CSC 465

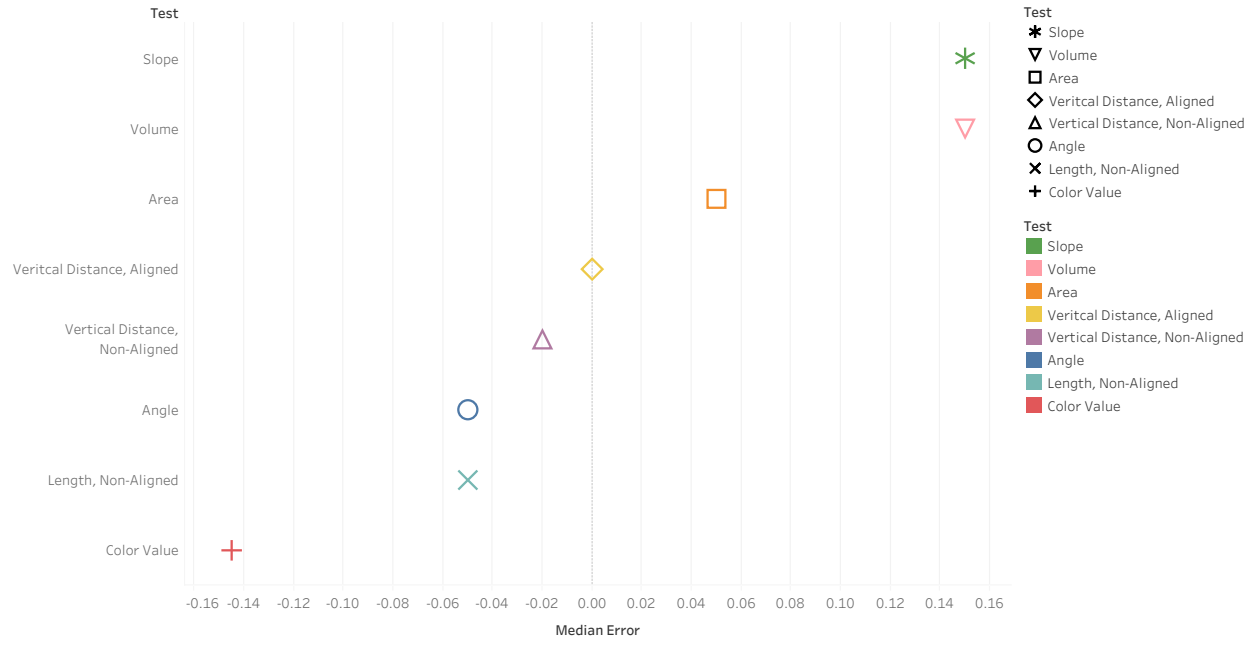
Homework 2

Chaonan Shi
1901412

Problem 3:

a. From Tableau:

data



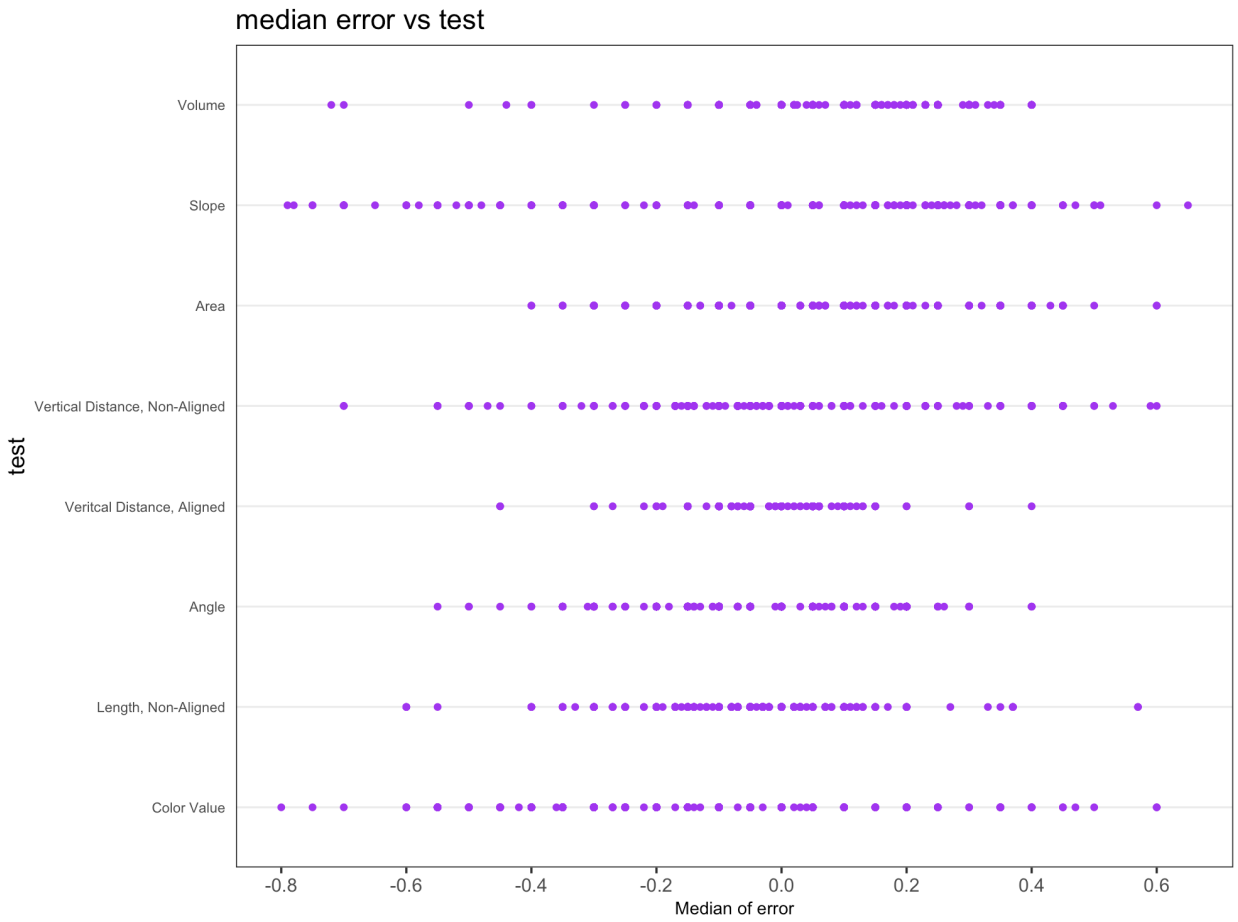
Median of Error for each Test. Color shows details about Test. Shape shows details about Test.

For Tableau, I I choose color of the background as white, and categories as different colors and shapes;

In this case, I sort the test by **descending,** from -0.16 to 0.16 (+/- 1) scaled by 0.02;

It makes sense to the audience that which test categories have higher error, and color emphasized positive and negative values and shapes make categories easily to differential.

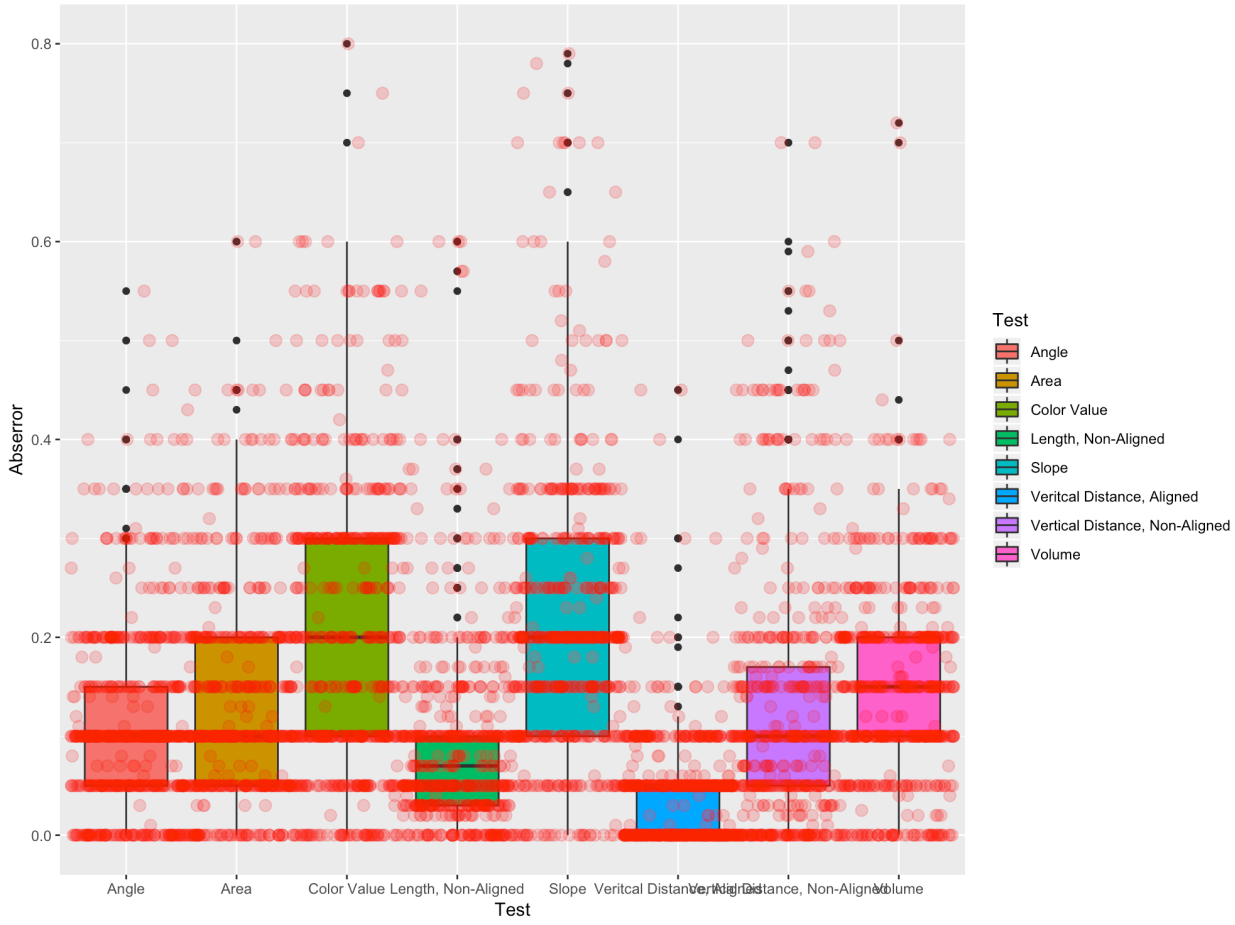
For R(failed)



(It seems like I messed up the error, so that the all plot distributed among x-axe, complete it by tableau, will try it again in R later…)

