Lab 5

Lab 5

Due Tuesday Feb1st - Recommended to complete this before starting the midterm

This lab we will look at some data from the plastic trash picked up during clean-up events around the world. I took this data set from the Tidy Tuesday website. You can read the documentation here, including the references and description of the different column names.

I have done some pre-processing of the data for you for this lab, to create two more easy-to-use data frames.

First read in the countrytotals.csv data frame

```
setwd("C:\\Users\\Benny Panjaitan\\Documents\\GitHub\\esp106-Naomi\\W5 Lab\\ESP106_week5_data\\")
tidytue <- read.csv("countrytotals.csv")
```

Have a look at the data frame. Then column "total" gives the total number of pieces of plastic picked up in that country in 2020. The columns "num_events" and "volunteers" give the number of trash pick-up events and the number of volunteers in that country. We are going to use this to investigate where the plastic trash problem is worst.

1. What 5 countries had the worst plastic problem as measured by the number of pieces of trash picked up?

Answer:

```
head (tidytue[order(tidytue$total,
                    decreasing=TRUE),])
```

```
##     X     country empty hdpe  ldpe     o   pet    pp   ps  pvc total num_events
## 35 35     Nigeria   316 7532 15470  8284 27390  2234 2011   16 63253       6030
## 37 37 Philippines   344   82  3939  4272  2949 41987  600 1011 55184       2040
## 46 46 Switzerland  2342  102  1528 45612  1018   902  722   51 