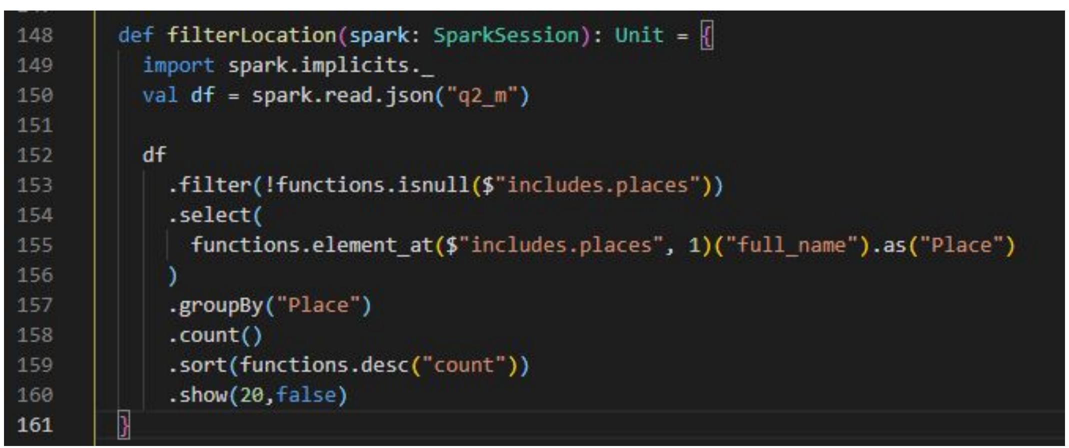
Project 2

Geography and Sports Terminology on Twitter

Group 1 -
Jacob Boullion, Allie Burkeen, Delaney Lekien

What is the most currently referenced country? Does that change based on the origin country of the tweet? How much conversation about countries on twitter is done by residents vs. foreigners?

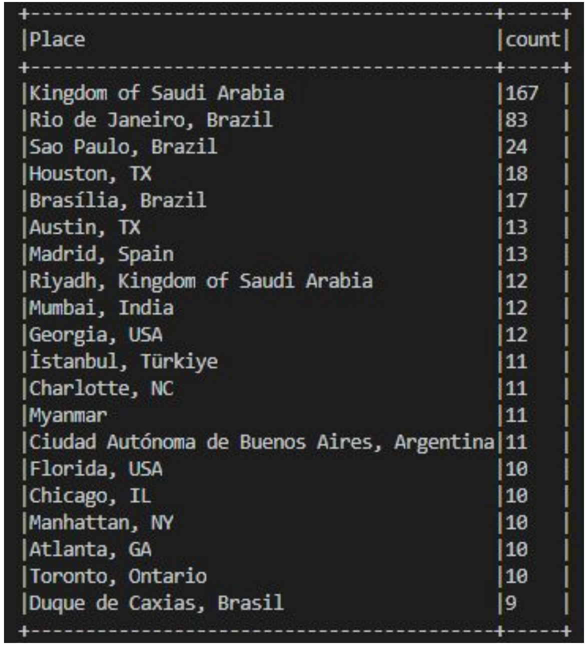
What individual location is referenced the most in a day? (Monday-Friday)



```scala
def filterLocation(spark: SparkSession): Unit = {
  import spark.implicits._
  val df = spark.read.json("q2_m")

  df
    .filter(!functions.isnull($"includes.places"))
    .select(
      functions.element_at($"includes.places", 1)("full_name").as("Place")
    )
    .groupBy("Place")
    .count()
    .sort(functions.desc("count"))
    .show(20,false)
}
```

