Perception data set analysis

Figure 1 Univariant Scatterplot with jitter

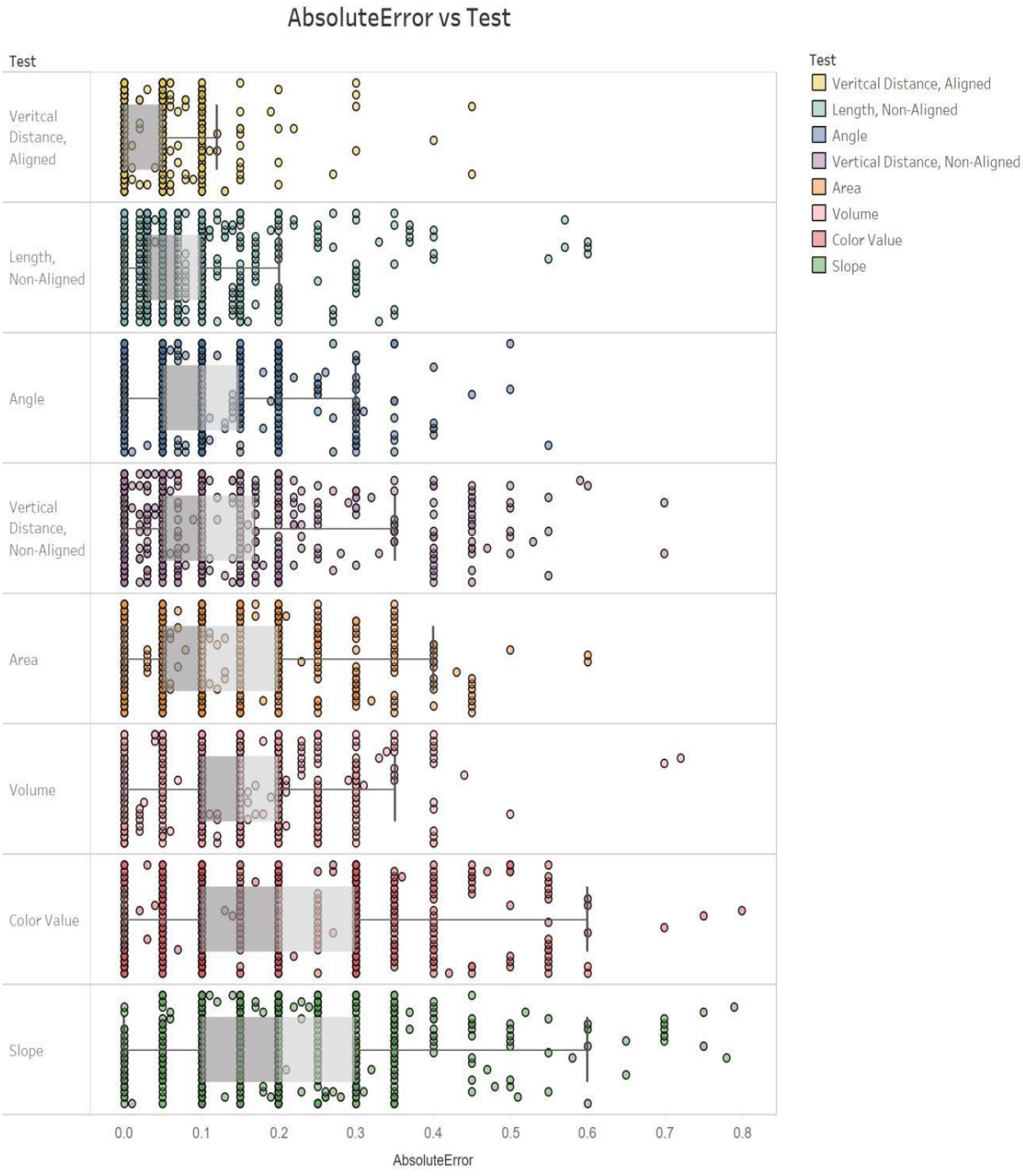
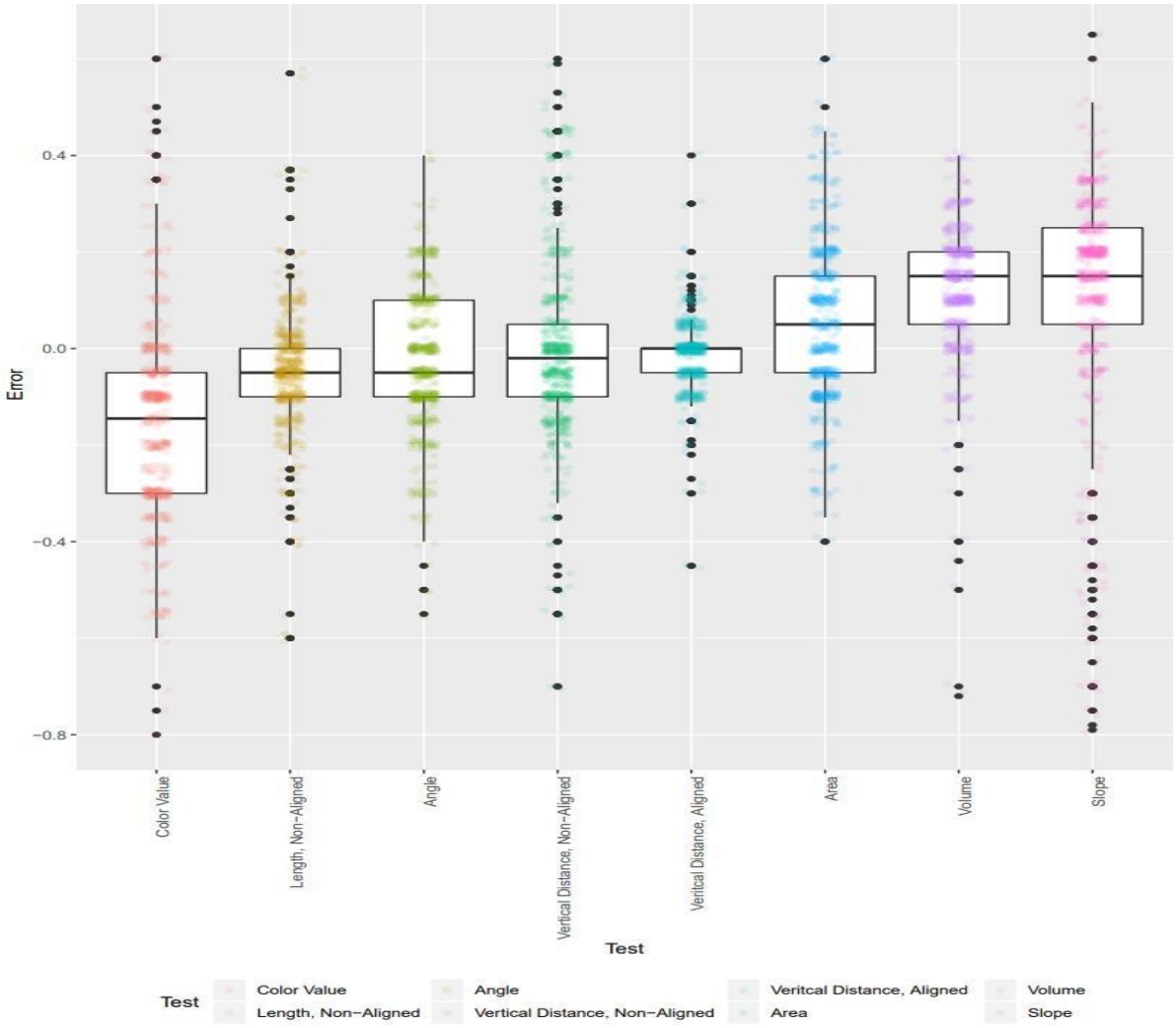


Figure 1(the univariant scatterplot) shows the majority of the Absolute Error data falls in the range from 0 to 0.4. The Vertical distance Aligned test data clumped between 0 to 0.15, which has the smallest absolute error. The Slope data clumped between 0.1 and 0.3, which has the largest absolute error. Color Value data spreads the longest range with the lowest density. Slope and Color Value have the most outliers.

Figure 2 Error vs Test box jitter plot



The Error of the response field will be used to the graph. Figure 2 above (the jittered box plot) can clearly show the possibility of under/overestimated data. Color value test shows the most likely underestimated because the majority of the data are under zero, with the largest negative mean. Length non-aligned shows that more likely to underestimate since the median is negative and the majority of the data under zero. Vertical distance aligned appears that people are more

likely to underestimate than overestimate, and its Error median is zero. One the other hand, Slope and Volume show the most overestimated because the majority of the data are greater than zero with positive medians. Also, Angle shows more overestimated than underestimated.

