Este documento contém um relatório que responde a algumas perguntas realizadas pelo time de negócios, bem como a execução de alguns dos quesitos bônus

Douglas Raimundo de Oliveira Silva
linkedin.com/in/dellonath/

Quesitos mínimos

• Qual a distância média percorrida por viagens com no máximo 2 passageiros;

A distância média percorrida para viagens com até dois passageiros é de 2.02Km (ou milhas, não especificado).

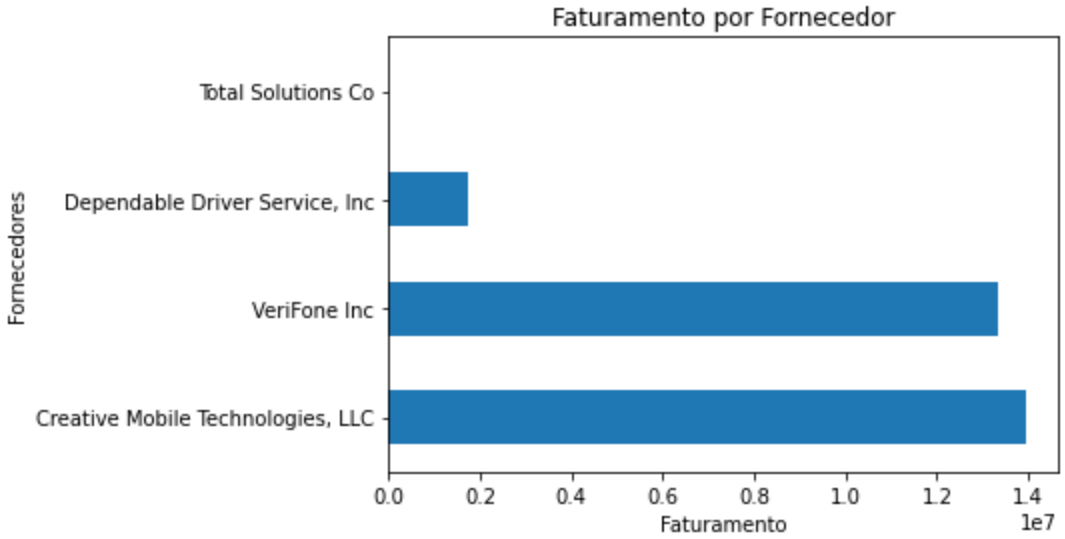
Obs: não é especificado a unidade de medida (se é milhas ou quilômetro), assumirei quilômetros.

```
[30]: data.query('qt_passenger <= 2').nr_trip_distance.mean()

[30]: 2.024365161879035
```

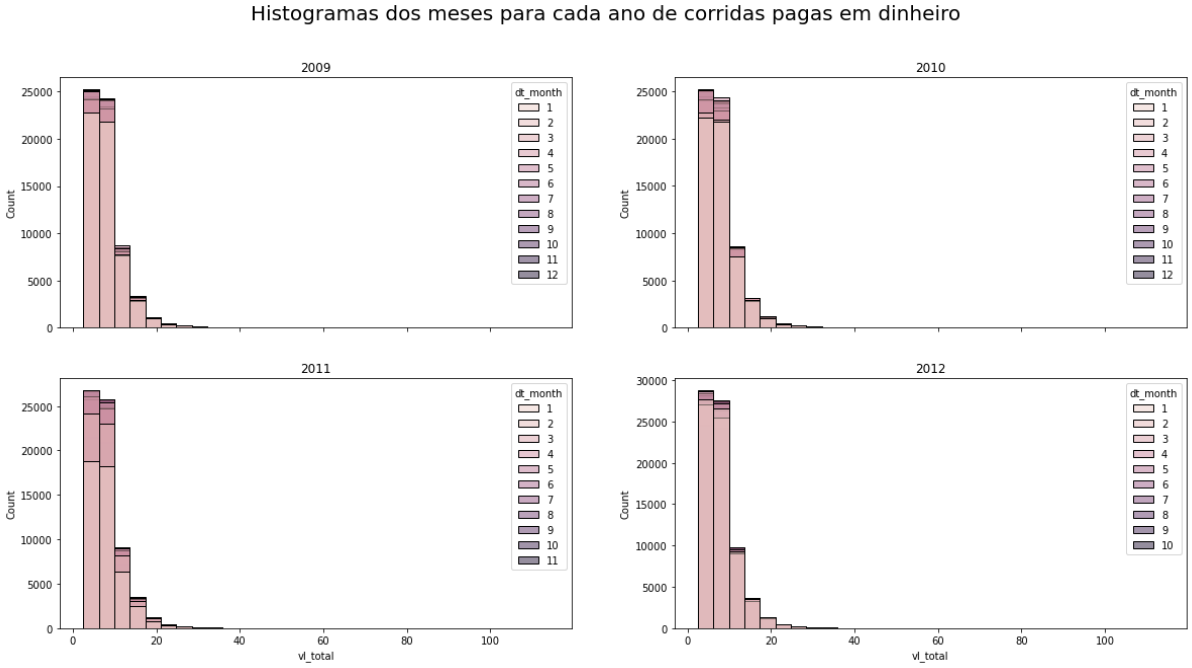
• Quais os 3 maiores vendors em quantidade total de dinheiro arrecadado;

