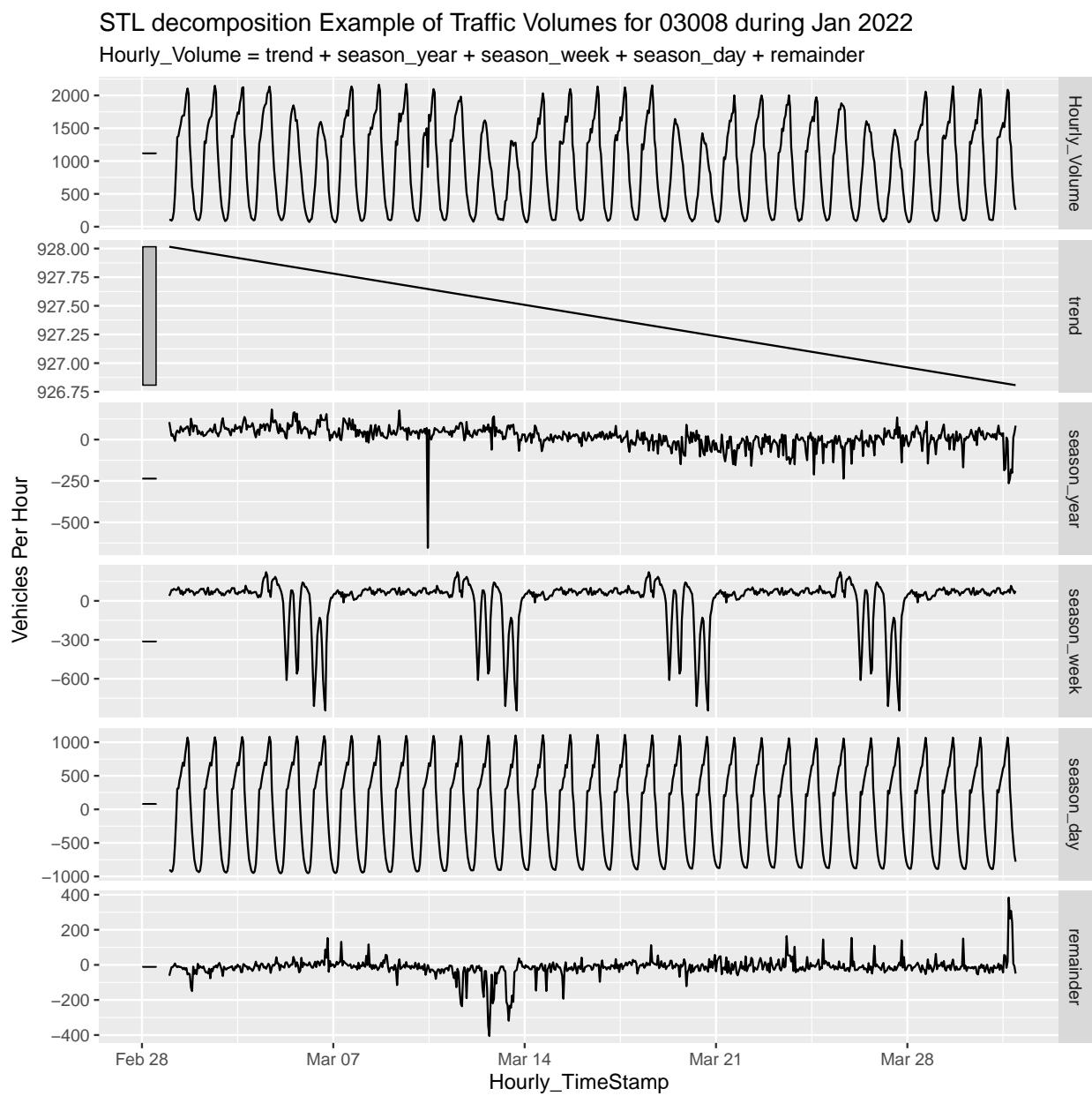


Final Project Part 2: Predicting Automobile Traffic

Sreeti Ravi, Shawn Strasser

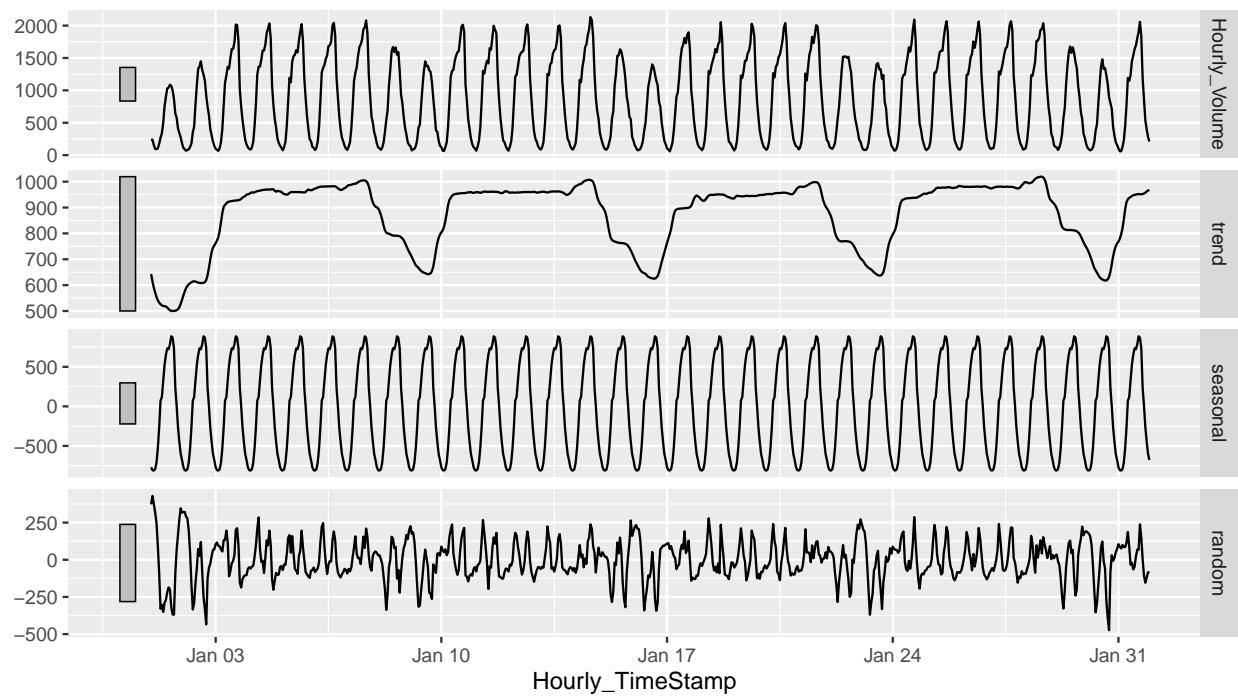
4/3/2022

1. Time Series Decomposition



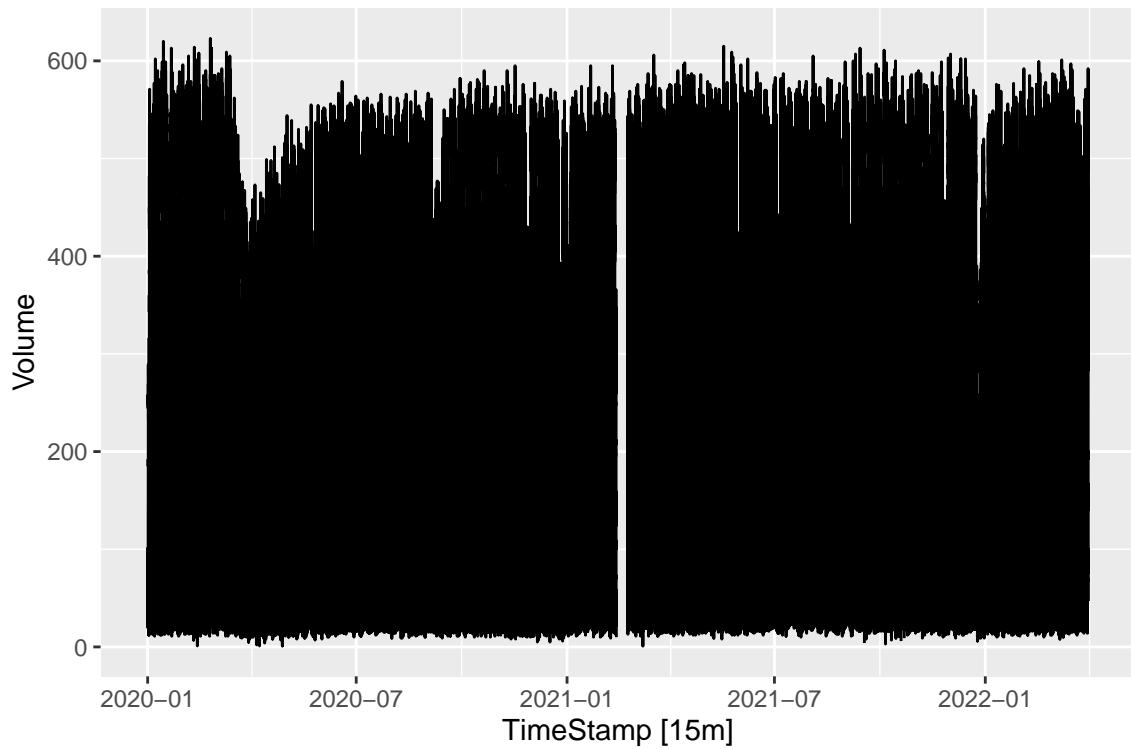
Classical Decomposition of Traffic Volumes for Jan 2022

