Exercicio Hive

Ricardo...

**Índice**

[Dataset 2](#_bookmark0)

[Estudando o dataset 2](#_bookmark1)

[Preparando o dataset 4](#_bookmark2)

[Preguntas a resolver 7](#_bookmark3)

[Creación de estrutura e carga de datos 8](#_bookmark4)

[Consultas Hive 8](#_bookmark5)

[Consulta 1: Cantos rexistros correspondentes a viaxes contén o dataset 8](#_bookmark6)

[Consulta 2: Cales son as 5 estacións con maior número de saídas 8](#_bookmark7)

[Consulta 3: Cales son as 5 estacións con maior número de chegadas 8](#_bookmark8)

[Consulta 4: Cal é a viaxe coa maior distancia percorrida 8](#_bookmark9)

[Consulta 5: Canta distancia percorreuse en total entre tódalas viaxes 8](#_bookmark10)

[Consulta 6: Cal é a distancia media percorrida por viaxe 9](#_bookmark11)

[Consulta 7: Cal é a viaxe de maior duración 9](#_bookmark12)

[Consulta 8: Cal é a duración media por viaxe 9](#_bookmark13)

[Consulta 9: Cal é a velocidade media da viaxes 9](#_bookmark14)

[Consulta 10: Cantas viaxes se fixeron cando a temperatura era menor de 0ºC 9](#_bookmark15)

[Consulta 11: Cantas viaxes fixéronse cando a temperatura estaba entre 15 e 25 ºC 9](#_bookmark16)

[Consulta 12: Datos da primeira viaxe segundo data 9](#_bookmark17)

[Resultado da execución das consultas Hive 10](#_bookmark18)

[Consulta 1: hive\_01\_viaxes\_count 10](#_bookmark19)

[Consulta 2: hive\_02\_estacions\_mais\_saidas 10](#_bookmark20)

[Consulta 3: hive\_03\_estacions\_mais\_chegadas 10](#_bookmark21)

[Consulta 4: hive\_04\_viaje\_max\_distancia 10](#_bookmark22)

[Consulta 5: hive\_05\_total\_distancia 10](#_bookmark23)

[Consulta 6: hive\_06\_distancia\_media 10](#_bookmark24)

[consulta 7: hive\_07\_viaxe\_duracion\_maior 10](#_bookmark25)

[consulta 8 : hive\_08\_duracion\_media\_viaxes 10](#_bookmark26)

[consulta 09: hive\_09\_velocidade\_media\_viaxes 10](#_bookmark27)

[consulta 10: hive\_10\_viaxes\_frio 11](#_bookmark28)

[consulta 11: hive\_11\_viaxes\_temp\_ok 11](#_bookmark29)

[consulta 12: hive\_12\_datos\_primeira\_viaxe 11](#_bookmark30)