Visualizando o gráfico acima, tem-se que **Creative Mobile Technologies, LLC** é o fornecedor que mais arrecadou, com um somatório, nos 4 anos, de 13950565.36 dólares, seguido pela **VeriFone Inc** (13356427.32 dólares), uma diferença de, aproximadamente, 2,58%. A terceira posição fica com a **Dependable Driver Service, Inc**, registrando um somatório de 1733328.68 dólares.



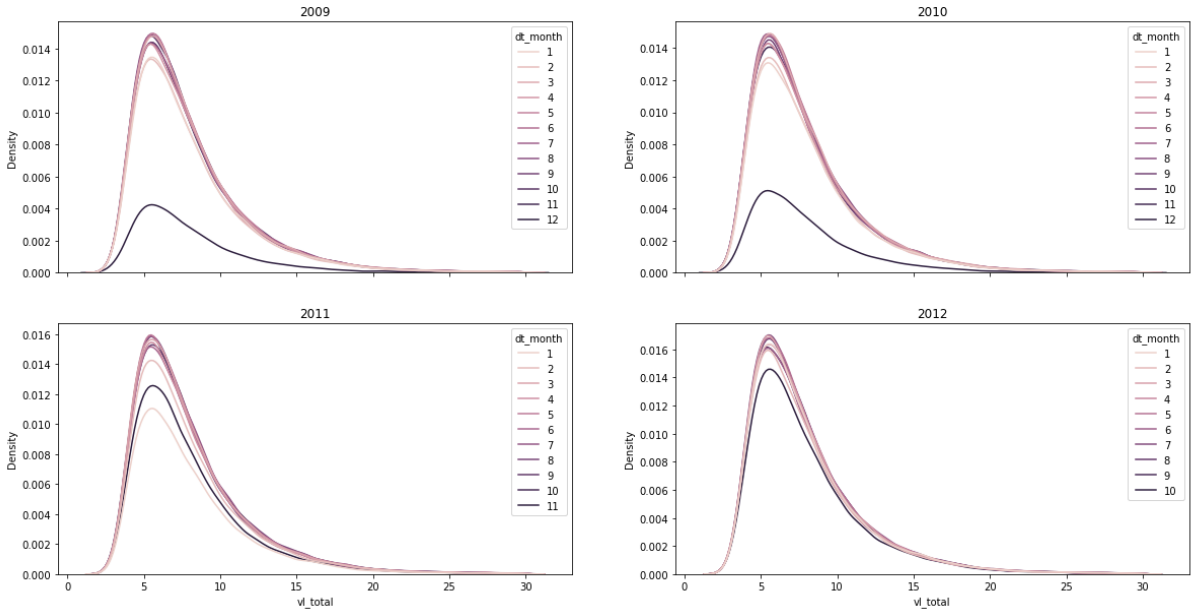
• Faça um histograma da distribuição mensal, nos 4 anos, de corridas pagas em dinheiro;

Para responder a este quesito, tomei o ano e mês dos dt_pickup e utilizei o subplots para separar cada um dos anos.



Julguei que a visualização não ficou muito amigável, por isso fiz um filtro retirando vl_total > 30 e utilizei o kdeplot para plotar a densidade ao invés do histograma em si. O resultado está abaixo.

