**Base de dados**

Baixei um csv, da sessão série histórica mensal a partir de 2013(MENSAL\_ESTADOS-DESDE\_Fev2013), do site: <http://www.anp.gov.br/precos-e-defesa-da-concorrencia/precos/levantamento-de-precos/serie-historica-do-levantamento-de-precos-e-de-margens-de-comercializacao-de-combustiveis>

**Seleção, Pré-Processamento, Transformação, Mineração**

Usei o *profile\_report* do pandas para obter um relatório detalhado dos dados, e pude constatar uma grande correlação entre as colunas *PREÇO MÉDIO REVENDA, DESVIO PADRÃO REVENDA, PREÇO MÍNIMO REVENDA,* e *PREÇO MÁXIMO REVENDA*. Foram encontrados alguns atributos com valores ausentes ou zerados, porém não relevantes. Foi preciso converter o tipo de algumas colunas que estavam como *OBJECT* para *FLOAT/INT*, a fim de facilitar o tratamento e enxergar novas correlações. Após esse procedimento, usei o *profile\_report* novamente, e encontrei novas correlações entre os campos de valores max, min, e médio de revenda e distribuição. Após analisar os dados visualmente, constatei que esses não eram relevantes para conseguir visualizar os dados dos valores médios dos combustíveis nos estados/regiões. O atributo *MÊS* é concatenado com o *ANO*(mês/ano) e foi dividido em suas respectivas colunas(MES, ANO). Ao final a quantidade de atributos foi reduzida de 17 para 8, com 3 numéricos e 5 categóricos.

**Resultado**

Desde 2013, as regiões Norte e Nordeste são as que possuem em média os combustíveis, ETANOL e GASOLINA, mais caros do país, enquanto a SUL possui os menos caros. Da região Nordeste, Ceará e Alagoas possuem os maiores preços, enquanto Paraíba e Pernambuco os menores desde 2017.