Part – II

Data pre- processing:

(1) "region_from" is 1
(2) "region_to" is 5
(3) "v0_num_traj" is used to calculates the flow for the hour
(4) The number of dates in the dataset is 28
(5) The number of hours in dataset is 24 hours or less are for each date
(6) The number of records should be (28 * 24). However, the dimension of the data frame is (576,3)
(7) There are 96(672 - 576) records missing, assuming 28 dates and 24 hours (672) for each date. These records are assumed to be zero for all the purposes of this assignment.
(8) Data cleaning: As the main interest of translink is the busier periods of the day, I remove the hours with more than 50% of time zero passengers. The table below shows that hours 1,2,3,4,5 can be removed.

Hours	Number/Percentage of zeroes
1	16 (57%)
2	20 (71%)
3	20 (71%)
4	20 (71%)
5	20 (71%)

(9) In conclusion, after data cleaning we have 19 hours and 28 dates

Additive or Multiplicative series:

(1) To find the additive or multiplicative series, the ACF of the reminder of time series is compared
(2) For the time series with a frequency of 19, the ACF of the reminder component from the additive and multiplicative series is calculated.
(3) For Additive series, reminder component is calculated from 'stl' component
(4) For the Multiplicative series, reminder component is calculated from the log of time series, using the 'stl' component
(5) The ACF for the reminder component of Additive series is 2.02, while from Multiplicative series is 2.22.
(6) Due to the relative similar ACF values, I will be choosing multiplicative series, due to better residuals statistics of the forecast models.
(7) Source: Link

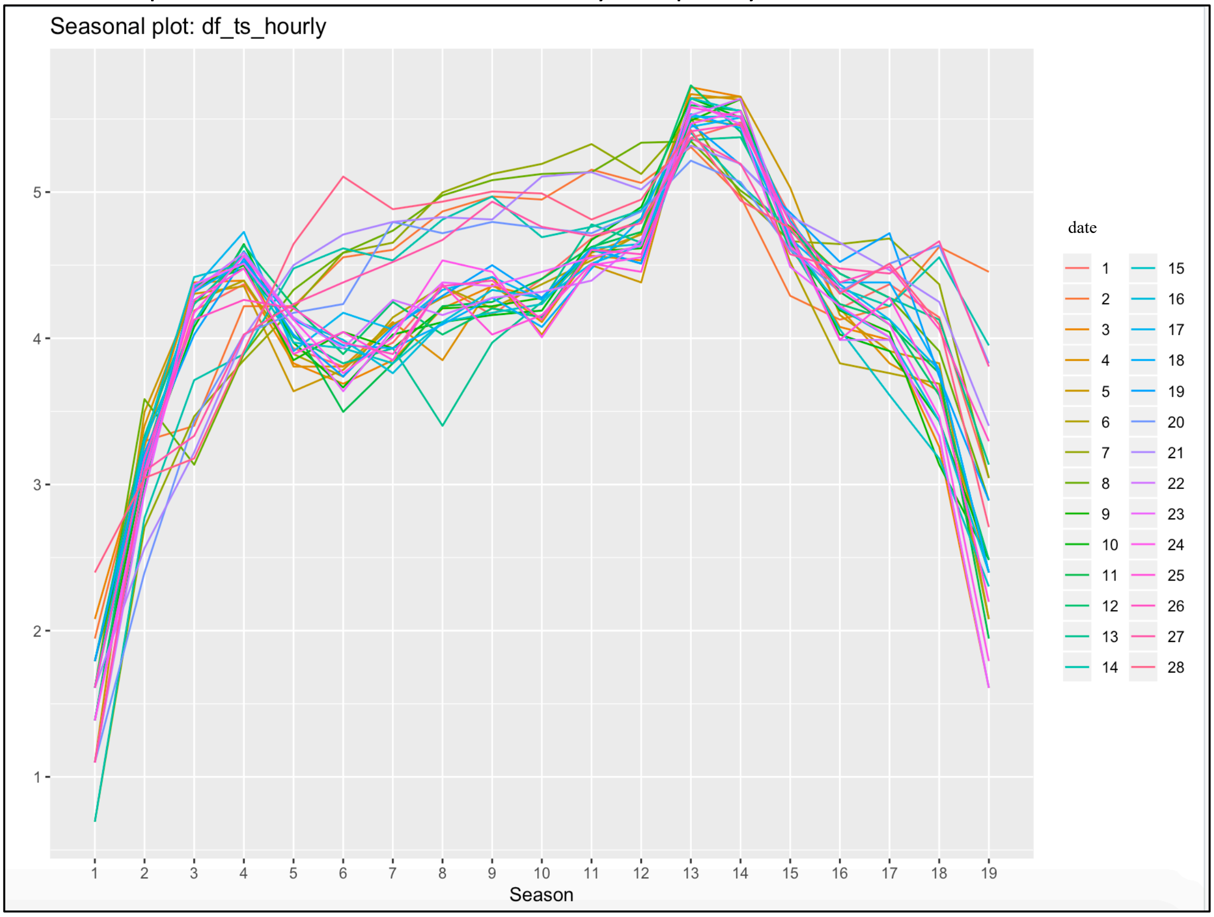
Question - 3

Check for stationarity and seasonality of the time-series data. Explain how you have done this and include relevant graphs and numerical summaries