# Dataset

## Estudando o dataset

#### Nome orixinal do ficheiro:

database. csv

#### Renomeado como:

bikes\_ database. csv

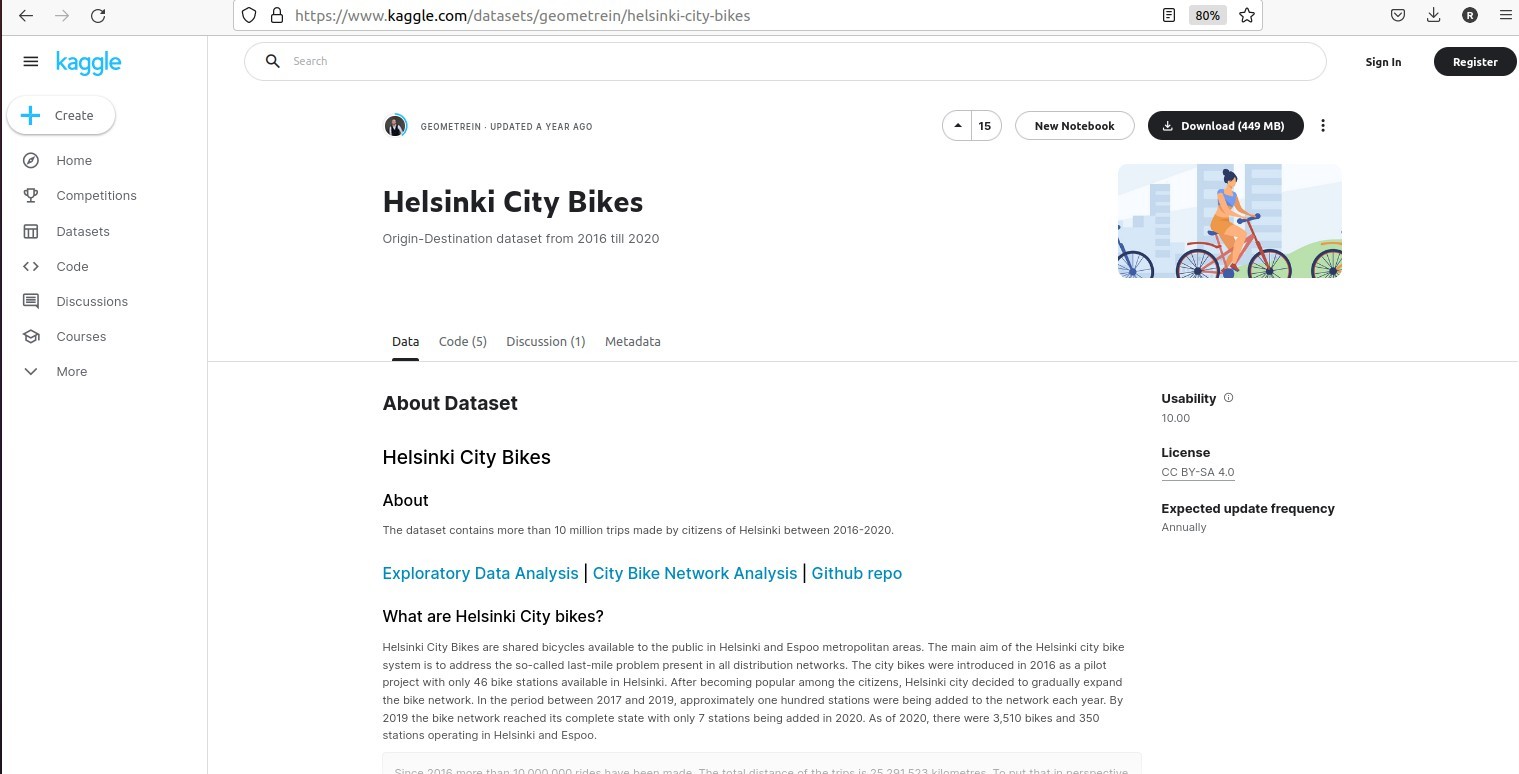
#### Orixe:

link: <https://www.kaggle.com/geometrein/helsinki-city-bikes>

#### Descrición:

Helsinqui City Bikes é o sistema de rede de bicicletas públicas compartidas nas áreas metropolitanas de Helsinqui e Espoo (Finlandia). En 2020 había funcionando 350 estacións e 3.510 bicicletas.

Entre 2016 e 2020 realizáronse máis de 10 millóns de desprazamentos. A distancia total dos devanditos traxectos foi de 25.291.523 Km.



#### Tamaño do dataset :

2,0 GB (1985327688 bytes)

#### Tamaño comprimido zip :

439,9 MB (439875920 bytes)

#### Tamaño comprimido tar. xz :

232,5 MB (232489532 bytes)

#### Contido:

O dataset contén máis de 10 millóns de rexistros, cada un con datos de desprazamentos en bicicleta

realizados por cidadáns de Helsinqui entre 2016 e 2020, usando a rede de bicicletas compartidas da área Helsinqui e Espoo.

