ABSTRACT

São Paolo du Brazil is the most populated city of Brazil with over than 10 millions of citizen.

Furthermore, São Paulo is the largest city in all of South America and among the largest cities in the world.

Over the years the information has had an increase in movement speed. On the same wavelength, the needs of large districts need to increase the movement of goods and people both in terms of quantity and quality. Like online traffic, urban city traffic also needs efficiency and speed.

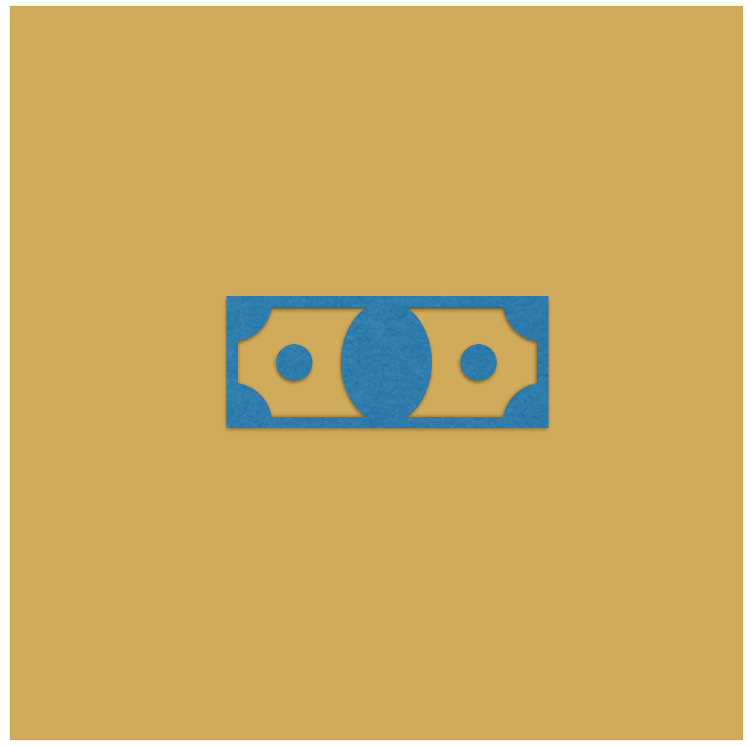
Traffic behavior in cities requires optimization to fulfil what globalization processes and requests of the market are.

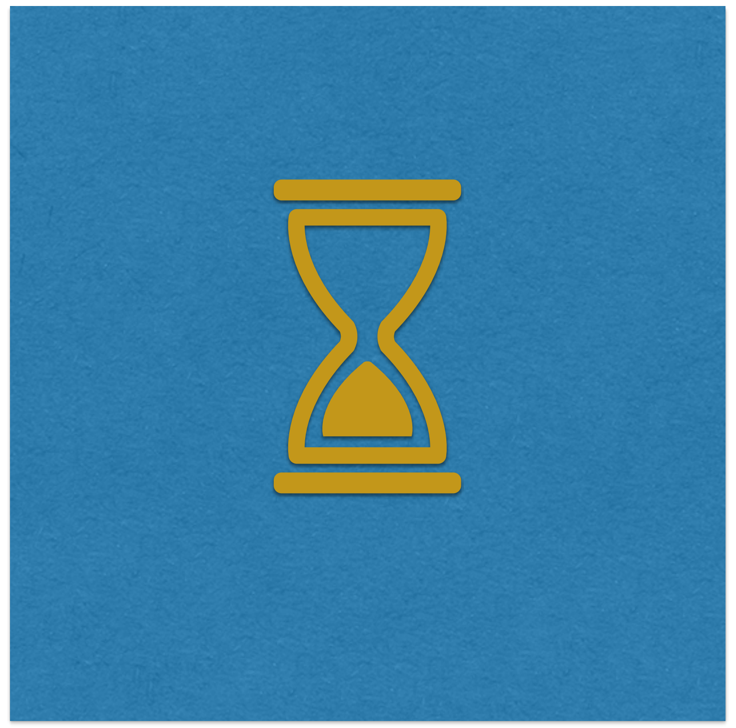
16 milions

TRAFFIC COSTS

Monetary

i.e.: Money, Parking Fees, Motorway Tolls…





Non-Monetary

i.e.: Travel Times onboard a Vehicle, Waiting Times at a Stopped Bus, Parking Search Times.…

BOTH DEPEND ON THE FACTORS THEMSELVES AND ON THE NETWORK IN WHICH THEY ACT

Measurable charges are essentially quantifiable factors that the user must provide to equipe himself in order to his movement

Non-measurable charges are essentially non-quantifiable factors that the user can perceive as negative elements when making a move. It subtract from other activities

ITS: INTELLIGENCE TRANSPORT SYSTEM

Citizens suffer from the decreasing in travel efficiency due to the growing number of people on the move. So, traffic flow's temporal dependence is crucial to the effectiveness of a prediction regarding the shifting traffic conditions (Smith).
In other words, if we can observe traffic's evolution with a temporal pattern we should better predict traffic pattern (Kerner).

HUMAN

DRIVER ASSISTANCE

AUTONOMOUS DRIVING

Smith, B. L. & Oswald, R. K. (2003), Meeting real-time traffic flow forecasting requirements with imprecise computations, Computer Aided Civil and Infrastructure Engineering, 18(3), 201– 13.
Kerner, B. S. (2004a), Three-phase traffic theory and highway capacity, Physica A, 333, 379– 440.

CONTRIBUTION

There are no studies which provide or analyze the traffic of São Paolo with regressive machine learning techniques: the aim of the project is to verify the possibility to develop other kinds of prediction with the dataset proposed, and verify the possibility of prediction with two others Neural Networks

Fit different Machine Learning Regression and Neural Networks Techniques

ORGANIZATION

02 Related Work:
Discussion of previous work and reports

SECTION 2

03 Method:
Description of the system for achieving accurate

SECTION 3

04 Results and Evaluation:
Showing experimental results

SECTION 4

05 Conclusion and Practical Suggestions
Summary of the project and discuss the future possibility

SECTION 5

RELATED WORK

T. Gindele, S. Brechtel and R. Dillmann, "A probabilistic model for estimating driver behaviors and vehicle trajectories in traffic environments," 13th International IEEE Conference on Intelligent Transportation Systems, Funchal, 2010, pp. 1625-1631.

