Como treinamento para a disciplina de Visualização de Dados do curso de especialização em Ciência de Dados e BIG Data da Universidade PUC Minas, resolvi fazer um estudo sobre a distribuição de veículos, seus tipos em todo país.

Para este estudo, estou fazendo uso da base de dados do portal do DENATRAN consolidado para dezembro de 2019. Nenhum motivo especial para escolher esta base de dados, foi uma decisão aleatória.

Para este artigo utilizei as seguintes bases de dados:

* Frota por UF e Tipo de Veículo
* Frota por Município e Tipo

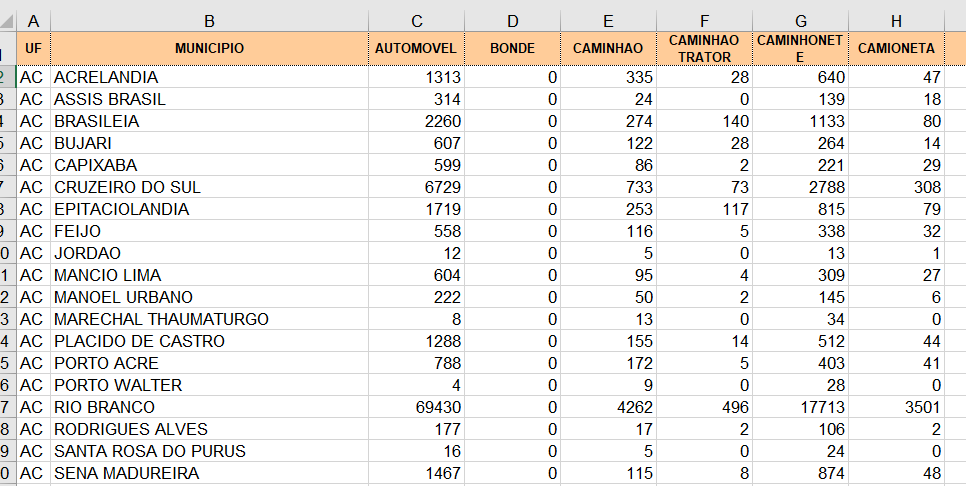
Vamos tentar responder 5 questões, que numa primeira visão podem parecer bastante simples, acerca da distribuição da frota de veículos no país:

* Qual a quantidade total de veículos registrados pelo DENATRAN?
* Qual o tamanho da frota por estado em todo País?
* Ranking dos 5 municípios com maior frota.
* Comparativo de quantidade de veículos por habitantes.
* Municípios com maior população e menor quantidade de veículos.

Partindo deste ponto, vamos preparar as nossas bases de dados para tentar responder a estes questionamentos. Nesta fase, eu ainda não havia me decidido por quais ferramentas utilizar. Por enquanto foco na origem dos dados.

A primeira fonte de dados é um arquivo no formato XLS (Pasta de trabalho do Excel) e possui a seguinte estrutura de campos:

|  |  |
| --- | --- |
| UF | Sigla do Estado |
| MUNICIPIO | Nome completo do município |
| AUTOMOVEL | Quantidade para cada tipo de veículo registrado |
| BONDE, CAMINHAO,TRATOR, CAMINHONETE, CAMIONETA, CHASSI PLATAF, CICLOMOTOR,... | |
| TOTAL | Total de veículos para um Município |

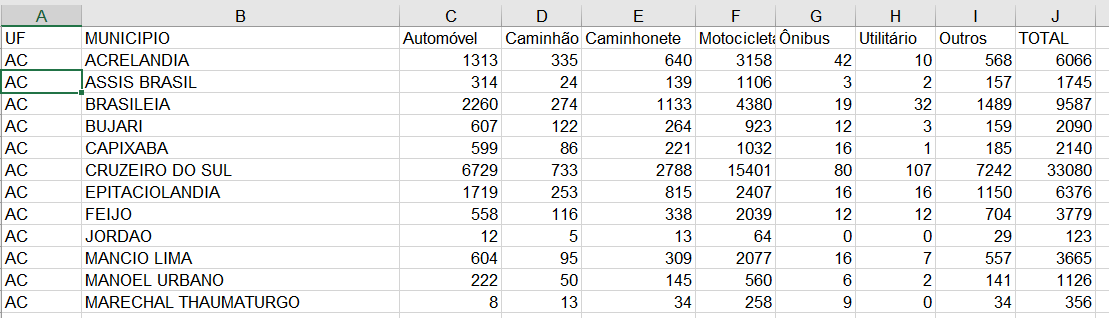


Entre a identificação do nome do Município e o total existem 22 colunas para comportar os totais de cada categoria de veículo tratada pelo órgão. Acredito que não interessa para este estudo quantificar, por exemplo quantos tratores de esteira foram registrados ou quantos veículos “Chassi Plataforma” existem. Por isso, vamos transformar esses dados priorizando um grupo de veículos mais comuns e consolidando os demais em uma coluna que chamaremos de “OUTROS”.

