Yuyi Zhang

May 12, 2019

Spring 2019 CPS Quarter Term A

Instructor: Steward Huang

Assignment 5: Time Series Analysis

ALY 6015\_Intermediate Analytics

# **Introduction**

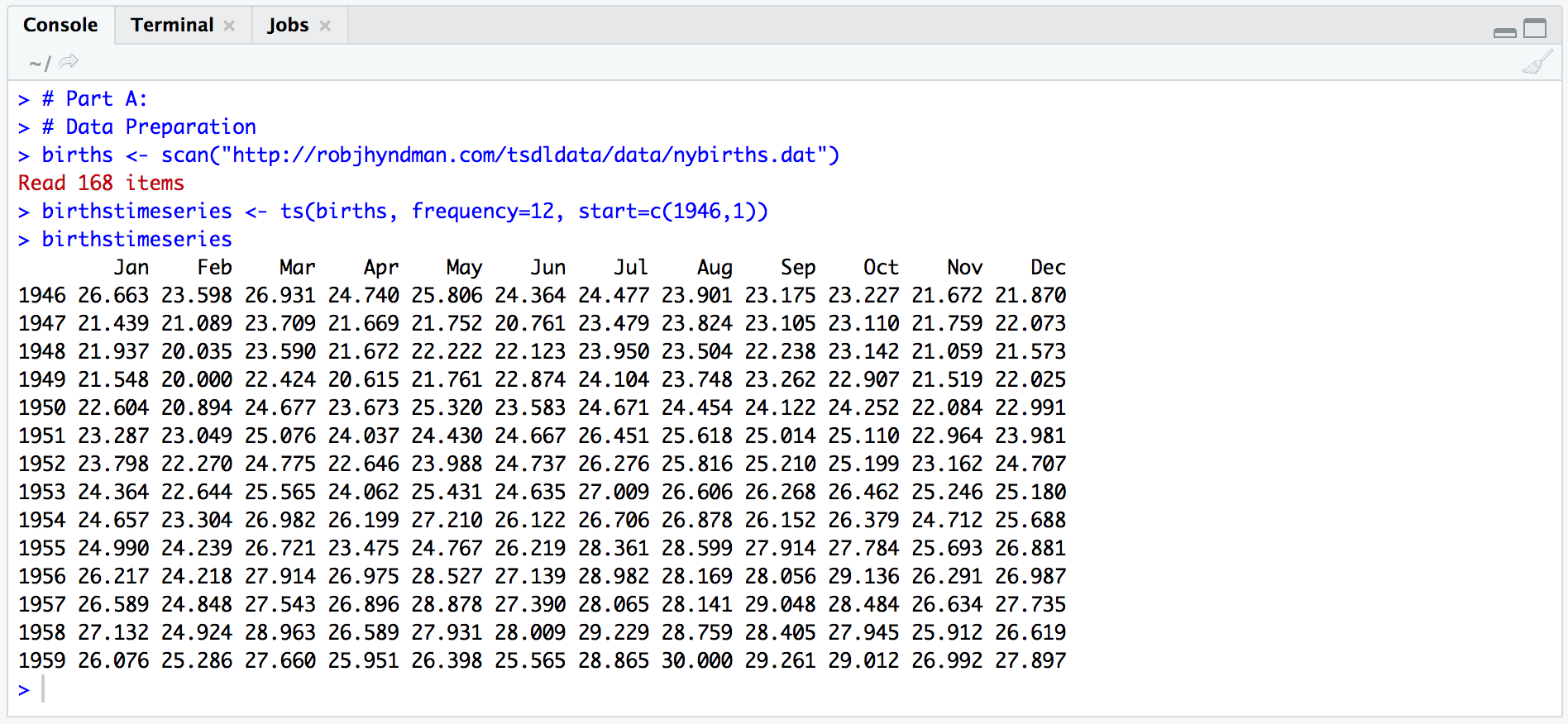
In this report, we are going to go through two different dataset and apply time series analysis to them by decomposing and ARIMA preparation. The first dataset called “births”, which is the number of births per month in New York City. The second dataset called “valcanodust” is the data on the volcanic dust veil index in the northern hemisphere from 1500-1969.

