MASTER OF SCIENCE IN INFORMATION TECHNOLOGY

PETER IRUNGU MWANGI

SCT321-C004-2079/2018

MIT 3102 BUSINESS INTELLIGENCE

Discuss the dataset

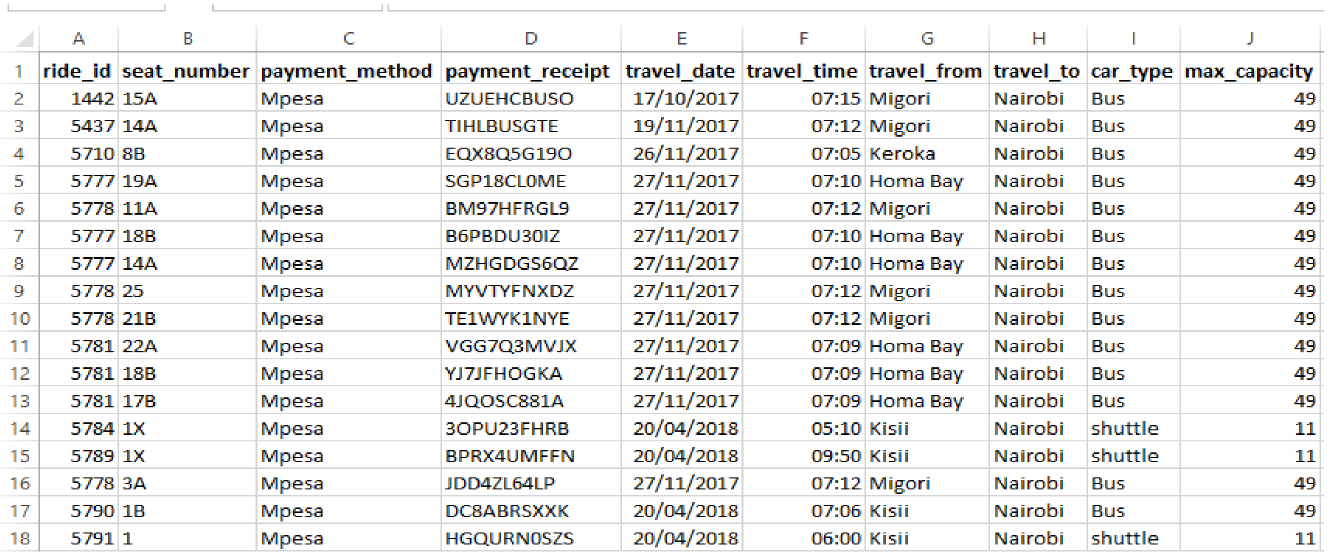
Introduction

There are two dataset types provided; train_revised.csv (as a training dataset) and test_question.csv (as a testing dataset) datasets.

The train_revised.csv contains dataset of tickets purchased from Mobiticket for fourteen (14) routes from “up country” into the city of Nairobi. The dataset runs for a period between 17 October 2017 and 20 April 2018.

Training dataset (train_revised.csv)

The train or training data is data for building a model, it’s used in a supervised learning as it contains outcomes to train the machine/model. The variables and sample data for this dataset are as shown in the figure 1.1.



