projeto

October 13, 2019

1 Projeto da Disciplina CAP394 - Introdução à Data Science

1. Professores:

- Rafael Santos
- Gilberto Ribeiro de Queiroz

2. Aluno:

- Henrique Eduardo de Macedo

1.0.1 Sobre o Conjunto de dados

Os dados a serem analisados são o conjunto de registros obtidos a partir do movimento dos veículos da Uber em suas viagens pela cidade de São Paulo. O dados podem ser obtidos no site https://movement.uber.com/cities.

O dataset possui 9 colunas contendo basicamente um inteiro como código da origem e do destino a hora os tempos médios do deslocamento.

Como conjunto auxiliar de dados, será utilizado uma tabela json contendo as áreas da cidade e seus códigos correspondentes também disponível no site.

1.0.2 Perguntas básicas a serem respondidas:

1. Qual o melhor horário para efetuar deslocamentos para o aeroporto de Guarulhos ? **RESPONDIDA**
2. No horário de pico, qual o sentido de deslocamento predominante do tráfego ? **RESPONDIDA**
3. É possível estimar o tempo de delocamento entre dois bairros com algum modelo de apreendizagem de máquina ? **EM PROGRESSO**
4. É possível estimar o aumento no tempo de deslocamento devido a alguma interrupção no tráfego ? **EM PROGRESSO**

[1]:
```
import pandas as pd
```

[2]:
```
df = pd.read_csv(r'C:
↪\Users\hp\Downloads\uber\sao_paulo-od_zones_2017-2018-4-All-HourlyAggregate.
↪csv')
```

[3]:
```
df.shape
```

[3]:
```
(2950937, 7)
```

1.0.3 Sobre a lista de variáveis

- **sourceid:** Código da região de partida da viagem
- **dstid:** Código da região de destino da viagem
- **hod:** Hora do dia
- **mean_travel_time:** Tempo médio da viagem
- **standard_deviation_travel_time:** Desvio padrão
- **geometric_mean_travel_time**: Média geométrica
- **geometric_standard_deviation_travel_time**: Desvio padrão da média geométrica

[4]:
```
df.columns
```

[4]:
```
Index(['sourceid', 'dstid', 'hod', 'mean_travel_time',
       'standard_deviation_travel_time', 'geometric_mean_travel_time',
       'geometric_standard_deviation_travel_time'],
      dtype='object')
```

[5]:
```
df.head()
```

[5]:

	sourceid	dstid	hod	mean_travel_time	standard_deviation_travel_time	geometric_mean_travel_time	geometric_standard_deviation_travel_time
0	84	119	1	1700.07	556.19	1631.09	1.31
1	418	456	5	1405.84	445.70	1346.89	1.33
2	463	99	20	2481.23	288.07	2464.21	1.13
3	72	239	1	1416.94	238.94	1398.21	1.17
4	37	70	1	1794.60	482.13	1736.27	1.29

1.0.4 Como traduzir o sourceid e dstid:

A fonte dos dados também fornece um json com os atributos relativos a código.

Como exemplo, para o código 0: “NumeroZona”:1, “NomeZona”:“Sé”, “NumDistrit”:80, “NomeDistri”:“Sé”, “NumeroMuni”:36, “NomeMunici”:“São Paulo”, “Paulo”,“MOVEMENT_ID”:“1”,“DISPLAY_NAME”:“Sé” “geometry”:{“type”:“Polygon”,“coordinates”:[[[-46.6291313,-23.5503351],[-46.6276154,-23.5518791],[-46.6275214,-23.5522241],[-46.6285723,-23.5521811],[-46.6295655,-23.5524881],[-46.6331403,-23.5524752],[-46.6353474,-23.55149],[-46.6354194,-23.5511471],[-46.6380082,-23.5499721],[-46.6388124,-23.549088],[-46.6388743,-23.548648],[-46.6352436,-23.5429958],[-46.6343978,-23.5438421],[-46.6338618,-23.5440981],[-46.6341168,-23.5444601],[-46.6336778,-

23.5451841],[-46.6332584,-23.5474111],[-46.6322882,-23.5463551],[-46.6310282,-23.5491991],[-46.6304198,-23.5487961],[-46.6300858,-23.5494561],[-46.6291313,-23.5503351]]]}}

Pergunta 01: Qual o melhor dia/horário para efetuar deslocamentos para os aeroportos ?

Para responder essa pergunta vamos inicialmente retirar do dataframe as colunas com com dados que não iremos usar. O dataset resultante também irá servir para os questionamentos seguintes.

após análise, excluímos as colunas abaixo com comando df.drop: 1. 'standard_deviation_travel_time' 2. 'geometric_mean_travel_time' 3. 'geometric_standard_deviation_travel_time'

[6]:
```
df = df.drop(['standard_deviation_travel_time', 'geometric_mean_travel_time',
       'geometric_standard_deviation_travel_time'],axis=1)
```

A partir deste dataframe criamos outro, DF_GRU, com todas as viagens com dstid=374, Aeroporto Internacional de Guarulhos.

