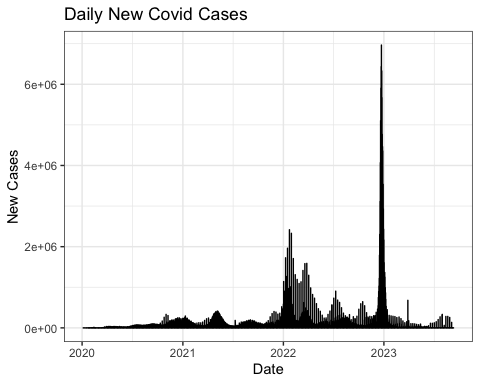
Exploratory Data Analysis (EDA)

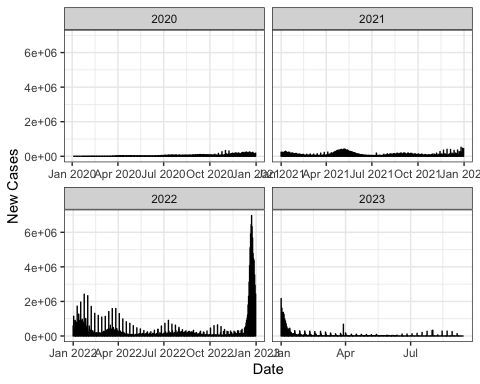
October 25

Cindy Ha, Willie Xie, Erica Zhang

# Temporal Visualization

We can look at new\_cases from 2020 to 2023.





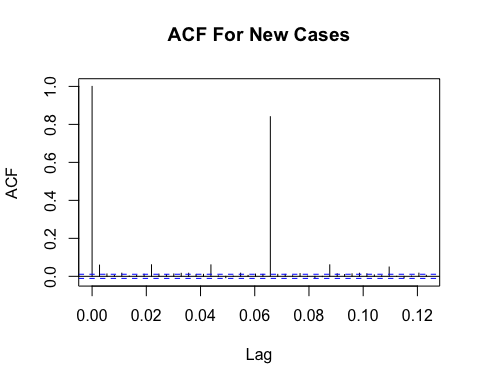
# Stationarity

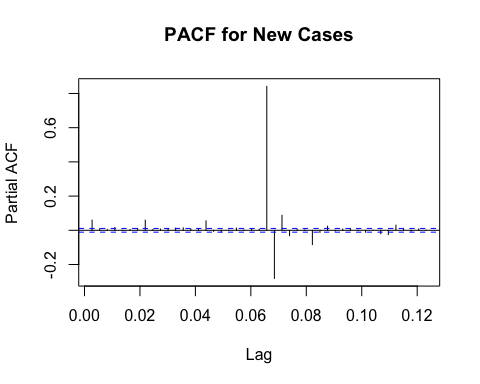
Next, we can check if the dataset is stationary or not. This will be done with the Augmeneted Dickey-Fuller Test.

Augmented Dickey-Fuller Test  
  
data: data\_ts  
Dickey-Fuller = -12.743, Lag order = 31, p-value = 0.01  
alternative hypothesis: stationary

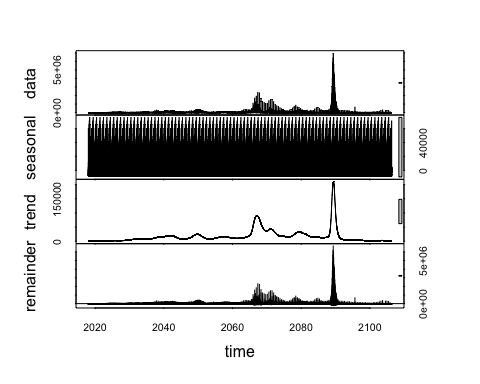
# Correlation

We can also explore the autocorrelation and partial autocorrelation.





