MATH2250 Time Series Final Project

Li, Yihao

December 13, 2020

1 Abstract

This project aim to describe the techniques I have seen in the course on a time series data in R. I will look at the linear regression on the data set, visual inspection of the data, the forecast model to forecast one year data and the conclusion about it.

2 Data Description

In this project, I use Shanghai Car License Plate Auction Price data set from `www.kaggle.com`. Shanghai uses an auction system to sell a limited number of license plates to fossil-fuel car buyers every month. The average price of this license plate is about 13,000 US dollars and it is often referred to as ”the most expensive piece of metal in the world.” So, my goal is to predict the avg price or the lowest price for the next month. In this project, I picked the average price as my data to predict the future price in next 12 months.

3 Methodology

In this project, my main objective was to find a model to efficiently forecast the price of Shanghai car license plate. The suitable forecasting methods were chosen for finding the method that was suitable for short term analysis in monthly. Therefore, I built three models and compare them to find the best model to forecast the average price for Shanghai Car License plate. The three models are ARIMA model, Holt linear method, and Holt Winter method.

For this project, I can explain in detail of each step as follows.

3.1 Data Prepossessing

This project used data set about Shanghai Car License plate price in each month over a long period of time from the years 2002 to 2019. I used R and Rstudio for building the model. The first step is to read the csv file with the following R commands:

```
sh_plate<-read.csv(”shanghai_plate_price.csv”)
sh_plate<-sh_plate[,c(-2,-3,-5)]
```

The raw data are not ready for constructing forecasting model because based on the figure 1 it shows that it is not a stationary data set. Moreover, since the original data set is a monthly data set, we add the each date for every month from Jan 2002 to March 2019. The data set has 204 observations, I used the following R code to implement this and the result is showing in the figure 1 as well:

```
d <- as.Date(”2001-12-1”) + seq(0,6100,30)
next.month <- function(d) as.Date(as.yearmon(d) + 1/12) +
  as.numeric(d - as.Date(as.yearmon(d)))
next.month(d)
avg_price<-data.frame(date = next.month(d)+0:203,avgprice)
```

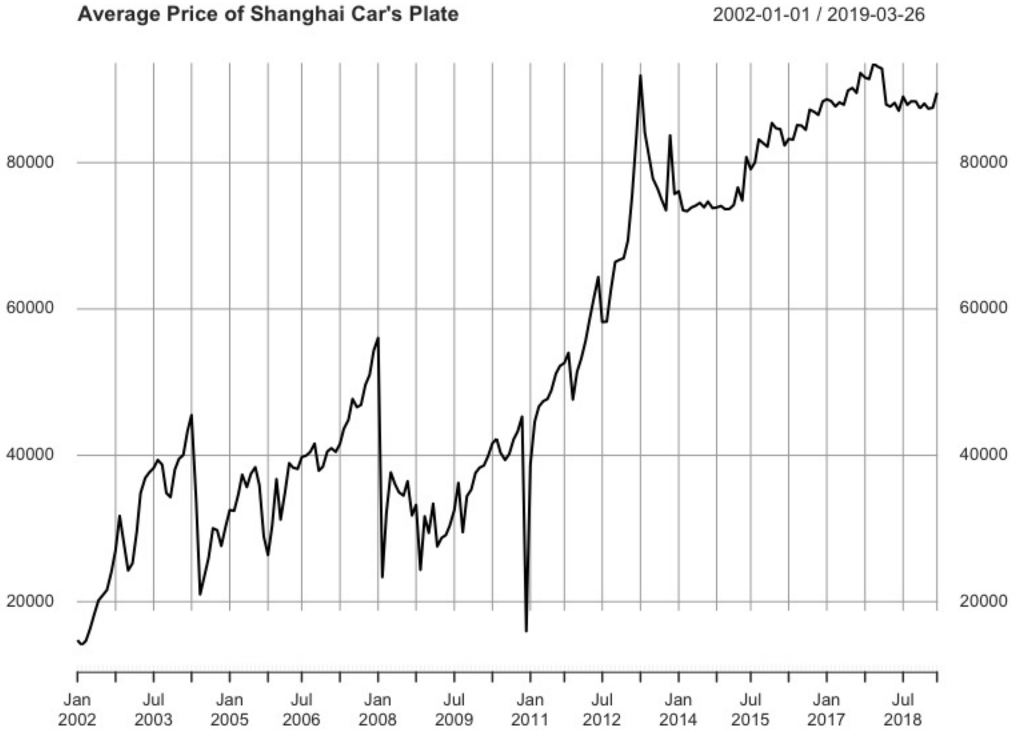


figure 1: The Average Price of Shanghai Car Plate

3.2 Constructing and decomposing time series format

This step uses the data series that are in the appropriate format for time series construction and decomposition. I used the ts() function in R library for construction of a time series. This function must be specifying a frequency of time series. In this project, since I only used the monthly data, I used 12 as my frequency to indicate that the time series is a monthly series.

```
a<-ts(avg_price ,start = c(2002,1),frequency = 12)
```

