Data Analysis Report of Plastic Pollution Data Set

Amrinder Sehmbi

9/04/2021

Data Background

It is well known that plastic is incorporated in most of the products used in society due to its low cost and easy manufacturing. Unfortunately, the overuse of plastic makes it a significant cause of the world's pollution problems. This issue has inclined our group to study some issues related to plastic pollution more thoroughly. In this data analysis we will explore and analyze the world plastic pollution count data set from the 2021 collection of data sets on the tidytuesday git hub page. This data is explored from the organization "Break Free from Plastic", which is a global movement envisioning a future free from plastic pollution. The main purpose to collect this data was to raise awareness about the growing concern of plastic pollution around the world. The data set includes information about the total plastic count for each country, plastic company and type of plastic in the year 2019 and 2020. It also includes details about the pickup events and volunteers used to collect the data.

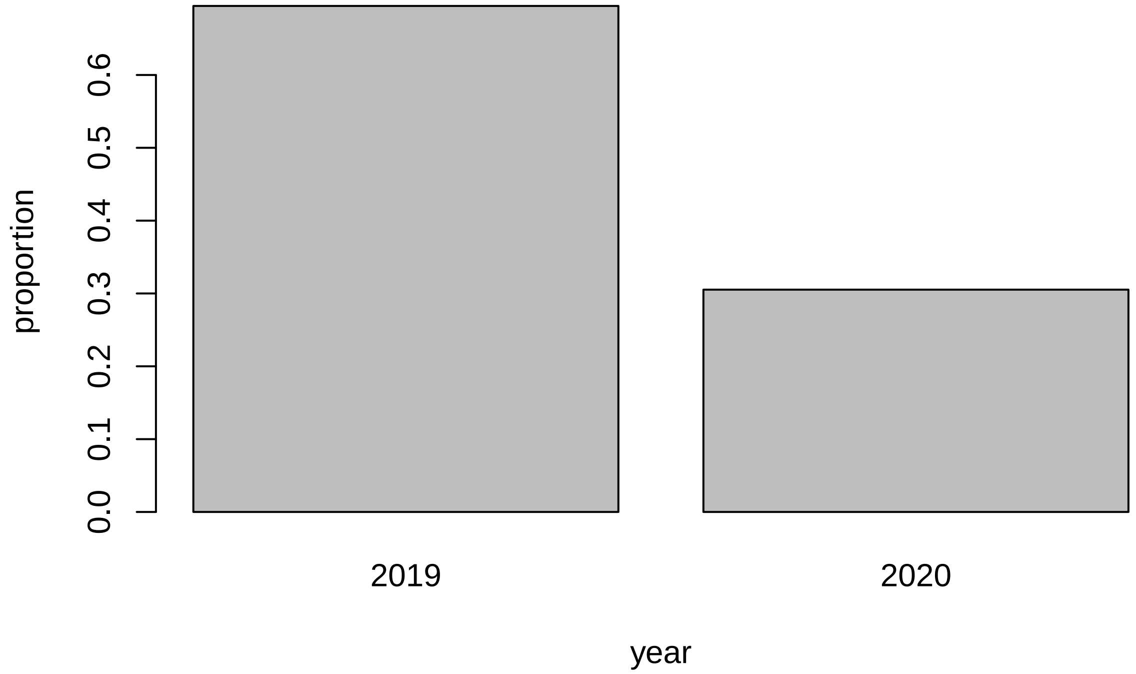
Exploring Variables

Year of Data Recorded

```r
attach(plastics)
# Table of proportion for years
year <- table(plastics$year)
prop.table(year)
```

```
##
##      2019      2020
## 0.6947683 0.3052317
```

```r
# Barplot of proportion of years
barplot(prop.table(year), xlab = "year", ylab = "proportion")
```



This data set only includes data for the year 2019 and 2020. From the bar plot above, we see that there is significantly more data for the year 2019 then 2020. Approximately, 70% of the data is from the year 2019 and 30% is from the year 2020.

Country

```
attach(plastics)
```

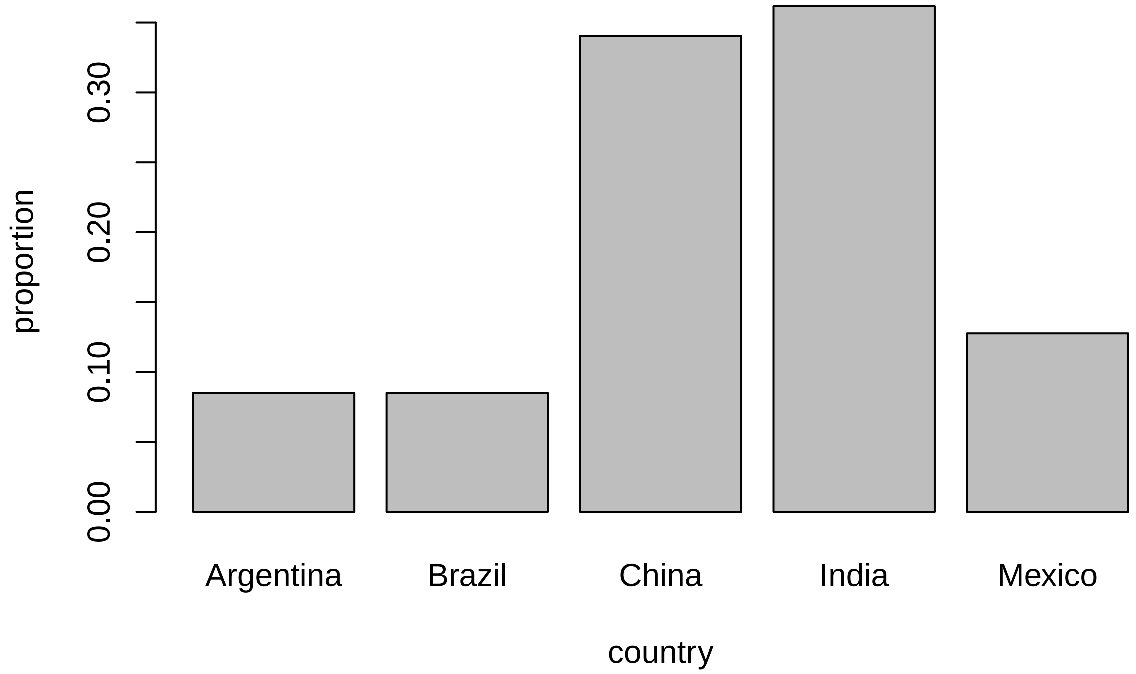
```
## The following object is masked _by_ .GlobalEnv:
##
##     year
```

```
## The following objects are masked from plastics (pos = 3):
##
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year
```

```
sub <- subset(plastics, country == c("Argentina", "India", "China", "Brazil", "Mexico") & grand_total >
# Table of proportion of the countries
countries <- table(sub$country)
prop.table(countries)
```

```
##
##  Argentina     Brazil      China      India     Mexico
## 0.08510638 0.08510638 0.34042553 0.36170213 0.12765957
```

```
# Barplot of proportion of the countries
barplot(prop.table(countries), xlab = "country", ylab = "proportion")
```



For the analysis involving countries, we are only looking at a few countries which are known to have large populations. From the bar plot above, it is clear that the majority of the data is from India(36%) and China(34%) and the least data is from Argentina(8.5%), Brazil(8.5%) and Mexico(13%).

Grand Total : Total Count of Plastic Collected

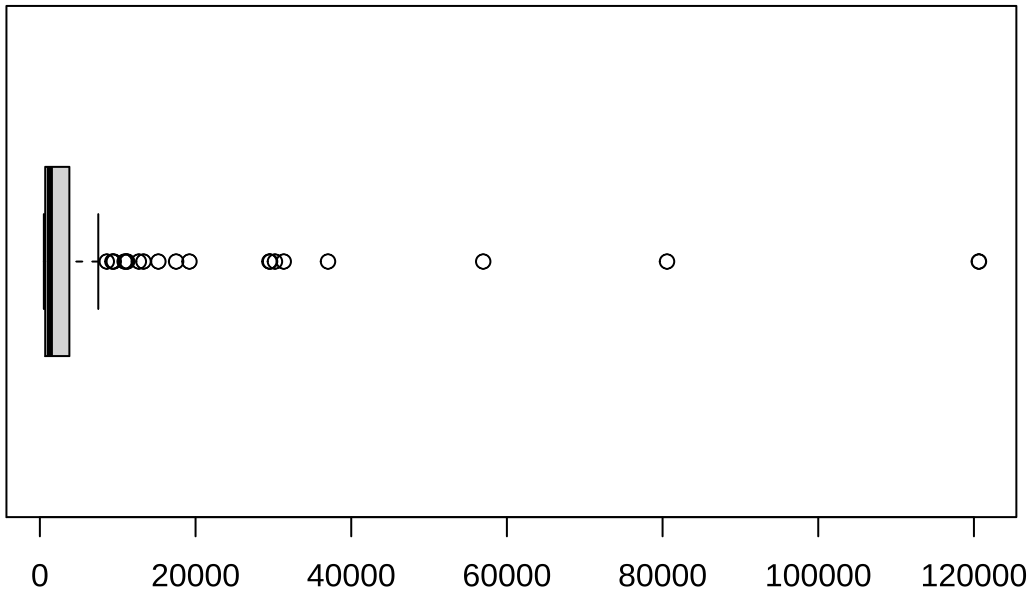
```
attach(plastics)
```

```
## The following object is masked _by_ .GlobalEnv:
##
##     year

## The following objects are masked from plastics (pos = 3):
##
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year

## The following objects are masked from plastics (pos = 4):
##
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year
```

```
summary(grand_total[grand_total > 500])
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's 
##    509.0    685.5   1288.0   5035.8   3789.5 120646.0       14
```

```
boxplot(grand_total[grand_total > 500], horizontal = TRUE)
```



For the analysis of this data set, we have decided to only consider total plastic counts greater than 500 since we think it is not an accurate representation of the true count which could have been affected by various factors. To begin with, we see the max count of plastic was 120646 and min count was 509 and the mean count was 5035. From the box plot, we see that distribution for grand_total seems to be right skewed. The median seem to be towards the left of the box which implies more than 50% of the total counts of plastic are above 1288 and less than 50% of the total counts are below 1288. The middle 50% of total counts spreads across a range of Q3 - Q1 = 3789.5 - 685.5 = 3104. There seems to a few extreme potential outliers like the max which is 120646 and others like 80000, 57000 and etc. These outlier could be due to some particular countries having more garbage collection events and volunteers.

Number of Events

```r
attach(plastics)
```

```
## The following object is masked _by_ .GlobalEnv:
##
##     year
```

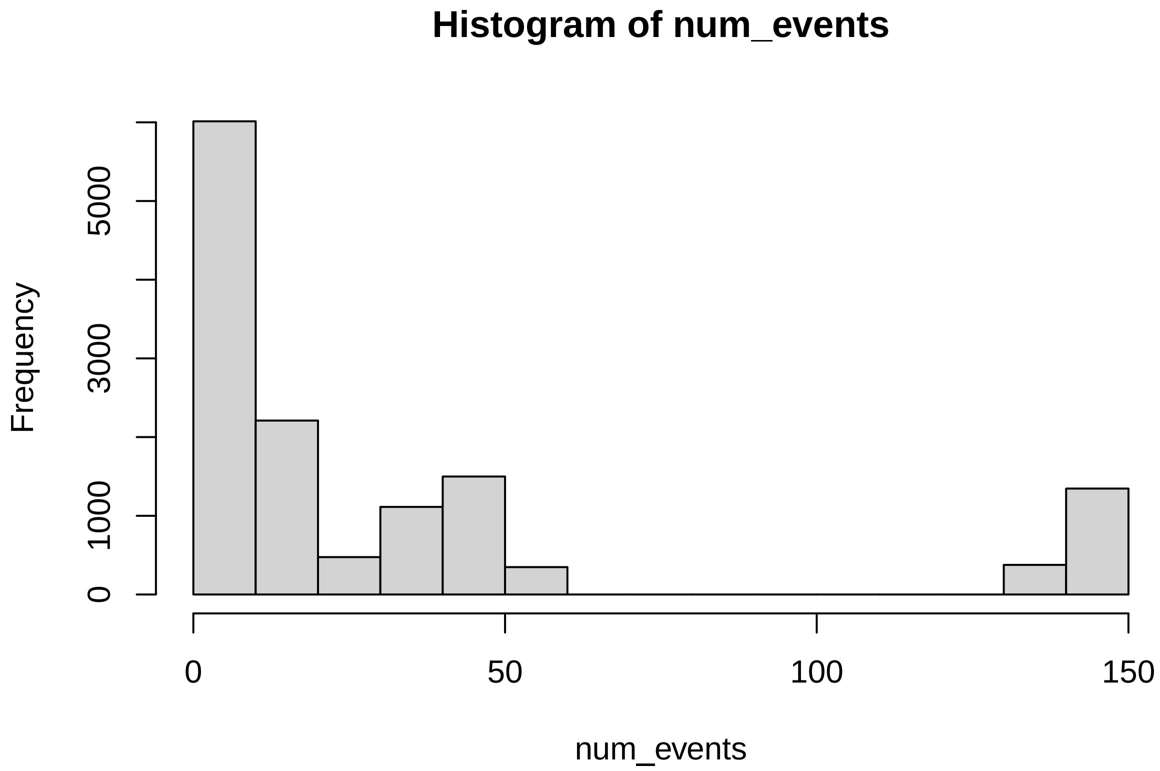
```
## The following objects are masked from plastics (pos = 3):
##
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year
```

```
## The following objects are masked from plastics (pos = 4):
##
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year
## The following objects are masked from plastics (pos = 5):
##
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year
```

```
summary(num_events)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    4.00   15.00   33.37   42.00  145.00
```

```
hist(num_events)
```



From the summary statistic of number of events given above, on average there were 33 cleanup event with max 145 and min 1. The histogram of number of events given above seems to right skewed and bimodal such that majority of countries had only few cleanup events.

