DATA 7202 Assignment 2 Report

Name: Yupeng Wu
No: 45960600

Part One:

1.

From the ?family command in Rstudio, we can find out there are several families can be used in the generalized linear model. They are “binomial”, “gaussian”, “gamma”, “inverse-gaussian”, “poisson”, “quasi”, “quasi-binomial”, “quasi-poisson” and “negative-binomial” from the lecture.

In order to use a non-negative distribution, I chose poisson, gaussian, gamma and negative-binomial with different link function to find a suitable one. The results are in the following table.

Distribution	Link Function	AIC	R squared
Poisson	sqrt	228651442	0.1131
Gaussian	identity	836541	0.0209
Gamma	log	701268	0.1630
Negative-binomial	sqrt	700442	0.1720

From the above data, we can find out the negative-binomial has the minimum AIC and a maximum R squared, which shows that this model is the best one.

```
glm.nb(formula = shares ~ ., data = ds, link = "sqrt", init.theta = 1.054884963)

Deviance Residuals:
                              Min       1Q      Median      3Q       Max
                          -3.9023  -0.9527   -0.5951   -0.0937   20.8643

Coefficients:
                               Estimate    Std.Error   z value   Pr(>|z|)
(Intercept)                    2.89E+01    2.36E+00    12.222   <2.0E-16
n_tokens_title                 7.77E-01    6.39E-02    12.155   <2.0E-16
n_tokens_content              -1.87E+00    2.01E-01    -9.327   <2.0E-16
num_hrefs                      2.48E-01    1.73E-02    14.349   <2.0E-16
num_self_hrefs                -1.96E+00    2.17E-01    -9.003   <2.0E-16
num_imgs                       2.54E-01    2.13E-02    11.954   <2.0E-16
num_videos                     3.61E-01    4.37E-02     8.248   <2.0E-16
num_keywords                   3.76E-01    8.45E-02     4.452    8.50E-06
data_channel_is_lifestyle     -6.87E+00    1.01E+00    -6.813    9.53E-12
data_channel_is_entertainment -1.29E+01    6.61E-01   -19.571   <2.0E-16
data_channel_is_bus           -1.05E+01    9.32E-01   -11.263   <2.0E-16
data_channel_is_socmed        -5.08E+00    9.55E-01    -5.317    1.06E-07
data_channel_is_tech          -5.60E+00    8.99E-01    -6.231    4.64E-10
data_channel_is_world         -8.76E+00    8.99E-01    -9.741   <2.0E-16
kw_min_max                    -1.97E+00    8.10E-02   -24.346   <2.0E-16
kw_max_max                    -2.26E+00    4.48E-01    -5.056    4.27E-07
```

kw_avg_max	-7.71E-01	4.77E-01	-1.618	0.105673
kw_min_avg	8.84E-03	3.53E-04	25.007	<2.0E-16
kw_max_avg	7.43E+00	3.17E-01	23.408	<2.0E-16
self_reference_min_shares	3.98E-04	2.23E-05	17.87	<2.0E-16
self_reference_max_shares	9.22E-05	7.33E-06	12.58	<2.0E-16
weekday_is_monday	-1.70E+00	6.38E-01	-2.666	0.007675
weekday_is_tuesday	-4.42E+00	6.22E-01	-7.107	1.18E-12
weekday_is_wednesday	-4.17E+00	6.22E-01	-6.701	2.06E-11
weekday_is_thursday	-4.25E+00	6.23E-01	-6.815	9.43E-12
weekday_is_friday	-3.74E+00	6.45E-01	-5.8	6.62E-09
weekday_is_saturday	2.17E+00	8.13E-01	2.669	0.007605
LDA_00	5.38E+00	1.40E+00	3.835	0.000126
LDA_01	2.59E+00	1.49E+00	1.743	0.081355
LDA_02	-8.41E+00	1.33E+00	-6.349	2.16E-10
LDA_03	6.23E+00	1.47E+00	4.238	2.26E-05
global_subjectivity	3.35E+01	2.59E+00	12.924	<2.0E-16
global_sentiment_polarity	1.24E+01	1.62E+01	0.768	0.44273
global_rate_positive_words	-7.09E+01	1.05E+01	-6.778	1.22E-11
global_rate_negative_words	-5.40E+00	1.60E+01	-0.338	0.735161
avg_positive_polarity	-1.20E+01	2.61E+00	-4.611	4.00E-06
min_positive_polarity	-1.11E+01	2.44E+00	-4.532	5.83E-06
max_positive_polarity	3.20E+00	9.60E-01	3.338	0.000845
avg_negative_polarity	3.00E+00	2.71E+00	1.11	0.267196
min_negative_polarity	-2.83E+00	9.95E-01	-2.844	0.004453
max_negative_polarity	-6.77E+00	2.41E+00	-2.811	0.004939
title_subjectivity	2.12E-01	6.50E-01	0.327	0.743899
title_sentiment_polarity	9.43E-01	6.08E-01	1.552	0.120571
abs_title_subjectivity	4.45E+00	8.50E-01	5.235	1.65E-07
abs_title_sentiment_polarity	5.00E+00	9.53E-01	5.246	1.55E-07

2.

Variables:

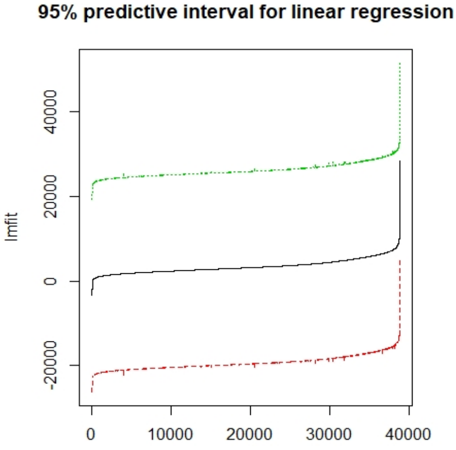
From the model summary in the negative-binomial, there are 34 variables which p-value is very small and has three stars. While the multiple linear regression model only has 8 variables. This shows that there are more significant variables in the generalized linear model.

