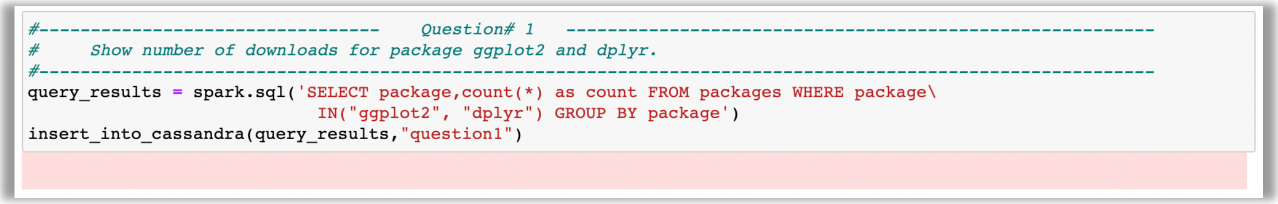
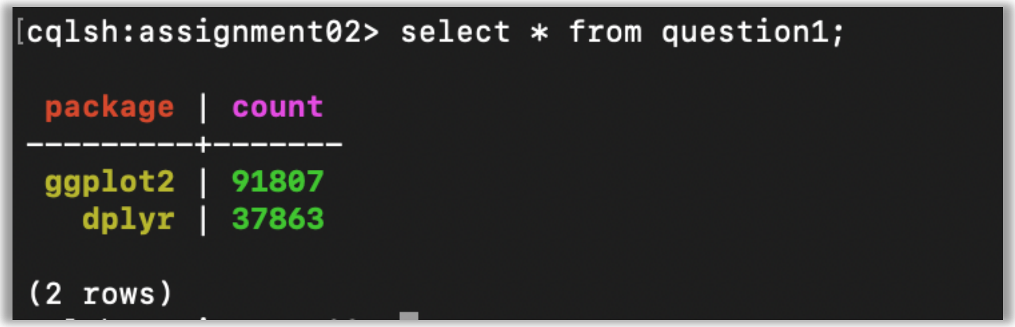
Question#1:

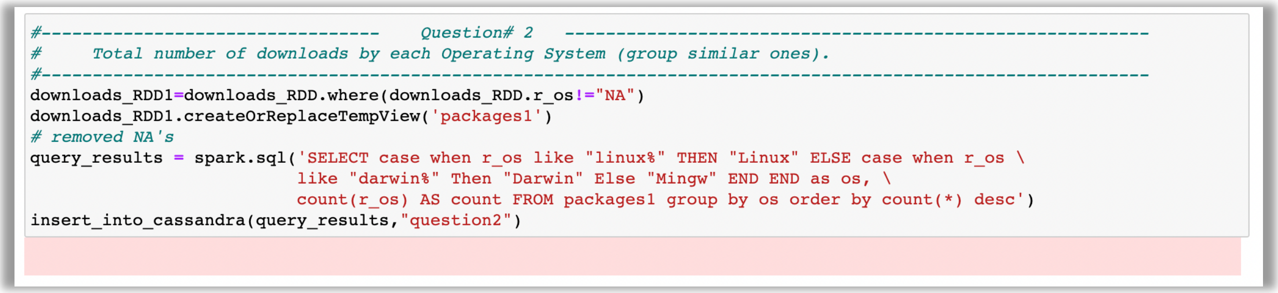


```
#----------------------------------    Question# 1   ----------------------------------------------------------
#     Show number of downloads for package ggplot2 and dplyr.
#-----------------------------------------------------------------------------------------------------------
query_results = spark.sql('SELECT package,count(*) as count FROM packages WHERE package\
                          IN("ggplot2", "dplyr") GROUP BY package')
insert_into_cassandra(query_results,"question1")
```