* Automóvel
* Caminhão
* Caminhonete
* Motocicleta
* Ônibus
* Utilitário
* Outros (consolidado de todos os demais veículos)

O primeiro desafio no tratamento da base de dados será consolidar todas as demais colunas em uma única coluna “Outros”. Não chega a ser exatamente um desafio porque utilizei os recursos do Excel. Eu realmente poderia ter feito isso na própria ferramenta onde vou criar a apresentação, mas decidi deixar a base original o mais perfeita possível antes da importação.

Sendo assim, após o tratamento no Excel, a base de dados ficou assim:



Informação consolidada e pronta para ser manipulada.

Utilizando o Power BI

Decidi contar minha análise utilizando o software Microsoft Power BI. Eu poderia utilizar o Tableau que foi apresentado em meu curso ou até mesmo o Excel, onde eu tenho uma maior facilidade de trabalhar. Pode ser até mesmo que eu experimente os outros dois em uma oportunidade futura. Mas neste momento vamos aprender a trabalhar com a ferramenta da Microsoft.

Base de origem pesada

A primeira coisa que pude notar é que uma base de dados com 5.570 demorou um pouco para carregar. Eu acabei fazendo algumas transformações na base e gerando outras tabelas. Uma solução bem simples foi exportar apenas a planilha final para um arquivo do tipo CSV (Comma Separated Values). O formato texto carregou rapidamente (eu diria até instantaneamente).

Analisando a base de dados importada eu percebo que posso responder 4 das 5 perguntas iniciais. Porém não tenho dado para analisar o questionamento:

* Comparativo de quantidade de veículos por habitantes.

Me falta aqui exatamente a quantidade de habitantes por cidade com base no levantamento mais próximo a data da nossa base de dados (dezembro de 2019).

Eu encontrei uma base de informações na wikipedia



Eu sei que preciso do município e da população. Esta página parece ter tudo isso e eu posso tanto importar esses dados no Power BI quanto trata-los no Excel e depois exportar os dados preparados.

Novo problema, chave fraca

Porém agora eu tenho um novo problema. A única forma de ligar essa tabela da Wikipedia com a tabela que eu baixei do DENATRAN é utilizado o nome do município. Mas sei também que podemos ter municípios com mesmo nome no Brasil e até problemas com a grafia do nome do município. A base dos veículos possui uma coluna com a sigla do estado e esta aqui possui o nome do estado por extenso.

A solução perfeita seria utilizar o identificador que o IBGE dá para cada município, mas infelizmente a base consolidada dos veículos não a possui.

Acontece que nesta mesma página da Wikipédia existe um link para a fonte de referência no site do IBGE. Esta fonte de referência é um arquivo em formato PDF que possui algumas tabelas e uma delas parece ser bastante promissora, possui o nome do município e até a sigla do estado, além da população estimada que é o que queremos para responder ao questionamento.



Importar uma tabela de um arquivo PDF não é tarefa complicada para o Power BI ou para o Excel, que é onde estou preparando minha base de dados de estudo.

Validando se os dados

Eu preciso que os nomes dos estados e das cidades sejam iguais nas duas tabelas de origens são a minha ligação entre elas não poderá funcionar. Ou seja, se o município “Cachoeiro de Itapemirim” na tabela do DENATRAN estiver grafado desta forma e na tabela da quantidade populacional estiver grafado “Cachoeiro **do** Itapemirim”, esta sutil diferença representará um problema no processamento e análise daqueles dados.

Tratando os nomes acentuados

Uma comparação superficial nas duas bases já me revela uma diferença importante, a base do DENATRAN não possui caracteres de acentos e cedilha enquanto a base do IBGE as possui. Mesmo sabendo que o acento e a cedilha são parte da grafia completa do nome do município, é mais simples remover o acento do que colocar acento na base que não possui.

