ANWB Pre-processing (all modifications)

- **Queried the data**
- **Converted the fetched data into a DataFrame**
- **Converted the `event_start` column to a datetime format, extracted the date from it, and used it as an index**
- **Created a new feature `avg_speed` using `speed_kmh` and `end_speed_kmh`**
- **Converted the `maaxwarde` column from `float64` to `int64` because exact measurements are not needed for the project goal**
- **Removed columns**:
  - `event_end` (as long as we have the event duration information, this column is unnecessary, especially since the `event_start` column has been converted and renamed to only have the date)
  - `municipality_name` (municipality is Breda)
  - `latitude` (we want the general location)
  - `longitude`
  - `speed_kmh` (we want the average speed)
  - `end_speed_kmh`
  - `road_manager_name` (not needed for our project goal)
  - `is_valid` (after using it to filter the data)
- **Sorted the DataFrame by year (descending)**
- **Created a new feature to classify the incident types based on the `incident_severity` column with the categories: low risk, medium risk, and high risk**
- **Removed incident severity types that have a low number of samples**
- **Changed the data type of `incident_severity` from `varchar` to `string`**
- **Removed outliers in the `duration_seconds` column**
- **Removed outliers in the `avg_speed` column**
- **Plotted the average speed of the roads with the highest number of outliers**
- **Plotted the average duration of the outlier incidents by road names**
- **Normalized the `duration_seconds` column using log transformation**
- **Normalized the `avg_speed` column using a RobustScaler**
- **Transformed the numerical features using a StandardScaler**
- **One-hot encoded categorical features**

Findings:

1. Most common incident types were speedings.
2. The road with the highest number of incidents is ‘Franklin Rooseveltlaan’ - 73053 (most incident types are speed - 31846).
3. February 2023 was the month with the highest number of incidents.

KNMI Pre-processing

- **Removed unnecessary columns**: `[NAME, latitude, longitude, dr_pws_10, ww_cor_10, ri_pws_10]`
- **Split the `dtg` column into date and time columns**
- **Filtered dates between 01-01-2018 and 29-02-2024**
- **Removed missing values**
- **Filled missing values with the mean, median, mode, or a fixed value (e.g., 0)**
- **Converted object columns and interpolated missing values**
- **Identified outliers using the Z-scores method for the `dr_regenm_10` and `ri_regenm_10` columns**
- **Transformed data using log transformation and square root transformation**
- **Split the dataset into training, validation, and test sets**
- **Created a new feature that determines the risk level of the rain (Low-Mid-High), using the rain intensity and duration columns**
- **Created visuals showing the distribution of risk levels and risk levels over time**