**GOAL1**

Sberbank… blah blah

**GOAL2**

* Introduction

Sberbank, one of Russia’s largest bank, would like to predict average house prices based on historic trend. We used the time series analysis technic on aggregated monthly data to solve this problem

* Data Wrangling
  1. The timestamp if each house purchase record are written in the number of days since 12/30/1899. So we transform the data into the familiar YYYY-MM-DD format with R

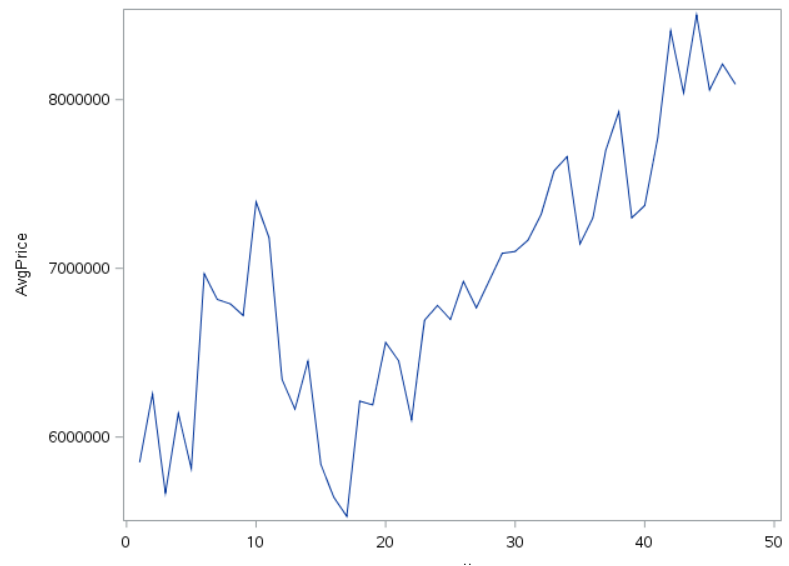
|  |
| --- |
| # Select ID, timestamp and price, and transform to YYYY-MM-DD  goal2\_data <- modelingData[c("id", "timestamp", "price\_doc")]  goal2\_data$timestamp <- as.Date(goal2\_data$timestamp, origin = "1899-12-30") |

* 1. With Aggregate function in R, we could get the average monthly price. And we also define sequential month number to make the time series analysis easier

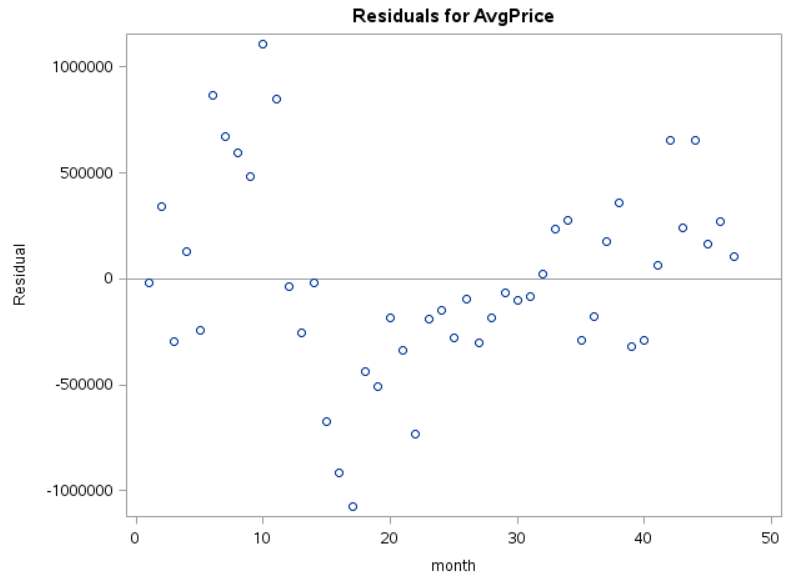
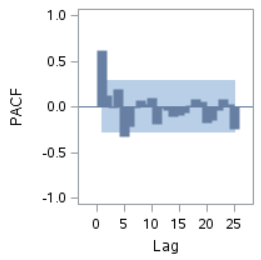
|  |
| --- |
| # Aggregate by month-year  goal2\_agg <- aggregate(goal2\_data, by=list(goal2\_data$monthYear), FUN = mean)  goal2\_agg <- goal2\_agg[c("Group.1", "price\_doc")]  names(goal2\_agg) <- c("monthYear", "AvgPrice")  goal2\_agg$monthNumber <- seq.int(nrow(goal2\_agg)) |

* Plotting Time Series

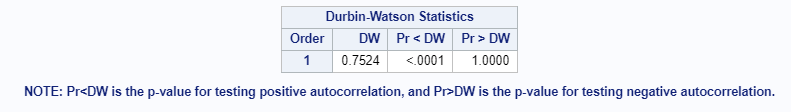
The plotting and subsequent analysis is conducted with the SAS tool. We can see the overall trend of the mean price increase accompanied with oscillating ups and downs.



