Trend Analysis on Illinois water daily flow data

1. Data:
   I use the gate number from “Copy of Illinois gages 2022-3225.xlsx” to find the data in the USGS. Then, I only use the gate which has over 50 years of record (note: after filtering 10 missing days in a year)
   I use gate number “03336900” for the test and its “year” and “flow” data for analysis.
2. Augmented Dickey-Fuller test
   The test can only be used to inform the degree to which a null hypothesis can be rejected or fail to be reject.

- **Null Hypothesis (H0):** If failed to be rejected, it suggests the time series has a unit root, meaning it is non-stationary. It has some time dependent structure.
- **Alternate Hypothesis (H1):** The null hypothesis is rejected; it suggests the time series does not have a unit root, meaning it is stationary. It does not have time-dependent structure.
  We use p-value to decide whether we should reject H0 or not.
- **p-value > 0.05**: Fail to reject the null hypothesis (H0), the data has a unit root and is non-stationary.
- **p-value <= 0.05**: Reject the null hypothesis (H0), the data does not have a unit root and is stationary.
  ADF test: The more negative of ADF Statistic, the more likely we are to reject the null hypothesis (means we have a stationary dataset). Also, the number is less than 1% (-3.431), so that we can reject the null hypothesis

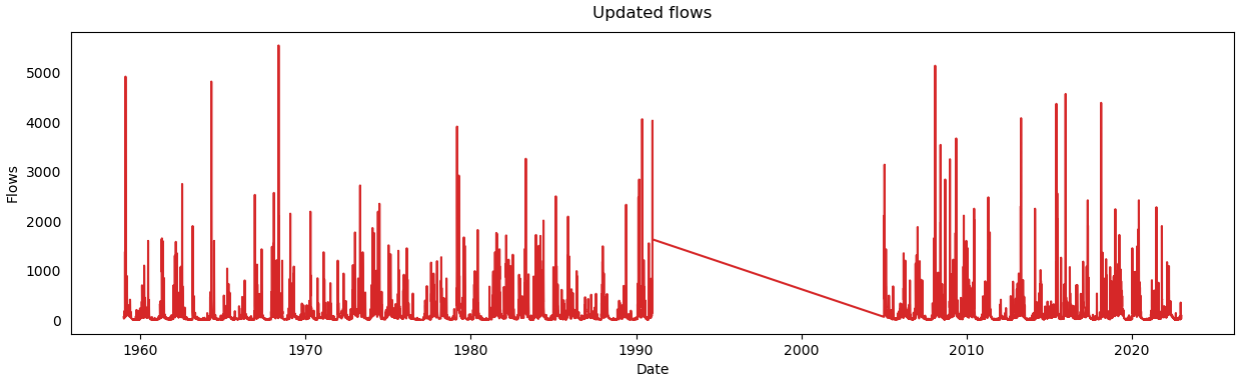
```
ADF Statistic: -16.991512
p-value: 0.000000
Critical Values:
    1%: -3.431
    5%: -2.862
    10%: -2.567
```

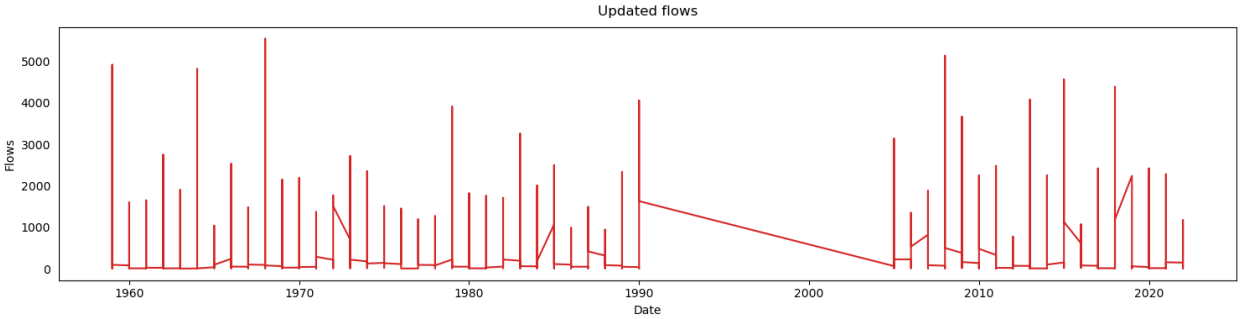
3. KPSS test
   The KPSS test is also commonly used to analyze the stationarity of series. However, it is quite different than the ADF test. For the interpretation of **p-value**, it is opposite to each other.