	A	B	C	D	E	F	G	H	I	J
1	**ride_id**	**seat_number**	**payment_method**	**payment_receipt**	**travel_date**	**travel_time**	**travel_from**	**travel_to**	**car_type**	**max_capacity**
2	1442	15A	Mpesa	UZUEHCBUSO	17/10/2017	07:15	Migori	Nairobi	Bus	49
3	5437	14A	Mpesa	TIHLBUSGTE	19/11/2017	07:12	Migori	Nairobi	Bus	49
4	5710	8B	Mpesa	EQX8Q5G19O	26/11/2017	07:05	Keroka	Nairobi	Bus	49
5	5777	19A	Mpesa	SGP18CL0ME	27/11/2017	07:10	Homa Bay	Nairobi	Bus	49
6	5778	11A	Mpesa	BM97HFRGL9	27/11/2017	07:12	Migori	Nairobi	Bus	49
7	5777	18B	Mpesa	B6PBDU30IZ	27/11/2017	07:10	Homa Bay	Nairobi	Bus	49
8	5777	14A	Mpesa	MZHGDGS6QZ	27/11/2017	07:10	Homa Bay	Nairobi	Bus	49
9	5778	25	Mpesa	MYVTYFNXDZ	27/11/2017	07:12	Migori	Nairobi	Bus	49
10	5778	21B	Mpesa	TE1WYK1NYE	27/11/2017	07:12	Migori	Nairobi	Bus	49
11	5781	22A	Mpesa	VGG7Q3MVJX	27/11/2017	07:09	Homa Bay	Nairobi	Bus	49
12	5781	18B	Mpesa	YJ7JFHOGKA	27/11/2017	07:09	Homa Bay	Nairobi	Bus	49
13	5781	17B	Mpesa	4JQOSC881A	27/11/2017	07:09	Homa Bay	Nairobi	Bus	49
14	5784	1X	Mpesa	3OPU23FHRB	20/04/2018	05:10	Kisii	Nairobi	shuttle	11
15	5789	1X	Mpesa	BPRX4UMFFN	20/04/2018	09:50	Kisii	Nairobi	shuttle	11
16	5778	3A	Mpesa	JDD4ZL64LP	27/11/2017	07:12	Migori	Nairobi	Bus	49
17	5790	1B	Mpesa	DC8ABRSXXK	20/04/2018	07:06	Kisii	Nairobi	Bus	49
18	5791	1	Mpesa	HGQURN0SZS	20/04/2018	06:00	Kisii	Nairobi	shuttle	11

Figure 1.1

Variable definitions

a) ride_id – this is a unique identification representing trip made from an up country,

b) seat_number – this stores the seat numbers for a specific car,

c) payment_method – stores type of payment method used e.g. Mpesa or Cash,

d) payment_receipt – store the receipt number as a prove of payment and it’s unique,

e) travel_date – stores the commuting date when the ride happened,

f) travel_time – stores commuting time for the ride,

g) travel_from – store the start or originating town for the ride,

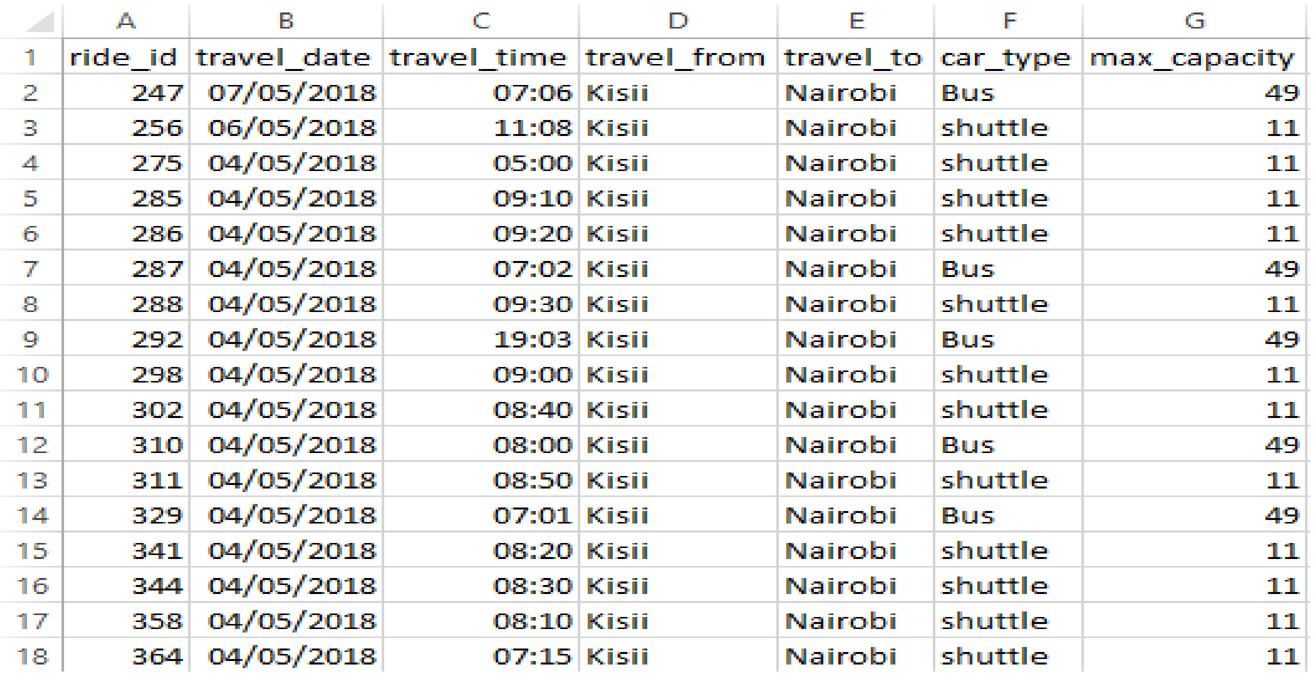
h) travel_to – stores the destination of the ride i.e. Nairobi,