Bryan E. Porter, Kelli J. England, Predicting Red-Light Running Behavior: A Traffic Safety Study in Three Urban Settings, Journal of Safety Research, Volume 31, Issue 1, 2000, Pages 1-8, ISSN 0022-4375

G. Gualtieri, M. Tartaglia, Predicting urban traffic air pollution: A gis framework, Transportation Research Part D: Transport and Environment, Volume 3, Issue 5, 1998, Pages 329-336, ISSN 1361-9209.

Anil Namdeo, Gordon Mitchell, Richard Dixon, TEMMS: an integrated package for modelling and mapping urban traffic emissions and air quality, Environmental Modelling & Software, Volume 17, Issue 2, 2002, Pages 177-188, ISSN 1364-8152

Sassi, R. J., Affonso, C., & Ferreira, R. P. (2011, August). Rough Neuro-Fuzzy Network Applied to Traffic Flow Breakdown in the City of São Paulo. In Management and Service Science (MASS), International Conference on (pp. 1-5). IEEE, 2011.

Affonso, C., Sassi, R. J., & Ferreira, R. P. (2011, July). Traffic flow breakdown prediction using feature reduction through rough-neuro fuzzy networks. In Neural Networks (IJCNN), The International Joint Conference Neural Networks (pp. 1943-1947). IEEE, 2011

Decision Making

Situation and context used with Dynamic Bayesian Network realted to the human safety (Gindele)
Violation of the red-traffic lights (Porter)

Air Pollution

Framework related to the air pollution phenomenon (Gualtieri). That’s kind of prediction can be minimized if exceed traffic time decrease (Namdeo).

Artificial Neural Network

Neuro-Fuzzy Network using Backpropagation (Sassi)
Rough-Fuzzy Sets to use Dynamic Routing without human supervision (Affonso)

DATASET

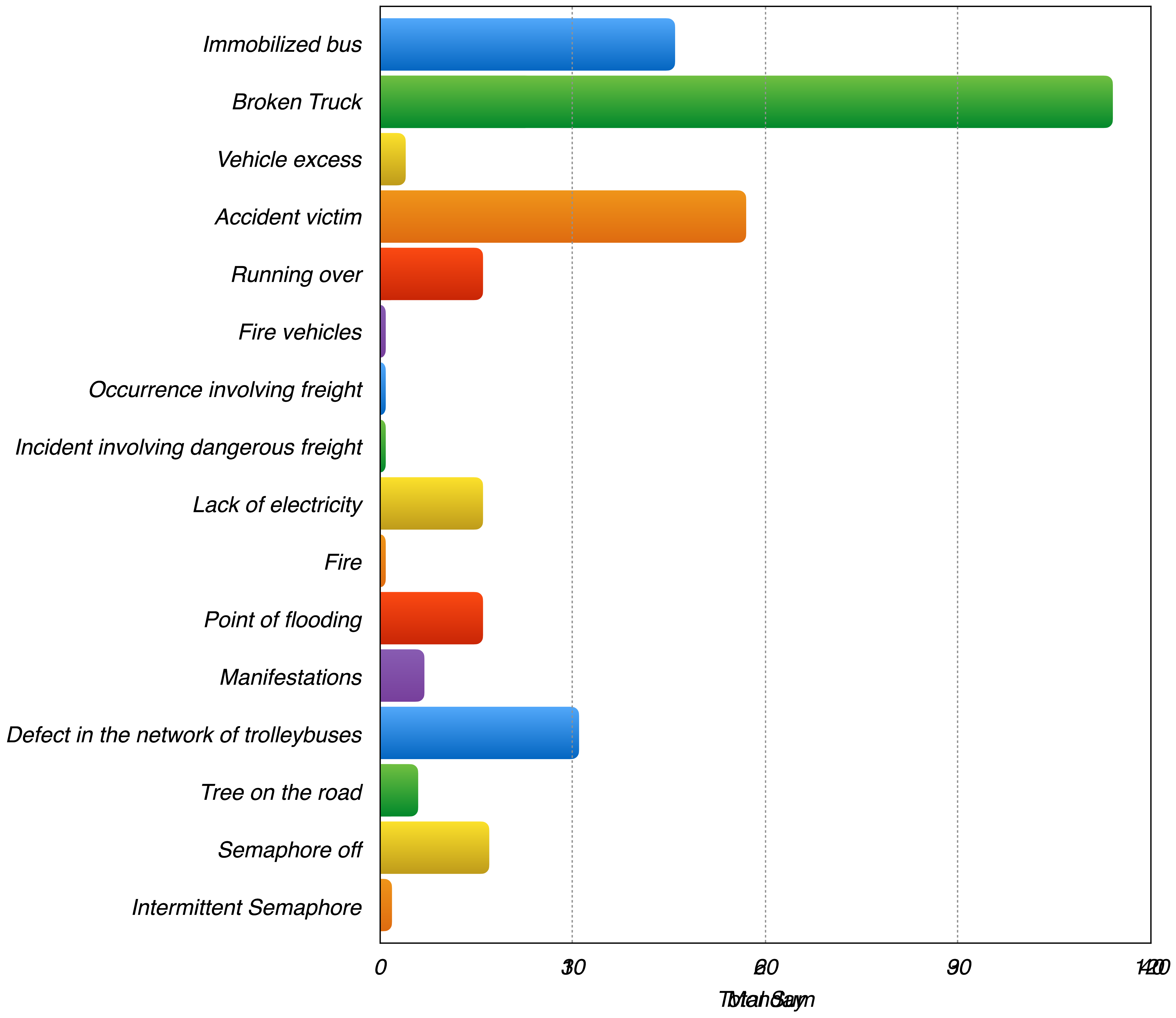
135 Rows and 19 Columns
Surveys divided into 5 days from Monday, December 14, 2009, till Friday, December 18, 2009. The surveys are 27 per day.
The collection provides a survey every 30 minutes, from 7.00 am to 8.00 pm.

