Project 2: Forecasting NYC Noise Complaints

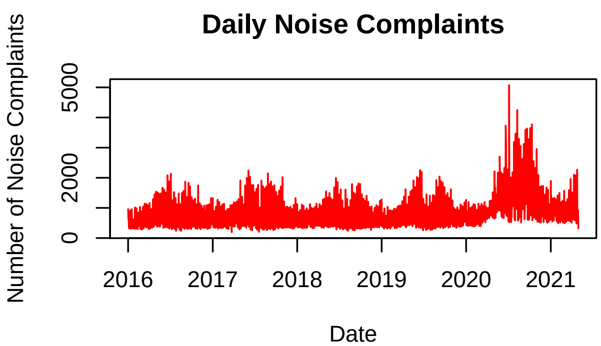
Elizabeth Combs (eac721)

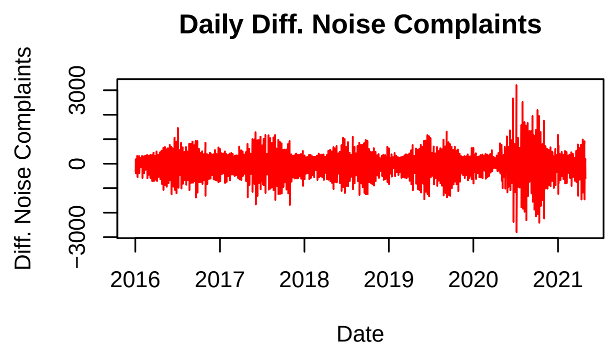
05/04/2021

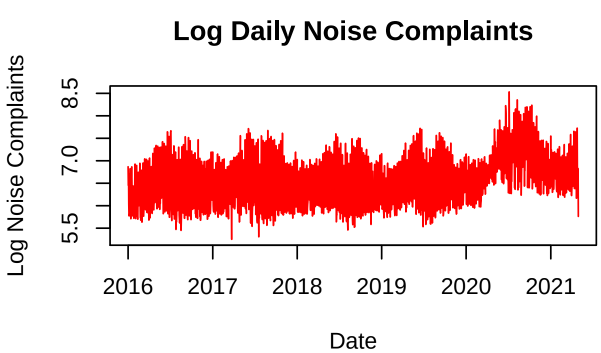
0. Introduction

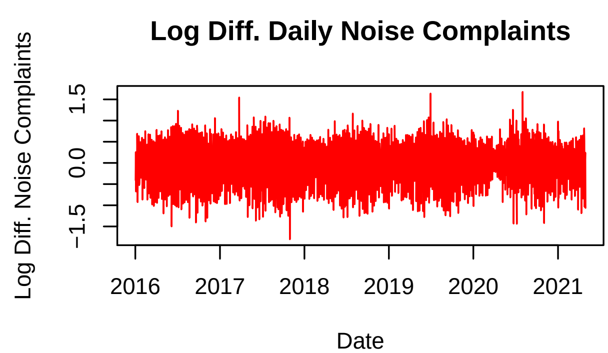
The data for this project was retrieved from NYC Open Data which hosts all of the 311 Service Requests from 2010 to Present. The data is filtered to create about five years of data from Jan. 1, 2016 through Apr. 30, 2021 for "Noise - Residential" complaint types to reduce the size of the data pulled from the API and focus on a shorter, more recent period of time. The data was filtered using this documentation. There are a total of 1,410,942 complaints during this time period with an average of 724 complaints per day. A simple tally (sum by date) of the complaints was used for this project so that the data could be converted to a daily time series for modeling purposes to make evenly spaced time intervals. *Note: All code for this project is available on GitHub.*

1. Data Review: Take Logs, Take Differences









Based on the graphs above, we can see the level-dependent volatility in the upper two plots. Thus, we will log the data so that we can eliminate the level-dependent volatility (which we can see in the lower graphs).

Additionally, there appears to be a weekly seasonality component of the residential noise complaints with larger values on weekends. We will remove the weekly seasonality as described in Project 1 by subtracting the seasonal mean from each value to remove this structure.