- **If p-value <=0.05**, then the series is non-stationary.
- KPSS test around a deterministic trend. “The word ‘deterministic’ implies the slope of the trend in the series does not change permanently. That is, even if the series goes through a shock, it tends to regain its original path.” In KPSS test,

```
KPSS Statistic: 0.07829827447654009
p-value: 0.1
num lags: 45
Critial Values:
   10% : 0.119
   5% : 0.146
   2.5% : 0.176
   1% : 0.216
Result: The series is stationary
```

4. The trend of the dataset:
   It seems there is no seasonality.



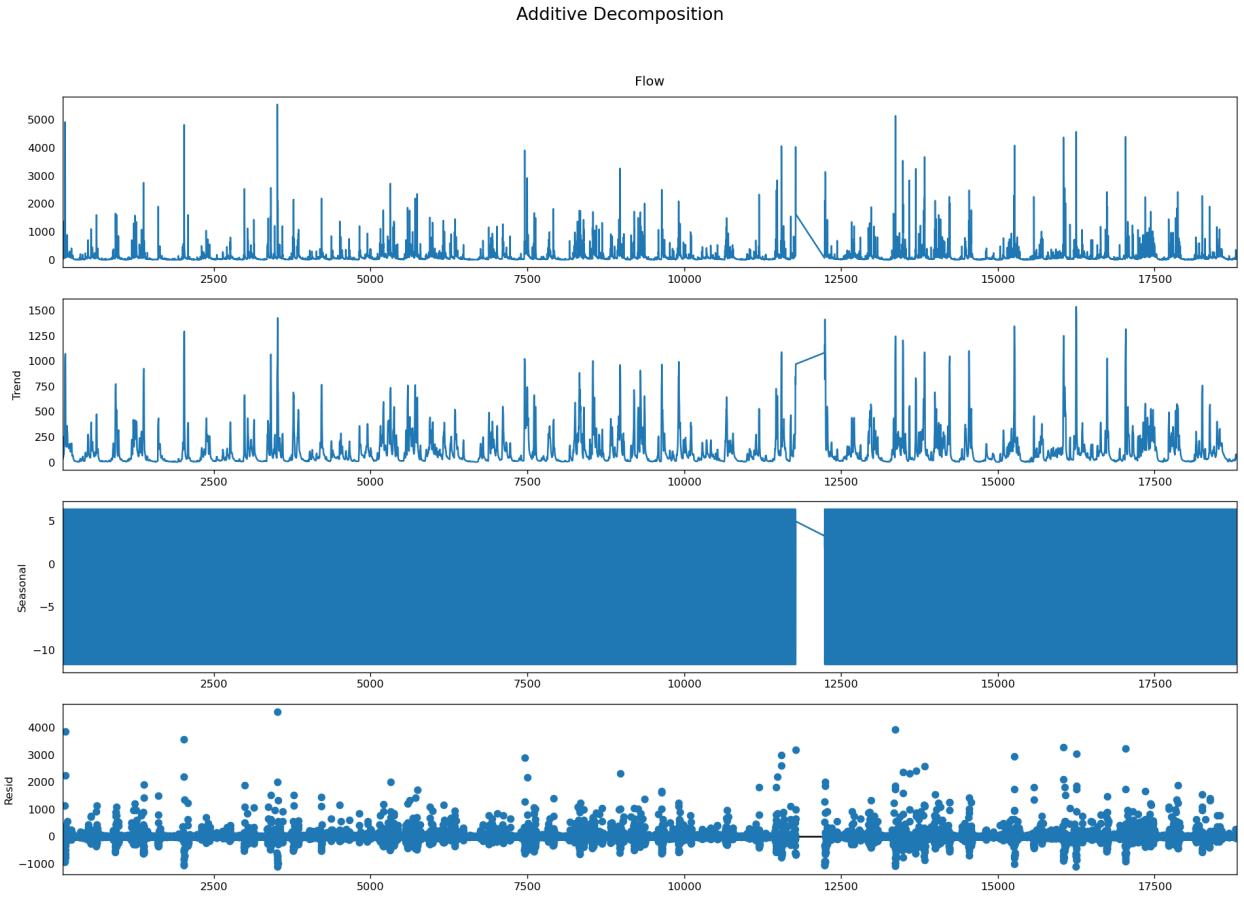


