INTRODUCTION

The models described in this report were created with the aim to analyse seismological data of earthquakes and answer some questions of interest.

An earthquake is a natural event caused by the movement of tectonic plates that creates compression or extension forces of the external layer of the earth called the *crust*. These forces then generate fractures which produce earthquakes. These are called tectonic earthquakes. There are a second type called volcanic earthquakes created by the rising of magma. They both generate a train of wave recorded by an instrument called a *geophone*. The graphs obtained by the geophone are called *seismogram* and these graphs make it possible to determine many values which some of them are reported in the dataset for this report.

The data set is composed of 12 variables reporting geographical information of earthquakes such as the country where they happened, the latitude and longitude, and the values of different magnitude scales used by seismologists to better describe the magnitude of a telluric event. The variables in the dataset are:

- id = ID of record
- lat = Latitude of earthquake (degrees)
- long = Longitude of earthquake (degrees)
- dist = Distance travelled by earthquake in a particular direction (km)
- depth = Depth of earthquake (km)
- md = Magnitude of earthquake, estimated from the duration of seismic wave-train (Md)
- richter = Intensity of earthquake (Richter)
- mw = Moment magnitude scale value of earthquake (Mw)
- ms = Surface-wave magnitude scale value of earthquake (Ms)
- mb = Bodywave magnitude value, measured using P-waves and a short-period seismograph in the first few seconds of an earthquake (mb)
- country = Country of earthquake
- direction = Direction of earthquake

After a brief exploratory analysis which provides information about the general features of the dataset, there will be an analysis to evaluate if the mean of *xm* (see page 6 for a definition of *xm*) is statistically different from 4.1 and if there also is a difference of the average of the moment magnitude (*mw*) in different countries where the earthquakes happened. Afterwards, a model will be created with *richter* as the response variable and all the other as explanatory variables. Automatic model selection technique has been applied to determine which variables are related to *richter*. Finally, a new variable called *serious* will be created which has value of 1 if *richter* is equal or greater than 5 and 0 if *richter* is less than 5. This will be used as response to create two regression models, for one of them all variables will be used as explanatory variables except for *id, mw, richter* and *xm*, and for the other one, *xm* will be the only explanatory variable.

EXPLORATORY ANALYSIS

Table 1 reports all the descriptive statistics of the variables. It is possible to notice for example that they all are positive and that the different scales used to measure an earthquake have as minimum value 0 and maximum between 7.1 and 7.9.

Variable	Mean	Std Dev	Minimum	Maximum	Lower 95% CL for Mean	Upper 95% CL for Mean	Lower Quartile	Median	Upper Quartile
id	11871.00	6853.58	1.0000000	23741.00	11783.82	11958.18	5936.00	11871.00	17806.00
lat	37.9521937	2.1944648	29.7400000	46.3500000	37.9242779	37.9801094	36.2200000	38.2100000	39.3600000
long	30.7065322	6.5638114	18.3400000	48.0000000	30.6230340	30.7900303	26.1600000	28.2400000	33.7300000
dist	3.1750149	4.7154610	0.1000000	95.4000000	3.0828677	3.2671622	1.4000000	2.3000000	3.6000000
depth	18.4424076	23.2267930	0	225.0000000	18.1469400	18.7378753	5.0000000	10.0000000	22.0000000
md	1.9076071	2.0593288	0	7.4000000	1.8814104	1.9338038	0	0	3.8000000
richter	2.2003875	2.0805645	0	7.2000000	2.1739207	2.2268543	0	3.5000000	4.0000000
mw	4.4775758	1.0487482	0	7.7000000	4.4483529	4.5067986	4.1000000	4.7000000	5.0000000
ms	0.6789478	1.6764715	0	7.9000000	0.6576214	0.7002742	0	0	0
mb	1.6954888	2.1466149	0	7.1000000	1.6681818	1.7227959	0	0	4.1000000

Table 1

Table of descriptive statistics

Variable	N Miss
id	0
lat	0
long	0
dist	13679
depth	0
md	0
richter	0
mw	18791
ms	0
mb	0

Table 2

Missing values for each variable

Table 2 shows that only *mw* and *dist* have missing values of 18791 and 13679 respectively. These correspond to a 79% and 58% of the data for the two variables respectively. Computing missing data using mean or mode would mean to alter the two variables, therefore the missing values will be maintained for the both of them.

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

lat	lat	depth	long	dist	mw	ms	md	richter	mb
	1.00000	-0.24222	0.23749	0.07934	-0.05950	0.05564	0.05506	-0.03522	0.01178
		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0696
	23741	23741	23741	10062	23647	23741	23741	23741	23741
long	long	lat	richter	mw	md	depth	ms	dist	mb
	1.00000	0.23749	-0.14168	0.13424	0.10749	-0.06622	0.04412	0.02833	0.00441
		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0045	0.4965
	23741	23741	23741	23647	23741	23741	23741	10062	23741
dist	dist	lat	depth	long	ms	richter	mw	md	mb
	1.00000	0.07934	0.02869	0.02833	-0.00906	0.00630	-0.00591	-0.00431	-0.00009
		<.0001	0.0040	0.0045	0.3636	0.5278	0.5533	0.6653	0.9927
	10062	10062	10062	10062	10062	10062	10062	10062	10062
depth	depth	mb	ms	lat	richter	mw	long	md	dist
	1.00000	0.31320	0.25969	-0.24222	0.15044	0.14640	-0.06622	0.04289	0.02869
		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0040
	23741	23741	23741	23741	23741	23647	23741	23741	10062
md	md	ms	mw	richter	long	lat	depth	mb	dist
	1.00000	0.46177	0.32139	-0.23659	0.10749	0.05506	0.04289	-0.02287	-0.00431
		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0004	0.6653
	23741	23741	23647	23741	23741	23741	23741	23741	10062
richter	richter	ms	mb	md	depth	long	mw	lat	dist
	1.00000	0.41895	0.24000	-0.23659	0.15044	-0.14168	0.04732	-0.03522	0.00630
		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.5278
	23741	23741	23741	23741	23741	23741	23647	23741	10062
mw	mw	ms	md	mb	depth	long	lat	richter	dist
	1.00000	0.39672	0.32139	0.31603	0.14640	0.13424	-0.05950	0.04732	-0.00591
		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.5533
	23647	23647	23647	23647	23647	23647	23647	23647	10062
ms	ms	mb	md	richter	mw	depth	lat	long	dist
	1.00000	0.58826	0.46177	0.41895	0.39672	0.25969	0.05564	0.04412	-0.00906
		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.3636
	23741	23741	23741	23741	23647	23741	23741	23741	10062
mb	mb	ms	mw	depth	richter	md	lat	long	dist
	1.00000	0.58826	0.31603	0.31320	0.24000	-0.02287	0.01178	0.00441	-0.00009
		<.0001	<.0001	<.0001	<.0001	0.0004	0.0696	0.4965	0.9927
	23741	23741	23647	23741	23741	23741	23741	23741	10062

Table 5

Table of correlation of all variables except id and the two character variables country and direction

Table 5 shows the correlation between all the numeric variables except *id*, and they present weak correlation between each other, the highest value of 0.588 is between *ms* and *mb*.

Figure 1 to 5 are histograms of all the variables representing a scale to measure an earthquake. Zero values for some of these variables have been removed because there are too many, making the histogram concentrated only on these values and not showing the pattern of the other values as shown in figure 1a and 1b. They are the same histogram but in figure 1b all the zeros have been removed in order to make the pattern of the other values clearer. All the histograms show that the numeric variables in the dataset are characterised by a non-normal distribution.