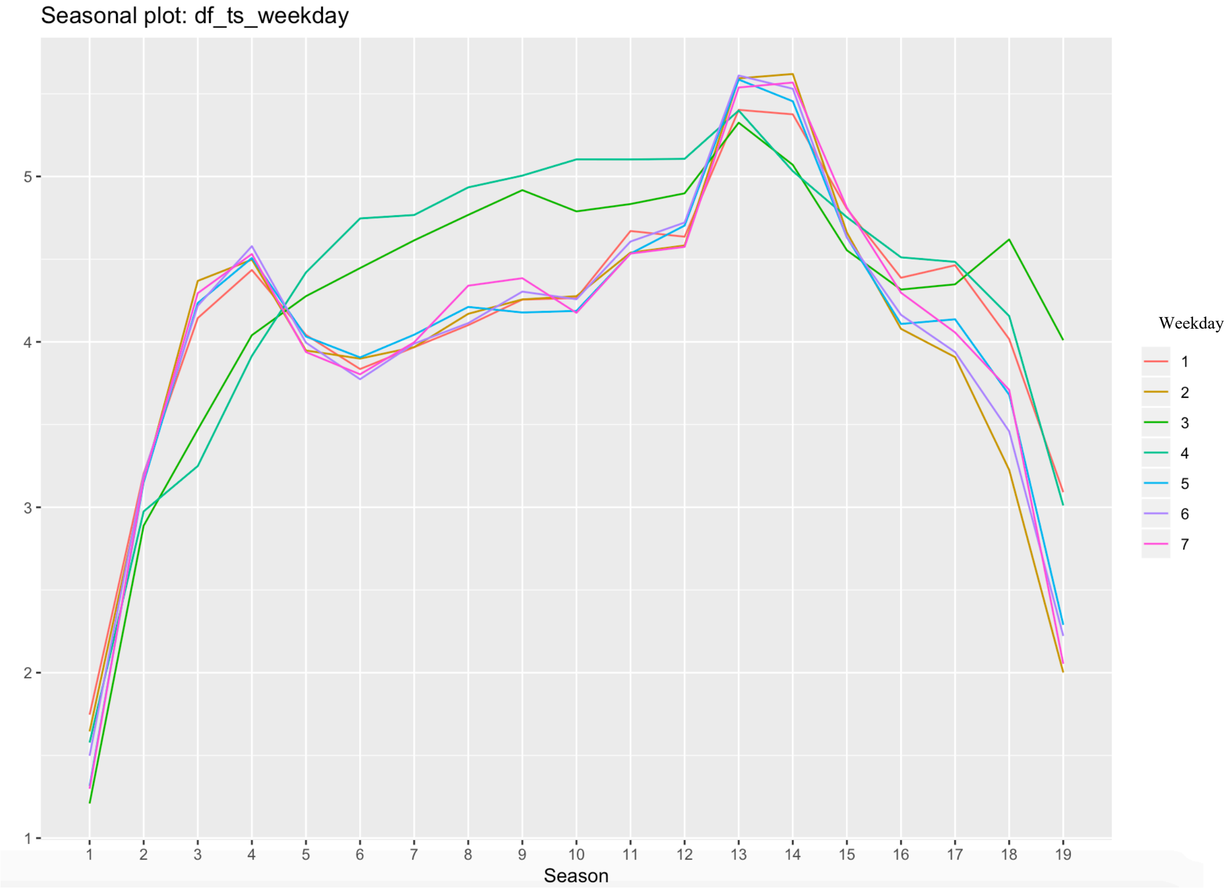
Answer - 3

3.1 Seasonality analysis

1. The time plot of the flow data at hourly frequency of 19 can be seen as:



2. On converting dates into days, the following curve can be obtained (frequency = 19):



3. From 1 and 2 it is clear that among day and date the seasonality of the time series is weekly while the frequency is hourly.
4. This can be verified by calculating the strength of the seasonality. The strength of the seasonality can be measured as follows:
    a. Strength of seasonality can be calculated as
    $$F_s = \max\left(0, 1 - \frac{Varience(R_t)}{(Varience(R_t) + Varience(S_t))}\right)$$
    Closer F_s value is to 1, more seasonality
    b. Strength of seasonality of dates: 0.89
    c. Strength of seasonality of days: 0.91
    d. Hence, seasonality is weekly. As the strength of seasonality is greater in weekly.
5. Post removing the weekly seasonality, the strength of seasonality for remaining data is 0.016. The seasonality is less than 5%, hence on performing this function we can say that there is no seasonality.
6. **In conclusion, the seasonality of the time series is at weekly**

3.2 Trend Analysis

The series has no trend. This is can be seen as the trend component of the series is NULL.

```
> trend(ts.stl)
NULL
```

3.3 Stationary Analysis

1. KPSS test is used to find if the time series is stationary. The null hypothesis is that the data is stationary, while the alternate hypothesis is that the data is not stationary.
2. On doing the KPSS test, the following is the result: The value of test statics is less than 0.05. This concludes that the **series is stationary**.

```
#######################
# KPSS Unit Root Test #
#######################

Test is of type: mu with 6 lags.

Value of test-statistic is: 0.0172

Critical value for a significance level of:
                10pct  5pct 2.5pct  1pct
critical values 0.347 0.463  0.574 0.739
```

Question – 4

Choose and detail each type of model used (including mathematical form and explanation of notation) and some details of how it was fitted. Explain why each model may be suitable for this type of data

4.1 Stochastic model

The first step for having a stochastic model, is to choose a stochastic model based on AICc value, RMSE value and the residual plots. In stochastic mode, I will compare four kind of models: Arima (ARIMA), Seasonal Arima (SARIMA), Holt-winter(HW), Exponential Smoothening State space (ETS). To test the results from the above algorithms the following factors will be checked:

1. Second order AIC (Akaike Information Criterion) or AICc: The model with lowest AICc value should be selected
2. Mean of residue: The model with mean of the residue close to zero must be selected
3. Ljung-Box test: This test is done to check the independence of residuals. The p-value to pass this test must be greater than 0.05
4. Variance of residuals: The model constant residual plot must be selected
5. Normal residual plots: The distribution of residual plots should be normal for prediction interval
6. Root mean square error (RMSE): The model with minimum RMSE value should be selected

Model	ARIMA*	SARIMA*	HW	ETS
AICc	52.75	51.39	1481.28	1426.37
Mean of Residual	Close to zero	Close to zero	Close to zero	Close to zero
Ljung-Box test	0.127	0.124	1.434e-08	0.006
Variance	Constant variance	Constant variance	Constant variance	Constant variance
Residue plot	Approximately normal	Approximately normal	Approximately normal	Approximately normal
RMSE	0.32	0.31	0.31	0.27

*Tested from auto-arima model

From the above table, it can be deduced that ARIMA and SARIMA model has the lowest AICc values, close to mean residual; acceptable p-value from Ljung-Box test (should be greater than 0.05); constant residue variance; approximately normal residue plot; lowest RMSE value. For further analysis SARIMA model will be used, as it has the best model statistics

4.1.1 Explanation

SARIMA or Sesonal autoregressive Integrated Moving Average aims to describe the autocorrelation in the data. SARIMA is formed by including the additional seasonal terms in ARIMA.

ARIMA has two components – Autoregressive (AR) and Moving Average (MA). In Autoregressive model (AR), the time series is forecasted using linear combination of past values of variable. In Moving average (MA) model, the time series is forecasted using the linear combination of past forecast error. ARIMA is a combination of AR and MA and can be represented as:

$$ARIMA\ (p, d, q)$$

Where, p is the order of AR model; q is the order of MA mode; d is the differentiation for stationary time series.

The SARIMA can be defined as,

$$SARIMA\ (p, d, q) \quad (P, D, Q)_m$$

Where, P, D, Q is the seasonal part of ARIMA model defined above; m is the seasonality.
The seasonal part of SARIMA is same as the non-seasonal part but includes backshift for a period of m.

4.1.2 Mathematical Detail

Mathematically ARIMA, this can be written as:

$$y'_t = c + \underbrace{\emptyset_1 \, y'_{t-1} \ldots \ldots \ldots \ldots \ldots \ldots \emptyset_p \, y'_{t-p}}_{\text{AR}} + \overbrace{\theta_1 \varepsilon_{t-1} \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \theta_q \varepsilon_{t-q}}^{\text{MA}} + \varepsilon_t$$

Where, y'_t is the differenced series at time t. This series have been differenced d times; ε_t is the white noise at time t;
p is the number of lags in AR model
q is the number of lags in the MA model.
$\emptyset_i$ and θ_i are coefficient values.

This can also be written as,

$$\underbrace{(1 - \emptyset_1 B \ldots \ldots \emptyset_1 B^p)}_{\text{AR}} \overbrace{(1 - B)^d y_t}^{\text{Stationary Series}} = c + \overbrace{(1 + \theta_1 B \ldots \theta_q B^q)}^{\text{MA}} \varepsilon_t$$

Where,

$$B_{y_t} \, or \, B = y_{t-1}$$

and,

$$B(B_{y_t}) = y_{t-2}$$

SARIMA can be written as:

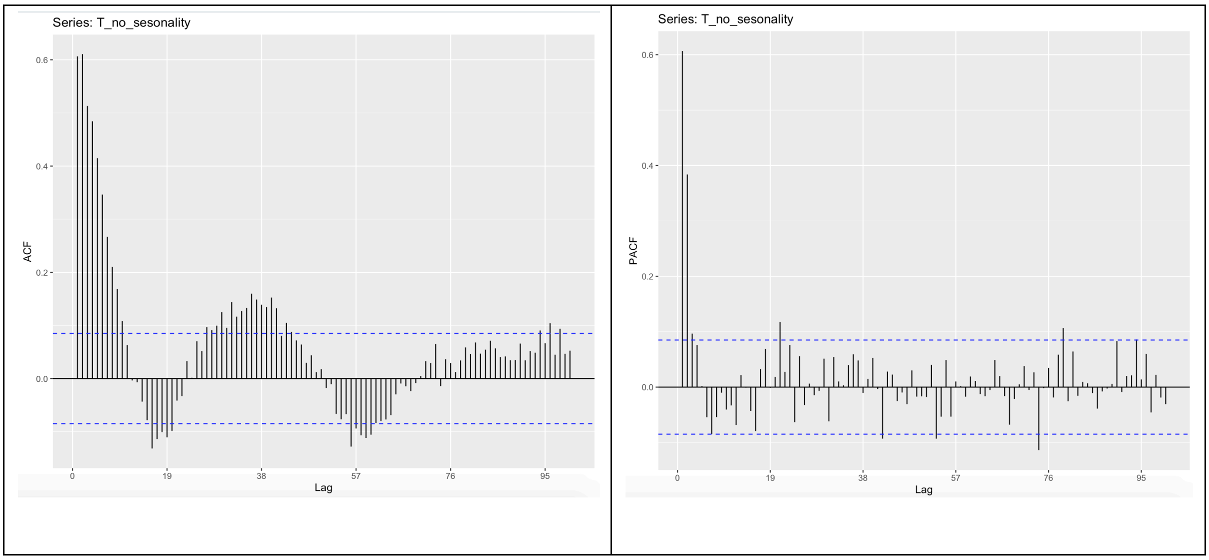
$$(1 - \emptyset_1 B \ldots \ldots \emptyset_1 B^p)\,(1 - B)^d\,(1 - \phi_1 B_1^m \ldots \ldots \phi_P B_P^m)\,(1 - B^m)^d\, y_t$$
$$= c + (1 + \theta_1 B \ldots \theta_q B^q)\left(1 + \varphi_1 B_1^m \ldots \ldots \varphi_Q B_Q^m\right) \varepsilon_t$$

The values of p and q can also be determined using the ACF and PACF plots. ACF plots or Auto-correlation plots measures the correlation between y_t and y_{t-k} i.e. correlation of the series with k lagged terms. For moving averages (MA), the value of q is obtained from ACF plot, and it is the value at which ACF plot crosses the upper confidence interval. The PACF or Partial or auto-correlation plot is used to measure correlation plots between y_t and y_{t-k} after removing effects from the values between these time series. For autoregressive models (AR), the value of p is obtained from PACF plot, and it is the value at which PACF plot crosses the upper confidence interval.

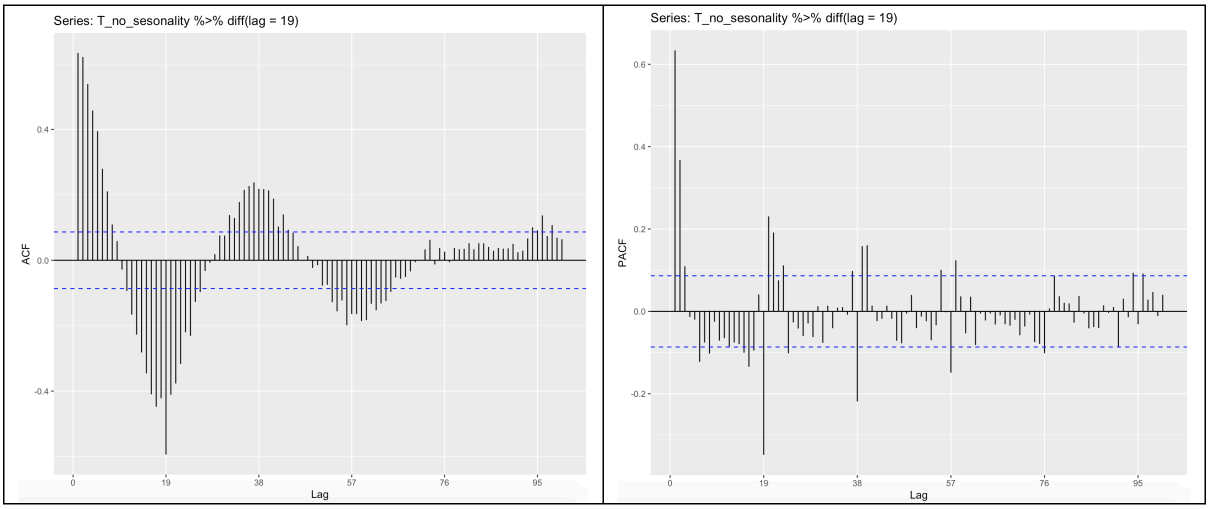
4.1.3 Implementation Detail

SARIMA is suitable for the current data at hand, because of the following reason:

1. The 1.6% of seasonal component left even after removing weekly seasonal component as specified in [3.1 Seasonality analysis], can be taken care by the SARIMA model.
2. The time series at hand has AR and MA components, as shown in ACF and PACF plots below



3. The time series also has AR in the seasonal component i.e. data with a lag of 19 (seasonal period of hourly; m = 19), as shown in ACF and PACF plot below



4. SARIMA has the best model statistics, as specified in [Stochastic model]

4.1.4 Fitting Details

The following steps are followed to get an ARIMA model:

1. Divide the data into test and train, such that test data contains the last 7 * 19 values i.e. 19 values for each day of week

2. On applying auto.arima to t his dataset without removing the weekly seasonality, the p-value of Ljung-Boxtest was less than 0.05. As shown below

```
        Ljung-Box test

data:  Residuals from ARIMA(4,0,0)(2,1,0)[19]
Q* = 52.155, df = 32, p-value = 0.01367

Model df: 6.   Total lags used: 38
```

