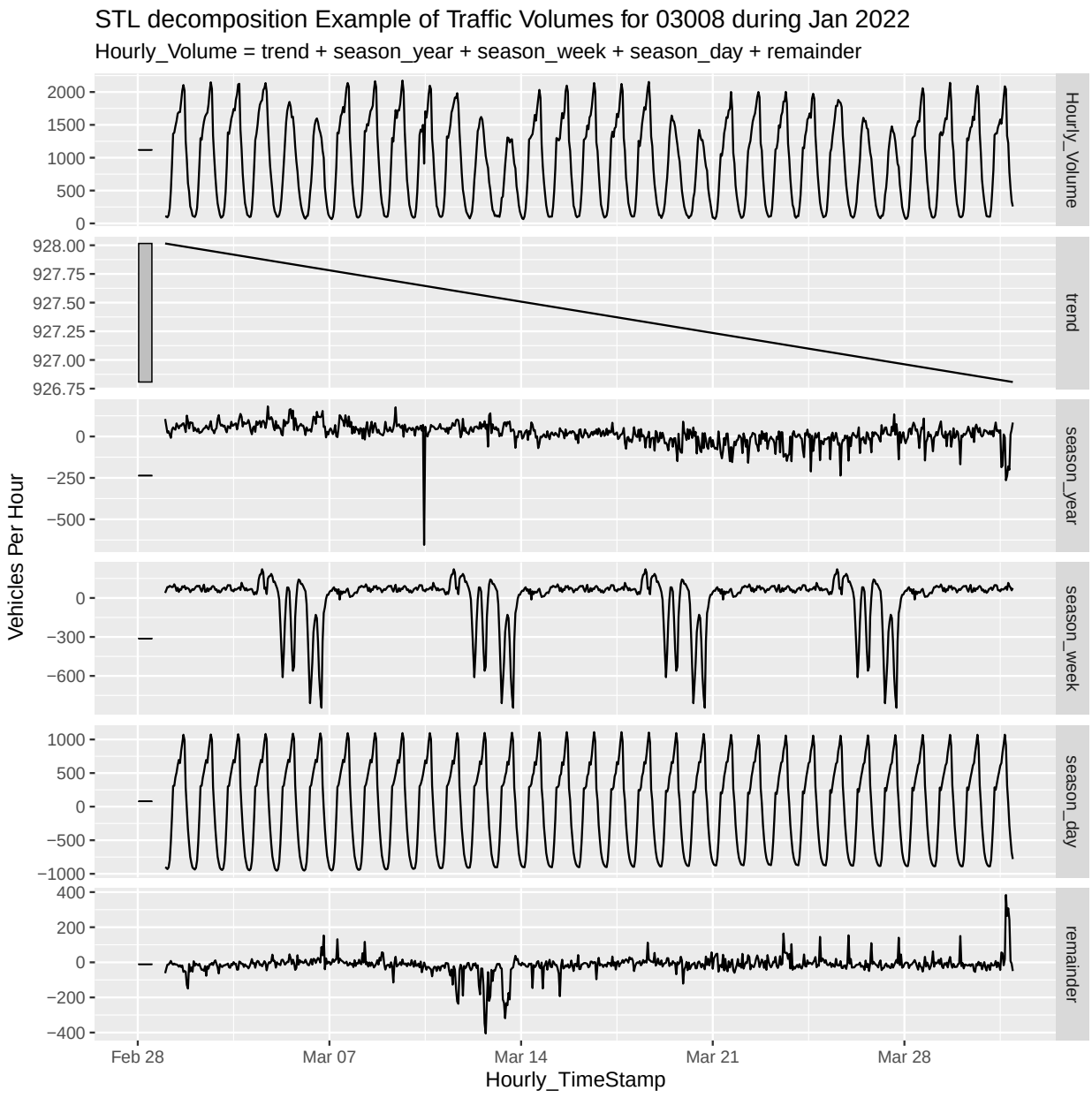
Final Project Part 2: Predicting Automobile Traffic