i) car_type – stores the type of car used for the ride, Bus or shuttle,

j) max_capacity – stores the maximum number of passengers for each car type.

Testing dataset (test_revised.csv)

Test_question.csv dataset is the testing dataset, the outcomes are not known, it depends on the model created from the training (train_revised.csv) dataset to predict its outcomes. In this case it contains most of the variables from the training dataset with an exception of seat_number, payment_method and payment_receipt. The figure 1.2 shows sample records for the test dataset.



	A	B	C	D	E	F	G
1	ride_id	travel_date	travel_time	travel_from	travel_to	car_type	max_capacity
2	247	07/05/2018	07:06	Kisii	Nairobi	Bus	49
3	256	06/05/2018	11:08	Kisii	Nairobi	shuttle	11
4	275	04/05/2018	05:00	Kisii	Nairobi	shuttle	11
5	285	04/05/2018	09:10	Kisii	Nairobi	shuttle	11
6	286	04/05/2018	09:20	Kisii	Nairobi	shuttle	11
7	287	04/05/2018	07:02	Kisii	Nairobi	Bus	49
8	288	04/05/2018	09:30	Kisii	Nairobi	shuttle	11
9	292	04/05/2018	19:03	Kisii	Nairobi	Bus	49
10	298	04/05/2018	09:00	Kisii	Nairobi	shuttle	11
11	302	04/05/2018	08:40	Kisii	Nairobi	shuttle	11
12	310	04/05/2018	08:00	Kisii	Nairobi	Bus	49
13	311	04/05/2018	08:50	Kisii	Nairobi	shuttle	11
14	329	04/05/2018	07:01	Kisii	Nairobi	Bus	49
15	341	04/05/2018	08:20	Kisii	Nairobi	shuttle	11
16	344	04/05/2018	08:30	Kisii	Nairobi	shuttle	11
17	358	04/05/2018	08:10	Kisii	Nairobi	shuttle	11
18	364	04/05/2018	07:15	Kisii	Nairobi	shuttle	11

Figure 1.2

Observations

The training dataset (train_revised.csv), the variable ride_id appears multiple times in the dataset which is different from the testing dataset (test_question.csv).

The training dataset has more columns compared to the testing dataset.

In both dataset there is no variable which stores the total number of tickets sold for each ride from all routes. This build the question and defines what the model should do, predict the number of tickets for each ride.

Assumptions in creating the model

To get the number of tickets for each ride in the training dataset, we count/aggregate how many times the ride_id has been repeated in the dataset. To achieve this, we consolidate the ride by count of the ride_id and store them in a variable no_of_tickets. Then, merge the aggregate to the training dataset and remove the duplicates.

