Univariate Data Analysis Case I: Amazon Host Case

We will try and answer the following questions:

1. What are the major hosts contributing towards the count?
2. How valuable are these hosts?
3. Which one deserves the highest pay per click and which one the lowest?
4. How can you quantify the importance?
5. How much is the size of the effect?

```
> names(amazon)
[1] "Host"       "Count"      "Proportion"
> str(amazon)
'data.frame':	22 obs. of  3 variables:
 $ Host      : Factor w/ 22 levels "24hour-mall.com",..: 21 14 22 11 20 2 13 4 5 7 ...
 $ Count     : int  89919 7258 6078 4381 4283 1639 1573 1289 1285 1166 ...
 $ Proportion: num  0.4758 0.0384 0.0322 0.0232 0.0227 …
> tail(amazon)
              Host Count Proportion
17    netscape.com   544 0.00287837
18    dealtime.com   543 0.00287308
19         att.net   533 0.00282017
20   postcards.org   532 0.00281487
21 24hour-mall.com   503 0.00266143
22           Other 63229 0.33455205
```

The function factor is used to encode a vector as a factor. If the argument ordered = TRUE, the factor levels are assumed to be ordered. We can combine the last 11 sources also into other hosts as they are not varied much. We need to first choose the required rows to remain and then concatenate them into other hosts. Then we need to fix the count variable, by aggregating the remaining and then the same for proportion. Finally let us order the data set by count.

```
> sorted
                  Host Count  Proportion
11            imdb.com   886 0.004687930
10 daily-blessings.com  1166 0.006169443
9          bmezine.com  1285 0.006799086
8           atwola.com  1289 0.006820250
7             iwon.com  1573 0.008322927
6              aol.com  1639 0.008672141
5     recipesource.com  4283 0.022661855
4           google.com  4381 0.023180385
3            yahoo.com  6078 0.032159411
2              msn.com  7258 0.038402929
12         other hosts 69239 0.366351669
1     Typed amazon.com 89919 0.475771974
```

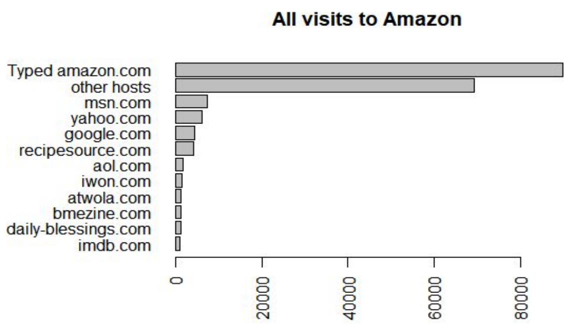
1. What are the major hosts contributing towards the count?

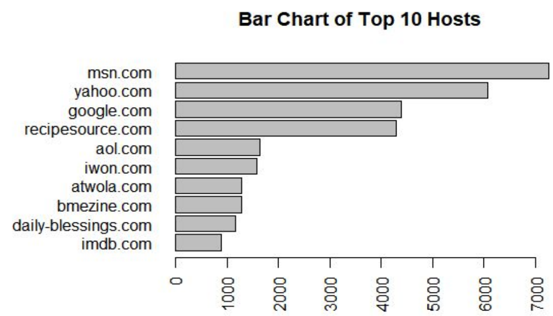
Now let us plot a bar graph to see where we stand:

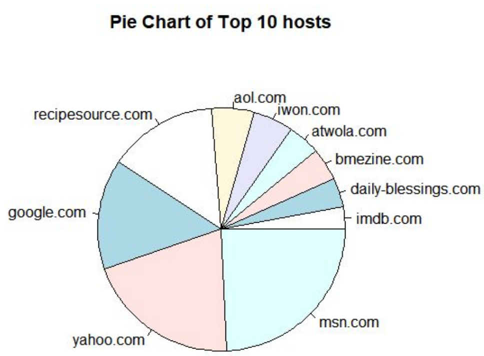
```
> barplot(sorted$Count, names.arg = sorted$Host, horiz = TRUE, las = 2, main = "All visits to Amazon")
```

From this plot you can't really tell much as other plots are dominant. Hence we try and ignore 'other hosts'.

```
> barplot(sorted$Count[1:10], names.arg = sorted$Host[1:10], horiz = TRUE, las=2, main='Bar Chart of Top 10 Hosts')
> pie(sorted$Count[1:10], sorted$Host[1:10], main='Pie Chart of Top 10 hosts')
```







The above charts explain the major hosts. Since the dataset consists of categorical variables let us look into mosaic plot. We will be using the **library(gmodels)** library for the same. For this we will use a different dataset specific to understanding how host purchases differ.

```
> names(host_purchase)
[1] "host"     "purchase"
> str(host_purchase)
'data.frame':	17619 obs. of  2 variables:
 $ host    : Factor w/ 3 levels "msn.com","recipesource.com",..: 1 1 1 3 3...
 $ purchase: Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 …
> summary(host_purchase)
               host       purchase
 msn.com         :7258   Yes:  516
 recipesource.com:4283   No :17103
 yahoo.com       :6078
```

Now let us create a contingency table to feed into the mosaic plot.