Sreeti Ravi, Shawn Strasser

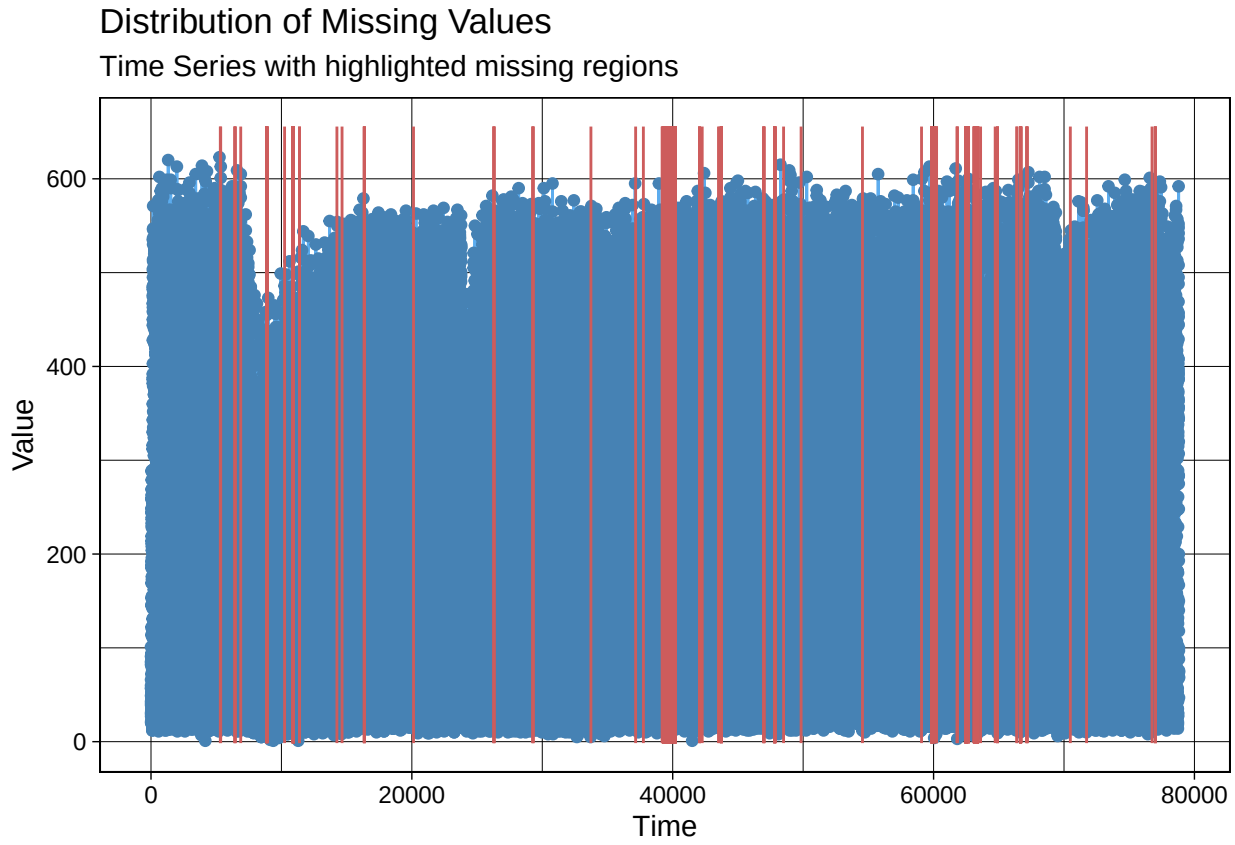
3/26/2022

1. Time Series Decomposition

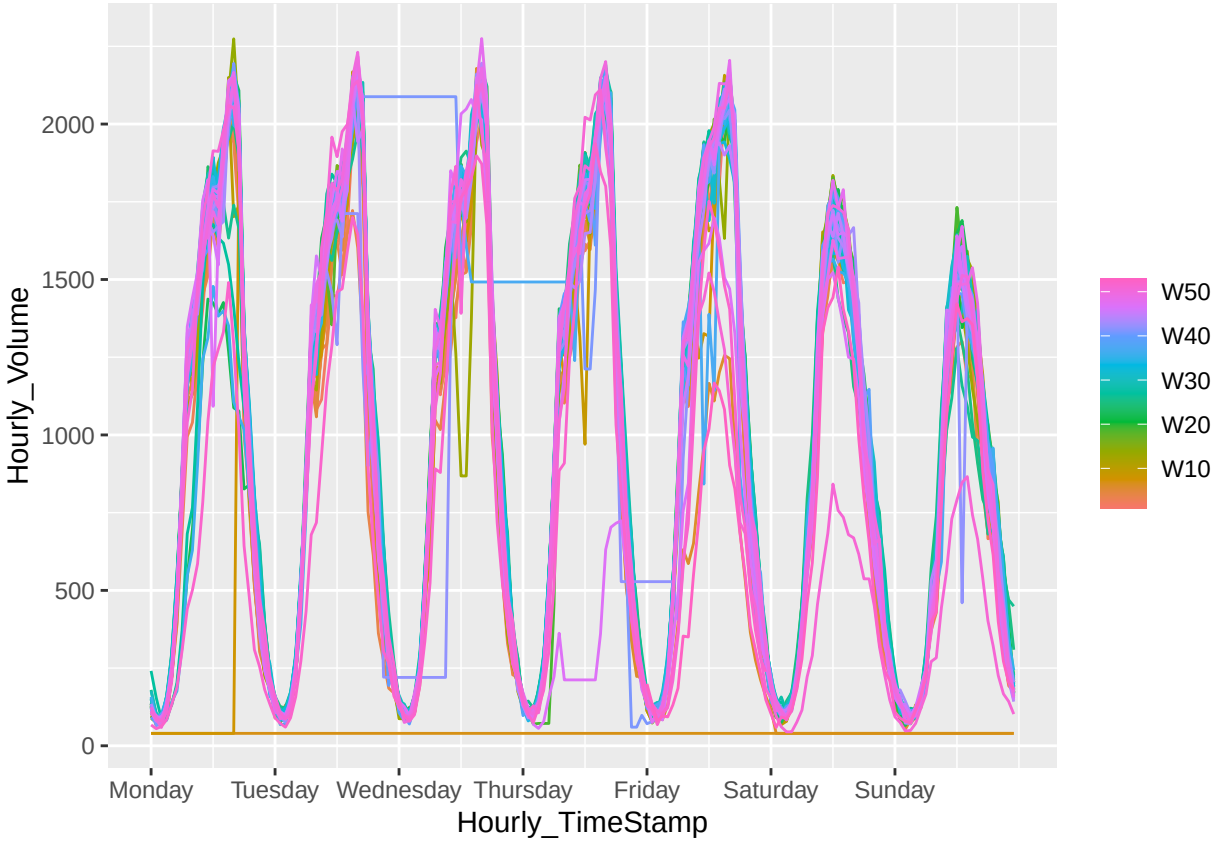


2. Time Series Visualization

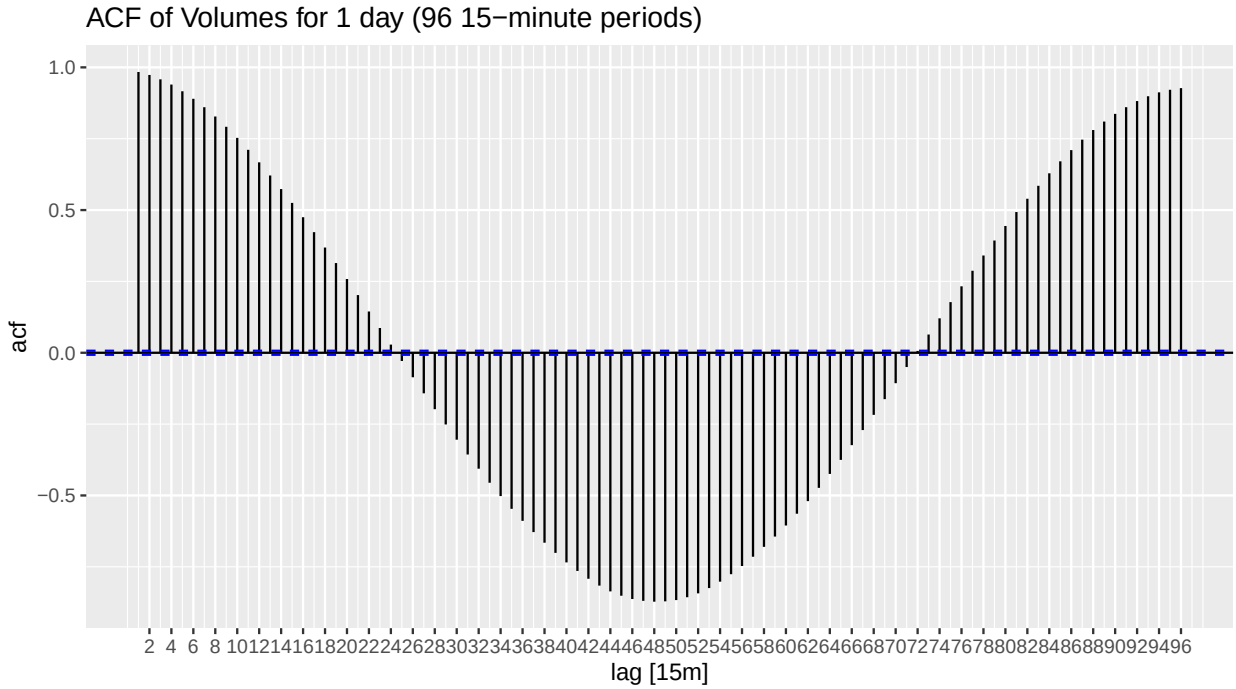
This plot reveals some gaps in the data.

