Projecte final

Anàlisi de dades del Bicing Barcelona

Mel Aubets Serra

Índex

Introducció 3

1. Consultes 3

1.1. Considerant rendibles aquelles estacions en les que hi ha més d'un 50% de bicicletes usades de mitjana, quantes estacions rendibles hi ha? Quantes no ho son? 3

Codi associat a la consulta 3

1.2. Quins són els districtes postals amb més i menys rendibilitat? 4

Codi associat a la consulta 5

1.3. Quins són els districtes postals amb més i menys estacions per població? 5

Codi associat a la consulta 6

1.4. A quins districtes postals s'haurien d'afegir estacions? A quins en sobren? 6

Codi associat a la consulta 7

1.5. Representació gràfica de rendibilitat i estacions per població per districte postal 7

2. Optimització 8

2.1. Pandas versus PySpark 8

2.2. Configuració de PySpark 8

Introducció

En aquest projecte es realitzarà l'anàlisi de les dades del Bicing Barcelona 2021 extretes del web https://opendata-ajuntament.barcelona.cat/

Es realitzaran una sèrie de consultes referents a la rendibilitat i s'analitzarà el nombre d'estacions en funció de la població per districte postal per acabar concloent en quins districtes fan falta més estacions. També s'adjuntarà un gràfic recollint les dades obtingudes.

Per acabar, es farà una breu comparativa del temps d'execució en PySpark o en Pandas.

1. Consultes

1.1. Considerant rendibles aquelles estacions en les que hi ha més d'un 50% de bicicletes usades de mitjana, quantes estacions rendibles hi ha? Quantes no ho son?

Per poder respondre a aquesta consulta s'ha de resoldre quantes bicicletes s'utilitzen de mitjana en cada estació, per fer-ho s'hauran d'agrupar les dades per identificador d'estació, calcular la mitjana de bicicletes utilitzades per cada una d'elles i, finalment, dividir aquesta mitjana per la capacitat de l'estació.

Un cop realitzades aquestes operacions es realitza un comptatge damunt d'aquelles que superin el 50% de mitjana i s'obtenen els següents resultats:

Si es consideren rendibles aquelles estacions amb més d'un 50% de bicicletes utilitzades de mitjana :
Hi ha 412 estacions rendibles i 93 estacions no rendibles.

L'estació més rendible és la:C/ LLOBREGÓS, 115 amb un ratio de 0.9070498322206834 de bicicletes usades de mitjana

L'estació menys rendible és la:PL. JOAQUIM XIRAU I PALAU, 1 amb un ratio de 0.2525915963496568 de bicicletes usades de mitjana

Codi associat a la consulta

```
bicing_status =
bicing_status.groupBy('station_id').agg(avg('num_bikes_available').alias('num_bikes_available'))
bicing = bicing_status.join(bicing_info,
['station_id']).orderBy('station_id')
```

In [5]:

```
rentable = bicing.select('station_id',
                         'name',
                         'num_bikes_available',
                         'capacity',
                         'post_code')

rentable = rentable.withColumn("ratio",
                               ((1-
(col("num_bikes_available")/col("capacity"))))).withColumn("post_code"
, rentable['post_code'].cast(IntegerType()))
```

In [6]:

```
ren=rentable.filter("ratio >= 0.5").count()
noRen=rentable.count()-ren
mesRen=rentable.orderBy('ratio', ascending = False).first()
mesRenNom = mesRen.name
mesRenRatio = mesRen.ratio
menysRen=rentable.orderBy('ratio').first()
menysRenNom = menysRen.name
menysRenRatio = menysRen.ratio
```

1.2. Quins són els districtes postals amb més i menys rendibilitat?

Per resoldre aquesta qüestió s'agruparà la base de dades *rentable*, obtinguda en la qüestió anterior, mitjançant els codis postals, se sumarà el ratio associat a cada estació per districtes i es dividirà pel nombre d'estacions per districte, segons el valor obtingut per aquesta divisió es podrà saber en quins districtes s'utilitzen més o menys bicicletes.