Name the new dataset as train_aggregate.csv, which has an addition variable which answers our question on number_of_tickets sold per ride. The new train dataset should now have the following variables; ride_id, travel_date, travel_time, travel_from, travel_to, car_type, max_capacity, number_of_tickets. The new dataset will be used to create a predictive model to predict the outcomes of the testing dataset.

Conclusions

From the problem description, to build a model to predict the number of seats that Mobitickets can expect to sell for each ride for a specific route on a specific date and time, the aggregated training dataset with already determined or known outcomes will be ideal training dataset to solve this problem.

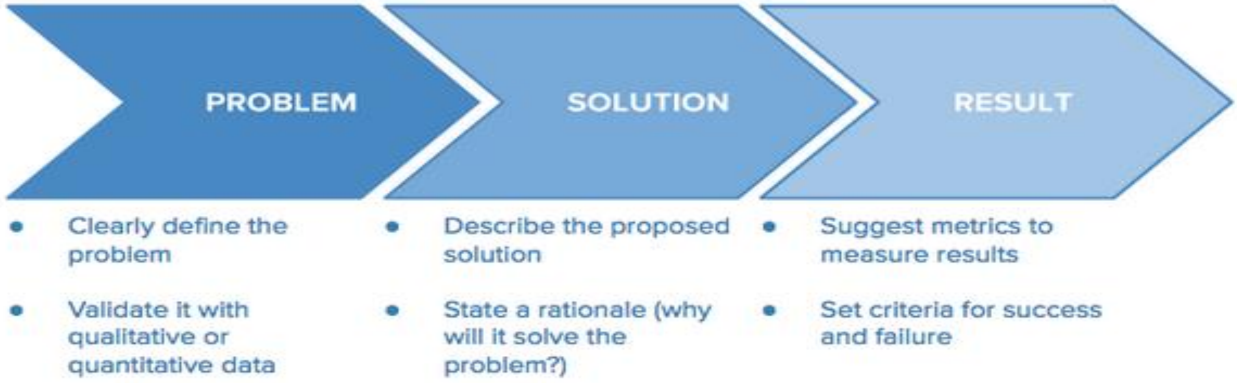
Model using regression

Regression analysis – is a set of statistical processes for estimating the relationship among variables and understanding which among the independent variables are related to the dependent variables and explore the nature of the identified relationships.

Regression model – it's used to investigate the relationship between two or more variables and use it to estimate one variable based on the others.

Hypothesis for this analysis

What is meaning of hypothesis - A hypothesis is a prediction you create prior to running an experiment. The common format is: if (cause), then (effect), because (rationale).


PROBLEM
SOLUTION
RESULT
Clearly define the problem
Validate it with qualitative or quantitative data
Describe the proposed solution
State a rationale (why will it solve the problem?)
Suggest metrics to measure results
Set criteria for success and failure

Basic view of a hypothetical question and its purpose

Problem definition:

Nairobi city traffic directly affects the movement of people to and from the up-country, thus affecting the commuter service providers business.

