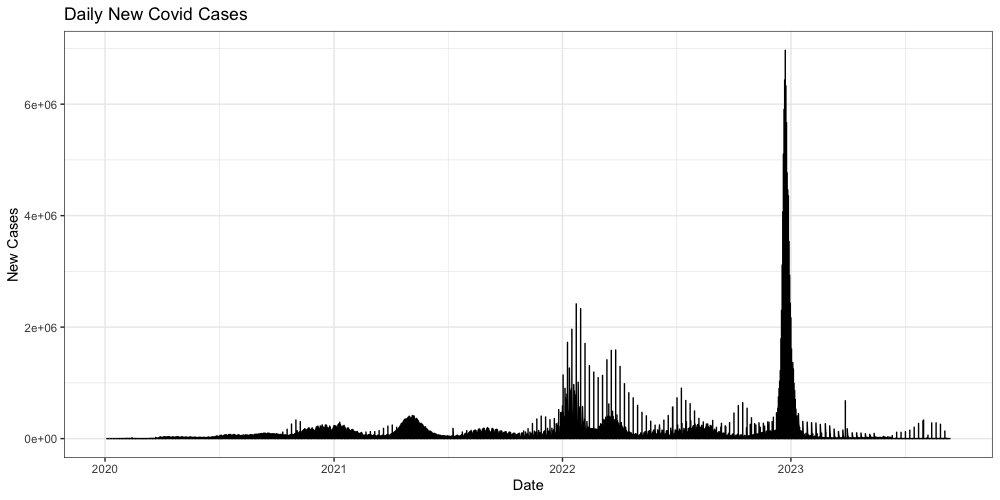
Exploratory Data Analysis (EDA)