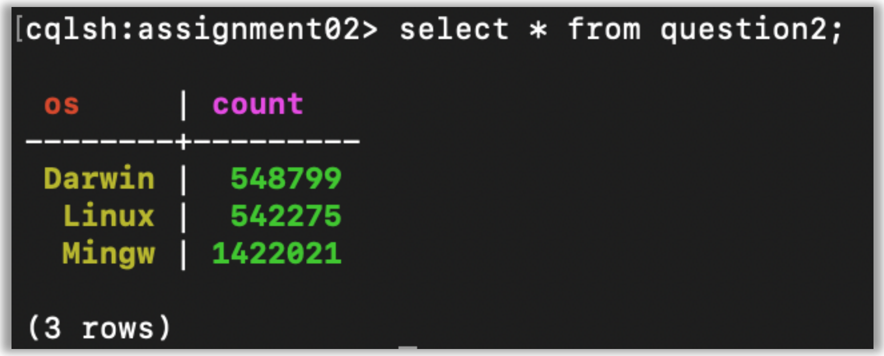


```
cqlsh:assignment02> select * from question1;

 package | count
---------+-------
 ggplot2 | 91807
   dplyr | 37863

(2 rows)
```

Question#2:

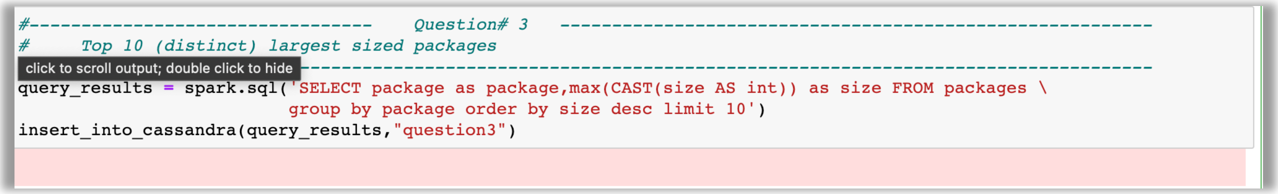


```
#----------------------------------    Question# 2   ----------------------------------------------------------
#     Total number of downloads by each Operating System (group similar ones).
#-----------------------------------------------------------------------------------------------------------
downloads_RDD1=downloads_RDD.where(downloads_RDD.r_os!="NA")
downloads_RDD1.createOrReplaceTempView('packages1')
# removed NA's
query_results = spark.sql('SELECT case when r_os like "linux%" THEN "Linux" ELSE case when r_os \
                          like "darwin%" Then "Darwin" Else "Mingw" END END as os, \
                          count(r_os) AS count FROM packages1 group by os order by count(*) desc')
insert_into_cassandra(query_results,"question2")
```

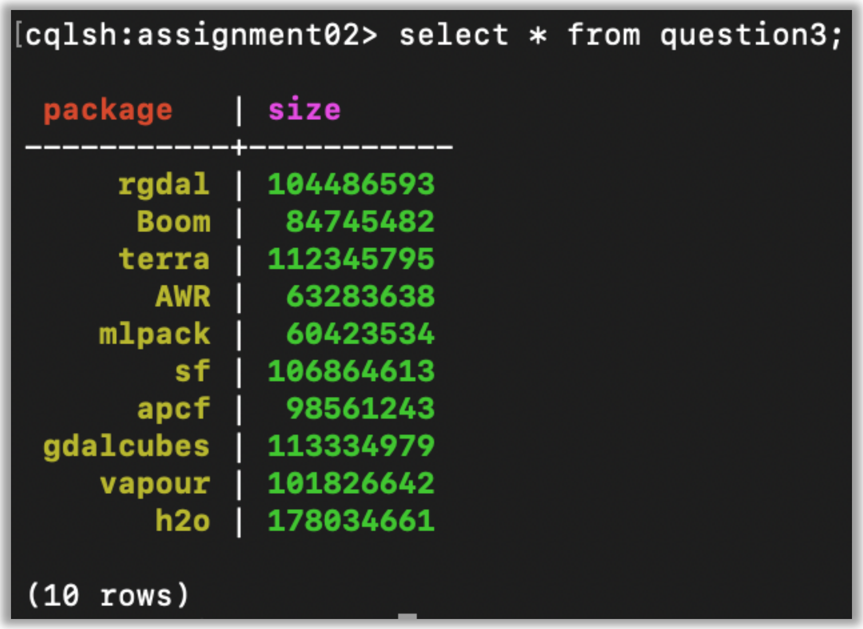


```
cqlsh:assignment02> select * from question2;

 os     | count
--------+---------
 Darwin |  548799
  Linux |  542275
  Mingw | 1422021

(3 rows)
```

Question#3:

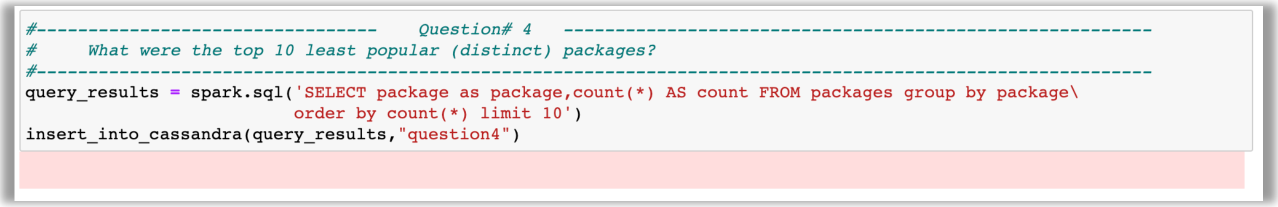


```
#----------------------------------    Question# 3   ---------------------------------------------------------
#     Top 10 (distinct) largest sized packages
click to scroll output; double click to hide ------------------------------------------------------------------------
query_results = spark.sql('SELECT package as package,max(CAST(size AS int)) as size FROM packages \
                          group by package order by size desc limit 10')
insert_into_cassandra(query_results,"question3")
```

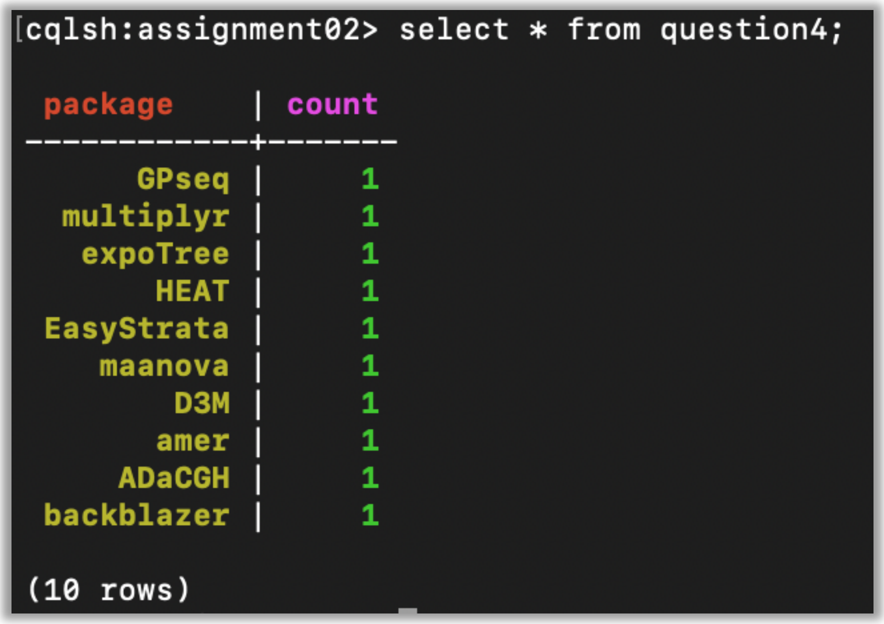


```
cqlsh:assignment02> select * from question3;

 package     | size
-------------+-----------
       rgdal | 104486593
        Boom |  84745482
       terra | 112345795
         AWR |  63283638
      mlpack |  60423534
          sf | 106864613
        apcf |  98561243
   gdalcubes | 113334979
      vapour | 101826642
         h2o | 178034661

(10 rows)
```

Question#4:

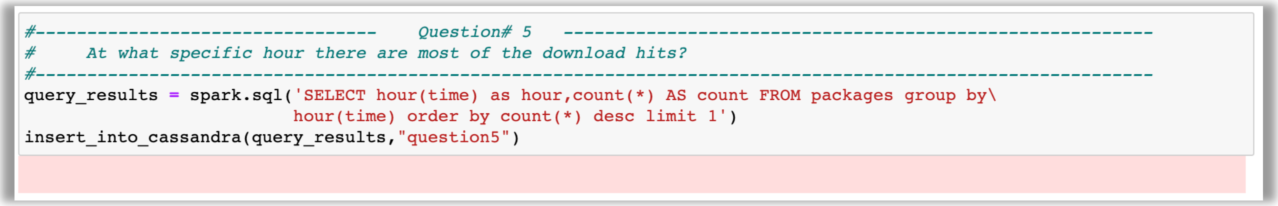


```
#----------------------------------    Question# 4   ---------------------------------------------------------
#     What were the top 10 least popular (distinct) packages?
#----------------------------------------------------------------------------------------------------------
query_results = spark.sql('SELECT package as package,count(*) AS count FROM packages group by package\
                          order by count(*) limit 10')
insert_into_cassandra(query_results,"question4")
```