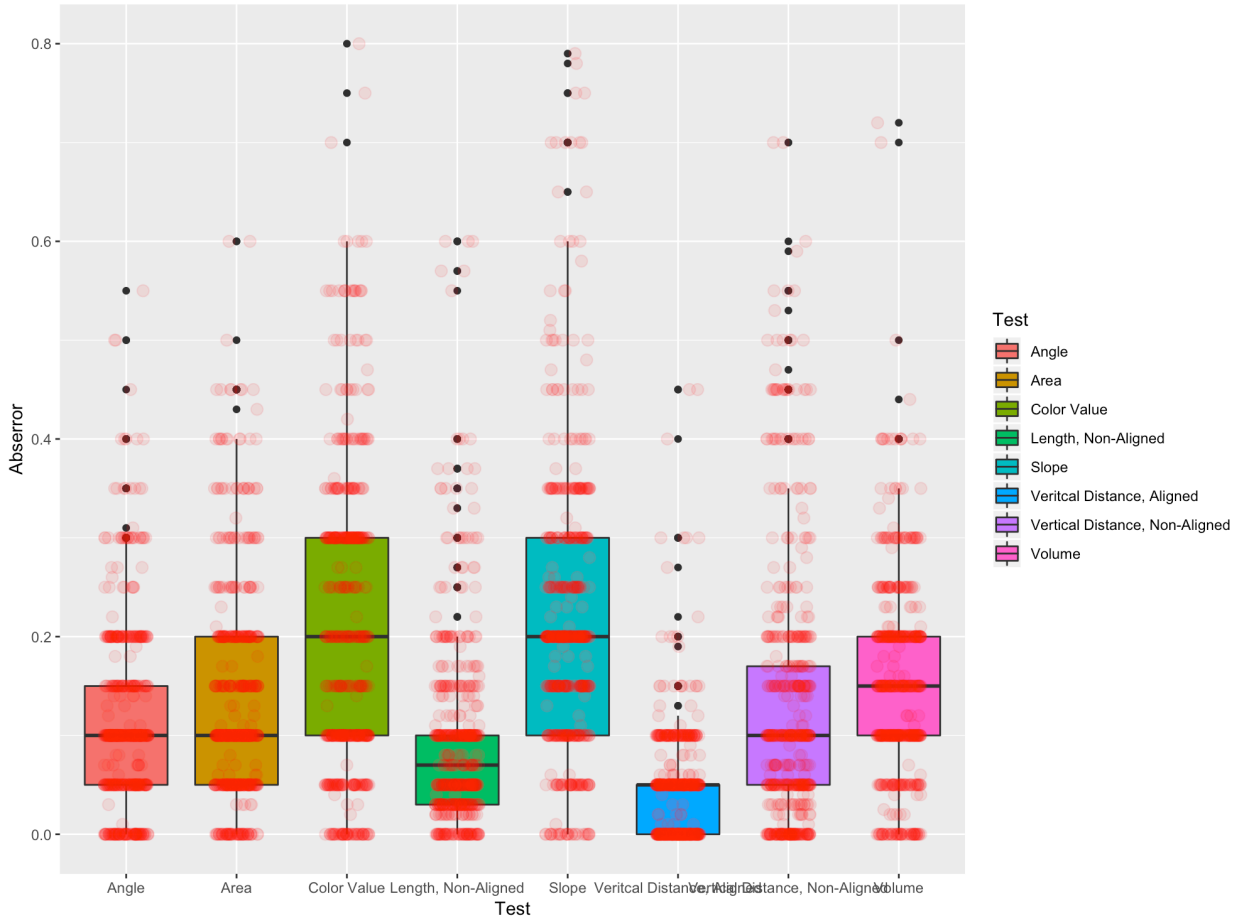
b. From R:

Jitter without distribution:



In this case, I use default setting to visualize the clusters and and data. However, it obviously messed up since the width and the transparency. So I decide to change the 'width' and 'alpha' in this case, also adding the distribution to the jitter, see next graph;

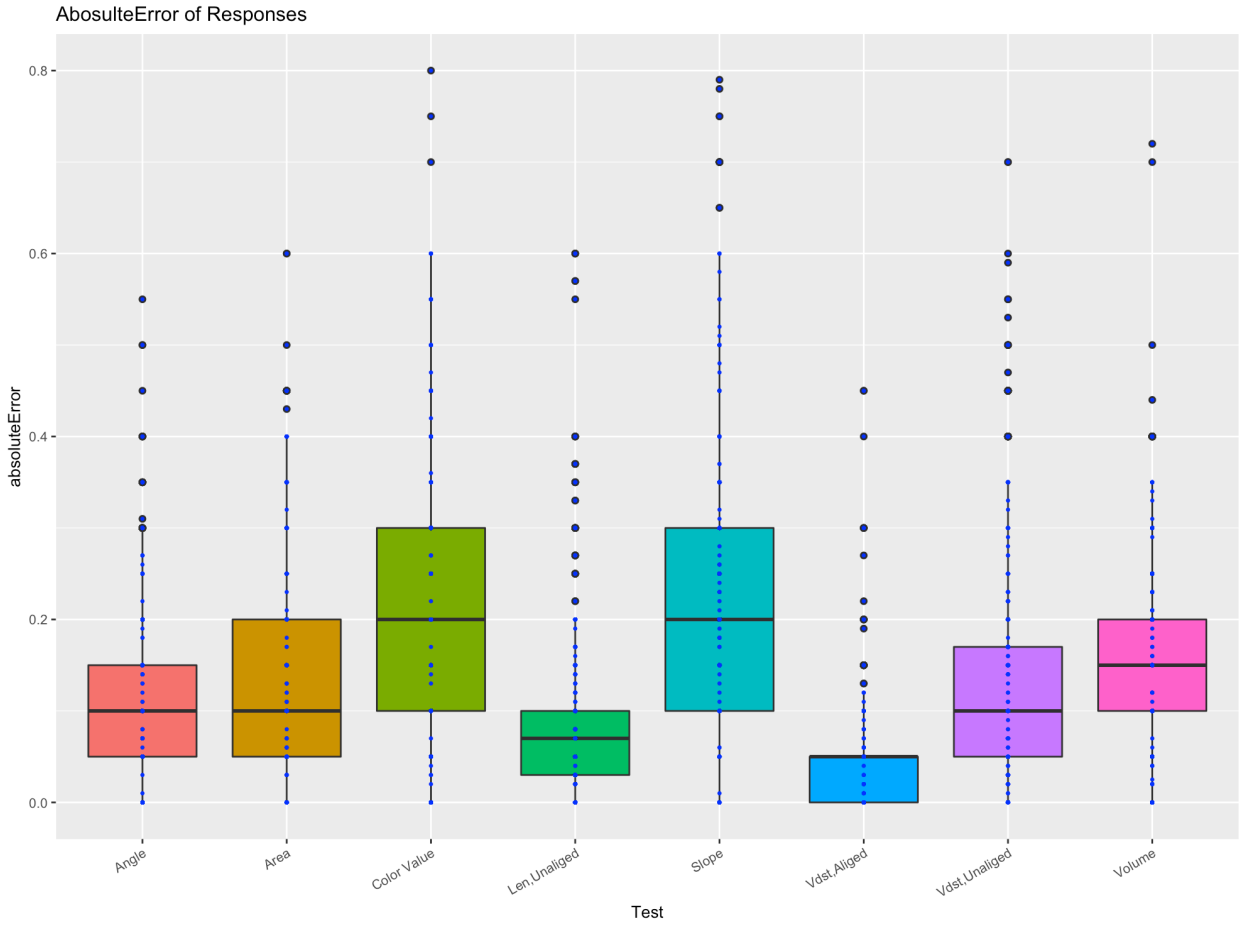
Jitter with distribution:



For jitter with distribution, I narrowed down the width and alpha as well for better view;
The result tells us that the data for 'test' close to normal distribution but definitely existing some of outliers for example the volume. In this case, the distribution gives the sense to us by using clusters and spreading the data into the box plot which make us can easily to observe whether the data is close to normal distribution or not;
Moreover, noticeable clumping existing in every 'test' categories. If most of clusters around the box, then we can see it close to normal distribution, otherwise, it's not.

In this case, without the distribution, it is hard to observe noticeable clumping of responses, we can use violin plot/method to add the distribution to this.

c. From R:

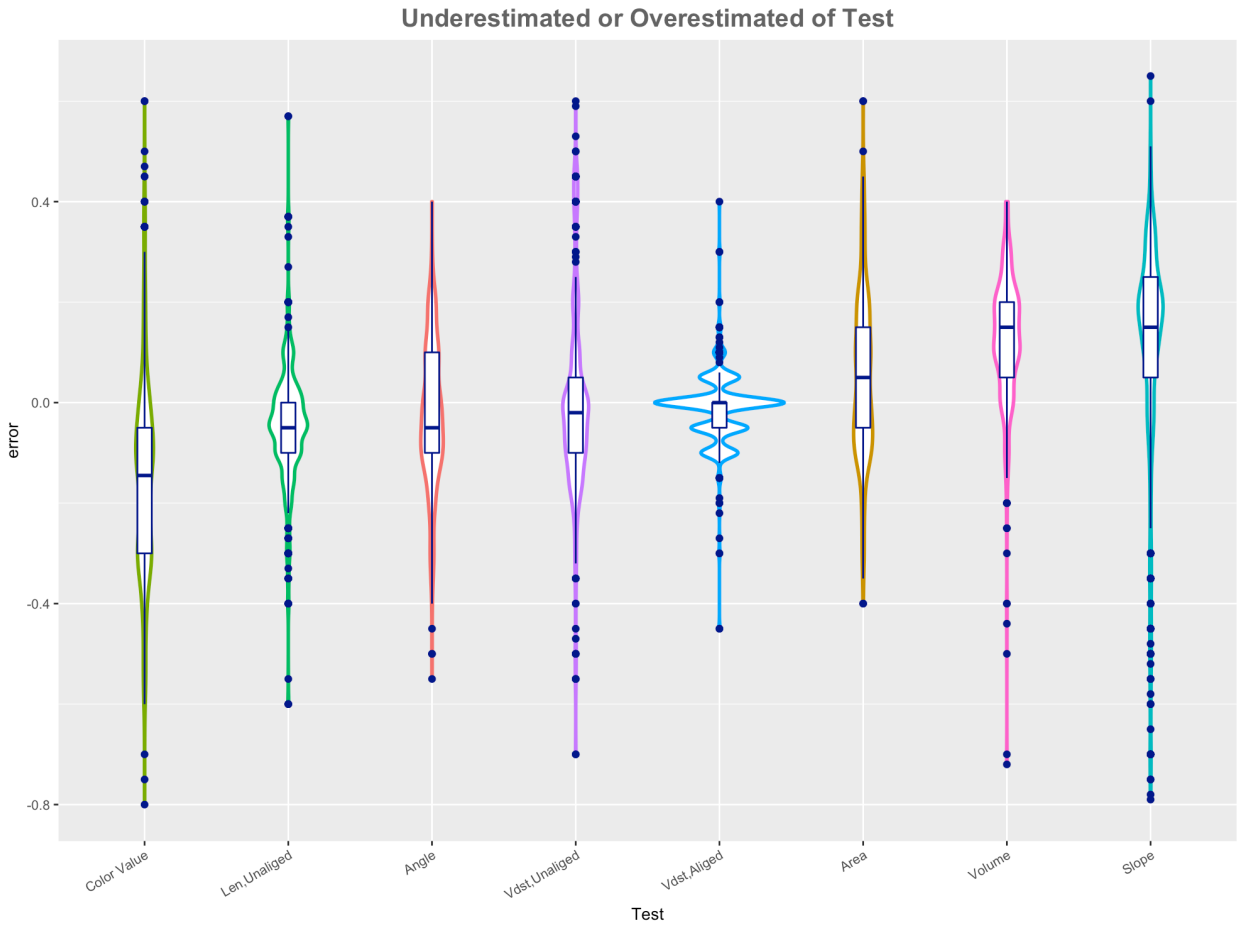


For quantization in the responses on multiple of 5% experiment with jittering along the direction of the response values (absoluteError) by a small amount. It will not distort the data but visualization, it the dataset is large enough. The term 'jittering' covers other distributions than uniform, and it is typically used to better visualize overlapping values, such as integer covariates. This helps grasp where the density of observations is high. In this case, since the absoluteError is pretty small, so that if we are not changing the quantized range of the response, it is possible that all the value clustered into one or several groups. From statistic point of view, for instance, the outliers. Even the small among of data what we are not changing the quantized range of the response, it could bias the distribution for the total data set. In this case, people will have high probability to omit that part and which could bias the determination from that.

