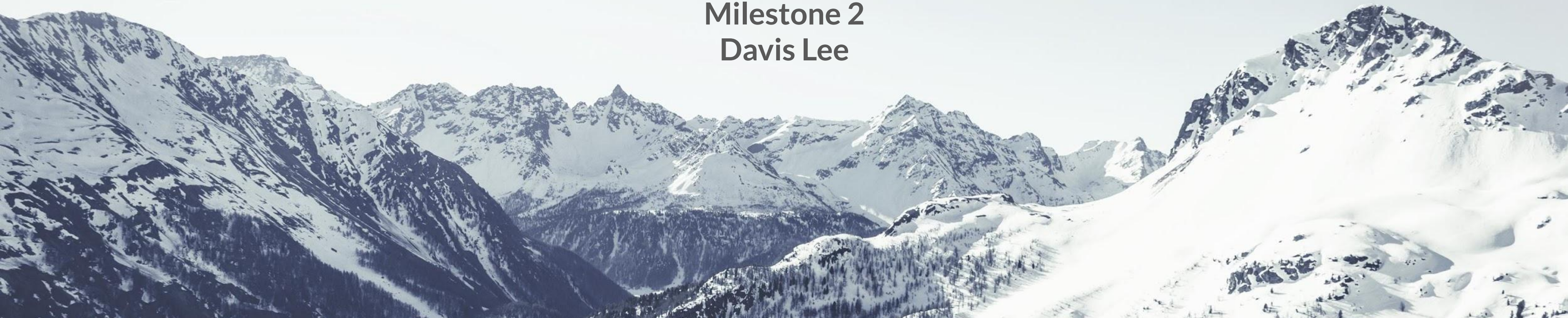
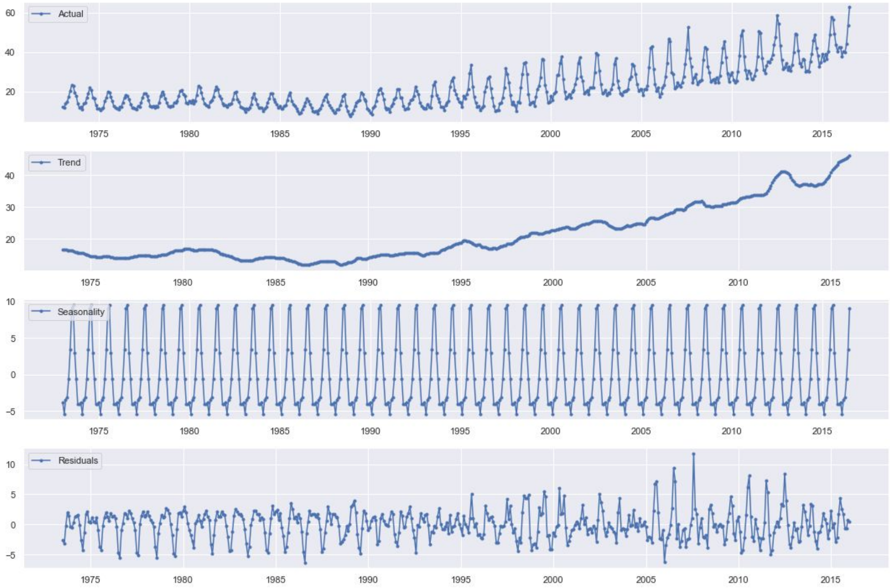
Carbon Emissions Forecasting

Milestone 2
Davis Lee



Refined Insights



1

CO2 emission peaks significantly every summer while it is at its annual minimum on winter season

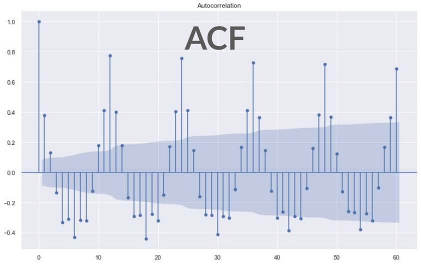
While one may argue this is simply because energy demand peaks every summer (as electricity is the major source of cooling in summer), it cannot explain then why CO2 emission is at its lowest every winter. Finding out reason behind low CO2 emission in winter seasons may lead us to a way reduce overall CO2 emission.