Figure 3 Histogram for Compare Error in Display 1 and Display 2

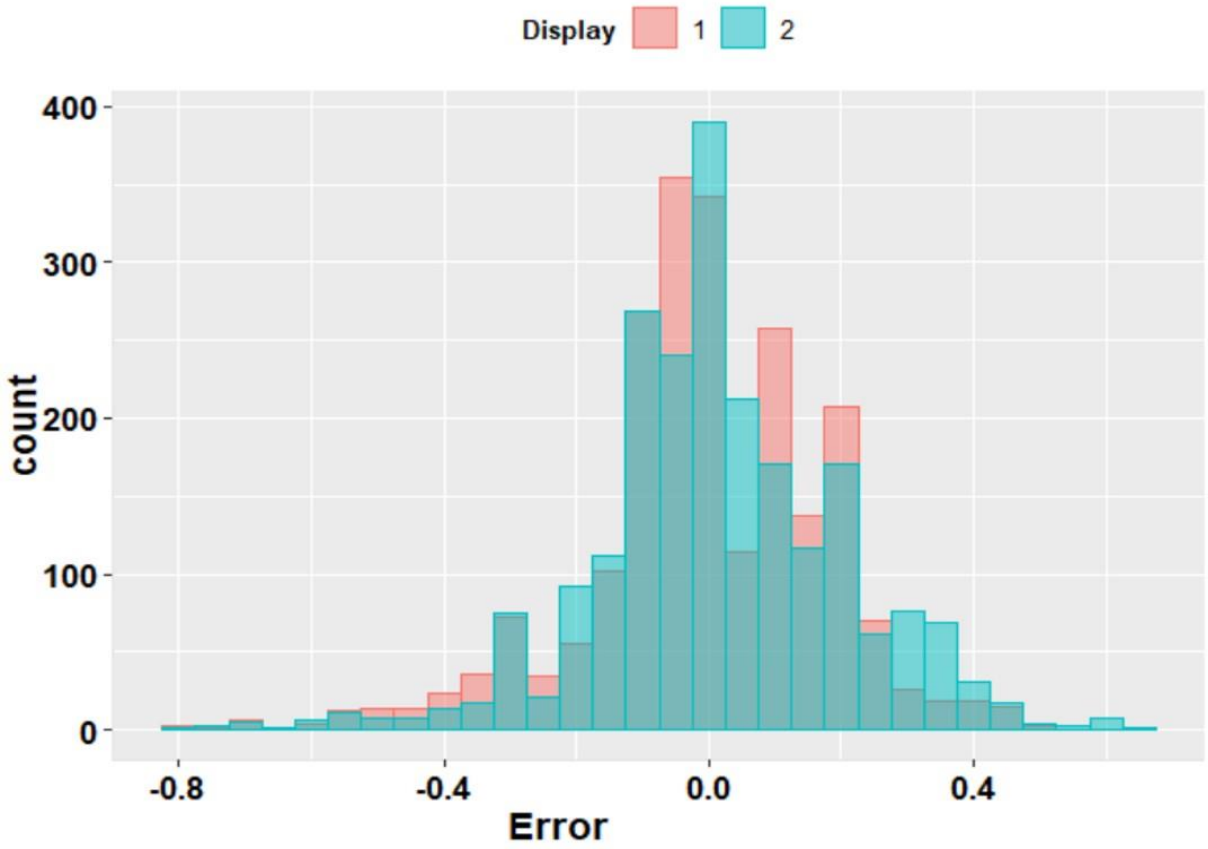
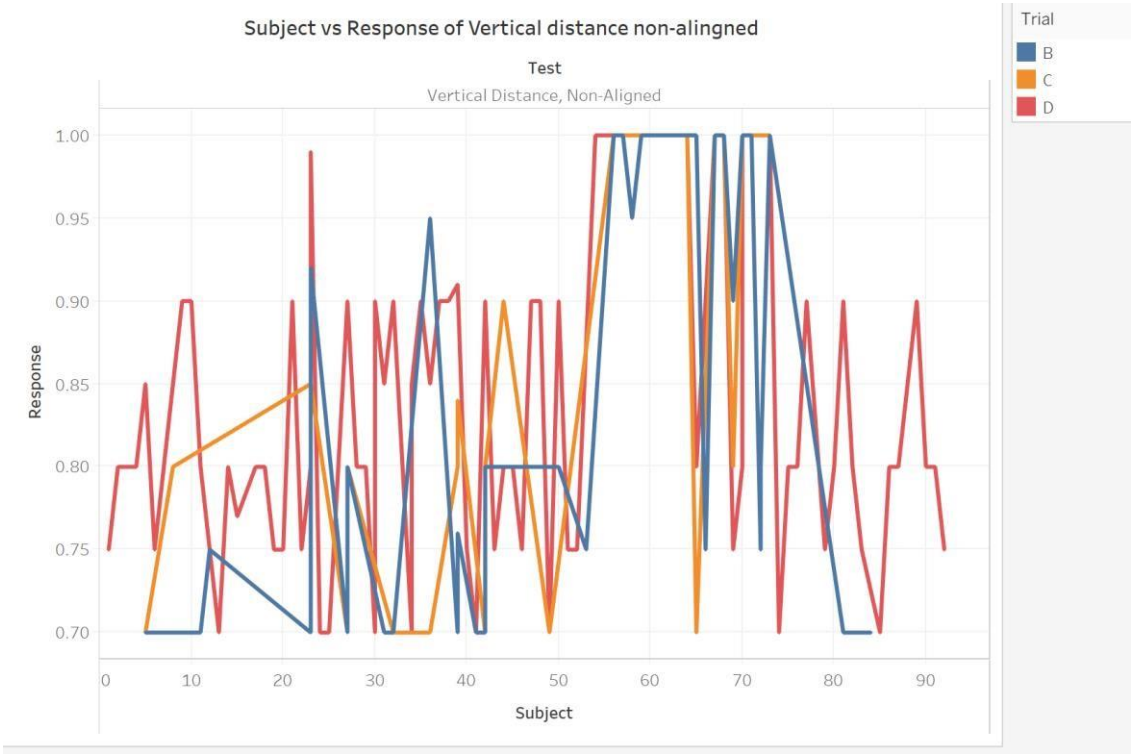


