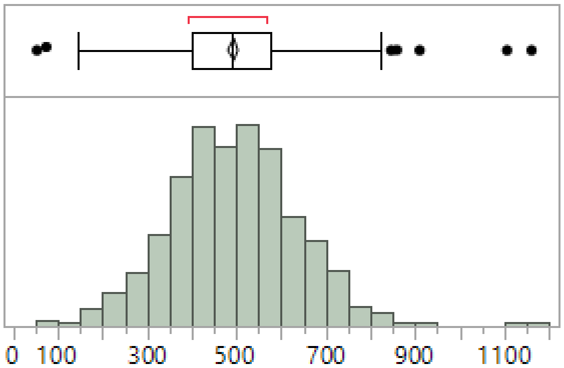
Step 1: Check the distribution of number of trips since it is a count variable. It could possibly require log/sqrt transform

Distributions

Trips



Quantiles

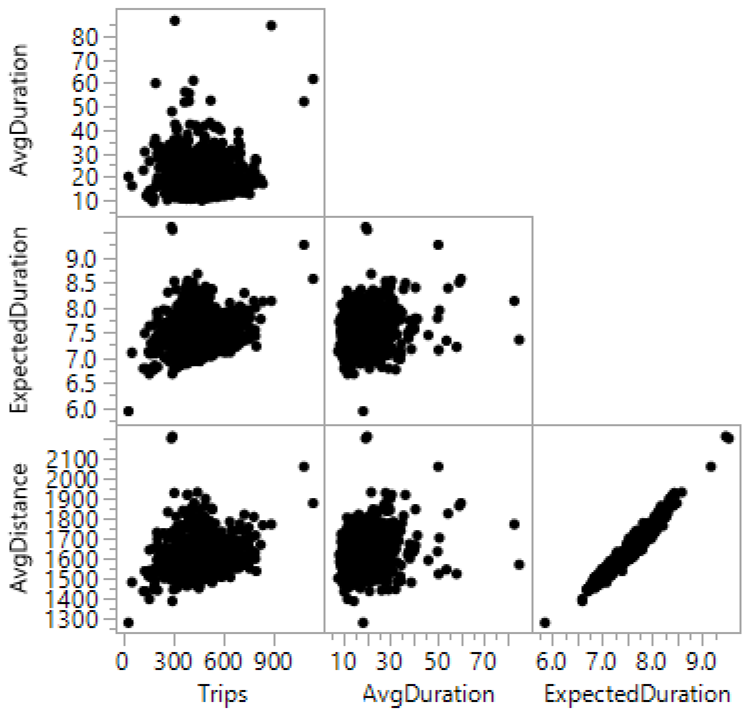
100.0%	maximum	1161
99.5%		876.595
97.5%		770.625
90.0%		669
75.0%	quartile	575.25
50.0%	median	492
25.0%	quartile	399.5
10.0%		323
2.5%		217.925
0.5%		148.24
0.0%	minimum	53

Summary Statistics

Mean	492.33973
Std Dev	138.66895
Std Err Mean	5.1323681
Upper 95% Mean	502.41571
Lower 95% Mean	482.26374
N	730

Conclusion: Does not require any transform. The right tail could be outliers

Step2: Check for multi-collinearity



Conclusion: There isn't a significant collinearity between trips and other 3 variables. This is expected since they are average values. Average duration isn't collinear with average distance and expected duration which is interesting to explore

Step 3: To get an idea of the importance of variables, we run a stepwise fit model including all predictors. Since I don't want to lose too many variables at this point of time, I choose AIC as the stopping rule, and the direction as forward.