#### Número total de rexistros:

12.157.458 rexistros

#### Campos:

14 campos

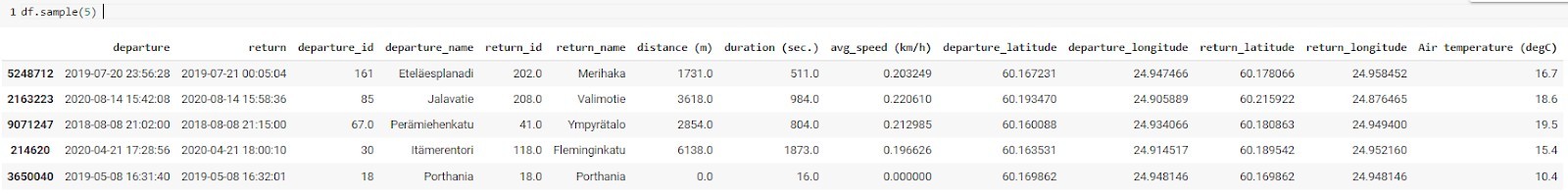
#### Nome, tipo e contido dos campos:

float64(8), object(6)

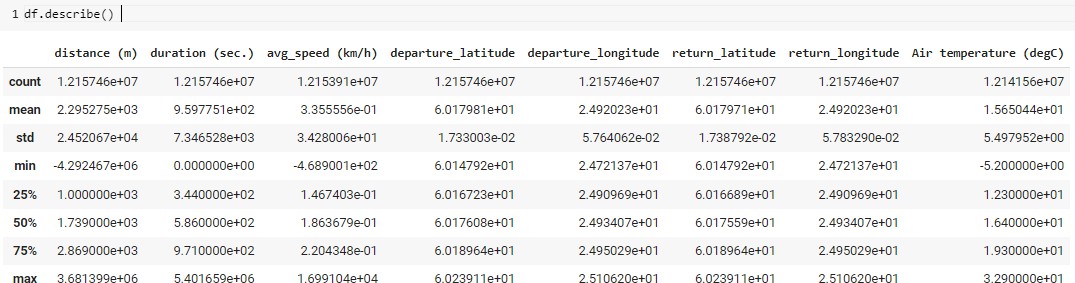
# Colum Dtype Contido

1. departure object data e hora de partida
2. return object data e hora de chegada
3. departure\_ide object id de saída
4. departure\_name object nome estación de saída
5. return\_ide object id de chegada
6. return\_name object nome estación de chegada
7. distance (m) float64 distancia (estimada) do traxecto en metros
8. duration (sec.) float64 duración do traxecto en segundos
9. avg\_speed (km/ h) float64 velocidade media do traxecto (km/h)
10. departure\_latitude float64 latitude estación saída
11. departure\_longitude float64 lonxitude estación saída
12. return\_latitude float64 latitude estación chegada
13. return\_longitude float64 lonxitude estación chegada
14. Air temperature (degC) float64 temperatura do aire en data do traxecto (ºC)

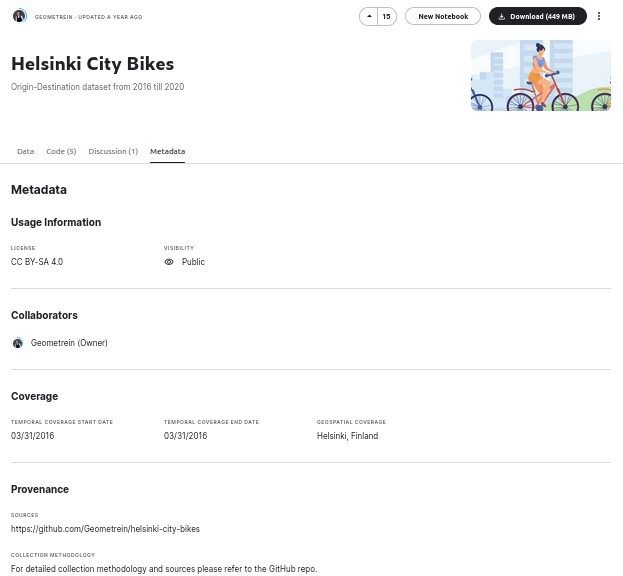
#### Mostra:



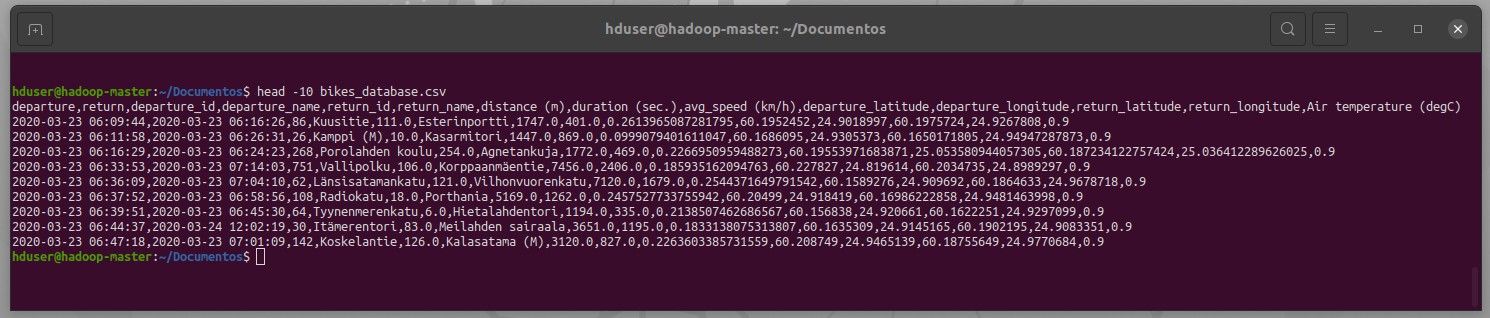
#### Resumo estatístico:



#### Metadatos:



#### Head:



Obsérvase que o dataset ten ringleira de encabezado e que os campos están separados por coma.

## Preparando o dataset

Antes de empeza-las consultas, adecuamo-lo dataset, para o que usaremos *python* e *pandas*, por exemplo.

import pandas as pd df=pd.read\_csv('ruta/bikes\_database.csv')

Hai unha serie de columnas que, alomenos de momento, non se van a usar:

departure\_id return\_id

departure\_latitude departure\_longitude return\_latitude return\_longitude

Quitamos con ***drop*** as columnas que non usaremos:

df.drop('nome columna', inplace = True, axis = 1)

Renomeamo-las restantes columnas:

df.rename(columns={'departure':'fecha\_salida', 'return':'fecha\_llegada', 'departure\_name':'estacion\_salida', 'return\_name':'estacion\_llegada', 'distance(m)':'distancia\_m', 'duration(sec.)':'duracion\_sec', 'avg\_speed (km/h)':'velocidad\_media', 'Air temperature (degC)':'temperatura'},

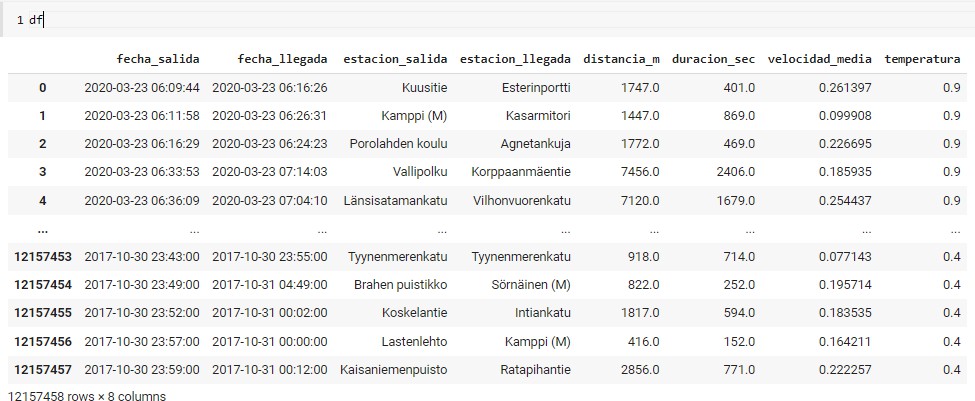
inplace= True)

