Time Series Analysis - Final Project

Predicting Traffic Patterns

Tuesday Team 1: Vamshi Gadepally, Jose Gerala, Nitin Gupta, Jack Murray, Irem Pamuksuz

Spring 2023

Agenda

- Business Case and Objective
- Data Overview
- Transformations, Feature Engineering, and Assumptions
- Data Exploration
- Model Selection
- Forecast Evaluation
- Conclusion and Next Steps

Business Case and Objective

Problem : It is a concern for the Minneapolis especially during holiday seasons that the traffic accidents on highways have been increasing drastically. Some argue that this may be a result of increased traffic volume despite the fact that average daily traffic has been decreased by 5% within the city (Lee, 2019). The location is roughly between Minneapolis and St Paul.

Objective: To develop a forecasting model that accurately predicts future traffic volume, enabling better traffic management and planning

Data Overview

Variables Description:

- holiday: Categorical US National holidays plus regional holiday, Minnesota State Fair
- temp: Numeric Average temp in kelvin
- rain_1h: Numeric Amount in mm of rain that occurred in the hour
- snow_1h: Numeric Amount in mm of snow that occurred in the hour
- clouds_all: Numeric Percentage of cloud cover
- weather_main: Categorical Short textual description of the current weather
- weather_description: Categorical Longer textual description of the current weather
- date_time: DateTime Hour of the data collected in local CST time
- traffic_volume: Numeric Hourly I-94 ATR 301 reported westbound traffic volume

Observation Count & Time Period:

- Total Observations: 48,204
- Duration: Oct 2012 - Sept 2018
- Cadence: Hourly

Data Source:

UCI Machine Learning Repository - "Metro Interstate Traffic Volume Data Set"
(https://archive.ics.uci.edu/ml/datasets/Metro+Interstate+Traffic+Volume)

Citation

Traffic data from MN Department of Transportation
Weather data from OpenWeatherMap

Transformations / Feature Engineering

- Replacing the missing temp values (zero degree Kelvin entries) with the previous recorded temp value
- Convert temperature from Kelvin to Fahrenheit
- Create a variable for day of the week
- Create binary variable for holiday
- Create binary variable for rain
- Dropped variable for snow as the data is sparse. Observations of snow limited to Dec 2015 and Jan 2016 only

Assumptions

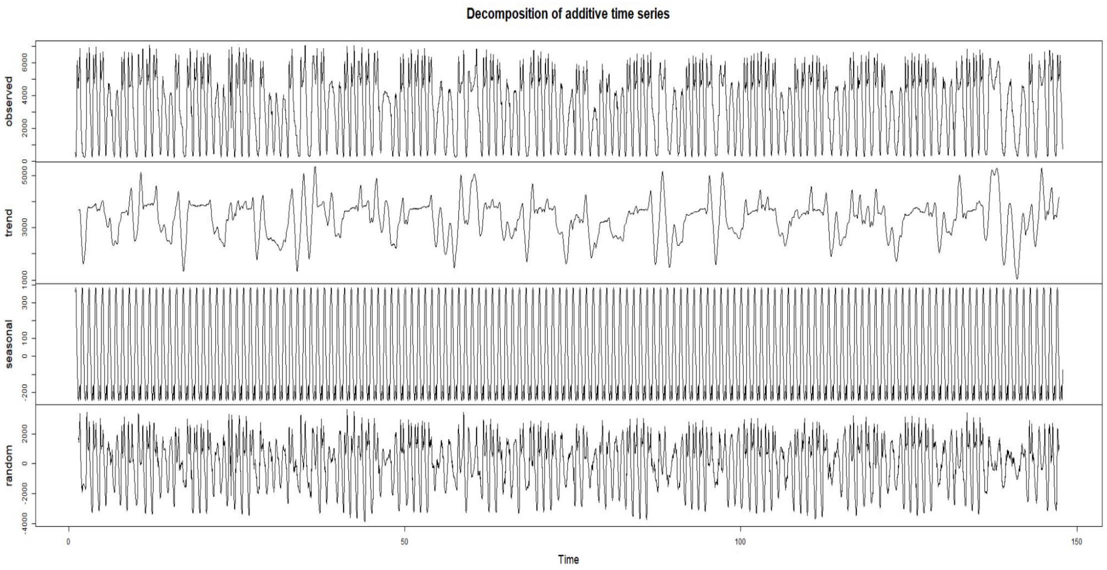
- **Data considered:** Split the dataset into a train set consisting of data from May 1 to August 31, 2018, and a test set consisting of data from September 1 to September 30, 2018.
- **Forecast timeframe:** All predictions were made based on a weekly time frame to capture the dynamic nature of traffic and provide more reliable forecasts that align with the actual traffic patterns observed in the real world.
- **Augmented Dickey-Fuller Test** resulted in p-values <0.01 and **KPSS Test** for Level Stationarity resulted in p-values of 0.1, for all variables, indicating stationary for all variables.
- **BoxCox transformation:** Suggested an appropriate lambda value for the data meaning a Box Cox was required, suggesting the original data did have certain changes (increase or decrease) in variation with the level of the time series.
- Despite stationarity **1st order differencing** showed to have a reduction in the seasonality for this time series data.

Decomposition shows strong seasonality

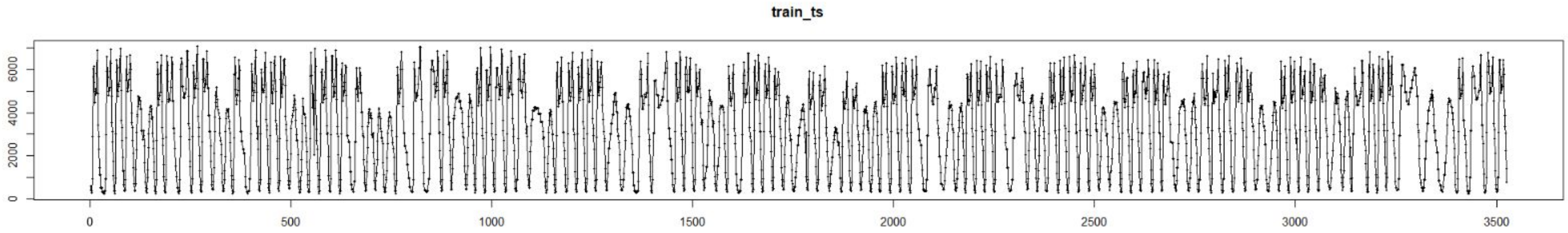
Data suggests **seasonality** occurring every 24 hours

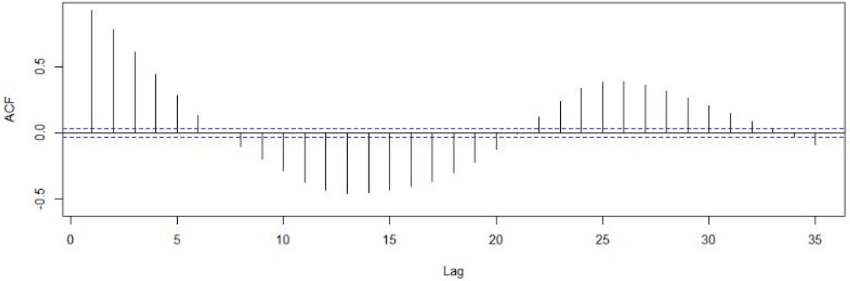
No trend is apparent in the data

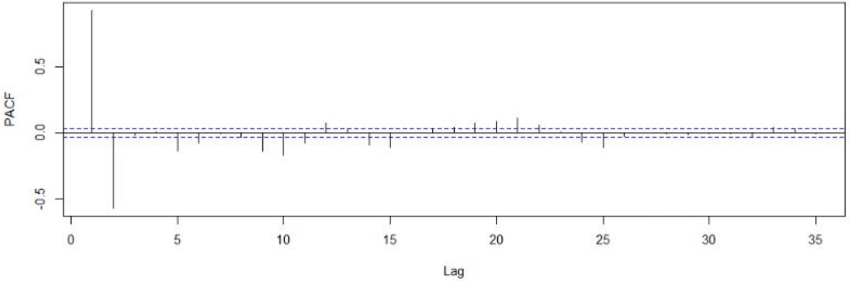
The data appears to be **non-stationary** and will require additional transformations