The last variable “Slowness in traffic” is the dependent variable, all of others are independent

	count	mean	std	min	25%	50%	75%	max
Day	135.0	3.000000	1.419481	1.0	2.0	3.0	4.00	5.0
Hour (Coded)	135.0	14.000000	7.817890	1.0	7.0	14.0	21.00	27.0
Immobilized bus	135.0	0.340741	0.659749	0.0	0.0	0.0	1.00	4.0
Broken Truck	135.0	0.874074	1.102437	0.0	0.0	1.0	1.00	5.0
Vehicle excess	135.0	0.029630	0.170195	0.0	0.0	0.0	0.00	1.0
Accident victim	135.0	0.422222	0.696116	0.0	0.0	0.0	1.00	3.0
Running over	135.0	0.118519	0.346665	0.0	0.0	0.0	0.00	2.0
Fire vehicles	135.0	0.007407	0.086066	0.0	0.0	0.0	0.00	1.0
Occurrence involving freight	135.0	0.007407	0.086066	0.0	0.0	0.0	0.00	1.0
Incident involving dangerous freight	135.0	0.007407	0.086066	0.0	0.0	0.0	0.00	1.0
Lack of electricity	135.0	0.118519	0.504485	0.0	0.0	0.0	0.00	4.0
Fire	135.0	0.007407	0.086066	0.0	0.0	0.0	0.00	1.0
Point of flooding	135.0	0.118519	0.712907	0.0	0.0	0.0	0.00	7.0
Manifestations	135.0	0.051852	0.222554	0.0	0.0	0.0	0.00	1.0
Defect in the network of trolleybuses	135.0	0.229630	0.818998	0.0	0.0	0.0	0.00	8.0
Tree on the road	135.0	0.044444	0.206848	0.0	0.0	0.0	0.00	1.0
Semaphore off	135.0	0.125926	0.464077	0.0	0.0	0.0	0.00	4.0
Intermittent Semaphore	135.0	0.014815	0.121261	0.0	0.0	0.0	0.00	1.0
Slowness in traffic (%)	135.0	10.051852	4.363243	3.4	7.4	9.0	11.85	23.4

DATA EXPLORATION

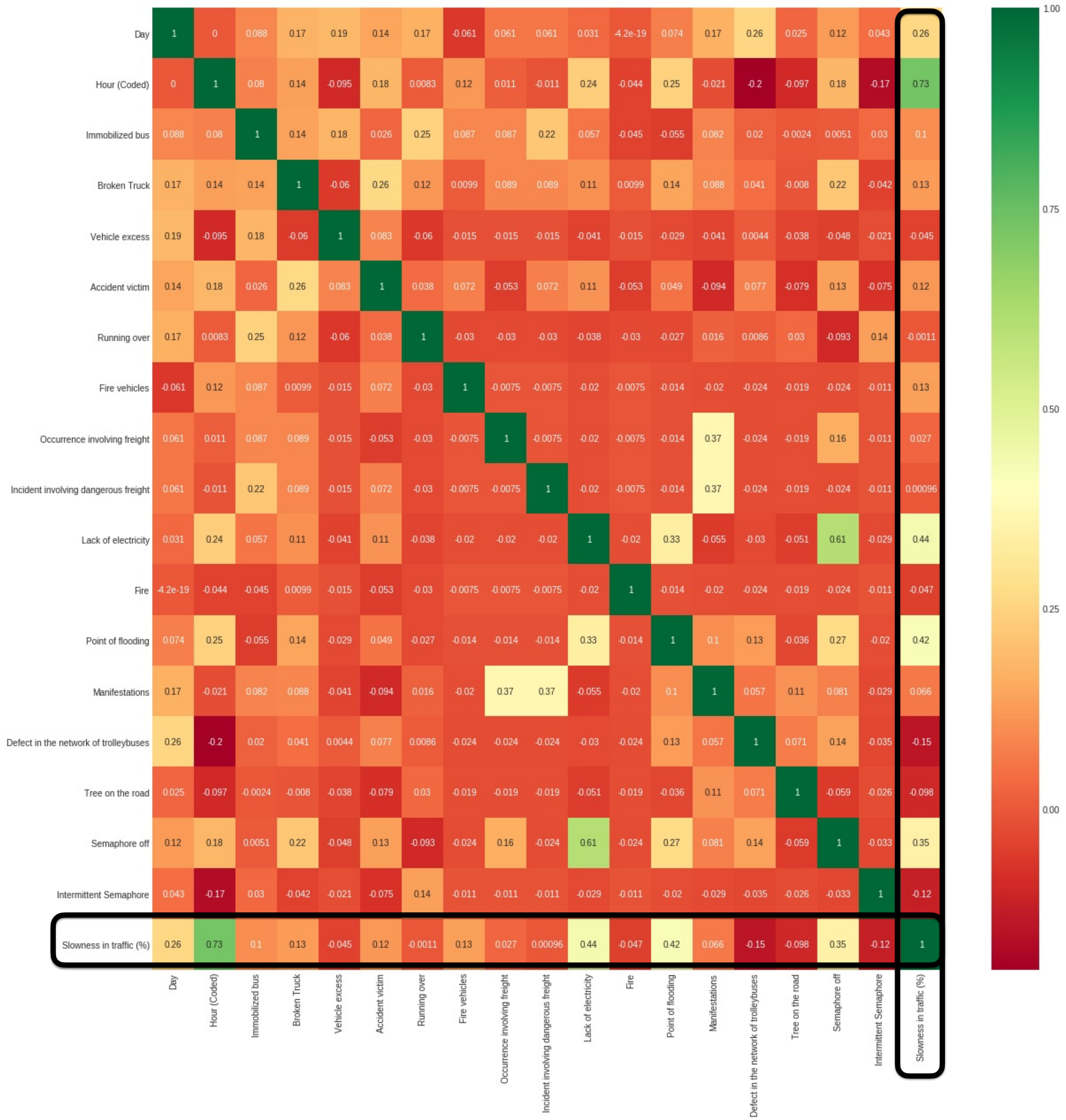
This graph provides the behavior of accidents during different days which cause are the variable considered in all five days registered.



