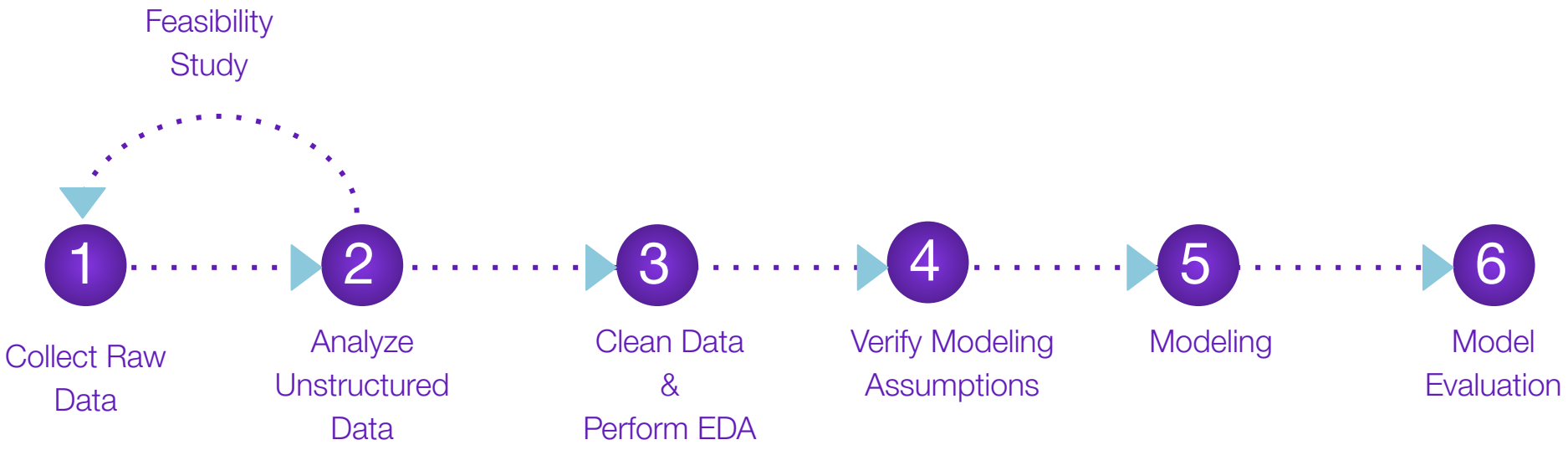
Understanding Load Impact and Power Outages

Michael Roberts, Prasoon Karmacharya and Stacey Andreadakis

Our Objective:
Predict Duration of Power Outages

The Reality:
Finding Good Data is Difficult

METHODOLOGY



ITERATIONS OF DATA

Phase 1	Columbia University IP Address Data Set	PowerOutage.Us Data Set	Collected Tweets using twitterscraper
Phase 2	**NYISO Load Data Set**	Collected Tweets using twitterscraper	
Phase 3	**NYISO Load Data Set**	**Energy.Gov Outage Data (OE-417)**	Tweets Collected from 2nd Phase
Phase 4	**NYISO Load Data Set**	**Energy.Gov Outage Data (OE-417)**	**Weather Data Collected using API**

ITERATIONS OF MODELS

Phase 1	Logistic Regression Model	Features in Regression Model	Sentiment Analysis
Phase 2	Logistic Regression Model	Sentiment Analysis	
Phase 3	Logistic Regression Model	Features in Regression Model	Sentiment Analysis
Phase 4	**Univariate + Multivariate Time Series Models**	**Determine Duration of Power Outages**	**Variables in Multivariate Model**

ANALYSIS OF OUTAGE DATA

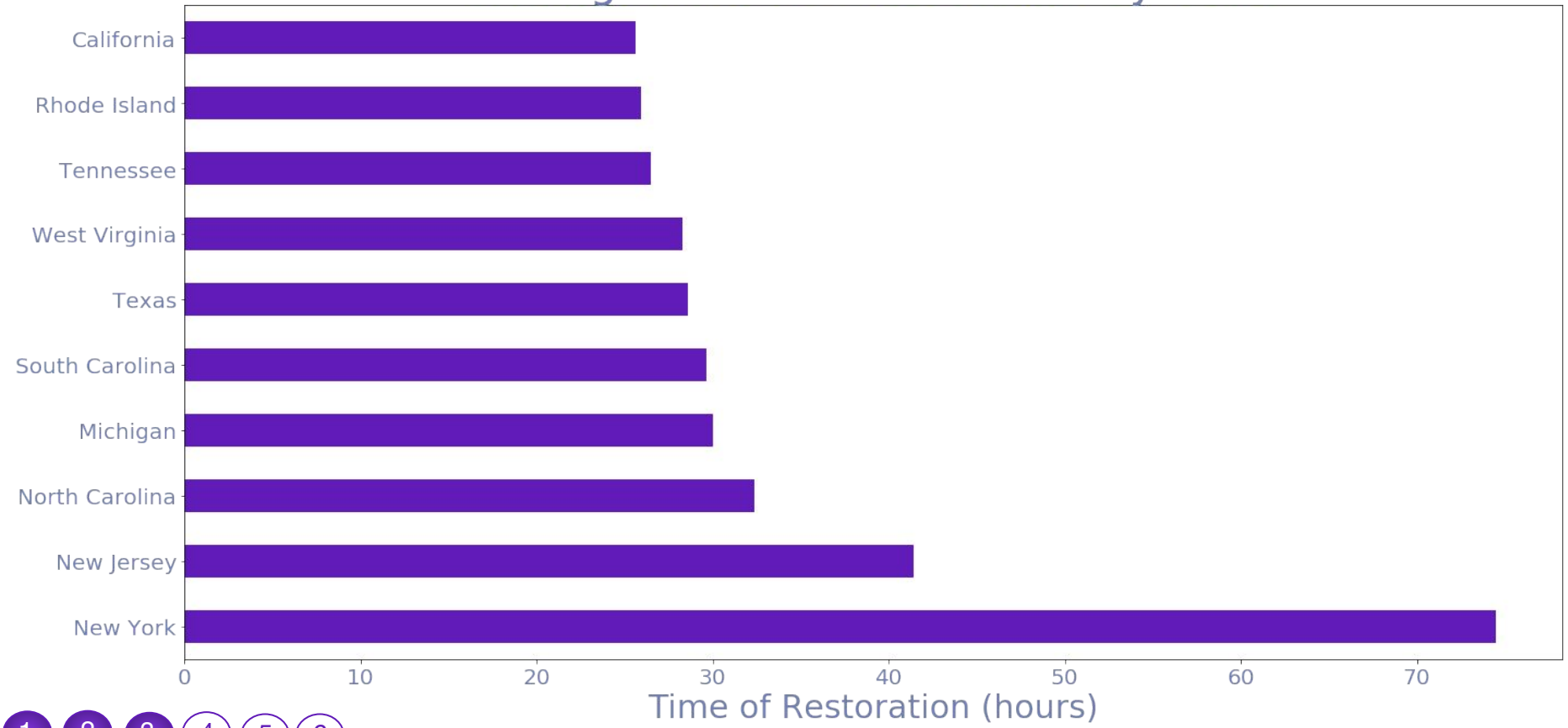
- Collecting and Cleaning
- EDA
- Visualizations
- Incompatibility with
  - Time Series
  - ISO Load Data

