Signal Analysis and Processing Task 1 (Autocorrelation)

Justinas Lekavičius

April 22, 2022

Abstract

This is the report for Signal Analysis and Processing task 1 – autocorrelation (variant number 5). The purpose is to demonstrate the performance of the algorithm, giving examples of how the results can be interpreted in different cases.

Contents

1 List of signals analysed **2**

2 Signal preprocessing **3**

3 Signal analysis and autocorrelation implementation **4**

3.1 Temperature in Seoul 4

3.2 House Prices in United Kingdom 9

3.3 Natural Gas Prices 12

3.4 Daily Vaccinations in Lithuania 15

3.5 Unemployment Levels in United States (March) 18

4 Conclusions **20**

5 Code fragment **21**

1 List of signals analysed

This section describes the signals analysed and used for this task. The signal origins are detailed, their source (which database and source was used), as well as whether the signals were subjected to any preprocessing or editing.

1. Temperature in Seoul (1973-01-05 – 2022-04-09) – air temperature in Seoul, South Korea. The data was taken from National Centers for Environmental Information, National Oceanic and Atmospheric Administration database [3]. The download link: `https://www.ncei.noaa.gov/access/past-weather/KSM00047108/data.csv`

2. House prices in the United Kingdom (1952-11-01 – 2017-11-01) – UK house prices since 1952. The data was taken from Datahub database [1]. The download link: `https://datahub.io/core/house-prices-uk/r/data.csv`

3. Natural gas prices (1997 January – 2020 August) – natural gas monthly prices (US Henry Hub), measured in dollars per million British thermal units. Data originally comes from U.S. Energy Information Administration and was taken from Datahub database [2]. The download link: `https://datahub.io/core/natural-gas/r/monthly.csv`

4. US monthly unemployment rate (1948 - 2019) – percentage of unemployment level in the United States for each month from 1948 to 2019. The data was acquired from Kaggle [5] (originally from the US Bureau of Labor Statistics). The download link: `https://www.kaggle.com/datasets/tunguz/us-monthly-unemployment-rate-1948-present/download`

5. COVID-19 vaccination progress in Lithuania (2020-12-27 – 2022-04-19) – the data of vaccination progress in Lithuania. The dataset was acquired from Our World in Data GitHub repository, and the database also contains the data of other countries of the world [4]. The download link: `https://github.com/owid/covid-19-data/raw/master/public/data/vaccinations/vaccinations.csv`

2 Signal preprocessing

Some databases were preprocessed via the main Python code to either produce more relevant or more coherent signals.

The first signal (temperature in Seoul) was truncated by grouping data by months, and getting max value of each month. This resulted in a signal of highest average air temperature for each month from 1973 January to 2022 April. This produces 592 points (rows). The second signal (house prices in the United Kingdom) was also similarly processed, however, grouping data by years. This produces the max average house price for each year from 1952 to 2017, and produced 66 points (rows). The third signal (natural gas prices), just like the first signal, was condensed by grouping data by months, producing 284 rows (points), indicating the max monthly natural gas prices from 1997 January to 2020 August. The fourth signal (US monthly unemployment rate) was filtered by only selecting points for the month of March from 1948 to 2019, result – 72 points (rows). The fifth signal (COVID-19 vaccination progress in Lithuania) was produced by filtering the data set – only selecting the data for Lithuania. The data was not grouped by months or years and was kept as originally acquired, resulting in 479 rows (points) – vaccination data from December 17th 2020 to April 19th 2022.

3 Signal analysis and autocorrelation implementation

This section covers the analysis of selected signals and the display of autocorrelation performance. Graphs are also provided for illustration of the original signals and implemented algorithm performance.

3.1 Temperature in Seoul

One of the signals analysed was the air temperature in Seoul, South Korea from 1973 to 2022. The selected features of the analysed signal are the date and average temperature in degrees Fahrenheit. Originally, the signal contained the average temperature for each day from 1973 to 2022, which produced a large number of points. Therefore, it was decided to group the data by months, selecting the max value for each month (day with the highest temperature) to display the highest average temperature for each month from 1973 to 2022. This produces 592 points (rows). The original produced signal is displayed in Figure 1.

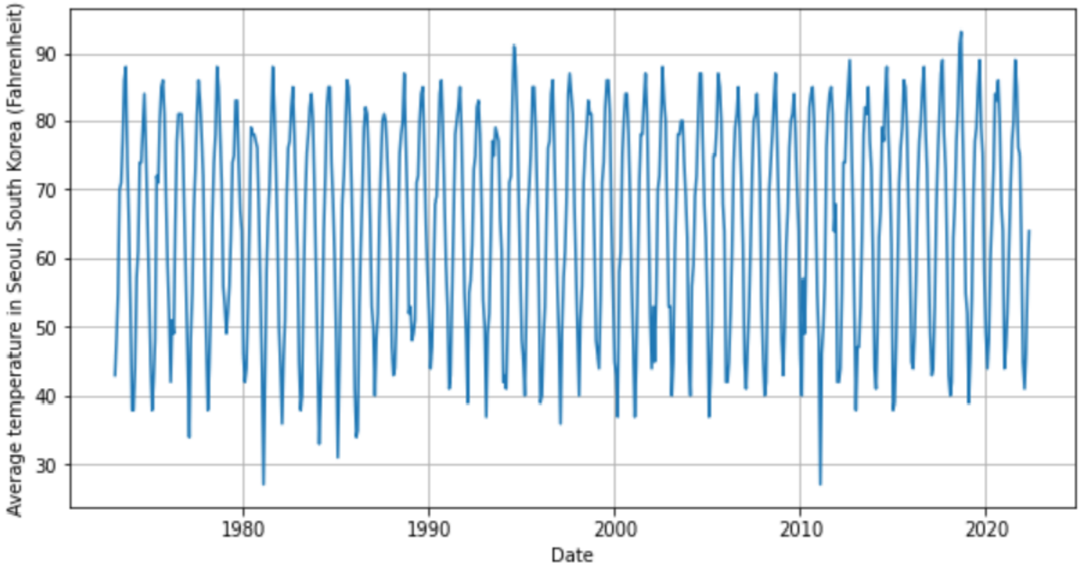

Figure 1: Highest average temperatures in Seoul, South Korea (Fahrenheit) for each month from 1973 to 2022 (original signal).