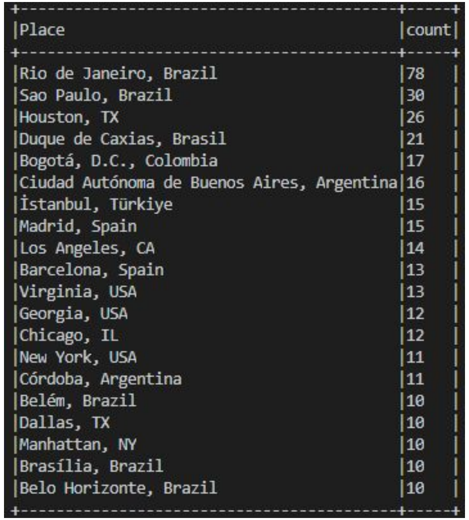
Monday



```
+------------------------------------------+-----+
|Place                                     |count|
+------------------------------------------+-----+
|Kingdom of Saudi Arabia                   |167  |
|Rio de Janeiro, Brazil                    |83   |
|Sao Paulo, Brazil                         |24   |
|Houston, TX                               |18   |
|Brasília, Brazil                          |17   |
|Austin, TX                                |13   |
|Madrid, Spain                             |13   |
|Riyadh, Kingdom of Saudi Arabia           |12   |
|Mumbai, India                             |12   |
|Georgia, USA                              |12   |
|İstanbul, Türkiye                         |11   |
|Charlotte, NC                             |11   |
|Myanmar                                   |11   |
|Ciudad Autónoma de Buenos Aires, Argentina|11   |
|Florida, USA                              |10   |
|Chicago, IL                               |10   |
|Manhattan, NY                             |10   |
|Atlanta, GA                               |10   |
|Toronto, Ontario                          |10   |
|Duque de Caxias, Brasil                   |9    |
+------------------------------------------+-----+
```

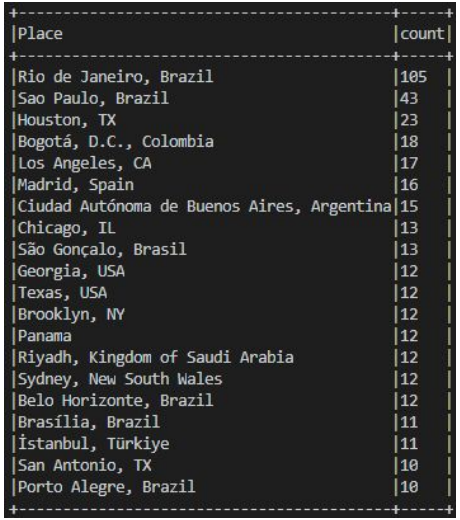
What individual location is referenced the most in a day? (Monday-Friday)

Tuesday



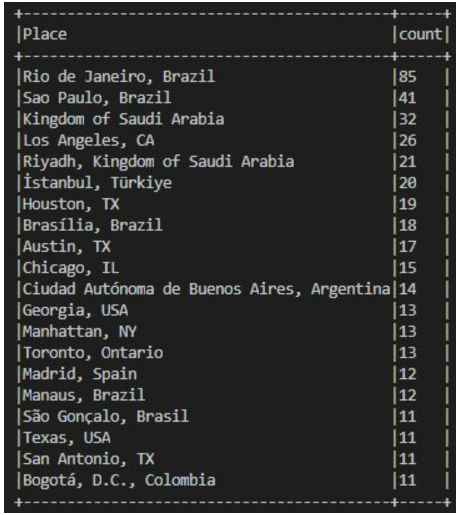
```
+------------------------------------------+-----+
|Place                                     |count|
+------------------------------------------+-----+
|Rio de Janeiro, Brazil                    |78   |
|Sao Paulo, Brazil                         |30   |
|Houston, TX                               |26   |
|Duque de Caxias, Brasil                   |21   |
|Bogotá, D.C., Colombia                    |17   |
|Ciudad Autónoma de Buenos Aires, Argentina|16   |
|İstanbul, Türkiye                         |15   |
|Madrid, Spain                             |15   |
|Los Angeles, CA                           |14   |
|Barcelona, Spain                          |13   |
|Virginia, USA                             |13   |
|Georgia, USA                              |12   |
|Chicago, IL                               |12   |
|New York, USA                             |11   |
|Córdoba, Argentina                        |11   |
|Belém, Brazil                             |10   |
|Dallas, TX                                |10   |
|Manhattan, NY                             |10   |
|Brasília, Brazil                          |10   |
|Belo Horizonte, Brazil                    |10   |
+------------------------------------------+-----+
```