Figure 4 Display anomalous data



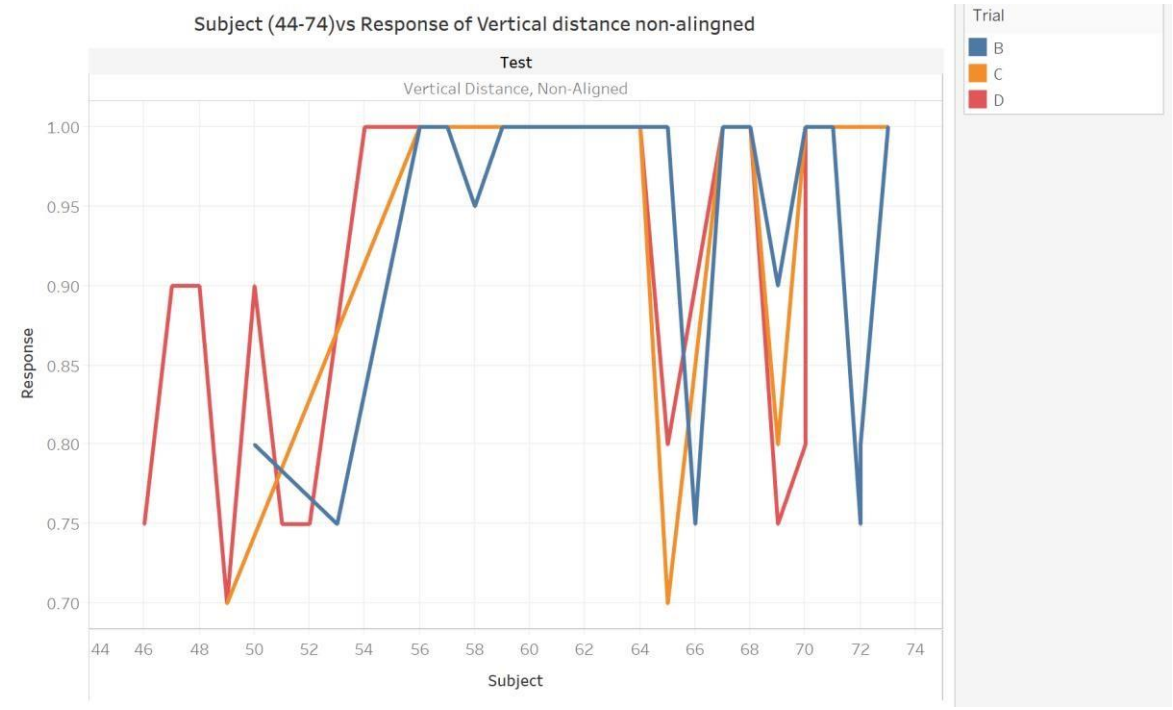


Figure 5 Absolute Error vs. Test violin

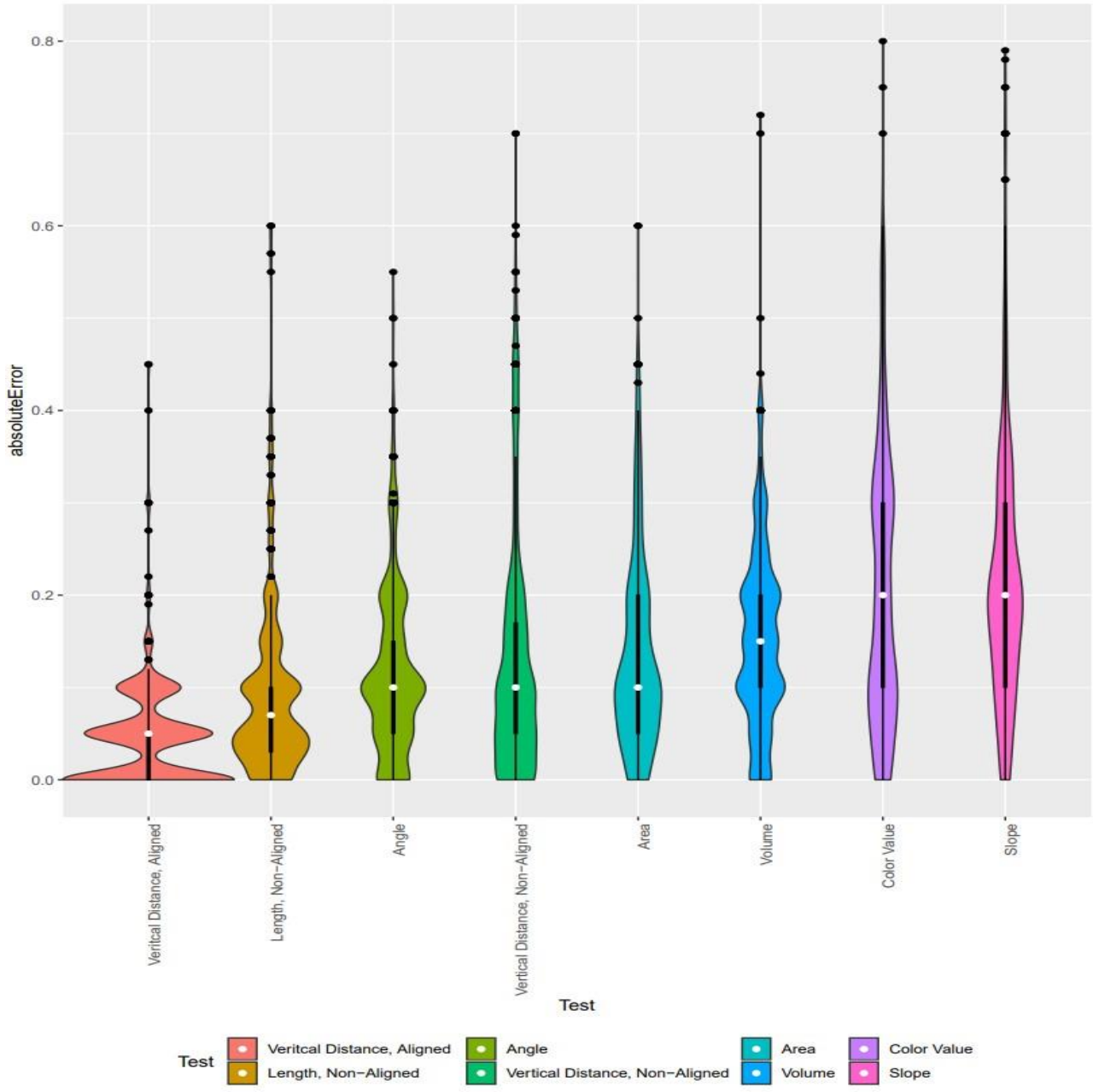


Figure 5 (the Absolute violin plot) shows that most data are clumped between 0 to 0.2. The Vertical Distance, the Aligned test produced the best result because the median which is about 0.05 is the smallest and at least 50% the data falls between 0 to 0.05 which is the smallest among all tests. The Slope and Color Value methods produced the most incorrected results because their

data medians are the highest and majority data falls in the range 0 - 0.4, which are longer than the others. as well as the outliers. The means of Angle, Vertical distance non-aligned, and Area are equal at 0.1.

Figure 6 Absolute Error vs. Test split by Display violin plot

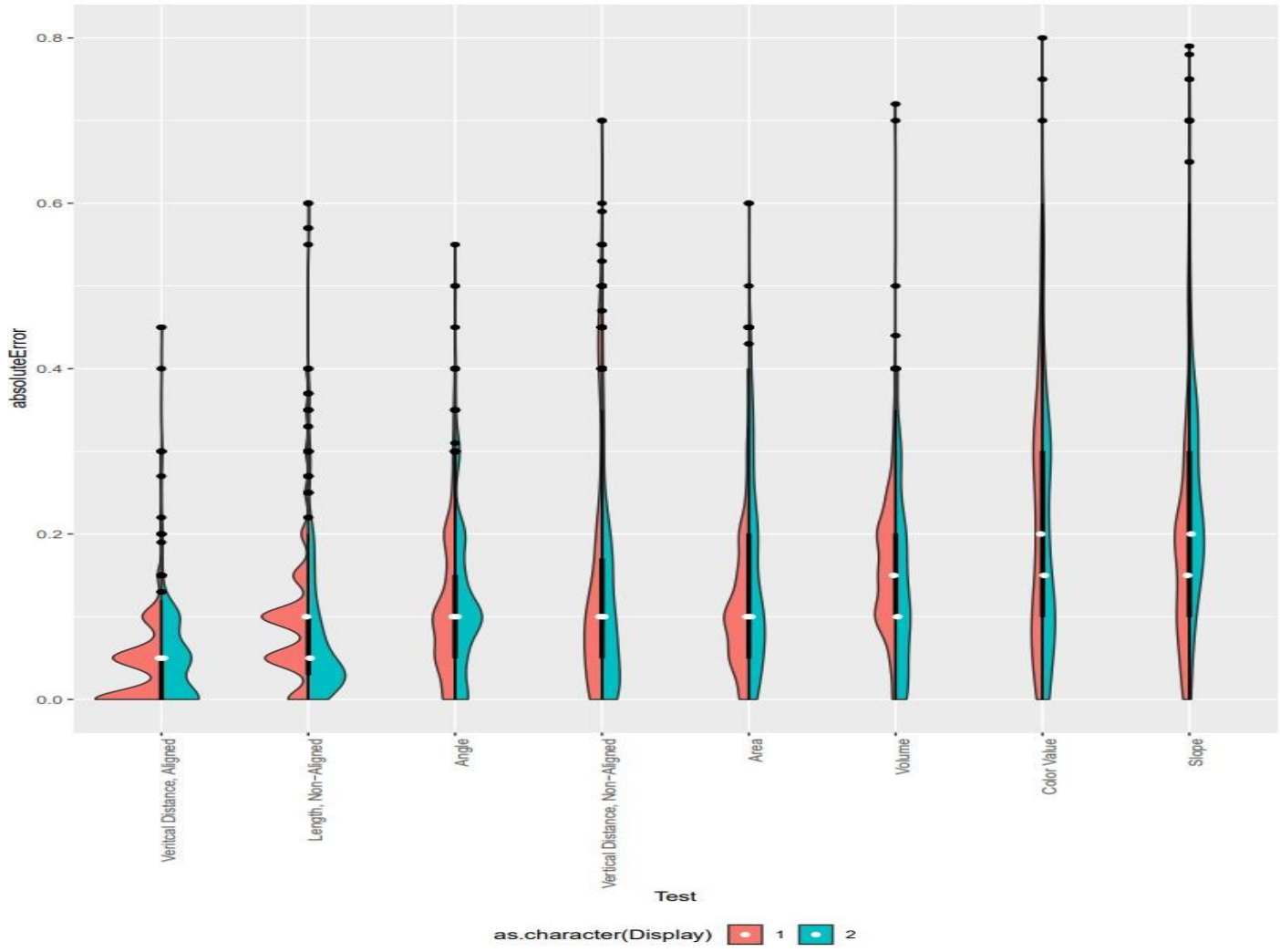


Figure 6(the split violin plot) shows that overall Display 2 produced better results than Display 1. The means of test Length Non-Aligned, Volm, and Color Value all decreased on the Display

2. There is no change of the means of Angle, Vertical Non-Aligned, and Area, but the mean of Slop increased from 0.15 to 0.2. The curve of the Length Non-Aligned on Display 2 become more normally distributed with a mean decrease from 0.1 to 0.05. The curve of the Vertical Distance Aligned also become more smooth than the Display 2. Also, the means of Color Value reduced from 0.2 to 0.15 after the second display. The mean of Volume test reduced from 0.15 to 0.1. Even though there is no change of mean of the Angle, but the curve becomes more smoothly centered around the means than Display1.