# **Analysis**

## **Part A: Decomposing**

### **Data preparation**

In this phase, we are going use dataset “births” and get to see the output of the snippet.



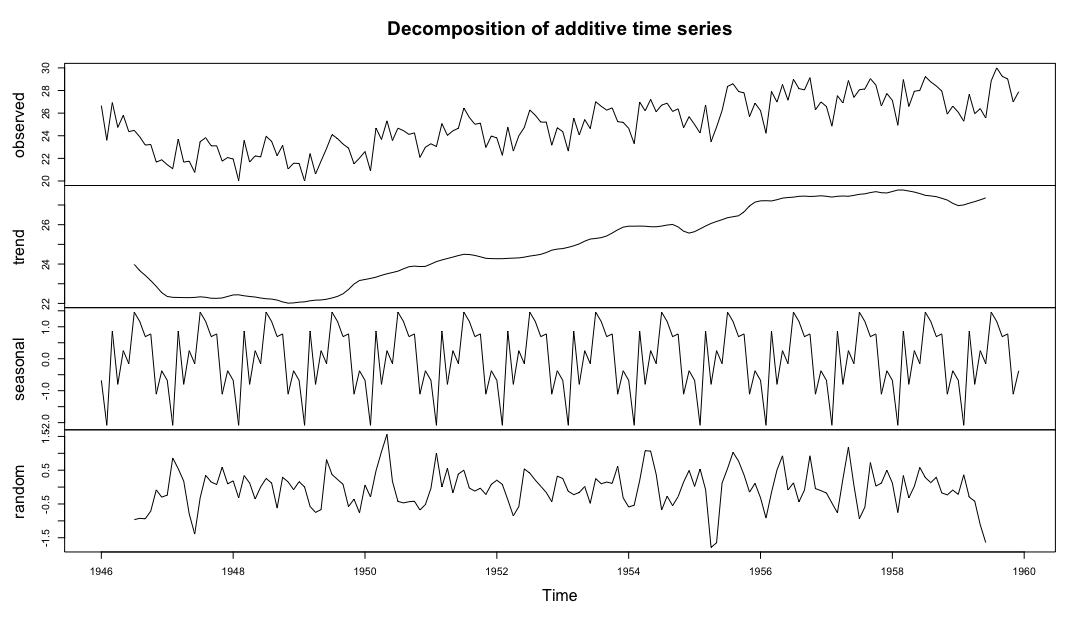
*Figure 1*. Data preparation

### **Decomposing**

This dataset is a seasonal time series consists of a trend component, a seasonal component and an irregular component. So in this process, we are going to separate the time series into three components: trend, seasonal and irregular.