Stepwise Fit for Trips

Stepwise Regression Control

Stopping Rule: Minimum AICc

Direction: Forward

Rules: Combine

SSE	DFE	RMSE	RSquare	RSquare Adj	Cp	p	AICc	BIC
6165748.2	719	92.603723	0.5602	0.5540	4.9053265	11	8696.363	8751.044

Current Estimates

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
☑	☑	Intercept	82.0973193	1	0	0.000	1
☐	☑	Month{Jan&Feb&Mar&Dec&Apr-Nov&Jul&May&Jun&Aug&Sep&Oct}	-78.463746	5	4144339	96.656	7.1e-78
☐	☑	Month{Jan&Feb-Mar&Dec&Apr}	-33.284663	2	456146.8	26.596	7.2e-12
☐	☐	Month{Jan-Feb}	0	1	7.549395	0.001	0.97635
☐	☑	Month{Mar-Dec&Apr}	-20.648497	1	70004.11	8.163	0.0044
☐	☐	Month{Dec-Apr}	0	1	23.30584	0.003	0.95847
☐	☑	Month{Nov&Jul&May&Jun-Aug&Sep&Oct}	-24.084101	2	225555.7	13.151	2.46e-6
☐	☐	Month{Nov&Jul&May-Jun}	0	1	9583.648	1.118	0.29076
☐	☐	Month{Nov&Jul-May}	0	1	1946.383	0.227	0.63411
☐	☐	Month{Nov-Jul}	0	3	20657.29	0.802	0.49279
☐	☑	Month{Aug&Sep-Oct}	-11.503713	1	21244.36	2.477	0.11594
☐	☐	Month{Aug-Sep}	0	1	2997.047	0.349	0.55477
☐	☑	Year{2017-2018}	-35.294903	1	827668.5	96.516	1.8e-21
☐	☑	DayofWeek{Mon&Sun-Tue&Thu&Fri&Wed&Sat}	-49.510337	3	2151424	83.627	2e-46
☐	☑	DayofWeek{Mon-Sun}	-20.467065	1	87902.29	10.250	0.00143
☐	☑	DayofWeek{Tue-Thu&Fri&Wed&Sat}	-25.326976	1	212531.3	24.784	8.04e-7
☐	☐	DayofWeek{Thu&Fri-Wed&Sat}	0	1	5044.486	0.588	0.44348
☐	☐	DayofWeek{Thu-Fri}	0	1	0.012356	0.000	0.99904
☐	☐	DayofWeek{Wed-Sat}	0	2	7010.656	0.408	0.66507
☐	☐	Weekend{Yes-No}	0	1	5946.103	0.693	0.40539
☐	☐	AvgDuration	0	1	4686.5	0.546	0.46013
☐	☐	ExpectedDuration	0	1	6275.297	0.732	0.39268
☐	☑	AvgDistance	0.22393156	1	250716	29.236	8.73e-8