COLLECTING, CLEANING & ANALYZING

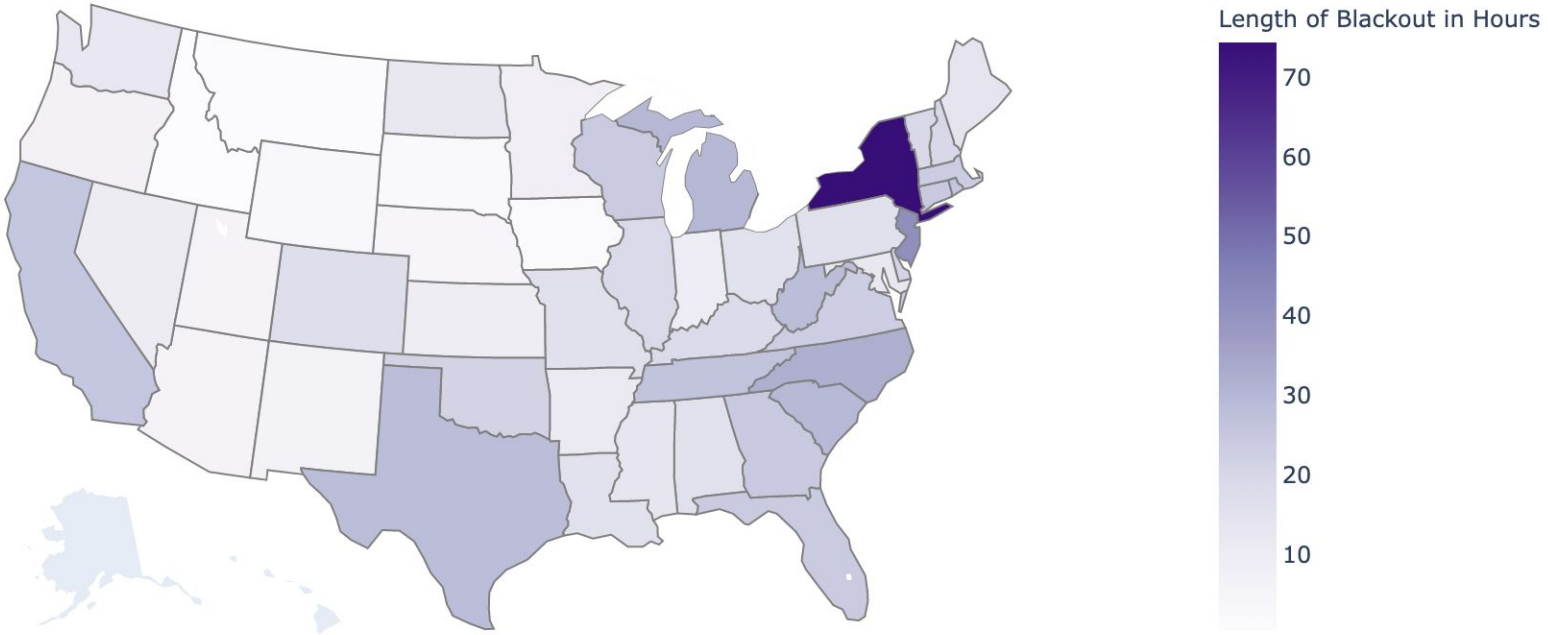
- What was in the data
  - Date/time of outage and restoration
  - Area Affected
  - Alert Criteria
  - 2015-2020
- Creating timedelta objects
  - Calculating restoration time
- Duplicating rows for observations with multiple states