Hourly_Volume = trend + seasonal + random



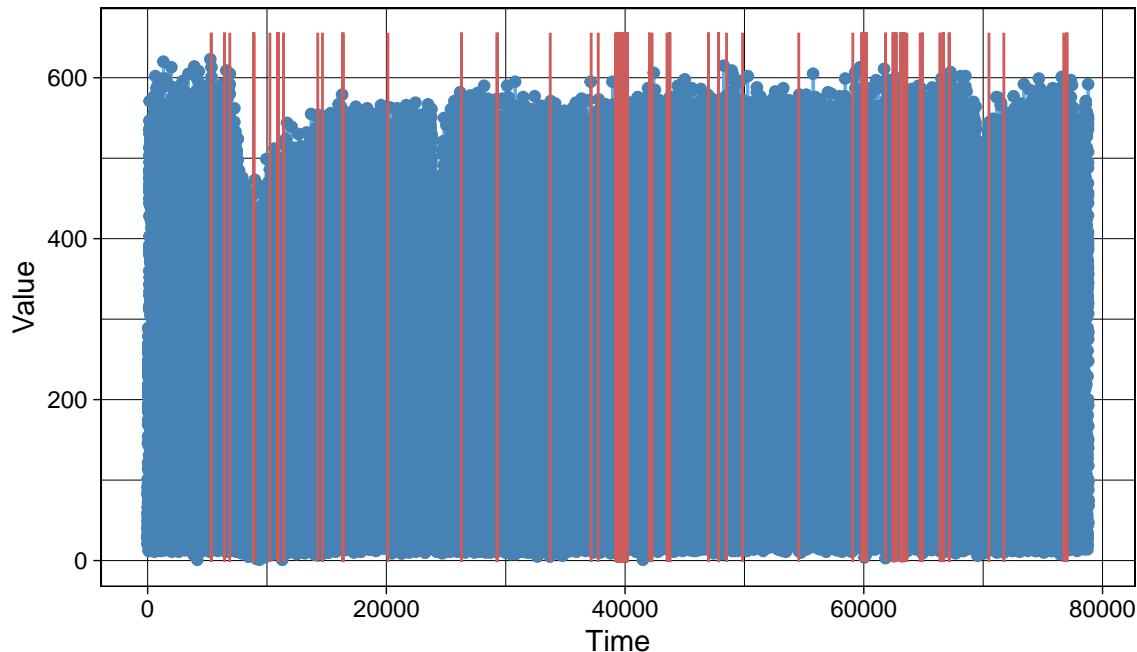
2. Time Series Visualization

Autoplot of a subset of the data reveals a large gap right in the middle.

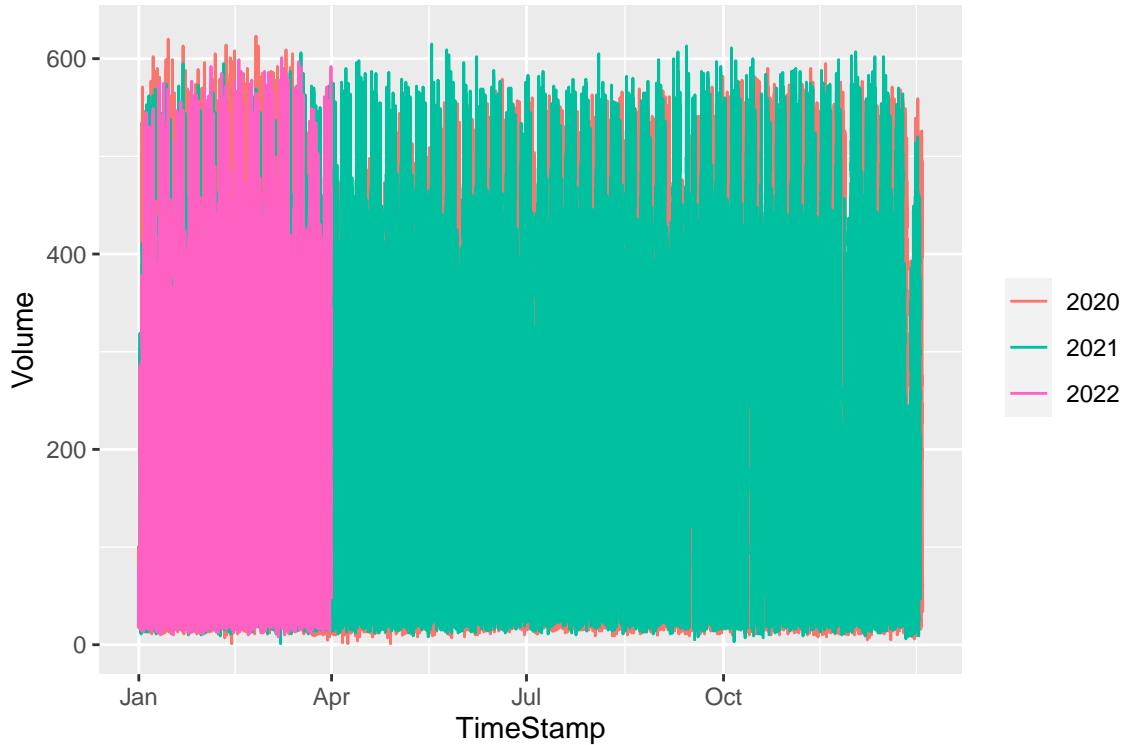


ggplot_na_distribution function shows other gaps as well, which will need to be filled in.

