Air Quality in São Paulo

Dataset:

O dataset contém medições de poluentes realizadas de hora em hora pela CETESB em algumas estações de monitoramento de qualidade do ar no estado de São Paulo (Brasil), abrangendo o período de 5 de agosto de 2013 a 9 de setembro de 2020.

Disponível em: https://www.kaggle.com/datasets/amandalk/sp-air-quality

Objetivo:

O objetivo geral deste projeto é analisar a variação do nível de dióxido de nitrogênio (NO2), medido em microgramas por metro cúbico (µg/m³), presente no ar da cidade de São Paulo ao longo do dia. Entre os objetivos específicos estão a identificação das principais causas da concentração desse poluente, a análise de seus padrões de comportamento e o estudo das particularidades relacionadas a cada região da cidade.

Linguagem e Bibliotecas:

O projeto foi desenvolvido utilizando a linguagem de programação Python. Para a visualização dos dados, foram empregadas as bibliotecas Matplotlib e Seaborn, conhecidas por sua flexibilidade e capacidade de gerar gráficos informativos e esteticamente agradáveis. A manipulação de dados tabulares foi realizada com a biblioteca Pandas, amplamente utilizada devido à sua eficiência e versatilidade. Por fim, para lidar com dados no formato datetime, foi utilizada a biblioteca Datetime, que oferece ferramentas robustas para a manipulação e formatação de dados temporais.

Análise Exploratória:

O dataset é composto por 11 colunas divididas em três categorias principais: horário de coleta, cidade e poluentes. A coluna que representa o horário de coleta é denominada Datetime e apresenta os dados no formato internacional ISO 8601 (‘YYYY-MM-DD HH:MM:SS’), indicando a data e a hora de cada medição realizada. A localização das estações de monitoramento é identificada pela coluna Station, que é composta por 59 estações diferentes. As colunas restantes representam diferentes poluentes monitorados, sendo eles: ‘Benzene’, ‘CO’, ‘PM10’, ‘PM2.5’, ‘NO2’, ‘O3’,

‘SO2’, ‘Toluene’ e ‘TRS’. Os valores registrados estão expressos em microgramas por metro cúbico (µg/m³).

Ademais, foi constatada a presença de valores ausentes (NaN) em algumas colunas de poluentes, incluindo a coluna ‘NO2’. Para minimizar os efeitos desses dados ausentes sobre as análises, avaliou-se a possibilidade de exclusão dos registros incompletos. No entanto, como os valores ausentes representavam mais de 5% do total de registros, decidiu-se pela imputação dos dados faltantes. A imputação foi realizada substituindo os valores ausentes de ‘NO2’ pela mediana dos valores dessa mesma variável para cada estação de coleta identificada na coluna ‘Station’. A escolha da mediana como método de imputação dos valores ausentes na coluna NO2 foi baseada em sua robustez frente a valores extremos (outliers). Após essa etapa, os outliers foram identificados considerando os grupos da coluna ‘Station’, utilizando o intervalo interquartil (IQR) para cada estação de coleta. Os limites, inferior (Q1 – 1,5 x IQR) e superior (Q3 + 1,5 x IQR), foram calculados separadamente para cada grupo, garantindo uma identificação mais precisa dos valores extremos, que foram posteriormente removidos da análise.

Figura - print(air_quality.head(3))

	Datetime	Station	Benzene	CO	PM10	PM2.5	NO2	O3	SO2	Toluene	TRS
0	2013-05-08 01:00:00	Araçatuba	NaN	NaN	30.0	NaN	NaN	7.0	NaN	NaN	NaN
1	2013-05-08 02:00:00	Araçatuba	NaN	NaN	30.0	NaN	NaN	6.0	NaN	NaN	NaN
2	2013-05-08 03:00:00	Araçatuba	NaN	NaN	31.0	NaN	NaN	6.0	NaN	NaN	NaN

Fonte: Air Quality in São Paulo.py

Manipulação dos Dados:

Após a análise exploratória dos dados, a etapa de manipulação foi realizada utilizando as bibliotecas Pandas para manipulação de dados tabulares, Datetime para trabalhar com informações temporais e Matplotlib/Seaborn para a visualização dos dados. Primeiramente, foi criada uma nova coluna chamada Hour, a partir da qual foram extraídos os horários de coleta registrados na coluna Datetime. Essa transformação foi realizada utilizando as funcionalidades do módulo Datetime, permitindo a conversão e o acesso eficiente aos elementos de tempo.

