US Box Office Forecast

By Leng, Zhou, Liang

Problem Statement

- Goal: forecast US movie market based on monthly total gross box office revenue
- Data:
  - Box Office
    - Source: Box Office MOJO
    - Variables : Date, Total Gross
  - Economic Indicator: Monthly US Unemployment Rate

Assumptions

- Hypothesis:
  - Box office revenue is seasonal
  - It is correlated with economy
  - Historical trend is predictive for future values

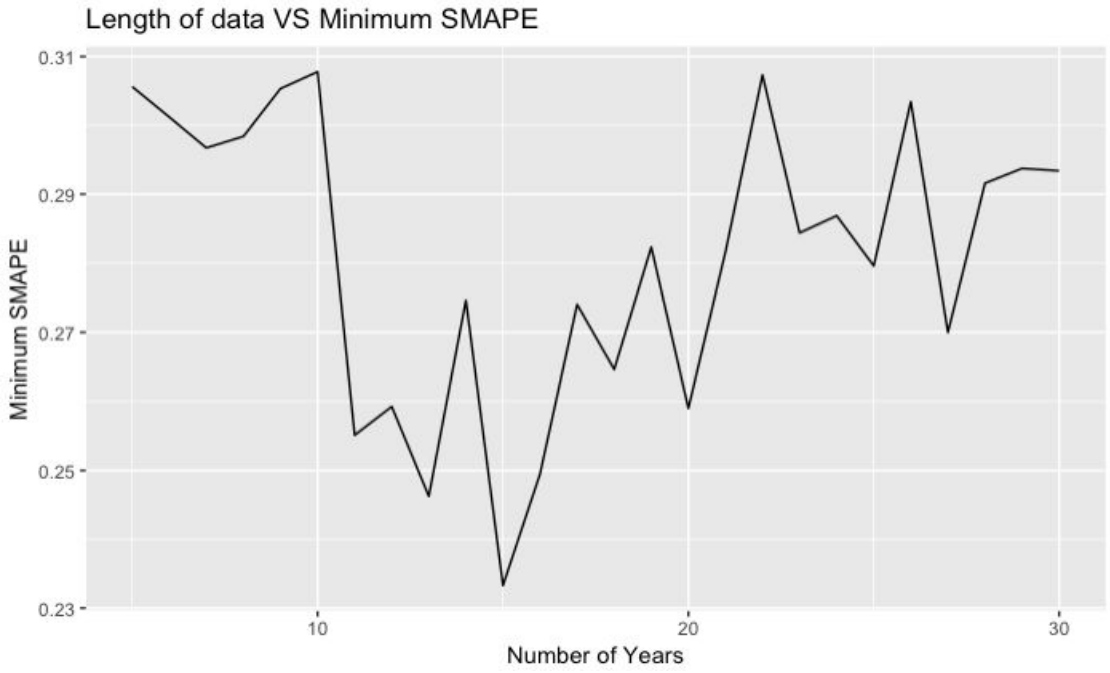
Data Properties and Transformation

- Daily total gross data of US Box office from Jan 1982 to Apr 2018
- Transform daily total gross into monthly total gross
- Split into training set and testing set:
  - Training set:
    - ? - Apr 2017
  - Testing set:
    - May 2017 - Apr 2018

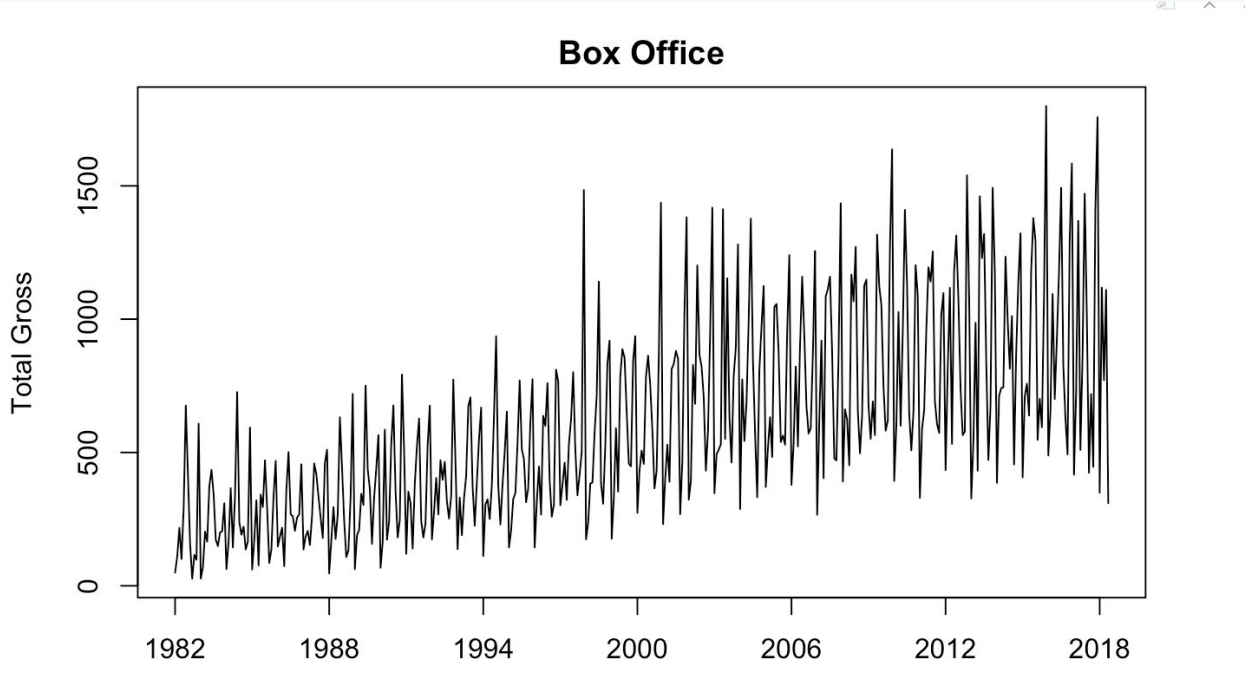
Date	Day	Movies	TotalGross
01/1/2	Tue	20	10248284
01/2/2	Wed	20	4616413
01/3/2	Thu	20	3767510
01/4/2	Fri	20	6783240
01/5/2	Sat	20	10114063
01/6/2	Sun	20	6109144
01/7/2	Mon	19	1710810
01/8/2	Tue	19	1978472
01/9/2	Wed	19	1525350
01/10/2	Thu	19	1395114
01/11/2	Fri	23	5693000
01/12/2	Sat	23	7268117
1/13/02	Sun	23	4692094
1/14/02	Mon	23	1137194
1/15/02	Tue	23	1305321
1/16/02	Wed	23	1100225
1/17/02	Thu	23	1112025
1/18/02	Fri	24	9914214
1/19/02	Sat	24	10024101