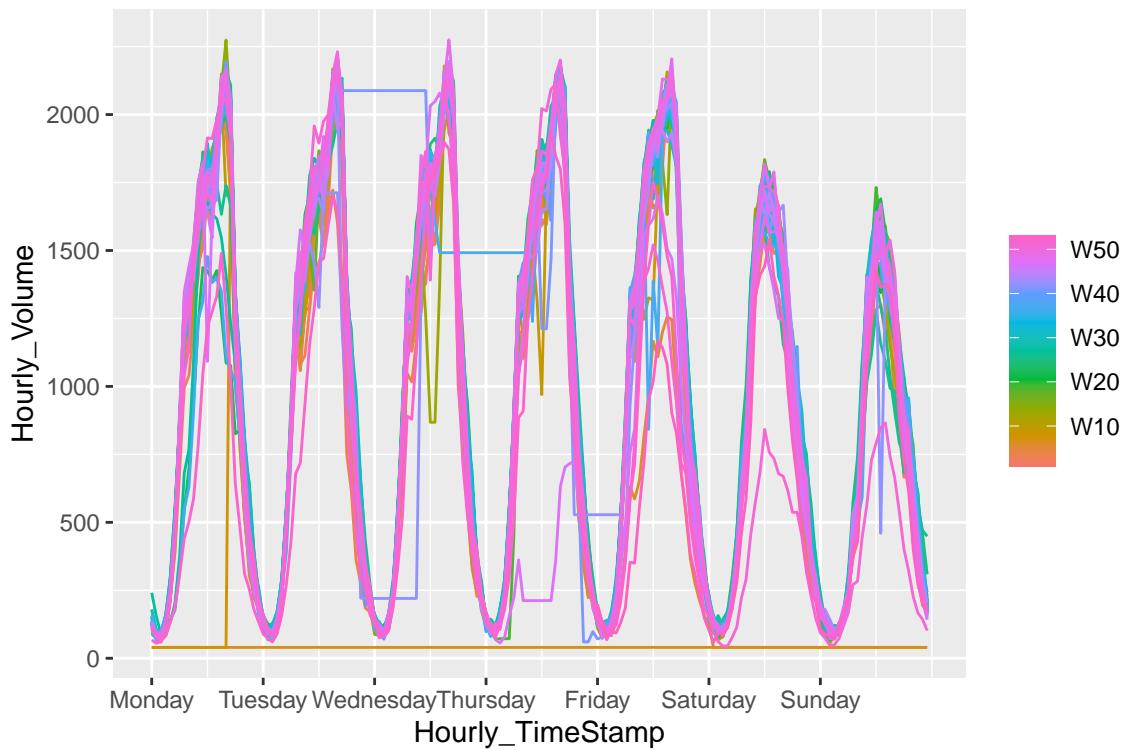
Distribution of Missing Values
Time Series with highlighted missing regions



Here, the years 2020, 2021 and 2022 are plotted together. With the exception of holidays and unexpected events, the traffic at this locations is pretty consistent throughout the year.

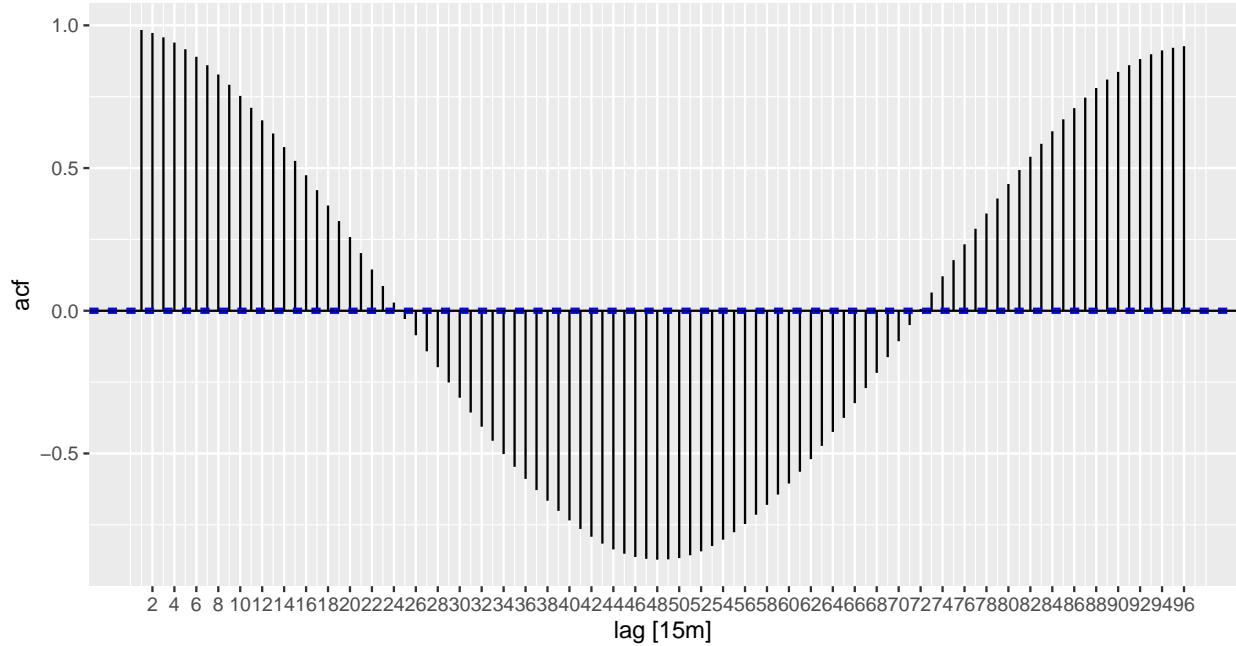


Here, each week of 2021 is plotted.