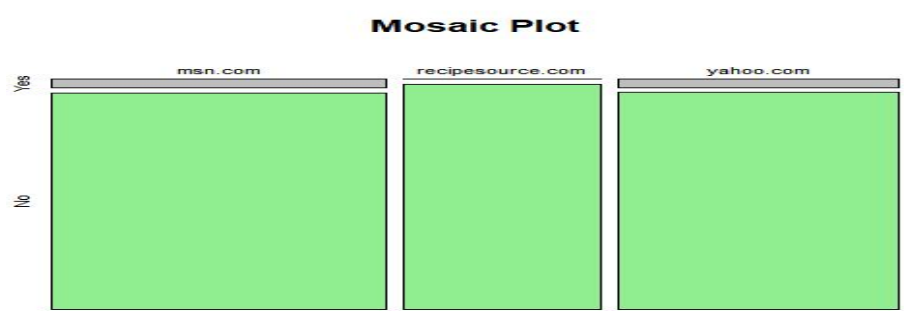
2. How valuable are these hosts?
3. Which one deserves the highest pay per click and which one the lowest?

```
> host_table = CrossTable(host_purchase$purchase, host_purchase$host,
prop.chisq = F, prop.c = F, prop.t = F)
   Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
|-------------------------|


Total Observations in Table:  17619


                       | host_purchase$host
host_purchase$purchase |           msn.com | recipesource.com |         yahoo.com |         Row Total |
-----------------------|-------------------|------------------|-------------------|-------------------|
                   Yes |               285 |                1 |               230 |               516 |
                       |             0.552 |            0.002 |             0.446 |             0.029 |
-----------------------|-------------------|------------------|-------------------|-------------------|
                    No |              6973 |             4282 |              5848 |             17103 |
                       |             0.408 |            0.250 |             0.342 |             0.971 |
-----------------------|-------------------|------------------|-------------------|-------------------|
          Column Total |              7258 |             4283 |              6078 |             17619 |
-----------------------|-------------------|------------------|-------------------|-------------------|
```

```
> mosaicplot(host_table$t, color = c("grey","lightgreen"), xlab = "Host", ylab
= "Purchase", main = "Mosaic Plot")
```



4. How can you quantify the importance?

We need to quantify this impact to see if it is real or by chance by conducting hypothesis testing:

$$H_0 : \textit{Host does not affect Purchase}$$
$$H_0 : \textit{Host affects Purchase}$$

```
> chisq.test(host_purchase$host, host_purchase$purchase)
      Pearson's Chi-squared test

data:  host_purchase$host and host_purchase$purchase
```

```
X-squared = 168.24, df = 2, p-value < 2.2e-16
```

Since the p-value is lower than 0.5% significance level we need to reject the null hypothesis and conclude that the host does indeed affect the Purchase. Now let us further explore as to which host contributes more.

```
> chisq.test(msnrecipe$host, msnrecipe$purchase)
      Pearson's Chi-squared test with Yates'
      continuity correction

data:  msnrecipe$host and msnrecipe$purchase
X-squared = 168.2, df = 1, p-value < 2.2e-16
```

Since the p-value is lower than 0.5% significance level we need to reject the null hypothesis and conclude that the host does indeed affect the Purchase. Now let us look into msn and yahoo.

```
> chisq.test(msnyahoo$host, msnyahoo$purchase)
      Pearson's Chi-squared test with Yates'
      continuity correction

data:  msnyahoo$host and msnyahoo$purchase
X-squared = 0.14472, df = 1, p-value = 0.7036
```

In this case, we need to accept the null hypothesis as the p-value is above the significance level of 0.5%.

Hence these tests reveal that:

1. Msn and Recipesource have different impacts on purchase.
2. There is no statistically significant difference between msn and yahoo.

Now let us look into the size of effect by looking at their confidence intervals. Let us look into one sample t-test to test whether a population mean is significantly different from some hypothesized value.

```
> t.test(as.numeric(host_msn$purchase=="Yes"))
      One Sample t-test

data:  as.numeric(host_msn$purchase == "Yes")
t = 17.222, df = 7257, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.03479753 0.04373650
sample estimates:
 mean of x
0.03926702
> t.test(as.numeric(host_recipesource$purchase=="Yes"))
      One Sample t-test

data:  as.numeric(host_recipesource$purchase == "Yes")
t = 1, df = 4282, p-value = 0.3174
```

```
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.0002242629  0.0006912253
sample estimates:
   mean of x
0.0002334812
> t.test(as.numeric(host_yahoo$purchase=="Yes"))
      One Sample t-test

data:  as.numeric(host_yahoo$purchase == "Yes")
t = 15.46, df = 6077, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.03304299 0.04263980
sample estimates:
mean of x
0.0378414
```

5. How much is the size of the effect?

We cannot use p-value for size. We need to use Confidence Intervals.

```
> t.test(msnrecipe.purchase ~ msnrecipe$host)
      Welch Two Sample t-test

data:  msnrecipe.purchase by msnrecipe$host
t = 17.031, df = 7408.6, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.03454069 0.04352638
sample estimates:
         mean in group msn.com
                  0.0392670157
mean in group recipesource.com
                  0.0002334812
```

Hence the 95% CI for Prob(msn) - Prob(recipesource)

```
> t.test(msnyahoo.purchase ~ msnyahoo$host)
      Welch Two Sample t-test
data:  msnyahoo.purchase by msnyahoo$host
t = 0.42618, df = 13001, p-value = 0.67
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.005131295  0.007982536
sample estimates:
  mean in group msn.com mean in group yahoo.com
             0.03926702              0.03784140
```