Exposure notes:

* Reasoning for excluding exposure > 1

 They make up only **0.18 %** of the dataset.

 They distort the claim rate downward (0.26 → 0.04).

 One-year exposure is the standard actuarial unit used for frequency modelling.

**Claim rate difference**

* For 1-year (or shorter) policies, the **average claim rate per year is ≈ 0.26**.
* For > 1-year policies, it’s **≈ 0.04** — about **6–7 times lower**.
* This gap is *too large* to be a real difference in risk; it’s almost certainly a **measurement distortion** caused by how exposure is defined.

Why? Because:

* Claim rate = Claims / Exposure.
* When exposure doubles (2 years), the denominator doubles, but the number of claims doesn’t necessarily double — so the rate looks artificially small.
* That means the > 1-year records break the assumption that “claim rate ≈ risk per unit
* time.”

# EDA

1.Target variable

A screenshot of a black screen

AI-generated content may be incorrect.

The target variable (ClaimNb) is extremely sparse, with approximately 95% of policies having zero claims and an average claim count of only 0.05.  
This strong zero-inflation and likely overdispersion indicate that a simple Poisson model would suffer from high bias, as it assumes equal mean and variance.  
Therefore, more flexible count models such as the Negative Binomial or Tweedie distribution are expected to better capture the data structure without excessive variance.

A screenshot of a black screen

AI-generated content may be incorrect.

Hetamap.

A screenshot of a black screen

AI-generated content may be incorrect.

A screenshot of a black and white screen

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a black screen

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.