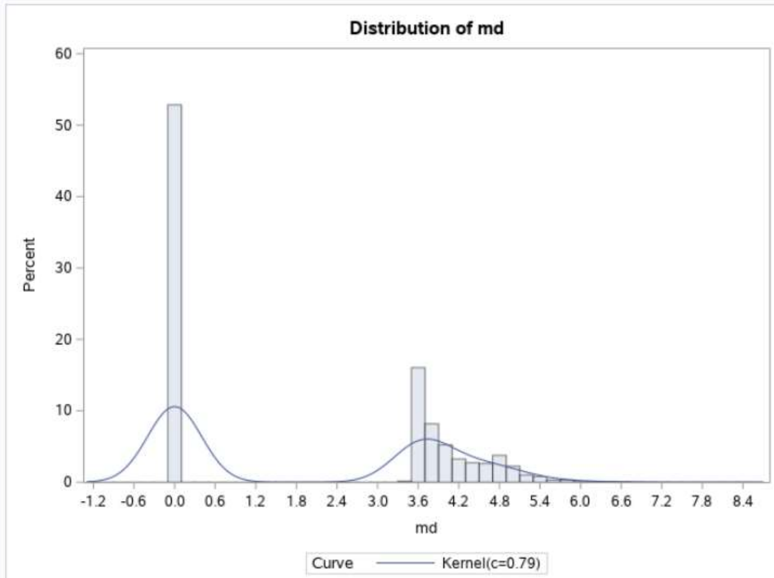

Figure 1a
Histogram of the variable md with zero values

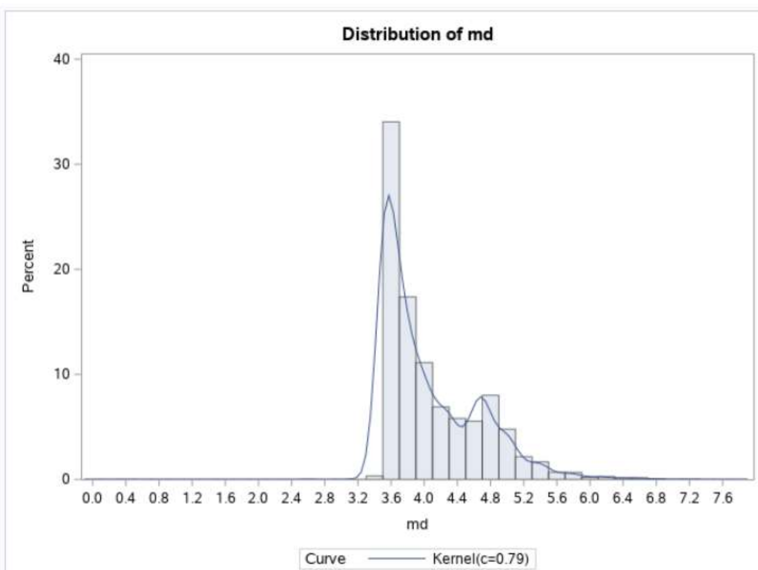

Figure 1b
Histogram of the variable md without zero values

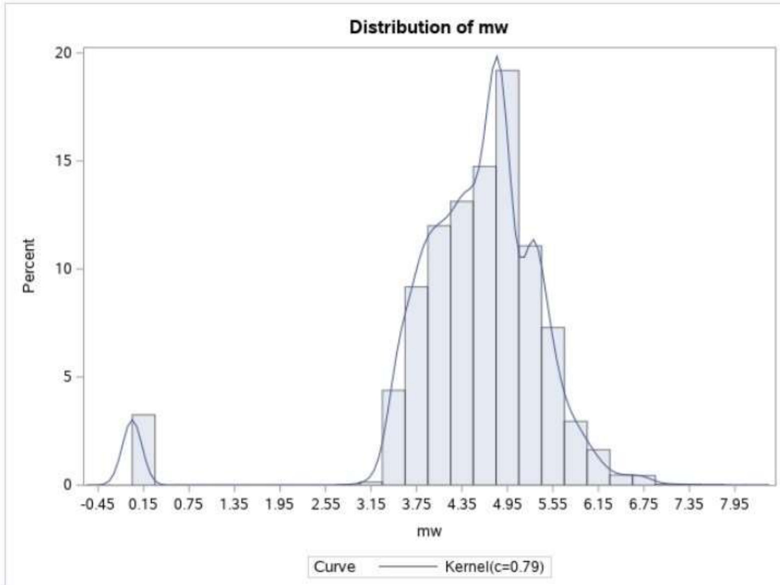

Figure 2
Histogram of the variable mw

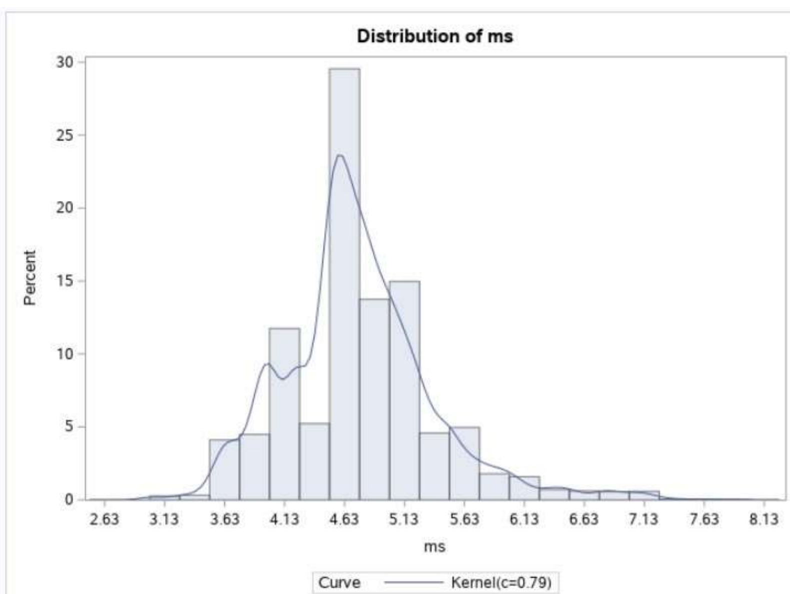

Figure 3
Histogram of the variable ms

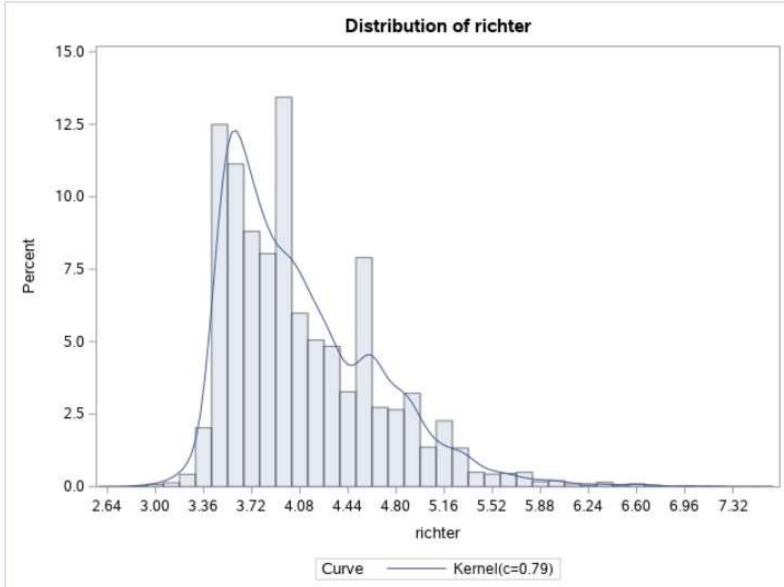

Figure 4

Histogram of the variable richter

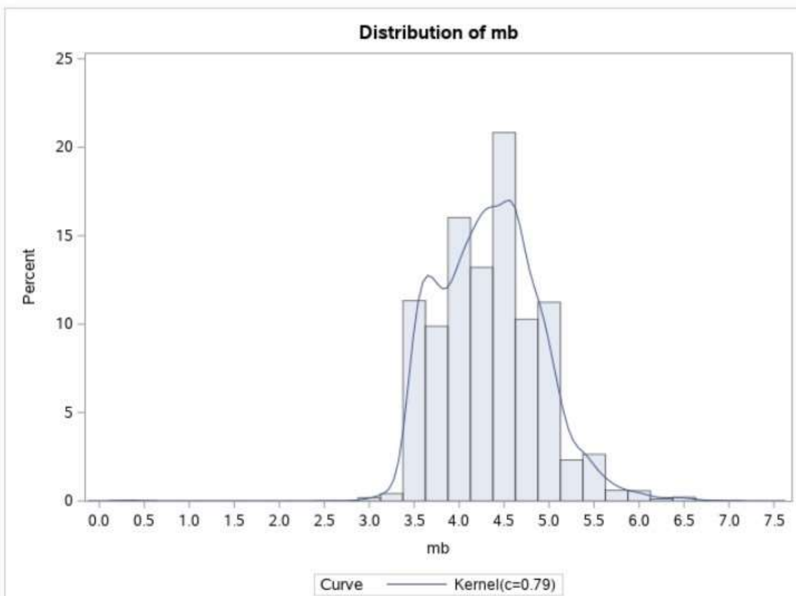

Figure 5

Histogram of the variable mb

Figure 6 shows a bar chart with the frequency of telluric events happening in different countries. Turkey, with a number of events around 1200, is the country experiencing the major number of earthquakes, amongst those taken into consideration in this report. It is followed by the Mediterranean area with a much lower number around 500, then Greece with a number around 400 and Aegean see around 200. All the other remaining countries are less than 100.

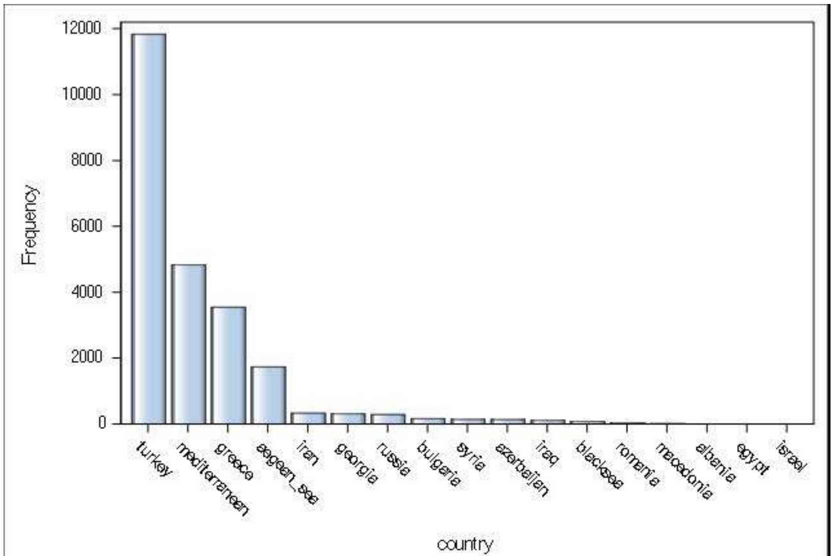

Figure 6
Bar chart

ANALYSIS

When an earthquake happens, different values are calculated through seismograms and the highest of these values is taken to describe the event. A new variable called *xm* is created and represents this highest value amongst all the scales used to determine the magnitude of an earthquake which for this report are: *md, richter, mw, ms* and *mb*. It is now of interest to understand if the mean of *xm* is statistically different from the value of 4.1. Therefore, a one sample T test using the unmodified dataset has been performed, with the null hypothesis that the mean of *xm* is equal to 4.1.

