2 - Limpeza de dados

O dataset `Air Quality NYC` possui os índices de gases emitidos na cidade de NY e suas principais fontes. Também possui dados hospitalares causados por esses poluentes, tráfegos de caminhões e carros. Ha duas separações de dados para os indicadores, aqueles medidos sazonalemente (verão e inverno), e aqueles medidos anualmente:

Sazonais (Verão/Inverno):

Indicator ID	Name
375	Nitrogen dioxide (NO2)
386	Ozonone (O3)
365	Fine particles (PM 2.5)

Anuais:

Indicator ID	Name
647	Outdoor Air Toxics - Formaldehyde
646	Outdoor Air Toxics - Benzene
651	Cardiovascular hospitalizations due to PM2.5 (age 40+)
652	Cardiac and respiratory deaths due to Ozone
650	Respiratory hospitalizations due to PM2.5 (age 20+)
659	Asthma emergency departments visits due to Ozone
661	Asthma hospitalizations due to Ozone
657	Asthma emergency department visits due to PM2.5
639	Deaths due to PM2.5
653	Asthma emergency departments visits due to Ozone
655	Asthma hospitalizations due to Ozone
648	Asthma emergency department visits due to PM2.5
644	Annual vehicle miles traveled (cars)
645	Annual vehicle miles traveled (trucks)
643	Annual vehicle miles traveled
642	Boiler Emissions- Total NOx Emissions
641	Boiler Emissions- Total PM2.5 Emissions
640	Boiler Emissions- Total SO2 Emissions

Estratégias para limpeza de dados

Missing values: Não é todo indicador que possui dado de todos os anos (2005-2023). Dados sazonais possuem de 2008-2023, e indicadores anuais possuem de 2005-2019. Infelizmente não temos o que fazer, alguns dados são medidas de um período maior (2012-2014), apenas distribuimos a mesma média para os anos 2012, 2013 e 2014.

Outliers: Não identificamos nenhum outlier no dataset;

Inconsistência: Não identificamos inconsistências no dataset;

Padronização: Os indicadores sazonais terão a coluna inicial `Time Period` quebrada em duas: `Season` e `Year`. Como o inverno começa em um ano e termina em outro, consideraremos o ano em que ele começa para manter a padronização. Os indicadores anuais possuem dois formatos: YYYY e YYYY-YYYY. Os anos que estão em um intervalo, separamos e copiamos a coluna `Data Value` para todos o intervalo.

Os dados foram coletados em diferentes regiões da cidade, mas para o nosso uso, agruparemos apenas na cidade. A coluna `Geo Type Name` diz qual a granularidade da informação. Alguns são agregados de outras regiões, `Citywide` por exemplo, representa a média para a cidade como um todo, já a média dos 5 `Borough` é o equivalente à `Citywide`, e assim por diante. Por esse motivo, utilizaremos apenas os registros de `Citywide`.