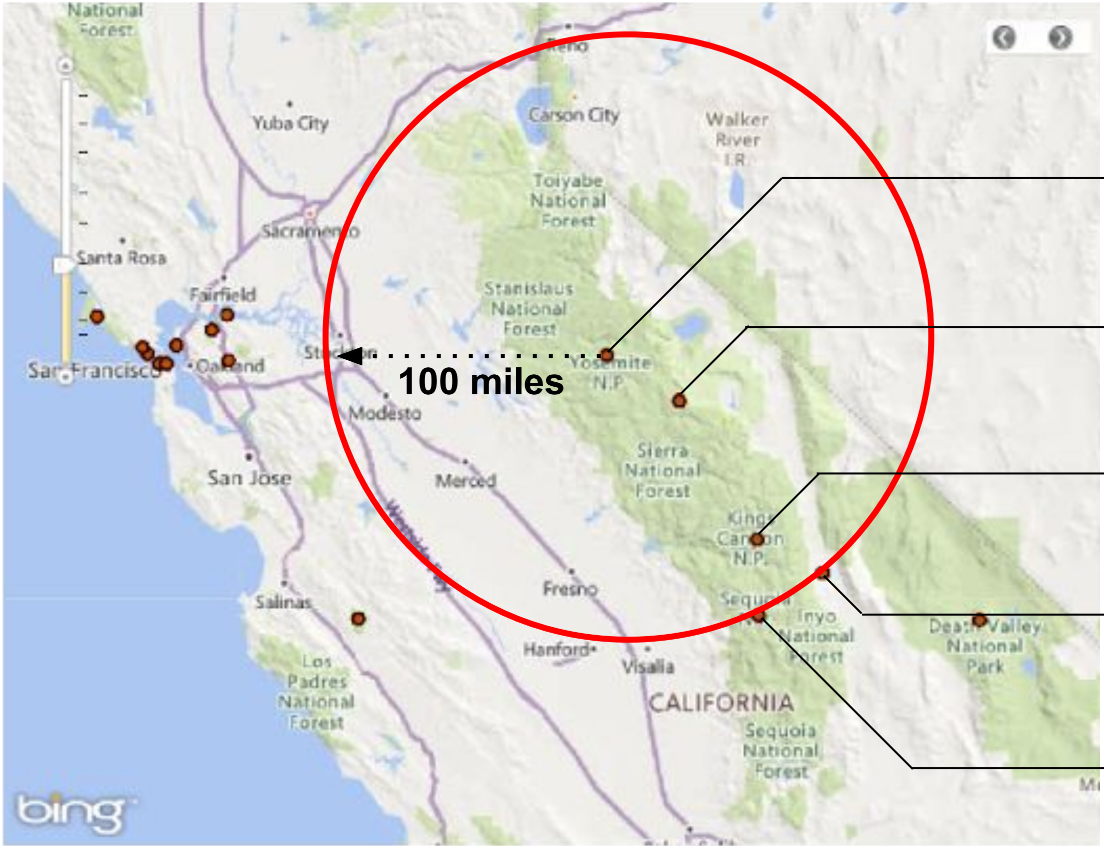
Is Yosemite National Park Stealing Tourists from Surrounding Parks?

Casey Ellis, Meenu Ravi, Kentaro Hama

Yosemite NP and its surrounding parks



Yosemite NP

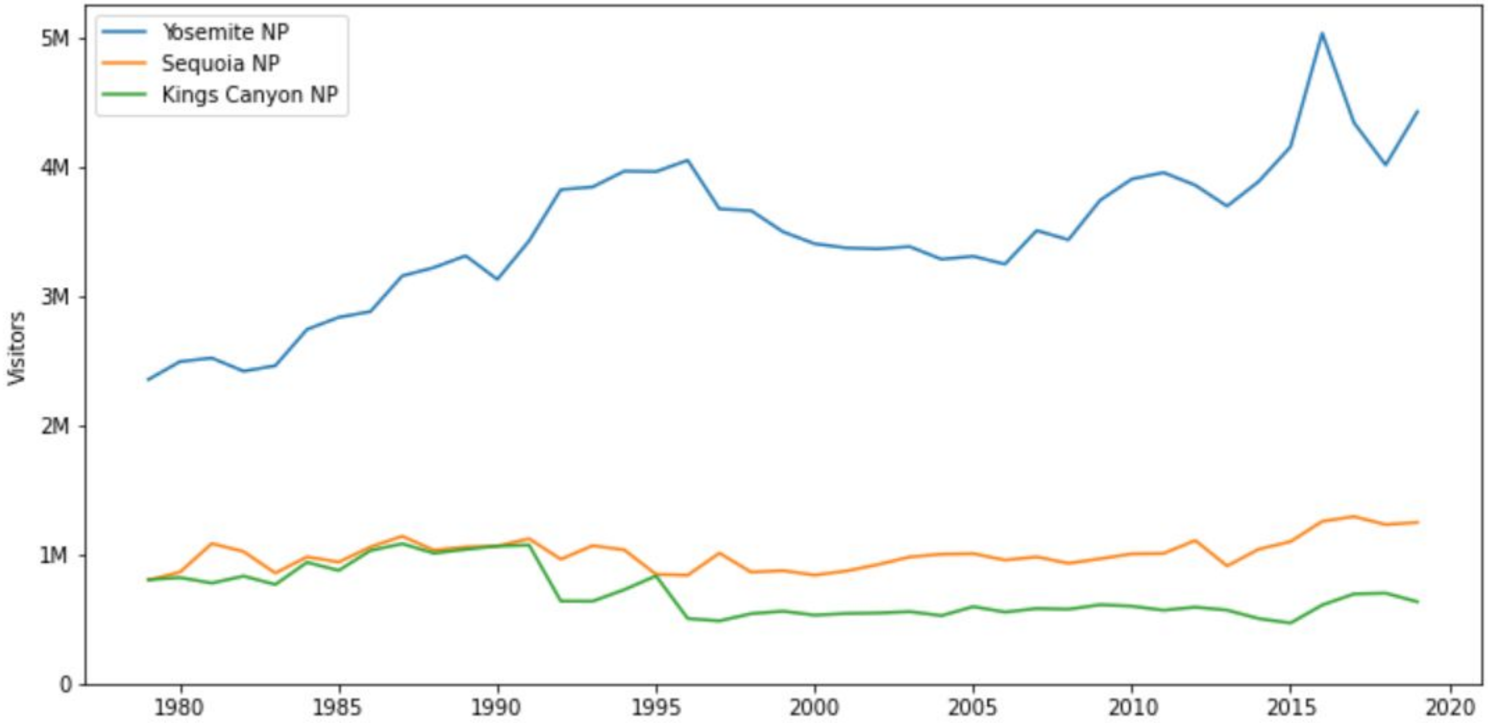
Devils Postpile National Monument (Not included in analysis)

Kings Canyon NP

Manzanar National Historic Site (Not included in analysis)

Sequoia NP

Annual number of visitors by park



Research Question

Suppose you plan to go to California for your next vacation.

- Would you visit not only Yosemite but also Sequoia and Kings Canyon?
- Or, do you think that if you can enjoy the nature in Yosemite, you don't need to visit the surrounding parks?

In other words

- If the number of visitors to Yosemite increases, will the number of visitors to the surrounding parks also increase, or will the number of visitors to the surrounding parks decrease as Yosemite takes away the tourists?
- In addition, is the relationship between Sequoia and Yosemite different from the relationship between Kings Canyon and Yosemite?

Why does it matter?

Suppose the hiking trails in Yosemite are renovated, and the number of visitors to Yosemite is expected to increase.

How should the surrounding parks react to this?

- If the number of visitors to the surrounding parks is also expected to increase, the surrounding parks should increase the number of staff.
- If visitors to the surrounding parks are expected to decrease, they might want to increase their advertising.

The best strategy for the surrounding park depends on its relationship with Yosemite.

Methodology

Models:

- SARIMAX
- Auto SARIMAX
- VARMAX

Endogenous variables:

- Number of visitors to Sequoia NP
- Number of visitors to Kings Canyon NP

Exogenous variables:

- <u>Number of visitors to Yosemite NP</u>
- Number of visitors to the other surrounding park (When Sequoia is the dependent variable, Kings Canyon is used as the explanatory variable, and vice versa.)
- Gasoline price
- Temperature

⇒ Select the best model based on the performance on the test set and interpret the results of the best model

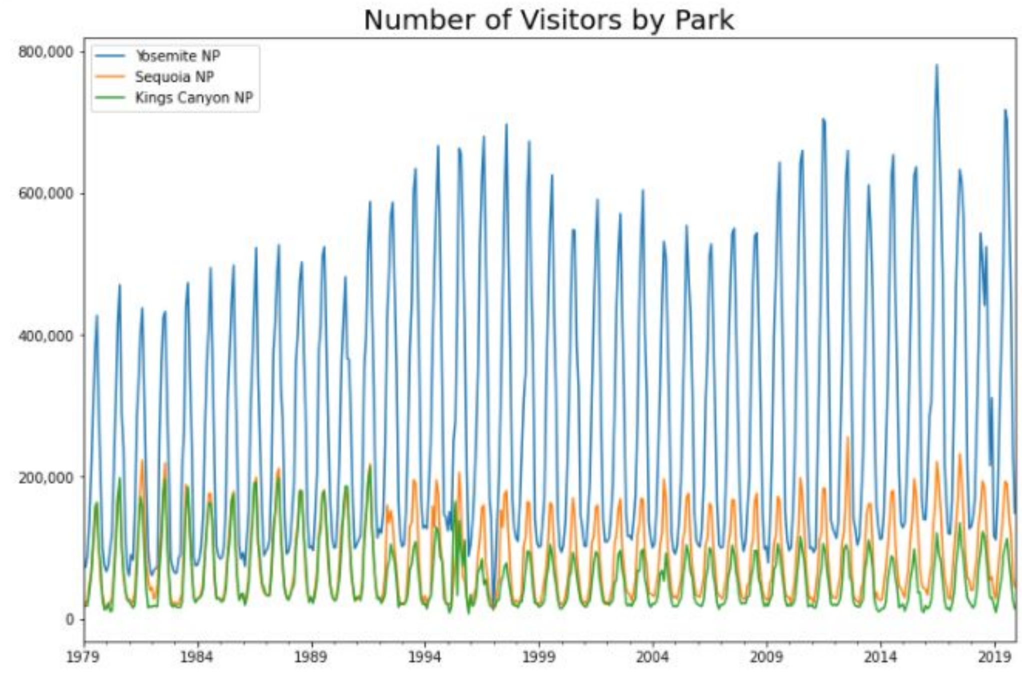
Data Description

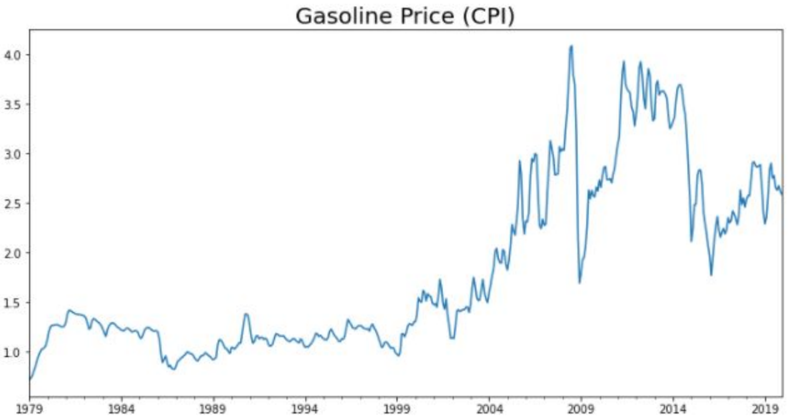
- Number of visitors by park and by month from Jan 1979 to Dec 2019 (Dropped the data for 2020 as an outlier)
  Source: Visitor Use Statistics, the U.S. National Park Service