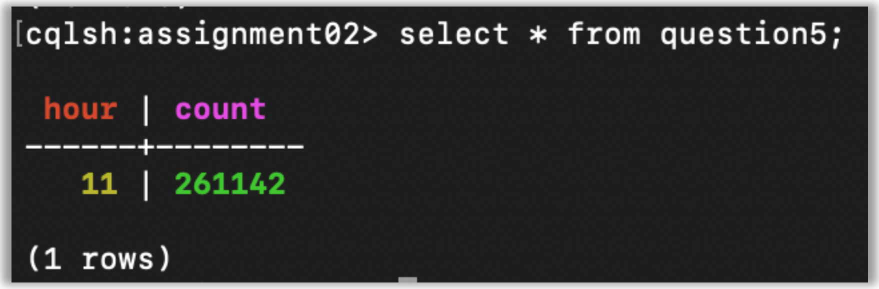


```
cqlsh:assignment02> select * from question4;

 package      | count
--------------+-------
        GPseq |     1
    multiplyr |     1
     expoTree |     1
         HEAT |     1
  EasyStrata |     1
      maanova |     1
          D3M |     1
         amer |     1
       ADaCGH |     1
   backblazer |     1

(10 rows)
```

Question#5:

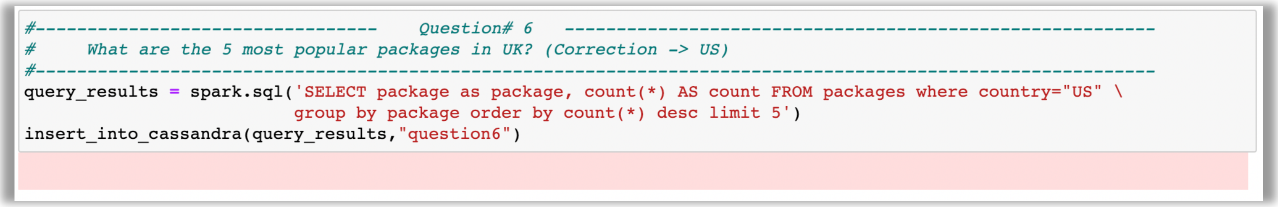


```
#----------------------------------    Question# 5   ----------------------------------------------------------
#     At what specific hour there are most of the download hits?
#-------------------------------------------------------------------------------------------------------------
query_results = spark.sql('SELECT hour(time) as hour,count(*) AS count FROM packages group by\
                          hour(time) order by count(*) desc limit 1')
insert_into_cassandra(query_results,"question5")
```

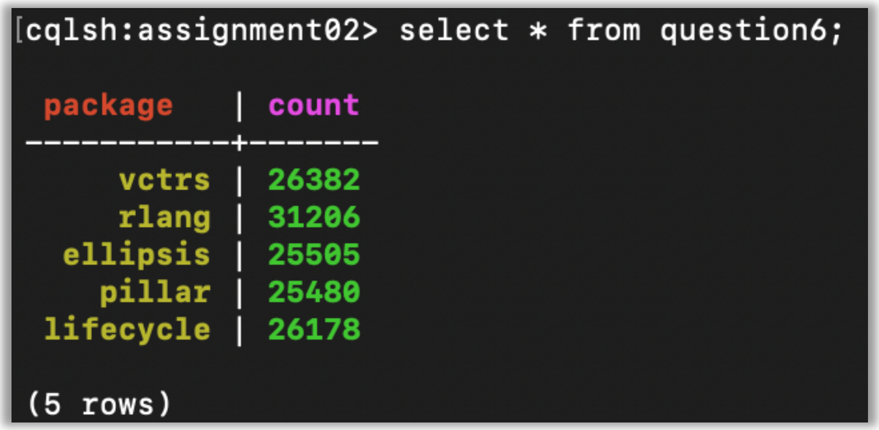


```
cqlsh:assignment02> select * from question5;

 hour | count
------+--------
   11 | 261142

(1 rows)
```