Figure 7 Error vs Test violin plot

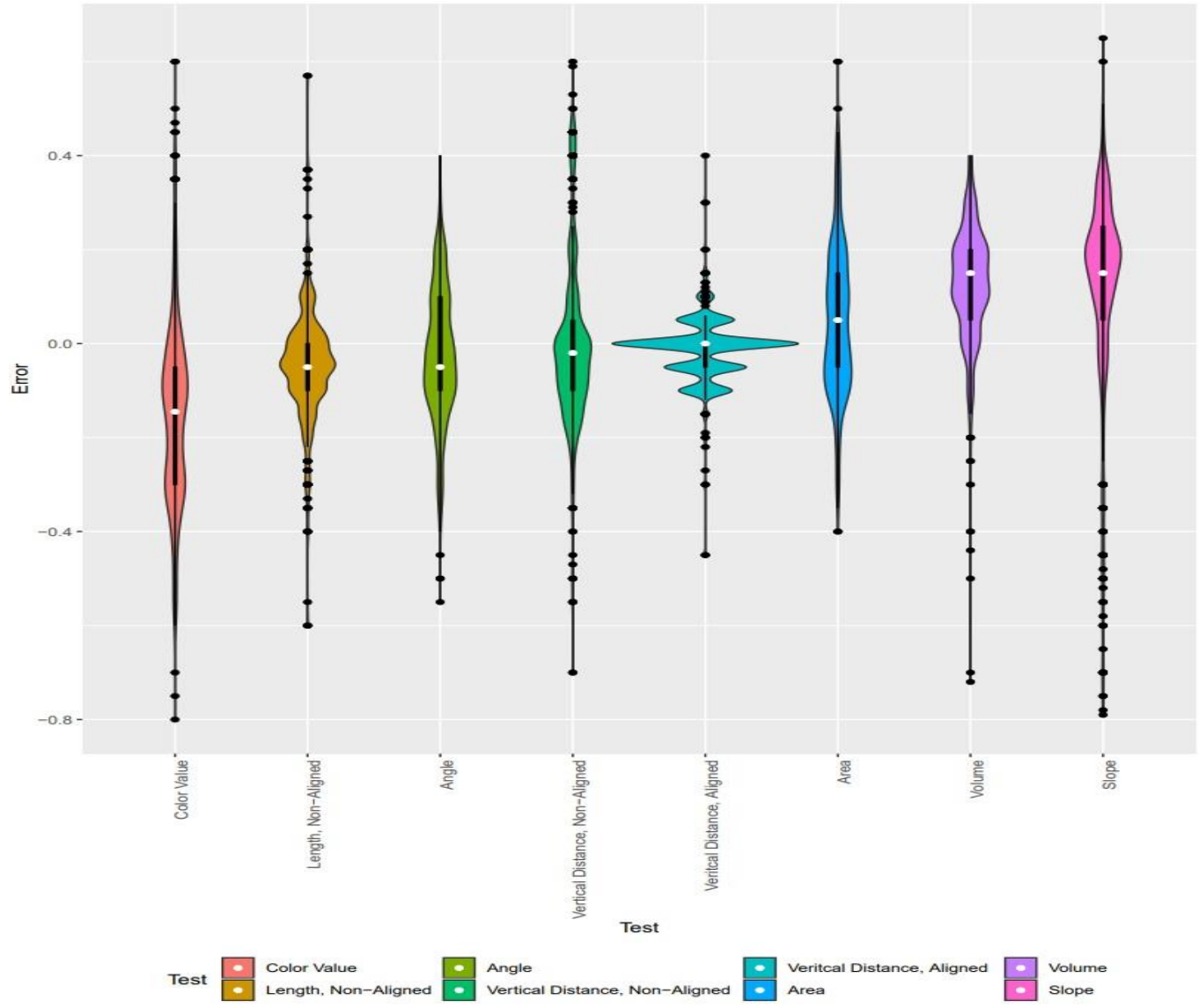


Figure 7 (the Error violin plot) shows more information than the jittered box plot because it shows more details about the distribution. The curve of the Vertical Distance aligned violin shows that there is the highest frequency around zero, which means it has the most accurate perception. Also, the Error violin shows more information about the data range from positive to negative. The median of the Vertical distance of Absolute Error in

Figure 5 is 0.05, which is greater the Error median, which is 0. The Slope data has the longest span than the others. Error violin plot shows more information about the distribution of negative and positive outliers than the Absolute Error plot.

Figure 8 Error vs. Test split by Display

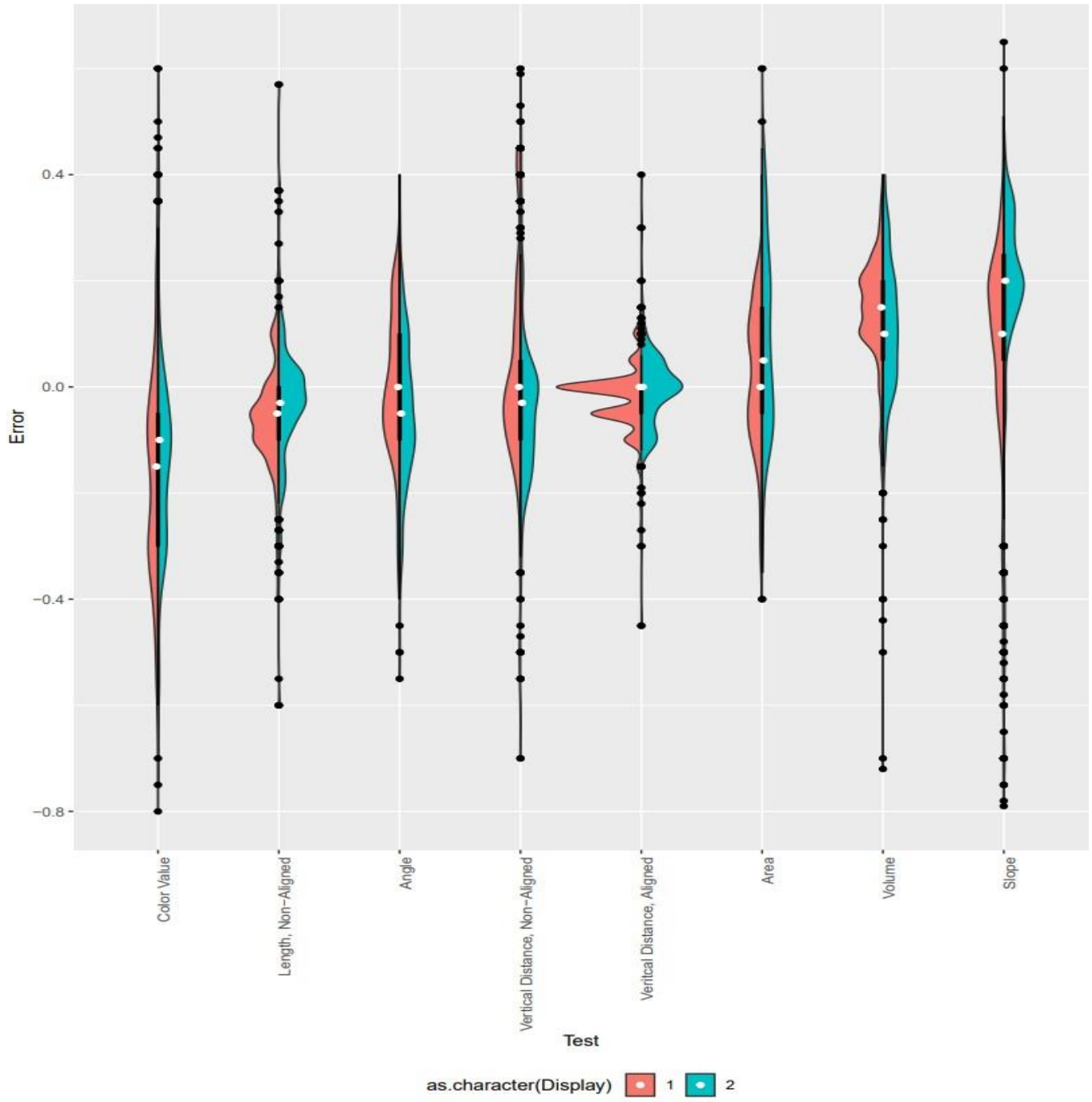


Figure 8 (the split violin plot) shows that the means of Color Value, Length Non-Aligned, and Volume all decreased on Display 2. There is no change of the mean of Angle, and there is some increasing of the means of Angle, Area. The mean of Slop increased the most, from 0.1 to 0.2. Also, the mean of Color Value reduced from 0.2 to 0.15 after the second display.

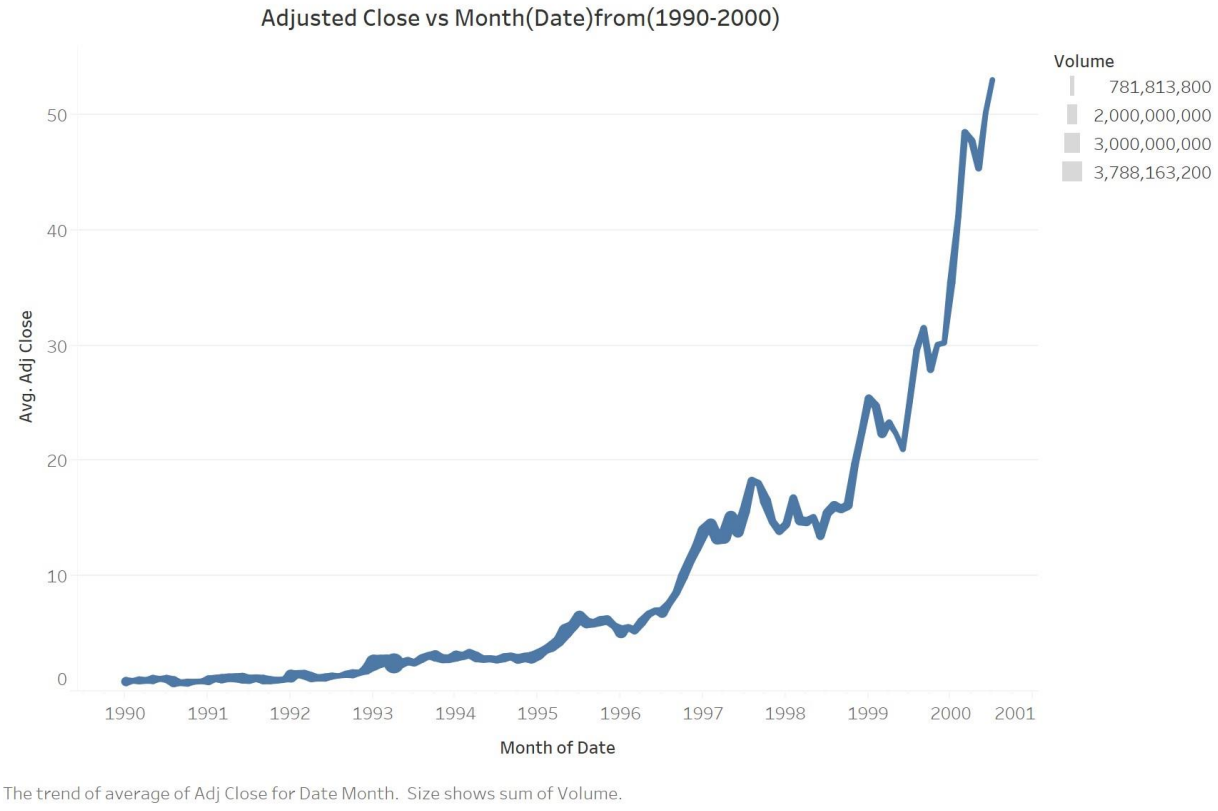
R code for violin and box plot:

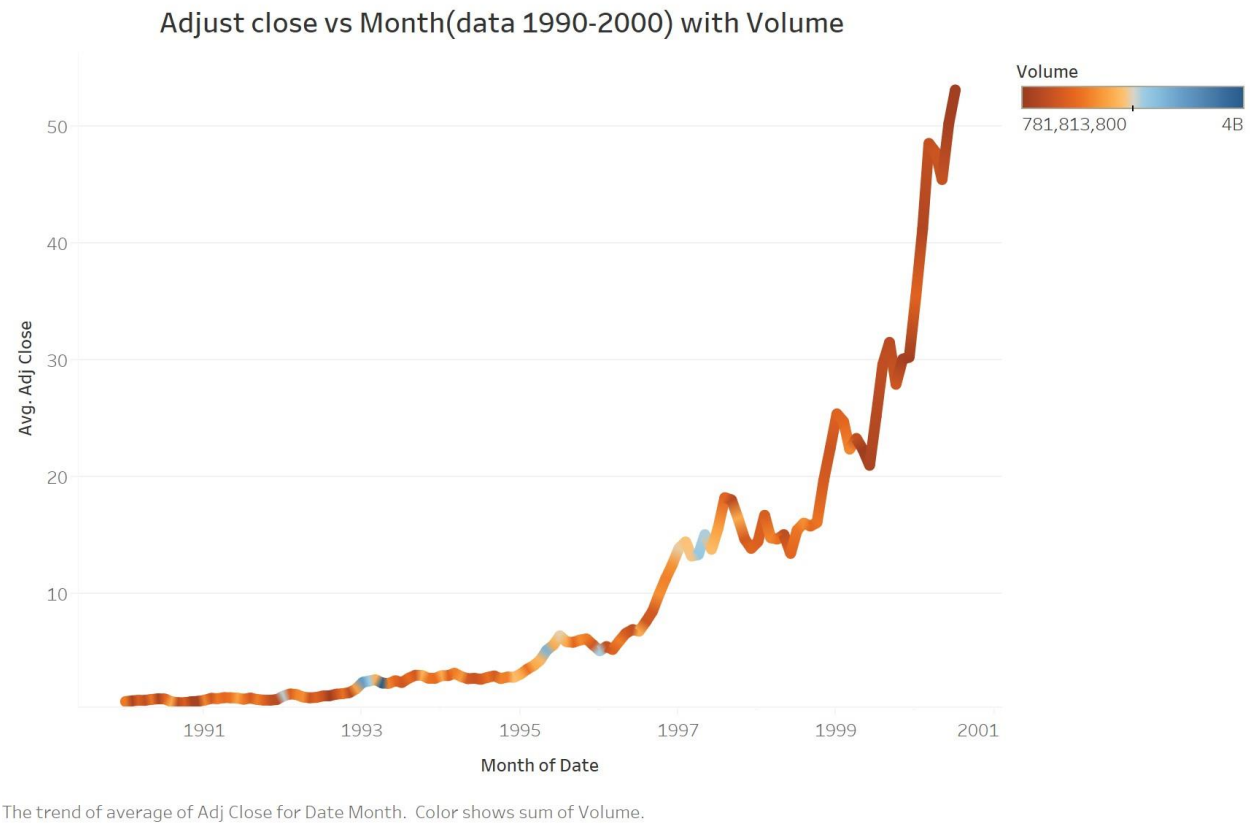
```
setwd("C:/Users/Yingping Li/Documents/csc465/hw1/Datasets")

percep = read.table("PerceptionExperiment.csv", header=T, sep=',') head(percep)
percep$Error <- percep$Response - percep$TrueValue
write.csv(percep, file = "percepNew.csv")

head(percep)
percep$absoluteError<-abs(percep$Error)

# Error Median sort---------------------------------------------- percep$Test
<- reorder(percep$Test, percep$Error, median)

# Error boxplot----------------------------------------------
ggplot(percep,aes(Test,Error))+geom_boxplot()+geom_point(aes(color=Test),alpha = 0.1,positio
n=position_jitter(width=0.1,height=0.01))+ theme(legend.position="bottom",axis.text.x = eleme
nt_text(angle = 90, hjust = 1))

# Error violin plot----------------------------------------------
ggplot(percep, aes(x=Test, y=Error, fill=Test,width=1.4)) + geom_violin() +  geom_boxplot(col
or="black", fill="black",width=0.02) + theme(legend.position="bottom",axis.text.x = element_te
xt(angle = 90, hjust = 1)) + stat_summary(fun.y = median, fun.ymin = median, fun.ymax = medi
an,geom = "point", color="white",width = 0.02)

# Error violin split plot----------------------------------------------
ggplot(percep, aes(x=Test, y=Error, fill=as.character(Display),width=1.4)) + geom_split_violin()
+  geom_boxplot(color="black", fill="black",width=0.02,position = position_dodge(width=0.0
2)) + theme(legend.position="bottom",axis.text.x = element_text(angle = 90, hjust = 1)) + stat_s
ummary(fun.y = median, fun.ymin = median, fun.ymax = median,geom = "point", color="white
",width = 0.05,position = position_dodge(width=0.05))


# Absolute Median sort---------------------------------------------- percep$Test
<- reorder(percep$Test, percep$absoluteError, median)

# AbsoluteError box plot----------------------------------------------
ggplot(percep,aes(Test,absoluteError))+geom_boxplot()+geom_point(aes(color=Test),alpha = 0.
1,position=position_jitter(width=0.1,height=0.01)) + theme(legend.position="bottom",axis.text.x
= element_text(angle = 90, hjust = 1))

# AbsoluteError Violin plot----------------------------------------------
```

```
ggplot(percep, aes(x=Test, y=absoluteError, fill=Test, width=1.4)) + geom_violin() + geom_box
plot(color="black", fill="black",width=0.02) + theme(legend.position="bottom",axis.text.x = ele
ment_text(angle = 90, hjust = 1)) + stat_summary(fun.y = median, fun.ymin = median, fun.ymax
= median,geom = "point", color="white",width = 0.02)

# AbsoluteError Violin split plot---------- ggplot(percep, aes(x=Test,
y=absoluteError, fill=as.character(Display))) + geom_split_violin() +
geom_boxplot(color="black", fill="black",width=0.02) + theme(legend.position="bottom",axis.t
ext.x = element_text(angle = 90, hjust = 1)) + stat_summary(fun.y = median, fun.ymin = median,
fun.ymax = median,geom = "point", color="white",width = 0.05,position = position_dodge(widt
h=0.05))
#Split Violin Function----------
GeomSplitViolin <- ggproto("GeomSplitViolin", GeomViolin,
                    draw_group = function(self, data, ..., draw_quantiles = NULL) {
                      data <- transform(data, xminv = x - violinwidth * (x - xmin), xmaxv = x + viol
inwidth * (xmax - x))
                      grp <- data[1, "group"]
                      newdata <- plyr::arrange(transform(data, x = if (grp %% 2 == 1) xminv else x
maxv), if (grp %% 2 == 1) y else -y)
                      newdata <- rbind(newdata[1, ], newdata, newdata[nrow(newdata), ], newdata
[1, ])
                      newdata[c(1, nrow(newdata) - 1, nrow(newdata)), "x"] <- round(newdata[1, "x
"])

                      if (length(draw_quantiles) > 0 & !scales::zero_range(range(data$y))) {
stopifnot(all(draw_quantiles >= 0), all(draw_quantiles <=
                                                  1))
                        quantiles <- ggplot2:::create_quantile_segment_frame(data, draw_quantiles)
aesthetics <- data[rep(1, nrow(quantiles)), setdiff(names(data), c("x", "y")), dr op = FALSE]
                        aesthetics$alpha <- rep(1, nrow(quantiles))
both <- cbind(quantiles, aesthetics)
                        quantile_grob <- GeomPath$draw_panel(both, ...)
                        ggplot2:::ggname("geom_split_violin", grid::grobTree(GeomPolygon$draw_
panel(newdata, ...), quantile_grob))
                      }
else {
                        ggplot2:::ggname("geom_split_violin", GeomPolygon$draw_panel(newdata,
 ...))
                      }
                    })

geom_split_violin <- function(mapping = NULL, data = NULL, stat = "ydensity", position = "id
entity", ...,
```