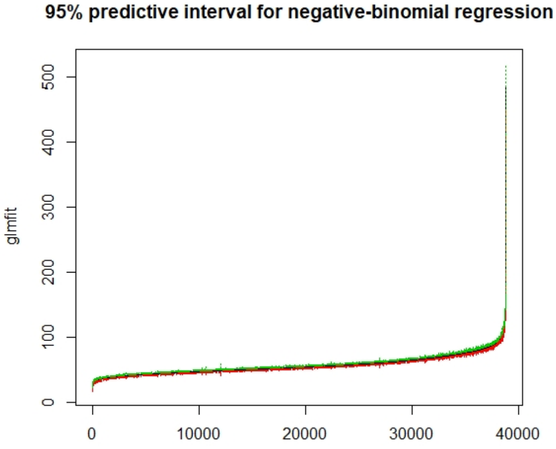
Goodness of fit:

	AIC	R squared
Negative binomial	700442	0.1720
Multiple linear regression	836542	0.0209

The generalized linear model with higher AIC and lower R squared is a better model.

Predictive Intervals:

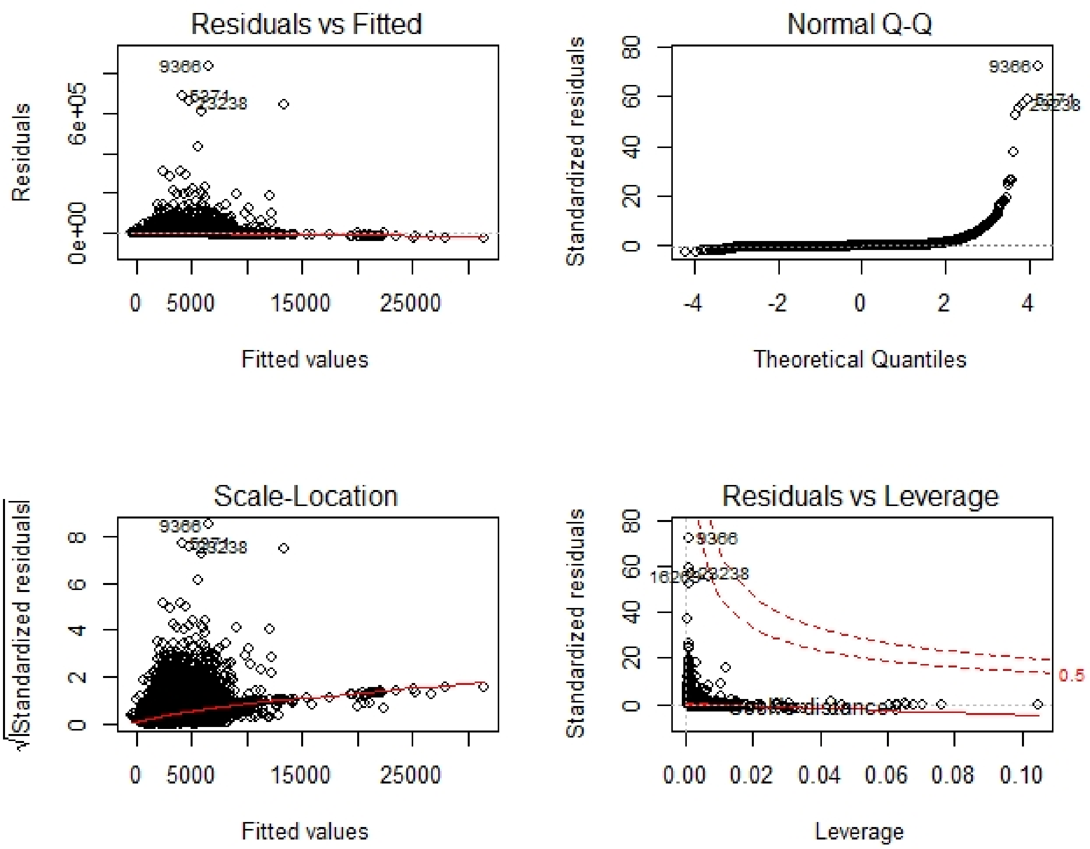




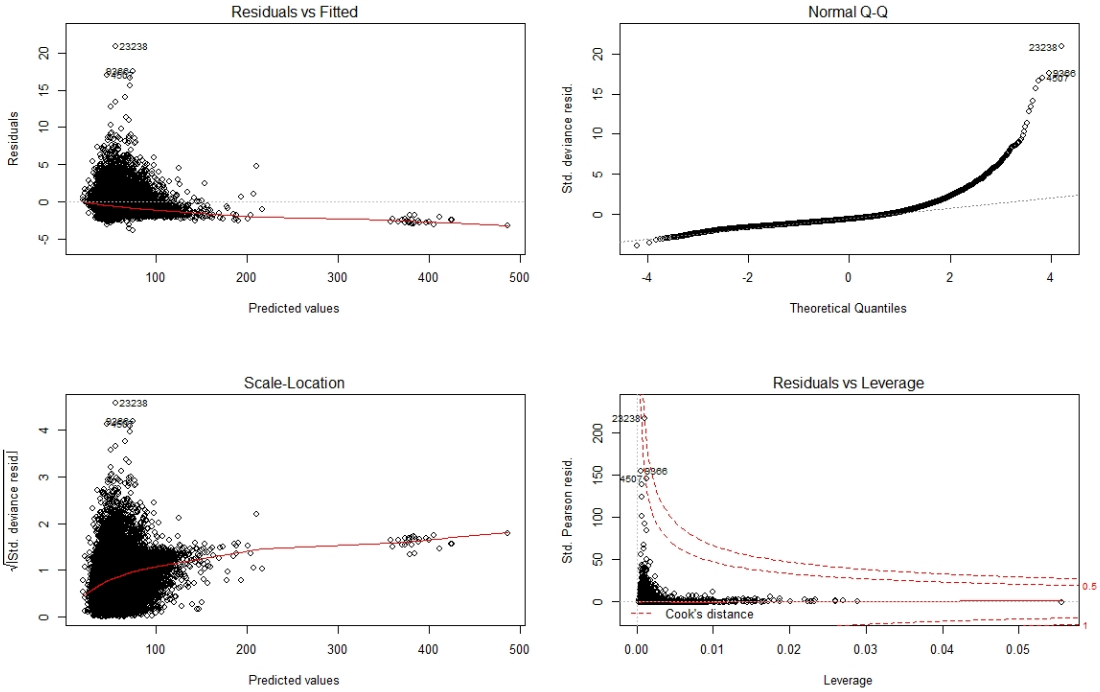
The width of the predictive interval of the negative-binomial regression is narrower than the width of PI of the multiple linear regression. This is because the negative-binomial regression has a more accurate prediction.

