Lab report 2020

UNDERSTANDING EARTHQUAKE DATA

Introduction

Earthquakes are responsible for large scale destruction and loss of life every year. Understanding them better gives us an opportunity to anticipate the impact of future earthquakes and put in preventative measures in high risk areas to alleviate potential damage.

This lab report aims to explore data from earthquakes over a number of years to help gain that understanding. The data contains information on 23,741 independently observed earthquakes. This is available with the below fields:

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

To help our understanding, this lab report explores the following challenges:

a) Taking the largest magnitude value from each earthquake, is the average value different from 4.1?
b) Is there a difference between the average moment magnitude scale value and the country the earthquake occurred in?
c) Using a regression model to predict richter, how does the model perform and what variables have the biggest impact?
d) Using a regression model to predict whether an earthquake is serious or not (defined as having a richter value of 5 or more), how does the model perform and what variables have the biggest impact?
e) Comparing the above model to one that only uses the largest magnitude value, which is most likely to extrapolate best on to new data?

The lab report will start by exploring the available data to get an understanding of the underlying structure and patterns, as well as to help identify anything that may skew our view of the data. Following the exploratory analysis, a more formal analysis will be conducted that will use statistical

modelling techniques to answer each question. Finally, the report will conclude by outlining the main findings from the formal analysis and reflecting on the approach taken.

Exploratory analysis

Understanding numeric fields

Checking for missing values

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
id	23741	11871.00	6853.58	1.0000000	23741.00
lat	23741	37.9521937	2.1944648	29.7400000	46.3500000
long	23741	30.7065322	6.5638114	18.3400000	48.0000000
dist	10062	3.1750149	4.7154610	0.1000000	95.4000000
depth	23741	18.4424076	23.2267930	0	225.0000000
md	23741	1.9076071	2.0593288	0	7.4000000
richter	23741	2.2003875	2.0805645	0	7.2000000
mw	4950	4.4775758	1.0487482	0	7.7000000
ms	23741	0.6789478	1.6764715	0	7.9000000
mb	23741	1.6954888	2.1466149	0	7.1000000

The table shows a count of populated numeric fields ('N') and some summary information for each field.

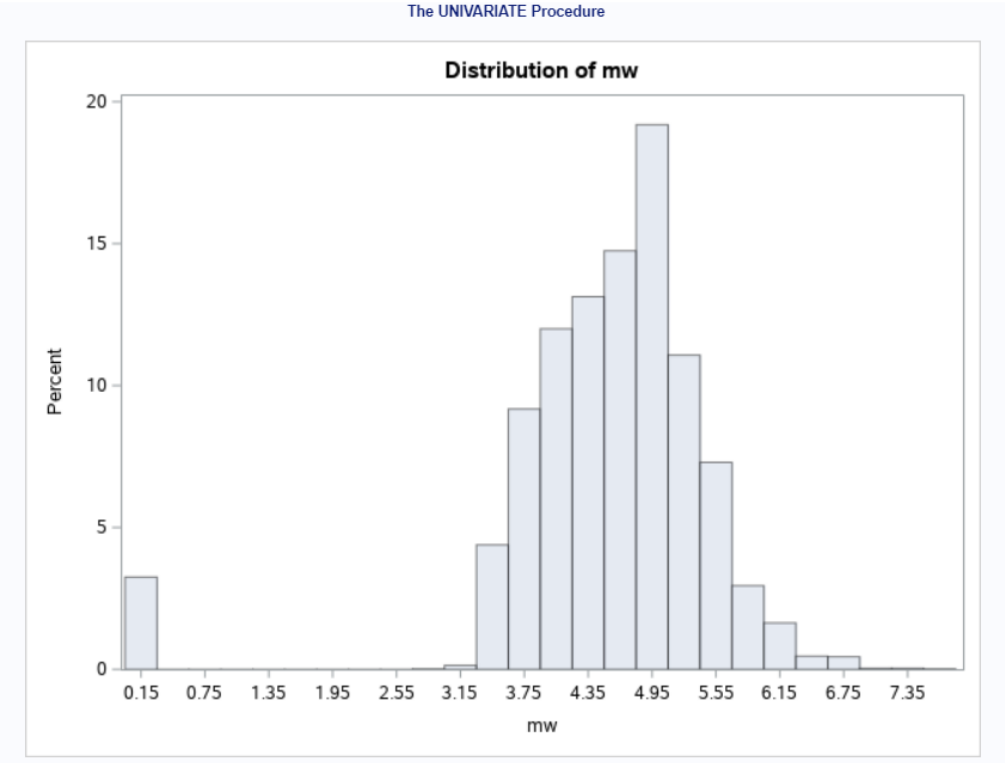
The count of records show that the 'dist' and 'mw' fields contain a large amount of missing values. Many of the numeric value fields have a minimum of 0 which is likely also missing, this is shown below for the fields identified as containing 0 values.

Value of 0s by numeric field

records	depth_0	md_0	richter_0	mw_0	ms_0	mb_0
23741	1497	12548	10968	161	20337	14469

The table shows how many 0 values are present in each field as well as an overall count of records in the table for comparison ('records').

All of the fields except for 'mw' have a high volume of 0's which suggests they are missing values. As 'mw' only contains 161, and already contains missing values, these may be accurate missing values, to test this, the below histogram shows how these compare to other values:



The histogram above plots the distribution of the 'mw' variable.

This shows that the 0 values sit away from the rest of the distribution and are therefore likely missing values. These will be corrected as missing for future analysis in addition to the other fields identifying as having missing values recorded as 0 ('depth', 'md', 'richter', 'ms', 'mb').

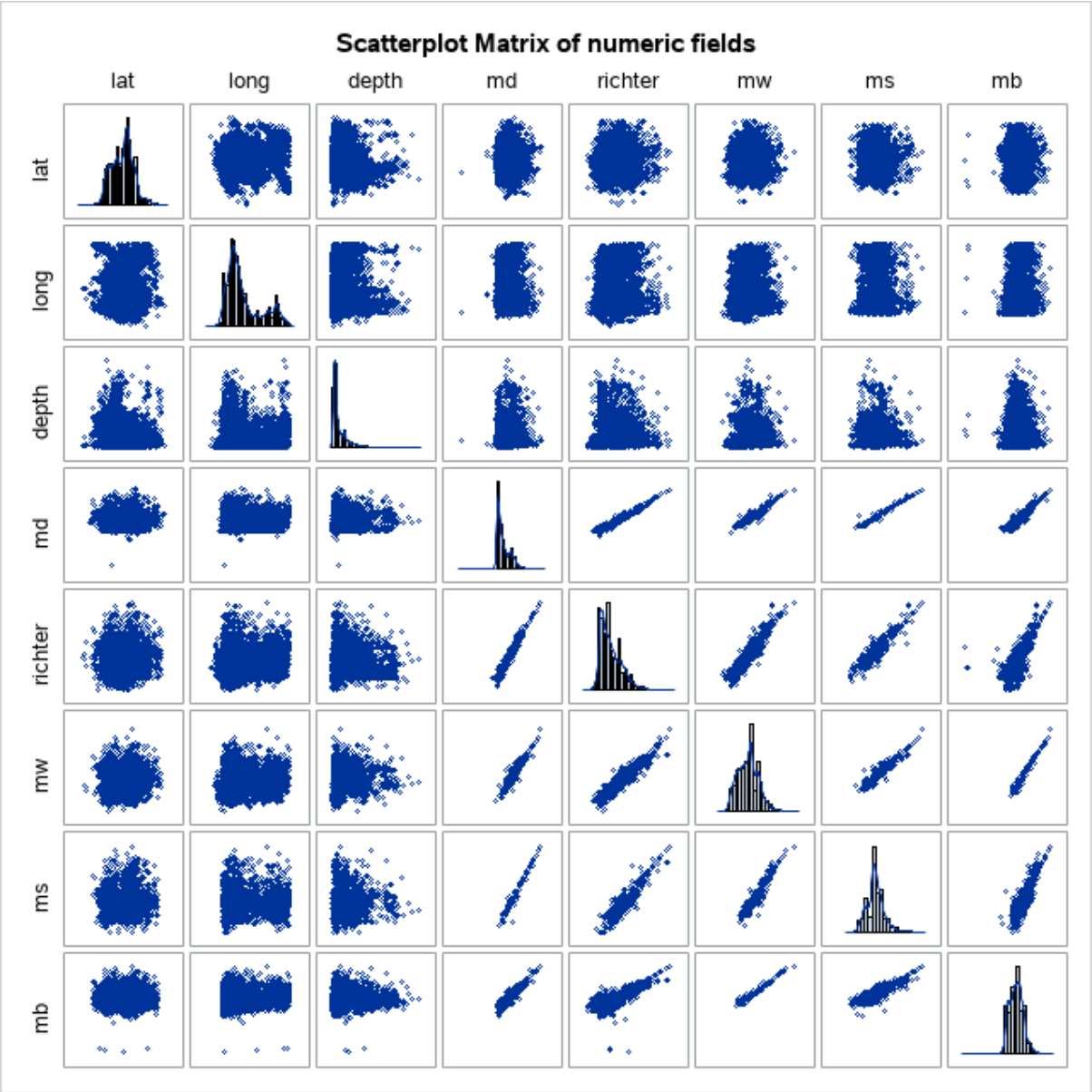
Understanding the scale, distribution and correlation of numeric values, and checking if they contain outliers