5. Additive and Multiplicative Time Series:
   We may have different combinations of trends and seasonality. Depending on the nature of the trends and seasonality, a time series can be modeled as an additive or multiplicative time series.

- **Additive time series:**
  Value = Base Level + Trend + Seasonality + Error

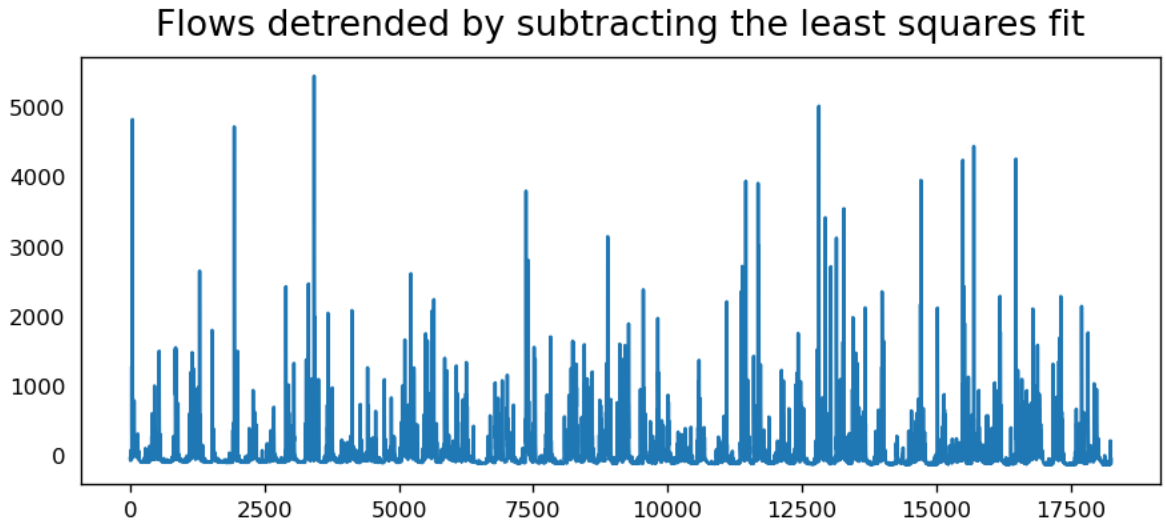
- **Multiplicative Time Series:**
  Value = Base Level x Trend x Seasonality x Error

Since we have zero values in Flow, we could not use Multiplicative Decomposition. If we look at the residuals of the additive decomposition closely, it has some pattern.



