Yellowstone National Park

Time Series and Forecasting – Winter Quarter 2020

Claire Zhang | Daniela Matinho | Sneha Vasudevan | Tim Chang

Agenda

- Yellowstone & Business Problem
- EDA & Data visualization
- Modeling:
  - Exponential Smoothing: ETS & Holt-Winters Method
  - Arima/sArima
  - Tslm
  - ArimaX (regression)
  - VAR
- Model evaluation
- Future work

Yellowstone Park & Business Problem

The park is open 24/7 and it receives over 4 million visitors a year.

Problem Statement

Forecast the number of visitors of the Yellowstone park receives yearly:

- To efficient allocate resources allowing the park management to provide better care to its assets and thereby increase visitors' satisfaction

Curiosity Park's budget is $34 million.

This translates into more than $380 million a year in visitor spending in communities near the park. That spending supports 5,300 jobs in the local area, according to a recent economic impact report.

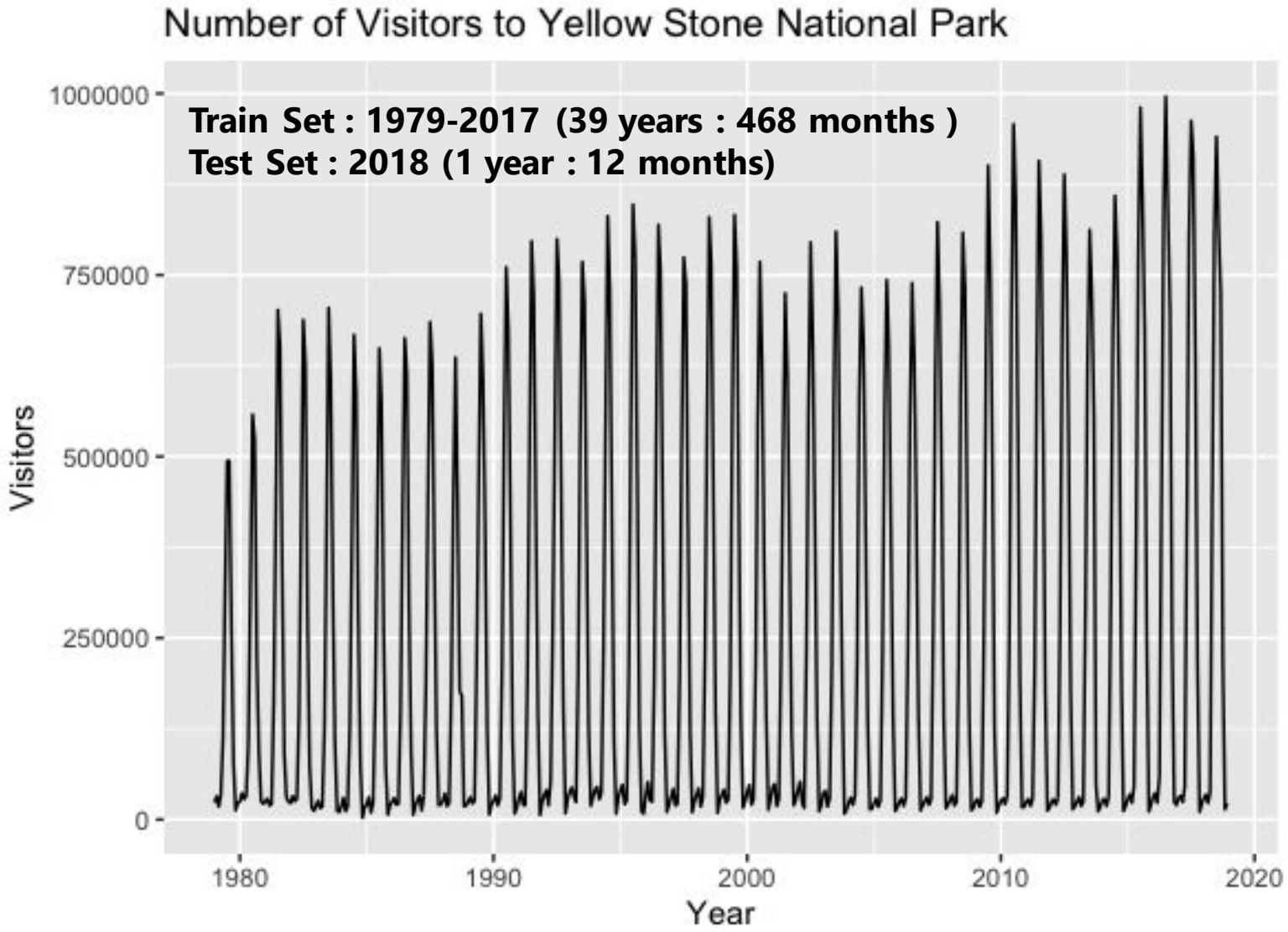
Exploratory Data Analysis (EDA)

yellowstone.csv

```
YEAR,MONTH,Visitors
1979,JAN,"23,605"
1979,FEB,"31,992"
1979,MAR,"17,813"
1979,APR,"34,095"
1979,MAY,"108,952"
1979,JUN,"313,924"
1979,JUL,"493,902"
1979,AUG,"493,915"
1979,SEP,"262,786"
1979,OCT,"76,542"
1979,NOV,"12,698"
```

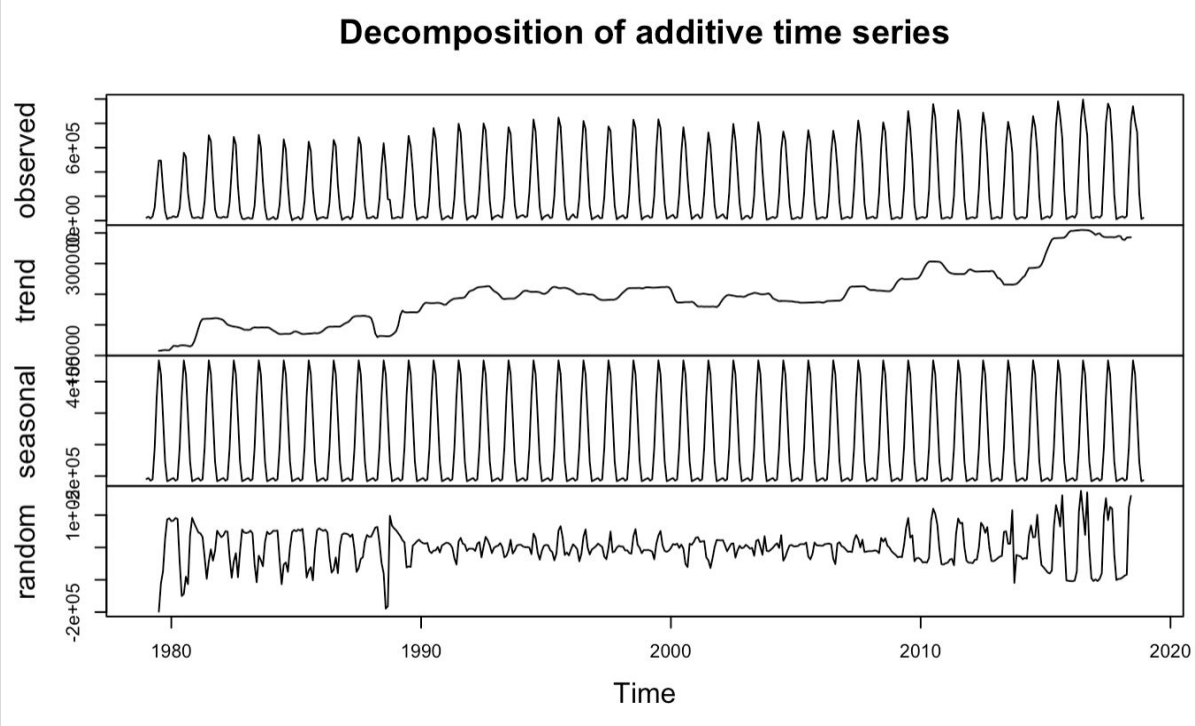
- ❑ Strong Seasonal
- ❑ Variance increases over time
- ❑ Tiny upward trend

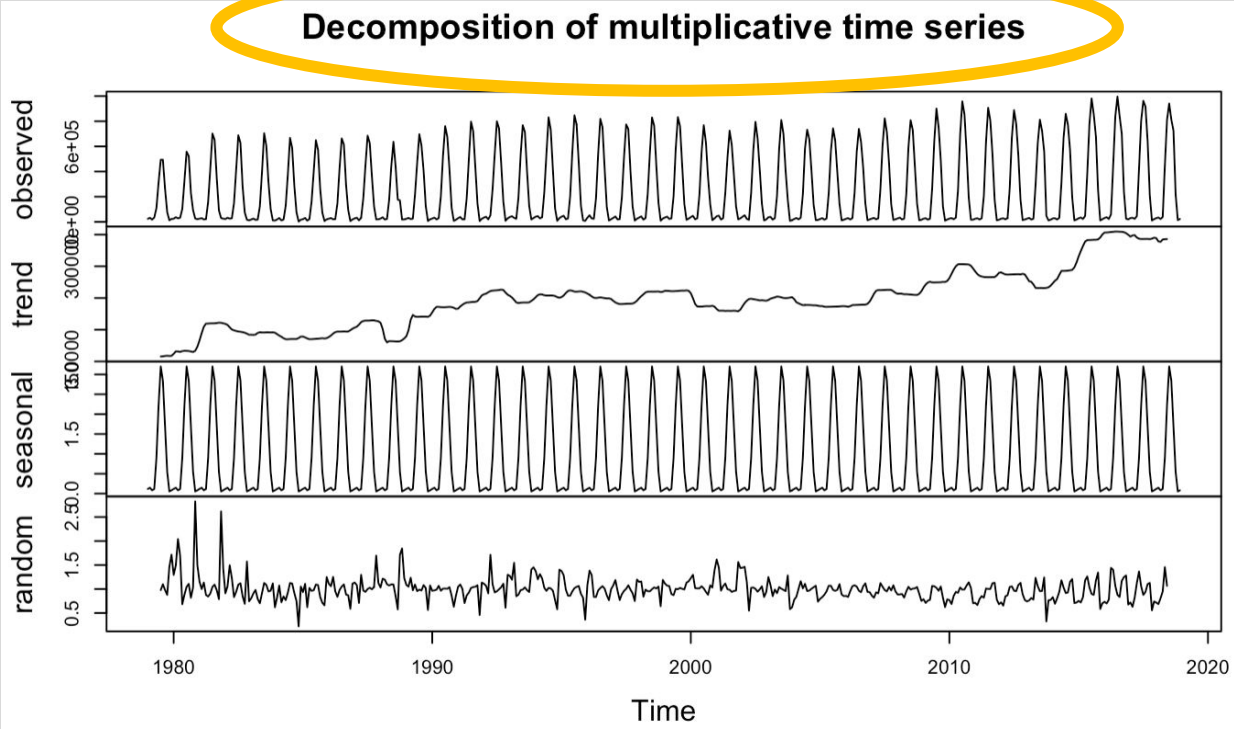
Source: U.S. National Park Service



Decomposition

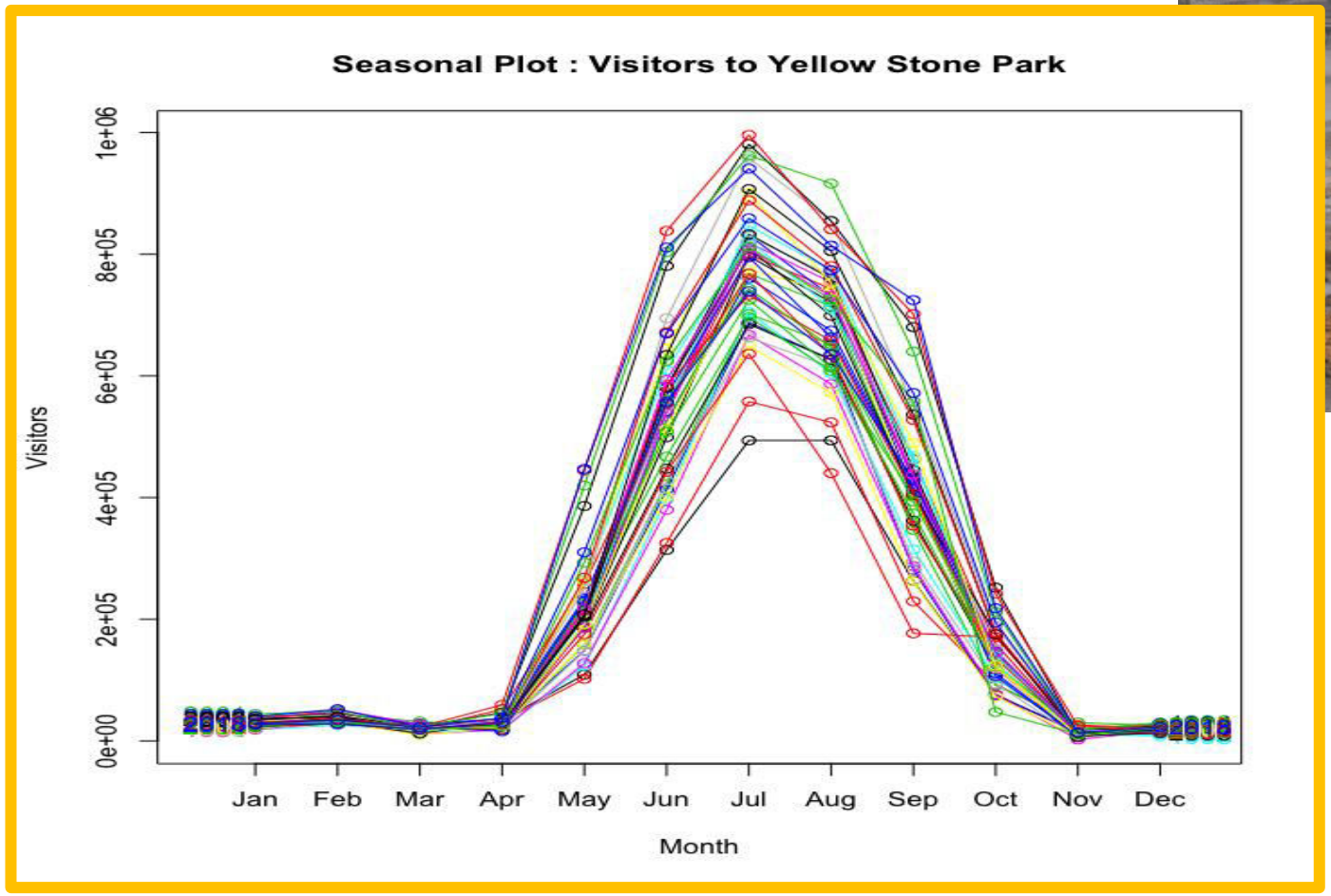
This decomposition represents better our data