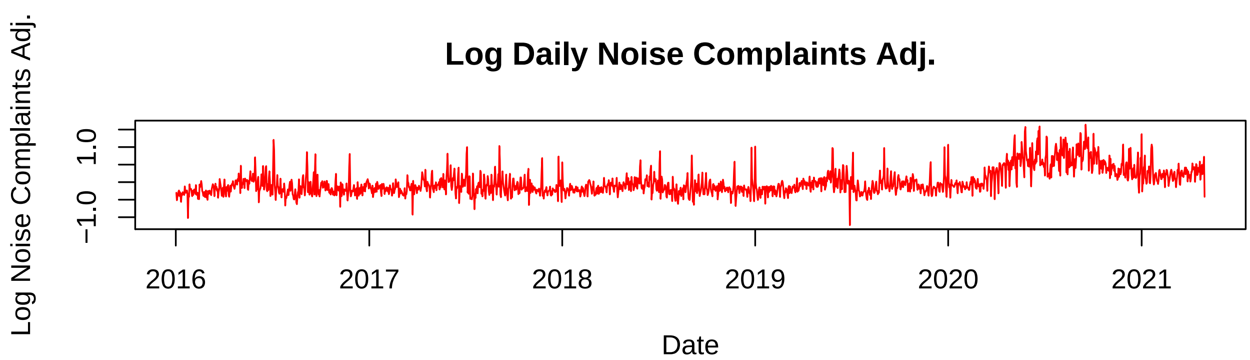
Show 7 entries Search:

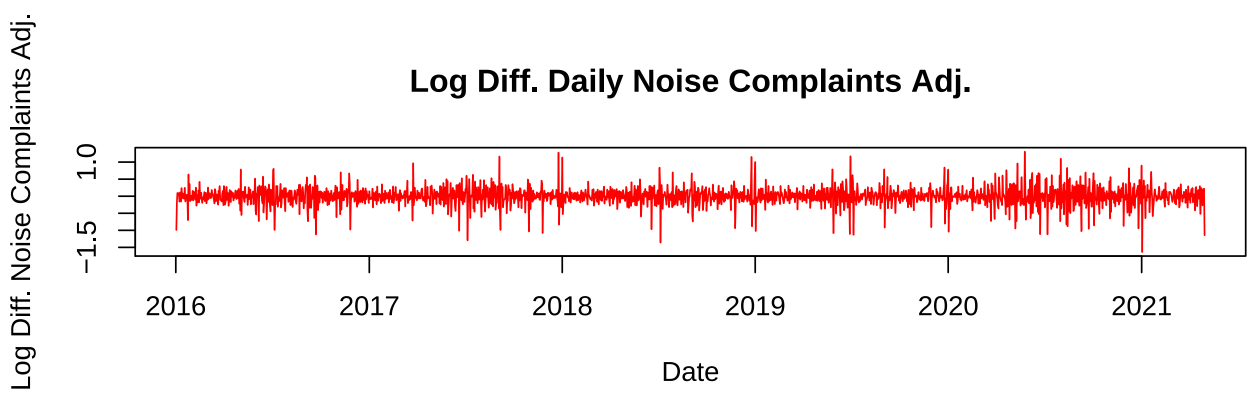
	day	log_weekly_noise_complaints
1	Friday	6.18
2	Monday	6.46
3	Saturday	6.76
4	Sunday	7.23
5	Thursday	6.1
6	Tuesday	6.14
7	Wednesday	6.08