DATASET EXPLORATION

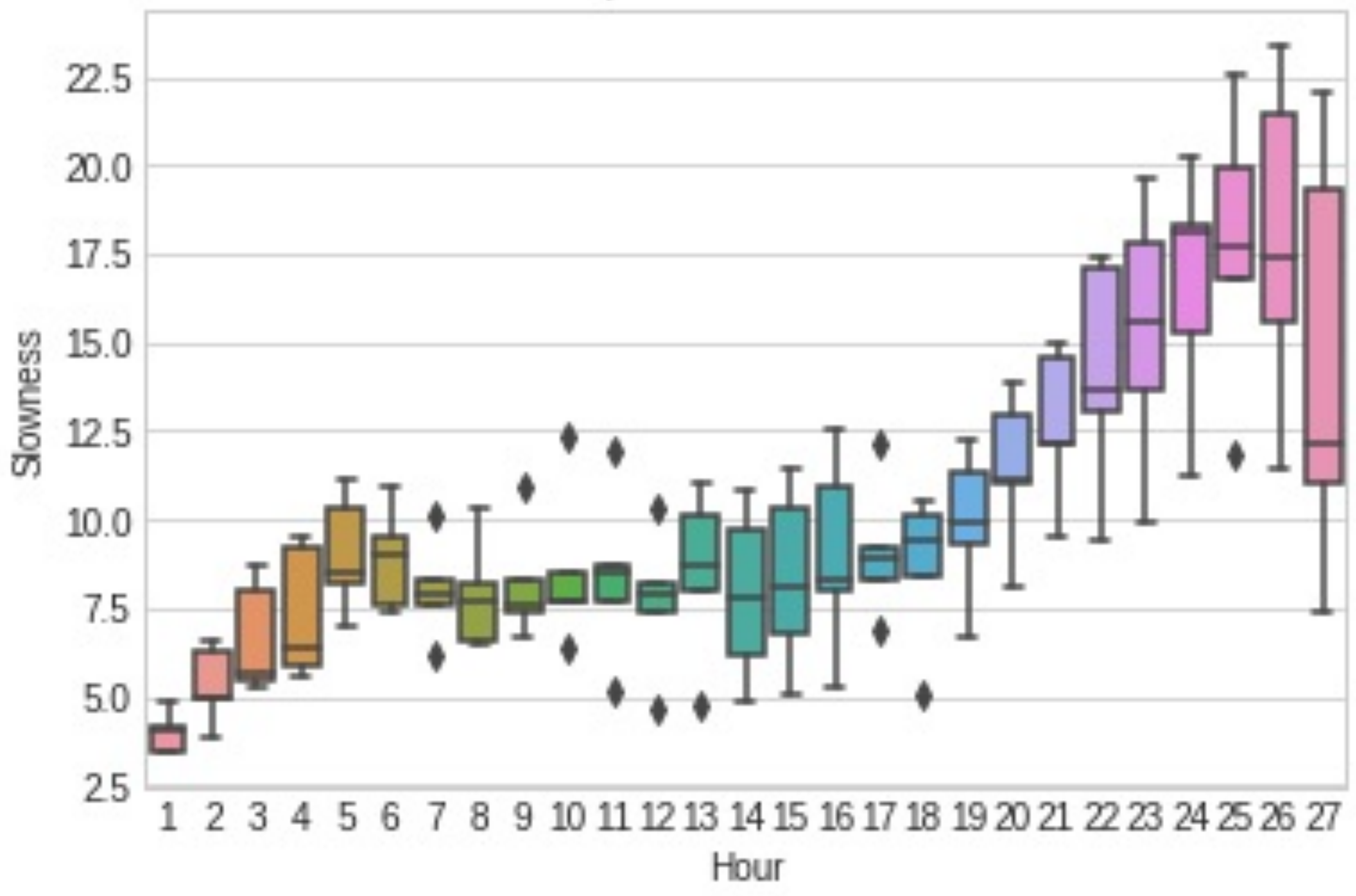
This graph provides the correlazione between the variable, some in the following bullet with respect to the depedent variable:
(e.g. Slowness in traffic)