52277       1746
## 21 21       India    22  489  6684  7369   895  1107   67  340 16973      19488
## 49 49        Togo     0    0     0 11233   759     0    2    0 11994          5
## 22 22   Indonesia     7 1304  1461  1473  2818  2912  145   36 10156      14700
##    volunteers
## 35     703165
## 37     109800
## 46      52380
## 21     122844
## 49        460
## 22     115248
```

Another way to show this using dplyr package is by using top_n function to pick 5 countries with highest "total" value, then using arrange function to sort the countries in descending order.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
top5 <- arrange(top_n(tidytue, 5, total), desc(total))
top5
```

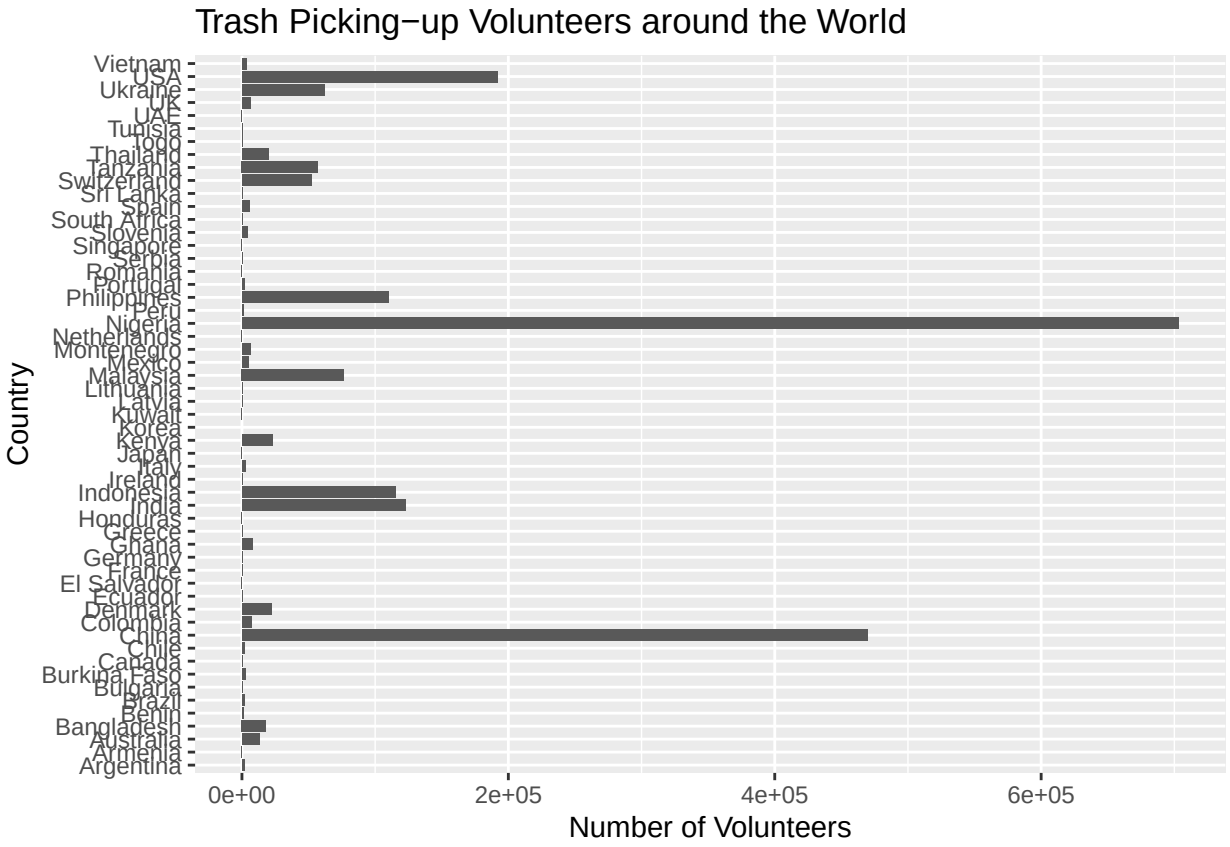
```
##    X     country empty hdpe  ldpe     o   pet    pp   ps  pvc total num_events
## 1 35     Nigeria   316 7532 15470  8284 27390  2234 2011   16 63253       6030
## 2 37 Philippines   344   82  3939  4272  2949 41987  600 1011 55184       2040
## 3 46 Switzerland  2342  102  1528 45612  1018   902  722   51 52277       1746
## 4 21       India    22  489  6684  7369   895  1107   67  340 16973      19488
## 5 49        Togo     0    0     0 11233   759     0    2    0 11994          5
##   volunteers
## 1     703165
## 2     109800
## 3      52380
## 4     122844
## 5        460
```

Both ways show that 5 countries that have the worst plastic problem are:

1. Nigeria
2. Philippines
3. Switzerland
4. India
5. Togo

2. Make a plot showing the distribution of volunteers across countries