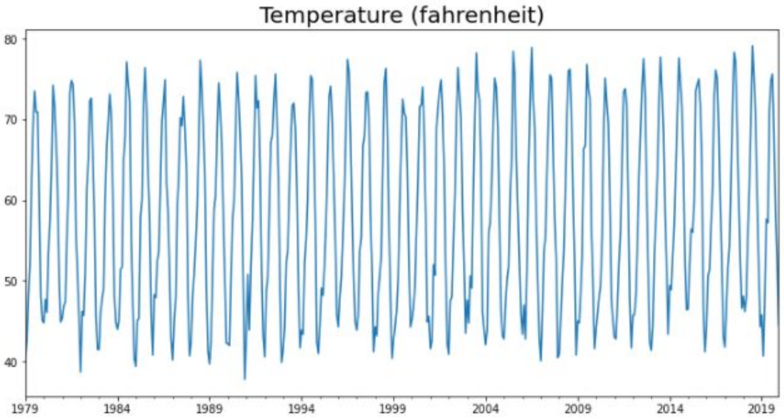
- Monthly gasoline prices (CPI Average Price Data, U.S. city average)
  Source: U.S. Bureau of Labor Statistics

- Monthly Temperature Data (°F)
  Source: National Oceanographic and Atmospheric Administration

Original series



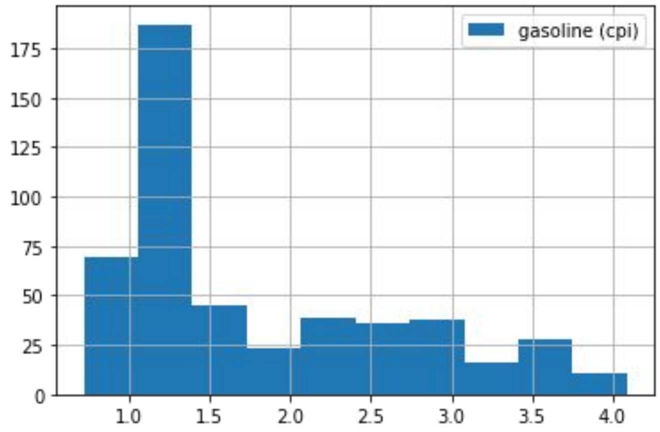


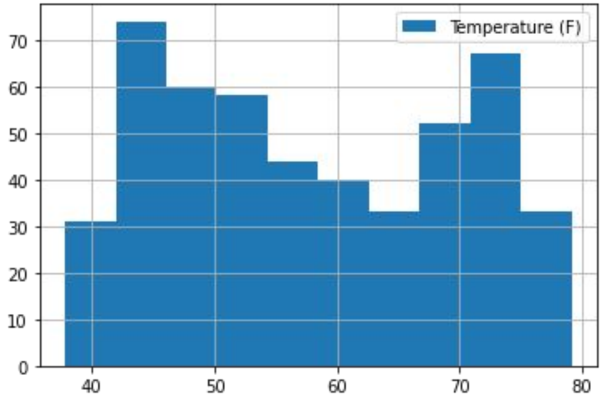


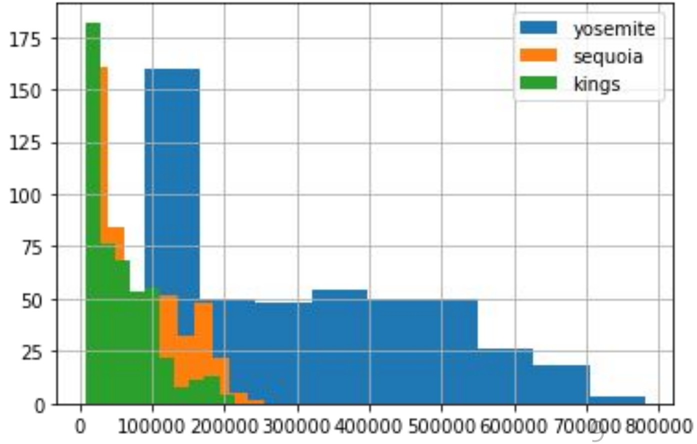
EDA

Columns	Min	Max	Mean
Yosemite	12520	780728	28981
Sequoia	14085	255984	83577
Kings *	7195	213332	58032
Gasoline (CPI)	.716	4.09	1.8
Temperature (F)	37.8	79.1	57.83