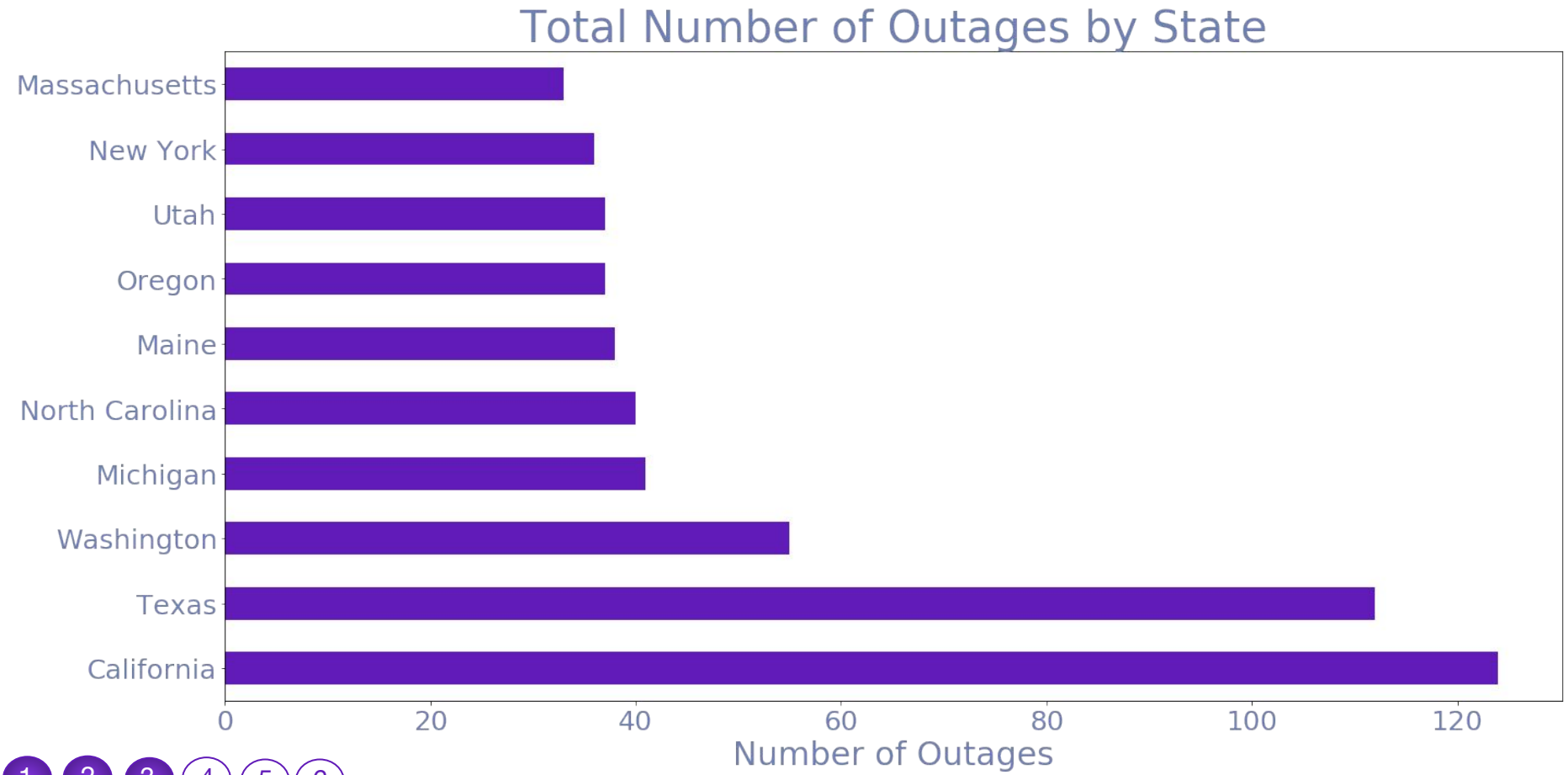
state	NA
[Tennessee]	2/10/2002 21:00
	2/27/2002 11:35
[Alabama, Georgia, Kentucky, Missouri, North C...	
	3/11/2002 12:00
[Oregon]	
	9-Apr