Data Exploration







Sinusoidal ACF plot further indicates seasonality in our time series data

Data Exploration

```
]: ts.isnull().sum()
```

```
]: holiday                0
   temp                   0
   rain_1h                0
   snow_1h                0
   clouds_all             0
   weather_main           0
   weather_description    0
   date_time              0
   traffic_volume         0
   dtype: int64
```

No null values observed

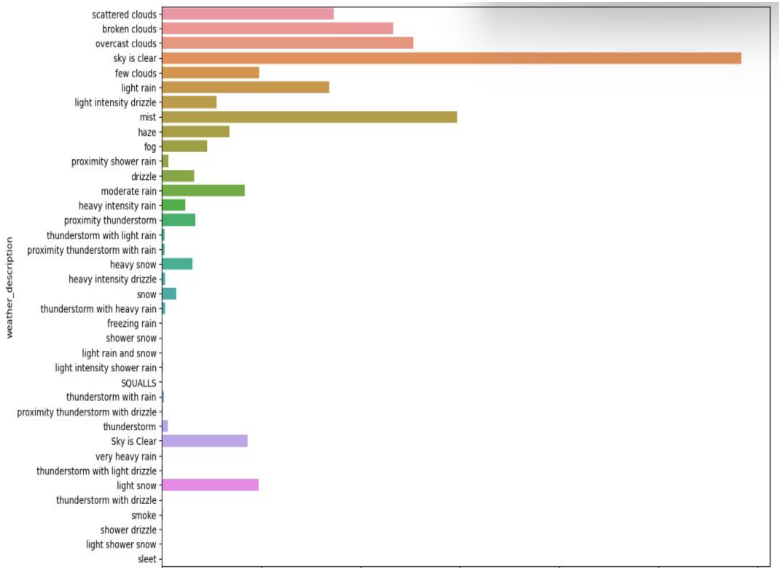
```
2]: ts.describe()
```

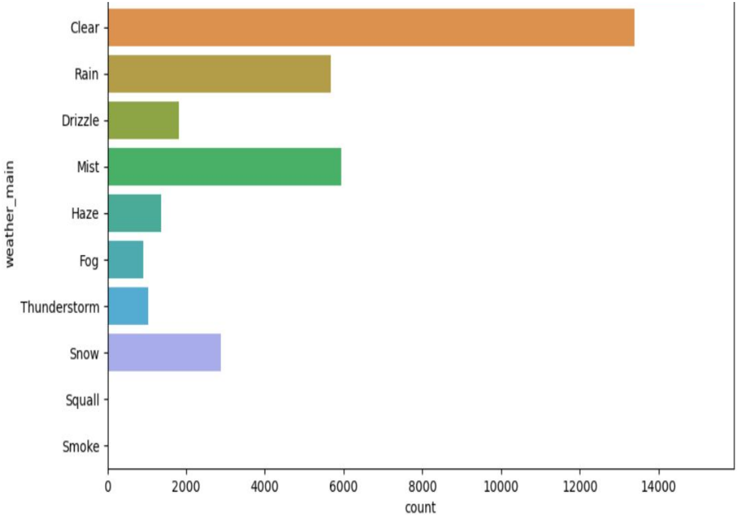
2]:

	temp	rain_1h	snow_1h	clouds_all	traffic_volume
count	48178.000000	48178.000000	48178.000000	48178.000000	48178.000000
mean	281.205439	0.334439	0.000223	49.342791	3260.149840
std	13.341764	44.801217	0.008170	39.016968	1987.020666
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	272.160000	0.000000	0.000000	1.000000	1193.250000
50%	282.450000	0.000000	0.000000	64.000000	3380.000000
75%	291.810000	0.000000	0.000000	90.000000	4933.000000
max	310.070000	9831.300000	0.510000	100.000000	7280.000000

Summary of numeric columns

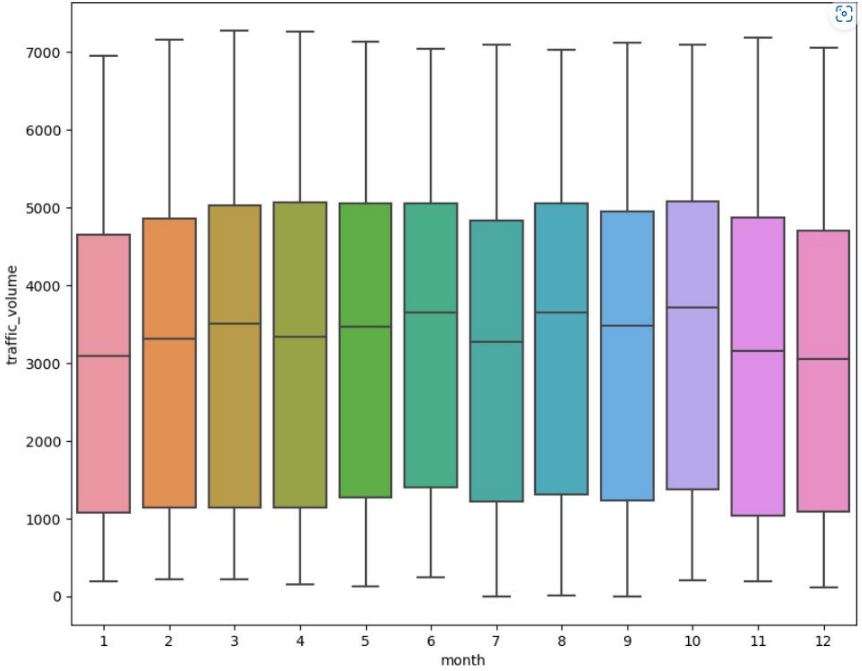
Data Exploration





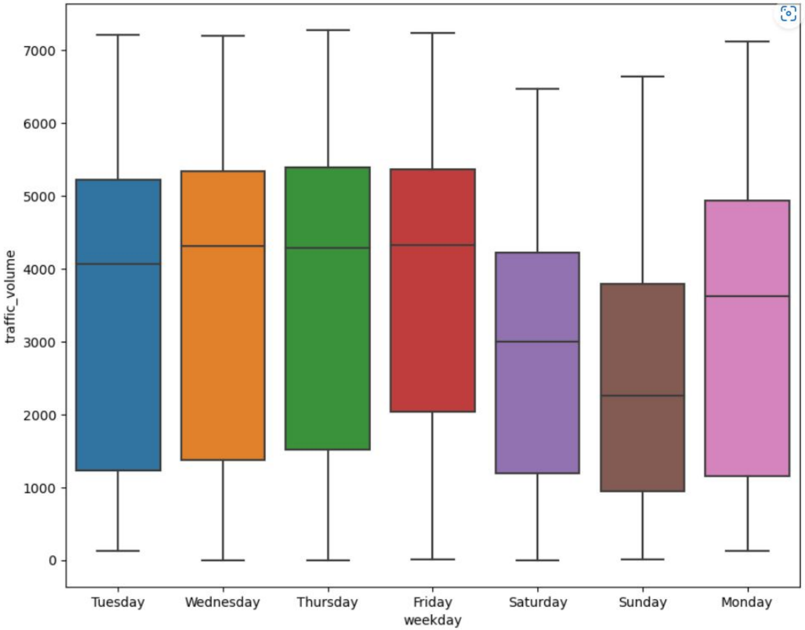
Most recorded days had clear skies with no precipitation