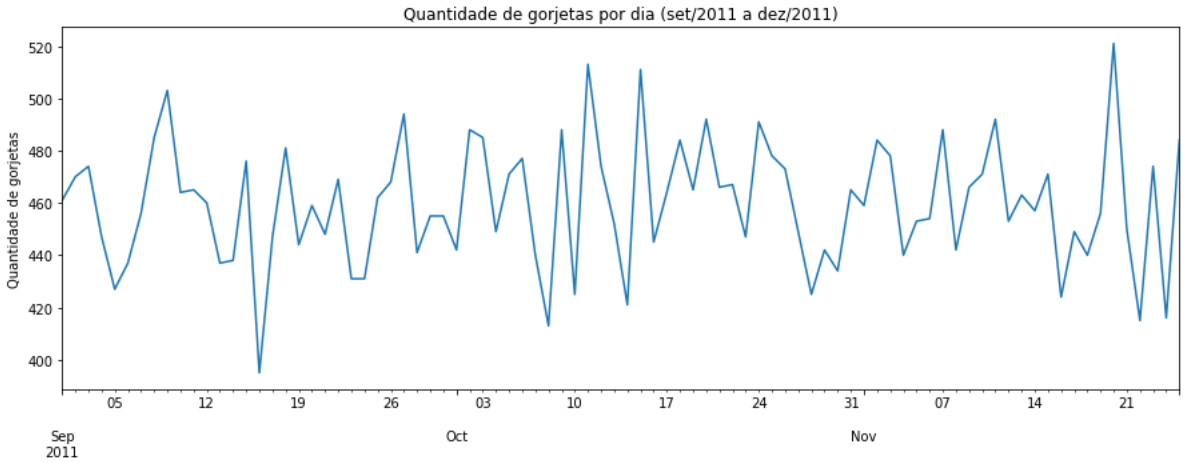
Histogramas dos meses para cada ano de corridas pagas em dinheiro

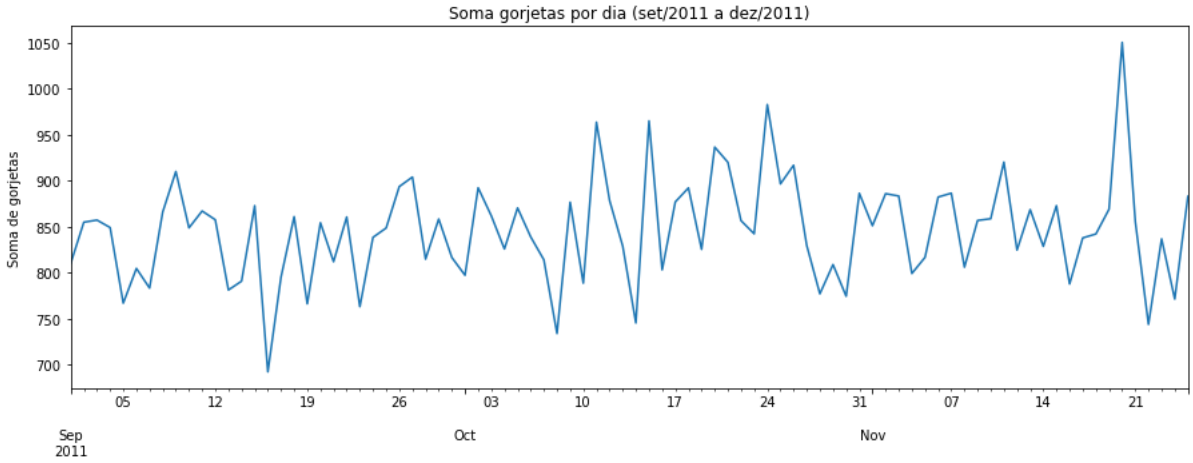


• Faça um gráfico de série temporal contando a quantidade de gorjetas de cada dia, nos últimos 3 meses de 2011.

O gráfico abaixo contém a quantidade de gorjetas, seguido pela soma total das gorjetas por dia. Para gerar a série temporal utilizei o método .resample() do pandas.

Para calcular a quantidade, eu apenas fiz uma condição lógica verificando, para cada viagem, se o vl_tip era maior que zero. Caso sim, eu atribuia 1, senão 0. No final eu realizava a soma desse campo (mais detalhes no código).





- Qual o tempo médio das corridas nos dias de semana;

Para calcular o tempo médio, fiz a diferença entre dt_dropoff - dt_pickup. Os resultados estão abaixo. O tempo médio de corridas na semana (de segunda a sexta) é de, aproximadamente, 8 minutos e 45 segundos.