Showing 1 to 7 of 7 entries Previous 1 Next

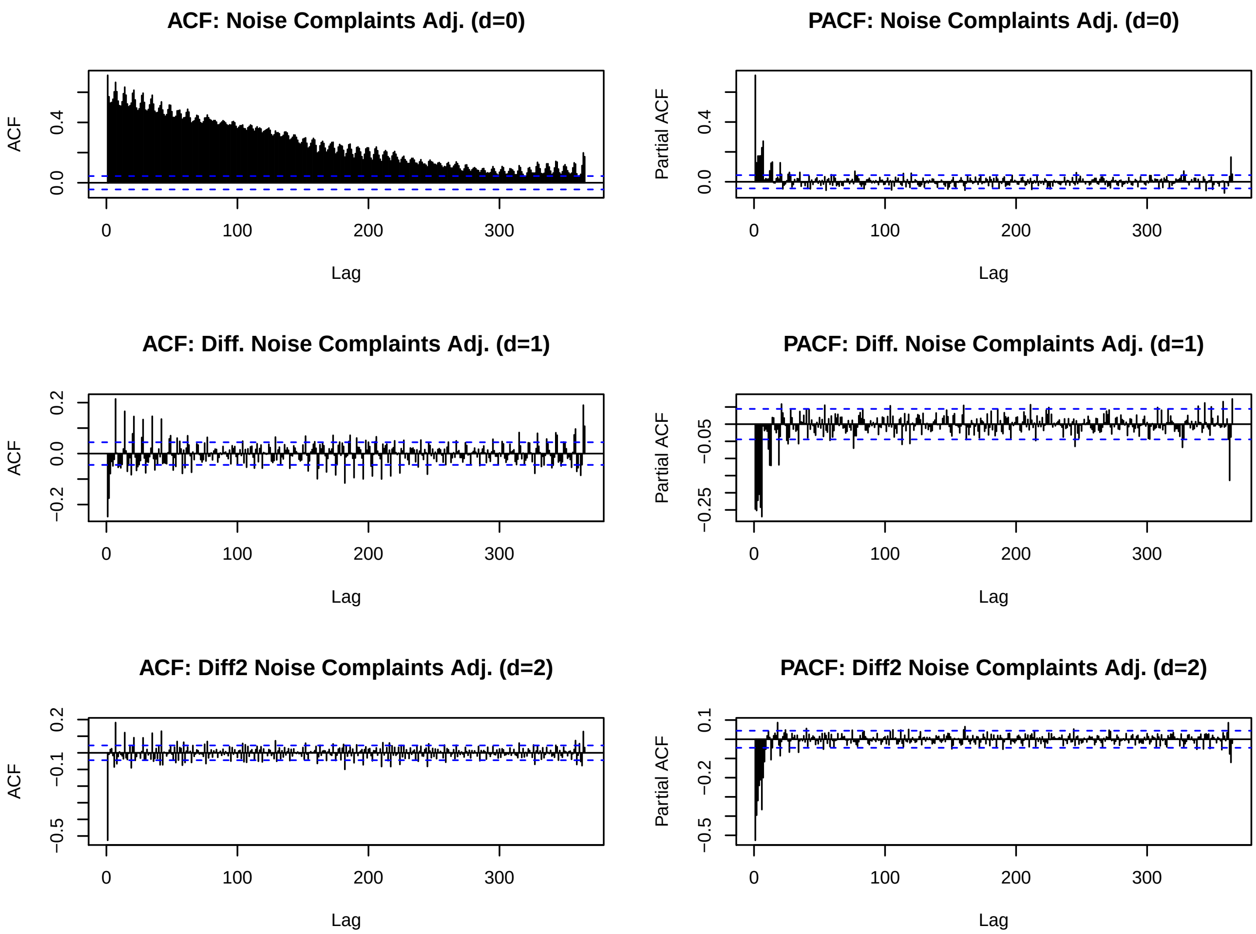
After the seasonality adjustment, we also remove the last value x_{n+1} of the dataset for validation purposes. This way, we can check the performance of each model we run (i.e. ARIMA vs. ARIMA-ARCH). The log of the final value (April 30) is 5.76 and the seasonally adjusted final value is $x_{n+1} = -0.42$.

Now, we can view the seasonally adjusted series and the difference of the series again now that we have removed the weekly structure:





Viewing the seasonally adjusted data, we can see that the data is mostly mean reverting around approximately 0. However, there was a surge in noise complaints and volatility around the time period when COVID started. Thus, we can check the ACF and PACF plots for values of d between 0 and 2 to select a value for the parameter:



The main purpose of differencing is to make the data stationary. We want to make sure to difference only as many times as is necessary. We wish to use an integer value for d, so there is a tough decision to make between d=0 and d=1. We will select d=0 for the following reasons: There is a long die down in the series' ACF plot, which could indicate long memory. Furthermore, the die down does not start at autocorrelation=1. Another reason is that the ACF plot of the difference (d=1) starts out around -0.3, which is approaching -0.5.

```
##
## Call:
##   fracdiff::fracdiff(x = noise_complaints$log_n_adj)
##
## Coefficients:
##         d
## 0.3982114
## sigma[eps] = 0.237325
## a list with components:
##  [1] "log.likelihood"  "n"                "msg"              "d"
##  [5] "ar"              "ma"               "covariance.dpq"   "fnormMin"
##  [9] "sigma"           "stderror.dpq"     "correlation.dpq" "h"
## [13] "d.tol"           "M"                "hessian.dpq"      "length.w"
## [17] "residuals"       "fitted"           "call"
```

Should we wish to use a d between 0 and 1, we can run the `fracdiff::fracdiff` command (above) which generated a d of 0.4, which is closer to 0 than 1. Finally, I tried forecast::ndiffs to help make my decision with the result of 0. For these reasons and that we want to be conservative and guard against overdifferencing, we select d=0. *Note: Even though we adjusted the data, it seems like there is still some seasonality structure in* *the data (each 7 days), so there may be better ways to improve our model (i.e. use SARIMA). This may be another reason that differencing is not straightforward in this case.*

By selecting d=0, we will assume that the data is stationary. While in practice, it is difficult to find a truly stationary series, it is the best assumption we can make at this time. We use these plots to identify the d for our modeling, but we cannot use them to identify an ARIMA(p,d,q). Instead, we can use AICC as a metric to select an ARIMA model.