Volunteers

```
attach(plastics)
```

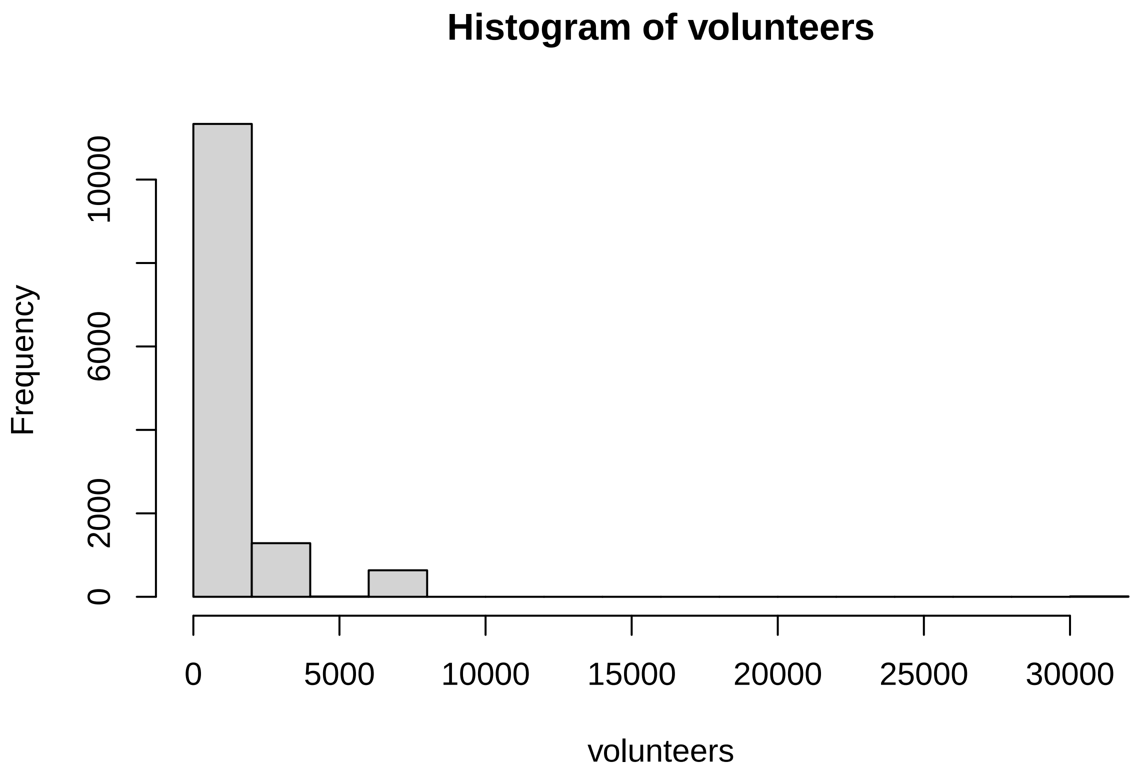
```
## The following object is masked _by_ .GlobalEnv:
##
##     year
```

```
## The following objects are masked from plastics (pos = 3):
##
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year
## The following objects are masked from plastics (pos = 4):
##
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year

## The following objects are masked from plastics (pos = 5):
##
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year

## The following objects are masked from plastics (pos = 6):
##
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year
```

```
summary(volunteers)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##       1     114     400    1118    1416   31318     107
```

```
hist(volunteers)
```



From the summary statistics of volunteers given above, it seems like there were 1118 people at each clean up event on average with min 1 and max 31318. By observing the histogram of volunteers, the distribution seems to be right skewed such that majority of events had only few volunteers.

Exploring Relationships

Year vs Grand Total:Total Plastic Count

```
total_2019 <- subset(plastics, select = c(grand_total, year))
total_2019 <- subset(total_2019, grand_total > 500 & year <= 2019)
attach(total_2019)
```

```
## The following object is masked _by_ .GlobalEnv:
##
##     year

## The following objects are masked from plastics (pos = 3):
##
##     grand_total, year

## The following objects are masked from plastics (pos = 4):
##
##     grand_total, year

## The following objects are masked from plastics (pos = 5):
##
##     grand_total, year

## The following objects are masked from plastics (pos = 6):
##
##     grand_total, year

## The following objects are masked from plastics (pos = 7):
##
##     grand_total, year
```

```
grand_total_2019 <- grand_total
total_2020 <- subset(plastics, select = c(grand_total, year))
total_2020 <- subset(total_2020, grand_total > 500 & year >= 2020)
attach(total_2020)
```

```
## The following object is masked _by_ .GlobalEnv:
##
##     year

## The following objects are masked from total_2019:
##
##     grand_total, year

## The following objects are masked from plastics (pos = 4):
##
##     grand_total, year

## The following objects are masked from plastics (pos = 5):
##
##     grand_total, year

## The following objects are masked from plastics (pos = 6):
##
##     grand_total, year

## The following objects are masked from plastics (pos = 7):
##
##     grand_total, year
## The following objects are masked from plastics (pos = 8):
##
##     grand_total, year
```

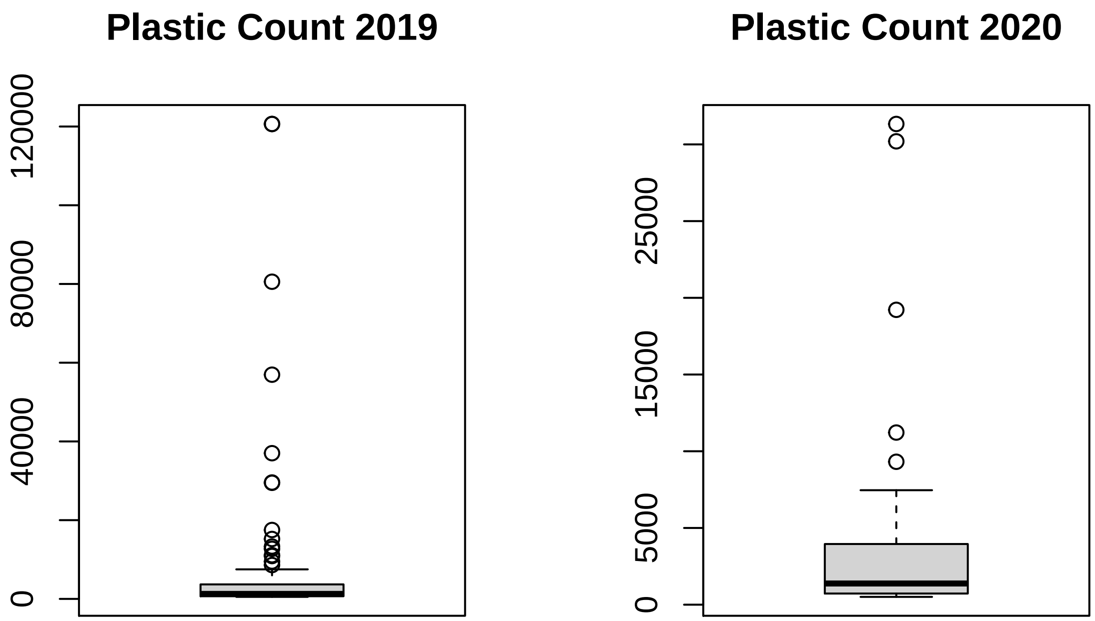
```
grand_total_2020 <- grand_total
#Summary of 2019 total plastic count.
favstats(grand_total_2019)
```

```
##  min    Q1 median     Q3    max     mean       sd   n missing
##  515 676.5   1263 3684.5 120646 6240.309 17898.82 123       0
```

```
#Summary of 2020 total plastic count.
favstats(grand_total_2020)
```

```
##  min  Q1 median   Q3   max     mean       sd  n missing
##  509 727 1380.5 3911 31331 3271.952 5182.038 84       0
```

```
par(mfrow = c(1,2))
#Histogram of 2019 total plastic count.
boxplot(grand_total_2019, main = "Plastic Count 2019")
#Histogram of 2020 total plastic count.
boxplot(grand_total_2020, main = "Plastic Count 2020")
```



Before analyzing the relationship between year and total count of plastic, we clean the data for outlier total values less than 500 since we think it is not an accurate representation of the total count for a country in a year. Using the box plot and summary statistic above for plastic count in 2019 and 2020, we can infer some interesting relationships. From the summary statistics, we see the mean count for 2019 count(6240.309) is almost double the mean count for 2020(3271.952). A possible reason for a significant difference could be the max count for 2019(120649) which is significantly larger than the max count for 2020(31331). We also see that there was more data for 2019 (n=123) then for 2020 (n=84). From the two box plots, both distribution of counts seem to be right skewed such that most of the counts in 2019 and 2020 seem to be on the lower end of counts. The middle line of both boxes seem to be on the lower end such that less than 50% of the counts are lower than the respective median counts. The middle 50% of counts in 2019 extends across a range of Q3 - Q1 = 3684.5 - 676.5 = 3008 which is slightly less than the range of middle 50% of counts in 2020 Q3 - Q1 = 3911 - 727 = 3182 which implies the middle 50% of counts in 2020 is slight more spread out than 2020. In addition, we see that the 2019 count has more potential outlier than the 2020 count.

Country vs Grand Total : Total Plastic Count

```
sub <- subset(plastics, select = c(grand_total, country))
sub <- subset(sub,country == c("Argentina", "India", "China", "Brazil", "Mexico") & grand_total > 10 & g

sub1 <- subset(sub, country == "Argentina")
attach(sub1)
```

```
## The following object is masked from total_2020:
##
##     grand_total

## The following object is masked from total_2019:
##
##     grand_total

## The following objects are masked from plastics (pos = 5):
##
##     country, grand_total

## The following objects are masked from plastics (pos = 6):
##
##     country, grand_total

## The following objects are masked from plastics (pos = 7):
##
##     country, grand_total

## The following objects are masked from plastics (pos = 8):
##
##     country, grand_total

## The following objects are masked from plastics (pos = 9):
##
##     country, grand_total
```

```
# Summary Statistics for Argentina Total Plastic Count
summary(grand_total)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12.00   15.75   23.00   25.50   32.75   44.00
```

```
sub2 <- subset(sub, country == "India")
attach(sub2)
```

```
## The following objects are masked from sub1:
##
##     country, grand_total

## The following object is masked from total_2020:
##
```

```
##     grand_total
```

```
## The following object is masked from total_2019:
##
##     grand_total
```

```
## The following objects are masked from plastics (pos = 6):
##
##     country, grand_total
```

```
## The following objects are masked from plastics (pos = 7):
##
##     country, grand_total
```

```
## The following objects are masked from plastics (pos = 8):
##
##     country, grand_total
```

```
## The following objects are masked from plastics (pos = 9):
##
##     country, grand_total
```

```
## The following objects are masked from plastics (pos = 10):
##
##     country, grand_total
```

```
# Summary Statistics for India Total Plastic Count
summary(grand_total)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   13.00   18.00   32.00   56.71   48.00  261.00
```

```
sub3 <- subset(sub, country == "China")
attach(sub3)
```

```
## The following objects are masked from sub2:
##
##     country, grand_total
```

```
## The following objects are masked from sub1:
##
##     country, grand_total
```

```
## The following object is masked from total_2020:
##
##     grand_total
```

```
## The following object is masked from total_2019:
##
##     grand_total
```

```
## The following objects are masked from plastics (pos = 7):
##
##     country, grand_total
```

```
## The following objects are masked from plastics (pos = 8):
##
##     country, grand_total
```

```
## The following objects are masked from plastics (pos = 9):
##
##     country, grand_total
## The following objects are masked from plastics (pos = 10):
##
##     country, grand_total

## The following objects are masked from plastics (pos = 11):
##
##     country, grand_total
```

```
# Summary Statistics for China Total Plastic Count
summary(grand_total)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   11.00   17.25   34.50   62.81   56.75  392.00
```

```
sub4 <- subset(sub, country == "Brazil")
attach(sub4)
```

```
## The following objects are masked from sub3:
##
##     country, grand_total

## The following objects are masked from sub2:
##
##     country, grand_total

## The following objects are masked from sub1:
##
##     country, grand_total

## The following object is masked from total_2020:
##
##     grand_total

## The following object is masked from total_2019:
##
##     grand_total

## The following objects are masked from plastics (pos = 8):
##
##     country, grand_total

## The following objects are masked from plastics (pos = 9):
##
##     country, grand_total

## The following objects are masked from plastics (pos = 10):
##
##     country, grand_total

## The following objects are masked from plastics (pos = 11):
##
##     country, grand_total

## The following objects are masked from plastics (pos = 12):
##
##     country, grand_total
```

```
# Summary Statistics for Brazil Total Plastic Count
summary(grand_total)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12.00   18.75   21.00  108.50  110.75  380.00
```

```r
sub5 <- subset(sub, country == "Mexico")
attach(sub5)
```

```
## The following objects are masked from sub4:
##
##     country, grand_total
```

```
## The following objects are masked from sub3:
##
##     country, grand_total
```

```
## The following objects are masked from sub2:
##
##     country, grand_total
```

```
## The following objects are masked from sub1:
##
##     country, grand_total
```

```
## The following object is masked from total_2020:
##
##     grand_total
```

```
## The following object is masked from total_2019:
##
##     grand_total
```

```
## The following objects are masked from plastics (pos = 9):
##
##     country, grand_total
```

```
## The following objects are masked from plastics (pos = 10):
##
##     country, grand_total
```

```
## The following objects are masked from plastics (pos = 11):
##
##     country, grand_total
```