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
id	23741	11871.00	6853.58	1.0000000	23741.00
lat	23741	37.9521937	2.1944648	29.7400000	46.3500000
long	23741	30.7065322	6.5638114	18.3400000	48.0000000
dist	10062	3.1750149	4.7154610	0.1000000	95.4000000
depth	22244	19.6835641	23.4810736	0.2000000	225.0000000
md	11193	4.0461449	0.5847001	0.3000000	7.4000000
richter	12773	4.0898301	0.5638908	2.9000000	7.2000000
mw	4789	4.6281061	0.6633560	2.9000000	7.7000000
ms	3404	4.7352820	0.6272705	3.0000000	7.9000000
mb	9272	4.3413072	0.5584969	0.2000000	7.1000000

The table above is an updated statistical summary of the numeric fields after setting the 0 values as missing.

This shows that the mean and standard deviation for the main measurement variables ('md', 'mw', 'ms', 'mb and 'richter') are very similar so it is assumed the variables are on a comparable scale.



The visual above plots the distribution of each of the numeric variables and shows a scatterplot for each combination of numeric variable.

The scatterplot matrix demonstrates three things;

1. The majority of variables look to have an approximately normal distribution with the some right skew observed for value metrics – particularly 'md' and 'richter'
2. The value metrics ('md', 'richter', 'mw', 'ms', 'mb') appear to have a strong linear positive correlation
3. There are some extreme low values that may be erroneous, particularly for the 'mb' field

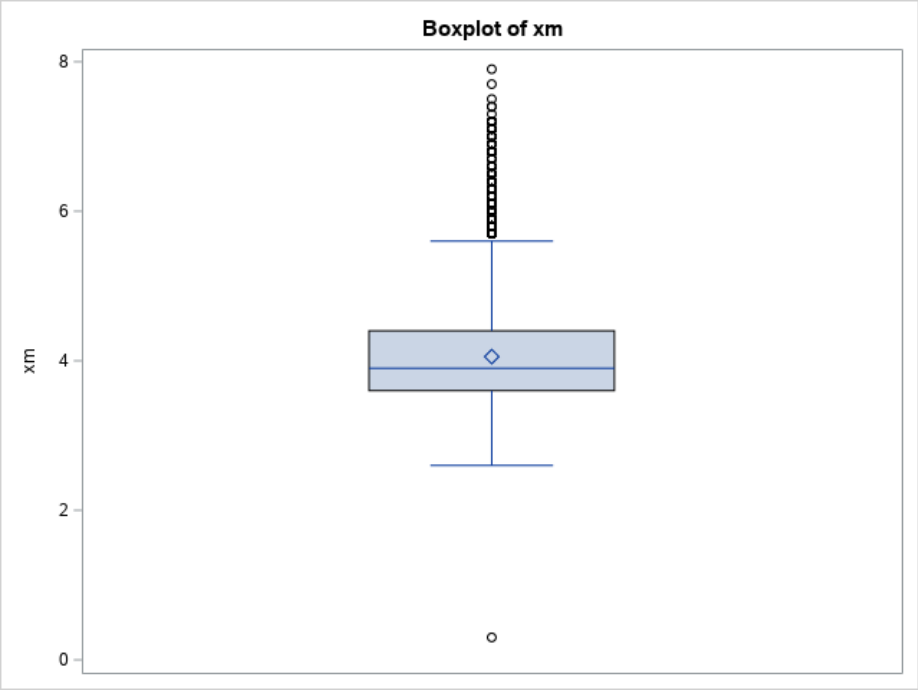
Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations								
	lat	**long**	**depth**	**md**	**richter**	**mw**	**ms**	**mb**
lat	1.00000 23741	0.23749 <.0001 23741	-0.25300 <.0001 22244	0.00694 0.4626 11193	0.03326 0.0002 12773	0.06707 <.0001 4789	-0.00498 0.7716 3404	-0.02784 0.0073 9272
long	0.23749 <.0001 23741	1.00000 23741	-0.08194 <.0001 22244	-0.01104 0.2429 11193	0.10636 <.0001 12773	0.09949 <.0001 4789	0.01875 0.2742 3404	0.20219 <.0001 9272
depth	-0.25300 <.0001 22244	-0.08194 <.0001 22244	1.00000 22244	0.36008 <.0001 10670	0.22948 <.0001 12368	0.23797 <.0001 4789	-0.00540 0.7530 3397	0.15450 <.0001 8572
md	0.00694 0.4626 11193	-0.01104 0.2429 11193	0.36008 <.0001 10670	1.00000 11193	0.98466 <.0001 3522	0.97836 <.0001 3020	0.99578 <.0001 2997	0.96802 <.0001 3307
richter	0.03326 0.0002 12773	0.10636 <.0001 12773	0.22948 <.0001 12368	0.98466 <.0001 3522	1.00000 12773	0.96019 <.0001 4669	0.96992 <.0001 3084	0.87520 <.0001 5666
mw	0.06707 <.0001 4789	0.09949 <.0001 4789	0.23797 <.0001 4789	0.97836 <.0001 3020	0.96019 <.0001 4669	1.00000 4789	0.98044 <.0001 3021	0.97846 <.0001 3059
ms	-0.00498 0.7716 3404	0.01875 0.2742 3404	-0.00540 0.7530 3397	0.99578 <.0001 2997	0.96992 <.0001 3084	0.98044 <.0001 3021	1.00000 3404	0.94233 <.0001 3382
mb	-0.02784 0.0073 9272	0.20219 <.0001 9272	0.15450 <.0001 8572	0.96802 <.0001 3307	0.87520 <.0001 5666	0.97846 <.0001 3059	0.94233 <.0001 3382	1.00000 9272

The above correlation matrix shows the strength of the correlation (using Pearson Correlation Coefficients) and whether each correlation is statistically significant.

