Unilever Case

Nurefsan Davulcu

2022-03-14

Part 1: Data Import, Cleaning and Merge

Data Import

```r
volume_data <- read_xlsx('volume_data.xlsx')
product_data <- read_xlsx('product_data.xlsx')

str(volume_data)
```

```
## tbl_df [51,557 x 8] (S3: tbl_df/tbl/data.frame)
##  $ Ex Factory Start Week: num [1:51557] 201701 201701 201701 201701 201716 ...
##  $ # Ex-Factory Weeks   : num [1:51557] 2 2 2 2 2 2 2 2 2 2 ...
##  $ Year                 : chr [1:51557] "2017" "2017" "2017" "2017" ...
##  $ Scanning volume      : num [1:51557] 50428 620719 299081 88404 26063 ...
##  $ Ex Factory volume    : num [1:51557] 84000 784476 351072 60480 42000 ...
##  $ Promo Name           : chr [1:51557] "Elliot O'Brien" "Elliot O'Brien" "Elliot O'Brien" "Elliot O
##  $ Plan Accounts        : chr [1:51557] "Megan Howard" "Megan Howard" "Megan Howard" "Megan Howard"
##  $ MRDR                 : chr [1:51557] "MCCI52734221488349" "NAAF07835189817320" "YVWD4592258431021
```

```r
str(product_data)
```

```
## tbl_df [2,593 x 5] (S3: tbl_df/tbl/data.frame)
##  $ Cluster     : chr [1:2593] "Cluster 1" "Cluster 1" "Cluster 1" "Cluster 1" ...
##  $ Category    : chr [1:2593] "Category 1" "Category 1" "Category 1" "Category 1" ...
##  $ Brand       : chr [1:2593] "Brand 1" "Brand 1" "Brand 1" "Brand 1" ...
##  $ Product Name: chr [1:2593] "Kieran Bishop" "Jennifer Smith" "Miss Abbie Holland" "Dr. Irene Hart"
##  $ MRDR        : chr [1:2593] "MCCI52734221488349" "NAAF07835189817320" "YVWD45922584310213" "ZQYM21
```

```r
head(volume_data)
```

```
## # A tibble: 6 x 8
##   `Ex Factory St~` `# Ex-Factory ~` Year  `Scanning volu~` `Ex Factory vo~`
##              <dbl>            <dbl> <chr>            <dbl>            <dbl>
## 1           201701                2 2017             50428            84000
## 2           201701                2 2017            620719           784476
## 3           201701                2 2017            299081           351072
## 4           201701                2 2017             88404            60480
## 5           201716                2 2017             26063            42000
## 6           201716                2 2017            191005           277200
## # ... with 3 more variables: `Promo Name` <chr>, `Plan Accounts` <chr>,
## #   MRDR <chr>
```

```
head(product_data)
```

```
## # A tibble: 6 x 5
##   Cluster   Category   Brand   `Product Name`     MRDR
##   <chr>     <chr>      <chr>   <chr>              <chr>
## 1 Cluster 1 Category 1 Brand 1 Kieran Bishop      MCCI52734221488349
## 2 Cluster 1 Category 1 Brand 1 Jennifer Smith     NAAF07835189817320
## 3 Cluster 1 Category 1 Brand 1 Miss Abbie Holland YVWD45922584310213
## 4 Cluster 1 Category 1 Brand 1 Dr. Irene Hart     ZQYM21358460104493
## 5 Cluster 1 Category 1 Brand 1 Stewart Wood       YUSO25895760670703
## 6 Cluster 1 Category 1 Brand 1 Miss Laura May     XUMC70001280070651
```

