traffic data

Number of rows before cleaning 5050
Number of rows after cleaning 2951

Summary of Fixes Applied

remove 50 duplicate rows

replace missing traffic id with index +1
The dataset contained multiple inconsistent date formats such as:

- `15/01/2024 8AM`
- `2024-01-15T08:00Z`
- Invalid entries like `TBD` or `2099-00-00 99:99`

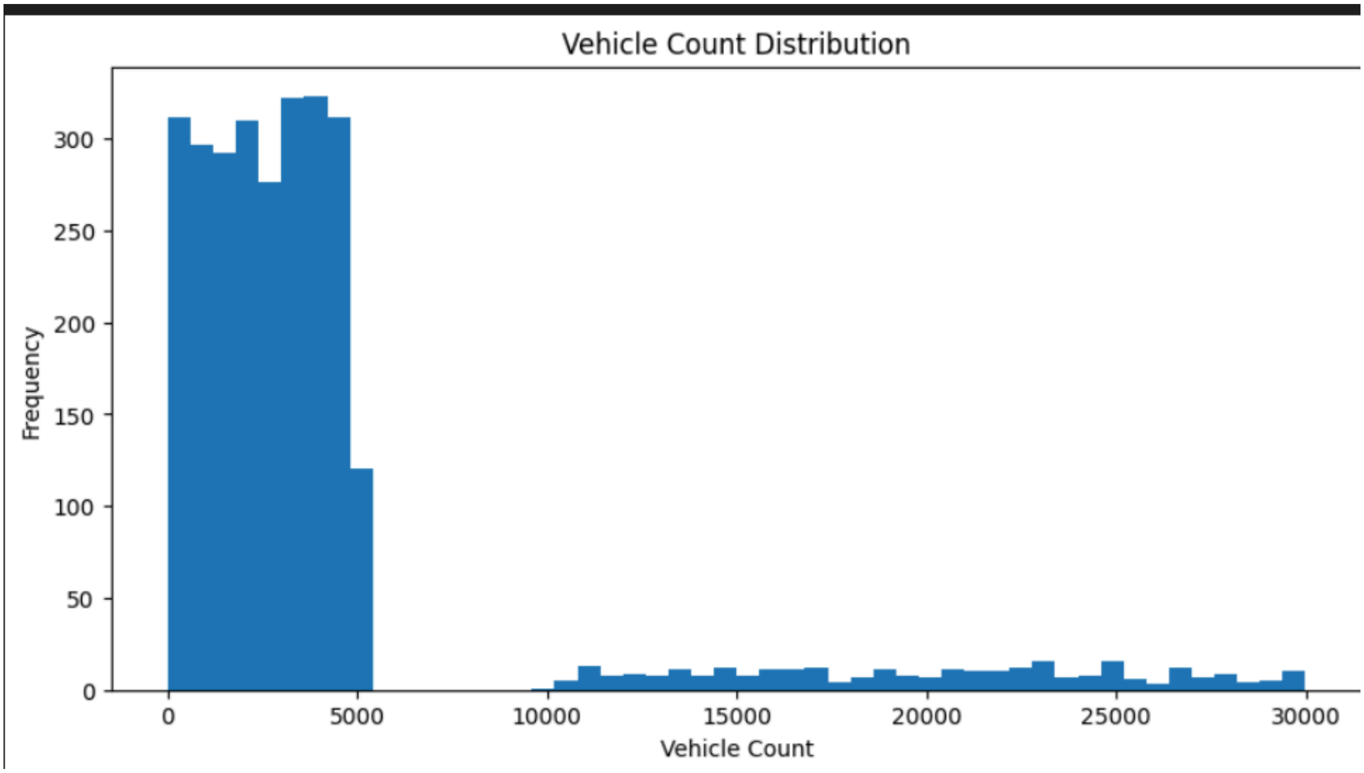
All valid timestamps were converted into a consistent format:
YYYY-MM-DD HH:MM
Invalid timestamps were removed.