A média por dia da semana é a mesma, independente do dia. Vide abaixo.

```
data.query("nm_weekday in ('Monday', 'Tuesday', 'Wednesday', 'Thuesday', 'Friday')").dt_trip_duration.mean()
```

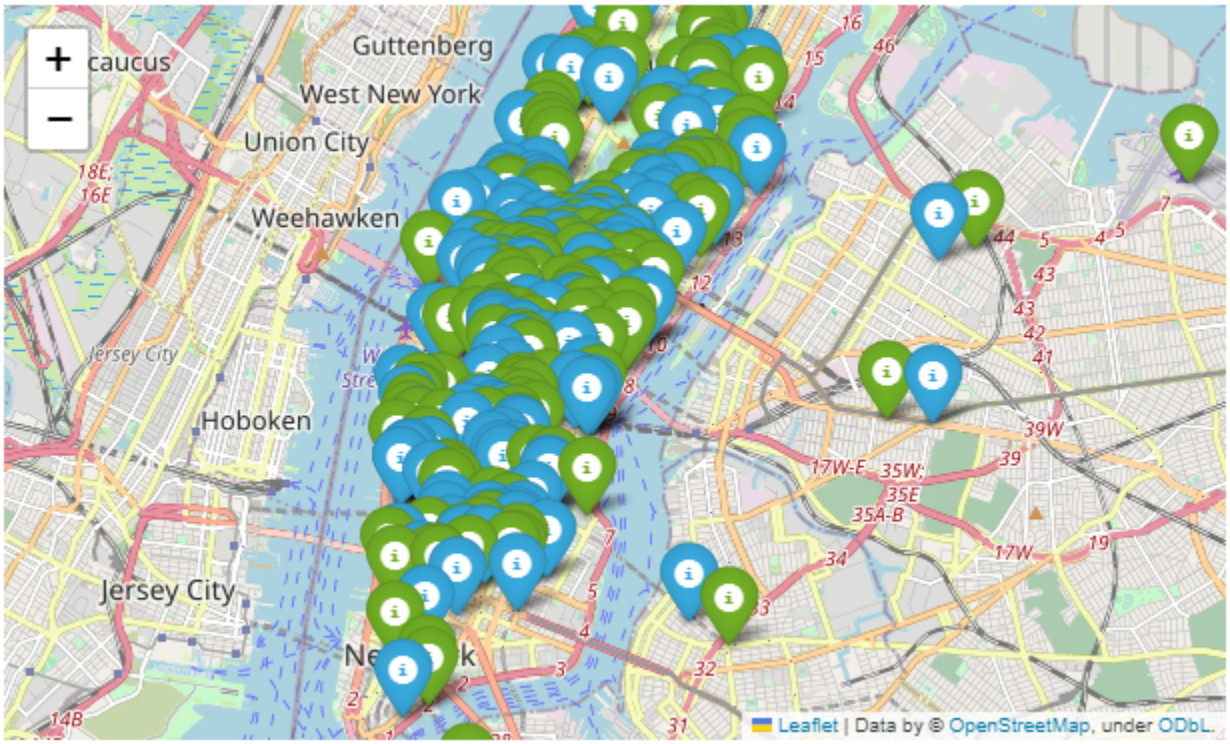
```
Timedelta('0 days 00:08:44.900675431')
```

```
data.groupby('nm_weekday').dt_trip_duration.mean().to_frame()
```