The moving averages were calculated for the original signal, choosing the parameters of 30 months and 60 months. This is illustrated in Figure 2.

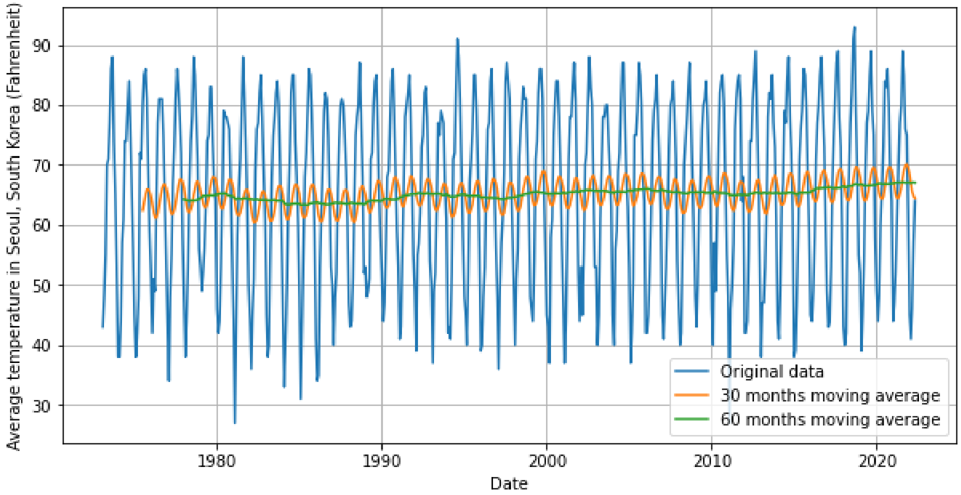

Figure 2: The original signal of average temperature in Seoul, the 30 month moving average and 60 month moving average

Firstly, autocorrelation function is applied for the original signal, and quasi-periodicity of the signal can be observed which may not have been apparent before the application of the function. This is indicated by the autocorrelation coefficient fluctuating consistently in the range of 1 and -1, as illustrated in Figure 3.

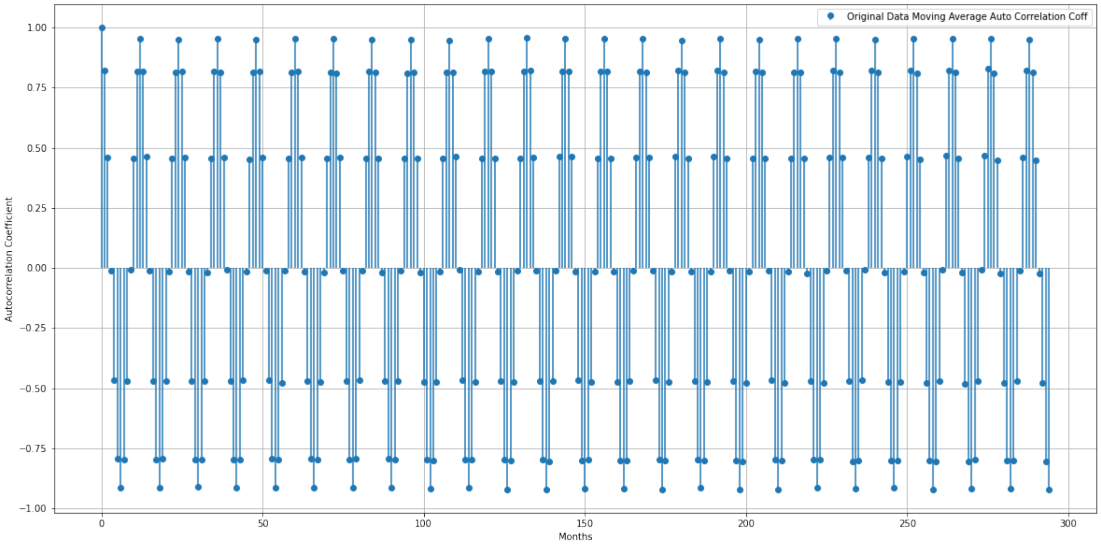

Figure 3: Autocorrelation function for the original signal (Seoul air temperature).

Displaying the autocorrelation coefficient in Figure 4 for melted original signal (signal of 12 month moving average) shows that the fluctuations are only very slight, and occur every 4-5 years.

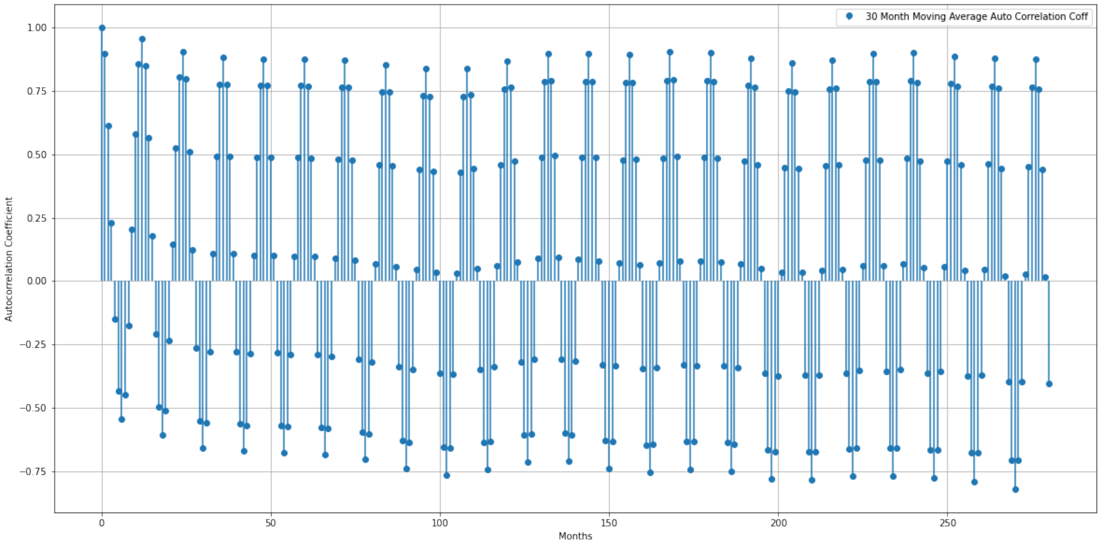

Figure 4: Autocorrelation coefficient for signal of 12 month moving average (air temperature in Seoul).