*Figure 2*. Decomposing

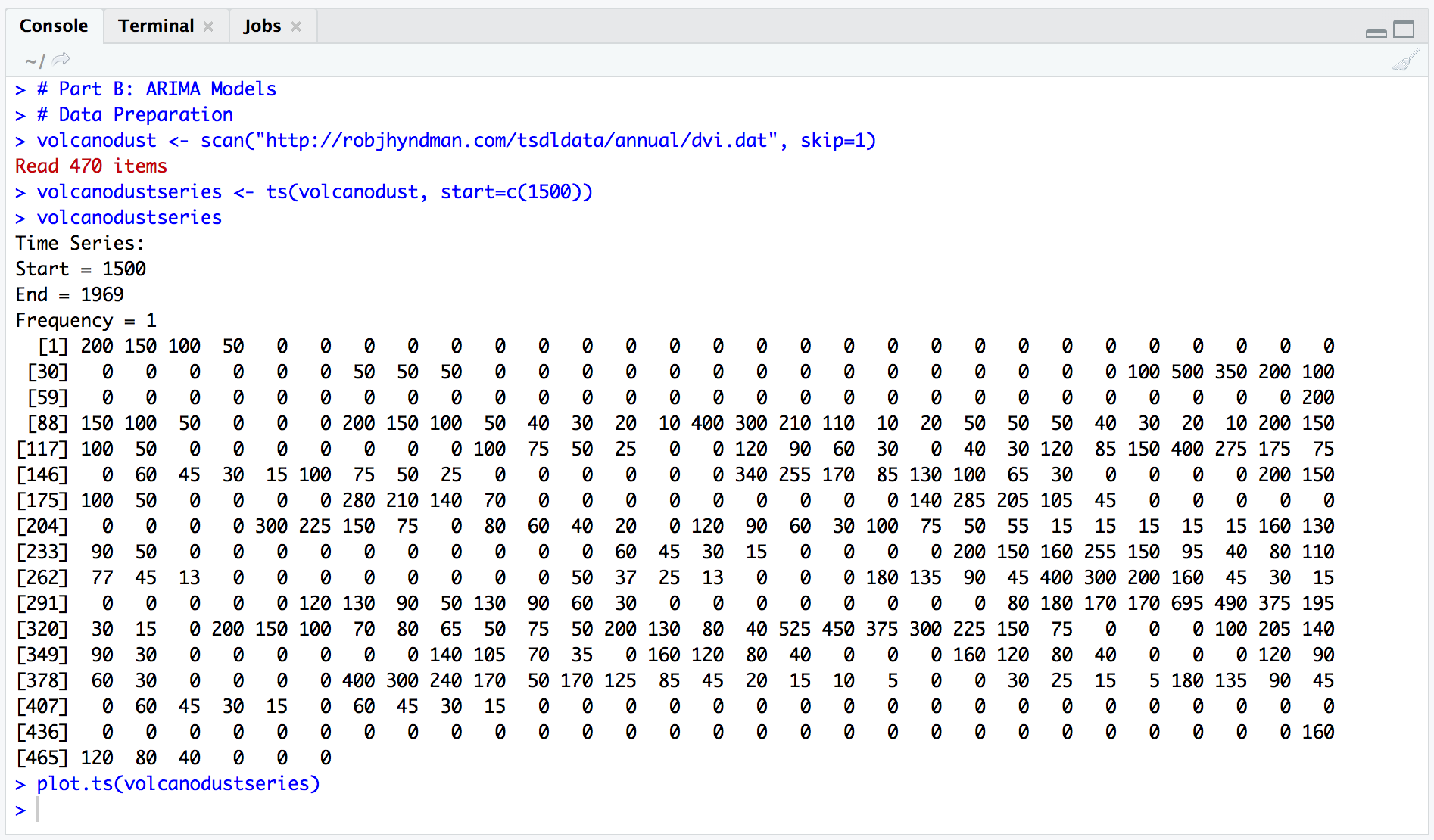


*Figure 3*. Decomposition plots

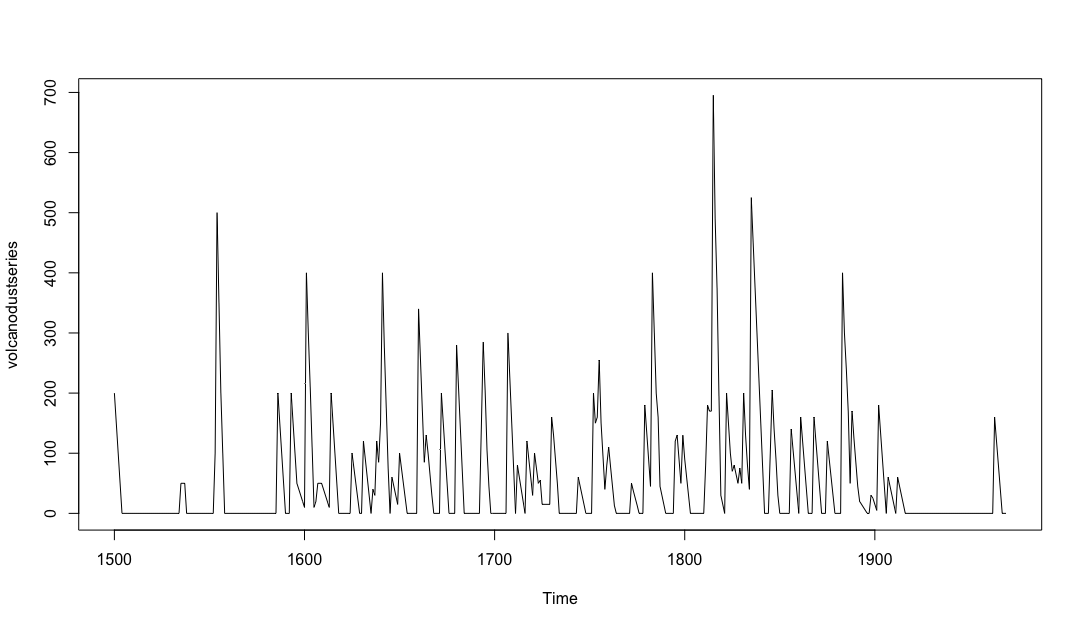
This plot shows the original time series and three components we mentioned before. As expected, the trend shows overall increase. There is a highly consistent seasonal cycle as shown in the third plot from top. Irregular noise shown in the bottom.