Residual:

Multiple Linear Regression



Negative-binomial Regression (link='sqrt')



The main difference in the previous plots are in the Q-Q plot. From the negative-binomial regression, more data is located near the red line, which means these data are more likely belongs to Normal distribution.

However, the red line in the residuals plot in negative-binomial does not horizontal at zero. This shows that in generalized linear regression the data does not fit the linearity assumption so much. In theory, it should be fitted to the linearity assumption and most of the points should equally separate in the both sides of the red line. So, this model still has a low R squared value and can not predict the goals with high accuracy.

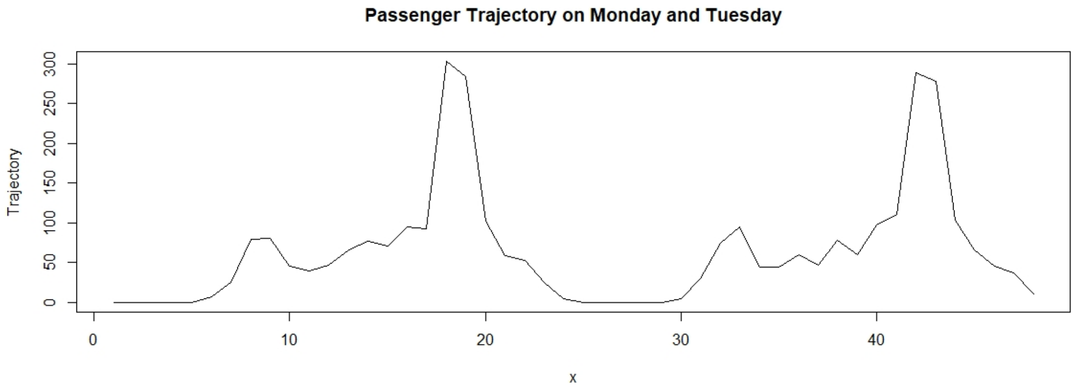
Part Two:

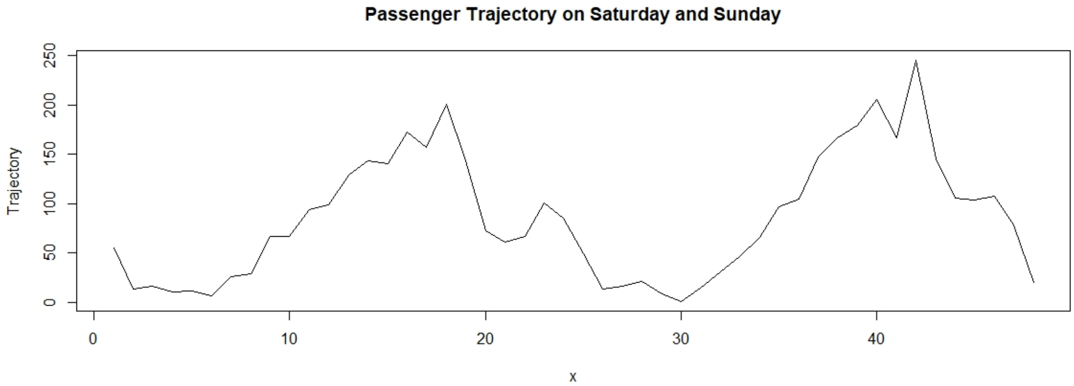
3.

Data preparation:

After reading the v0_num_traj data from region 1 to region 5, there are two important observations. ① The data only has 19 hours for one day. ② There is a big difference between the workday plot and the weekend plot.

For the first observation, I tried to make up some zeros to the dataset to make sure each day has a 24 hours data record. But this is not a good idea since there will be some negative forecast points from the ARIMA model. Now I'm using a data with 19 records per day and no missing data.