On the other hand, small quantized range of the responses will summarize the data more precise which allowing us to reveal more detail from the data and statistic. However, of course the weakness of doing this, if the dataset are fairly large, then it could messed up our perception test, in other works, paying more attention in the detail

which not quite efficient for analysis.

d. From R:

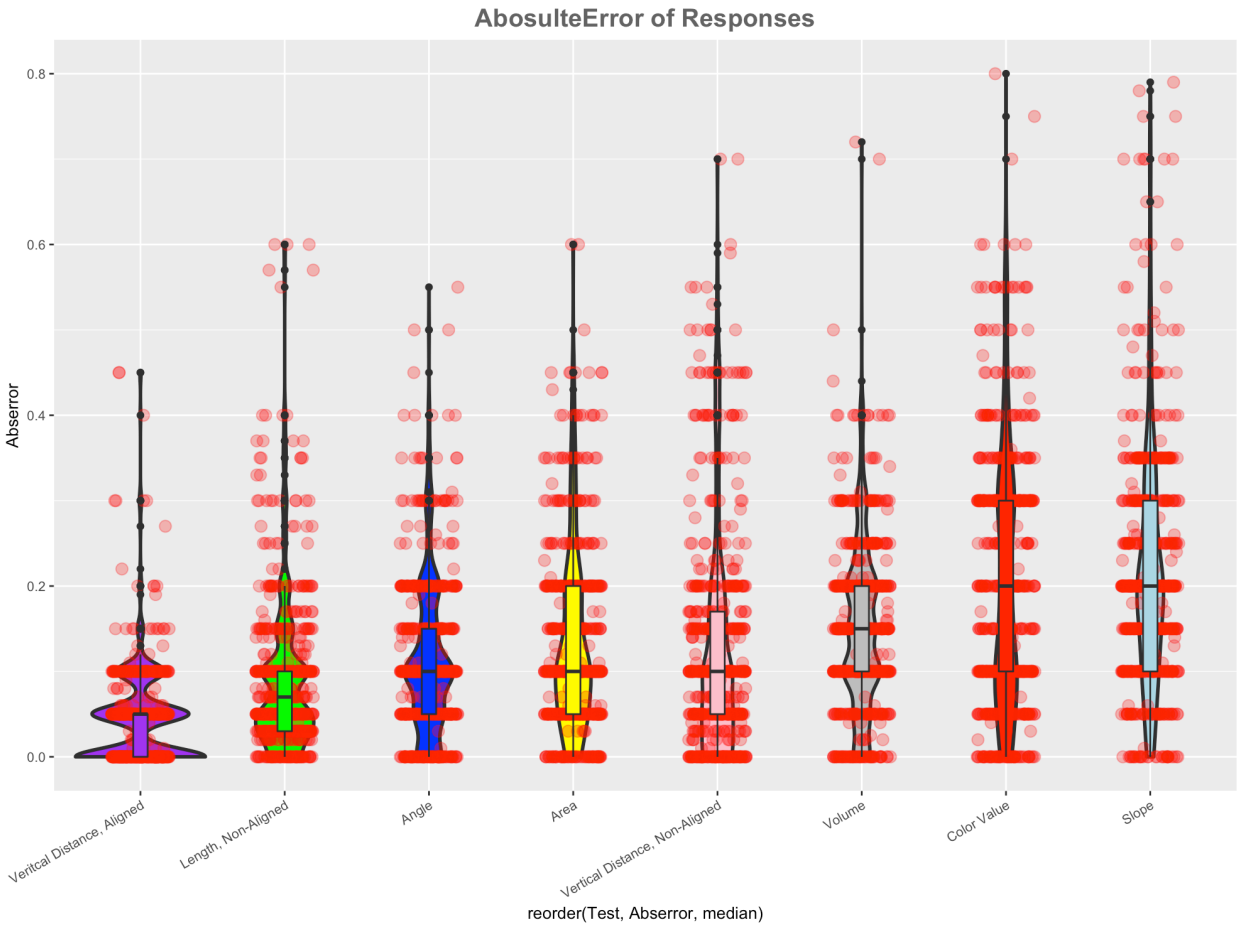


Method: R I used geom_violin() and geom_boxplot() to plot x= Test, y= Error and set the color= Test without legend. And then I sorted error value by median and bolded and center the title. What's more, the x labels were shortened.

Analysis: Based on looking at the median value of error, we can see that the tests of color value, unaligned length, angle, unaligned vertical distance are generally underestimated by people. The tests of area, volume and slope are generally overestimated by people. I used the error and test to graph this test. I have to mention that the absolute error can't be used to get this result because it cannot differentiate the overestimation and underestimation. The violin and boxplot could reveal this clearly

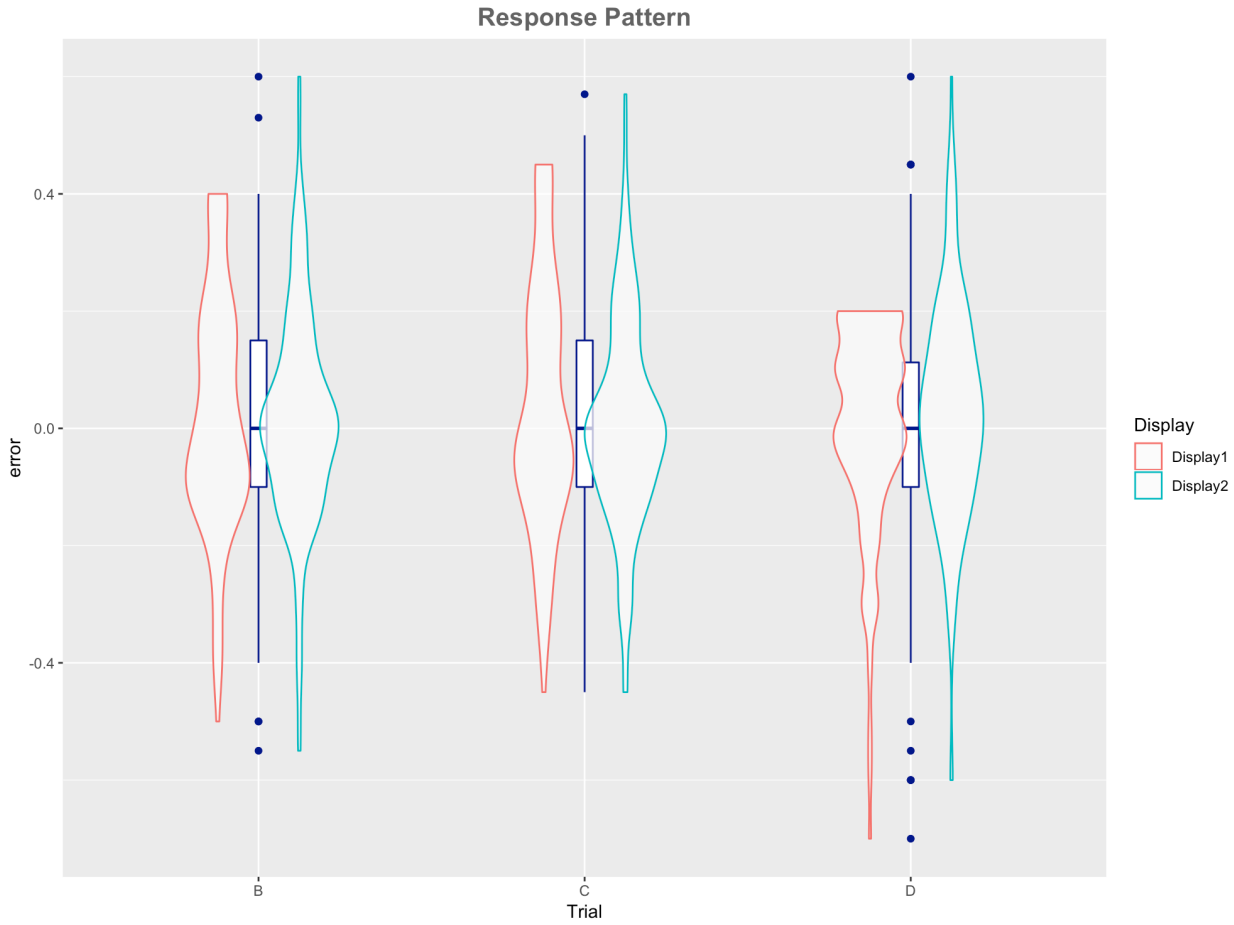
as you see on the graph as above. If the value of error is negative standing for underestimation and the value of error is positive standing of overestimation. What's more, we can see some peaks in each test. It means that more people will make that specific error with different value of error.

For adding jittering in this case:



In this case, the jittered scatter plot overlay would help us to better understanding the distribution, as I mentioned above, The term 'jittering' covers other distributions than uniform, and it is typically used to better visualize overlapping values, such as integer covariates. This helps grasp where the density of observations is high. So adding jittering, we could observe the density of observation more clearly.

e. From R:



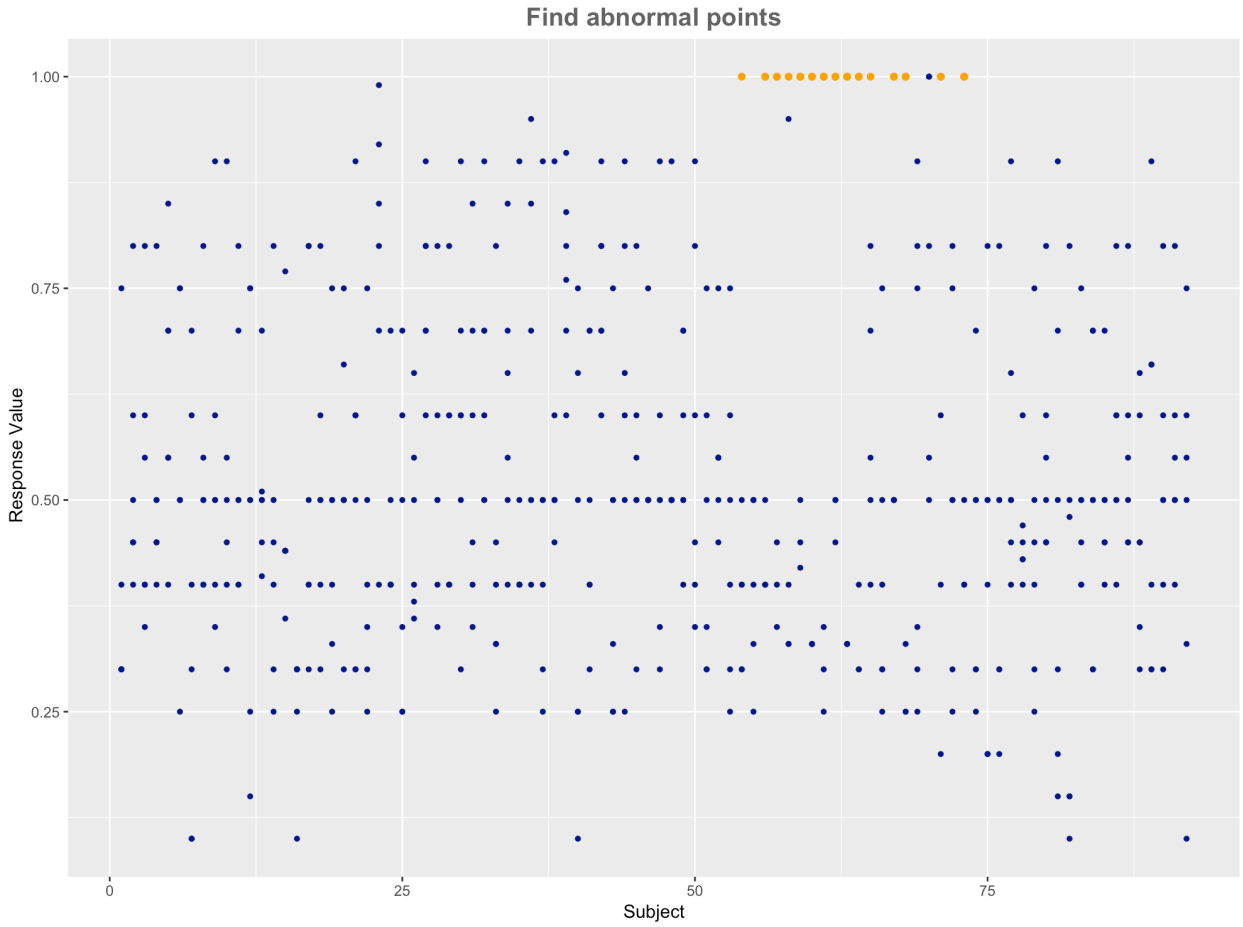
Method: R First, I subset the data to get subjects 56 to 73 using [student$Subject >=56 & student$Subject<=73,] and reset the value of Display using Subjects$Display[Subjects$Display== 1] <- "Display1" and Subjects$Display[Subjects$Display== 2] <- "Display2". And I also used geom_violin() in my graph.