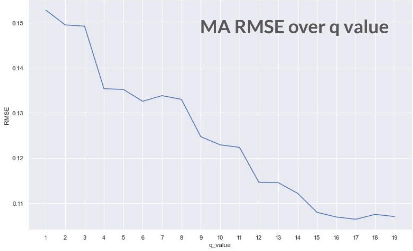
2

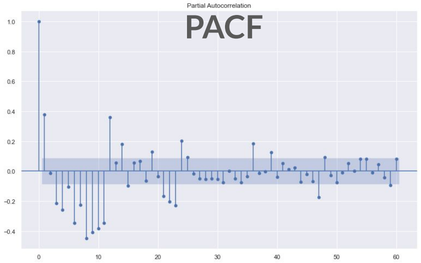
There was a significant event in the mid 1990's in terms of natural gas energy production

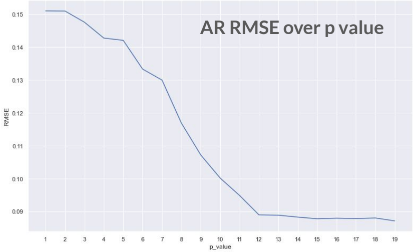
Not only that was the point when CO2 emission from natural gas started to show annual increase, that was also when the residual plot started show completely different patterns from the pattern it had kept for decades. Finding out what happened during this period (mid 1990's) may lead us to an insight that will help us derive to better solution

Techniques Used



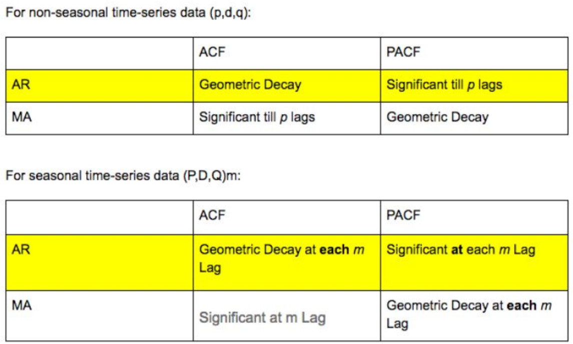






ACF/PACF Cheat Sheet

Source :
https://www.linkedin.com/pulse/reading-acf-pacf-plots-missing-manual-cheatsheet-saqib-ali/



For non-seasonal time-series data (p,d,q):

	ACF	PACF
AR	Geometric Decay	Significant till *p* lags
MA	Significant till *p* lags	Geometric Decay

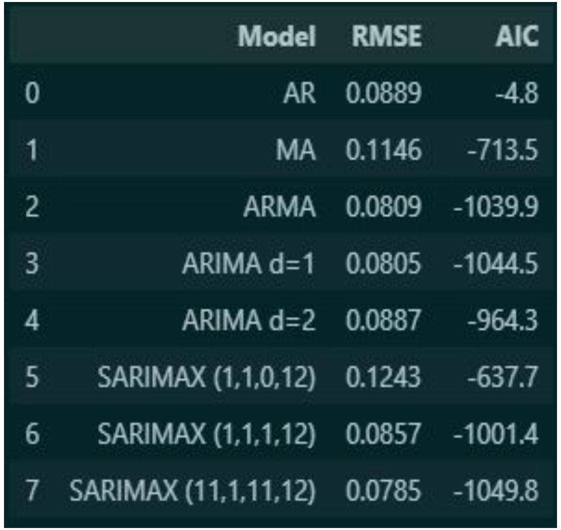
For seasonal time-series data (P,D,Q)m:

	ACF	PACF
AR	Geometric Decay at **each** *m* Lag	Significant **at** each *m* Lag
MA	Significant at m Lag	Geometric Decay at **each** *m* Lag

- To interpret the ACF and PACF plots, i referred to the cheat sheet above.
- ACF plot showed slow but gradual decay every 12 lag, and from PACF plot, we can observe significant spike every 12 lag. According to the cheat sheet, this indicates our data is seasonal arima model with seasonal AR and non seasonal AR.
- As for p and q values for my models, I plotted RMSE over p and q value and chose where there was significant ankle pattern (p=12,q=12) to get the best performance. My plan was to reduce p and q values in case my model is overfitting the training data. (it did not happen)