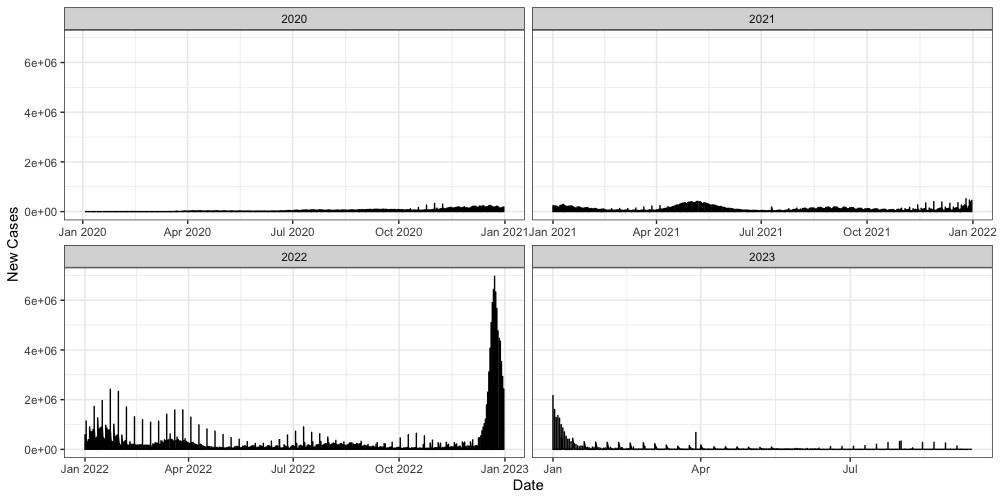
October 25

Cindy Ha, Willie Xie, Erica Zhang

# Temporal Visualization

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





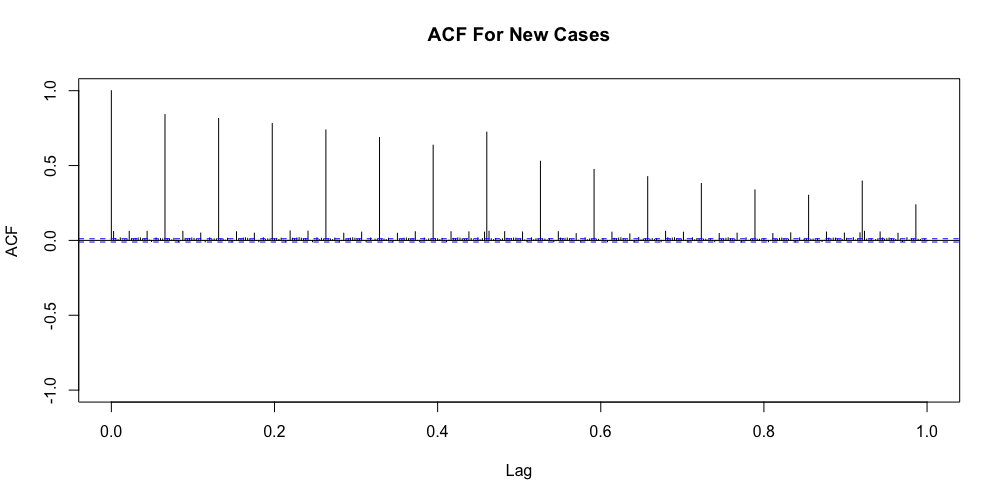
# Stationarity

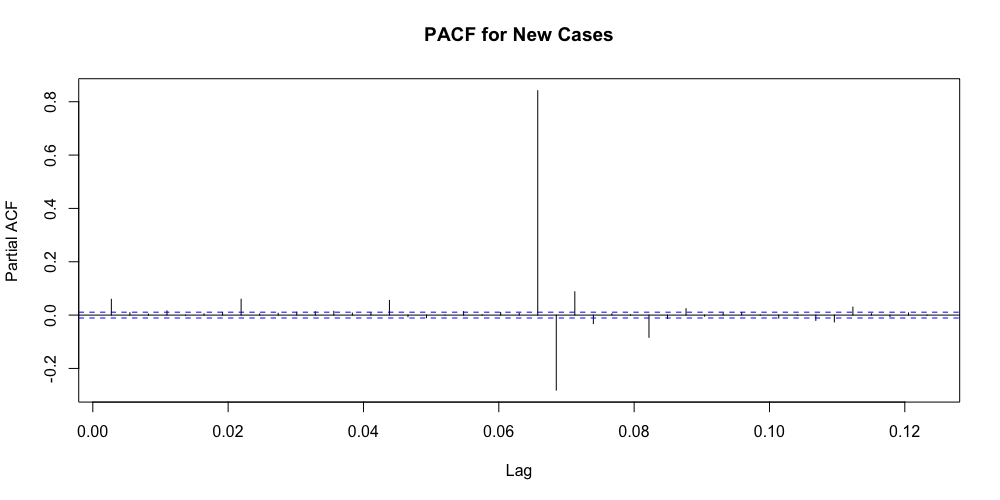
Next, we can check if the dataset is stationary or not. This will be done with the Augmeneted Dickey-Fuller Test.

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

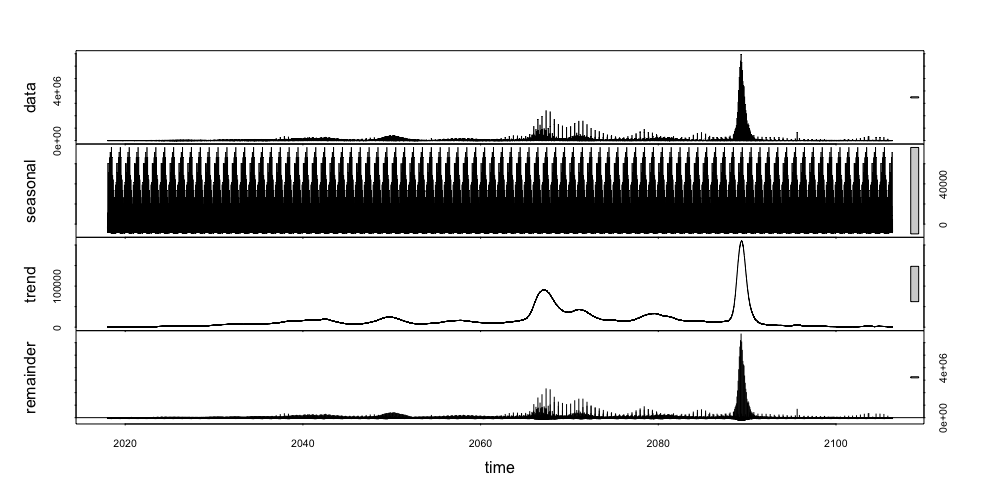
# Correlation

We can also explore the autocorrelation and partial autocorrelation.





# Seasonal Decomposition



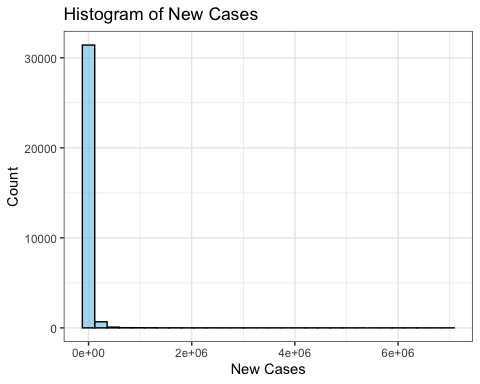
# Univariate Analysis

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

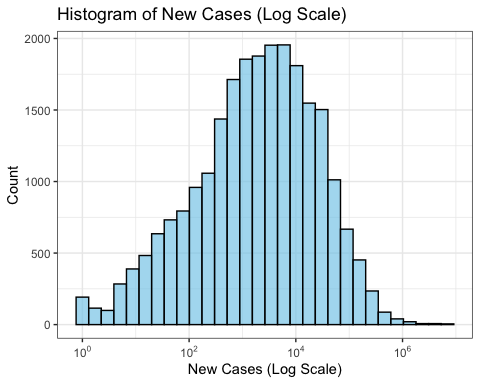
There are **161** missing response values, mainly at the beginning of the COVID outbreak before 2020/9 or more recently after 2023/5

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:

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.