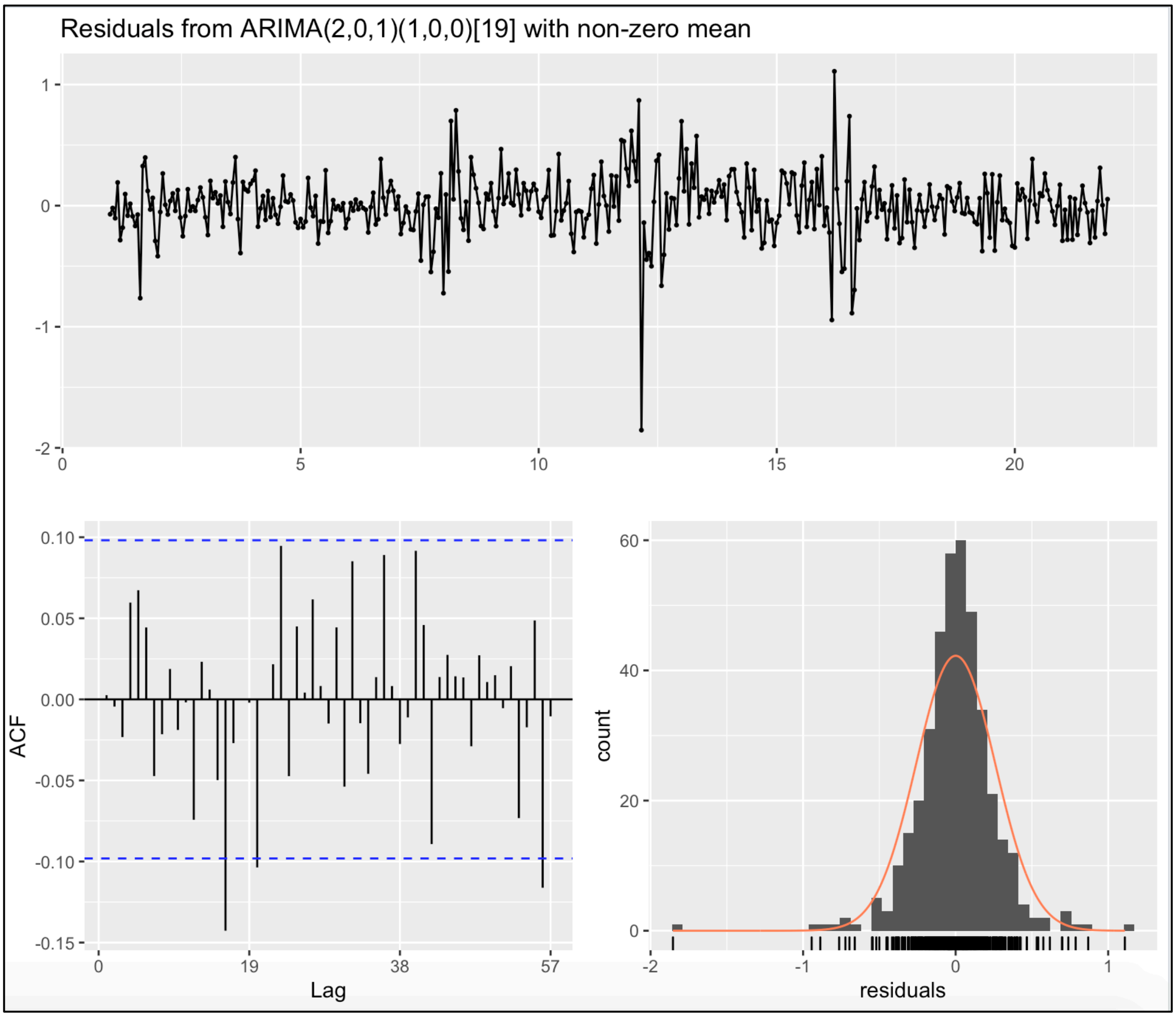
On removing the weekly seasonality from the data, the Ljung-Box test is as follows:

```
        Ljung-Box test

data:  Residuals from ARIMA(2,0,1)(1,0,0)[19] with non-zero mean
Q* = 40.454, df = 33, p-value = 0.1744

Model df: 5.   Total lags used: 38
```

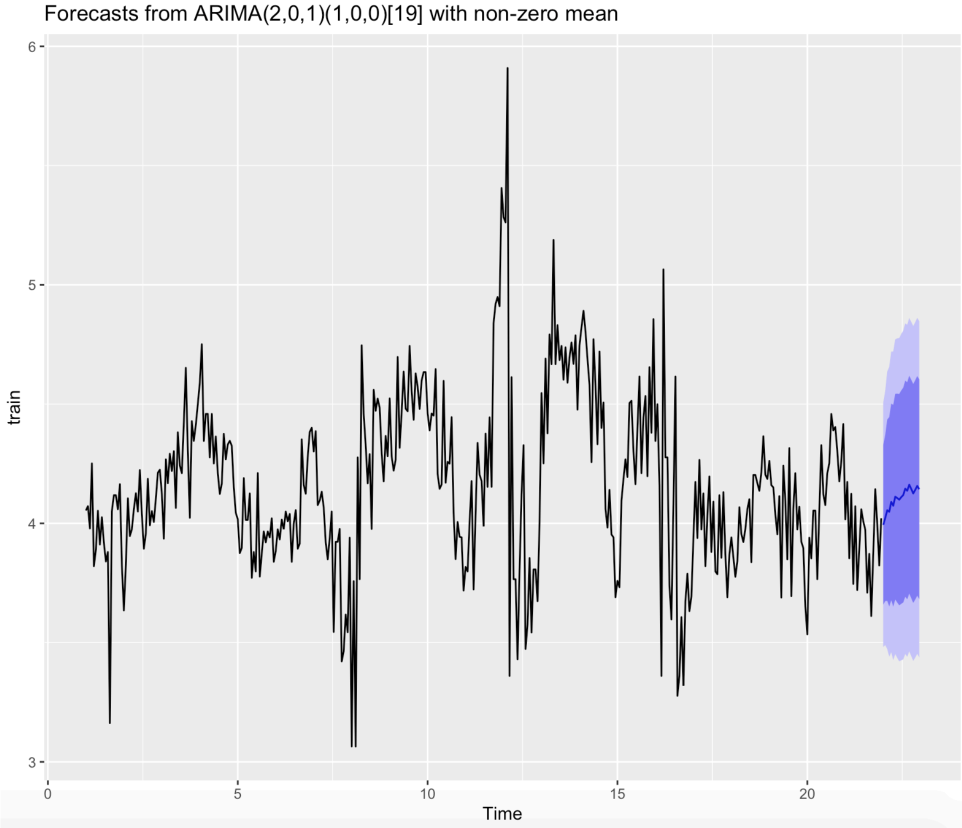
3. After applying the auto.arima on the model, the following statistics is obtained.

```
Series: train
ARIMA(2,0,1)(1,0,0)[19] with non-zero mean

Coefficients:
         ar1     ar2      ma1     sar1    mean
      0.5658  0.2903  -0.2733  -0.0823  4.1512
s.e.  0.1046  0.0787   0.1076   0.0509  0.0597

sigma^2 estimated as 0.06806:  log likelihood=-28
AIC=68.01   AICc=68.22   BIC=91.94
```

4. The residual plots can be seen below, they satisfy all the condition specified in [Stochastic model]. **NOTE:** The ACF plot of the residual can be seen out of the threshold, however it should be noted that the magnitude of out of threshold ACF is very less.