2. Arima Model Selection Using AICc

We can use the AICc criteria to select which ARIMA(p, 0, q) model we should choose:

Show 18 entries Search:

	p	d	q	constant	aicc
12	1	0	2	FALSE	-190.2
18	2	0	2	FALSE	-190.17
11	1	0	2	TRUE	-188.19
17	2	0	2	TRUE	-188.16
16	2	0	1	FALSE	-170.45
15	2	0	1	TRUE	-168.44
10	1	0	1	FALSE	12.15
9	1	0	1	TRUE	14.15
14	2	0	0	FALSE	176.01
13	2	0	0	TRUE	178
8	1	0	0	FALSE	205.24
7	1	0	0	TRUE	207.23
6	0	0	2	FALSE	462.19
5	0	0	2	TRUE	464.16
4	0	0	1	FALSE	716.16
3	0	0	1	TRUE	718.12
2	0	0	0	FALSE	1582.78
1	0	0	0	TRUE	1584.72

Showing 1 to 18 of 18 entries Previous 1 Next

We select the ARIMA(1, 0, 2) with no constant and the minimum AICc=-190.2 among candidates. The model summary is printed below:

```
## Series: noise_complaints$log_n_adj
## ARIMA(1,0,2) with zero mean
##
## Coefficients:
##          ar1      ma1      ma2
##       0.9969  -0.5880  -0.3158
## s.e.  0.0018   0.0208   0.0204
##
## sigma^2 estimated as 0.05291:  log likelihood=99.11
## AIC=-190.22   AICc=-190.2   BIC=-167.93
##
```

```
## Training set error measures:
##                       ME      RMSE       MAE      MPE     MAPE      MASE
## Training set 0.002804038 0.2298354 0.1621016 235.1742 685.1727 0.8580025
##                    ACF1
## Training set 0.01133023
```

The formula of the model we fitted is is $x_t = 0.9969t - 1 + \varepsilon_t - 0.5880\varepsilon_{t-1} - 0.3158\varepsilon_{t-2}$, where x_t is the number of daily adjusted residential noise complaints in NYC. *Note: In Minitab, the MA coefficients have the sign flipped compared to the R output, so we don't need to change the sign here.*

The one-step ahead forecast of this model is, including 95% confidence intervals:

Show 10 entries   Search:

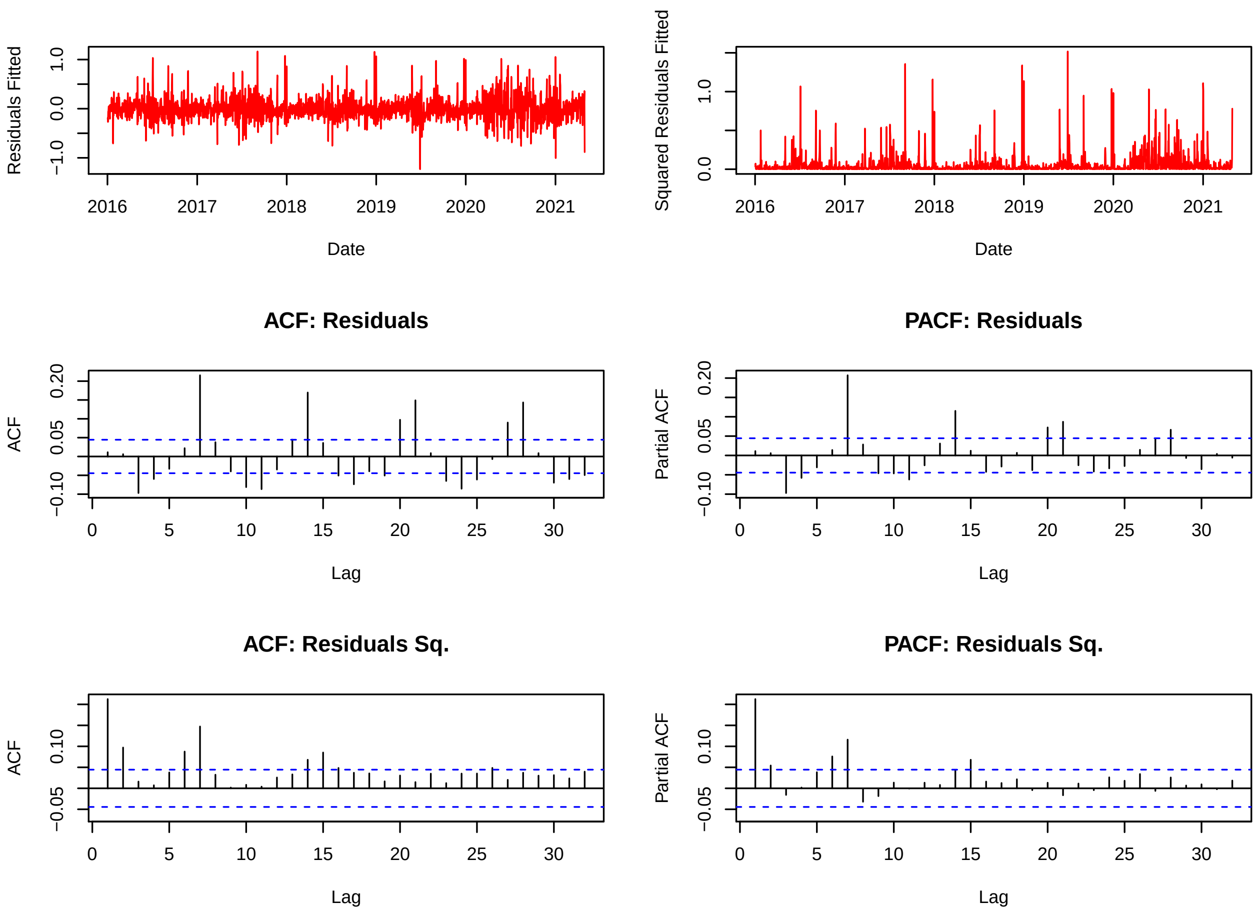
	Point Forecast	Lo 95	Hi 95
1947	-0.011	-0.462	0.44