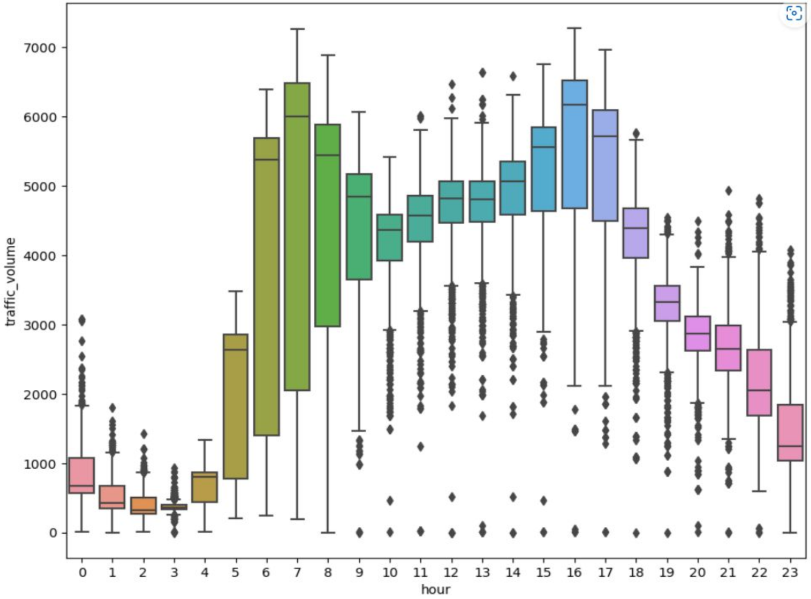
Data Exploration



Traffic volume is higher in the summer months

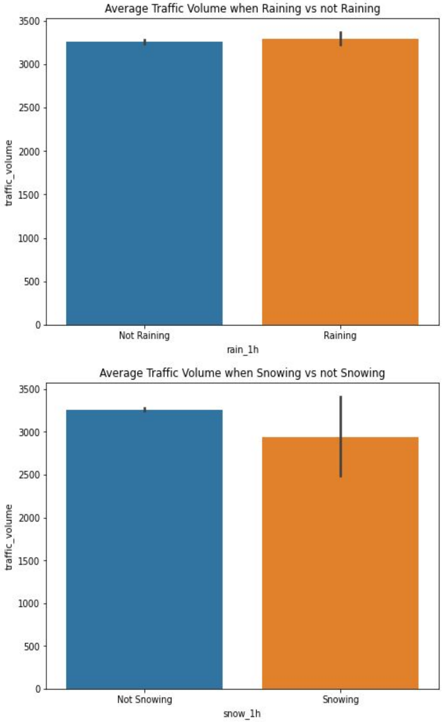
Data Exploration



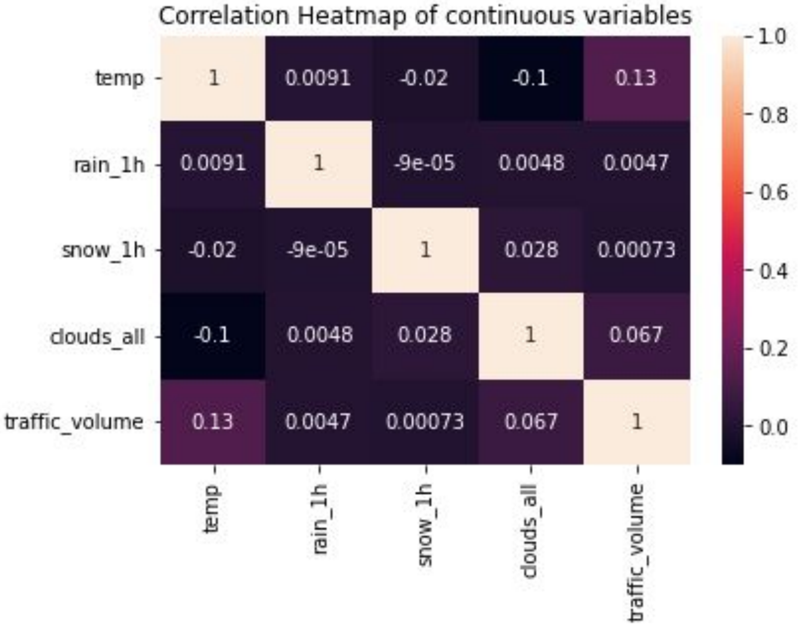


Traffic seems to be higher on weekdays and during commute hours of morning and early evening

Data Exploration



Average traffic volume decreases when it's snowing



Traffic volume is most correlated with temperature

Model Evaluation & Selection

Following are the family of time series models evaluated:

1. Naive and Seasonal Naive
2. ARIMA family
3. ARIMAX
4. LSTM

Model Selection Criteria

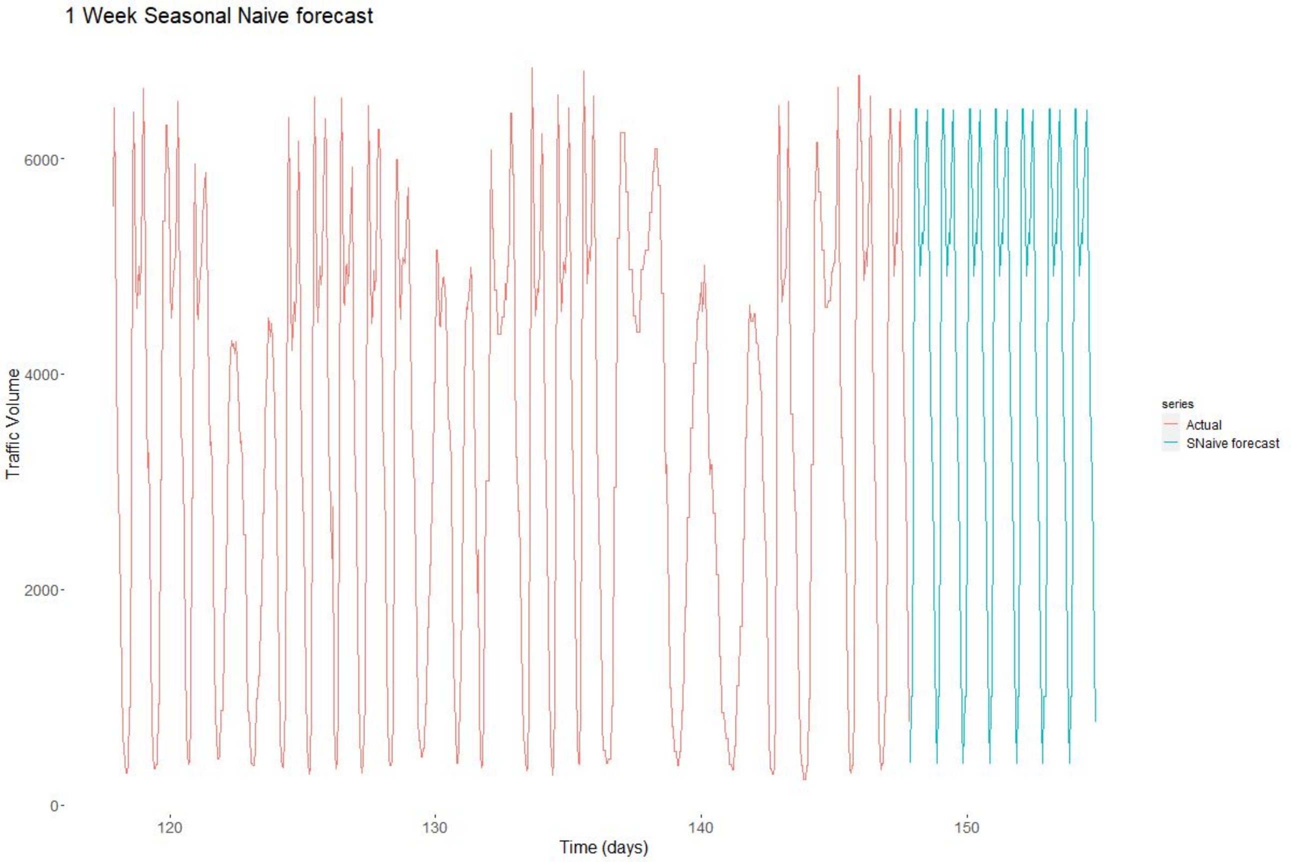
R squared, AIC, BIC, AICc for same family of models