## **Part B: ARIMA Models**

### **Data preparation**



*Figure 4*. Data Preparation

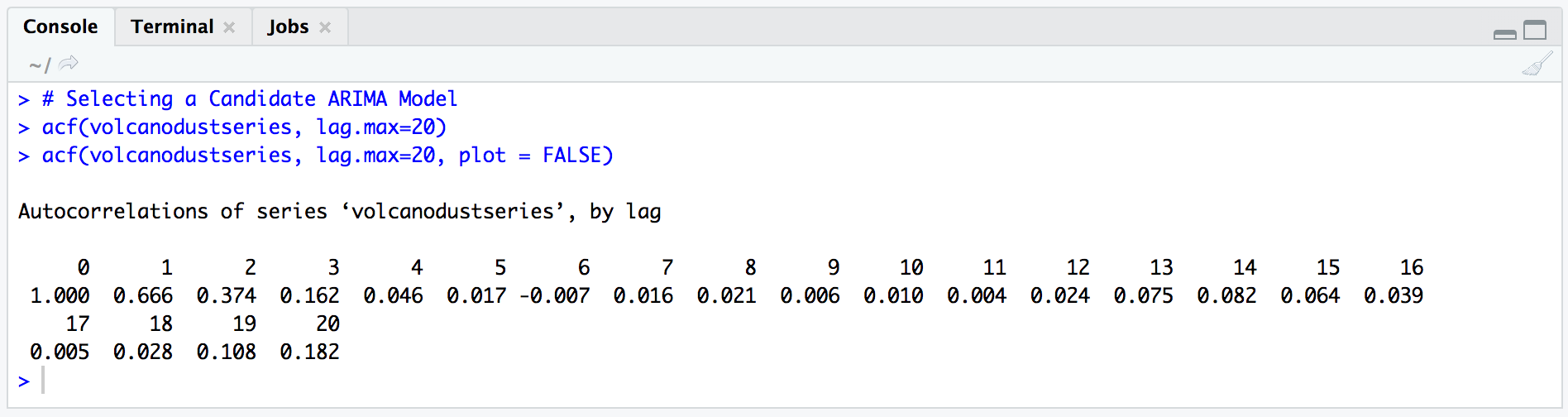


*Figure 5*. Volcano Dust Series

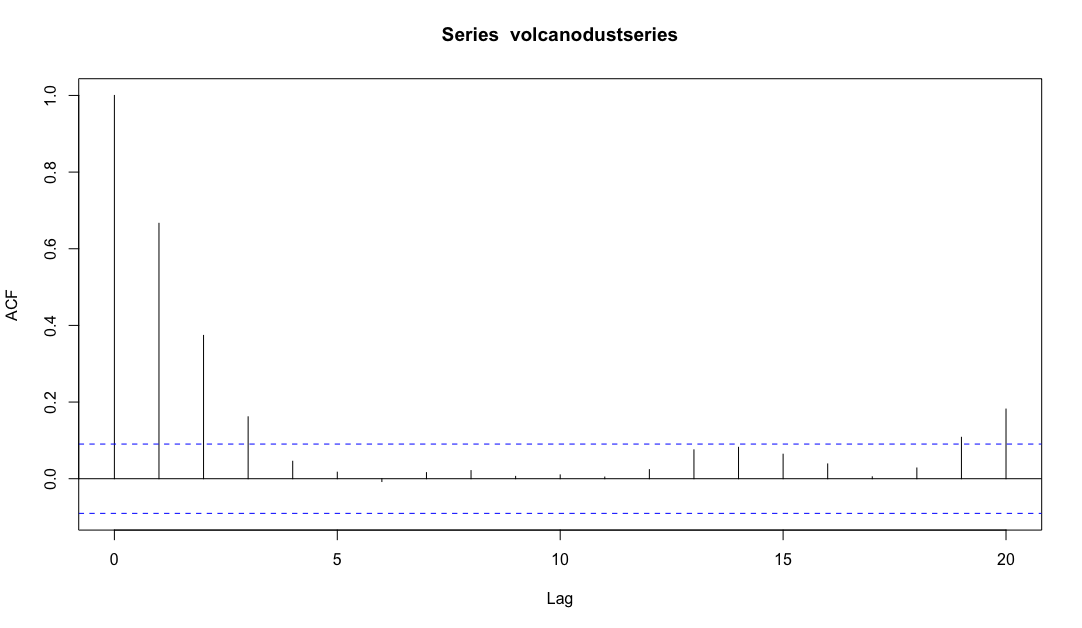
This time series appears to be stationary in mean and variance. Therefore we don’t need to different the series in order to fit an ARIMA model. In another word, for ARIMA(p, d, q) model, the value of d equals 0. Then we are going to examine the correlogram and partial correlogram of the stationary time series.

### **Selecting a Candidate ARIMA Model**

In this step, we are going to plot a correlogram for lags 1-20.