5. The forecast of the time-series value can be plotted as:

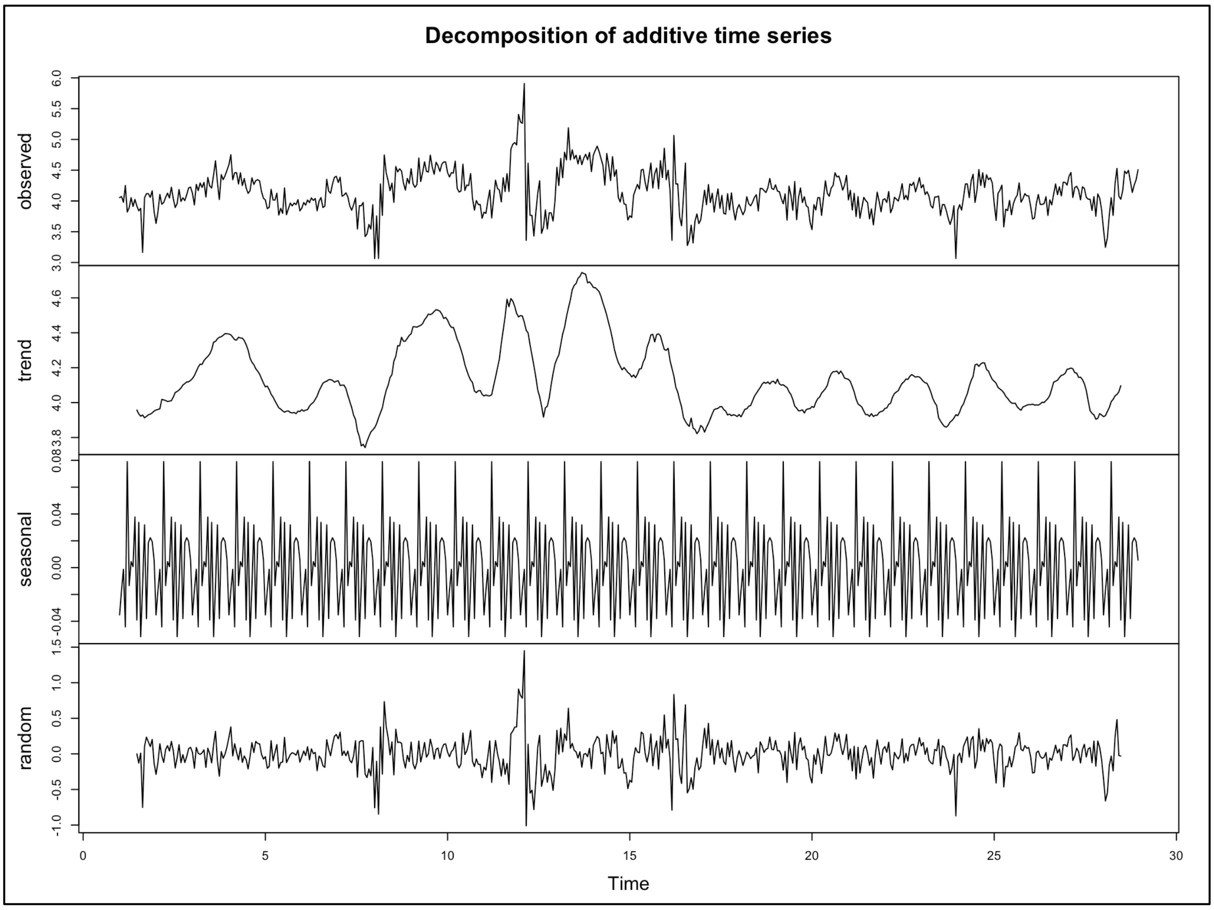


4.2 Non-stochastic mode

Seasonal Naïve method will be used as the non-stochastic model

4.2.1 Explanation

Seasonal Naïve method as this current data has hourly seasonal patterns. In this method, the forecast is set to the value of last observed value of the season that the current season is similar to. i.e. if the seasonality is monthly, the current value is set to the value of last month. In this case ,the seasonality is hourly, hence the values from the same last hour will be considered as the same as this hour.



4.2.2 Mathematical Detail

$$Y_T = Y_{T-m}$$

Where, m is the seasonal period; Y_T is the value that needs to be predicted

4.2.3 Implementation Detail

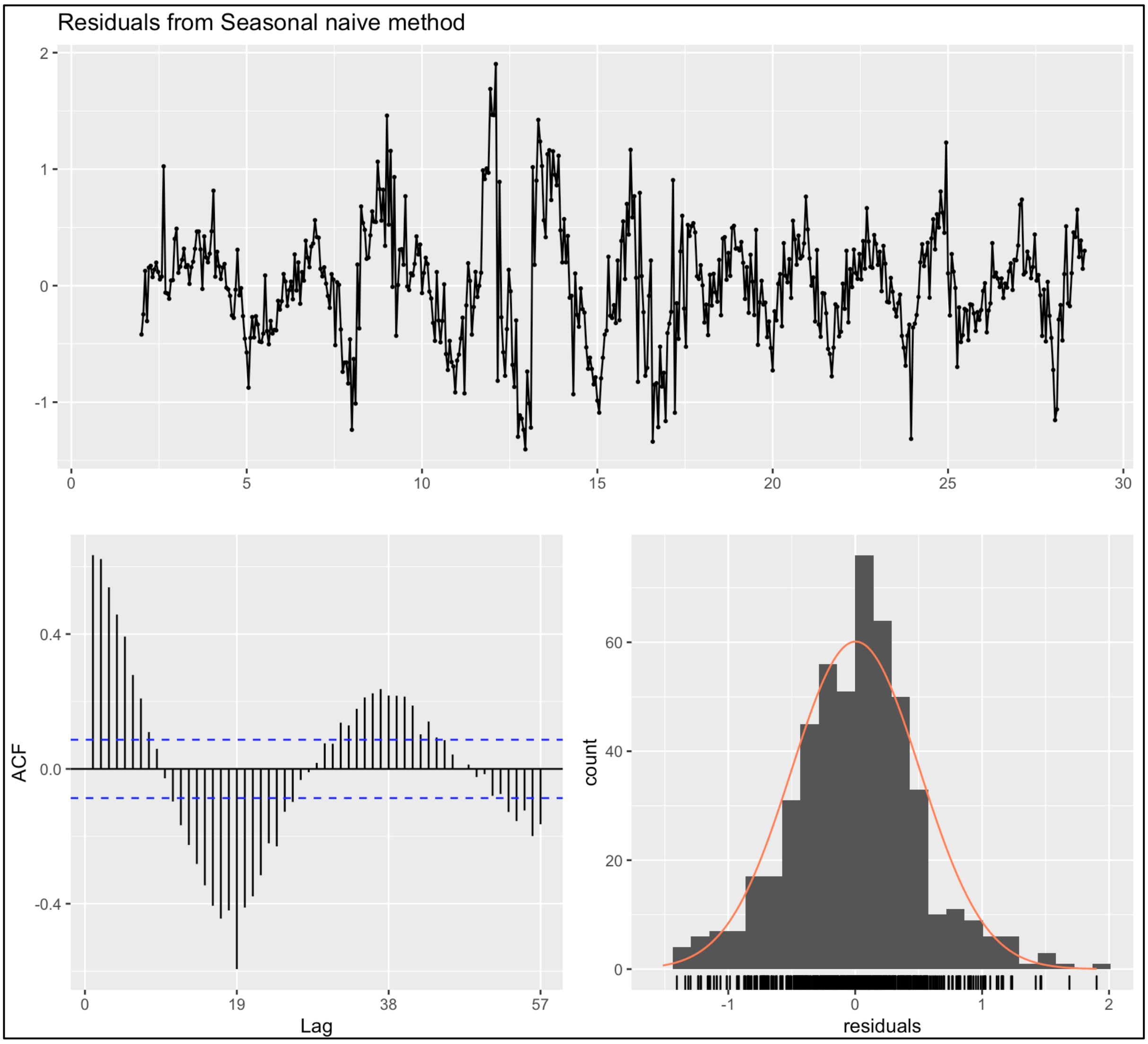
Seasonal naïve is suitable for the current data at hand, because of the following reason:

1. The series has frequency of data as 19, as shown in the seasonality plot
2. The data is repetitive at frequency of 19, as seen from the seasonality plot.