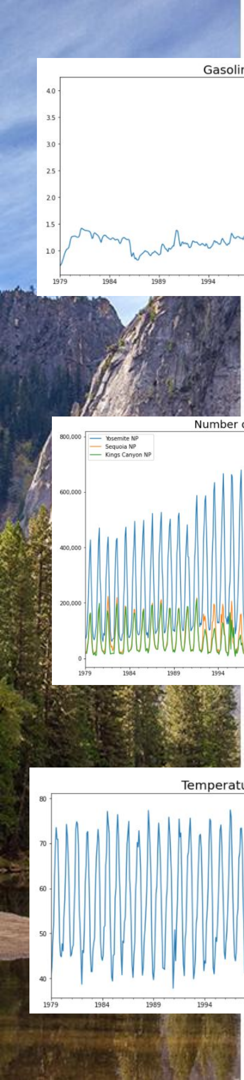
* Outliers Present

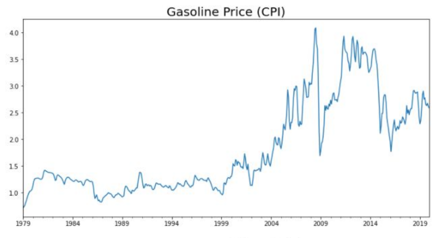


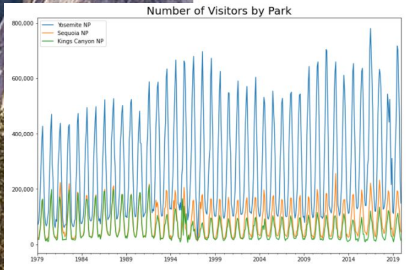


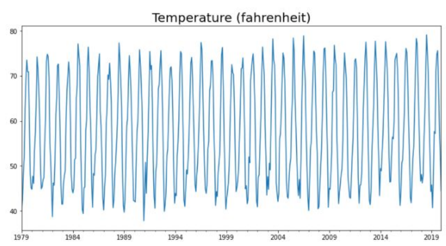


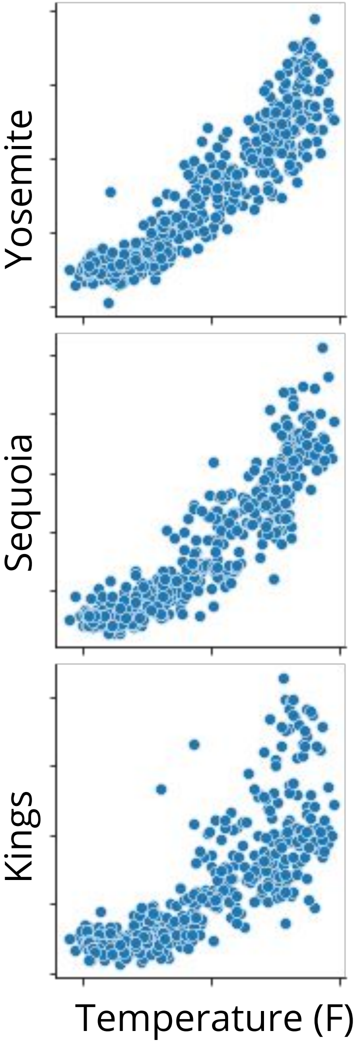
EDA Extended

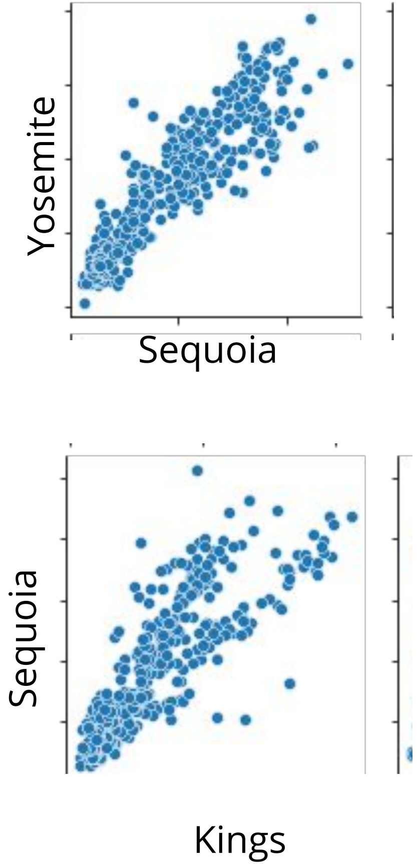


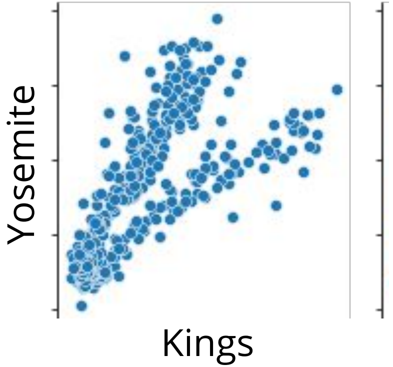




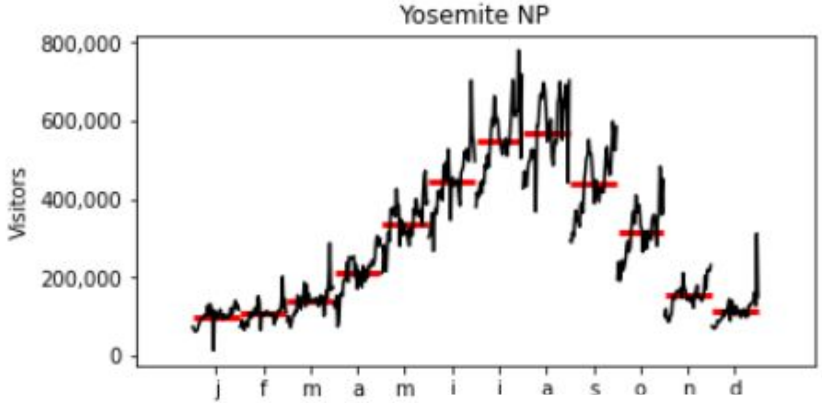


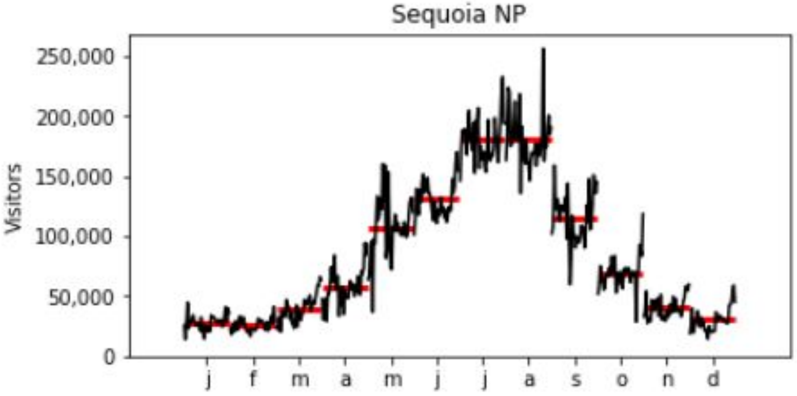


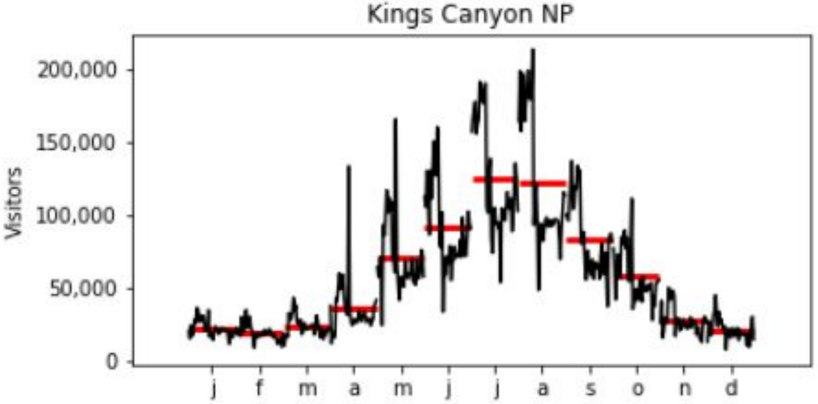




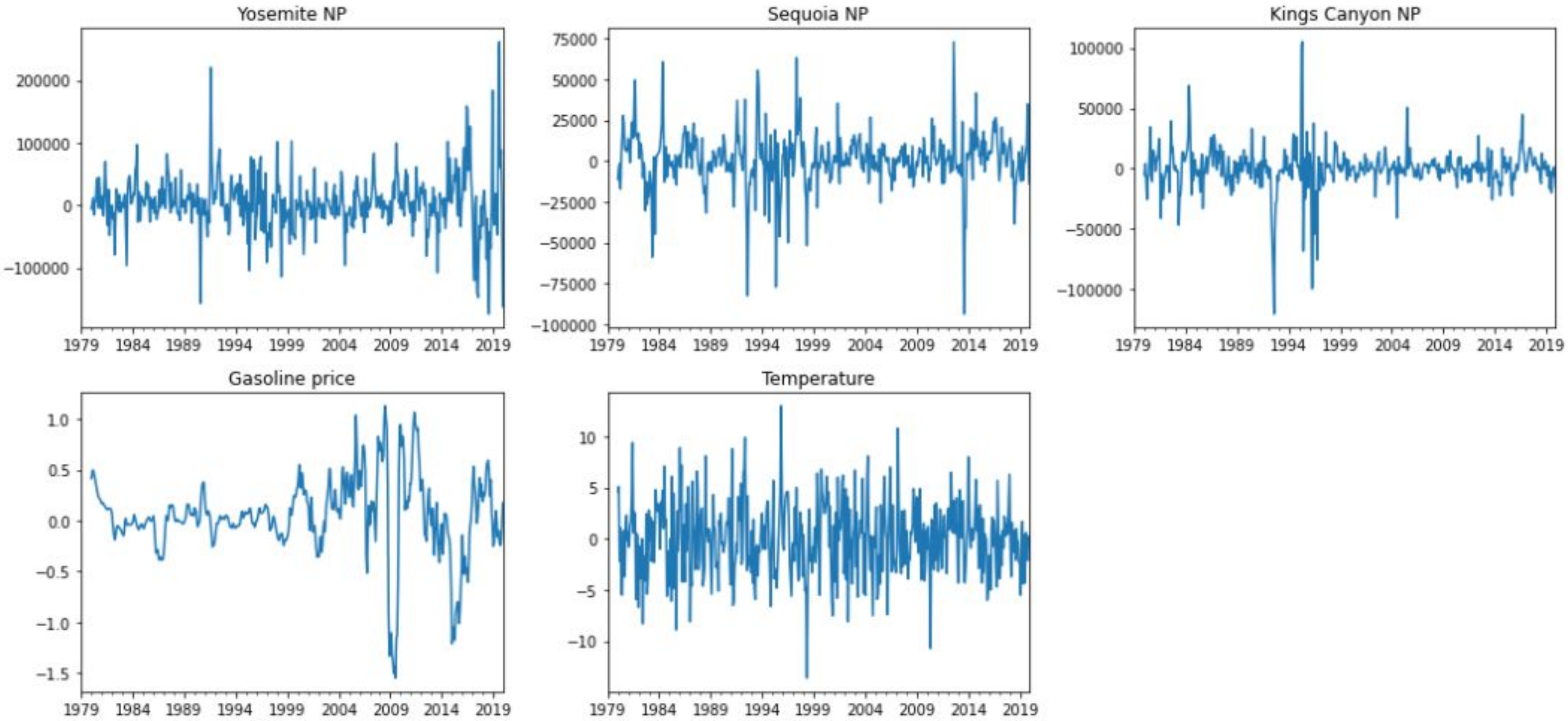
Seasonal subseries plots





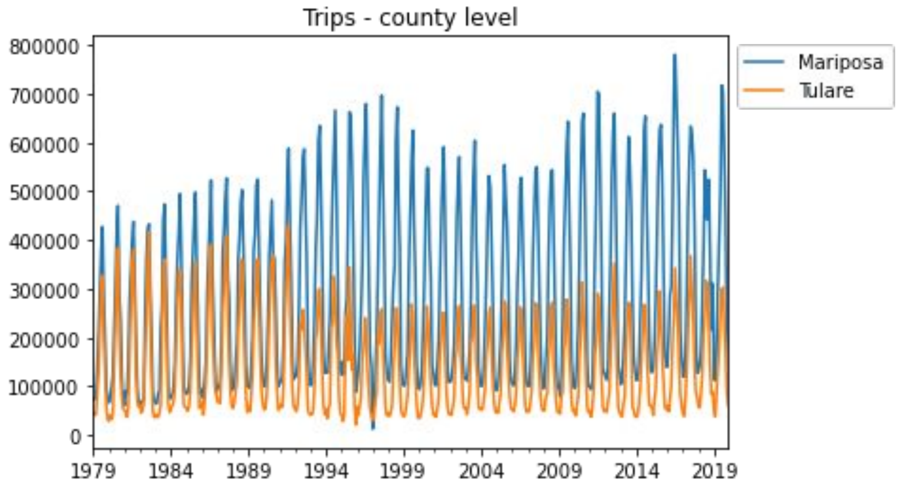


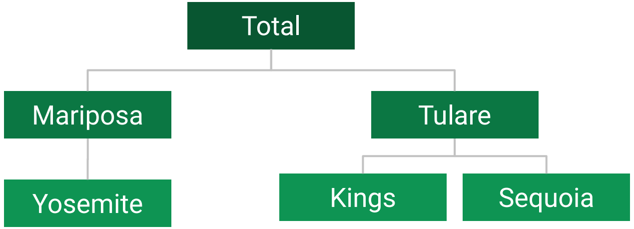
EDA: Stationarity check



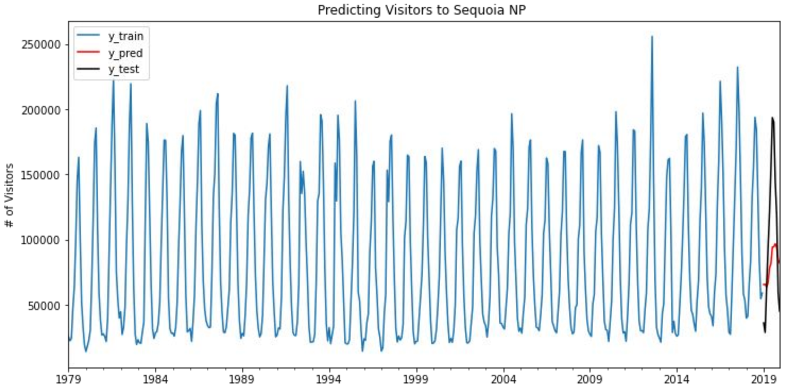
Both the ADF and KPSS test results indicate that all seasonally adjusted series are stationary.