Determining the training set

- For data of length 5 to 30 years, find the minimum sMAPE achieved across all models.
- Result:
  - Training set:
  - Apr 2002 - Apr 2017

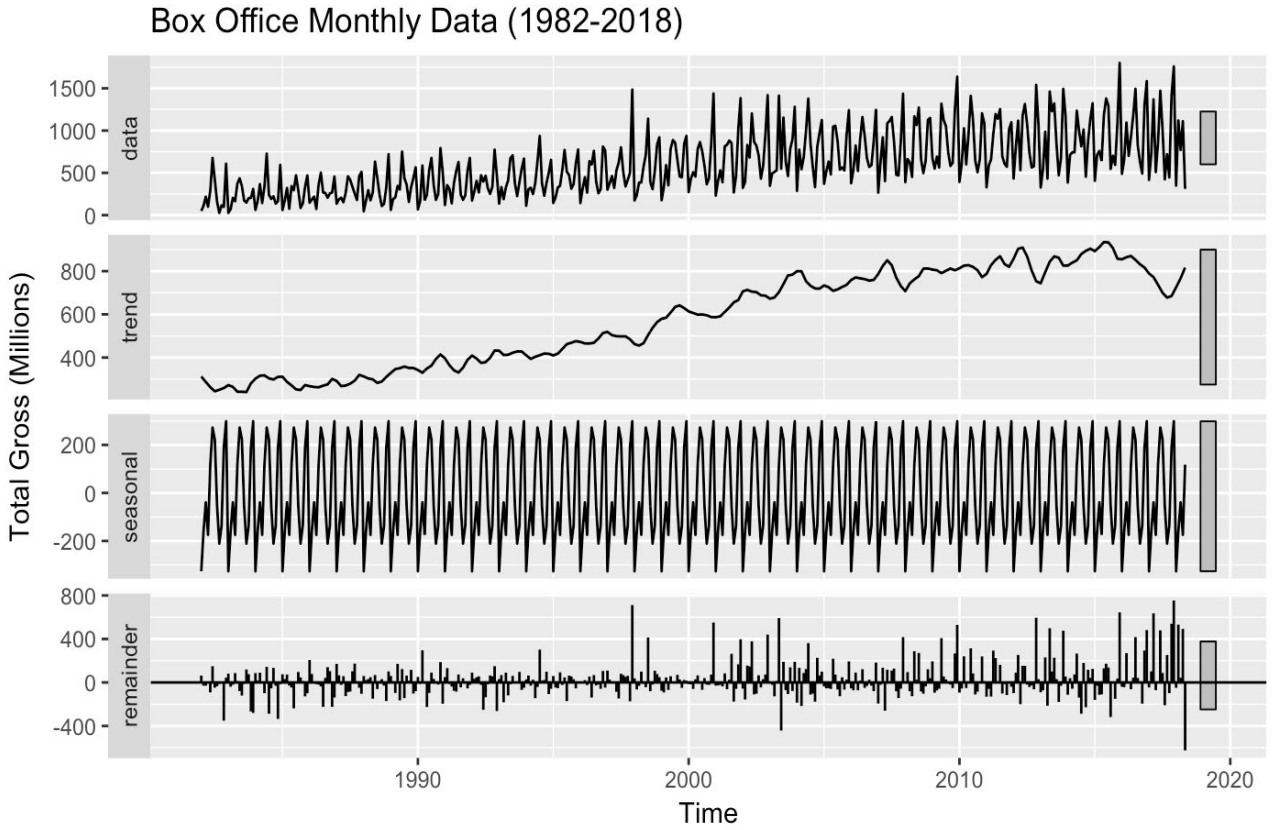


Exploration for Trend



- Clear linearly increasing trend
- Increasing variance

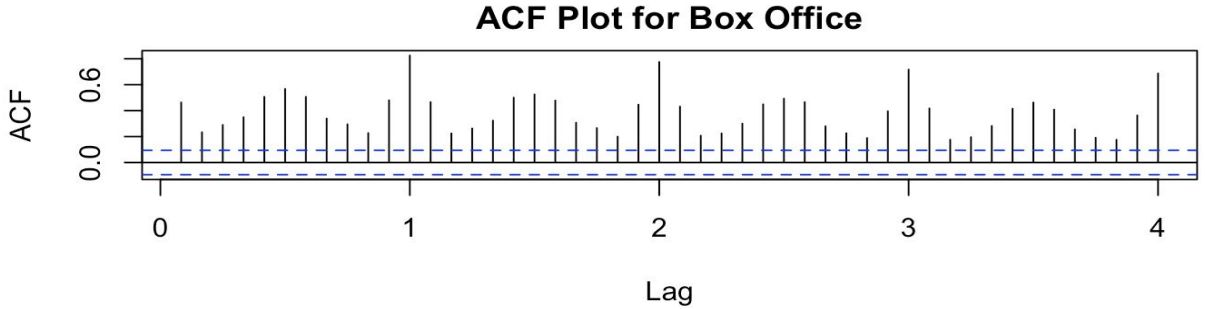
Decomposition

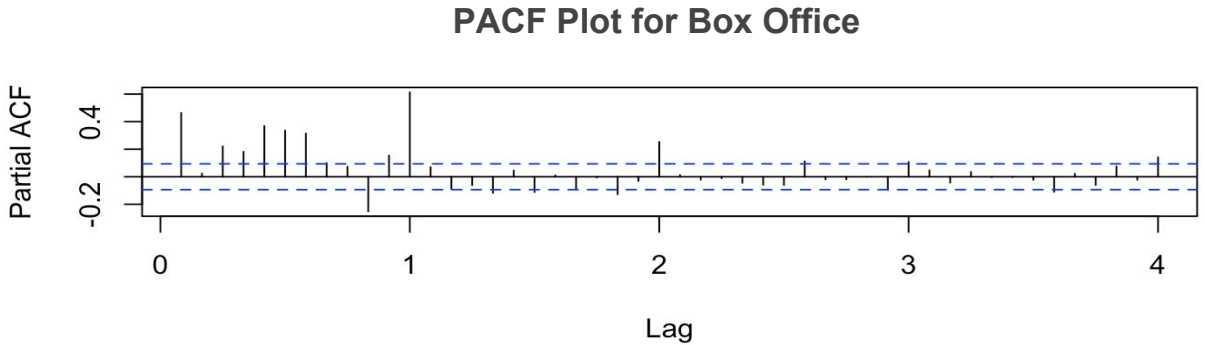


ADF Test for Stationarity

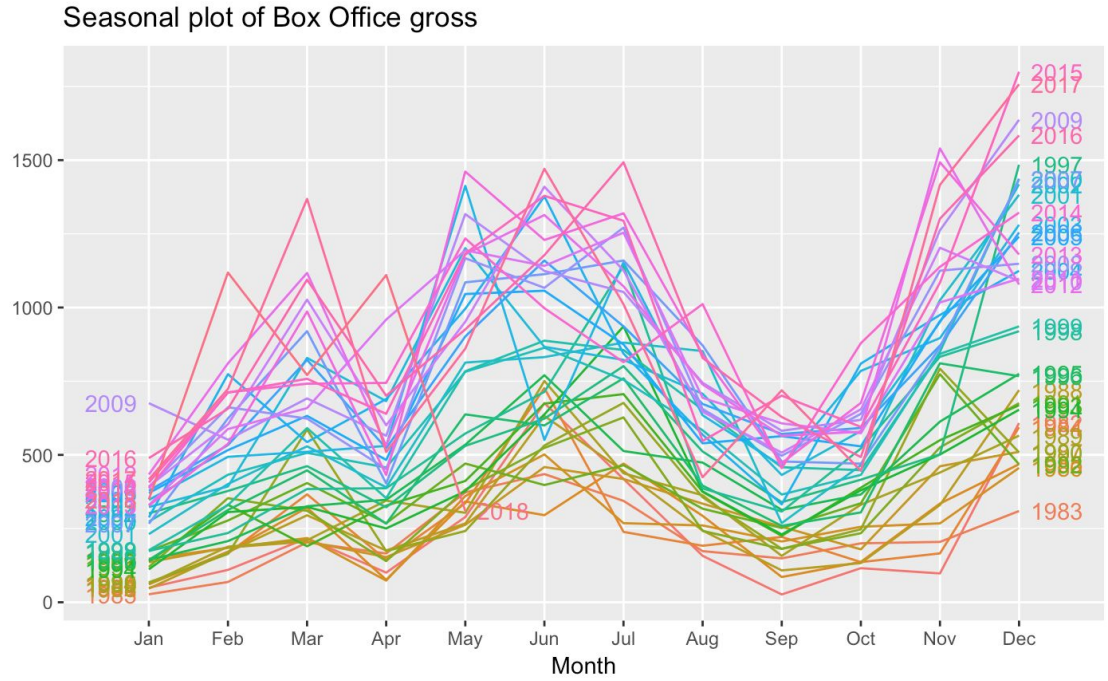
```
        Augmented Dickey-Fuller Test

data:  data.ts[, 1]
Dickey-Fuller = -6.7304, Lag order = 7, p-value = 0.01
alternative hypothesis: stationary
```

ACF and PACF


ACF Plot for Box Office
ACF
0.0
0.6
0
1
2
3
4
Lag


PACF Plot for Box Office
Partial ACF
-0.2
0.4
0
1
2
3
4
Lag

Exploration for Seasonality



Models

1.Linear regression with ARMA errors

2.Exponential Smoothing Model

3.Holt Winters

4. ARIMA, SARIMA, ARFIMA, and GARCH

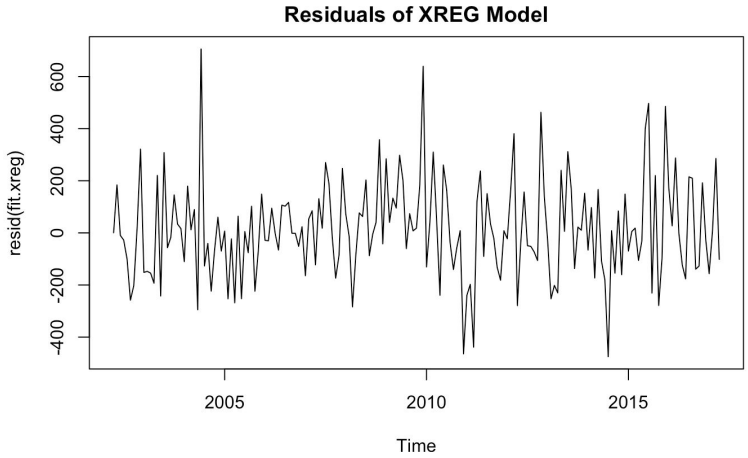
5. NNAR

6.TBATS

xreg Result

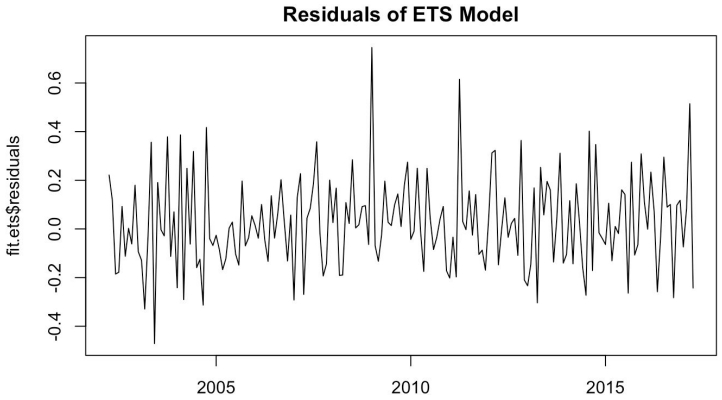
```
Series: train[, 1]