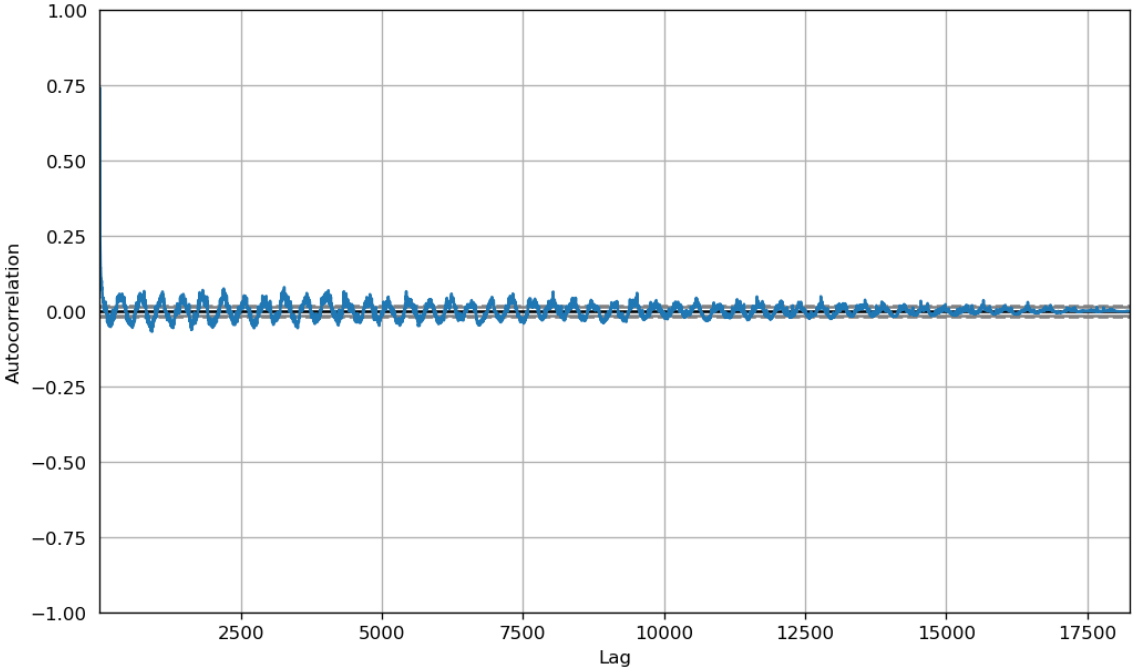
6. Detrend time series data:

In order to observe subtrends in the data that are seasonal or cyclical, detrending time series data would help to remove an underlying trend in the data.



7. Test for seasonality:

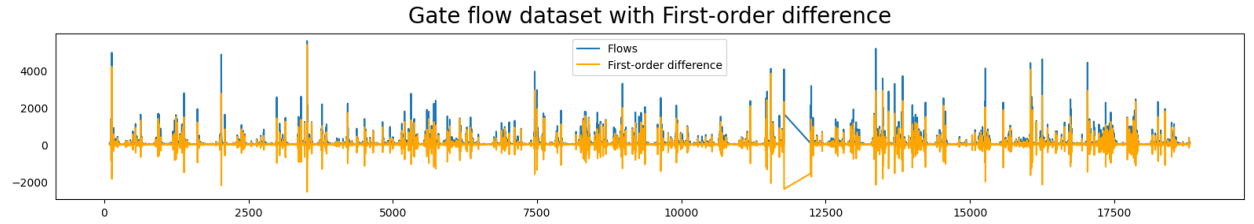
The correlation looks average from the beginning and gradually becomes a line.



8. Autocorrelation and Partial Autocorrelation Functions:

- **Autocorrelation** is simply the correlation of a series with its own lags. If a series is significantly autocorrelated, that means, the previous values of the series (lags) may be helpful in predicting the current value.
- **Partial Autocorrelation** also conveys similar information but it conveys the pure correlation of a series and its lag, excluding the correlation contributions from the intermediate lags.