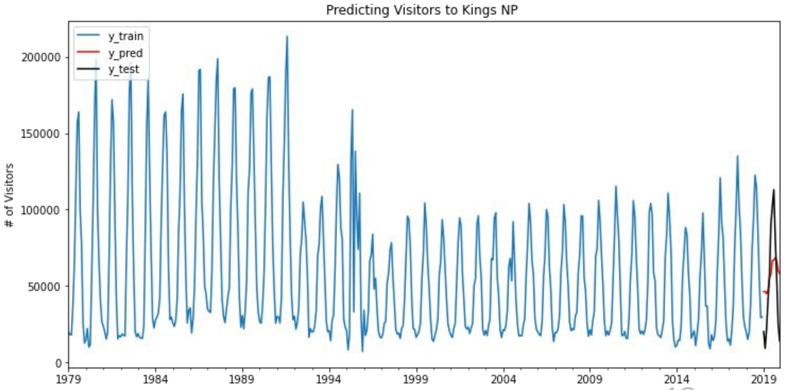
Hierarchical Time Series





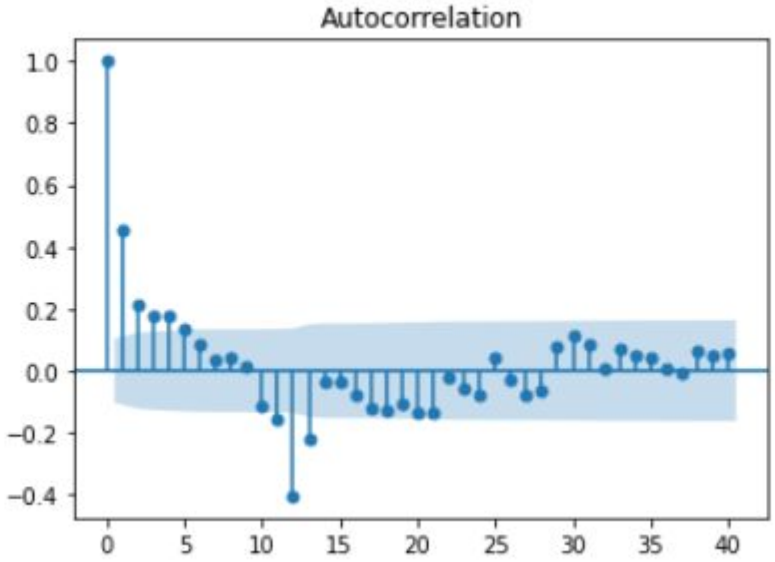
	SMAPE	MASE
Sequoia	0.47	1.69
Kings	0.56	1.29

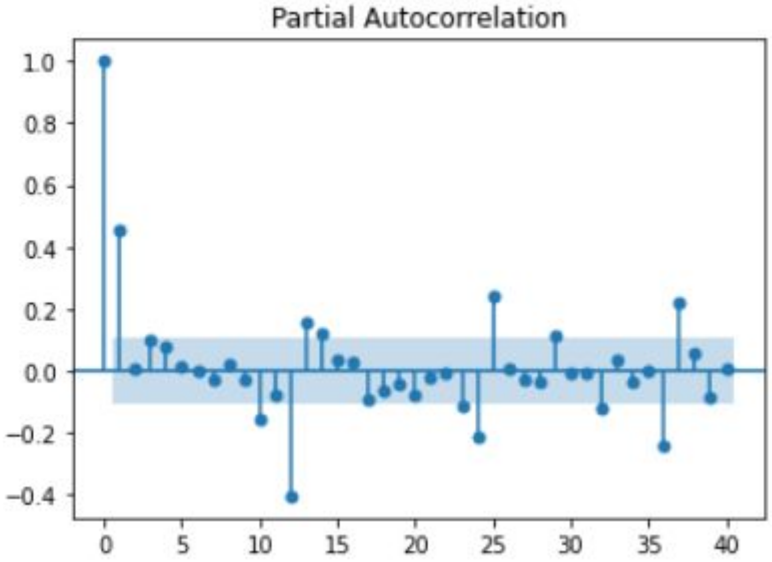




SARIMAX Model: order selection

Sequoia NP





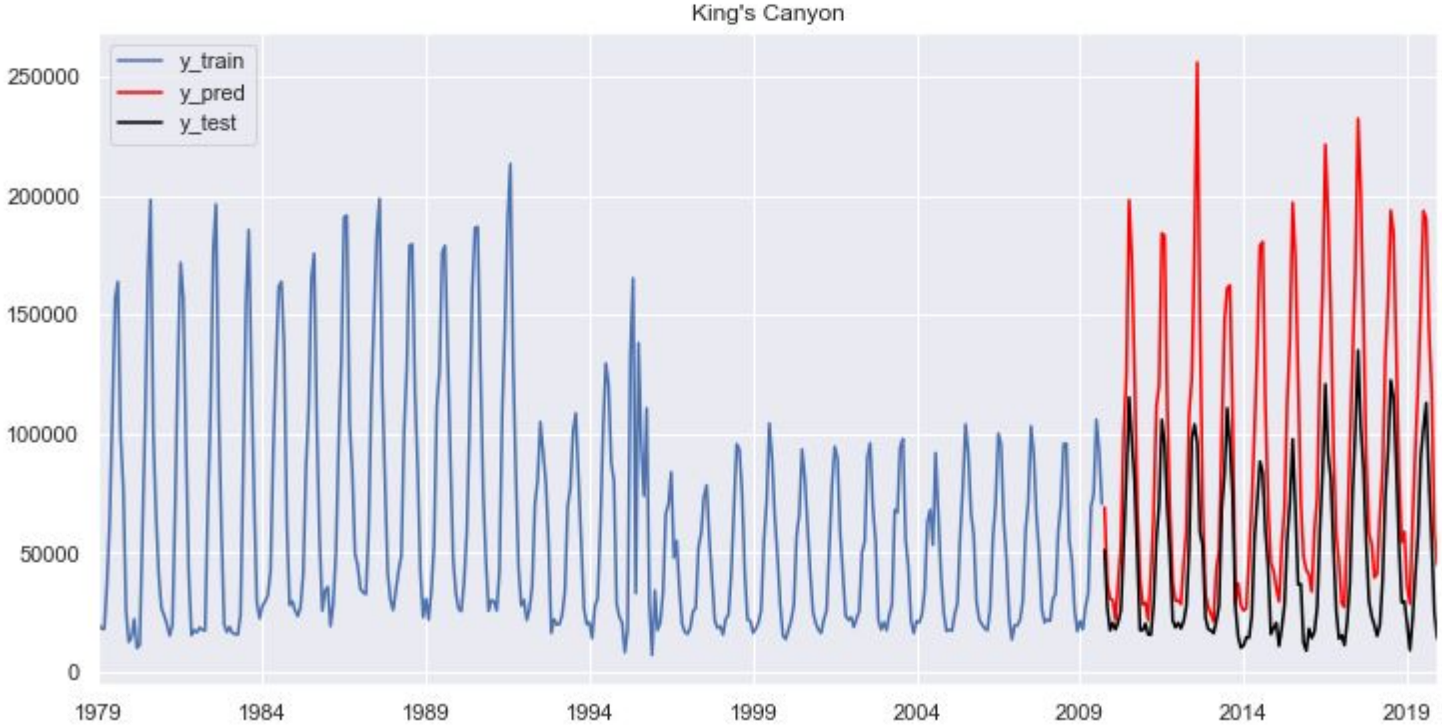
⇒ SARIMAX(0, 0, 4) (0, 1, 1) 12

Considering Exogenous Variables

⇒ including gasoline cost and temperature

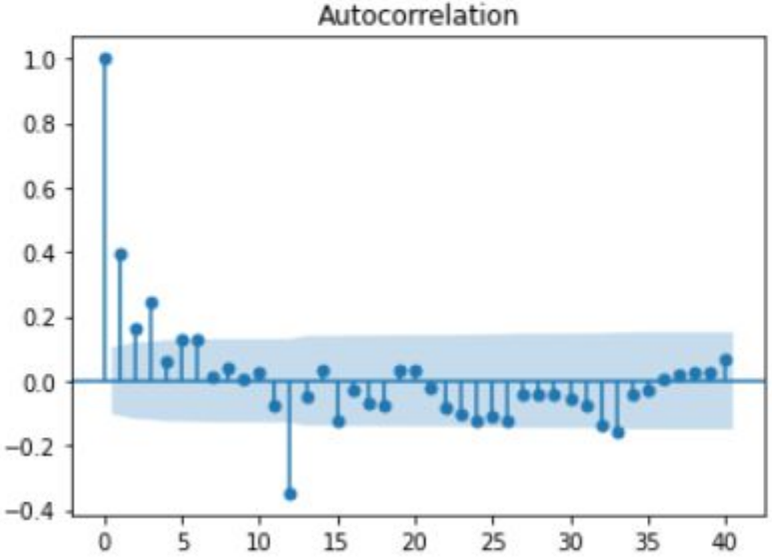
```
                                      SARIMAX Results
==========================================================================================
Dep. Variable:                              sequoia   No. Observations:              369
Model:             SARIMAX(0, 0, 3)x(0, 1, [1], 12)   Log Likelihood             359.288
Date:                              Tue, 25 May 2021   AIC                       -698.577
Time:                                      17:55:32   BIC                       -659.799
Sample:                                  01-01-1979   HQIC                      -683.153
                                       - 09-01-2009
Covariance Type:                                opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
yosemite     1.39e-16   4.74e-07   2.93e-10      1.000   -9.29e-07    9.29e-07
sequoia        1.0000   8.26e-07   1.21e+06      0.000       1.000       1.000
kings      -6.668e-17   1.28e-06   -5.2e-11      1.000   -2.51e-06    2.51e-06
gasoline   -1.648e-12    3.2e-10     -0.005      0.996   -6.29e-10    6.26e-10
temp        1.528e-13   6.69e-09   2.28e-05      1.000   -1.31e-08    1.31e-08
ma.L1          0.3804   2.27e-06   1.68e+05      0.000       0.380       0.380
ma.L2          0.1921   1.73e-06   1.11e+05      0.000       0.192       0.192
ma.L3          0.1551   2.76e-07   5.62e+05      0.000       0.155       0.155
ma.S.L12      -0.3006   3.62e-06  -8.29e+04      0.000      -0.301      -0.301
sigma2          1e-10   1.07e-10      0.932      0.352    -1.1e-10     3.1e-10
===================================================================================
Ljung-Box (L1) (Q):                   0.04   Jarque-Bera (JB):            345324.66
Prob(Q):                              0.84   Prob(JB):                         0.00
Heteroskedasticity (H):               0.00   Skew:                            11.79
Prob(H) (two-sided):                  0.00   Kurtosis:                       153.53
===================================================================================
```