Product Data Cleaning

```
# check duplicates - MRDRs should be unique
length(unique(product_data$MRDR)) #2576 hmm this should theoretically be 2593.
```

```
## [1] 2576
```

```
length(unique(volume_data$MRDR)) # 2576
```

```
## [1] 2576
```

```
# look at duplicates
id1 <- which(duplicated(product_data$MRDR))
product_data[id1,]
```

```
## # A tibble: 17 x 5
##    Cluster   Category   Brand   `Product Name`         MRDR
##    <chr>     <chr>      <chr>   <chr>                  <chr>
##  1 Cluster 1 Category 4 Brand 2 Derek Morris           OZKU05070811942737
##  2 Cluster 1 Category 7 Brand 7 Natasha Fox            OZKU05070811942737
##  3 Cluster 1 Category 4 Brand 2 Louise Gray            OZKU05070811942737
##  4 Cluster 1 Category 7 Brand 7 Connor Hayes           OZKU05070811942737
##  5 Cluster 1 Category 7 Brand 7 Dr. Maurice Smith      OZKU05070811942737
##  6 Cluster 1 Category 7 Brand 7 Dr. Martyn Lynch       OZKU05070811942737
##  7 Cluster 1 Category 7 Brand 7 Mr. Dominic Mann       OZKU05070811942737
##  8 Cluster 1 Category 7 Brand 7 Shane Holland-Booth    OZKU05070811942737
##  9 Cluster 1 Category 7 Brand 7 Mr. Duncan Thomas      OZKU05070811942737
## 10 Cluster 1 Category 7 Brand 7 Alexandra Lewis        OZKU05070811942737
## 11 Cluster 1 Category 7 Brand 7 Charles McDonald       OZKU05070811942737
## 12 Cluster 1 Category 7 Brand 7 Dr. Janice Coates      OZKU05070811942737
## 13 Cluster 1 Category 7 Brand 7 Gillian Taylor-Smith   OZKU05070811942737
## 14 Cluster 1 Category 7 Brand 7 Dr. Barry Harris       OZKU05070811942737
## 15 Cluster 1 Category 7 Brand 7 Mrs. Vanessa Duffy     OZKU05070811942737
## 16 Cluster 1 Category 7 Brand 7 Mohammed Young         OZKU05070811942737
## 17 Cluster 1 Category 7 Brand 7 Jodie Chambers-Bradley OZKU05070811942737
```

```r
# check 0ZKU05070811942737 in volume data
id2 <- which(volume_data$MRDR=='0ZKU05070811942737')
volume_data[id2,]
```

```
## # A tibble: 18 x 8
##    `Ex Factory St~` `# Ex-Factory ~` Year  `Scanning volu~` `Ex Factory vo~`
##               <dbl>            <dbl> <chr>             <dbl>            <dbl>
##  1           201909                2 2019               7510             9504
##  2           201725                2 2017              39265            38880
##  3           201937                2 2019              39440            40820
##  4           201904                2 2019              41040            41040
##  5           201808                2 2018               7462            18924
##  6           201808                2 2018               5536            14040
##  7           201808                2 2018               9455            23976
##  8           201808                2 2018               6955            17634
##  9           201808                2 2018              15136            38376
## 10           201808                2 2018               6853            17376
## 11           201808                2 2018               5626            14268
## 12           201808                2 2018               4684            11880
## 13           201808                2 2018                406             1032
## 14           201808                2 2018               4110            10422
## 15           201808                2 2018               3821             9690
## 16           201808                2 2018                336              852
## 17           201808                2 2018                524             1326
## 18           202007                2 2020               3600             570.
## # ... with 3 more variables: `Promo Name` <chr>, `Plan Accounts` <chr>,
## #   MRDR <chr>
```

```r
#Conclusion: MRDR 0ZKU05070811942737 has multiple product names associated (likely error), we will just
# as we don't want the same volume data repeated when we merge
product_data<-product_data[-id1,]
length(unique(product_data$MRDR))  # now 2576 as expected
```

```
## [1] 2576
```

Data Merge

```r
# merge on MRDR
dat <- merge(product_data, volume_data, by = 'MRDR', all=FALSE)
```