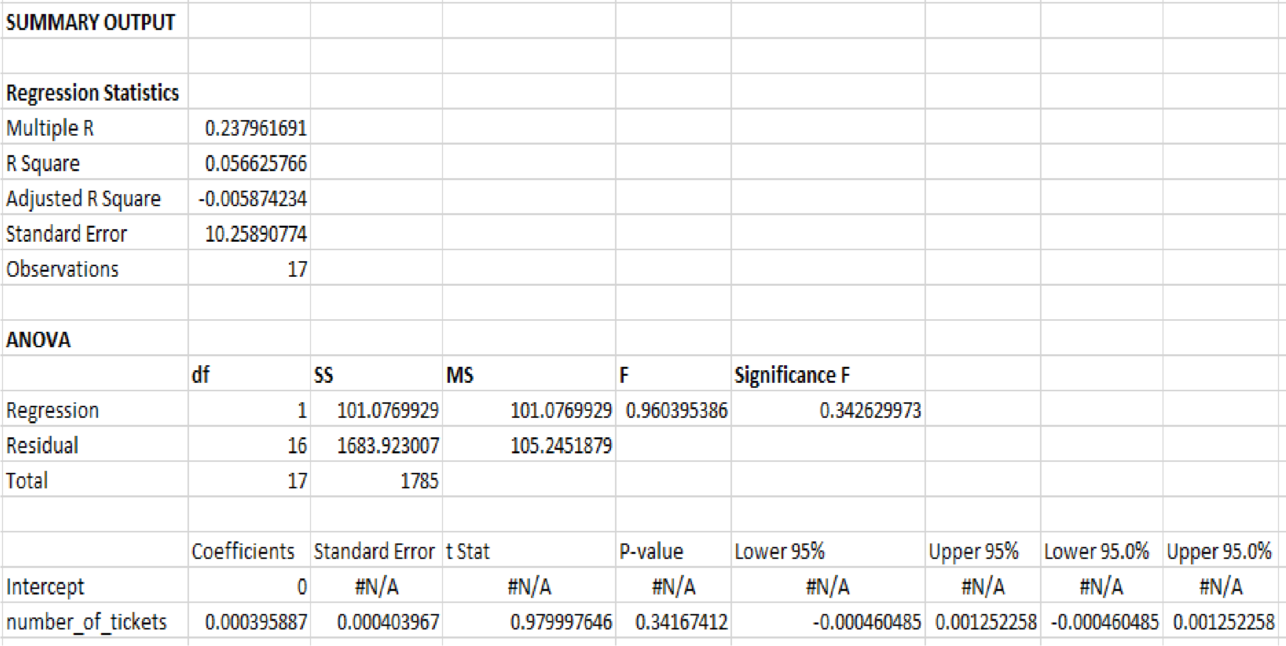
Data validation:

Using previous data on tickets sold for each route to the upcountry, the aggregate it to find total number of the tickets sold for each ride. Using regression analysis, ticketing company can predict the tickets sales for each route to upcountry.