[7]:
```
df_gru = df.loc[(df['dstid'] == 374)] #374 é o código do aeroporto de guarulhos
df_gru.head()
```

[7]:

	sourceid	dstid	hod	mean_travel_time
104	162	374	22	1048.84
328	154	374	0	1590.82
402	422	374	23	2972.89
629	245	374	8	2893.96
701	383	374	14	728.90

Após agrupar as linhas conforme os seus respectivos horários e extraindo a média, temos o resultado abaixo:

[18]:
```
grouped = df_gru[['mean_travel_time']].groupby(df_gru['hod'])
grouped.mean().sort_values(by='mean_travel_time',ascending=False)
```

[18]:

hod	mean_travel_time
16	3724.025952
17	3573.893348
15	3500.565882
18	3280.338013
14	3204.440852
19	2947.892181
13	2937.003742
6	2895.338858
7	2816.832721
12	2779.574286
10	2766.095956
11	2763.788226
9	2751.188606
8	2699.930675
5	2695.899636
20	2592.876720
21	2406.226325

```
4          2347.029893
22         2307.366348
23         2187.713661
3          2138.696491
2          2123.264146
1          2077.011418
0          2044.779897
```

Resultado Considerando todas as regiões da cidade, o horário com tempo de viagem mais lento é o das 16hs. O oposto é o horário da meia noite.

Pergunta 02: Nos horários de pico, qual o sentido de deslocamento predominante do tráfego ?

Para resoler esse problema vamos descobrir quais horários específicos apresentam o maior volume de corridas (tráfego) para refinir qual será o nosso "horário de pico".

CORRIDAS POR HORÁRIO DO DIA

[21]:
```
df['hod'].value_counts().sort_values(ascending=False)
```

[21]:
```
16    132369
17    131821
15    130918
18    130765
19    129865
14    128441
20    128206
8     127788
7     127646
6     127419
13    127068
21    126963
9     126909
11    126880
10    126633
12    126442
22    125647
23    121931
5     121508
0     114321
4     107338
1     105716
2      99287
3      99056
Name: hod, dtype: int64
```

CORRIDAS POR HORÁRIO DO DIA NORMALIZADO

[23]:
```
df['hod'].value_counts(normalize=True).sort_values(ascending=False)
```

[23]:
```
16    0.044857
17    0.044671
15    0.044365
18    0.044313
19    0.044008
14    0.043525
20    0.043446
8     0.043304
7     0.043256
6     0.043179
13    0.043060
21    0.043025
9     0.043006
11    0.042997
10    0.042913
12    0.042848
22    0.042579
23    0.041319
5     0.041176
0     0.038741
4     0.036374
1     0.035825
2     0.033646
3     0.033568
Name: hod, dtype: float64
```

Podemos verificar que o maior volume de tráfego (considerando as corridas de Uber) estão nos seguintes horários: 7-8:00hs e 16-17:00hs.

[45]:
```
pico_manha = df[(df["hod"] == 7) | (df["hod"] == 8)]
pico_manha["dstid"].value_counts(normalize=True) # normalize=True
```

[45]:
```
365    0.003633
364    0.003629
374    0.003610
366    0.003559
131    0.003547
161    0.003516
4      0.003484
9      0.003473
39     0.003469
2      0.003457
162    0.003433
5      0.003429
62     0.003406
92     0.003398
91     0.003367
270    0.003351
271    0.003335
```

```
132    0.003335
27     0.003328
169    0.003316
68     0.003316
57     0.003300
108    0.003300
73     0.003296
26     0.003292
170    0.003277
168    0.003273
6      0.003269
63     0.003265
60     0.003261
         ...
402    0.000317
408    0.000317
442    0.000305
392    0.000266
394    0.000258
412    0.000251
413    0.000235
511    0.000223
381    0.000196
355    0.000196
146    0.000168
444    0.000168
143    0.000164
414    0.000141
356    0.000141
473    0.000133
443    0.000090
471    0.000086
411    0.000070
513    0.000059
385    0.000051
509    0.000047
475    0.000047
517    0.000047
353    0.000031
434    0.000031
299    0.000027
357    0.000023
474    0.000008
510    0.000004
Name: dstid, Length: 514, dtype: float64
```

No horário de pico entre 7 e 8hs os destinos mais procurados foram 365, 364, 374 e 366 (região de guarulhos)

[46]:
```
pico_tarde = df[(df["hod"] == 16) | (df["hod"] == 17)]
pico_tarde["dstid"].value_counts(normalize=True) # normalize=True
```

[46]:
```
364    0.003524
365    0.003524
374    0.003490
366    0.003460
131    0.003437
161    0.003429
4      0.003335
9      0.003327
162    0.003304
132    0.003289
39     0.003289
2      0.003263
5      0.003244
92     0.003236
170    0.003236
169    0.003233
62     0.003225
91     0.003225
136    0.003202
27     0.003198
26     0.003183
378    0.003161
168    0.003153
270    0.003145
93     0.003134
111    0.003115
57     0.003111
40     0.003104
108    0.003104
271    0.003092
         ...
352    0.000337
381    0.000326
390    0.000314
442    0.000307
389    0.000303
355    0.000280
402    0.000273
413    0.000246
511    0.000235
356    0.000216
412    0.000201
146    0.000189
143    0.000185
```

```
444    0.000170
473    0.000159
414    0.000129
443    0.000102
513    0.000098
411    0.000087
471    0.000076
517    0.000068
475    0.000045
353    0.000045
509    0.000045
385    0.000038
299    0.000030
357    0.000026
434    0.000023
474    0.000011
510    0.000008
Name: dstid, Length: 514, dtype: float64
```

No horário de pico entre 16 e 17hs os destinos mais procurados foram 364, 365, 374 e 366 (região de guarulhos)

Resultado Nos horários de pico da manhã e da tarde os quatro maiores destinos são regiões de guarulhos.