Wednesday



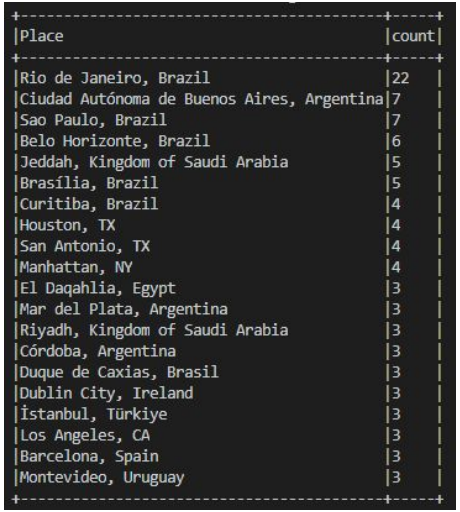
```
+------------------------------------------+-----+
|Place                                     |count|
+------------------------------------------+-----+
|Rio de Janeiro, Brazil                    |105  |
|Sao Paulo, Brazil                         |43   |
|Houston, TX                               |23   |
|Bogotá, D.C., Colombia                    |18   |
|Los Angeles, CA                           |17   |
|Madrid, Spain                             |16   |
|Ciudad Autónoma de Buenos Aires, Argentina|15   |
|Chicago, IL                               |13   |
|São Gonçalo, Brasil                       |13   |
|Georgia, USA                              |12   |
|Texas, USA                                |12   |
|Brooklyn, NY                              |12   |
|Panama                                    |12   |
|Riyadh, Kingdom of Saudi Arabia           |12   |
|Sydney, New South Wales                   |12   |
|Belo Horizonte, Brazil                    |12   |
|Brasília, Brazil                          |11   |
|İstanbul, Türkiye                         |11   |
|San Antonio, TX                           |10   |
|Porto Alegre, Brazil                      |10   |
+------------------------------------------+-----+
```

Thursday



```
+------------------------------------------+-----+
|Place                                     |count|
+------------------------------------------+-----+
|Rio de Janeiro, Brazil                    |85   |
|Sao Paulo, Brazil                         |41   |
|Kingdom of Saudi Arabia                   |32   |
|Los Angeles, CA                           |26   |
|Riyadh, Kingdom of Saudi Arabia           |21   |
|İstanbul, Türkiye                         |20   |
|Houston, TX                               |19   |
|Brasília, Brazil                          |18   |
|Austin, TX                                |17   |
|Chicago, IL                               |15   |
|Ciudad Autónoma de Buenos Aires, Argentina|14   |
|Georgia, USA                              |13   |
|Manhattan, NY                             |13   |
|Toronto, Ontario                          |13   |
|Madrid, Spain                             |12   |
|Manaus, Brazil                            |12   |
|São Gonçalo, Brasil                       |11   |
|Texas, USA                                |11   |
|San Antonio, TX                           |11   |
|Bogotá, D.C., Colombia                    |11   |
+------------------------------------------+-----+
```

Friday

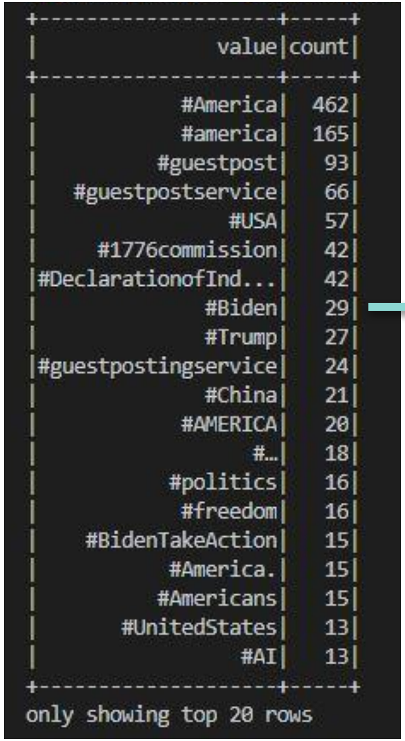


```
+------------------------------------------+-----+
|Place                                     |count|
+------------------------------------------+-----+
|Rio de Janeiro, Brazil                    |22   |
|Ciudad Autónoma de Buenos Aires, Argentina|7    |
|Sao Paulo, Brazil                         |7    |
|Belo Horizonte, Brazil                    |6    |
|Jeddah, Kingdom of Saudi Arabia           |5    |
|Brasília, Brazil                          |5    |
|Curitiba, Brazil                          |4    |
|Houston, TX                               |4    |
|San Antonio, TX                           |4    |
|Manhattan, NY                             |4    |
|El Daqahlia, Egypt                        |3    |
|Mar del Plata, Argentina                  |3    |
|Riyadh, Kingdom of Saudi Arabia           |3    |
|Córdoba, Argentina                        |3    |
|Duque de Caxias, Brasil                   |3    |
|Dublin City, Ireland                      |3    |
|İstanbul, Türkiye                         |3    |
|Los Angeles, CA                           |3    |
|Barcelona, Spain                          |3    |
|Montevideo, Uruguay                       |3    |
+------------------------------------------+-----+
```

