Time Series Analysis

Capstone II

Priya Theru

Springboard April 2019
Data Science Career Track

Background

- Forecasting the future values has always been one of the most challenging problems in the field. More specifically, the primary goal here is to focus on the problem of forecasting future values of number of ads being watched per hour in a real mobile game.

Dataset

- The Ads dataset consists of approximately 215 time series. Each of these time series represents ads being watched per hour starting from September 13 2017 to September 22 2017. The currency dataset consists of 300 time series. Each of the time series represents in-game currency being spent per day starting from May 1st 2017 to February 24th 2018.

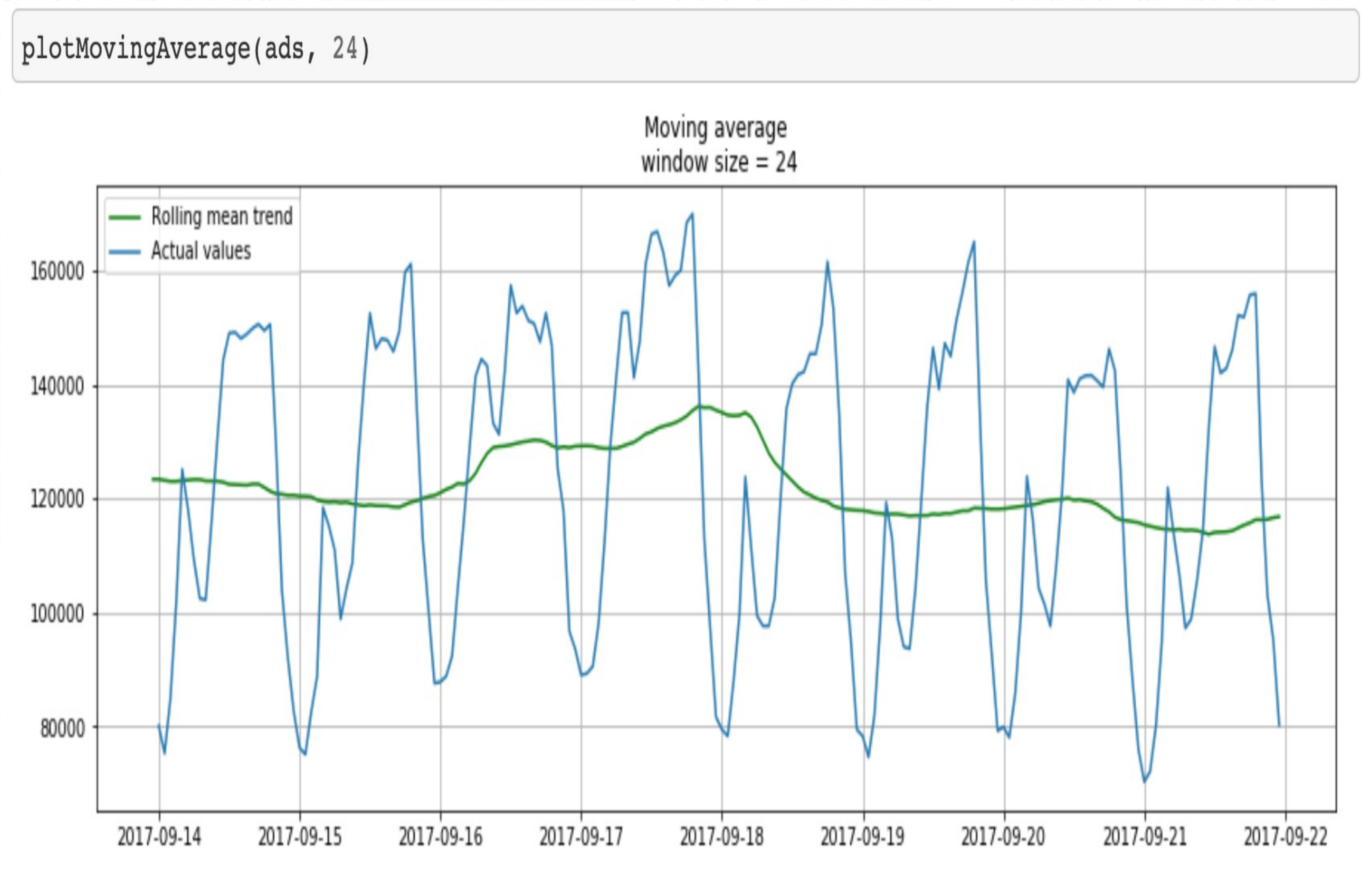
Steps

- Moving Average
- Stationarity
- Forecasting using ARIMA

Moving Average

- The initial approach towards forecasting is the Moving Average which begins with a naive hypothesis : “tomorrow will be the same as today”. The future value of the dependent variable depends on the average of its previous values.
- When daily smoothing was applied on hourly data, the values are higher during the weekends as more time to play on the weekends while fewer ads on weekdays.