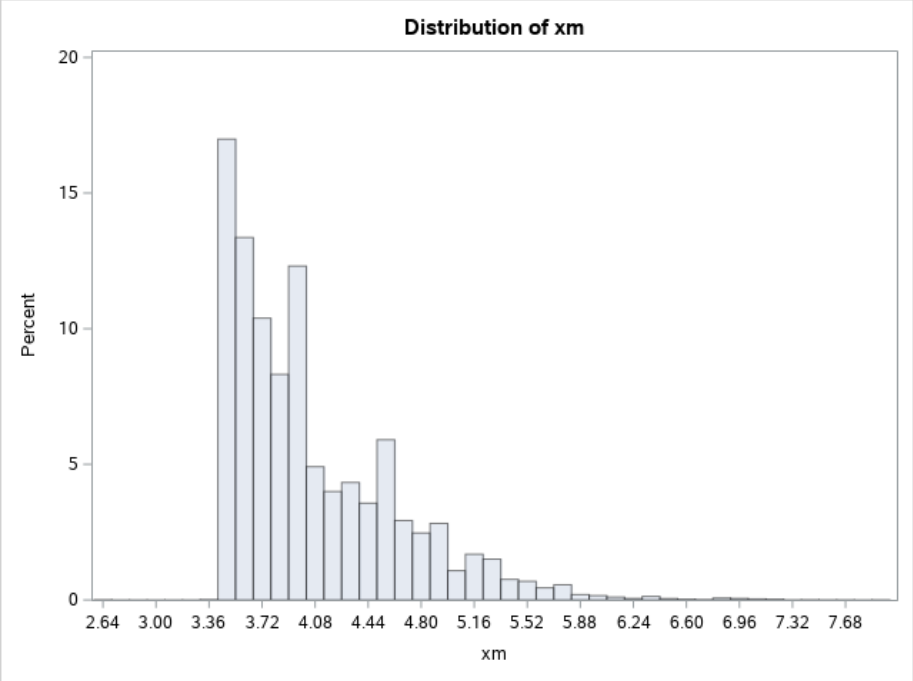
From this we can observe:

- The correlation between ‘md’, ‘richter’, ‘mw’, ‘ms’ and ‘mb’ are very strong and including more than one variable risks skewing any statistical model
- Latitude, longitude and depth all have a low (but significant) correlation
- Other correlations can be observed between location fields (latitude, longitude and depth) and value fields but these are small and not consistently significant
    - This will be further explored as part of the modelling process

As the exploratory analysis has shown that all of the value metrics have a strong positive correlation, a new metric is developed that takes the largest magnitude value – named as ‘xm’.



The boxplot of ‘xm’ shows one extreme low value which is removed from further analysis by removing any values less than 1 (this removed one value only).



The above chart shows a histogram of the newly formed ‘xm’ variable.

The distribution has a positive (right) skew. Further transformation of this may be required if this is found to affect the statistical models.

Understanding character fields

Checking for missing values

The FREQ Procedure

country	Frequency	Percent	Cumulative Frequency	Cumulative Percent
turkey	11850	49.91	11850	49.91
mediterranean	4843	20.40	16693	70.31
greece	3560	15.00	20253	85.31
aegean_sea	1748	7.36	22001	92.67
iran	346	1.46	22347	94.13
georgia	322	1.36	22669	95.48
russia	303	1.28	22972	96.76
bulgaria	176	0.74	23148	97.50
syria	154	0.65	23302	98.15
azerbaijan	150	0.63	23452	98.78
iraq	122	0.51	23574	99.30
blacksea	90	0.38	23664	99.68
romania	44	0.19	23708	99.86
macedonia	28	0.12	23736	99.98
albania	2	0.01	23738	99.99
egypt	2	0.01	23740	100.00
israel	1	0.00	23741	100.00

The table shows a count of records associated with each value in the ‘country’ field (including missing values which would be shown in a separate line).

The field shows no evidence of missing values but does show a number of values with low frequency.

To reduce noise, all values with less than 300 records have been grouped up as ‘other’ (this accounts for 3.24% of all records) leaving ‘Turkey’, ‘Mediterranean’, ‘Greece’, ‘Aegan Sea’, ‘Iran’, ‘Georgia’ and ‘Russia’.

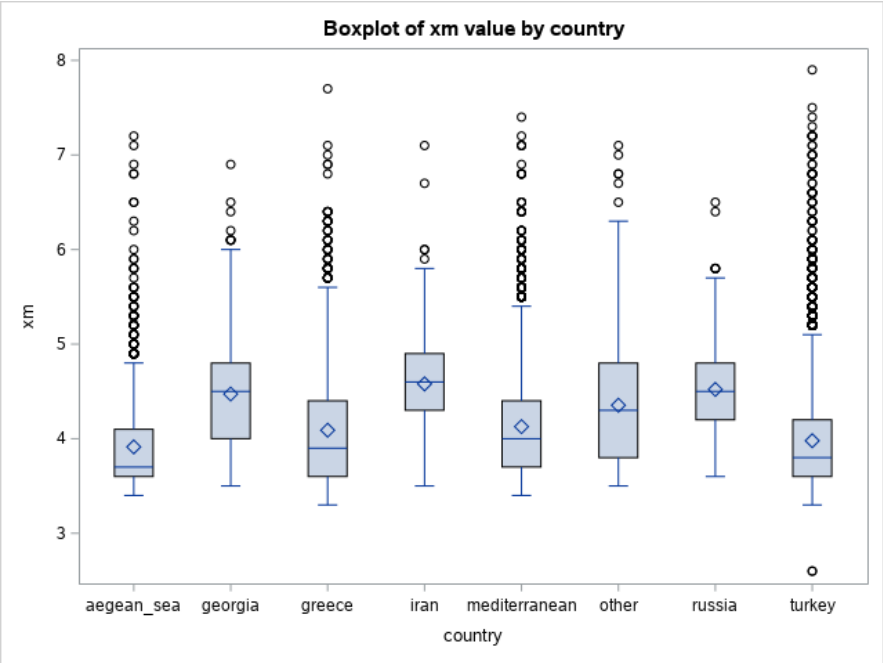
The FREQ Procedure

direction	Frequency	Percent	Cumulative Frequency	Cumulative Percent
	13679	57.62	13679	57.62
north_west	2019	8.50	15698	66.12
south_west	2011	8.47	17709	74.59
south_east	1917	8.07	19626	82.67
north_east	1901	8.01	21527	90.67
south	605	2.55	22132	93.22
north	576	2.43	22708	95.65
east	536	2.26	23244	97.91
west	497	2.09	23741	100.00

The table shows a count of records associated with each value in the ‘direction’ field (including missing values which are shown in a separate line).