Showing 1 to 1 of 1 entries   Previous 1 Next

While this forecast interval seems large, it needs to be this large to fit the data 95% of the time. The interval does contain the real value, -0.418, for the one-step ahead prediction (see problem 10).

3. Arima Residuals Analysis

Here is a plot of the residuals as well as ACF and PACF of both the residuals and the squared residuals:



The ACF and PACF plots of the residuals do not show significant autocorrelations and partial autocorrelations especially at low lags; thus, they appear to be uncorrelated. There does appear to be some pattern at ~7 days between lags, which may be associated with the day of week which has some structure even though we made a seasonality adjustment. Even if the residuals are uncorrelated and centered at zero (the expectation is zero), they do not look to be independent from the plots of the residuals and squared residuals. The residuals squared ACF plot seems to die down, which might be an indication that there is conditional volatility in our data that could be captured using a different model (i.e. ARCH). More specifically, when the volatility is high, the ARIMA model fits less well. This problem is evidence of conditional heteroscedasticity since when volatility is high, it tends to stay high and when it is low it tends to stay low.

4. ARCH Model Selection & GARCH(1,1)

Using the residuals from the ARIMA model, we find the log likelihood values and AICC values for ARCH(q) models where q ranges from 0 to 10. The log likelihood for the ARCH(0) model is calculated by hand. Here are the AICc values for the ARCH(q):

Show 11 entries Search:

	q	loglik	aicc
9	8	273.22	-528.34
10	9	272.96	-525.81
11	10	273.33	-524.52
8	7	268.15	-520.23
7	6	255.51	-496.96
6	5	249.25	-486.46
4	3	246.44	-484.87
3	2	244.95	-483.89
5	4	246.06	-482.08
2	1	220.96	-437.91
1	0	100.13	-198.25

Showing 1 to 11 of 11 entries Previous 1 Next

The best model based on q ranges from 0 to 10 is ARCH(8), which has the lowest AICC of -528.34.

Next, we consider a GARCH (1,1) model and evaluate AICC for the GARCH (1,1) model, using q=2.

Show 10 entries Search:

	loglik	aicc
1	249.12	-494.24

Showing 1 to 1 of 1 entries Previous 1 Next

We could select the model with the lowest AICc=-528.34, which is the ARCH(8) model. The model results below indicate that omega and some alphas (lower from 1-2 and weekly from 6-8) are significant:

```
##
## Call:
## garch(x = resid, order = c(0, 9), trace = FALSE)
##
## Model:
## GARCH(0,9)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -5.40083 -0.58258 -0.04623  0.53404  5.56409
##
## Coefficient(s):
##     Estimate  Std. Error  t value Pr(>|t|)
## a0  0.018348    0.001032   17.780  < 2e-16 ***
## a1  0.373279    0.031099   12.003  < 2e-16 ***
## a2  0.097413    0.019414    5.018 5.23e-07 ***
## a3  0.004593    0.011552    0.398   0.6909
## a4  0.002835    0.013468    0.211   0.8333
## a5  0.019117    0.015982    1.196   0.2316
## a6  0.046047    0.017949    2.565   0.0103 *
## a7  0.100230    0.021302    4.705 2.54e-06 ***
## a8  0.074524    0.015598    4.778 1.77e-06 ***
## a9  0.017345    0.013035    1.331   0.1833
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Diagnostic Tests:
##  Jarque Bera Test
##
## data:  Residuals
## X-squared = 929.8, df = 2, p-value < 2.2e-16
##
##
##  Box-Ljung test
##
## data:  Squared.Residuals
## X-squared = 9.4791e-05, df = 1, p-value = 0.9922
```

```
## 'log Lik.' 272.9643 (df=10)
```