Then, the data is ready for decomposition. This step is to decompose a time series into trend, seasonal, cyclical and irregular components by applying the decomposed() function as below:

```
plot(decompose(a))
```

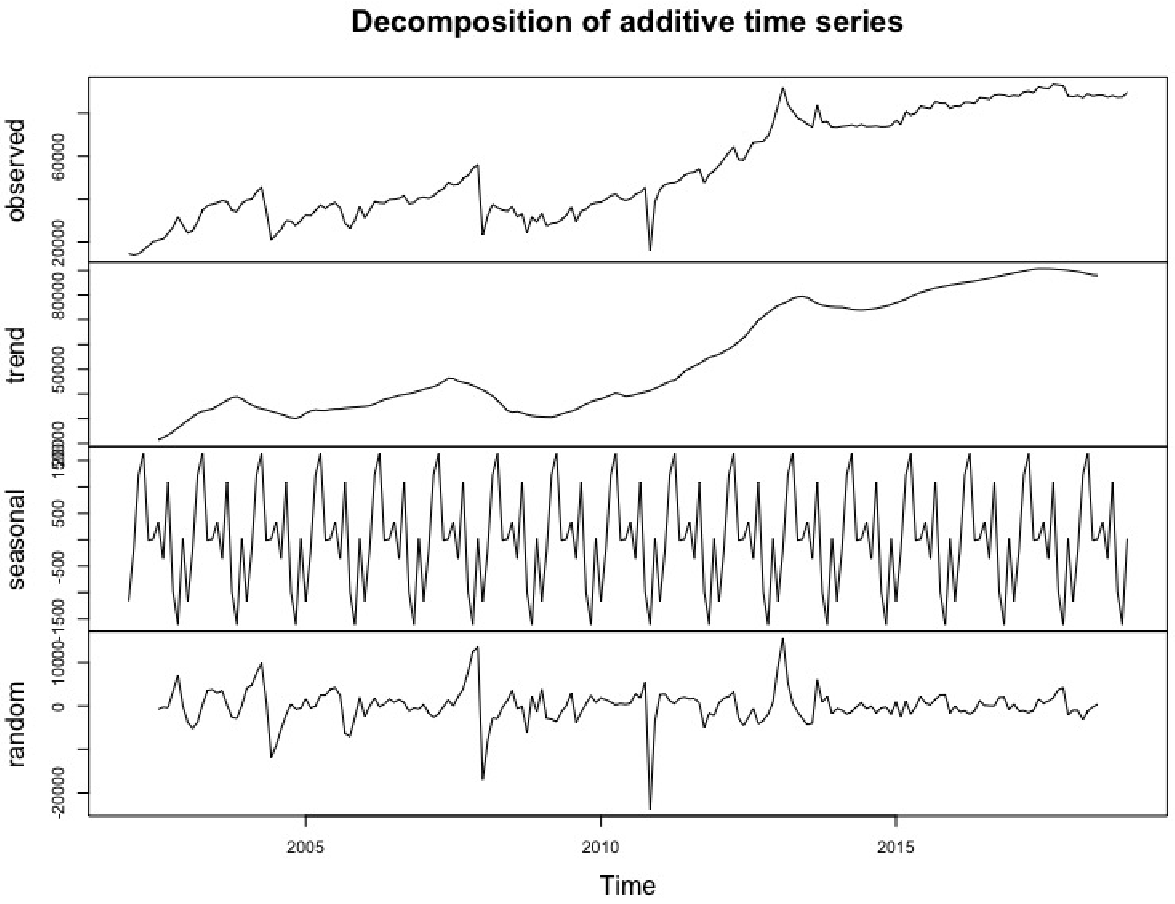


figure 2: Time Series Decomposition.

As we could see from figure 2, it is a seasonal data which has a significantly increasing trend and it has been successfully transform into a time series data.