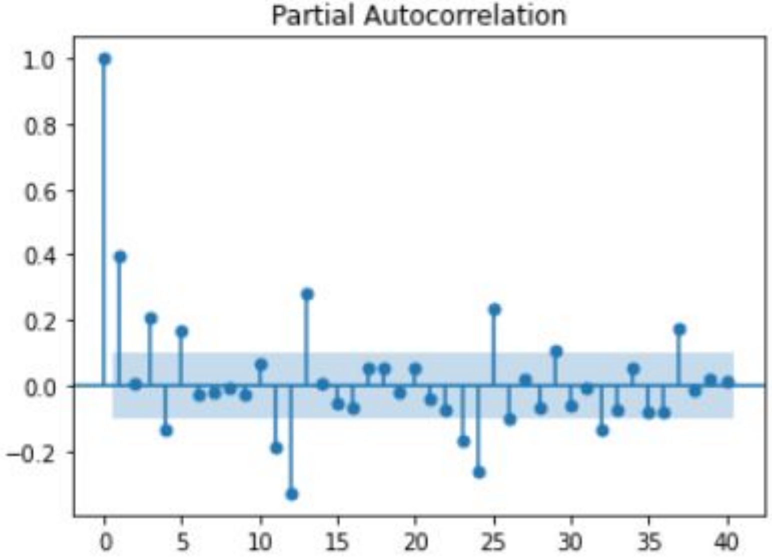
Considering Exogenous Variables



SARIMAX Model: order selection

Kings Canyon NP





⇒ SARIMAX(0, 0, 3) (0, 1, 1) 12

SARIMAX Model (Sequoia NP)

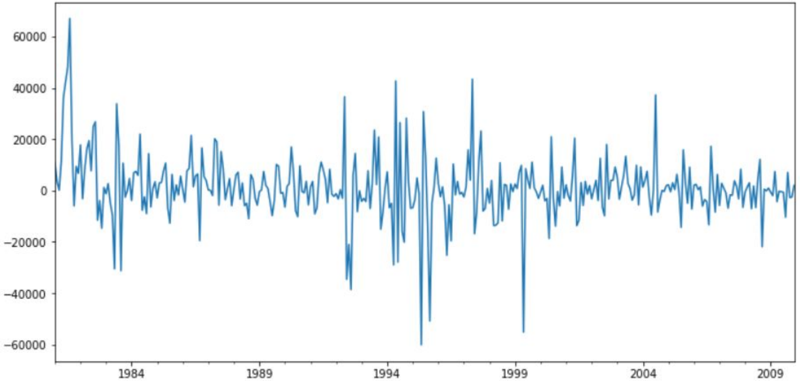
SARIMAX Results

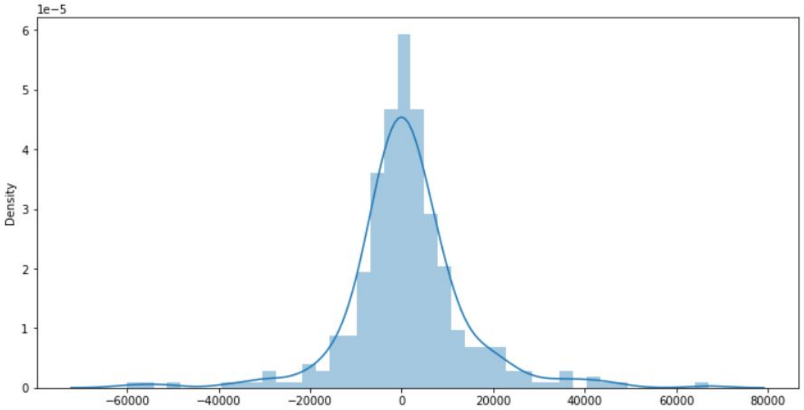
Dep. Variable:	sequoia	No. Observations:	360
Model:	SARIMAX(0, 0, 4)x(0, 1, [1], 12)	Log Likelihood	-3791.574
Date:	Wed, 19 May 2021	AIC	7603.147
Time:	15:22:17	BIC	7641.669
Sample:	01-01-1980	HQIC	7618.484
	- 12-01-2009		
Covariance Type:	opg		

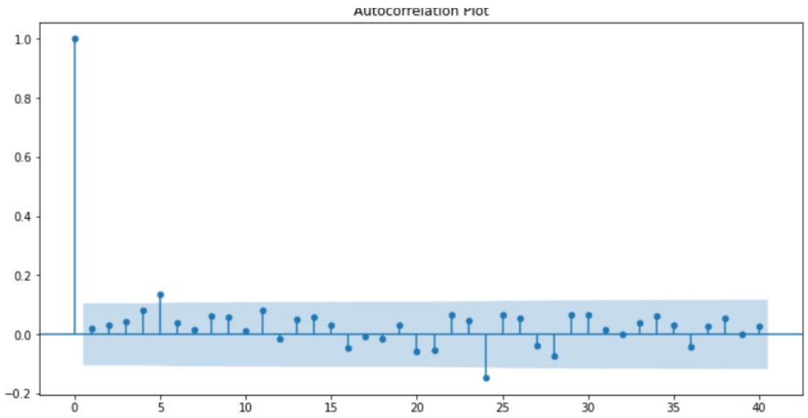
	coef	std err	z	P>|z|	[0.025	0.975]
yosemite_seasdiff	0.0688	0.016	4.209	0.000	0.037	0.101
kings_seasdiff	0.0382	0.024	1.611	0.107	-0.008	0.085
gasoline_seasdiff	-806.7442	6498.161	-0.124	0.901	-1.35e+04	1.19e+04
temp_seasdiff	16.0816	154.078	0.104	0.917	-285.905	318.069
ma.L1	0.6195	0.040	15.472	0.000	0.541	0.698
ma.L2	0.3232	0.043	7.434	0.000	0.238	0.408
ma.L3	0.2413	0.052	4.656	0.000	0.140	0.343
ma.L4	0.1840	0.062	2.987	0.003	0.063	0.305
ma.S.L12	-0.5375	0.036	-14.940	0.000	-0.608	-0.467
sigma2	1.898e+08	0.843	2.25e+08	0.000	1.9e+08	1.9e+08

Ljung-Box (L1) (Q):	0.04	Jarque-Bera (JB):	431.11
Prob(Q):	0.84	Prob(JB):	0.00
Heteroskedasticity (H):	0.34	Skew:	0.06
Prob(H) (two-sided):	0.00	Kurtosis:	8.45

Residual diagnosis (SARIMAX Sequoia NP)





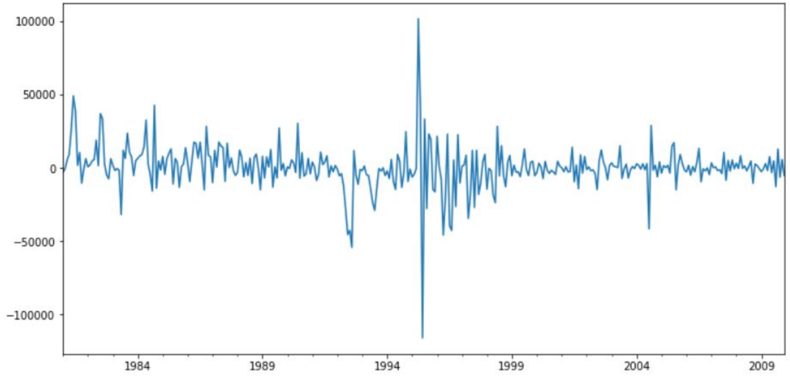


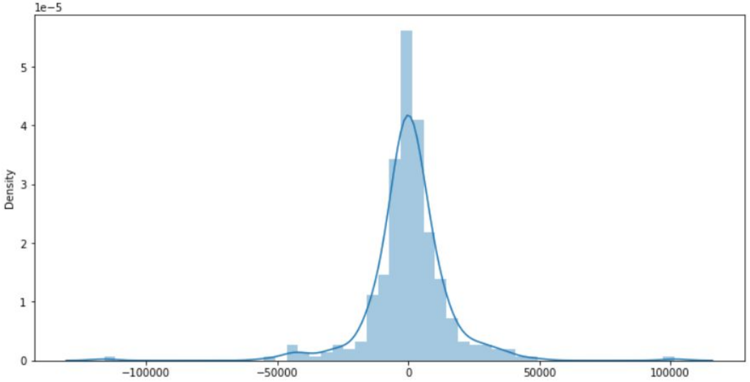
	lb_stat	lb_pvalue	bp_stat	bp_pvalue
12	16.067196	0.188178	15.689377	0.205881

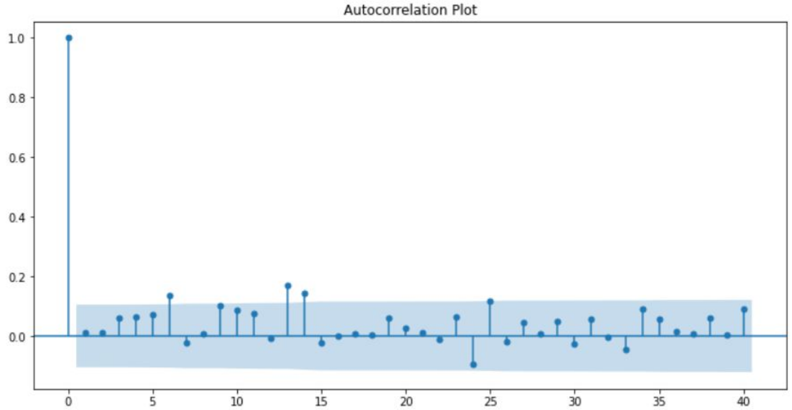
SARIMAX Model (Kings Canyon NP)