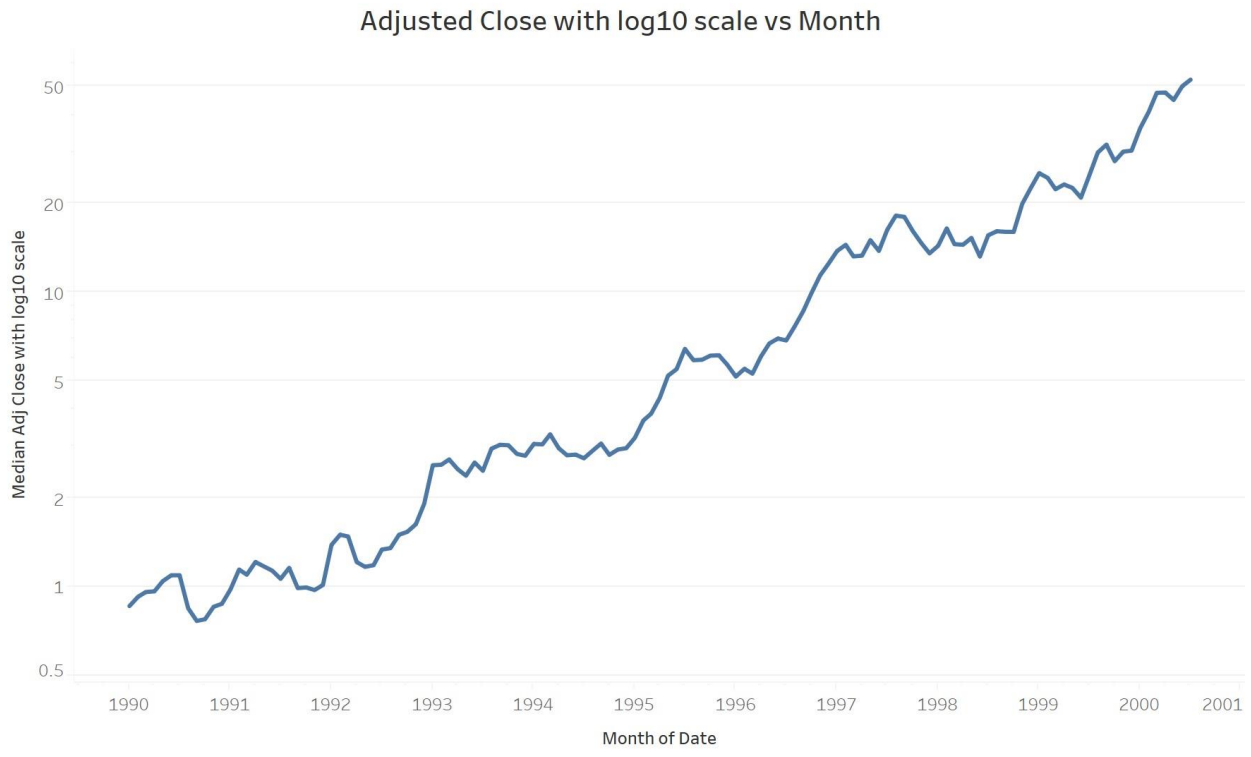
```
                draw_quantiles = NULL, trim = TRUE, scale = "area", na.rm = FALSE,
show.legend = NA, inherit.aes = TRUE) {
  layer(data = data, mapping = mapping, stat = stat, geom = GeomSplitViolin,
position = position, show.legend = show.legend, inherit.aes = inherit.aes,          params
= list(trim = trim, scale = scale, draw_quantiles = draw_quantiles, na.rm = na.rm,
 ...))
}
```

stock data analysis



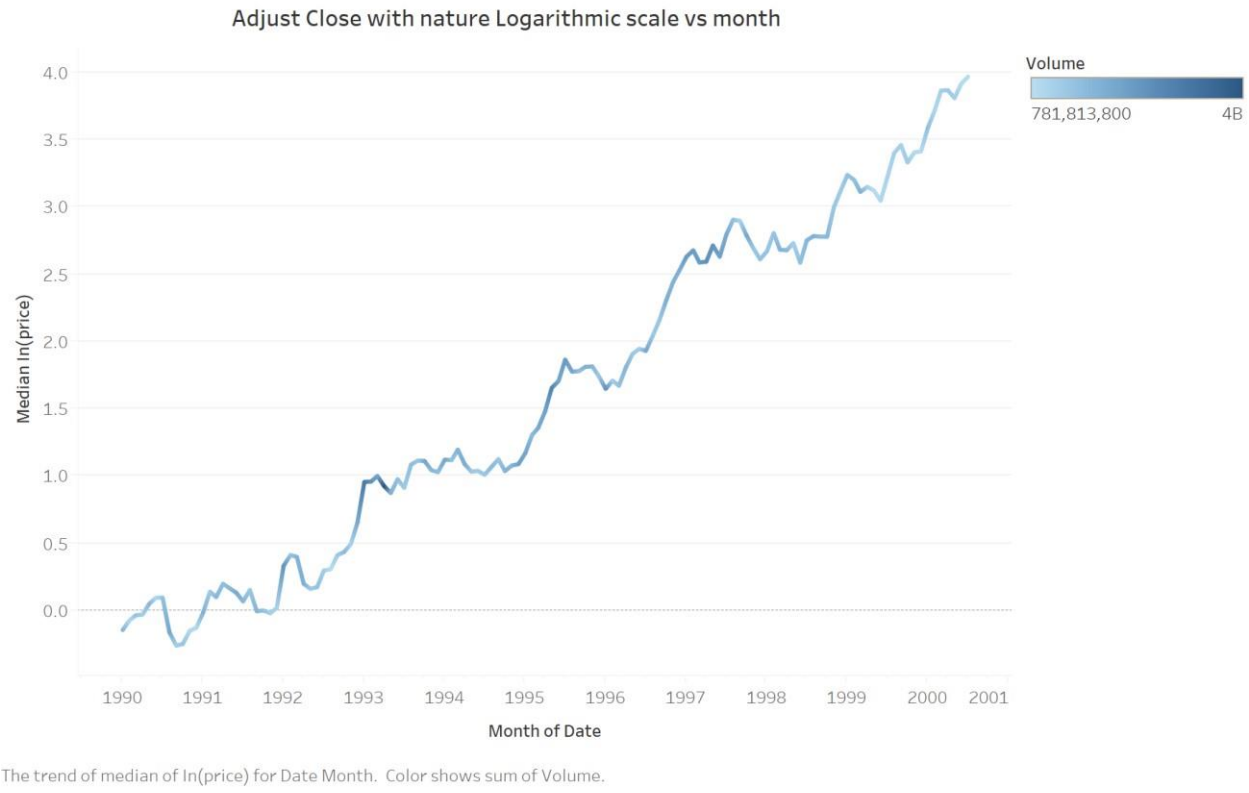


I think graph b can communicate the Volume data more effiectively. The color can express the gradual change of the Volume more clearly than the thickness of the line, but the thickness of the line cannot very clearly to display the small change of the volume.



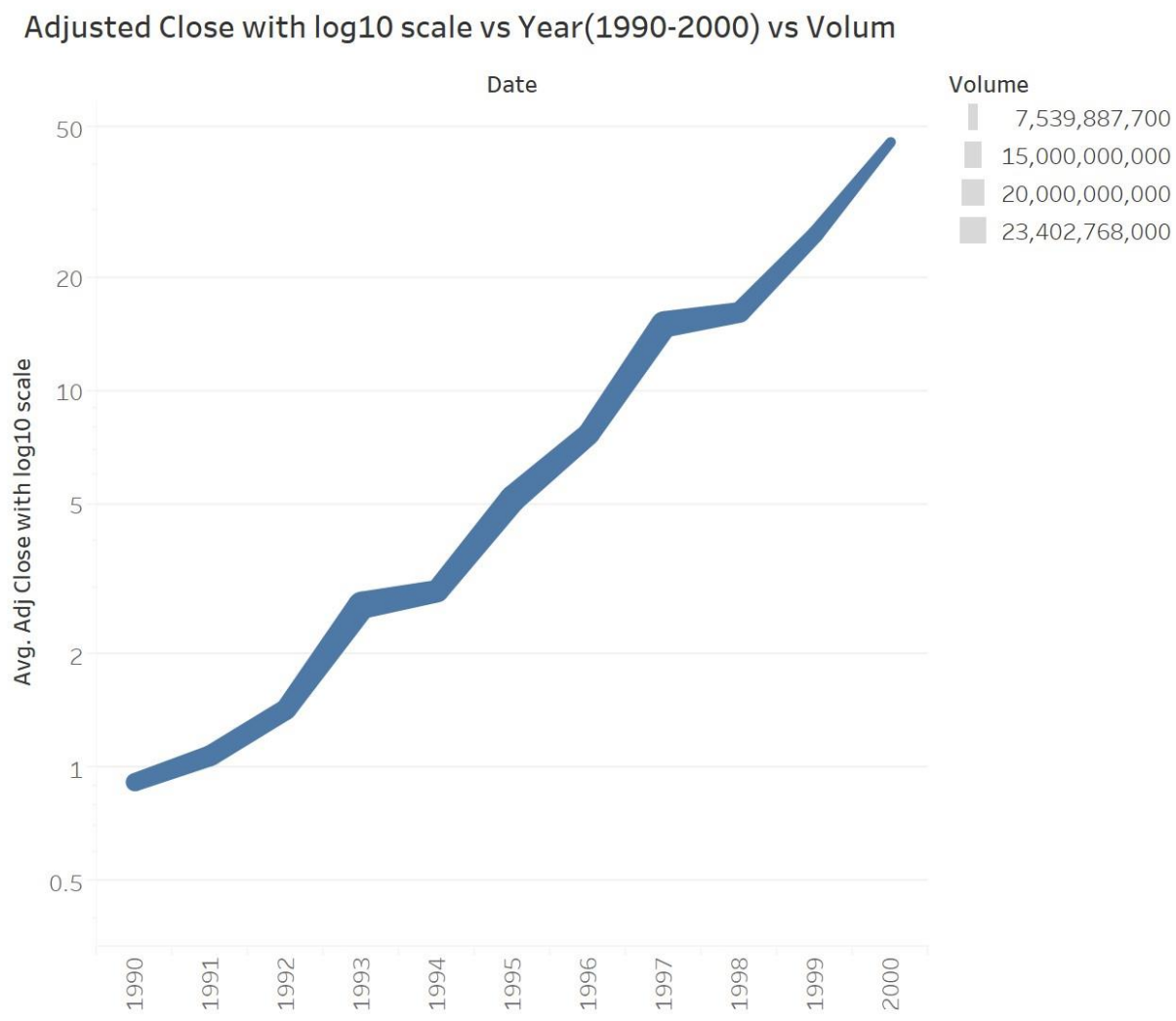
The trend of median of Adj Close for Date Month.

