Module 5 Homework

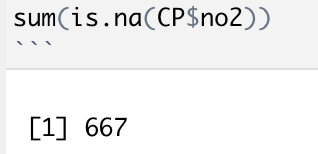
For this homework assignment, you will be exploring a real data set in preparation for modeling. The data set includes hourly air pollution information from the nationally controlled Changping air-quality monitoring site in China. The data were provided by the Beijing Municipal Environmental Monitoring Center. The meteorological data are sourced from the weather station nearest the monitoring site and were provided by the China Meteorological Administration. The data are from the time period March 1, 2013 to February 28, 2017. Missing data are denoted as NA.

Data Dictionary:

|  |  |
| --- | --- |
| Variable Name | Variable Description |
| No | Row number |
| year | Year |
| month | Month |
| day | Day |
| hour | Hour |
| PM2.5 | PM2.5 concentration (ug/m^3) |
| PM10 | PM10 concentration (ug/m^3) |
| SO2 | SO2 concentration (ug/m^3) |
| NO2 | NO2 concentration (ug/m^3) |
| CO | CO concentration (ug/m^3) |
| O3 | O3 concentration (ug/m^3) |
| TEMP | Temperature (degrees Celsius) |
| PRES | Barometric pressure (hPa) |
| DEWP | Dew point temperature (degrees Celsius) |
| RAIN | Precipitation (mm) |
| wd | Wind direction |
| WSPM | Wind speed (m/s) |
| Station | Name of the air quality monitoring site |

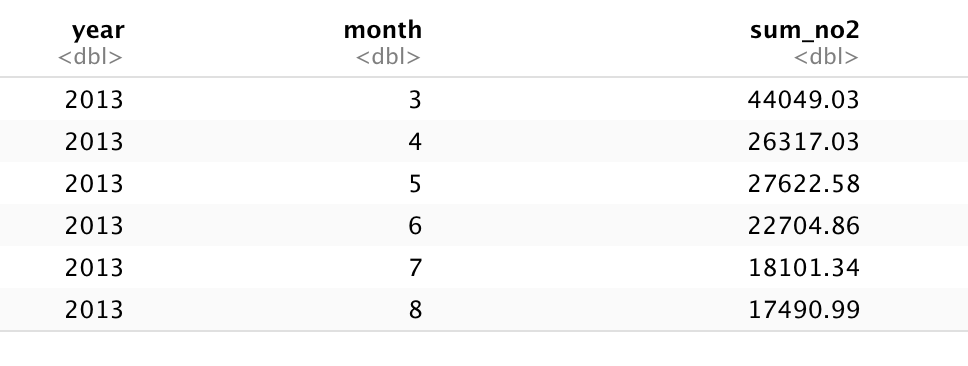
You will need to apply all of the data preparation and exploratory analysis techniques that you have learned up to this point in order to understand the patterns that are present in the data.