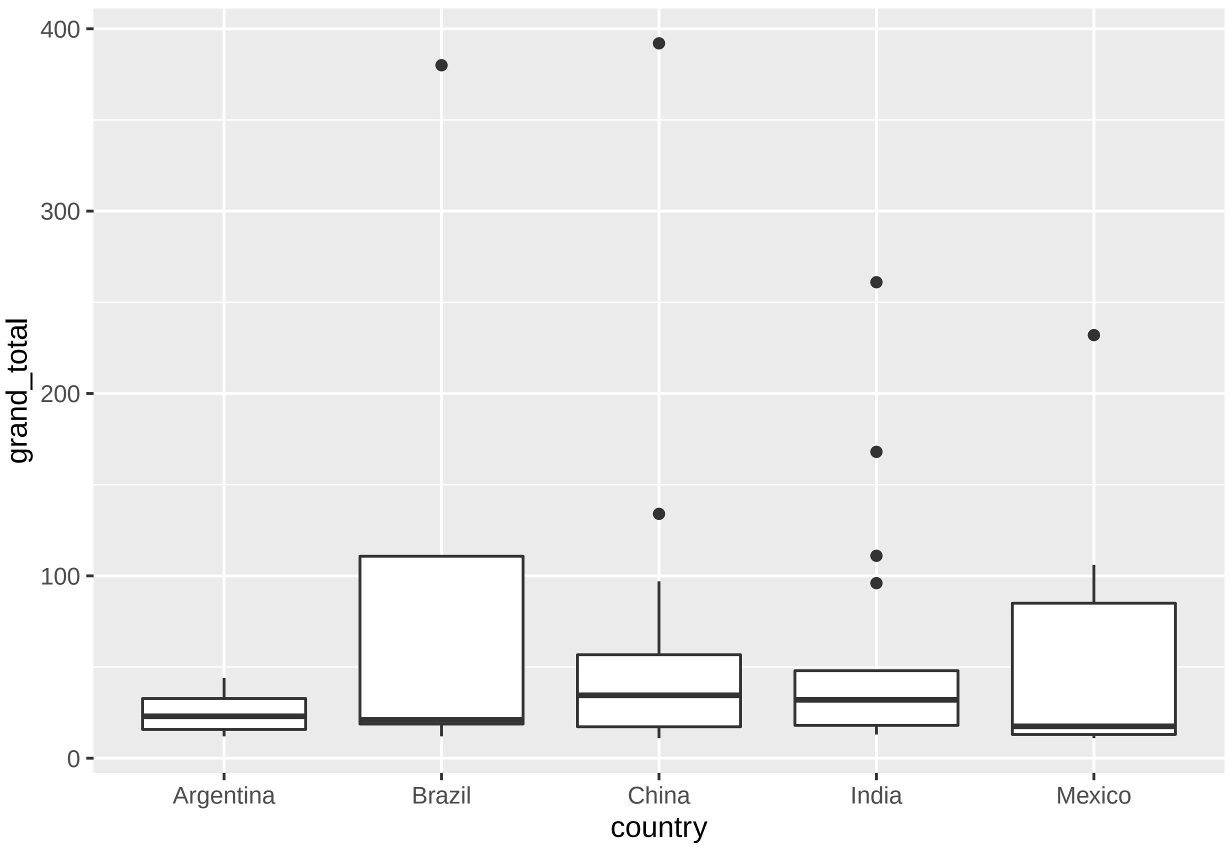
```
## The following objects are masked from plastics (pos = 12):
##
##     country, grand_total
```

```
## The following objects are masked from plastics (pos = 13):
##
##     country, grand_total
```

```r
# Summary Statistics for Mexico Total Plastic Count
summary(grand_total)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   11.00   13.00   17.50   66.17   85.00  232.00
```

```r
# Box Plot of Total Plastic Count for Contries Above.
ggplot(sub, aes(x = country, y = grand_total)) + geom_boxplot()
```



In order to describe the relationship between countries and the count of plastic from different companies, we have selected a few countries which have known large populations and we will only only look at counts of plastic greater than 10 for the companies in each of the countries selected. Using the box plot constructed above, we can infer a lot of relations between different countries' plastic count. For instance, we quickly see that China seems to have the max single count of plastic of roughly 390 in this data set for these groups of countries. We can also infer that the distribution for plastic count for Argentina, China and Mexico seem to be right skewed such that most of the plastic count for these countries seem to be on the lower side of the counts. Also, the distribution for India and Brazil seem to be left skewed such that most of their plastic counts seem to be on the higher side of the counts. For Argentina, China and India, the median seems to be in the middle of the 50% box which implies that around 50% of their counts of plastic are above and below their respective median. On the other hand, the median line is in the lower end of the 50% box for Brazil and Mexico, which implies less than 50 percent of their plastic counts are lower than their respective medians. The middle 50% of the plastic counts for Brazil and Mexico seem to be most spread out then the other countries such the plastic count extend across the range of roughly 90 for Brazil and 70 for Mexico. From the summary statistics, we can infer that the mean count for India(56), China(62) and Mexico(66) have a small difference while the mean count for Argentina(25) and Brazil(108) have a significant difference then the other means. A possible reason for such difference could be the proportion of each country's data we explore above or possible outliers.

Type of Plastic vs Grand Total : Total Plastic Count

To investigate the relationship between plastic pollution and grand_total, we choose 3 types of plastic pollution which are hdpe, pvc and ps to analyze.

HDPE(High density polyethylene) vs Grand Total

```
tuesd <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/202
```

```
## Rows: 13380 Columns: 14

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (2): country, parent_company
## dbl (12): year, empty, hdpe, ldpe, o, pet, pp, ps, pvc, grand_total, num_eve...

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
attach(tuesd)
```

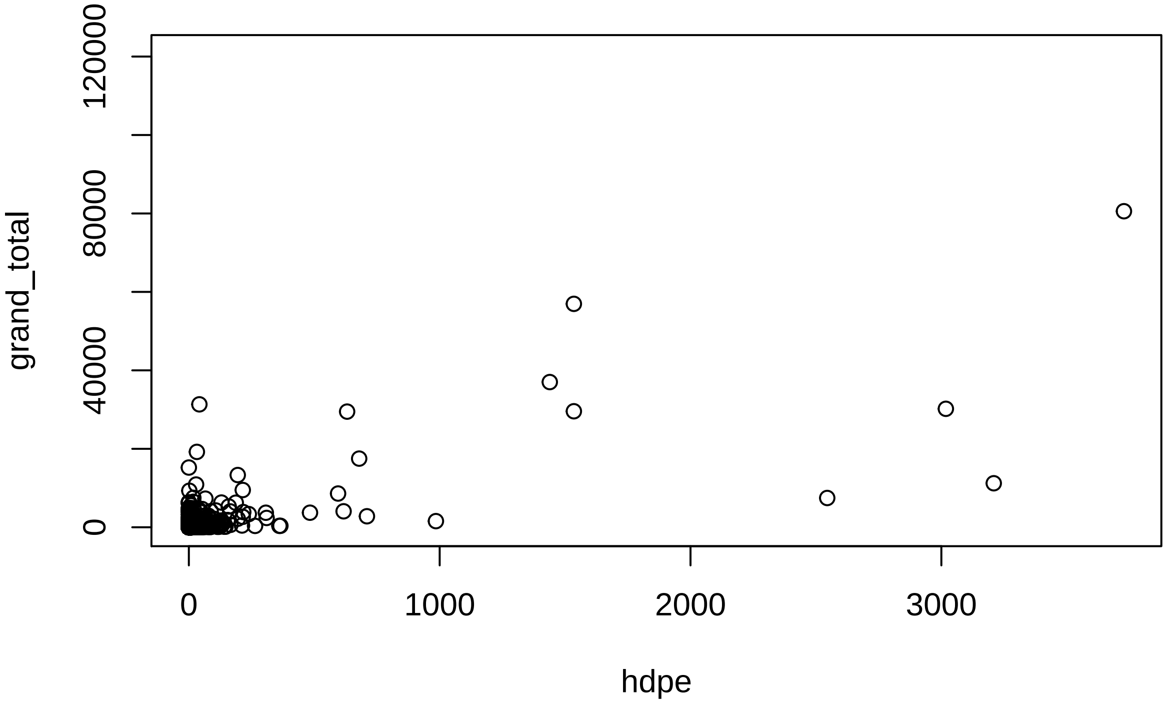
```
## The following object is masked _by_ .GlobalEnv:
##
##     year

## The following objects are masked from sub5:
##
##     country, grand_total

## The following objects are masked from sub4:
##
##     country, grand_total

## The following objects are masked from sub3:
##
##     country, grand_total

## The following objects are masked from sub2:
##
##     country, grand_total

## The following objects are masked from sub1:
##
##     country, grand_total

## The following objects are masked from total_2020:
##
##     grand_total, year

## The following objects are masked from total_2019:
##
##     grand_total, year

## The following objects are masked from plastics (pos = 10):
##
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year

## The following objects are masked from plastics (pos = 11):
##
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year

## The following objects are masked from plastics (pos = 12):
##
```

```
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year

## The following objects are masked from plastics (pos = 13):
##
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year

## The following objects are masked from plastics (pos = 14):
##
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year
```

```
summary(hdpe)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
##    0.000    0.000    0.000    3.046    0.000 3728.000     1646
```

```
plot(hdpe,grand_total)
```



From the summary statistics above for hdpe, the mean of hdpe pollution is 15.29 which is the largest amount in these 3 variables. Companies in some countries created the highest pollution of hdpe which is 3728.00, while some companies reached minimum amount of pollution that is 0.00. By observing the scatter plot between hdpe and grand_total above, we find it to be a positive linear relationship if we remove the outliers.

PVC vs Grand Total

```
tuesd <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/202
## Rows: 13380 Columns: 14

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (2): country, parent_company
## dbl (12): year, empty, hdpe, ldpe, o, pet, pp, ps, pvc, grand_total, num_eve...

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
attach(tuesd)
```

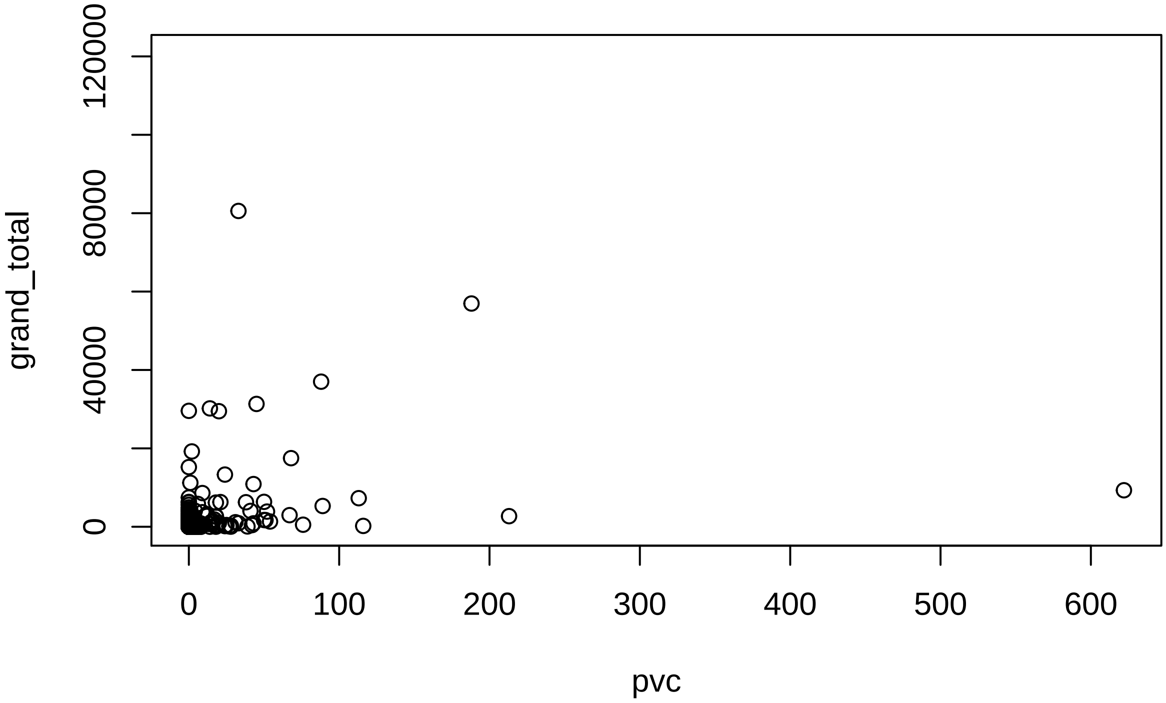
```
## The following object is masked _by_ .GlobalEnv:
##
##     year

## The following objects are masked from tuesd (pos = 3):
##
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year

## The following objects are masked from sub5:
##
##     country, grand_total

## The following objects are masked from sub4:
##
##     country, grand_total

## The following objects are masked from sub3:
##
##     country, grand_total

## The following objects are masked from sub2:
##
##     country, grand_total

## The following objects are masked from sub1:
##
##     country, grand_total

## The following objects are masked from total_2020:
##
##     grand_total, year

## The following objects are masked from total_2019:
##
##     grand_total, year

## The following objects are masked from plastics (pos = 11):
##
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year

## The following objects are masked from plastics (pos = 12):
##
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year

## The following objects are masked from plastics (pos = 13):
```

```
##
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year
## The following objects are masked from plastics (pos = 14):
##
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year
## The following objects are masked from plastics (pos = 15):
##
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year
```

```
summary(pvc)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00    0.00    0.00    0.35    0.00  622.00    4328
```

```
plot(pvc,grand_total)
```



From the summary satistics of pvc is given above, the minimum amount of pvc is 0 as well but the maximum amount is 1183. Also, pvc owns the smallest mean in 3 variables that is 0.635. By observing the scatter plot for pvc above, we find a negative linear relationship between pvc and grand_total.