Question#6:

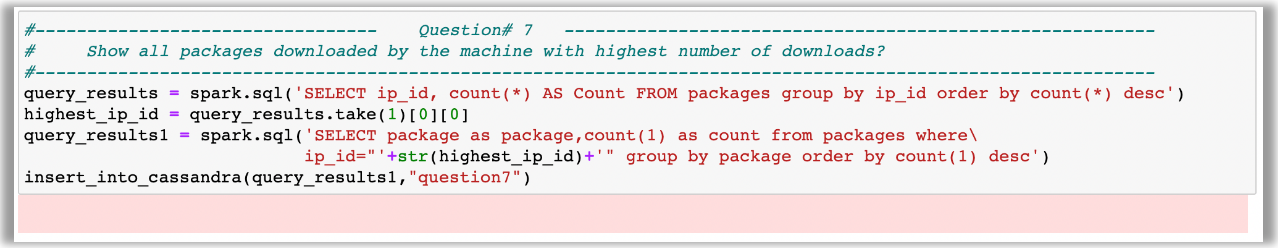


```
#----------------------------------    Question# 6   ----------------------------------------------------------
#     What are the 5 most popular packages in UK? (Correction -> US)
#-------------------------------------------------------------------------------------------------------------
query_results = spark.sql('SELECT package as package, count(*) AS count FROM packages where country="US" \
                          group by package order by count(*) desc limit 5')
insert_into_cassandra(query_results,"question6")
```

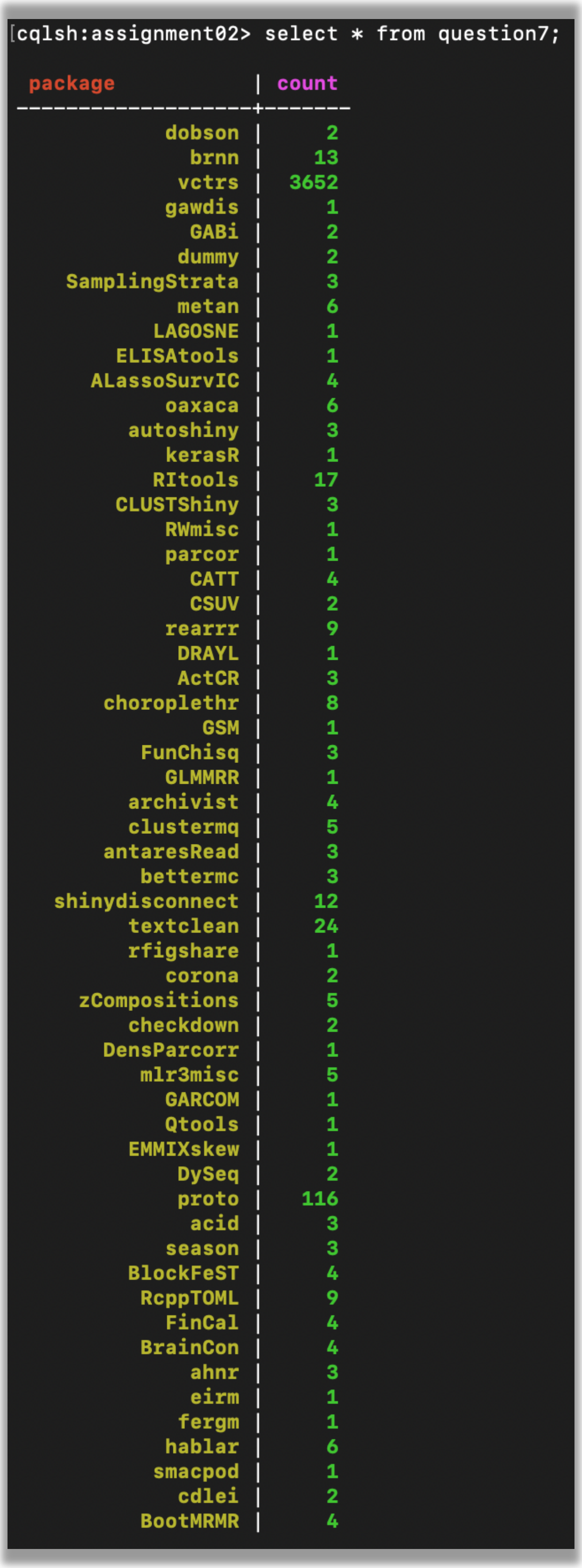


```
cqlsh:assignment02> select * from question6;

 package   | count
-----------+-------
     vctrs | 26382
     rlang | 31206
  ellipsis | 25505
    pillar | 25480
 lifecycle | 26178

(5 rows)
```