Convertemo-lo tipo de dato object (string) a data/hora nas columnas que conteñen a data e hora de chegada e partida:

df['fecha\_salida'] = df['fecha\_salida'].astype('datetime64') df['fecha\_llegada'] = df['fecha\_llegada'].astype('datetime64')

#### Info do dataset resultante:

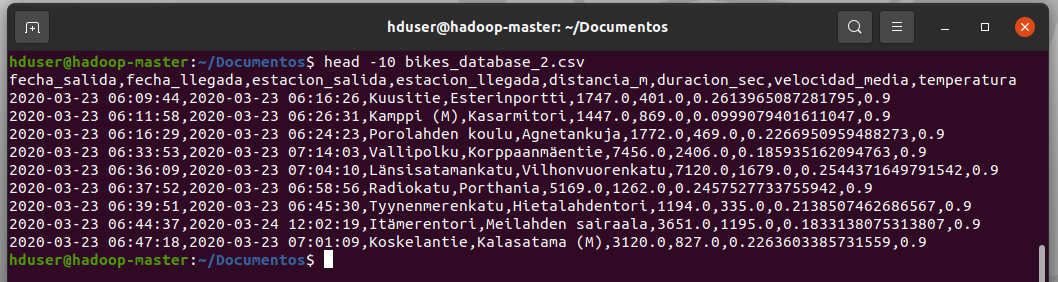
#### Mostra do dataset resultante:



Gárdase o dataset modificado con outro nome, en formato csv:

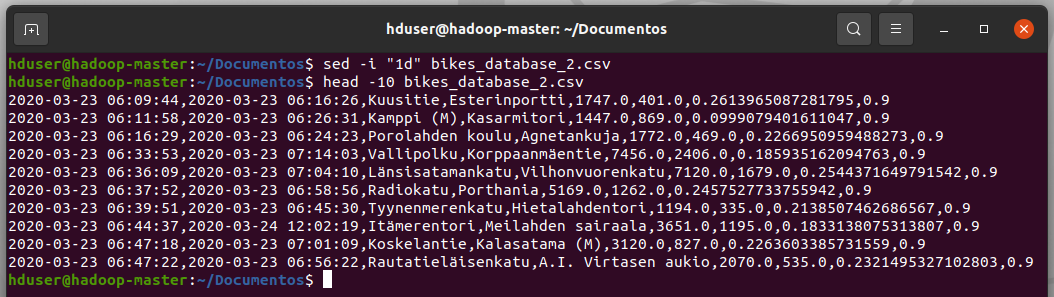
df.to\_csv('bikes\_database\_2.csv' ,index= False)

#### Head:



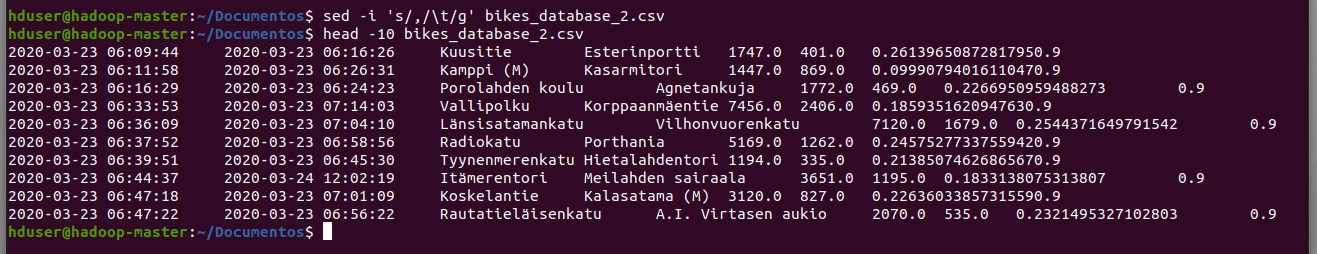
Quitamo-los encabezados para traballar con só os datos en Hive:

sed -i "1d" bikes\_database\_2.csv



Para evitar problemas, cambiamo-la coma que fai de separadora de columnas polo tabulador:

sed -i 's/,/\t/g' bikes\_database\_2.csv



Unha vez que o dataset xa está preparado e depurado, podemos realiza-las consultas.

# Preguntas a resolver

1. Cantos rexistros correspondentes a viaxes contén o dataset
2. Cales son as 5 estacións con maior número de saídas
3. Cales son as 5 estacións con maior número de chegadas
4. Cal é a viaxe coa maior distancia percorrida
5. Canta distancia percorreuse en total entre todas as viaxes
6. Cal é a distancia media percorrida por viaxe
7. Cal é a viaxe de maior duración
8. Cal é a duración media por viaxe
9. Cal é a velocidade media das viaxes
10. Cantas viaxes fixéronse cando a temperatura era menor de 0º C
11. Cantas viaxes fixéronse cando a temperatura estaba entre 15 e 25ºC
12. Datos da primeira viaxe segundo data

Máis ideas:

* + Preguntas de viaxes por datas (por meses, días da semana, por horas, etc)
  + Preguntas de viaxes segundo temperaturas (duración de viaxes a temperaturas inferiores a 0ºC, etc)

Pendente:

Join con dataset días choiva e preguntas de viaxes segundo días choiva <https://en.ilmatieteenlaitos.fi/download-observations>

## Creación de estrutura e carga de datos

CREATE EXTERNAL TABLE IF NOT EXISTS viaxes (

data\_ini TIMESTAMP, data\_fin TIMESTAMP, saida STRING, chegada STRING, distancia FLOAT, duracion FLOAT, velocidade FLOAT, temperatura FLOAT) ROW FORMAT DELIMITED

FIELDS TERMINATED BY '\t' STORED AS TEXTFILE;

LOAD DATA INPATH ‘/user/hduser/input/bikes\_2.csv’ INTO TABLE viaxes;

## Consultas Hive

Para que as saídas das consultas se vaian gardando no sistema hdfs, iniciamos cada consulta con:

INSERT OVERWRITE DIRECTORY '/user/hduser/output/nome\_ficheiro' ....

### Consulta 1: Cantos rexistros correspondentes a viaxes contén o dataset

INSERT OVERWRITE DIRECTORY '/user/hduser/output/hive\_01\_viaxes\_count' SELECT COUNT(\*) FROM viaxes

### Consulta 2: Cales son as 5 estacións con maior número de saídas

INSERT OVERWRITE DIRECTORY 'output/hive\_02\_estacions\_mais\_saidas' SELECT saida, COUNT(\*) AS num\_rex FROM viaxes GROUP BY saida ORDER BY num\_rex DESC LIMIT 5;

### Consulta 3: Cales son as 5 estacións con maior número de chegadas

INSERT OVERWRITE DIRECTORY 'output/hive\_03\_estacions\_mais\_chegadas' SELECT chegada, COUNT(\*) AS num\_rec FROM viaxes GROUP BY chegada ORDER BY num\_rex DESC LIMIT 5;

### Consulta 4: Cal é a viaxe coa maior distancia percorrida

INSERT OVERWRITE DIRECTORY 'output/hive\_04\_max\_distancia' SELECT distancia FROM viaxes ORDER BY distancia DESC LIMIT 1;