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                 kings   No. Observations:               360
Model:             SARIMAX(0, 0, 3)x(0, 1, [1], 12)   Log Likelihood            -3855.192
Date:                               Wed, 19 May 2021   AIC                        7728.385
Time:                                       15:22:23   BIC                        7763.054
Sample:                                   01-01-1980   HQIC                       7742.187
                                        - 12-01-2009
Covariance Type:                                 opg
=====================================================================================
                        coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------
yosemite_seasdiff     0.0540      0.027      2.004      0.045       0.001       0.107
sequoia_seasdiff     -0.0195      0.039     -0.494      0.621      -0.097       0.058
gasoline_seasdiff -1213.5728   1.19e+04     -0.102      0.919   -2.45e+04    2.21e+04
temp_seasdiff       -20.3418    276.228     -0.074      0.941    -561.738     521.055
ma.L1                 0.6098      0.042     14.411      0.000       0.527       0.693
ma.L2                 0.2958      0.031      9.415      0.000       0.234       0.357
ma.L3                 0.3132      0.048      6.518      0.000       0.219       0.407
ma.S.L12             -0.5742      0.043    -13.352      0.000      -0.659      -0.490
sigma2              3.09e+08      0.614   5.03e+08      0.000    3.09e+08    3.09e+08
===================================================================================
Ljung-Box (L1) (Q):                   0.01   Jarque-Bera (JB):              3275.42
Prob(Q):                              0.90   Prob(JB):                         0.00
Heteroskedasticity (H):               0.31   Skew:                            -0.61
Prob(H) (two-sided):                  0.00   Kurtosis:                        17.98
===================================================================================
```

Residual diagnosis (SARIMAX Kings Canyon)







	lb_stat	lb_pvalue	bp_stat	bp_pvalue
12	19.794268	0.071079	19.272155	0.082169

Auto SARIMAX Model (Sequoia NP)

Best model: SARIMAX(2,0,2)(0,1,2)12

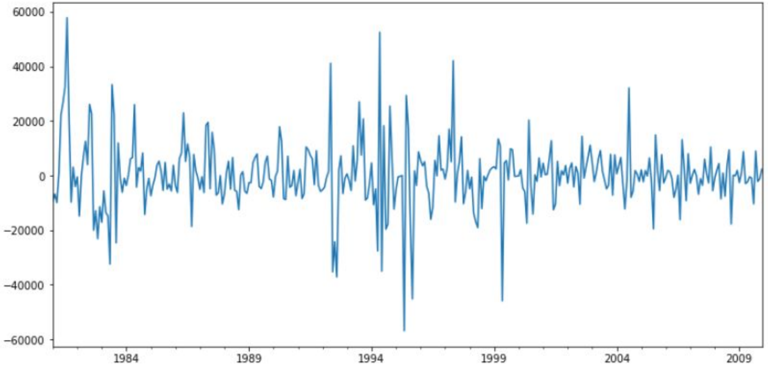
SARIMAX Results

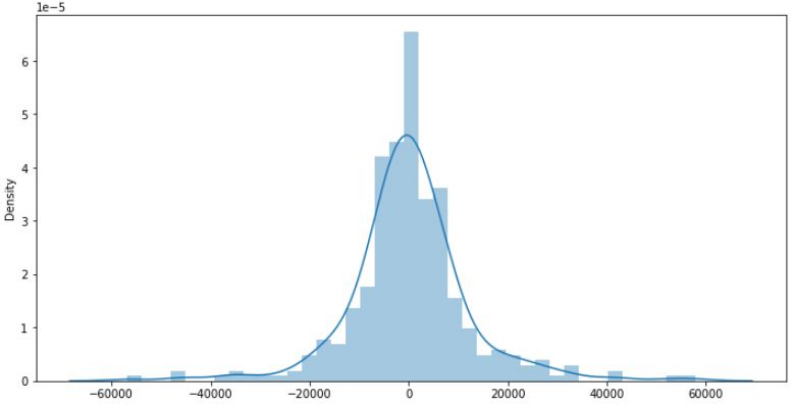
Dep. Variable:	y	No. Observations:	360
Model:	SARIMAX(2, 0, 2)x(0, 1, 2, 12)	Log Likelihood	-3774.757
Date:	Wed, 19 May 2021	AIC	7571.514
Time:	15:27:39	BIC	7613.888
Sample:	01-01-1980	HQIC	7588.384
	- 12-01-2009		
Covariance Type:	opg		

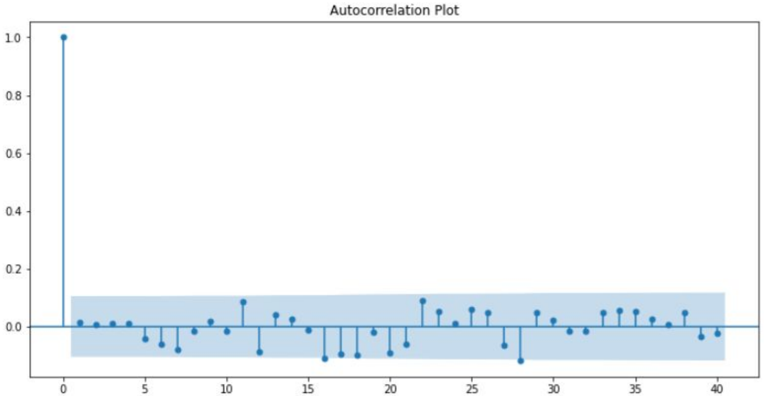
	coef	std err	z	P>|z|	[0.025	0.975]
yosemite_seasdiff	0.0617	0.019	3.225	0.001	0.024	0.099
kings_seasdiff	0.0552	0.024	2.349	0.019	0.009	0.101
gasoline_seasdiff	-806.7450	6672.755	-0.121	0.904	-1.39e+04	1.23e+04
temp_seasdiff	16.1063	171.374	0.094	0.925	-319.781	351.994
ar.L1	1.3747	0.120	11.427	0.000	1.139	1.610
ar.L2	-0.3841	0.116	-3.323	0.001	-0.611	-0.158
ma.L1	-0.8025	0.129	-6.239	0.000	-1.055	-0.550
ma.L2	-0.1123	0.101	-1.108	0.268	-0.311	0.086
ma.S.L12	-0.5157	0.046	-11.237	0.000	-0.606	-0.426
ma.S.L24	-0.2235	0.047	-4.730	0.000	-0.316	-0.131
sigma2	1.87e+08	0.641	2.92e+08	0.000	1.87e+08	1.87e+08

Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	307.69
Prob(Q):	0.99	Prob(JB):	0.00
Heteroskedasticity (H):	0.37	Skew:	0.07
Prob(H) (two-sided):	0.00	Kurtosis:	7.60