Obs	Variable	N	Mean	LowerCLMean	UpperCLMean	StdDev	LowerCLStdDev	UpperCLStdDev	UMPULowerCLStdDev	UMPUUpperCLStdDev	StdErr	Minimum	Maximum
1	xm	23741	4.0552	4.0479	4.0625	0.5744	0.5692	0.5796	0.5692	0.5796	0.00373	0.3000	7.9000

Obs	Variable	Mean	LowerCLMean	UpperCLMean	StdDev	LowerCLStdDev	UpperCLStdDev	UMPULowerCLStdDev	UMPUUpperCLStdDev
1	xm	4.0552	4.0479	4.0625	0.5744	0.5692	0.5796	0.5692	0.5796

Obs	Variable	tValue	DF	Probt
1	xm	-12.02	23740	<.0001

Table 6
Statistics and T test

From table 6 it is possible to observe that the model is statistically significant because the p-value is lower than the threshold of 0.05, hence the null hypothesis can be rejected. However, looking at the diagnostic plots in figure 7 and 8, the assumption of a normal distribution of the data seems to be not reasonable. For this reason, transformations of the data have been applied. Log and square root transformations do not resolve this issue. Taking instead the inverse of *xm*, the normal distribution seems to be respected as shown in figure 9 and 10. Figure 10 is the QQplot and it is clear that the data are following the diagonal line. From table 7 it is possible to deduce that the model is statistically significant because the p-value is still less than 0.05. The assumption of independency of the observations is reasonable because the dataset is a collection of values describing different telluric events.

