Report for Apprehensions at the US-Mexico border from 2000 to 2017

Hongyi He, Glib Dolotov

February 19, 2018

1.Introduction

In this document, we provide several graphs for comparing US-Mexico border apprehension data. First, you will see charts that facilitate comparison of apprehension data in each sector and for each month during the years of 2010 and 2017. Next, we provide two t-tests analyzing the same data. Finally, we drive a time-series and averages of total apprehensions across all sectors from 2000 to 2017. For all segments, we include the R code that was used to compare and plot the data.

2.Comparison between 2010 and 2017

```
library(ggplot2)
library(tidyverse)
```

```
## Warning: replacing previous import by 'tidyr::%>%' when loading 'broom'
```

```
## Warning: replacing previous import by 'tidyr::gather' when loading 'broom'
```

```
## Warning: replacing previous import by 'tidyr::spread' when loading 'broom'
```

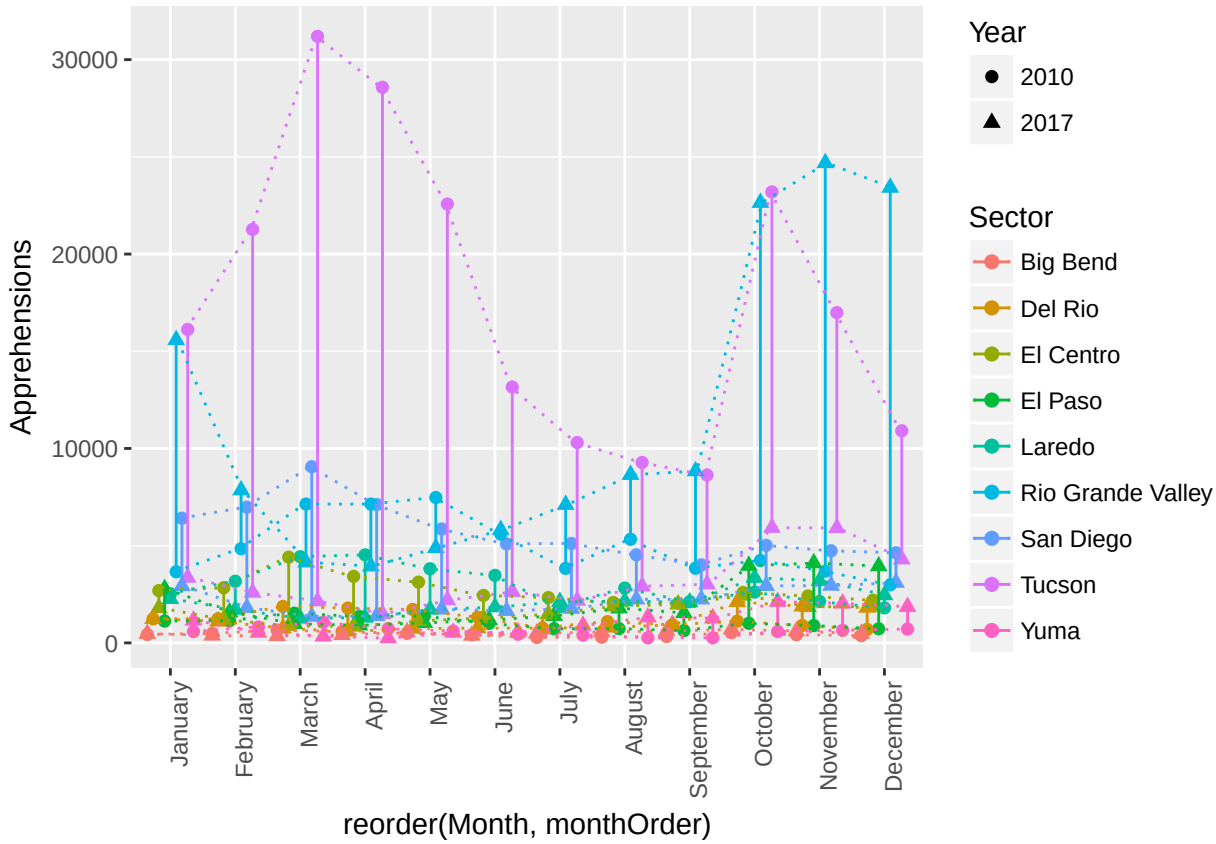
```
## -- Attaching packages ---------------------------------------------------------------------- tidyverse
```

```
## ✓ tibble  1.4.2     ✓ purrr   0.2.4
## ✓ tidyr   0.8.0     ✓ dplyr   0.7.4
## ✓ readr   1.1.1     ✓ stringr 1.2.0
## ✓ tibble  1.4.2     ✓ forcats 0.2.0
```