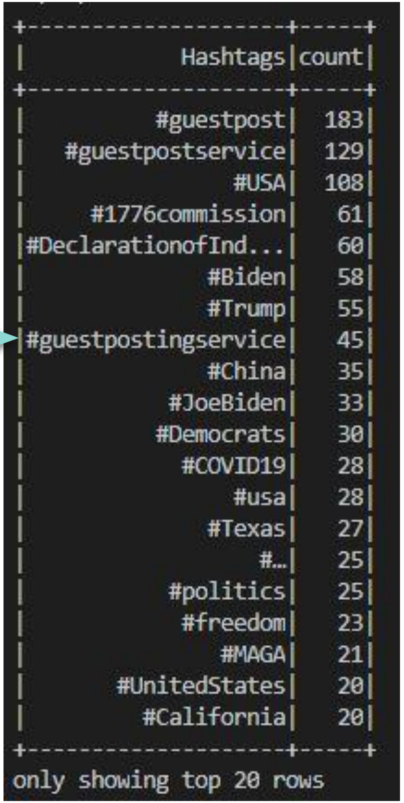
What other hashtags are associated with the America hashtag?

In order to answer this question we need to look at filtered streams and create a rule for only including tweets that have “#America”

- Take tweets and filter out all words except ones beginning with Hashtags
- Then group by and count the resulting Hashtags and compare the data



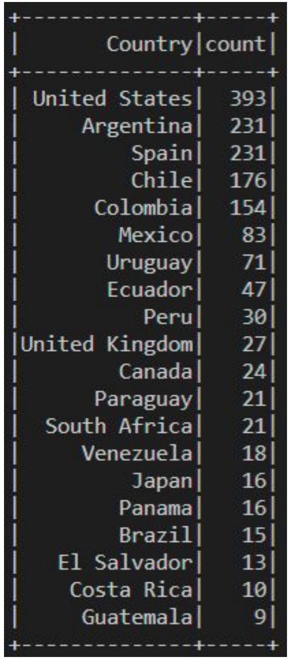
```
+--------------------+-----+
|               value|count|
+--------------------+-----+
|            #America|  462|
|            #america|  165|
|          #guestpost|   93|
|   #guestpostservice|   66|
|                #USA|   57|
|      #1776commission|   42|
|#DeclarationofInd...|   42|
|              #Biden|   29|
|              #Trump|   27|
|#guestpostingservice|   24|
|              #China|   21|
|            #AMERICA|   20|
|                  #…|   18|
|           #politics|   16|
|            #freedom|   16|
|     #BidenTakeAction|   15|
|           #America.|   15|
|          #Americans|   15|
|       #UnitedStates|   13|
|                 #AI|   13|
+--------------------+-----+
only showing top 20 rows
```



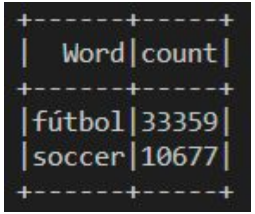
```
+--------------------+-----+
|            Hashtags|count|
+--------------------+-----+
|          #guestpost|  183|
|   #guestpostservice|  129|
|                #USA|  108|
|      #1776commission|   61|
|#DeclarationofInd...|   60|
|              #Biden|   58|
|              #Trump|   55|
|#guestpostingservice|   45|
|              #China|   35|
|           #JoeBiden|   33|
|          #Democrats|   30|
|            #COVID19|   28|
|                #usa|   28|
|              #Texas|   27|
|                  #…|   25|
|           #politics|   25|
|            #freedom|   23|
|               #MAGA|   21|
|       #UnitedStates|   20|
|         #California|   20|
+--------------------+-----+
only showing top 20 rows
```

Which country is talking about soccer/ fútbol more out of the countries that use these terms?