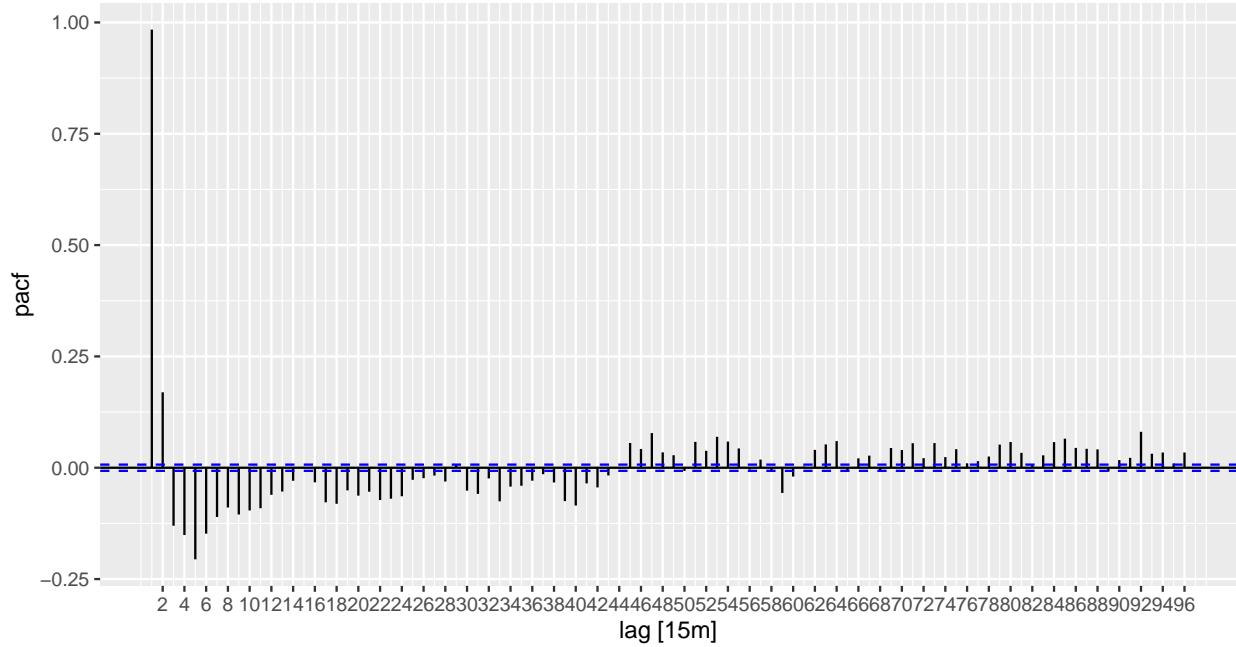


ACF and PACF plots reveal strong and persistent autocorrelation over days and weeks

ACF of Volumes for 1 day (96 15–minute periods)



PACF of Volumes for 1 day (96 15–minute periods)



3. Description of Time Series

This plot gives a good overview of how traffic volumes changed since 2020. The location where this data comes from is a downtown traffic signal in Salem, Oregon.

7-day and 1-day Moving Average Volume for 03008



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 traffic signal locations.

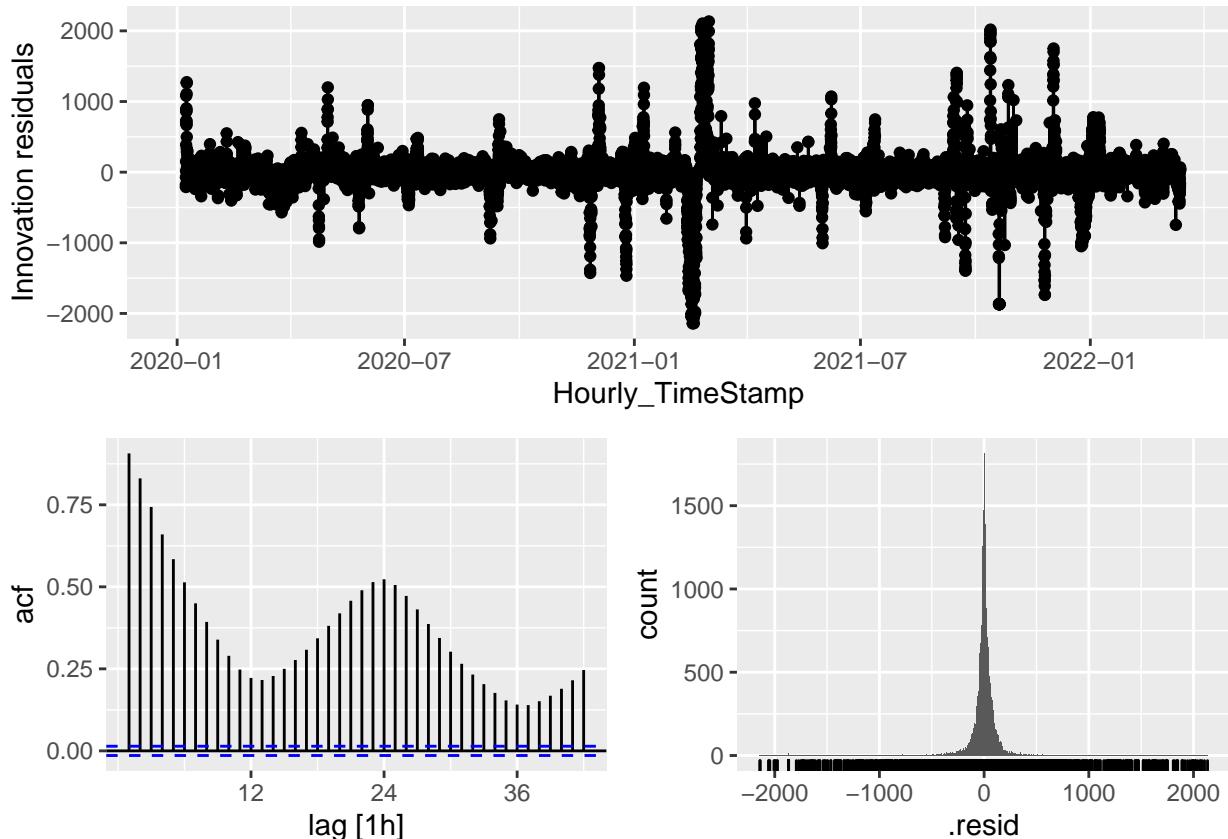
There is strong seasonality at the levels of day, week, and year. There tends to be more traffic on average during the summer than winter. Weekday traffic is higher than weekend traffic, and midday traffic is higher than night time.