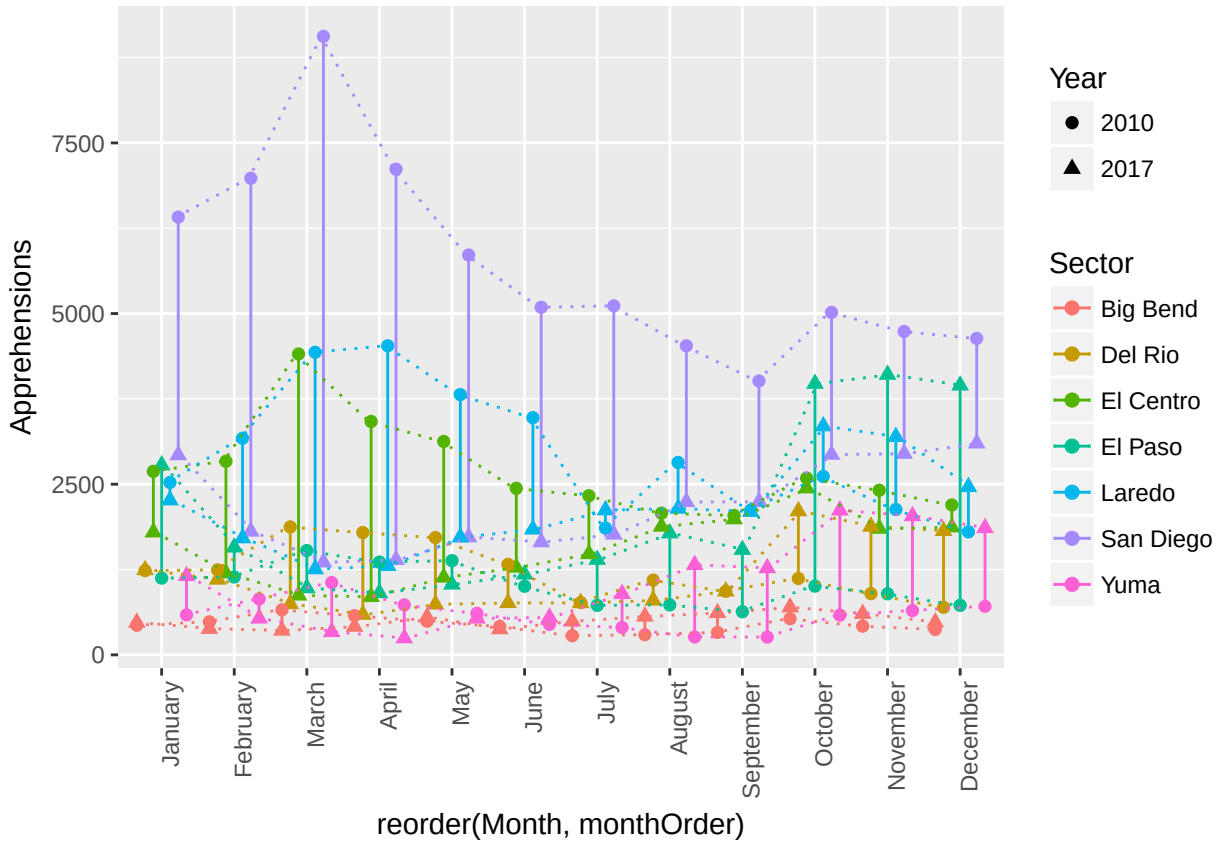
```
## -- Conflicts ------------------------------------------------------------------------- tidyverse_confl
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
#function that turn the sector col to row name
rowname <- function(x){
  rownames(x) <- x[,1]
  x<-subset(x,select=-Sector)
  x
}

itemize_df <- function(df,yr){
  tibble("Year" = rep(yr,
                      times = nrow(df)*ncol(df)),
         "Sector" = rep(rownames(df),
                        each = ncol(df)),
         "Month" = rep(colnames(df),
                       times = nrow(df)),
         "Apprehensions" = c(t(df)))
}
```