Regression with ARIMA(2,1,2)(1,0,0)[12] errors

Coefficients:
         ar1      ar2      ma1     ma2    sar1    xreg
      0.1468  -0.0964  -1.3998  0.4059  0.8266  4.2112
s.e.  0.2417   0.0990   0.2336  0.2333  0.0441  1.7035

sigma^2 estimated as 37263:  log likelihood=-1207.74
AIC=2429.48   AICc=2430.13   BIC=2451.83

Training set error measures:
                    ME     RMSE      MAE       MPE     MAPE      MASE         ACF1
Training set 12.01577 189.266 142.426 -3.883848 18.34319 0.8749033 -0.009272678
```

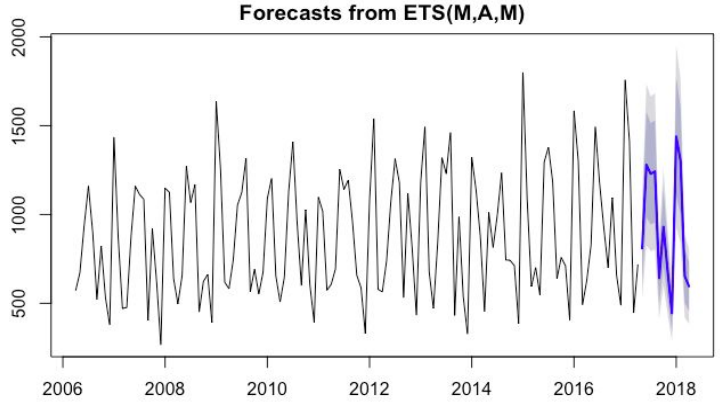
```
Call:
lm(formula = TotalGross ~ Movies, data = data.ts)

Residuals:
    Min      1Q  Median      3Q     Max
-723.00 -220.67  -85.91  219.60 1102.38

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 199.8769    31.7187   6.302 7.23e-10 ***
Movies       11.5395     0.8081  14.280  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 304.8 on 435 degrees of freedom
Multiple R-squared:  0.3192,    Adjusted R-squared:  0.3176
F-statistic: 203.9 on 1 and 435 DF,  p-value: < 2.2e-16
```



