Exploratory Data Analysis (EDA)

October 25

Cindy Ha, Willie Xie, Erica Zhang

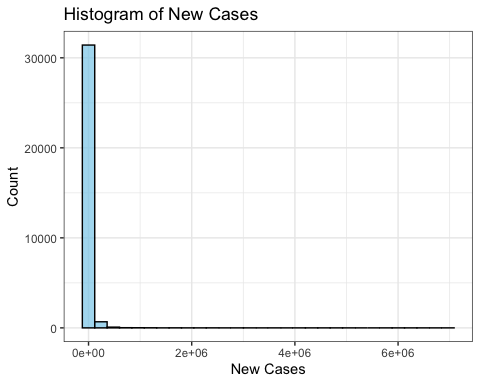
# Univariate Analysis

We start with response variable new\_cases and check for missingness:

# A tibble: 161 × 3  
 continent location date   
 <chr> <chr> <date>   
 1 South America Argentina 2020-01-01  
 2 North America Mexico 2020-01-01  
 3 South America Argentina 2020-01-02  
 4 North America Mexico 2020-01-02  
 5 Asia Sri Lanka 2020-01-28  
 6 Asia India 2020-02-02  
 7 North America Canada 2020-02-07  
 8 Asia India 2020-03-02  
 9 Asia Sri Lanka 2020-03-11  
10 Asia Sri Lanka 2020-03-14  
# ℹ 151 more rows

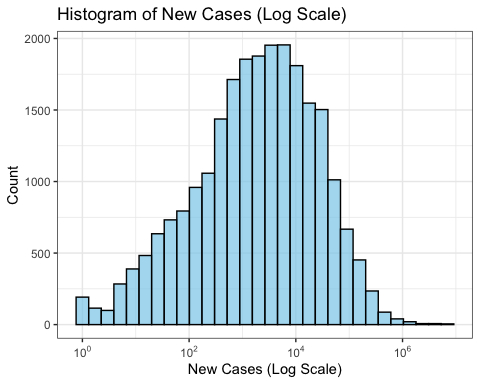
There are **161** missing response values, mainly at the beginning of the COVID outbreak before 2020/9 or more recently after 2023/5

Then, looking at the distribution of the response variable:



The distribution is heavily skewed to the right.

We thus **log-transform** new\_cases and look at the distribution after transformation:



It looks much more normally distributed now. When it comes to model training, we should probably consider log transforming the response variable first and then de-log when making predictions.

Also, a quick overview of the distribution of significant predictor variables:

A collage of graphs

Description automatically generated

We see that most of the predictor variables are also heavily **positively skewed**. Features such as female\_smokers, male\_smokers, and life\_expectancy do have a more even distribution and its time-independence make them good features to use for clustering imputation.