However, since the ARCH model is not parsimonious and many coefficients are not significant, the GARCH(1,1) model may be a more simple model that we could use for similar results in forecasting. Furthermore, the GARCH(1,1) model has three coefficients that are very statistically significant p<2e-16:

```
##
## Call:
## garch(x = resid, order = c(1, 1), trace = FALSE)
##
## Model:
## GARCH(1,1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -6.19223 -0.57838 -0.04491  0.53337  5.07151
##
## Coefficient(s):
##     Estimate  Std. Error  t value Pr(>|t|)
## a0  0.014210    0.001171    12.13   <2e-16 ***
## a1  0.434731    0.030557    14.23   <2e-16 ***
## b1  0.368945    0.032581    11.32   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Diagnostic Tests:
##  Jarque Bera Test
##
## data:  Residuals
## X-squared = 959.51, df = 2, p-value < 2.2e-16
##
##
##  Box-Ljung test
##
## data:  Squared.Residuals
## X-squared = 0.078361, df = 1, p-value = 0.7795
```

```
## 'log Lik.' 249.1244 (df=3)
```

In this case, we prefer the more parsimonious model. Thus, the selected model has the form: $h_t = 0.014210 + 0.434731\varepsilon_{t-1}^2 + 0.368945h_{t-1}$. Based on the model output, all estimates are statistically significance with p<2e-16. *Note: The parameters are significant based on the above summary output. Since the output gives us the two-sided p-values, we should divide the given p-values by two to get the one-sided value. This further signifies the significance of these parameters. However, we should still be wary of putting too much stake in the statistical significance of these parameters since we are most focused on getting the best forecast rather than the interpretability of these coefficients.*

While the Box-Ljung test p-value is high so the model seems to fit based on this criteria, the model fails the test for the residuals. There may be a better fitting model that we could use to capture this variation.

The unconditional (marginal) variance of the shocks can be computed with the formula $\text{var}(\varepsilon_t) = \frac{\omega}{1-\sum_{j=1}^{q} \hat{\alpha_j}}$. So, we can compute our unconditional variance as $\frac{0.014210}{1-(0.434731+0.368945)} = \frac{0.014210}{0.196324} = 0.07238035$.

5. Forecast of ARIMA-ARCH model

Construct a 95% one step ahead forecast interval for the log adj. noise complaints, based on your ARIMA-ARCH model. We use the formula $h_t = f_1 + \sqrt(h_1)$, where h_1 is given by the model formula above and f_1 is the ARIMA forecast.

```
f1 <- fcasts$`Point Forecast`

ht <- fit.var$fit[,1]^2
h1 <- fit.var$coef[1] + tail(ht, n=1) %>% as.numeric()

f1 + c(-1, 1) * 1.96 * sqrt(h1)
```

```
## [1] -0.6266791  0.6048702
```

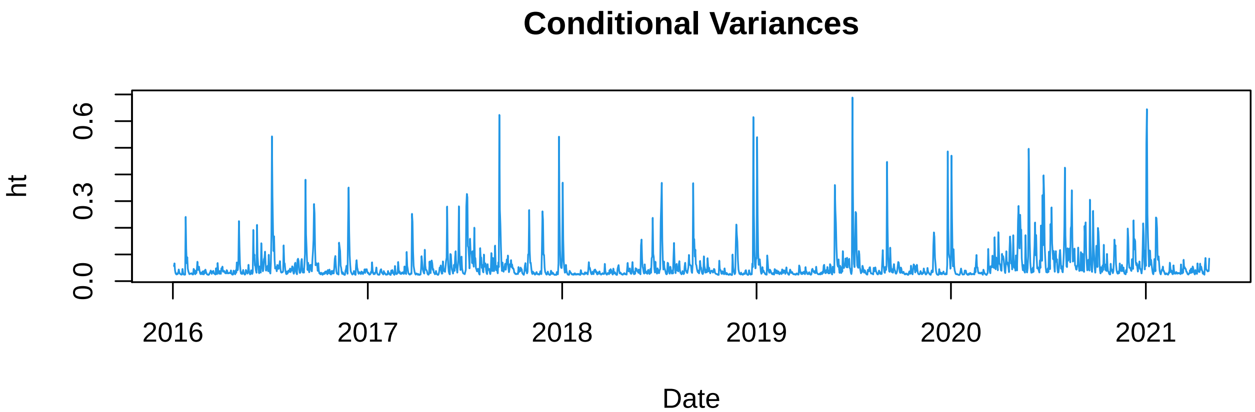
This interval is wider than the interval from problem 2 which was thinner. Our second model may be able to better capture the volatility so that we can be more sure during the more volatile time period and still be 95% confident in our forecast. Since the volatility lately has been higher (perhaps due to COVID), the ARIMA-ARCH forecast has a wider interval to accommodate it.