Data Cleaning - Merged Data

```r
# format variables
dat$Cluster <- factor(dat$Cluster)
dat$Category <- factor(dat$Category)
dat$Brand <- factor(dat$Brand)
#dat$Year <- factor(dat$Year)
dat$Year <- as.numeric(dat$Year)
```

```r
dat$`Plan Accounts`<- factor(dat$`Plan Accounts`)

summary(dat$`# Ex-Factory Weeks`)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   2.000   2.366   2.000   9.000
```

```r
summary(dat$Cluster)
```

```
## Cluster 1 Cluster 2
##     36332     15225
```

```r
summary(dat$Category)
```

```
##  Category 1 Category 10 Category 11  Category 2  Category 3  Category 4
##         965        4027        2053          60        4820       10050
##  Category 5  Category 6  Category 7  Category 8  Category 9
##        3275        4974       12188        3617        5528
```

```r
summary(dat$Brand)
```

```
##  Brand 1 Brand 10 Brand 11 Brand 12 Brand 13 Brand 14 Brand 15 Brand 16
##     5065     3085      117      323      249        9     3036      179
## Brand 17 Brand 18 Brand 19  Brand 2 Brand 20 Brand 21 Brand 22  Brand 3
##      358      651     8285     8985      110     2025       24        3
##  Brand 4  Brand 5  Brand 6  Brand 7  Brand 8  Brand 9
##      717     1148      199    11489      226     5274
```

```r
summary(dat$Year)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2016    2018    2019    2019    2020    2021
```

```r
summary(dat$`Plan Accounts`)
```

```
##  Megan Howard    Oliver Fry Ricky Wallace
##         29575          9814         12168
```

```r
# duplicates - this was the one MRDR up there.. error need to fix

# Format Date columns, Make Start Week into Dates, extract day of the year as a seperate column
# extract days
day_of_Year <- as.numeric(str_sub(dat$`Ex Factory Start Week`, start= -2))
origin <- sapply(dat$Year-1,paste,'12','31',sep = '-')
start_date<- as.Date(day_of_Year, origin = origin)

dat$day_of_Year<-day_of_Year
dat$start_date<-start_date
```

Part 2: Creating Metric of Interest (Proportion of scanning/factory volume)

```
# Factory and Scanning volume anomalies
summary(dat$`Ex Factory volume`)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -24480    2362    6600   14746   16128 1619152
```

```
# theres negatives - these should be errors
sum(dat$`Scanning volume`<0 )
```

```
## [1] 3
```

```
sum(dat$`Ex Factory volume`<0)
```

```
## [1] 40
```

```
# zero scanning volume - not being sold
# zero factory volume - not being stocked
sum(dat$`Scanning volume` == 0)
```

```
## [1] 1752
```

```
sum(dat$`Ex Factory volume`== 0)
```

```
## [1] 4256
```

```
# not being stocked and not being sold
sum(dat$`Scanning volume` == 0 & dat$`Ex Factory volume`== 0)
```

```
## [1] 997
```

```
# proportion of scan/factory
dat$prop <- dat$`Scanning volume`/dat$`Ex Factory volume`*100
summary(dat$prop)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's
## -24845.46     58.21     78.26       Inf    100.00       Inf       997
```

```
# We will filter to valid data and also set cutoff for upperbound
id <- dat$prop >=0 & is.finite(dat$prop) & dat$prop <200
sum(id) # 45,248 "valid" data from 51,557
```