3.3 Building Models

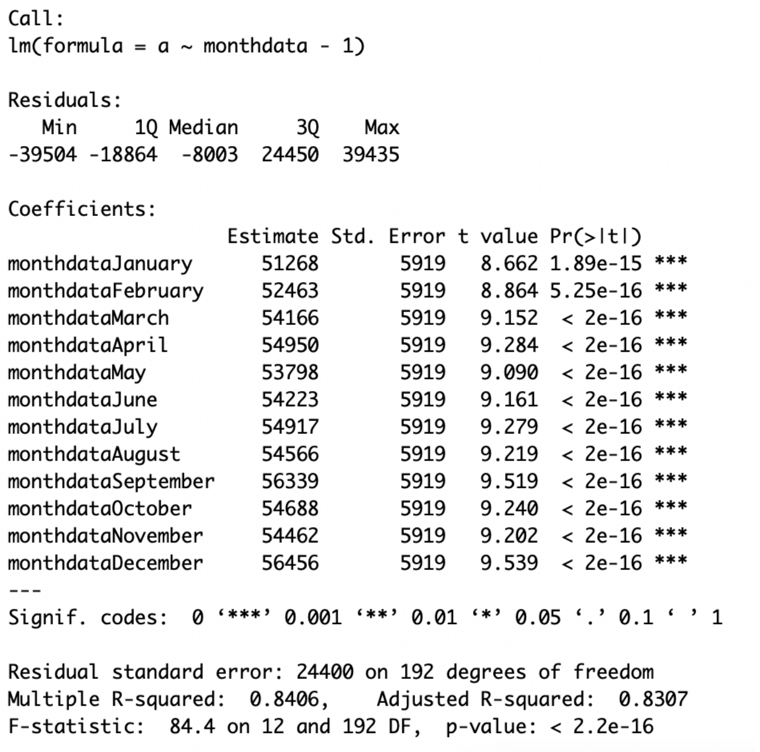
The forecasting method in this project is to build the three models to find the suitable model for this project.

3.3.1 Linear Regression

The linear regression utilizes least-squares regression to create a line of best fit to the data set. As I mentioned previously after decomposing the dataset, it has the seasonality. Hence, the model fits multiple linear regression models to the data set so that there is a regression for each different month as follows:

```
monthdata <- season(a)
modellinear <- lm(a~monthdata-1)
summary(modellinear)
```

The summary of such linear regression model has obstained as follows:



```
Call:
lm(formula = a ~ monthdata - 1)

Residuals:
   Min     1Q Median     3Q    Max 
-39504 -18864  -8003  24450  39435 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
monthdataJanuary      51268       5919   8.662 1.89e-15 ***
monthdataFebruary     52463       5919   8.864 5.25e-16 ***
monthdataMarch        54166       5919   9.152  < 2e-16 ***
monthdataApril        54950       5919   9.284  < 2e-16 ***
monthdataMay          53798       5919   9.090  < 2e-16 ***
monthdataJune         54223       5919   9.161  < 2e-16 ***
monthdataJuly         54917       5919   9.279  < 2e-16 ***
monthdataAugust       54566       5919   9.219  < 2e-16 ***
monthdataSeptember    56339       5919   9.519  < 2e-16 ***
monthdataOctober      54688       5919   9.240  < 2e-16 ***
monthdataNovember     54462       5919   9.202  < 2e-16 ***
monthdataDecember     56456       5919   9.539  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 24400 on 192 degrees of freedom
Multiple R-squared:  0.8406,	Adjusted R-squared:  0.8307 
F-statistic:  84.4 on 12 and 192 DF,  p-value: < 2.2e-16
```

figure 3: Summary of Linear Regression Model.

The column of estimates is the corresponding coefficient for each single month. Since p-values from both individual variable and overall model is strictly small, it is reasonable to conclude that the model is significant and all coefficents are non-zero. Furthermore, by looking at the value of r^2 , we can conclude that this correlation can explain 84.06% of the variation and the model.