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

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

Seasonal Naive model with lags at 1 day, 1 week, and 1 year

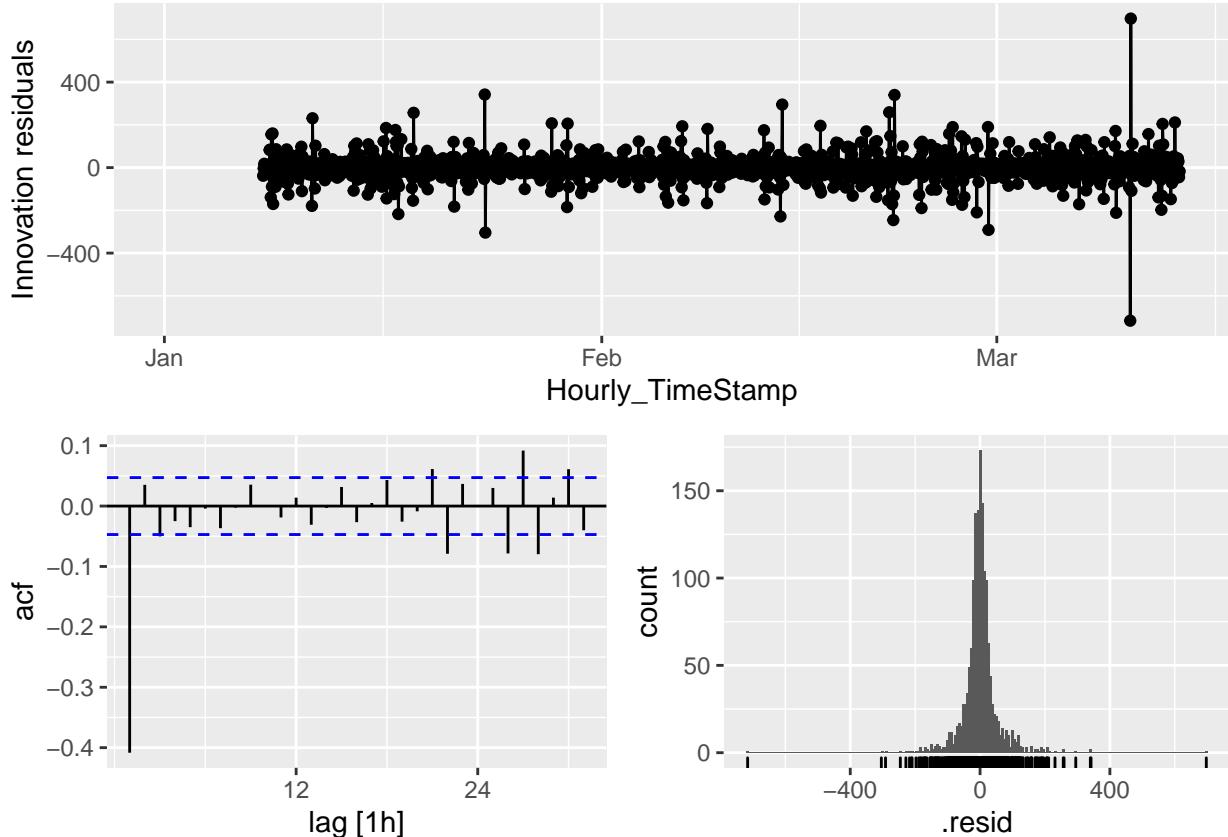
The residuals are highly correlated so this model could be improved.



STL Decomposition

STL works well because it automatically handles the complex seasonal periods present in the data, which include day, week, and year.

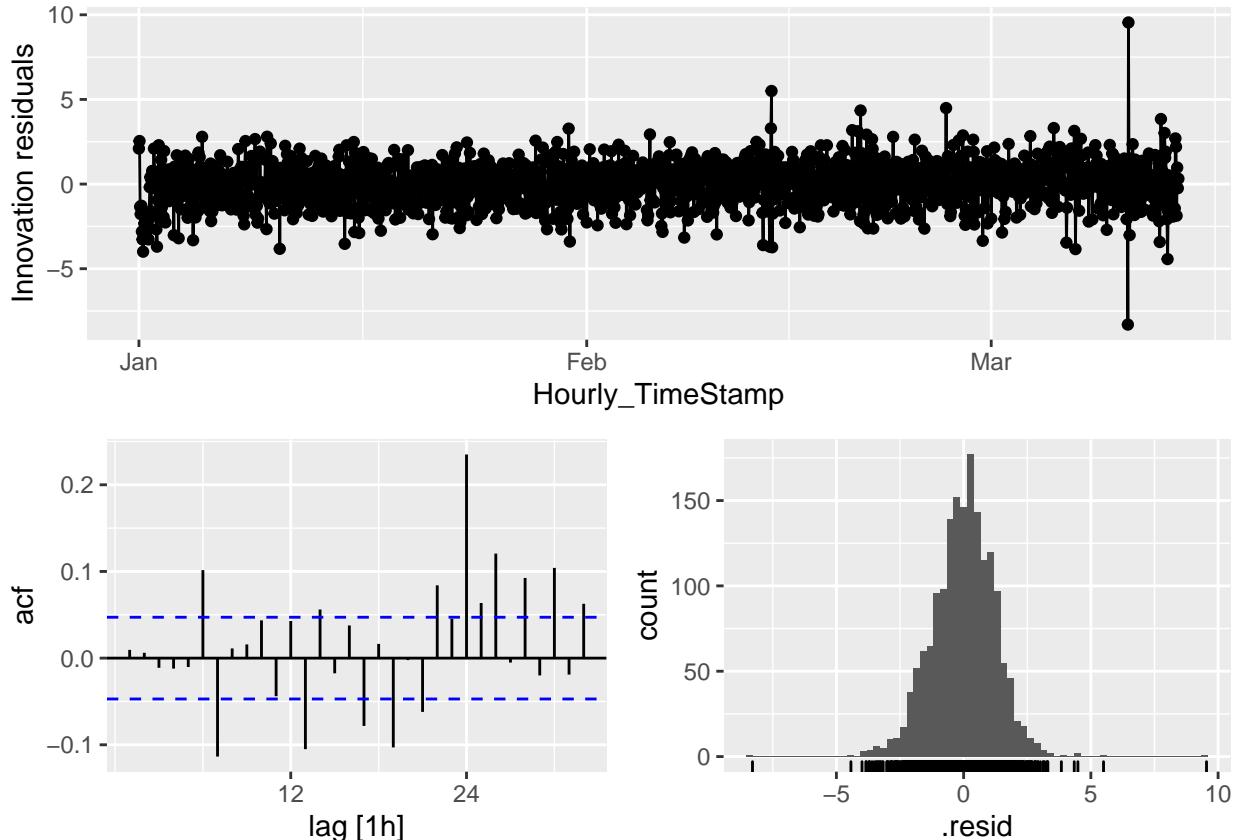
The ACF of residuals shows a significant correlation only at lag 1, so this model could be improved perhaps with some ARIMA terms, but overall this is pretty good. This model will be a good option to use.