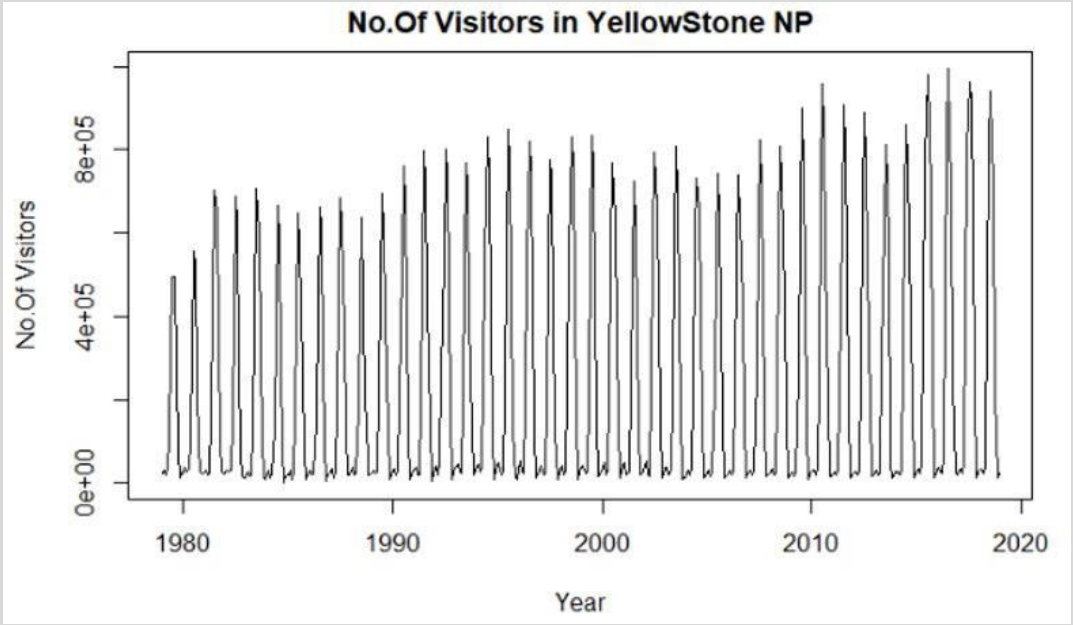
Seasonal/Monthly Plot

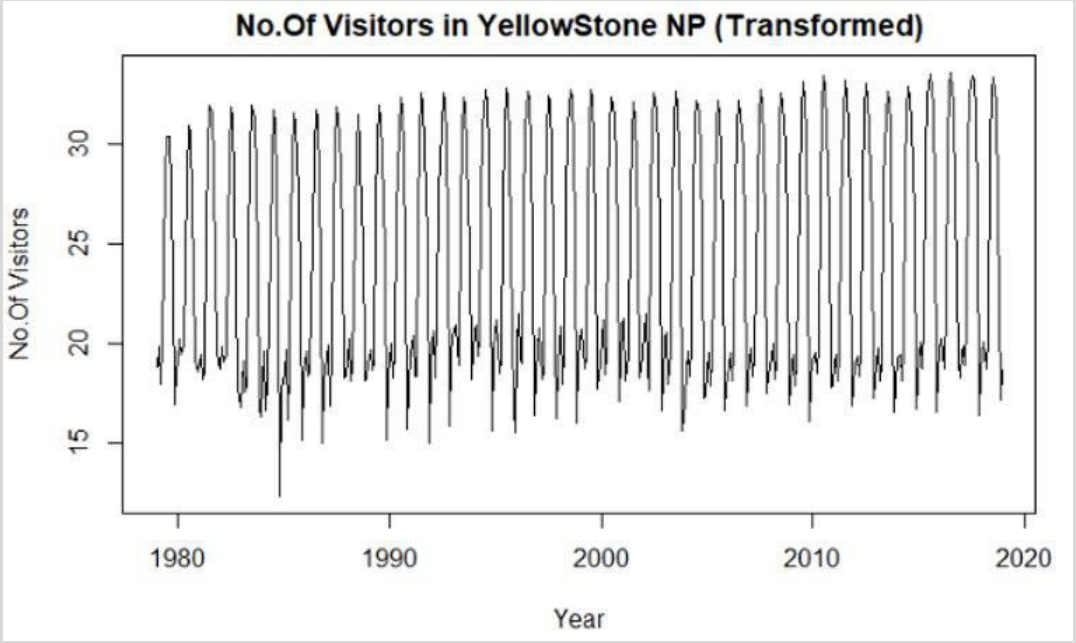
- Strong seasonality within each year
- November to April : low season
- May to October : busy season with peak in July



Box Cox

Lambda value: 0.1140685





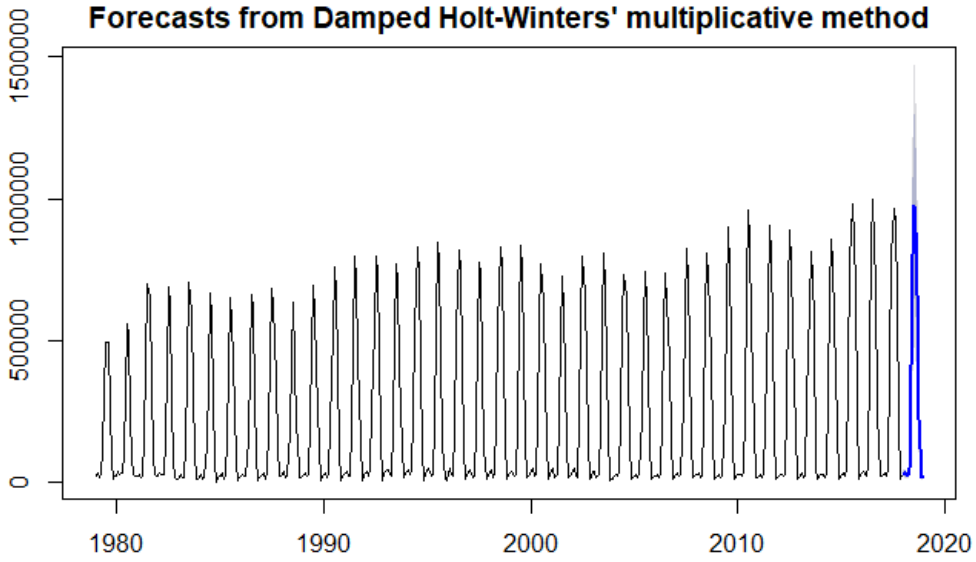
Exponential Smoothing Models

- Holt-Winters
- ETS

Holt-Winters

AICC	
Additive Damped	12770.92
Multiplicative	12242.32
Multiplicative Damped	12362.12

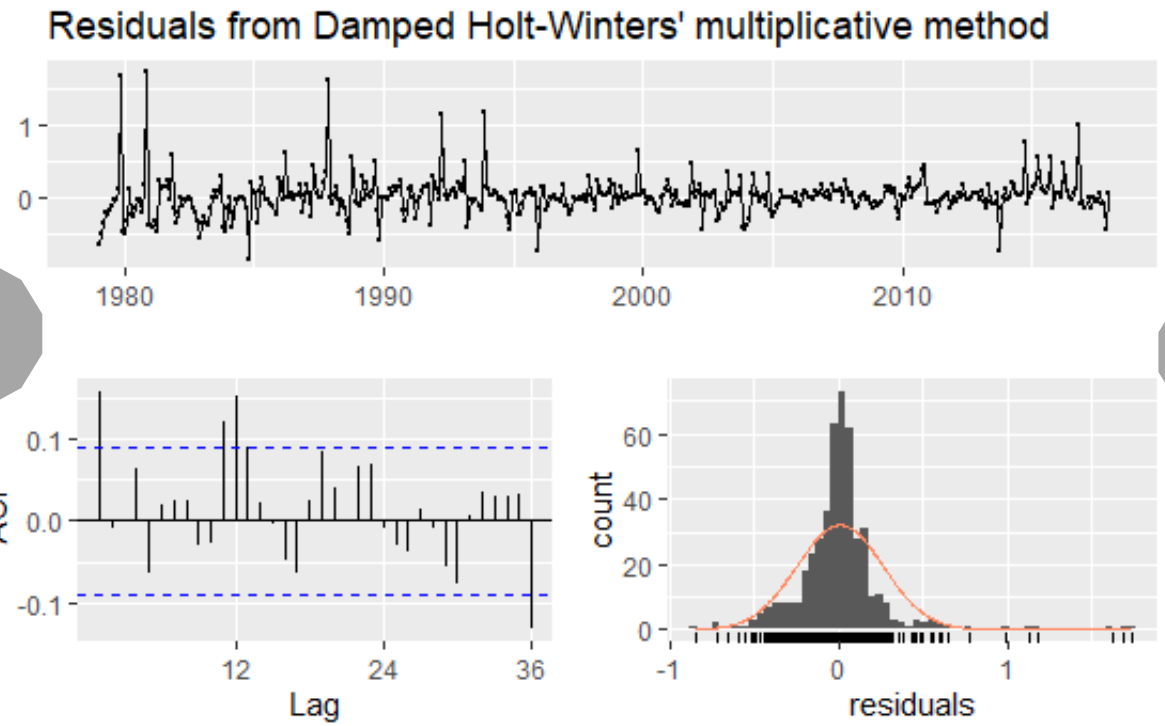
ERRORS		
	MAPE	RMSE
Additive Damped	10.329	31452.33
Multiplicative	8.7739	43541.78
Multiplicative Damped	6.3551	31175.87



Best Model: Multiplicative with Damped

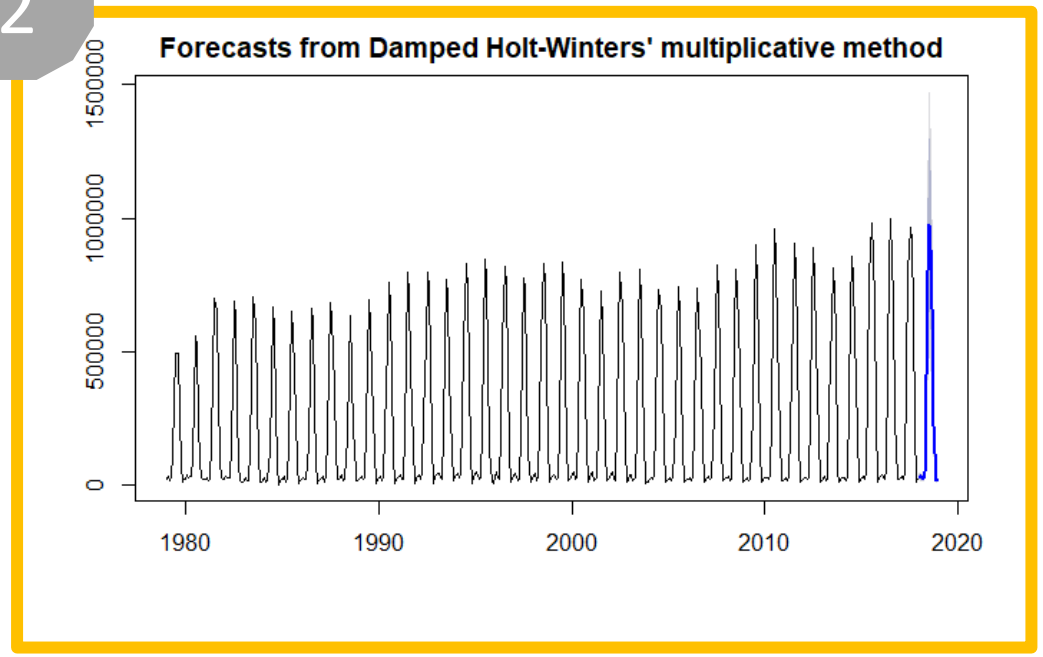
Holt-Winters

1




```
        Ljung-Box test

data:  Residuals from Damped Holt-Winters' multiplicative method
Q* = 51.941, df = 7, p-value = 5.997e-09

Model df: 17.   Total lags used: 24
```

2



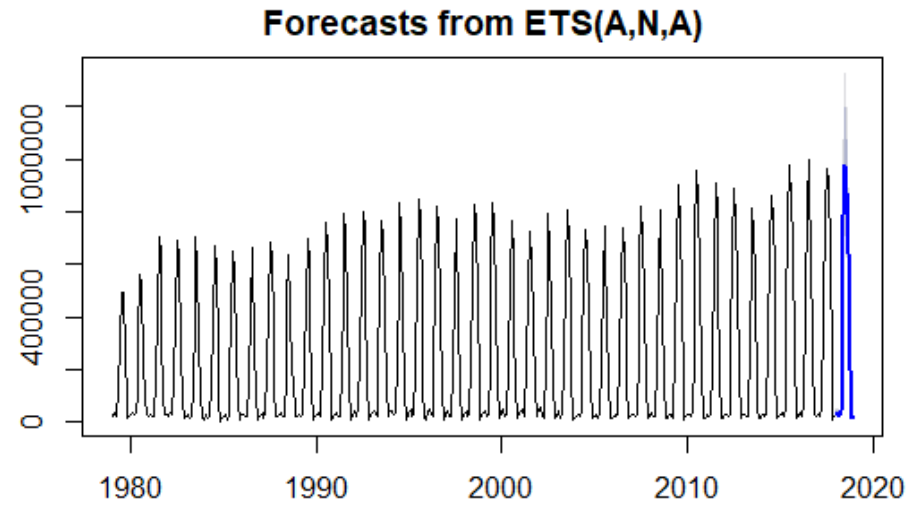
Exponential Smoothing (ETS)

```
ETS(A,N,A)

Call:
 ets(y = train, lambda = lambda_ts)

  Box-Cox transformation: lambda= 0.1141

  Smoothing parameters:
    alpha = 0.0983
    gamma = 0.257

  Initial states:
    l = 23.157
    s = -5.1802 -6.0519 -0.2159 4.7843 7.6906 8.1356
           6.0903 1.7034 -3.9509 -5.1951 -3.2089 -4.6013

  sigma:  0.7546

     AIC     AICc      BIC
2629.784 2630.846 2692.011
```

AICC	
Ets(MNM)	12118.12
Ets(ANA)	2630.84

Notes: Error, Trend, Seasonality

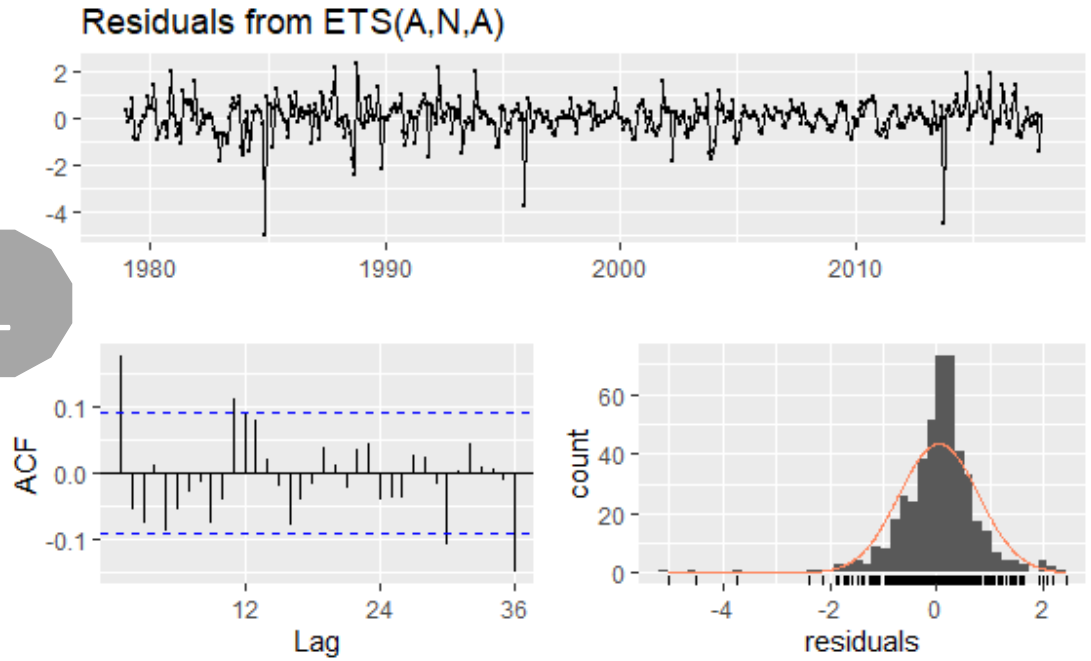
ERRORS		
	MAPE	RMSE
Ets(MNM)	8.2744	42872.09
Ets(ANA)	8.0576	38256.02