- Used the filtered stream to collect the data
- Counted the number of times users used the word fútbol vs soccer
- Grouped and counted the country to see which country was talking about the sport the most



```
+--------------+-----+
|       Country|count|
+--------------+-----+
| United States|  393|
|     Argentina|  231|
|         Spain|  231|
|         Chile|  176|
|      Colombia|  154|
|        Mexico|   83|
|       Uruguay|   71|
|       Ecuador|   47|
|          Peru|   30|
|United Kingdom|   27|
|        Canada|   24|
|      Paraguay|   21|
|  South Africa|   21|
|     Venezuela|   18|
|         Japan|   16|
|        Panama|   16|
|        Brazil|   15|
|   El Salvador|   13|
|    Costa Rica|   10|
|     Guatemala|    9|
+--------------+-----+
```

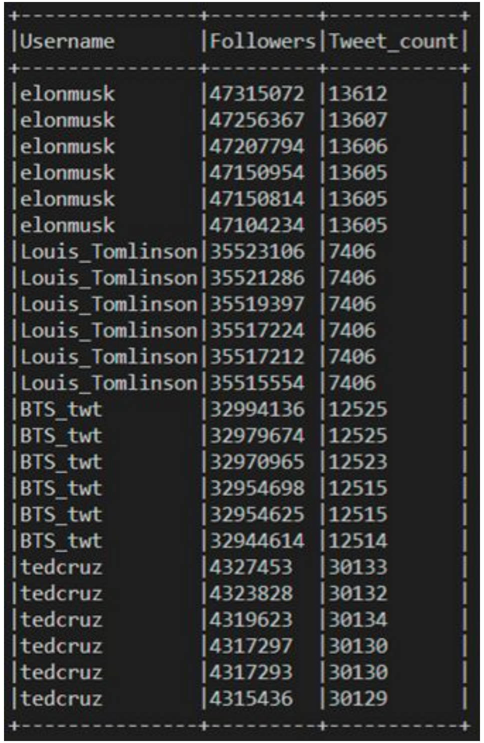


```
+------+-----+
|  Word|count|
+------+-----+
|fútbol|33359|
|soccer|10677|
+------+-----+
```

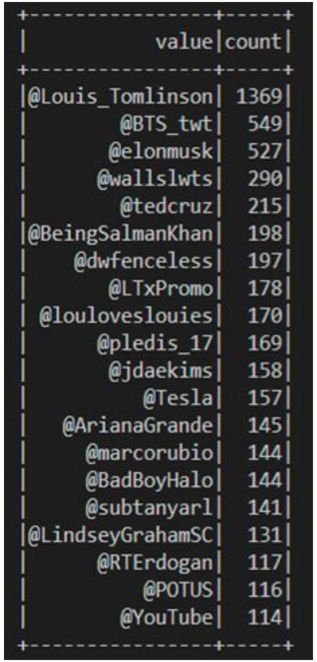
Does the activity on a Twitter page translate to more followers?

- To determine whether these factors had influence over the other, we looked at the top mentions from the sampled stream and chose four from that list to compare activity and followers. This data was recorded throughout the following week.