RMSE for inter-family model comparison

Model 1 - Seasonal Naive

1 Week Forecast

Daily seasonality

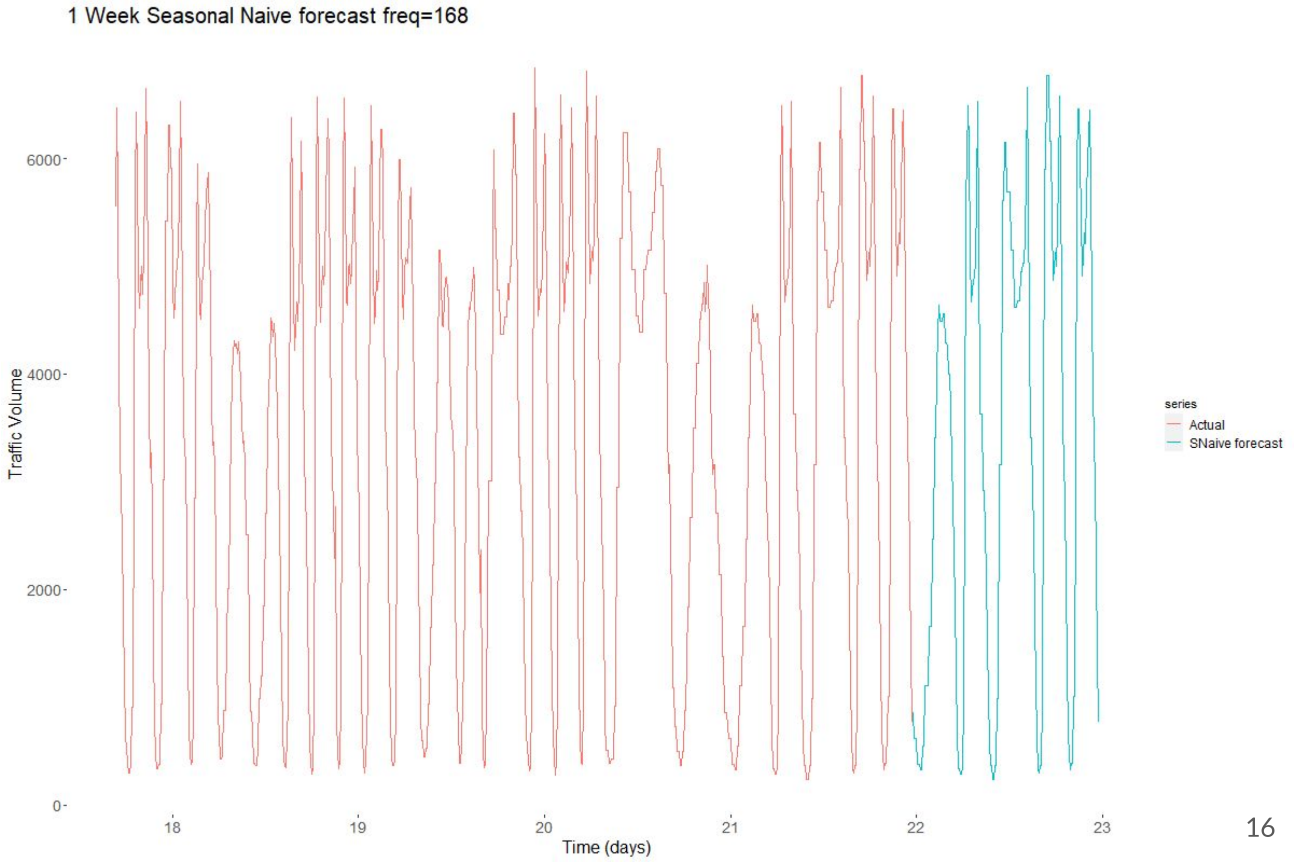


"MAE: 2747.76785714286"
"RMSE: 3212.96942373518"

Model 1 - Seasonal Naive

1 Week Forecast

Weekly seasonality

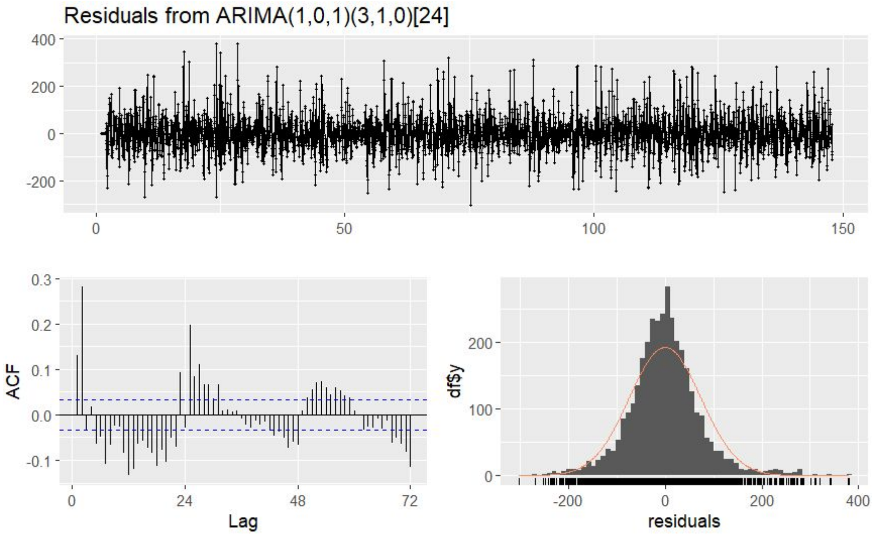


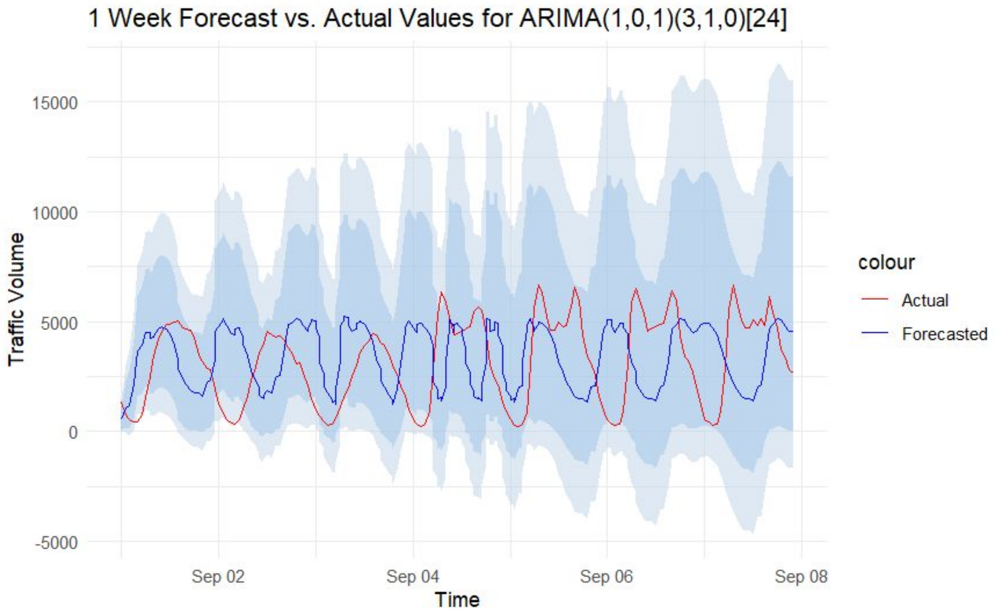
"MAE: 2115.84523809524"
"RMSE: 2512.85841319441"