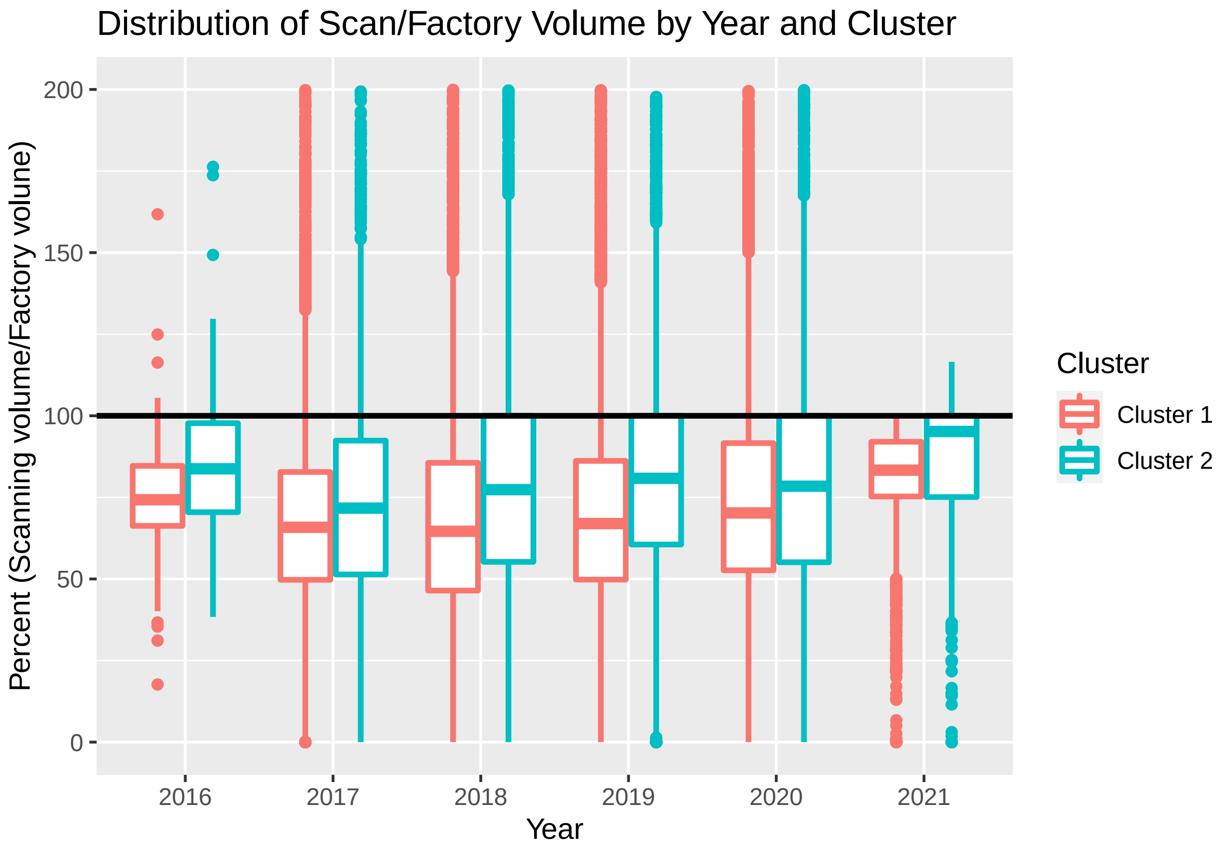
```
## [1] 45248
```

```
dat <- dat[id,] # filter to valid data, 4288 observations eliminated

## PROP > 1 UNDERSTOCKING
## PROP < 1 OVERSTOCKING
```

Part 3: Visualizations

```
# Boxplots of scan/factory by Year and Cluster
dat$Year <- factor(dat$Year)
plot0 <- ggplot(data = dat) +
  geom_boxplot(mapping = aes(x = Year, y = prop, color = Cluster),
               size = 1) +
  geom_hline(yintercept=100, size = 1) +
  ylab('Percent (Scanning volume/Factory volume)') +
  ggtitle('Distribution of Scan/Factory Volume by Year and Cluster')
# add values of the medians in here
plot0
```