Dynamic Harmonic Regression

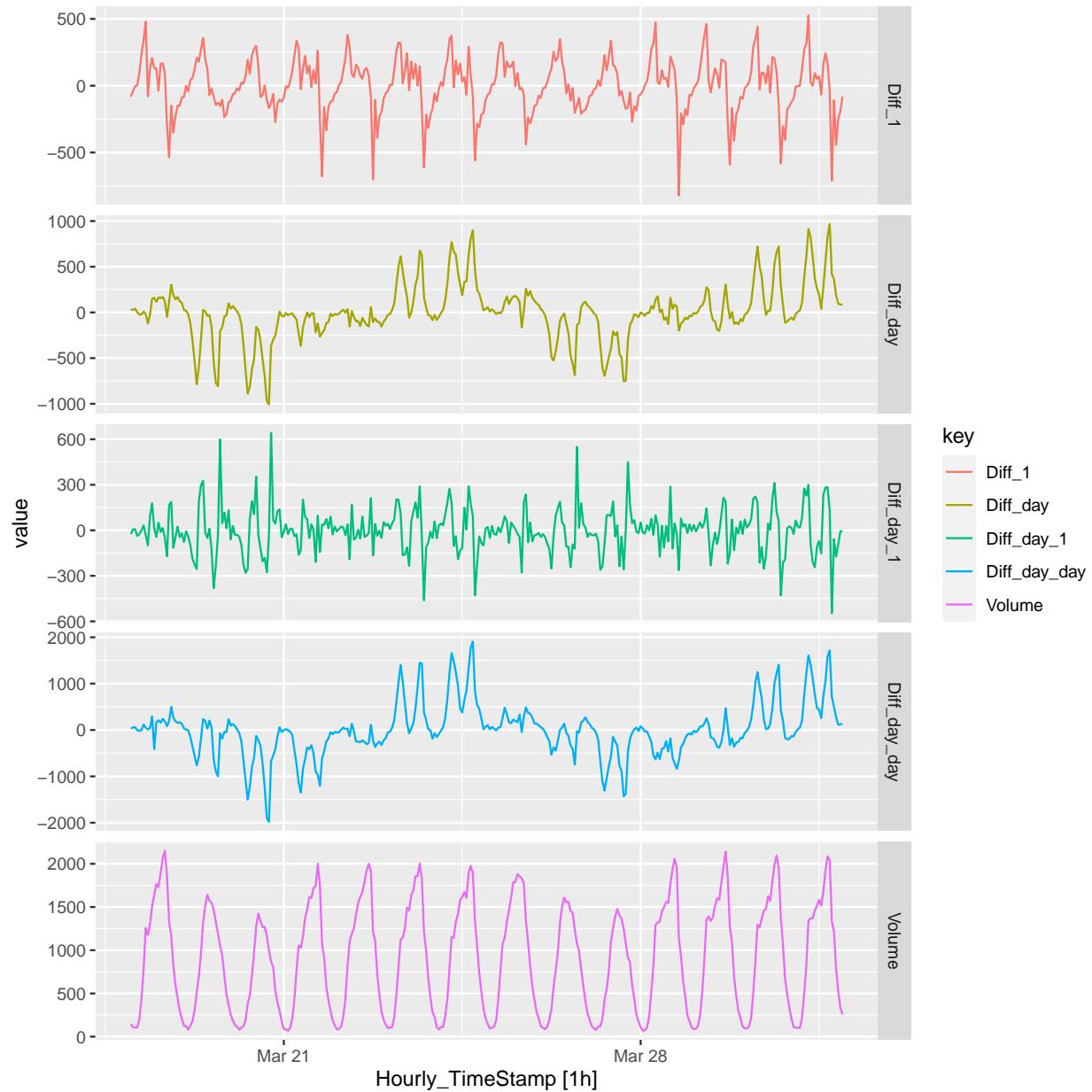
This is a dynamic harmonic regression model fit to an ARIMA error structure, with 10 Fourier terms for the daily seasonal period, and 5 Fourier terms for the weekly seasonal period. The square root is used to ensure predictions remain positive.

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.



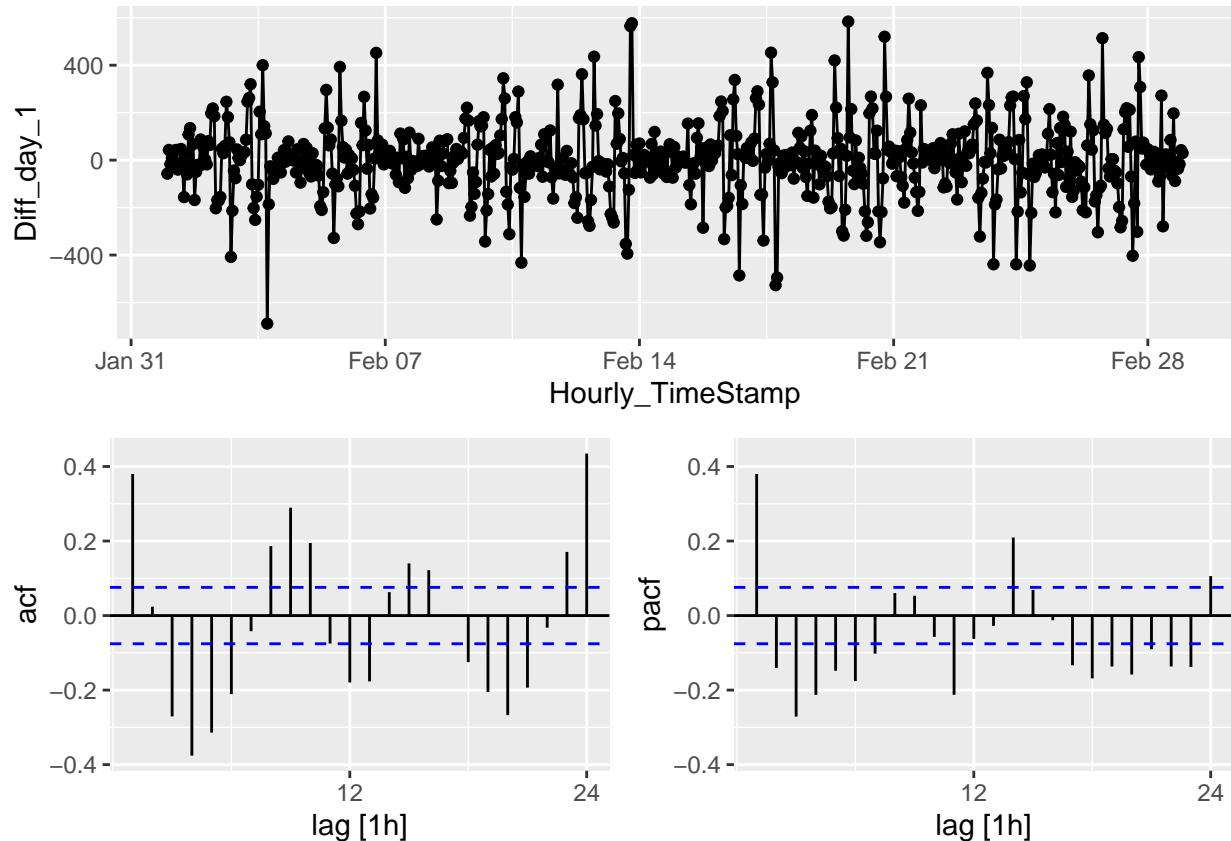
ARIMA

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 the plot of the 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.