Then to verify whether the residual of such model is normal, we can further plot the qq-plot as follows:

```
plot(modellinear)
```

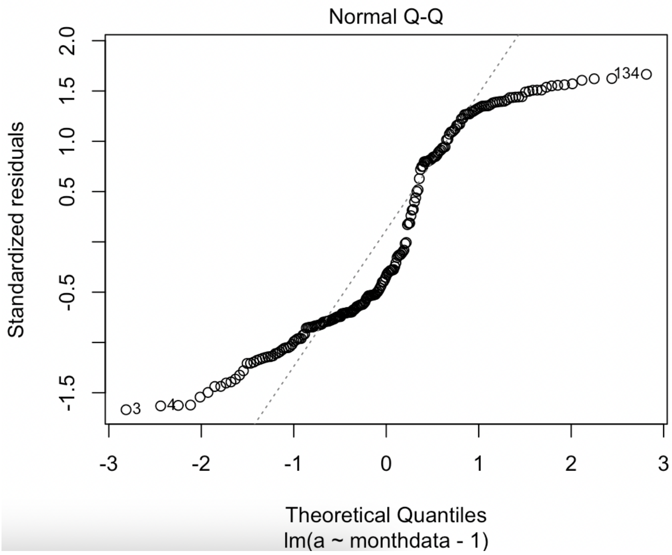


figure 4: QQ-plot of Linear Regression Model.

It is quite obvious to observe that residuals fall off the straight line, particularly at the head and tail of the data. Therefore, the residuals are not normally distributed. We can further prove such finding by looking at the histogram and density line of residuals as follows:

```
hist(modellinear$residuals, prob=TRUE)
lines(density(modellinear$residuals))
```

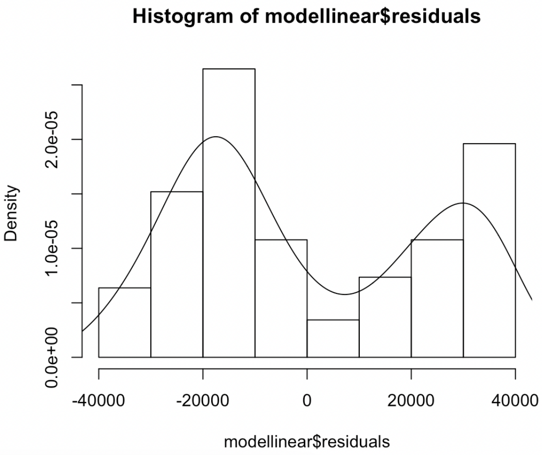


figure 5: Histogram of Residuals with Density Line.

It is clear now that residuals are not normally distributed.

3.3.2 ARIMA

First of all, I built ARIMA model. The steps to constructing ARIMA model and forecasting time series are as follows:

1. Determine the suitable order for ARIMA(p,d,q), which can be considered from Autocorrelation function (ACF) and Partial autocorrelation function (PACF). In R, there is a function to find order(p,d,q) automatically, and it can be shown as follows:

```
monavg_price<-auto.arima(a)
monavg_price
```

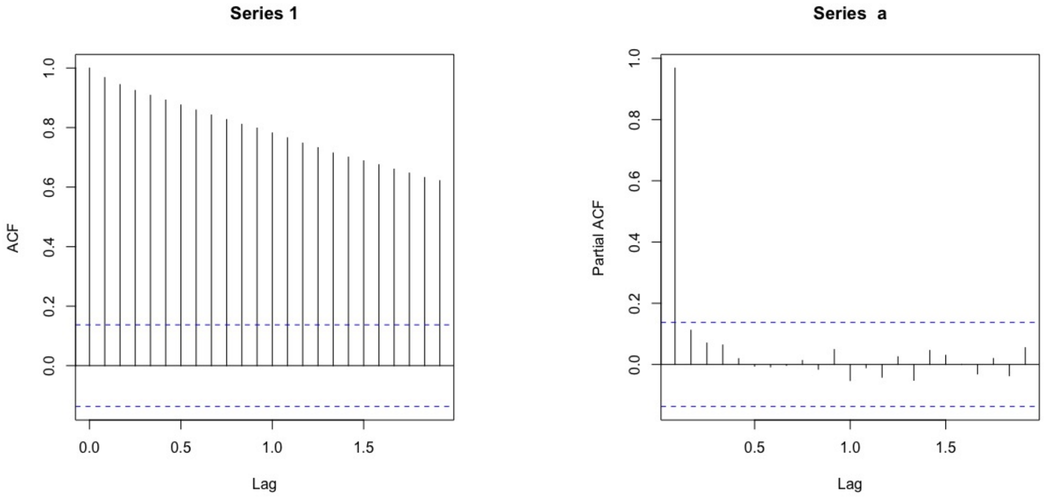


figure 6: ACF figure 7: PACF

Orders that are suitable for this dataset is ARIMA(0,1,2) based on the auto.arima() function.