Moving Average Contd.

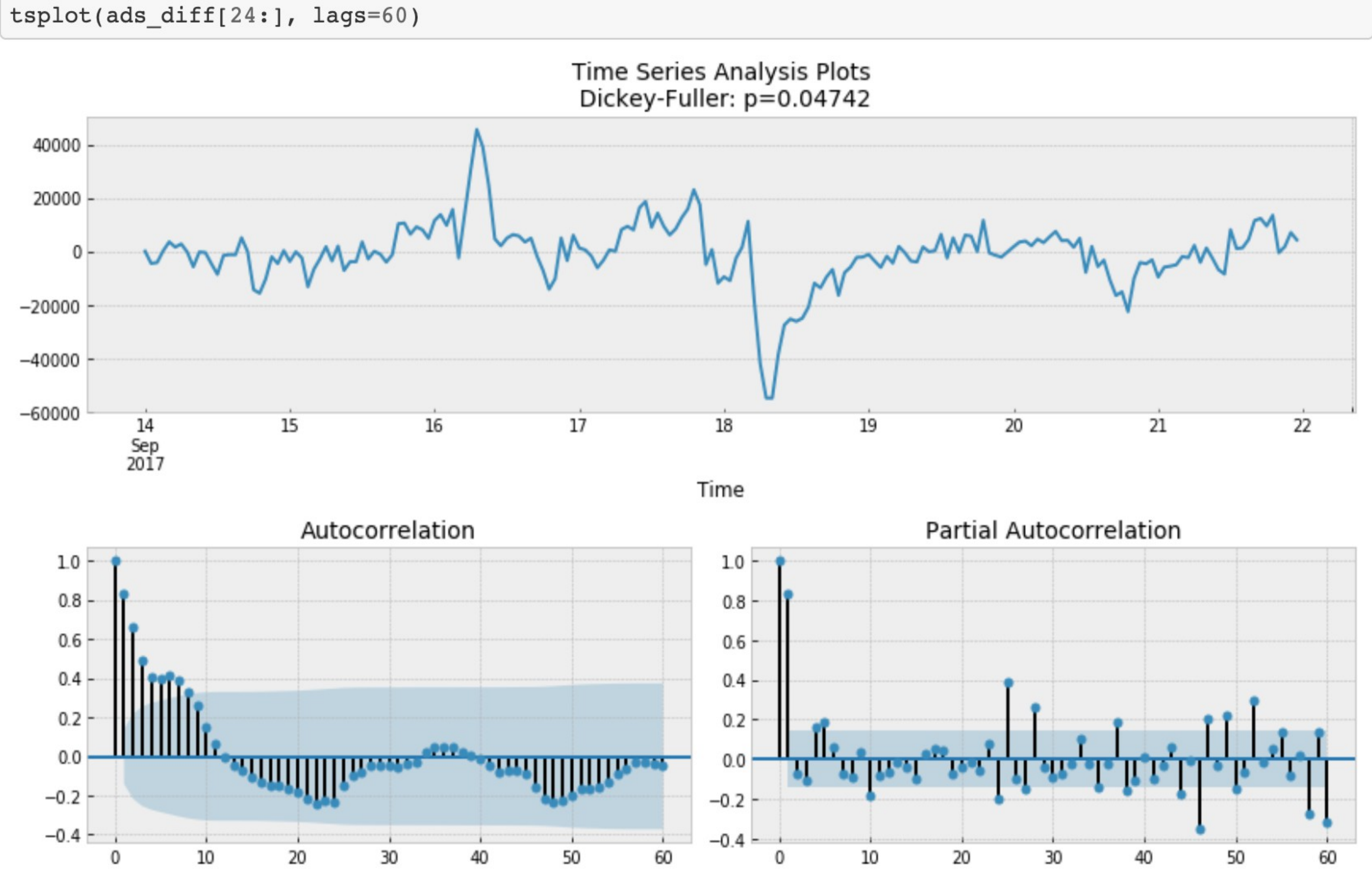


```
plotMovingAverage(ads, 24)
```

Stationarity

- An important property of time series is stationarity which basically means that it doesn't not change its statistical properties over time specifically, the mean and the variance. And we can fight non-stationarity using one such popular method – differences , which is the main idea behind Dickey-Fuller test for stationarity. A custom code is written to conduct the stationarity test.

Stationarity Contd.