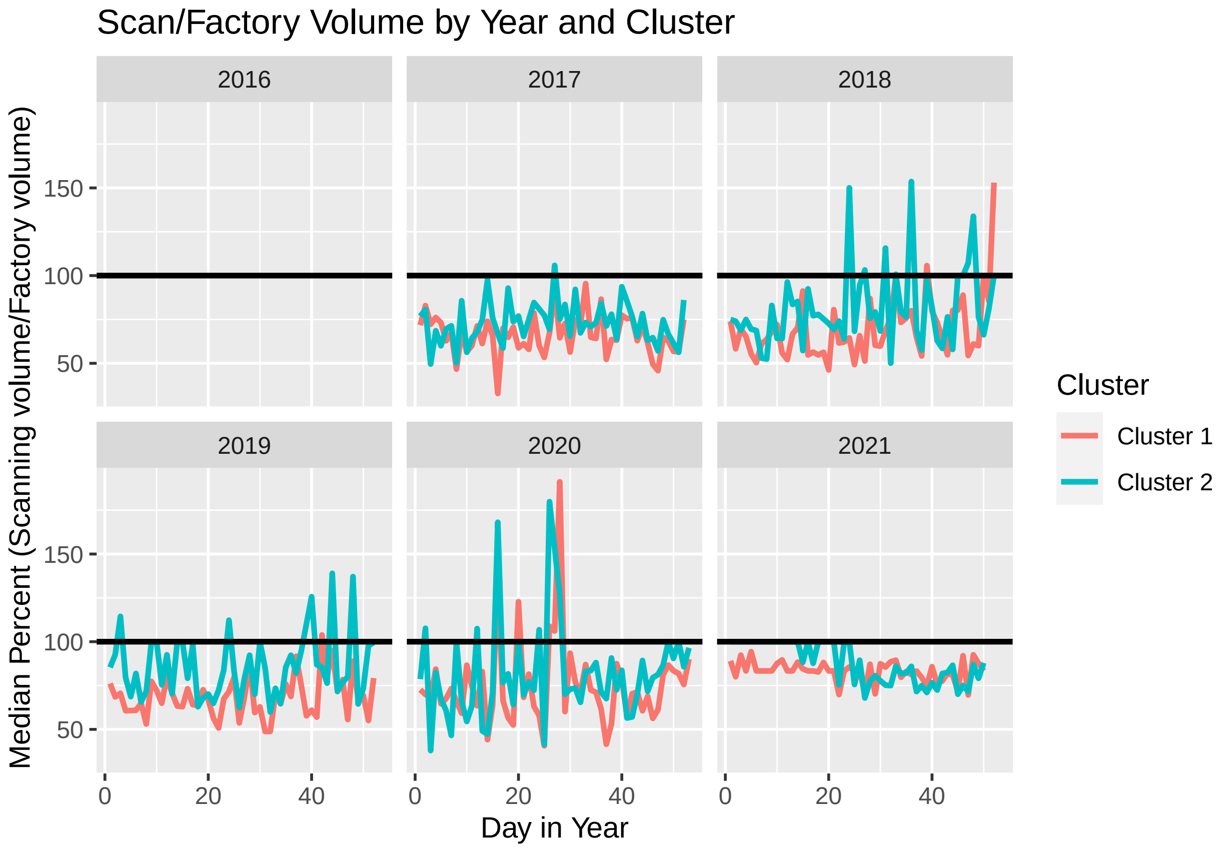
```
# MEDIAN of scan/factory over Time - LINE PLOTS

temp<- dat %>% select(day_of_Year,Cluster,Year,prop) %>%
  group_by(day_of_Year,Cluster,Year) %>% summarise(med = median(prop), sd = sd(prop))
```

```
## `summarise()` has grouped output by 'day_of_Year', 'Cluster'. You can override
## using the `.groups` argument.
```