Alternatively, visualising the autocorrelation coefficient in Figure 5 for signal of 60 month moving average) shows the trend more clearly visually. The lowest value of autocorrelation coefficient is reached at around 250 months (20-21 years) and then picks up again, as seen by the increase at the end of the graph.

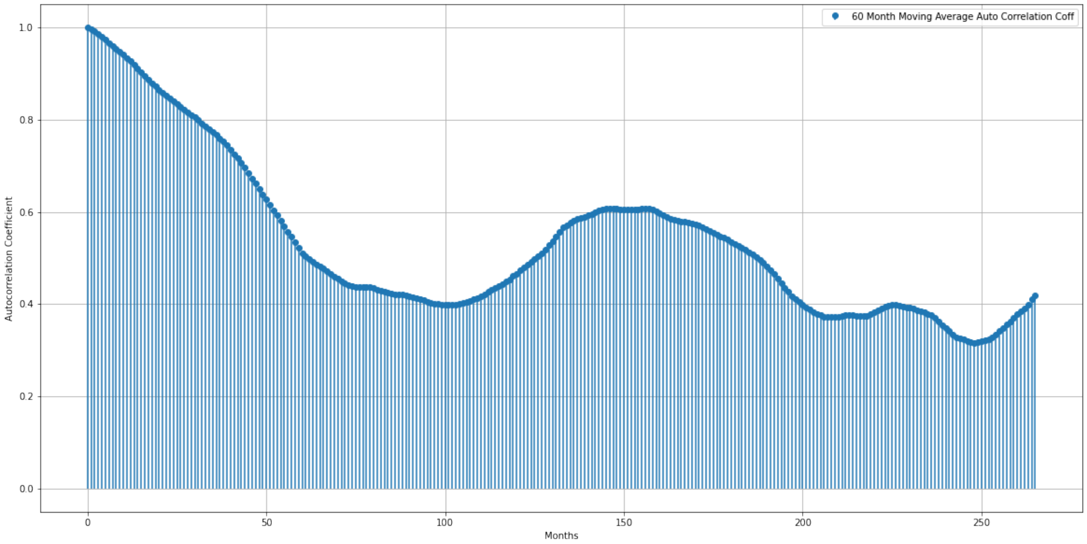

Figure 5: Autocorrelation coefficient for signal of 60 month moving average (air temperature in Seoul).

This signal was selected for artificial noising (as required by the task). The signal was noised by randomly noise of three predefined intensities, i.e., standard deviations – 1, 5 and 15. The mean of the chosen normal distribution was selected equal to 0. For production of the noise, np.random.normal function was utilised. The result is displayed in Figure 6.

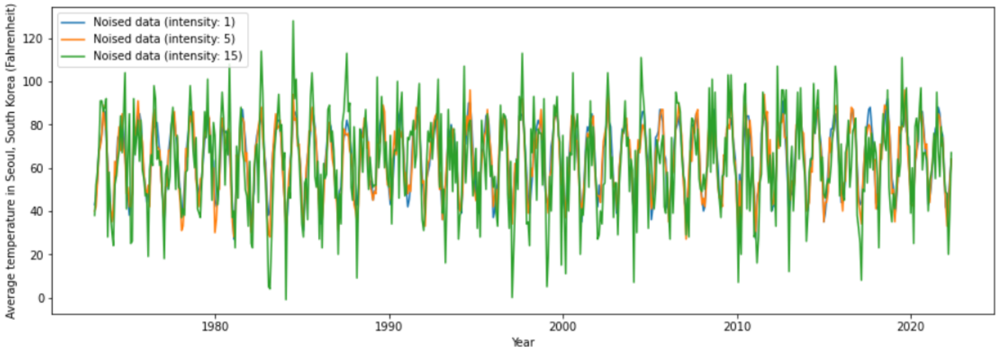

Figure 6: Signal of highest average temperatures in Seoul, South Korea noised by 1, 5 and 15 standard deviations (intensities).

Autocorrelation was applied to the noised signal of three different intensities (1, 5 and 15) and the same quasi-periodicity can be seen in the illustration below. Unfortunately, no new insights can be discovered from the application of autocorrelation to noised signals – the coefficient retains its consistency, and on the minimum and maximum values of the autocorrelation coefficient seem to change.

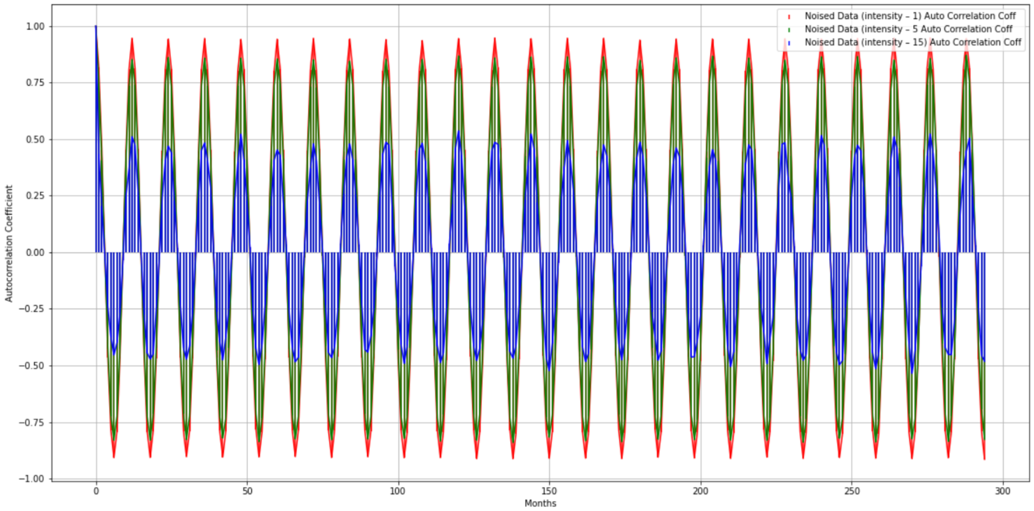

Figure 7: Noised signals (temperature in Seoul) and applied autocorrelation to them.