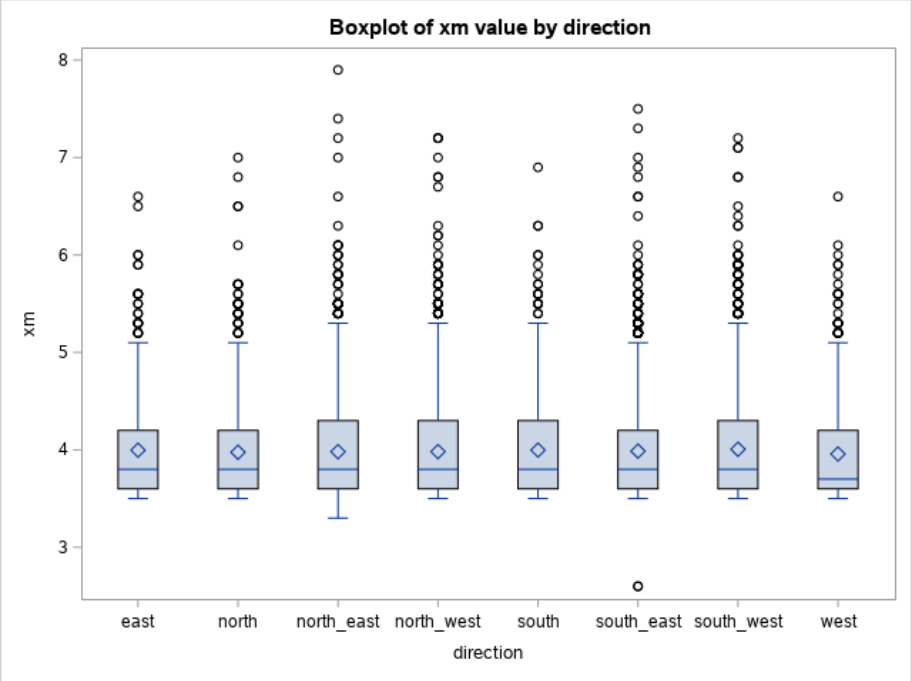
The field contains many missing values that will need to be addressed when creating a statistical model. Values are also much more likely to be two part (i.e. north west instead of north or west) suggesting recording may be skewed.

Understanding whether numeric values change significantly based on character values



The above chart shows the distribution of the 'xm' magnitude value split by country.

The chart shows some differences in the distribution of xm between the countries. This pattern is worth keeping in mind as we look to group values within the formal analysis.



The above chart shows the distribution of the 'xm' magnitude value split by direction.

No significant difference is observed across directions indicating that this variable is unlikely to add significant information to the statistical models.

Summarising Exploratory Data analysis findings and actions taken

- The data contained missing values across all variables except for id, latitude and longitude
    - For the metric values, these were sometimes recorded as 0
        - The 0 values were set as missing so as not to skew analysis
    - Many of the variables contain a large amount of missing variables and will have limited scope for modelling
- The magnitude metric values ('md', 'richter', 'mw', 'ms', 'mb') had a comparable mean and standard deviation so are assumed to be on a similar scale
- There is a strong positive correlation between all of the magnitude metric values
    - Smaller correlations were observed with location fields (latitude, longitude and depth)
- While the majority of variables had an approximately normal distribution, a positive (right) skew was observed for some of the value metrics which may impact our statistical models
- A new 'xm' field was derived from the maximum value across all magnitude measurements for each record
    - Plotting the distribution of this identified one record with a low magnitude value (less than 1) which was removed to not skew subsequent analysis
- Some countries with low frequency (defined as less than 300) have been grouped as 'other' to reduce noise with the data
- The distribution of the 'xm' value different by country but no significant difference was observed by direction

Formal analysis

Taking the largest magnitude value from each earthquake, is the average value different from 4.1?

One Sample T-test. Is the largest magnitude value mean different to 4.1?

The TTEST Procedure

Variable: xm

N	Mean	Std Dev	Std Err	Minimum	Maximum
23740	4.0554	0.5738	0.00372	2.6000	7.9000

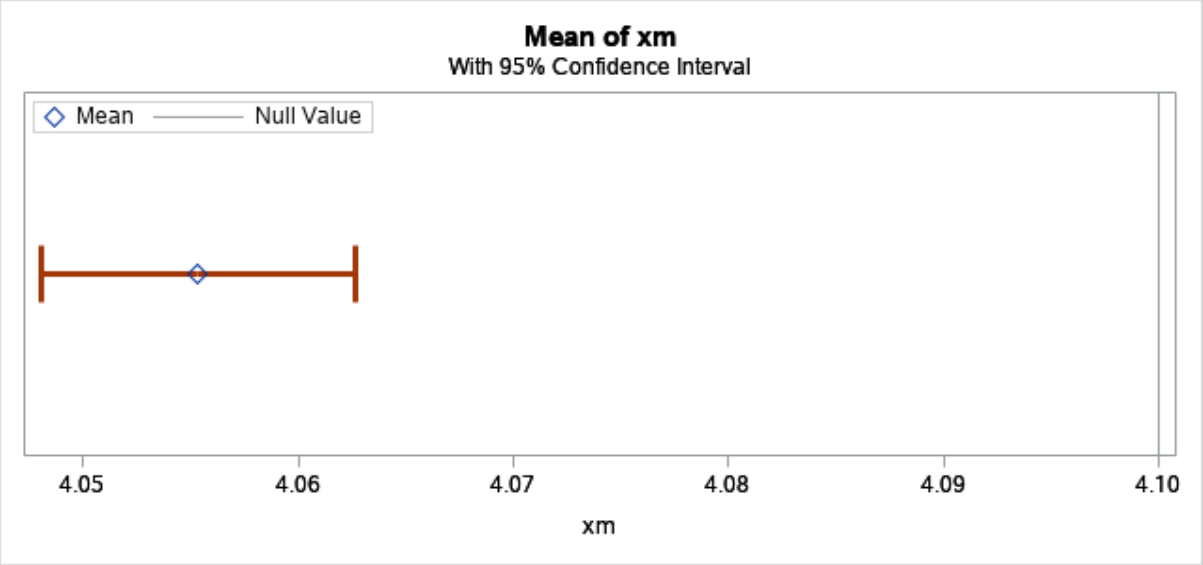
Mean	95% CL Mean		Std Dev	95% CL Std Dev	
4.0554	4.0481	4.0627	0.5738	0.5687	0.5791

DF	t Value	Pr > |t|
23739	-11.99	<.0001

The above tables show the output of a one sample t-test comparing the mean of the 'xm' variable to 4.1.

Based on the results ($t=-11.99$; $p<0.05$), we would infer that the average value of the largest magnitude value ('xm') is significantly different to 4.1 and we would reject the null hypothesis at a 95% confidence level.

The actual value is likely to sit between 4.0481 and 4.0627 as shown in the confidence interval chart below:

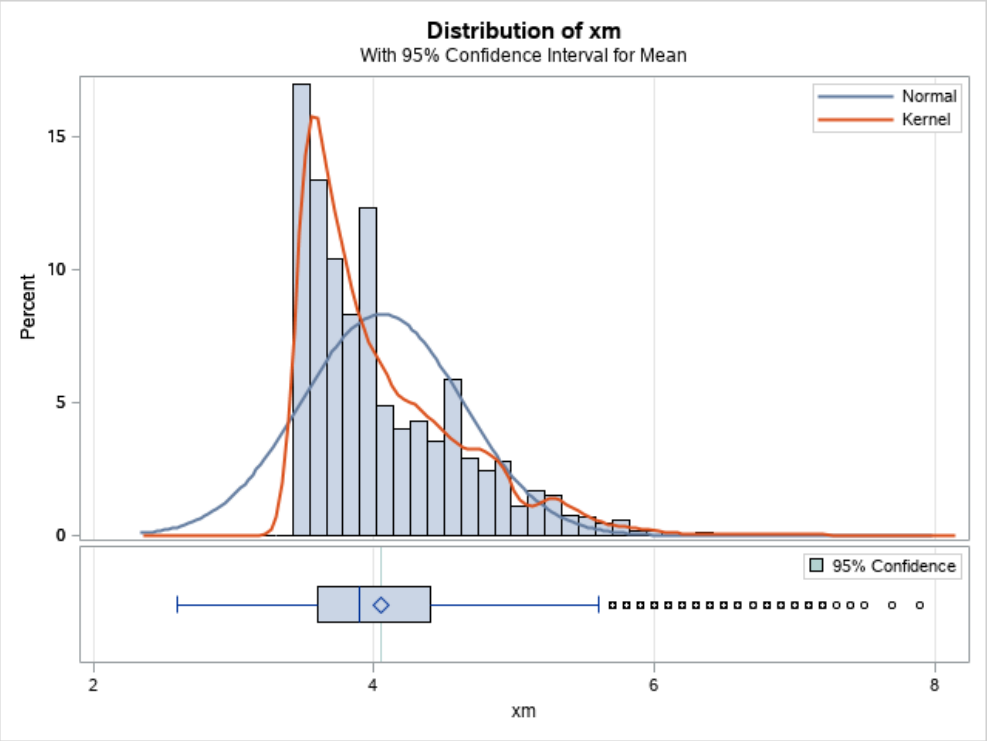


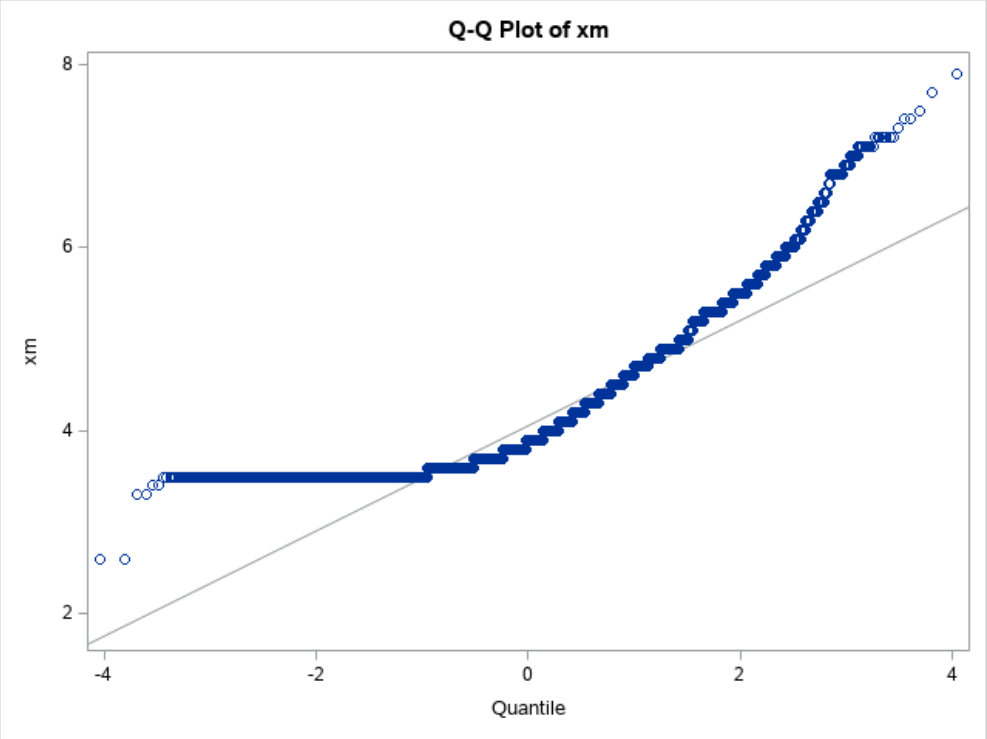
For this test to be valid, two assumptions must be met:

1. The observations are independent
2. The 'xm' value must be Normally distributed

The observations are assumed to be independent though this would be void if some earthquakes were caused by other earthquakes (e.g. aftershocks).

To test if 'xm' can be classed as normally distributed the below diagnostic charts were produced:





As observed with the exploratory analysis, the magnitude value has a right skew and is not strictly normal. The integrity of the test may be improved by repeating the analysis with a non-parametric test.

Is there a difference between the average moment magnitude scale value and the country the earthquake occurred in?

One-Way ANOVA with Country as Explanatory

The GLM Procedure

Dependent Variable: mw

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	76.821679	10.974526	25.85	<.0001
Error	4781	2030.095243	0.424617		
Corrected Total	4788	2106.916922			

R-Square	Coeff Var	Root MSE	mw Mean
0.036462	14.07977	0.651627	4.628106

Source	DF	Type I SS	Mean Square	F Value	Pr > F
country	7	76.82167934	10.97452562	25.85	<.0001

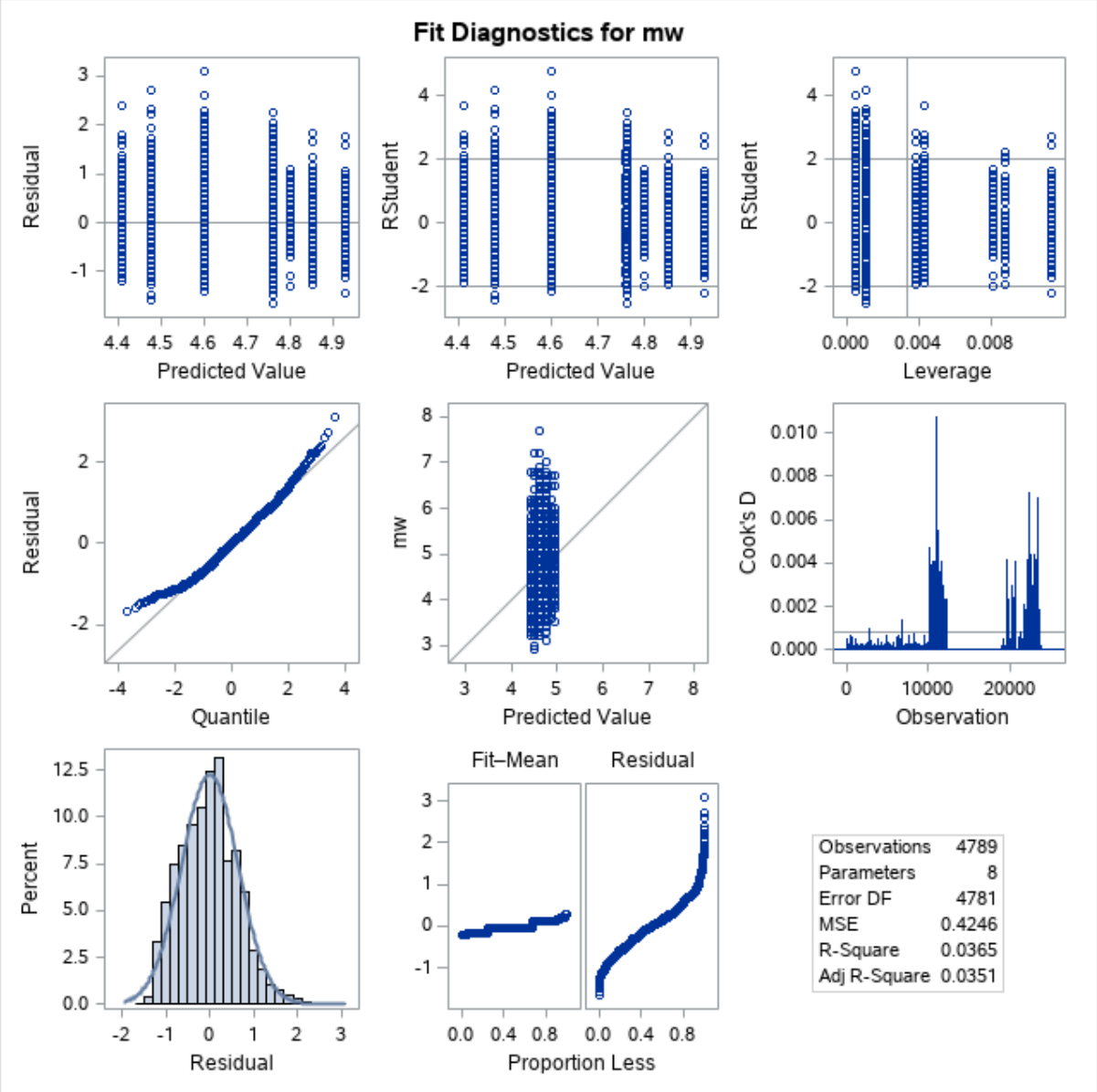
Source	DF	Type III SS	Mean Square	F Value	Pr > F
country	7	76.82167934	10.97452562	25.85	<.0001

The above tables show the results of a one-way ANOVA test looking to see if the 'mw' variable is significantly different by country.