# 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.

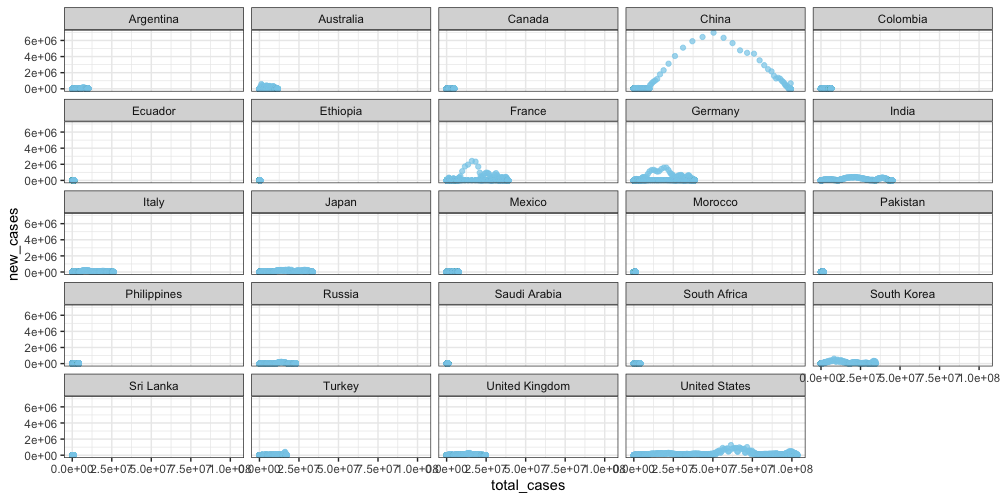
A screenshot of a graph

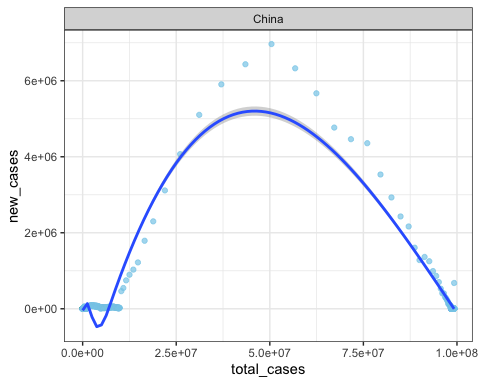
Description automatically generated

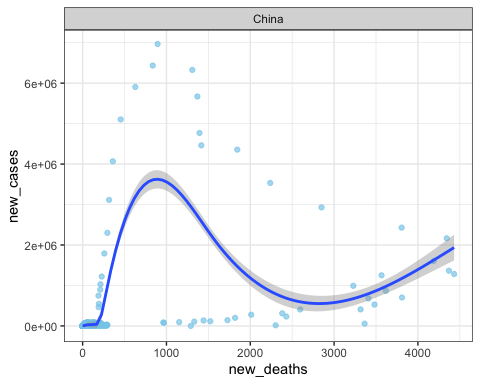
We can see that there are interesting trends for total\_cases, new\_deaths, and reproduction\_rate.

# Multivariate Analysis

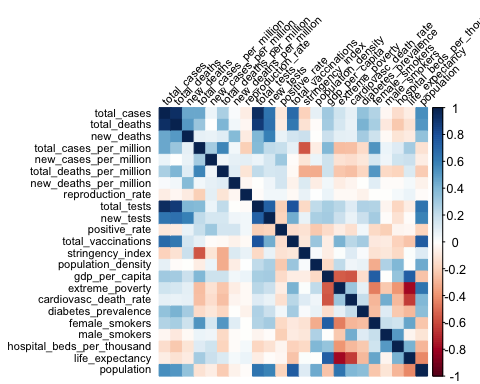
Here are more plots for those three predictors, but this time accounting for location. We can also notice that many countries stop reporting after a certain time and that China’s graph also shows a clear pattern for total\_cases and new\_deaths.







Lastly, this is the correlation plot.



An initial glance shows that there are strong correlations between:

* total\_cases & total\_tests
* gdp\_per\_capita & life\_expectancy