Best Model: Additive

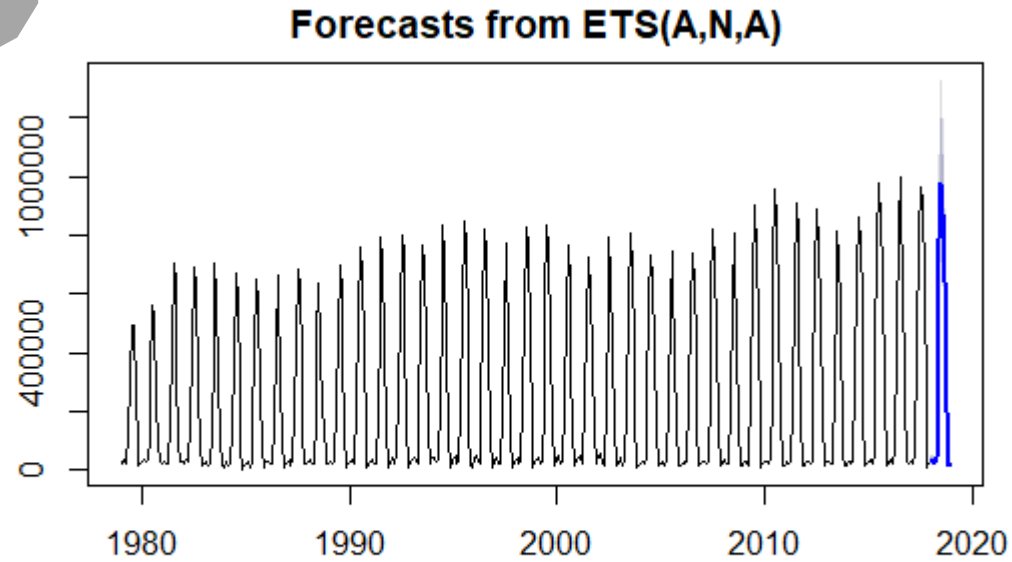
Exponential Smoothing (ETS)

1




```
        Ljung-Box test

data:  Residuals from ETS(A,N,A)
Q* = 49.684, df = 10, p-value = 3.051e-07

Model df: 14.   Total lags used: 24
```

2



Auto Regression Interated Moving Average Models

- sARIMA

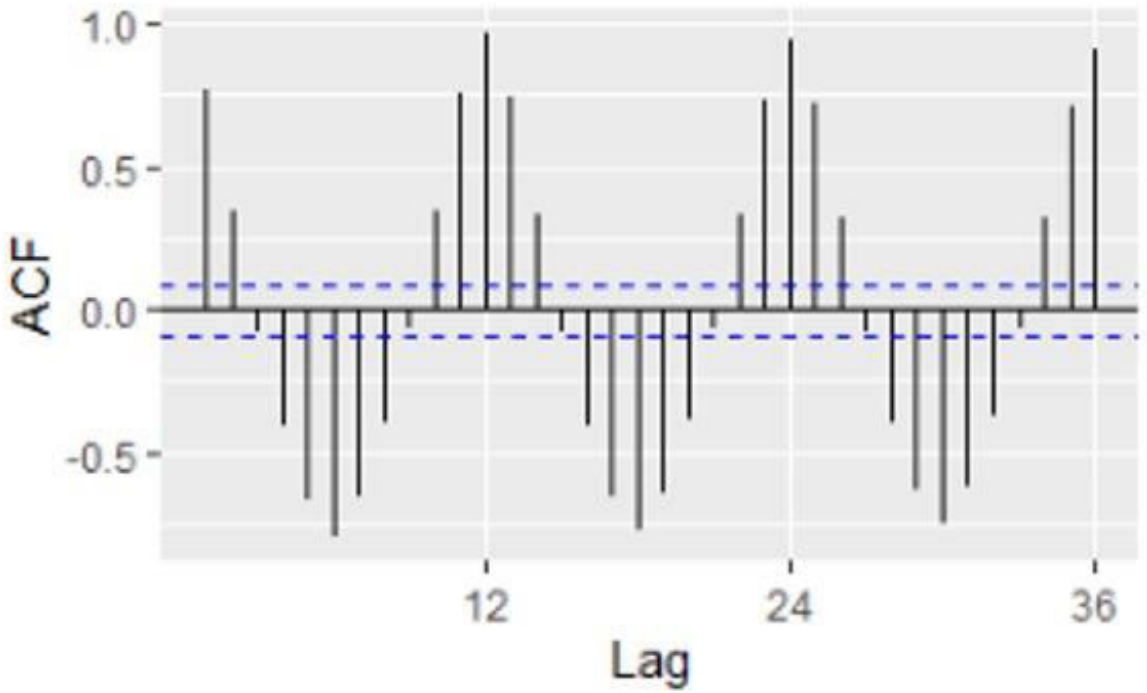
```
        Augmented Dickey-Fuller Test

data:  dfts
Dickey-Fuller = -22.15, Lag order = 7, p-value = 0.01
alternative hypothesis: stationary

p-value smaller than printed p-value[1] 0.01
```

```
        KPSS Test for Level Stationarity

data:  dfts
KPSS Level = 0.47569, Truncation lag parameter = 5, p-value = 0.04714

[1] 0.04714106
```

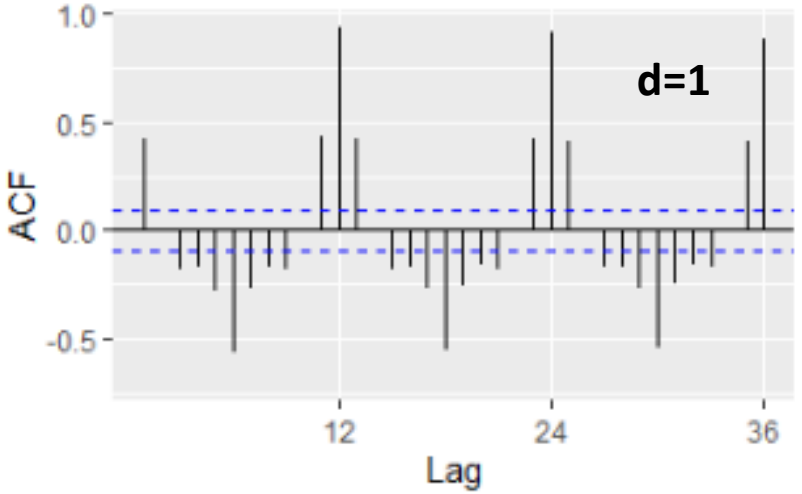
Is data Stationary?

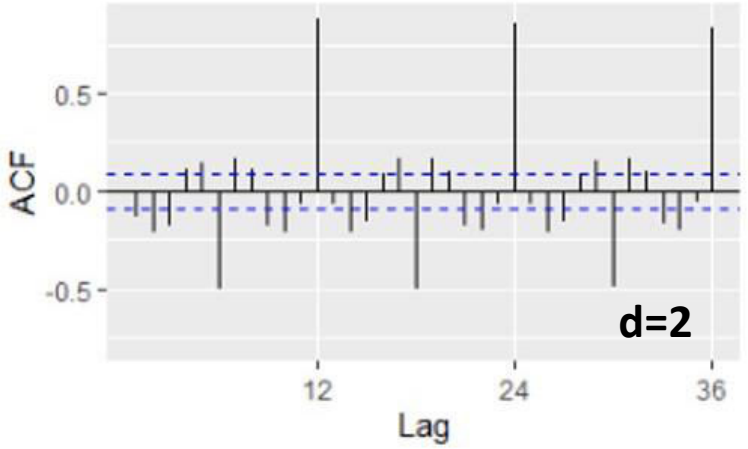
Conflict Conclusions:

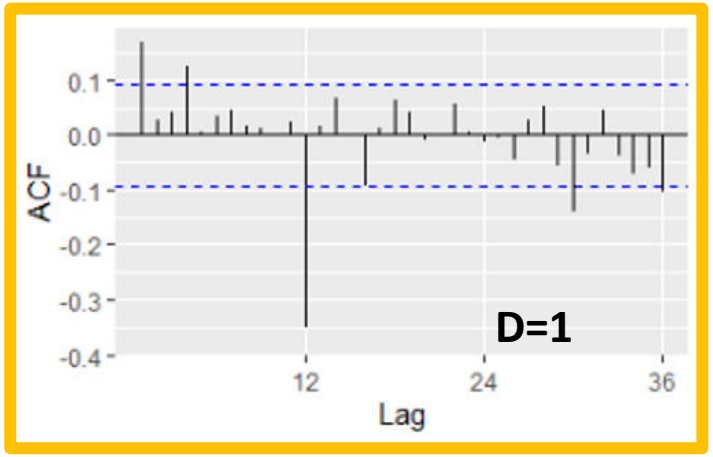
- ADF Test : Data is stationary
- KPSS Test: Data is not stationary



Differencing

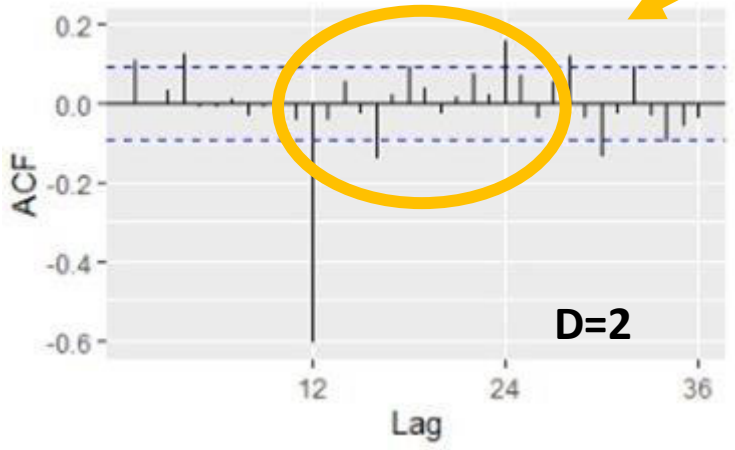


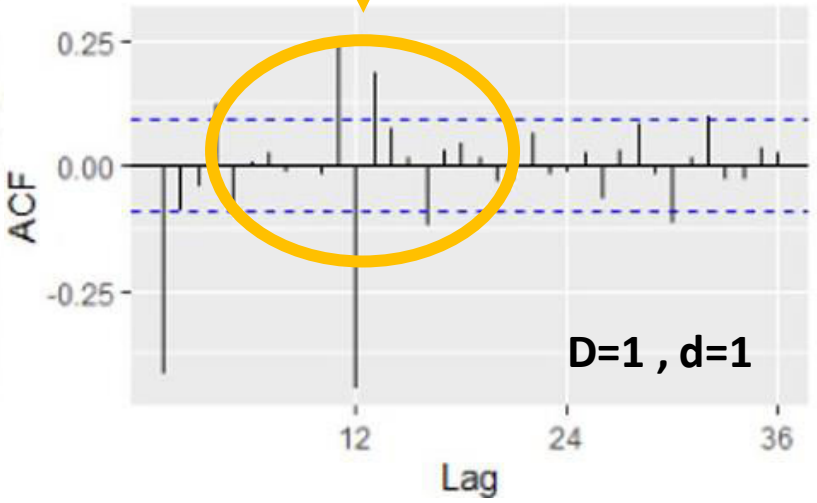




Preferred order of differencing

Over differencing





Stationarity Checks – After Differencing

```
p-value smaller than printed p-value
        Augmented Dickey-Fuller Test

data:  sdiff1_ndiff0
Dickey-Fuller = -6.4529, Lag order = 7, p-value = 0.01
alternative hypothesis: stationary
```

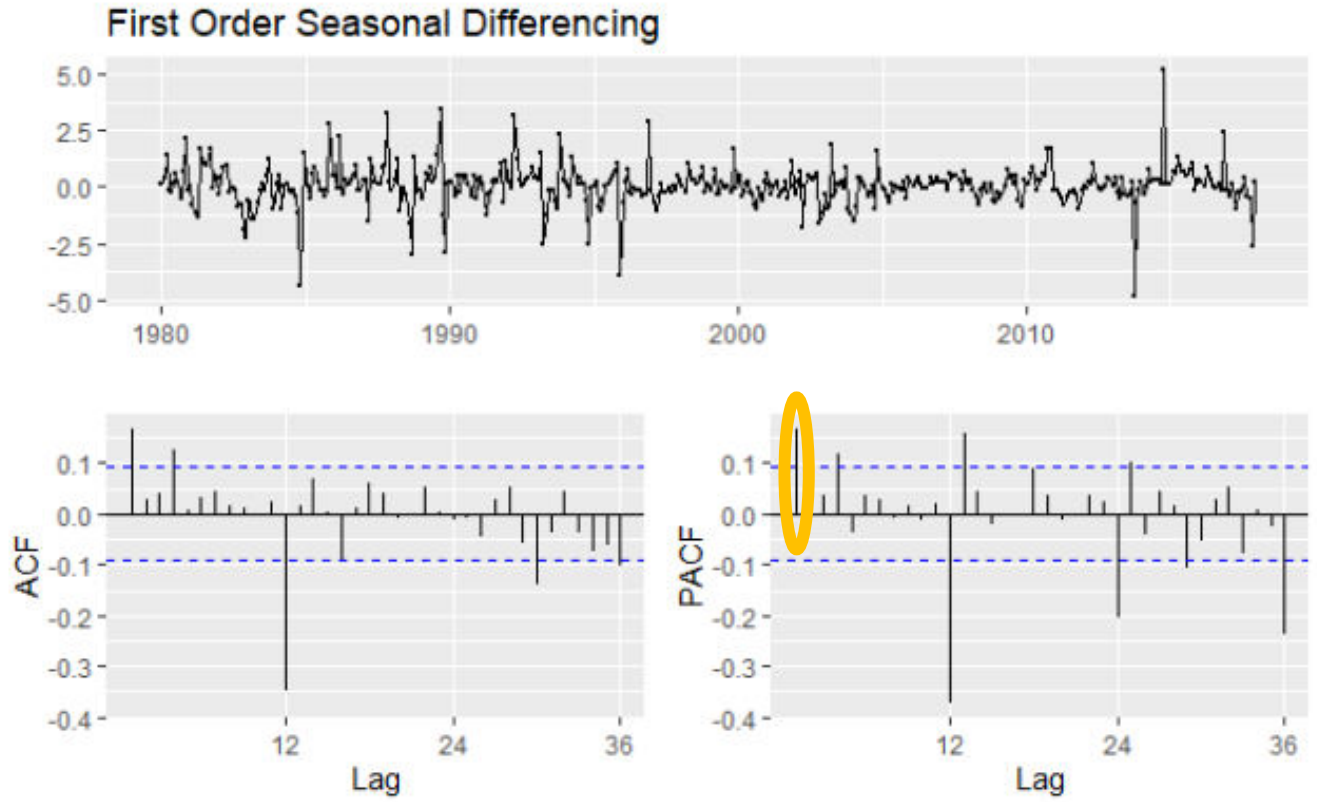
Same Conclusions:

- ADF Test : Data is stationary
- KPSS Test: Data is stationary

```
p-value greater than printed p-value
        KPSS Test for Level Stationarity

data:  sdiff1_ndiff0
KPSS Level = 0.050777, Truncation lag parameter = 5, p-value = 0.1
```

sARIMA

Seasonal Differencing – First Order



	ACF	PACF	Interpretation
Non-Seasonal	Exponential decay	Cut Off after lag=1	AR(1) model
Seasonal	Cut Off after lag =12 (12*1)	Exponential Decay	MA(1) model
Model interpreted based on preferred differencing: sARIMA{(1,0,0)(0,1,1)[12]}			

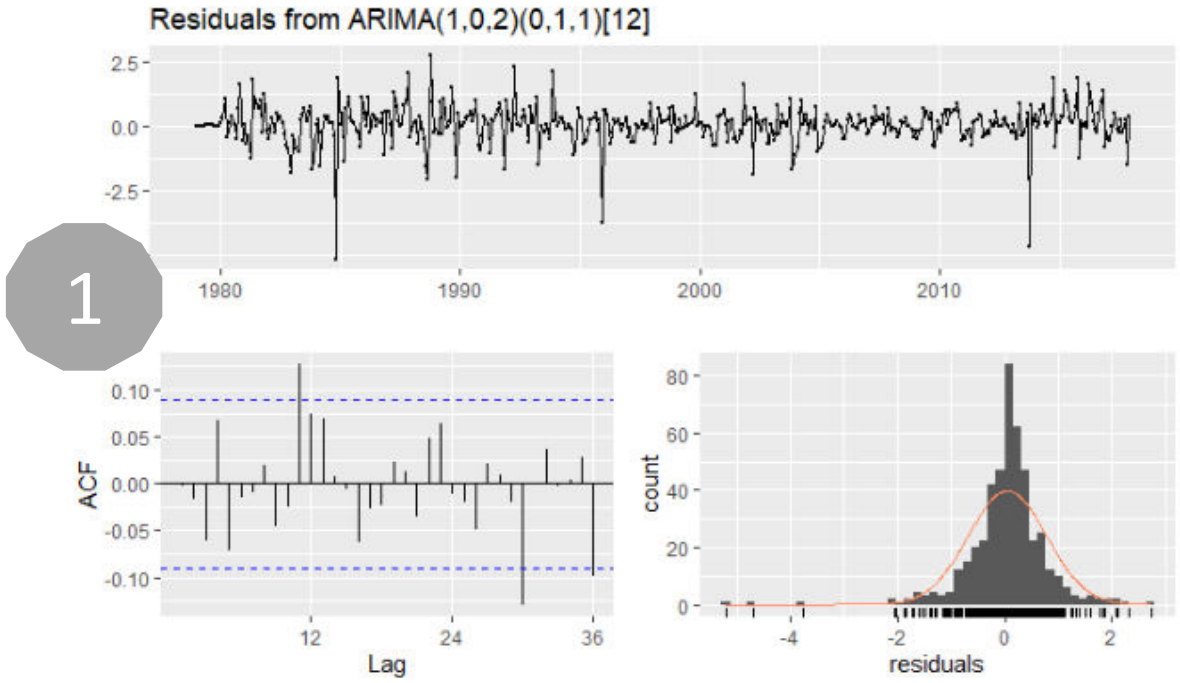
SARIMA Model Selection

S.No	Model Type	AICc	Co-efficients							Ljung-Box test	Forecasting Errors (h-12)	
			ar1	ar2	ma1	ma2	sar1	sma1	sma2	p value	RMSE	MAPE
Model1	sARIMA{(1,0,0)(0,1,1)[12]}	1050.2	0.3246	NA	NA	NA	NA	-0.6644	NA	0.004487	39817.83	7.001718
Model2	sARIMA{(1,0,0)(1,1,1)[12]}	1046.91	0.3097	NA	NA	NA	0.1628	-0.7555	NA	0.004635	46856.63	7.984333
Model3	sARIMA{(1,0,0)(0,1,2)[12]}	1047.42	0.3124	NA	NA	NA	NA	-0.6005	-0.1001	0.004097	32048.23	15.19112
Model4	sARIMA{(1,0,1)(0,1,1)[12]}	1049.04	0.9341	NA	-0.8026	NA	NA	-0.6882	NA	0.001153	38167.52	6.433607
Model5	sARIMA{(2,0,1)(0,1,1)[12]}	1036.44	1.1802	-0.2043	-0.9042	NA	NA	-0.7199	NA	0.0924	37625.66	7.528159
Model6	sARIMA{(1,0,2)(0,1,1)[12]}	1035.08	0.9682	NA	-0.6809	-0.1925	NA	-0.7227	NA	0.1169	37815.34	7.712205
Model7	sARIMA{(1,0,2)(1,1,1)[12]}	**1031.66**	0.9725	NA	-0.6998	-0.1857	0.1512	-0.7966	NA	0.174	42635.58	7.597139
Model8	sARIMA{(1,0,2)(1,1,2)[12]}	1033.73	0.9723	NA	-0.6995	-0.1857	0.1694	-0.815	0.0138	0.1377	42622.42	7.589048

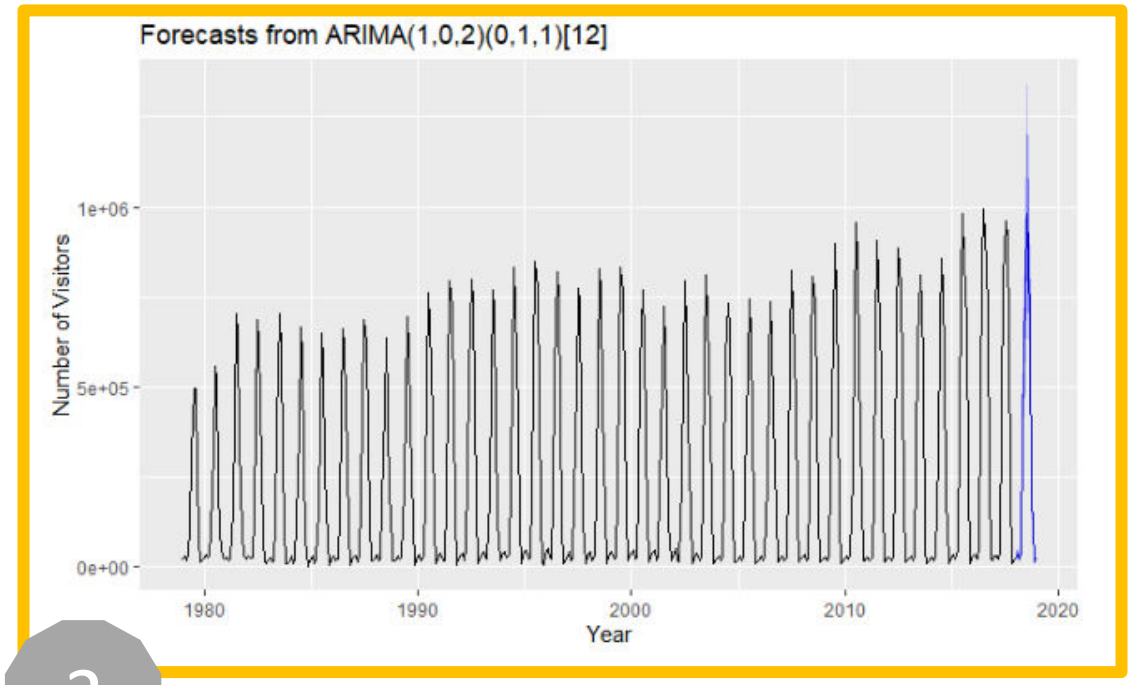
Selected best model

Best model selected by auto.arima

sARIMA



1



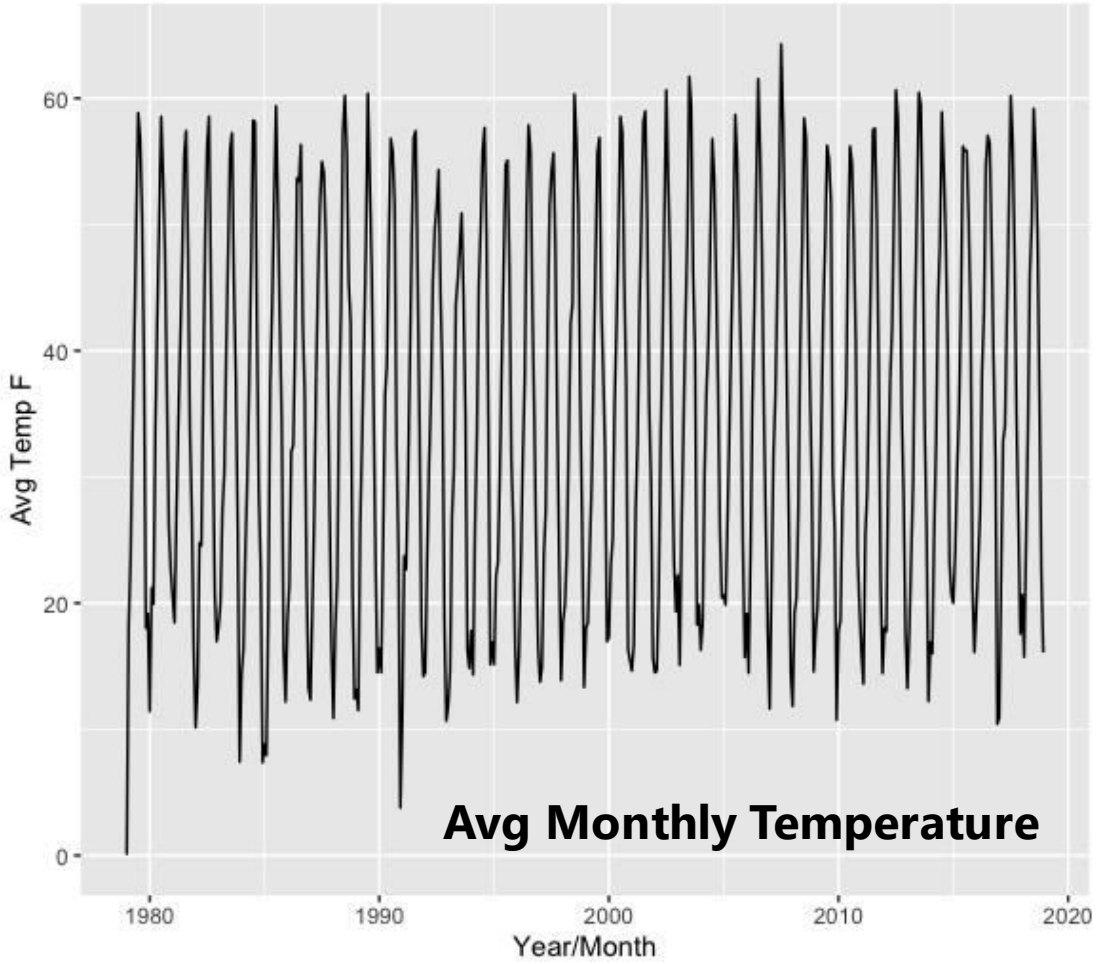
2

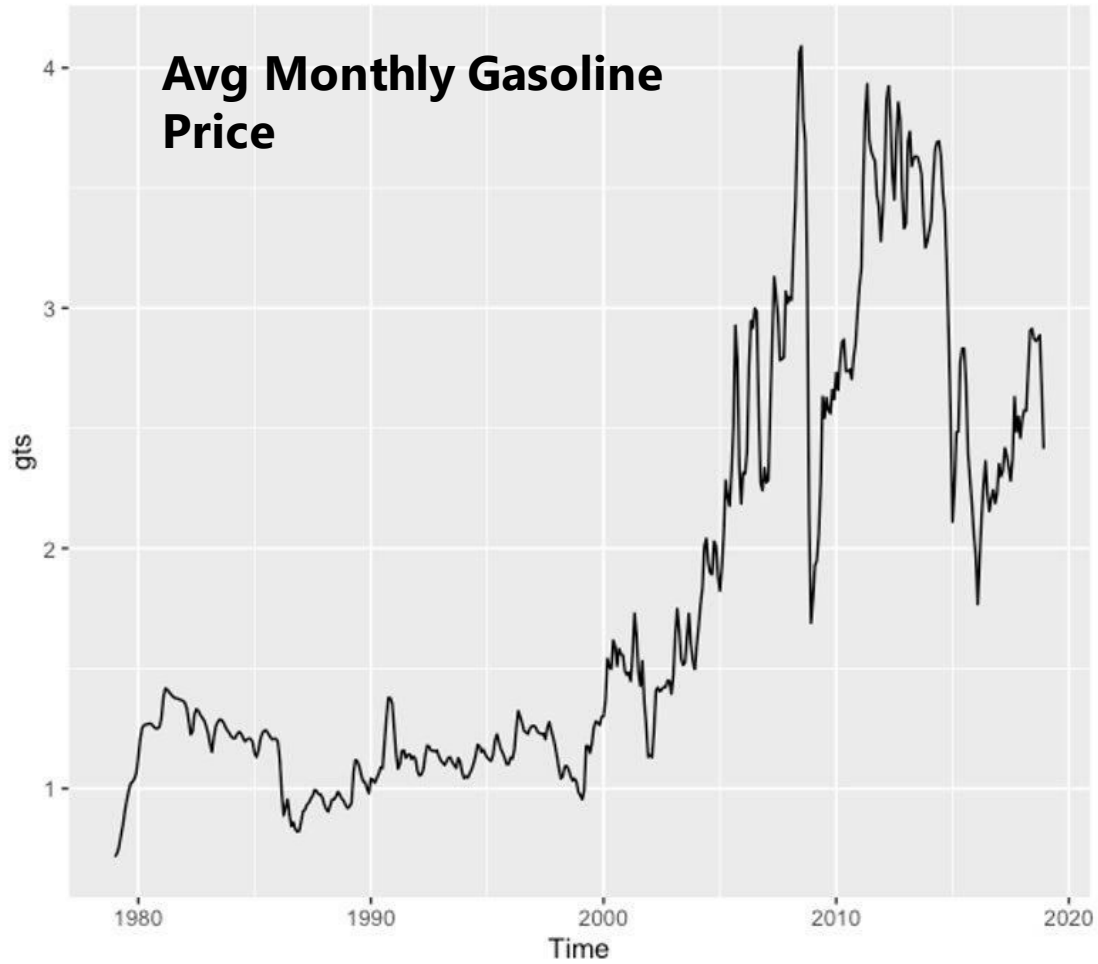
```
        Ljung-Box test

data:  Residuals from ARIMA(1,0,2)(0,1,1)[12]
Q* = 27.694, df = 20, p-value = 0.1169

Model df: 4.   Total lags used: 24
```