Based on the results (F=25.85; p<0.05), we would infer that there is a difference in the 'mw' value by country and we would reject the null hypothesis at a 95% confidence level.

For the test to be valid, three assumptions must be met:

1. The observations are independent (as per the t-test, it is assumed that they are)
2. The errors are normally distributed
3. The groups have equal variances



The above shows the residual charts for the ANOVA test. Examining the histogram and the QQ plot show that the distribution is approximately normal.

Levene's Test for Homogeneity of mw Variance ANOVA of Squared Deviations from Group Means					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
country	7	12.5101	1.7872	4.90	<.0001
Error	4781	1743.8	0.3647		

The table above shows the output from the Levene's test for unequal variances.

Based on the results (F=4.90; p<0.05), we reject the null hypothesis (that the variances are equal) at a 95% confidence level and infer that the variances are not equal.

As the variances are unequal, the test is repeated using Welch's variance-weighted one-way ANOVA:

Welch's variance-weighted one-way ANOVA with Country as Explanatory

The GLM Procedure

Dependent Variable: mw

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	76.821679	10.974526	25.85	<.0001
Error	4781	2030.095243	0.424617		
Corrected Total	4788	2106.916922			

R-Square	Coeff Var	Root MSE	mw Mean
0.036462	14.07977	0.651627	4.628106

Source	DF	Type I SS	Mean Square	F Value	Pr > F
country	7	76.82167934	10.97452562	25.85	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
country	7	76.82167934	10.97452562	25.85	<.0001

Similar to the previous test, we can reject the null hypothesis at a 95% confidence level and infer that country has a significant impact on 'mw'.

Using a regression model to predict richter, how does the model perform and what variables have the biggest impact?

To tell the full story data of the data and avoid bias, data is excluded where richter is not populated:

How many values are populated for candidate variables?

lat_populated	long_populated	dist_populated	depth_populated	md_populated	mw_populated	ms_populated	mb_populated	country_populated	direction_populated
12773	12773	4074	12368	3522	4669	3084	5666	12773	4074

The table shows a view of how many values are populated in fields when missing richter values are removed.

The variables 'dist', 'md', 'mw', 'ms', 'mb', and 'direction' are not considered for the model for three reasons:

1. The variables contain a large amount of missing values and imputing them risks heavily biasing the model
2. The exploratory analysis showed that the 'md', 'mw', 'ms' and 'mb' variables have a very high correlation with 'richter' as alternate magnitude measures suggesting a large amount of target leakage that, again, would bias the model
3. The 'direction' field showed little difference on the magnitude measures in the exploratory analysis

This leaves 'lat', 'long', 'depth', and 'country' as possible candidates.

The 'depth' field contains some missing values. There is potential to impute these values using the median however, as the volumes are small (and there is a general preference to avoid imputing unless necessary) then these records are just excluded from the model.

The GLM Procedure

Dependent Variable: richter

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	402.391517	40.239152	139.10	<.0001
Error	12357	3574.672827	0.289283		
Corrected Total	12367	3977.064343			

R-Square	Coeff Var	Root MSE	richter Mean
0.101178	13.11288	0.537851	4.101698

The tables above show a summary of the performance of the ANCOVA regression model with all of the candidate exploratory variables included.

The model is shown to be statistically significant at a 95% confidence level ($F = 139.10$, $p < 0.05$). The R-Square value of 0.10 shows that the model only explains a low amount of variation in the model.

Source	DF	Type I SS	Mean Square	F Value	Pr > F
country	7	190.3998676	27.1999811	94.03	<.0001
lat	1	8.6621483	8.6621483	29.94	<.0001
long	1	4.0307701	4.0307701	13.93	0.0002
depth	1	199.2987306	199.2987306	688.94	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
country	7	113.7209840	16.2458549	56.16	<.0001
lat	1	19.1667764	19.1667764	66.26	<.0001
long	1	0.9593705	0.9593705	3.32	0.0686
depth	1	199.2987306	199.2987306	688.94	<.0001

The tables above show the sum of squares and corresponding p-values associated with the introduction of each variable.

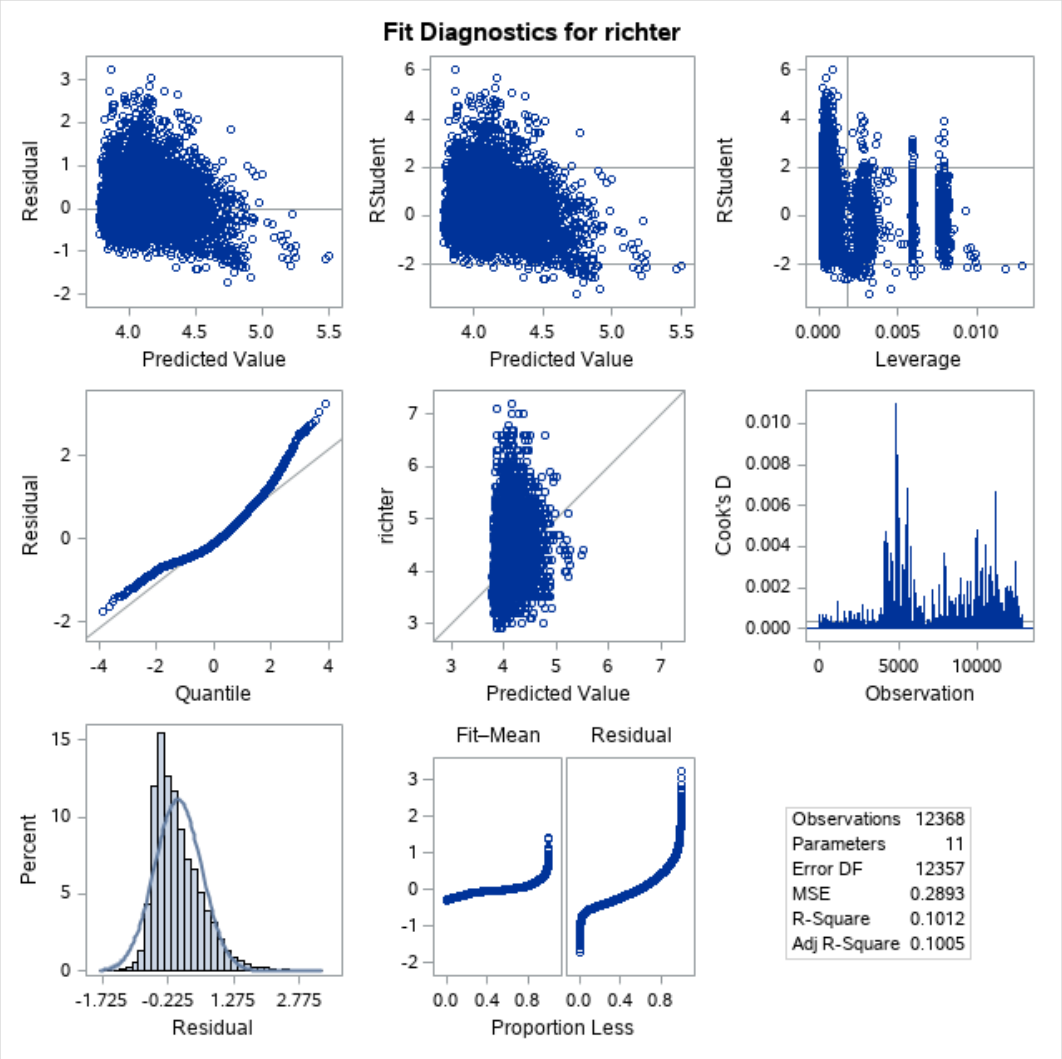
Depth and country are responsible for most of the information gain and are shown to be statistically significant at a 95% confidence level. The latitude and longitude variable have a smaller impact and it is not proven that the addition of longitude improves the model ($p > 0.05$).