ETS Result

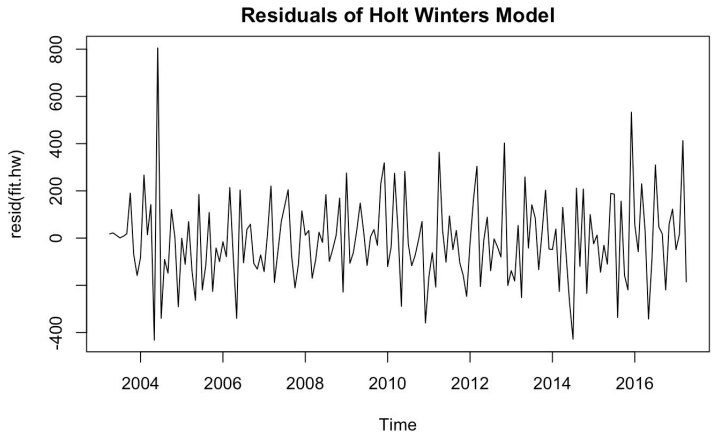
```
ETS(M,N,M)

Call:
 ets(y = train[, 1])

  Smoothing parameters:
    alpha = 0.0413
    gamma = 1e-04

  Initial states:
    l = 776.392
    s=0.9863 0.7305 0.4919 1.5553 1.3069 0.748
           0.6247 0.8349 1.2807 1.3512 1.37 0.7196

  sigma:  0.19

     AIC     AICc      BIC
2775.659 2778.568 2823.636

Training set error measures:
                    ME     RMSE      MAE        MPE     MAPE     MASE       ACF1
Training set 22.56404 157.1561 118.8698 -0.9573125 14.45335 0.730201 -0.2403196
```

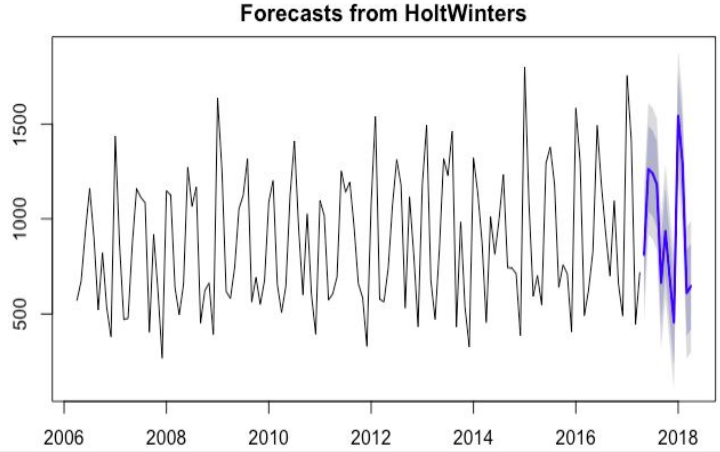




Holt Winters Result

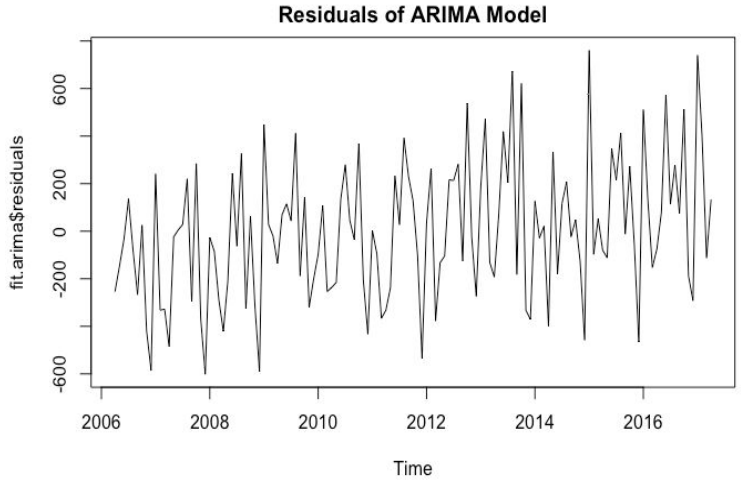
	Length	Class	Mode
fitted	676	mts	numeric
x	181	ts	numeric
alpha	1	-none-	numeric
beta	1	-none-	numeric
gamma	1	-none-	numeric
coefficients	14	-none-	numeric
seasonal	1	-none-	character
SSE	1	-none-	numeric
call	2	-none-	call

