ANWB pre-processing (all modifications)

* Querried 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 as an index
* Created new feature ‘avg\_speed’ using ‘speed\_kmh‘ and ‘end\_speed\_kmh’
* Converted ‘maaxwarde’ column from float64 to int64 because we do not need the exact measurements for the project goal
* Columns removed:
  + Event\_end (as long as we have the event duration information, we do not need this column, especially since the event\_start column has been converted and renamed so as 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 low number of samples
* Changed 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.