2. The appropriate order was used for constructing model and forecasting time series. For example, in ARIMA model, the R commands are as the following and the running result was shown in the blow figures.

```
acf(a)
pacf(a)
monavg_price<-auto.arima(a)
plot(forecast(monavg_price,12))
```

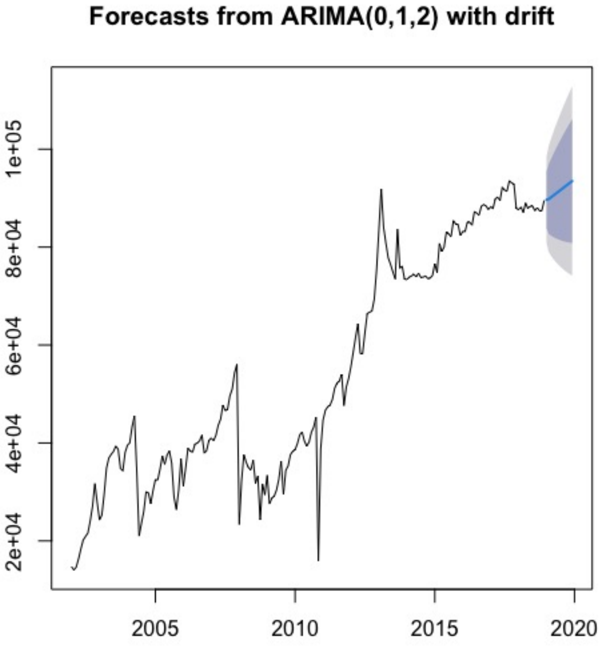


figure 8: ARIMA model forecasting for next 12 months

From figure 3, the black line is the actual data, whereas the blue line shows the forecast value and the area above and below the blue lines are error bounds at confidence level 80%.

3.3.3 Holt Linear method

The second model I built for this project is using Holt Linear method. The R command for constructing Holt's linear model in a monthly series are as follows:

```
holtavg_price<-holt(a,h=50)
```

Running results of a monthly model is shown in figure 4. Then the time period for forecasting in monthly has been set to next 12 months, and the running results are shown in figure 5:

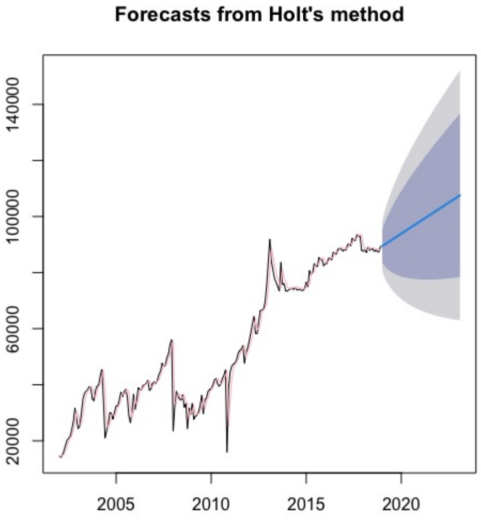


figure 9: Fitted Value in Holt's method

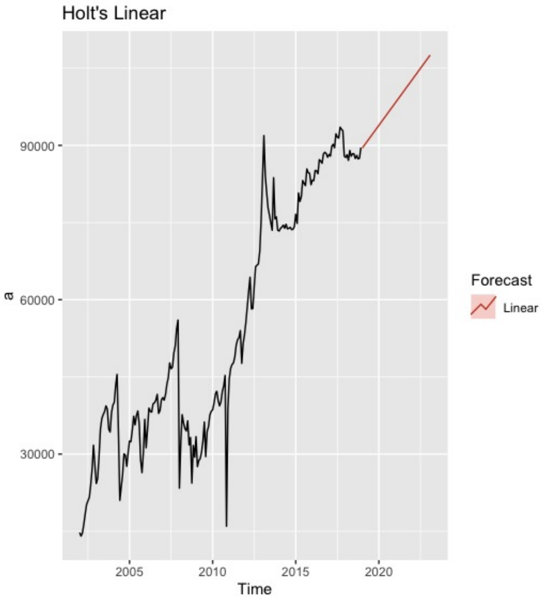


figure 10: Holt's Linear Method

3.3.4 Holt Winters Method

The third model I built was Holt Winters Method. In this method, I set the same forecasting value and set the seasonal equal to multiplicative in HoltWinters() function. The R command for constructing Holt Winters Method in a monthly series are as follows:

```
holtWavg_price<-HoltWinters(a,seasonal = "multiplicative")
summary(holtWavg_price)
```