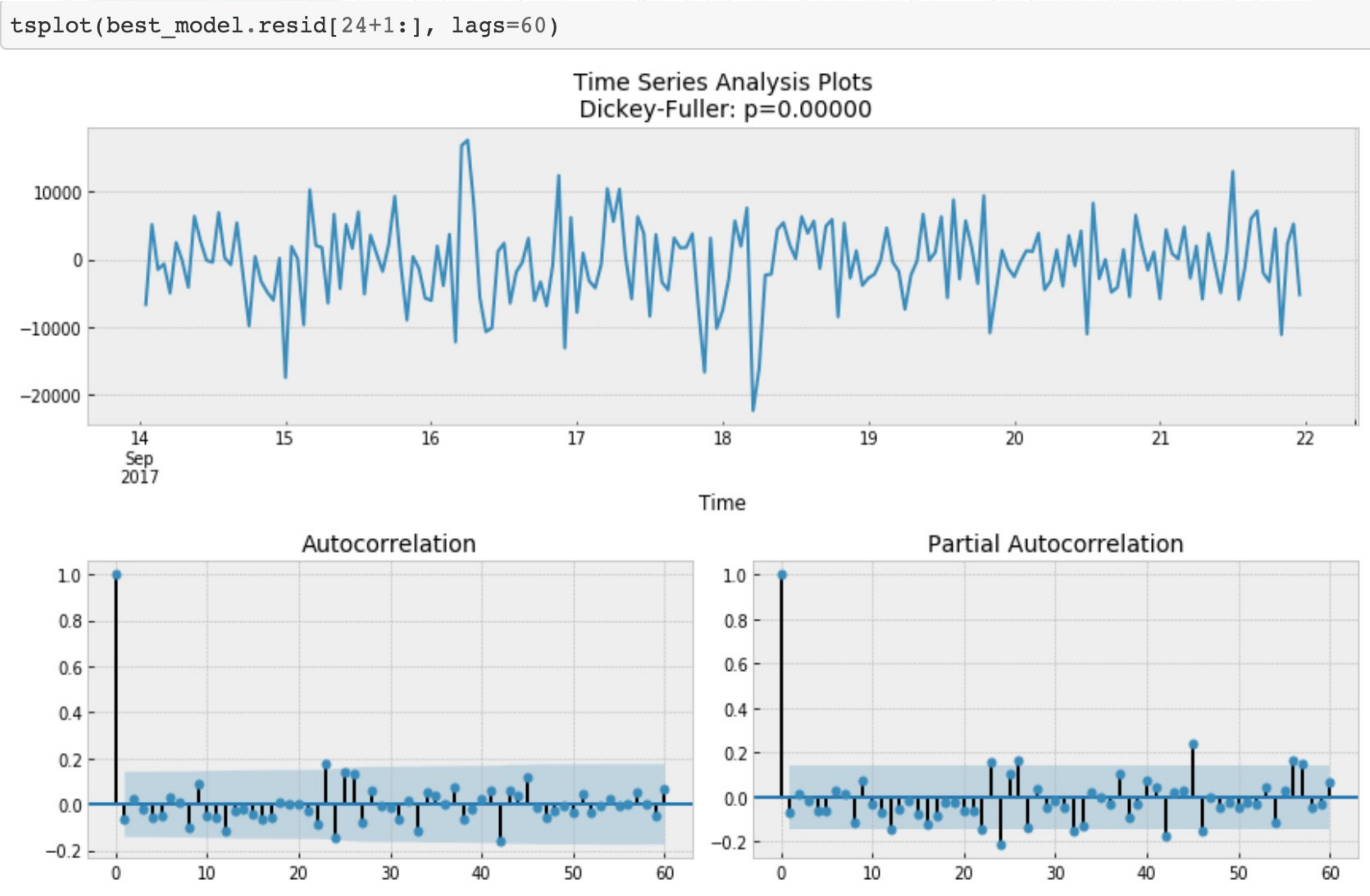
```
tsplot(ads_diff[24:], lags=60)
```

ARIMA

- SARIMA(p, d, q)(P, D, Q, s), - Seasonal Autoregression Moving Average model

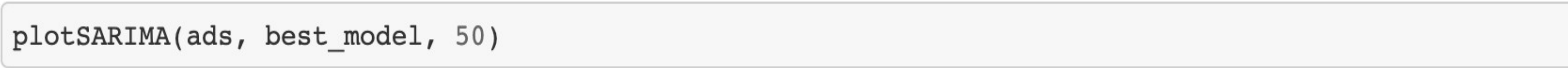
- AR(p) - autoregression model i.e. regression of the time series onto itself. The basic assumption is that the current series values depend on its previous values with some lag (or several lags). The maximum lag in the model is referred to as p. To determine the initial p, you need to look at the PACF plot and find the biggest significant lag after which most other lags become insignificant.
- MA(q)- moving average model. Without going into too much detail, this models the error of the time series, again with the assumption that the current error depends on the previous with some lag, which is referred to as q. The initial value can be found on the ACF plot with the same logic as before.