Average Time of Restoration by State

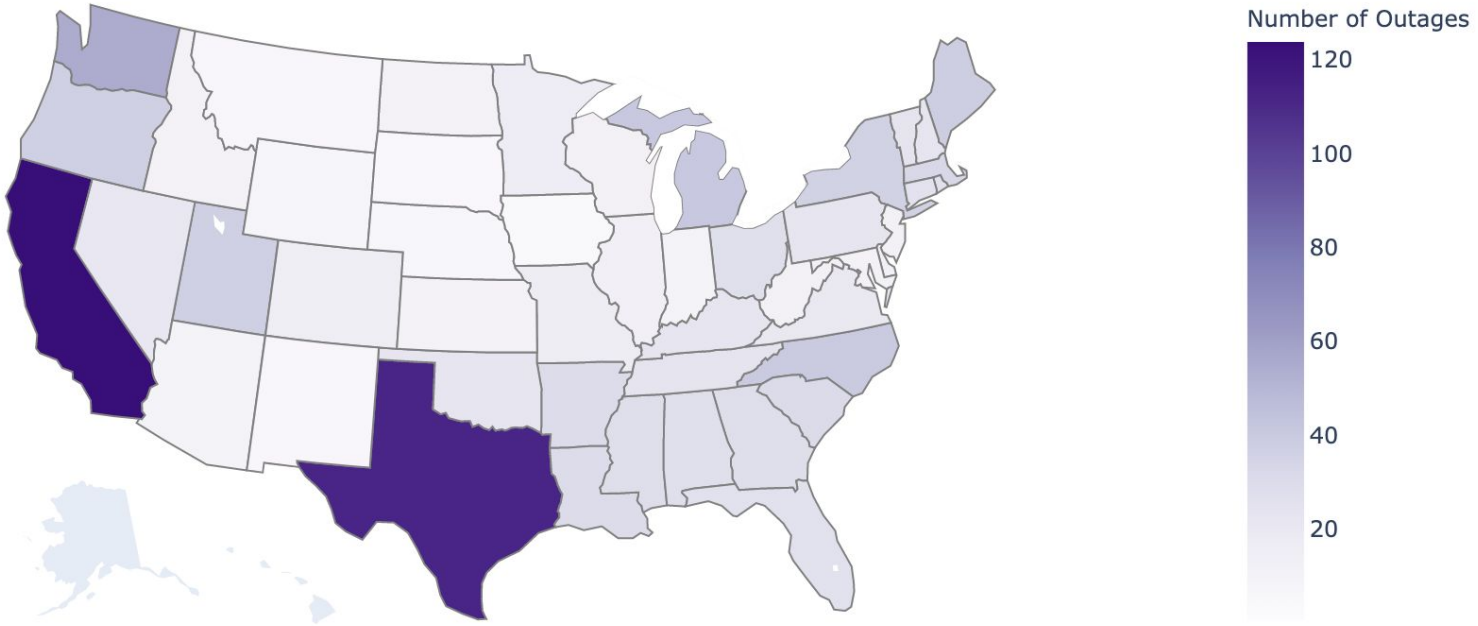


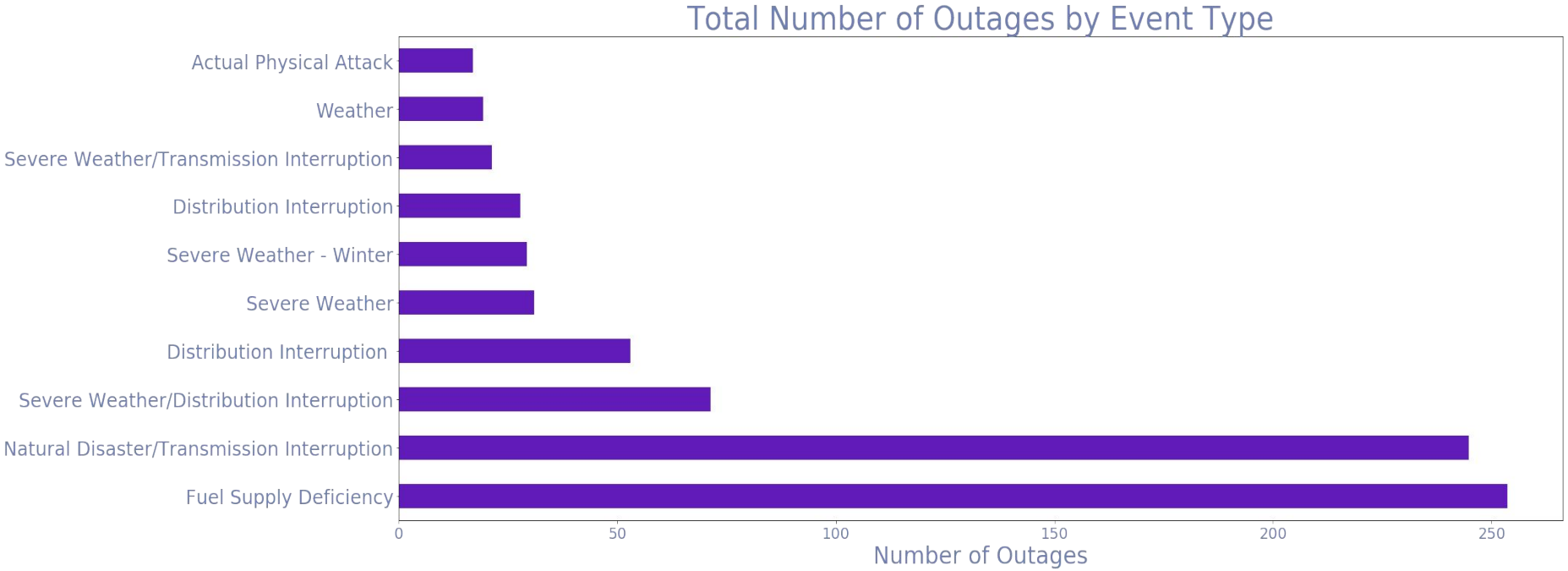
Average Length of Power Outages by State




Total Number of Outages by State
Massachusetts
New York
Utah
Oregon
Maine
North Carolina
Michigan
Washington
Texas
California
0
20
40
60
80
100
120
Number of Outages

Number of Outages by State




Total Number of Outages by Event Type
Actual Physical Attack
Weather
Severe Weather/Transmission Interruption
Distribution Interruption
Severe Weather - Winter
Severe Weather
Distribution Interruption
Severe Weather/Distribution Interruption
Natural Disaster/Transmission Interruption
Fuel Supply Deficiency
0
50
100
150
200
250
Number of Outages

INCOMPATIBILITY OF DATA

ISO Load Data

There was not enough observations in the Energy.Gov Data for New York to either engineer features or create target columns for that dataset to accompany Load Data.

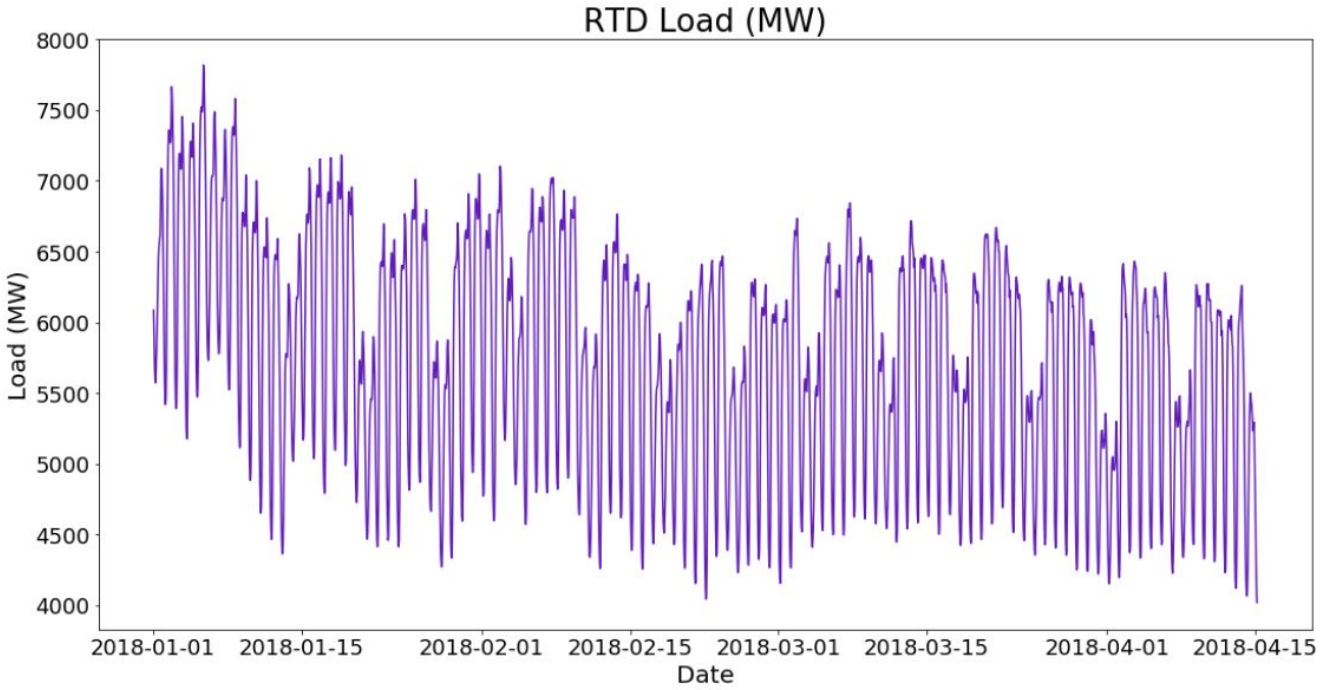
Time Series

While our knowledge of time series models was fairly new, after EDA it became clear that the Energy.Gov Outage Data could not be used in conjunction with ARIMA or VAR models