```
## Warning: Removed 1 rows containing missing values (`position_stack()`).
```



3. Notice that there is a lot of variation across countries in the number of volunteers involved in trash pickup. What problem might that cause for the interpretation of your answer to question 1?

```
summary(tidytue$volunteers)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
##     12.0    302.5   1875.0  39330.3  16803.0 703165.0        1
```

Answer: As shown from the summary of Tidy Tuesday's Volunteers above, it is seen that the mean of 39,330 cannot represent the number of volunteers involved across countries because the range is to wide with minimum value of 12 and maximum value of 703,165.

4. Add a column to the data frame creating a variable that should be more closely related to the presence of plastic pollution in the country

To better illustrate the the condition of plastic pollution in each countries, I add the variable of number of plastic trash picked up per picking-up events under the column of 'ratio'.

5. What 5 countries have the worst plastic pollution, as measured by this new variable?

```
##     X      country empty hdpe ldpe     o  pet  pp  ps pvc total num_events
## 49 49         Togo     0    0    0 11233  759   0   2   0 11994          5
## 8   8 Burkina Faso     0  267   20  3623 4664  68  90  42  8774         15
```

```
## 17 17      Germany    10    2  210  4790  102  95   7   0  5216         19
## 15 15  El Salvador     0    2  180    53  185 318 182   0   920          5
## 18 18        Ghana     0  220  291     0 8272   0   0   0  8783         52
## 43 43 South Africa     0   79    0     8  770   0   0   0   857          6
##    volunteers     ratio
## 49        460 2398.8000
## 8        2445  584.9333
## 17        456  274.5263
## 15         50  184.0000
## 18       8060  168.9038
## 43        120  142.8333
```

Answer:

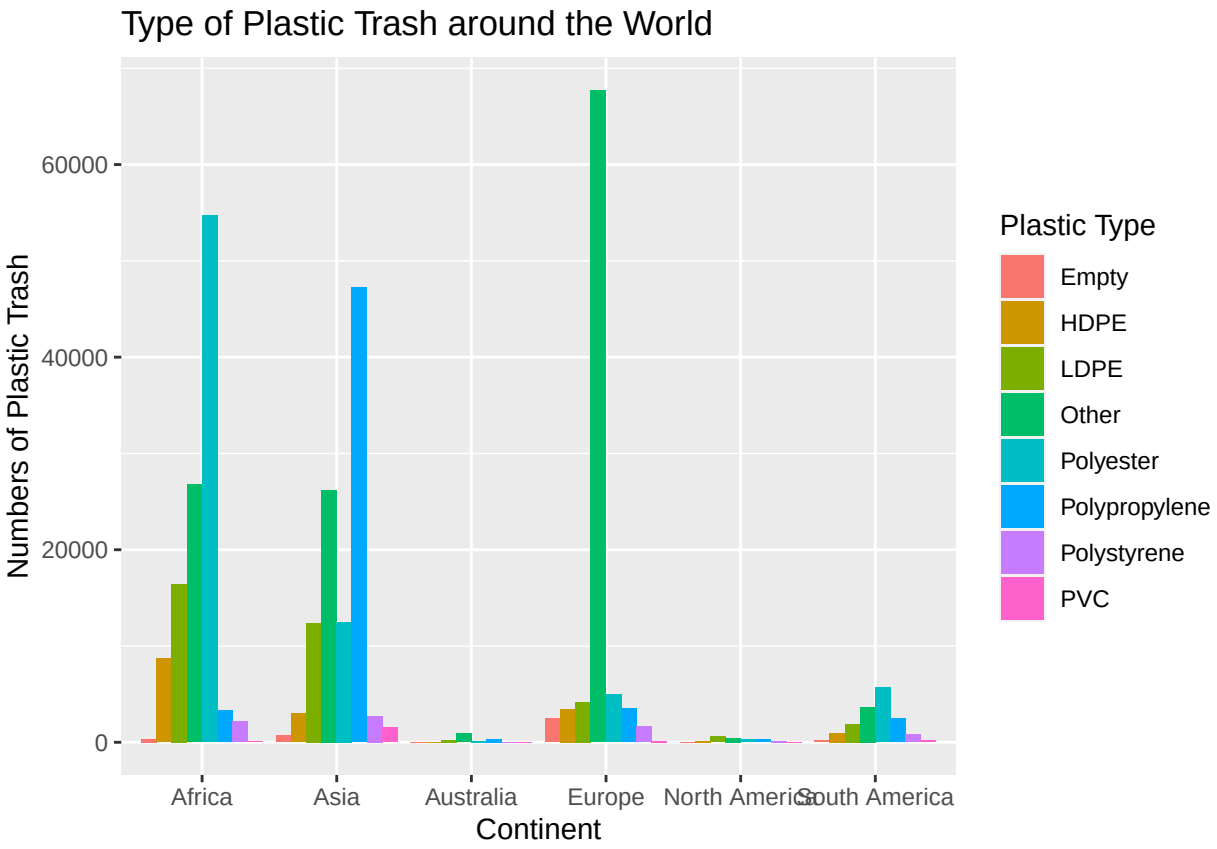
After including the new variable of average number of trash picked up per events, 5 countries that have the worst plastic pollution are:

1. Togo
2. Burkina Faso
3. Germany
4. El Salvador
5. Ghana

Now we will make a plot of the variation in the types of trash and how it differs around the world. Read in the continenttypes.csv data frame. This gives the breakdown of the different types of plastic collected on each continent in 2020 and the total number of pick up events.

