# Data set description

This document describes the steps for the application of NMF from receiving the data in “VOC sampling window data 20240601’.

Merged dataset ("../DataProcessing/Trailer\_hourly\_merge\_20240905.rds"):

This dataset contains the trailer data, well data, and flare data.

Trailer data (“../DataProcessing/TrailerProcessed-20240601.rds”)

Trailer data from 2023-04-15 00:05:00 to 2024-06-01 05:05:00 received in “VOC sampling window data 20240601’.

Well data (“../DataProcessing/WellsProcessed0301.rds"):

Contains the monthly oil and gas production data within 1km of the trailer.

Flare data (“../DataProcessing/FlaringProcessed.rds"):

Filters for flares where temp\_bb >= 1600 and is within a 100km radius from the trailer. Also includes average flare characteristic variables.

# NMF analysis (“NMF\_voc\_norm.rmd”)

Five component NMF without Ozone

1. Remove hourly observation with missing observation for any chemical

2. Remove background noise level using min values (except for chemicals with minimum value < 2\*LOD and maximum value > 100\*LOD)

3. Zero values are converted to a random value between 0 and 0.5\*LOD

(non-VOC LOD based on [LNM LOD and Uncertainty Table.xlsx - Google Sheets](https://docs.google.com/spreadsheets/d/1cL9QzJVxVQu9c-2kOuEO3Prrfh-t_89G/edit?gid=501798226#gid=501798226))

(VOC LOD based on mean LDL in '../data/LNM\_VOC\_Uncertainties.xlsx')

4. Normalize using min & max

~~5. Compute weight matrix according to Guha's paper, without LOQ~~ (Not required for svd seed)

6. Remove ozone (Step to do this wouldn’t affect # of observations)

7. run NMF using ` nmf(normalized\_matrix\_less\_o3, rank = 5, method = "KL", seed='nndsvd')`