Els resultats obtinguts són els següents:

```
Els deu districtes postals on les estacions son més rendibles son els
següents:

+---------+------------------+
|post_code|    rentable_ratio|
+---------+------------------+
|     8022| 0.863675165299612|
|     8023|0.8538471649190068|
|     8042|0.8389067044491677|
|     8006|0.8253970535673893|
|     8035|0.8228713662239416|
|     8034|0.8142623837704457|
|     8031|0.8073166819548427|
|     8021|0.8042107511431367|
|     8017|0.7943923542670357|
|     8016|0.7928681424709142|
+---------+------------------+


Els deu districtes postals on les estacions son menys rendibles son el
s següents:

+---------+-------------------+
|post_code|     rentable_ratio|
+---------+-------------------+
|     8930| 0.3999327766761839|
|     8039|0.41874030510298893|
|     8019|0.45585316610564053|
|     8002| 0.4808707605915257|
|     8001|  0.483718681754345|
|     8038| 0.4935458895792438|
|     8003| 0.5015208624620618|
|     8005| 0.5051067519346677|
|     8018| 0.5493721503180977|
|     8013| 0.5786848062259456|
+---------+-------------------+
```

Codi associat a la consulta

```
rentable_zones = rentable.groupBy('post_code')\
                         .agg((sum('ratio')/count('post_code'))
                              .alias('rentable_ratio'))\
                         .orderBy('rentable_ratio', ascending = False)

a = int(rentable_zones.count()/2)
m_rentable_zones = rentable_zones.limit(a+1)
l_rentable_zones = rentable_zones.orderBy('rentable_ratio').limit(a)
```

1.3. Quins són els districtes postals amb més i menys estacions per població?

Per a aquesta consulta s'ha utilitzat la base de dades *post_population,* que conté els valors de població per districte postal, i s'ha ajuntat amb la base de dades *bicing*, s'ha agrupat per codi postal i població i s'ha dividit el nombre d'estacions per el nombre d'habitants. Aquesta dada s'ha normalitzat a 1 per ser més llegible.

Els resultats obtinguts són els següents:

```
Els deu districtes postals amb menys estacions per població son els se
güents:

+---------+--------------------+
|post_code|    population_ratio|
+---------+--------------------+
|    08040|0.004914438090200875|
|     8001| 0.04227888899421746|
|    08035|0.054142640364188165|
|    08038|  0.1107399214482416|
|    08022| 0.13268379755308468|
|    08039|  0.1563236296368416|
|    08033| 0.15701115536095403|
|    08004|  0.1814975583288117|
|    08023| 0.18673458706953702|
|    08032| 0.18884524005551506|
+---------+--------------------+

Els deu districtes postals amb més estacions per població son els següe
nts:

+---------+-------------------+
|post_code|   population_ratio|
+---------+-------------------+
|    08002|                1.0|
|    08007| 0.8354993560473013|
|    08010| 0.8300569966267303|
|    08011| 0.8064289301858363|
|    08015| 0.6910137228862329|
|    08003| 0.6532623169107856|
|    08037| 0.5993012631013169|
|    08005| 0.5511740253904056|
|    08013| 0.5063336520076482|
|    08025|0.49122549379543284|
+---------+-------------------+
```

Codi associat a la consulta

```
stations_population = bicing.join(post_population, ['post_code'])

population = stations_population.groupBy('post_code','population')\
                    .agg((count('post_code')/col('population'))
                         .alias('ratio'))\
                    .orderBy('ratio', ascending = False)

max_z = population.first().ratio

population_zones =
population.withColumn('population_ratio',(population['ratio']/max_z))\
            .select('post_code','population_ratio')\
            .orderBy('population_ratio')

m_population_zones = population_zones.limit(a+1)
l_population_zones = population_zones.orderBy('population_ratio',
ascending = False).limit(a)
```

1.4. A quins districtes postals s'haurien d'afegir estacions? A quins en sobren?

Per resoldre aquesta darrera qüestió s'ha dividit les bases de dades en dues fraccions, una per sobre de la meitat i l'altra per sota. S'ha seleccionat aquells districtes postals més rendibles i amb més població per estació (o menys estacions per població) com aquells districtes que caldria reforçar. El mateix plantejament per aquells que quedin per sota la meitat serviran per conèixer els districtes on hi sobren estacions.