Step History

Step	Parameter	Action	"Sig Prob"	Seq SS	RSquare	Cp	p	AICc	BIC	
1	Month{Jan&Feb&Mar&Dec&Apr-Nov&Jul&May&Jun&Aug&Sep&Oct}	Entered	0.0000	3700876	0.2640	466.9	2	9053.76	9067.51	○
2	DayofWeek{Mon&Sun-Tue&Thu&Fri&Wed&Sat}	Entered	0.0000	1648770	0.3816	278.27	3	8928.67	8946.99	○
3	Year{2017-2018}	Entered	0.0000	1230905	0.4694	137.94	4	8818.9	8841.78	○
4	Month{Jan&Feb-Mar&Dec&Apr}	Entered	0.0000	402352.4	0.4981	93.422	5	8780.33	8807.77	○
5	AvgDistance	Entered	0.0000	275166.3	0.5178	63.607	6	8753.24	8785.24	○
6	DayofWeek{Tue-Thu&Fri&Wed&Sat}	Entered	0.0000	208454.2	0.5326	41.504	7	8732.42	8768.97	○
7	Month{Nov&Jul&May&Jun-Aug&Sep&Oct}	Entered	0.0000	205933.3	0.5473	19.694	8	8711.16	8752.25	○
8	DayofWeek{Mon-Sun}	Entered	0.0015	88313.82	0.5536	11.483	9	8702.98	8748.61	○
9	Month{Mar-Dec&Apr}	Entered	0.0044	70235.21	0.5586	5.3617	10	8696.81	8746.96	○
10	Month{Aug&Sep-Oct}	Entered	0.1159	21244.36	0.5602	4.9053	11	8696.36	8751.04	○
11	Month{Nov&Jul&May-Jun}	Entered	0.2908	9583.648	0.5608	5.7972	12	8697.3	8756.5	○
12	Weekend{Yes-No}	Entered	0.4069	5906.607	0.5613	7.1143	13	8698.68	8762.39	○
13	AvgDuration	Entered	0.3917	6302.023	0.5617	8.3856	14	8700.01	8768.24	○
14	ExpectedDuration	Entered	0.3953	6210.909	0.5622	9.6675	15	8701.37	8774.09	○
15	Month{Aug-Sep}	Entered	0.5003	3906.922	0.5624	11.216	16	8703	8780.22	○
16	Month{Nov-Jul}	Entered	0.5760	9497.398	0.5631	14.118	18	8706.08	8792.28	○
17	DayofWeek{Thu&Fri-Wed&Sat}	Entered	0.7430	926.959	0.5632	16.01	19	8708.08	8798.76	○
18	Month{Jan-Feb}	Entered	0.9269	72.70983	0.5632	18.002	20	8710.19	8805.34	○
19	Month{Dec-Apr}	Entered	0.9639	17.66152	0.5632	20	21	8712.32	8811.93	○
20	DayofWeek{Thu-Fri}	Entered	0.9965	0.16225	0.5632	22	22	8714.45	8818.53	○
