Curso de Ciência da Computação, Câmpus Palmas
Disciplina: Aprendizado de Máquina
Aluno: Romeu Miranda Borges
Data: 18/05/2025

1. A BASE DE DADOS

O nome da base de dados utilizada é *Metro Interstate Traffic Volume*. Consiste em dados sobre o volume de tráfego por hora do dia em um dado ponto de uma estrada interestadual dos Estados Unidos da América, entre as cidades de Minneapolis e St. Paul, ambas em Minnesota. Junto ao volume de tráfego, o dataset fornece informações sobre o clima e o feriado corrente no horário do registro. Pode ser acessada pelo link: https://archive.ics.uci.edu/dataset/492/metro+interstate+traffic+volume.

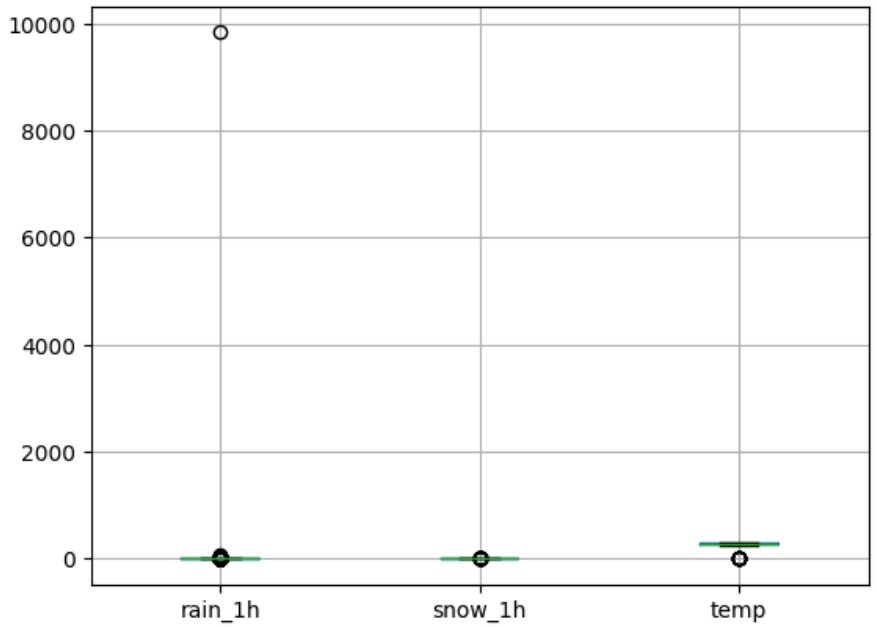
1.1. DICIONÁRIO DE DADOS

Nome da variável	**Descrição**	**Tipo**	**Unidade**	**Valores possíveis/Intervalo numérico**	**Observações**
holiday	Nome do feriado (nacional ou estadual)	Categórico *(string)*			
temp	Temperatura média	Contínuo	Kelvin		
rain_1h	Quantidade de chuva que ocorreu na hora do registro	Contínuo	Milímetros (mm)	0 a 9831,3	
snow_1h	Quantidade de neve que ocorreu na hora do registro	Contínuo	Milímetros (mm)	0 a 0,51	
clouds_all	Porcentagem de nebulosidade céu	Inteiro	Porcentagem (%)	0 - 100%	
weather_main	Descrição curta do tempo na hora do registro	Categórico *(string)*			

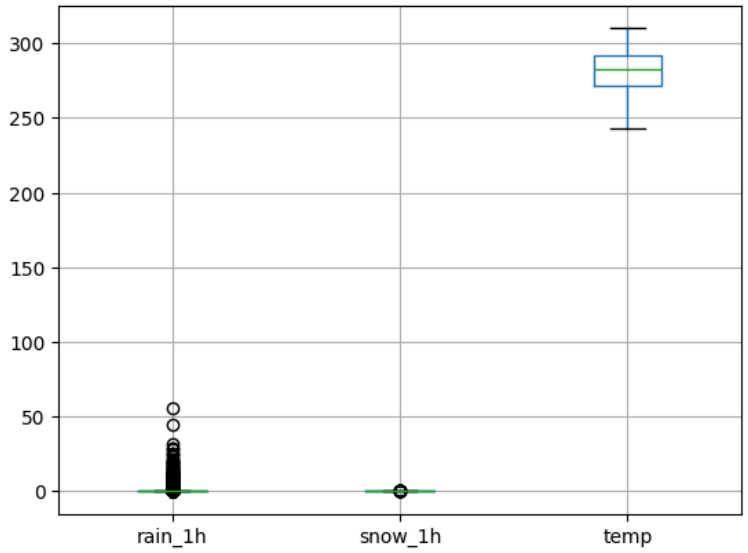
weather_description	Descrição pouco mais extensa do tempo na hora do registro	Categórico *(string)*			
date_time	Data e hora do registro	Data e hora		2012-10-02 09:00:00 a 2018-09-30 23:00:00	Valores faltantes entre o mês 8 de 2014 e o mês 6 de 2015
traffic_volume	Volume de tráfego no sentido oeste	Inteiro		0 a 7280	

2. REMOÇÃO DE OUTLIERS

Com os dados explorados, observou-se que, a princípio, não seria necessária a aplicação de um método de remoção de outliers que envolva cálculo do desvio padrão de uma coluna, por exemplo. Tomemos o boxplot a seguir:



Notamos dois valores destoantes: um registro com temperatura 0 (em Kelvin) e outro indicando quase 10 mil milímetros de chuva. Tendo sido removidos os valores através de filtragem simples, obtemos:



Para a coluna *rain_1h*, observamos muitos pontos como potenciais outliers. Porém, é interessante observar que uma grande porção do conjunto de dados possui 0 como valor para *rain_1h*, o que faz com que medidas razoáveis sejam indicadas como outliers nesse gráfico. Tendo observado isso, tais registros foram mantidos.

Em sequência, foi aplicada uma filtragem nas colunas *weather_main* e *weather_description*, para remover classificações com quantidade desprezível de registros, ou com inconsistência nos dados.

3. CONCLUSÃO

Foram identificados duas colunas com outliers, *rain_1h* e *temp*. Os dados destoantes foram removidos. Numa tentativa de aprofundar o tratamento da base, foi feita uma análise para remover classes irrelevantes ou cuja descrição não condizia com os valores das colunas numéricas. Nesta segunda filtragem, as colunas afetadas foram *weather_main* e *weather_description*. O conjunto original possui 48204 linhas, e o final, 40528. Foram removidos, portanto, 7676 registros.

Os detalhes da implementação estão disponíveis em https://github.com/romeuborges19/trabalho-ml/blob/main/trabalho-01.ipynb.