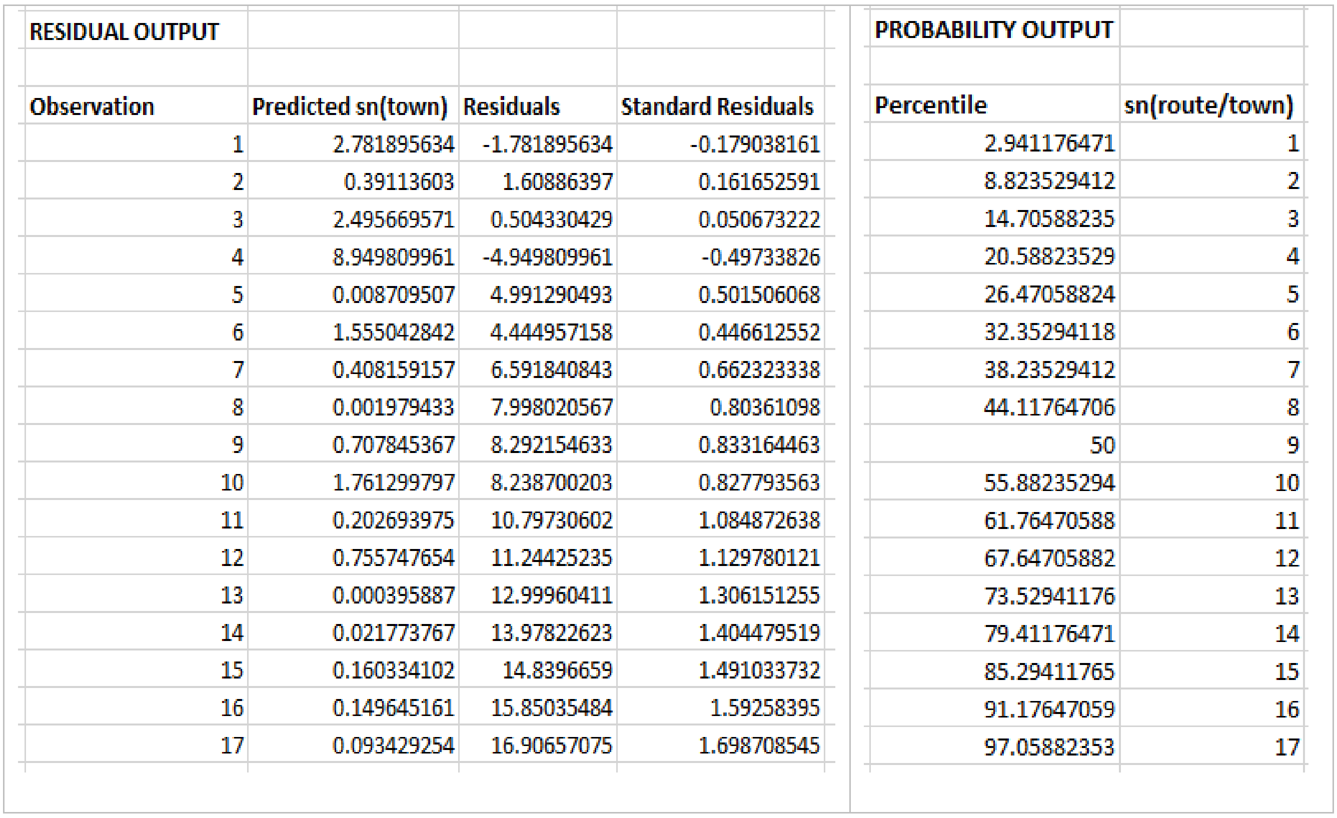
Solution:

From the aggregated training dataset (*train_aggregate.csv*), identify the independent and dependent variables. The dependent variables in this case will be the number of tickets for each ride, this is because it's dependent to the route on a specific date and time. This makes route as an independent variable.

The summary output between routes against the number of tickets



SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.237961691							
R Square	0.056625766							
Adjusted R Square	-0.005874234							
Standard Error	10.25890774							
Observations	17							
ANOVA								
	df	**SS**	**MS**	**F**	**Significance F**			
Regression	1	101.0769929	101.0769929	0.960395386	0.342629973			
Residual	16	1683.923007	105.2451879					
Total	17	1785						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	0	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
number_of_tickets	0.000395887	0.000403967	0.979997646	0.34167412	-0.000460485	0.001252258	-0.000460485	0.001252258