Model 2 - Seasonal Arima

1 Week Forecast - ARIMA(1,0,1)(3,1,0)

```
AIC=40128.32   AICc=40128.34   BIC=40165.28
```





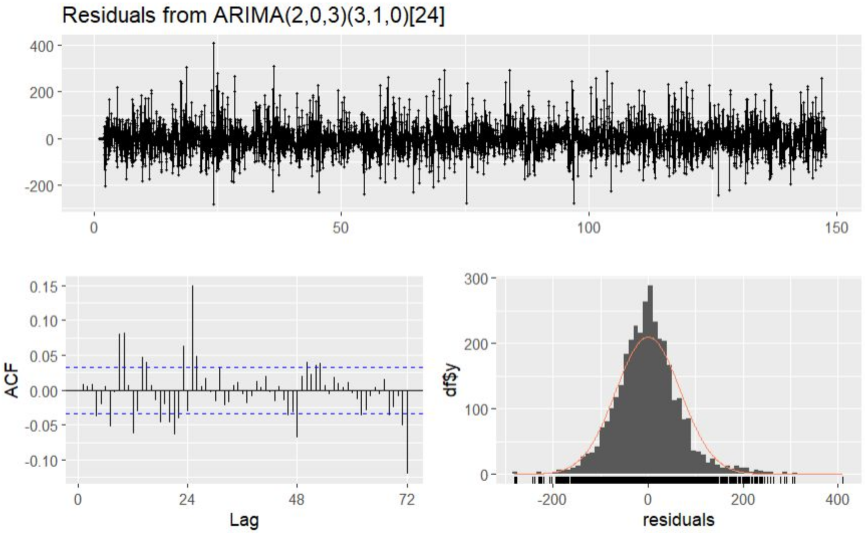
Ljung-Box test:
p-value = 2.2e-16 < 0.05

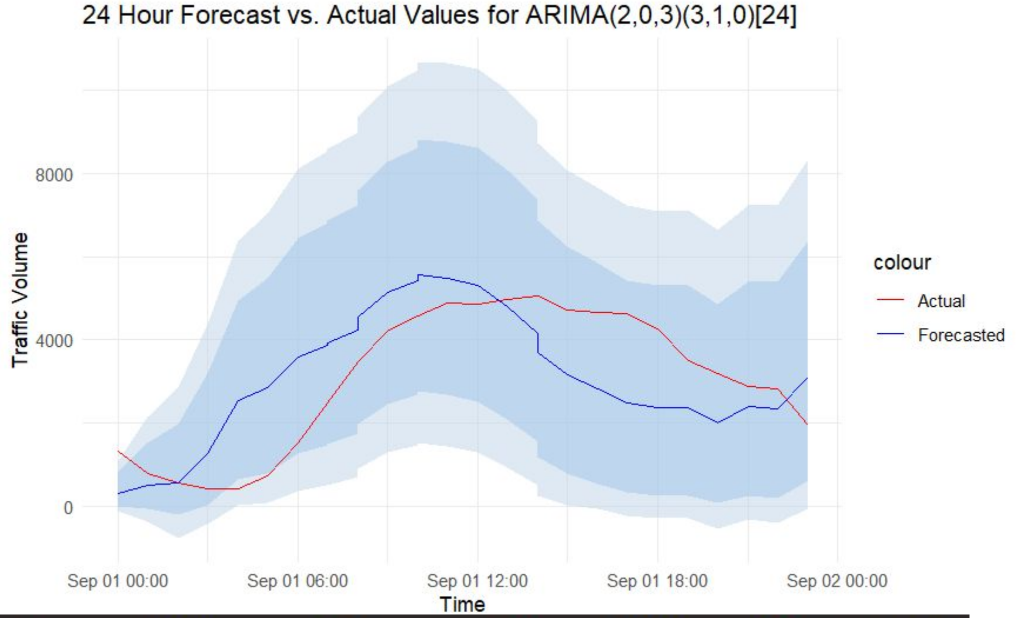
	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	29.0101	656.0169	467.4803	-11.6703	26.1889	0.9682	0.0847
Test set	-32.7696	1879.4608	1437.5892	-106.9043	135.8468	2.9775	NA

Model 2 - Seasonal Arima

24 hours Forecast - ARIMA(2,0,3)(3,1,0)

AIC=39547.55 AICc=39547.61 BIC=39603





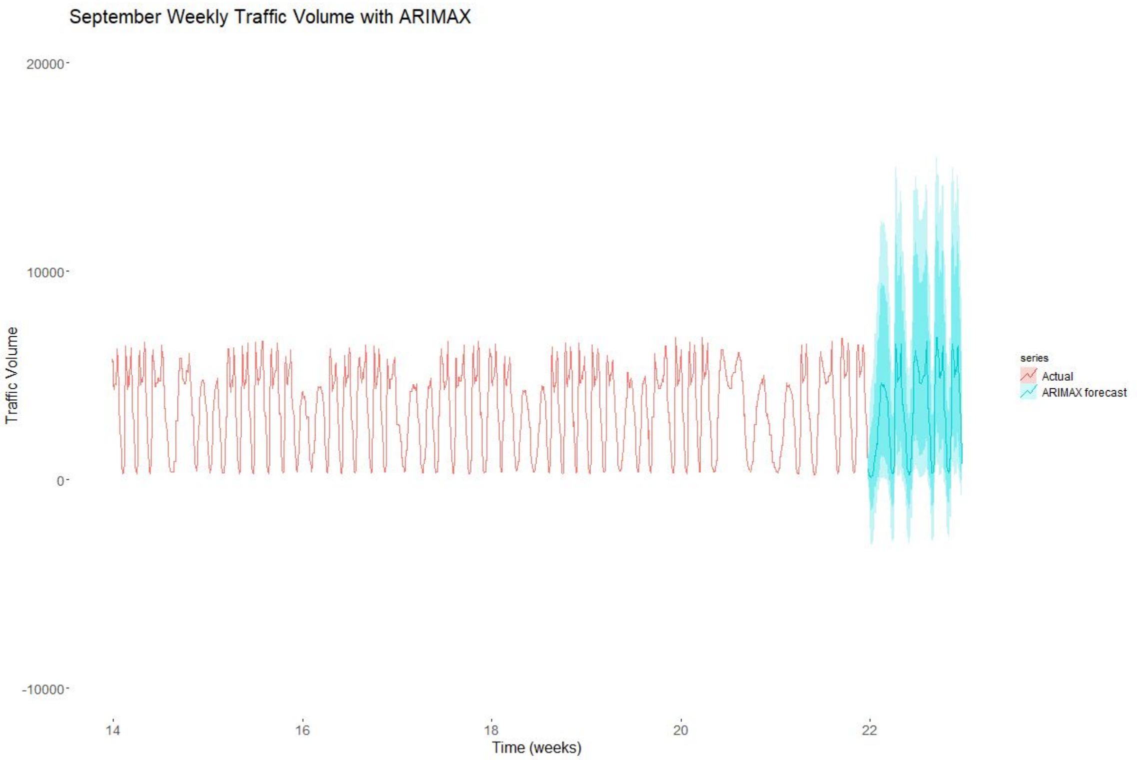
Ljung-Box test:
p-value = 2.2e-16 < 0.05

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	21.0092	612.9287	446.3463	-8.7821	23.3328	0.9245	-0.023
Test set	-233.1865	587.2364	452.2778	-13.1336	21.8350	0.9368	NA