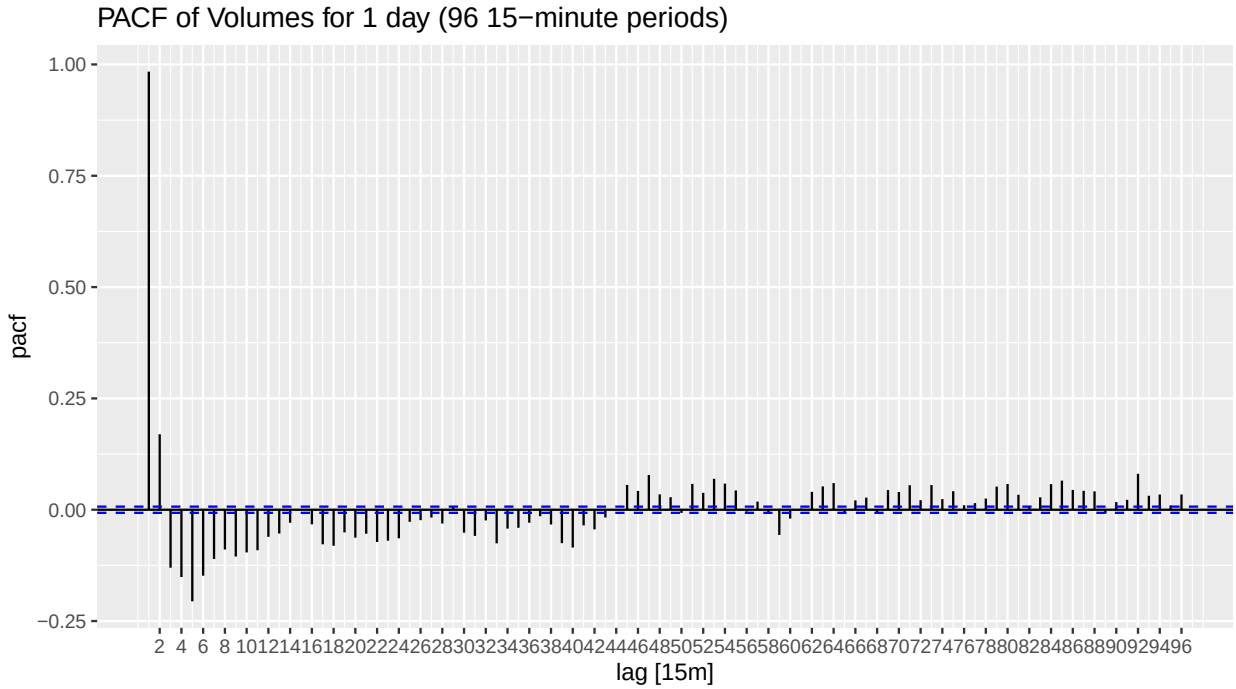


With the exception of holidays and unexpected events, the traffic at this locations is pretty consistent throughout the year. This visual also makes the gaps in data become more apparent.



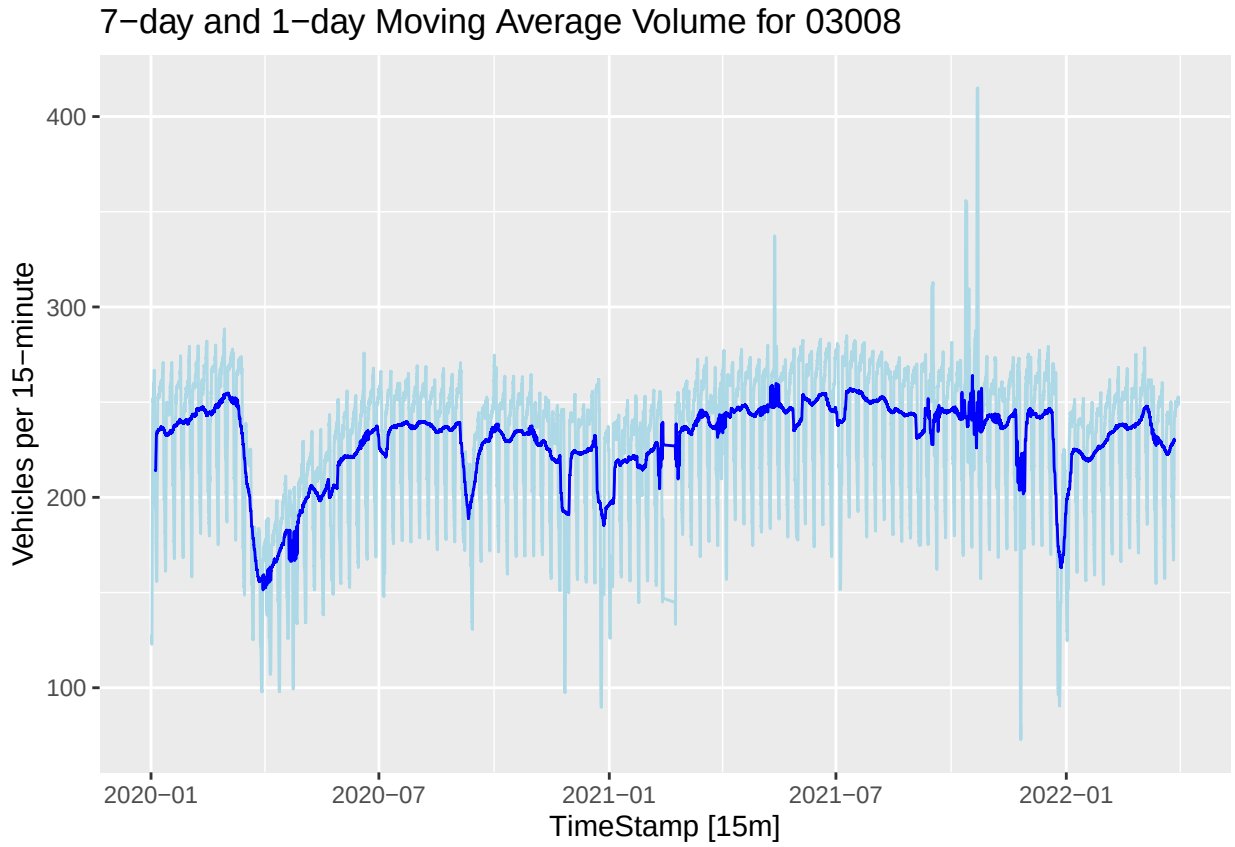
ACF plots reveal strong and persistent autocorrelation over days and weeks





