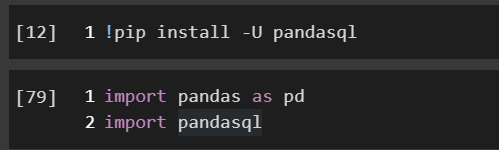
Limpeza dos dados

A tarefa desse módulo consiste em limpar alguns dados faltosos ou irrelevantes no dataframe.

Primeiro passo: Importar as bibliotecas:

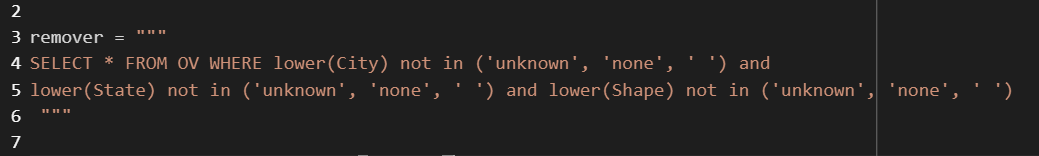


Segundo passo: Carregar o arquivo **OVNIS.csv** em um dataframe:



Em sequência, criei uma consulta SQL atribuída à uma variável chamada de **remover**. Nessa consulta removemos todos os tipos de dados inválidos das colunas **City, State e Shape**, como:

*unknow, none* e *espaços em branco.* É importante utilizar a função **lower** para deixar os dados em minúsculo tornando-os uniformes para alcançar todos os dados.



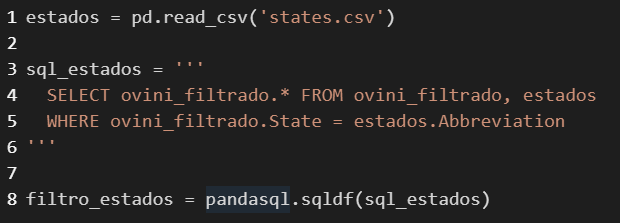
Após isso, atribuí essa consulta à uma nova variável chamada de **ovni\_filtrado:**



Feito isso, precisamos manter somente os registros referentes aos 51 estados citados no **states.csv**. Para isso, atribuí o **states.csv** à uma variável chamada **estados:**



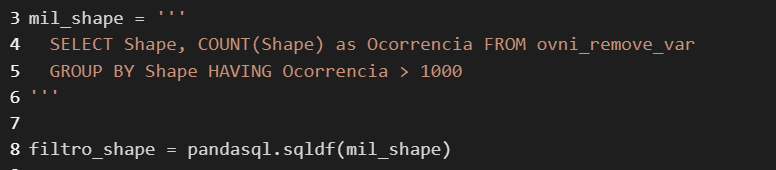
Em seguida, criei uma variável chamada **sql\_estados** para receber a consulta SQL que compara os estados do **states.csv** com o nosso dataframe e guarda somente os estados solicitados na variável **filtro\_estados:**

****

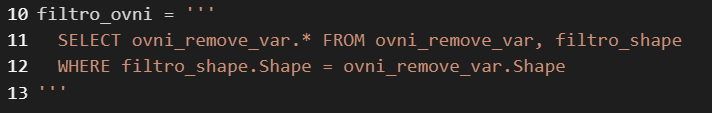
Em sequência removi as variáveis irrelevantes (**Duration, Summary e Posted)** e atribuindo a variável **óvni\_remove\_var:**

****

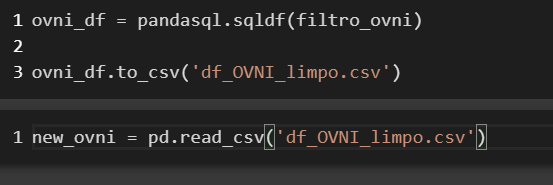
E para manter somente os registros de Shapes mais populares (com mais de 1000 ocorrências), foi criado mais uma consulta SQL chamada de **mil\_shape** e salvamos o resultado da consulta na variável **filtro\_shape:**

****

Em seguida, comparamos o resultado da consulta com nosso dataframe, para de fato manter somete os Shapes mais populares, salvando na variável **filtro\_ovni**:



E por ultimo, foi salvo o dataframe final em um arquivo CSV com o nome **"df\_OVNI\_limpo":**



<https://github.com/HugoCalisto/datascience>