21	Best	Specific	.	.	0.5602	4.9053	11	8696.36	8751.04	◉

Conclusion: We see a bit of all predictor variables coming in – Year, Month, Day, AvgDistance.

Step 4: Then, we try backward selection.

Stepwise Fit for Trips

Stepwise Regression Control

SSE	DFE	RMSE	RSquare	RSquare Adj	Cp	p	AICc	BIC
6178534.3	720	92.635293	0.5592	0.5537	4.3836893	10	8695.807	8745.963

Current Estimates

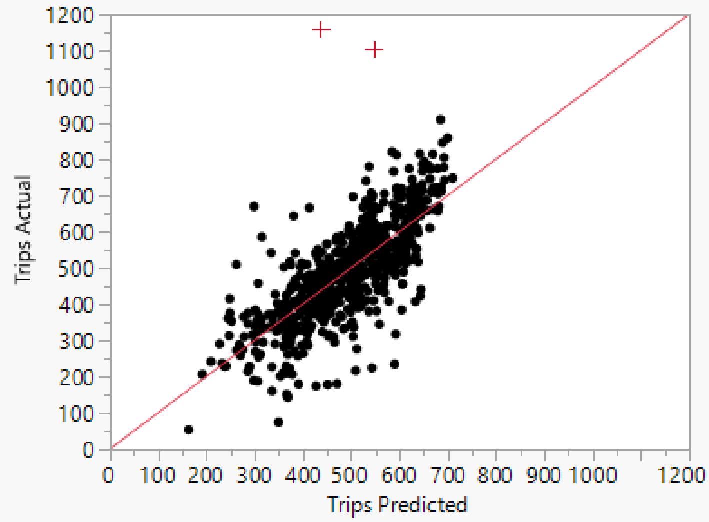
Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
☑	☑	Intercept	-5.0596306	1	0	0.000	1
☐	☑	Month{Jan&Feb&Mar&Dec&Apr-Nov&Jul&May&Jun&Aug&Sep&Oct}	-77.838469	4	4195423	122.226	1.4e-79
☐	☑	Month{Jan&Feb-Mar&Dec&Apr}	-34.114982	2	469087	27.332	3.6e-12
☐	☐	Month{Jan-Feb}	0	1	284.9933	0.033	0.85554
☐	☑	Month{Mar-Dec&Apr}	-20.211153	1	67098.58	7.819	0.00531
☐	☐	Month{Dec-Apr}	0	1	293.849	0.034	0.85334
☐	☑	Month{Nov&Jul&May&Jun-Aug&Sep&Oct}	-22.974836	1	221104.7	25.766	4.91e-7
☐	☐	Month{Nov&Jul&May-Jun}	0	1	7810.583	0.910	0.34042
☐	☐	Month{Nov&Jul-May}	0	1	475.9231	0.055	0.81401
☐	☐	Month{Nov-Jul}	0	3	18089.88	0.702	0.55114
☐	☐	Month{Aug&Sep-Oct}	0	1	17342.64	2.024	0.15528
☐	☐	Month{Aug-Sep}	0	1	2958.637	0.344	0.55745
☐	☑	Year{2017-2018}	-36.476387	1	921011.9	107.328	1.5e-23
☐	☑	DayofWeek{Mon&Sun-Tue&Thu&Fri&Wed&Sat}	-47.706082	3	2093883	81.335	2.5e-45
☐	☑	DayofWeek{Mon-Sun}	-20.264253	1	86186.41	10.044	0.00159
☐	☑	DayofWeek{Tue-Thu&Fri&Wed&Sat}	-25.447851	1	214419.4	24.987	7.26e-7
☐	☐	DayofWeek{Thu&Fri-Wed&Sat}	0	1	5337.21	0.622	0.4307
☐	☐	DayofWeek{Thu-Fri}	0	1	3.025292	0.000	0.98504
☐	☐	DayofWeek{Wed-Sat}	0	2	8130.497	0.473	0.6233
☐	☐	Weekend{Yes-No}	0	1	7266.551	0.847	0.35782
☐	☐	AvgDuration	0	1	7565.66	0.882	0.34811
☐	☑	ExpectedDuration	59.4909073	1	288114.1	33.575	1.03e-8
☐	☐	AvgDistance	0	1	902.5046	0.105	0.74596