3. Description of Time Series

This plot gives an overview of how traffic volumes change seasonally at a particular location in Salem, OR.



Traffic went down during COVID lockdowns March 2020, and again dropped during a wildfire in September 2020. There are regular dips during holidays in the winter and summer, with occasional unexpected dips due to snow on the road. There is not a strong long term trend present for this location, but there will likely be a trend at other locations.

There is strong seasonality at the levels of day, week, and year. There tends to be more traffic on average during the summer than winter. Weekday traffic is higher than weekend traffic, and midday traffic is higher than night time.

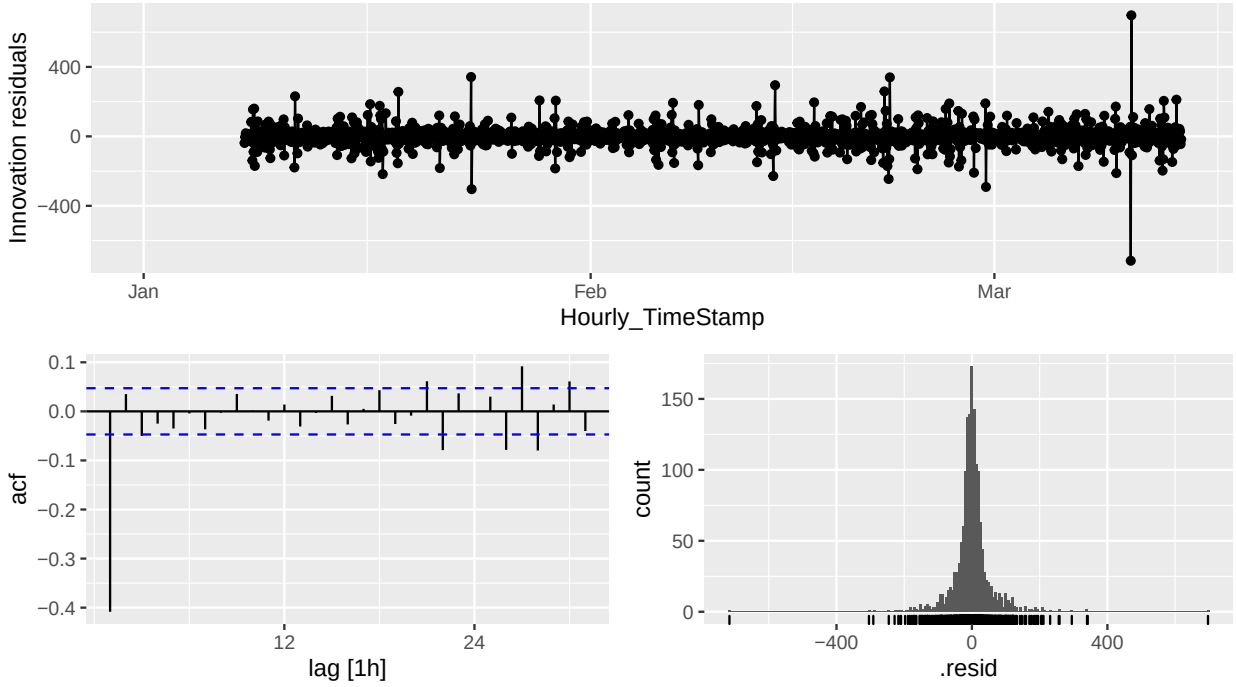
This time series is not stationary and will require some differencing to become stationary.

4. Model – apply several models (transformations if needed), explain decisions

Forecast using STL Decomposition

This data includes complex seasonality which STL handles automatically, but will be more difficult for other types of models. For that reason, I'll start with STL.