Em seguida, uma tabela adicional foi criada a partir de dois dicionários. O primeiro foi construído a partir dos valores únicos das estações de coleta encontrados na

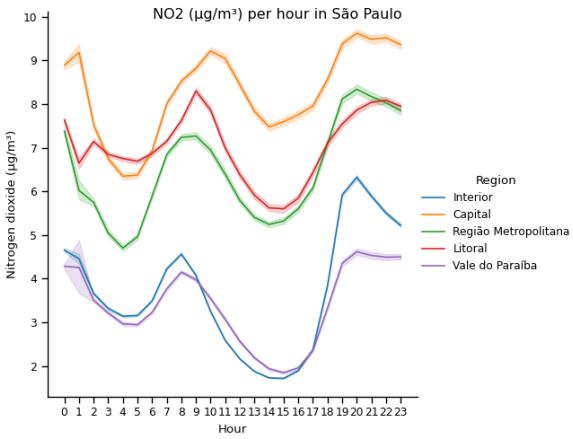
coluna Station do DataFrame original. O segundo dicionário categorizou essas estações com base em suas localizações geográficas, atribuindo-as a cinco regiões do Estado de São Paulo: Capital, Região Metropolitana, Interior, Vale do Paraíba e Litoral.

Posteriormente, utilizando as funcionalidades de junção da biblioteca Pandas, realizou-se uma operação de "merge" à esquerda entre o DataFrame original e a nova tabela. Isso permitiu a adição da coluna Region, com as respectivas classificações regionais, ao DataFrame principal. Esse procedimento enriqueceu os dados, incorporando informações geográficas úteis para análises posteriores.

Por fim, as bibliotecas Matplotlib e Seaborn foram empregadas para a visualização do DataFrame manipulado. Foi criado um gráfico relacional em linha, no qual a coluna Hour foi representada no eixo horizontal, enquanto os valores da coluna NO2 foram plotados no eixo vertical. O gráfico foi segmentado por região, permitindo a análise das tendências de concentração de dióxido de nitrogênio em diferentes localidades do Estado de São Paulo ao longo do tempo.

Análise dos Dados:

Gráfico - NO2 (µg/m³) per hour in São Paulo



Fonte: Air Quality in São Paulo.py

A análise dos dados revela um padrão diurno e noturno na concentração de dióxido de nitrogênio (NO_2) na atmosfera de São Paulo. Durante o período matutino, observa-se um pico significativo de concentração entre 6h e 10h, coincidindo com o aumento do tráfego veicular característico do horário de pico da manhã.

O segundo pico ocorre entre 17h e 20h, novamente associado ao tráfego intenso no final do expediente. Por outro lado, os períodos de menor concentração de NO_2

ocorrem na madrugada, entre 1h e 5h, e no início da tarde, entre 11h e 15h, reflexo de uma menor atividade veicular nesses horários e de uma maior dispersão dos poluentes devido ao aumento da temperatura e da radiação solar.

Quanto às diferenças regionais, as áreas da Região Metropolitana, Capital e Litoral apresentam as maiores concentrações de NO_2 ao longo do dia. Esse fenômeno pode ser atribuído à maior densidade populacional, ao tráfego intenso e às atividades industriais nessas regiões.

Em contrapartida, as regiões do Vale do Paraíba e do Interior registram as menores concentrações de NO_2, uma vez que possuem menor densidade populacional e atividades industriais menos intensas. Dessa forma, os dados corroboram as variações comportamentais do poluente em diferentes períodos do dia e regiões, atendendo aos objetivos de compreender as causas, padrões e particularidades das concentrações de NO_2.

Conclusão:

Os resultados indicam que o tráfego veicular é o principal responsável pela emissão de dióxido de nitrogênio em São Paulo, com influência direta nos picos de concentração. Além disso, as diferenças regionais evidenciam a necessidade de políticas locais direcionadas para mitigar os impactos do poluente, especialmente nas áreas mais afetadas, como a Região Metropolitana, Capital e Litoral. Nesse contexto, a adoção de combustíveis menos poluentes se mostra fundamental para a redução das emissões. Alternativas como veículos elétricos, que não emitem poluentes diretamente, e o estímulo ao uso de GNV (Gás Natural Veicular), que emite menos NO_2, podem contribuir significativamente para a melhoria da qualidade do ar. Adicionalmente, é essencial o incentivo ao transporte público sustentável, a promoção de biocombustíveis e a implementação de controle rigoroso das emissões industriais, buscando uma redução substancial da poluição atmosférica e o aprimoramento da saúde pública.