PS(Polystyrene) vs Grand Total

```
tuesd <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/202
## Rows: 13380 Columns: 14
## -- Column specification ------------------------------------------------------
## Delimiter: ","
## chr  (2): country, parent_company
## dbl (12): year, empty, hdpe, ldpe, o, pet, pp, ps, pvc, grand_total, num_eve...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
attach(tuesd)
```

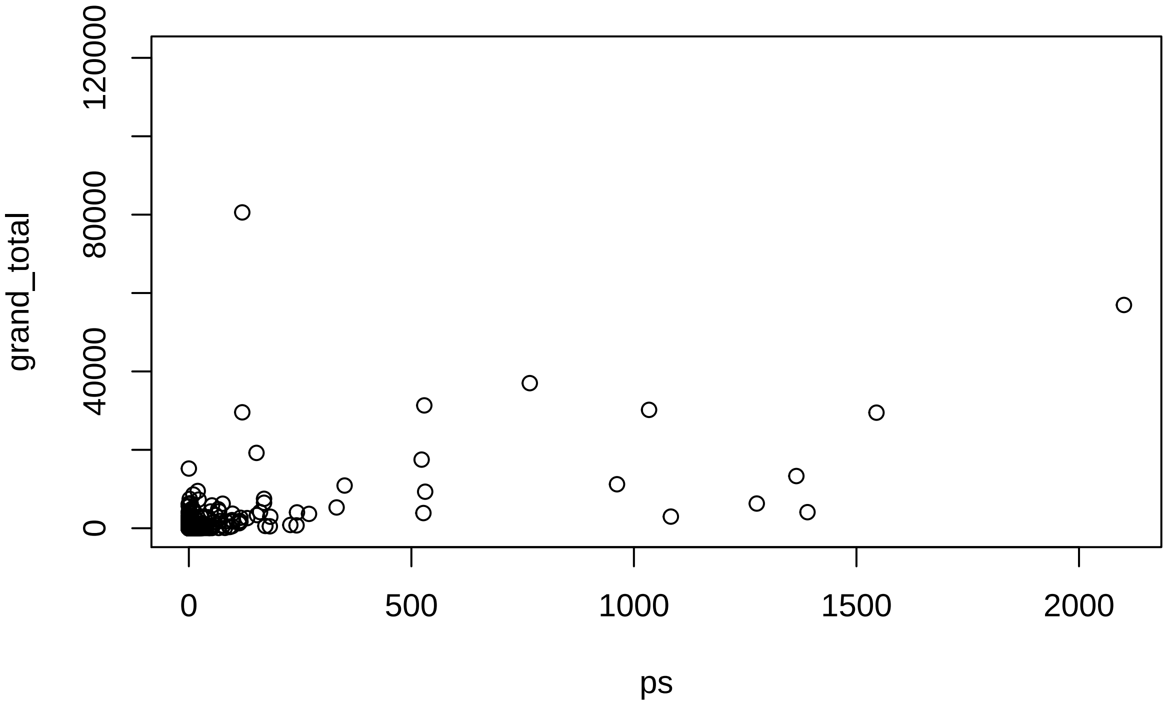
```
## The following object is masked _by_ .GlobalEnv:
##
##     year
## The following objects are masked from tuesd (pos = 3):
##
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year
## The following objects are masked from tuesd (pos = 4):
##
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year
## The following objects are masked from sub5:
##
##     country, grand_total
## The following objects are masked from sub4:
##
##     country, grand_total
## The following objects are masked from sub3:
##
##     country, grand_total
## The following objects are masked from sub2:
##
##     country, grand_total
## The following objects are masked from sub1:
##
##     country, grand_total
## The following objects are masked from total_2020:
##
##     grand_total, year
## The following objects are masked from total_2019:
##
##     grand_total, year
## The following objects are masked from plastics (pos = 12):
##
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year
## The following objects are masked from plastics (pos = 13):
```

```
##
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year

## The following objects are masked from plastics (pos = 14):
##
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year

## The following objects are masked from plastics (pos = 15):
##
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year

## The following objects are masked from plastics (pos = 16):
##
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year
```

```r
summary(ps)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's 
##    0.000    0.000    0.000    1.862    0.000 2101.000     1972
```

```r
plot(ps,grand_total)
```



From the summary statistics of ps is given above, we observe that minimum count of ps is 0 and maximum count of ps is 2101. By observing the scatter plot for ps above, it is hard to analyze how the linear relationship is even though we remove the outliers.

```r
tuesd <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/202
```

```
## Rows: 13380 Columns: 14

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (2): country, parent_company
## dbl (12): year, empty, hdpe, ldpe, o, pet, pp, ps, pvc, grand_total, num_eve...

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
attach(tuesd)
```

```
## The following object is masked _by_ .GlobalEnv:
##
##     year

## The following objects are masked from tuesd (pos = 3):
##
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year

## The following objects are masked from tuesd (pos = 4):
##
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year

## The following objects are masked from tuesd (pos = 5):
##
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year

## The following objects are masked from sub5:
##
##     country, grand_total

## The following objects are masked from sub4:
##
##     country, grand_total

## The following objects are masked from sub3:
##
##     country, grand_total

## The following objects are masked from sub2:
##
##     country, grand_total

## The following objects are masked from sub1:
##
##     country, grand_total

## The following objects are masked from total_2020:
##
##     grand_total, year

## The following objects are masked from total_2019:
##
```

```
##     grand_total, year
```

```
## The following objects are masked from plastics (pos = 13):
## 
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year
```

```
## The following objects are masked from plastics (pos = 14):
## 
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year
```

```
## The following objects are masked from plastics (pos = 15):
## 
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year
```

```
## The following objects are masked from plastics (pos = 16):
## 
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year
```

```
## The following objects are masked from plastics (pos = 17):
## 
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year
```

```r
summary(grand_total)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's 
##      0.00      1.00      1.00     90.15      6.00 120646.00        14
```

From the summary statistics of grand_total given above, the maximum amount in a certain countries whole year is 120646.00 but the minimum amount is 0.00.

Interval Estimations

One population mean for Grand Total

```
plastics <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/
```

```
## Rows: 13380 Columns: 14

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (2): country, parent_company
## dbl (12): year, empty, hdpe, ldpe, o, pet, pp, ps, pvc, grand_total, num_eve...

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
attach (plastics)
```

```
## The following object is masked _by_ .GlobalEnv:
##
##     year

## The following objects are masked from tuesd (pos = 3):
##
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year

## The following objects are masked from tuesd (pos = 4):
##
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year

## The following objects are masked from tuesd (pos = 5):
##
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year

## The following objects are masked from tuesd (pos = 6):
##
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year

## The following objects are masked from sub5:
##
##     country, grand_total

## The following objects are masked from sub4:
##
##     country, grand_total

## The following objects are masked from sub3:
##
##     country, grand_total

## The following objects are masked from sub2:
##
##     country, grand_total

## The following objects are masked from sub1:
```

```
## 
##     country, grand_total
```

```
## The following objects are masked from total_2020:
## 
##     grand_total, year
```

```
## The following objects are masked from total_2019:
## 
##     grand_total, year
```

```
## The following objects are masked from plastics (pos = 14):
## 
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year
```

```
## The following objects are masked from plastics (pos = 15):
## 
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year
```

```
## The following objects are masked from plastics (pos = 16):
## 
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year
```

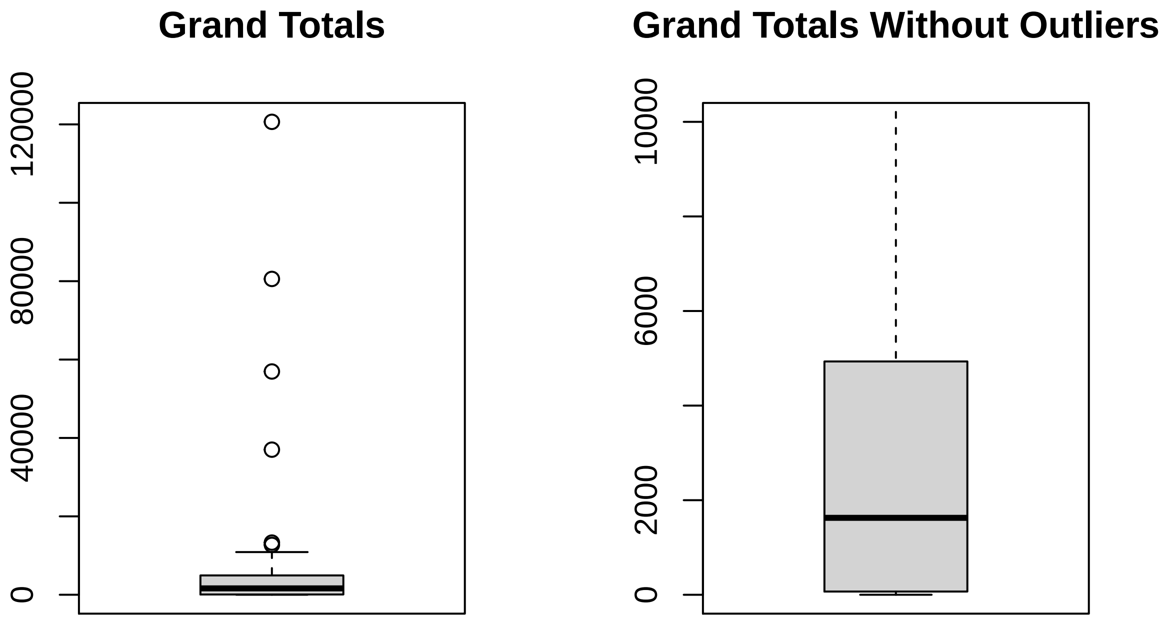
```
## The following objects are masked from plastics (pos = 17):
## 
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year
```

```
## The following objects are masked from plastics (pos = 18):
## 
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year
```

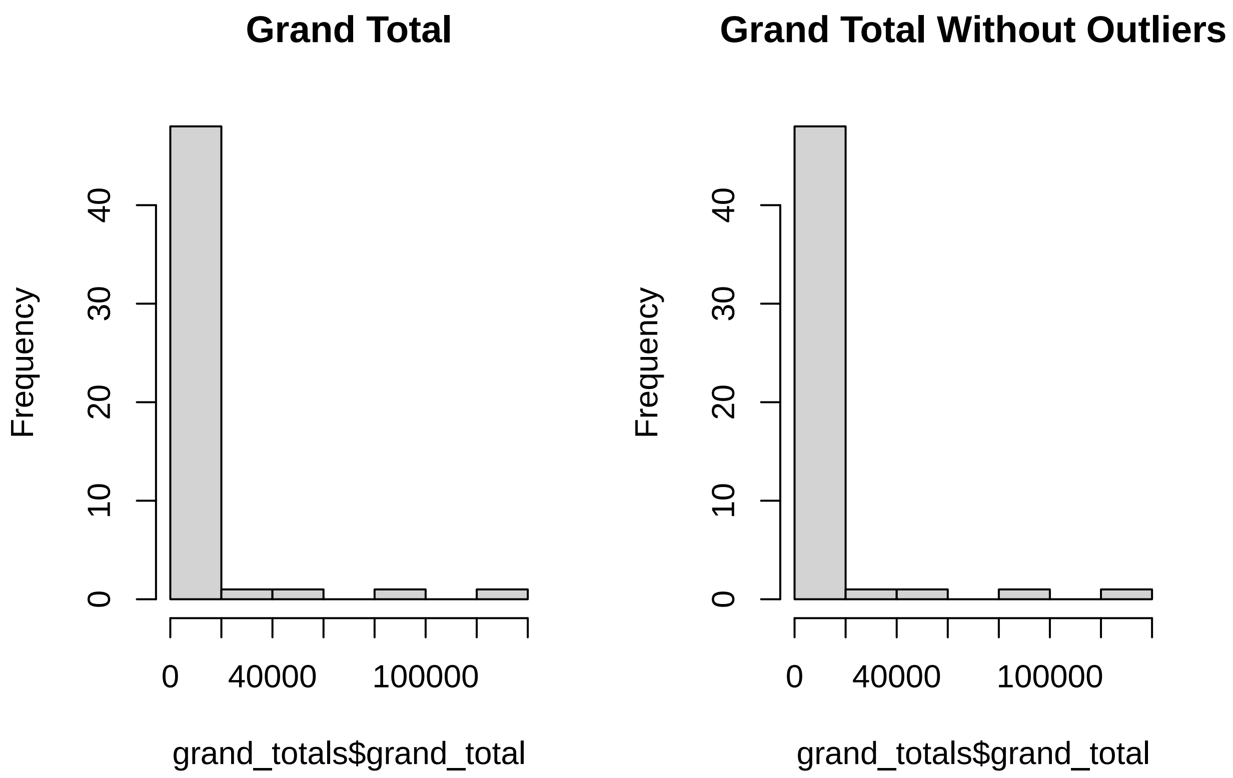
```r
grand_totals = subset(plastics, parent_company == "Grand Total")
summary(grand_totals$grand_total)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max. 
##      1.00     76.25   1627.00   8254.44   4907.00 120646.00
```

```r
par(mfrow=c(1,2))
boxplot(grand_totals$grand_total, main="Grand Totals")
boxplot(grand_totals$grand_total, ylim=c(1,10000), main="Grand Totals Without Outliers")
```



```r
par(mfrow=c(1,2))
hist(grand_totals$grand_total, main="Grand Total")
hist(grand_totals$grand_total, main="Grand Total Without Outliers")
```



```
grand_totals = subset(plastics, parent_company == "Grand Total")
x = mean(grand_totals$grand_total)
n = length(grand_totals$grand_total)
p = x/n
p
```

```
## [1] 158.7393
```

```
SE_p=sqrt(-1*(p*(1-p))/n)
SE_p
```

```
## [1] 21.94373
```

```
conf.level=0.95
alpha=1-conf.level
alpha
```

```
## [1] 0.05
```

```
z.score=qnorm(alpha/2,mean=0,sd=1,lower.tail = FALSE)
z.score
```

```
## [1] 1.959964
```

```
Margin.Error=z.score*SE_p
Margin.Error
```

```
## [1] 43.00892
```

```
lower.bound=p-Margin.Error
lower.bound
```

```
## [1] 115.7304
```

```
upper.bound=p+Margin.Error
upper.bound
```

```
## [1] 201.7482
```

```
Ninetyfive.confidence.interval=data.frame(lower.bound,upper.bound)
Ninetyfive.confidence.interval
```

```
##   lower.bound upper.bound
## 1    115.7304    201.7482
```

We began analyzing the grand total of plastics by cleaning and filtering the data. The data was filtered by removing the redundant information like the consecutive count of each plastic and only focusing on the total of plastic found in each country. This was done by using a `subset` function on our dataset to only select the rows that contained `Grand total` in the `parent_company` variable. We found that when removing key outliers, the data provided a clearer understanding of the general spread of more low plastic producing countries. From our summary and plots, we can see grand_total is very right skewed with a significant difference between the median and mode. The unimodal shape doesn't not initially indicate a normal distribution, however this is expected in real world results as we know few key countries produce the largest amount of plastic waste. Moving onto our confidence interval, we are 95% confident that the grand total parameter is between 115.7304 and 201.7482 from our plastics dataset.