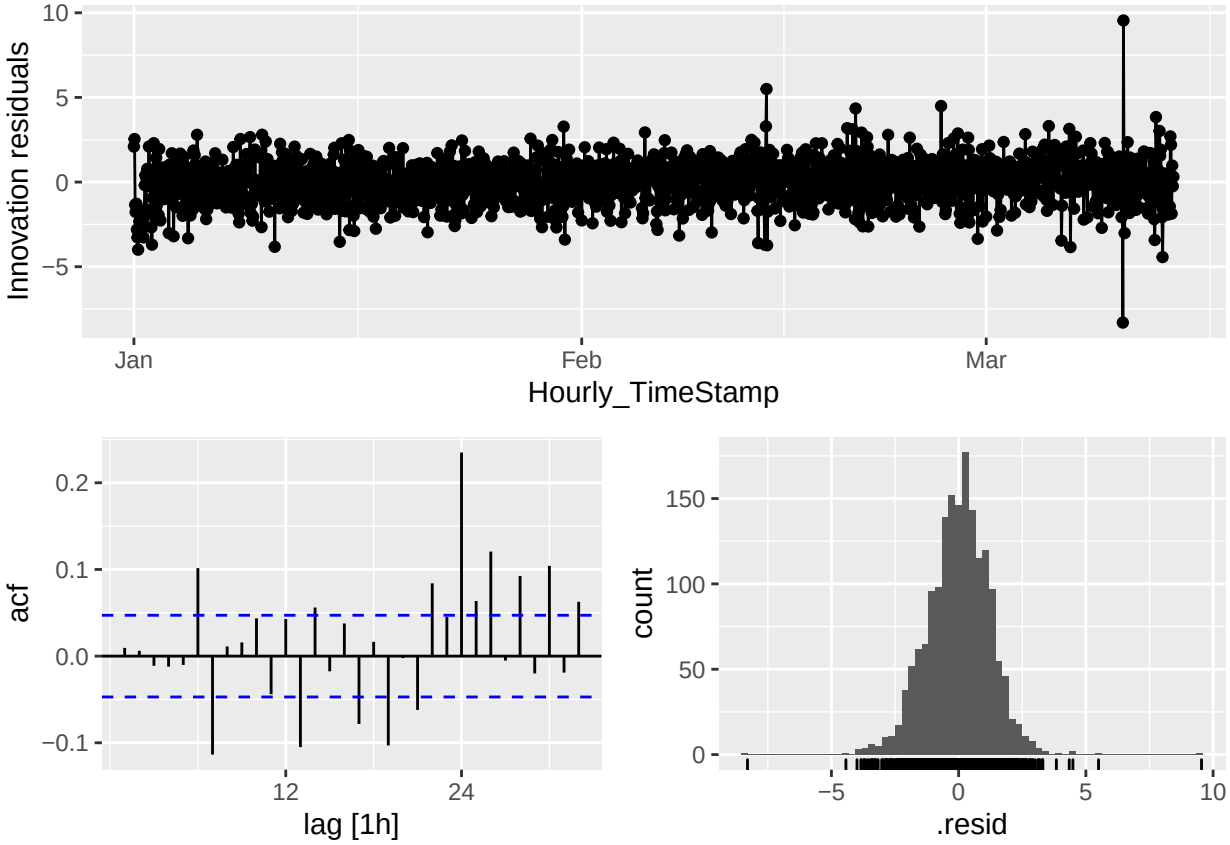
Below, the residuals are shown for a naive STL model with a trend window over the past week, and it does a good job of capturing the seasonality. The ACF of residuals still shows a significant correlation at lag 1, so this model could be improved perhaps with some ARIMA terms, but overall this is pretty good.



Forecast using Dynamic Harmonic Regression

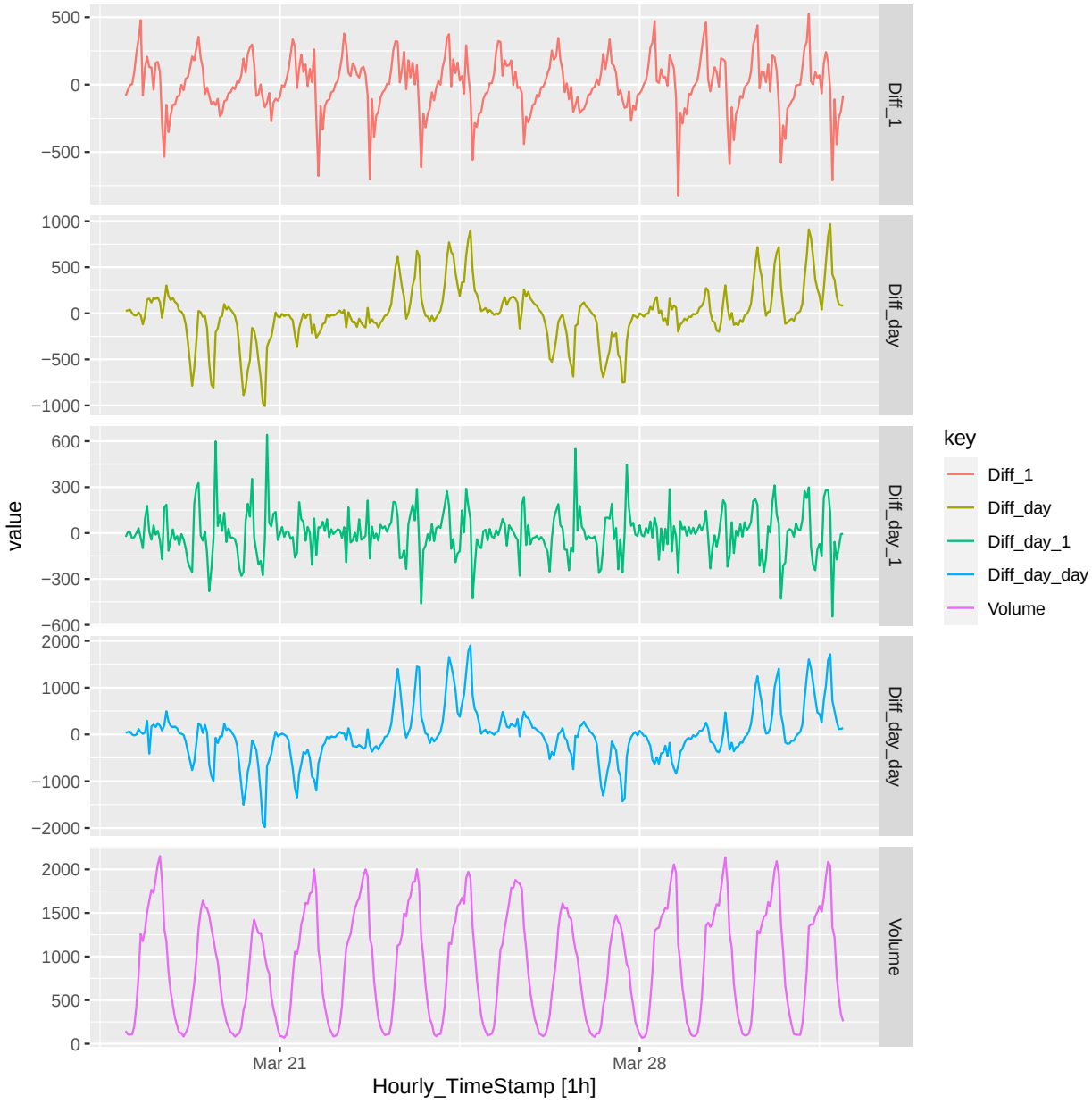
This is a dynamic harmonic regression model fit to an ARIMA error structure, with 10 Fourier terms for the daily seasonal period, and 5 Fourier terms for the weekly seasonal period. The square root is used to ensure predictions remain positive.