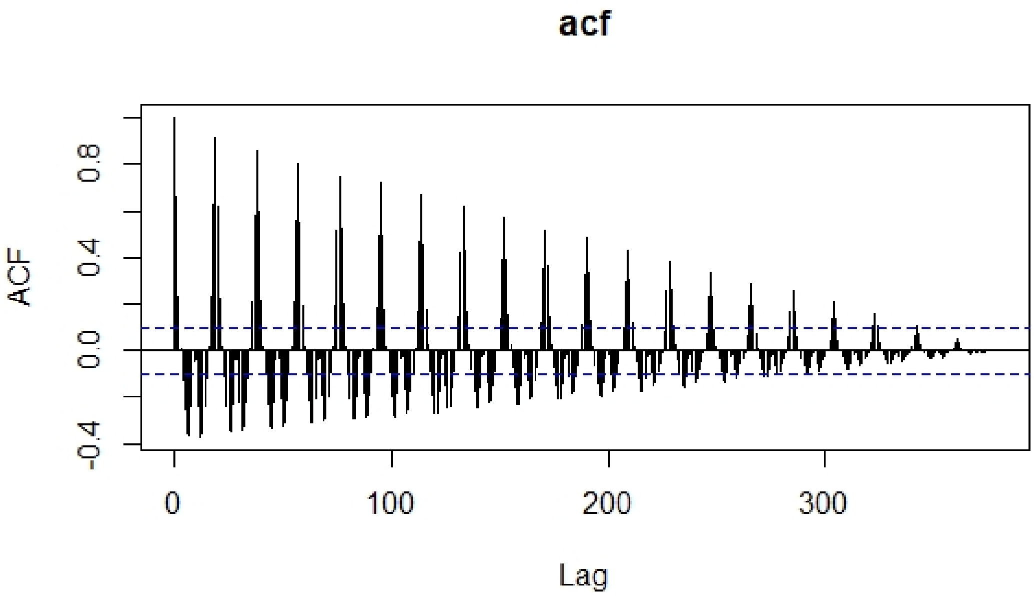




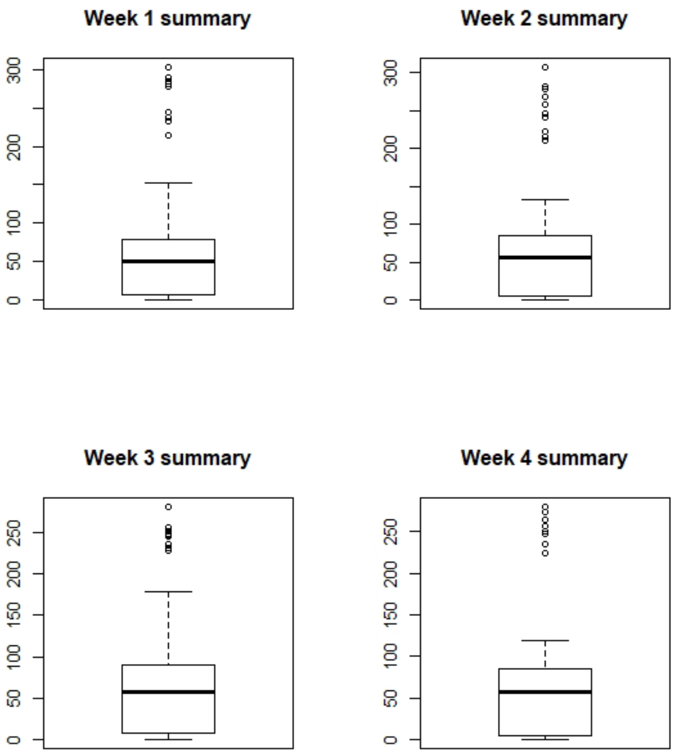
As shown in the above Figure, the times series for the weekdays and weekends have a big difference, which may have a big impact on the model. So, I decided to drop all the data from Saturdays and Sundays in order to have a more accurate prediction on the weekday.

Finally, my training data has $19 * 5 * 4 = 380$ rows.

Stationarity:



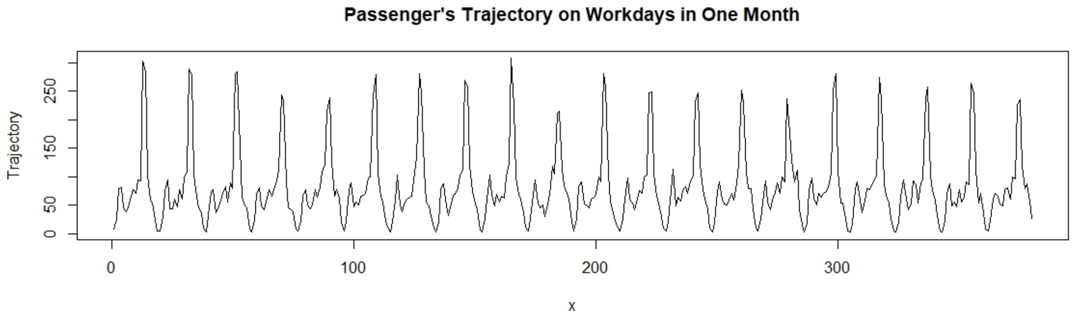
From the ACF plot, we can see that finally it is converges to zero, which shows this data is stable.



Also, from the boxplot for the four weeks' data, the mean and variance are very similar. This shows that each week's data is almost the same, which represents the stationarity.

Finally, the adf.test shows the time series data have a very small p-value, which means it has stationary.

Seasonality:



From the trajectory plot, we can easily find out this time series has a seasonality. Each day is a period, so we have a frequency of 19 in the data.

Besides, two functions isSeasonal(·) and wo(·) from library "seastests" also show that this time series has a seasonality.

4.

In order to fit the seasonal continuous data, I choose SARIMA, seasonal naïve and holt winters.

Seasonal ARIMA:

Definition: A time series $\{X_t\}$ is called an autoregressive integrated moving average (ARIMA) process with order p, d, and q, denoted $\{X_t\} \sim ARIMA(p,d,q)$, if its d-th order difference

$$Z_t = (1-B)^d X_t$$

is a stationary $ARMA(p,q)$ process, where $d \geq 1$ is an integer:

$$b(B)(1-B)^d X_t = a(B)\varepsilon_t$$

$\{X_t\}$ is called ARIMA as it is the integration of a differenced series $\{Z_t\}$.

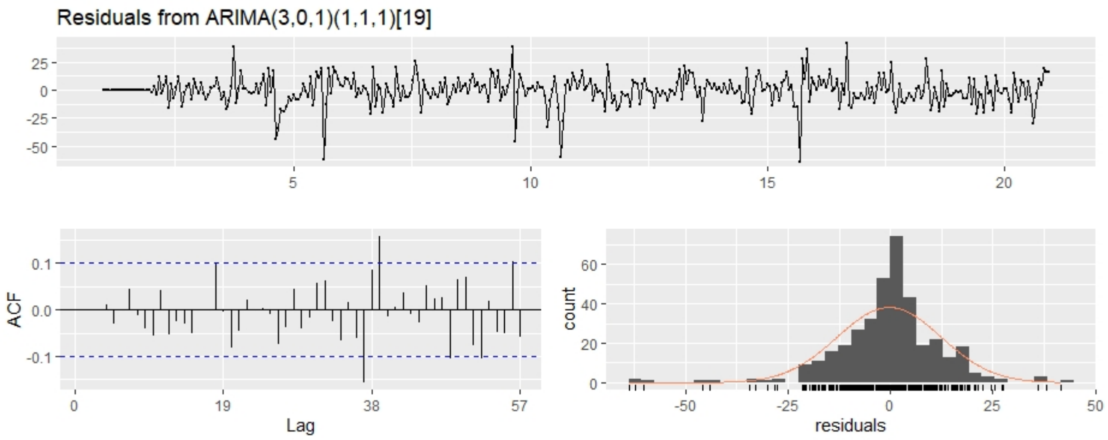
The $ARIMA(p,d,q) \times (ps,ds,qs)_s$ is a seasonal model, which has a process:

$$(1-\beta_1 B - \cdots - \beta_p B^p) \times (1-\beta_1^* B^s - \cdots - \beta_{p_s}^* (B^s)^{p_s} \times (\delta^d(\delta_s^{d_s} X_t) - \mu)$$

$$= (1+\alpha_1 B + \cdots + \alpha_q B^q) \times \left(1+\alpha_1^* B^s + \cdots + \alpha_{q_s}^* (B^s)^{q_s}\right)\varepsilon_t$$