3.2 House Prices in United Kingdom

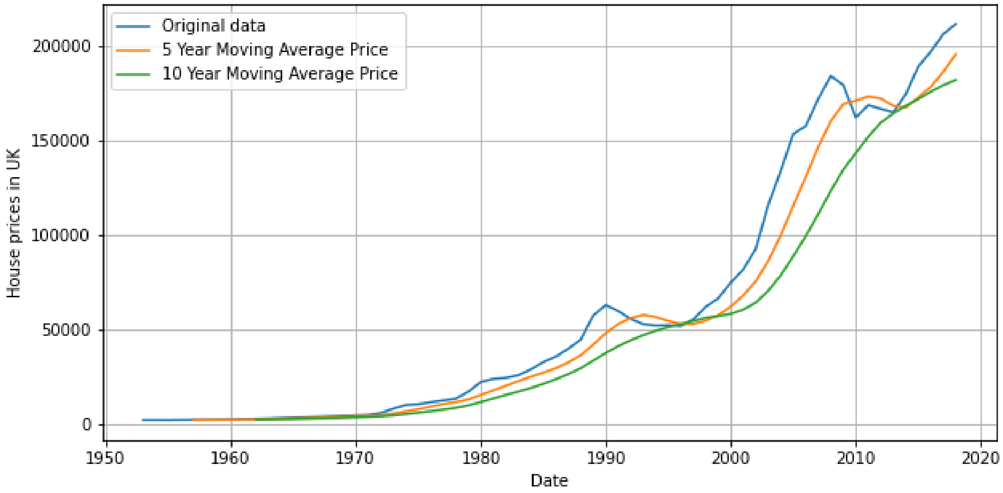

Figure 8: The original signal of house prices in UK, the 5 year and 10 year moving average signals.

Autocorrelation coefficient for house price date is very high during the whole period (Figure 9). Similar results are for melted data (Figure 14 and Figure 15). Thus, noise could potentially be added to this signal in order to see whether that would impact autocorrelation coefficient.

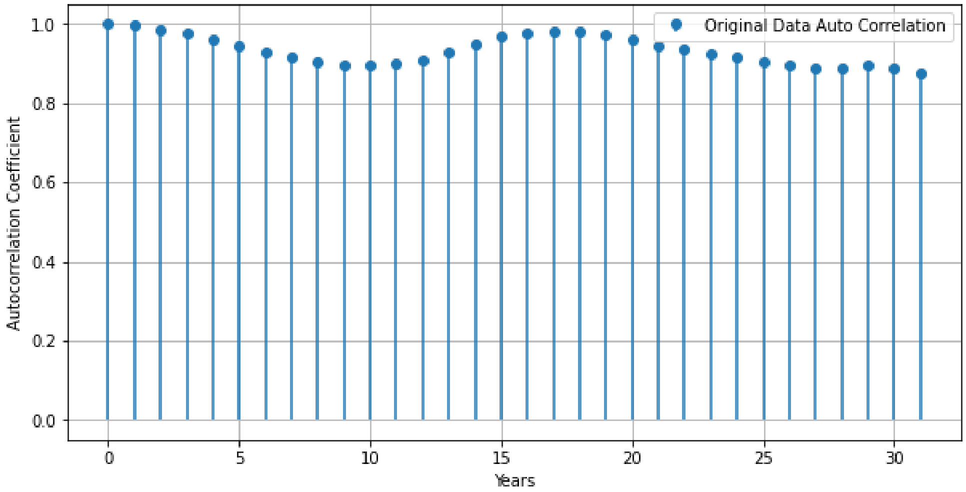

Figure 9: The original signal of house prices in UK with the autocorrelation function applied – coefficient and years.

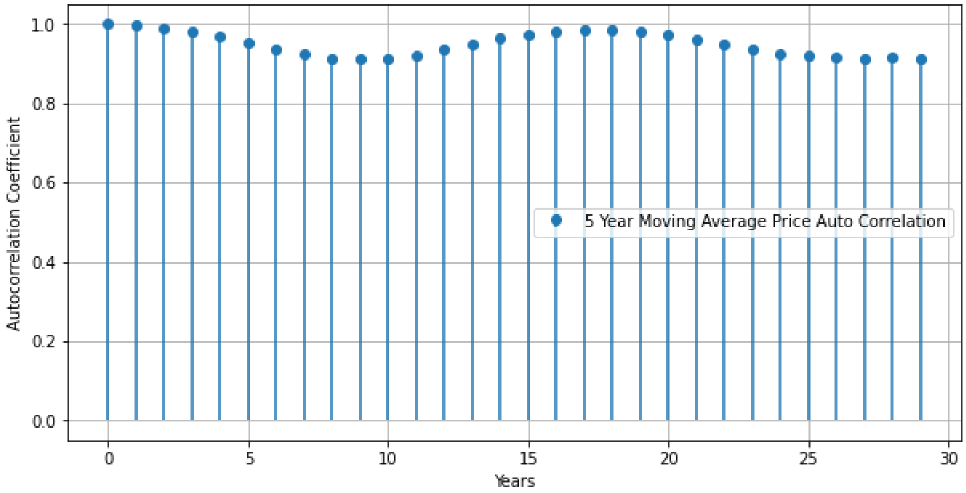

Figure 10: The 5 year moving average signal of house prices in UK with the autocorrelation function applied – coefficient and years.

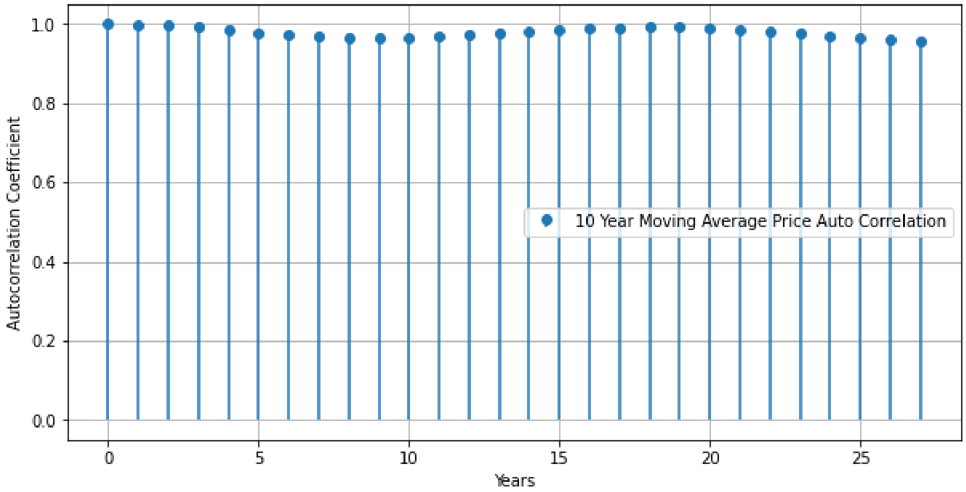

Figure 11: The 10 year moving average signal of house prices in UK with the autocorrelation function applied – coefficient and years.