The ACF of residuals shows a significant correlation at lag 24, which is the seasonal day period. This means the model could be improved still, but it does seem pretty good.



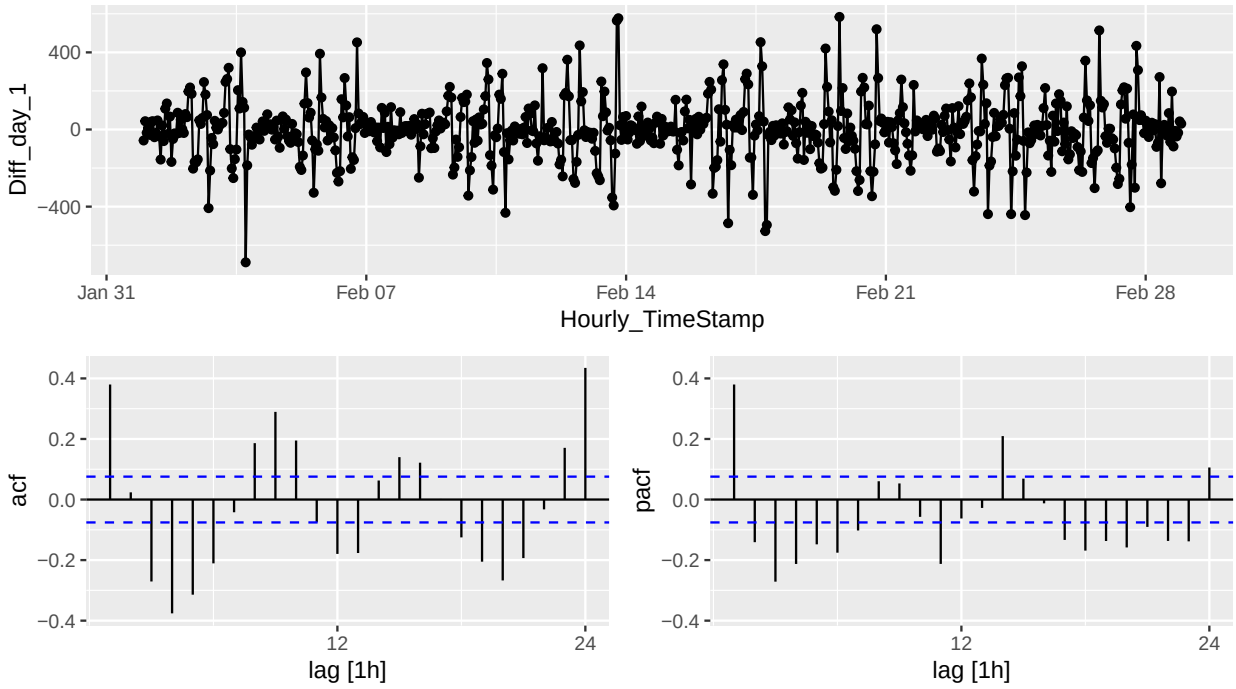
ARIMA

ARIMA has been commonly used to forecast traffic, but the complex seasonality makes setting up a model difficult. The first step is to find a differencing combination that results in stationary data.



A KPSS test shows the double difference of the day period and 1 lag creates stationary data, but from plotin this double differenced data and the ACF/PACF it's actually clear that there is still seasonality present at the weekly level. The ACF and PACF both appear sinusoidal.

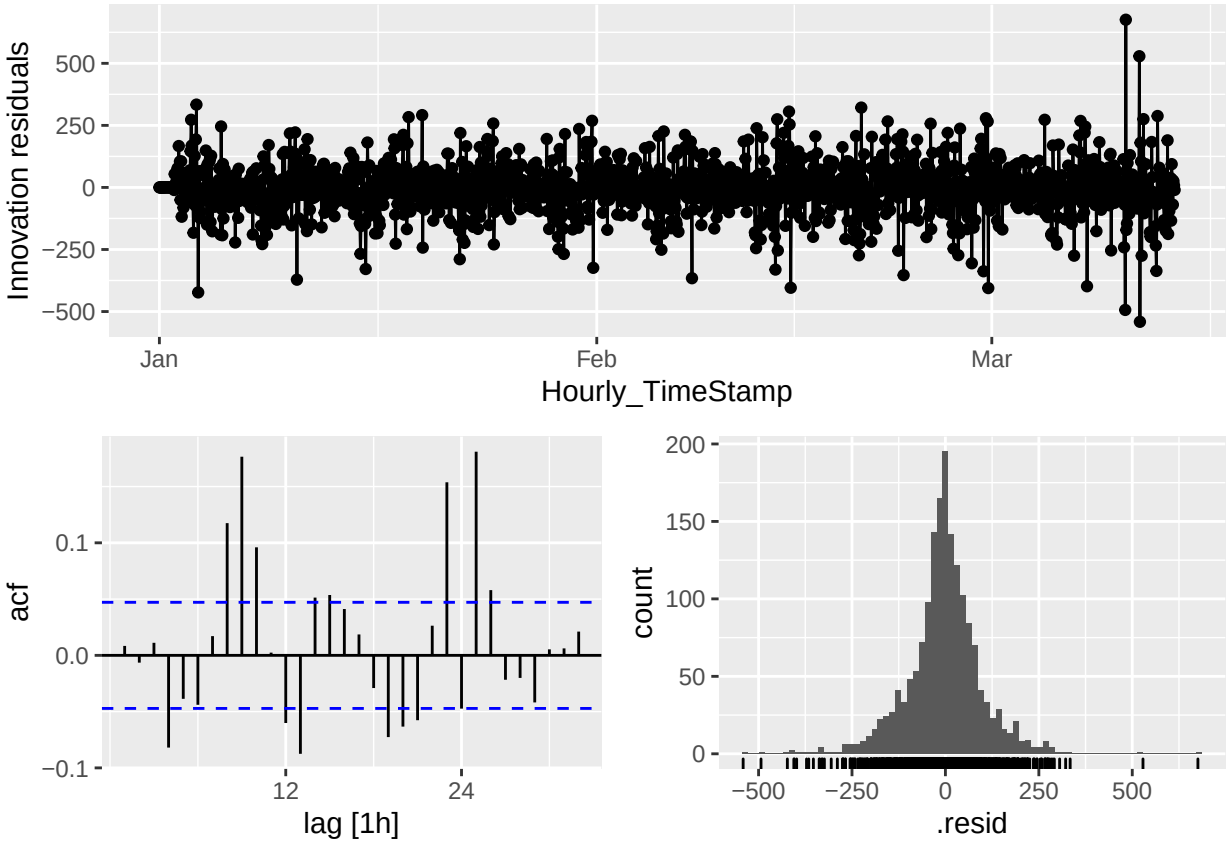
```
##    kpss_stat  kpss_pvalue
## 0.0005244991 0.1000000000
```