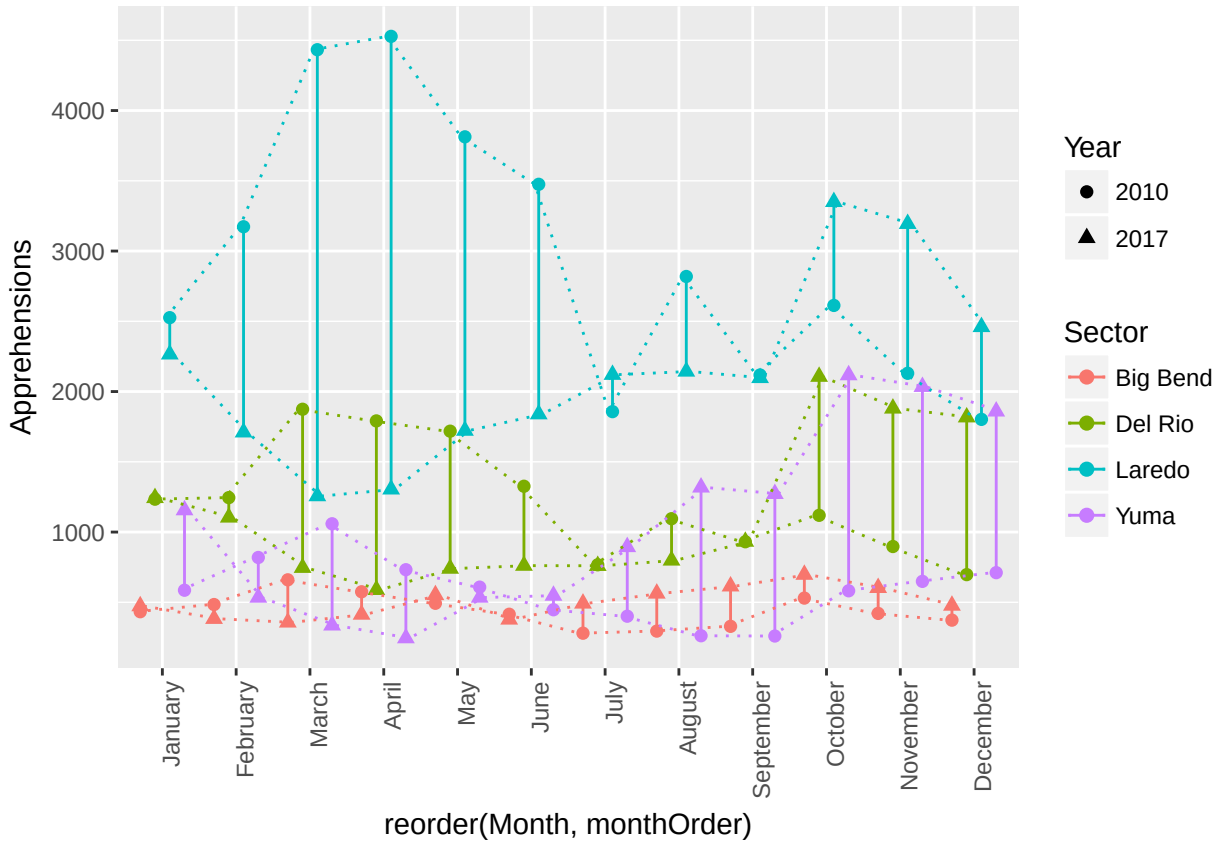
```r
#load the data and clean them
year2010 <- read.csv("BP Apprehensions 2010.csv")
year2017 <- read.csv("PB Apprehensions 2017.csv")
year2010 <- rowname(year2010)
year2017 <- rowname(year2017)

# Rearranging Columns into calendar order
year2010 <- cbind(year2010[4:12], year2010[1:3])
year2017 <- cbind(year2017[4:12], year2017[1:3])

sector <- rownames(year2010)
month <- colnames(year2010)

x <- itemize_df(year2010,"2010")
y <- itemize_df(year2017,"2017")
z <- rbind(x,y)

# This function creates the desired plot given a data input
ourplot <- function(dset){
  monthOrder <- rep(1:12, times = length(dset$Month)/12)
  ggplot(data = dset, aes(reorder(Month,monthOrder),
                          Apprehensions,
                          color = Sector,
                          group = interaction(Sector,Month),
                          shape = Year)) +
    geom_point(size = 2,
               position = position_dodge(width = 0.8)) +
    geom_line(position = position_dodge(width=0.8)) +
    geom_line(aes(group = interaction(Year, Sector)),
              position = position_dodge(width = 0.8),
              linetype = "dotted") +
    theme(axis.text.x = element_text(angle = 90, hjust = 1))
}
ourplot(z)
```




```
# Since it is difficult to visualize smaller data, we take a subset excluding
# the sectors that have the most extreme apprehension tallies. This effectively
# "zooms the graph in."
z2 = subset(z, !(Sector %in% c("Tucson","Rio Grande Valley")))
ourplot(z2)
```




```
# Again we "zoom in" by excluding the sectors with the most apprehensions per
# month.
z3 = subset(z2, !(Sector %in% c("San Diego", "El Centro", "El Paso")))
ourplot(z3)
```