MODEL 1: UNIVARIATE TIME SERIES

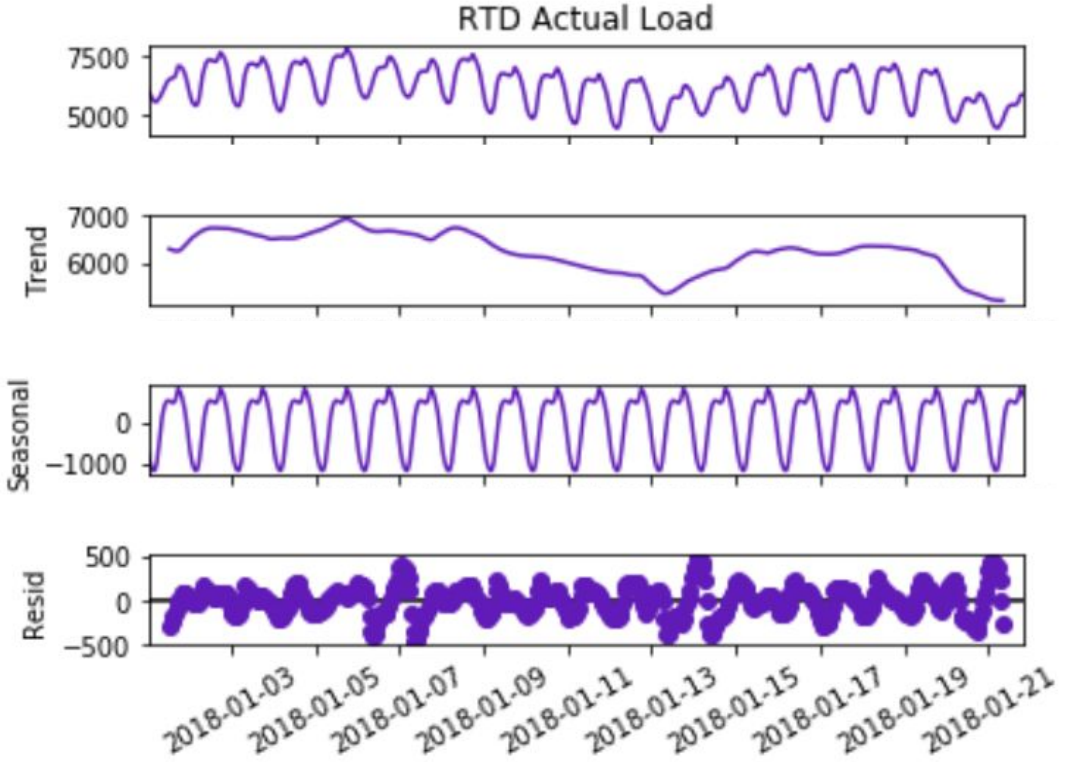
- MODEL : ARIMA (AutoRegressive Integrated Moving Average)
- Checking Seasonality, Trends and Stationarity
- Identifying the lag time
- Grid search on p, q, d
- Forecasting
- Evaluating the model

POWER LOAD DATA



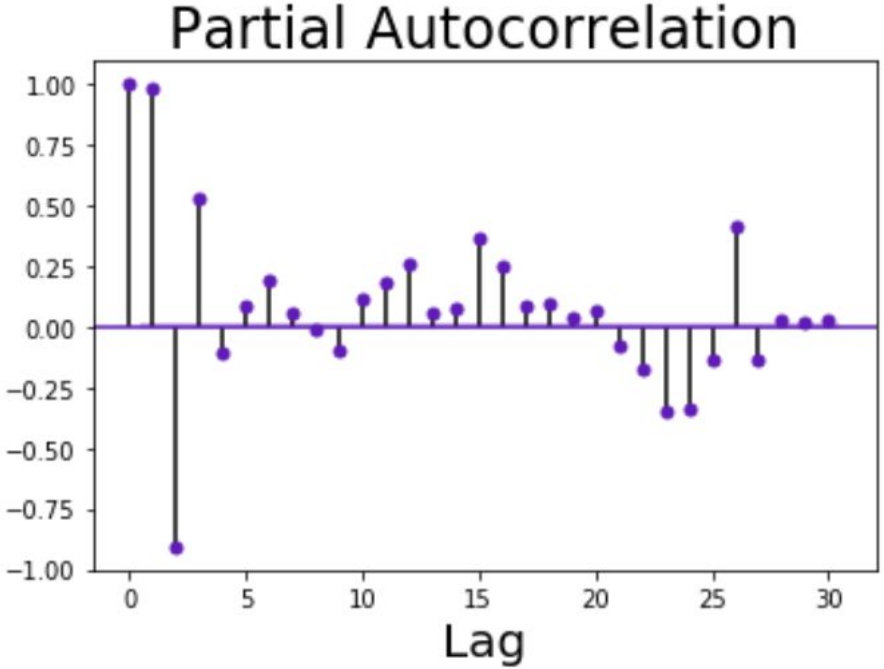
- Load Data for NYC (2018, 2019)
- Source: NYISO
- Test/Train Split: 90/10

TREND, SEASONALITY AND STATIONARITY - UNIVARIATE



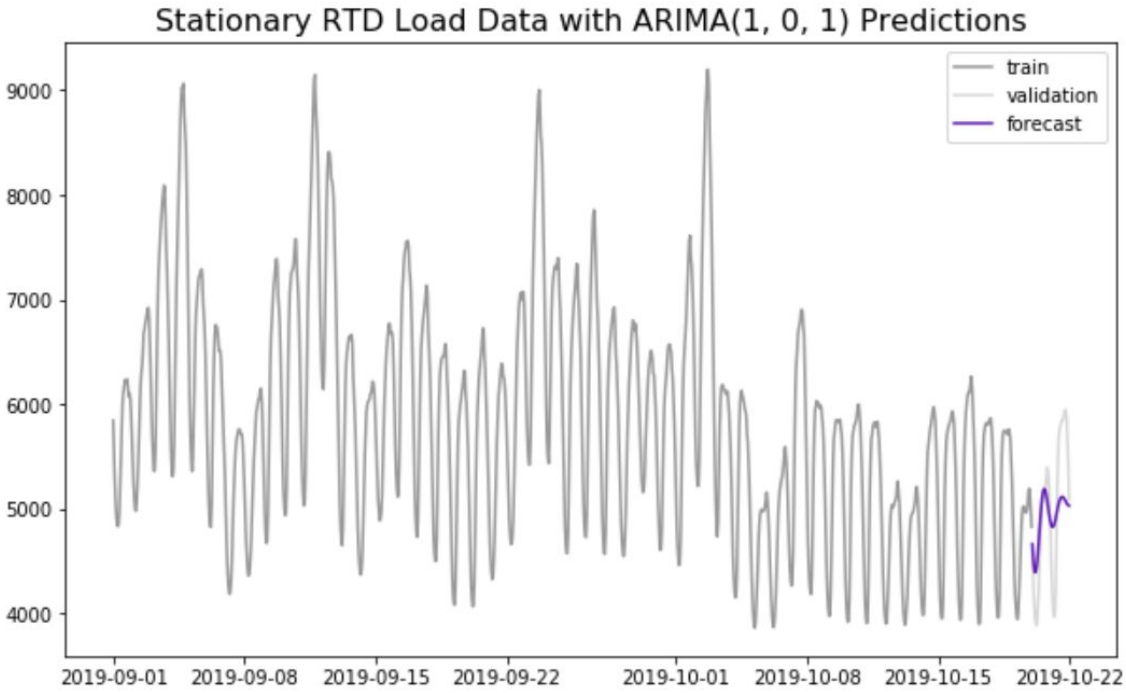
- Augmented Dickey-Fuller Test: (p-value = 0.0)