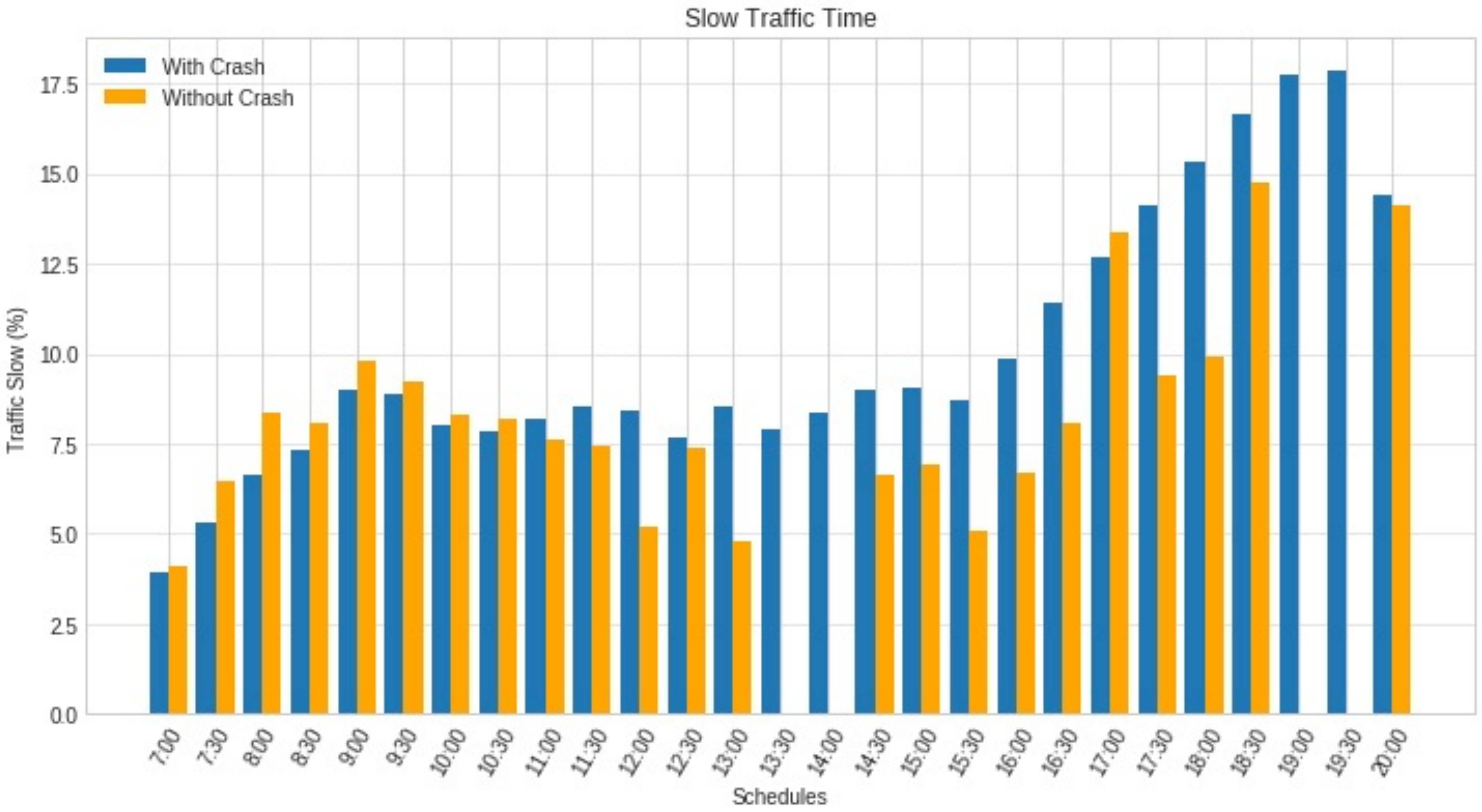
Hour (Coded): 0.729962
Lack of electricity: 0.436569
Point of flooding: 0.420016
Semaphore off: 0.347242
Day: 0.261948
Fire vehicles: 0.134103
Broken Truck: 0.131998
Accident victim: 0.121730
Immobilized bus: 0.101143
Manifestations: 0.066377
Occurrence involving freight: 0.026791
Incident involving dangerous freight: 0.000957
Running over:-0.001133
Vehicle excess: -0.045297
Fire: -0.046737
Tree on the road: -0.098489
Intermittent Semaphore: -0.119942
Defect in the network of trolleybuses: -0.147035



DATASET EXPLORATION

Below you can see how the Slowness behavior in different hour (from 1 to 27 calculate every 30 minutes from 7 am to 8 pm)





Above you can see the effects of the low amount of data. For at some times, the slowness is greater without incident than when there are incidents. However, taking into account the data, it is possible that *until 11:00 am the occurrence of incidents does not affect traffic slowness so much*. *From 14:30, the incidents have a greater impact on the slow traffic*.

Transformation of the Hour variable that means an observation every 30 minutes from 7 am to 8 pm:

originally it is compose by 27 ordinal value, from 1 to 27, every of them with a specific time (e.g. 1=7.00, 2=7.30, 3=8.00 and so far and so on).

The variabile is transformed from ordinal to binary value assigned to the respective new column (27).

DATA PROCESSING

The X vectors are all of the independent variables.
The y vector is the dependent variable “Slowness in traffic”

The dataset is now ready to the be split in training and test set.

Hour (Coded)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
1	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
23	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
24	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
25	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
26	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
27	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0

MODELS

POLYNOMIAL REGRESSION

Polynomial regression is a form of regression analysis in which the relationship between the independent variable x and the dependent variable y is modelled as an n^th degree polynomial in x.

DECISION TREE

Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees. Growing a tree involves deciding on which features to choose and what conditions to use for splitting, along with knowing when to stop.

RANDOM FOREST

The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees.

MODELS

MLP

A MultiLayer Perceptron (MLP) is a class of feedforward artificial neural network (ANN).
An MLP consists of at least three layers of nodes: an input layer, a hidden layer and an output layer.