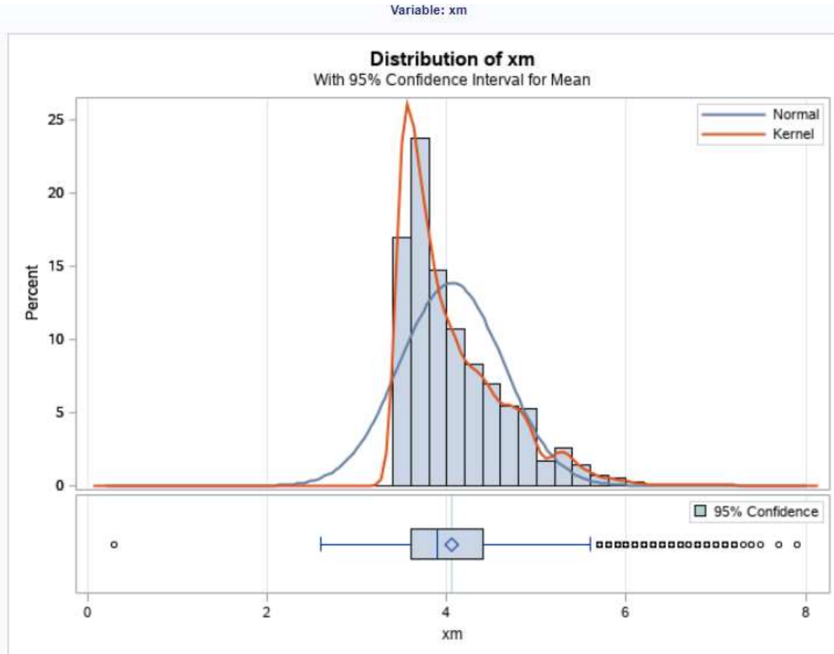

figure 7
Histogram of xm with overimposed the Normal and Kernel distributin

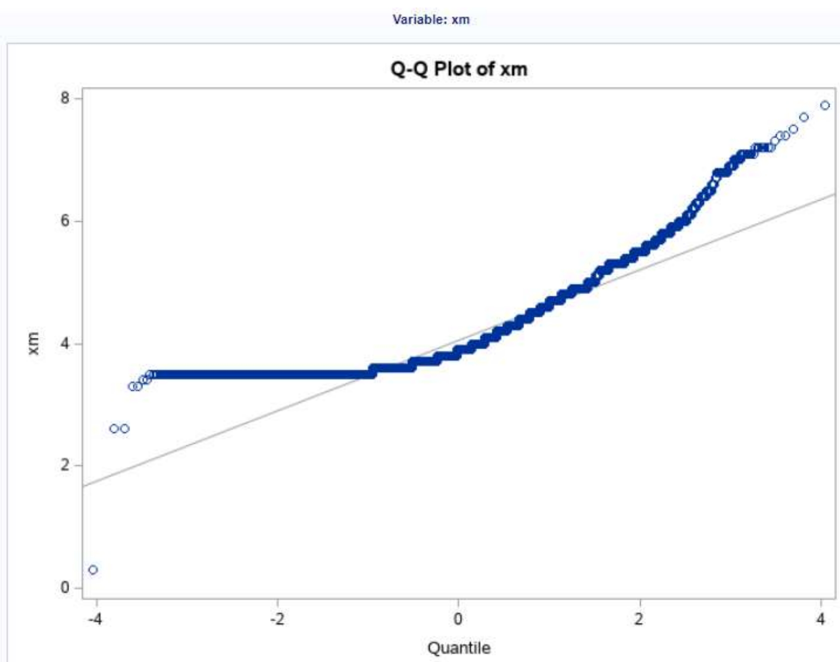

figure 8
QQ plot of xm

Obs	Variable	N	Mean	LowerCLMean	UpperCLMean	StdDev	LowerCLStdDev	UpperCLStdDev	UMPULowerCLStdDev	UMPUUpperCLStdDev	StdErr	Minimum	Maximum
1	Invxm	23741	0.2510	0.2506	0.2515	0.0368	0.0365	0.0371	0.0365	0.0371	0.000239	0.1266	3.3333

Obs	Variable	Mean	LowerCLMean	UpperCLMean	StdDev	LowerCLStdDev	UpperCLStdDev	UMPULowerCLStdDev	UMPUUpperCLStdDev
1	Invxm	0.2510	0.2506	0.2515	0.0368	0.0365	0.0371	0.0365	0.0371

Obs	Variable	tValue	DF	Probt
1	Invxm	-16118	23740	<.0001

Table 7
Statistics and T test of the inverse of xm

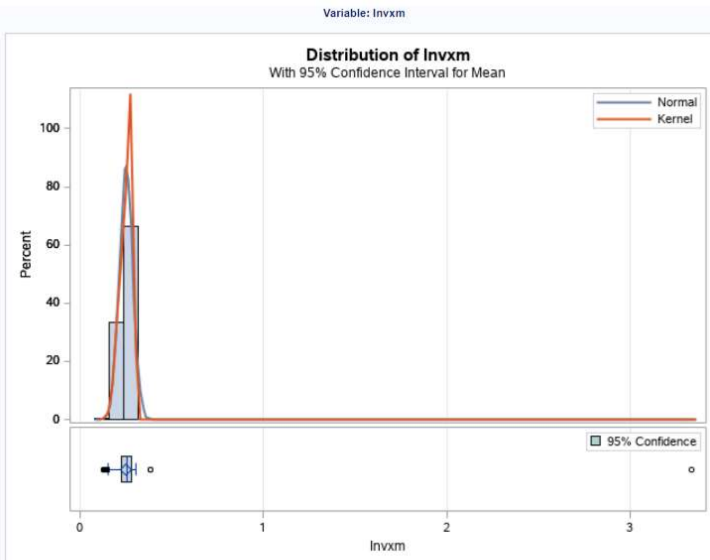

figure 9
Histogram of the incverse of xm with overimposed the Normal and Kernel distributin

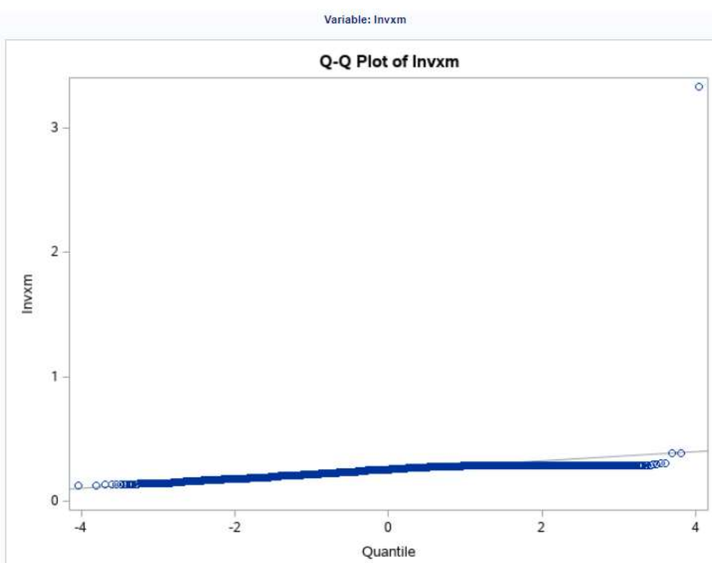

figure 10
QQ plot of the inverse of xm

It is also of interest to investigate if the mean of one of the scales to measure the magnitude of an earthquake, in this case *mw*, is different between countries in which the earthquake happened.

Obs	Class	Levels	Values
1	country	17	aegean_sea albania azerbaijan blacksea bulgaria egypt georgia greece iran iraq israel macedonia mediterranean romania russia syria turkey

Obs	Label	N	NObsUsed	NObsRead	SumFreqsUsed	SumFreqsRead	Label1	nvalue1	DependentVariables
1	Number of Observations Read	23741	4950	23741	4950	23741	Number of Observations Read	23741	mw
2	Number of Observations Used	4950	4950	23741	4950	23741	Number of Observations Used	4950	mw

Obs	Dependent	Source	DF	SS	MS	FValue	ProbF
1	mw	Model	15	218.931125	14.595408	13.78	<.0001
2	mw	Error	4934	5224.339784	1.058845	_	_
3	mw	Corrected Total	4949	5443.270909	_	_	_

Obs	Dependent	RSquare	CV	RootMSE	DepMean
1	mw	0.040221	22.98123	1.029002	4.477576

Obs	Dependent	HypothesisType	Source	DF	SS	MS	FValue	ProbF
1	mw	1	country	15	218.9311248	14.5954083	13.78	<.0001
2	mw	3	country	15	218.9311248	14.5954083	13.78	<.0001

Table 8
Statistics and ANOVA model

Table 8 reports this ANOVA model and it is possible to notice that the p-value of the hypothesis type 3 (last table in table 8) is less than the threshold of 0.05, confirming that there is a statistical difference between the different countries taken into consideration for this report.

Figure 11 is the diagnostic plots. QQ-plot and the histogram of the residuals show a sufficient normal distribution. However, the QQ-plot shows a strange pattern in its left side which would be worthy a further investigation but for time reason it is not possible to effectuate.

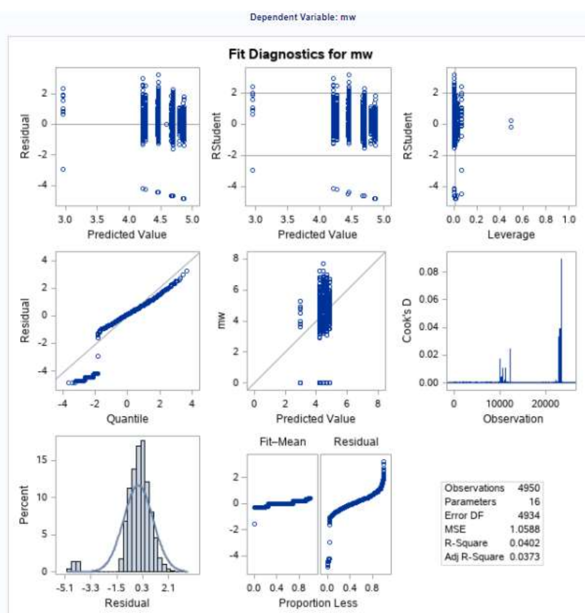

Figure 11
Diagnostic plots

The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

country	mw LSMEAN	LSMEAN Number
aegean_sea	4.26528926	1
albania	4.60000000	2
azerbaijan	4.85000000	3
blacksea	4.66585366	4
bulgaria	4.87435897	5
egypt	4.25000000	6
georgia	4.67672414	7
greece	4.21114488	8
iran	4.87303371	9
iraq	4.84137931	10
macedonia	2.95333333	11
mediterranean	4.69255429	12
romania	4.44666667	13
russia	4.80081301	14
syria	4.44444444	15
turkey	4.45932282	16