3.3 Natural Gas Prices

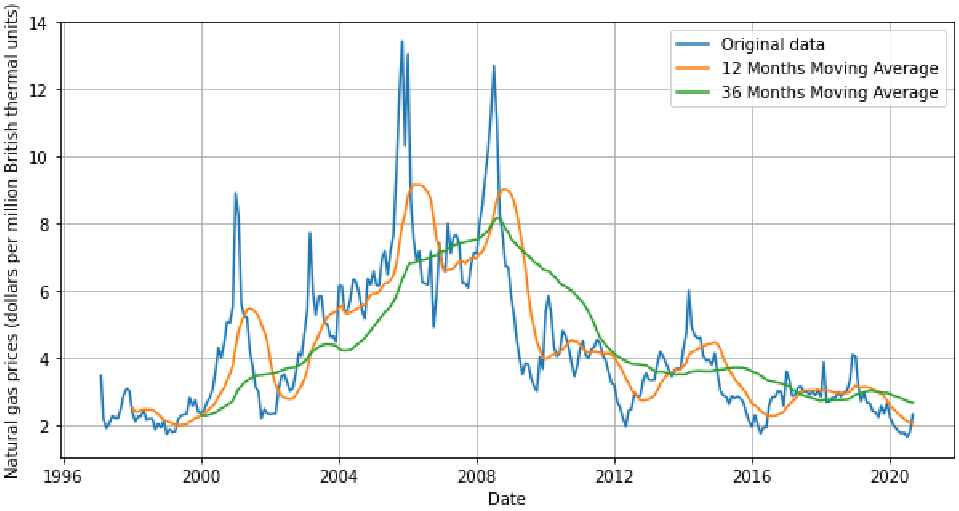

Figure 12: The original signal of natural gas prices, the 12 month moving average and 36 month moving average

In order to see trends for gas prices - moving average algorithm was applied on the provided data. Autocorrelation function results show that for original time series (Figure 13) and for melted data (Figure 14 and Figure 15) coefficient behaves a little bit differently. In original time series there are more data fluctuations and coefficient drops below zero after 58th month whereas in the second and third graphs it drops below zero after 49th and 51st month, respectively.

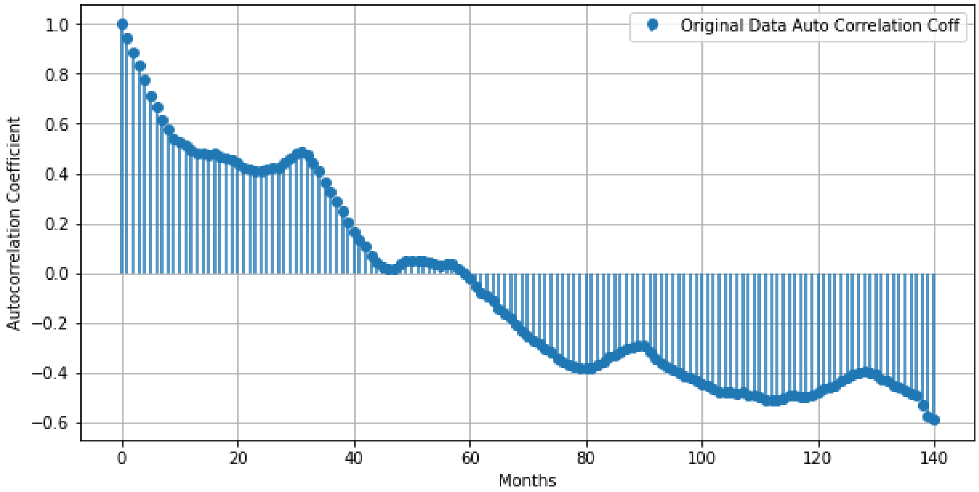

Figure 13: The autocorrelation coefficient for the original signal of natural gas prices.

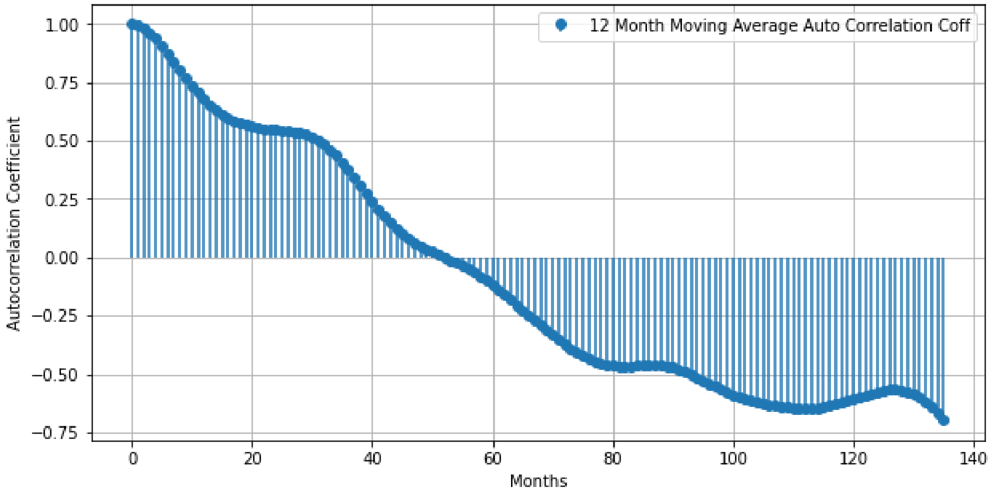

Figure 14: The autocorrelation coefficient for the 6 month moving average signal of natural gas prices.