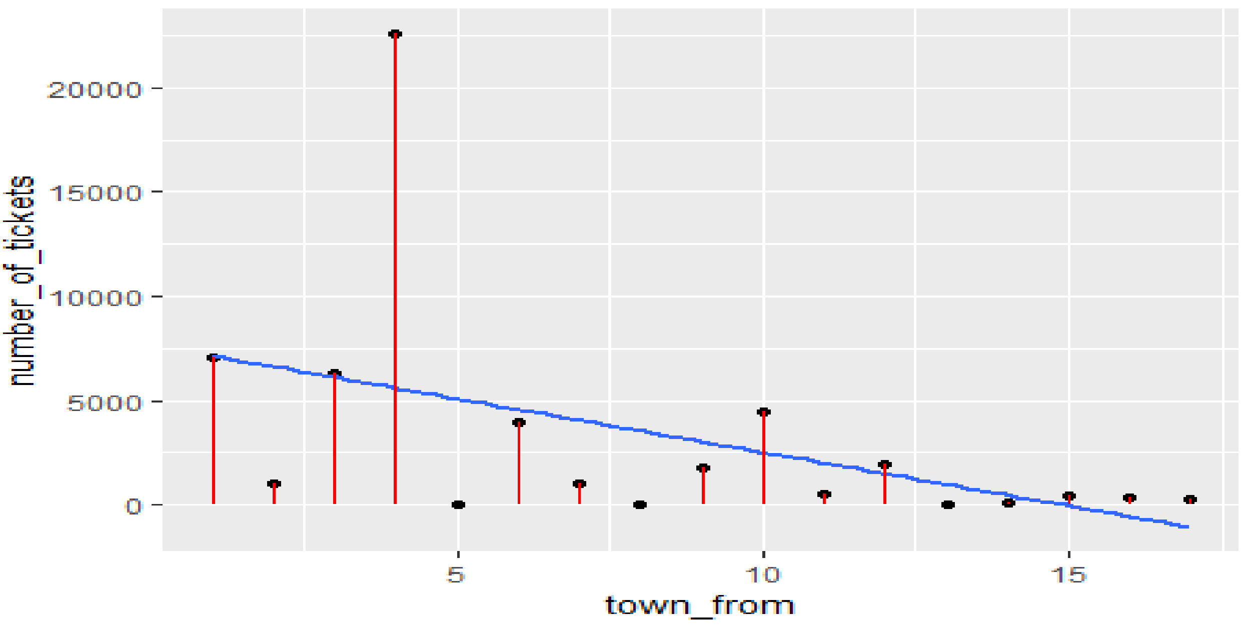


RESIDUAL OUTPUT			
Observation	Predicted sn(town)	Residuals	Standard Residuals
1	2.781895634	-1.781895634	-0.179038161
2	0.39113603	1.60886397	0.161652591
3	2.495669571	0.504330429	0.050673222
4	8.949809961	-4.949809961	-0.49733826
5	0.008709507	4.991290493	0.501506068
6	1.555042842	4.444957158	0.446612552
7	0.408159157	6.591840843	0.662323338
8	0.001979433	7.998020567	0.80361098
9	0.707845367	8.292154633	0.833164463
10	1.761299797	8.238700203	0.827793563
11	0.202693975	10.79730602	1.084872638
12	0.755747654	11.24425235	1.129780121
13	0.000395887	12.99960411	1.306151255
14	0.021773767	13.97822623	1.404479519
15	0.160334102	14.8396659	1.491033732
16	0.149645161	15.85035484	1.59258395
17	0.093429254	16.90657075	1.698708545

PROBABILITY OUTPUT	
Percentile	sn(route/town)
2.941176471	1
8.823529412	2
14.70588235	3
20.58823529	4
26.47058824	5
32.35294118	6
38.23529412	7
44.11764706	8
50	9
55.88235294	10
61.76470588	11
67.64705882	12
73.52941176	13
79.41176471	14
85.29411765	15
91.17647059	16
97.05882353	17

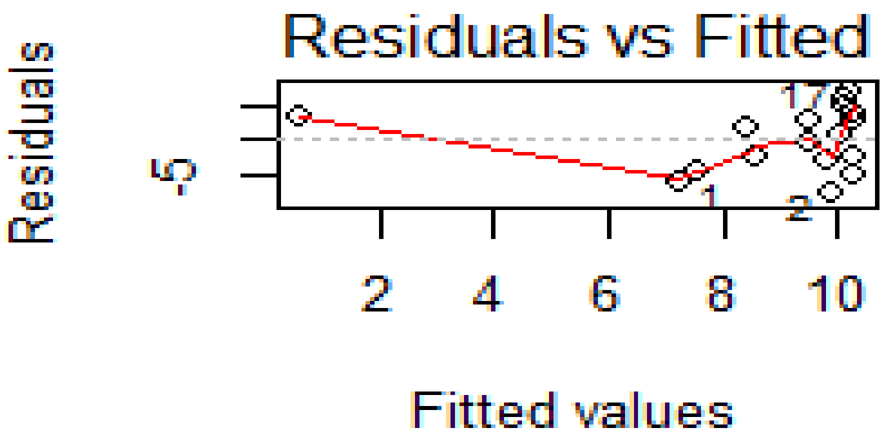
Model using R.

```
Coefficients:

      (Intercept)  number_of_tickets
       10.3040303         -0.0004292

# A tibble: 6 x 9
  town_from number_of_tickets .fitted .se.fit .resid   .hat .sigma .cooksd
.std.resid
      <int>             <int>   <dbl>   <dbl>  <dbl>  <dbl>  <dbl>   <dbl>
<dbl>
1         1              7027    7.29     1.39  -6.29 0.0914   4.43   0.103       -
1.43
2         2               988    9.88     1.20  -7.88 0.0674   4.24   0.114       -
1.77
3         3              6304    7.60     1.31  -4.60 0.0807   4.59   0.0476      -
1.04
4         4             22607    0.600    4.23   3.40 0.844    4.18   9.39
1.87
5         5                22   10.3      1.28  -5.29 0.0775   4.53   0.0602      -
1.20
6         6              3928    8.62     1.13  -2.62 0.0604   4.71   0.0111      -
0.587
```