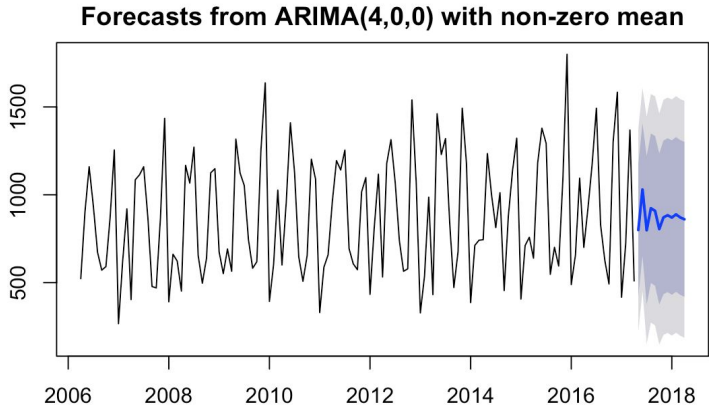




ARIMA Result

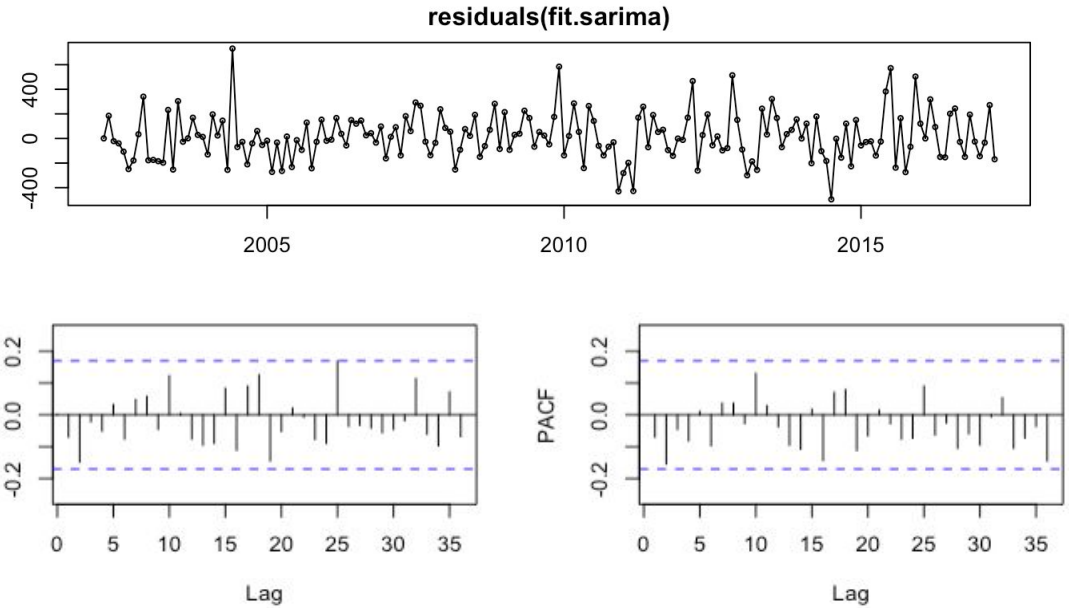
```
Series: train[, 1]
ARIMA(4,1,1)

Coefficients:
          ar1      ar2      ar3      ar4      ma1
      -0.1427  -0.4900  -0.2741  -0.3971  -0.9371
s.e.   0.0705   0.0676   0.0674   0.0706   0.0218

sigma^2 estimated as 80075:  log likelihood=-1271.42
AIC=2554.84   AICc=2555.33   BIC=2574

Training set error measures:
                    ME     RMSE      MAE       MPE     MAPE     MASE         ACF1
Training set 31.2711 278.2448 216.5319 -9.200128 31.16821 1.330126 -0.06627176
```

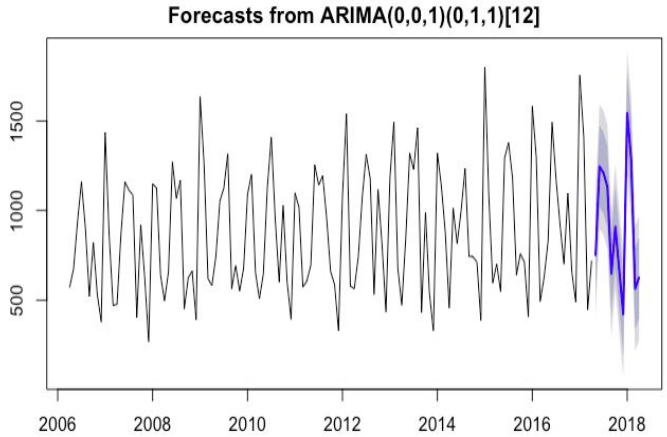




SARIMA Result

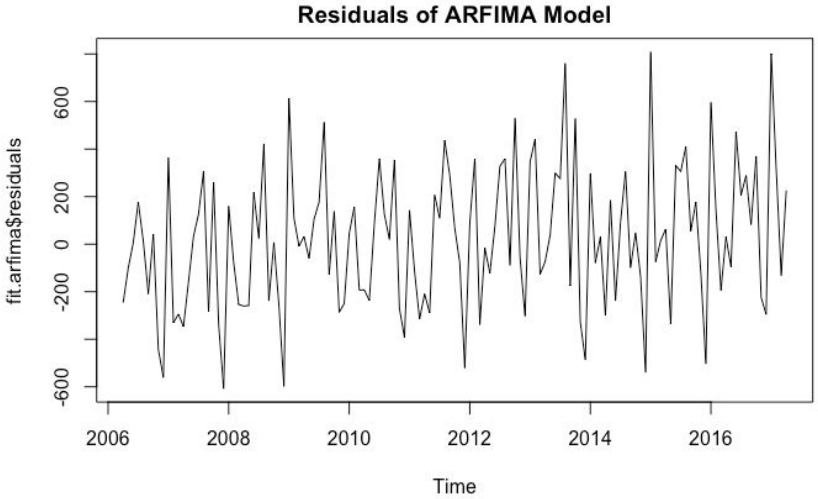
```
Series: train[, 1]
ARIMA(0,1,3)(1,0,0)[12]

Coefficients:
          ma1     ma2     ma3    sar1
      -1.2730  0.1610  0.1237  0.8142
s.e.   0.0741  0.1184  0.0721  0.0444

sigma^2 estimated as 38356:  log likelihood=-1211.03
AIC=2432.06   AICc=2432.4   BIC=2448.02

Training set error measures:
                    ME     RMSE      MAE       MPE   MAPE      MASE        ACF1
Training set 15.80153 193.1235 146.7109 -3.687826 18.962 0.9012251 -0.00596499
```

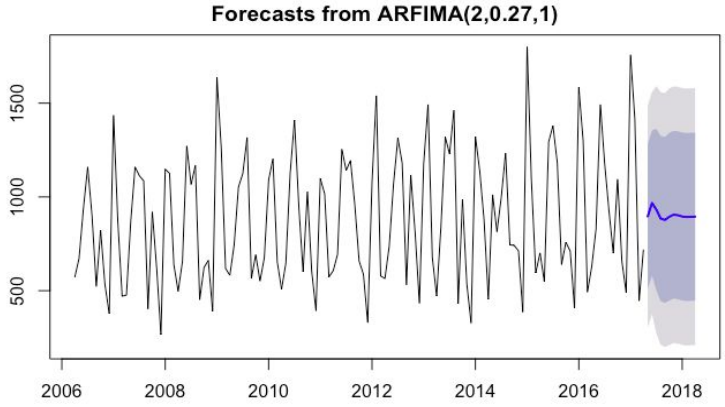




ARFIMA Result

```
Call:
  arfima(y = train[, 1])

*** Warning during (fdcov) fit: unable to compute correlation matrix; maybe change 'h

Coefficients:
       Estimate
d         0.268
ar.ar1    0.490
ar.ar2   -0.443
ma.ma1    0.776
sigma[eps] = 300.2181
[d.tol = 0.0001221, M = 100, h = 9.996e-06]
Log likelihood: -947.7 ==> AIC = 1905.372 [5 deg.freedom]
```