Meu primeiro tratamento foi remover todos os acentos nos nomes das cidades. Usei uma função em VBA que eu mesmo criei no Excel para agilizar, mas poderia ter feito o mesmo utilizando o DAX no Power BI, a lógica é a mesma com algumas adaptações. Mas este é assunto para um outro artigo.

Minha análise pode passar bem sem os acentos desta vez, mas para o futuro preciso rever essa origem de dados.

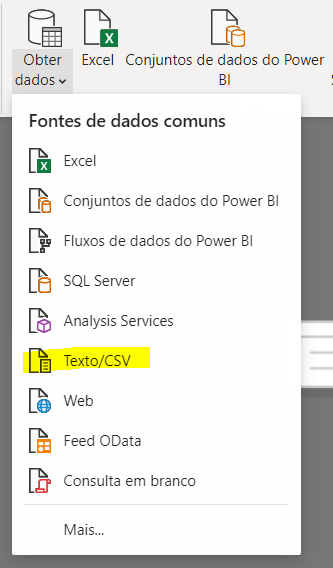
Tratando municípios com grafia diferente

Para garantir que os nomes dos municípios são iguais nas duas tabelas, utilizei o Excel para fazer as comparações e me deparei com uma grande disparidade na grafia nos seus nomes.

Esta fase tomou bastante tempo porque exigiu bastante atenção e pesquisa. Afinal são mais de 5 mil municípios. E encontrei diferenças bem interessantes, como nomes grafados com S em uma base e com Z em outra, muitos, como PORTO DE PEDRAS no Alagoas aparecia como PORTA DA PEDRA em uma das bases, o que valeu uma verificada no site da prefeitura para decidir qual das duas deveria prevalecer. E o correto é LAJEDO DO TABOCAL ou LAGEDO DO TABOCAL? No site oficial do município está grafado com J e é esta fonte que adotar. No fim eu fiquei agradecido por serem somente pouco mais de 5 mil municípios.

Importação da base de dados para o Power BI

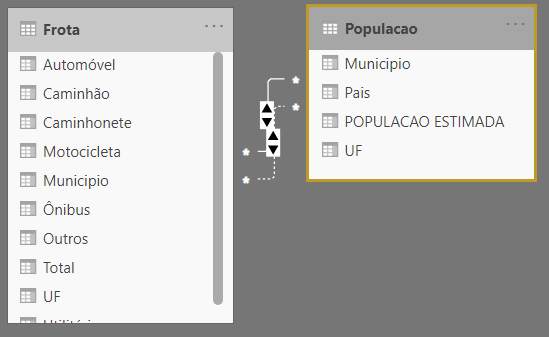
Esta parte é bastante simples e o recurso é bastante intuitivo na ferramenta. A importação das duas bases, agora em formato CSV, foi simples e rápida.



A funcionalidade de importação de arquivos CSV

Para quem está acostumado a trabalhar com banco de dados, a visão é bem familiar. As duas bases de origens são exibidas com entidades contendo seus campos. Porém não foi criado automaticamente nenhum relacionamento. O Power BI espera que você diga qual a relação há entre as duas estruturas.

Definir o relacionamento é também bastante intuitivo. Bastou arrastar o nome do campo de uma base para o correspondente na base do lado. Eu havia tomado o cuidado de deixar esses dois nomes com a mesma grafia.