Question#7:

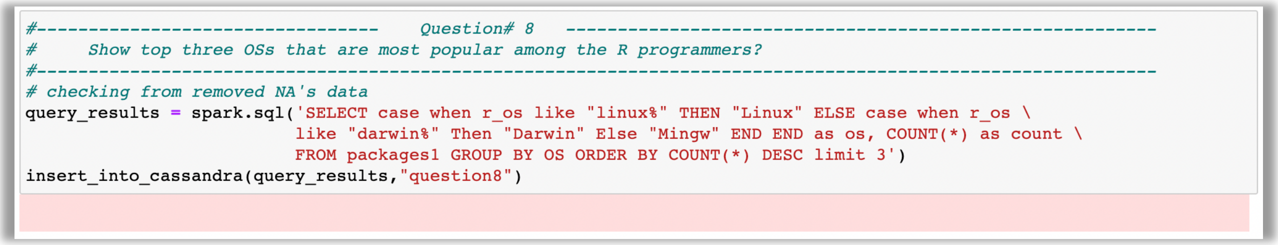


```
#----------------------------------    Question# 7   ----------------------------------------------------------
#     Show all packages downloaded by the machine with highest number of downloads?
#-------------------------------------------------------------------------------------------------------------
query_results = spark.sql('SELECT ip_id, count(*) AS Count FROM packages group by ip_id order by count(*) desc')
highest_ip_id = query_results.take(1)[0][0]
query_results1 = spark.sql('SELECT package as package,count(1) as count from packages where\
                           ip_id="'+str(highest_ip_id)+'" group by package order by count(1) desc')
insert_into_cassandra(query_results1,"question7")
```

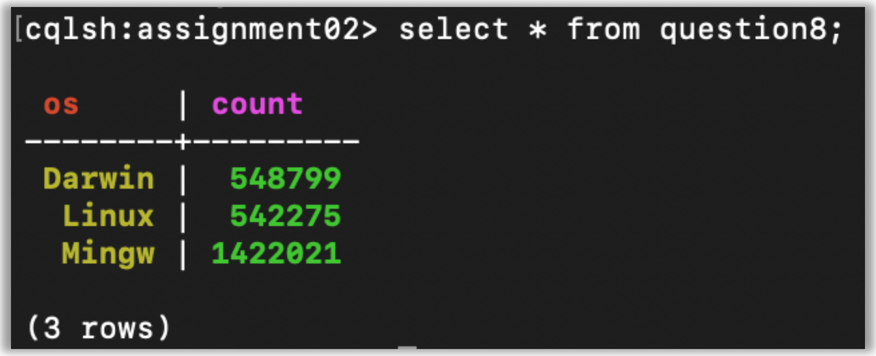


```
[cqlsh:assignment02> select * from question7;

 package           | count
-------------------+-------
            dobson |     2
              brnn |    13
             vctrs |  3652
            gawdis |     1
              GABi |     2
             dummy |     2
    SamplingStrata |     3
             metan |     6
           LAGOSNE |     1
        ELISAtools |     1
      ALassoSurvIC |     4
            oaxaca |     6
         autoshiny |     3
            kerasR |     1
           RItools |    17
        CLUSTShiny |     3
            RWmisc |     1
            parcor |     1
              CATT |     4
              CSUV |     2
            rearrr |     9
             DRAYL |     1
             ActCR |     3
       choroplethr |     8
               GSM |     1
          FunChisq |     3
            GLMMRR |     1
         archivist |     4
         clustermq |     5
       antaresRead |     3
          bettermc |     3
   shinydisconnect |    12
         textclean |    24
         rfigshare |     1
            corona |     2
     zCompositions |     5
         checkdown |     2
       DensParcorr |     1
          mlr3misc |     5
            GARCOM |     1
            Qtools |     1
         EMMIXskew |     1
             DySeq |     2
             proto |   116
              acid |     3
            season |     3
         BlockFeST |     4
          RcppTOML |     9
            FinCal |     4
          BrainCon |     4
              ahnr |     3
              eirm |     1
             fergm |     1
            hablar |     6
           smacpod |     1
             cdlei |     2
          BootMRMR |     4
```