The graph with logarithmic scale becomes approximately a line instead of the curve shape of the standared line gragh, and the logarithmic scale highlighted each fluctuation.

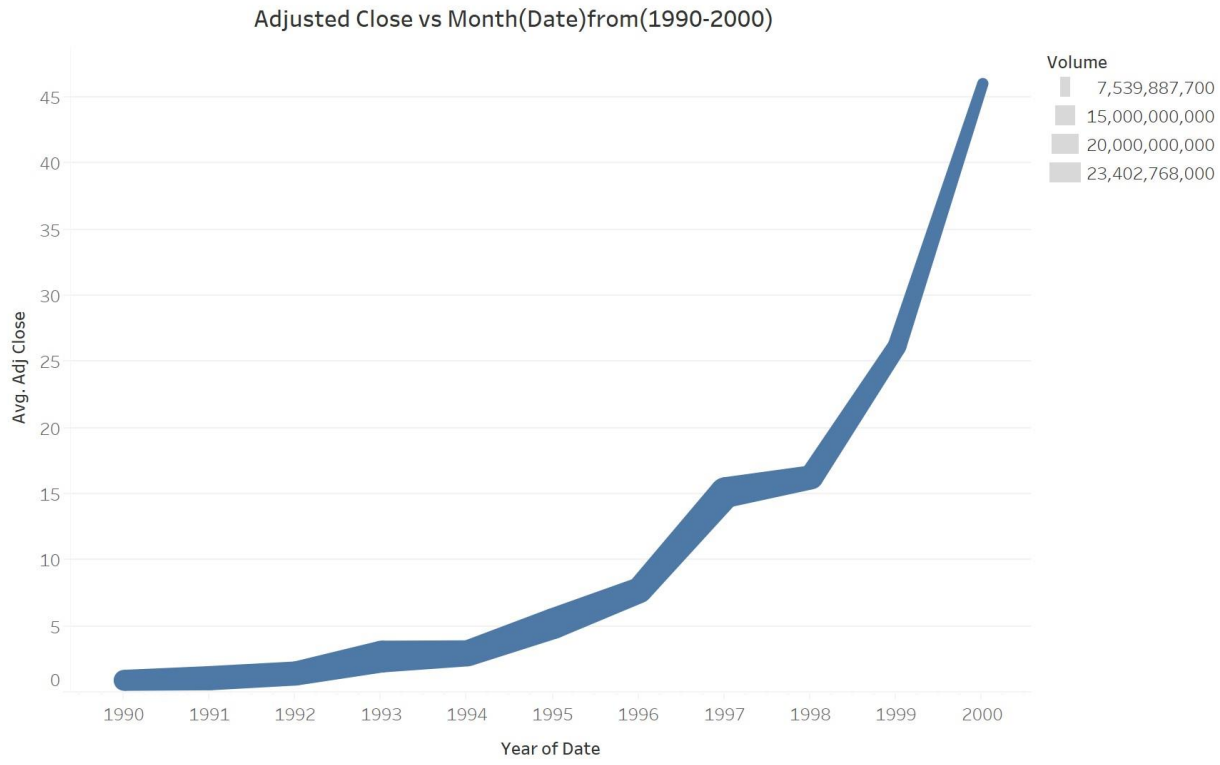


The three surges in price that are between 15% and 20% are May 1992 to March 1993, October 1994 to July 1995, and July 1996 to February 1997.

At the 1991 year point over these years was the yearly%-wise change the fastest.

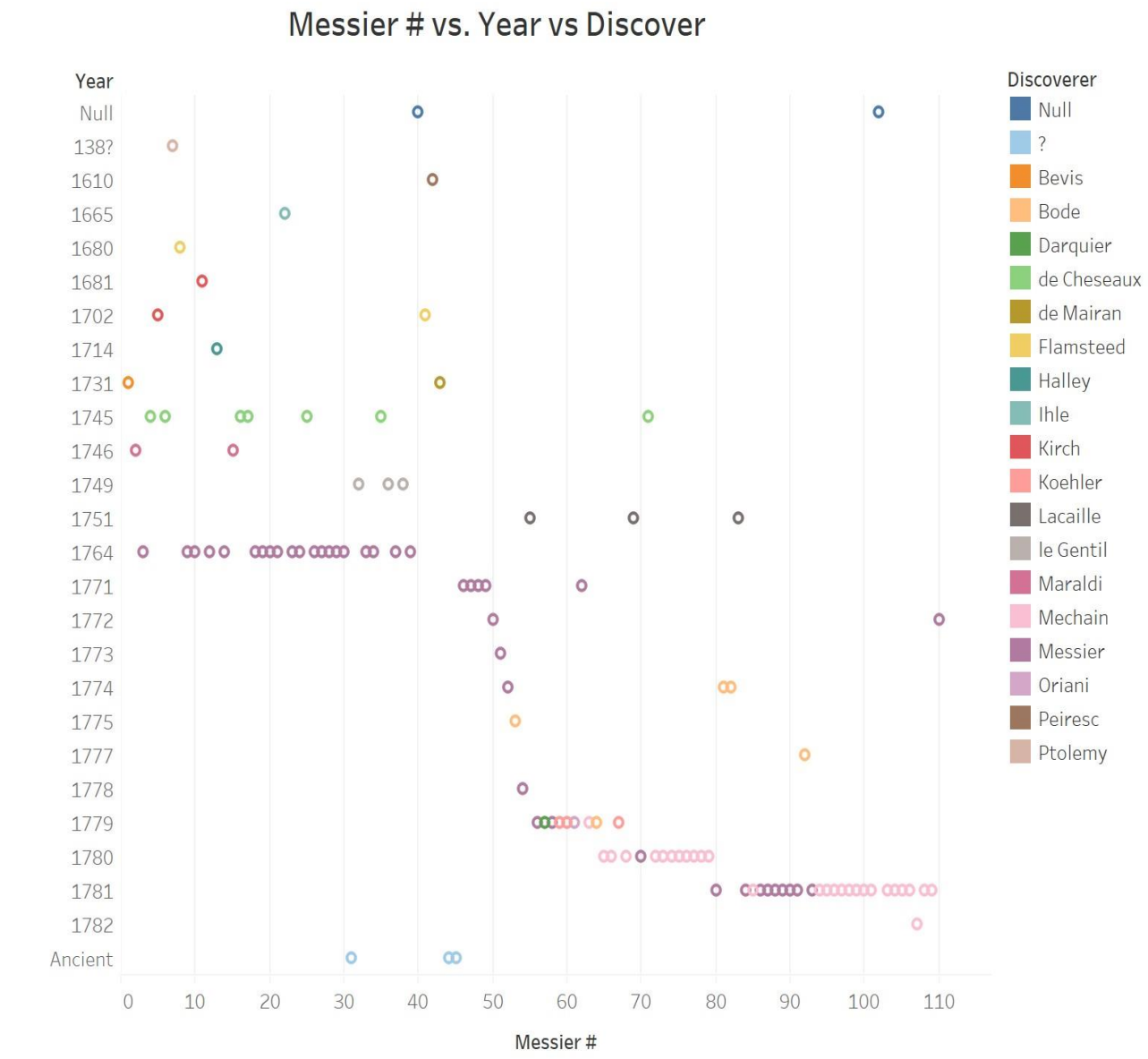


The trend of average of Adj Close for Date Year. Size shows sum of Volume.