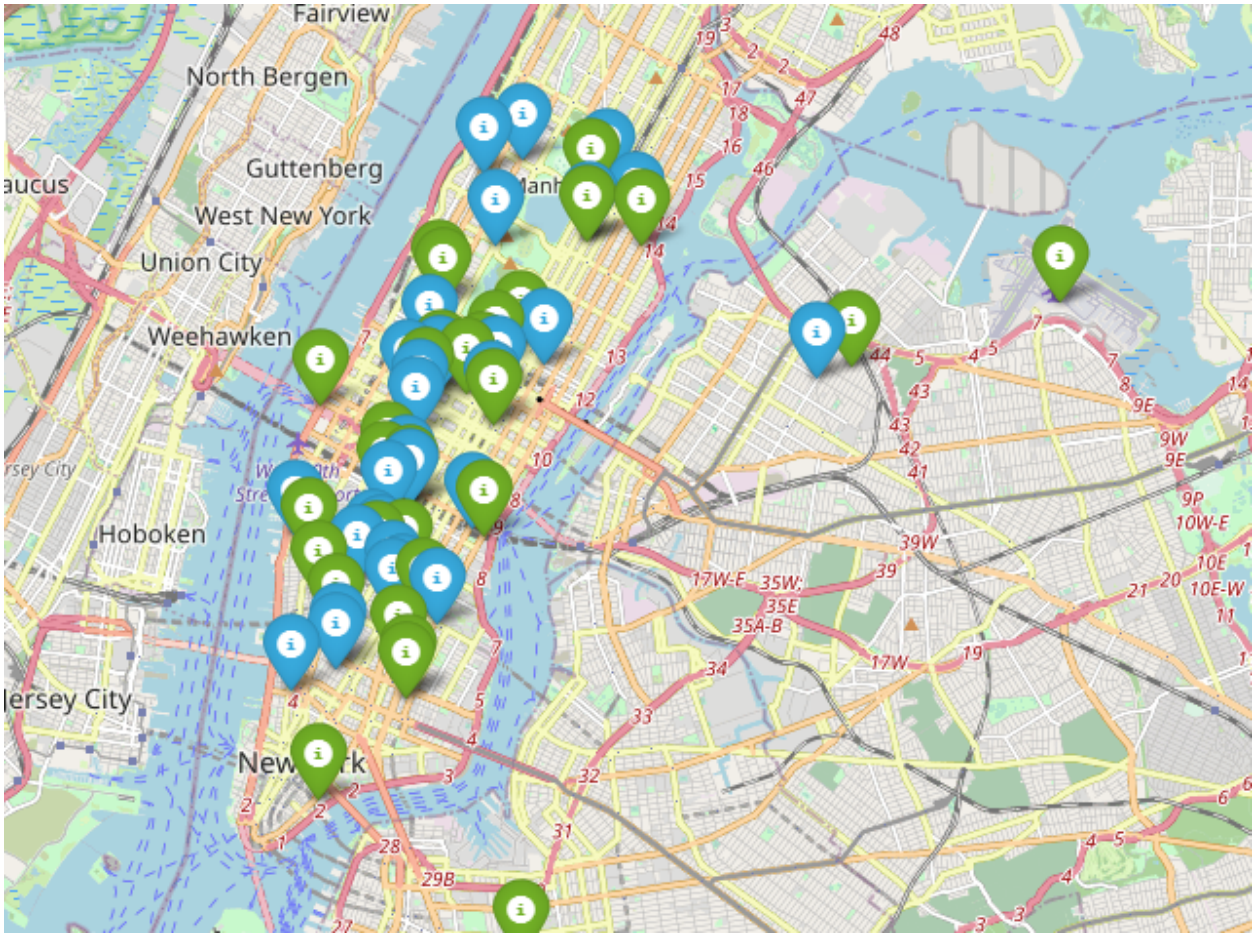
nm_weekday	dt_trip_duration
Friday	0 days 00:08:44.636195883
Monday	0 days 00:08:44.491497072
Saturday	0 days 00:08:45.123094297
Sunday	0 days 00:08:44.405574516
Thursday	0 days 00:08:46.152692091
Tuesday	0 days 00:08:45.313223140
Wednesday	0 days 00:08:44.123057885

• Fazer uma visualização em mapa com latitude e longitude de pickups and dropoffs no ano de 2010;

Abaixo têm-se um mapa com as coordenadas, como requisitado. Os pontos azuis representam os pickups, enquanto os verdes os dropoofs. Apenas cosiderado o ano de 2010.



Para melhor visualização, diminui o números de pontos plotados no mapa. Vale ressaltar que este mapa é interativo, a biblioteca folium (utilizada para gerar o mapa) permite aproximações e ajustes do mapa.



• Simular um streaming dos dados dos JSON e fazer uma visualização acompanhando uma métrica em tempo-real;

Não foi implementado.

• Conseguir provisionar todo seu ambiente em uma cloud pública, de preferência AWS; Implementação de alguma arquitetura de pipeline de engenharia de dados para suas análises.

O ambiente inteiro foi construído na AWS. Para o pipeline de engenharia de dados utilizei o S3 e o AWS Glue. Para as análises utilizei o Sagemaker. A parte de engenharia está melhor descrita no README.md.