The next 12 months forecasting value I used the following R commands and the running results for this method are shown in the following figure respectfully:

```
plot(forecast(holtWavg_price,12))
fore<-forecast(holtWavg_price,12)
accuracy(fore)
```

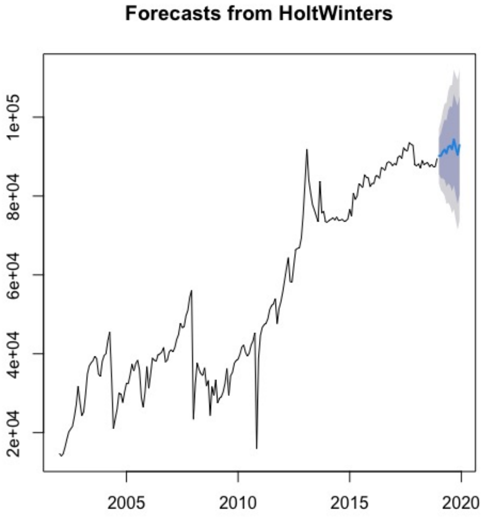


figure 11: Holt Winters Method forecasting

From the figure 5, the blue line is the forecasting value for the next 12 months and the area above and below the line are the 80% and 95% confident level.

4 Experimental Evaluation

This project used Shanghai Car Licence Plate dataset obtained from Kaggle.com for evaluating the time series models. This project has 204 observations. The time series data was divided into two groups I divided the data into two groups. The first group was training dataset which contain data from Jan,2002 to May, 2016 for construction the forecasting models. The second group was test data set which contain data from June 2016 to March 2019 for finding the most suitable forecasting period.

The three forecasting methods that were used in this project are ARIMA, Holt Linear and Holt Winters models. The suitable model that can be represented the real dataset when considerinh from the minimum value of AIC and RMSE value. First of all, I created three histogram charts of each models residuals and I created a table of AIC and RMSE value for each model are shown in the below:

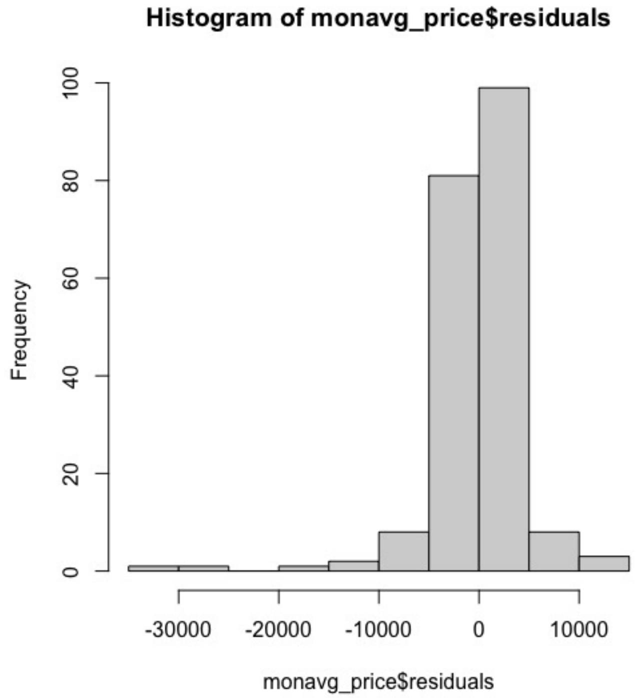


figure 12: Histogram of ARIMA Residuals

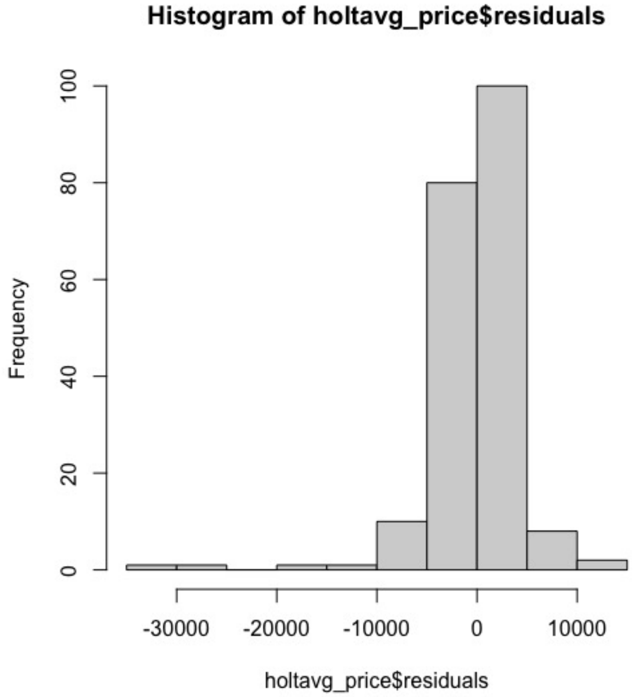


figure 13: Histogram of Holt Linear Residuals

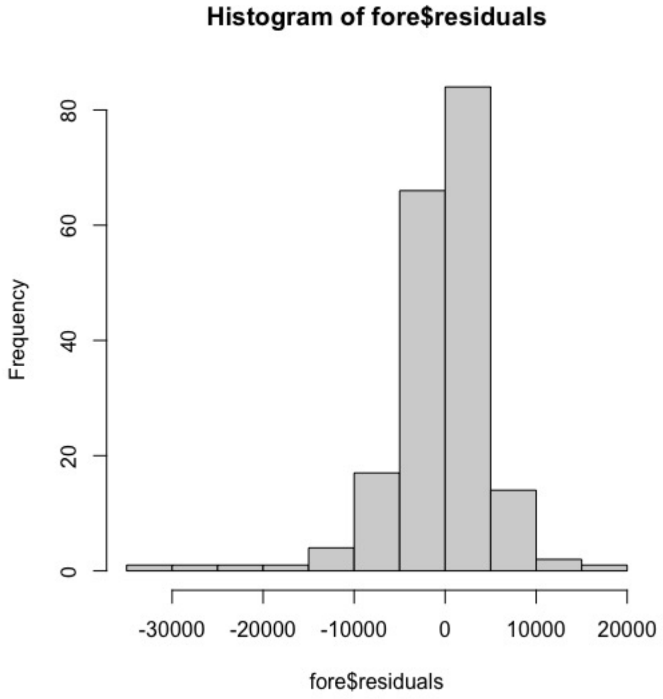
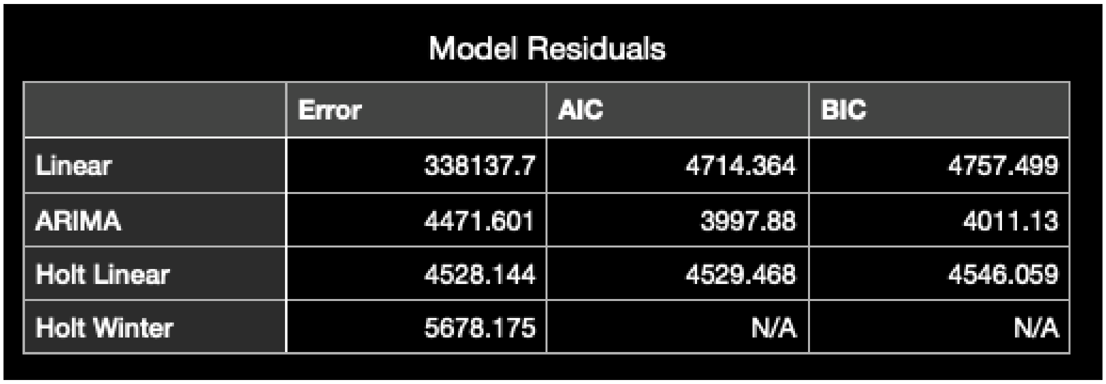


figure 14: Histogram of Holt Winters Residuals



Model Residuals

	Error	AIC	BIC
Linear	338137.7	4714.364	4757.499
ARIMA	4471.601	3997.88	4011.13
Holt Linear	4528.144	4529.468	4546.059
Holt Winter	5678.175	N/A	N/A

table 1: Model Residuals

From the previous figures and the Model Residuals table, ARIMA model has the smallest error and AIC value for this dataset which are 4471.601 and 3997.88. Therefore, in this project ARIMA model is the most suitable model to forecasting the average price for Shanghai Car Plate in next 12 months.