### Consulta 5: Canta distancia percorreuse en total entre tódalas viaxes

INSERT OVERWRITE DIRECTORY 'output/hive\_05\_total\_distancia' SELECT SUM(distancia) FROM viaxes;

### Consulta 6: Cal é a distancia media percorrida por viaxe

INSERT OVERWRITE DIRECTORY 'output/hive\_06\_distancia\_media' SELECT AVG(distancia) FROM viaxes;

### Consulta 7: Cal é a viaxe de maior duración

INSERT OVERWRITE DIRECTORY 'output/hive\_07\_max\_duracion' SELECT duracion FROM viaxes ORDER BY duracion DESC LIMIT 1;

### Consulta 8: Cal é a duración media por viaxe

INSERT OVERWRITE DIRECTORY 'output/hive\_08\_duracion\_media' SELECT AVG(duracion) FROM viaxes;

### Consulta 9: Cal é a velocidade media da viaxes

INSERT OVERWRITE DIRECTORY 'output/hive\_09\_velocidade\_media' SELECT AVG(velocidad) FROM viaxes;

### Consulta 10: Cantas viaxes se fixeron cando a temperatura era menor de 0ºC

INSERT OVERWRITE DIRECTORY 'output/hive\_10\_viaxes\_frio' SELECT COUNT(\*) FROM viaxes WHERE temperatura < 0;

### Consulta 11: Cantas viaxes fixéronse cando a temperatura estaba entre 15 e 25 ºC

INSERT OVERWRITE DIRECTORY 'output/hive\_11\_viaxes\_temp\_ok'

SELECT COUNT(\*) FROM viaxes WHERE temperatura>=15 AND temperatura<26;

### Consulta 12: Datos da primeira viaxe segundo data

INSERT OVERWRITE DIRECTORY 'output/hive\_12\_datos\_primeira\_viaxe' SELECT \* FROM viaxes ORDER BY data\_ini LIMIT 1;

# Resultado da execución das consultas Hive

## Consulta 1: hive\_01\_viaxes\_count

12157458

## Consulta 2: hive\_02\_estacions\_mais\_saidas

Itämerentori 330397

Töölönlahdenkatu 242555

Kamppi (M) 201560 Rautatientori / länsi 175358 Ympyrätalo 172776

## Consulta 3: hive\_03\_estacions\_mais\_chegadas

Itämerentori 332453

Töölönlahdenkatu 243592

Kamppi (M) 195787 Rautatientori / länsi 178954 Ympyrätalo 177476

## Consulta 4: hive\_04\_viaje\_max\_distancia

359383.34

(359,38 km)

## Consulta 5: hive\_05\_total\_distancia

2.5475649018170776 E9

(25.475.649. km)

## Consulta 6: hive\_06\_distancia\_media

2587.072079543791

(2,587 km)

## consulta 7: hive\_07\_viaxe\_duracion\_maior

5012669.0

(58 días)

## consulta 8 : hive\_08\_duracion\_media\_viaxes

1060.442647813547

(17 min. 40 sec.)

## consulta 09: hive\_09\_velocidade\_media\_viaxes

10.870777369952668

(10,87 km/h)

## consulta 10: hive\_10\_viaxes\_frio

3769

## consulta 11: hive\_11\_viaxes\_temp\_ok

260036

## consulta 12: hive\_12\_datos\_primeira\_viaxe

2016-05-02T09:31:00.000Z;2016-05-02T09:32:00.000Z;Erottajan aukio;Erottajan aukio;2.0;68.0;0.0017647059;14.1

data\_ini 2016-05-02T09:31:00.000Z; data\_fin 2016-05-02T09:32:00.000Z;

saida Erottajan aukio; chegada Erottajan aukio; distancia 2.0;

duracion 68.0;

velocidade 0.0017647059;

temperatura 14.1