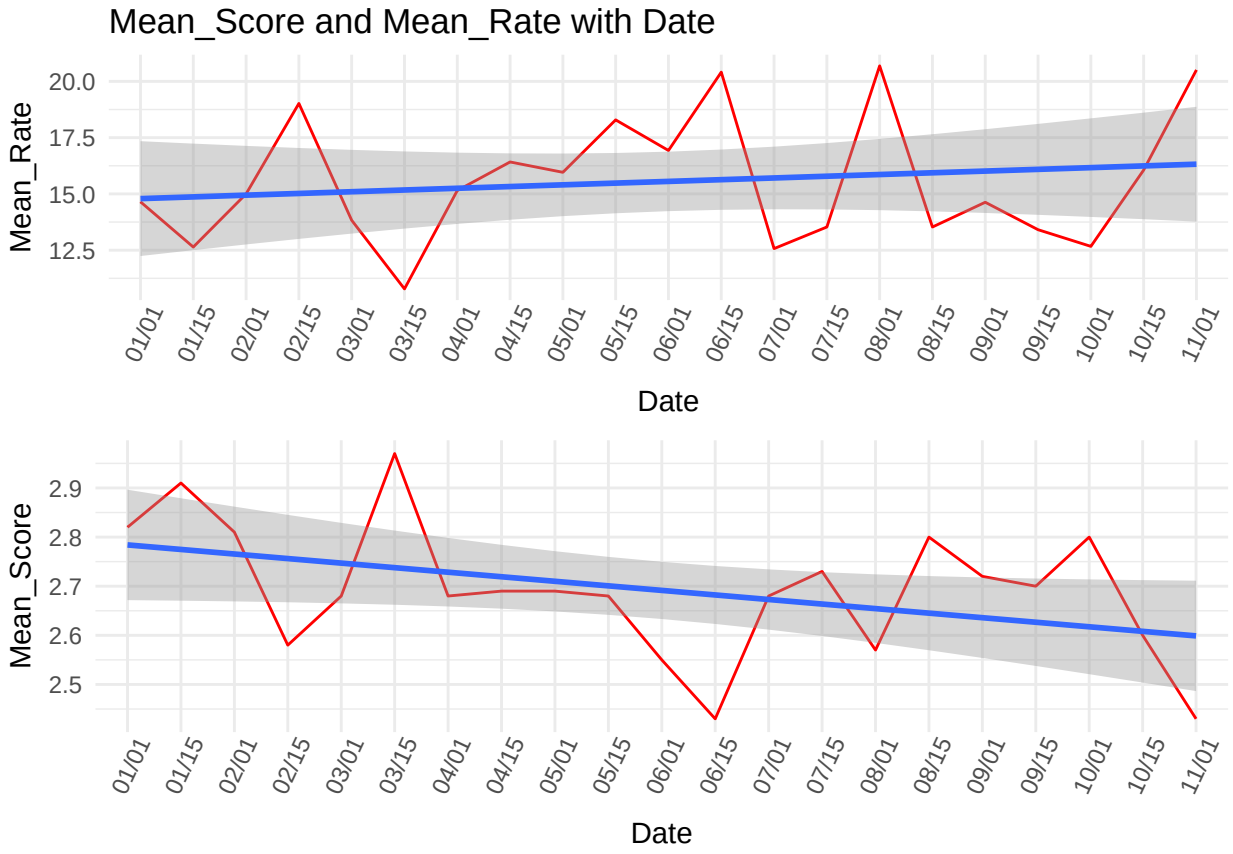
Group A's Mean_Score is more focus on the higher score, the median is almost the same as third quartile; the Mean_Rate is more equally distributed. Group B's Mean_Score and Mean_Rate own the same min and Max, while Mean_Score have more lower point and Mean_Rate owns more higher point. Group C's Mean_Score have an outlier and the rest of the data is more concentrate; the Group C's data slightly more focus on the lower points.

```
g1<-ggplot(data, aes(Group_Size,Mean_Score,color = Group_Label)) + geom_point() + geom_line() + theme_s
g2<-ggplot(data, aes(Group_Size,Mean_Rate,color = Group_Label)) + geom_point() + geom_line() + theme_so

grid.arrange(g1,g2,nrow=2,ncol=1)
```