Analysis:

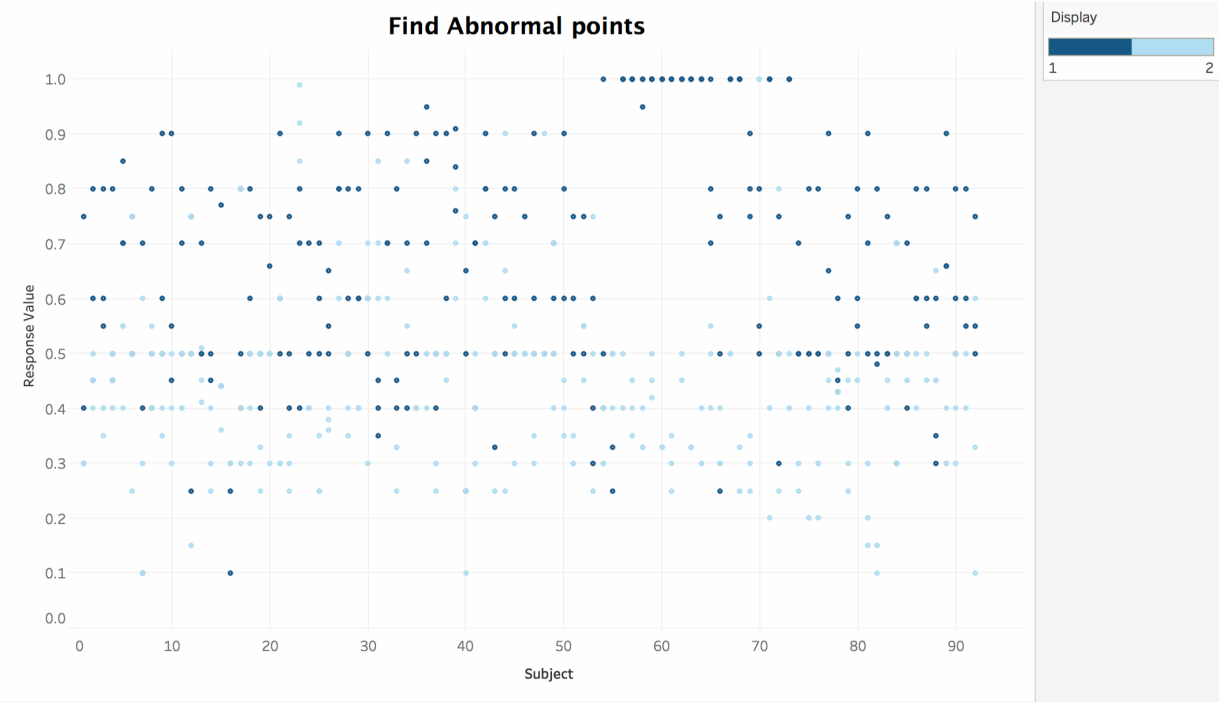
Yes, there is some differences in the values for Display 1 and 2 As you can see as above, for trial B, the peak of Display1 marked in red is about -0.15 while peak of Display2 marked in blue is about 0. It means that people participants get better at judging after having done trial B once. For trial C, the peak of Display1 marked in red is about -0.13 while peak of Display2 marked in blue is about 0. It means that people participants get better at judging after having done trial C once. For trial D, the peak of Display1 marked in red is 0,0.1, and 0.2, respectively, while peak of Display2 marked in blue is about 0. We can see that blue curve spread wider than the red

curve. It means that people participants do not get better at judging after having done trial D once, the result even gets worse.

f. From R:



From Tableau:

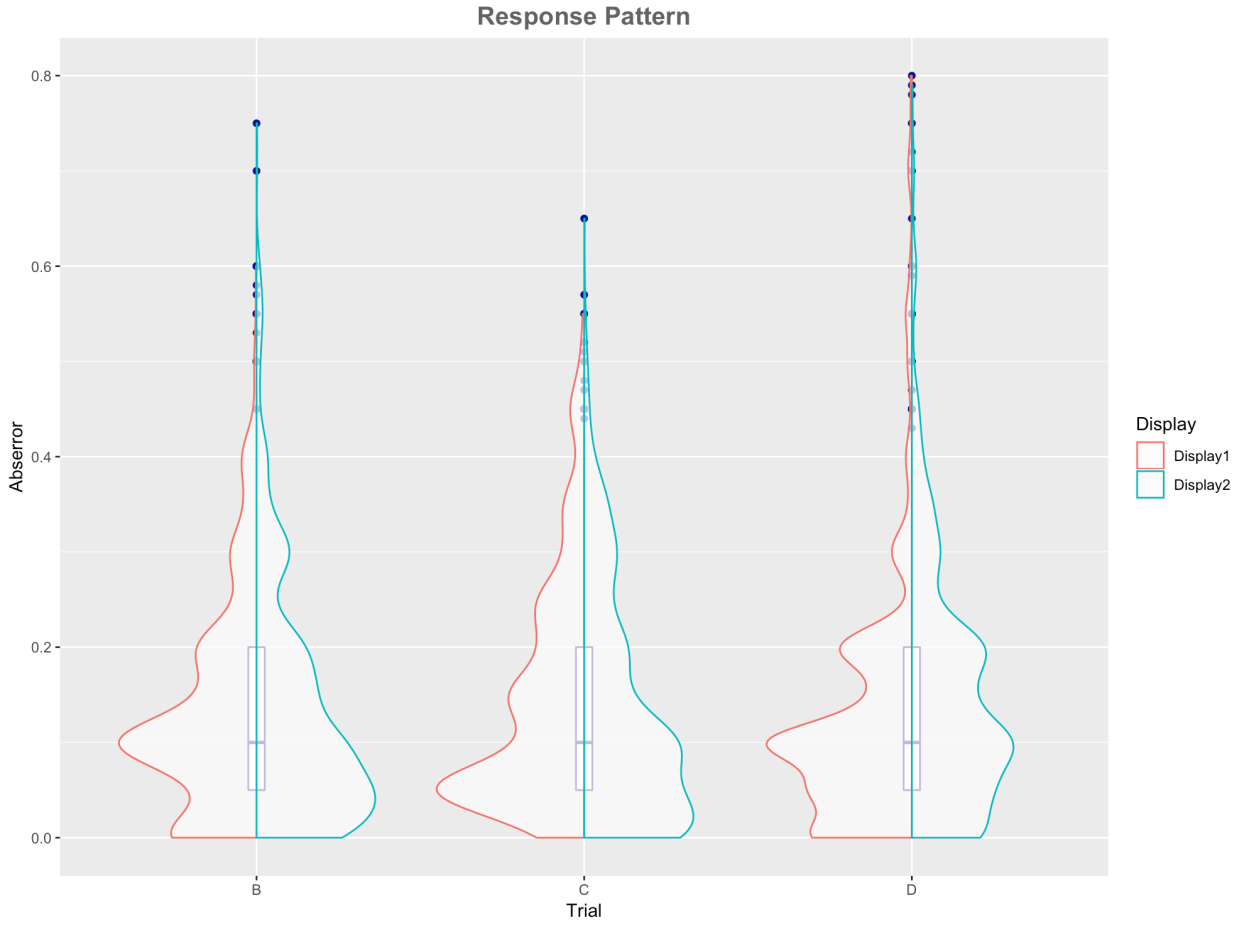


Method: R & Tableau First, I used Tableau(filter Test: Vertical distance, non-aligned) to find those abnormal points and found the corresponding subjects and then used R to highlight those points in orange. After that, I reset the value of Display using Subjects$Display[Subjects$Display== 1] <- "Display1" and Subjects$Display[Subjects$Display== 2] <- "Display2" , get the first subset data using subset(pe, Test=='Vertical Distance, Non-Aligned') , and get the second subset data using data[data$Subject %in% c(56:68,54,71,73) & data$Test=='Vertical Distance, Non- Aligned' & data$Display=='Display1'& data$Response==1,] , and finally set x= Subject, y= Response and bolded and centered the title.

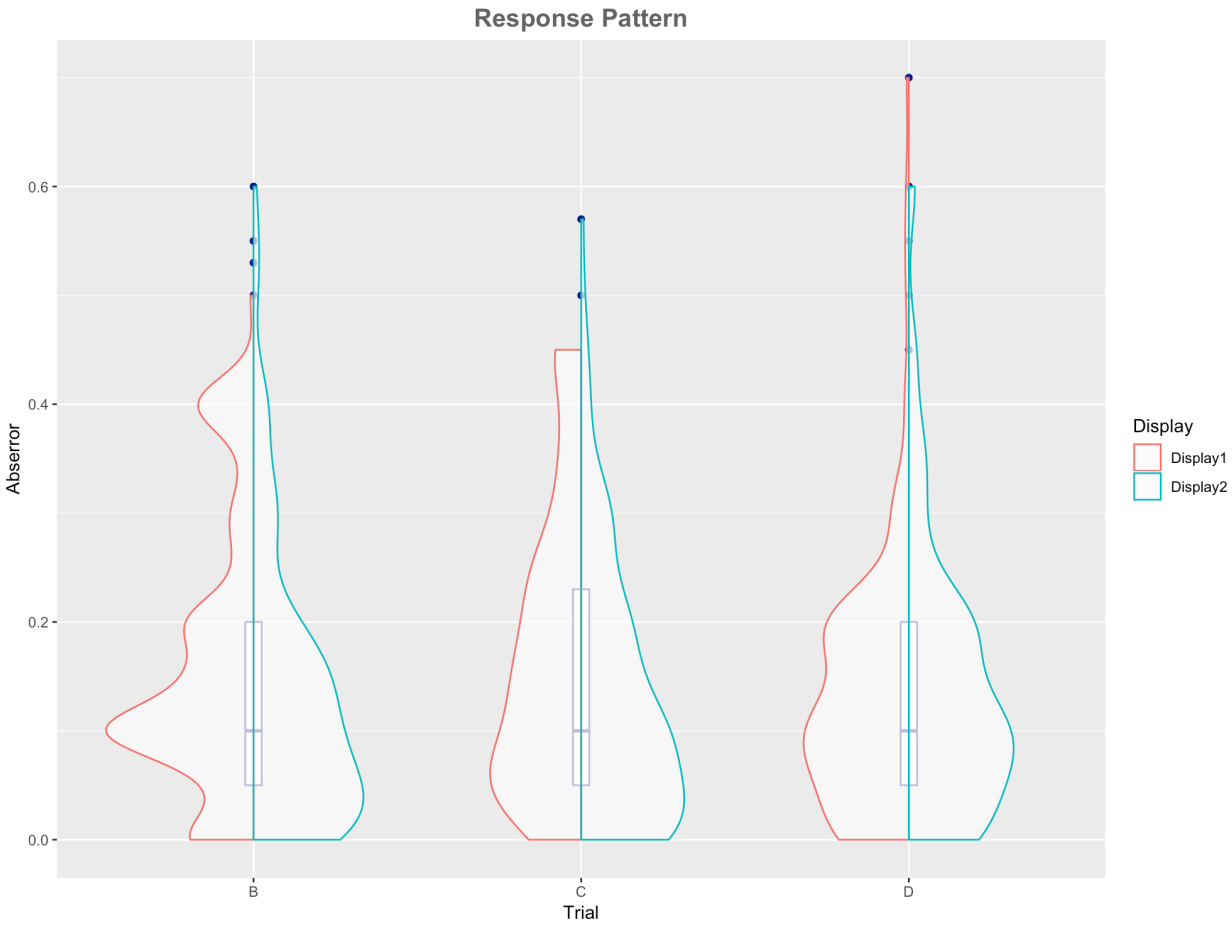
Analysis: Abnormal points: Subject 54,56, 57, 58,59,60,61,62,63,64,65,66,67,71,73 All true values gave by those responses are 1, no matter the trial is B, C, and D for Display 1.

g. From R:

All subjects:



Subjects: 56-73



Method: R

I used geom_split_violin () and geom_boxplot() to plot x= Trial, y= absulteError and set the color= Trial with legend. There are two data sets I used, one is the whole dataset, another one is the subset of dataset (subjects: 56-73) using Subjects=pe[pe$Subject >=56 & pe$Subject<=73,]. And then I bolded and center the title.