```
setwd("C:\\Users\\Benny Panjaitan\\Documents\\GitHub\\esp106-Naomi\\W5 Lab\\ESP106_week5_data\\")
contype <- read.csv('continenttypes.csv')

library(ggplot2)
ggplot(contype, aes(fill=plastic_type, y=total, x=continent)) +
    geom_bar(position="dodge", stat="identity") +
    labs(title="Type of Plastic Trash around the World", y = "Numbers of Plastic Trash", x = "Continent
    scale_fill_discrete(name = "Plastic Type", labels = c("Empty", "HDPE", "LDPE", "Other", "Polyester
```

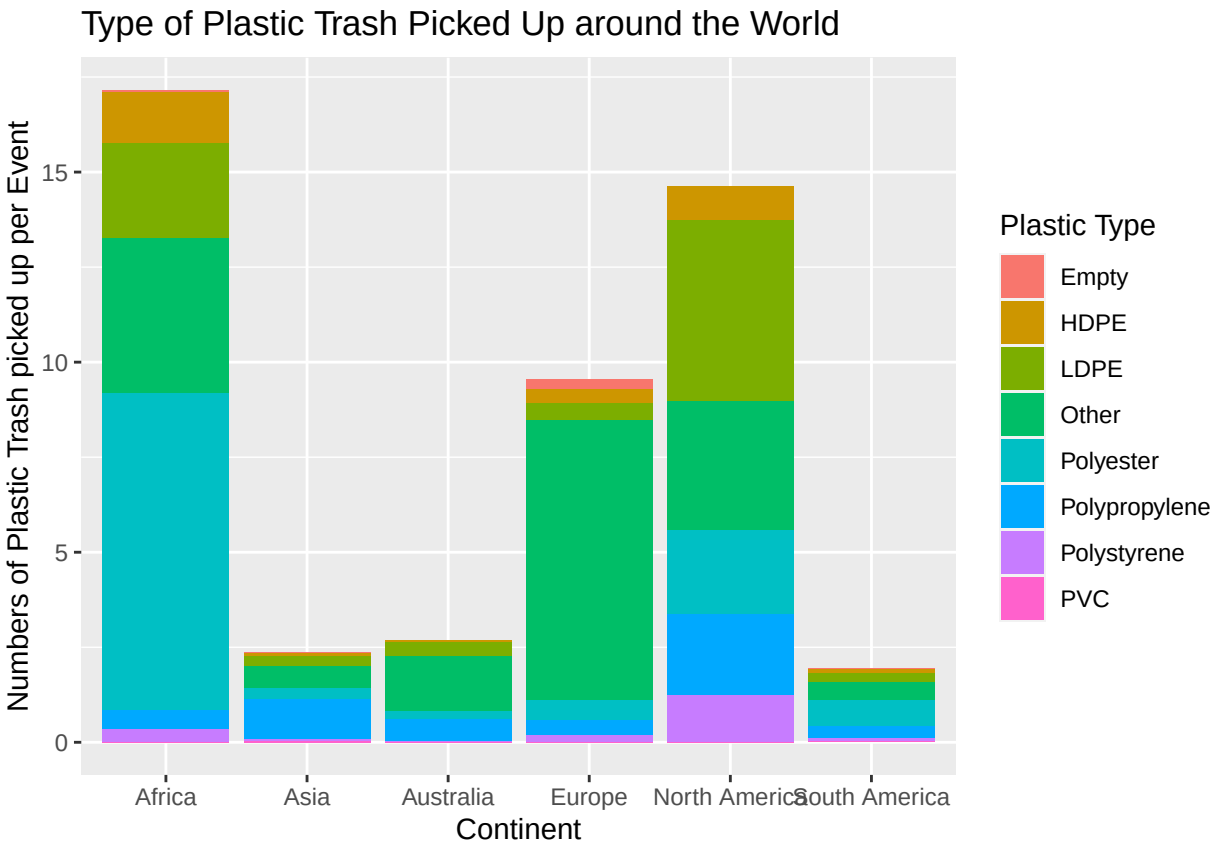


6. Add a column to this data frame with a variable that captures the existence of different types of plastic trash, controlling for the intensity of the pick-up effort in different continent

I add new column of 'te' which shows how many plastic trash (for each type) picked up per event.

7. Make a plot using ggplot showing both the total amount and distribution of types of plastic picked up in each continent in the average pick-up event.

Hint: Check out options in the R graph gallery



8. Try uploading your R markdown file and plot to your Git Hub repository. Don't put your knitted HTML file in Github - these are large files that are not designed to be stored on Github