## One Population Proportion for HDPE(High density polyethylene) Count

```
tuesd <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/202
```

```
## Rows: 13380 Columns: 14
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (2): country, parent_company
## dbl (12): year, empty, hdpe, ldpe, o, pet, pp, ps, pvc, grand_total, num_eve...
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
attach(tuesd)
```

```
## The following object is masked _by_ .GlobalEnv:
##
##     year
```

```
## The following objects are masked from plastics (pos = 3):
##
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year
```

```
## The following objects are masked from tuesd (pos = 4):
##
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year
## The following objects are masked from tuesd (pos = 5):
##
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year

## The following objects are masked from tuesd (pos = 6):
##
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year

## The following objects are masked from tuesd (pos = 7):
##
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year

## The following objects are masked from sub5:
##
##     country, grand_total

## The following objects are masked from sub4:
##
##     country, grand_total

## The following objects are masked from sub3:
##
##     country, grand_total

## The following objects are masked from sub2:
##
##     country, grand_total

## The following objects are masked from sub1:
##
##     country, grand_total

## The following objects are masked from total_2020:
##
##     grand_total, year

## The following objects are masked from total_2019:
##
##     grand_total, year

## The following objects are masked from plastics (pos = 15):
##
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year

## The following objects are masked from plastics (pos = 16):
##
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year

## The following objects are masked from plastics (pos = 17):
##
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year

## The following objects are masked from plastics (pos = 18):
##
```

```
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year
```

```
## The following objects are masked from plastics (pos = 19):
##
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year
```

```
x=15.29
n=13307
p=x/n
p
```

```
## [1] 0.001149019
```

```
SE_p=sqrt((p*(1-p))/n)
SE_p
```

```
## [1] 0.0002936797
```

```
conf.level=0.95
alpha=1-conf.level
alpha
```

```
## [1] 0.05
```

```
z.score=qnorm(alpha/2,mean=0,sd=1,lower.tail = FALSE)
z.score
```

```
## [1] 1.959964
```

```
Margin.Error=z.score*SE_p
Margin.Error
```

```
## [1] 0.0005756016
```

```
lower.bound=p-Margin.Error
lower.bound
```

```
## [1] 0.0005734177
```

```
upper.bound=p+Margin.Error
upper.bound
```

```
## [1] 0.001724621
```

```
Ninetyfive.confidence.interval=data.frame(lower.bound,upper.bound)
Ninetyfive.confidence.interval
```

```
##    lower.bound upper.bound
## 1 0.0005734177 0.001724621
```

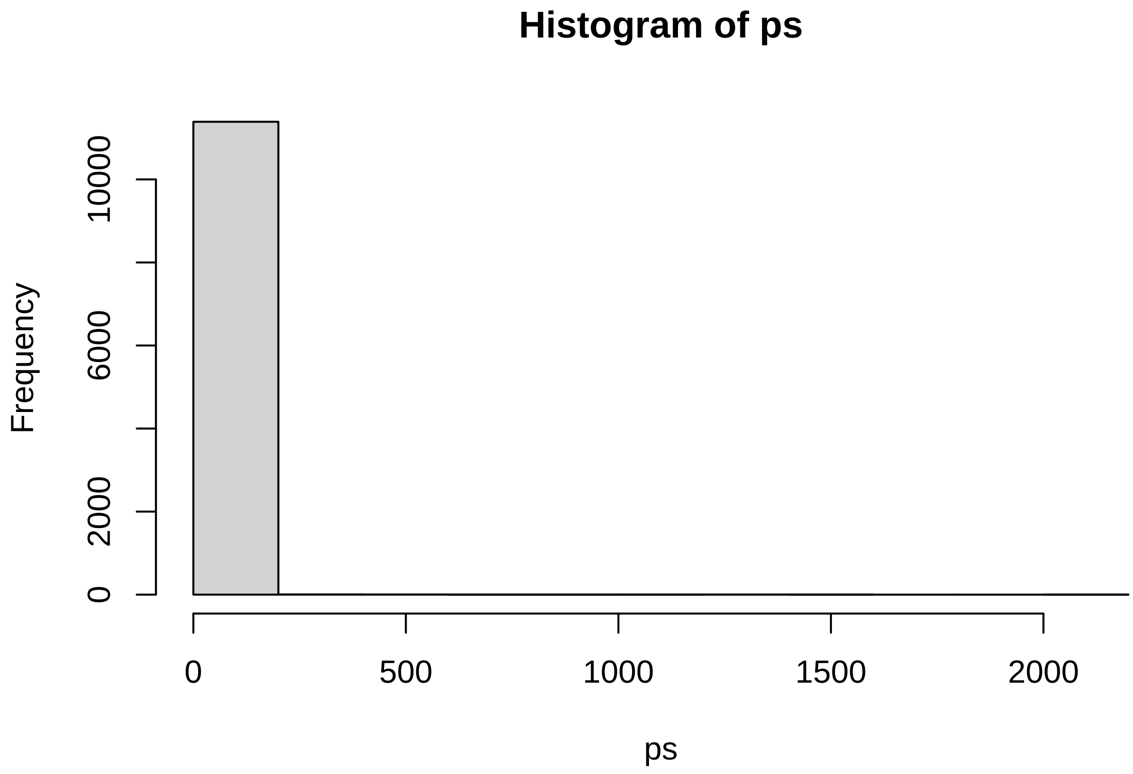
Then, the 95% confidence intervals for population proportion of hdpe is given below. We are 95% confident that the true proportion hdpe is between 0.0005734177 and 0.001724621.

Ratio of the Two Population Variances of PS(Polystyrene count) and HDPE(High density polyethylene count)

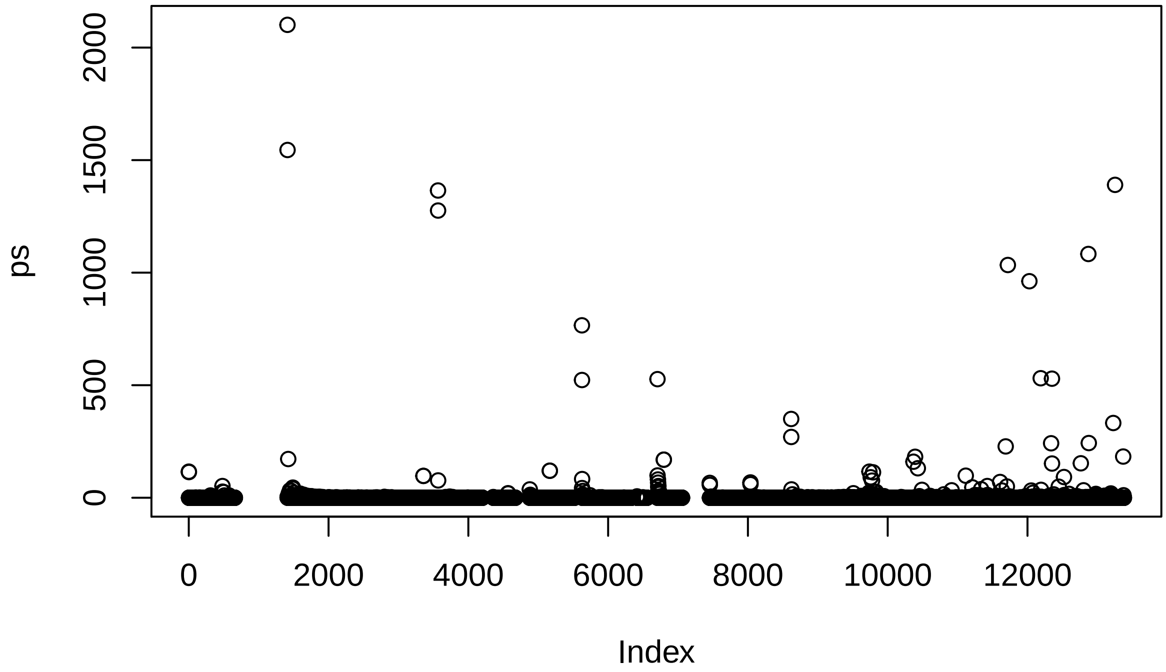
```
favstats(ps)
```

```
##  min Q1 median Q3  max     mean       sd     n missing
##    0  0       0  0 2101 1.862114 39.73706 11408    1972
```

```
hist(ps)
```



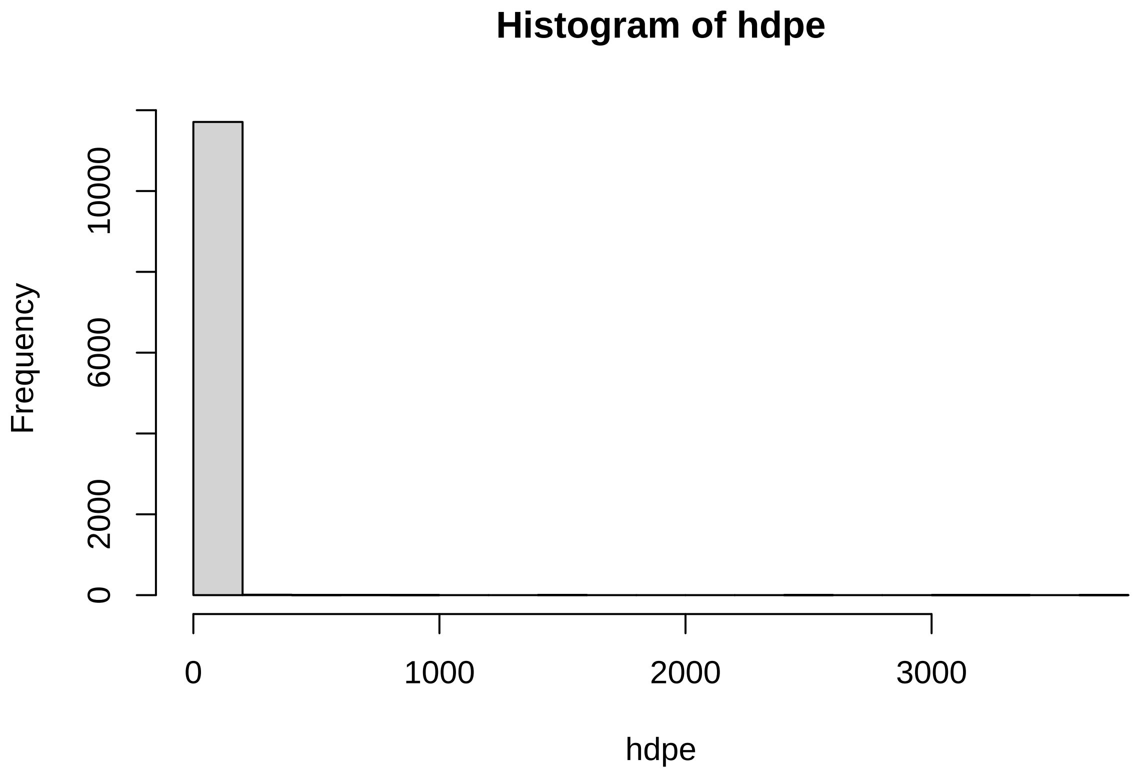
```
plot(ps)
```



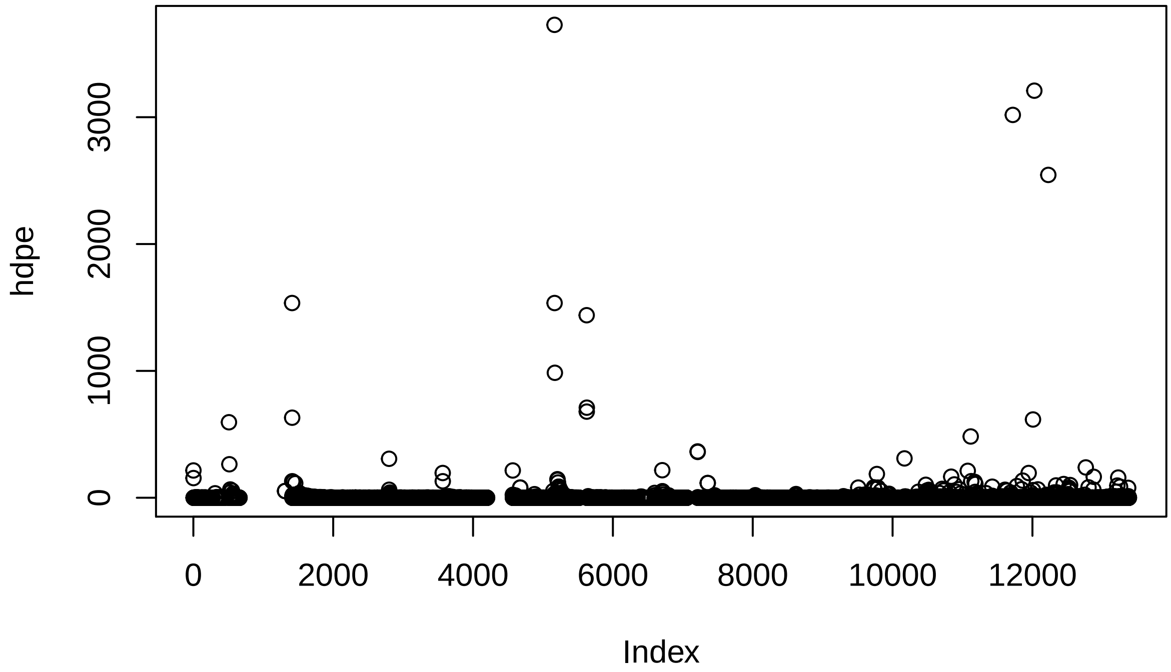
```
favstats(hdpe)
```

```
##  min Q1 median Q3  max    mean       sd     n missing
##    0  0      0  0 3728 3.04602 66.12304 11734    1646
```

```
hist(hdpe)
```