Conclusion: More or less the same set of variables except Expected Duration. This is expected due to its collinearity with Avg Distance

Step 5: Use these important predictors (Month, Year, DayOfWeek, either AvgDistance or ExpectedDuration) and do a Mixed Model to capture time variable. I included AR(1) structure to check for higher order correlation

Fit Mixed

Actual by Predicted Plot

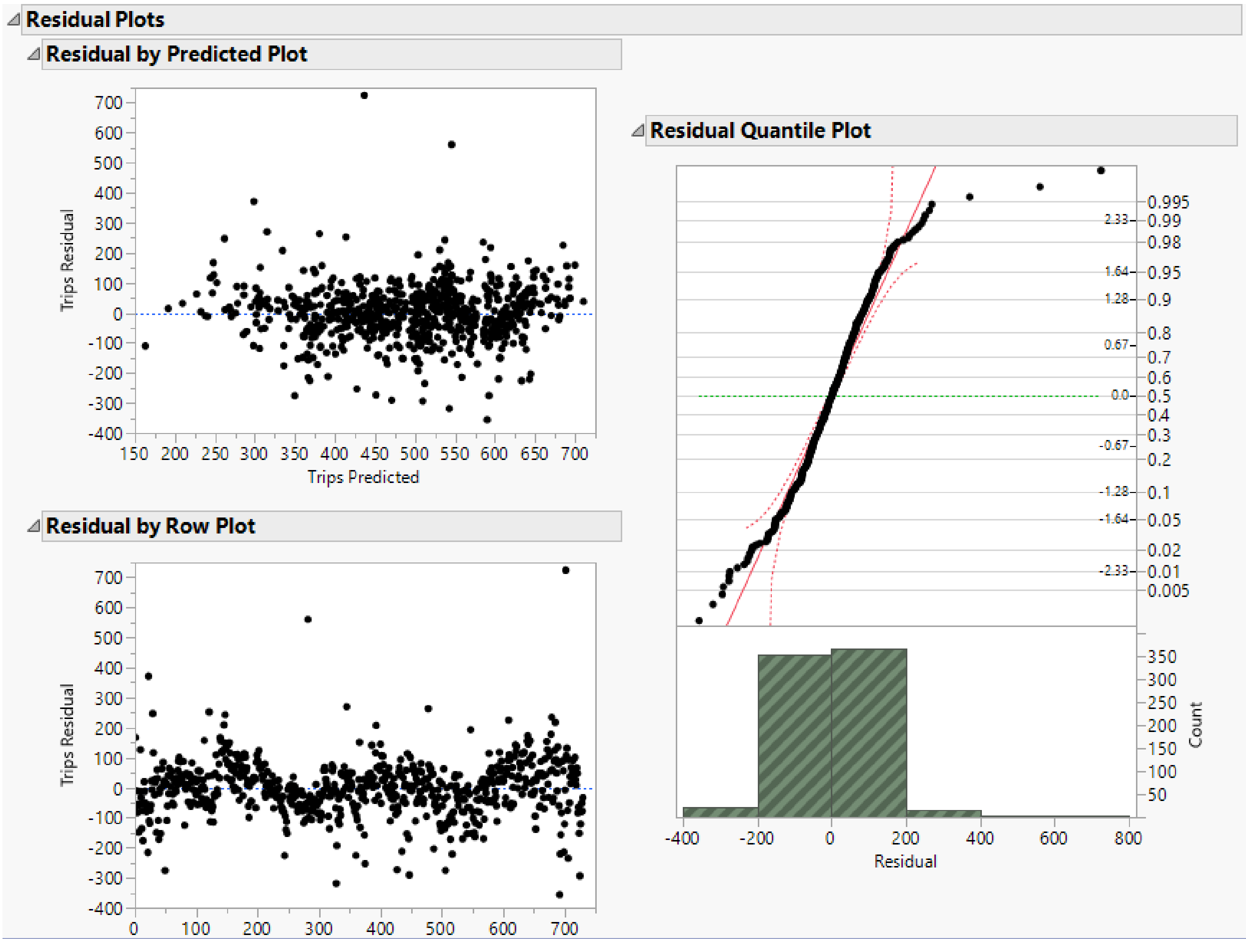


Fit Statistics

-2 Residual Log Likelihood	8434.6962
-2 Log Likelihood	8555.0813
AICc	8600.5127
BIC	8700.1283

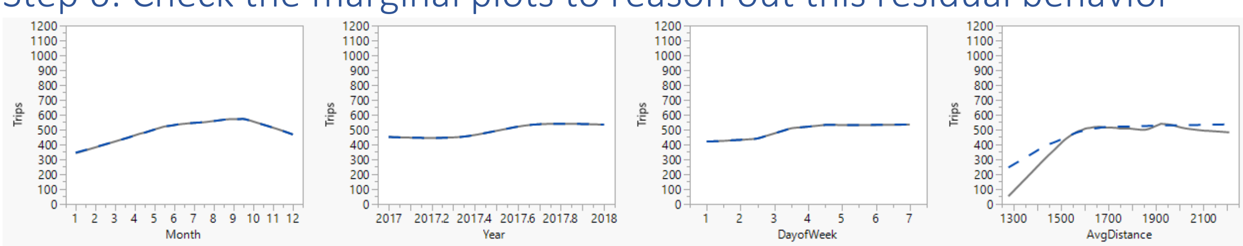
Repeated Effects Covariance Parameter Estimates

Covariance Parameter	Estimate	Std Error	95% Lower	95% Upper
AR(1)	0.4052765	0.0357006	0.3353046	0.4752485
Residual	8848.3702	569.1835	7830.7649	10079.174