Handling Missing Values

- Missing `area` values were imputed using the most frequent category (mode).
- Missing values in categorical fields like `congestion_level` and `road_condition` were filled.

in city we replace nan with London because there is only value on that column

in vehicle_count



we draw histogram

use IQR to get the outlier and replace it with median

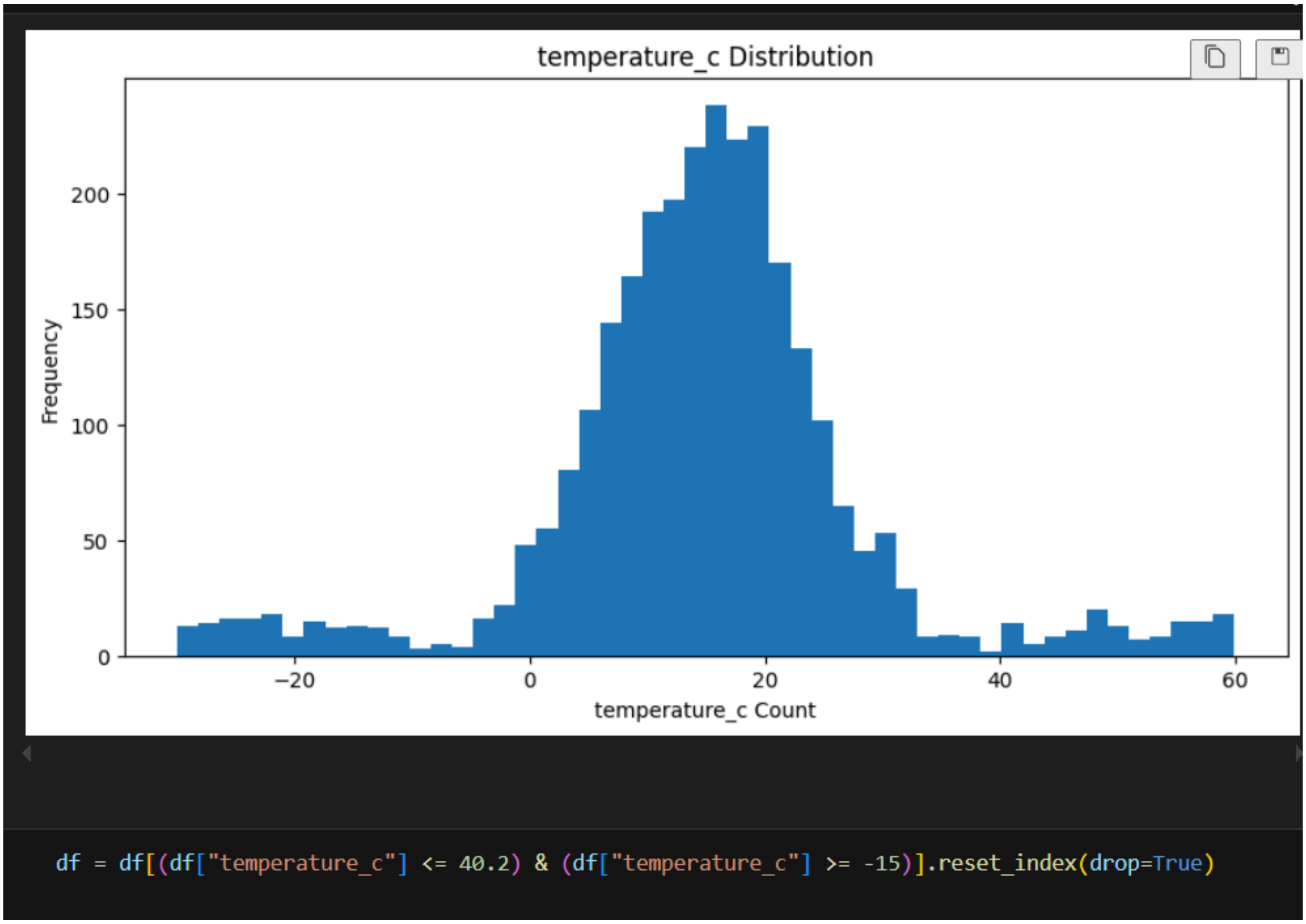
and finally we convert ["traffic_id", "vehicle_count", "accident_count", "visibility_m"] data type into integer