Table 9
Index for each level of the variable country

The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

Least Squares Means for effect country
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: mw

i/j	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1		1.0000	0.0067	0.6201	0.0006	1.0000	0.0346	1.0000	0.0002	0.2417	0.0002	<.0001	1.0000	0.0003	1.0000	0.2800
2	1.0000		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9766	1.0000	1.0000	1.0000	1.0000	1.0000
3	0.0067	1.0000		1.0000	1.0000	1.0000	0.9996	0.0002	1.0000	1.0000	<.0001	0.9988	0.9934	1.0000	0.9448	0.1923
4	0.6201	1.0000	1.0000		0.9996	1.0000	1.0000	0.2853	0.9996	1.0000	<.0001	1.0000	1.0000	1.0000	1.0000	0.9968
5	0.0006	1.0000	1.0000	0.9996		1.0000	0.9956	<.0001	1.0000	1.0000	<.0001	0.9826	0.9854	1.0000	0.8891	0.0397
6	1.0000	1.0000	1.0000	1.0000	1.0000		1.0000	1.0000	1.0000	1.0000	0.9536	1.0000	1.0000	1.0000	1.0000	1.0000
7	0.0346	1.0000	0.9996	1.0000	0.9956	1.0000		0.0005	0.9938	1.0000	<.0001	1.0000	1.0000	0.9999	0.9996	0.6854
8	1.0000	1.0000	0.0002	0.2853	<.0001	1.0000	0.0005		<.0001	0.0857	0.0003	<.0001	1.0000	<.0001	0.9988	<.0001
9	0.0002	1.0000	1.0000	0.9996	1.0000	1.0000	0.9938	<.0001		1.0000	<.0001	0.9715	0.9843	1.0000	0.8783	0.0188
10	0.2417	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.0857	1.0000		<.0001	1.0000	0.9982	1.0000	0.9882	0.8333
11	0.0002	0.9766	<.0001	<.0001	<.0001	0.9536	<.0001	0.0003	<.0001	<.0001		<.0001	0.0071	<.0001	0.0008	<.0001
12	<.0001	1.0000	0.9988	1.0000	0.9826	1.0000	1.0000	<.0001	0.9715	1.0000	<.0001		0.9999	0.9994	0.9977	<.0001
13	1.0000	1.0000	0.9934	1.0000	0.9854	1.0000	1.0000	1.0000	0.9843	0.9982	0.0071	0.9999		0.9972	1.0000	1.0000
14	0.0003	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	<.0001	1.0000	1.0000	<.0001	0.9994	0.9972		0.9632	0.0303
15	1.0000	1.0000	0.9448	1.0000	0.8891	1.0000	0.9996	0.9988	0.8783	0.9882	0.0008	0.9977	1.0000	0.9632		1.0000
16	0.2800	1.0000	0.1923	0.9968	0.0397	1.0000	0.6854	<.0001	0.0188	0.8333	<.0001	<.0001	1.0000	0.0303	1.0000	

Table 10
Significance of each level of the variable country

In table 9 are reported all the levels of the variable *country* which can then be used in table 10 to distinguish for which levels, or country, there is a statistically difference of the mean of the moment magnitude *mw*. The mean of *mw* is statistically different for all countries except those with index equal to 6 and 10, corresponding to Egypt and Iraq.

Another interesting aspect to analyse is to create a regression model with *richter* as response and all the other variables as explanatory variables. *id* and *xm* will be not included because *id* represents only the identification number of each earthquake and *xm* would be redundant because it is created based on the values of the other variables. An initial general model considering the remaining variables has been created.

The GLM Procedure

Dependent Variable: richter

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	15	408.744867	27.249658	49.19	<.0001
Error	1715	950.090569	0.553989		
Corrected Total	1730	1358.835436			

Table 11
ANOVA table of the general model

The GLM Procedure

Dependent Variable: richter

R-Square	Coeff Var	Root MSE	richter Mean
0.300805	16.86861	0.744304	4.412363

Table 12
Statistics of the general model

The GLM Procedure

Dependent Variable: richter

Parameter	Estimate		Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	0.6392065950	B	0.59552870	1.07	0.2833	-.5288325515	1.8072457415
lat	0.0410839185		0.01478184	2.78	0.0055	0.0120915821	0.0700762548
long	0.0084083073		0.00329634	2.55	0.0108	0.0019430332	0.0148735815
dist	0.0061231325		0.00436529	1.40	0.1609	-.0024387244	0.0146849893
depth	-.0009763068		0.00099498	-0.98	0.3266	-.0029278165	0.0009752029
md	-.2353735517		0.02633044	-8.94	<.0001	-.2870167069	-.1837303966
mw	0.3667320564		0.02580654	14.21	<.0001	0.3161164394	0.4173476735
ms	0.3494568225		0.04078559	8.57	<.0001	0.2694620719	0.4294515731
mb	-.0601678246		0.03653303	-1.65	0.0998	-.1318218126	0.0114861635
country turkey	0.0000000000	B	.	.	.	.	.
direction east	0.0536124823	B	0.11301866	0.47	0.6353	-.1680564651	0.2752814296
direction north	0.0970145179	B	0.11242837	0.86	0.3883	-.1234966711	0.3175257070
direction north_east	0.0595560862	B	0.09420740	0.63	0.5274	-.1252174251	0.2443295975
direction north_west	-.0219753739	B	0.09254105	-0.24	0.8123	-.2034805951	0.1595298473
direction south	0.0901786826	B	0.11235817	0.80	0.4223	-.1301948096	0.3105521748
direction south_east	0.0586206347	B	0.09396674	0.62	0.5328	-.1256808631	0.2429221324
direction south_west	0.0809034186	B	0.09260258	0.87	0.3824	-.1007224852	0.2625293224
direction west	0.0000000000	B	.	.	.	.	.

Table 13
Parameters estimates of the general model

From the p-value of table 11 it is possible to deduce that this model is statistically significant, even though it only explains 30% of the variance of the response as indicated by the R-square value of table 12. Table 13 reports the parameters estimates and it is possible to notice from the p-values that not all the variables are statistically significant for this model. Figure 13 is the diagnostic plots which show that the residuals are sufficiently normally distributed, they have their mean around zero and their variance seems to be constant. However, some strange patterns are distinguishable in more than one plot and they should be further investigated but for time reason it will not be possible.