2.1

3. Statistical testing on data from 2010 and 2017

T-test Results and Interpretation

3.1 T-test between sectors with the most apprehensions in 2010 and 2017

```
#load the data
year2010 <- read.csv("BP Apprehensions 2010.csv")
year2017 <- read.csv("PB Apprehensions 2017.csv")

#function that turn the sector col to row name
operation_row <- function(x){
  rownames(x) <- x[,1]
  x<-subset(x,select=-Sector)
  x <- rbind(x, colSums(x))
  -length(rownames(x))
  rownames(x) <- c(rownames(x)[-length(rownames(x))], "Total")
  x <- cbind(x,rowSums(x))
  colnames(x) <- c(colnames(x)[-length(colnames(x))], "Total")
  x
}

#clean the data
year2010 <- operation_row(year2010)
```

```r
year2017 <- operation_row(year2017)

#select the sector with the most apprehensions in 2010 and 2017
year2010_rowtotal<-year2010[-nrow(year2010),]
year2017_rowtotal<-year2017[-nrow(year2017),]
most_sec_2010<-(year2010_rowtotal[year2010_rowtotal$Total==max(year2010_rowtotal$Total),])
most_sec_2017<-(year2017_rowtotal[year2017_rowtotal$Total==max(year2017_rowtotal$Total),])
most_sec_2010<-most_sec_2010[,-ncol(most_sec_2010)]
most_sec_2017<-most_sec_2017[,-ncol(most_sec_2017)]

#do the t-test, compare the sector with the most apprehensions in two years
most_sec_2010<-t(most_sec_2010)
most_sec_2017<-t(most_sec_2017)
print(paste(colnames(most_sec_2010),"is sector with most apprehension in 2010"))
```

```
## [1] "Tucson is sector with most apprehension in 2010"
```

```r
print(paste(colnames(most_sec_2017),"is sector with most apprehension in 2017"))
```

```
## [1] "Rio Grande Valley is sector with most apprehension in 2017"
```

```r
t.test(most_sec_2010,most_sec_2017,paired=T)
```

```
##
##  Paired t-test
##
## data:  most_sec_2010 and most_sec_2017
## t = 1.7602, df = 11, p-value = 0.1061
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1557.42 13997.42
## sample estimates:
## mean of the differences
##                    6220
```

t test result: p= 0.1061 > 0.05

Explanation: We know that the sector with most apprehension in 2010 is Tucson and the sector with most apprehension in 2017 is the Rio Grande Valley. From the results of t-test, we cannot reject the null hypothesis that the apprehension in Tucson in 2010 has an equal estimated mean as Rio Grande Valley in 2017. Though the sector with the most apprehension changed from 2010 to 2017, but the maximum number of apprehension did not change.