Residual diagnosis (Auto SARIMAX Sequoia NP)







	lb_stat	lb_pvalue	bp_stat	bp_pvalue
12	10.208639	0.597663	9.884237	0.626116

Auto SARIMAX Model (Kings Canyon NP)

Best model: SARIMAX(5,0,5)(0,1,2)12

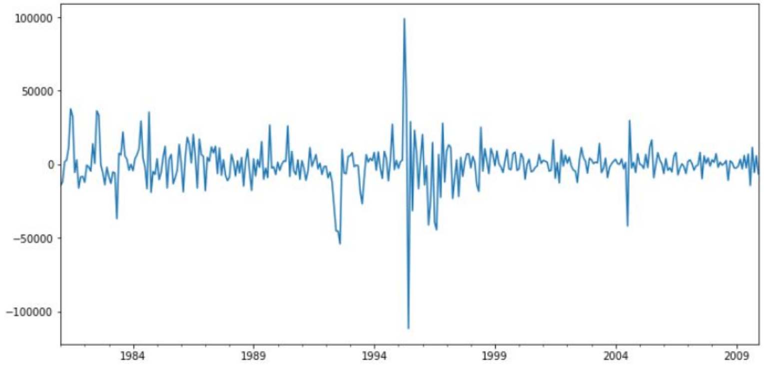
SARIMAX Results

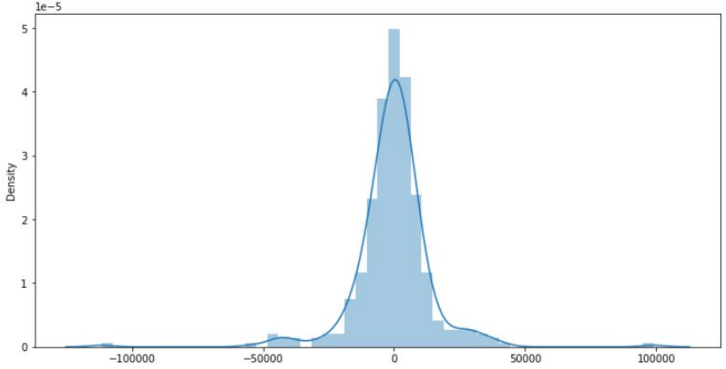
Dep. Variable:	y	No. Observations:	360
Model:	SARIMAX(5, 0, 5)x(0, 1, [1, 2], 12)	Log Likelihood	-3839.782
Date:	Wed, 19 May 2021	AIC	7713.564
Time:	15:33:40	BIC	7779.052
Sample:	01-01-1980	HQIC	7739.636
	- 12-01-2009		
Covariance Type:	opg		

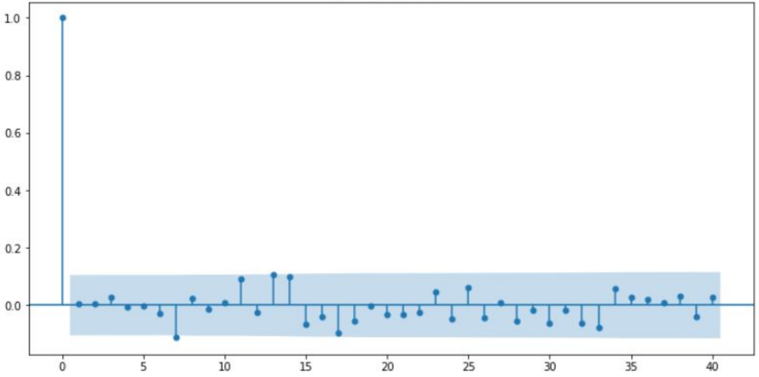
	coef	std err	z	P>|z|	[0.025	0.975]
yosemite_seasdiff	0.0506	0.033	1.511	0.131	-0.015	0.116
sequoia_seasdiff	-0.0056	0.043	-0.132	0.895	-0.090	0.078
gasoline_seasdiff	-1213.5729	1.16e+04	-0.105	0.916	-2.39e+04	2.14e+04
temp_seasdiff	-20.3389	263.429	-0.077	0.938	-536.651	495.973
ar.L1	0.0685	0.433	0.158	0.874	-0.779	0.916
ar.L2	0.1676	0.288	0.582	0.561	-0.397	0.732
ar.L3	0.3334	0.181	1.847	0.065	-0.020	0.687
ar.L4	0.5160	0.235	2.194	0.028	0.055	0.977
ar.L5	-0.1293	0.146	-0.884	0.377	-0.416	0.157
ma.L1	0.5205	0.432	1.206	0.228	-0.326	1.367
ma.L2	0.0607	0.119	0.510	0.610	-0.172	0.294
ma.L3	-0.1642	0.145	-1.135	0.257	-0.448	0.119
ma.L4	-0.8012	0.137	-5.858	0.000	-1.069	-0.533
ma.L5	-0.2785	0.270	-1.032	0.302	-0.807	0.250
ma.S.L12	-0.5600	0.056	-10.015	0.000	-0.670	-0.450
ma.S.L24	-0.1205	0.072	-1.671	0.095	-0.262	0.021
sigma2	2.843e+08	0.532	5.34e+08	0.000	2.84e+08	2.84e+08

Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	3680.20
Prob(Q):	0.99	Prob(JB):	0.00
Heteroskedasticity (H):	0.35	Skew:	-0.65
Prob(H) (two-sided):	0.00	Kurtosis:	18.88

Residual diagnosis (Auto Sarimax Kings Canyon)

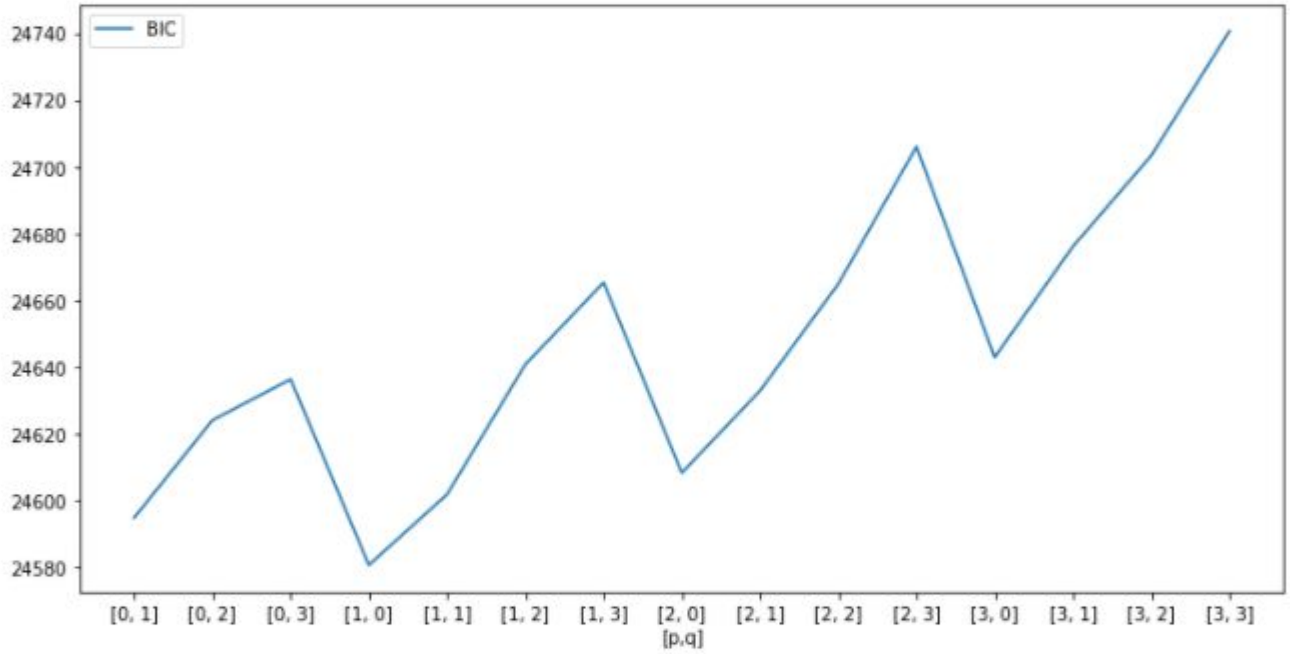






	lb_stat	lb_pvalue	bp_stat	bp_pvalue
12	8.432232	0.750507	8.181838	0.770764

VARMAX Model: order selection



VARMAX Model

Results for equation yosemite_seasdiff

	coef	std err	z	P>|z|	[0.025	0.975]
intercept	2836.2821	1944.339	1.459	0.145	-974.551	6647.116
L1.yosemite_seasdiff	0.3040	0.055	5.515	0.000	0.196	0.412
L1.sequoia_seasdiff	0.1923	0.095	2.022	0.043	0.006	0.379
L1.kings_seasdiff	-0.2359	0.091	-2.594	0.009	-0.414	-0.058
beta.gasoline_seasdiff	-7701.1569	6561.648	-1.174	0.241	-2.06e+04	5159.436
beta.temp_seasdiff	1568.3239	614.988	2.550	0.011	362.970	2773.678

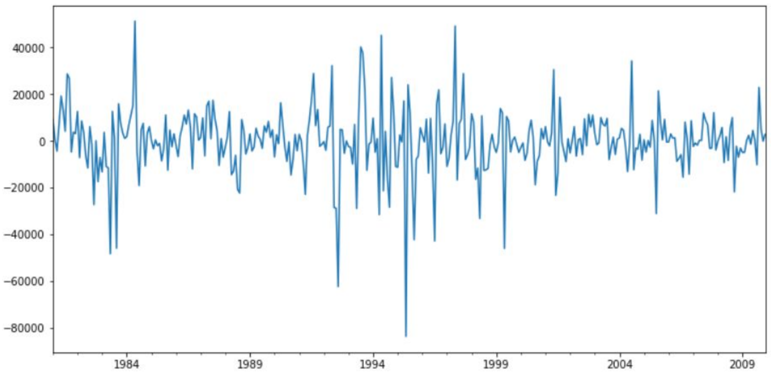
Results for equation sequoia_seasdiff

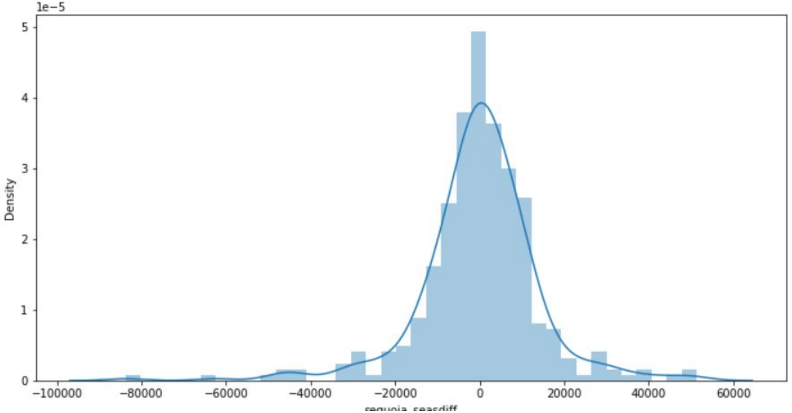
	coef	std err	z	P>|z|	[0.025	0.975]
intercept	562.0929	883.090	0.637	0.524	-1168.732	2292.918
L1.yosemite_seasdiff	-0.0705	0.023	-3.051	0.002	-0.116	-0.025
L1.sequoia_seasdiff	0.4883	0.045	10.901	0.000	0.400	0.576
L1.kings_seasdiff	-0.0146	0.035	-0.418	0.676	-0.083	0.054
beta.gasoline_seasdiff	-1634.6504	3639.765	-0.449	0.653	-8768.458	5499.158
beta.temp_seasdiff	428.4809	209.561	2.045	0.041	17.749	839.212

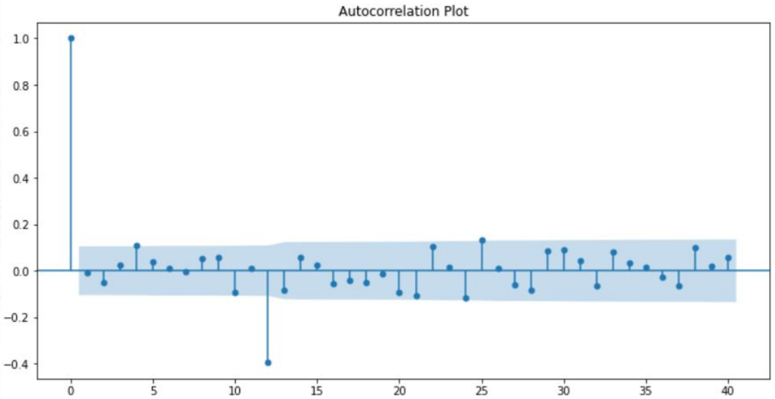
Results for equation kings_seasdiff

	coef	std err	z	P>|z|	[0.025	0.975]
intercept	-161.6678	1202.900	-0.134	0.893	-2519.308	2195.973
L1.yosemite_seasdiff	-0.0574	0.029	-1.981	0.048	-0.114	-0.001
L1.sequoia_seasdiff	0.2188	0.053	4.116	0.000	0.115	0.323
L1.kings_seasdiff	0.3621	0.029	12.480	0.000	0.305	0.419
beta.gasoline_seasdiff	-2199.3159	6400.756	-0.344	0.731	-1.47e+04	1.03e+04
beta.temp_seasdiff	173.9135	342.397	0.508	0.612	-497.173	845.000

Residual diagnosis (VARMAX Sequoia NP)

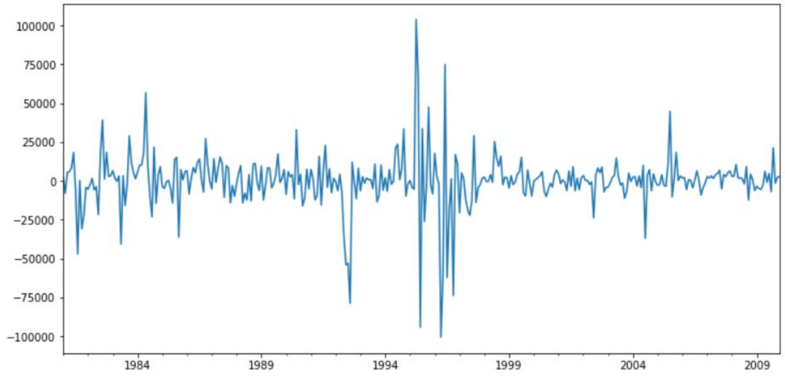