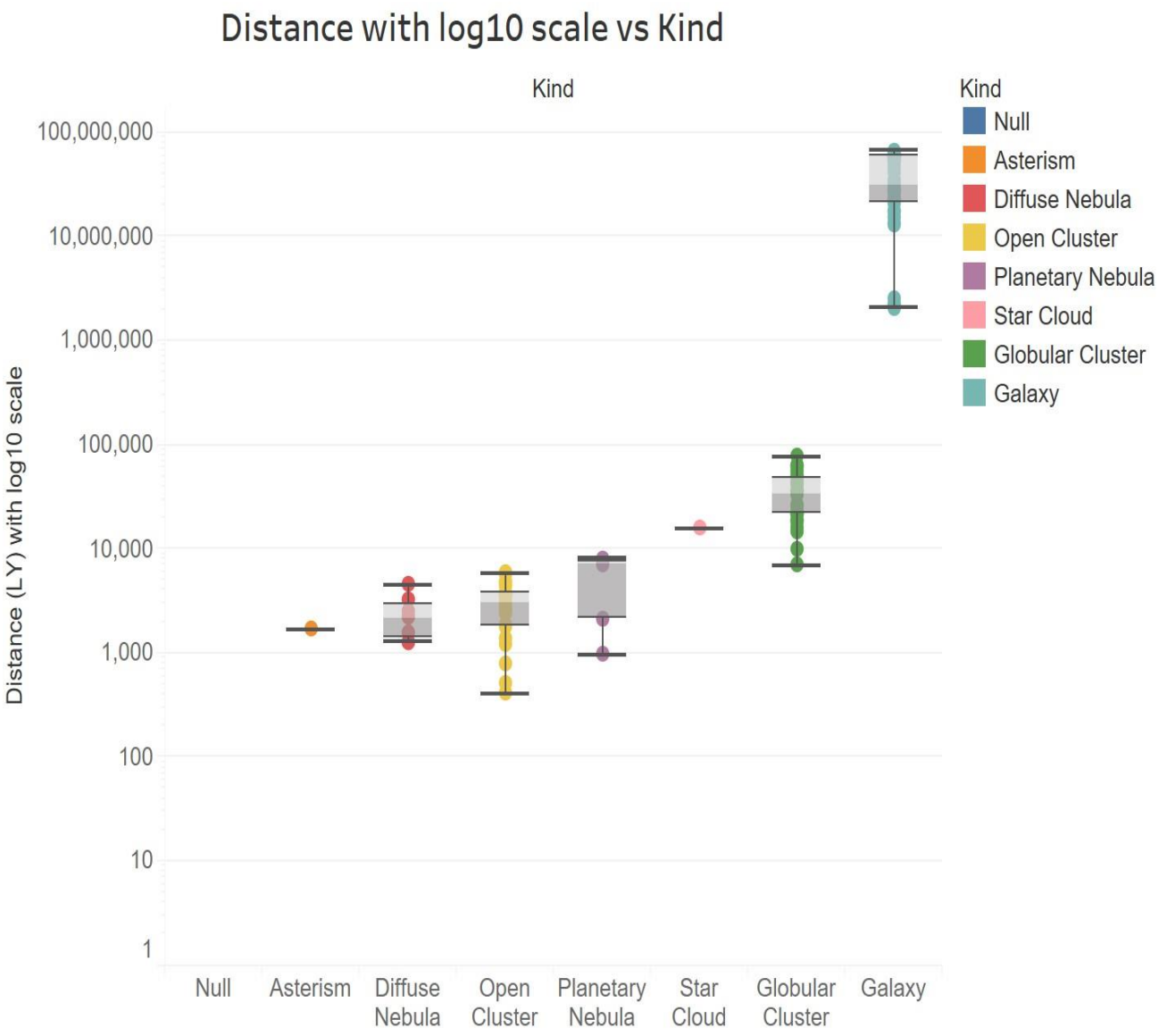


It workd much better when the curve has fewer data points.

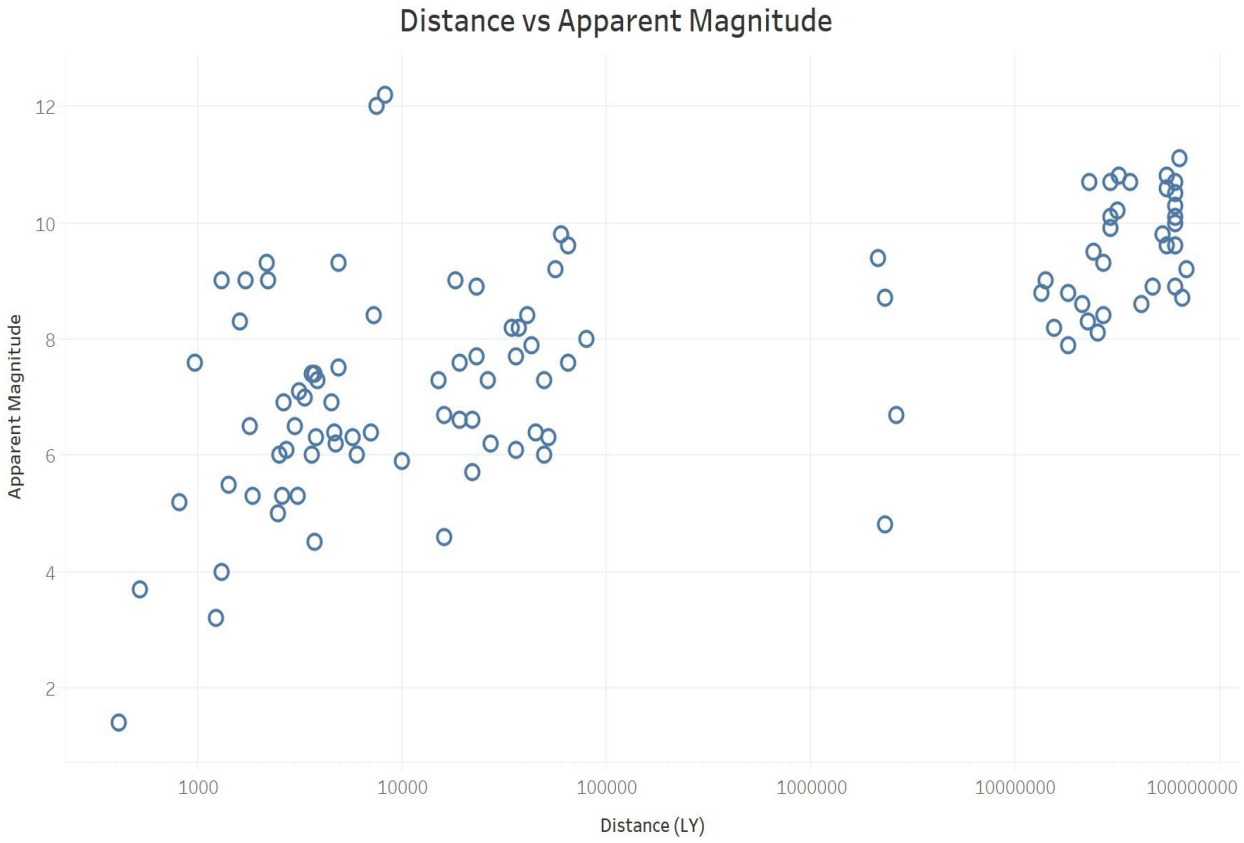
The astronomical data set analysis



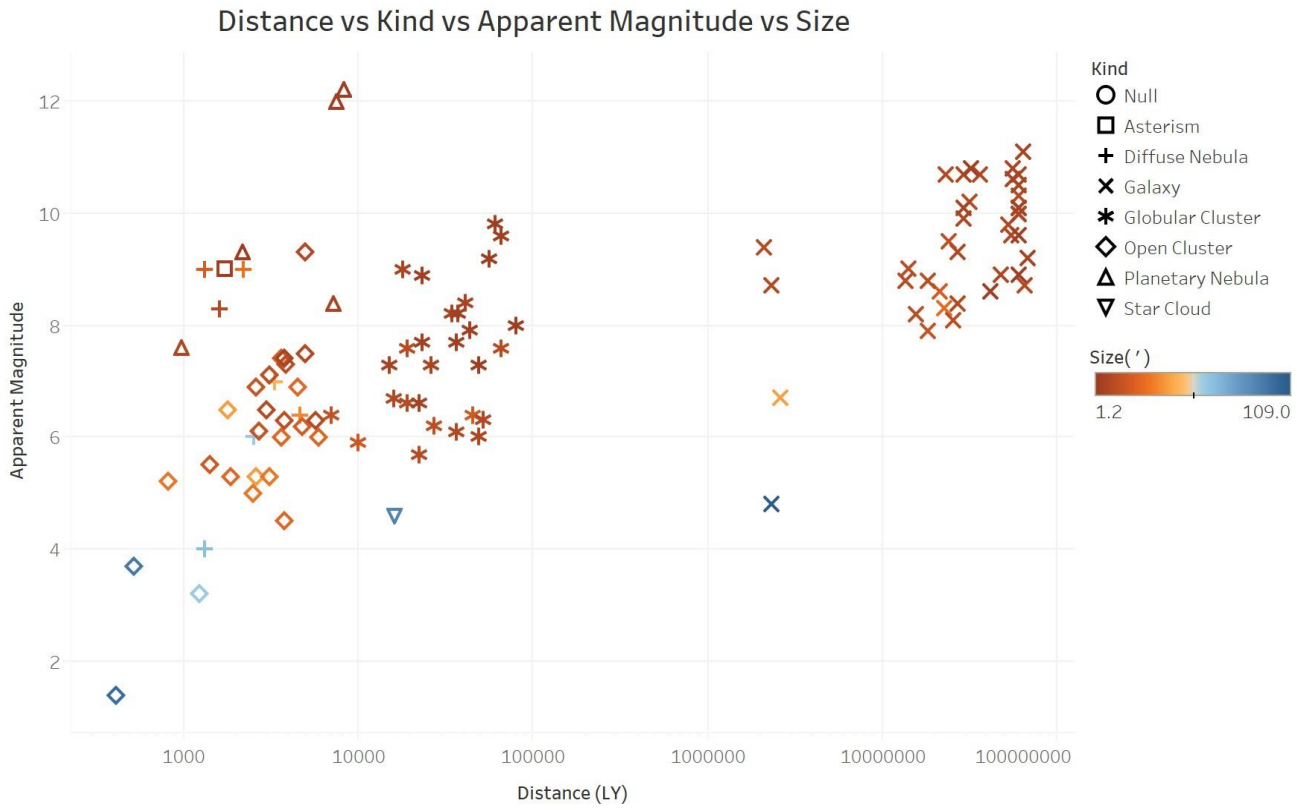
Messier # for each Year. Color shows details about Discoverer.



Median Distance (LY) for each Kind. Color shows details about Kind.



Distance (LY) vs. Apparent Magnitude.



Distance (LY) vs. Apparent Magnitude. Color shows details about Size('). Shape shows details about Kind.