Els resultats obtinguts són els següents:

```
A continuació es mostren els districtes postals en els que convindria
construir estacions:
+---------+------------------+--------------------+
|post_code|    rentable_ratio|    population_ratio|
+---------+------------------+--------------------+
|     8035|0.8228713662239416|0.054142640364188165|
|     8022| 0.863675165299612| 0.13268379755308468|
|     8033|0.7213261817156337| 0.15701115536095403|
|     8023|0.8538471649190068| 0.18673458706953702|
|     8032|0.7777921779959808| 0.18884524005551506|
|     8017|0.7943923542670357| 0.18995756073074643|
|     8006|0.8253970535673893| 0.19317812669193288|
|     8034|0.8142623837704457| 0.23670680333034796|
|     8042|0.8389067044491677|  0.2626464872504564|
|     8024|0.7720508720527861|  0.2830602354027233|
|     8016|0.7928681424709142|   0.300944669365722|
|     8031|0.8073166819548427|   0.314239413795533|
|     8028|0.7551573413659856| 0.33711816224920715|
+---------+------------------+--------------------+
```

```
A continuació es mostren els districtes postals en els que hi sobren e
stacions:
+---------+------------------+-------------------+
|post_code|    rentable_ratio|   population_ratio|
```

```
+---------+------------------+-------------------+
|     8002|0.4808707605915257|                1.0|
|     8010|0.6386914998384332| 0.8300569966267303|
|     8011|0.6022215355491175| 0.8064289301858363|
|     8015|0.5909462325557184| 0.6910137228862329|
|     8003|0.5015208624620618| 0.6532623169107856|
|     8005|0.5051067519346677| 0.5511740253904056|
|     8013|0.5786848062259456| 0.5063336520076482|
|     8018|0.5493721503180977| 0.4770719002883453|
|     8001| 0.483718681754345| 0.4650677789363921|
|     8009|0.6396749609647617|0.43353584447144594|
|     8014|0.6425283776016352|  0.401259559154296|
|     8026|0.6145421112402386|0.35426950354609926|
+---------+------------------+-------------------+
```

