**Nome:** Vanessa de Andrade Formiga

**Limpeza dos Dados**

**Problema:**

O problema proposto nesta terceira etapa do trabalho, partiu da necessidade que precisava fazer uma limpeza dos dados retirando campos ou colunas, que não precisava conter no arquivo. Teve como solução encontrada a utilização de consultas feitas através de comandos sqls.

**Solução:**

A solução encontrada para solucionar este problema, ocorreu com a divisão das tarefas em partes menores, que encontrasse descritas a seguir. Para realizar este trabalho precisou fazer a instalação da biblioteca pandasql, necessária para fazer as consultas. Na primeira imagem descreve como foram feitas as importações e a instalação.

**Uma imagem contendo objeto

Descrição gerada automaticamente**

A segunda imagem mostra o resultado que foi obtido após realizar a instalação da biblioteca.

Uma imagem contendo captura de tela

Descrição gerada automaticamente

Neste trabalho para realizar a primeira etapa, ocorreu com o carregamento do arquivo em um dataframe. Para carregar esse arquivo usou a variável denominada df e usou o read\_csv para fazer a leitura do arquivo Ovnis.csv. E a terceira imagem está mostrando como foram feitos este carregamento.



O resultado desse comando teve como saída a quarta imagem mostrada abaixo.



A quinta imagem mostra como realizou o carregamento do arquivo States.csv, que contém os estados americanos e as suas abreviações. Para separar as colunas utilizou-se a função sep, que serve para separar as colunas.

****

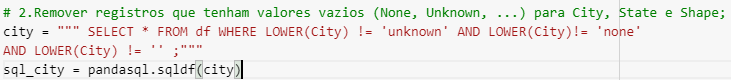
Nesta sexta imagem, mostra a saída do carregamento do arquivo States.csv.

Uma imagem contendo texto, recibo, captura de tela

Descrição gerada automaticamente

Foram realizadas nesta segunda etapa a remoção dos registros das colunas City, State e Shape que contém valores None, Unknown ou estejam vazias.

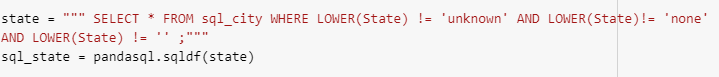
A sétima imagem, exibe como realizou-se a remoção na coluna City. Para isso utilizou-se a variável city que realiza uma consulta completa, selecionando todos os campos da tabela, localizado na variável denominada df, onde a cidade for diferente de “Unknown”, “None” e vazio. E utilizou o Lower, porque pode ter campos maiúsculos como minúsculos.

****

A oitava imagem imprime na tela o resultado da limpeza em City.

****

A nona imagem resulta na seleção de todas as colunas da query obtida no comando anterior, onde o estado seja diferente de Unknown, None e Vazio.

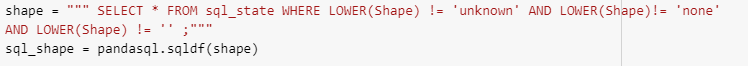
****

A décima imagem mostra a resposta realizada para fazer a limpeza da coluna State.

**Uma imagem contendo captura de tela

Descrição gerada automaticamente**

A décima primeira imagem demostrar, como realizou a seleção de todas as colunas, da consulta gerada anteriormente, onde o Shape for diferente de None, Unknown e vazio.

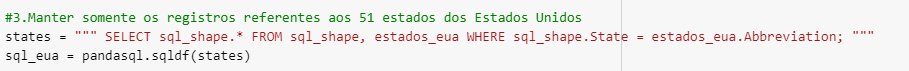
****

A décima segunda imagem exibe o resulta da limpeza em Shape.

**Uma imagem contendo captura de tela

Descrição gerada automaticamente**

Na terceira etapa precisou realizar uma limpeza, onde seriam mantidos somente os estados americanos. Retirando estados que estão em outros países. Nesta décima terceira imagem pode fazer a visualização de como realizado está consulta.

****

A décima quarta imagem exibe o resultado da remoção dos estados que não fazem parte dos estados americanos.

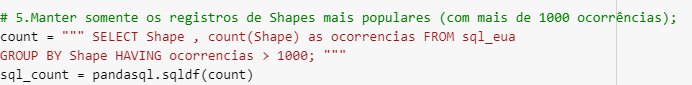
**Uma imagem contendo texto

Descrição gerada automaticamente**

Na quarta etapa realizou a extração das colunas dos dataframe, que não eram utilizadas no momento das consultas. Para fazer está remoção utilizou a variável utilizada na consulta anterior, fazendo um drop nas colunas Duration, Summary e Posted e Unnamed, utiliza-se o axis porque que fazer um drop na coluna que sejam verdadeiros. A décima quinta imagem, mostra essa operação foi realizada.

****

Na quinta etapa precisava fazer uma limpeza que obtivesse como resultado os shapes mais como populares que estivesse nos registros com mais 1000 ocorrências que estes deveriam ser mantidos. Para fazer essa query, utilizou o select que seleciona todos shapes e depois conta estes shapes, colocando o apelido nele com o nome ocorrências do comando sql anterior, agrupando por shapes que possui ocorrências com mais de 1000. A consulta foi realizada que será exibida na décima sexta imagem exibida a seguir.

****

O resultado está sendo imprimido na décima sétima imagem.

****

Depois de fazer a consulta dos shapes, precisou outra consulta de todos os campos de sql\_eua, das variáveis sql\_eua e sql\_count, onde o sql\_count.Shape é igual a sql\_eua.Shape. A décima oitava imagem, mostra como a consulta ocorreu.

**Uma imagem contendo objeto

Descrição gerada automaticamente**

A décima nona imagem demostra o resultado da consulta.

****

Na última etapa a tarefa que deveria ser feita era salvar em um novo arquivo do formato.csv colocando o seu nome como “df\_OVNI\_limpo”, o resultado obtido depois de fazer todas essas consultas necessárias. Fez o uso do resultado obtido da etapa anterior e passando to\_csv que é para escrever um novo arquivo passando o nome “df\_OVNI\_limpo” e a sua extensão. Como isso aconteceu e mostrado na vigésima imagem abaixo.

****

**Link para o Github:**

<https://github.com/vanessaformiga/trabalho-de-ci-ncia-de-dados/tree/master/limpeza%20dos%20Dados>

**Referências:**

**Pandas:**

<https://pandas.pydata.org/pandas-docs/stable/>

**Drop:**

<https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.drop.html?highlight=drop#pandas.DataFrame.drop>

**Colab:**

<https://colab.research.google.com>