Here I used $ARIMA(3,0,1) \times (1,1,1)$ model to fit this data set. And the summary of this model is shown as below.

```
Coefficients:
         ar1      ar2      ar3      ma1     sar1
      0.7753  -0.1874  -0.1117  -0.5774   0.0080
s.e.  0.2069   0.0850   0.0696   0.2047   0.0648
        sma1
     -0.8712
s.e.  0.0539

sigma^2 estimated as 168.4:  log likelihood = -1450.94,
 aic = 2915.88
```



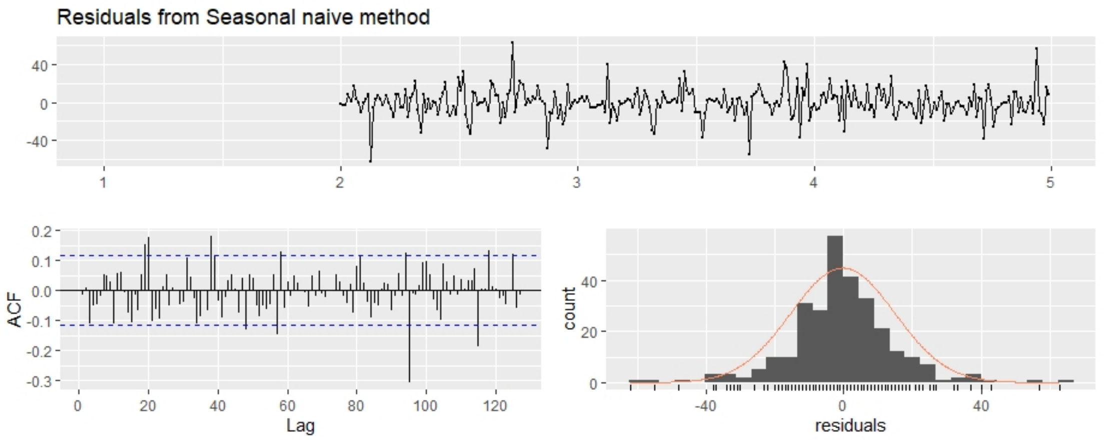
Seasonal naïve:

A useful method for highly seasonal data is seasonal naïve. We set each forecast to be equal to the last observed value from the same season of the year. Formally, the forecast for time $T + h$ is written as

$$\hat{y}_{T+h|T} = y_{T+h-m(k+1)}$$

Where m is the seasonal period and k is the integer part of $\frac{h-1}{m}$. And the summary of this model is shown as below.

```
Forecast method: Seasonal naive method

Model Information:
Call: snaive(y = ts0)

Residual sd: 17.9113

Error measures:
                     ME     RMSE      MAE       MPE
Training set 0.03878116 17.88653 12.34349 -3.450466
                 MAPE MASE      ACF1
Training set 20.78761    1 0.1993623
```



Holt Winters:

The component form for the additive method is:

$$\hat{y}_{t+h|t} = l_t + hb_t + s_{t+h-m(k+1)}$$
$$l_t = \alpha(y_t - s_{t-m}) + (1-\alpha)(l_{t-1} + b_{t-1})$$
$$b_t = \beta^*(l_t - l_{t-1}) + (1-\beta^*)b_{t-1}$$
$$s_t = \gamma(y_t - l_{t-1} - b_{t-1}) + (1-\gamma)s_{t-m}$$

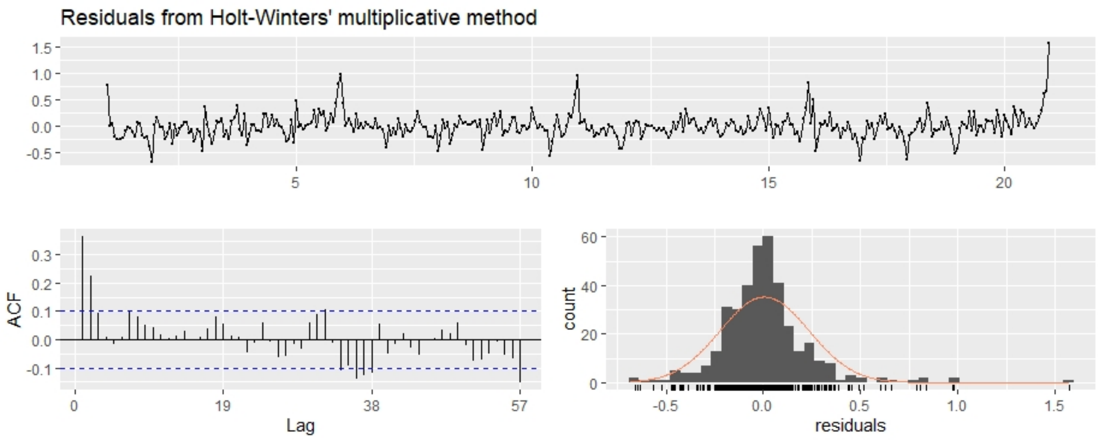
Where k is the integer part of $\frac{h-1}{m}$, which ensures that the estimates of the seasonal indices used for forecasting come from the final year of the sample. The level equation shows a weighted average between the seasonally adjusted observation $y_t - s_{t-m}$ and the non-seasonal forecast $l_{t-1} + b_{t-1}$ for time t. The trend equation is identical to Holt's linear method. The seasonal equation shows a weighted average between the current seasonal index, $y_t - l_{t-1} - b_{t-1}$, and the seasonal index of the same season last year (i.e., mm time periods ago).