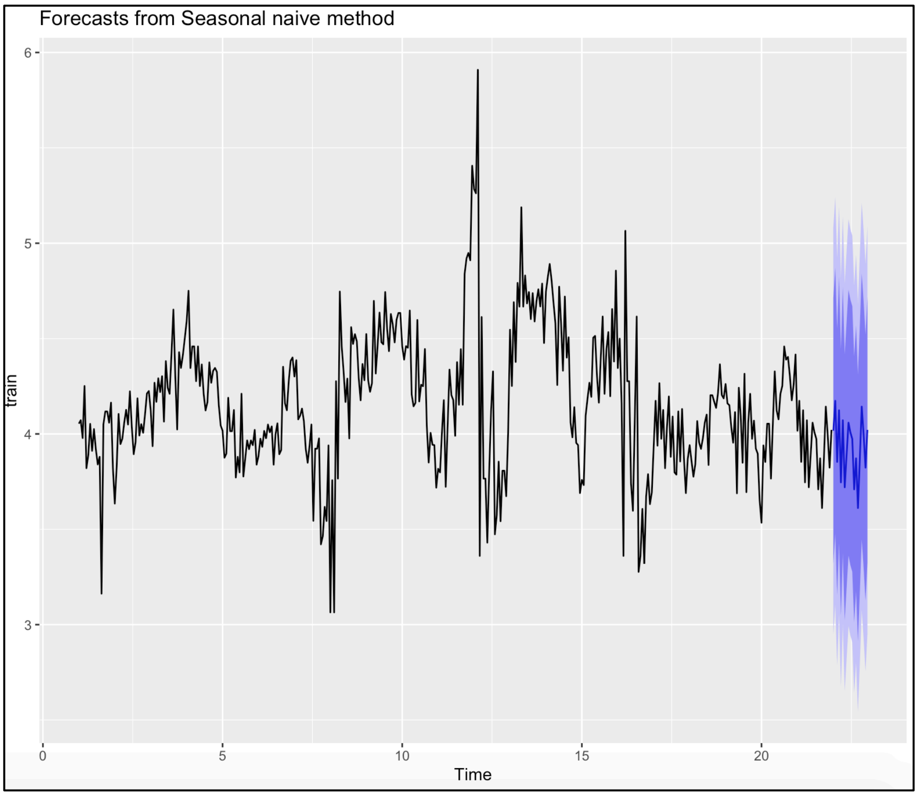
4.2.4 Fitting Details

The following steps are followed to get a seasonal naive model:

1. Divide the data into test and train, such that test data contains the last 7 * 19 values i.e. 19 values for each day of week
2. Ljung-Box test is not checked for this model as it is non-stocastic model
3. The residual plot is as follows:



4. The forecast can be seen as below:



Question – 5

Report full details of each fitted model

Answer – 5

Comparison and fitting detail of the two models:

Detail	SARIMA	SNAIVE
Type of model	Stochastic	Non-stochastic
Root mean square error (RMSE)	0.26	0.30
Mean percentage error (MPE)	-2.30	1.63
Mean of Residual	Close to zero	Close to zero
Variance	Constant	Constant
Residue plot	Approximately normal	Approximately normal

From the above table we can see that SARIMA is better as compared to SNAIVE model, as it has better accuracy measure(RMSE and MPE). The additional detail for SARIMA model.

1. Second order AIC (Akaike Information Criterion) or AICc: 68.22
2. Mean of residue: Close to zero
3. Ljung-Box test: 0.174
4. Variance of residuals: Constant residual plot
5. Normal residual plots: Normal distribution
6. Root mean square error (RMSE): 0.26
7. Residual plot can be seen in [Fitting Details]

Question – 6

Discuss the limitations of each model with respect to this dataset.

Answer- 6

SARIMA:

1. The ACF plot is not always within the threshold limits, indicating that the residuals are not completely white noise.
2. The p-value is about 0.127. As p-value is > 0.05 which does not confirm that the data is independent

SNAÏVE:

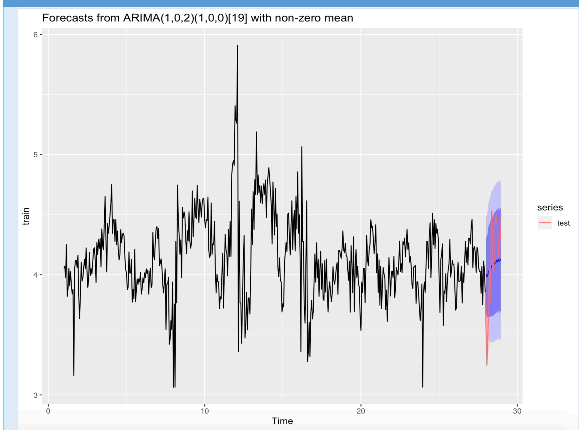
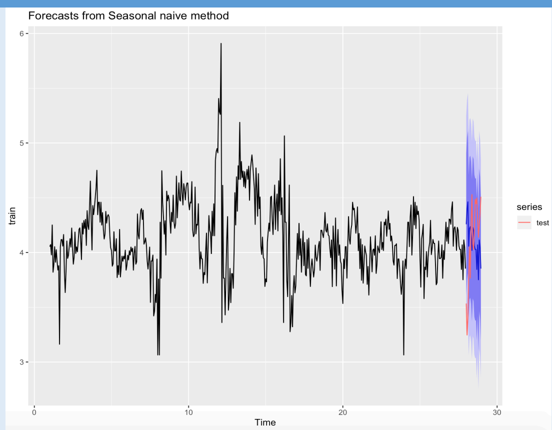
1. This model does not capture the effect of random white noise of the data
2. This model only depends on the value from the last similar hour. There could be a case when the entire data was not captured by translink, in such as case the forecast will be very different than the actual

Question – 7

Give and plot model predictions for each model over the observed data range (include the observed values somehow for comparison). Also give 95% predictive intervals for the stochastic model over this range

Answer – 7

For this question, the test data is the last 19 values of the time series.

SARIMA *	SNAIVE *
	

*These plots are before transforming the data back to normal

SARIMA:

Hour	Observed Data	Point Forecast	Low 95%	High 95%
1	10	4	2	6
2	20	19	12	32
3	23	47	27	81
4	49	70	40	124
5	103	54	30	97
6	164	52	29	96
7	131	61	33	113
8	138	75	40	140
9	148	82	43	154
10	146	80	42	151
11	122	104	55	198
12	140	112	59	213
13	226	237	124	454
14	139	215	112	413
15	115	107	56	206
16	73	72	37	138
17	90	65	34	126
18	59	46	24	89
19	14	15	8	28

Question – 8

Evaluate accuracy 1 and 2 hours ahead via final day.

Answer – 8

Details of calculation:

1. To do this calculation, the train data is take 1 hour or 2 hour before the values that needs to be predicted
2. The test data is taken to be, 1 hour or 2 hour that needs to be predicted
3. Using the SARIMA and SNAIVE model, the next 1 hour or 2 hours is predicted
4. These steps are performed in a loop for 1 hour or 2 hour
5. RMSE is calculated on comparing the test data and the forecasted data

Details of output:

1. In accuracy for hours ahead, SARIMA has an average of 25% error in RMSE. On the other hand, SNAIVE has an average of 43% error.
2. For 2-hours ahead, SARIMA has an average of 31% error in RMSE. On the other hand, SNAIVE has an average of 45% error.
3. There is 35% of the results where SNAIVE performs better as compared to SARIMA

Accuracy hours ahead:

Hour	RMSE SARIMA	RMSE SNAIVE	SARIMA < SNAIVE
1	47%	72%	TRUE
2	59%	115%	TRUE
3	13%	106%	TRUE
4	31%	29%	FALSE
5	50%	17%	FALSE
6	0%	47%	TRUE
7	45%	10%	FALSE
8	54%	51%	FALSE
9	15%	15%	FALSE
10	20%	18%	FALSE
11	3%	11%	TRUE
12	35%	46%	TRUE
13	18%	42%	TRUE
14	11%	65%	TRUE
15	11%	25%	TRUE
16	24%	39%	TRUE
17	3%	14%	TRUE
18	14%	30%	TRUE
19	24%	66%	TRUE

Accuracy 2 hours ahead

Hour	RMSE SARIMA	RMSE SNAIVE	SARIMA < SNAIVE
1, 2	62%	96%	TRUE
3,4	21%	78%	TRUE
5,6	38%	35%	FALSE
7,8	59%	37%	FALSE
9,10	21%	16%	FALSE
11,12	26%	33%	TRUE
13,14	17%	55%	TRUE
15,16	24%	32%	TRUE
17,18	11%	24%	TRUE

Question – 9

Make predictions for test day, which will be evaluated by RMSE vs true counts for that day (held by lecturer)

Answer – 9

1. From the above questions, it can be deduced that the performance of SARIMA model is better than that of SNAÏVE model.
2. To predict the values from this models, first SARIMA model is used
3. From the prediction obtained from above, seasonality is added for Monday at hourly basis. As specified in [3.1 Seasonality analysis]
4. From the output of 3, the series is raised to the power of exponent, as initially I have applied log to make the series multiplicative.