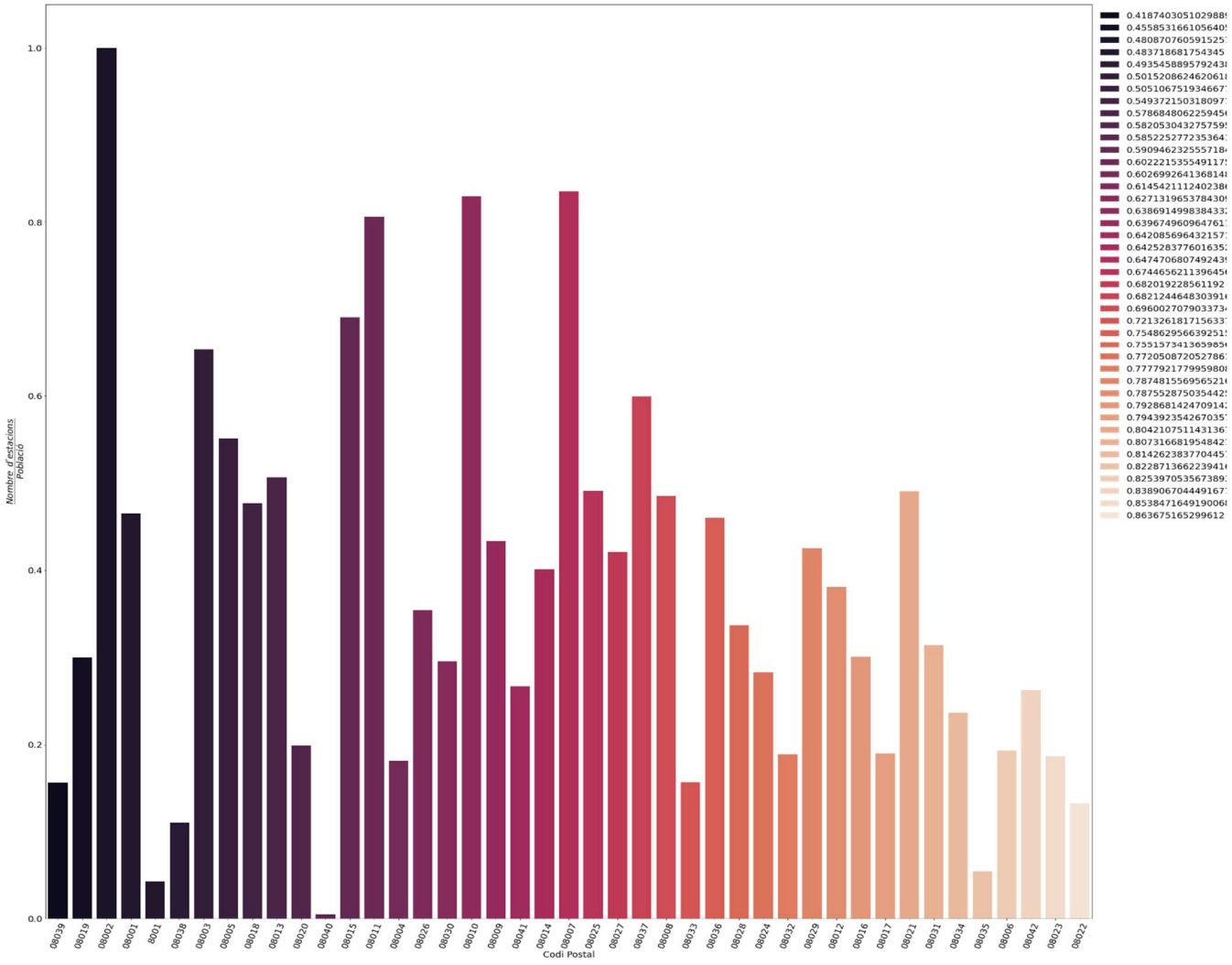
Codi associat a la consulta

```
bikes_needed = m_rentable_zones.join(m_population_zones,
['post_code'], "inner")\

.select('post_code','rentable_ratio','population_ratio')


bikes_leftover = l_rentable_zones.join(l_population_zones,
['post_code'], "inner")\

.select('post_code','rentable_ratio','population_ratio')
```

1.5. Representació gràfica de rendibilitat i estacions per població per districte postal

S'ha associat un gradient de color a la rendibilitat, els colors més foscos són les estacions menys rendibles i els clars els més. S'ha representat un gràfic de barres amb les estacions per població a l'eix y i els districtes postals a l'eix x.

Els districtes més clars i amb les barres més baixes son aquells on fa falta reforç.



Gràfic 1: Nombre d'estacions per població i rendibilitat en funció del districte postal

2. Optimització

2.1. Pandas versus PySpark

S'ha realitzat la primera consulta utilitzant el mòdul pandas de python per poder fer una comparativa del temps emprat en cada cas, utilitzant PySpark s'obté el resultat per a la primera consulta en 77.369s, en canvi, utilitzant Pandas tarda 222.519s, és a dir que hi ha una diferencia de 145.15s, o, dit d'una altra manera, Pandas ha tardat casi el triple ($\frac{222.519s}{77.369s} = 2.876$).

2.2. Configuració de PySpark

Tot i que s'ha anat modificant la configuració de PySpark (nombre de nuclis = 1, 4, 8 , memòria utilitzada = 2, 4) no s'ha apreciat cap diferencia notable entre els temps d'execució.