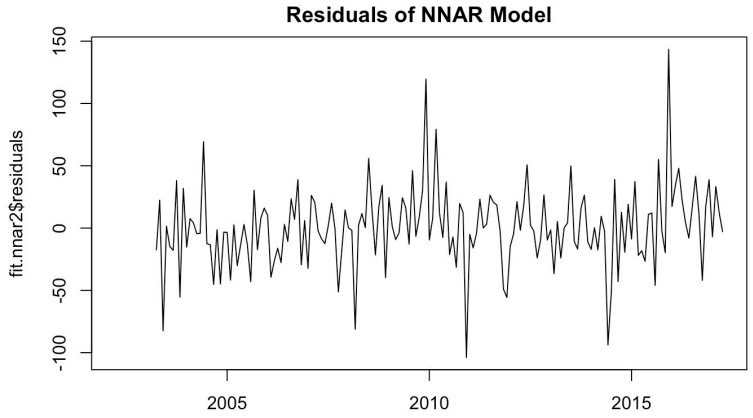
ARMA and GARCH Result

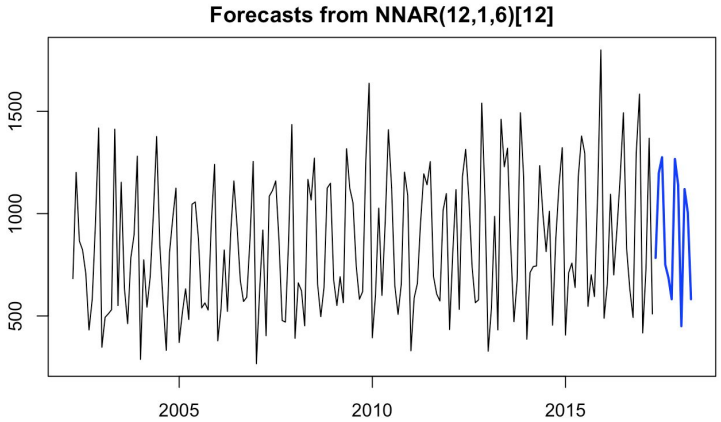
```
Series: log.bo
ARIMA(2,0,1)(1,0,0)[12] with zero mean

Coefficients:
          ar1      ar2      ma1    sar1
      -0.2662  -0.1221  -0.9875  0.8426
s.e.   0.0763   0.0755   0.0111  0.0390

sigma^2 estimated as 0.05538:  log likelihood=-1.32
AIC=12.64   AICc=12.98   BIC=28.6

Training set error measures:
                     ME      RMSE       MAE      MPE     MAPE      MASE         ACF1
Training set 0.01576539 0.2326989 0.1835699 80.96233 163.3329 0.5808288 -0.02103611

        Box-Ljung test

data:  fit.garch.arma$residuals^2
X-squared = 26.3, df = 12, p-value = 0.009732
```

```
Title:
 GARCH Modelling

Call:
 garchFit(formula = ~arma(1, 2) + garch(1, 1), data = log.bo,
    cond.dist = "std", trace = F)

Mean and Variance Equation:
 data ~ arma(1, 2) + garch(1, 1)
<environment: 0x11e4e3ed8>
 [data = log.bo]

Conditional Distribution:
 std

Coefficient(s):
         mu          ar1          ma1          ma2        omega       alpha1        beta1
 0.00078797  -0.54439968  -0.18924317  -0.75300034   0.05554917   0.00000001   0.70263499
      shape
10.00000000

Std. Errors:
 based on Hessian
```

NNAR Result

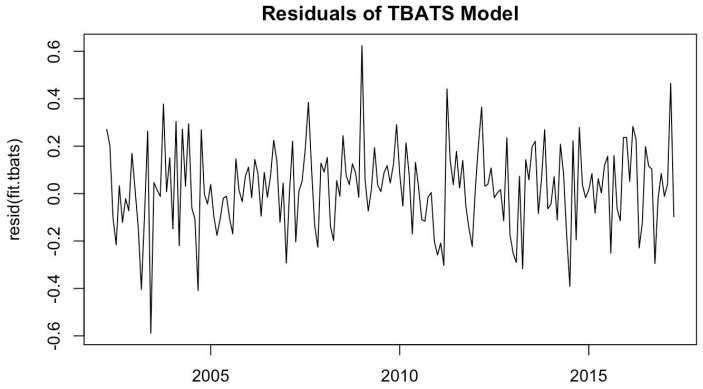
	Length	Class	Mode
x	181	ts	numeric
m	1	-none-	numeric
p	1	-none-	numeric
P	1	-none-	numeric
scalex	2	-none-	list
scalexreg	2	-none-	list
size	1	-none-	numeric
xreg	181	-none-	numeric
subset	181	-none-	numeric
model	20	nnetarmodels	list
nnetargs	0	-none-	list
fitted	181	ts	numeric
residuals	181	ts	numeric
lags	12	-none-	numeric
series	1	-none-	character
method	1	-none-	character
call	3	-none-	call