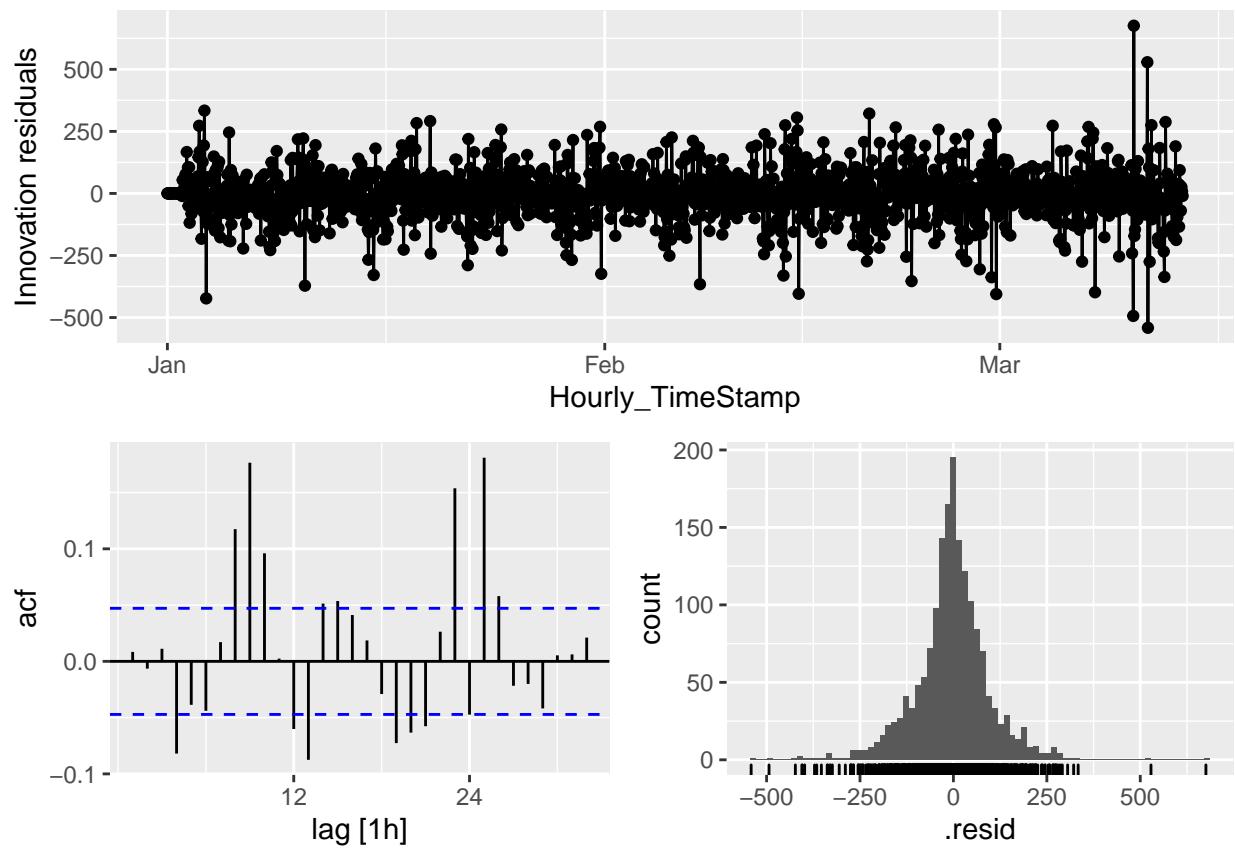
```
##      kpss_stat   kpss_pvalue
## 0.0005244991 0.1000000000
```



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 likely to work well.

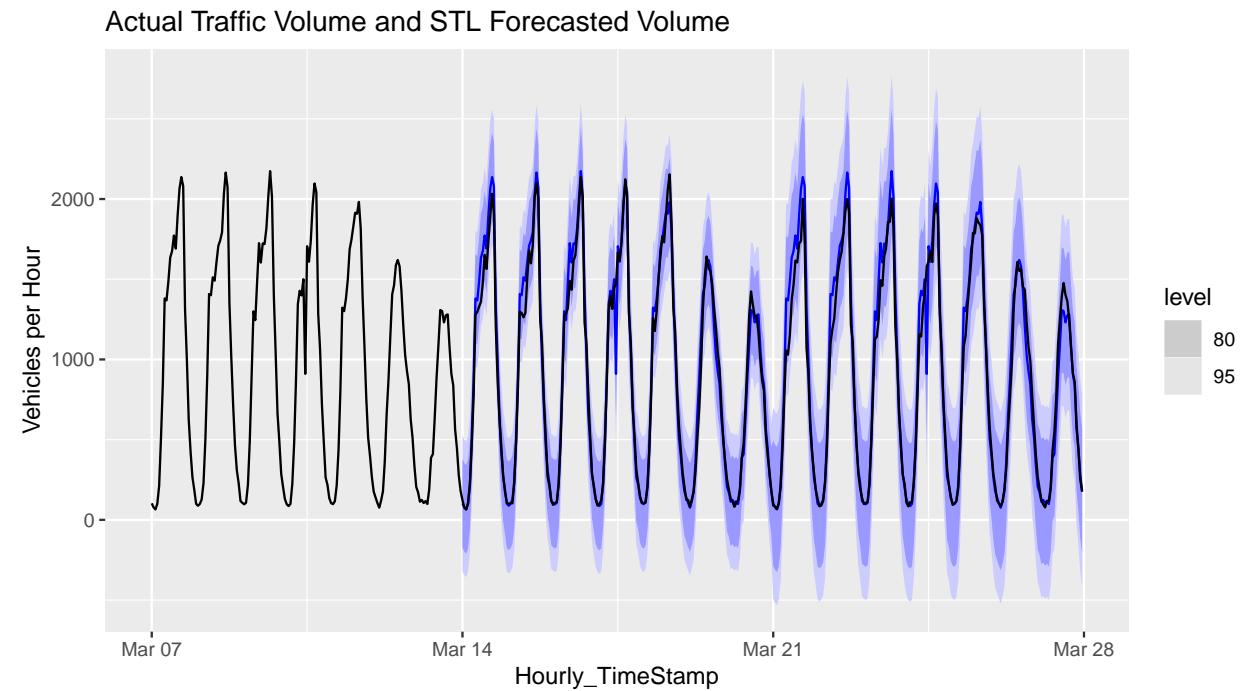
```
## # A mable: 4 x 2
## # Key:     Model name [4]
##   'Model name'          Orders
##   <chr>                <model>
## 1 a1                  <ARIMA(1,0,3)(2,1,0)[24]>
## 2 a2                  <ARIMA(1,1,3)(2,1,0)[24]>
## 3 b1                  <ARIMA(4,0,0)(2,1,0)[24]>
## 4 b2                  <ARIMA(4,1,0)(2,1,0)[24]>

## # A tibble: 4 x 8
##   .model sigma2 log_lik    AIC    AICc    BIC ar_roots  ma_roots
##   <chr>   <dbl>   <dbl>   <dbl>   <dbl> <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]>
```