An ARIMA (1,1,3)(2,1,0)[24] had the lowest AIC, so it was selected for use. There is however autocorrelation in the residuals, so this model is not likley to work well.

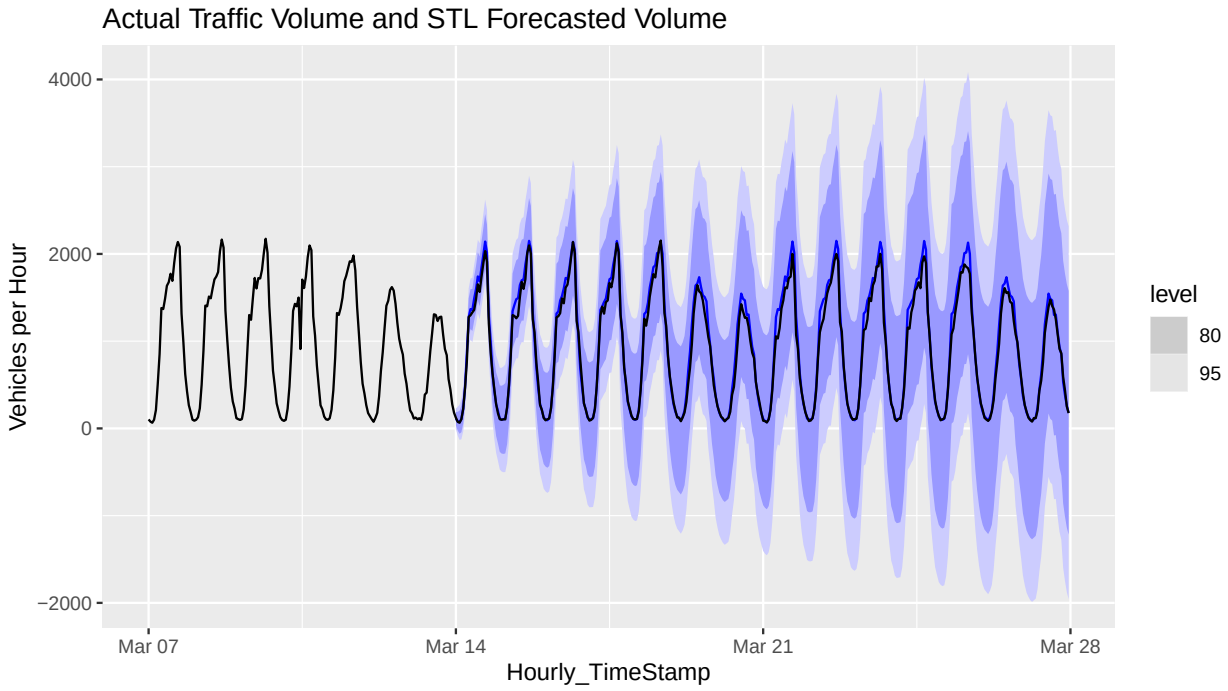
```
## # A mable: 4 x 2
## # Key:     Model name [4]
##   `Model name`                    Orders
##   <chr>                          <model>
## 1 a1           <ARIMA(1,0,3)(2,1,0)[24]>
## 2 a2           <ARIMA(1,1,3)(2,1,0)[24]>
## 3 b1           <ARIMA(4,0,0)(2,1,0)[24]>
## 4 b2           <ARIMA(4,1,0)(2,1,0)[24]>
```

```
## # A tibble: 4 x 8
##   .model sigma2 log_lik    AIC   AICc    BIC ar_roots   ma_roots
##   <chr>   <dbl>   <dbl>  <dbl>  <dbl>  <dbl> <list>     <list>
## 1 a2      9928. -10256. 20525. 20525. 20563. <cpl [49]> <cpl [3]>
## 2 b1      9885. -10256. 20526. 20526. 20564. <cpl [52]> <cpl [0]>
## 3 a1      9891. -10256. 20527. 20527. 20565. <cpl [49]> <cpl [3]>
## 4 b2     10264. -10281. 20576. 20576. 20614. <cpl [52]> <cpl [0]>
```


Innovation residuals
500
250
0
−250
−500
Jan
Feb
Mar
Hourly_TimeStamp
acf
0.1
0.0
−0.1
12
24
lag [1h]
count
200
150
100
50
0
−500
−250
0
250
500
.resid