Create a model and show regression modelling assumptions

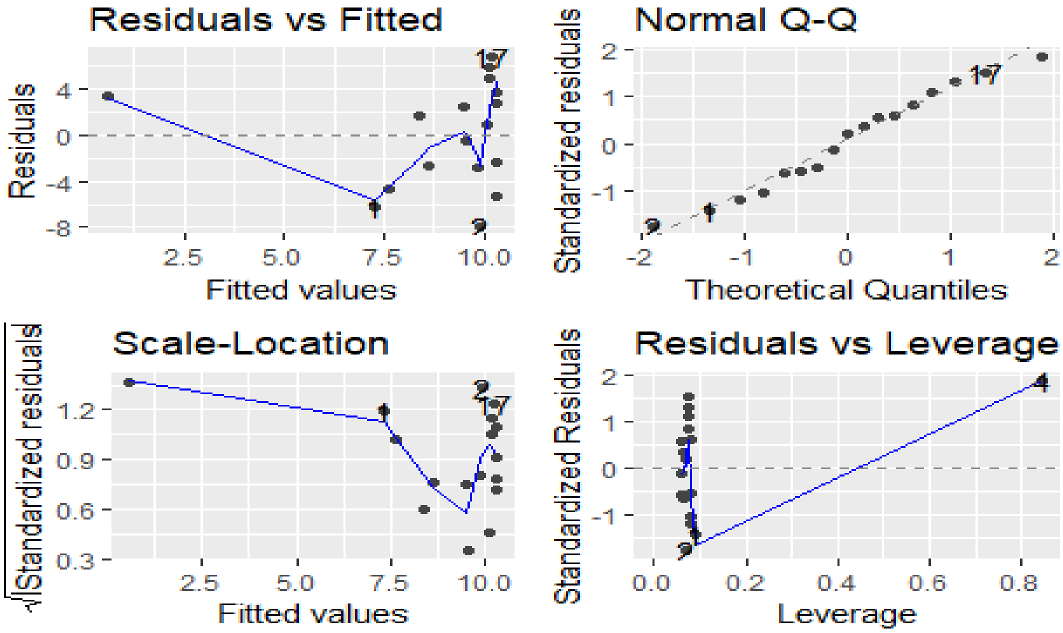
Linearity - The relationship between the predictor (x) and the outcome (y) is assumed to be linear. The linearity assumption can be checked by inspecting the *Residuals vs. fitted* plot (1st plot)



Normality of residuals. The residual errors are assumed to be normally distributed. Constant variance. The QQ plot of residuals can be used to visually check the normality assumption. The normal probability plot of residuals should approximately follow a straight line.

Homogeneity of residuals variance. The residuals are assumed to have a constant variance *(homoscedasticity)*

Independence of residuals error terms



Conclusion

The model can be used to predict the number of tickets in the testing dataset