Assignment 1
Subject : BDM
Student number: 2982006

1.

In [8]:
```
#Question 1: Show number of downloads for package ggplot2 for each day (the 1st and 2nd of March).
package_download_count = downloads_RDD.filter(lambda x: "ggplot2" in x)
package_download_count = package_download_count.map(lambda x: (x[0], 1))
package_download_count = package_download_count.reduceByKey(lambda a,b: a+b)
package_download_count.collect()
```

Out[8]:
```
[('2019-03-01', 25385), ('2019-03-02', 13344)]
```

2.

In [9]:
```
#Question 2: Highest number of downloads by a country - on both days.
package_download_count = downloads_RDD.map(lambda x: (x[8], 1))
package_download_count = package_download_count.reduceByKey(lambda a,b: a+b)
package_download_count = package_download_count.sortBy(lambda a: a[1], ascending= False)
package_download_count.collect()
```

Out[9]:
```
[('US', 1776597),
 ('NA', 352318),
 ('CA', 324601),
 ('DE', 152275),
 ('GB', 146479),
 ('IN', 127671),
 ('HK', 111794),
 ('JP', 103439),
 ('CN', 101058),
 ('FR', 83768),
 ('ES', 81804),
 ('NL', 67038),
 ('AU', 63396),
 ('CO', 56621),
 ('CH', 54046),
 ('MX', 46061),
 ('IT', 44640),
 ('BR', 39300),
 ('KR', 38021),
 ('PL', 35978),
```

3.

In [3]:
```
#Question 3:Show top 10 largest sized packages.
x = downloads_RDD.map(lambda x: x[2])
y = downloads_RDD.map(lambda x: x[6])
z = x.zip(y)
z.distinct()
z = z.sortBy(lambda a:a[0], ascending=False)
z.take(11)
```

Out[3]:
```
[('999989', '"BClustLonG"'),
 ('999983', '"BClustLonG"'),
 ('999983', '"BClustLonG"'),
 ('999983', '"BClustLonG"'),
 ('999983', '"BClustLonG"'),
 ('999983', '"BClustLonG"'),
 ('999983', '"BClustLonG"'),
 ('999940', '"data.table"'),
 ('999929', '"RcppEigen"'),
 ('999929', '"RcppArmadillo"'),
 ('999929', '"scales"')]
```

4.

In [11]:
```
#Question 4:What were the top 10 most popular packages on 2nd of March?
package_download_count = downloads_RDD.filter(lambda x: "2019-03-02" in x)
package_download_count = package_download_count.map(lambda x: (x[6],1))
package_download_count = package_download_count.reduceByKey(lambda a,b: a+b)
package_download_count = package_download_count.sortBy(lambda a:a[1], ascending=False)
package_download_count.take(10)
```

Out[11]:
```
[('rlang', 19600),
 ('Rcpp', 18384),
 ('tibble', 16290),
 ('pillar', 14957),
 ('yaml', 14630),
 ('openssl', 14407),
 ('stringr', 14112),
 ('R6', 13965),
 ('fansi', 13796),
 ('cli', 13678)]
```

5.

In [14]:
```
#Question 5:What OS is used for downloading the most popular package? - on both days.
package = downloads_RDD.map(lambda x: (x[6],1))
package = package.reduceByKey(lambda a,b: a+b)
package = package.sortBy(lambda a:a[1], ascending=False)
z = package.take(1)
os = downloads_RDD.filter(lambda x: (z[0][0]) in x)
os = os.map(lambda x: (x[5],1))
os = os.reduceByKey(lambda a,b: a+b)
os = os.sortBy(lambda a: a[1], ascending=False)
os.take(1)
```

Out[14]:
```
[('mingw32', 26172)]
```

6.

In [14]:
```
#Question 6: What is the most popular package in Ireland?
package_download_count = downloads_RDD.filter(lambda x: "IE" in x)
package_download_count = package_download_count.map(lambda x: (x[6],1))
package_download_count = package_download_count.reduceByKey(lambda a,b: a+b)
package_download_count = package_download_count.sortBy(lambda a:a[1], ascending=False)
package_download_count.take(1)
```

Out[14]:
```
[('ggplot2', 228)]
```

7.

In [18]:
```
#Question 7:What is the highest number of downloads by a single machine? What OS it has?
x = downloads_RDD.map(lambda x: (x[9],1))
x = x.reduceByKey(lambda a,b: a+b)
x = x.sortBy(lambda a:a[1], ascending=False)
z = x.take(1)
y = downloads_RDD.filter(lambda x: z[0][0] in x)
y = y.map(lambda x: (x[5],1))
y = y.reduceByKey(lambda a,b: a+b)
y = y.sortBy(lambda a:a[1], ascending=False)
os = y.take(1)
print(z)
print(os)
```

```
[('8', 228763)]
[('mingw32', 108025)]
```

8.

In [20]:
```
#Question 8:What OS is most popular among the R programmers?
os = downloads_RDD.map(lambda x: (x[5],1))
os = os.reduceByKey(lambda a,b: a+b)
os = os.sortBy(lambda a: a[1], ascending=False)
os.take(1)
```

Out[20]:
```
[('mingw32', 2000498)]
```

9.

In [18]:
```
#Question 9:How many R users still use 32 bit machines?
y = downloads_RDD.filter(lambda x: "i386" in x)
y = y.map(lambda x: (x[4],1))
y = y.reduceByKey(lambda a,b: a+b)
y.collect()
```

Out[18]:
```
[('i386', 153963)]
```

10.

In [21]:
```
#Question 10: List total number of incomplete records - lines which have missing values.
missing = downloads_RDD.filter(lambda x: "NA" in x)
missing.count()
```

Out[21]:
```
468436
```