Hour	Forecast
6	4
7	25
8	64
9	89
10	66
11	61
12	71
13	82
14	88
15	88
16	113
17	118
18	244
19	217
20	108
21	71
22	67
23	46
24	14

Question – 10

Include a paragraph aimed at a member of the Translink planning staff who may not have a statistics background, explaining how your modelling could potentially help them make decisions about how many bus/train/ferry services to run at various times to meet demand.

Answer – 10

Using this analysis, I made an attempt to model the flow of customers in TransLink services. From the various statistical test and checks performed(like Ljung-Box test, AICc value, plot diagnostics), it can be concluded that this model is statistically significant. This model will help Translink in future planning of the flow of customers. This model determines the number of customers that will use Translink services per hour. This model like any other machine learning model has some error. However, from [SARIMA: Answer – 7] it can be seen that the flow is always higher than the 95% lowest flow and always lower than the 95% highest flow. The model accuracy is close to about 70%, which means that the model is right about 70% of the time in predicting the approximate flow. Lastly, as seen from [Stochastic model] and [Answer – 5], it can be seen that SARIMA model is the best model across methods (stochastic and non-stochastic) for this time series.

Part – I

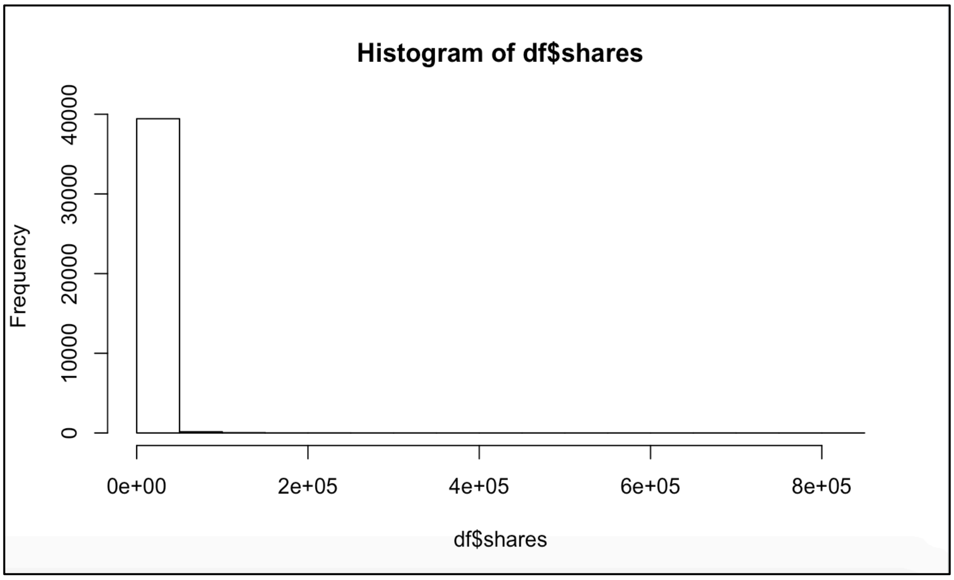
Question – 1

Choose a suitable generalised linear model and suitable link function and explain these decisions. Fit the model to the data and report summary results. [2 marks]

Answer – 1

Poisson & Negative binomial GLM are suitable link functions, as:

1. The value of prediction variable i.e. shares is always greater than or equal to zero
2. The value of prediction variable is discrete data
3. The distribution of prediction variable is not normal. The distribution of shares can be seen as:



4. The mean of the response variable is 3395.38, while the variance of response variable is 135185984. This suggests that the variance is much greater than the mean.

5. Preliminary model statistics also suggests presence of over-dispersion i.e. mean << variance. In a model this is calculated by (Residual deviance > Degree of freedom)
6. These GLM models the event occurring in the specific time frame. In this use case as well, the number of times an online news is shared is calculated during two years from 7th January 2013

Result comparison between the Poisson & Negative binomial GLM:

To calculate these results, the outlier treatment is done and the values having high variance is removed. Variable transformations have also been applied as done in Assignment -1

GLM Model	Goodness of fit	AIC	RMSE	Over-Dispersion (Residual deviance > Degree of freedom)
Poisson GLM	11.53%	233,800,000	7682.413	Ture
Negative Binomial GLM	14.87%	716,300	7667.736	True
*** Gaussian GLM** (Transformation as in Assignment -1)	11.73%	-153,100	7790.447	False

* Gaussian GLM gives the same results as the Linear regression model used in Assignment - 1
From the above table, it can be concluded that Negative Binomial GLM gives the best results, while comparing the GLM models. The model statistics from Gaussian GLM is obviously better, as all the explanatory variables are according to this model. Hence, comparing Gaussian GLM is not right.

Negative Binomial is better than Poisson GLM because the variance of this data(135185984) is very large as compared to the mean of the data(3395.38), this is also know as over-dispersion. Negative Binomial are more flexible to deal with over-dispersion as compared to Poisson GLM as they adjust the variance independent of variance

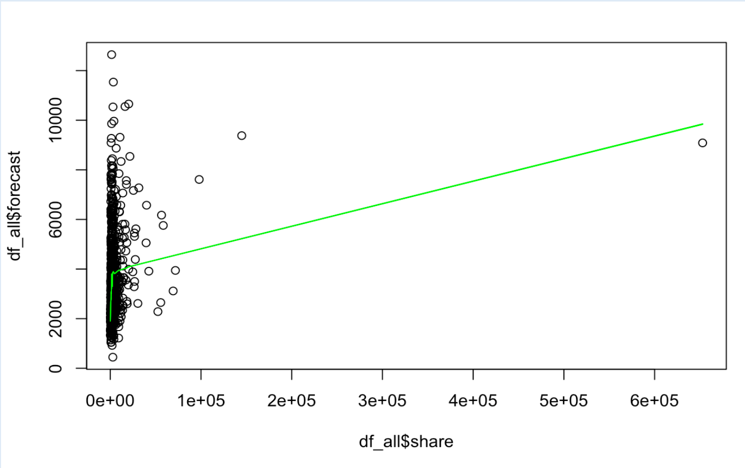
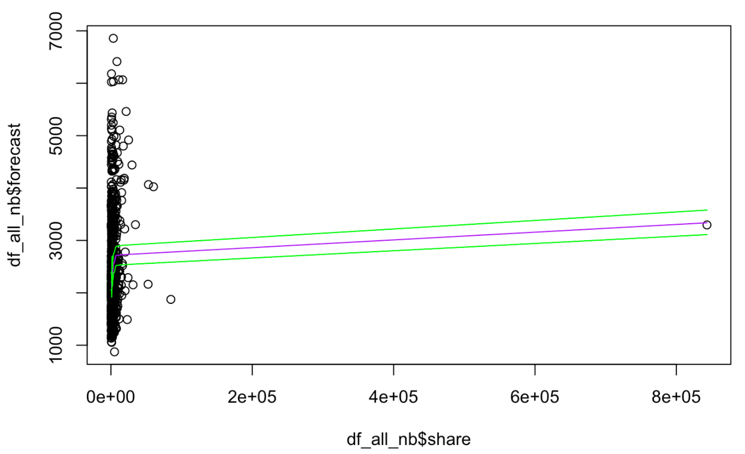
Question – 2