1. Read the .csv file into R. Are there any missing values present in the data?

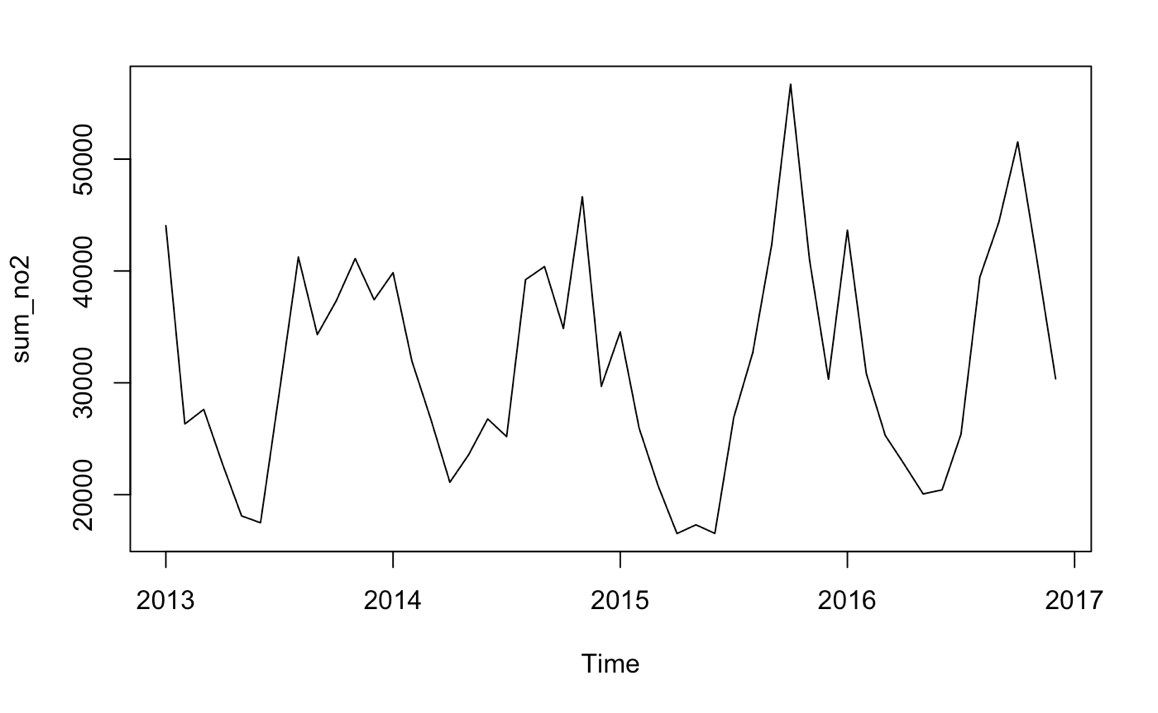


There are 667 missing values of no2 in dataset.

1. Your ultimate goal with these data is to produce a monthly forecast of the total NO2 concentration at the Changping air quality monitoring station. With this in mind, you will need to accumulate the data using a monthly time index before continuing with your analysis. Use the head function in R to print out the first few rows of the monthly data series and copy and paste them below.

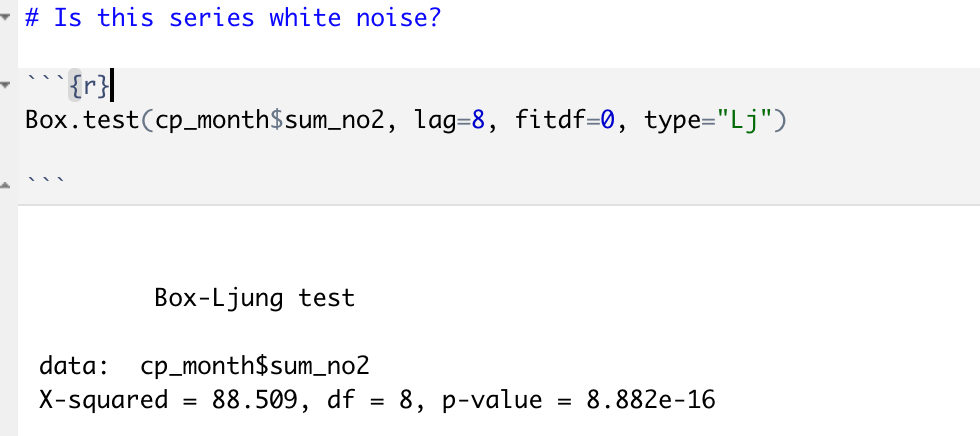


1. Create a time series object from the monthly data and generate a time plot of data. Copy and paste the plot below. How would you describe this series (e.g., are there any outstanding features of the series – outliers, trend, seasonality)?



The time series plot shows seasonality in NO2. Every year there is a trough around August and a peak around March.

1. Is this a white noise series? Paste the relevant output from R below. Be sure to state your hypotheses and along with your conclusion.