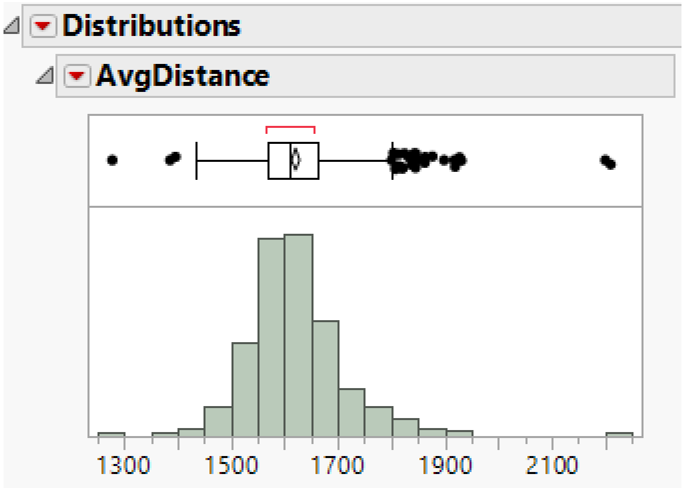


Conclusion: The model looks good to start with. Row 702 and Row 282 could be outliers. Most of the predictor variables chosen are statistically significant. The ANOVA table (Fixed Effects Test) is also significant. The residual plots are not so good. Q-Q plot is skewed, possibly because of the two outliers. From the trips-residual vs row number plot, we infer a wave like pattern. The row numbers are arranged in chronological pattern, date order. This is due to the serial correlation. So, AR(1) does not fix the problem.

Step 6: Check the marginal plots to reason out this residual behavior



Conclusion: It is doing well. AvgDistance could be checked for transformation.

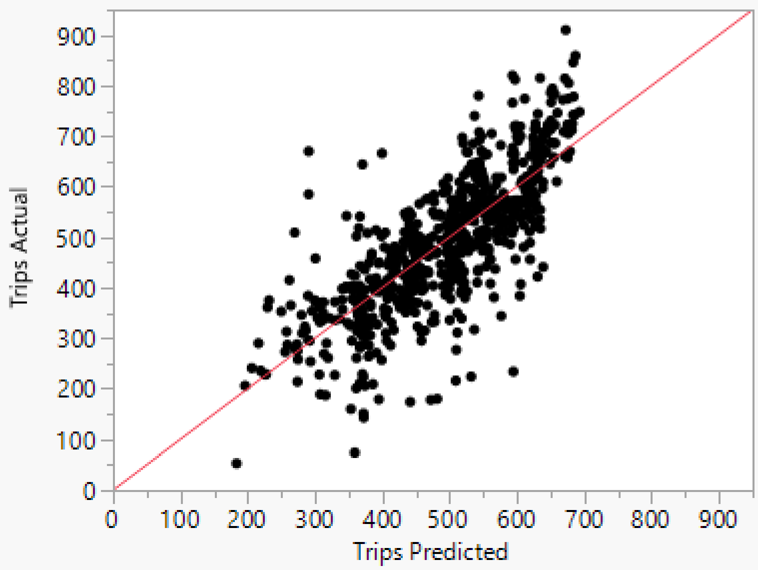


No transformation is required

Step 7: The two outliers mentioned above are removed and mixed model is tried again