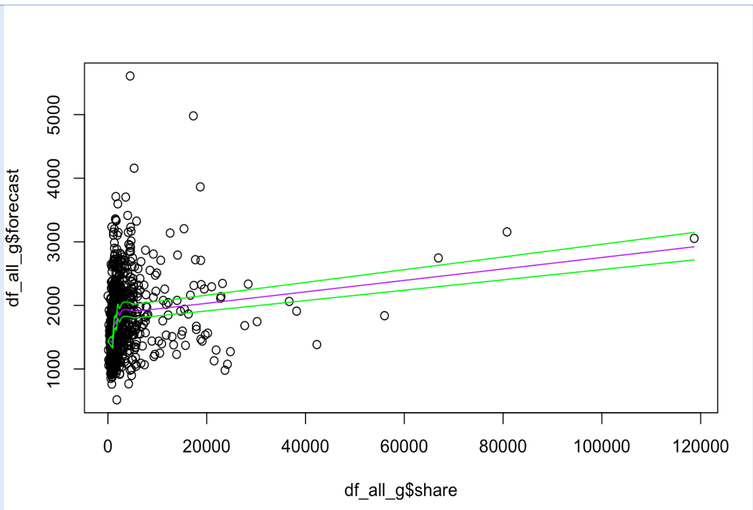
Produce a 95% predictive interval for each fitted model (ignore uncertainty with respect to model parameters). Compare the results from using this model with those from using a multiple linear regression model (= general linear model) which you would have run for assignment 1. Include consideration of significant variables, measures of goodness of fit, predictive intervals and related plots. Discuss theoretical and practical differences between the two models.

Answer – 2

Prediction interval

The 95% prediction interval for each fitted model can be seen as

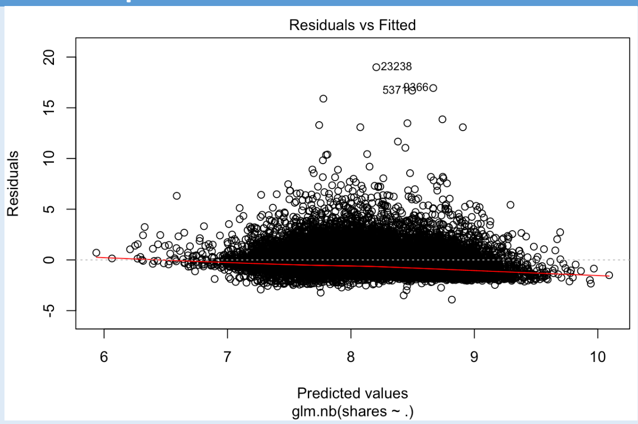
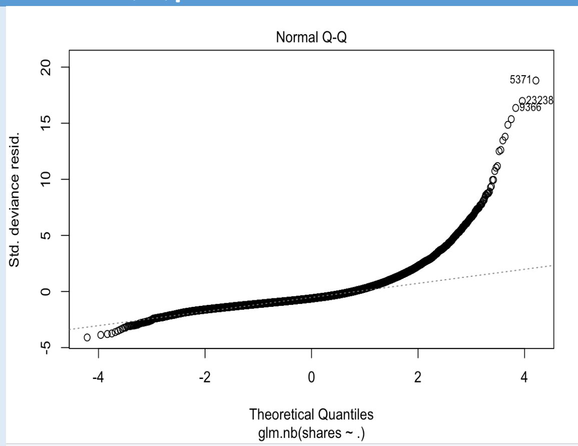
Model	Prediction interval
Poisson GLM **Note: Upper limit and Lower limits are approximately same. Which is why they seem to be overlapped in the image**	
Negative Binomial GLM	

Gaussian GLM (Transformation as in Assignment -1)	

Comparing results, the Negative Binomial GLM with Linear regression

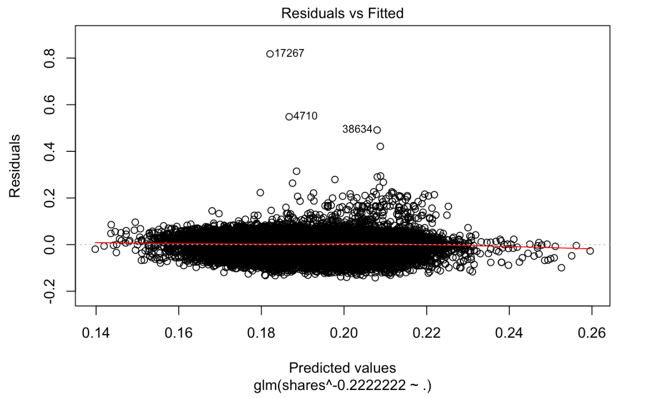
In the following table, the outlier treatment, high variance removal and data transform is done as detailed in Assignment – 1

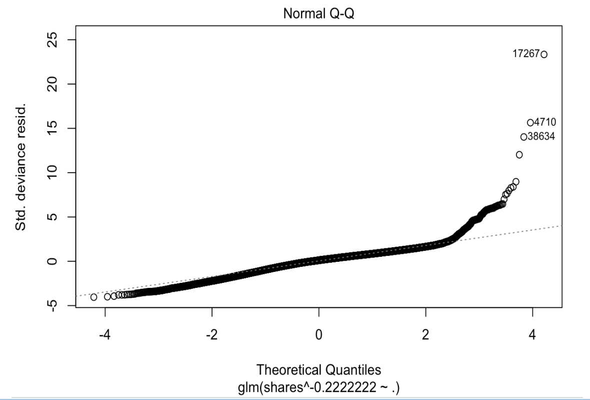
GLM Model	Goodness of fit	AIC	RMSE	Over-Dispersion (Residual deviance > Degree of freedom)
Negative Binomial GLM	14.87%	716,300	7,667.7	True
*** Gaussian GLM** (Transformation as in Assignment -1)	11.73%	-153,100	7, 790.4	False

Model	Residual plots	Normal Q-Q plot
Negative Binomial GLM		

***Gaussian GLM**

(Transformation as in Assignment -1)





From these two tables, the following conclusions can be made:

1. The goodness of fit for Negative binomial GLM(NB-GLM) is better as compared to Multiple Linear regression (MLR)
2. RMSE values for both the models are comparable
3. The AIC value for MLR is very less as compared to the AICc value of NR-GLM value. This is because, the data transformations are very specific to MLR
4. From the residual diagnostics, it can be seen that both NB-GLM and MLR have similar kind of plots for fitted residuals.
5. The Normal Q-Q plot in NB-GLM is highly influenced by the outliers as compared to the MLR Normal Q-Q plot
6. Significant variables: The significant variables are entirely different for the two models.
    a. NB-GLM: global_subjectivity, global_rate_negative_words, kw_max_avg, global_sentiment_polarity, abs_title_subjectivity
    b. MLR: weekday_is_Wednesday, n_unique_tokens , data_channel_is_entertainment, global_sentiment_polarity, LDA_02
7. Predictive intervals: The predictive interval graph does not give much information due to the spread of the data

Resources:

https://www.dataquest.io/blog/tutorial-poisson-regression-in-r/
https://stats.stackexchange.com/questions/60643/difference-between-binomial-negative-binomial-and-poisson-regression
https://otexts.com/fpp2/intro.html
https://www.youtube.com/watch?v=LaVyUoTqM90&t=5s
https://stats.idre.ucla.edu/r/dae/negative-binomial-regression/
https://stackoverflow.com/questions/3480388/how-to-fit-a-smooth-curve-to-my-data-in-r
https://www.r-bloggers.com/is-my-time-series-additive-or-multiplicative/
https://www.r-bloggers.com/prediction-intervals-for-poisson-regression/