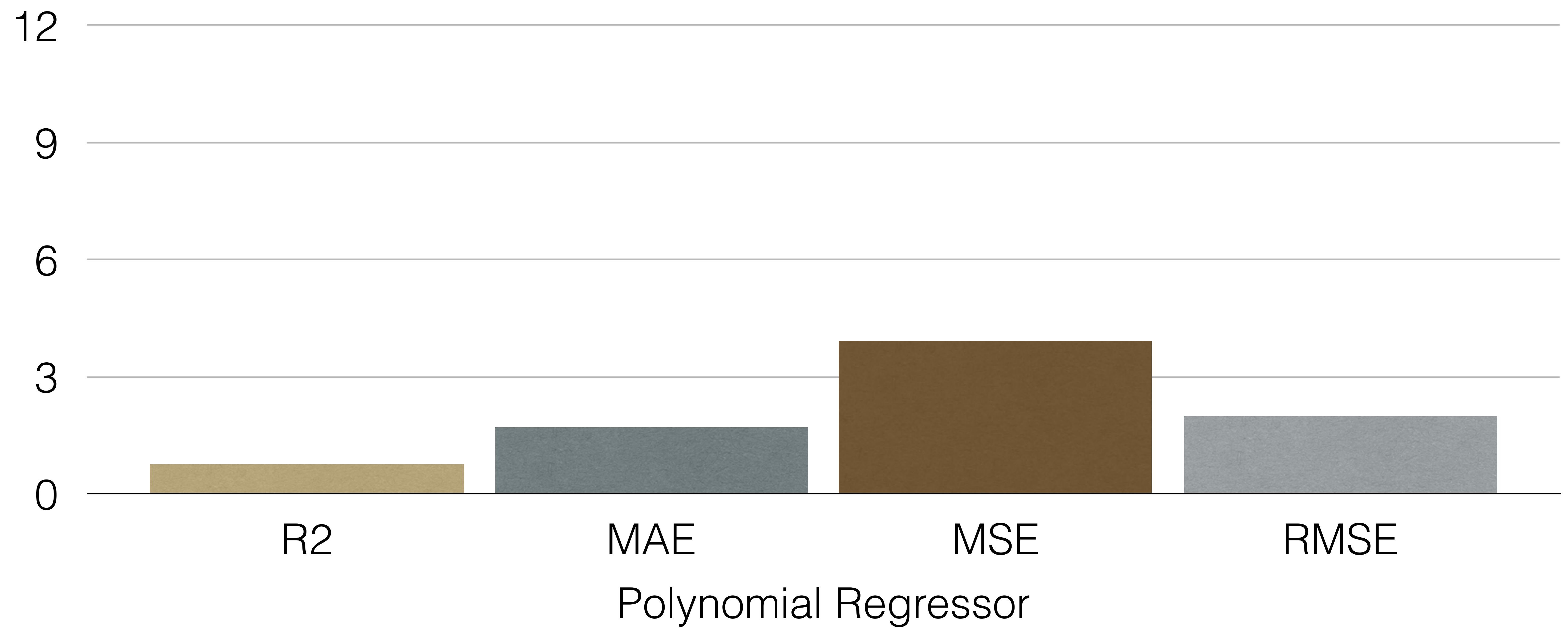
LSTM

A common LSTM unit is composed of a cell, an input gate, an output gate and a forget gate. The cell remembers values over arbitrary time intervals and the three gates regulate the flow of information into and out of the cell.

RESULTS

This graph provides the results of different models

The X vectors are all of the independent variables
The y vector is the dependent variable “Slowness in traffic”



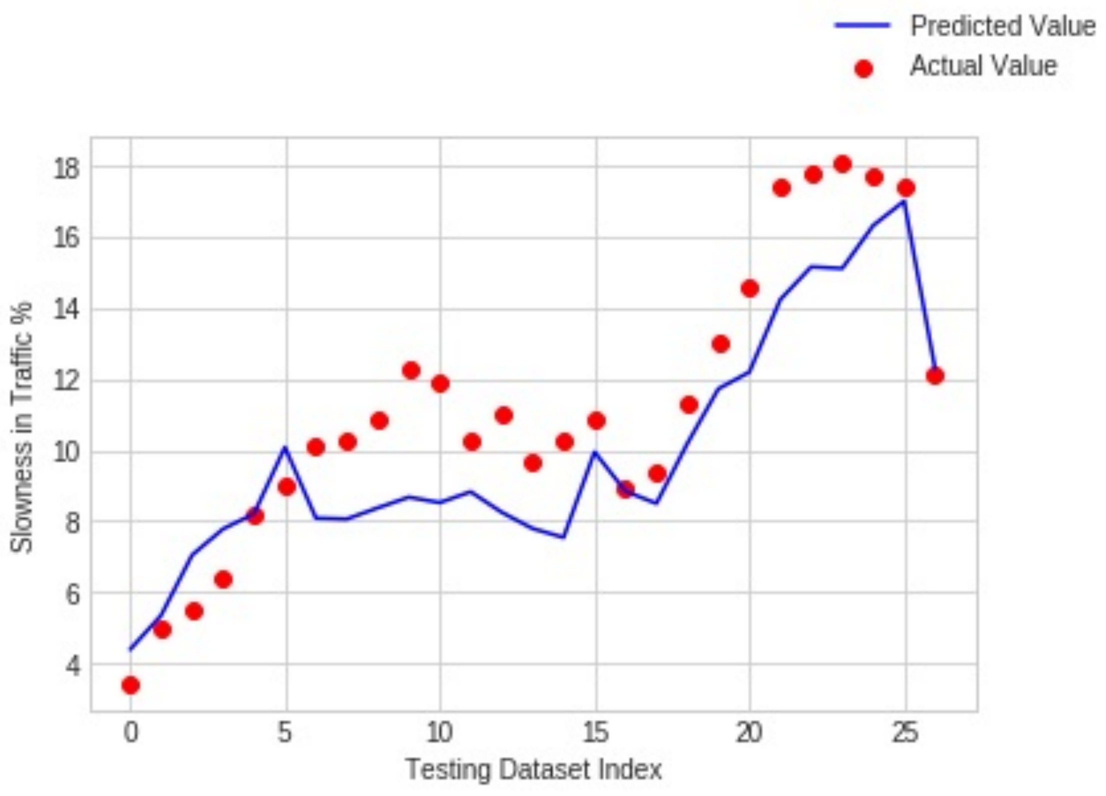
RESULTS

This table provides the results of different models

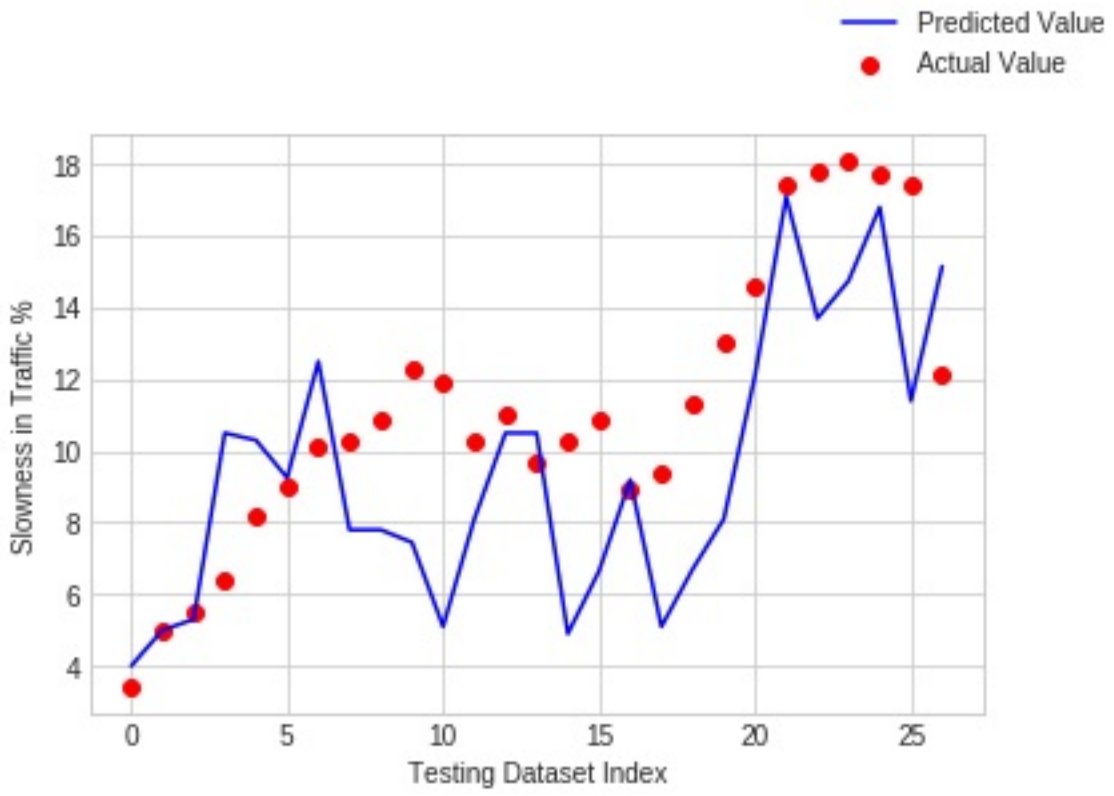
Model	R2	MAE	MSE	RMSE
LSTM			0,018	
Polynomial	0.74	1.68	3.91	1.97
Random Forest	0.51	2.28	7.32	2.70
MLP	0.49	2.39	7.61	2.76
Decision Tree	0.24	2.75	11.49	3.38