ARIMA Contd.



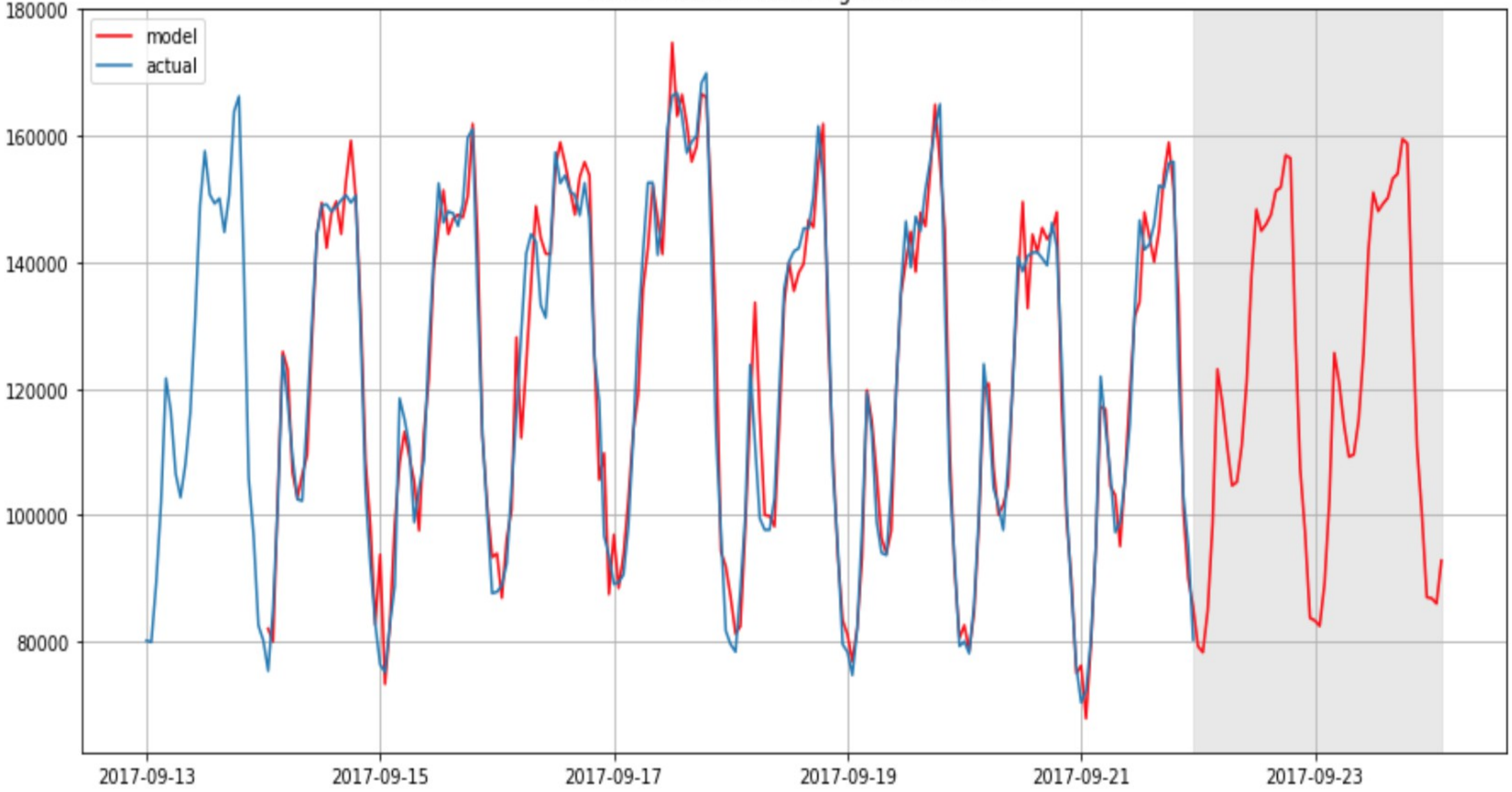
```
tsplot(best_model.resid[24+1:], lags=60)
```

ARIMA Contd.



```
plotSARIMA(ads, best_model, 50)
```

Mean Absolute Percentage Error: 3.94%



In the end, we got very adequate predictions. Our model was wrong by 3.94% on average, which is very good.