Oklahoma Earthquakes and Wastewater Disposal Wells

Final Report of the Term Project

DSA-5103, Intelligent Data Analytics

Prepared by: 1- Tien Phan

2- Tarek Mohamed

Group 18

**Executive Summary**

The frequency of earthquakes in Oklahoma has increased unpredictably in recent years compared to much less earthquakes in the past. Scientific studies found that the recent spike in earthquakes in Oklahoma is primary due to the injection of wastewater produced during oil production, but not from fracking or so-called “flow-back” water. Produced water that naturally coexists with oil and gas is reinjected into deeper zones via disposal wells that might be responsible for most of the quakes in the Oklahoma region.

This project studies the relationship between the earthquake’s frequency and magnitudes, and the wastewater that is being reinjected in underground wells for some years. Using data science and analytics different tools, the study aims at linking the different intensive relevant data, understanding, visualizing and exploring the data, and constructing different models to try and predict the earthquakes magnitudes based on different characteristics of the wastewater such as the injected water volume, the distances between the injection wells and the earthquakes locations, the dates of the injection, the dates of measuring the earthquakes magnitudes, and the formation used for reinjection. This study also tries to prove or disprove the beforementioned relationship and answer questions such as: does the reinjection of the wastewater cause the increasing earthquakes in Oklahoma? and what are the attributes that have the highest impact? Which factors should we focus on, so we can reduce that harmful effect?

Several regression models are constructed to evaluate the quake magnitude and frequency given other earthquake and wastewater attributes. The problem is a supervised learning problem and the linear and non-linear regression models used include ridge regression, lasso, elastic net regression, and SVM.

The study findings are proving the relationship between the injection wastewater and the earthquakes magnitude and frequency.

We recommend that to reduce the impact of the injected water on the earthquake’s possibilities, magnitudes, and frequencies, the injected water volumes should be reduced, and the injection depths should be reduced.

**Problem Description and background**

Earthquakes could have a huge harmful impact on all the life aspects and normally people try to avoid the earthquake regions for that reason. All shapes of development cannot be sustained or even planned properly if an earthquake is always a possibility. Over the past few years the earthquakes frequency in Oklahoma has increased significantly. Fingers usually point at the fracturing activities as the reason for that, but the studies found that there might be a strong relation between that undesirable phenomenon and the wastewater reinjected into the underground formations in the same state. In oil and gas industry, when the oil is extracted from the subsurface reservoir, some water is extracted with it. This water is useless and the way it gets disposed is by reinjecting it into subsurface formations.

The relationship between the reinjecting activity and the increasing earthquakes is established, but we try to investigate the most important attributes such as the injection depth, water volume, injection distance to the earthquake measurement, and the injection date.

Avoiding the reinjection could be very hard to achieve, so we aim at examining all the factors by using different models, so we can recommend different practices that would possibly reduce the impact to minimum.

**Exploratory data analysis**

**Used Data**

This study utilizes two sets of data as follows:

1- Earthquake Catalog

These catalogs are lists of earthquakes that have occurred within Oklahoma. Each earthquake is accompanied by detailed information regarding its location and size. The study uses the data from the year 2011 until 2017.

Data Source: Oklahoma Geological Survey, OGS

2- Wastewater disposal wells

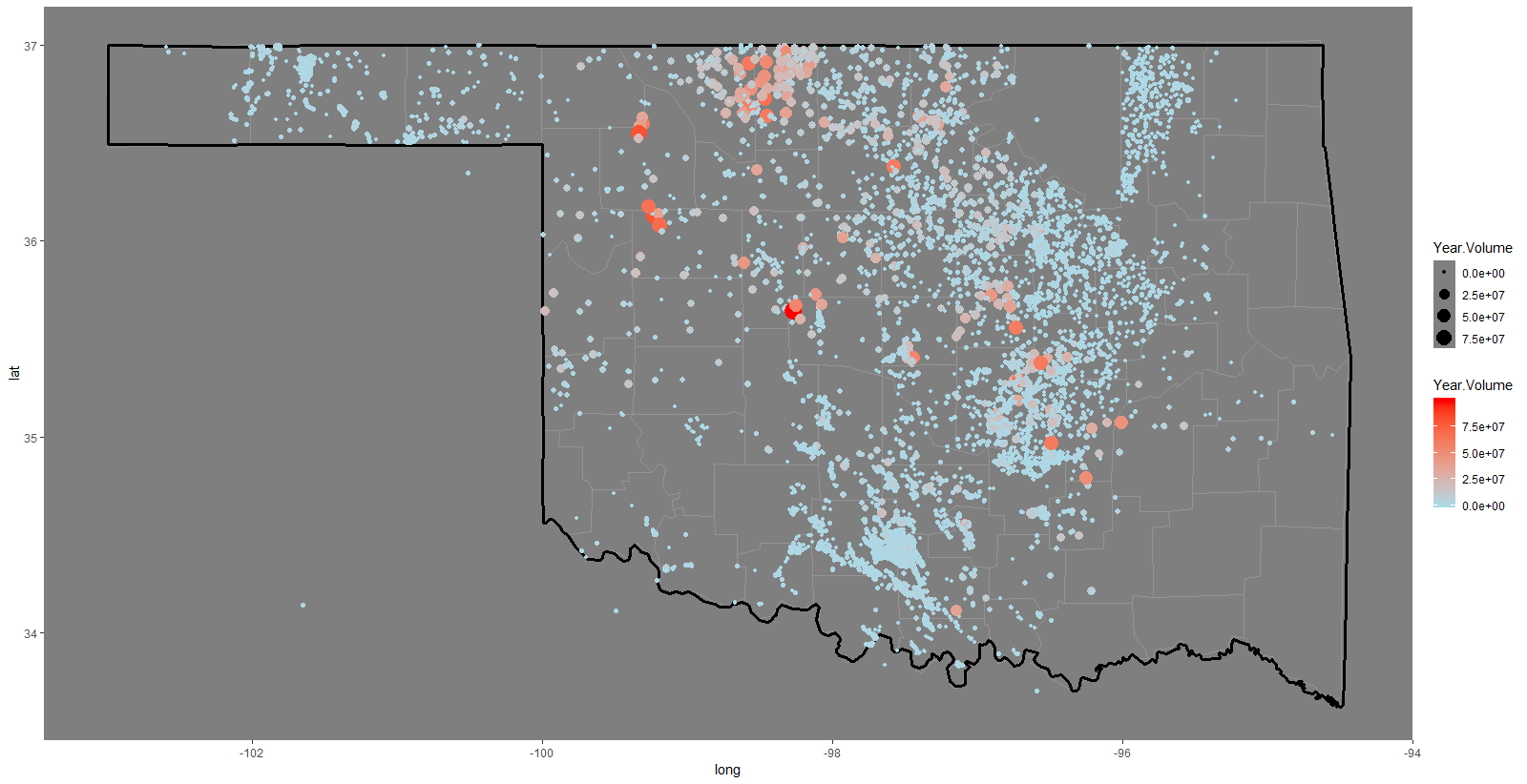
This data contains the wastewater injection wells locations and the water volumes that were injected alongside with many other attributes such as the date of injection; the operator name; the well API; formation code, name, and depth; and the status.