MODELS

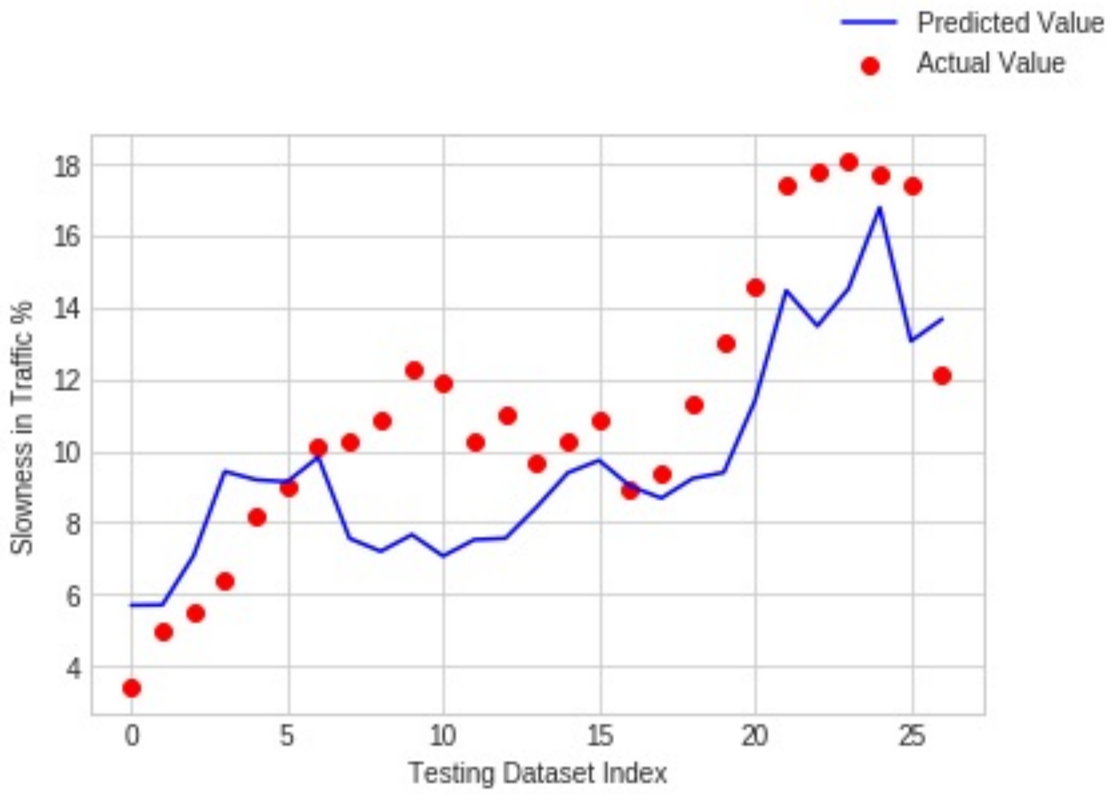
POLYNOMIAL REGRESSION



DECISION TREE

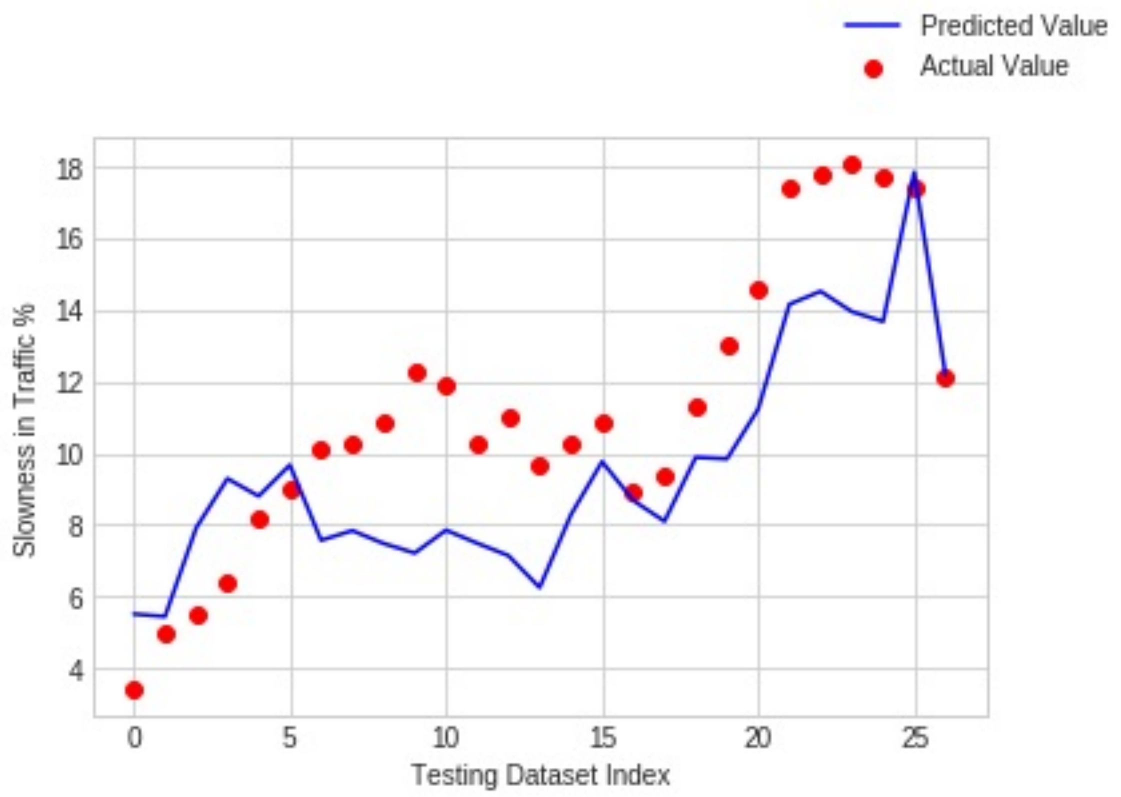


RANDOM FOREST

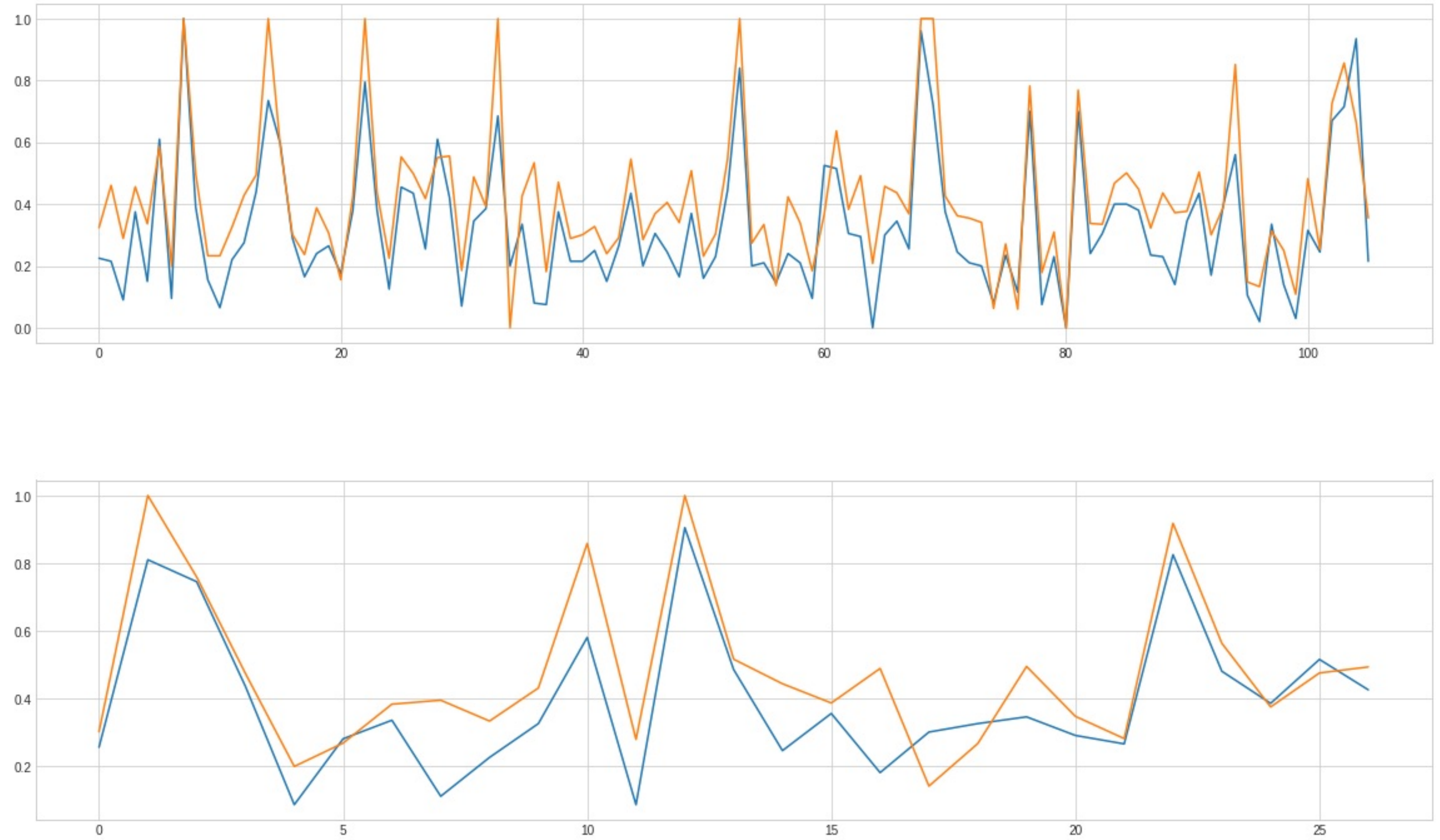


MODELS

MLP



LSTM



CONCLUSION

In this project, it is evaluated the prediction of the behavior of urban traffic in the city of São Paolo. The problem of traffic is known to everybody that has or not a car. The air pollution, the price of oil and the time spent in the car are some of the problems related to the congestion referred to the urban traffic.

After some transformation of the dataset, it is proposed different models:

- the best regression model results are achieved by the *Polynomial Regression* with *0.75 of R2, 1.68 of MAE and 3.91 of MSE*;
- the best Neural Networks results are achieved by *LSTM* with *1.83% MSE*

PRACTICAL SUGGESTIONS

1

Scalability &

Homogenization and scalability of the models that are used, be they neural networks, regressive or classification models.

2

Avoid Overfitting

A higher collection of values could be interesting to apply that's kind of prediction in a real scenario. The hyperparameters tuning is not used in this project because of the few data available for what concern regression models, different for Neural Networks

3

Other Models

This project is a practice prove of only three easy models, also without a grid search

THANK YOU!