3.2 T-test between the 3 month periods with the most apprehensions in 2010 and 2017

```r
#select the 3 month periods with the most apprehensions in 2010 and 2017
year2010_coltotal <- year2010[,-ncol(year2010)]
year2017_coltotal <- year2017[,-ncol(year2010)]

#sort the data in 2010,2017 with respect to total apprehension by month
year2010_coltotal <- year2010_coltotal[,order(year2010_coltotal["Total",],decreasing = T)]
year2017_coltotal <- year2017_coltotal[,order(year2017_coltotal["Total",],decreasing = T)]

#select the months with top three total apprehension
most_month_2010 <-  year2010_coltotal[,1:3]
most_month_2017 <-  year2017_coltotal[,1:3]
```

```r
most_month_2010 <- most_month_2010[-nrow(most_month_2010),]
most_month_2017 <- most_month_2017[-nrow(most_month_2017),]


a <- cbind(most_month_2010[,1],most_month_2010[,2],most_month_2010[,3])
b <- cbind(most_month_2017[,1],most_month_2017[,2],most_month_2017[,3])
#do the t-test, compare the sector with the most apprehensions in two years
print(paste(colnames(most_month_2010),
            "is one of the 3 month periods with the most apprehensions in 2010"))
```

```
## [1] "March is one of the 3 month periods with the most apprehensions in 2010"
## [2] "April is one of the 3 month periods with the most apprehensions in 2010"
## [3] "May is one of the 3 month periods with the most apprehensions in 2010"
```

```r
print(paste(colnames(most_month_2017),
            "is one of the 3 month periods with the most apprehensions in 2017"))
```

```
## [1] "November is one of the 3 month periods with the most apprehensions in 2017"
## [2] "October is one of the 3 month periods with the most apprehensions in 2017"
## [3] "December is one of the 3 month periods with the most apprehensions in 2017"
```

```r
t.test(a,b,paired=T)
```

```
##
##  Paired t-test
##
## data:  a and b
## t = 0.54504, df = 26, p-value = 0.5904
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2771.027  4770.804
## sample estimates:
## mean of the differences
##                999.8889
```

t test result: p= 0.5904 > 0.05

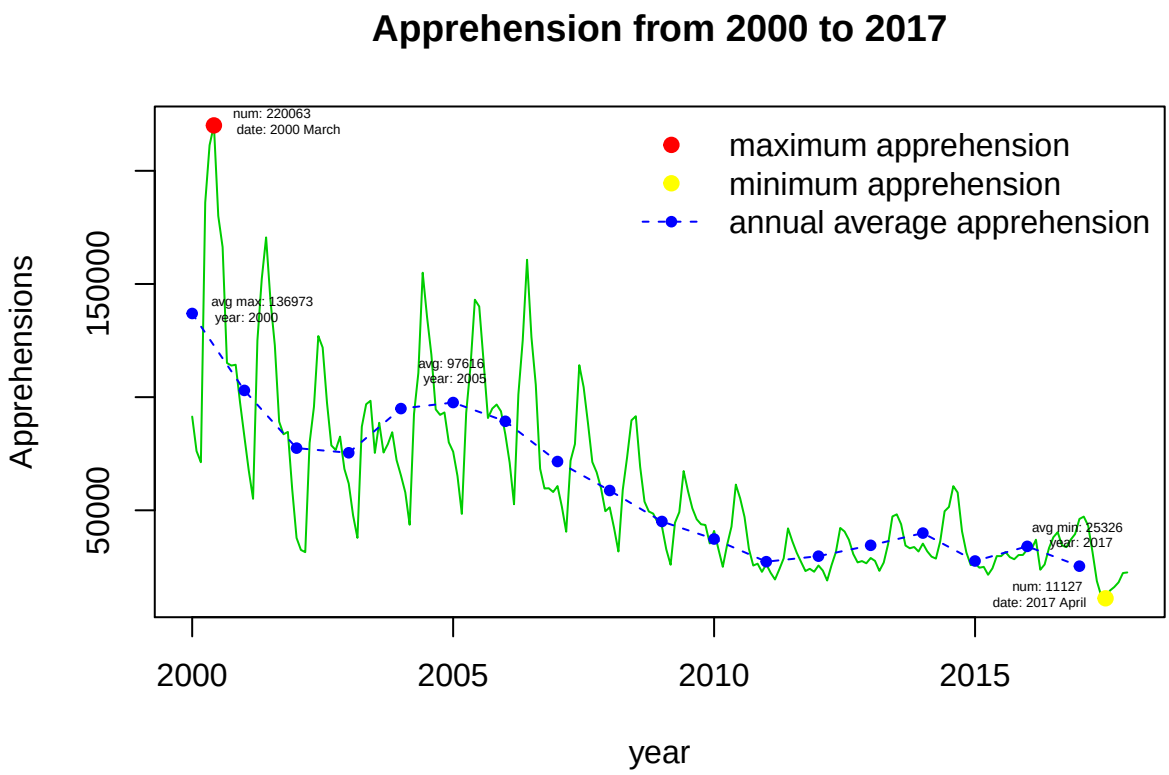
Explanation: We know that the 3-month period with the most apprehensions in 2010 is March, April and May, and the 3-month period with the most apprehensions in 2017 is November, October, December. From the result of t-test, we cannot reject the null hypothesis that the apprehension during March, April and May 2010 have an equal estimated mean as November, October, December in 2017. Though the period with most apprehensions changed, the maximum number of apprehensions didn't change much.