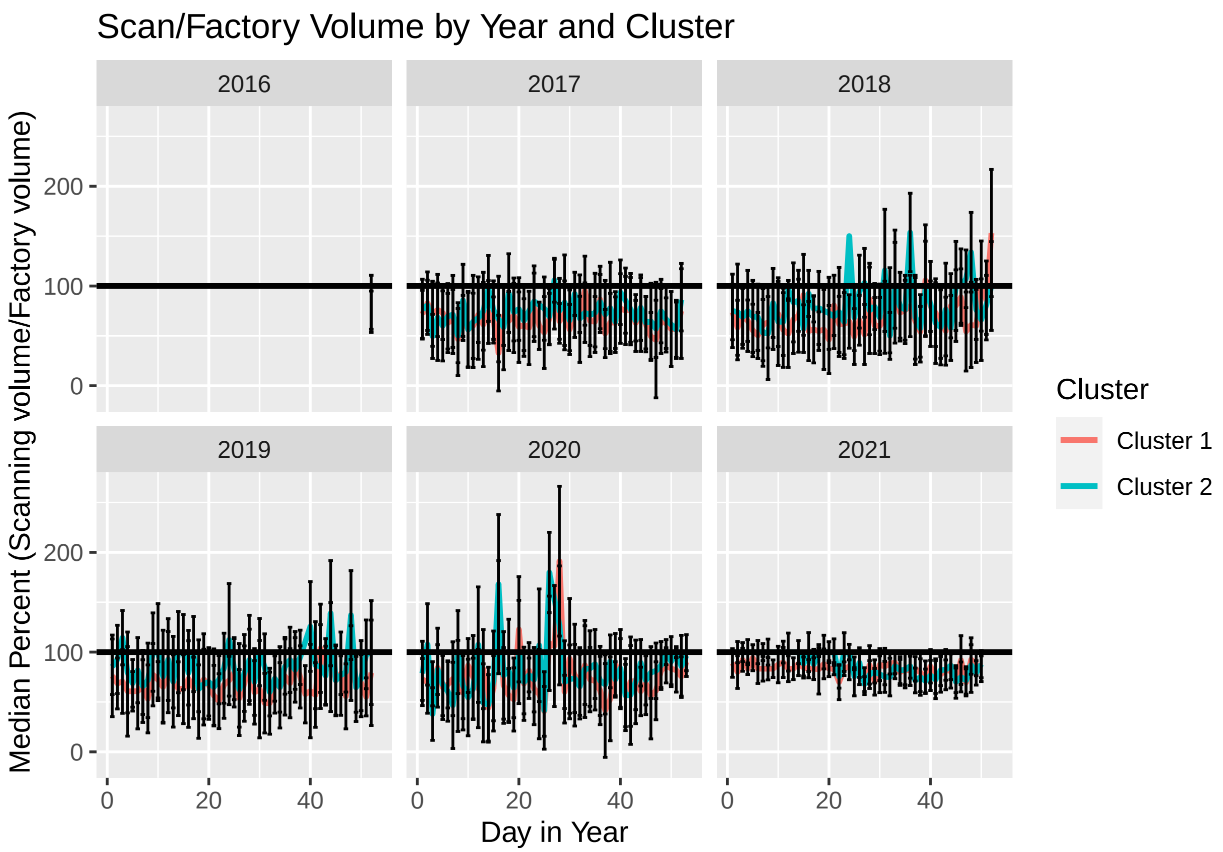
```
plot1 <- temp %>%
  ggplot() +
  geom_line(mapping = aes(x = day_of_Year, y = med, color = Cluster),
             size = 1) +
  geom_hline(yintercept=100, size = 1) +
  facet_wrap(facets = . ~ Year) +
  ylab('Median Percent (Scanning volume/Factory volume)') +
  xlab('Day in Year') +
  ggtitle('Scan/Factory Volume by Year and Cluster')
#ggplotly(plot1)
plot1
```

```
## geom_path: Each group consists of only one observation. Do you need to adjust
## the group aesthetic?
```



```
# also with error bars but messy to look at :)
plot1 + geom_errorbar(aes(x = day_of_Year,ymin=med-sd, ymax=med+sd))
```

```
## geom_path: Each group consists of only one observation. Do you need to adjust
## the group aesthetic?
```