The component form for the multiplicative method is:

$$\hat{y}_{t+h|t} = (l_t + hb_t)s_{t+h-m(k+1)}$$
$$l_t = \alpha\frac{y_t}{s_{t-m}} + (1-\alpha)(l_{t-1} + b_{t-1})$$
$$b_t = \beta^*(l_t - l_{t-1}) + (1-\beta^*)b_{t-1}$$
$$s_t = \gamma\frac{y_t}{l_{t-1} + b_{t-1}} + (1-\gamma)s_{t-m}$$

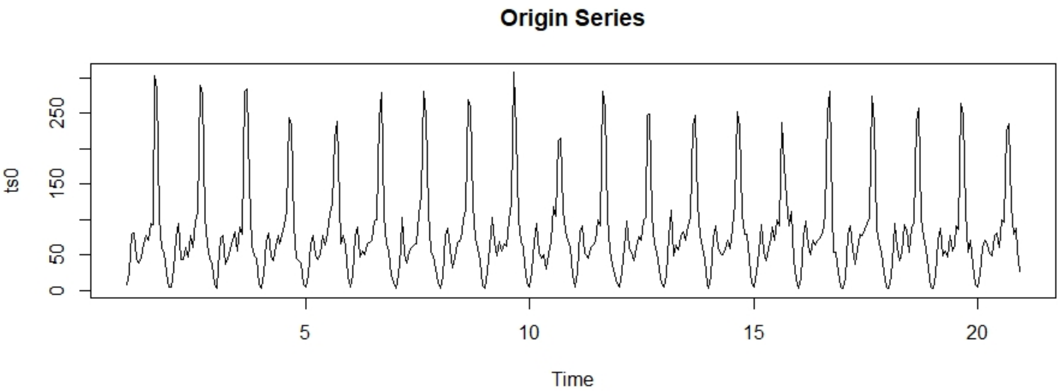
Here I choose the multiplicative method since it has a better performance. And the summary of this model is shown as below.

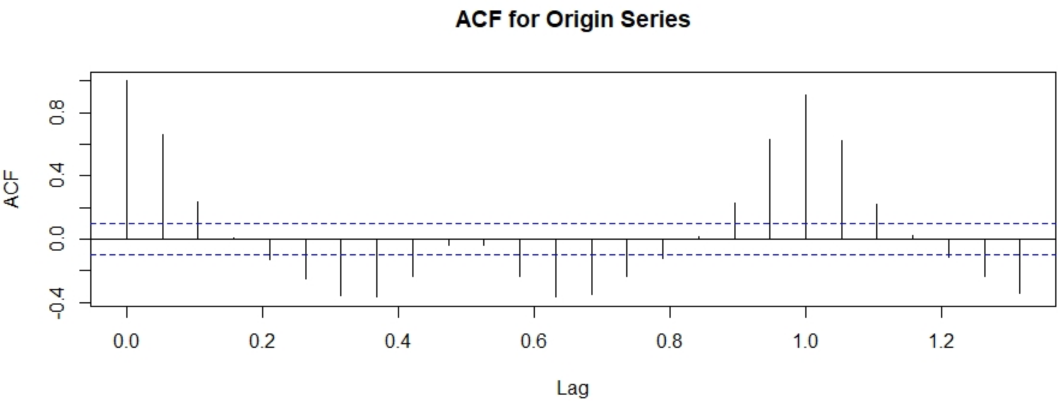
```
Forecast method: Holt-Winters' multiplicative method

Model Information:
Holt-Winters' multiplicative method

Call:
 hw(y = ts0, seasonal = "multiplicative")

  Smoothing parameters:
    alpha = 0.0147
    beta  = 0.0012
    gamma = 1e-04

  Initial states:
    l = 86.4597
    b = 0.1448
    s = 0.1347 0.4637 0.737 0.8298 1.3752 3.0591
           3.1871 1.2847 1.1939 0.838 0.8769 0.8077 0.65
81 0.5852 0.6662 1.1165 0.8562 0.2847 0.0452

  sigma:  0.2356

     AIC     AICc      BIC
4259.566 4262.946 4354.130
```



5.
Seasonal ARIMA:
Building the model:

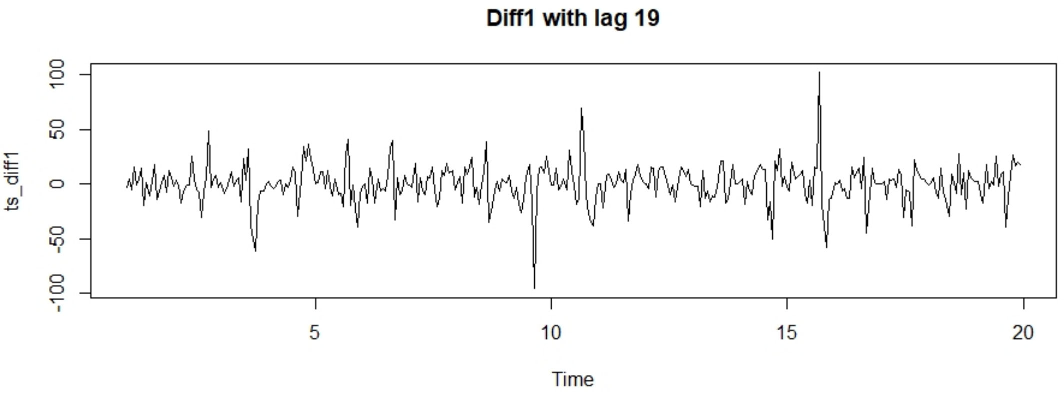
In order to get the parameters for the ARIMA model, we need acf and pacf plots.