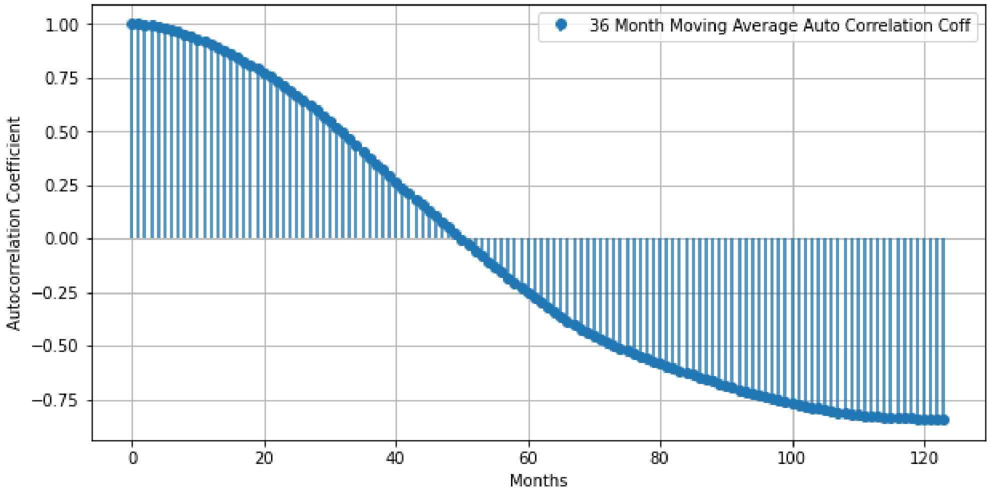

Figure 15: The autocorrelation coefficient for the 12 month moving average signal of natural gas prices

3.4 Daily Vaccinations in Lithuania

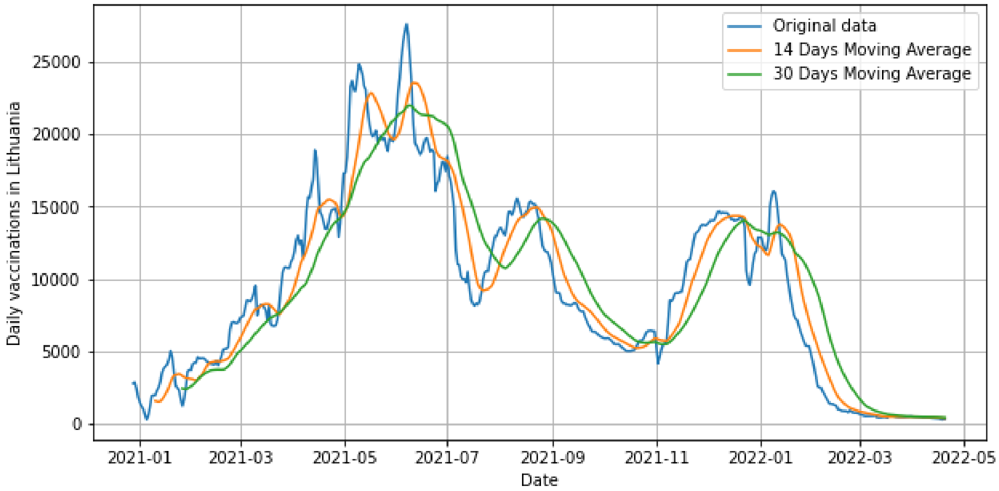

Figure 16: The original signal of daily vaccinations in Lithuania, the 14 days moving average and 30 days moving average.

Moving average algorithm was used for the original vaccination data in order to minimize fluctuations (Figure 19 and Figure 19) and see whether there are any changes in results of autocorrelation coefficient algorithm. Zero is reached on 64th, 62nd, 60th days), therefore, aside major fluctuations, moving average algorithm did not really have an impact on the time series.

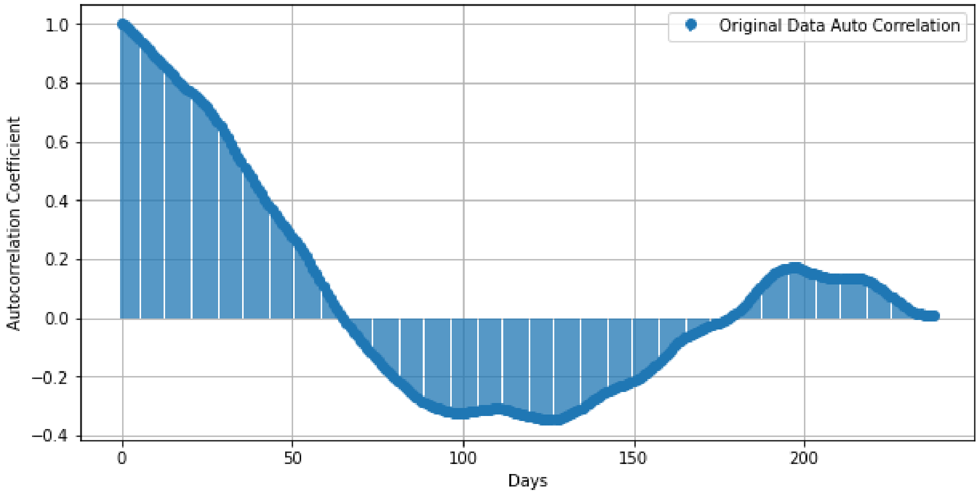

Figure 17: The autocorrelation coefficient of original signal (daily vaccinations in Lithuania).