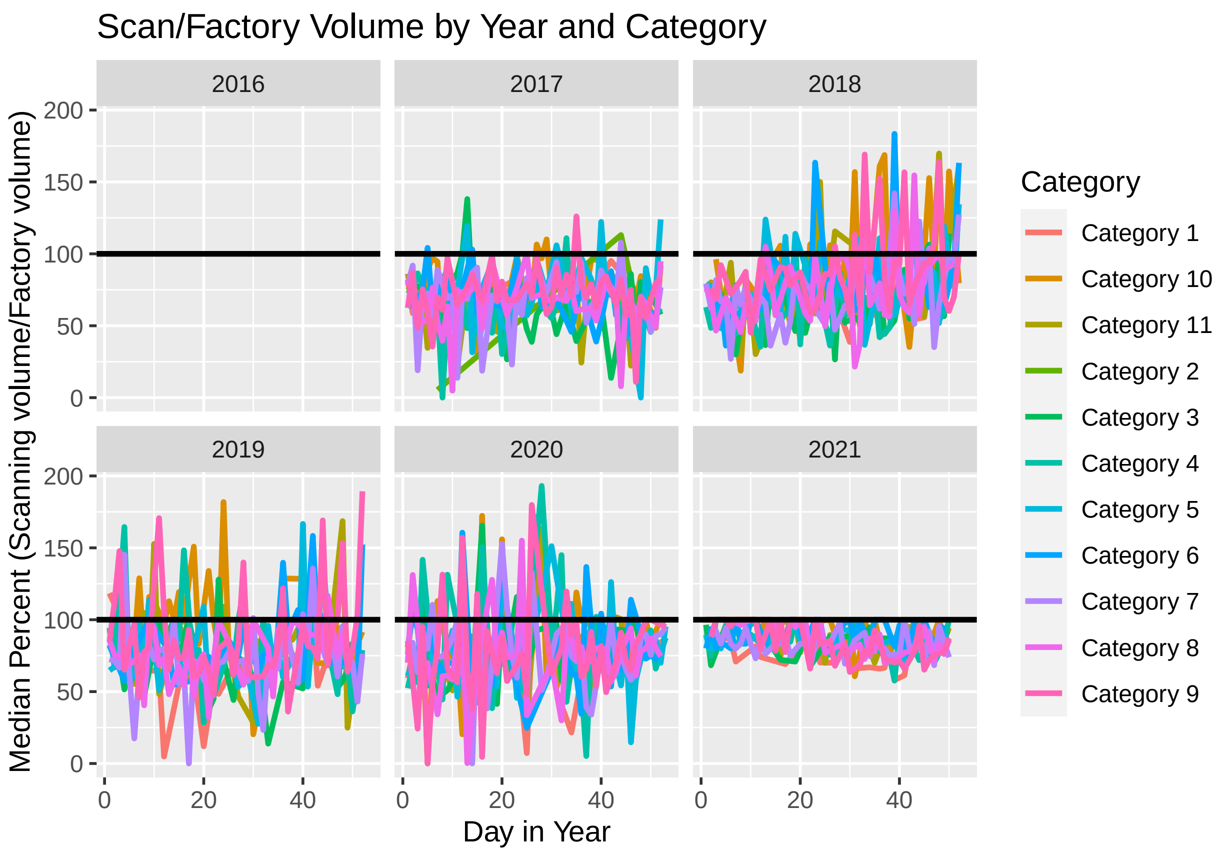


```
temp2<- dat %>% select(day_of_Year,Category,Cluster,Year,prop) %>%
  group_by(day_of_Year,Category,Year,Cluster) %>% summarise(med = median(prop), sd = sd(prop))
```

```
## `summarise()` has grouped output by 'day_of_Year', 'Category', 'Year'. You can
## override using the `.groups` argument.
```

```
plot2 <- temp2 %>%
  ggplot() +
  geom_line(mapping = aes(x = day_of_Year, y = med, color = Category, shape=Cluster),
            size = 1) +
  geom_hline(yintercept=100, size = 1) +
  facet_wrap(facets = . ~ Year) +
  ylab('Median Percent (Scanning volume/Factory volume)') +
  xlab('Day in Year') +
  ggtitle('Scan/Factory Volume by Year and Category')
#ggplotly(plot2)
plot2
```

```
## geom_path: Each group consists of only one observation. Do you need to adjust
## the group aesthetic?
```