Fit Mixed

Actual by Predicted Plot



Fit Statistics

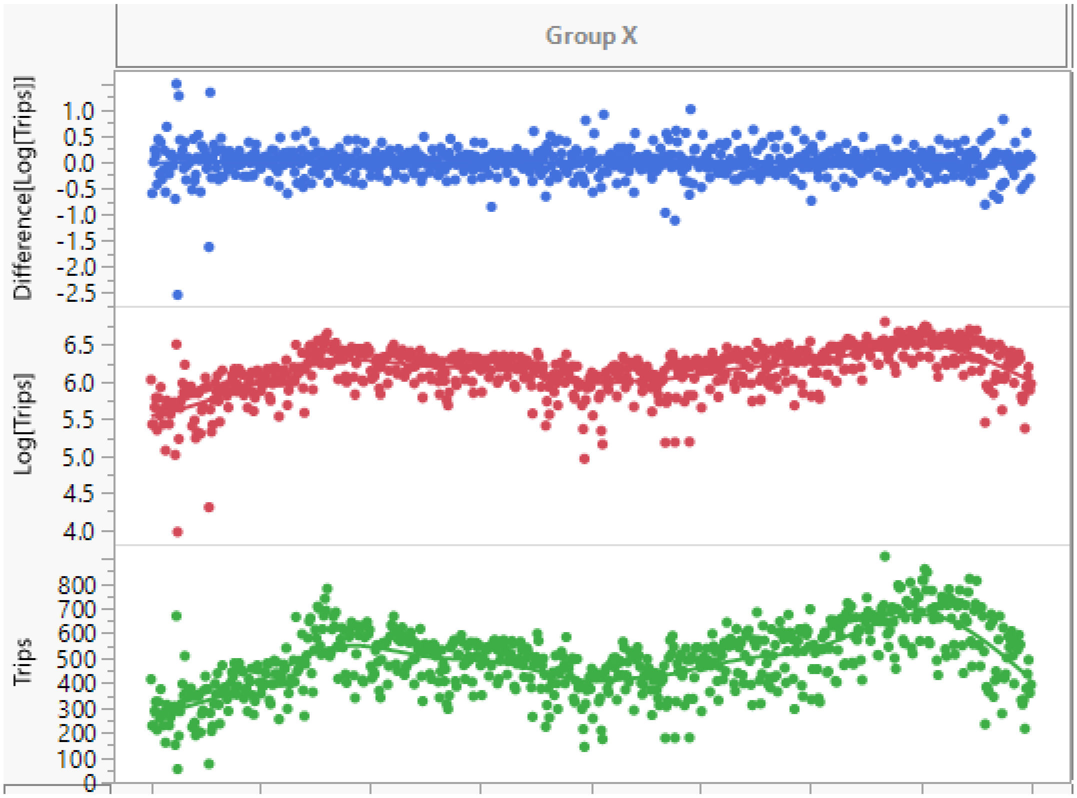
-2 Residual Log Likelihood	8247.8977
-2 Log Likelihood	8366.6998
AICc	8412.1353
BIC	8511.6864

Repeated Effects Covariance Parameter Estimates

Covariance Parameter	Estimate	Std Error	95% Lower	95% Upper
AR(1)	0.4828556	0.0347873	0.4146737	0.5510375
Residual	7687.983	538.66254	6732.5358	8863.4411

Conclusion: It can be seen that the AICc has improved from 8600 to 8412. Also, all the predictors are still significant.

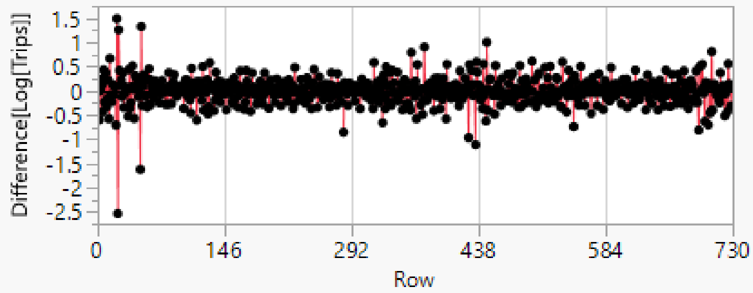
Step 7: Then I tried the Time Series Modelling in JMP under specialized modelling. Before that a look at the Trips pattern shows variability and seasonality. So, I log-transformed and differenced the data.



Step 8: I used this as the dependent variable and fit a time series model to check the ACF and PACF plots

Transfer Function Analysis

Time Series Difference[Log[Trips]]



Mean	-0.002694
Std	0.2835225
N	727
Zero Mean ADF	-38.03028
Single Mean ADF	-38.00297
Trend ADF	-37.98184