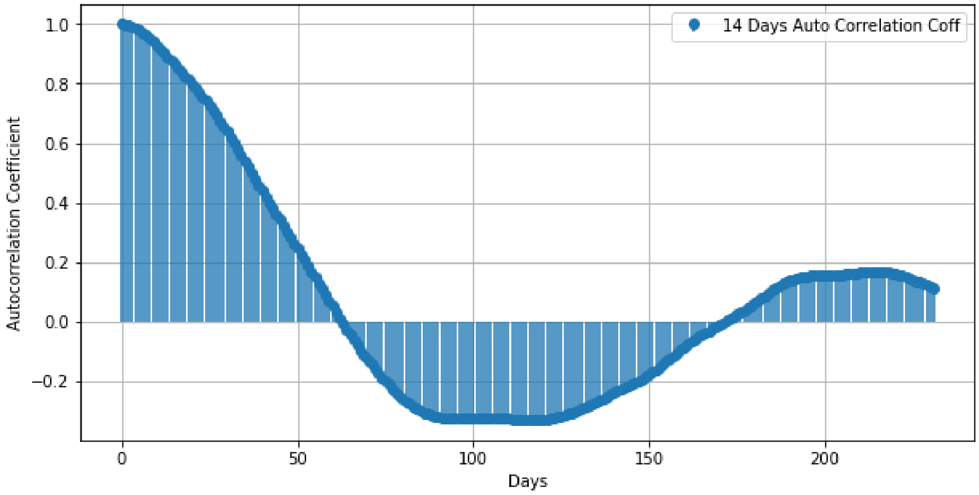

Figure 18: The autocorrelation coefficient of 14 days moving average signal (daily vaccinations in Lithuania).

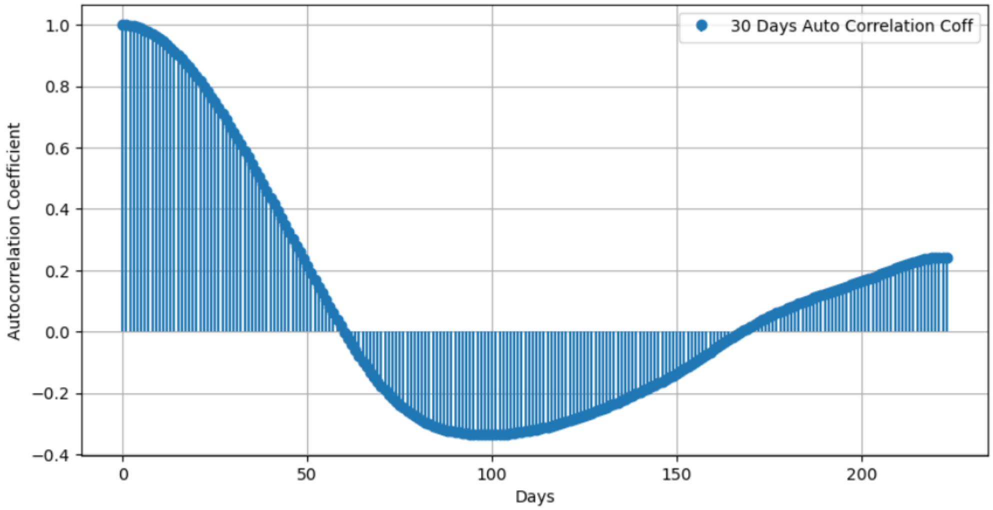

Figure 19: The autocorrelation coefficient of 30 days moving average signal (daily vaccinations in Lithuania).

3.5 Unemployment Levels in United States (March)

Figure 20 shows the original signal of unemployment levels in the United States for the month of March, also the 6 year and 12 year moving averages. Even though the amount of points is not large, the applied moving averages help show the trends of unemployment percentages throughout the years for the same month.

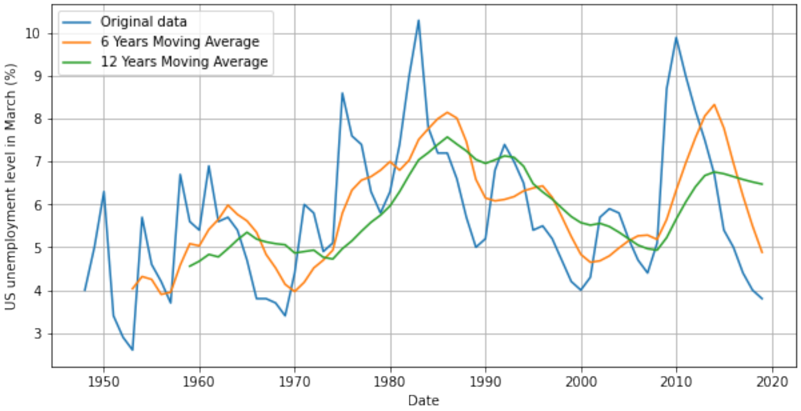

Figure 20: The original signal of unemployment levels (March) in US, the 6 years moving average and 12 years moving average.

Figure 21 displays the autocorrelation coefficient of the original signal of unemployment levels in the United States for the month of March. The coefficient does not reach the zero level immediately, and that could indicate the non-stationarity of the signal.

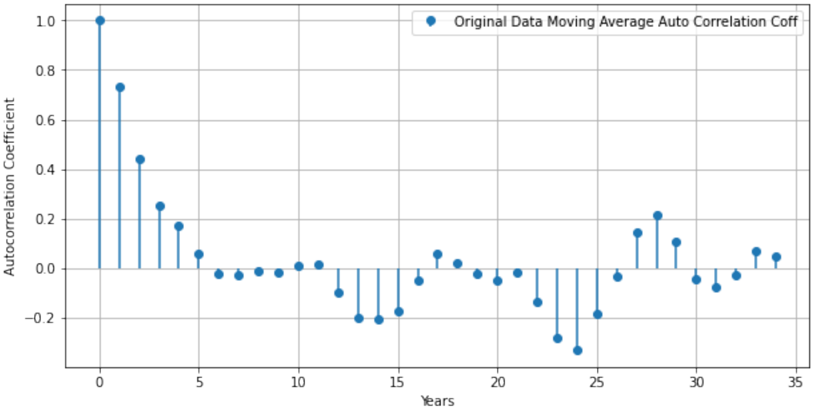

Figure 21: The autocorrelation coefficient of original signal of unemployment levels (March) in the US.