Model 3 - ARIMAX

1 Week Forecast

Weekly seasonality

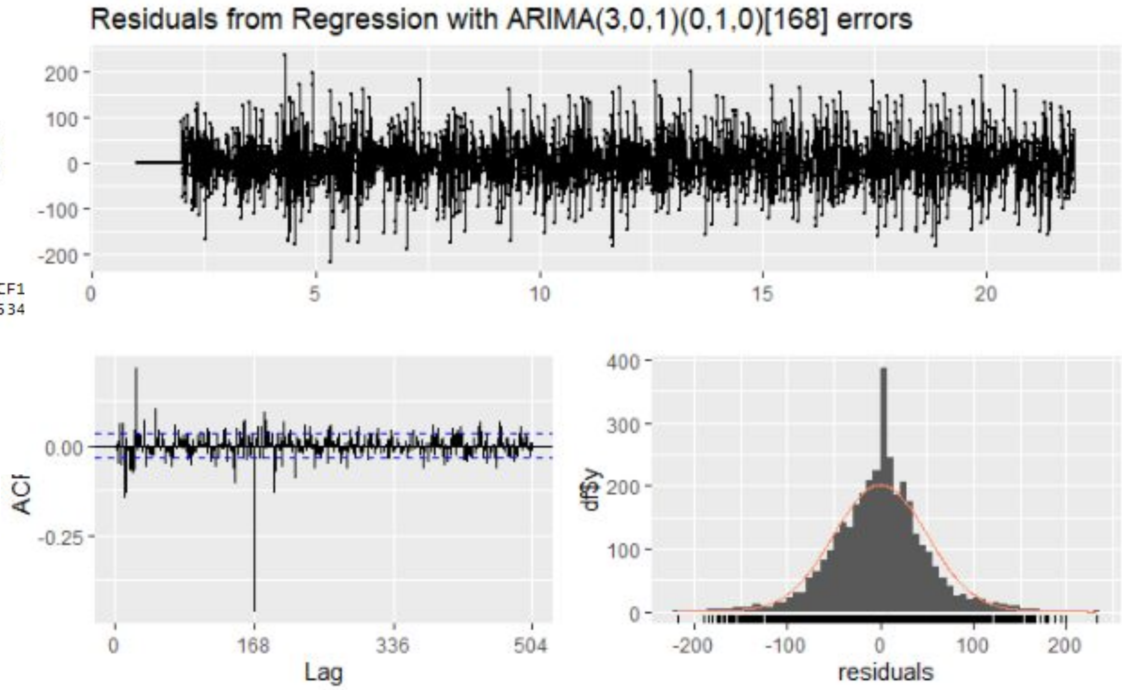


"MAE: 2126.83432032946"
"RMSE: 2509.54068129303"

Model 3 - ARIMAX

Weekly seasonality

```
Series: traffic
Regression with ARIMA(3,0,1)(0,1,0)[168] errors
Box Cox transformation: lambda= 0.6547738

Coefficients:
         ar1     ar2      ar3     ma1  temp_f  rain_1h_binary  holiday_binary
      0.7826  0.4972  -0.4687  0.6818  0.1706         -7.3828         -1.7616
s.e.  0.1095  0.1577   0.0617  0.1141  0.4540         15.0875         11.2910

sigma^2 = 2623:  log likelihood = -17974.31
AIC=35964.62   AICc=35964.67   BIC=36013.57

Training set error measures:
                   ME     RMSE      MAE       MPE     MAPE      MASE        ACF1
Training set 4.583371 803.8606 572.8828 -8.469858 27.83809 0.2701159 -0.01869534

        Ljung-Box test

data:  Residuals from Regression with ARIMA(3,0,1)(0,1,0)[168] errors
Q* = 2463.9, df = 332, p-value < 2.2e-16

Model df: 4.   Total lags used: 336
```



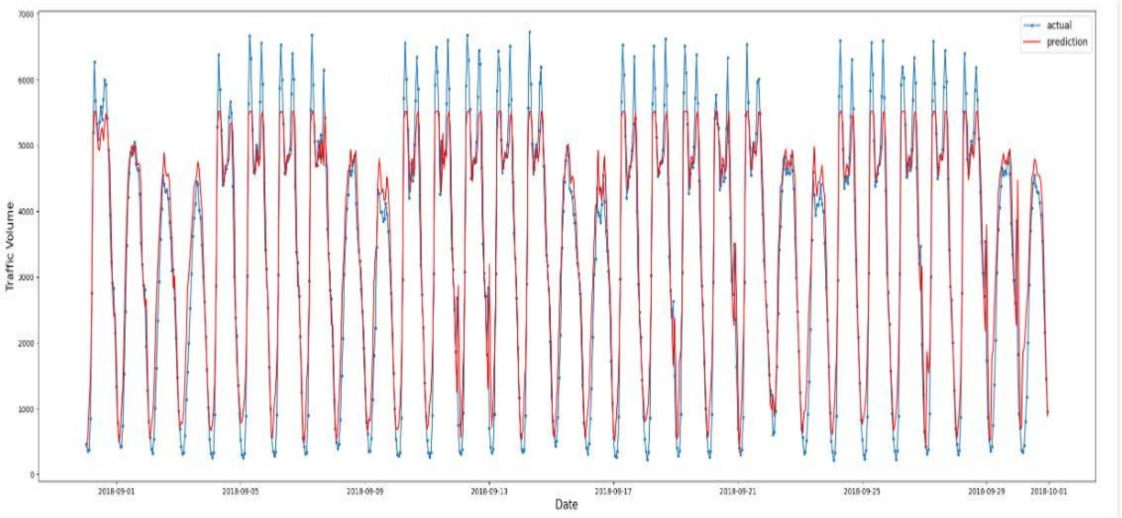
Model 4 - LSTM

1 Week Forecast

Weekly seasonality

Model - Bi-Directional LSTM
Activation - Relu
# of layers - 6

HyperParameters
Shuffle_buffer_size - for Randomness
Learning Rate
Epochs
EarlyStopping ReduceLROnPlateau
Optimizer - Adam



MAE: 436.73

RMSE: 572.48

Model Comparisons

	MAE	RMSE
Seasonal Naive	2115	2512
Seasonal Arima	1438	1879
ARIMAX	2126	2509
LSTM - Univariable	437	572
LSTM - Multivariable	687	945
ETS	5344	5697

Conclusion & Next Steps

Conclusion:

Overall, the model does a decent job in predicting the traffic for the horizon selection but the results can be improved further by experimenting more options or massaging the data into different windows.

Next Steps:

1. Forecast for Multiple horizon windows
2. Forecast on more data (beyond what we have chosen)
3. More variables or use derived/interaction variables
4. Experiment over various timing cadance for forecast - hourly/daily/weekly/monthly
5. Evaluate and fine tune deep learning models - LSTM, GRU, Bi-LSTM , Prophet

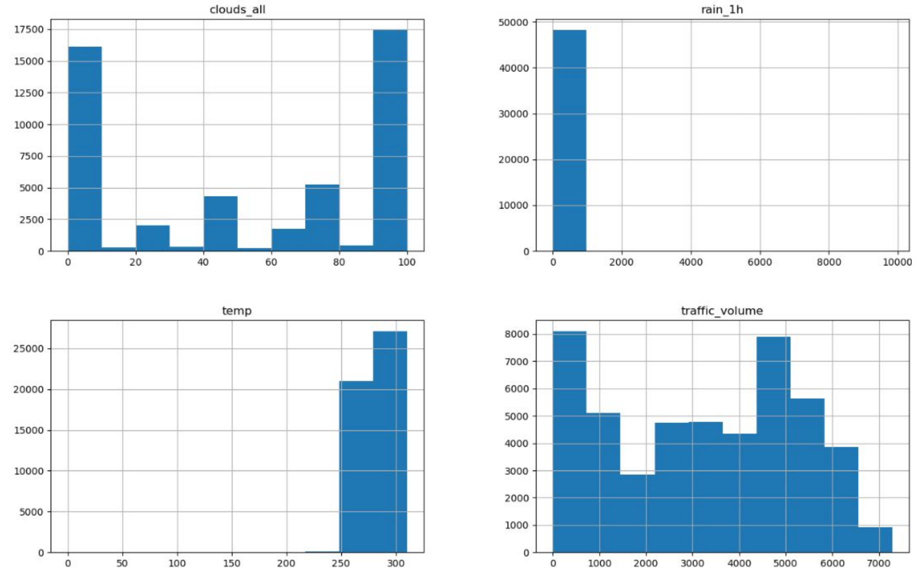
Thank You!