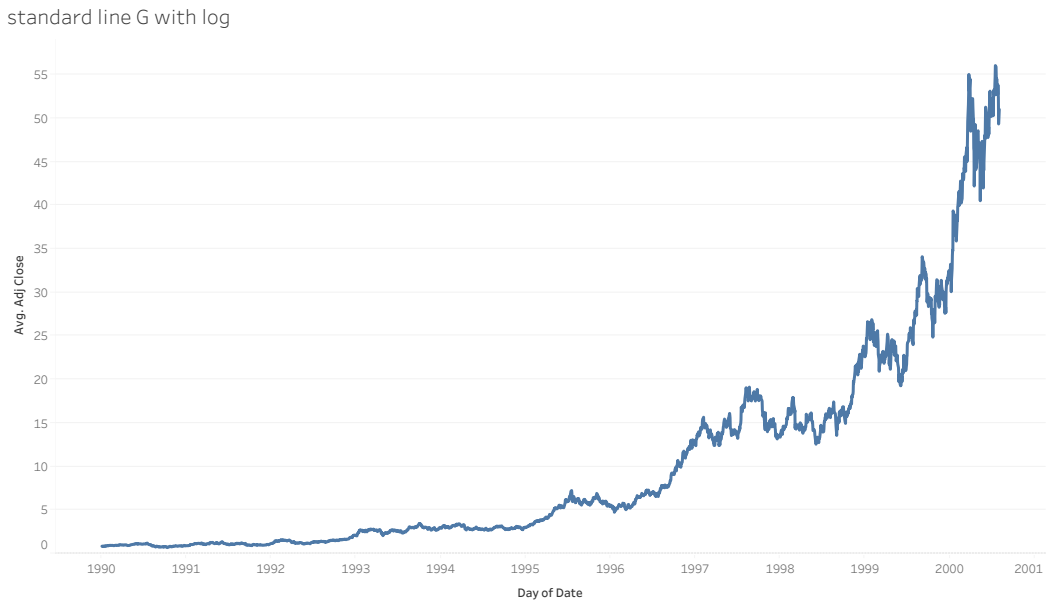
Analysis:

Those two graphs diver the information that a little bit different from that of b. does. I used the absolute error here. Generally, we can say that participants do get better at judging after having done trial once no matter from some of participants or all participants. The way we look at the curve should focus on the height and location of their peaks.

(I use absolute error instead of error and used distinct way, geom_split_violin (), to meet the requirement, and the information I gained from this way is different from the result shown previously)

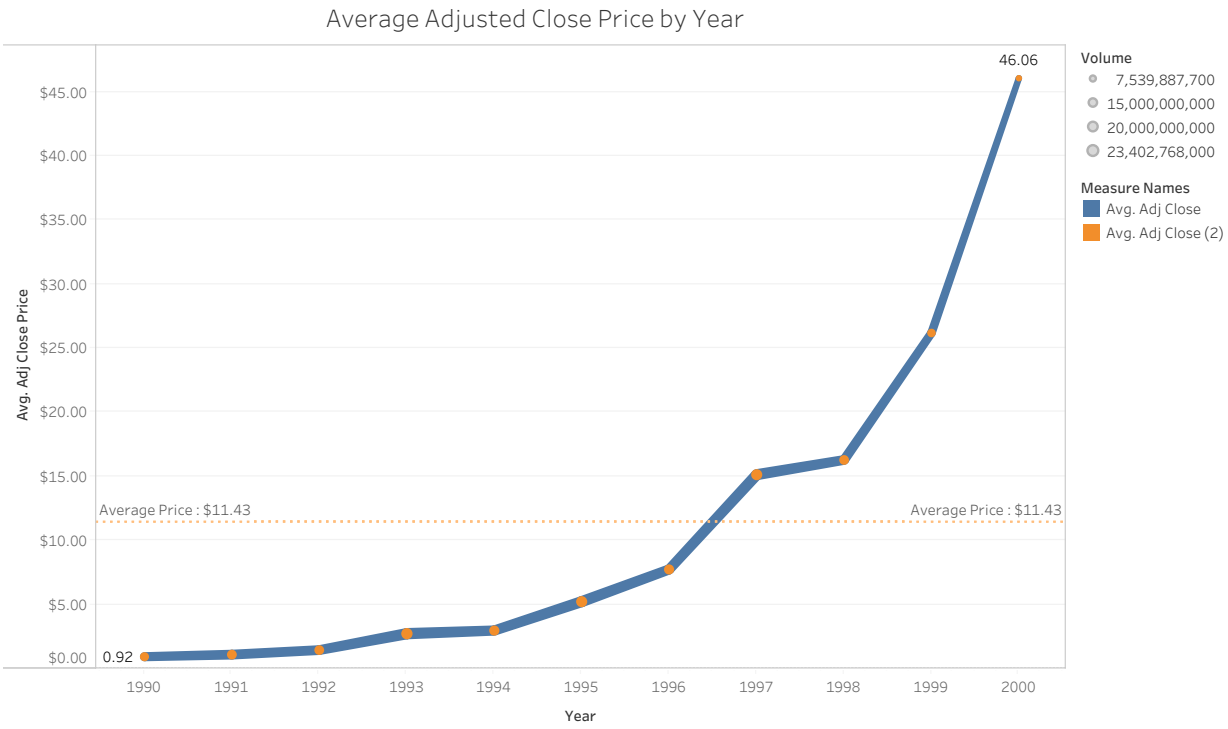
Problem 4:

a. From Tableau:



The trend of average of Adj Close for Date Day.

The trend of average of Adj Close for Date Day.



The trends of Avg. Adj Close and Avg. Adj Close for Date Year. Color shows details about Avg. Adj Close and Avg. Adj Close. Size shows sum of Volume.

Method: Tableau

Daily adjusted Closing Price

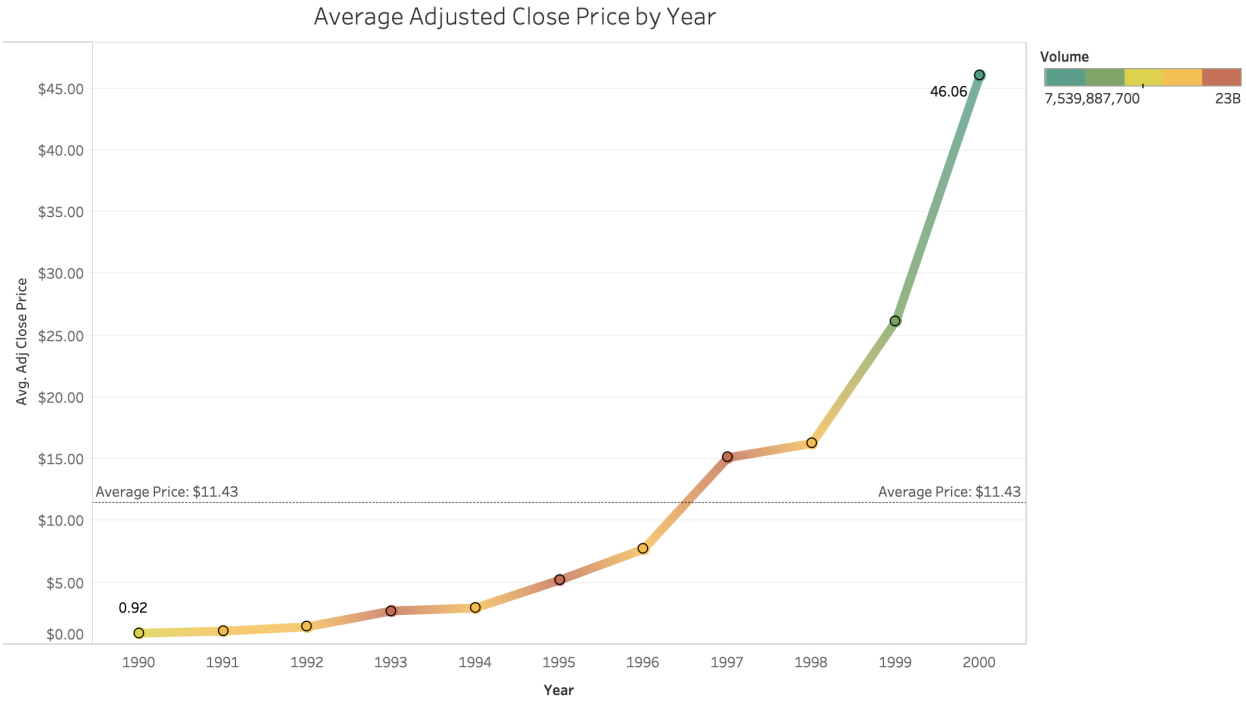
I used Tableau to answer this question. I set the x= Year (Using Exact Date), y= Adjust Close Price, size = Volume (SUM). The range of y is set fixed value from 0 to 60 and added dollar sign with value in y axis. The range of year is from 1990 to 2001 which is corresponding to the raw data. Because it shows daily price, so I did not choose any aggregation methods. However, it is difficult to see the daily price because there are too many data. It is hard to get extra information when I set the Volume measure to alter the thick along the curve. It looks wired that it event has some circles showing on the graph. That's why I display the **third** graph.

Average adjusted Closing Price by Year

I set the x= Year (Using Year), y= Adjust Close Price(AVG) & y= Adjust Close Price(AVG), size = Volume(SUM). What's more I used the dual axis for plotting the line marked in blue and plotting the scatter plot marked in orange. The range of y is I should set up to 50, however, there is an issue that I can't figure out. That's why I highlighted the minimum and maximum value of closing price. The range of year is from 1990 to 2000 because we use the average instead of daily price. I also set a

reference lined named average price (setting computation and value) and added dollar sign with value in y axis. Now we can easier see the change of volume by seeing the thickness of line compared with the daily graph. I have to say the result is better, but this is not the best way to display, I will show in following answer.

b. From Tableau:

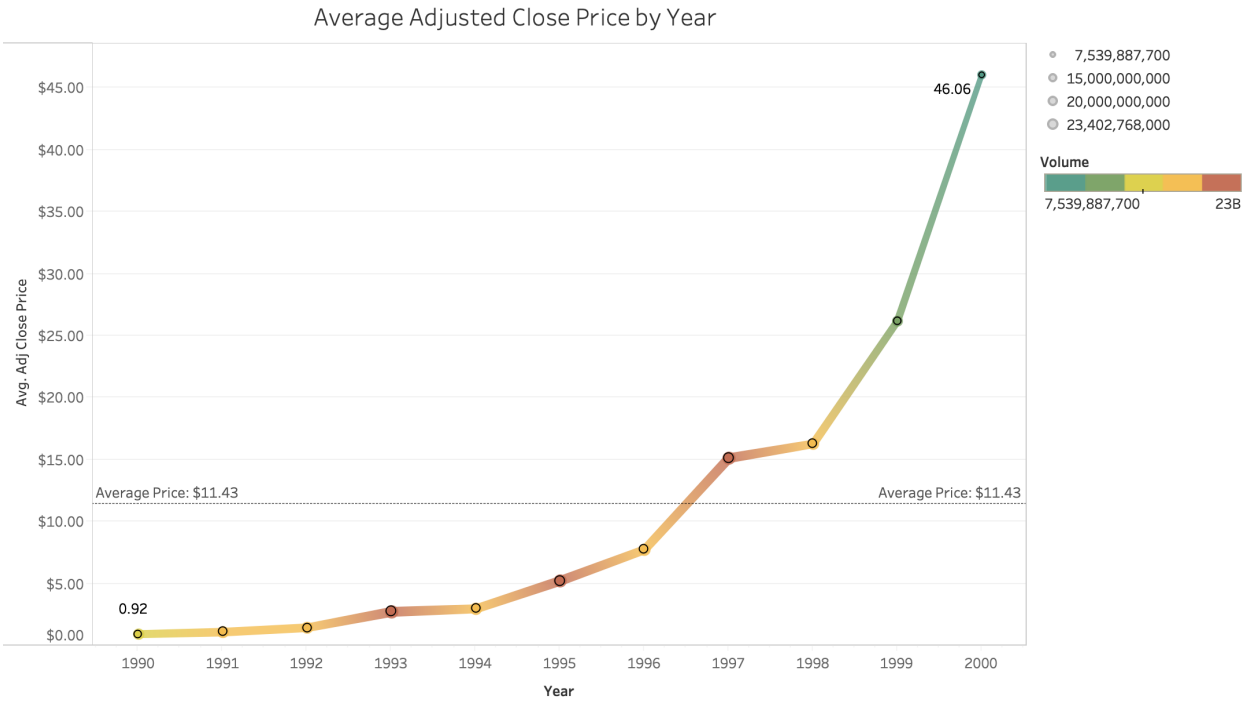


The trends of average of Adj Close and average of Adj Close for Date Year. Color shows sum of Volume.

Method: Tableau

I set the x= Year (Using Year), y= Adjust Close Price(AVG) & y= Adjust Close Price(AVG), color = Volume(SUM). What's more I used the dual axis for plotting the line marked in graduated color (Break= 5) and plotting the scatter plot marked with black border. The range of y is I should set up to 50, however, there is an issue that I can't figure out. That's why I highlighted the minimum and maximum value of closing price. The range of year is from 1990 to 2000 because we use the average instead of daily price. I also set a reference lined named average price (setting computation and value) and added dollar sign with value in y axis. Now we can easier see the change of volume by seeing the changing color of line compared with the daily graph.

c. From Tableau:



The trends of average of Adj Close and average of Adj Close for Date Year. Color shows sum of Volume. Size shows sum of Volume.