Question#8:

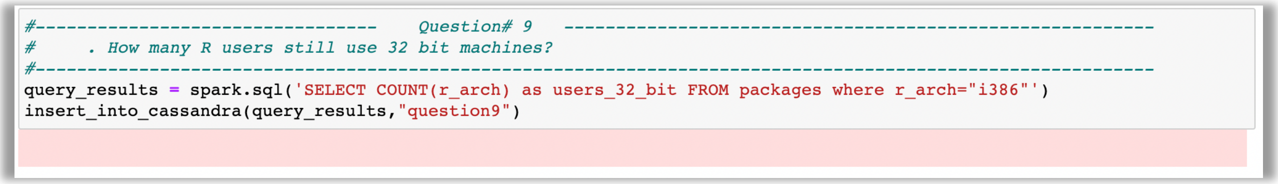


```
#---------------------------------    Question# 8    ---------------------------------------------------------
#     Show top three OSs that are most popular among the R programmers?
#-------------------------------------------------------------------------------------------------------------
# checking from removed NA's data
query_results = spark.sql('SELECT case when r_os like "linux%" THEN "Linux" ELSE case when r_os \
                          like "darwin%" Then "Darwin" Else "Mingw" END END as os, COUNT(*) as count \
                          FROM packages1 GROUP BY OS ORDER BY COUNT(*) DESC limit 3')
insert_into_cassandra(query_results,"question8")
```

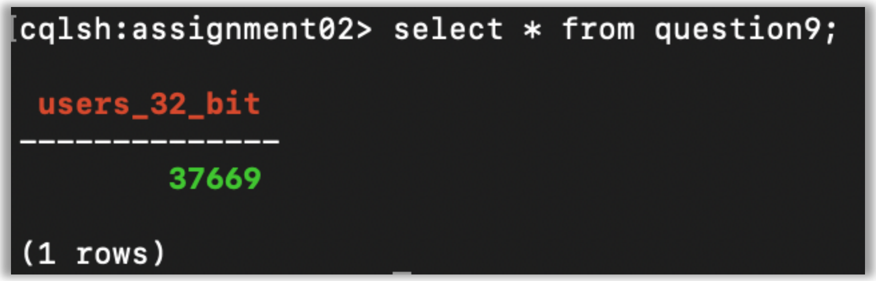


```
cqlsh:assignment02> select * from question8;

 os       | count
----------+---------
 Darwin |  548799
  Linux |  542275
  Mingw | 1422021

(3 rows)
```

Question#9:

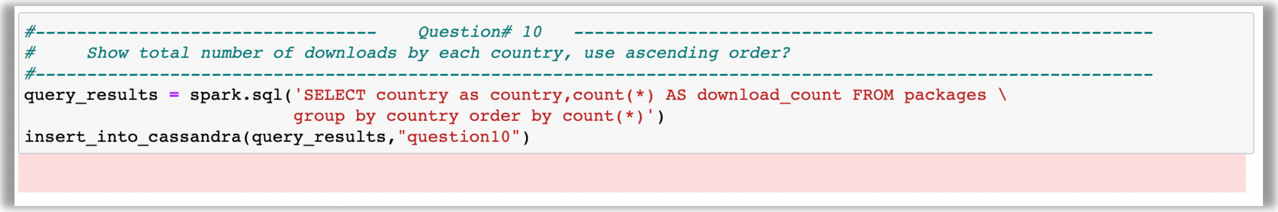


```
#---------------------------------    Question# 9    ---------------------------------------------------------
#     . How many R users still use 32 bit machines?
#-------------------------------------------------------------------------------------------------------------
query_results = spark.sql('SELECT COUNT(r_arch) as users_32_bit FROM packages where r_arch="i386"')
insert_into_cassandra(query_results,"question9")
```