Appendix

Workload Distribution

	Data Search	**EDA / Feature Engineering**	**Model Development**	**Presentation**
Irem Pamuksuz		X	X	X
Vamshi Gadepally	X	X	X	X
Jose Gerala		X	X	X
Nitin Gupta	X	X	X	X
Jack Murray	X	X		X

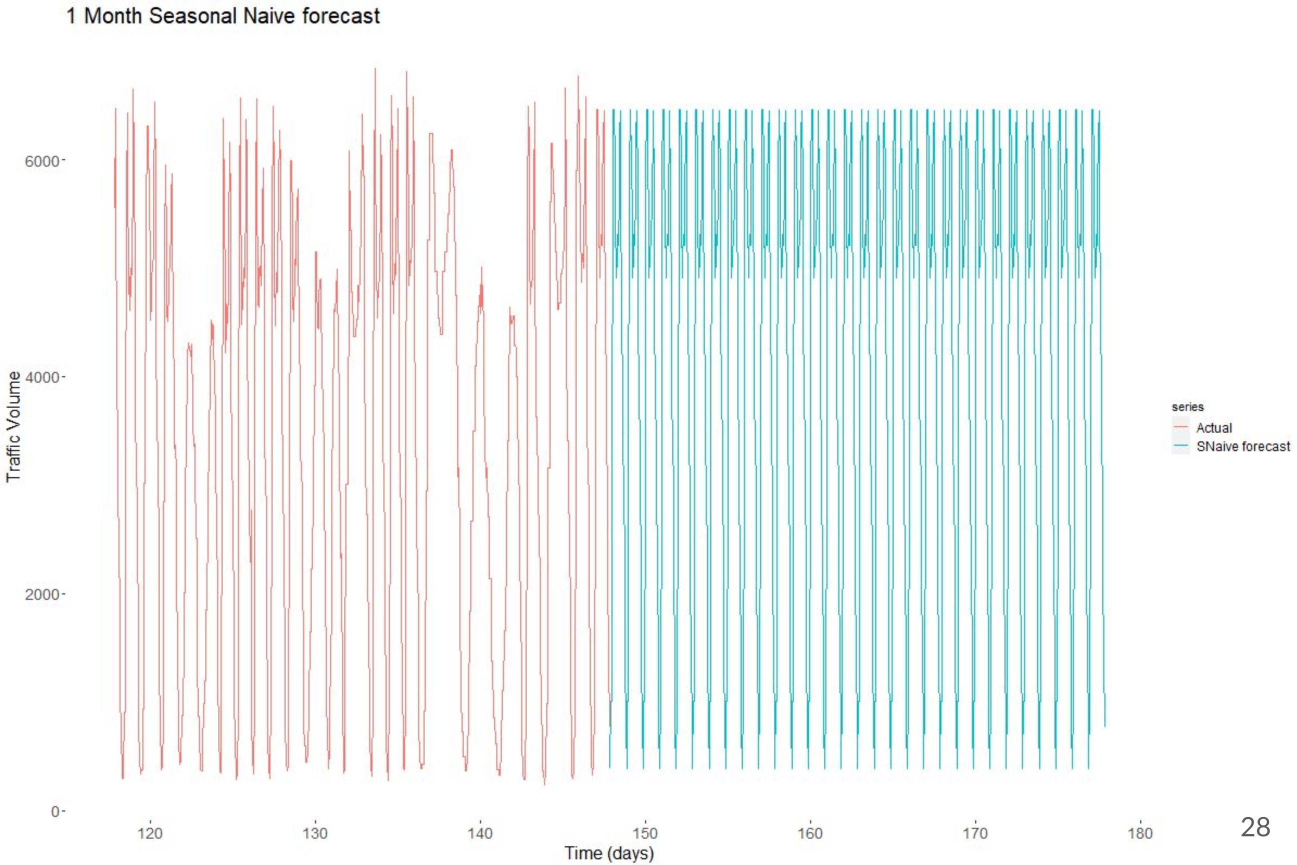
Data Exploration



Model 1 - Seasonal Naive

1 Month Forecast

Daily seasonality

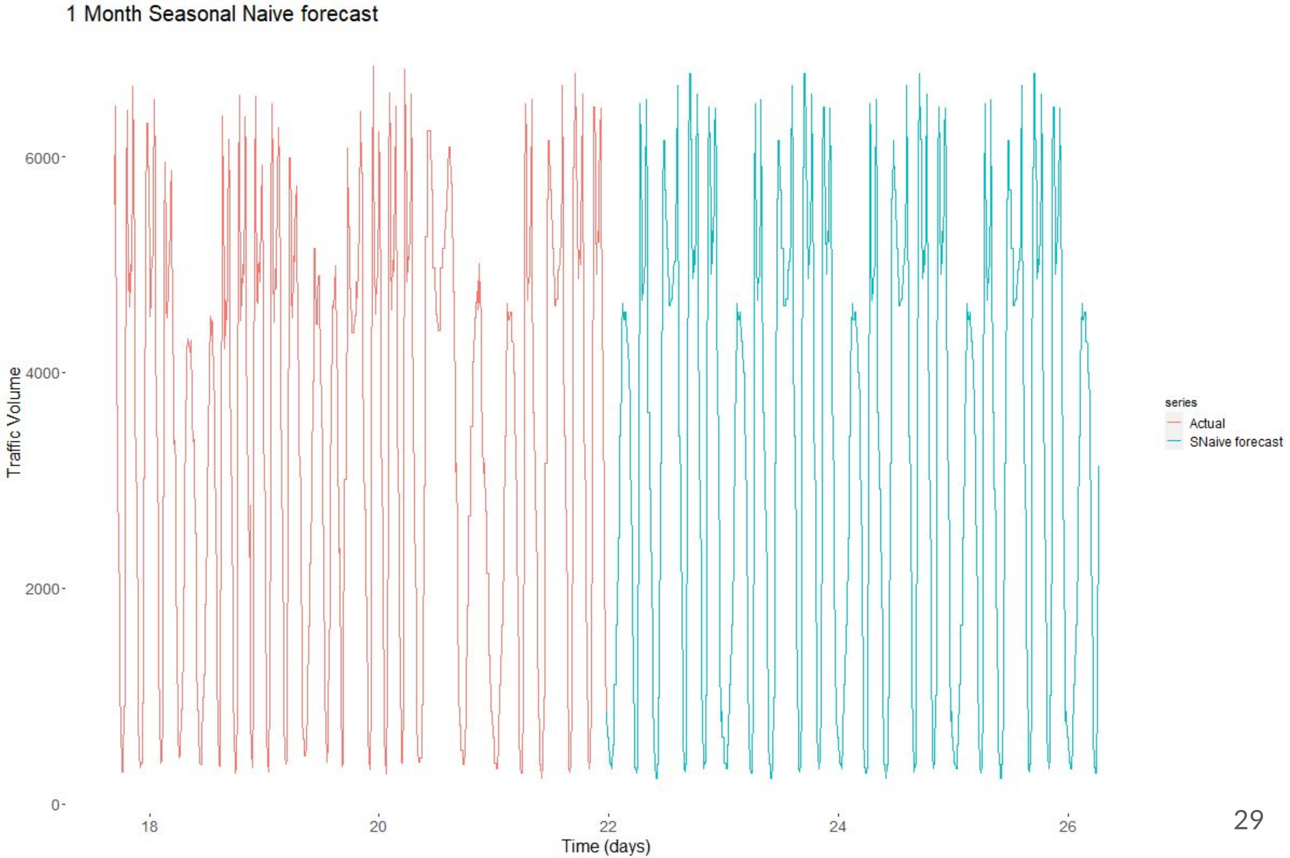


"MAE: 2722.17916666667"
"RMSE: 3213.77343524019"