Parameter	Estimate		Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	3.054239738	B	0.13806341	22.12	<.0001	2.783613914	3.324865561
country aegean_sea	-0.244613852	B	0.01948338	-12.55	<.0001	-0.282804324	-0.206423380
country georgia	0.148273294	B	0.04510399	3.29	0.0010	0.059862443	0.236684146
country greece	-0.157963054	B	0.01714114	-9.22	<.0001	-0.191562367	-0.124363741
country iran	0.396495259	B	0.05078264	7.81	<.0001	0.296953355	0.496037164
country mediterranean	0.033041535	B	0.01879610	1.76	0.0788	-0.003801755	0.069884826
country other	0.204271852	B	0.02666768	7.66	<.0001	0.151999030	0.256544674

Parameter	Estimate		Standard Error	t Value	Pr > |t|	95% Confidence Limits	
country russia	0.236251224	B	0.05202645	4.54	<.0001	0.134271261	0.338231187
country turkey	0.000000000	B	.	.	.	.	.
lat	0.027025794		0.00332021	8.14	<.0001	0.020517665	0.033533924
long	-0.001924498		0.00105678	-1.82	0.0686	-0.003995958	0.000146962
depth	0.005318205		0.00020262	26.25	<.0001	0.004921045	0.005715364

The table above shows parameter estimates and confidence intervals for each of the categorical levels (with 'turkey' as a reference category).

From this we can see that each of the levels are having some significant impact on response with the exception of 'country mediterranean' which may benefit from further grouping.



The above shows diagnostic plots for the linear regression model.

For the model to be valid, the errors must be independent, have a normal distribution, have a mean of zero, and have constant variance. The errors are assumed to be independent as the earthquakes are assumed to be independent. The QQ plot and histogram of residuals show that the errors are approximately normal. The residuals vs. predicted value chart shows little evidence of dispersion suggesting variance is constant and generally centred around 0.

Based on the above, it is inferred that the assumptions are not broken and that the model is valid.

Using a regression model to predict whether an earthquake is serious or not (defined as having a richter value of 5 or more), how does the model perform and what variables have the biggest impact?

A new variable 'serious' is defined as whether the richter value is great than or equal to 5 with values of either 1 (serious) or 0 (not serious).

As the richter value was used for the previous model, the candidate explanatory variable considerations continue to be true so only 'lat', 'long', 'depth', and 'country' are considered (with some records dropped where 'depth' is not populated).

The data is partitioned in to 'train' and 'test' sets (no validation partition is used as the models are not optimised towards validation performance).

Using the 'train' set, a logistic regression model is fit to predict serious using the candidate variables:

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

Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	4831.096	4635.766
SC	4838.162	4713.495
-2 Log L	4829.096	4613.766

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	215.3300	10	<.0001
Score	270.3322	10	<.0001
Wald	230.9464	10	<.0001

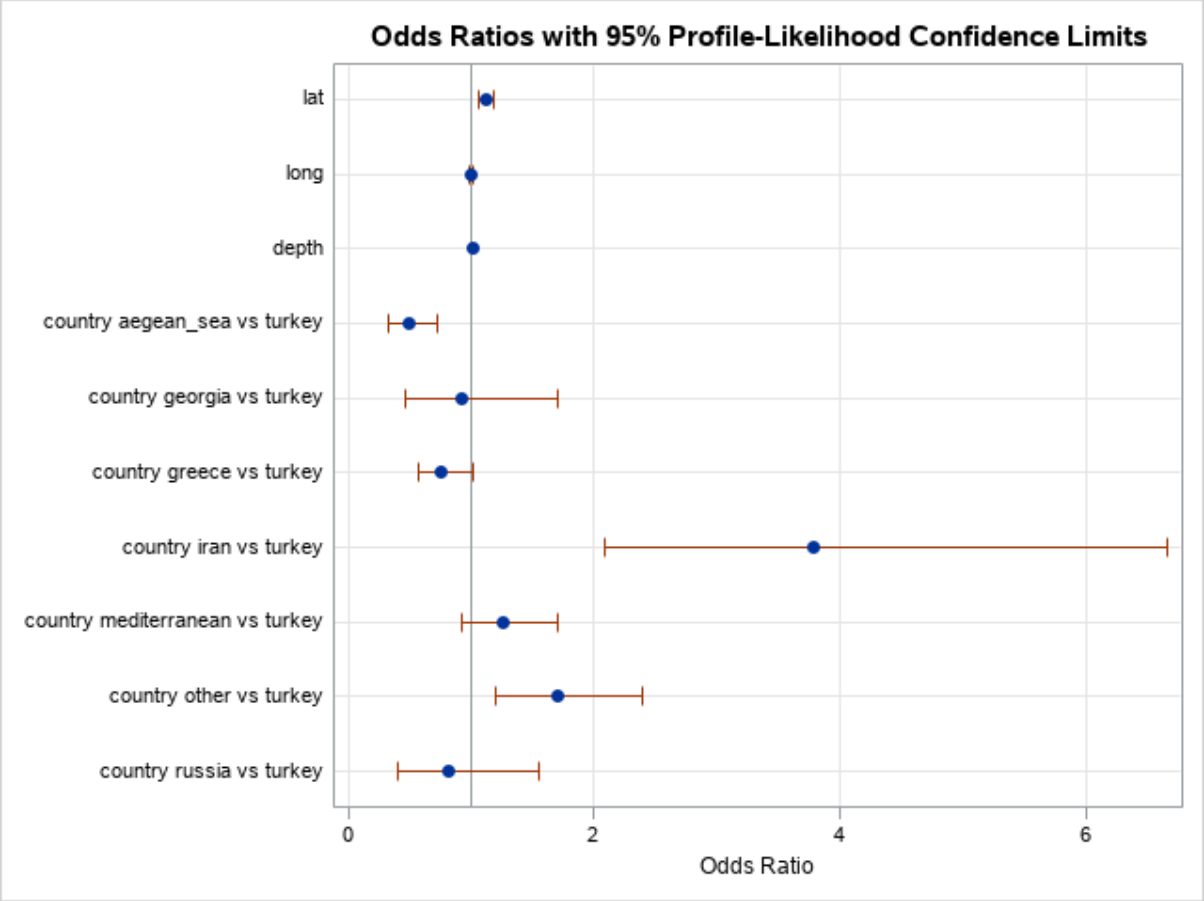
The above tables report on whether the model has converged and some overall model fit statistics.

Based on this, we can first confirm that the model has converged. The test statistics show that the null hypothesis can be rejected at a 95% confidence level and it can be inferred that the model does explain some variation in earthquakes likelihood to be serious.