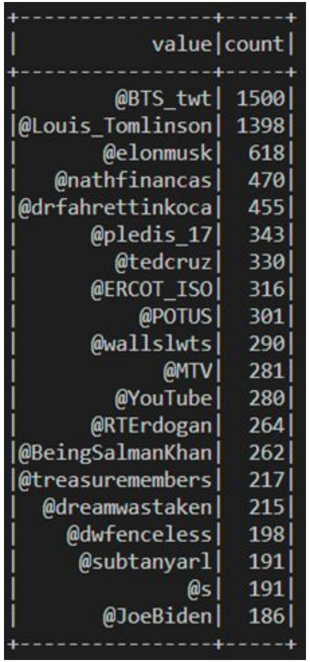
Question 5 Data



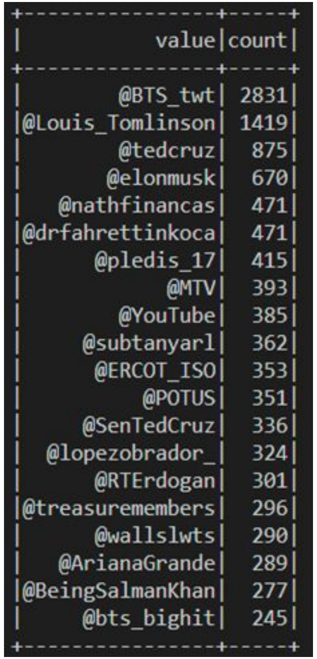
```
+---------------+---------+-----------+
|Username       |Followers|Tweet_count|
+---------------+---------+-----------+
|elonmusk       |47315072 |13612      |
|elonmusk       |47256367 |13607      |
|elonmusk       |47207794 |13606      |
|elonmusk       |47150954 |13605      |
|elonmusk       |47150814 |13605      |
|elonmusk       |47104234 |13605      |
|Louis_Tomlinson|35523106 |7406       |
|Louis_Tomlinson|35521286 |7406       |
|Louis_Tomlinson|35519397 |7406       |
|Louis_Tomlinson|35517224 |7406       |
|Louis_Tomlinson|35517212 |7406       |
|Louis_Tomlinson|35515554 |7406       |
|BTS_twt        |32994136 |12525      |
|BTS_twt        |32979674 |12525      |
|BTS_twt        |32970965 |12523      |
|BTS_twt        |32954698 |12515      |
|BTS_twt        |32954625 |12515      |
|BTS_twt        |32944614 |12514      |
|tedcruz        |4327453  |30133      |
|tedcruz        |4323828  |30132      |
|tedcruz        |4319623  |30134      |
|tedcruz        |4317297  |30130      |
|tedcruz        |4317293  |30130      |
|tedcruz        |4315436  |30129      |
+---------------+---------+-----------+
```



```
+----------------+-----+
|           value|count|
+----------------+-----+
|@Louis_Tomlinson| 1369|
|        @BTS_twt|  549|
|       @elonmusk|  527|
|      @wallslwts|  290|
|        @tedcruz|  215|
|@BeingSalmanKhan|  198|
|     @dwfenceless|  197|
|       @LTxPromo|  178|
|  @louloveslouies|  170|
|      @pledis_17|  169|
|       @jdaekims|  158|
|          @Tesla|  157|
|   @ArianaGrande|  145|
|     @marcorubio|  144|
|     @BadBoyHalo|  144|
|     @subtanyarl|  141|
|@LindseyGrahamSC|  131|
|      @RTErdogan|  117|
|          @POTUS|  116|
|        @YouTube|  114|
+----------------+-----+
```



```
+----------------+-----+
|           value|count|
+----------------+-----+
|        @BTS_twt| 1500|
|@Louis_Tomlinson| 1398|
|       @elonmusk|  618|
|    @nathfinancas|  470|
|@drfahrettinkoca|  455|
|      @pledis_17|  343|
|        @tedcruz|  330|
|      @ERCOT_ISO|  316|
|          @POTUS|  301|
|      @wallslwts|  290|
|            @MTV|  281|
|        @YouTube|  280|
|      @RTErdogan|  264|
|@BeingSalmanKhan|  262|
|@treasuremembers|  217|
|   @dreamwastaken|  215|
|     @dwfenceless|  198|
|     @subtanyarl|  191|
|              @s|  191|
|       @JoeBiden|  186|
+----------------+-----+
```



```
+----------------+-----+
|           value|count|
+----------------+-----+
|        @BTS_twt| 2831|
|@Louis_Tomlinson| 1419|
|        @tedcruz|  875|
|       @elonmusk|  670|
|    @nathfinancas|  471|
|@drfahrettinkoca|  471|
|      @pledis_17|  415|
|            @MTV|  393|
|        @YouTube|  385|
|     @subtanyarl|  362|
|      @ERCOT_ISO|  353|
|          @POTUS|  351|
|     @SenTedCruz|  336|
|   @lopezobrador_|  324|
|      @RTErdogan|  301|
|@treasuremembers|  296|
|      @wallslwts|  290|
|   @ArianaGrande|  289|
|@BeingSalmanKhan|  277|
|      @bts_bighit|  245|
+----------------+-----+
```

GitHub Link：

https://github.com/jacobbboullion/project2