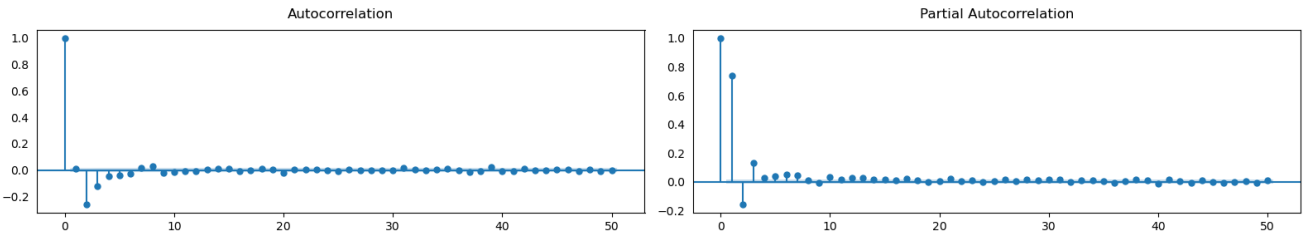
I would like to know how correlated is the number of flows this year with the number of flows in the previous year. Here, the previous year indicates the lag value of 1. Also, I use Autocorrelation and Partial Autocorrelation to find the values of p and q for Arima model. Before calculating autocorrelation, I should make the time series stationary. The easiest way to make time series stationary is by calculating the first-order difference.




```