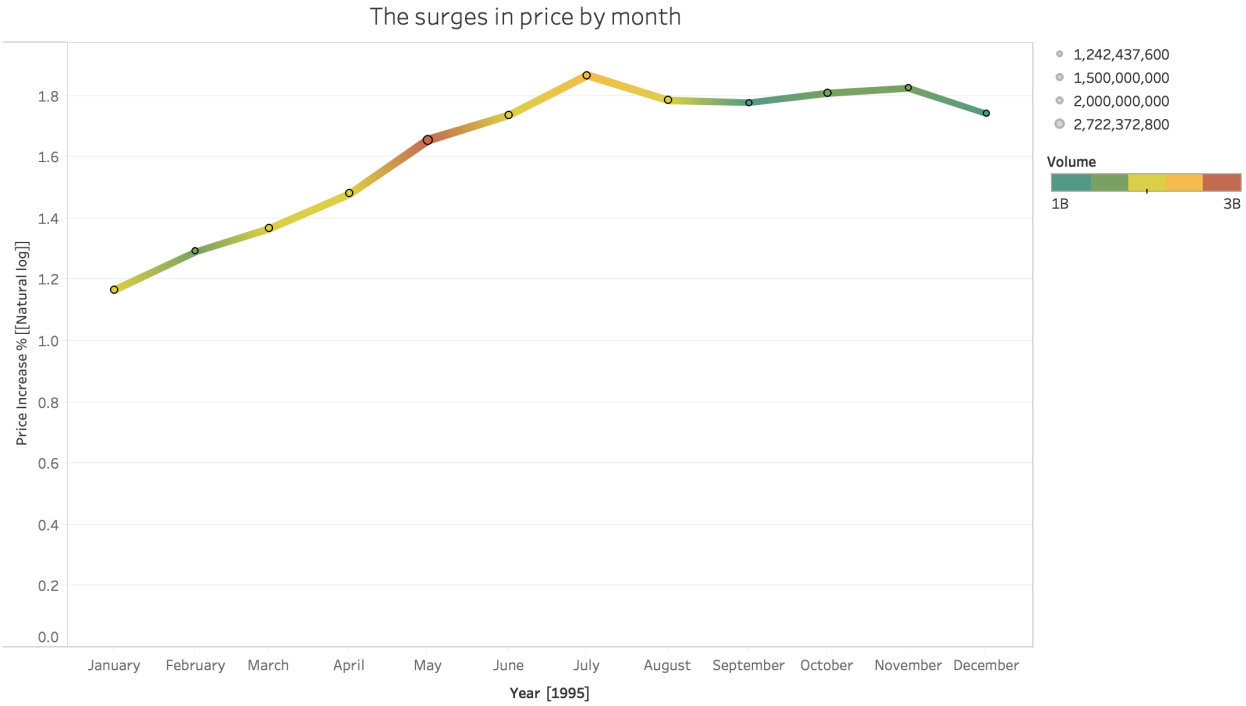
Method: Tableau

This time, I used Volume measure to alter both color and size and got the graph as above. The method is similar with that of 4b.

Analysis:

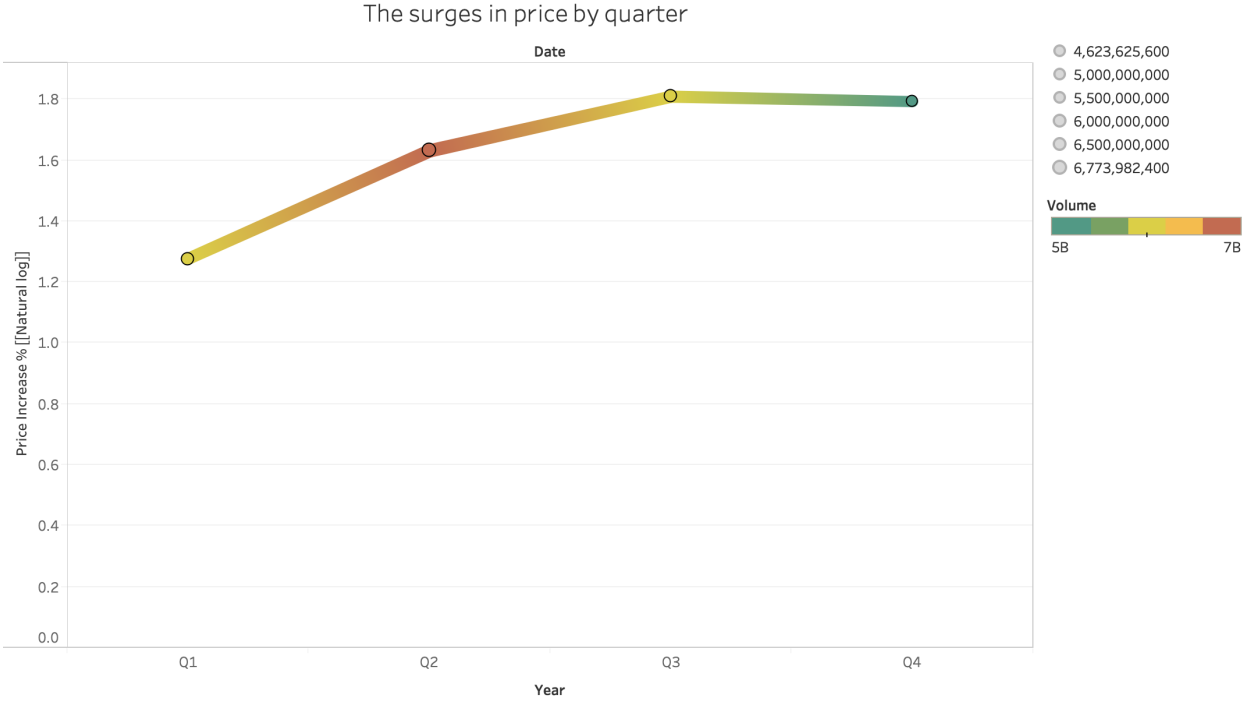
It is hard for us to see the change of volume based on the thickness of line, not very noticeable. In my opinion, using the Volume measure to alter the color of the line at each point communicates the Volume data more efficiently. What's more, combining two methods, size and color, may more efficient than that of only using color.

d. From Tableau:
(For Month)



The trends of average of ln(Price) and average of ln(Price) for Date Month. Color shows sum of Volume. Size shows sum of Volume. The data is filtered on Date Year, which keeps 1995. The view is filtered on average of ln(Price), which keeps all values.

(For Quarter)



The trends of average of ln(Price) and average of ln(Price) for Date Quarter. Color shows sum of Volume. Size shows sum of Volume. The data is filtered on Date Year, which keeps 1995. The view is filtered on average of ln(Price), which keeps all values.

Method: Tableau

The method I used for this answer is almost same with the method of 5c. There are two differences between those two graphs and the graph showing in 5c. I created a new column named In(price) which is natural log of adjusted closing price. I selected x = month and quarter, respectively, y=In(Price)(AVG). That's why you can see there are more circles showing in the graph by month than that of by quarter. To let audience clearly understand of the scale of y-axis, I added the subtitle to state that this wise percentage labeled as Price increase % with as [Natural log].

Analysis:

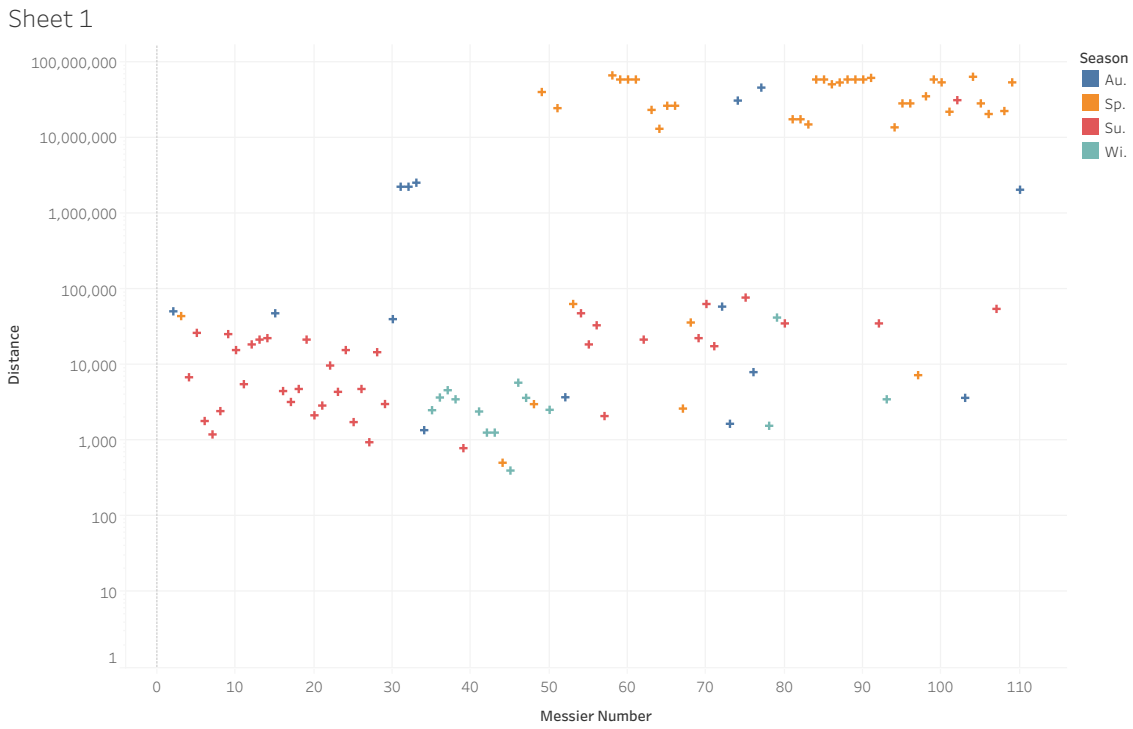
There are 4 points that are between 10% and 20% increases in my graph by month. There are 2 points that are between 10 and 20% increases in my graph by quarter. Because we are supposed to identify three surges in price, For answering this question from month graph. Those 4 surges occur on Jan 1995 to Feb 1995(1.18–1.30), Feb 1995 to Apr 1995(1.3–1.5), Apr 1995 to May 1995(1.5–1.63), and May 1995 to July 1995 (1.63–1.90).

For quarter: Those 2 surges occur on Q1 1995 to Q2 1995 (1.3–1.62) and Q2 1995 to Q3 1995(1.62–1.8).

Problem 5:

a. From Tableau:

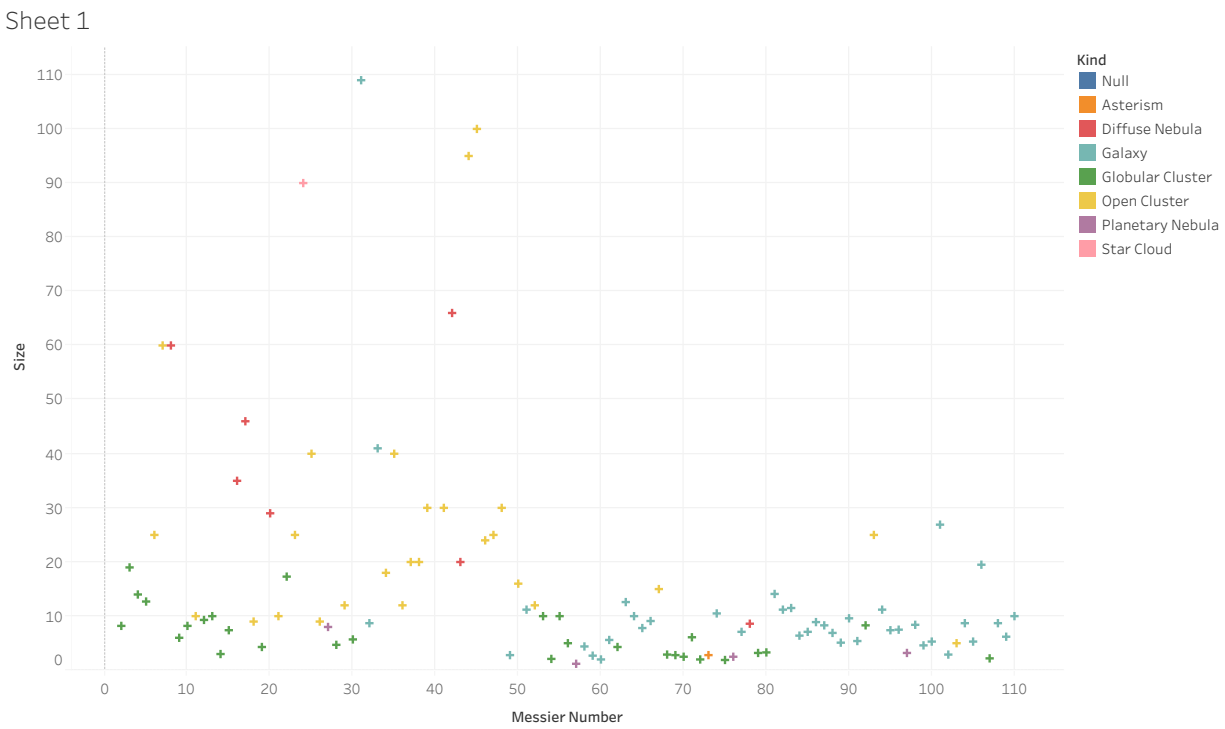
1st



1st: x=Messier Number, y= Distance_LY, color = Season.