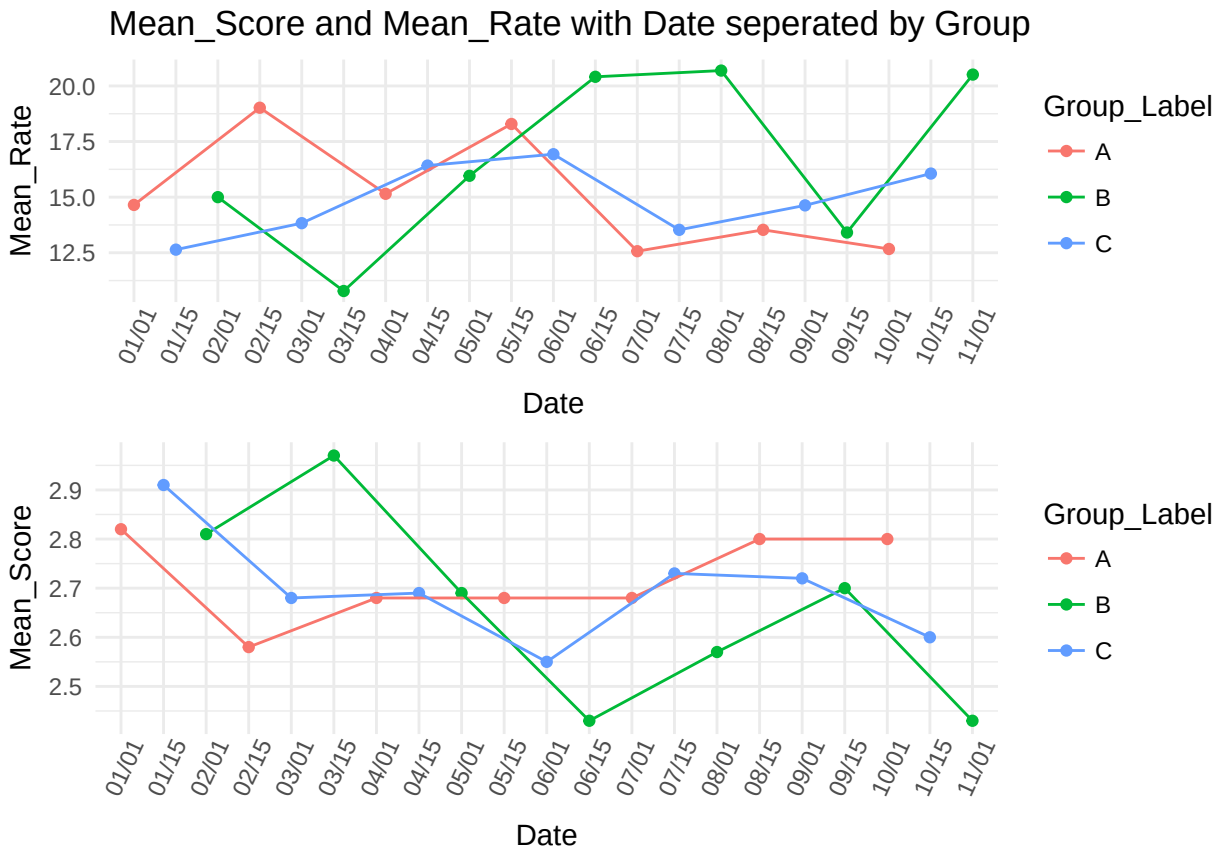
The Group_Size distribute the same as previously graph, the trend of the Mean_Rate and Mean_Score seems to be oppisite.

```
#subplot & timeline

g1<-ggplot(data, aes(Date,Mean_Rate,group=1)) + geom_line(color = 'red') + geom_smooth(method = 'lm') +
  theme_minimal() + theme(axis.text.x= element_text(angle = 65, vjust = 0.7)) + ggtitle("Mean_Score and
g2<-ggplot(data, aes(Date,Mean_Score,group=1)) + geom_line(color = 'red') + geom_smooth(method = 'lm') +
  theme_minimal() + theme(axis.text.x= element_text(angle = 65, vjust = 0.7))

grid.arrange(g1,g2,nrow=2,ncol=1)
```



If we analyze the overall linear regression model, the Mean_Score and Mean_Rate trend is totally opposite interval by every 15 days chronologically. Take a deeper look on the red line which represents the actaul data, the trend is more similar just the trend is opposite. the gray area stands for pointwise 95 percent confidence regions, the larger the area indicates the data is more closer to the linear regression trend, or we can say the range of standard error is smaller.

```
#timeline mean

g1 <- ggplot(data, aes(Date,Mean_Rate,group = Group_Label,color = Group_Label)) + geom_point() + geom_li
g2 <- ggplot(data, aes(Date,Mean_Score,group = Group_Label,color = Group_Label)) + geom_point() + geom_l

grid.arrange(g1,g2,nrow=2,ncol=1)
```



I seperate the Mean_Score and Mean_Rate into different groups, found that the Mean_Score of group A is more stable, not much alter. Both Mean_Score and Mean_Rate have a significant change in the period of 03/15 to 06/15.

```
#timeline group

g1 <- ggplot(data, aes(Date,Group_Label,color = Group_Label,group =1))  + geom_point() + geom_step() + s
g2 <- ggplot(data, aes(Date,Group_Size,group = Group_Label,color = Group_Label)) + geom_point() + geom_l

grid.arrange(g1,g2,nrow=2,ncol=1)
```

By analyzing the Group_Label and date graph, there is no bias in this dataset's sampling. As we can see the data sampling become smaller and smaller in overall view, there is an significant drop of the group B from 08/01 to 09/15.

Conclusion: By the above analyze, the change of Group B's Mean_Score base on time is stronger than other two groups; Group B and C shows an decrease trend while Group A is increasing. As for Mean_Rate is the opposite trend.

Nature of the Data?

The dataset is the measurement of 3 diffent group sources with different size group. Based on my opinion, the data could be the three different contractor bidding the contract of the building's remodeling. While Group_Size as the bids.