Histogram of hdpe
Frequency
0
2000
6000
10000
0
1000
2000
3000
hdpe

```
plot(hdpe)
```



```
favstats(ps ~ hdpe)
```

```
##      hdpe min     Q1 median     Q3 max       mean          sd     n
## 1       0   0   0.00    0.0   0.00 182  0.2017690   3.1285411 10175
## 2       1   0   0.00    0.0   0.00  97  0.3768473   4.9038514   406
## 3       2   0   0.00    0.0   0.00 531  5.4031008  47.4778763   129
## 4       3   0   0.00    0.0   0.00  77  1.3424658   9.0955105    73
## 5       4   0   0.00    0.0   0.00  17  1.0000000   3.2025631    40
## 6       5   0   0.00    0.0   0.00   3  0.1764706   0.6262243    34
## 7       6   0   0.00    0.0   0.00  11  1.0909091   3.2206047    22
## 8       7   0   0.00    0.0   0.00  60  3.5294118  14.5521375    17
## 9       8   0   0.00    0.0   0.00  27  1.7272727   5.9696201    22
## 10      9   0   0.00    0.0   0.00  11  1.3333333   3.2440422    15
## 11     10   0   0.00    0.0   0.00   2  0.1538462   0.5547002    13
## 12     11   0   0.00    0.0   0.00   0  0.0000000   0.0000000    11
## 13     12   0   0.00    0.0   0.00   2  0.1176471   0.4850713    17
## 14     13   0   0.00    1.0   8.00 172 24.1111111  56.4102040     9
## 15     14   0   0.00    0.0   0.00  10  1.6666667   4.0824829     6
## 16     15   0   0.00    0.0   6.00  15  3.8571429   6.6440091     7
## 17     16   0   0.00    0.0   0.00   0  0.0000000   0.0000000     2
## 18     17   0   0.00    0.0  67.50 270 67.5000000 135.0000000     4
## 19     18   0   0.25   46.0 149.50 169 71.6666667  83.1713092     6
## 20     19   0  16.50   33.0  49.50  66 33.0000000  46.6690476     2
## 21     20   0  17.00   34.0  51.00  68 34.0000000  48.0832611     2
## 22     21   0   0.00    0.0   0.00   0  0.0000000   0.0000000     5
## 23     22   0   0.50    1.0   1.50   2  1.0000000   1.0000000     3
## 24    23    0    0.00    0.0    0.00   19    2.5555556     6.3069626     9
## 25    24    0    0.00    0.0    0.00    1    0.2000000     0.4472136     5
## 26    25    0    0.00    0.0    0.00    0    0.0000000     0.0000000     2
## 27    26    0    0.00    0.0    3.50    7    2.3333333     4.0414519     3
## 28    28    0    9.25   18.5   27.75   37   18.5000000    26.1629509     2
## 29    29    0    0.00    0.0  175.00  350  116.6666667   202.0725942     3
## 30    30    0    0.00    0.0    0.00    0    0.0000000     0.0000000     4
## 31    31    0    0.00    0.0    0.00    0    0.0000000            NA     1
## 32    32    0   38.00   76.0  114.00  152   76.0000000   107.4802307     2
## 33    33    0    0.00    0.0    0.00    0    0.0000000     0.0000000     2
## 34    34    0    0.00    0.0    0.00    0    0.0000000     0.0000000     2
## 35    35    0    0.00    0.0    5.00   10    3.3333333     5.7735027     3
## 36    36    0    0.00    0.0    9.50   38    9.5000000    19.0000000     4
## 37    37    0    0.00    0.0    0.00    0    0.0000000            NA     1
## 38    38   NA      NA     NA      NA   NA          NaN            NA     0
## 39    39    0    0.00    0.0    0.00    0    0.0000000            NA     1
## 40    40    7    7.00    7.0    7.00    7    7.0000000            NA     1
## 41    41    0    0.00    0.0  114.00  228   76.0000000   131.6358614     3
## 42    42    0  132.25  264.5  396.75  529  264.5000000   374.0594872     2
## 43    43    0    0.50    1.0    1.50    2    1.0000000     1.4142136     2
## 44    44    0    0.00    0.0    0.00    0    0.0000000            NA     1
## 45    46    0    8.25   16.5   24.75   33   16.5000000    23.3345238     2
## 46    47    0   40.00   80.0  120.00  160   80.0000000   113.1370850     2
## 47    51    0   24.75   49.5   74.25   99   49.5000000    70.0035713     2
## 48    53    1    1.00    1.0    1.00    1    1.0000000            NA     1
## 49    54   26   26.00   26.0   26.00   26   26.0000000            NA     1
## 50    56    0    0.75   16.5   32.75   35   17.0000000    19.0962474     4
## 51    57    0    0.00    0.0    0.00    0    0.0000000            NA     1
## 52    58    0    0.00    0.0    0.00    0    0.0000000            NA     1
## 53    60    0    0.00    0.0    0.00    0    0.0000000            NA     1
## 54    62    0    0.00    0.0    0.00    0    0.0000000            NA     1
## 55    63    0   17.50   35.0   52.50   70   35.0000000    49.4974747     2
## 56    64 1083 1083.00 1083.0 1083.00 1083 1083.0000000            NA     1
## 57    65    0    0.00    0.0    0.00    0    0.0000000            NA     1
## 58    66   22   22.00   22.0   22.00   22   22.0000000            NA     1
## 59    67    0    0.00    0.0    0.00    0    0.0000000     0.0000000     4
## 60    69  116  116.00  116.0  116.00  116  116.0000000            NA     1
## 61    70    0    0.00    0.0    0.00    0    0.0000000            NA     1
## 62    75  113  113.00  113.0  113.00  113  113.0000000            NA     1
## 63    78    0    0.00    0.0   91.50  183   61.0000000   105.6550993     3
## 64    79   92   92.00   92.0   92.00   92   92.0000000            NA     1
## 65    80   14   15.50   17.0   18.50   20   17.0000000     4.2426407     2
## 66    82    0    8.25   16.5   24.75   33   16.5000000    23.3345238     2
## 67    87    0    0.00   26.0  386.50 1390  360.5000000   686.7709468     4
## 68    92    0    0.00    0.0    0.00    0    0.0000000            NA     1
## 69    96   10   10.00   10.0   10.00   10   10.0000000            NA     1
## 70    97  242  242.00  242.0  242.00  242  242.0000000            NA     1
## 71   100    6    6.00    6.0    6.00    6    6.0000000            NA     1
## 72   102    0    0.00    0.0    0.00    0    0.0000000            NA     1
## 73   105    0    0.00    0.0    0.00    0    0.0000000            NA     1
## 74   108    0   12.25   24.5   36.75   49   24.5000000    34.6482323     2
## 75   114    0    0.00    0.0    0.00    0    0.0000000            NA     1
## 76   116   NA      NA     NA      NA   NA          NaN            NA     0
## 77   120    0    0.00    0.0    0.00    0    0.0000000            NA     1
## 78   124   19   19.00   19.0   19.00   19   19.0000000           NA     1
## 79   125    0    0.00    0.0    0.00    0    0.0000000           NA     1
## 80   130    1  319.75  638.5  957.25 1276  638.5000000  901.5611460     2
## 81   131    2    2.00    2.0    2.00    2    2.0000000           NA     1
## 82   135    0    0.00    0.0    0.00    0    0.0000000           NA     1
## 83   141    0    0.00    0.0    0.00    0    0.0000000           NA     1
## 84   146    0    0.00    0.0    0.00    0    0.0000000           NA     1
## 85   155  114  114.00  114.0  114.00  114  114.0000000           NA     1
## 86   159  332  332.00  332.0  332.00  332  332.0000000           NA     1
## 87   165  243  243.00  243.0  243.00  243  243.0000000           NA     1
## 88   166    5    5.00    5.0    5.00    5    5.0000000           NA     1
## 89   187   76   76.00   76.0   76.00   76   76.0000000           NA     1
## 90   195    0  341.25  682.5 1023.75 1365  682.5000000  965.2007563     2
## 91   212    0    0.00    0.0    0.00    0    0.0000000           NA     1
## 92   215   20   44.00   68.0   92.00  116   68.0000000   67.8822510     2
## 93   217  527  527.00  527.0  527.00  527  527.0000000           NA     1
## 94   239  153  153.00  153.0  153.00  153  153.0000000           NA     1
## 95   264    0    0.00    0.0    0.00    0    0.0000000           NA     1
## 96   307    4    4.00    4.0    4.00    4    4.0000000           NA     1
## 97   310    0    0.00    0.0    0.00    0    0.0000000           NA     1
## 98   361   NA      NA     NA      NA   NA          NaN           NA     0
## 99   365   NA      NA     NA      NA   NA          NaN           NA     0
## 100  483   98   98.00   98.0   98.00   98   98.0000000           NA     1
## 101  595   10   10.00   10.0   10.00   10   10.0000000           NA     1
## 102  617    0    0.00    0.0    0.00    0    0.0000000           NA     1
## 103  631 1545 1545.00 1545.0 1545.00 1545 1545.0000000           NA     1
## 104  679  523  523.00  523.0  523.00  523  523.0000000           NA     1
## 105  710   43   43.00   43.0   43.00   43   43.0000000           NA     1
## 106  985    0    0.00    0.0    0.00    0    0.0000000           NA     1
## 107 1439  766  766.00  766.0  766.00  766  766.0000000           NA     1
## 108 1535  120  615.25 1110.5 1605.75 2101 1110.5000000 1400.7785335     2
## 109 2545    2    2.00    2.0    2.00    2    2.0000000           NA     1
## 110 3018 1034 1034.00 1034.0 1034.00 1034 1034.0000000           NA     1
## 111 3209  962  962.00  962.0  962.00  962  962.0000000           NA     1
## 112 3728  120  120.00  120.0  120.00  120  120.0000000           NA     1
##     missing
## 1       550
## 2         1
## 3         2
## 4         1
## 5         2
## 6         0
## 7         1
## 8         2
## 9         0
## 10        2
## 11        0
## 12        0
## 13        0
## 14        0
## 15        0
## 16        0
## 17        0
## 18        0
## 19        0
## 20        0
## 21        1
## 22        0
## 23        0
## 24        0
## 25        1
## 26        0
## 27        0
## 28        0
## 29        1
## 30        0
## 31        0
## 32        0
## 33        0
## 34        0
## 35        0
## 36        0
## 37        0
## 38        1
## 39        0
## 40        0
## 41        0
## 42        0
## 43        0
## 44        0
## 45        0
## 46        0
## 47        0
## 48        3
## 49        0
## 50        0
## 51        0
## 52        0
## 53        0
## 54        0
## 55        0
## 56        0
## 57        0
## 58        0
## 59        0
## 60        0
## 61        0
## 62        0
## 63        0
## 64        0
## 65        2
## 66        0
## 67        0
## 68        0
## 69        0
## 70        0
## 71        0
## 72        0
## 73         0
## 74         0
## 75         0
## 76         2
## 77         0
## 78         0
## 79         0
## 80         0
## 81         0
## 82         0
## 83         0
## 84         0
## 85         0
## 86         0
## 87         0
## 88         0
## 89         0
## 90         0
## 91         0
## 92         0
## 93         0
## 94         0
## 95         0
## 96         0
## 97         0
## 98         1
## 99         1
## 100        0
## 101        0
## 102        0
## 103        0
## 104        0
## 105        0
## 106        0
## 107        0
## 108        0
## 109        0
## 110        0
## 111        0
## 112        0
```

```
#alpha = 0.05
#alpha/2 = 0.025
#1 - alpha/2 = 1 - 0.025 = 0.975

n1 = 11408
n2 = 11734

lower.bound = qf(0.025, df1 = n1 -1, df2 = n2 - 1)
higher.bound = qf(0.975, df1 = n1 -1, df2 = n2 - 1)
Interval.Estimate = data.frame(lower.bound, upper.bound)
Interval.Estimate
```

```
##   lower.bound upper.bound
## 1   0.9642035 0.001724621
```

To find the difference in population variances ps and hdpe, we used the central limit theorem which states that as sample size gets larger, the variances of the sample sizes will approximate towards the variance of the popuation.The sample size for ps is 11408 and the sample size of hdpe is 11734 which is large enough for the CLT to be used. Since 1 is in this confidence interval, there is no evidence that the population variances of these two plastic types differ.

One Population Mean or Number of Events

```
attach(plastics)
```

```
## The following object is masked _by_ .GlobalEnv:
##
##     year

## The following objects are masked from tuesd (pos = 3):
##
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year

## The following objects are masked from plastics (pos = 4):
##
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year

## The following objects are masked from tuesd (pos = 5):
##
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year

## The following objects are masked from tuesd (pos = 6):
##
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year

## The following objects are masked from tuesd (pos = 7):
##
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year

## The following objects are masked from tuesd (pos = 8):
##
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year

## The following objects are masked from sub5:
##
##     country, grand_total

## The following objects are masked from sub4:
##
##     country, grand_total

## The following objects are masked from sub3:
##
##     country, grand_total

## The following objects are masked from sub2:
##
##     country, grand_total
## The following objects are masked from sub1:
##
##     country, grand_total
```

```
## The following objects are masked from total_2020:
##
##     grand_total, year
```

```
## The following objects are masked from total_2019:
##
##     grand_total, year
```

```
## The following objects are masked from plastics (pos = 16):
##
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year
```

```
## The following objects are masked from plastics (pos = 17):
##
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year
```

```
## The following objects are masked from plastics (pos = 18):
##
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year
```