The range of y is set fixed value from 0 to 110 and the range of x is set using log10. To let audience clearly understand of the scale of y-axis, I added the subtitle to state that this is logarithmic scale value labeled as [log10]. I kept the x not changed and tried putting the rest of variables into y one by one and colored them using categorical variables. I removed Null from my graph.

2nd:

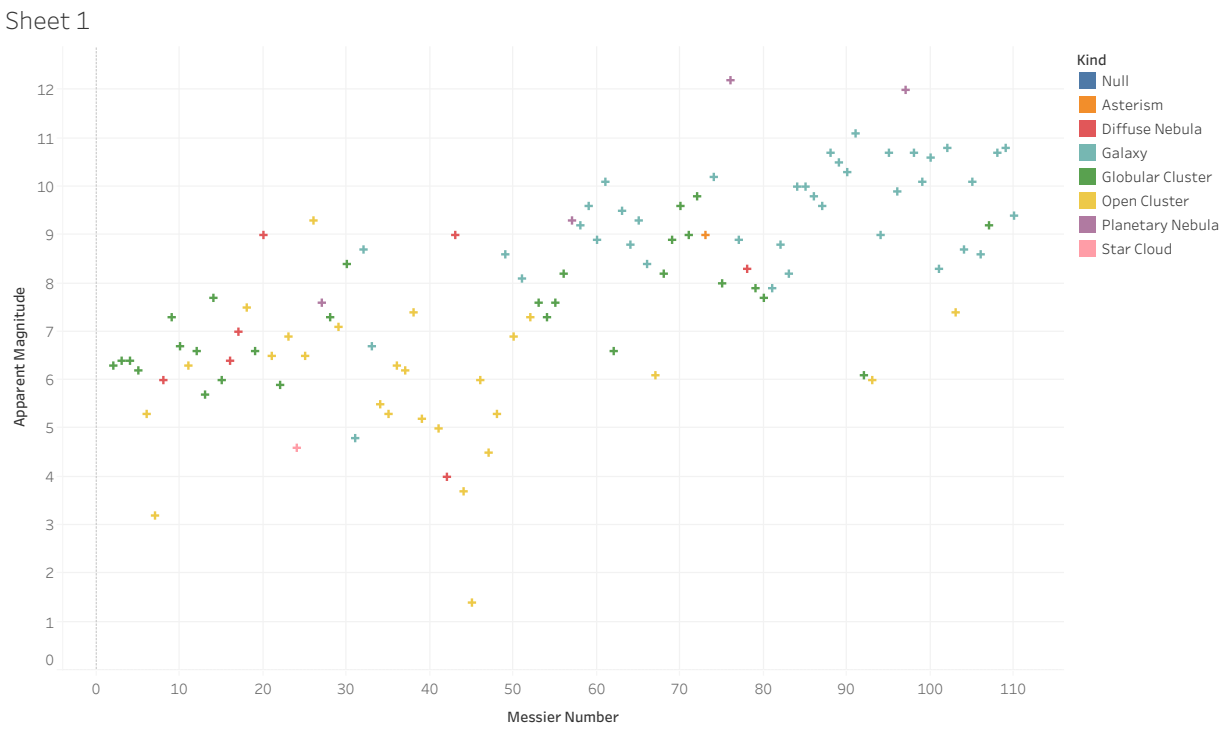


The plot of sum of Size for Messier Number. Color shows details about Kind.

2nd: x=Messier Number, y= size, color = Kind;

The range of y is set fixed value from 0 to 110 and the range of x is set fixed value from 0 to 110. I kept the x not changed and tried putting the rest of variables into y one by one and colored them using categorical variables. I removed Null from my graph.

3rd:



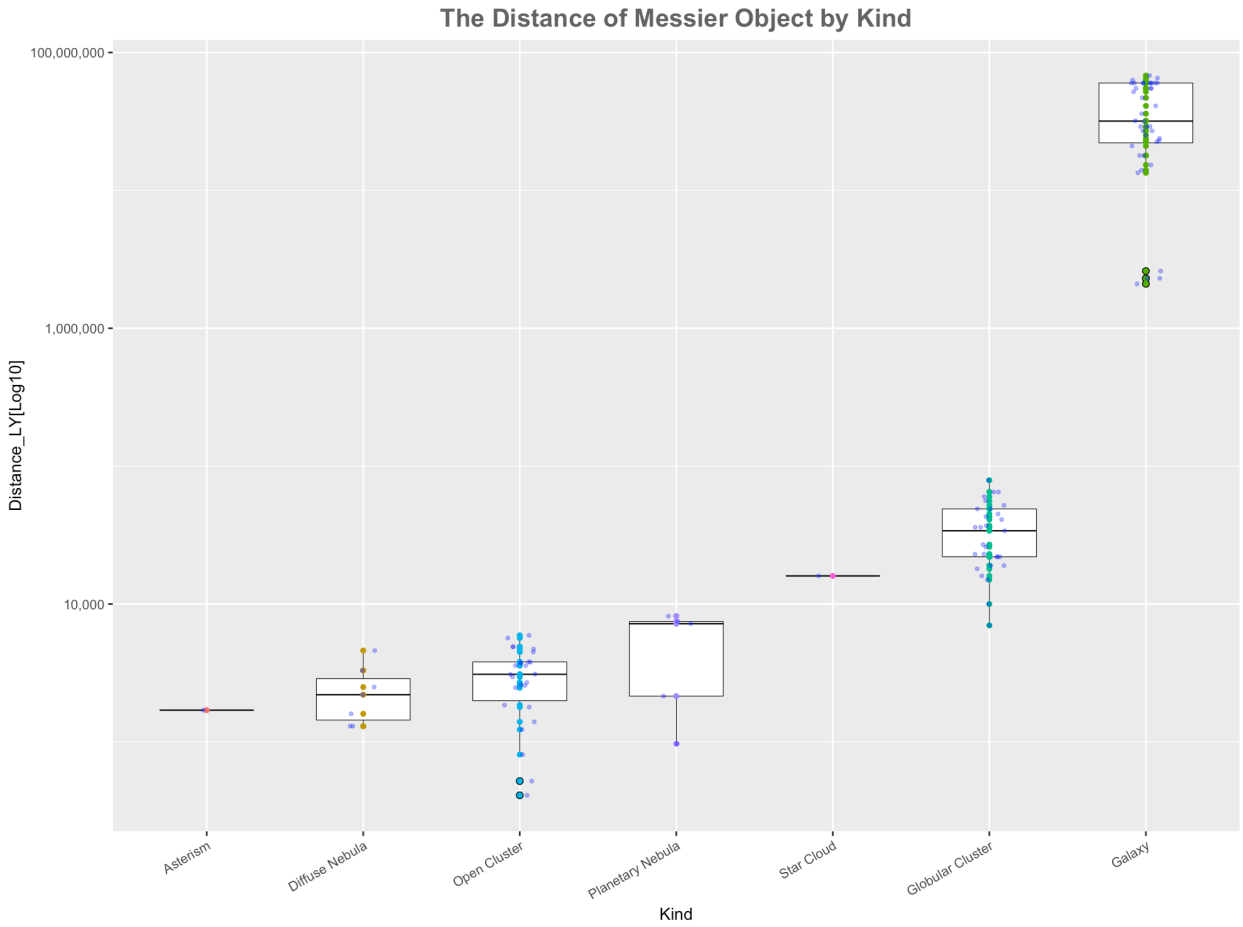
The plot of sum of Apparent Magnitude for Messier Number. Color shows details about Kind.

2nd: x=Messier Number, y= Apparent Magnitude, color = Kind;

Briefly Analysis:

1st:As you can see, objects marked in red standing for Su season having smaller messier number, mainly ranging from 0 to 30. Objects marked in green standing for Wi season having the moderate number, ranging from 30 to 50. Objects marked in orange standing for Sp season having larger messier number, mainly ranging from 50 to 70 and from 80 to 110. Objects marked in blue standing for Au season having small messier number, mainly ranging from 30 to 35. What's more, it also shows that the messier objects located far away distance are more likely to have large messier number.

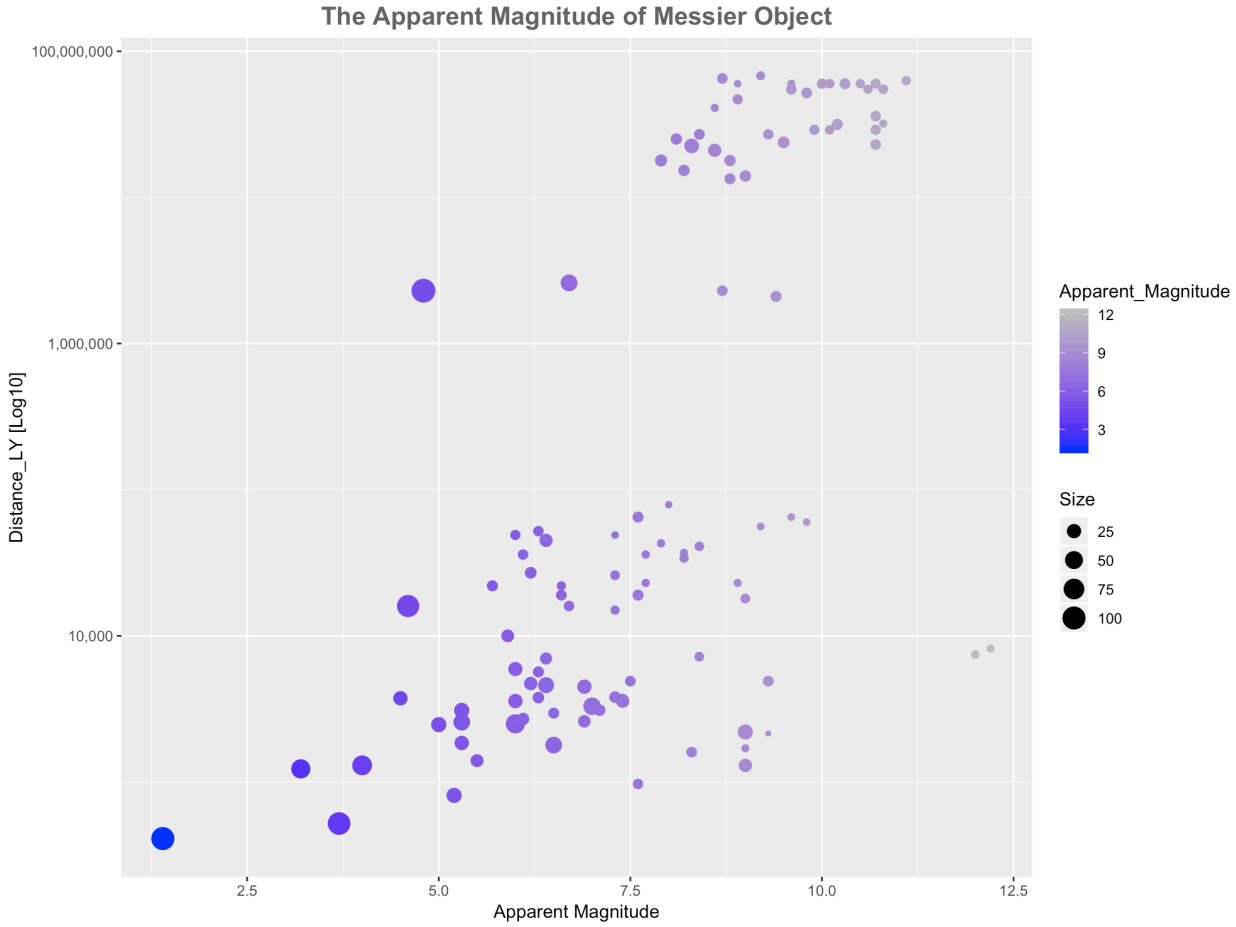
b. From R:



Method: R

I used geom_boxplot(),geom_point(), and geom_jitter() to plot this graph. The null value is removed by using ggplot(data=subset(Messier, !is.na(Kind)). The scale_y_log10(labels = comma) is used to log10 of y and force the tick marks not showing scientific notation. The boxplots are sorted by median value of distance using aes(x = reorder(Kind,Distance_LY,median), y=Distance_LY, color= Kind)). And then I bolded and center the title.