The 95% confidence interval is actually computing the 2.5% and 97.5% percentiles to get 5% of the data outside of the center. To get the 5th percentile we need to do the same operation as we did for 95% but with a different z-score. We use 1.96 for 95% and 1.65 for 90% confidence intervals:

```
## [1] -0.5292862  0.5074772
```

6. Conditional Variances Analysis:

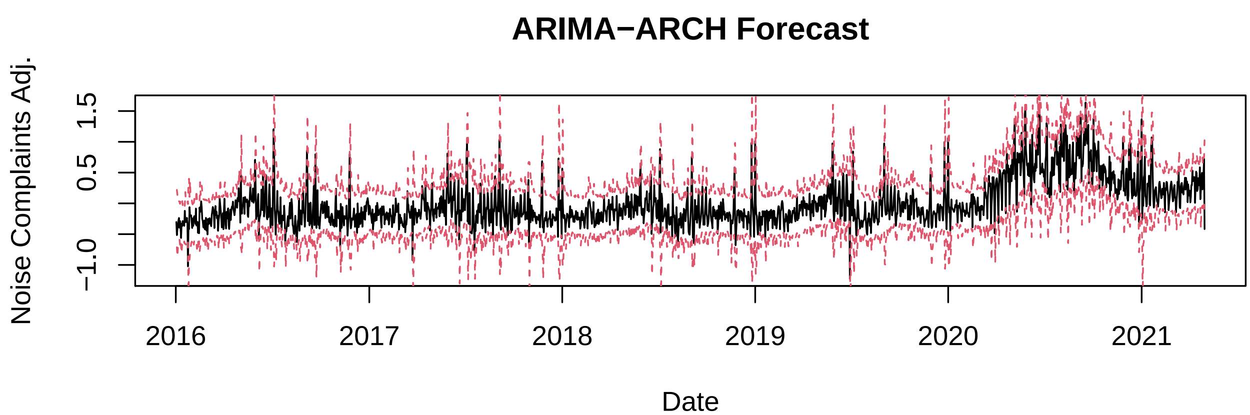
Here are the conditional variances, ht, for the fitted ARCH model from problem 4.



Volatility is especially high at the middle and end of years, which may correspond to July 4 celebrations and New Years celebrations. Prior to 2020, the volatility tended to be lower while in more recent times the volatility has been higher, perhaps due to COVID-19 and residents staying home more. The plot above seems to follow the same pattern as the time series plot in terms of volatility (problem #1). It also appears to capture a smoothed version of the residual plot from #3.

7. Visualize the ARIMA-ARCH Forecast

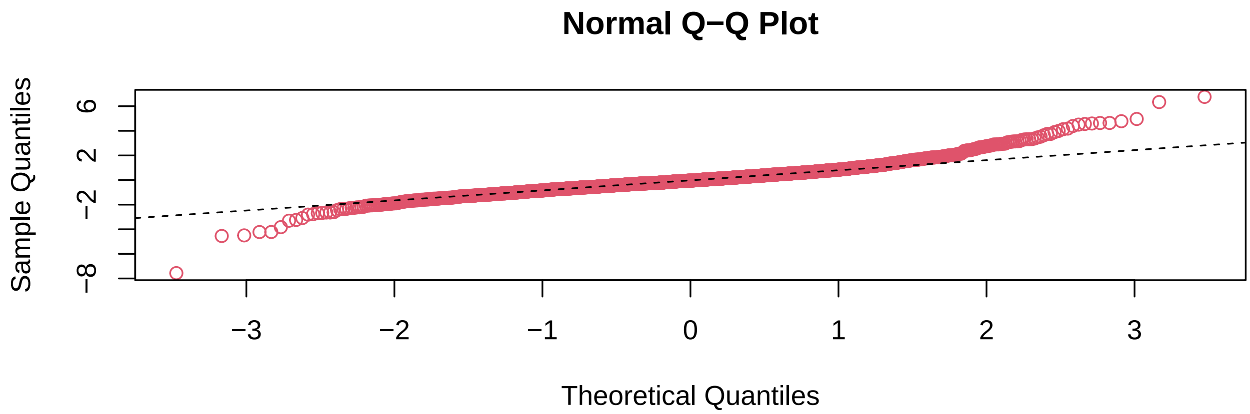
Here is a time series plot which simultaneously shows the log adj. noise complaints, together with the ARIMA-ARCH one-step-ahead 95% forecast intervals based on information available in the previous day:



While the forecast interval may seem wide, the past fitted values seem close together. The 95% interval follows the time series closely and appears to have only a few misses (when the interval does not capture the real data point). It appears that the forecast interval needs to be this wide to make sure we are able to forecast correctly 95% of the time, especially to capture the innate volatility of the dataset.

8. ARIMA-ARCH Model Adequecy

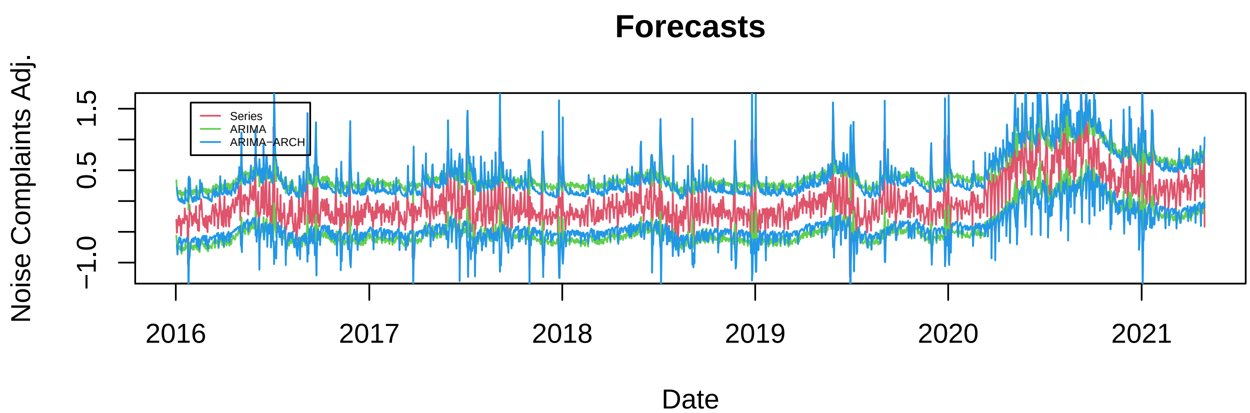
The residuals from your ARIMA-ARCH model are $e_t = \varepsilon_t/\sqrt{h_t}$. If the ARIMA-ARCH model is adequate, these residuals should be normally distributed with mean zero and variance 1. Here is a normal probability plot of the ARCH residuals:



There are a large number of data points that lie off of the x=y line towards the right and left side of the plot (tails of the distribution). This suggests that the model did not adequately describe the leptokurtosis in our data. If the model described it well, the data would sit more on the x=y line (normal data).

9. Forecast Interval Assessment

We can compare the forecasts on the same plot, where the main difference is that during periods of high volatility, the ARIMA-ARCH forecast is wider than the ARIMA forecast and narrower during low volatility:



We can compute the percentage of times we miss the intervals using the past data (abs(resid)>1.96). It is close to 5% ($\approx$ 4.78 for ARIMA-ARCH, $\approx$ 5.70 for ARIMA), so our interval is approximately wide enough in both cases. Capturing the volatility has allowed us to reach the 5% threshold and improved the forecast compared to the ARIMA only model.

```
## [1] "Percentage missed ARIMA-ARCH: 4.78"
```

```
## [1] "Percentage missed ARIMA: 5.7"
```

10. Predict and Compare 1-step ahead by Model:

```
## [1] "The real last data point is : -0.42."
```

The last point is -0.418 which is contained in both the ARIMA and ARIMA-ARCH intervals (ARIMA = c(-0.462, 0.440) and ARIMA-ARCH = c(-0.627, 0.605)). Even though the forecast intervals for the ARIMA-ARCH model are wider around the actualized values of noise complaints, we miss fewer times compared to the ARIMA. This might indicate that adding volatility explanation to our model with an ARCH model, we have gotten closer to the structure of the true data. Since the volatility remains high during COVID, the ARIMA-ARCH prediction interval remains wider than the ARIMA interval.