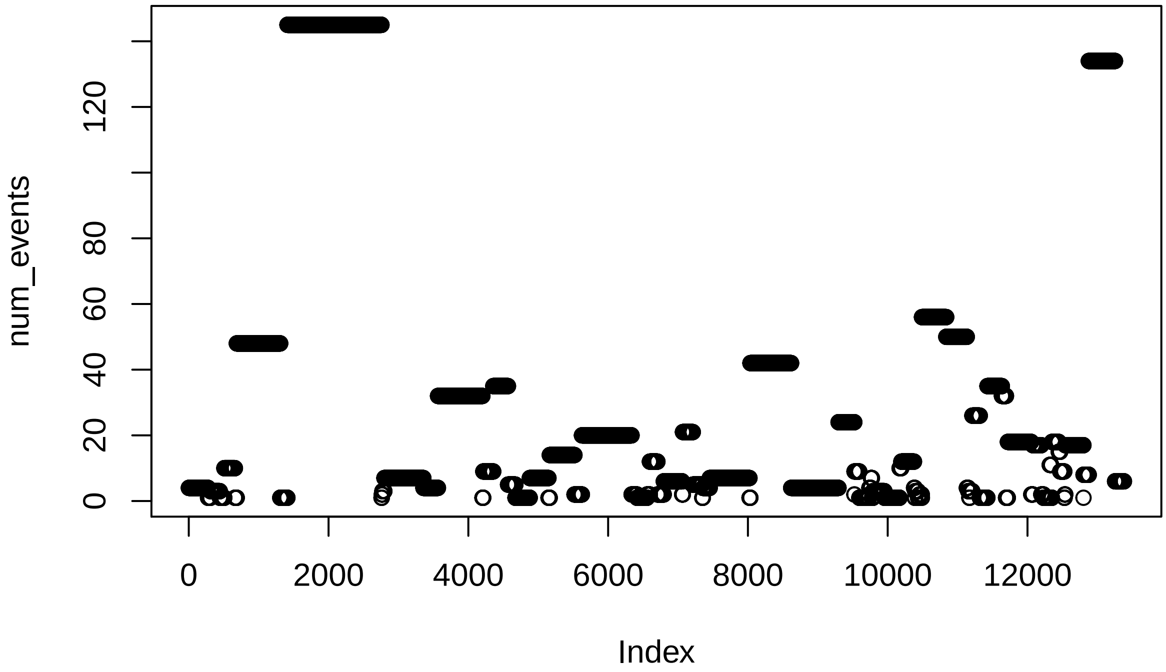
```
## The following objects are masked from plastics (pos = 19):
##
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year
```

```
## The following objects are masked from plastics (pos = 20):
##
##     country, empty, grand_total, hdpe, ldpe, num_events, o,
##     parent_company, pet, pp, ps, pvc, volunteers, year
```

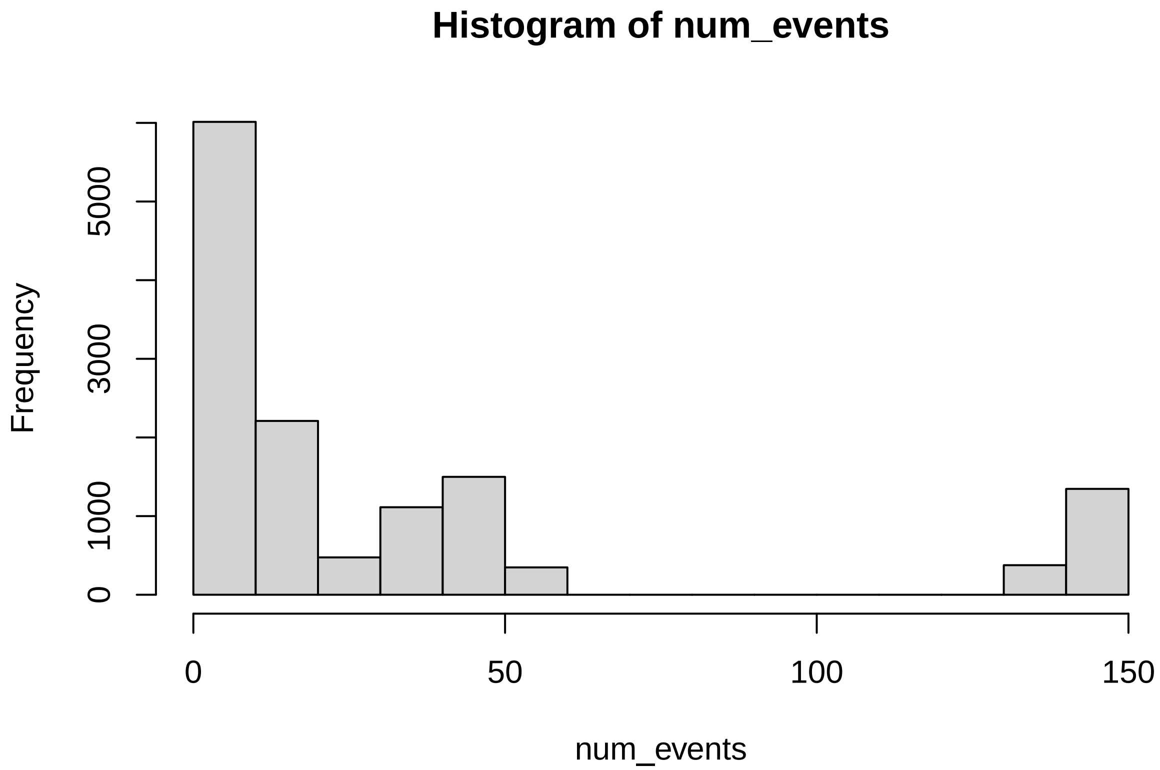
```
favstats(num_events)
```

```
##  min Q1 median Q3 max     mean       sd     n missing
##    1  4     15 42 145 33.36981 44.70864 13380       0
```

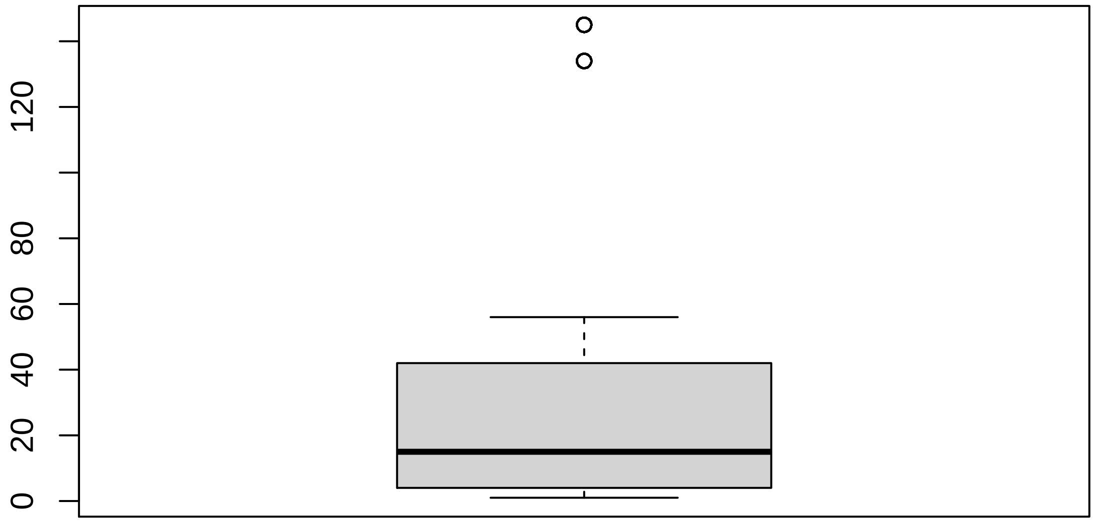
```
plot(num_events)
```



```
hist(num_events)
```



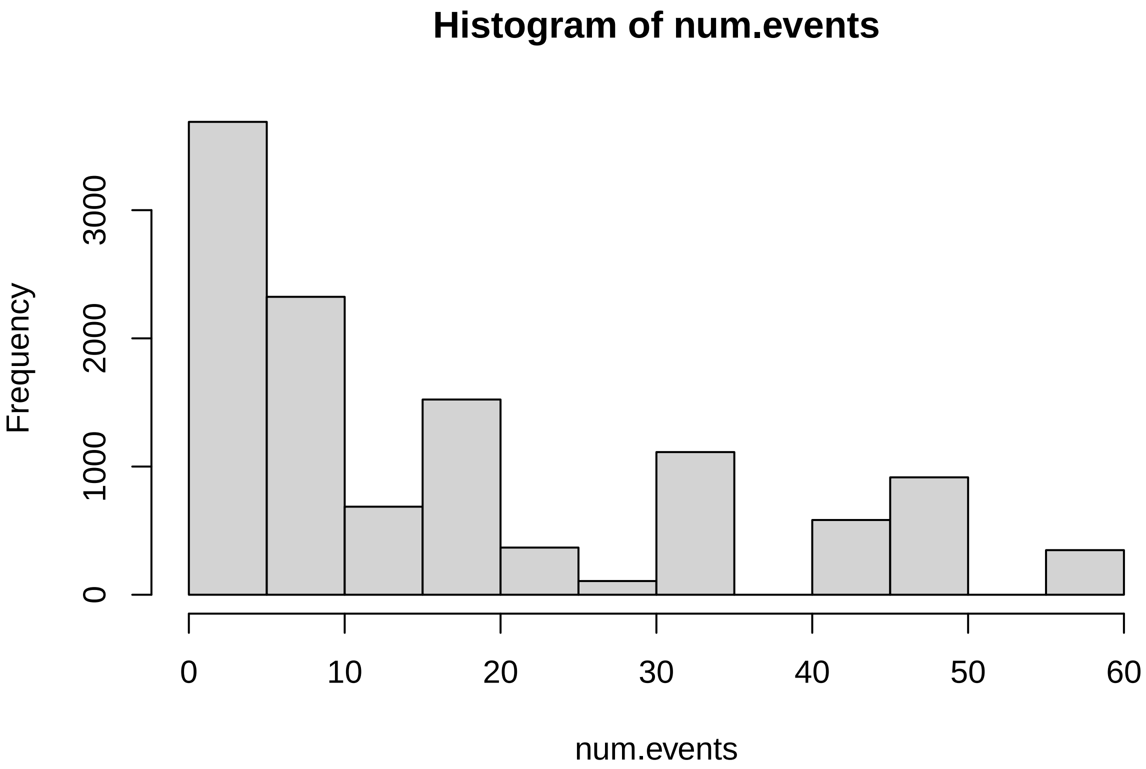
```r
boxplot(num_events)
```



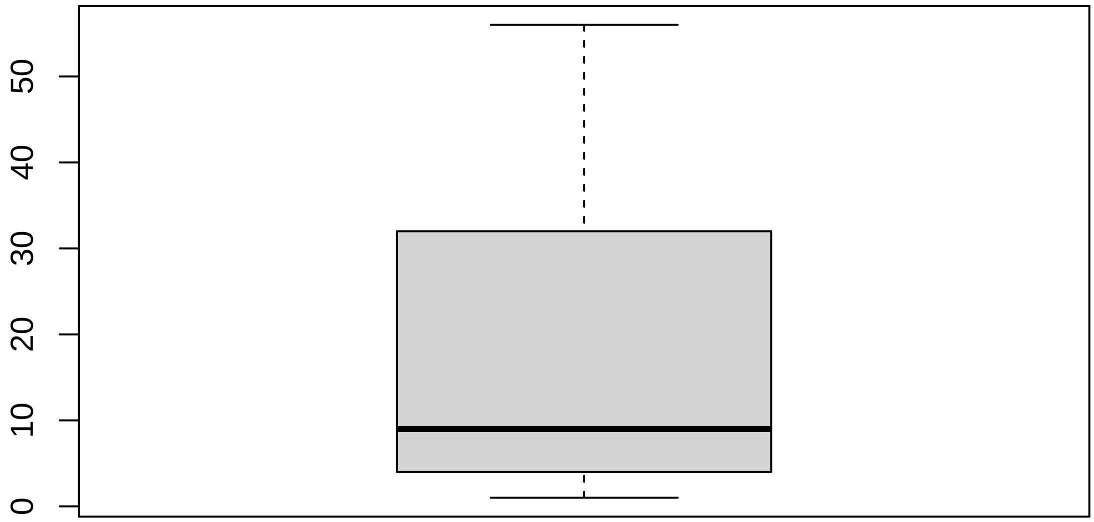
```
q1 = 4.00
q3 = 32.00
iqr = q3 - q1
x1 = num_events[q1 - (1.5)*(iqr) < num_events]
num.events = x1[x1 < q3 + (1.5)*(iqr) ]
favstats(num.events)
```

```
##  min Q1 median Q3 max     mean       sd     n missing
##    1  4      9 32  56 17.23572 16.38008 11658       0
```

```
hist(num.events)
```



```
boxplot(num.events)
```



```
s = 16.38008
n = 11658
x.bar = 17.23572
df = n - 1
t.score = qt(0.975, df)
t.score
```

```
## [1] 1.960168
```

```
SE = s/(sqrt(n))
SE
```

```
## [1] 0.1517064
```

```
ME = t.score*SE
ME
```

```
## [1] 0.29737
```

```
lower.bound = x.bar - ME
lower.bound
```

```
## [1] 16.93835
```

```
upper.bound = x.bar + ME
upper.bound
```

```
## [1] 17.53309
```

```
Interval.Estimate = data.frame(lower.bound, upper.bound)
Interval.Estimate
```

```
##   lower.bound upper.bound
## 1    16.93835    17.53309
```

Without the outliers, the data is still skewed but to a lesser degree from before. We can see from the histogram plots that there are more prominent peak. Because the assumption of normality is violated, the t distribution is robust and is able to work in this case without the prescence of strong outliers. We will now have to look at the assumptions and conditions of the t distribution

Independence Assumption: The data values of grand_total are independent from each other Randomization Condition: The data is from a random sample Normal Assumption: The data is slightly right skewed but t distribution works well for slightly less normal data

The Confidence Interval for population mean is:

x_bar += Margin of error of x_bar x_bar +=(t * _(n-1) * SE(X)) where standard error is s/ sqrt(n)

The standard error is 0.1517064 which is the amount of variation expected when we sample 11658 events The Mean estimate tells us that we are 95% confident that the true mean of the frequency of the number of events is 0.29737 % within 17.23572 We are confident that the true mean of the frequency of the number events is between 16.93835 and 17.53309. This confidence interval is fairly small as we had a large value of n.