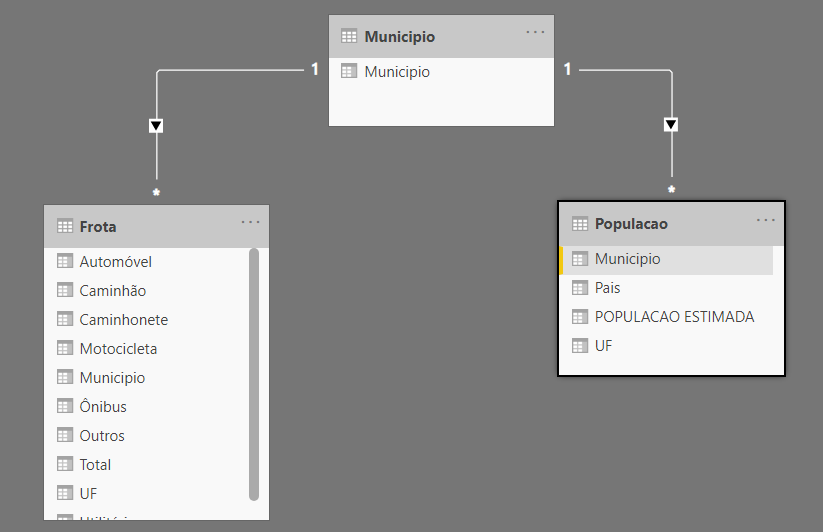
Tabelas Fato x Dimensão

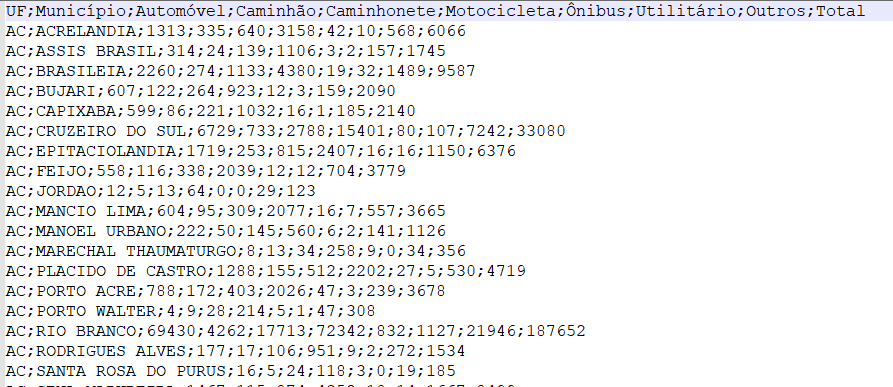
Da forma como está, para o nosso modelo, a nossa análise pode realmente funcionar como essa duas bases de dados porque existe exatamente a mesma quantidade de registros nas duas tabelas. Em outras palavras a tabela da Frota possui as totalizações das quantidades de veículos para cada um dos 5.570 municípios. O mesmo ocorre com a tabela população onde teríamos um registro com o total populacional para cada um deles.

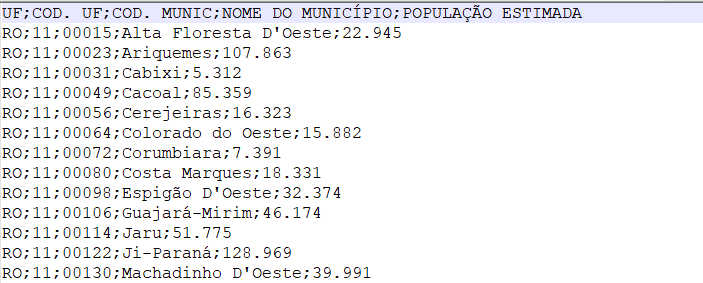
Porém tanto a tabela de frota, quanto a de população serão usadas aqui nesta análise como tabelas de “Fato”. As tabelas de fato armazenam as medidas utilizadas durante as análises. A “Frota” traz a quantidade de veículos por cidade e a “Populacao” a estimativa populacional.

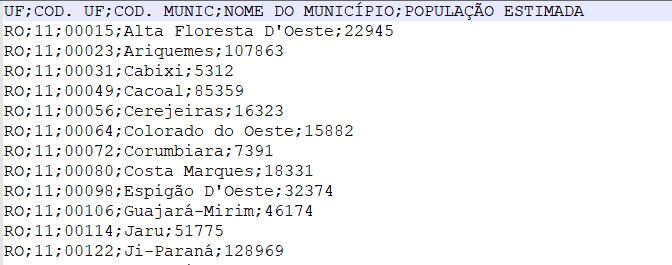
Preciso criar uma tabela de Dimensão. As tabelas de Dimensão representam o domínio da informação, que no caso desta análise, são os municípios. As tabelas de Dimensão não possuem dados que se repetem.

Eu poderia gerar uma nova base de informações mais completa contendo, Código do IBGE e nome do Município, mas deixaria para um trabalho futuro. O que farei aqui é gerar no Power BI uma nova tabela contendo somente registros dos nomes dos municípios.











(<https://antigo.infraestrutura.gov.br/component/content/article/115-portal-denatran/8559-frota-de-veiculos-2019.html>)

<https://pt.wikipedia.org/wiki/Lista_de_munic%C3%ADpios_do_Brasil_por_popula%C3%A7%C3%A3o>