[ 1.    0.01 -0.26 -0.12 -0.05 -0.04 -0.03  0.02  0.03 -0.02 -0.02 -0.01
 -0.01  0.    0.01  0.01 -0.01  0.    0.01  0.   -0.02  0.    0.    0.01
 -0.   -0.01  0.   -0.01 -0.   -0.   -0.    0.01  0.   -0.    0.    0.01
 -0.   -0.02 -0.01  0.02 -0.01]
```

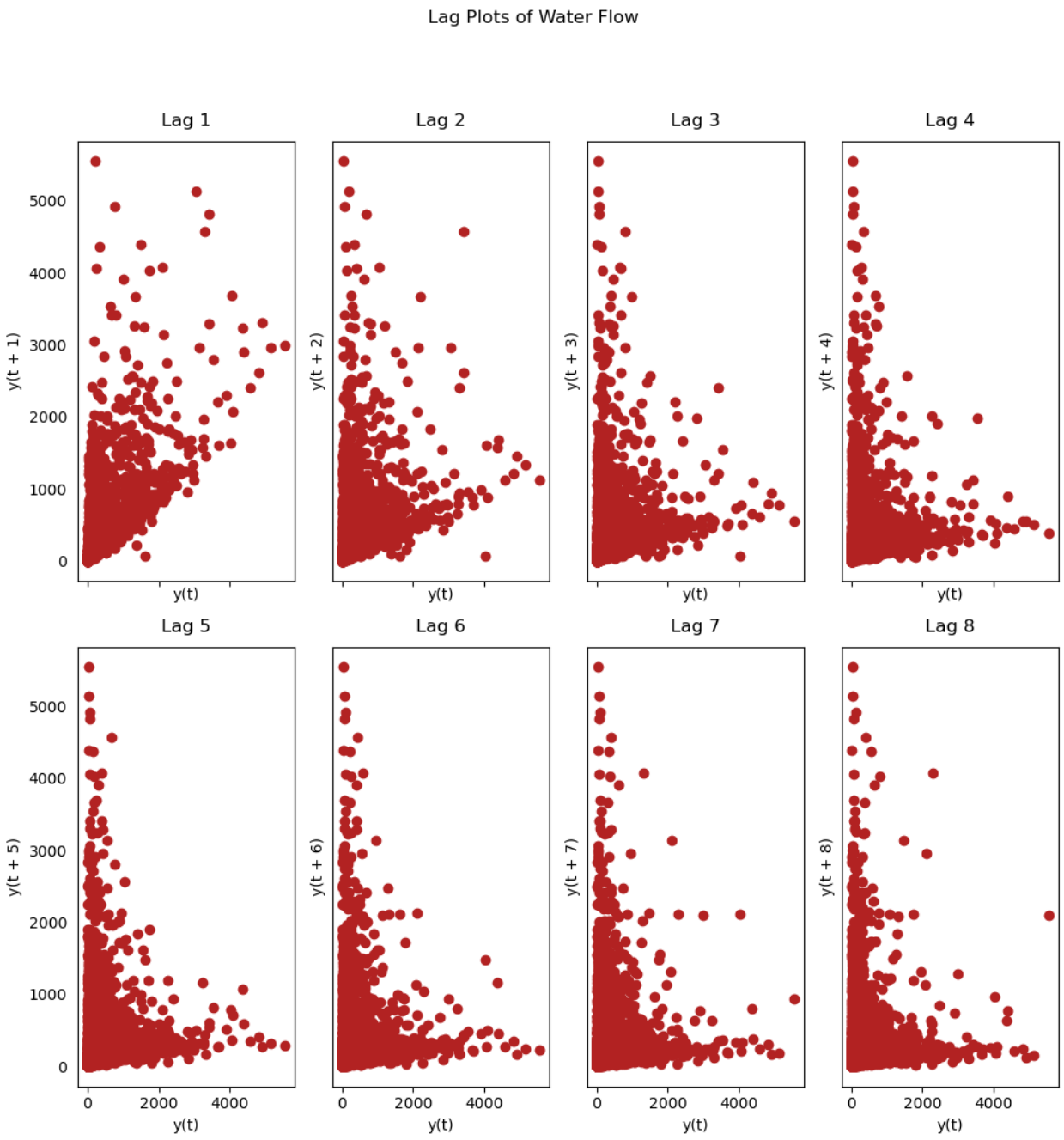
The first value is 1, because a correlation between two identical series was calculated. But take a look at as 8th period — autocorrelation value is 0.03. This tells you a value 8 periods ago has an impact compared with other lags on the value today.