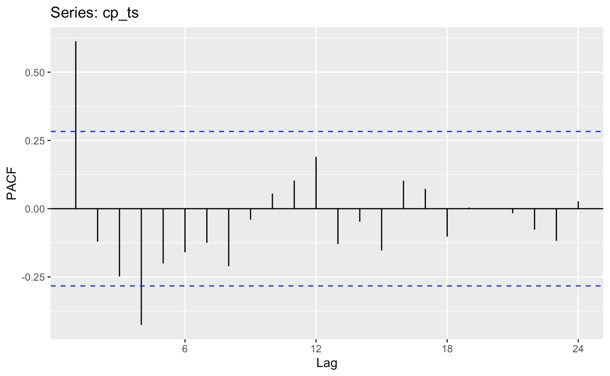
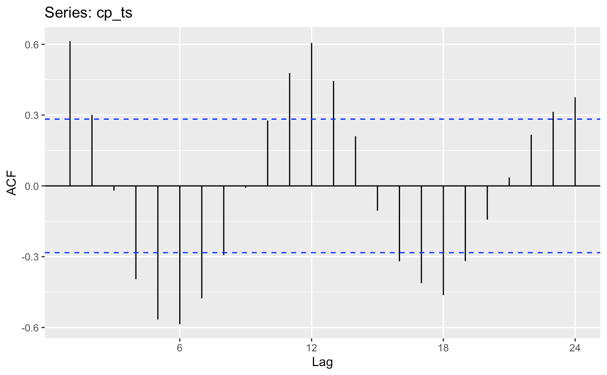


Ho: This series is white noise.

Ha: This series is not white noise.

It is not a white noise since the p-value<0.05, which reject the Ho.

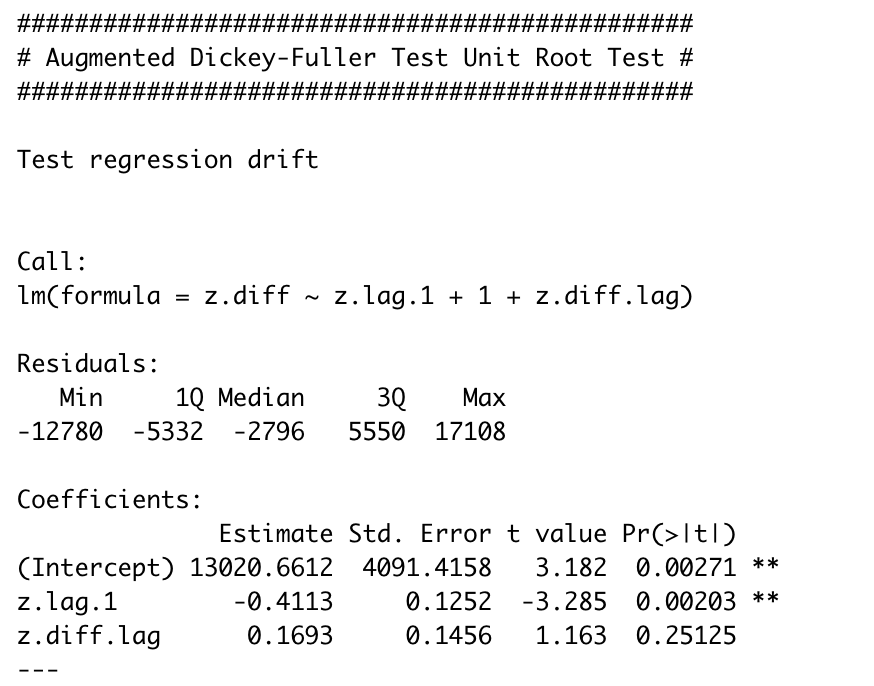
1. Generate plots of the autocorrelation function and the partial autocorrelation function for this series. Copy and paste the plots below. What do you learn about the series by looking at these plots?



ACF: The ACF shows a quick decay to 0. It appears to have a peak every 6 periods. Yt seems to have a strong relationship with Yt-6, Yt-12, etc. Also, it shows Yt has a strong relationship with Yt-1.

PACF: The PACF tells that Yt has a strong relationship with Yt-1.

1. Is this series stationary? Paste the relevant output from R below. Be sure to state your hypotheses along with your conclusion.



Ho: The series is non-stationary.

Ha: The series is stationary.

The p-value<0.05 rejects the null hypothesis and shows that series is stationary.

The next step in the process will be to build an ARIMA model for the data. We will come back to these data in a later modules to work through the model building and forecasting process.