Model 1 - Seasonal Naive

1 Month Forecast

Weekly seasonality

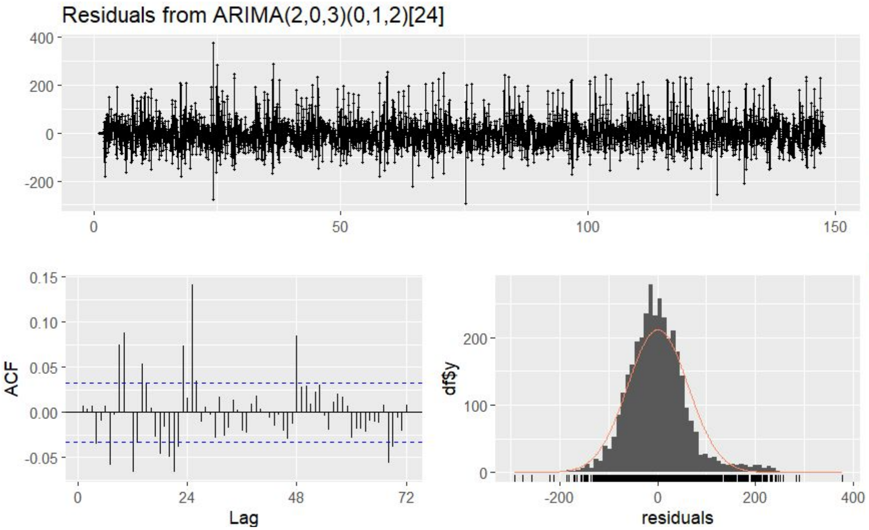


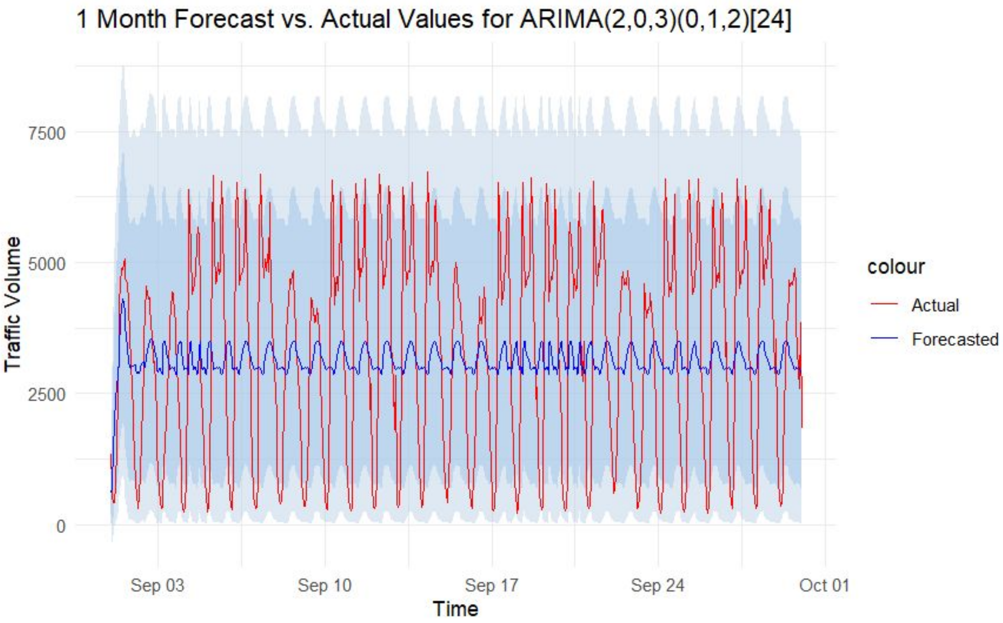
"MAE: 2257.88472222222"
"RMSE: 2777.78969561332"

Model 2 - Seasonal Arima

1 Month Forecast - ARIMA(2,0,3)(0,1,2)

```
AIC=38961.76   AICc=38961.8   BIC=39011.05
```





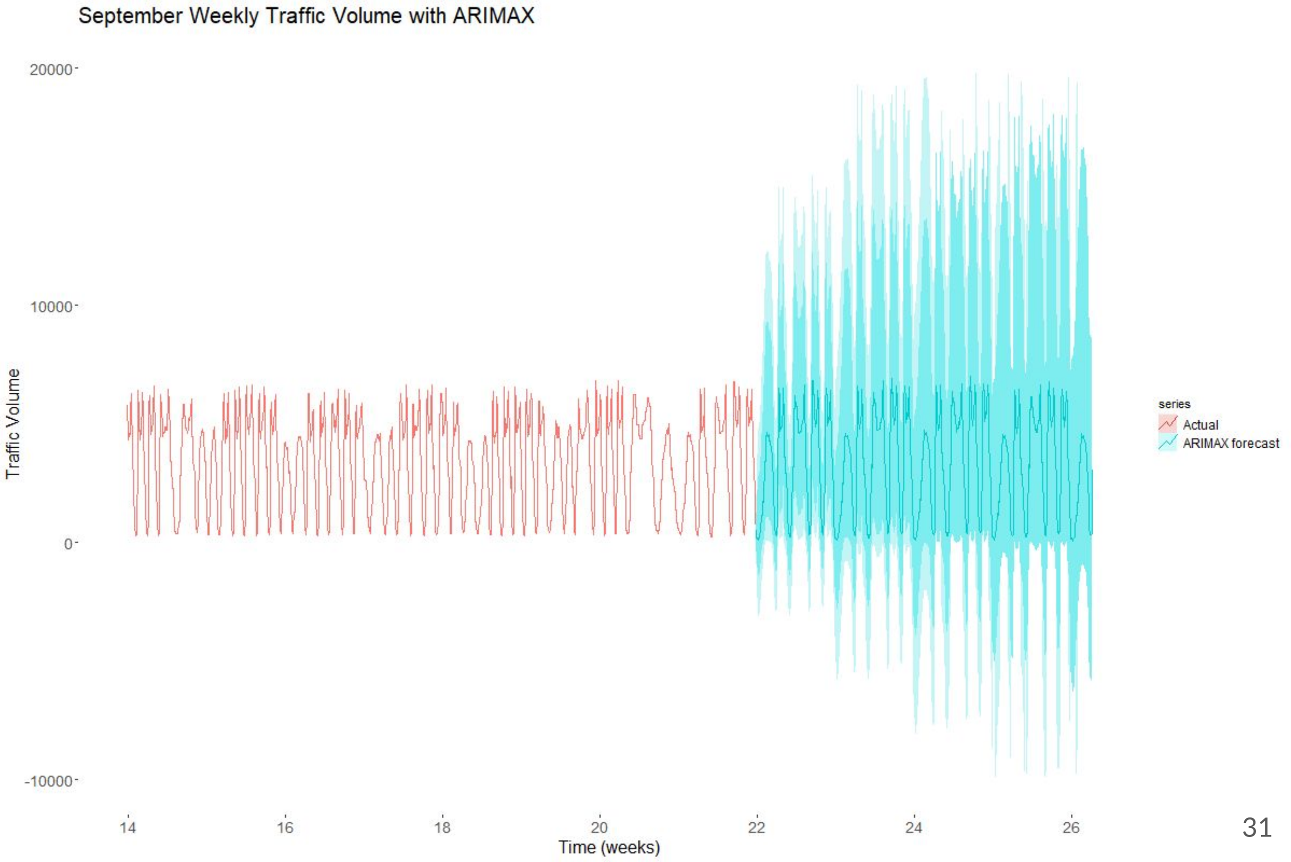
Ljung-Box test:
p-value = 2.2e-16 < 0.05

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	23.2860	564.5811	406.9974	-7.3572	19.8066	0.8430	-0.0303
Test set	179.9582	1940.3316	1684.2940	-124.7219	160.5191	3.4885	NA

Model 3 - ARIMAX

1 Month Forecast

Weekly seasonality




```
"MAE: 2284.22969966412"
"RMSE: 2796.16112478222"
```

Model 4 - LSTM Multivariate

1 Month Forecast

Weekly seasonality