# Seasonal Decomposition



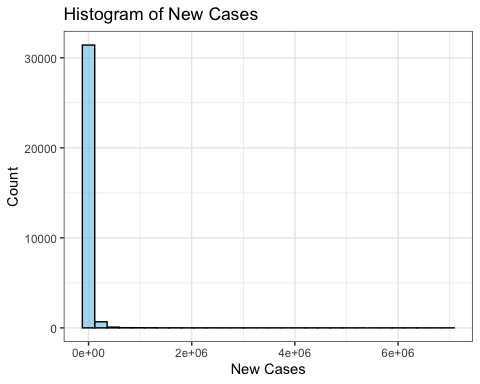
# Univariate Analysis

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

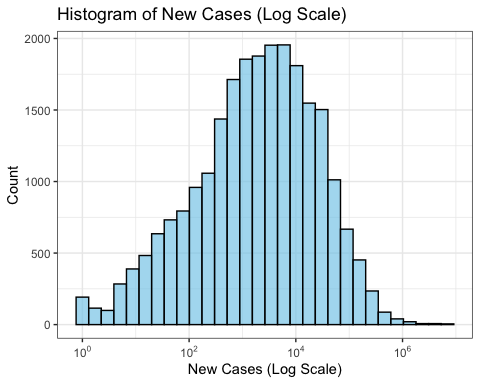
There are **161** missing response values, mainly at the beginning of the COVID outbreak before 2020/9 or more recently after 2023/5

Then, looking at the distribution of the response variable:



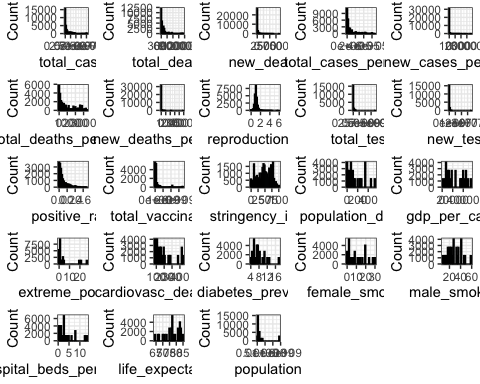
The distribution is heavily skewed to the right.

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.

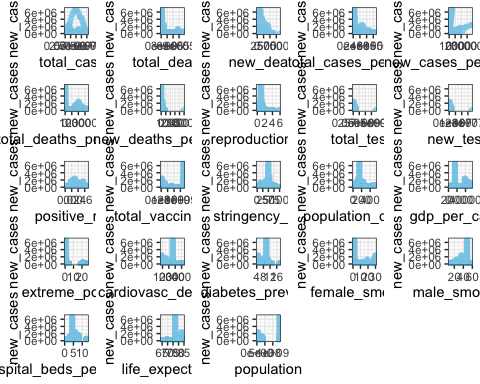
Also, a quick overview of the distribution of significant predictor variables:



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.

# Bivariate Analysis

Continuing on, we can also plot the significant predictor variables against new\_cases and look for any unique relationships that we could further explore.



TableGrob (5 x 5) "arrange": 23 grobs  
 z cells name grob  
total\_cases 1 (1-1,1-1) arrange gtable[layout]  
total\_deaths 2 (1-1,2-2) arrange gtable[layout]  
new\_deaths 3 (1-1,3-3) arrange gtable[layout]  
total\_cases\_per\_million 4 (1-1,4-4) arrange gtable[layout]  
new\_cases\_per\_million 5 (1-1,5-5) arrange gtable[layout]  
total\_deaths\_per\_million 6 (2-2,1-1) arrange gtable[layout]  
new\_deaths\_per\_million 7 (2-2,2-2) arrange gtable[layout]  
reproduction\_rate 8 (2-2,3-3) arrange gtable[layout]  
total\_tests 9 (2-2,4-4) arrange gtable[layout]  
new\_tests 10 (2-2,5-5) arrange gtable[layout]  
positive\_rate 11 (3-3,1-1) arrange gtable[layout]  
total\_vaccinations 12 (3-3,2-2) arrange gtable[layout]  
stringency\_index 13 (3-3,3-3) arrange gtable[layout]  
population\_density 14 (3-3,4-4) arrange gtable[layout]  
gdp\_per\_capita 15 (3-3,5-5) arrange gtable[layout]  
extreme\_poverty 16 (4-4,1-1) arrange gtable[layout]  
cardiovasc\_death\_rate 17 (4-4,2-2) arrange gtable[layout]  
diabetes\_prevalence 18 (4-4,3-3) arrange gtable[layout]  
female\_smokers 19 (4-4,4-4) arrange gtable[layout]  
male\_smokers 20 (4-4,5-5) arrange gtable[layout]  
hospital\_beds\_per\_thousand 21 (5-5,1-1) arrange gtable[layout]  
life\_expectancy 22 (5-5,2-2) arrange gtable[layout]  
population 23 (5-5,3-3) arrange gtable[layout]

We can see that there are interesting trend sfor total\_cases, new\_deaths, and reproduction\_rate.