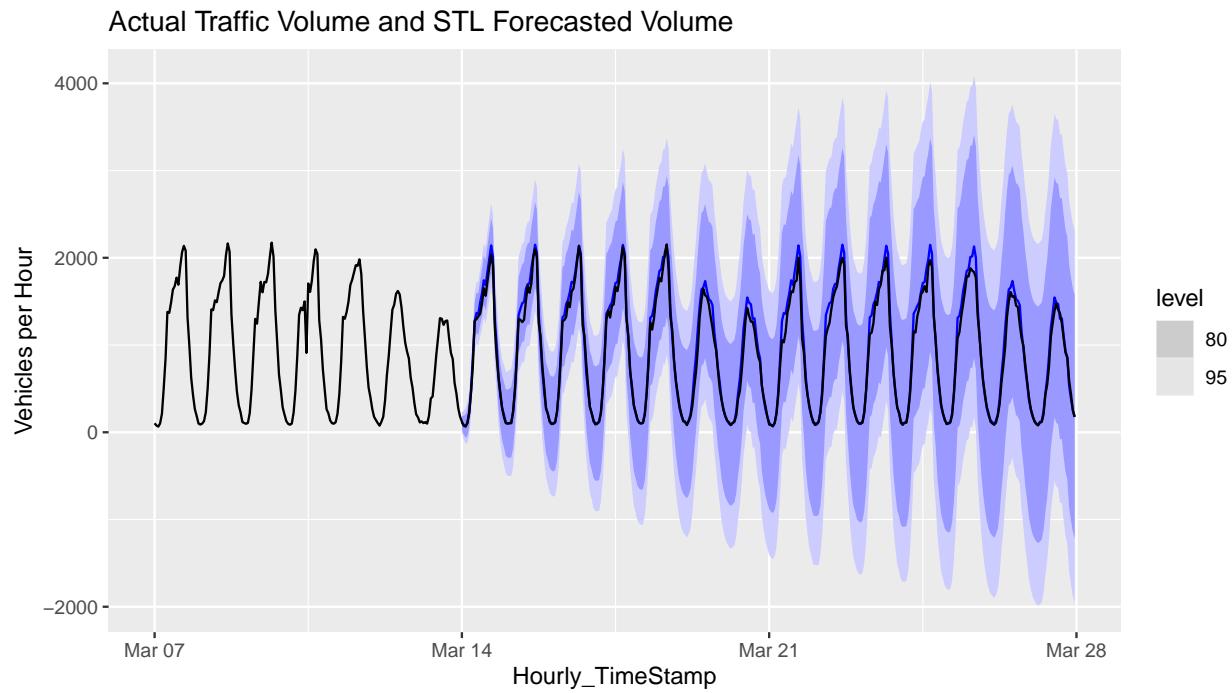
5. Prediction

Seasonal Naive 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 "SNAIVE(Hourly_Volume ~ Test"  Test -27.1 111.  73.1 -1.64  8.82  0.643 0.551 0.547
```

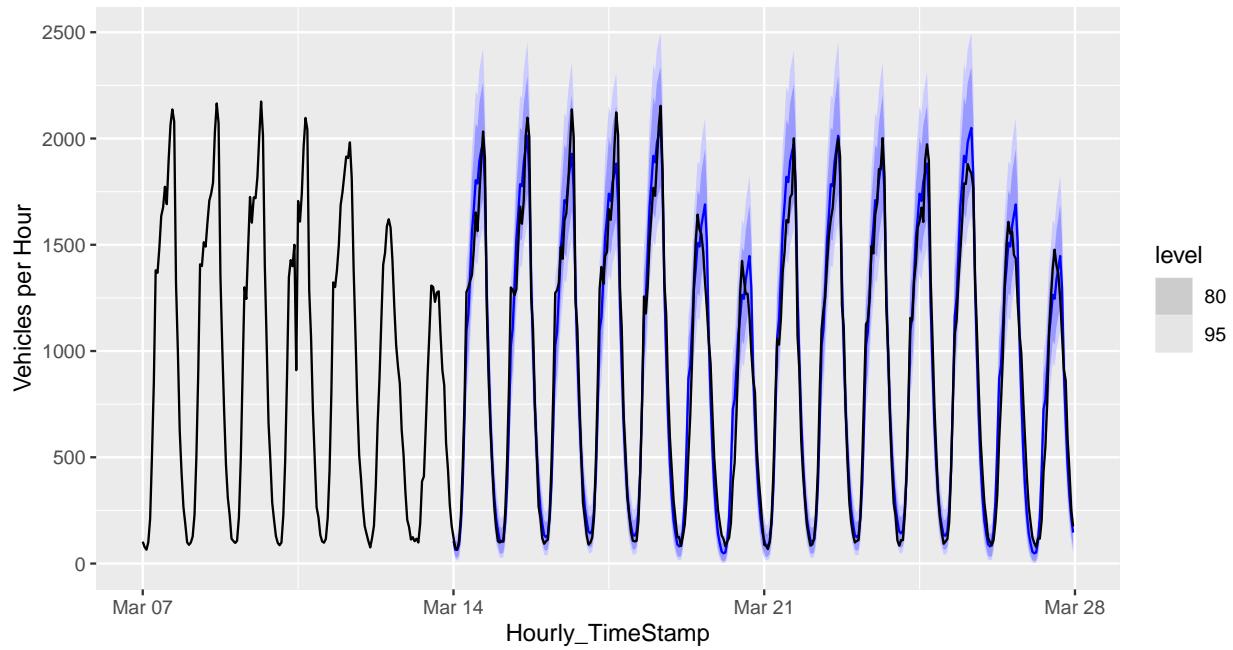
STL model



```
## # 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
```

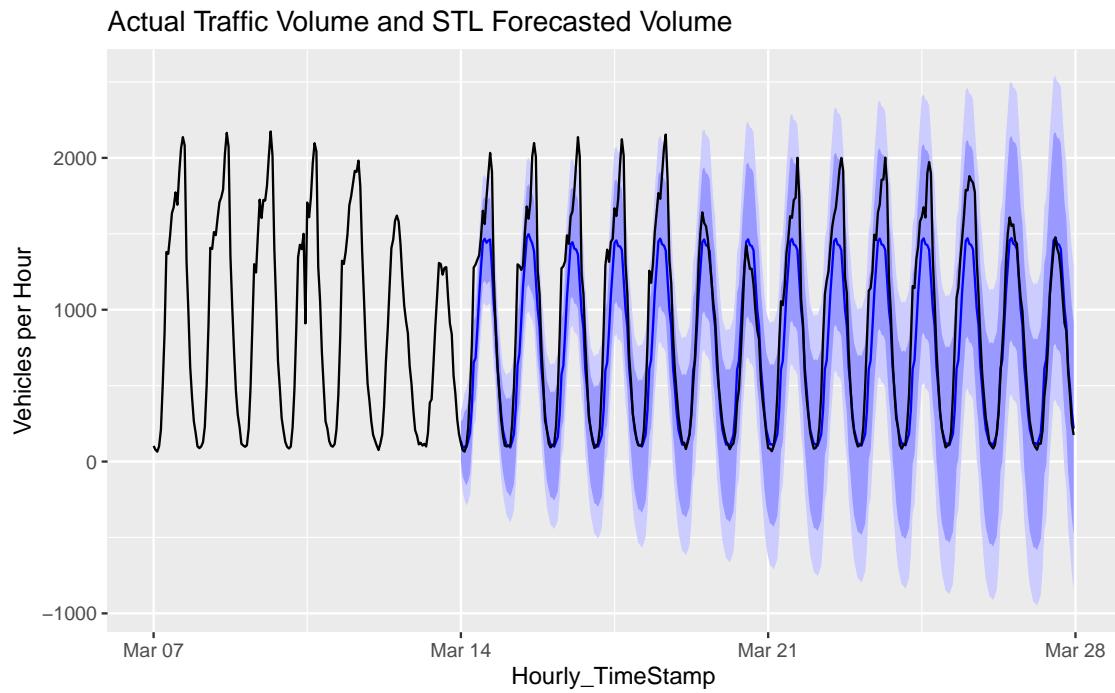
Dynamic Harmonic Regression

Actual Traffic Volume and STL Forecasted Volume



```
## # 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
```

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 8.8% for seasonal naive, 17.4% for dynamic harmonic regression, and 22.8% for ARIMA. The ARIMA model did not work well because it does not account for the multiple seasonal periods present in the data.