Regression Models

- ❑ TSLM (linear model)
- ❑ ARIMAX
- ❑ VAR

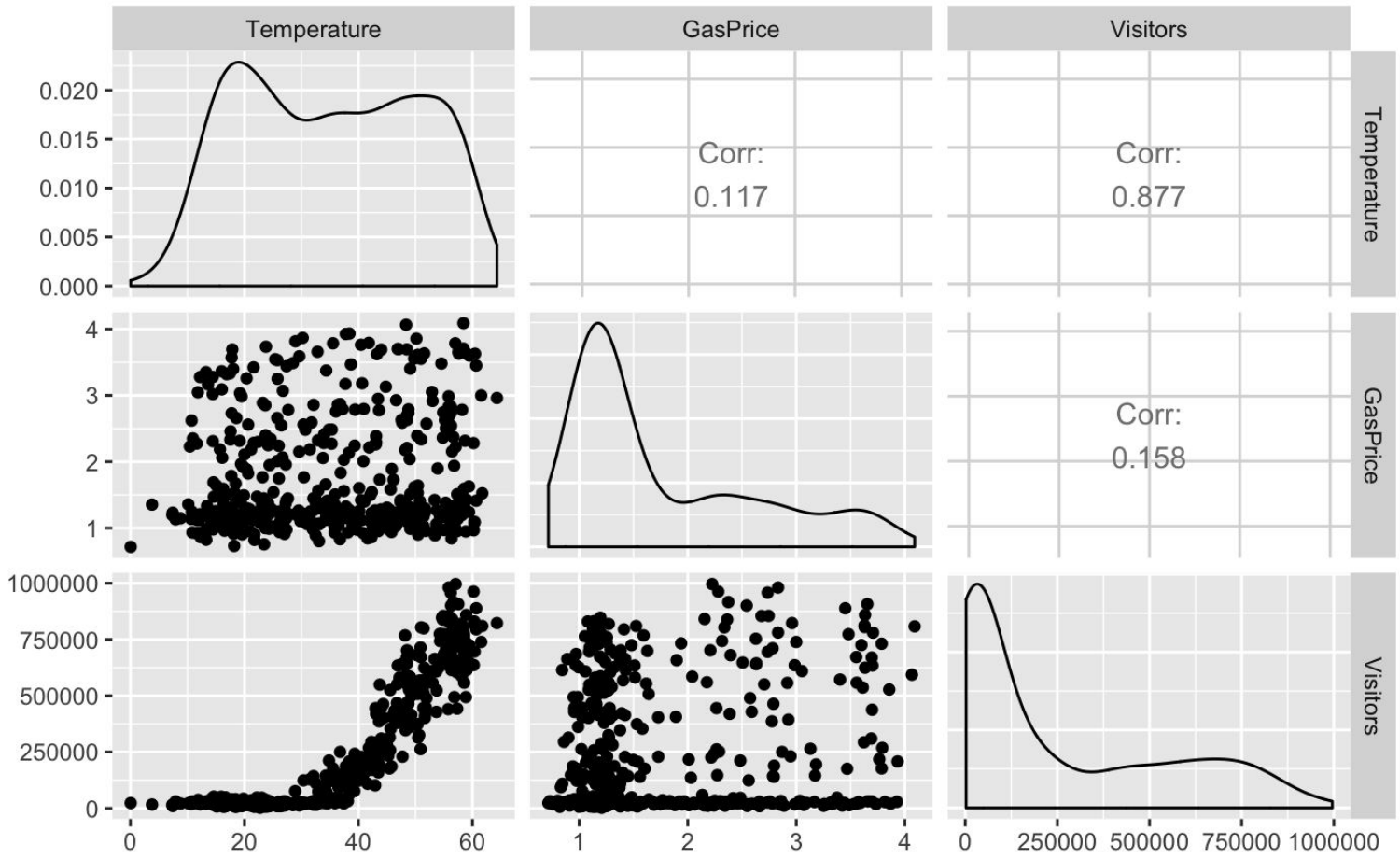
Predicting variables





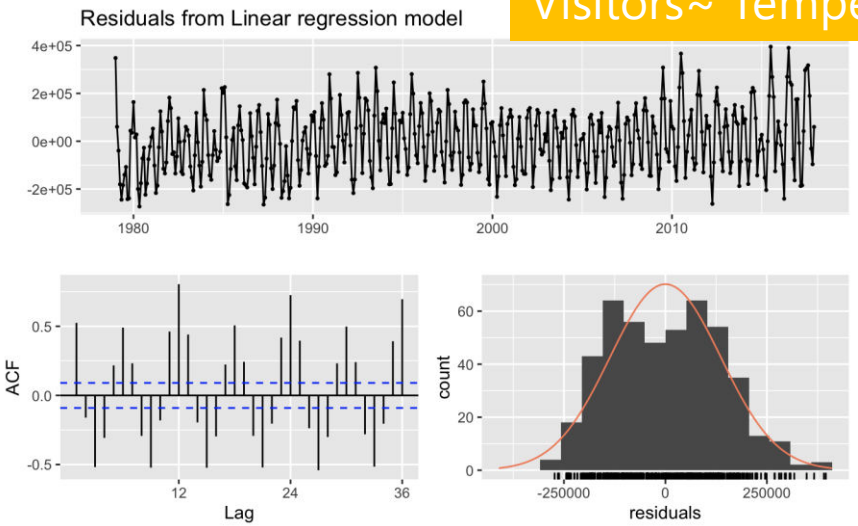
Sources: Temperature: National Center for Environmental Information ; Gas Price: U.S. Energy Information Administration

Correlation between variables

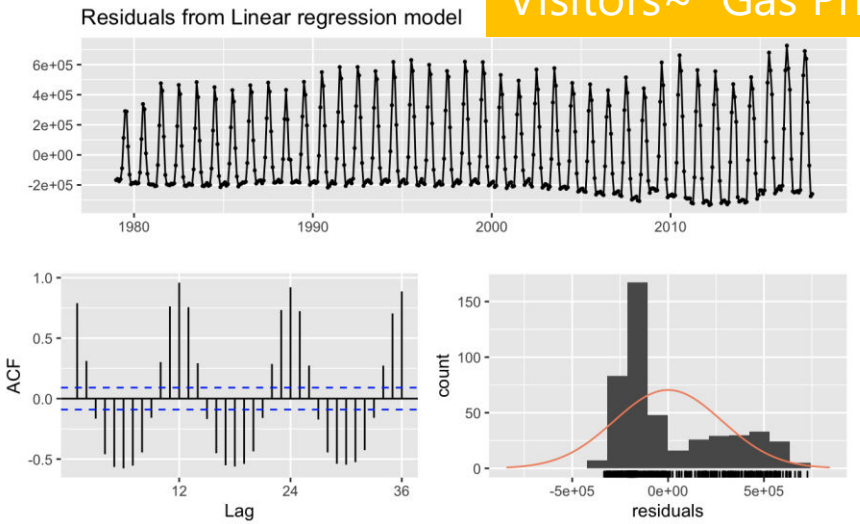


Time Series Linear Model

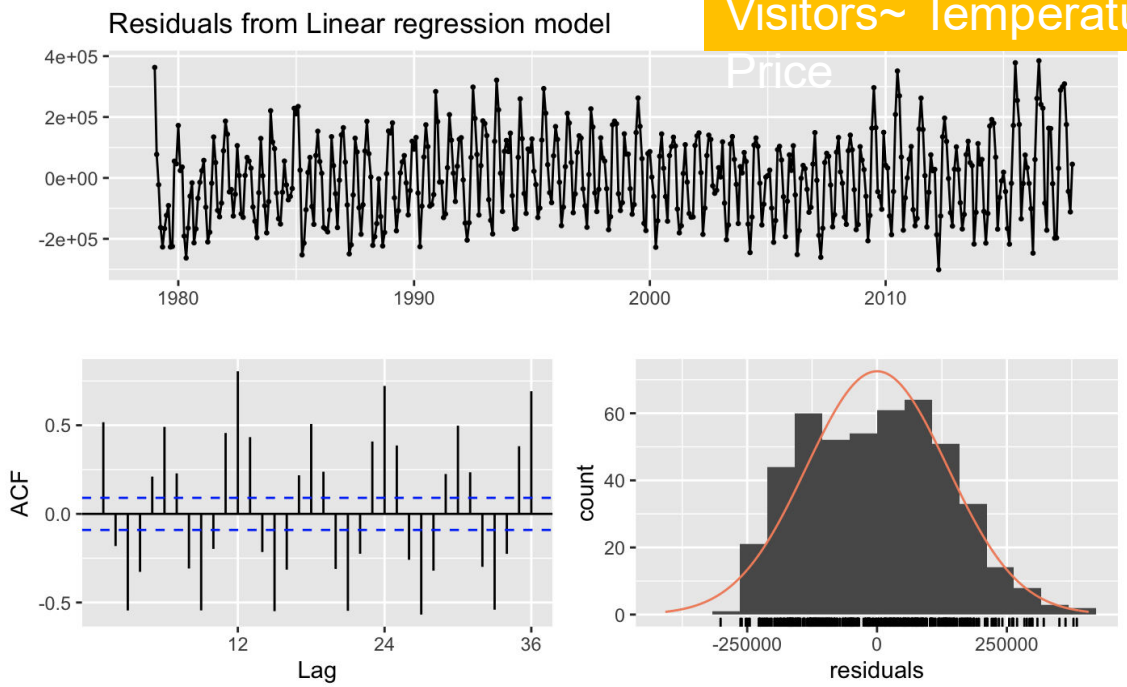
Visitors~ Temperature



Visitors~ Gas Price



Visitors~ Temperature + Gas Price



- Residuals are NOT white noise (Ljung-Box Test p-value < 2.2e-16)
- Residuals has ARIMA structure

ARIMAX

1) Auto.Arima : ARIMA{(1,0,2)(1,1,1)[12]}

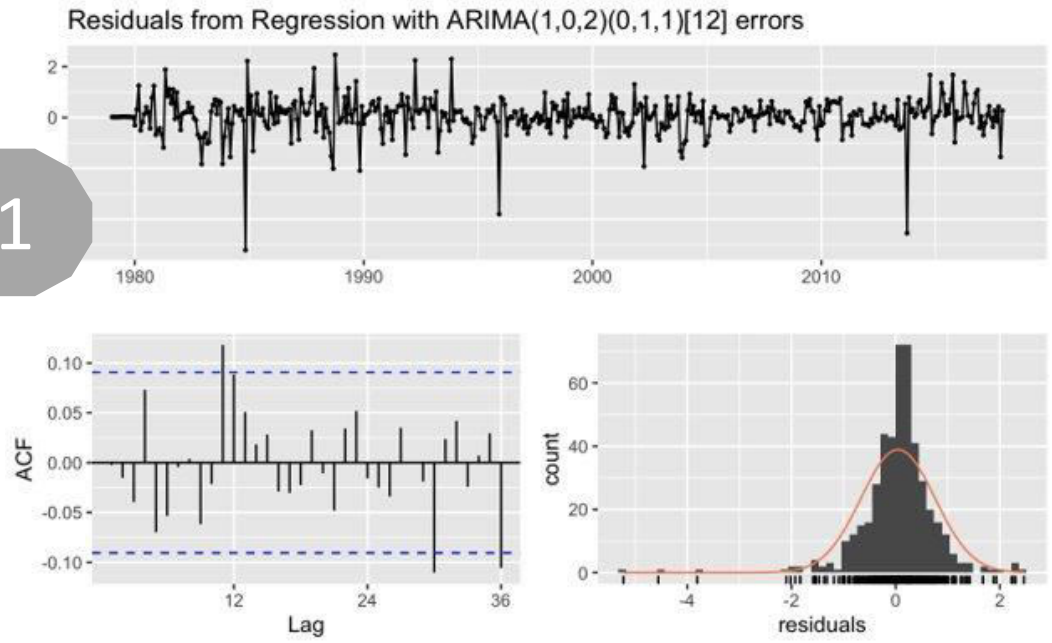
2) Arima: ARIMA{(1,0,2)(0,1,1)[12]}

- Variable Gas Price is not significant in 4 models that have Gas price as one of the predicting variables
- Residuals in all 6 models represent white noise

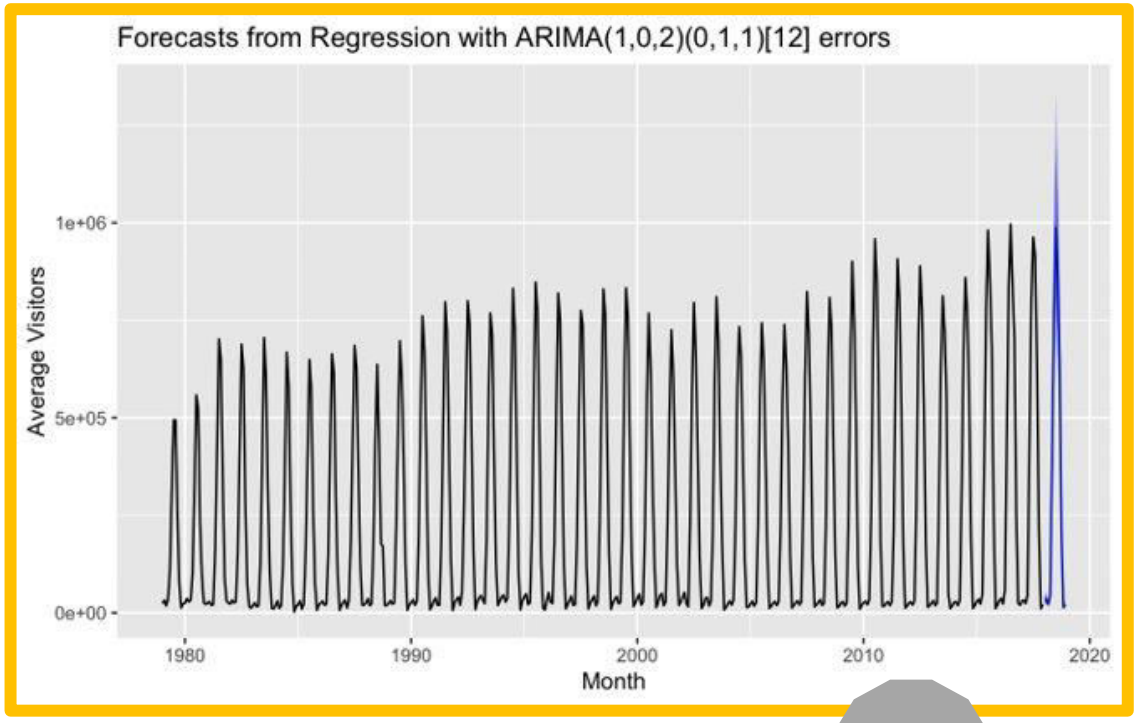
Models	RMSE	MAPE	AICc
auto.arima_both	40493.11	8.35	1007.7
auto.arima_temp	39979.91	8.26	1007.7
auto.arima_gas	43134.11	7.3	1029.76
Arima_both	36096.97	7.94	1014.16
Arima_temp	35658.78	7.69	1013.59
Arima_gas	38145.69	7.31	1035.15

ARIMAX

Visitors ~ Temperature , ARIMA(1,0,2)(0,1,1)[12]