4.The apprehensions change from 2000 to 2017

```r
#load, clean and rerrange the data
ts<-read.csv("PB monthly summaries.csv")
rownames(ts) <- ts[,1]
ts <- ts[order(ts$year),]
ts<-subset(ts,select=-year)

#turn it into timeseries format
ts_df<-ts
ts <- as.vector(t(ts))
ts<-ts(ts,start = c(2000,1), frequency=12)
```

```r
#plot the timeseries data
ts.plot(ts,
        col = 3,
        xlab="year",
        ylab="Apprehensions",
        lty=c(1:3),
        main ="Apprehension from 2000 to 2017")

#draw the avg
ts_avg<-apply(ts_df,1,mean)
ts_avg<-ts(ts_avg,start = c(2000), frequency=1)
points(ts_avg,col = 4,pch=20)
lines(ts_avg,col = 4,pch=20,lty=2)

#get the max/min month index
pos_max<-which(ts_df == max(ts_df), arr.ind=T)
pos_min<-which(ts_df == min(ts_df), arr.ind=T)

#get the month vector
month<-c("October","November","December","January","February","March","April","May","June","July","Augus

#get the exact time point of the maximum/minimum element in timeseries data
max_index <- time(ts)[ts==max(ts)]
min_index <- time(ts)[ts==min(ts)]
max_avg <- time(ts_avg)[ts_avg==max(ts_avg)]
min_avg <- time(ts_avg)[ts_avg==min(ts_avg)]

#label the maximum/minimum apprehension in the graph
text(max_index,max(ts),
     paste("num:",max(ts),"\n","date:",floor(max_index),month[as.vector(pos_max)[2]]),
     cex=0.4,pos=4)
text(min_index,min(ts),
     paste("num:",min(ts),"\n","date:",floor(min_index),month[as.vector(pos_min)[2]]),
     cex=0.4,pos=2)
text(max_avg,
     max(ts_avg),
     paste("avg max:",floor(max(ts_avg)),"\n","year:",floor(max_avg)),
     cex=0.4,
     pos=4)
text(min_avg,
     min(ts_avg),
     paste("avg min:",floor(min(ts_avg)),"\n","year:",floor(min_avg)),
     cex=0.4,
     pos=3)
text(2005,
     ts_avg[time(ts_avg)==2005],
     paste("avg:",floor(ts_avg[time(ts_avg)==2005]),"\n","year:",2005),
     cex=0.4,
     pos=3)

#pinpoint the maximum/minimum apprehension in the graph
points(max_index,max(ts),pch=19,col=2)
points(min_index,min(ts),pch=19,col=7)
```

```
#add a legend which denotes the meaning of the points
legend("topright",
       c("maximum apprehension",
         "minimum apprehension",
         "annual average apprehension"),
       col = c(2,7,4),
       pch =c(19,19,20),
       lty=c(0,0,2),
       bty = "n",
       bg = 'gray90')
```



As shown in the graph above, apprehensions at the US-Mexico border reached historic lows in 2017 and has shown a downward trend since 2000. Apprehensions have Dropped all the way from a peak of 220,063 in 2000 to a low of 11,127, a nearly 2000% decline.

As for average annual apprehensions, we can see they have declined from 136,973 in 2000 to 25,326 in 2017. Though the figure rose a little bit in 2005 to 97,616 , it went down to 25,326 in 2017.