The autocorrelation plot confirms our assumption about the correlation on lag 8, but on lag 16 seems not have an impact on the value today.



9. Lag Plots:

A **Lag plot** is a scatter plot of a time series against a lag of itself. It is normally used to check for autocorrelation. If there is any pattern existing in the series, the series is autocorrelated. If there is no such pattern, the series is likely to be random white noise.



10. Arima model:

A popular and widely used statistical method for time series forecasting is the ARIMA model. ARIMA is an acronym that stands for AutoRegressive Integrated Moving Average. It is a class of model that captures a suite of different standard temporal structures in time series data. The parameters of the ARIMA model are defined as follows and how to choose values:

- p: The number of lag observations included in the model, also called the lag order.

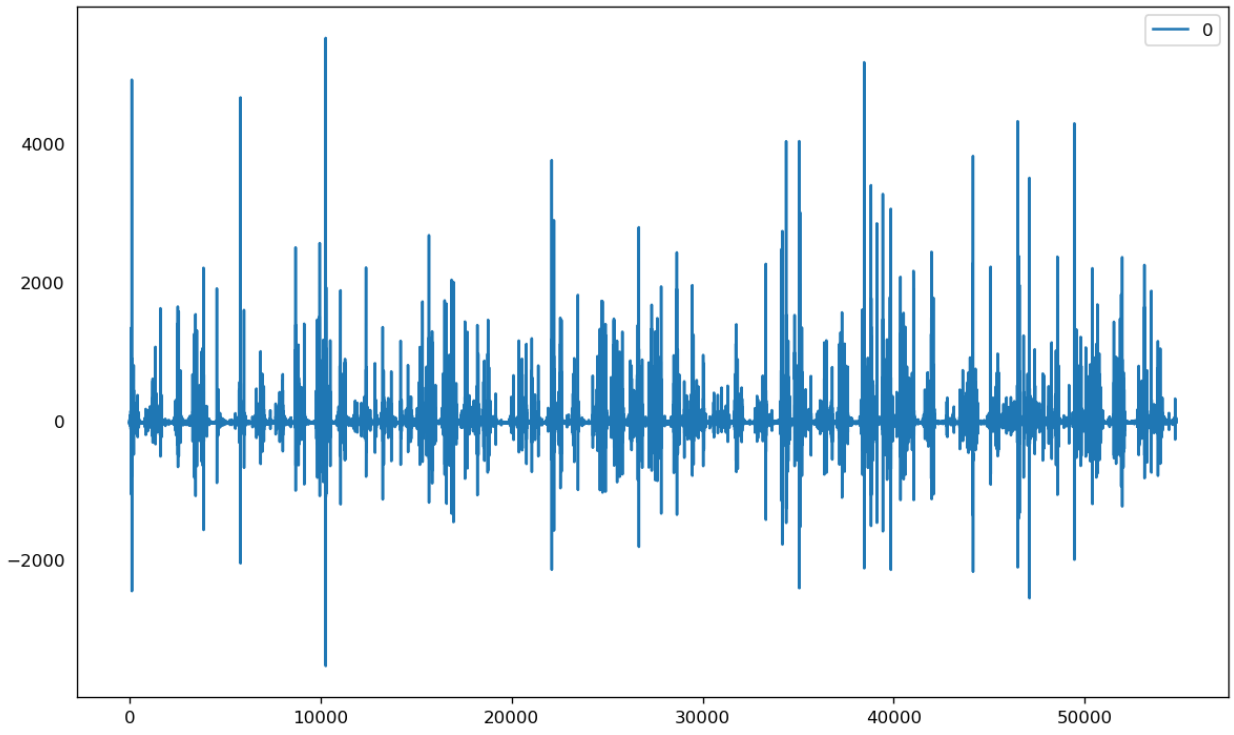
- d: The number of times that the raw observations are differenced, also called the degree of differencing.
- q: The size of the moving average window, also called the order of moving average.
- If the time series is stationary try to fit the ARMA model, and if the time series is non-stationary then seek the value of d.
- If the data is getting stationary then draw the autocorrelation and partial autocorrelation graph of the data.
- Draw a partial autocorrelation graph(ACF) of the data. This will help us in finding the value of p because the cut-off point to the PACF is p.
- Draw an autocorrelation graph(ACF) of the data. This will help us in finding the value of q because the cut-off point to the ACF is q.

I convert the Date into the float datatype. Therefore, the y_axis in the following photos are numbers, not the date format.
The year 1959-01-01 is converted into 0.0
The year 1959-01-02 is converted into 1,0