1



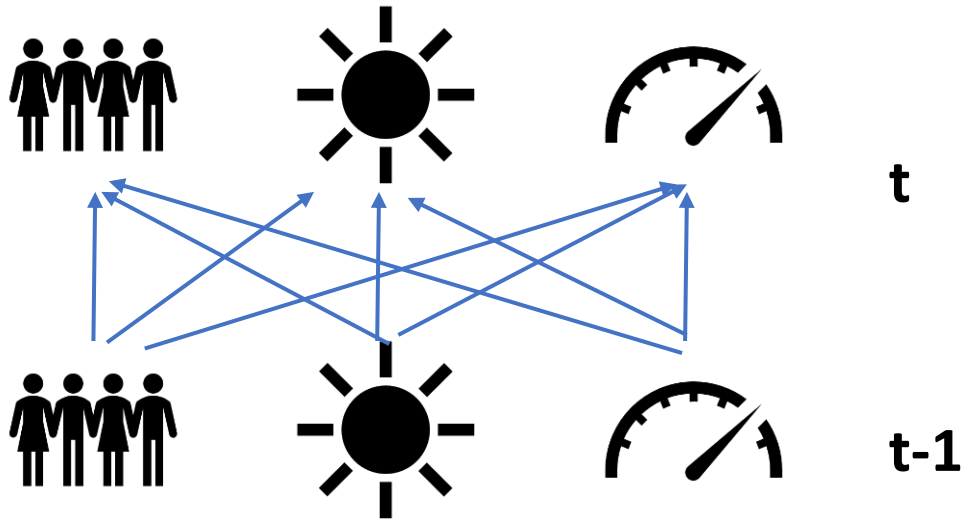
2

```
        Ljung-Box test

data:  Residuals from Regression with ARIMA(1,0,2)(0,1,1)[12] errors
Q* = 26.254, df = 19, p-value = 0.1233

Model df: 5.   Total lags used: 24
```

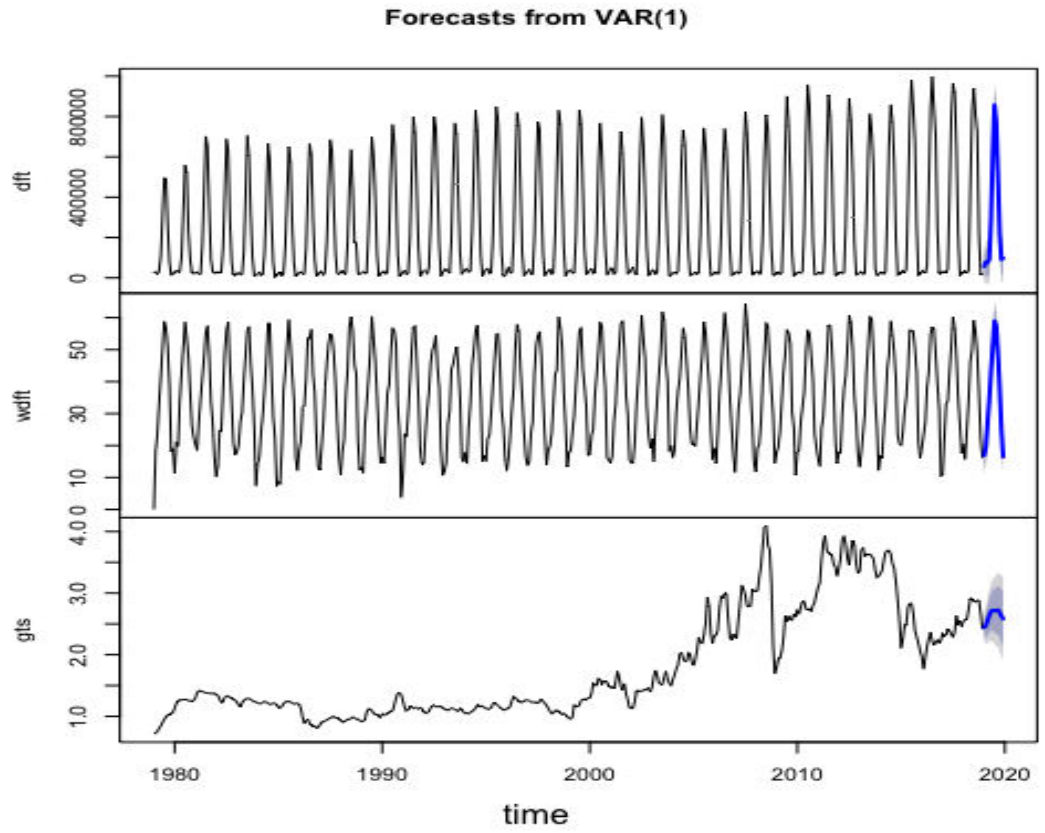
VAR (Vector Auto Regression)

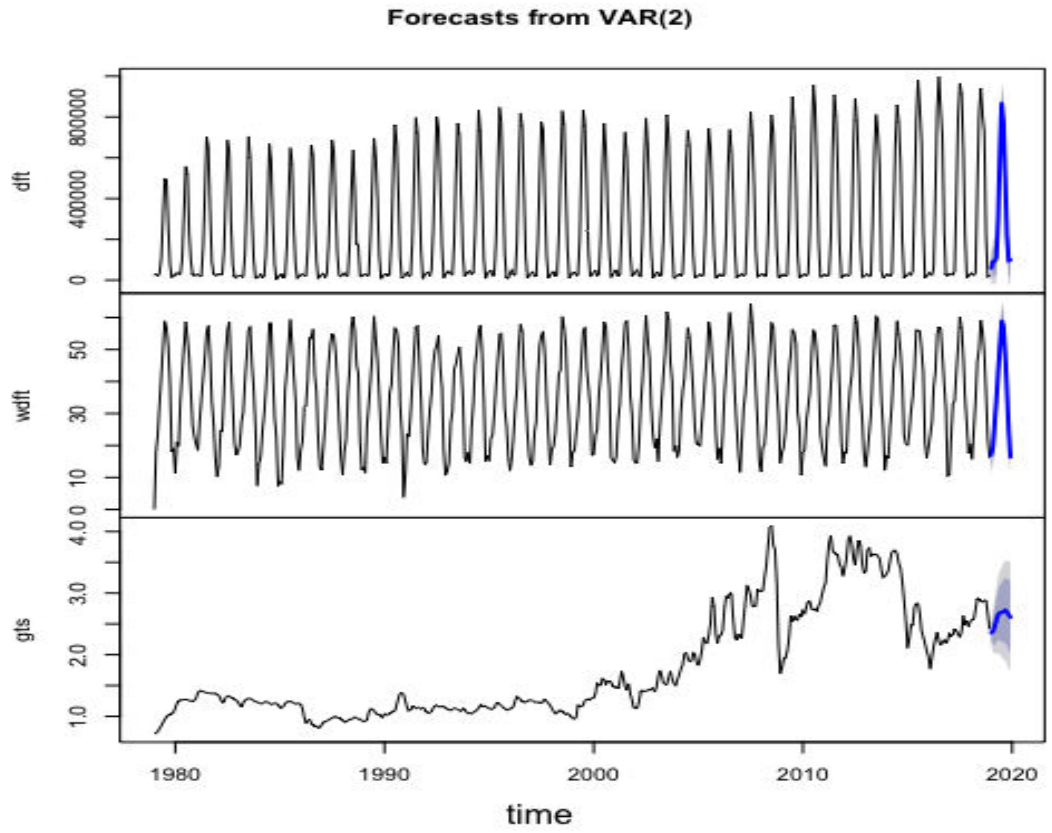
$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} + \begin{bmatrix} \phi_{11,1} & \phi_{12,1} \\ \phi_{21,1} & \phi_{22,1} \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-2} \end{bmatrix} + \begin{bmatrix} e_{1,t} \\ e_{2,t} \end{bmatrix}$$

- 3 Independent Variables (Visitors, Weather, Price of Gas)
- Monthly Average Temperature shares same seasonality as the amount of visitors
- Price of Gas is more difficult to tell if there is relationship although there is certainly a cyclical nature to the time series



VAR Model Comparison



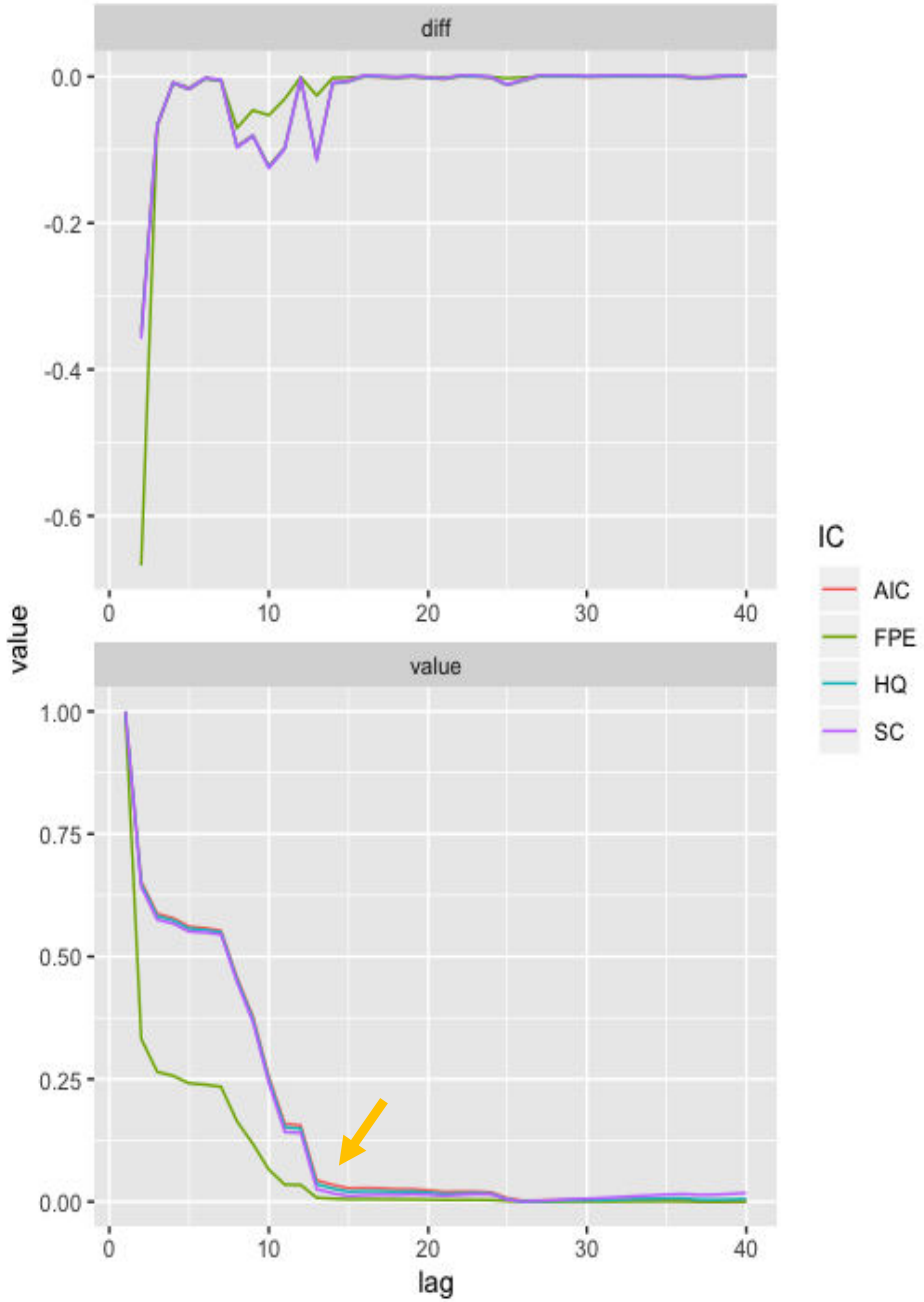


- Using Seasonal (12 periods) – creates dummy variables
- Eye test was not good enough to determine lag

VAR lag-value

- Using VARSelect package, we can use AIC to determine lag required
- AIC appears to get low enough at the 12 lag mark
- We can use more lag but the benefit decreases
- P-values were difficult to interpret

	Estimate	Std. Error	t value	Pr(>|t|)	
dft.l1	-0.00000001086	0.00000015381	-0.071	0.9438	
wdft.l1	-0.00062405845	0.00157141884	-0.397	0.6915	
gts.l1	1.47342736014	0.04903571348	30.048	< 2e-16	***
dft.l2	-0.00000020591	0.00000016346	-1.260	0.2085	
wdft.l2	0.00052606518	0.00158948402	0.331	0.7408	
gts.l2	-0.71608490805	0.08712735885	-8.219	2.59e-15	***
dft.l3	0.00000016649	0.00000016498	1.009	0.3135	
wdft.l3	-0.00230069882	0.00159173982	-1.445	0.1491	
gts.l3	0.20779724791	0.09353099115	2.222	0.0268	*
dft.l4	-0.00000001833	0.00000017177	-0.107	0.9151	
wdft.l4	0.00074261836	0.00159003734	0.467	0.6407	
gts.l4	0.04015295931	0.09432853361	0.426	0.6706	
dft.l5	-0.00000025134	0.00000017396	-1.445	0.1493	
wdft.l5	0.00130837043	0.00158805489	0.824	0.4105	
gts.l5	-0.07359490126	0.09422134002	-0.781	0.4352	
dft.l6	0.00000035316	0.00000017596	2.007	0.0454	*
wdft.l6	0.00105167786	0.00159192030	0.661	0.5092	
gts.l6	-0.01199724050	0.09400552877	-0.128	0.8985	
dft.l7	-0.00000036132	0.00000017564	-2.057	0.0403	*
wdft.l7	0.00369423503	0.00158707503	2.328	0.0204	*
gts.l7	0.02793728947	0.09319923328	0.300	0.7645	
dft.l8	0.00000030901	0.00000017646	1.751	0.0807	.
wdft.l8	-0.00048799199	0.00159843055	-0.305	0.7603	
gts.l8	0.01268315492	0.09259904968	0.137	0.8911	
dft.l9	-0.00000015527	0.00000017799	-0.872	0.3835	
wdft.l9	0.00004184269	0.00160105541	0.026	0.9792	
gts.l9	-0.08473390832	0.09238190800	-0.917	0.3596	
dft.l10	0.00000010465	0.00000017821	0.587	0.5574	