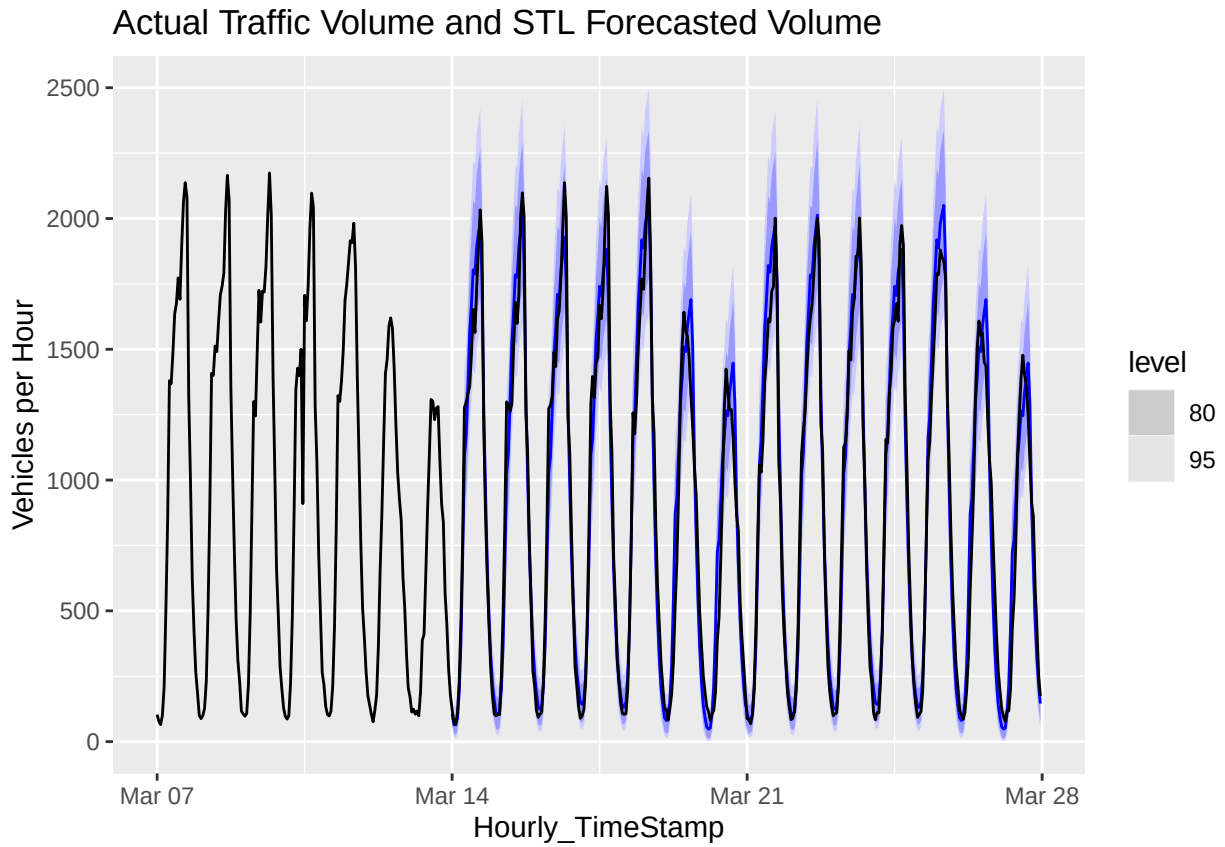
5. Prediction

Prediction using STL model



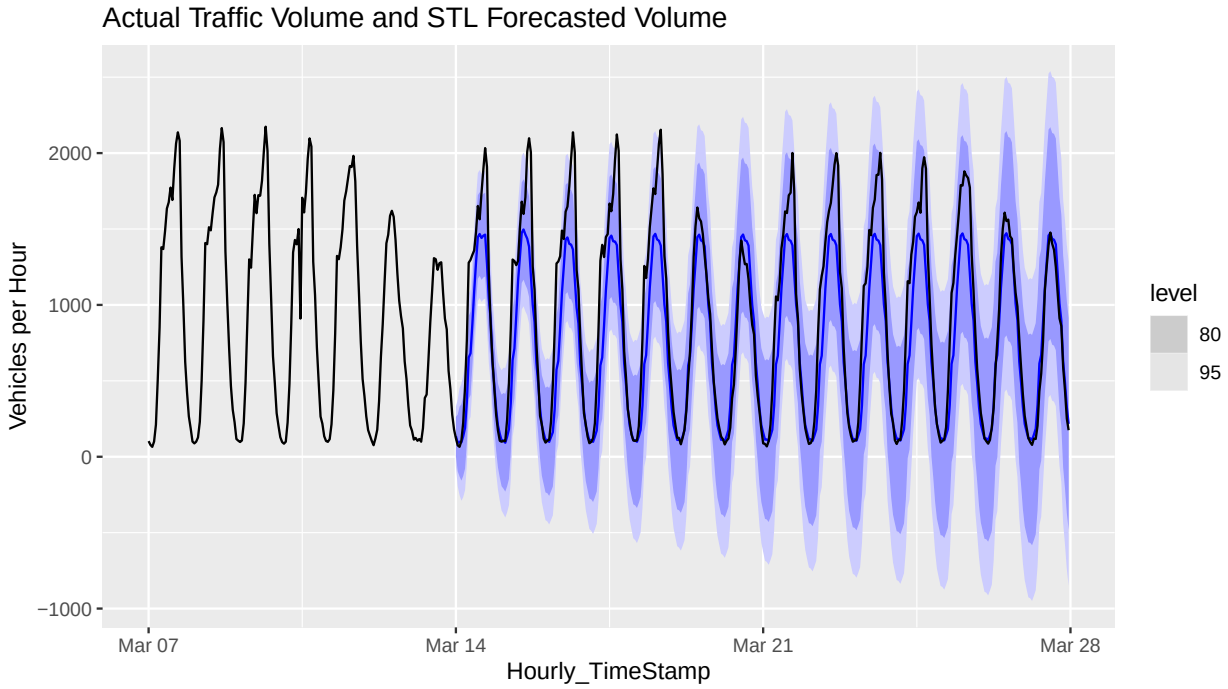

```
## # A tibble: 1 x 10
##   .model .type    ME  RMSE   MAE   MPE  MAPE  MASE RMSSE  ACF1
##   <chr>  <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 stlf   Test  -57.7  97.3  68.2 -5.33  7.85 0.600 0.481 0.625
```

Prediction using Dynamic Harmonic Regression




```
## # A tibble: 1 x 10
##   .model .type    ME  RMSE   MAE   MPE  MAPE  MASE RMSSE  ACF1
##   <chr>  <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 dhr    Test  -13.1  124.  96.4 -4.07  17.4 0.847 0.613 0.725
```

Prediction using ARIMA model




```
## # A tibble: 1 x 10
##   .model .type    ME  RMSE   MAE   MPE  MAPE  MASE RMSSE  ACF1
##   <chr>  <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 a2     Test   175.  292.  206.  11.5  22.8  1.81  1.45 0.856
```

Model Accuracy

The STL model was the most accurate, with a mean absolute percentage error of 7.8% vs 17.4% for dynamic harmonic regression, and 22.8% for ARIMA. The ARIMA model does did not work well because it does not account for the multiple seasonal periods present in the data.