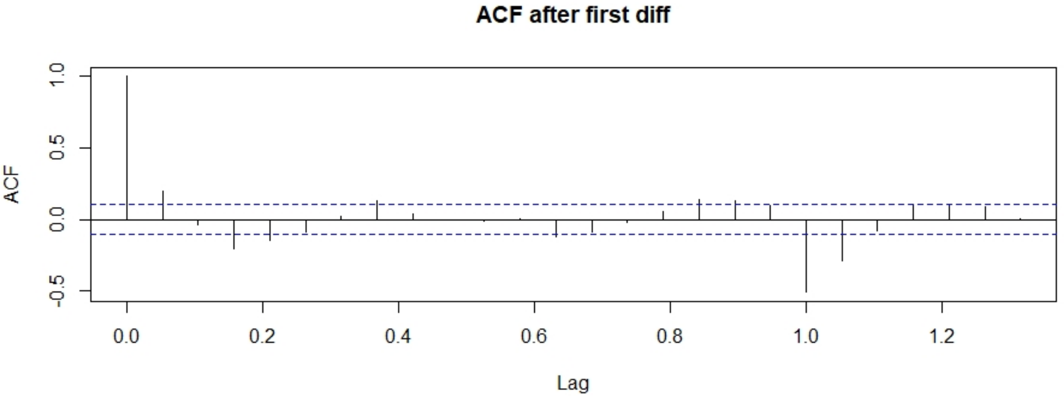


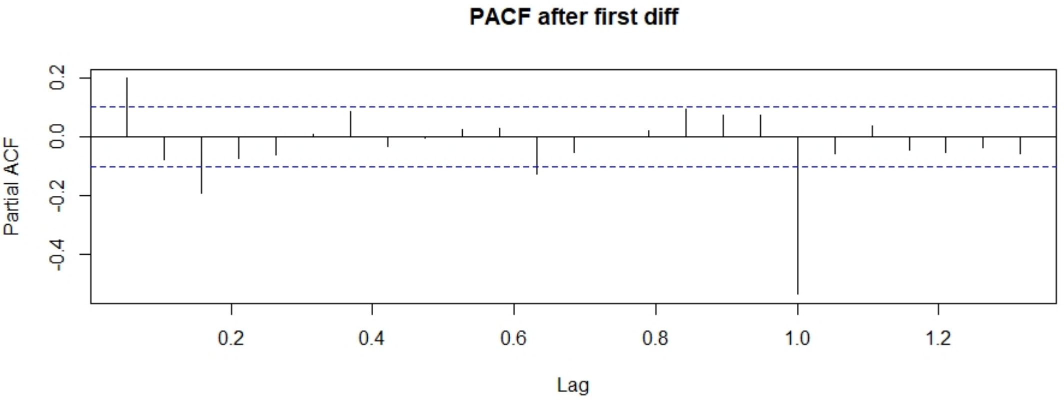


From the acf plot from the original series, we can find out that the period is 19. This is reasonable because there are 19 data points per day and we can see from the seasonality figure that every day has a very similar flow. So, this means every day is a period.

In order to remove the high seasonality of the data, we need use diff function to seasonal difference, the lag is 19.







After the first diff function, the time series seems stable, then $D = 1$. Then we need acf and pacf to confirm other parameters. In the ACF figure, there was two obvious spikes at 0 and 1 are outside the interval, which means $q = 1$ and $Q = 1$. In the PACF figure, the first and third spike is outside the interval, which means the value of p could range from $[1,3]$. The spike at 1 is outside the interval, so $P = 1$.

From the above analysis, there are many possible parameters for Seasonal ARIMA. Use AIC to measure those possibilities.

Order(p, d, q)	Seasonal(P, D, Q)	AIC
1,0,1	1,1,1	2927.88

2,0,1	1,1,1	2915.94
3,0,1	1,1,1	2915.88
1,1,1	1,1,1	2930.48
2,1,1	1,1,1	2929.65
3,1,1	1,1,1	2922.2

The best performance model is $ARIMA(3,0,1)(1,1,1)$ with $AIC = 2915.88$.

6.

Limitations of SARIMA:

The SARIMA model can get a very beautiful prediction based on the previous data. But if there is some accident, like the COVID-19, which have a huge impact on the translink data, the prediction will fail.

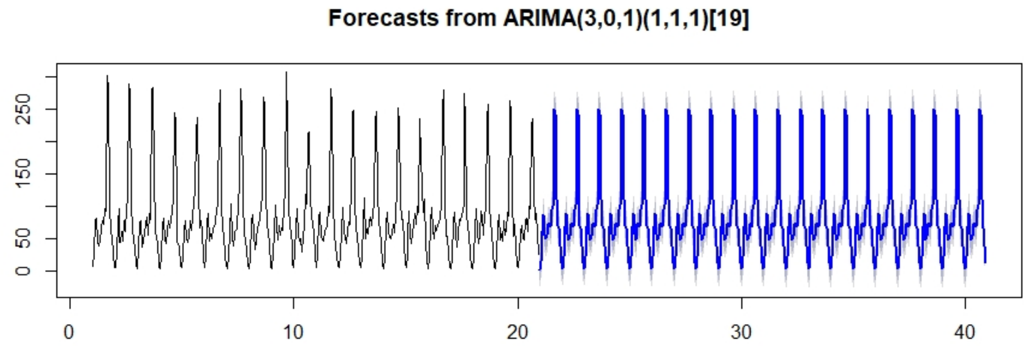
Limitations of Seasonal Naïve:

This model only predicts the result based on the last observed value from the same season, which means it may ignore some key information in the previous seasons.

Limitations of Holt Winters:

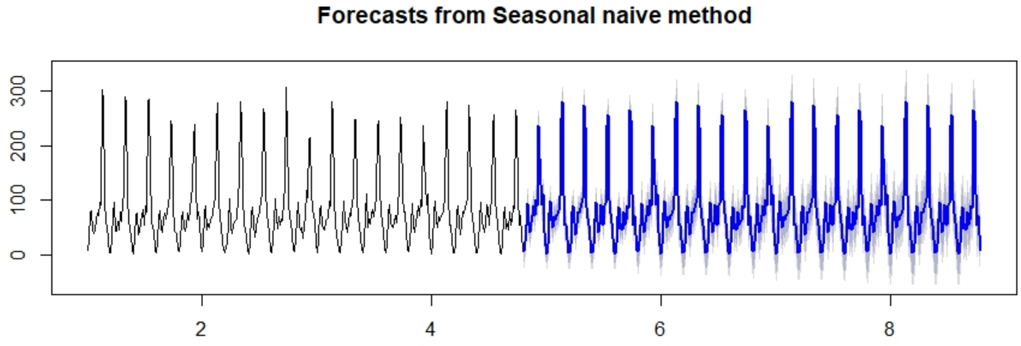
Similar with the limitation in seasonal naïve. The earlier data in this model will have a fewer impact on the final result. In this project, we supposed that every day is a period. But if the month is a period, the information in the early weeks may have a small impact on the prediction.