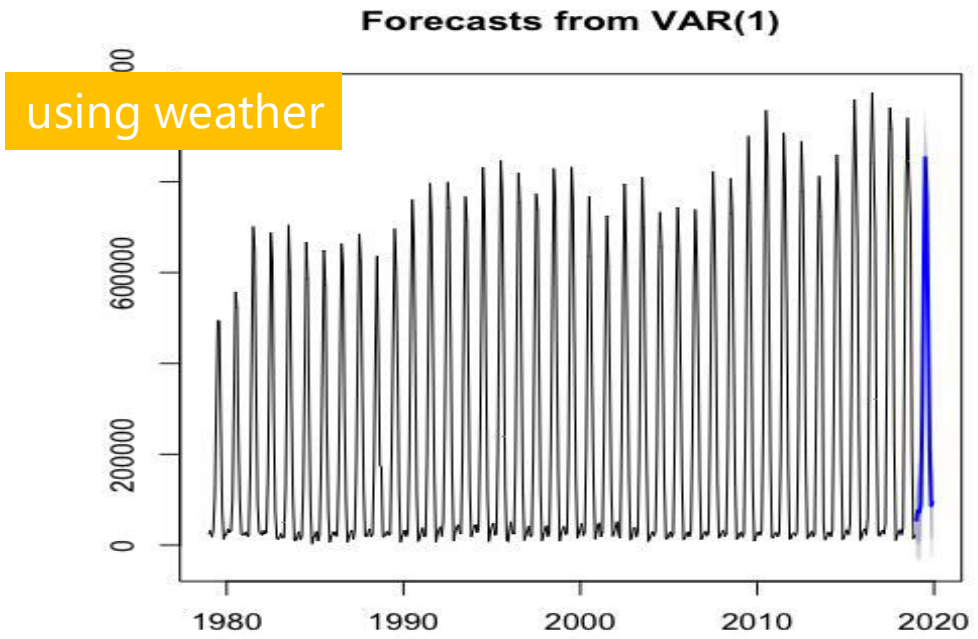


VAR Model

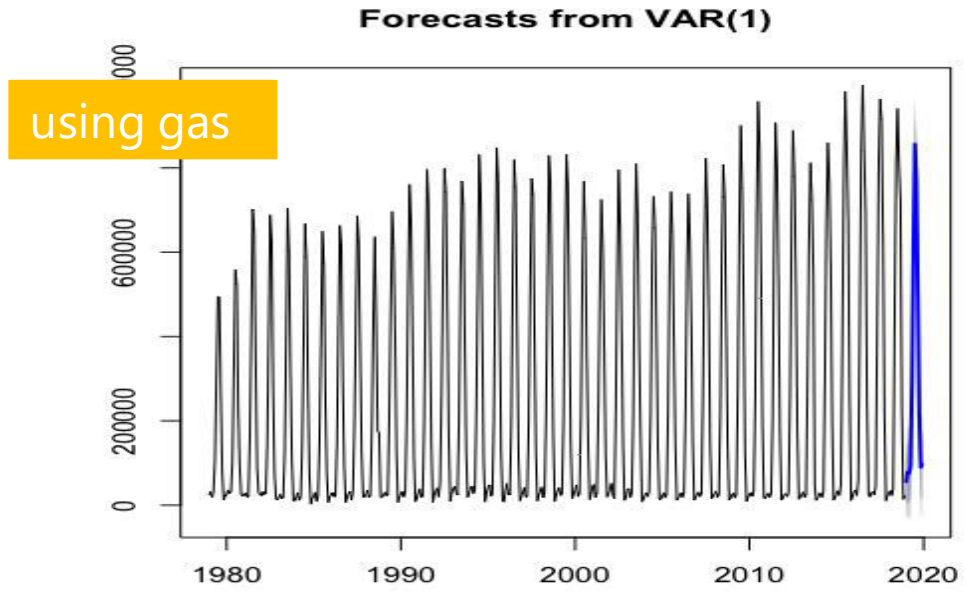
- When we split VAR model using only the weather or gasoline variable, AICC changes drastically
- The VAR model mostly depends on the visitor's lag values
- Consider using only gas variable

```
[1] "AIC For Var(1) with Weather/Gas"
[1] 13317.58
[1] "AIC For Var(1) with Weather"
[1] 14044.34
[1] "AIC For Var(1) with Gas"
[1] 10879.87
```

using weather



using gas

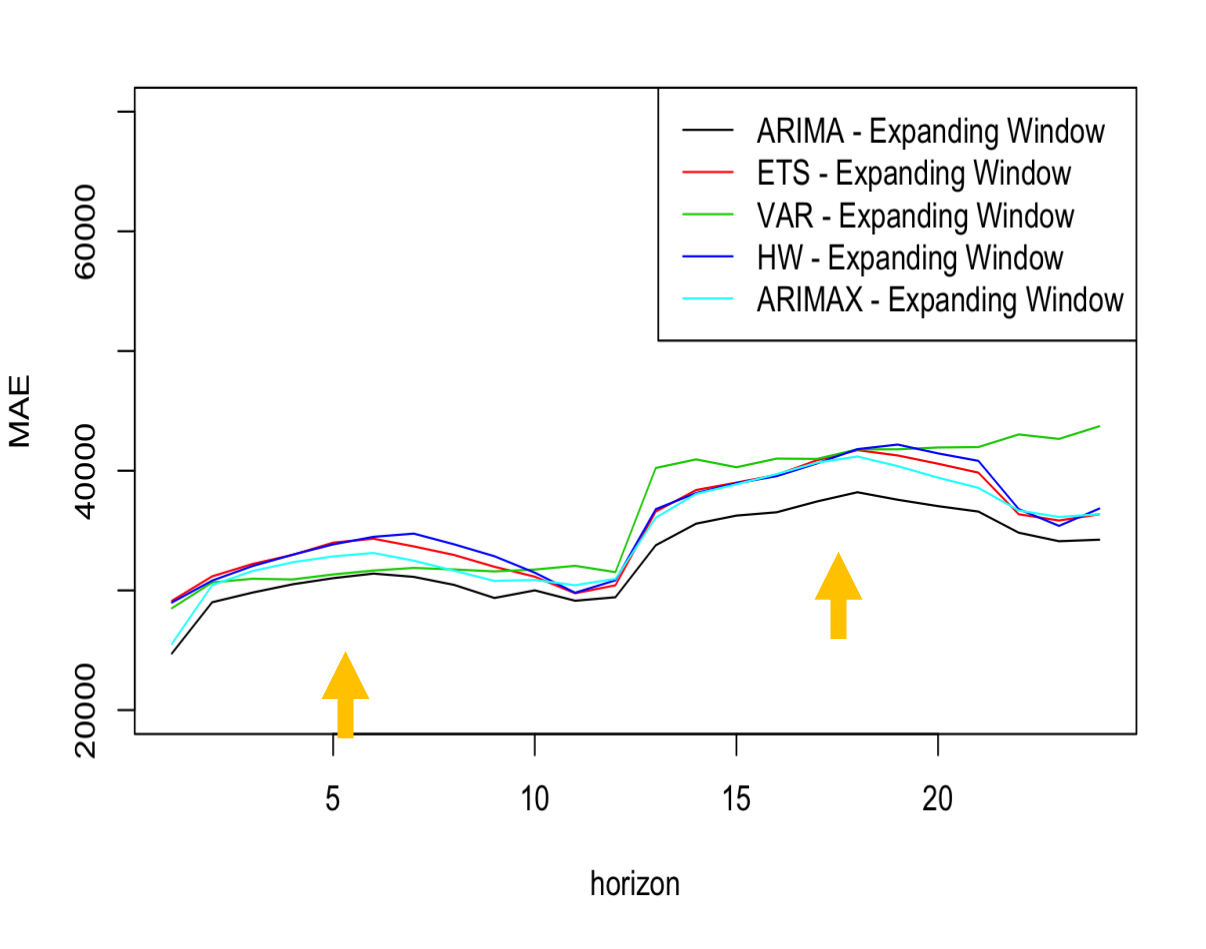


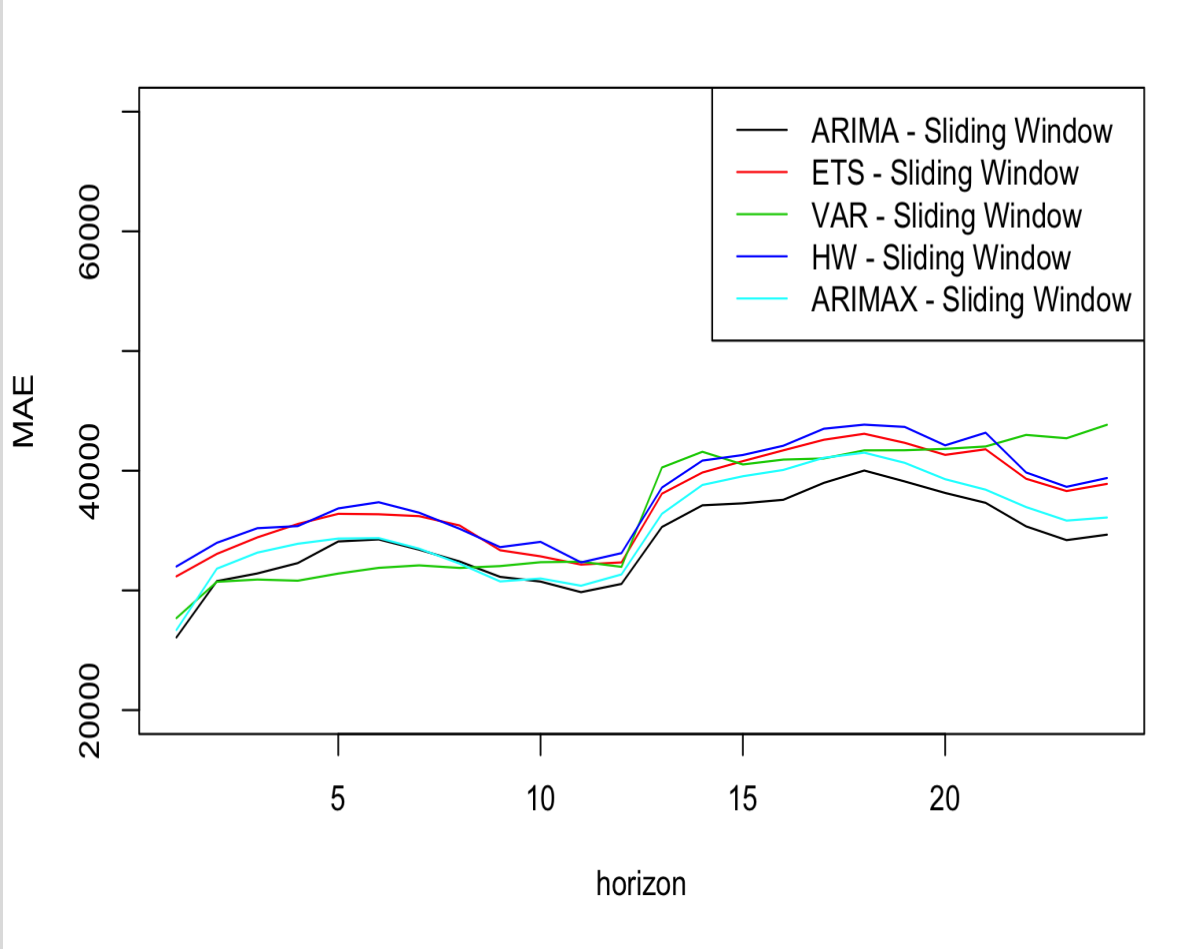
Forecast Models

- ARIMA
- VAR
- Holt-Winters
- ETS
- ARIMAX

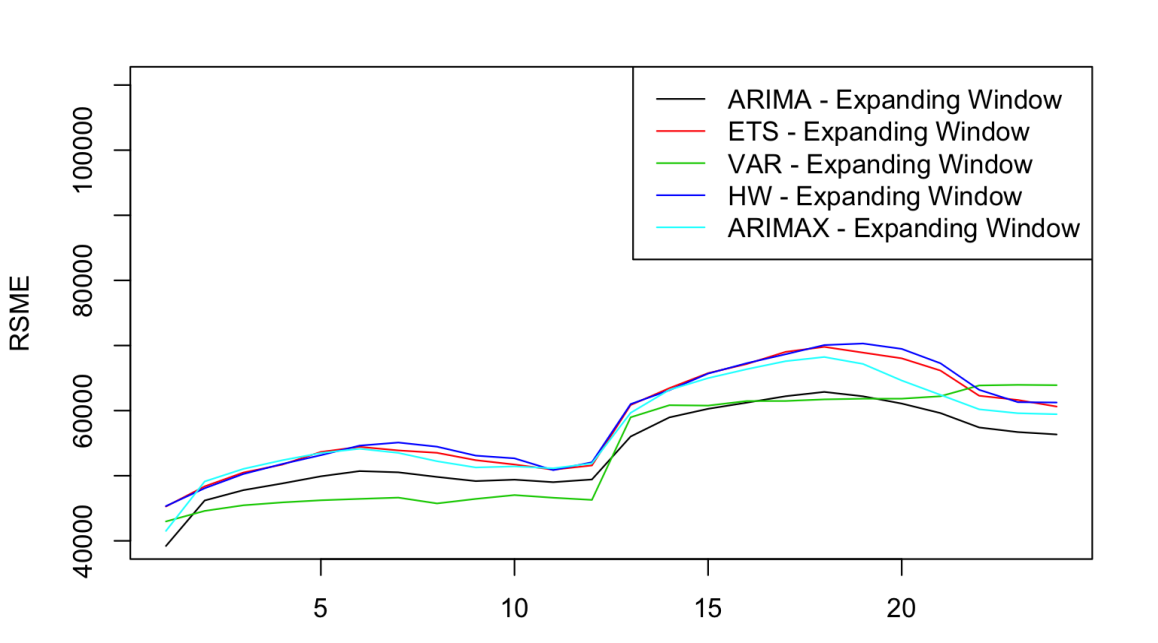
- Test Expanding & Sliding Window
  - Train (minimum size – 360)
  - Test (120 to 24)
  - MAE and RSME per month

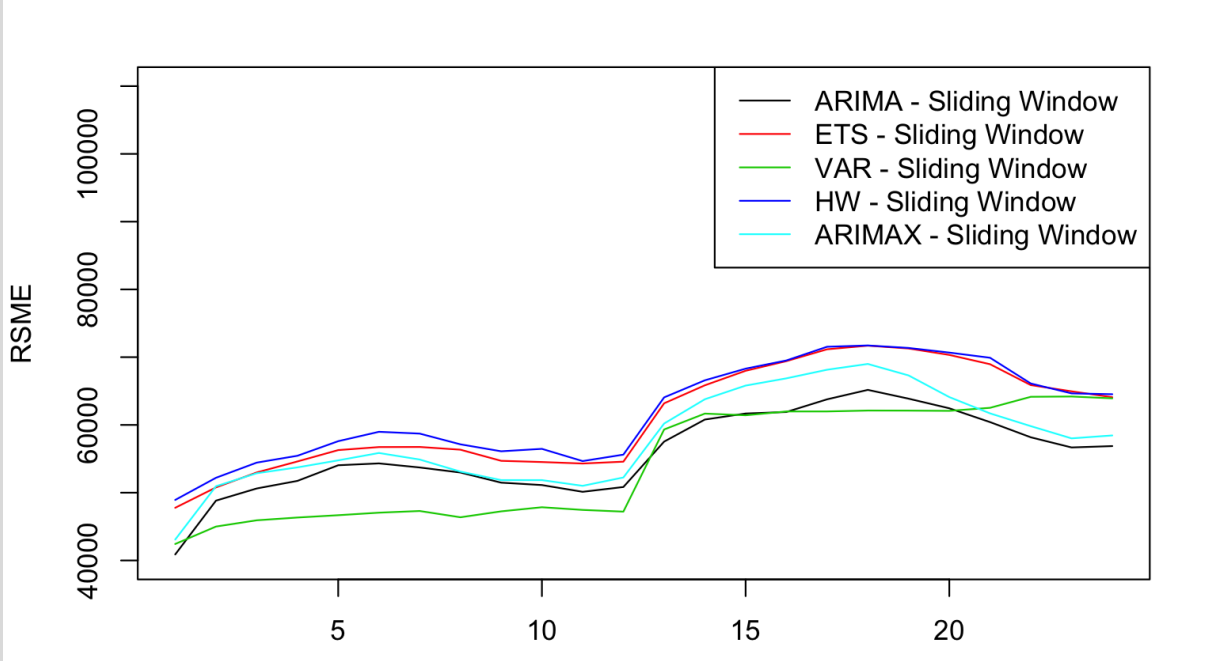
Forecast (Mean Absolute Error)