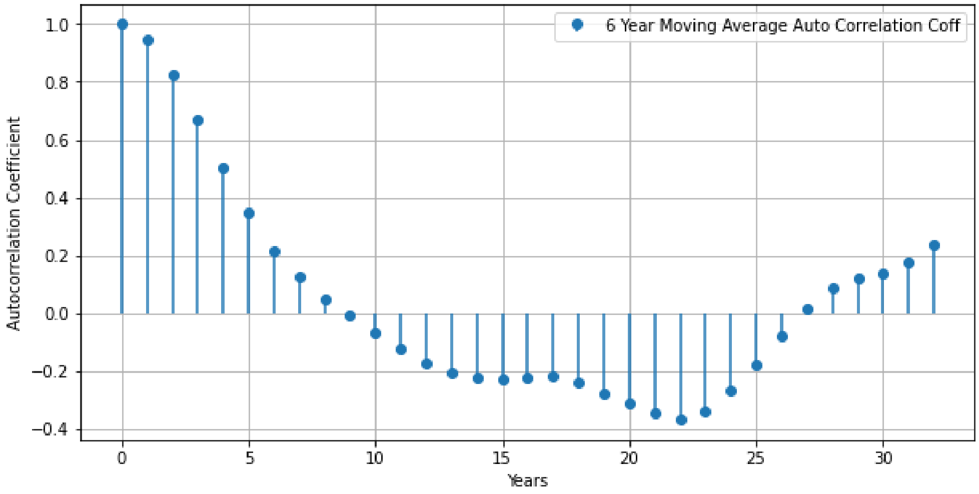

Figure 22: The autocorrelation coefficient of 6 years moving average signal of unemployment levels (March) in the US.

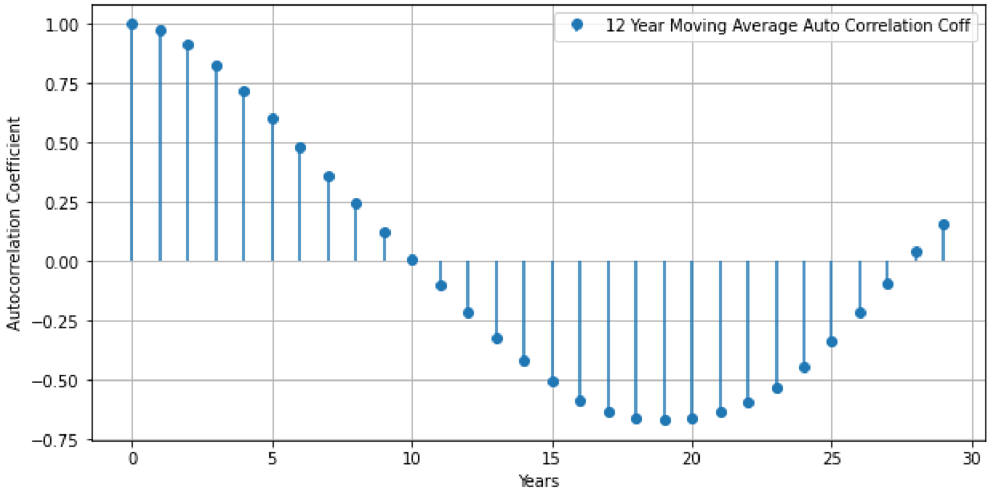

Figure 23: The autocorrelation coefficient of 12 years moving average signal of unemployment levels (March) in the US.

4 Conclusions

The application of autocorrelation algorithm for signals of various types displayed different properties, such as quasi-periodicity. The signals may also be melted to visually observe trends more clearly, and this was done using the moving average function – providing different parameters, i.e., different times (2 weeks and month, 12 months and 36 months, etc.). Concerning the noising of the signal and autocorrelation application (in this instance, the noising of air temperature in Seoul signal) did not provide any significant insights, and this may be due to insignificance of noising the selected signal. Other signal could have been selected, for example, the house prices in the UK.

5 Code fragment

The following code fragment is the implementation of the autocorrelation algorithm for the first task. The autocorrelation algorithm (depending on argument d) is defined as:

$$r(d) = \frac{\sum_{j=0}^{N-d}(f_j - \bar{f})(f_{d+j} - \bar{F})}{\sqrt{\sum_{j=0}^{N-d}(f_j - \bar{f})^2 \sum_{j=0}^{N-d}(f_{d+j} - \bar{F})^2}}, d = 0, 1, ..., [N/2]$$

$$\bar{f} = \frac{1}{N-d+1}\sum_{i=0}^{N-d} f_i, \bar{F} = \frac{1}{N-d+1}\sum_{i=0}^{N-d} f_{d+i}$$

```
# Autocorrelation algorithm
def calculate_f_mean(f, N, d):
  sum = 0
  for i in range (0, N - d + 1):
    sum += f[i]
  return 1 / (N - d + 1) * sum

def calculateTop(f, f_, F_, N, d):
  sum = 0
  for j in range(0, N - d):
    sum += (f[j] - f_)*(f[d+j] - F_)
  return sum

def calculateBottom(f, f_, F_, N, d):
  sum1 = 0
  sum2 = 0
  for j in range(0, N - d):
    sum1 += (f[j] - f_) ** 2
    sum2 += (f[d+j] - F_) ** 2
  return np.sqrt(sum1 * sum2)

def autocorrelation_function(series):

  N = np.array(series).size - 1 # Number of observations
  delays = int(N / 2)  # Number of delays (lags)
  delay = np.arange(start=0, stop = delays, step = 1) # Array of d
                                     values (0,1,2...N/2)
  corr = np.zeros(delays) # creating empty array for r(d) values

  for d in range (0, delays):
    # (34th equation). Since in python range last element is not
                                       included, adding + 1.
    f_ = calculate_f_mean(series[0 : N - d + 1], N, d)
    F_ = calculate_f_mean(series[d : N + 1], N, d)
    top = calculateTop(series, f_, F_, N, d)
    bottom = calculateBottom(series, f_, F_, N, d)
    corr[d] = top/bottom

  return delay, corr
```

References

[1] Datahub. House Prices in the UK since 1953. `https://datahub.io/core/house-prices-uk`, 2018.

[2] Datahub. Natural gas prices. `https://datahub.io/core/natural-gas`, 2018.

[3] National Centers for Environmental Information – SEOUL CITY KS (KSM00047108). Temperature in Seoul, South Korea. `https://www.ncei.noaa.gov/access/past-weather/seoul`, 2022.

[4] Ritchie H. Ortiz-Ospina E. et al) Mathieu, E. A global database of COVID-19 vaccinations. `https://doi.org/10.1038/s41562-021-01122-8`, 2021.

[5] US Bureau of Labor Statistics. US monthly unemployment rate from 1948 to 2019. `https://www.kaggle.com/datasets/tunguz/us-monthly-unemployment-rate-1948-present`, 2019.