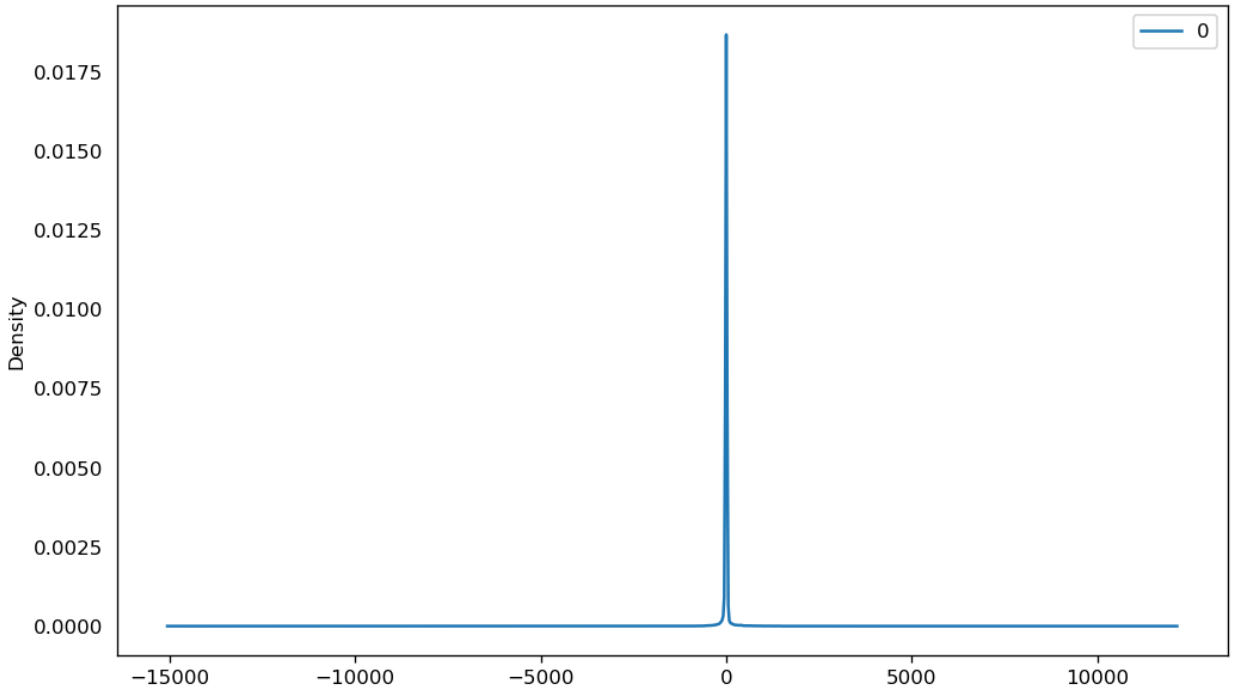
	Date	Flow
1959-01-01		57.0
1959-01-02		46.0
1959-01-03		31.0
1959-01-04		30.0
1959-01-05		28.0

	Date	Flow
0.0		57.0
1.0		46.0
2.0		31.0
3.0		30.0
4.0		28.0

Plot residual errors



Density plot of residuals



```
  warnings.warn(ARIMA_DEPRECATION_WARN, FutureWarning)
                              ARMA Model Results
==============================================================================
Dep. Variable:                      y   No. Observations:                36524
Model:                     ARMA(2, 1)   Log Likelihood             -232005.683
Method:                       css-mle   S.D. of innovations            138.793
Date:                Tue, 18 Apr 2023   AIC                         464021.366
Time:                        17:48:24   BIC                         464063.895
Sample:                             0   HQIC                        464034.885

==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const       5547.5057   2271.875      2.442      0.015    1094.712       1e+04
ar.L1.y    -8.146e-05   9.51e-05     -0.856      0.392      -0.000       0.000
ar.L2.y        0.9998   9.51e-05   1.05e+04      0.000       1.000       1.000
ma.L1.y       -0.0283      0.005     -5.458      0.000      -0.039      -0.018
                                    Roots
==============================================================================
                  Real          Imaginary           Modulus         Frequency
------------------------------------------------------------------------------
AR.1           -1.0000           +0.0000j            1.0000            0.5000
AR.2            1.0001           +0.0000j            1.0001            0.0000
MA.1           35.2987           +0.0000j           35.2987            0.0000
------------------------------------------------------------------------------
```

	0
count	36524.000000
mean	0.281453
std	148.225912
min	-8253.759937
25%	-2.804807
50%	0.104493
75%	2.527087
max	5340.008817

11. References:

https://www.kaggle.com/code/prashant111/complete-guide-on-time-series-analysis-in-python
https://www.simplilearn.com/tutorials/python-tutorial/time-series-analysis-in-python
https://analyticsindiamag.com/quick-way-to-find-p-d-and-q-values-for-arima/ (find p, d and q values for Arima model)
https://towardsdatascience.com/time-series-from-scratch-autocorrelation-and-partial-autocorrelation-explained-1dd641e3076f (autocorrelation and partial autocorrelation)
https://www.machinelearningplus.com/time-series/kpss-test-for-stationarity/ (KPSS test)