Forecast (RSME)





MAE Comparison

	Expanding Window	Sliding Window	Both
ARIMA	4342.84795	4435.95802	8778.80596
ARIMAX	4469.28143	4504.58346	8973.8649
ETS	4510.42948	4635.10539	9145.53487
VAR	4570.16483	4576.23464	9146.39947
Holt-Winters	4525.62987	4674.03473	9199.6646

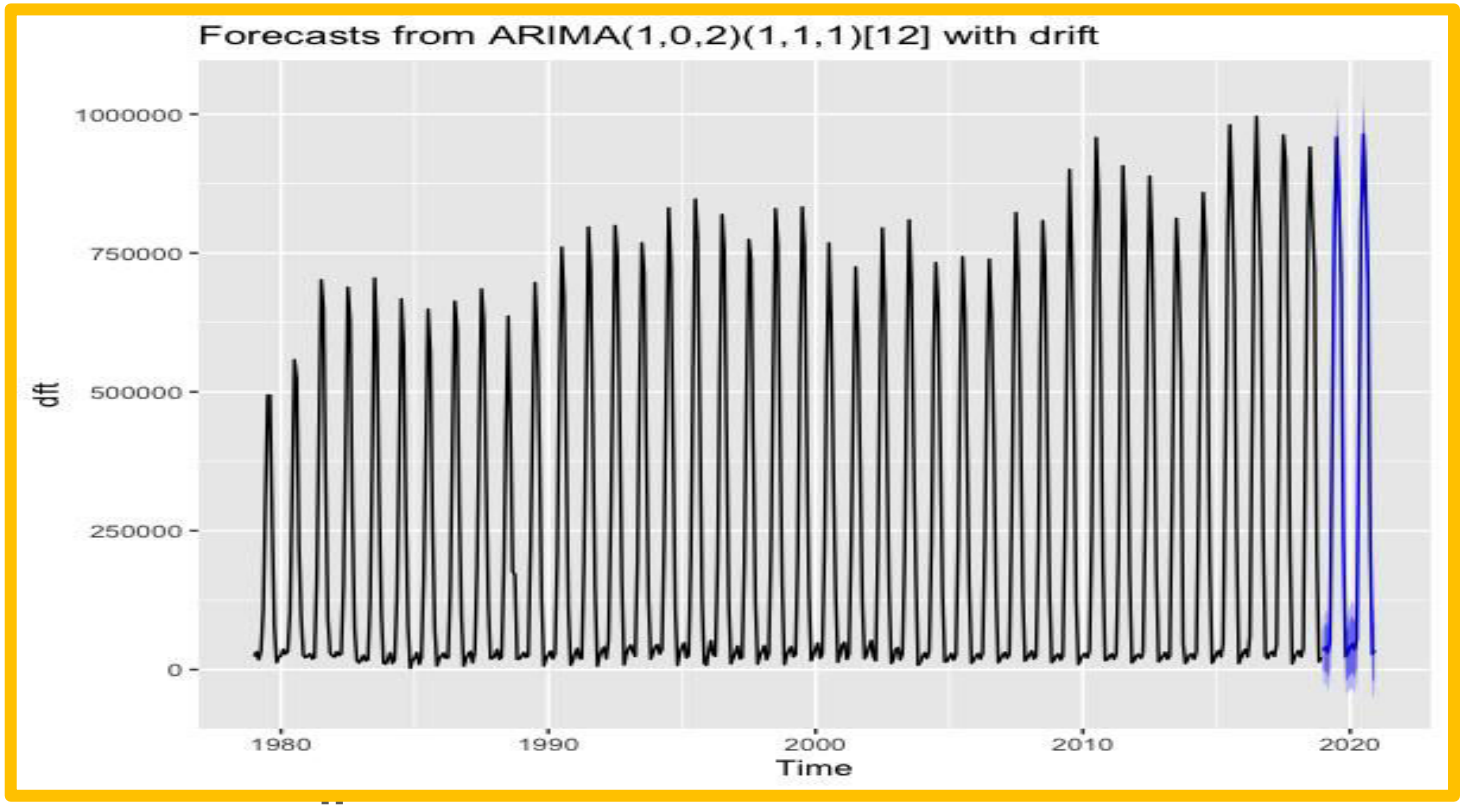
- Both Regression Models Performed best other than VAR

RSME Comparison

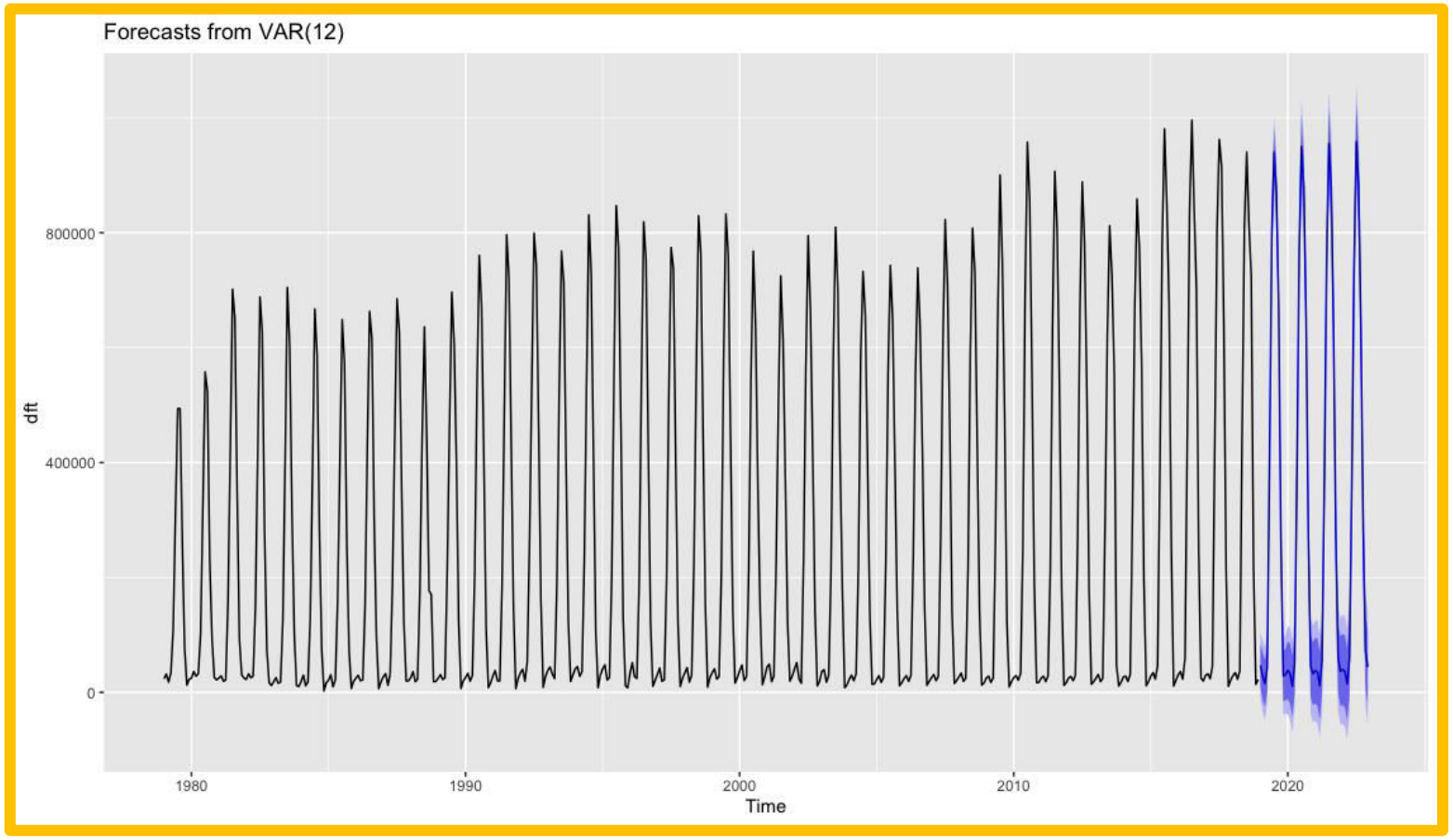
	Expanding Window	Sliding Window	Both
VAR	1293215.87	1304336.85	2597552.72
ARIMA	1294902.17	1340049.51	2634951.68
ARIMAX	1376646.47	1389443.64	2766090.11
HW	1410123.06	1485278.36	2895401.42
ETS	1401616.04	1465148.72	2866764.76

- VAR actually outperforms when using RSME
- Regardless of type, Exponential Smoothing and Regression models have very close RSME Values

Best Model (ARIMA) (MAE)



Best Model (VAR) (RSME)



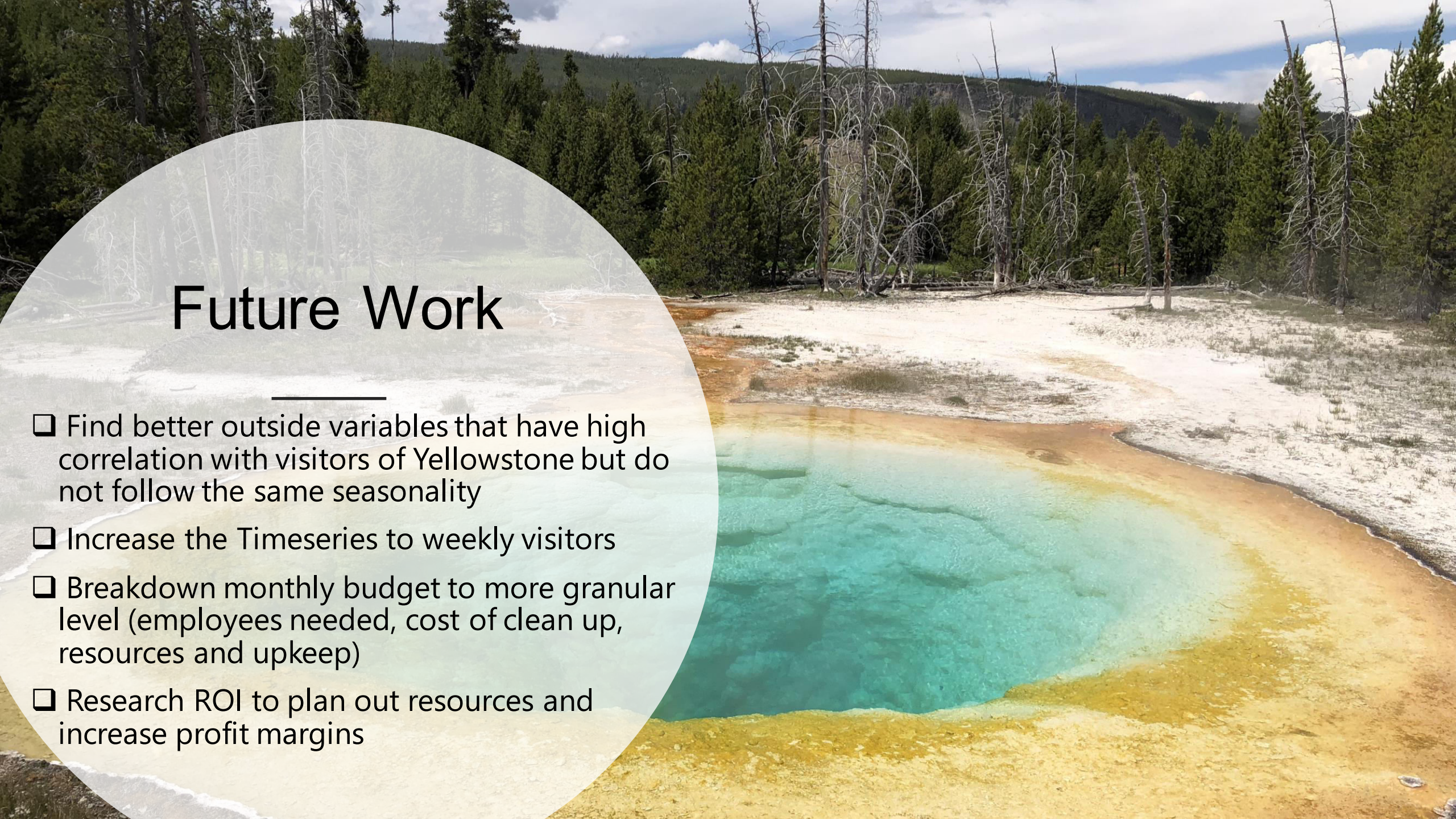
Proposed Budget

- This is very rough estimate based on yearly budget and number of visitors
- Number of visitors and budget peaks between June-August. 80% of Budget is spend between June-September
- Seasonality allows us to plan # of employees, resources and overhead

Month	Visitors	Budget ($mill)	Month	Visitors	Budget($mill)
Jan 2019	34408.49	0.28739137	Jan 2020	40720.13	0.34498214
Feb 2019	40725.29	0.34015142	Feb 2020	46209.2	0.3914857
Mar 2019	30710.81	0.25650709	Mar 2020	35797.32	0.30327595
Apr 2019	50359.070	0.42061602	Apr 2020	55339.6	0.46883872
May 2019	436709.62	3.64754674	May 2020	441041.26	3.73651456
June 2019	810820.06	6.77224391	June 2020	815338.31	6.9075702
July 2019	958576.73	8.00635769	July 2020	963498.2	8.16278515
Aug 2019	849776.41	7.09762055	Aug 2020	855107.73	7.24449789
Sep 2019	698843.1	5.8369744	Sep 2020	702707.05	5.95335484
Oct 2019	226523	1.89199686	Oct 2020	231193.66	1.9586795
Nov 2019	22787.46	0.19032859	Nov 2020	27468.17	0.23271115
Dec 2019	30202.96	0.25226536	Dec 2020	34856.31	0.29530369
TOTAL	4190443	~ 35 million	TOTAL	4249277	~ 36 million

Main Findings

- Yellowstone Park Visitors timeseries follows a strong seasonality making our forecasts relatively accurate
- Variables such as weather and gasolines (pulled from third party sources) did not perform better when testing against the train/test
- Error was greater in between the year where the number of visitors is significantly higher than the winter months
- ARIMA with seasonality performed the best making it easier for us since we did not need to manually difference our data
- Our predictions may not be perfect in the end but will give us a useful framework for yearly budgeting and preparation



Future Work

- ❑ Find better outside variables that have high correlation with visitors of Yellowstone but do not follow the same seasonality
- ❑ Increase the Timeseries to weekly visitors
- ❑ Breakdown monthly budget to more granular level (employees needed, cost of clean up, resources and upkeep)
- ❑ Research ROI to plan out resources and increase profit margins

Q&A

- Contribution

  Claire: EDA + ARIMAX

  Daniela: HW + ETS

  Sneha: ARIMA + sARIMA

  Tim: VAR +Test expanding & sliding