weather data

Number of rows before cleaning 5050
Number of rows after cleaning 2599

weather id ,date ,city we did like we did with traffic

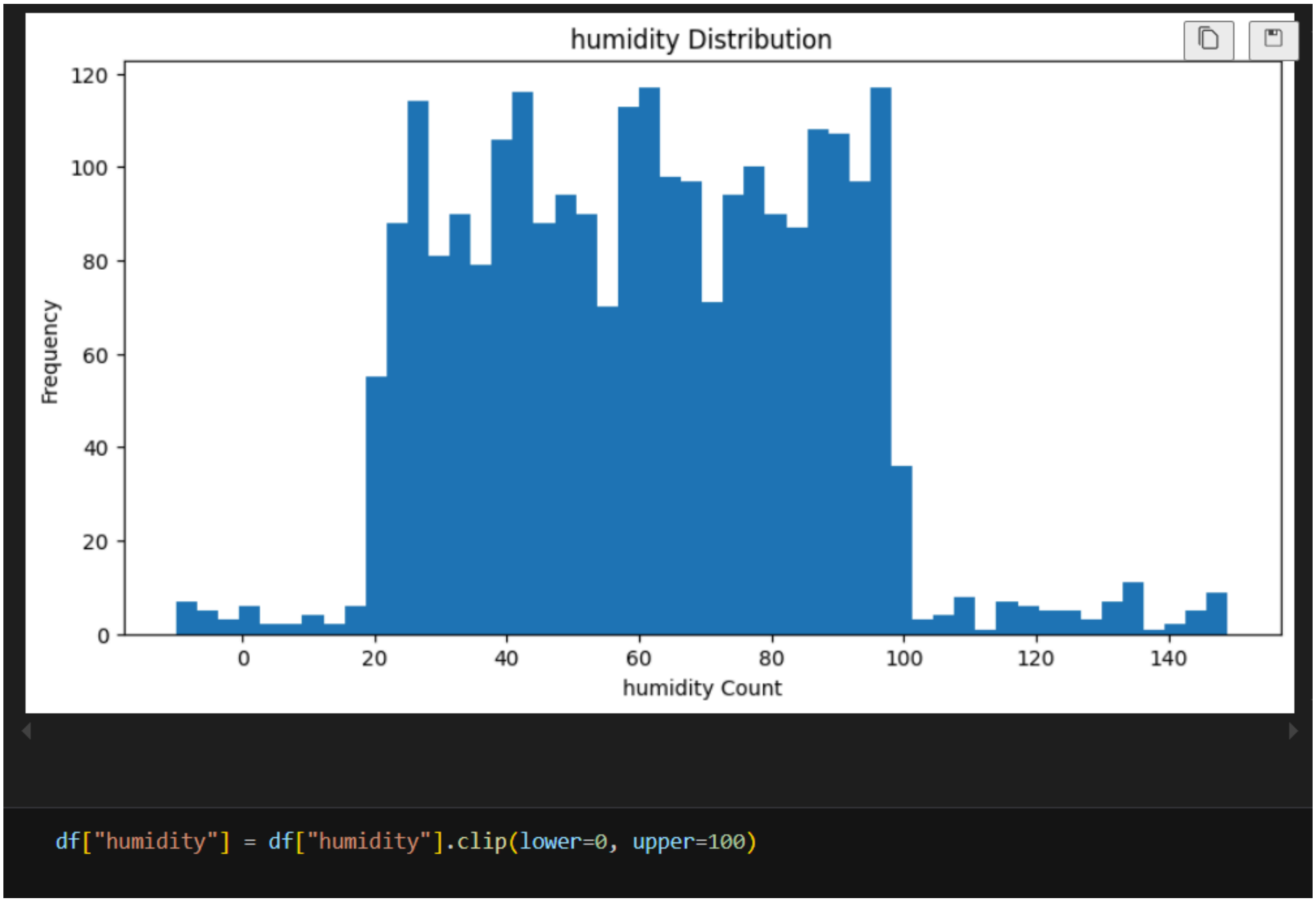
season column we replace missing value with mode



```
df = df[(df["temperature_c"] <= 40.2) & (df["temperature_c"] >= -15)].reset_index(drop=True)
```

we remove rows that temperature greater then 40.2 because this is the greater degree in weather of London was record and less then -15 also

and replace missing value with median



```
df["humidity"] = df["humidity"].clip(lower=0, upper=100)
```

humidity range is 0 100 so any values greater than 100 or less that 0 is invalid
and value greate than 100 will be replace with 100
and less than 0 will replace with 0
because we want to save rows and if remove those rows , data will decrease

in rain and wind

we reaplce missing with median also

["weather_id", "visibility_m"] convert data type into intger