IDENTIFYING LAG - UNIVARIATE

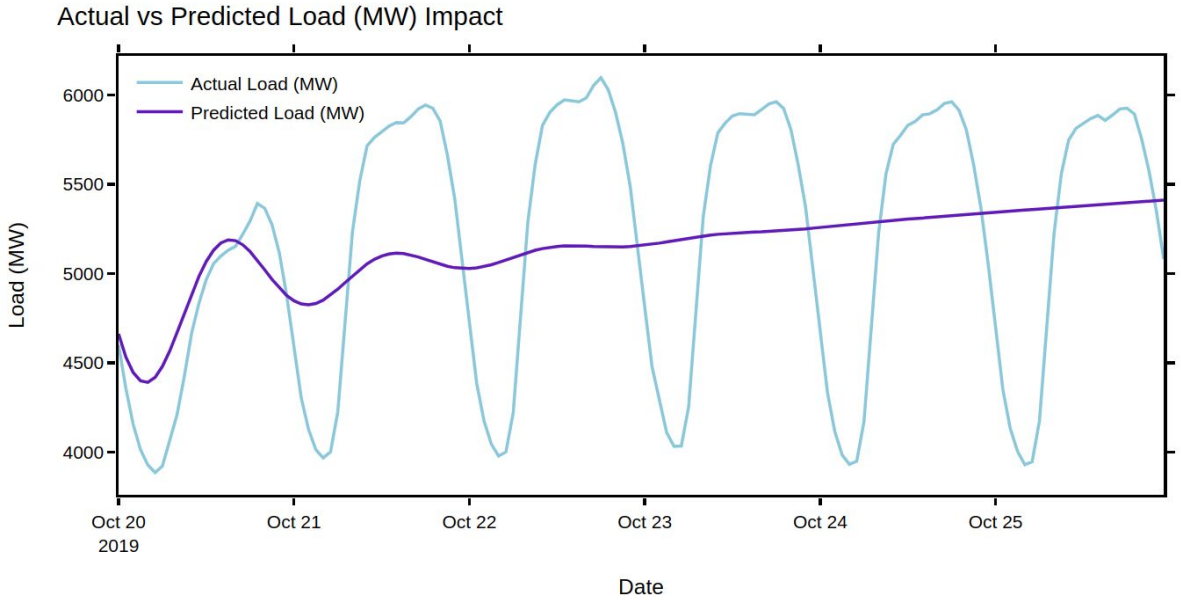


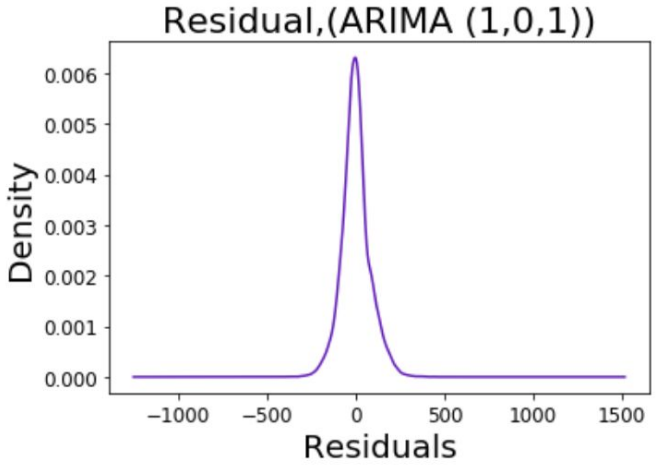
- Lag (p) = 7, based on PACF and Augmented Dickey-Fuller Test