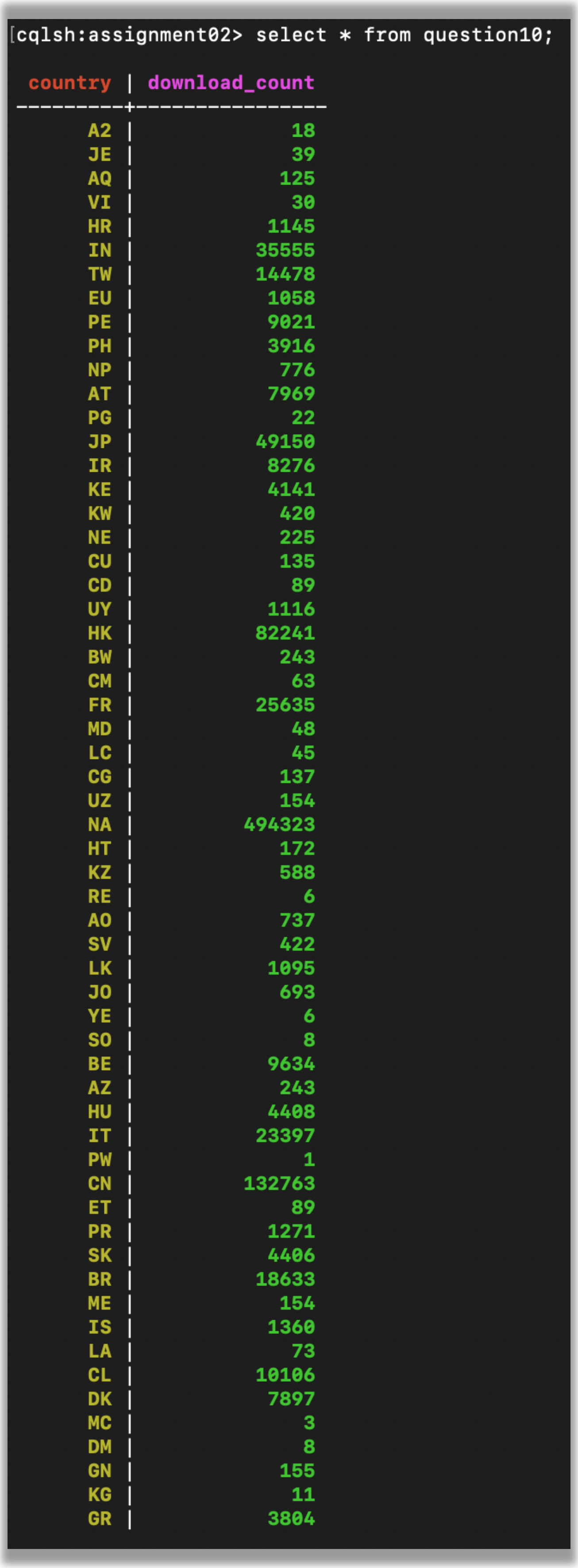


```
cqlsh:assignment02> select * from question9;

 users_32_bit
--------------
        37669

(1 rows)
```

Question#10:



```
#---------------------------------    Question# 10    --------------------------------------------------------
#     Show total number of downloads by each country, use ascending order?
#-------------------------------------------------------------------------------------------------------------
query_results = spark.sql('SELECT country as country,count(*) AS download_count FROM packages \
                          group by country order by count(*)')
insert_into_cassandra(query_results,"question10")
```



```
[cqlsh:assignment02> select * from question10;

 country | download_count
---------+----------------
      A2 |             18
      JE |             39
      AQ |            125
      VI |             30
      HR |           1145
      IN |          35555
      TW |          14478
      EU |           1058
      PE |           9021
      PH |           3916
      NP |            776
      AT |           7969
      PG |             22
      JP |          49150
      IR |           8276
      KE |           4141
      KW |            420
      NE |            225
      CU |            135
      CD |             89
      UY |           1116
      HK |          82241
      BW |            243
      CM |             63
      FR |          25635
      MD |             48
      LC |             45
      CG |            137
      UZ |            154
      NA |         494323
      HT |            172
      KZ |            588
      RE |              6
      AO |            737
      SV |            422
      LK |           1095
      JO |            693
      YE |              6
      SO |              8
      BE |           9634
      AZ |            243
      HU |           4408
      IT |          23397
      PW |              1
      CN |         132763
      ET |             89
      PR |           1271
      SK |           4406
      BR |          18633
      ME |            154
      IS |           1360
      LA |             73
      CL |          10106
      DK |           7897
      MC |              3
      DM |              8
      GN |            155
      KG |             11
      GR |           3804
```