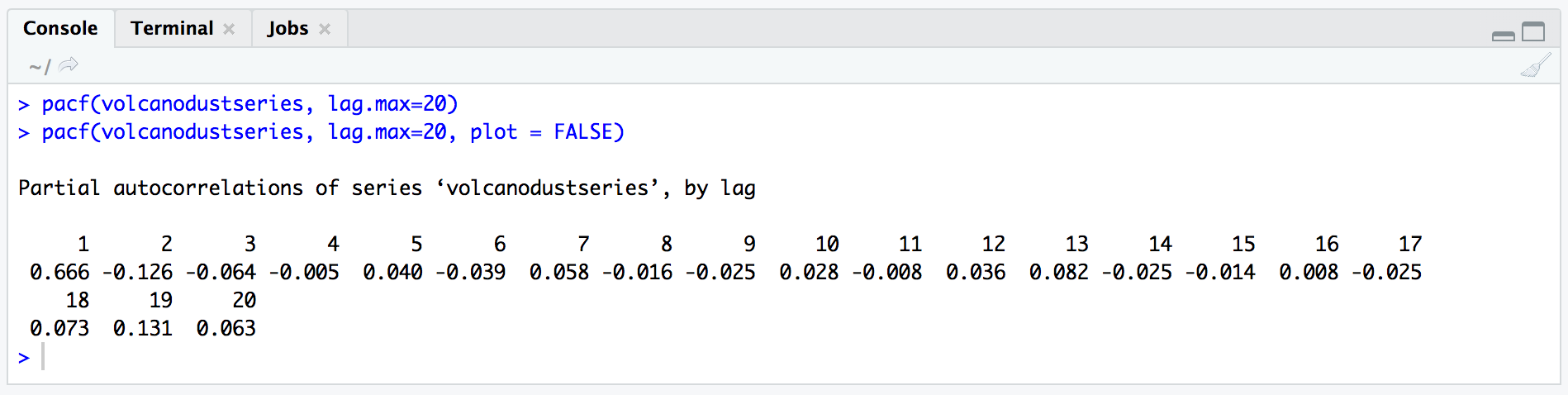
*Figure 6*. Autocorrelations



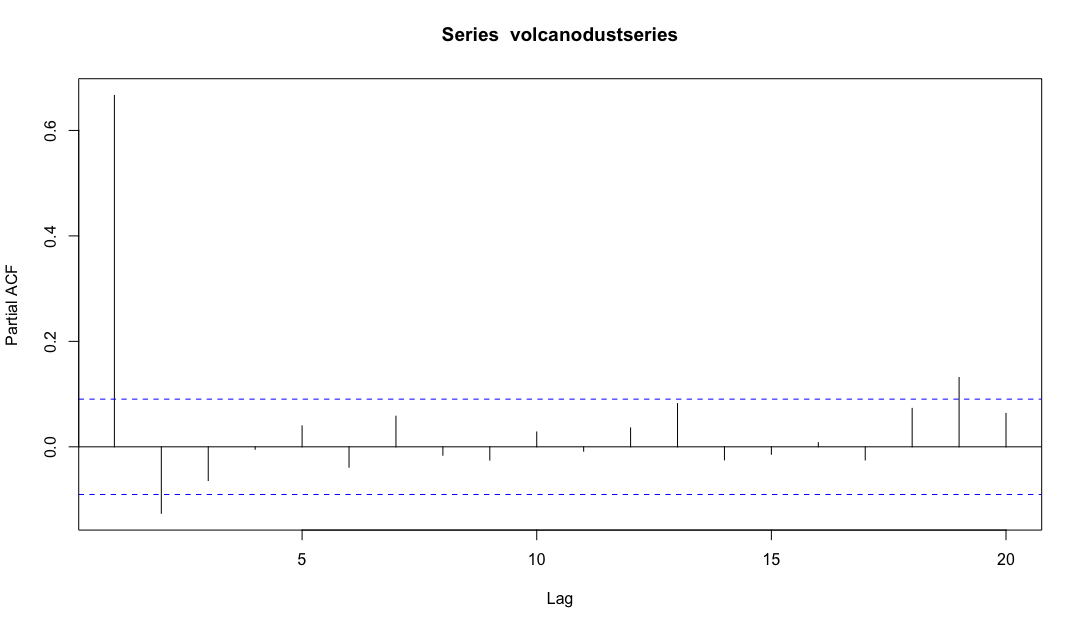
*Figure 7*. acf lag plot

Now we are going to find the value q of ARIMA(p, 0, q) model. As we can see that lag 0-2 exceed the significance bounds. So does lag 3, but we consider that this is due to chance since lag 3 and lag 19, 20 just exceed the 95% significance bounds quite a bit. Therefore, the q value we are going to go with 2, since after lag 2, since the autocorrelogram is near zero after lag 2.

Then we are going to exam the partial autocorrelogram.



*Figure 8*. Partial Autocorrelations



*Figure 9*. pacf lag plot

In this step, we could see that the partial autocorrelation at lag 1 is way exceed the significance bounds. Lag 2, however, is exceed the 95 significance bounds just a little bit like the ones in figure 7. So we take this one as an inbound number but went out by chance. Therefore the p value of ARIMA model is 1 there.

The way to final decide the most appropriate values for ARIMA model is to consider the fewest parameters is best. In this case, ARIMA (1, 0, 2) wins.

### **Auto Calculate ARIMA Model**

We could also use the auto.arima() function to double check what we choose manually.



*Figure 10*. Auto Calculate ARIMA model

As shown above, we’ve got ARIMA(1, 0, 2) with non-zero mean model. The same result as what we’ve got in 2.2.2.

# **Conclusion**

Both datasets are showing us the preparation steps for ARIMA forecasting. In part A, we generated a decompose plot to see whether the dataset is seasonal time series. In part B, we found the most appropriate p, d, q values for ARIMA model by both manual way and automatic way. And the calculation results matched.

Reference

1. Coghlan, A. (2018). *A Little Book of R for Time Series*. Retrieved from <https://media.readthedocs.org/pdf/a-little-book-of-r-for-time-series/latest/a-little-book-of-r-for-time-series.pdf>