The GLM Procedure

Dependent Variable: richter

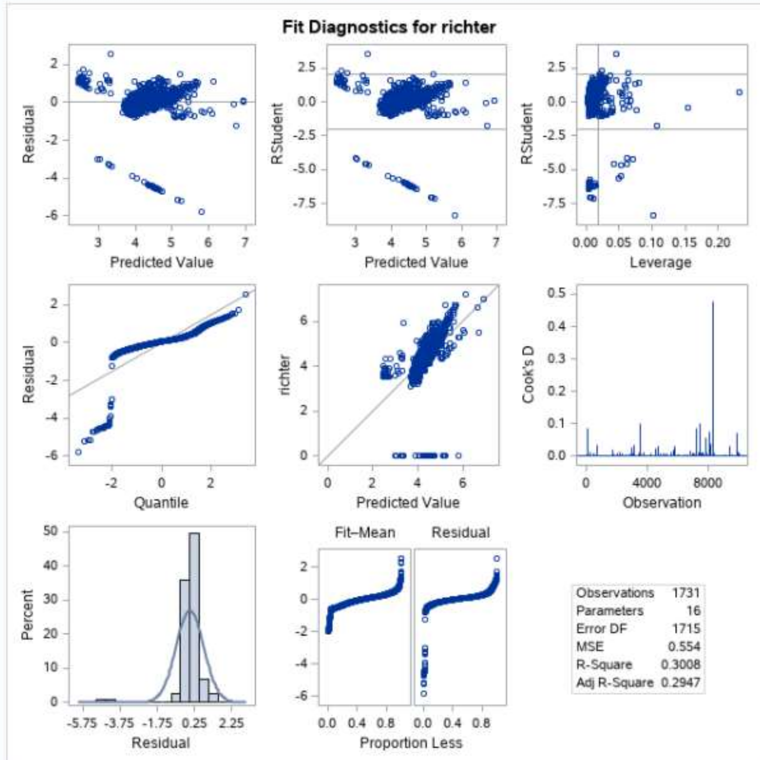

Figure 13
Diagnostic plots

In order to investigate which variable could be irrelevant in this model, the Hoeffding' s D and Spearman statistics technique will be applied. This technique calculates two indices and, if their values are different, it could indicate that the variables are not in a linear relationship and therefore, a deeper investigation is needed. More precisely, this technique will calculate the Spearman rank and Hoeffding rank and if the first has a low value and the second has a high value, this indicates that a further analysis is necessary.

Obs	Variable	ranksp	rankho	scoef	spvalue	hcoef	hpvalue
1	depth	5	4	0.23378	<.0001	0.01631	<.0001
2	dist	1	1	0.00121	0.9035	-0.00006	0.9989
3	lat	2	2	-0.03722	<.0001	0.00102	<.0001
4	long	4	3	-0.18753	<.0001	0.01296	<.0001
5	mb	6	7	0.37118	<.0001	0.04495	<.0001
6	md	3	5	-0.03929	<.0001	0.02910	<.0001
7	ms	7	6	0.51374	<.0001	0.03239	<.0001
8	mw	8	8	0.89103	<.0001	0.53706	<.0001

Table 14
Spearman and Hoeffding ranks

From table 14 and figure 14, it is possible to notice that only *md* has relatively low Spearman rank and a relatively high Hoeffding rank. That might indicate that *md* is not in a linear relationship with *richter* or that *md* is not significant for this model.

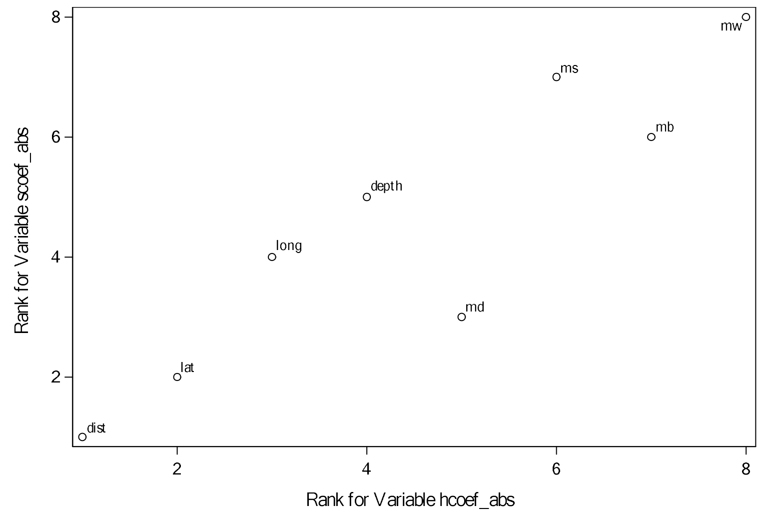

Figure 14
Spearman and Hoeffding ranks

A model selection technique using a forward selection and a significance level of 0.05 will be applied, in order to investigate if the variable *md* is really not statistically significant for this model.

Forward Selection Summary

Step	Effect Entered	Number Effects In	Number Parms In	F Value	Pr > F
0	Intercept	1	1	0.00	1.0000
1	mw	2	2	536.36	<.0001
2	ms	3	3	38.09	<.0001
3	md	4	4	87.18	<.0001
4	lat	5	5	9.38	0.0022
5	long	6	6	8.12	0.0044

Selection stopped as the candidate for entry has SLE > 0.05.

Stop Details

Candidate For	Effect	Candidate Significance		Compare Significance	
Entry	mb	0.0915	>	0.0500	(SLE)

Table 15
Variable selection

Table 15 shows that the variable *mb* should not be included in the model, whereas *md* seems to be significant. This is confirmed from the plots in figures 15 – 17. Figure 15 shows that the coefficients do not have a fluctuating behaviour when entering the model and the p-value maintains a value underneath 0.05 if the model is composed only of the variables *mw, ms, md, lat* and *long.* This is also confirmed by the plots of the fit criteria of figure 16, where AIC, AICC and SBC have their lowest values when the model has only the variables afore mentioned, and the adjust R-square is maximised. Furthermore, the average square error has its lowest value as shown in figure 17. From table 16 it is also possible to understand that this model is statistically significant as indicated by the low p-value, even though about 30% of the variation is explained by this model as the value of R-square indicates in the second table of table 16, which it also reports the values of all the fit criteria. In the last table of table 16 there are the parameters estimates and it is possible to notice, from the values of the p-value, that all the variables considered are statistically significant.