4.1 Training and Testing Groups

It is also good to train and test the most suitable model we found previously, which is ARIMA. To do so, the data is separated into two groups as 85% of it is the training group and the rest is the testing group as follows:

```
traingroup <- head(a, 173)
testgroup <- tail(a, 31)
finalmodel <- auto.arima(traingroup)
summary(finalmodel)
p15 <- forecast(finalmodel, 31)
err <- testgroup-p15$mean
MSE <- mean(err*err)
MSE
plot(a)
lines(p15$mean, col='red')
```

The MSE in this case is 22877108. And here is the plot:

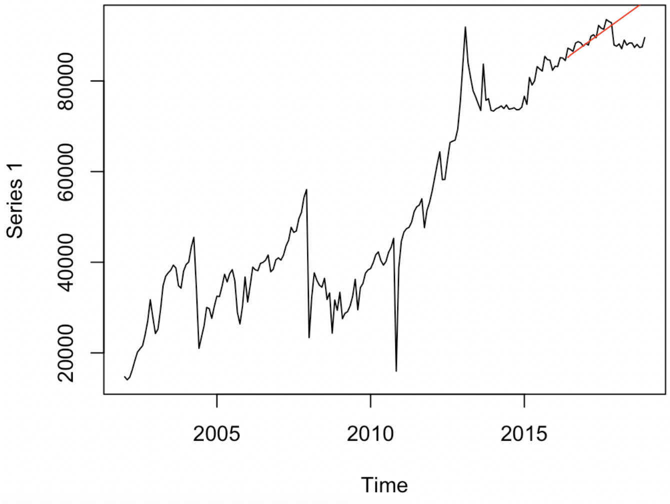


figure 15: The Predicted Values Against the Orignal Ones

4.2 Hypothesis Testing

4.2.1 Box Ljung Test

This Box Ljung test is used to investigate the lack of fit of a time series model. The corresponding hypothesis testing is to set the null hypothesis as the model does not exhibit the lack of fit while the alternative one as the model exhibits the lack of fit.

```
Box.test(finalmodel$residuals, lag = 12, type = ”Ljung−Box”)
```

The R commands returns p-value is 0.9926, and so we do not reject the null hypothesis and can conclude that does not exhibit the lack of fit.

4.2.2 Dickey-Fuller Unit-Root Test of Stationary

Looking at the plot is one of the approaches to check the existence of stationarity in a given time series, however the Dickey-Fuller test can perform the same function. The null hypothesis of Dickey-Fuller test is the data has a unit root while the alternative one is the opposite statement.

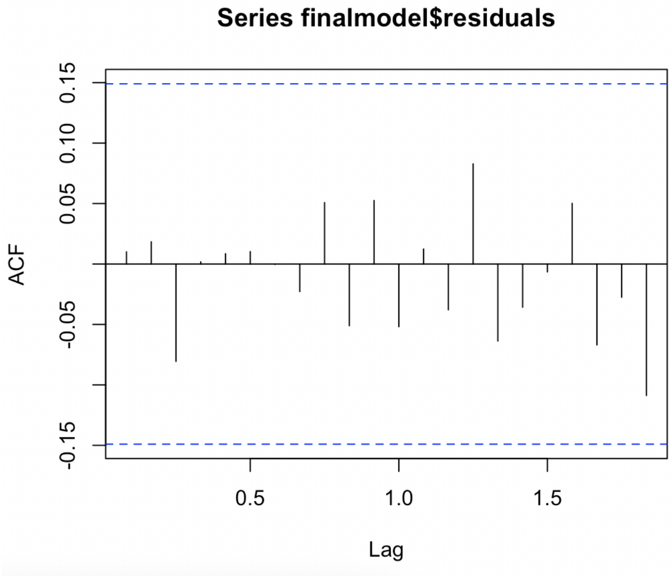


figure 16: ACF of Residuals.

From the graph above, it is clear to see that the lags of residuals after 0 are below the significant level. After running Dickey-Fuller test in R as follows:

```
adf.test(a, alternative=”stationary”)
```

The p-value is 0.5492 which allows us not to reject the null hypothesis. Therefore, there is no stationary existing in this time series, which supports the statement we found previously from the plot.

5 Conclusion

The purpose of this project is to find a suitable model to forecast the average price of Shanghai Car License Plate. I used ARIMA, Holt Linear and Holt Winter models for forecasting the average price of the Shanghai Car Plate from 2002 to 2020. Then, choosing the suitable forecasting method by considering the minimum values of AIC and RMSE value. The results showed that the ARIMA model can present the most suitable forecasting model for this monthly period.