UNIVARIATE TIME SERIES ANALYSIS



UNIVARIATE TIME SERIES ANALYSIS



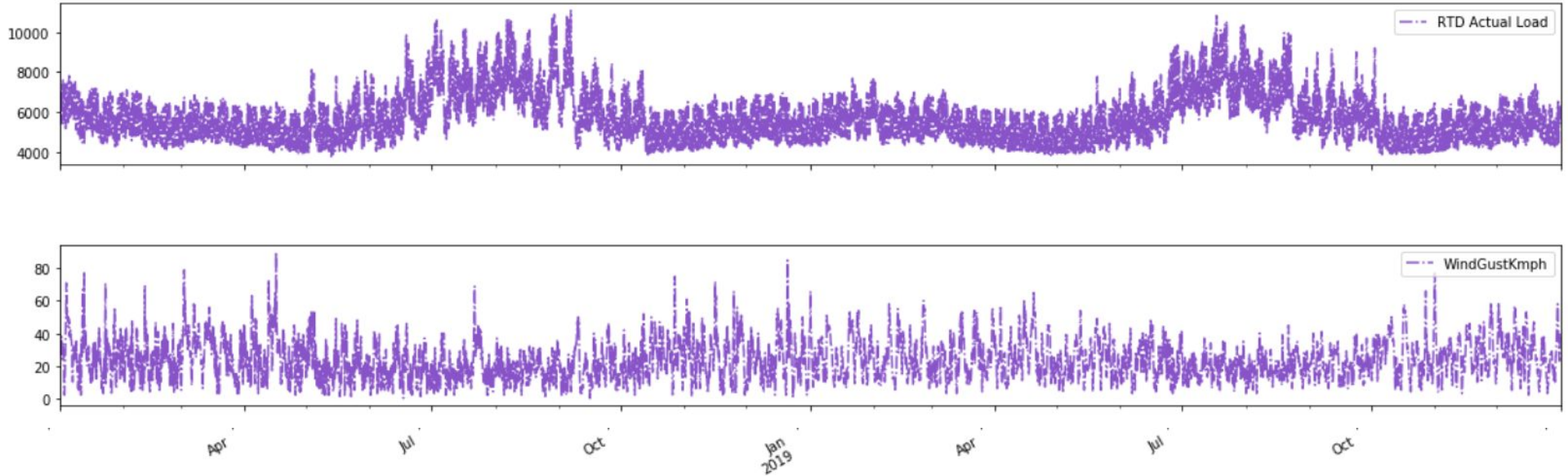


- RMSE = 888.4091 MW

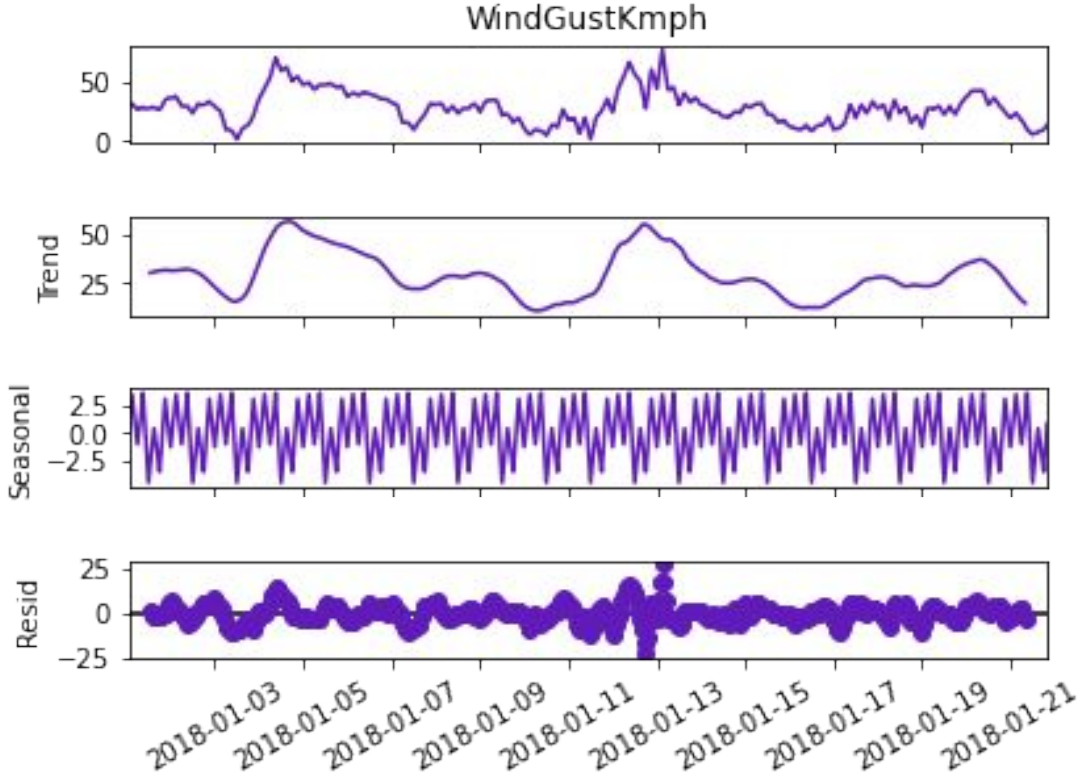
MODEL 2: MULTIVARIATE TIME SERIES

- MODEL : VAR (Vector AutoRegression)
- Endogenous and Exogenous Variables
- Checking Seasonality and Trends
- Cointegration and Stationarity Test
- Criteria to determine Lag
- Forecasting
- Evaluating the model

MULTIVARIATE TIME SERIES ANALYSIS: ENDOGENOUS VARIABLES



TREND AND SEASONALITY - MULTIVARIATE



CHECK FOR STATIONARITY: JOHANSEN TEST

VARIABLE	TYPE	EIG VALUE	STATIONARY
RTD LOAD (MW)	ENDOGENOUS	.2	YES
WIND GUST (KMPH)	ENDOGENOUS	.05	YES
WINDCHILL (C)	EXOGENOUS	.07	YES
HUMIDITY	EXOGENOUS	.03	YES
PRESSURE	EXOGENOUS	.02	YES
TEMP (C)	EXOGENOUS	.004	YES