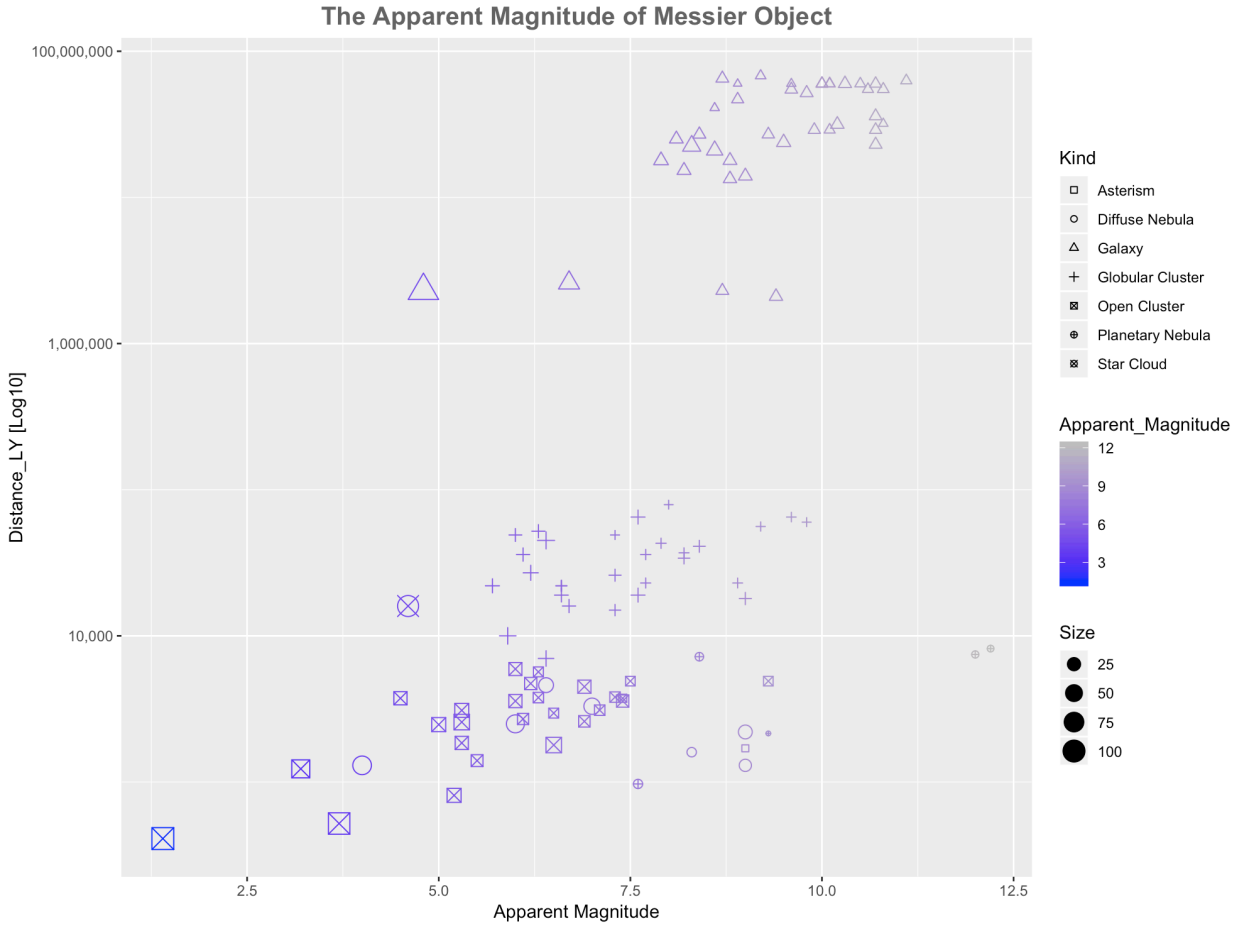
c. From R:



Method: R

I used geom_point () to plot this graph. The null value is removed by using ggplot(data=subset(Messier, !is.na(Kind)). The scale_y_log10(labels = comma) is used to log10 of y and force the tick marks not showing scientific notation. I used the scale_color_gradient(low="blue", high="gray") to visually differentiate the magnitude of objects, the higher the number of the fainter the object is in the sky. I set the blue standing for bright and gray standing for faint. I also set the size = Size which you can see different size of dot spread out in my graph. And then I bolded and centered the title.

d. From R:



Method: R

I used geom_point () to plot this graph. The null value is removed by using ggplot(data=subset(Messier, !is.na(Kind)). The scale_y_log10(labels = comma) is used to log10 of y and force the tick marks not showing as scientific notation. I used the scale_color_gradient(low="blue", high="gray") to visually differentiate the magnitude of objects, the higher the number of the fainter the object is in the sky. I set the blue standing for bright and gray standing for faint. I also set the size = Size which you can see different size of dot spread out in my graph. What's more, I manually set the shape using scale_shape_manual(values=c(0,1,2,3,7,10,13)) instead of use the shape= Kind which will only give 6 types of shape which is not enough for me since I need 7 types of shape.

Analysis:

In this graph, it delivers four information of messier objects. The color of shape

shows the level of apparent magnitude, the size of shape shows the size of objects, the type of shape shows the kind of objects, and the location of the shape shows the distance of objects. It is very easy to get information of messier object with legend.

Content:
R code
3.b:

```
pe$error = (pe$Response - pe$TrueValue)
pe$Abserror = abs(pe$Response - pe$TrueValue)
head(pe)

ggplot(pe, aes(x=Test, y=Abserror)) + geom_boxplot(aes(fill=Test)) +
  geom_jitter(color="red", alpha=.3, size=3, width=.2) +
  geom_point(aes(x=as.numeric(Test) + rnorm(n.each, 0, .03), y=Abserror),
color="red", alpha=.3, size=3)
```

3.c:

```
abr= c("Length, Non-Aligned" = "Len,Unaliged","Veritcal Distance,
Aligned"="Vdst,Aliged", "Vertical Distance, Non-Aligned"=
"Vdst,Unaliged")
ggplot(student, aes(x=Test, y=absoluteError)) +
  geom_boxplot(aes(fill=Test))  +
  geom_point(size=0.5,color='blue') +
  theme(legend.position='none')+
  theme(text = element_text(size=10),axis.text.x =
element_text(angle=30, hjust=1)) +
  labs(title = "AbosulteError of Responses")+
  scale_x_discrete(labels=abr)
```

3.d

```
ggplot(pe, aes(x=reorder(Test, Abserror, median), y=Abserror,
fill=factor(Test))) +
  geom_violin(size=1) +
  theme(legend.position='none')+
  theme(text = element_text(size=10),axis.text.x =
element_text(angle=30, hjust=1)) +
  labs(title = "AbosulteError of Responses")+
  geom_jitter(color="red", alpha=.3, size=3, width=.2) +
  theme(plot.title = element_text(hjust = 0.5,color="#666666",
face="bold", size=15))+
  geom_boxplot(width=.1) +
  scale_fill_manual(values=c("blue", "yellow", "red", "green",
"lightblue", "purple", "pink", "grey"))
```

3.e

```
# in this case, the second subjects is first subjects of subject
Subjects=pe[pe$Subject >=56 & pe$Subject<=73,]
Subjects$Display[Subjects$Display== 1] <- "Display1"
Subjects$Display[Subjects$Display== 2] <- "Display2"
ggplot(Subjects, aes(x=Test, y=error,color=Display))+
  geom_boxplot(width=0.05,color="darkblue")+
  geom_violin (alpha = 0.7, width =0.5) +
  #geom_split_violin(alpha = 0.7)
  theme(plot.title = element_text(hjust = 0.5,color="#666666",
face="bold", size=15))+
  labs(title = "Response Pattern")
```

3.f

```
## subset the original dataset
pe$Display[pe$Display== 1] <- "Display1"
pe$Display[pe$Display== 2] <- "Display2"

data= subset(pe, Test=='Vertical Distance, Non-Aligned')
data1=data[data$Subject %in% c(54,56:65,67,68,71,73) &
data$Test=='Vertical Distance, Non-Aligned' &
data$Display=='Display1'& data$Response==1,]

## highlight abnormal points
ggplot(data1,aes(x=Subject, y=Response)) +
  geom_point(data=data,aes(x=Subject, y=Response ),
color='darkblue',size=1) +
  geom_point(size=1.5,color='orange')+
  labs(title = "Find abnormal points") +
  theme(plot.title = element_text(hjust = 0.5,color="#666666",
face="bold", size=15)) +
  theme(legend.position='none') +
  ylab(" Response Value")
```

3.g

```
##Subject 56-73
ggplot(Subjects, aes(x=Trial, y=Abserror,color=Display))+
  geom_boxplot(width=0.05,color="darkblue")+
  geom_split_violin(alpha = 0.7)+
  theme(plot.title = element_text(hjust = 0.5,color="#666666",
face="bold", size=15))+
  labs(title = "Response Pattern")
#Subjects=student[student$Subject >=56 & student$Subject<=73,]
#Subjects$Display[Subjects$Display== 1] <- "Display1"
#Subjects$Display[Subjects$Display== 2] <- "Display2"

pe$Display[Subjects$Display== 1] <- "Display1"
pe$Display[Subjects$Display== 2] <- "Display2"
## All Subject
ggplot(pe, aes(x=Trial, y=Abserror,color=Display))+
  geom_boxplot(width=0.05,color="darkblue")+
  geom_split_violin(alpha = 0.7)+
  theme(plot.title = element_text(hjust = 0.5,color="#666666",
face="bold", size=15))+
  labs(title = "Response Pattern")
```

5.b

```
ggplot(data=subset(Messier, !is.na(Kind)), aes(x =
reorder(Kind,Distance_LY,median), y=Distance_LY, color= Kind)) +
  geom_boxplot(width=0.6,size=0.2,color="black")+
  geom_point(size=1)+
  theme(text = element_text(size=10),axis.text.x =
element_text(angle=30, hjust=1)) +
  labs(title = "The Distance of Messier Object by Kind")+
  theme(legend.position='none',plot.title = element_text(hjust =
0.5,color="#666666", face="bold", size=15))+
  geom_jitter(position=position_jitter(width=.1,
height=0),color="blue", size=0.7,alpha=.3)+
  xlab("Kind") +
  ylab("Distance_LY[Log10]") +
  scale_y_log10(labels = comma)
```

5.c

```
ggplot(data=subset(Messier, !is.na(Kind)), aes(x =
Apparent_Magnitude, y=Distance_LY ,color= Apparent_Magnitude)) +
  geom_point(aes(size= Size)) +
  scale_color_gradient(low="blue", high="gray") +
  labs(title = "The Apparent Magnitude of Messier Object")+
  theme(plot.title = element_text(hjust = 0.5,color="#666666",
face="bold", size=15)) +
  xlab("Apparent Magnitude") +
  ylab("Distance_LY [Log10]")+
  scale_y_log10(labels = comma)
```

5.d

```
ggplot(data=subset(Messier, !is.na(Kind)), aes(x =
Apparent_Magnitude, y=Distance_LY,color= Apparent_Magnitude)) +
  geom_point(aes(size= Size, shape=Kind)) +
  scale_color_gradient(low="blue", high="gray") +
```

```
  labs(title = "The Apparent Magnitude of Messier Object")+
  theme(plot.title = element_text(hjust = 0.5,color="#666666",
face="bold", size=15)) +
  xlab("Apparent Magnitude") +
  ylab("Distance_LY [Log10]") +
  scale_y_log10(labels = comma) +
  scale_shape_manual(values=c(0,1,2,3,7,10,13))
```