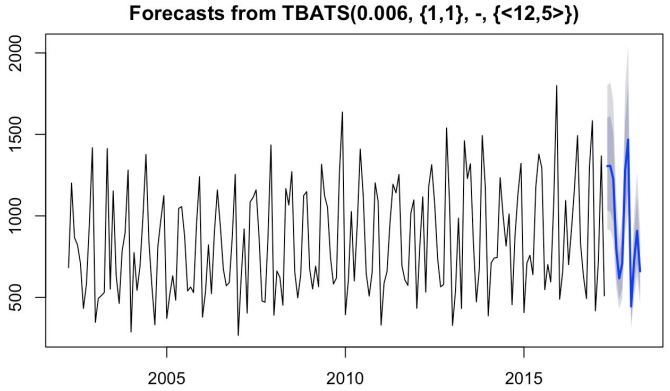




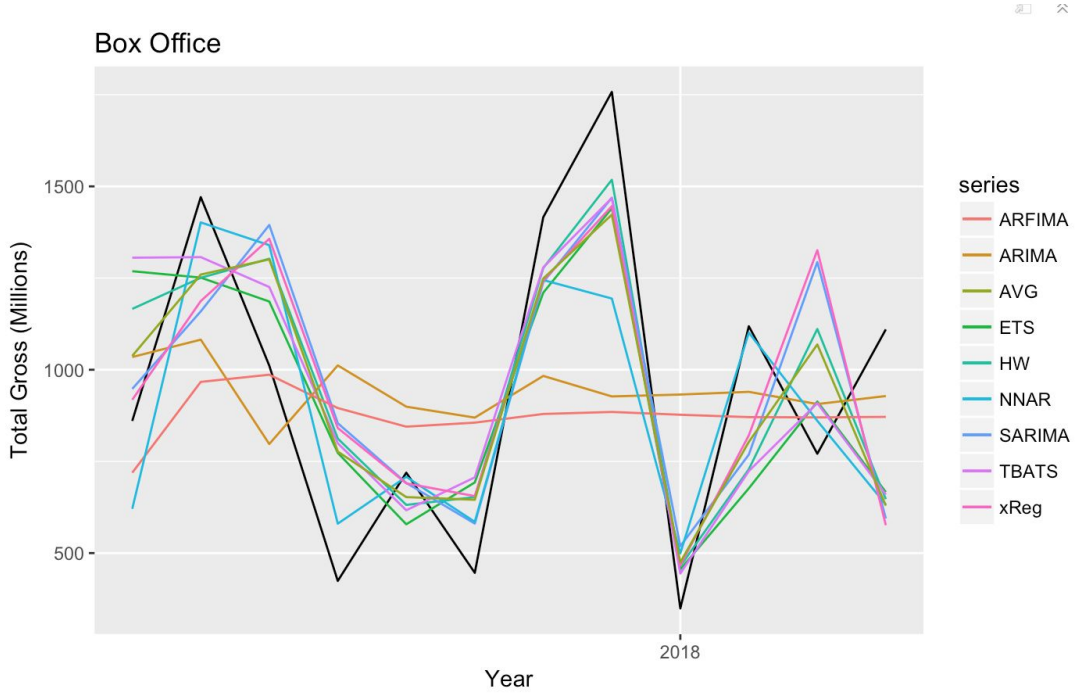
TBATS Result

lambda	1	-none-	numeric
alpha	1	-none-	numeric
beta	0	-none-	NULL
damping.parameter	0	-none-	NULL
gamma.one.values	1	-none-	numeric
gamma.two.values	1	-none-	numeric
ar.coefficients	1	-none-	numeric
ma.coefficients	1	-none-	numeric
likelihood	1	-none-	numeric
optim.return.code	1	-none-	numeric
variance	1	-none-	numeric
AIC	1	-none-	numeric
parameters	2	-none-	list
seed.states	13	-none-	numeric
fitted.values	181	ts	numeric
errors	181	ts	numeric
x	2353	-none-	numeric
seasonal.periods	1	-none-	numeric
k.vector	1	-none-	numeric
y	181	ts	numeric
p	1	-none-	numeric
q	1	-none-	numeric
call	3	-none-	call
series	1	-none-	character
method	1	-none-	character





Model Selection



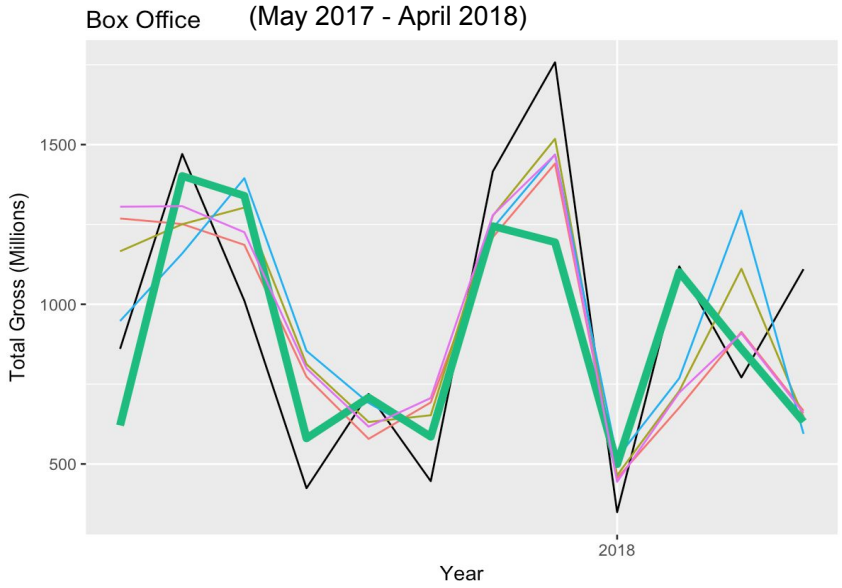
SMAPE result:

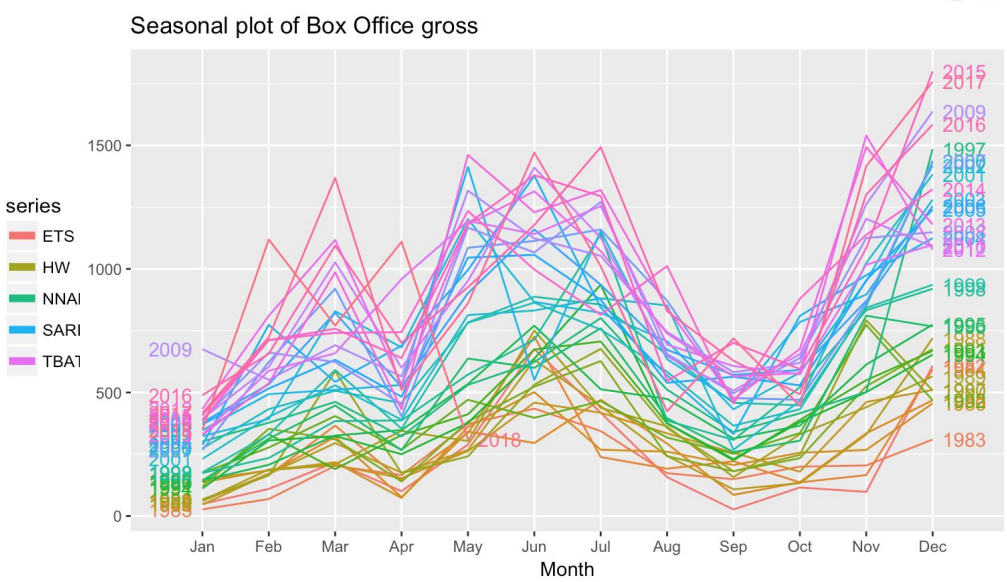
Xreg 0.3111
Ets 0.3098
Hw 0.3077
Arima 0.4011
Sarima 0.3177
Arfima 0.3934
Garch 0.4249
Nnar 0.2322
Tbats 0.2968
Avg 0.2907

The best model is:

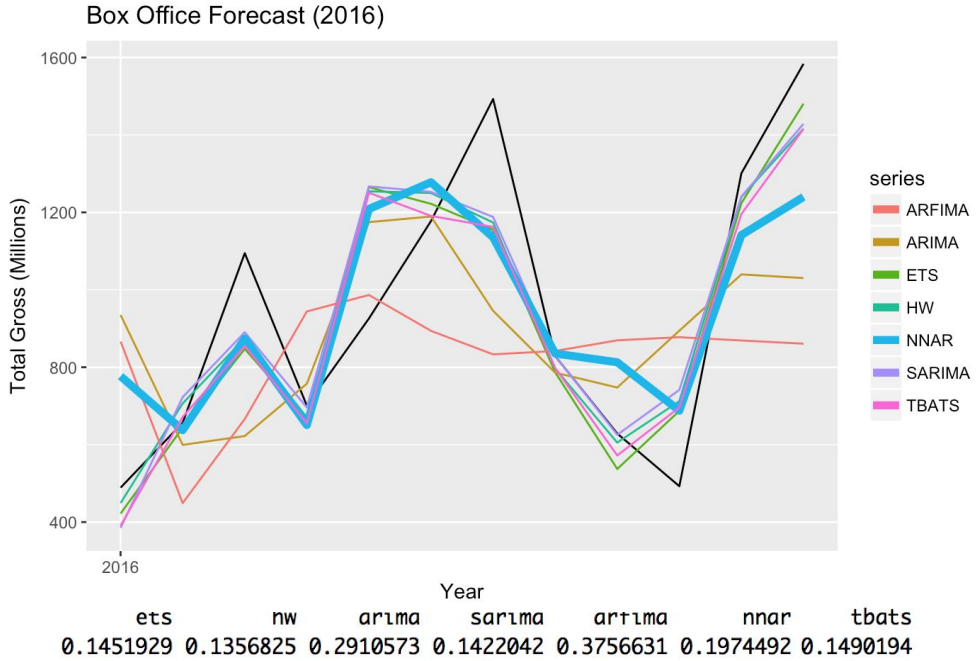
NNAR

Final Result - NNAR

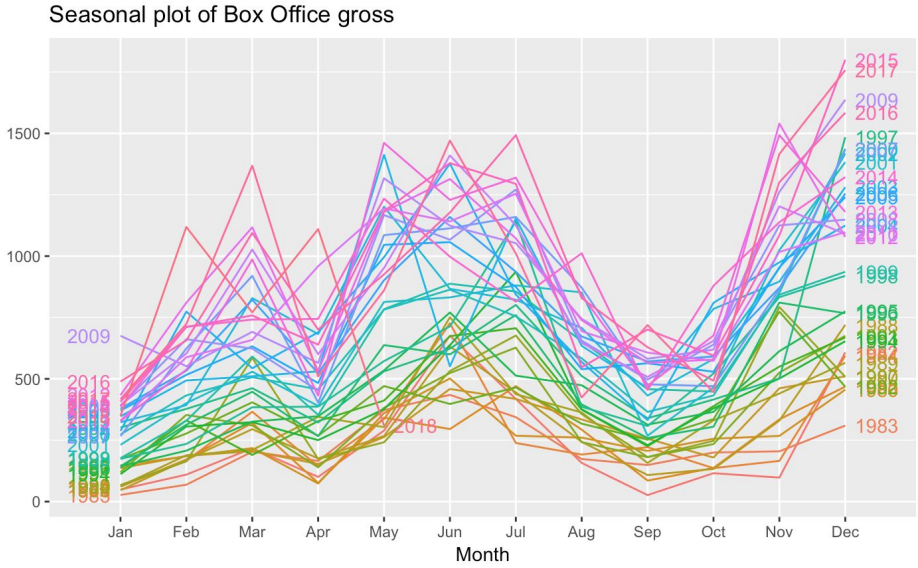




Final Result - NNAR?

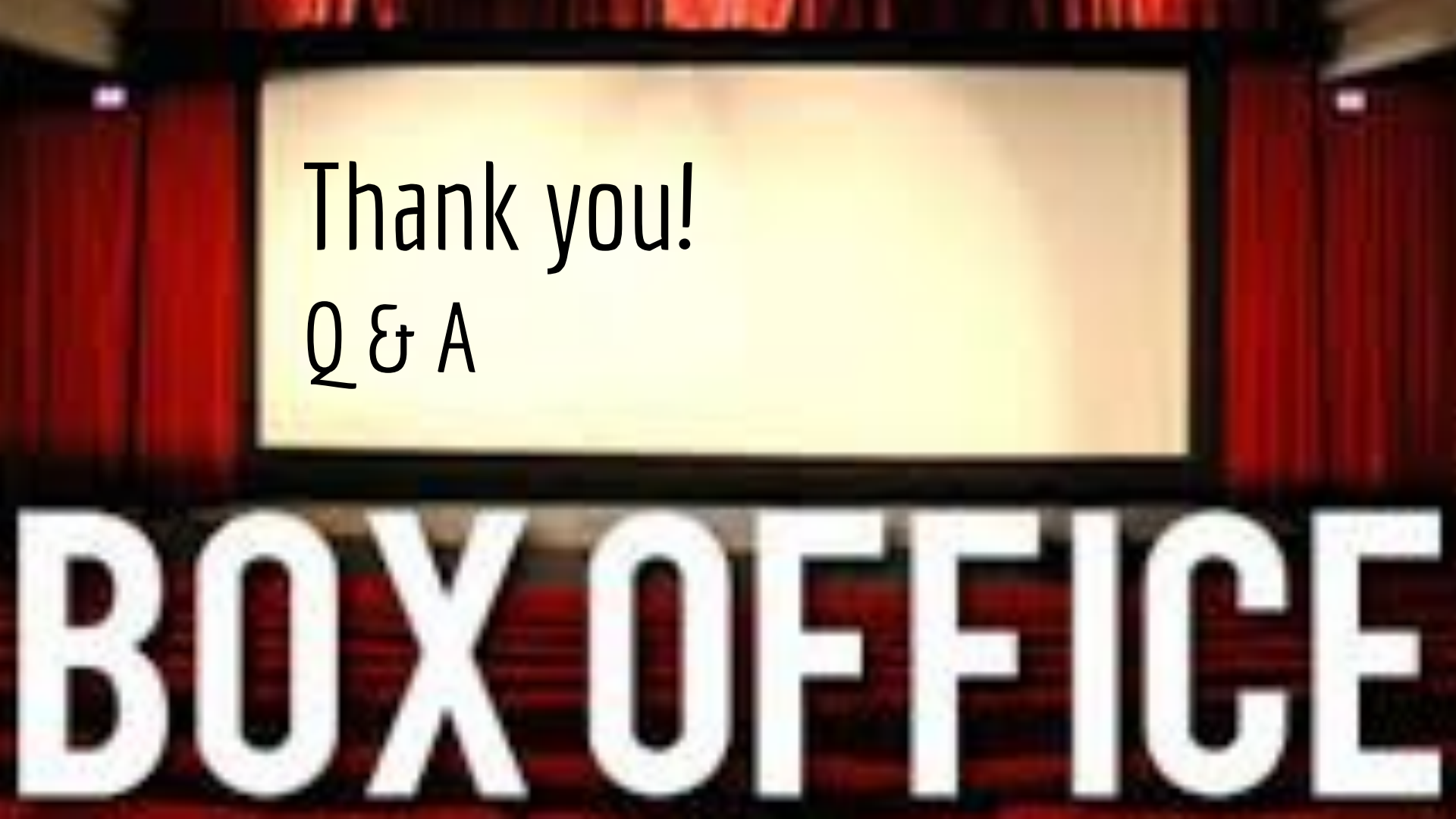



```
      ets        nw      arima    sarima    arfima      nnar     tbats
0.1451929 0.1356825 0.2910573 0.1422042 0.3756631 0.1974492 0.1490194
```



Future Work

- Try to forecast quarterly box office gross
- Add weight to historical data to balance forecasting power
- Utilize more independent variables to conduct multivariate analysis
- Ensemble models


Thank you!
Q & A
BOX OFFICE