7.
SARIMA:



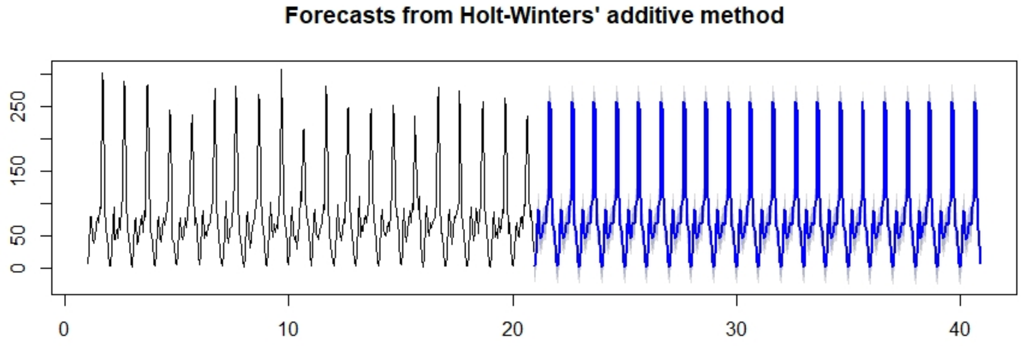
Point Forecast	Lo 95	Hi 95
0.689129	-24.7544	26.13268
14.41753	-11.5193	40.3543
61.26316	35.31195	87.21437
85.66637	59.3353	111.9974
55.65748	29.04568	82.26928
48.6556	21.96442	75.34679
54.52063	27.82682	81.21444
71.6234	44.92441	98.32239
73.35666	46.64829	100.065
67.17036	40.45746	93.88326
94.56504	67.85147	121.2786
100.4147	73.70107	127.1282
250.6791	223.9653	277.393
243.1544	216.4404	269.8684
110.0176	83.30357	136.7316
69.09102	42.37704	95.80499
66.70605	39.99216	93.41994
38.54505	11.83116	65.25894
11.33131	-15.3825	38.04509

Seasonal Naïve:



Point Forecast	Lo 95	Hi 95
5	-24.2507	34.25074
24	-5.25074	53.25074
55	25.74927	84.25074
93	63.74927	122.2507
55	25.74927	84.25074
41	11.74927	70.25074
58	28.74927	87.25074
72	42.74927	101.2507
89	59.74927	118.2507
71	41.74927	100.2507
100	70.74927	129.2507
90	60.74927	119.2507
236	206.7493	265.2507
178	148.7493	207.2507
128	98.74927	157.2507
91	61.74927	120.2507
111	81.74927	140.2507
41	11.74927	70.25074
17	-12.2507	46.25074

Holt Winters:



Point Forecast	Lo 95	Hi 95
2.771901	-23.313	28.85677
22.90321	-3.18166	48.98807
69.051	42.96613	95.13586
90.99059	64.90572	117.0755
54.25435	28.16948	80.33923
46.76299	20.67811	72.84787
54.17477	28.08989	80.25966
66.38104	40.29615	92.46593
71.15266	45.06775	97.23756
68.12959	42.04467	94.21451
96.22052	70.13559	122.3055
104.0119	77.92699	130.0969
258.2721	232.1871	284.3571
248.3091	222.2241	274.3941
110.0252	83.94017	136.1102
66.29939	40.21433	92.38445
59.84293	33.75783	85.92803
36.72466	10.63952	62.80981
9.502301	-16.5829	35.58749

8.

SARIMA:

Time ID	1 hour ahead pred	2 hours ahead pred	Ground Truth
6	3.77	4.31	4
7	23.06	23.00	23
8	70.53	70.54	61
9	90.87	92.83	70
10	50.95	55.43	66
11	53.20	50.16	52
12	56.45	56.69	48
13	68.20	69.95	77
14	74.19	72.37	80
15	69.87	68.70	61
16	91.85	93.63	99
17	102.58	101.11	96
18	254.09	255.40	224
19	238.57	244.97	235
20	110.80	111.53	119
21	74.05	72.39	79
22	67.70	66.72	88
23	40.57	36.54	57
24	10.39	6.81	26
RMSE:	12.6615	13.4739	

Seasonal Naïve:

Time ID	1 hour ahead pred	2 hours ahead pred	Ground Truth
6	5	5	4
7	24	24	23
8	55	55	61
9	93	93	70
10	55	55	66
11	41	41	52
12	58	58	48
13	72	72	77
14	89	89	80
15	71	71	61
16	100	100	99
17	90	90	96
18	236	236	224
19	178	178	235
20	128	128	119
21	91	91	79
22	111	111	88

23	41	41	57
24	17	17	26
RMSE:	17.223	17.223	

Holt Winters:

Time ID	1 hour ahead pred	2 hours ahead pred	Ground Truth
6	2.612247	2.653204	4
7	22.98996	23.00294	23
8	69.61976	69.57224	61
9	92.18742	92.21759	70
10	53.35393	53.64448	66
11	46.17652	46.20049	52
12	54.55135	54.29982	48
13	65.8648	65.93883	77
14	70.67085	70.71493	80
15	68.71341	68.65298	61
16	96.26998	96.39795	99
17	104.0785	104.0791	96
18	258.407	258.4358	224
19	248.2464	248.5216	235
20	109.5718	109.7394	119
21	65.90802	65.90951	79
22	59.35815	59.36944	88
23	36.60835	36.20762	57
24	9.497632	9.498298	26
RMSE:	15.00281	15.01611	

9.

Using SARIMA to predict:

Time ID	Prediction
10	53.14288
11	53.17516
12	60.69316
13	76.23539
14	75.38634
15	67.10001
16	93.66475
17	99.36655
18	249.7998
19	242.8571
20	110.1523
21	69.35065
22	67.06844
23	38.78322
24	11.47498

10.

Dear translink staff:

From the analysis to the one-month passenger trajectory data from region 1 to region 5, we can give you several suggestions.

First, please be aware of the different passenger flow on weekdays and weekends. People are more likely to take the public transportation in the afternoon or late night on weekend, while there are few passengers on weekdays. So, make sure there are enough bus/train/ferry services during that period. Maybe it is a good choice to have additional shifts.

Second, the traffic peak usually happens in 6 p.m in weekdays. This is time for people to get off work and go home, which will cause a huge demand in the transportation.

Third, the valley period of the traffic is usually in the early morning in weekdays. In order to save the operating cost, you may consider reduce the transportation shifts in this period.

Regards,

Yupeng Wu