TO BIC OR NOT TO BIC? LAG IS THE QUESTION

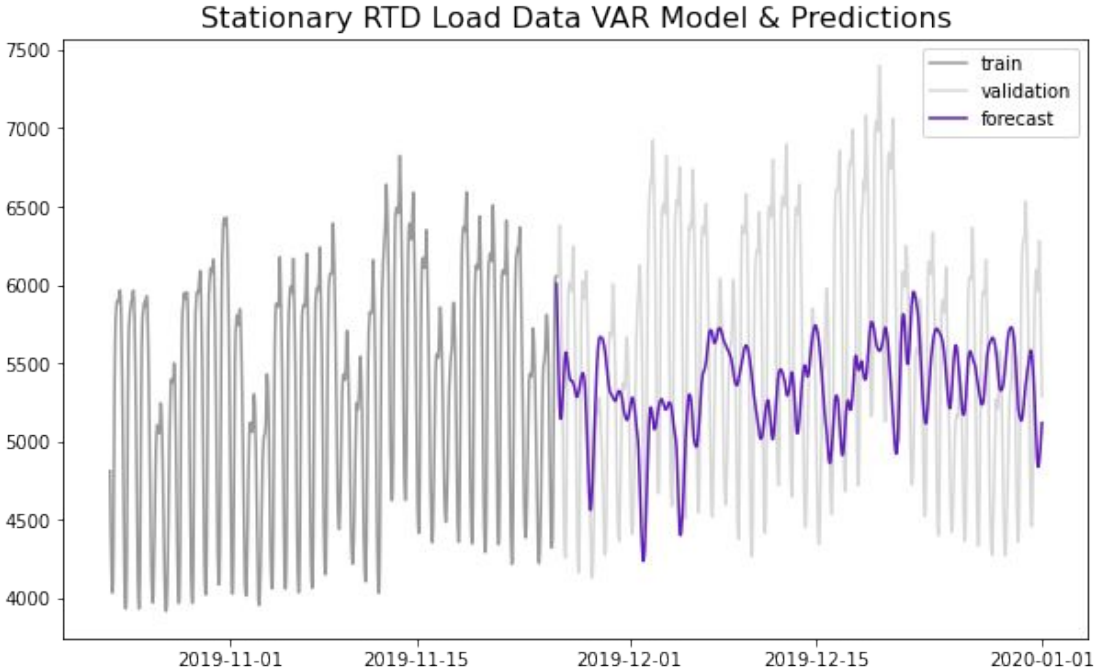
$$\text{AIC} = 2k - 2\ln(\hat{L})$$

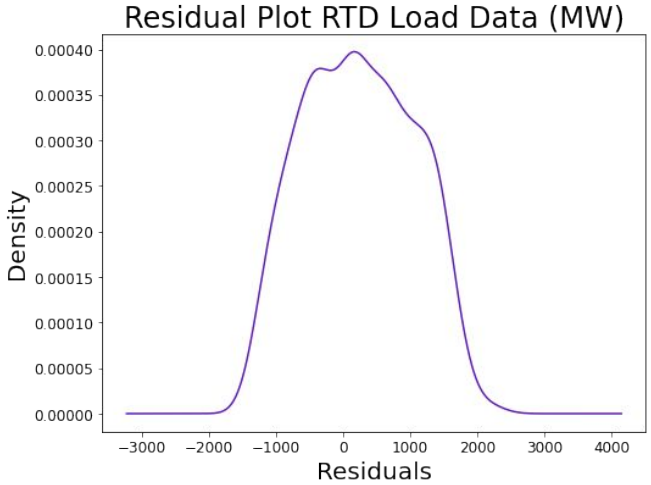
$$||y - X\beta||^2 + \alpha||\beta||^2$$

$$\text{BIC} = k\ln(n) - 2\ln(\hat{L})$$

$$||y - X\beta||^2 + \alpha||\beta||_1$$

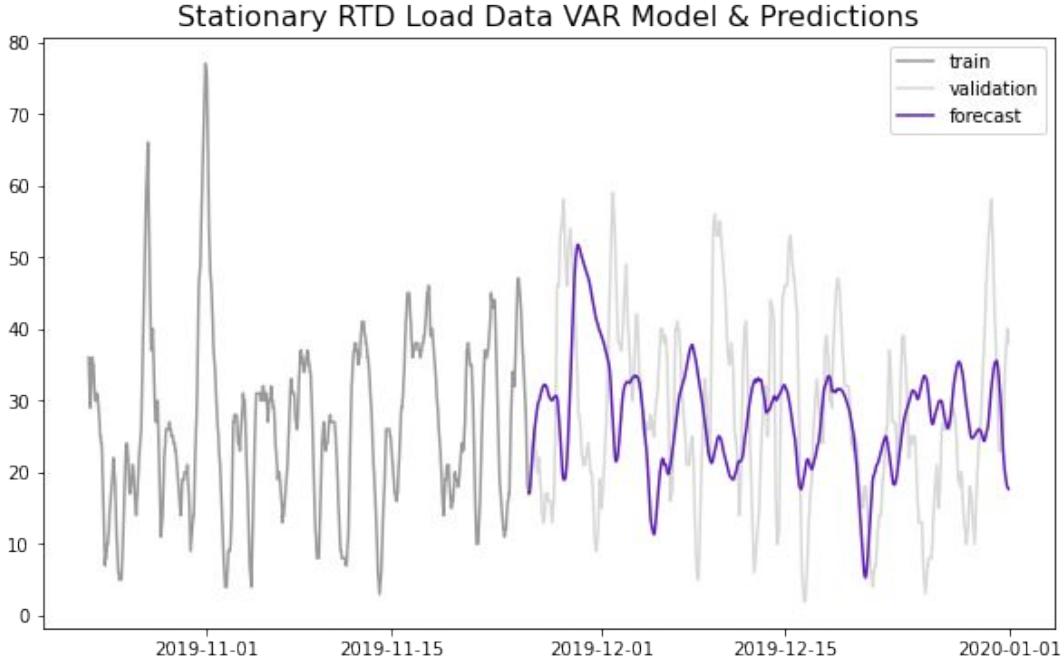
MULTIVARIATE TIME SERIES ANALYSIS

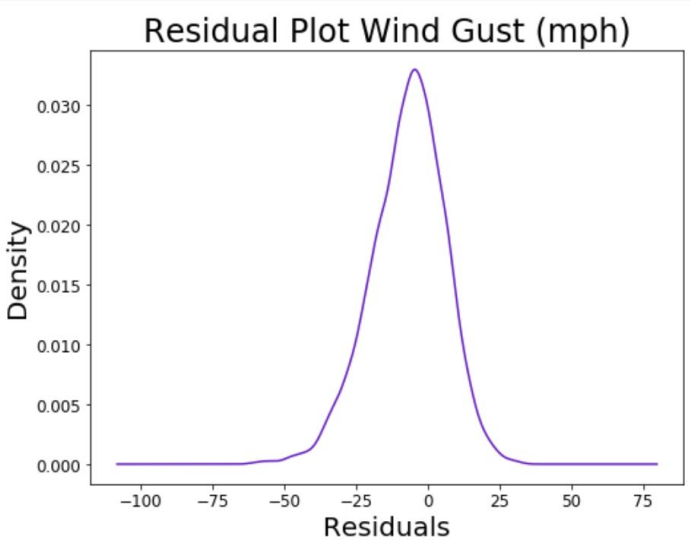




- RMSE RTD Load = 846.185 MW
- RMSE Wind Gust Actual = 15.0587 mph

MULTIVARIATE TIME SERIES ANALYSIS





- RMSE RTD Load = 846.185 MW
- RMSE Wind Gust Actual = 15.0587 mph

LESSONS LEARNED

- 80% of Data Science truly is about the collection of data, understanding the structure of the data and the rest is cake

- Working with a great team can go a long way in an otherwise arduous process

Acknowledgments

Travis Whalen

Noelle Brown

Dan Wilhelm

Niloofar Bayat

Kunal Mahajan

Vishal Misra

Dan Rubenstein

All my homies hate niloofar