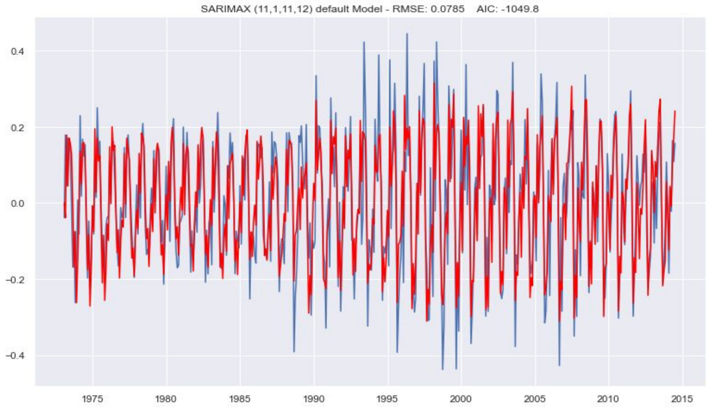
Final Solution Proposal

Tested models and their results



	Model	RMSE	AIC
0	AR	0.0889	-4.8
1	MA	0.1146	-713.5
2	ARMA	0.0809	-1039.9
3	ARIMA d=1	0.0805	-1044.5
4	ARIMA d=2	0.0887	-964.3
5	SARIMAX (1,1,0,12)	0.1243	-637.7
6	SARIMAX (1,1,1,12)	0.0857	-1001.4
7	SARIMAX (11,1,11,12)	0.0785	-1049.8

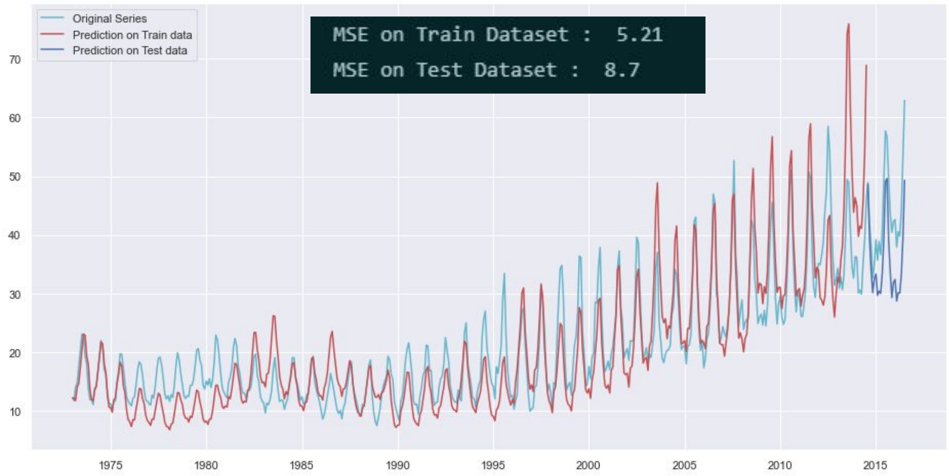
Proposed Model - SARIMAX



- AR, MA, ARMA, ARIMA d=1, ARIMA d=2, and SARIMAX with three variations of seasonal parameters were tested and compared
- As expected from my ACF/PACF plot analysis, Sarimax (p=11,d=1,q=11)(P=1,Q=1,D=1,M=12) showed the best performance on the training dataset with the lowest RMSE value as well as with the lowest AIC value among the models I tested
- Test results aside, SARIMAX would still be the most reasonable choice for the data as the data presented very prominent annual seasonality and models other than SARIMA/SARIMAX do not account for seasonality

Recommendations

Actual vs Predicted



```
MSE on Train Dataset :  5.21
MSE on Test Dataset :  8.7
```

Recommendations

- The model may perform better if we are to take data from 1995 instead from the whole dataset (from 1975). As i pointed out in the “Refined Insights” slide, I suspect there was a significant event that impacted CO2 emission from natural gas energy production that changed the emission pattern as well as the overall trend.

- I suggest performing extensive “Grid Search” for SARIMAX parameters to further improve model accuracy as I was only able to test a handful set of parameters as of yet

- Using regression models for trend and seasonal components and summing up forecast results may yield better result (just an idea)
  - linear / polynomial regression model for trend component
  - Multiple linear regression model for seasonality component
  - Time-series models for residual(noise) component