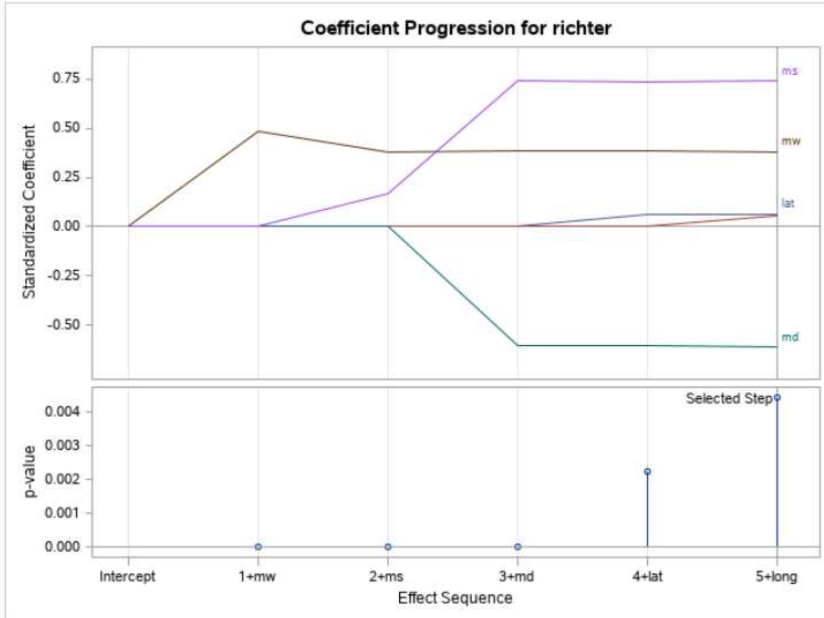

Figure 15
Coefficients at each iteration

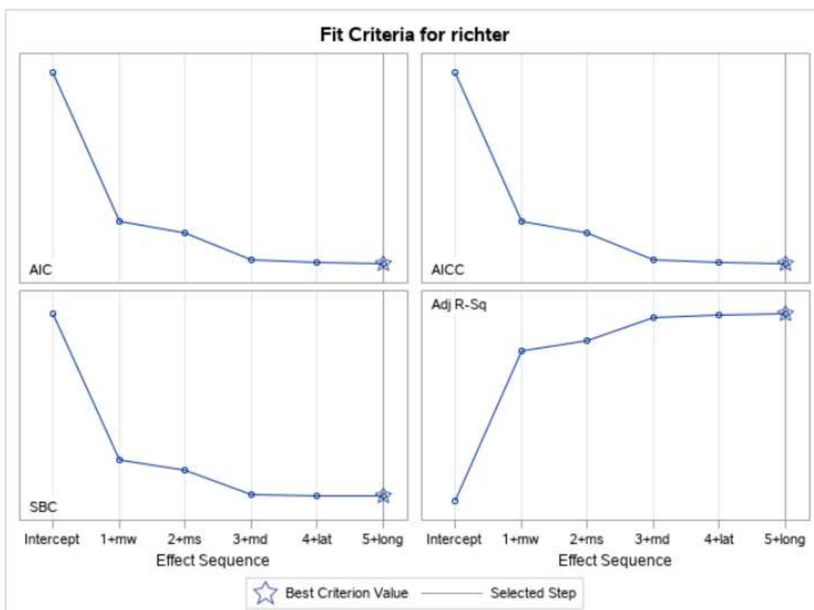

Figure 16
Fit criteria

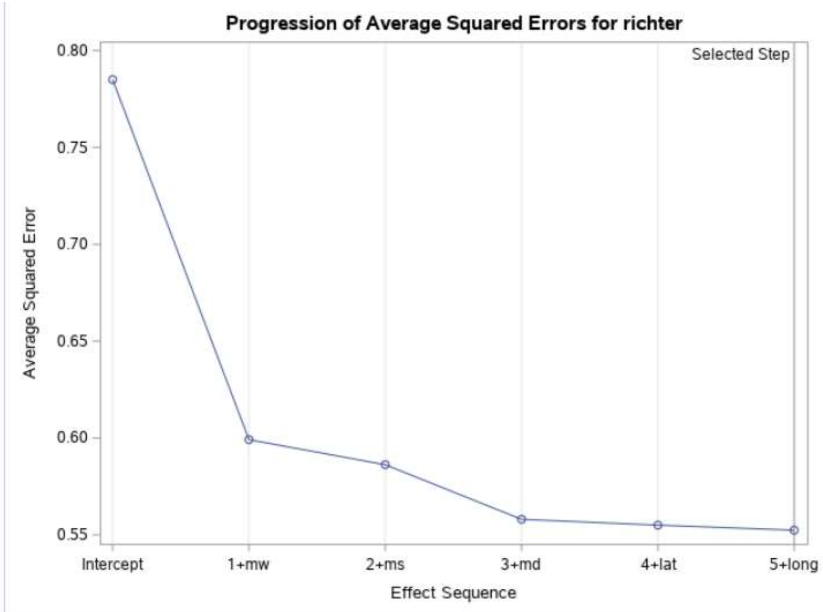

Figure 17
Average Square Error

Selected Model

Effects:	Intercept lat long md mw ms

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	402.57662	80.51532	145.24	<.0001
Error	1725	956.25881	0.55435		
Corrected Total	1730	1358.83544			

Root MSE	0.74455
Dependent Mean	4.41236
R-Square	0.2963
Adj R-Sq	0.2942
AIC	717.77967
AICC	717.84468
SBC	-982.48160

Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.587929	0.574883	1.02	0.3066
lat	1	0.042659	0.014445	2.95	0.0032
long	1	0.009208	0.003232	2.85	0.0044
md	1	-0.244064	0.025751	-9.48	<.0001
mw	1	0.369401	0.025703	14.37	<.0001
ms	1	0.295062	0.026465	11.15	<.0001

Table 16
Selected model

The final model can be written as follow:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon$$

Where y is the response variable in this case *richter*, $\beta_0 - \beta_5$ are the estimate values of table 16 and $X_1 - X_5$ are the parameters. ε represents the random part of the model that cannot be explained.

An earthquake with a magnitude of 5 or above can cause damage to infrastructures and people, with possible casualties. Therefore, a new variable called *serious* will be created which it has a value of 1 if *richter* is equal or greater than 5 and 0 if *richter* is less than 5. A new model will be created with *serious* as response variable and all the other as explanatory variables, excluding *id, mw, richter* and *xm.* However, before fitting the logistic regression model, the dataset will be split into *training* dataset, which is composed of the 70% of all data and will be used to train the model, and *test* dataset, composed of the remaining 30% of the data and used to make predictions. Then, by comparing the predicted values with the observed ones, it will be possible to establish how good this model can predict if an earthquake will be able to cause serious damages to infrastructures and people.

The LOGISTIC Procedure

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Table 17
Convergence status

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.5980	2.3661	0.4561	0.4994
lat		1	-0.1124	0.0618	3.3076	0.0690
long		1	0.0140	0.0141	0.9762	0.3231
dist		1	-0.00678	0.0182	0.1381	0.7101
depth		1	-0.0120	0.00425	7.9923	0.0047
md		1	-0.0413	0.0456	0.8198	0.3652
ms		1	0.7617	0.0739	106.2396	<.0001
mb		1	0.2622	0.0843	9.6770	0.0019
direction	east	1	-0.1702	0.3116	0.2983	0.5850
direction	north	1	0.2275	0.2750	0.6844	0.4081
direction	north_east	1	-0.2344	0.1828	1.6454	0.1996
direction	north_west	1	-0.2068	0.1756	1.3868	0.2389
direction	south	1	0.1669	0.2739	0.3715	0.5422
direction	south_east	1	-0.0193	0.1712	0.0127	0.9104
direction	south_west	1	-0.0708	0.1678	0.1781	0.6730

Table 18
Parameters estimates

Table 17 shows that the model has converged. Only the two variables *ms* and *mb* seems to be significant to explain the variable *serious* as shown from the p-values of table 18 which reports also the parameters estimates and the standard error. This is also confirmed by the plot of figure 18 which reports the confidence interval of the odds ratio and it is possible to notice that the only two variables not crossing the value of 1 are *ms* and *mb.* The value of 1 is chosen because the logistic regression uses the logit transformation which corresponds to the logarithmic of the odds express by the formula:

$$log\left(\frac{\theta_i}{1-\theta_i}\right) = \beta_0 + \beta_1 x_1 + \cdots \beta_p x_p$$

The LOGISTIC Procedure

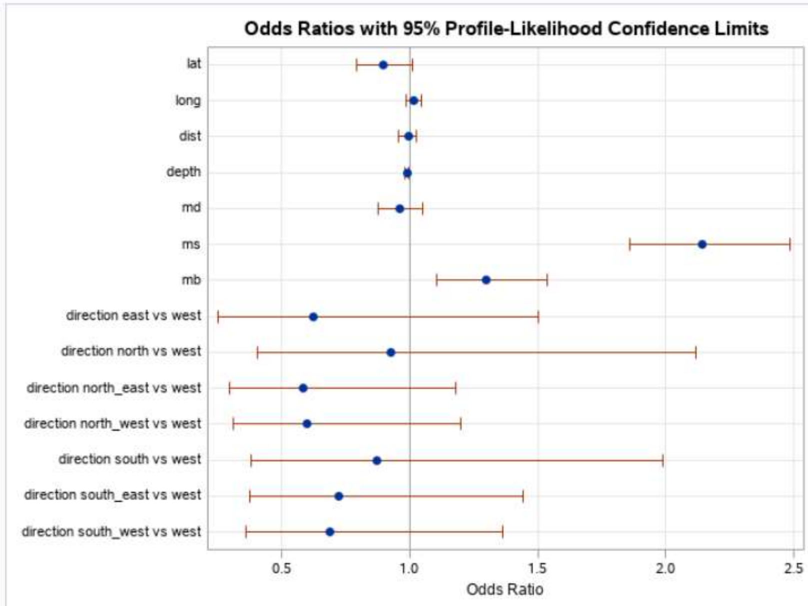

Figure 18
95% Confidence Interval of the ods ratios

The logistic regression model above discussed will be compared with a second logistic regression which it has *serious* as response variable and *xm* as the only explanatory variable.