Time Series Basic Diagnostics

Lag	AutoCorr	-.8-.6-.4-.2 0 .2 .4 .6 .8	Ljung-Box Q	p-Value	Lag	Partial	-.8-.6-.4-.2 0 .2 .4 .6 .8
0	1.0000		.	.	0	1.0000	
1	-0.3200		74.7369	<.0001*	1	-0.3200	
2	-0.0945		81.2653	<.0001*	2	-0.2193	
3	-0.0417		82.5360	<.0001*	3	-0.1740	
4	-0.0451		84.0260	<.0001*	4	-0.1759	
5	-0.0898		89.9472	<.0001*	5	-0.2546	
6	-0.0512		91.8747	<.0001*	6	-0.3165	
7	0.2824		150.594	<.0001*	7	0.0449	
8	0.0595		153.203	<.0001*	8	0.1643	
9	-0.2082		185.191	<.0001*	9	-0.0901	
10	0.0246		185.639	<.0001*	10	-0.0562	
11	-0.0500		187.493	<.0001*	11	-0.0888	
12	-0.0527		189.552	<.0001*	12	-0.1037	

Step 9: I see that it could be anything between AR(1) to AR(7) or MA or any combination of these. I played around with these, transfer functions and compared the significance of the predictors, and AIC values.

After these trail and errors, I came up with this model using following thumb rules:

1. Go for a parsimonious model
2. Look for significance of predictor terms
3. Go with the min AIC model

Report	Graph	Model	DF	Variance	AIC	SBC	RSquare	-2LogLH	We
[x]	[ ]	AR(7)	719	0.0545389	-42.27364	-5.562232	0.326	-58.27364	1.00
[x]	[ ]	ARMA(2, 2) No Intercept	723	0.0587908	7.990611	26.346317	0.270	-0.009389	0.00
[x]	[ ]	ARMA(1, 1) No Intercept	725	0.0594643	14.124032	23.301885	0.260	10.124032	0.00
[x]	[ ]	MA(2) No Intercept	725	0.0595282	14.934414	24.112267	0.259	10.934414	0.00
[x]	[ ]	MA(1) No Intercept	726	0.0626843	51.480434	56.069360	0.219	49.480434	0.00
[x]	[ ]	AR(2) No Intercept	725	0.0687288	118.73866	127.91651	0.146	114.73866	0.00
[x]	[ ]	AR(1) No Intercept	726	0.0721597	153.05513	157.64406	0.103	151.05513	0.00
[x]	[ ]	AR(1)	725	0.0722526	154.98887	164.16672	0.103	150.98887	0.00

Also, the transformation and first order differencing has already been done. The intercept is found insignificant always, so that was removed. I chose ARMA(1,1,1) since it is parsimonious and gives good results as well. AR(7) gives better results but is highly complex.

The best models's results is shown:

Model: ARMA(1, 1) No Intercept

Model Summary

DF	725	Stable	Yes
Sum of Squared Errors	43.1115846	Invertible	Yes
Variance Estimate	0.05946425		
Standard Deviation	0.24385294		
Akaike's 'A' Information Criterion	14.124032		
Schwarz's Bayesian Criterion	23.301885		
RSquare	0.26022339		
RSquare Adj	0.259203		
MAPE	.		
MAE	0.17538662		
-2LogLikelihood	10.124032		

Parameter Estimates

Term	Lag	Estimate	Std Error	t Ratio	Prob>|t|
AR1	1	0.27662068	0.0433556	6.38	<.0001*
MA1	1	0.85629845	0.0203017	42.18	<.0001*

Lag	AutoCorr	-.8-.6-.4-.2 0 .2 .4 .6 .8	Ljung-Box Q	p-Value	Lag	Partial	-.8-.6-.4-.2 0 .2 .4 .6 .8
0	1.0000		.	.	0	1.0000	
1	0.0184		0.2479	0.6186	1	0.0184	
2	-0.0164		0.4443	0.8008	2	-0.0167	
3	-0.0504		2.3005	0.5124	3	-0.0498	
4	-0.0735		6.2574	0.1807	4	-0.0721	
5	-0.0800		10.9563	0.0523	5	-0.0799	
6	0.0216		11.2992	0.0796	6	0.0188	
7	0.2944		75.1022	<.0001*	7	0.2881	
8	0.0987		82.2892	<.0001*	8	0.0909	
9	-0.1637		102.081	<.0001*	9	-0.1868	
10	-0.0393		103.224	<.0001*	10	-0.0288	