Primeiramente, é importado a biblioteca pandas e a datetime. Depois é lido os JSON's disponibilizados com os nomes de: 'data-sample_data-nyctaxi-trips-20XX-json_corrigido.json', onde os XX podem ser trocados de acordo com o ano.

Depois disso, foi feito a separação dos dados para análise de acordo com o case técnico.

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

Para isso, foi feito uma separação dos dados das corridas com 2 ou menos passageiros.

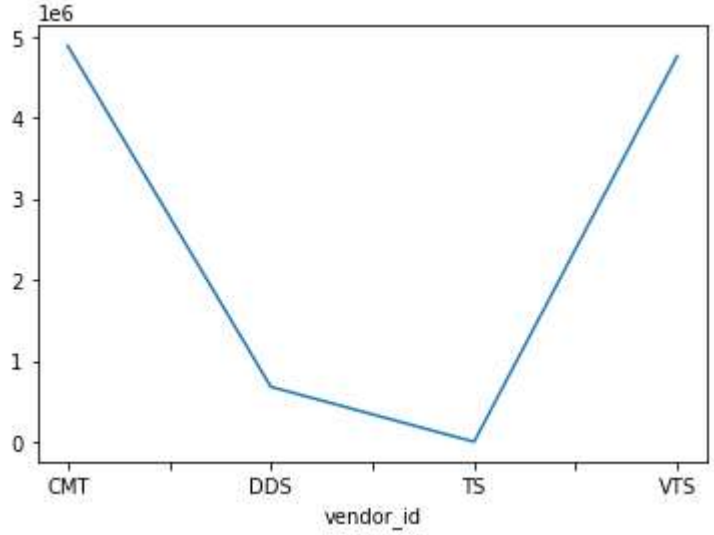
Depois foi feito em cada dataframe a média da distância percorrida, e depois novamente a média de todas as médias.

O resultado foi: 2.662526996203298. Como os dados são de Nova Iorque, deduzo que essa métrica esteja em milhas.

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

Para isso, foi agrupado os dados por 'vendor_id', e feito a soma das colunas, depois é gerado o gráfico com o total das vendas. Os 3 maiores vendedores em total de dinheiro arrecadado são: Creative Mobile Technologies, VeriFone Inc e Dependable Driver Service, nessa respectiva ordem para os 4 anos.

Os 4 anos obtiveram gráficos iguais que foi:



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

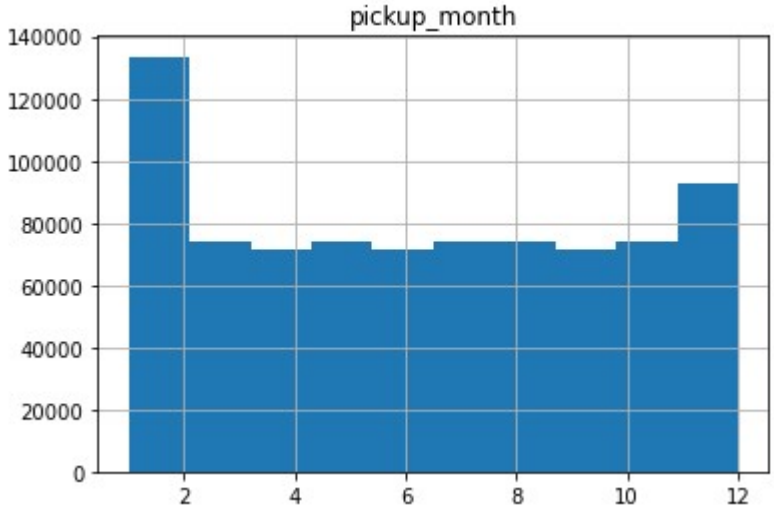
Começamos dividindo as datas em meses, para isso, foi transformado a coluna 'pickup_datetime', para o formato datetime para conseguirmos trabalhar com ela. Geramos uma nova coluna 'pickup_month', com a função .month do daytime, gerando uma coluna

com o mês, de 1 a 12, de acordo com o mês que foi realizado a corrida.

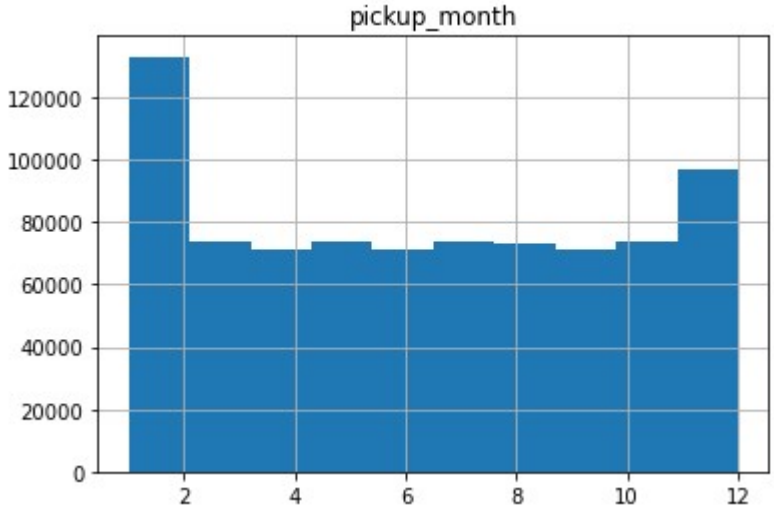
Depois disso, foi feito uma separação dos dados das corridas que foram feitas em dinheiro, para isso, foi utilizado o csv de pagamentos, nele, podemos ver que pagamento em dinheiro pode ser Cas, CAS, Cas, CASH ou CSH. Deixando apenas os pagamentos em dinheiro, utilizamos a função .hist() com relação da coluna 'pickup_month', criada anteriormente.