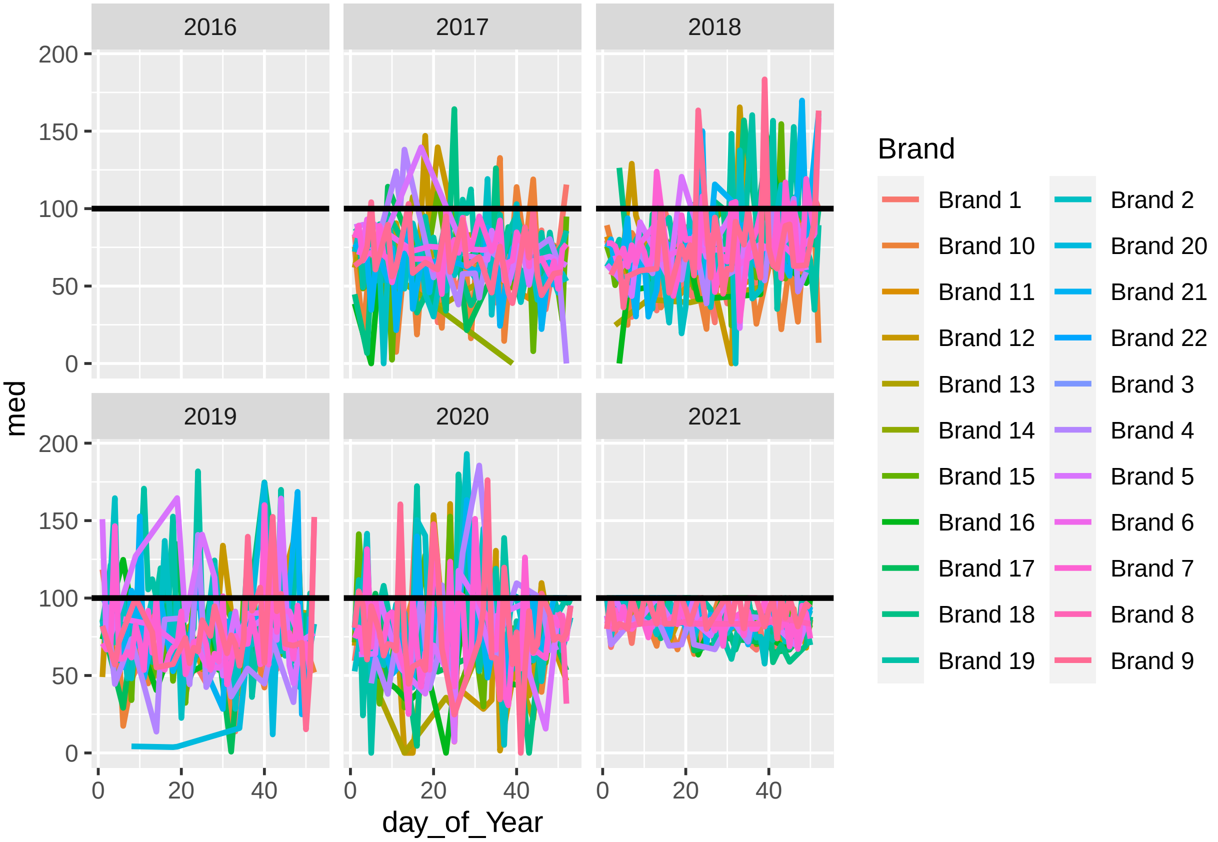


```r
# by brand.. too chaotic
temp3<- dat %>% select(day_of_Year,Brand,Year,prop) %>%
  group_by(day_of_Year,Brand,Year) %>% summarise(med = median(prop), sd = sd(prop))
```

```
## `summarise()` has grouped output by 'day_of_Year', 'Brand'. You can override
## using the `.groups` argument.
```

```r
plot3 <- temp3 %>%
  ggplot() +
  geom_line(mapping = aes(x = day_of_Year, y = med, color = Brand),
            size = 1) +
  geom_hline(yintercept=100, size = 1) +
  facet_wrap(facets = . ~ Year)
plot3
```

```
## geom_path: Each group consists of only one observation. Do you need to adjust
## the group aesthetic?
```


2016
2017
2018
2019
2020
2021
med
200
150
100
50
0
0
20
40
day_of_Year
Brand
Brand 1
Brand 10
Brand 11
Brand 12
Brand 13
Brand 14
Brand 15
Brand 16
Brand 17
Brand 18
Brand 19
Brand 2
Brand 20
Brand 21
Brand 22
Brand 3
Brand 4
Brand 5
Brand 6
Brand 7
Brand 8
Brand 9