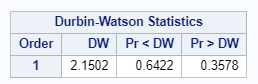
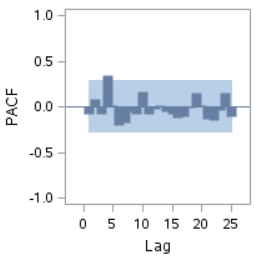
* Residual Series Modeling
  1. When fitting the data to OLS, we see the residuals are clustered along recent past. Positive residuals follow another positive residuals from previous observation, and so forth for negative residuals – hunch for autocorrelation
  2. The PACF result also suggest the autocorrelation. So we will Durbin-Watson statistics

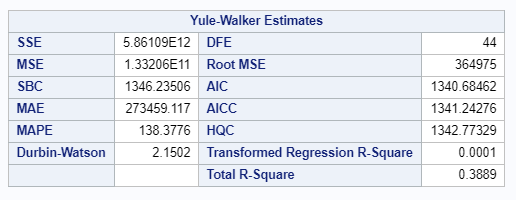
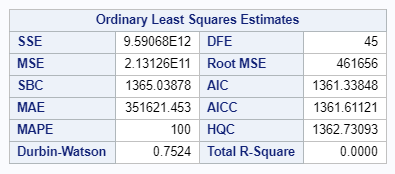
* 1. DW statistics value and Pr suggest there is a positive autocorrelation



* 1. Once corrected with AR(1), we see the residual autocorrelation is corrected based on DW value near to 2.0

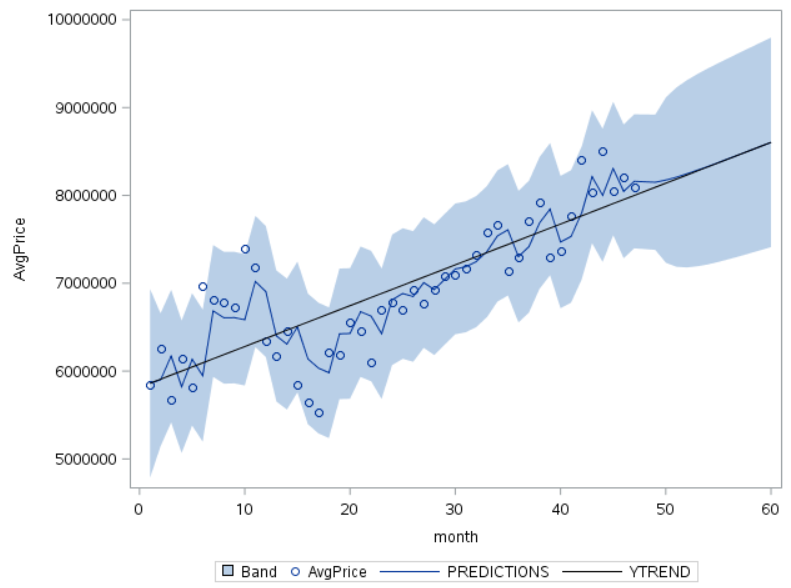
* 1. When comparing the AIC and SBC, we see corrected models are favored compared to the OLS. The figures are lower with Y-W estimates



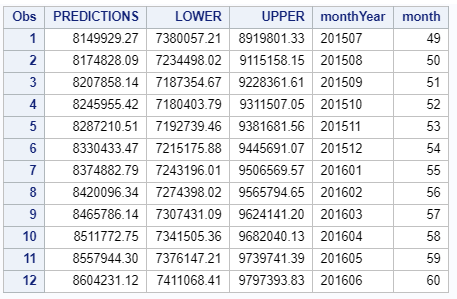
* Average House Price Prediction
  1. The monthly average price from July 2015 to June 2016 are predicted with the corrected estimates from SAS

|  |
| --- |
| PROC AUTOREG DATA=GOAL2\_PRED ;  MODEL AVGPRICE = MONTH / NLAG = (1) DWPROB;  OUTPUT OUT = PREDS\_PRICE P = PREDICTIONS LCL = LOWER UCL = UPPER PM = YTREND;  RUN; |

* 1. The SAS package help combine the trend estimate with the forecast residual analyzed in previous steps. The plot with 95% confidence interval, predictions and trend values are as following



* 1. Predictions of the housing price with 95% confidence interval



**Appendix 1: Data Dictionary**

Data dictionary

**Appendix 2: Code**

1. Goal 1 Codes
2. Goal 2 Codes