Difference Between Two Population Means of Total Plastic Count in 2019 and 2020

```r
total_2019 <- subset(plastics, select = c(grand_total, year))
total_2019 <- subset(total_2019, grand_total > 500 & year <= 2019)
attach(total_2019)
```

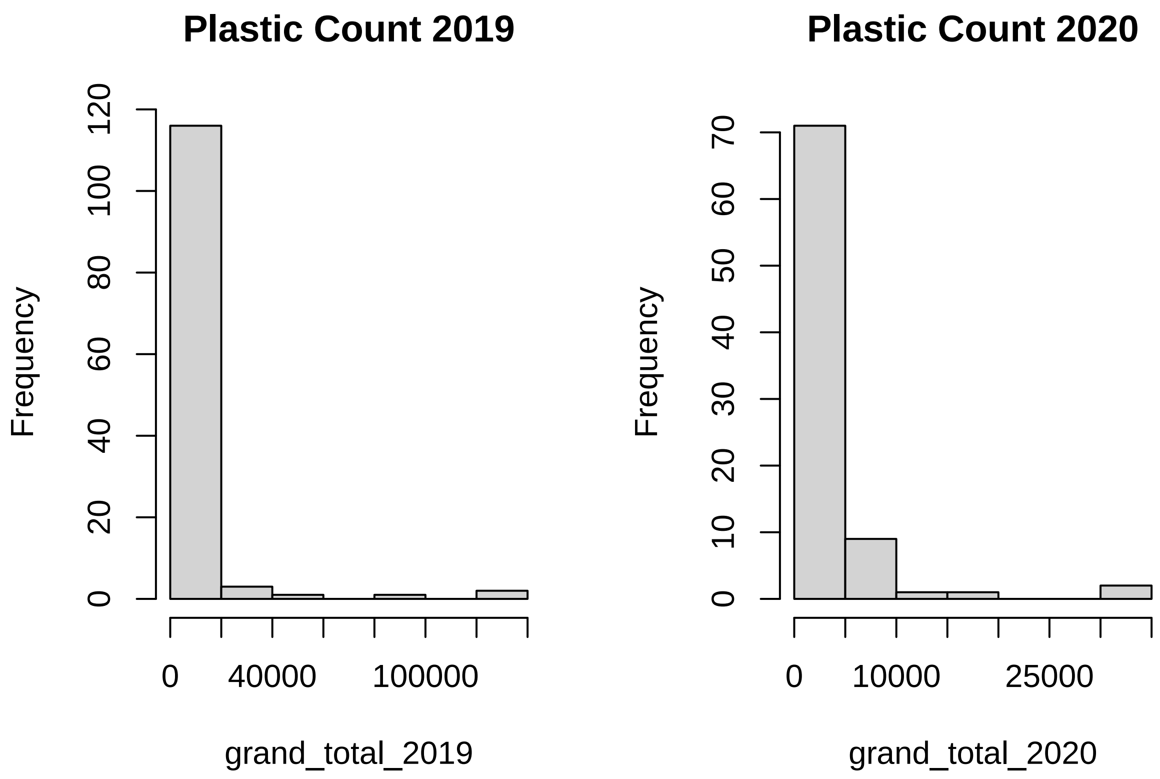
```
## The following object is masked _by_ .GlobalEnv:
##
##     year

## The following objects are masked from plastics (pos = 3):
##
##     grand_total, year

## The following objects are masked from tuesd (pos = 4):
##
##     grand_total, year

## The following objects are masked from plastics (pos = 5):
##
##     grand_total, year

## The following objects are masked from tuesd (pos = 6):
##
##     grand_total, year

## The following objects are masked from tuesd (pos = 7):
##
##     grand_total, year

## The following objects are masked from tuesd (pos = 8):
##
##     grand_total, year

## The following objects are masked from tuesd (pos = 9):
##
##     grand_total, year
## The following object is masked from sub5:
##
##     grand_total

## The following object is masked from sub4:
##
##     grand_total

## The following object is masked from sub3:
##
##     grand_total

## The following object is masked from sub2:
##
##     grand_total

## The following object is masked from sub1:
##
##     grand_total

## The following objects are masked from total_2020:
##
##     grand_total, year

## The following objects are masked from total_2019 (pos = 16):
##
##     grand_total, year

## The following objects are masked from plastics (pos = 17):
##
##     grand_total, year

## The following objects are masked from plastics (pos = 18):
##
##     grand_total, year

## The following objects are masked from plastics (pos = 19):
##
##     grand_total, year

## The following objects are masked from plastics (pos = 20):
##
##     grand_total, year

## The following objects are masked from plastics (pos = 21):
##
##     grand_total, year
```

```
grand_total_2019 <- grand_total
total_2020 <- subset(plastics, select = c(grand_total, year))
total_2020 <- subset(total_2020, grand_total > 500 & year >= 2020)
attach(total_2020)
```

```
## The following object is masked _by_ .GlobalEnv:
##
##     year

## The following objects are masked from total_2019 (pos = 3):
##
##     grand_total, year
## The following objects are masked from plastics (pos = 4):
##
##     grand_total, year

## The following objects are masked from tuesd (pos = 5):
##
##     grand_total, year

## The following objects are masked from plastics (pos = 6):
##
##     grand_total, year

## The following objects are masked from tuesd (pos = 7):
##
##     grand_total, year

## The following objects are masked from tuesd (pos = 8):
##
##     grand_total, year

## The following objects are masked from tuesd (pos = 9):
##
##     grand_total, year

## The following objects are masked from tuesd (pos = 10):
##
##     grand_total, year

## The following object is masked from sub5:
##
##     grand_total

## The following object is masked from sub4:
##
##     grand_total

## The following object is masked from sub3:
##
##     grand_total

## The following object is masked from sub2:
##
##     grand_total

## The following object is masked from sub1:
##
##     grand_total

## The following objects are masked from total_2020 (pos = 16):
##
##     grand_total, year

## The following objects are masked from total_2019 (pos = 17):
##
##     grand_total, year

## The following objects are masked from plastics (pos = 18):
##
##     grand_total, year

## The following objects are masked from plastics (pos = 19):
```

```
##
##     grand_total, year
## The following objects are masked from plastics (pos = 20):
##
##     grand_total, year
## The following objects are masked from plastics (pos = 21):
##
##     grand_total, year
## The following objects are masked from plastics (pos = 22):
##
##     grand_total, year
```

```
grand_total_2020 <- grand_total

par(mfrow = c(1,2))
#Histogram of 2019 total plastic count.
hist(grand_total_2019, main = "Plastic Count 2019")
#Histogram of 2020 total plastic count.
hist(grand_total_2020, main = "Plastic Count 2020")
```



```
t.test(grand_total_2019, grand_total_2020, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  grand_total_2019 and grand_total_2020
## t = 1.7358, df = 150.45, p-value = 0.08464
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -410.474 6347.187
## sample estimates:
## mean of x mean of y
##  6240.309  3271.952
```

As stated previously, before we found any confidence interval between year and grand_total, we cleaned the data such that we did not take into consideration of any outliers for the total count less than 500 since we think it would not be an accurate representation of total count of plastic for a country. To perform the t test to calculate the 95% confidence interval for difference in the means, we need to check the condition and make necessary assumptions. To begin with, we could assume that the groups are independent since it's trivial that the count of plastic in 2019 is unrelated to the count of plastic in 2020. Also, the randomization condition is met since it is reasonable to assume that the plastic counts are representative for the years 2019 and 2020. Looking at the histograms for total count in 2019 and 2020 above, we see they are slightly skewed to the right but we recognize that the t distribution works well for skewed data when the sample size is greater than 20 so the nearly normal condition is satisfied. Since all conditions were met, we were able to calculate the difference between two population means of total plastic count in 2019 and 2020. Unfortunately, since the value "0" is in this CI, we have no evidence to indicate that there is a difference in mean plastic count in 2019 and 2020.

Question

This data is explored from the organization "Break Free from Plastic", which is a global movement envisioning a future free from plastic pollution. The main purpose to collect this data was to raise awareness about the growing concern of plastic pollution around the world. In this statistical inference analysis we want to contribute to this awareness by reporting if there exists any difference in the plastic pollution count over the years. The question of interest that we are answering with this data set is whether there is any difference in the amount of plastic pollution produced on average in the years 2019 and 2020? We will use the independent t-test for conducting a hypothesis test with 0.05 significance level for comparing the mean of total plastic pollution count in 2019 and 2020 to analyze for any difference in pollution count.

Null & Alternative Hypotheses.

The null hypothesis is our current belief such that there is no difference in the total plastic pollution count on average in the years 2019 and 2020.. The alternative hypothesis is the question we proposed above such that there is some difference in the total plastic pollution count on average in the years 2019 and 2020.

Let μ_2019 denote the population mean for total count of plastic pollution in 2019. Let μ_2020 denote the population mean for total count of plastic pollution in 2020.

Null Hypothesis Ho : μ_2019 = μ_2020

Alternative Hypotheis Ha : μ_2019 $\neq$ μ_2020

Check Assumptions For Statistical Model

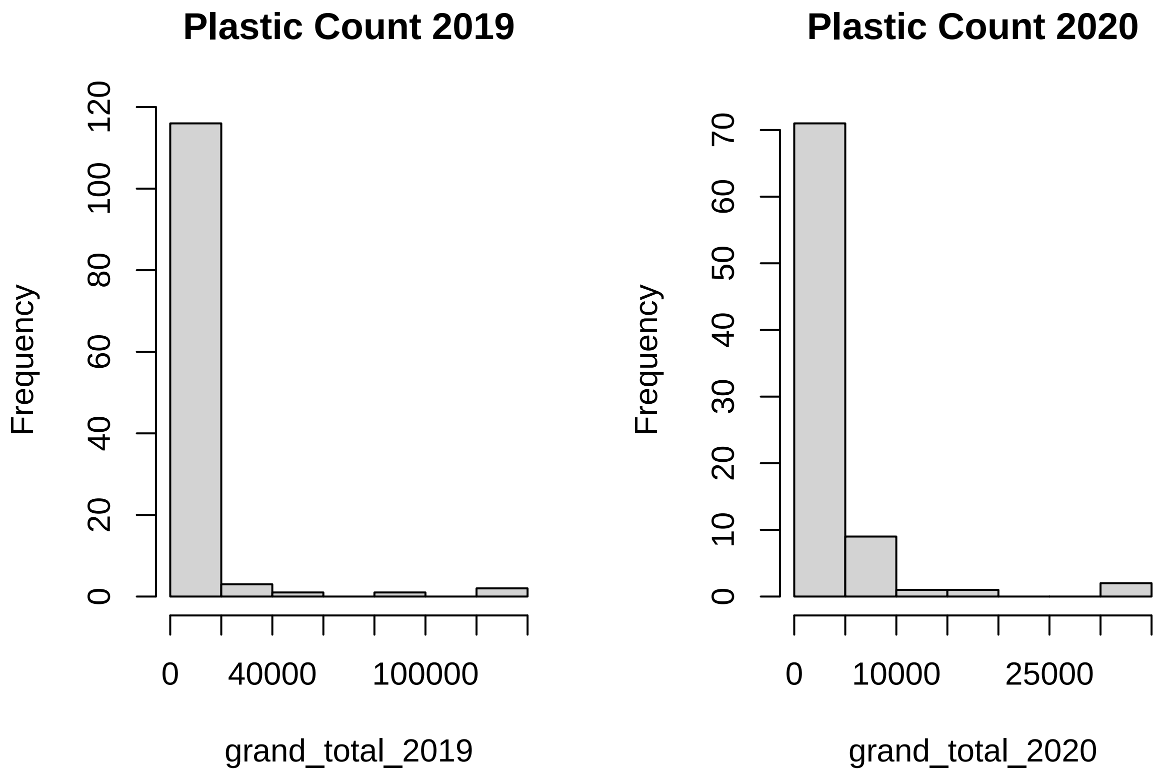
```
#Summary statistics for total plastic count in 2019
favstats(grand_total_2019)
```

```
##   min    Q1 median     Q3    max     mean       sd   n missing
##   515 676.5   1263 3684.5 120646 6240.309 17898.82 123       0
```

```
#Summary statistics for total plastic count in 2020
favstats(grand_total_2020)
```

```
##   min   Q1 median   Q3   max     mean       sd  n missing
##  509 727 1380.5 3911 31331 3271.952 5182.038 84       0
```

```r
par(mfrow = c(1,2))
#Histogram of 2019 total plastic count.
hist(grand_total_2019, main = "Plastic Count 2019")
#Histogram of 2020 total plastic count.
hist(grand_total_2020, main = "Plastic Count 2020")
```



In order to use the independent t-test, we need to check the necessary criteria and make valid assumptions to conduct the hypothesis test. To begin with, we could assume that the groups are independent since it's trivial that the count of plastic in 2019 is unrelated to the count of plastic in 2020. Also, the randomization condition is met since it is reasonable to assume that the plastic counts are representative for the years 2019 and 2020. Looking at the histograms for total count in 2019 and 2020 above, we see they are slightly skewed to the right but we recognize that the t distribution works well for skewed data when the sample size is greater than 20 so the nearly normal condition is satisfied. Therefore, we could use the t-test to determine if there is a difference in the mean plastic count in the years 2019 and 2020.

Test-Statistics

```r
t.test(grand_total_2019, grand_total_2020, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  grand_total_2019 and grand_total_2020
## t = 1.7358, df = 150.45, p-value = 0.08464
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -410.474 6347.187
## sample estimates:
## mean of x mean of y
##  6240.309  3271.952
```

T-Value & P-Value of Observed Test-Statistic

Since the assumptions and conditions were met above, the sampling distribution of the difference in sample means of two independent groups, divided by its standard error is modelled by a t-model. Assuming the null hypothesis is true, we get the value of the observed test statistic is t = 1.7358 with degrees of freedom df = 150.45. With the value of the observed test statistic, we get the p-value = 2 * P(t > 1.7358) = 0.08464. The p-value indicates the probability of observing something more extreme than what we have already found.

Decison

After conducting the t-test for the independent groups and observing the test statistic above, we could make a decision on whether we reject or fail to reject the null hypothesis. Since the p-value we observed is p-value = 0.08464 which is greater than the significance level of 0.05, we fail to reject the null hypothesis since we don't have any evidence against the null hypothesis. With the t-test we also found the 95% confidence interval (-410.474, 6347.187) for the true difference in mean of plastic count in 2019 and 2020. Since 0 is in the confidence interval, this provides additional support for not rejecting the null hypothesis.

Conclusion

In conclusion, since we fail to reject the null hypothesis, we do not have evidence to conclude there is a difference in the total plastic pollution count on average in the years 2019 and 2020.