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	-7.2400	1.1539	39.3707	<.0001
lat		1	0.1161	0.0271	18.3666	<.0001
long		1	0.00435	0.00841	0.2679	0.6047
depth		1	0.0143	0.00124	131.9960	<.0001
country	**aegean_sea**	1	-0.7987	0.1892	17.8242	<.0001
country	**georgia**	1	-0.1834	0.2900	0.4000	0.5271
country	**greece**	1	-0.3807	0.1534	6.1608	0.0131
country	**iran**	1	1.2279	0.2634	21.7275	<.0001
country	**mediterranean**	1	0.1246	0.1613	0.5967	0.4399
country	**other**	1	0.4308	0.1591	7.3303	0.0068
country	**russia**	1	-0.3170	0.3055	1.0764	0.2995

The above table looks at each explanatory variable in turn to provide a parameter estimate and understand if it is having a significant impact on the model (with ‘turkey’ as a reference category for the country variables.

Based on this, we can infer that, ‘lat’, ‘depth’ and some levels of country (‘aegean_sea’, ‘greece’, ‘iran’ and ‘other’) having a significant impact at the model within a 95% confidence level. We cannot reject the null hypothesis for other ‘country’ levels and for the ‘long’ variable.



The above chart shows the odds ratios for each of the explanatory variables.

From this we can see that 'latitude', 'depth', 'country iran', and 'country other' increase the likelihood of having a serious earthquake while 'country aegean_sea' decreases the likelihood.

As some variables are identified as having minimal impact, backwards stepwise regression (based on AIC) is applied to get a final model for comparison:

Summary of Backward Elimination

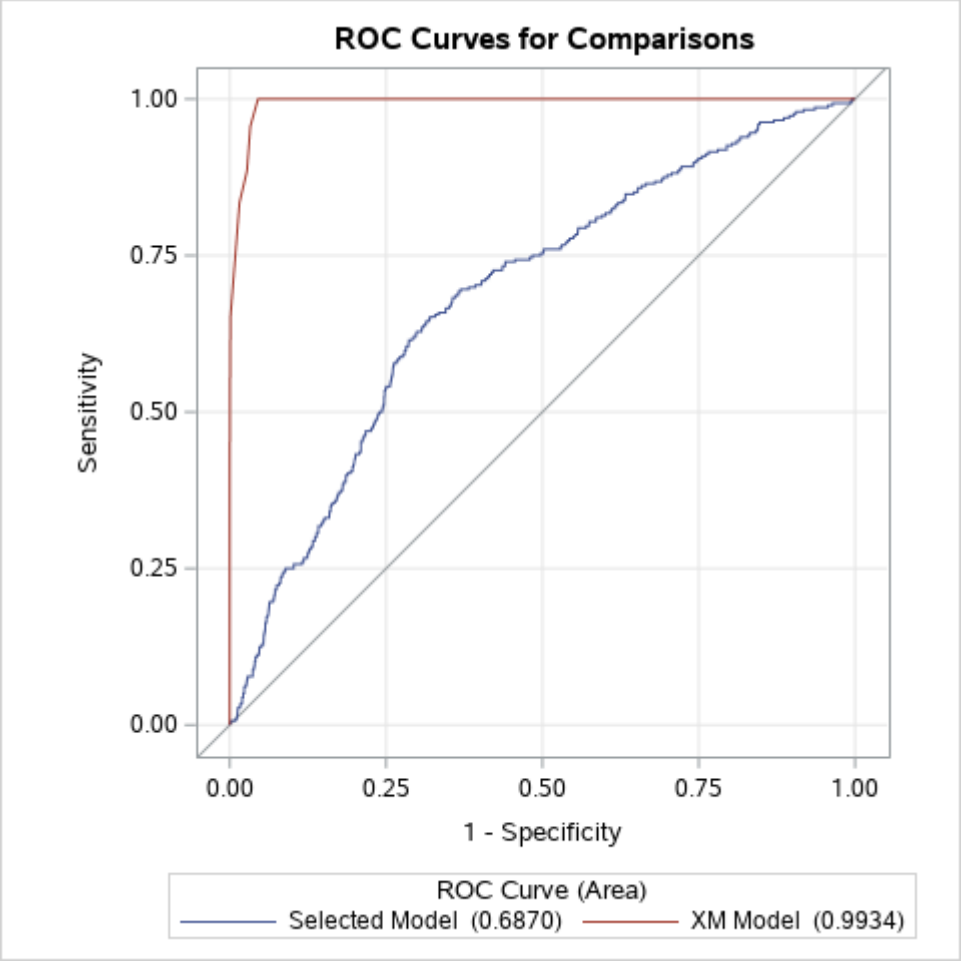
Step	Effect Removed	DF	Number In	Wald Chi-Square	Pr > ChiSq
1	long	1	3	0.2679	0.6047

The table above shows the removed variables from the backward elimination where 'longitude' is removed.

The final model is then confirmed to retain latitude, depth and country as explanatory variables.

Comparing the above model to one that only uses the largest magnitude value, which is most likely to extrapolate best on to new data?

An additional model is created using only the 'xm' variable as an explanatory variable and this is compared to our final selected model on the test data using a ROC curve:



ROC Contrast Test Results			
Contrast	**DF**	**Chi-Square**	**Pr > ChiSq**
Comparing Models	1	378.4853	<.0001

The above uses a ROC curve to assess model performance by plotting sensitivity and 1-specificity for each model over the test dataset and comparing the uplift.

The contrast results show a significant performance improvement for the 'xm' model vs. the selected model and that this is significant at a 95% confidence level. However, 'xm' was a partial product of the 'richter' variable that was used to derive the 'serious' variable so this contains a large amount of target leakage and should be used with caution.

Conclusion

With respect to the proposed lab report questions, the below has been found:

a) Taking the largest magnitude value from each earthquake, is the average value different from 4.1?

The model found that the average value was different to 4.1 (lower) but the model assumptions were not met as the population did not follow a normal distribution.

b) Is there a difference between the average moment magnitude scale value and the country the earthquake occurred in?

There was found to be a statistically significant difference in average moment magnitude by country at a 95% confidence level.

c) Using a regression model to predict richter, how does the model perform and what variables have the biggest impact?

Many of the variables were excluded as they contained a large amount of missing data to avoid target leakage. The resulting model tested how latitude, longitude, depth and country affected the richter value.

The model was found to be statistically significant at a 95% confidence level but only explained a small amount of variation. Depth and country had the biggest impact on the model and longitude was found not have a significant impact on the model when other variables are accounted for.

d) Using a regression model to predict whether an earthquake is serious or not (defined as having a richter value of 5 or more), how does the model perform and what variables have the biggest impact?

Similar to the previous model, the logistic regression model was found to be statistically significant at a 95% confidence level. Also similar to the previous model, depth and country were found to have the biggest impacts with some countries having a large impact (e.g. Iran) and others not shown to be statistically significant (e.g. Georgia).

Backwards stepwise regression removed 'longitude' as a variable but retained all other explanatory variables.

e) Comparing the above model to one that only uses the largest magnitude value, which is most likely to extrapolate best on to new data?

The model that used the largest magnitude value ('xm') as a sole explanatory variable was shown to massively improve the predictive performance on new data compared to the selected model. However, this should be treated with caution as this contains a large amount of target leakage as both the 'serious' and 'xm' variable were derived using the 'richter' variable.

To summarise, the approach taken tried to avoid target leakage as much as possible after the exploratory analysis found a high correlation between all of the magnitude variables. As a result, the models did not explain much variance.

Models could potentially have been improved by grouping some of the country levels further (potentially using Greenacre's method). The numeric variables may also be improved by applying transformations to normalise and standardise the variables at a cost of interpretation.

Generally, this kind of modelling is at its most useful when it could be used to predict situations where earthquakes cause significant risk. As such, future iteration could look to append more variables about where the earthquakes were located (beyond country) and certain conditions associated with those locations (e.g. proximity to a Faultline and time from last earthquake within X radius).