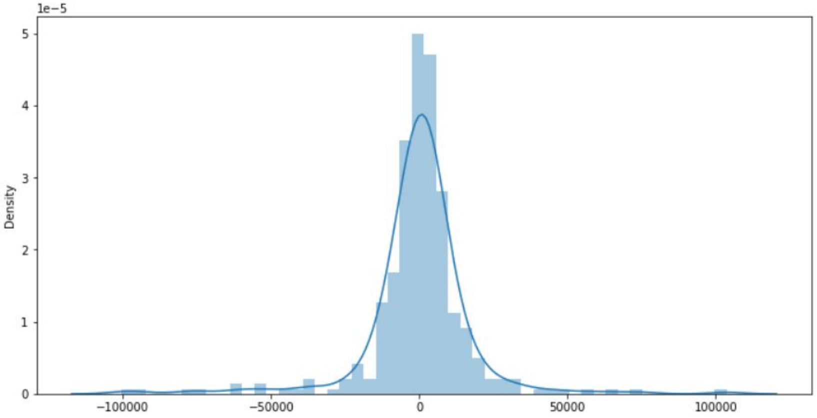


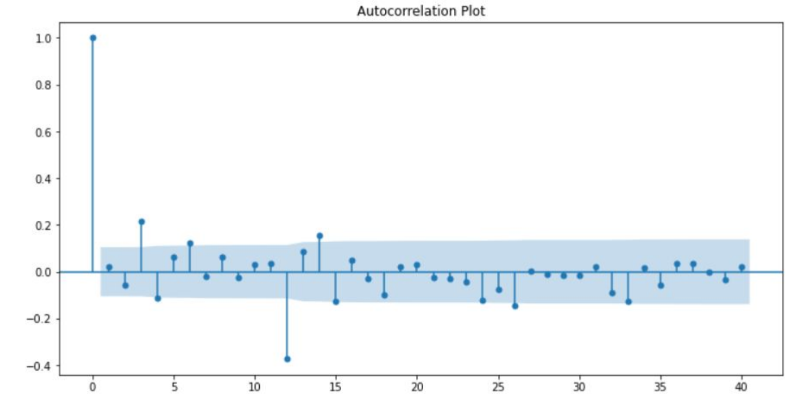


	lb_stat	lb_pvalue	bp_stat	bp_pvalue
12	67.267042	1.035801e-09	64.752406	3.029185e-09

Residual diagnosis (VARMAX Kings Canyon)







	lb_stat	lb_pvalue	bp_stat	bp_pvalue
12	82.211077	1.560167e-12	79.641411	4.831046e-12

Model Evaluation

Select the best model based on MAE, MSE, MAPE, sMAPE, MASE

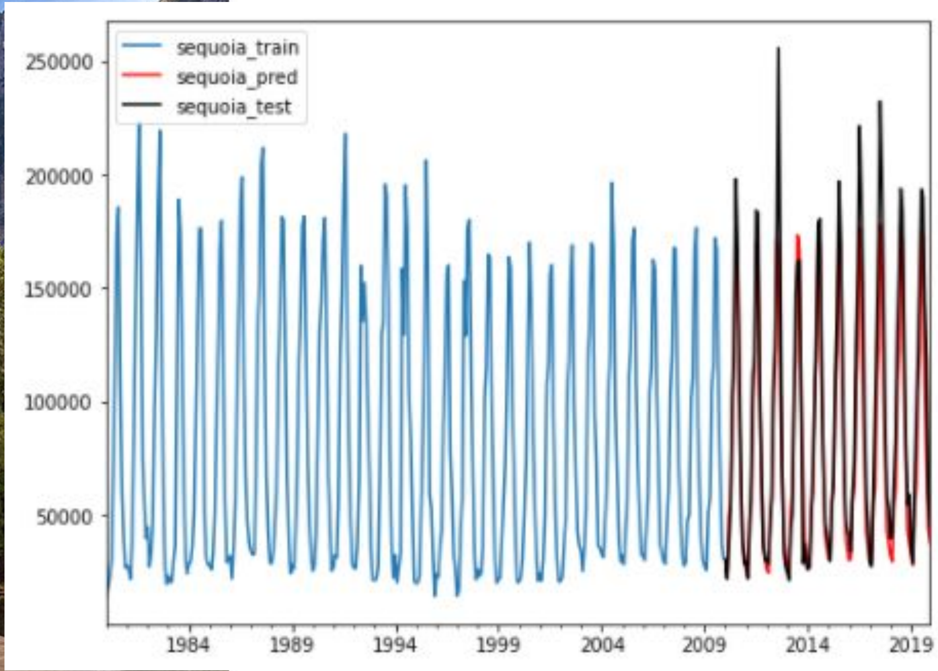
Sequoia NP

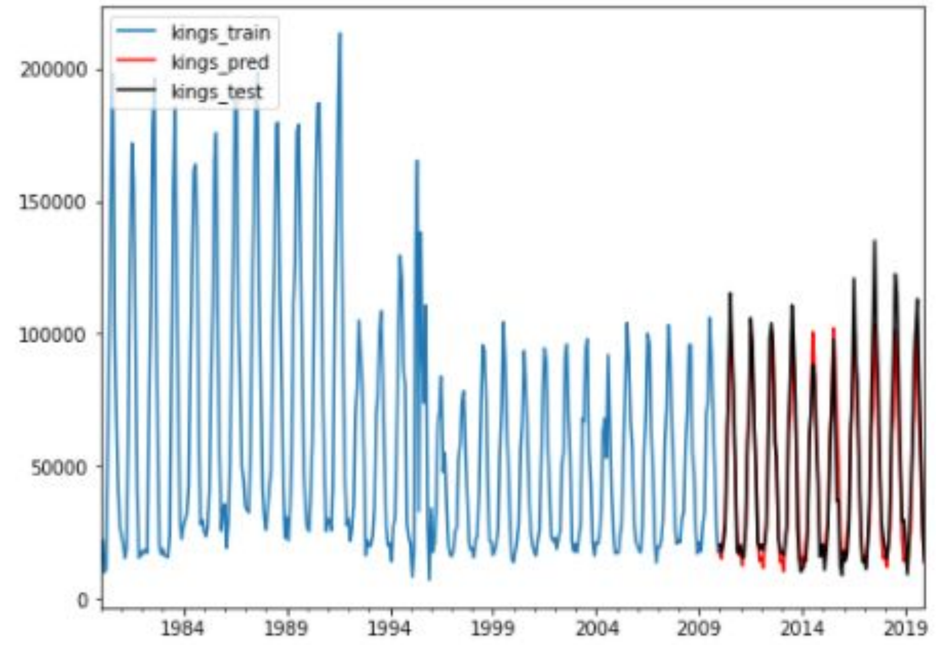
	MAE	MSE	MAPE	SMAPE	MASE
SARIMAX	16,026	527,155,772	17.41	0.19	0.59
Auto SARIMAX	16,175	524,275,566	17.76	0.20	0.59
VARMAX	**13,369**	**373,374,304**	**14.64**	**0.15**	**0.49**

Kings Canyon NP

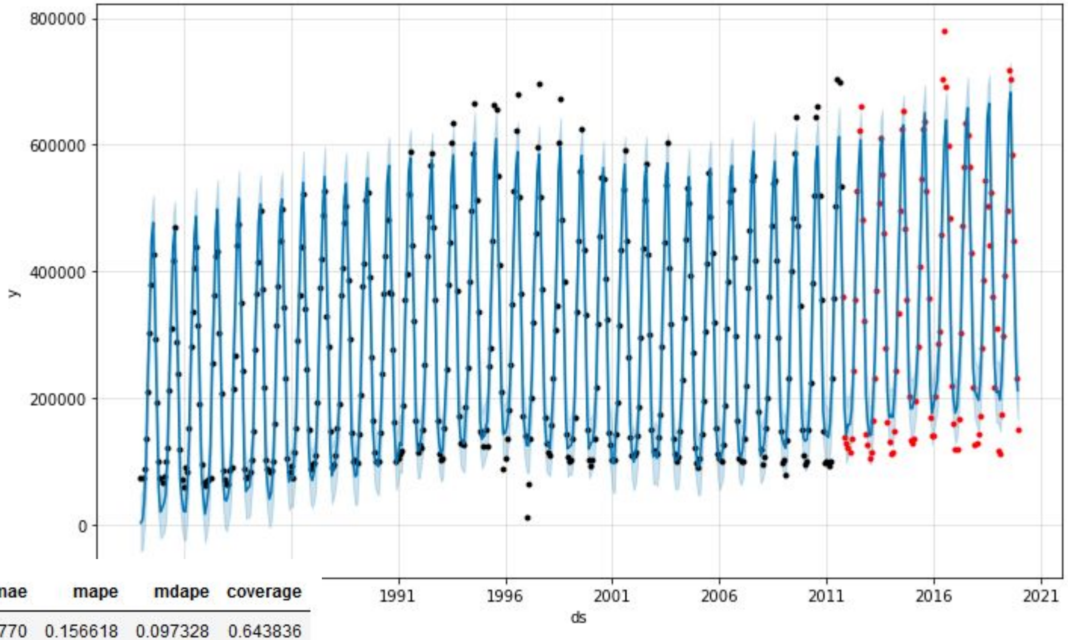
	MAE	MSE	MAPE	SMAPE	MASE
SARIMAX	8,417	136,249,669	23.08	0.20	0.40
Auto SARIMAX	8,320	134,778,362	22.75	0.20	0.39
VARMAX	**7,597**	**109,686,170**	**18.82**	**0.18**	**0.36**

Forecasting (VARMAX)





Prophet



	horizon	mse	rmse	mae	mape	mdape	coverage
0	36 days	1.347452e+09	36707.661975	28360.135770	0.156618	0.097328	0.643836
1	38 days	1.416623e+09	37638.053797	29130.543765	0.158219	0.097328	0.643836
2	39 days	1.534045e+09	39166.882276	29642.364328	0.154446	0.097328	0.650685
3	40 days	1.507362e+09	38824.754603	29167.159578	0.153800	0.097328	0.671233
4	41 days	1.528657e+09	39098.043432	29023.553136	0.153143	0.096774	0.671233
...	...	...	...	...	...	...	...
301	361 days	1.803849e+09	42471.738096	32788.446268	0.156537	0.134790	0.630137
302	362 days	1.809060e+09	42533.040821	32964.130305	0.158450	0.136363	0.625571
303	363 days	1.849077e+09	43000.897764	33343.491758	0.162305	0.136363	0.616438
304	364 days	1.860735e+09	43136.238295	33857.192782	0.162524	0.136363	0.598174
305	365 days	1.870501e+09	43249.293995	34212.815811	0.161867	0.142067	0.589041

Conclusion

- Interpretation
  - VARMAX performed best
  - Attendance at smaller parks is affected by visitorship at Yosemite
- Business application
  - Used to inform park maintenance about how many visitors to expect and therefore give insight into how much resources to allocate for certain time periods
  - Periods of low visitors can be predicted so major maintenance or renovations can be scheduled
  - Local businesses such as hotels and restaurants will be aware of how many visitor to expect
- Limitation
  - We are only looking at parks in a certain radius , so other factors such as travel conditions unique to the area, pricing, and unique characteristics of the parks (flora and fauna) can also have an effect, which we are not taking into consideration
  - Data is summarized monthly, so fluctuations at a smaller level are lost

Next Steps

- Try to find data specific to each of the larger and and smaller national parks such as parking costs, number of trails, and number of unique plant species to see if any characteristics might impact the visitation rates
- Looking at similar parks in other regions of the country and observe if a similar relationship is present between larger and smaller neighboring parks
- Determine if park visitorship affects attendance at other attractions such as theme parks or museums
- Look at other models such as fine tuning Prophet for more accurate predictions

Thanks!

Any questions?

Citation

- This presentation has been designed using resources from PoweredTemplate.com
- https://www.ncdc.noaa.gov/cag/divisional/time-series/0405/tavg/1/12/1979-2020?base_prd=true&begbaseyear=1979&endbaseyear=2020