A seguir estão os histogramas gerados, onde o eixo x é o mês, e o eixo y é o valor pago em dinheiro

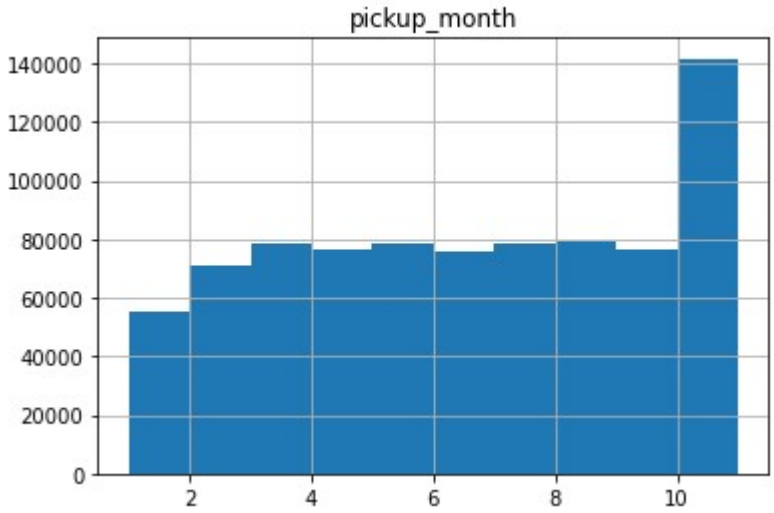
2009:



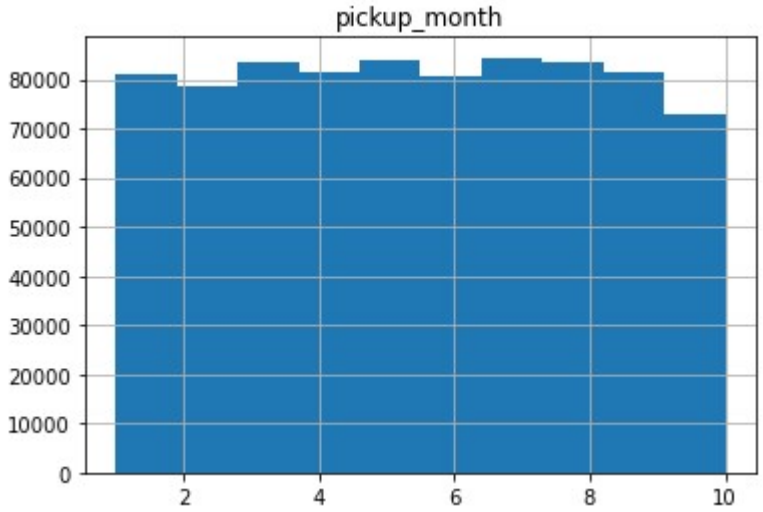
2010:



2011:



2012:

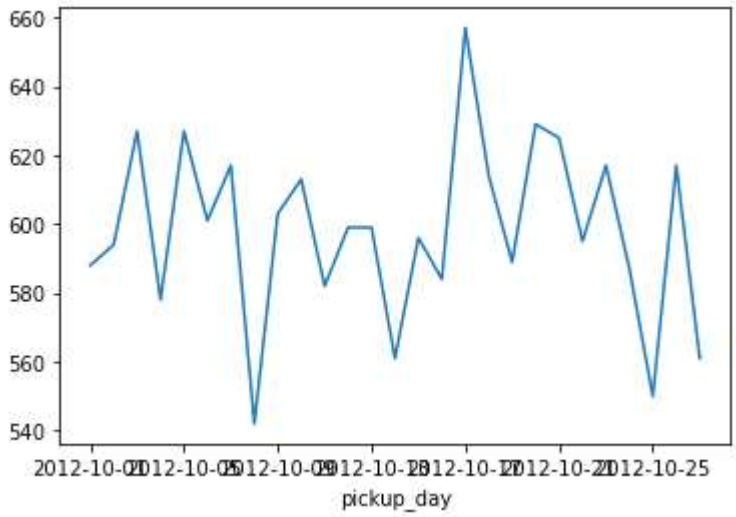


4. Faça um gráfico de série temporal contando a quantidade de gorjetas de cada dia, nos últimos 3 meses de 2012.

Foi feito o mesmo procedimento de dividir as datas em meses, porém, dessa vez, criando uma coluna com nome 'pickup_day', e utilizando a função date, nesse caso, ele gera uma nova coluna com a data sem as horas.

Depois separamos as corridas de 2012 que tiveram gorjeta, ou seja, a coluna 'tip_amount' maior do que 0 e também a coluna 'pickup_month' maior que 10, para pegar apenas os 3 últimos meses de 2012. Depois agrupamos pela coluna 'pickup_day', e utilizamos a função .count(), pois não queremos a soma das colunas, e sim a quantidade de gorjetas.

O gráfico gerado foi o seguinte:



Não foram feitos os bônus, pois eu avaliei, e não possuo conhecimento para realizá-los.