The LOGISTIC Procedure

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Table 19
Convergence status

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-33.8989	1.1473	872.9378	<.0001
xm	1	6.3250	0.2199	827.4074	<.0001

Table 20
Parameters estimates

Table 19 shows that also this model has converged and the p-value of table 20 shows that the variable *xm* is statistically significant for this model. This is also confirmed by the 95% confidence interval of the odds ratio reported in figure 19. As before the interval does not cross the value of 1.

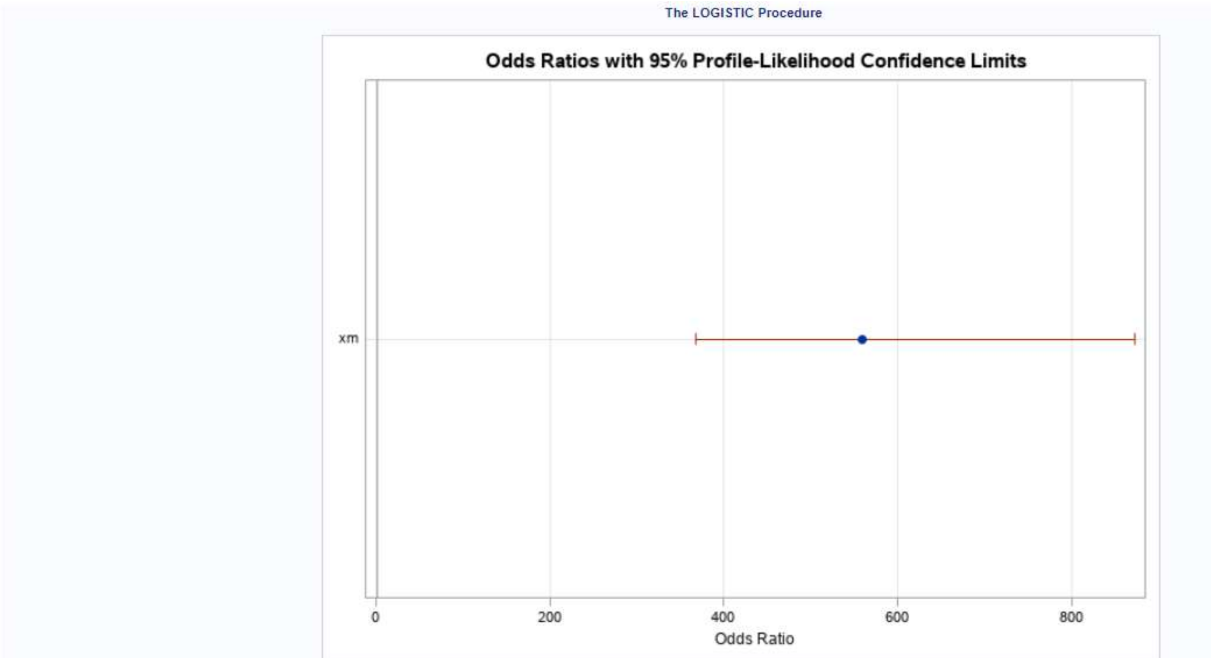

Figure 19
95% Confidence Interval of the ods ratios

The two logistic models described above have been trained using the *training* dataset. They have then been used to make predictions using the *test* dataset. In order to determine if the two models can predict whether an earthquake is *serious* or not, plots of the ROC curves are created and showed in figure 20. The ROC (Receiver Operating Characteristic) curve is a measure of the performance of a model in classifying observations, in this case as a *serious* earthquake or not. On the horizontal axis there is the rate of the observations wrongly classified as *serious* by the model and on the vertical axis there is the rate of the observations correctly classified as *serious*. The diagonal line represents the random guess, i. e. the model correctly classified as *serious* only 50% of the data. The value of AUC is the Area Under the Curve and in this case would be equal to 0.5. A model to perform better than a random guess has to have a value of AUC greater than 0.5.

The two models crated in this report have a high value of AUC: 0.8526 for the complex model, which is the first one created and that is composed of *serious* as response variable and all the other variables as explanatory variables, excluding *id, mw, richter* and *xm*; and 0.9928 for the simple one, which is composed of *serious* as response variable and only *xm* as response variable. Therefore, from the graph of figure 20 it possible to deduce that the second model is a better classifier than the more complex one.

Table 21 is a Chi-squared test with the null hypothesis that the two ROC curves are the same. From the value of the p-value it is possible to reject the null hypothesis, therefore, the two ROC curves are different.

The LOGISTIC Procedure

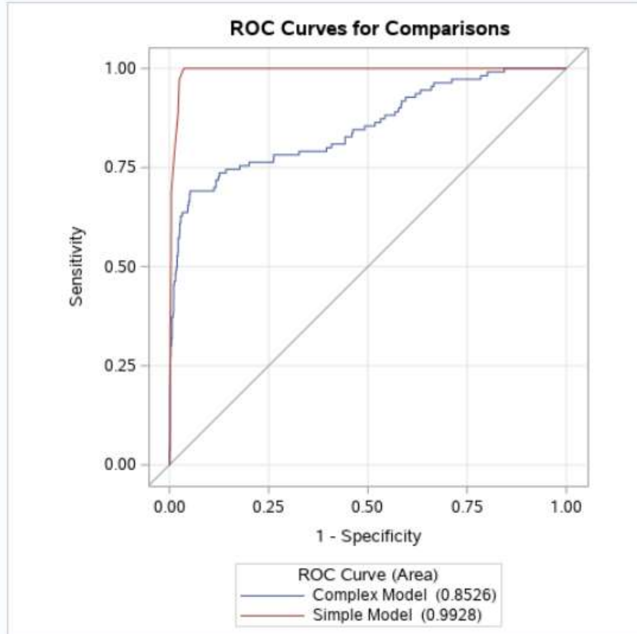

Figure 20
ROC curve and AUC value for the two logistic models

The LOGISTIC Procedure

ROC Contrast Test Results			
Contrast	DF	Chi-Square	Pr > ChiSq
Comparing two Models	1	38.7938	<.0001

Table 21
Chi-squared test for the two ROC curves

CONCLUSIONS

Earthquakes are natural phenomenon caused by geological structures called *faults*. They are the products of compressive or extensive forces caused by the movement of the tectonic plates. The dataset analysed in this report is a collection of values used by seismologists to describe a telluric event. To determine the magnitude or strength of an earthquake it is taken the highest value among all of those calculated by the seismologists. From the analysis described in this report has been determined that the magnitude determined in the way afore described is statistically different from the value of 4.1. Furthermore, it has been observed that one of the values calculated by the seismologist called moment magnitude (*mw*), is statistically different for all of the countries taken into consideration in this report except for Egypt and Iraq.

Fitting a linear regression model has determined that the variables *mw, ms, md, lat* and *long* can well explain the variance of the variable *richter*.

An earthquake with a magnitude higher than 5 for the *richter* scale is considered serious because can cause damage to infrastructures and people with possible victims. Therefore, the dataset has been augmented with a new variable called *serious* and which has value 1 if *richter* is equal or greater than 5 and 0 if *richter* is less than 5. This new variable has then been used as response variable to create two logistic models: the first one, more complex, which considers all the other variables of the dataset excluding *id, mw, richter* and *xm*; and a second one more simple which considers only *xm* as explanatory variable. Both models have been trained with the *train* dataset and then they have been used to make prediction using the *test* dataset. Comparing the performance of these two models using plots of the ROC curve and taking into consideration the value of AUC, has been observed that the simple one is a better predictor of an earthquake being serios then the more complex one, although both models are good predictors. They respectively have a value of AUC of 0.9928 and 0.8526.