Data Source: Oklahoma Corporation Commission, OCC

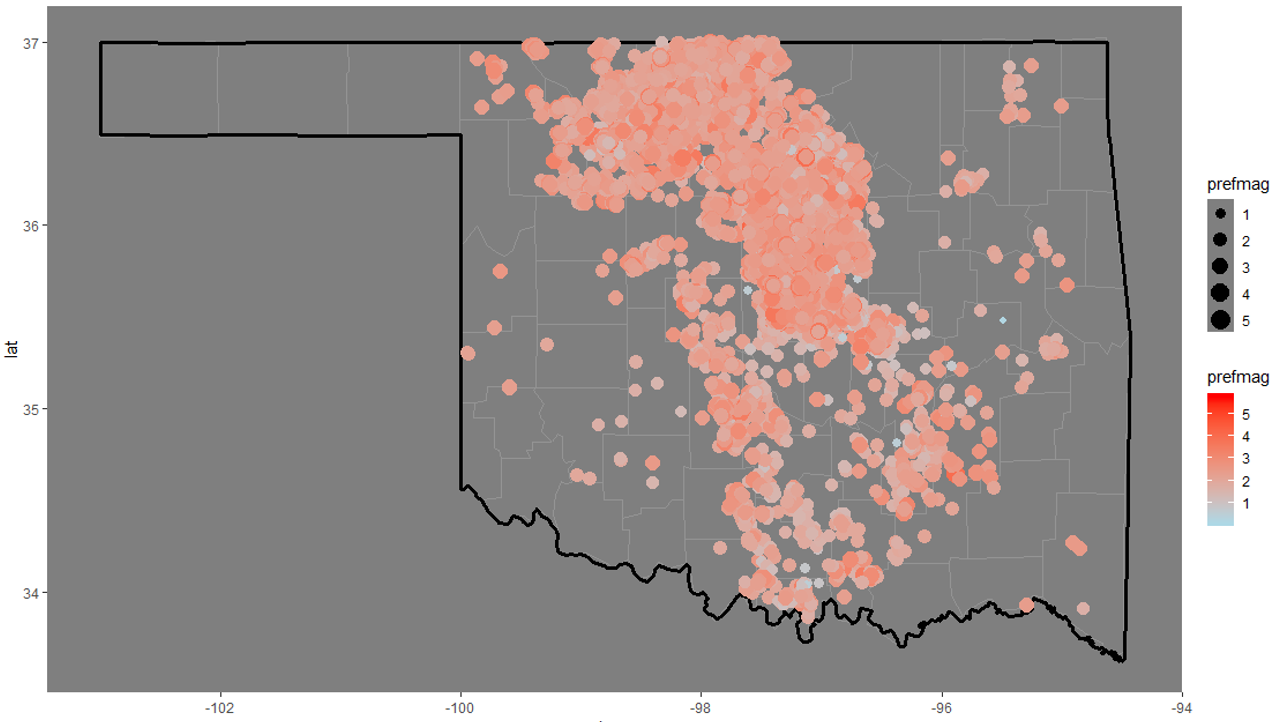
**Data preparation**

We worked on merging the data by mapping the locations of the earthquake’s measurements and the injection wells such that each injection well is surrounded by a group of earthquakes measurement locations and each of these injection well-quake measurements combination makes a node or a record that is used for our model.

Oklahoma map for UIC water injection wells locations and the color legend for cumulative volume of the injected water:



Oklahoma map for earthquakes locations and the color legend for the earthquake magnitudes:



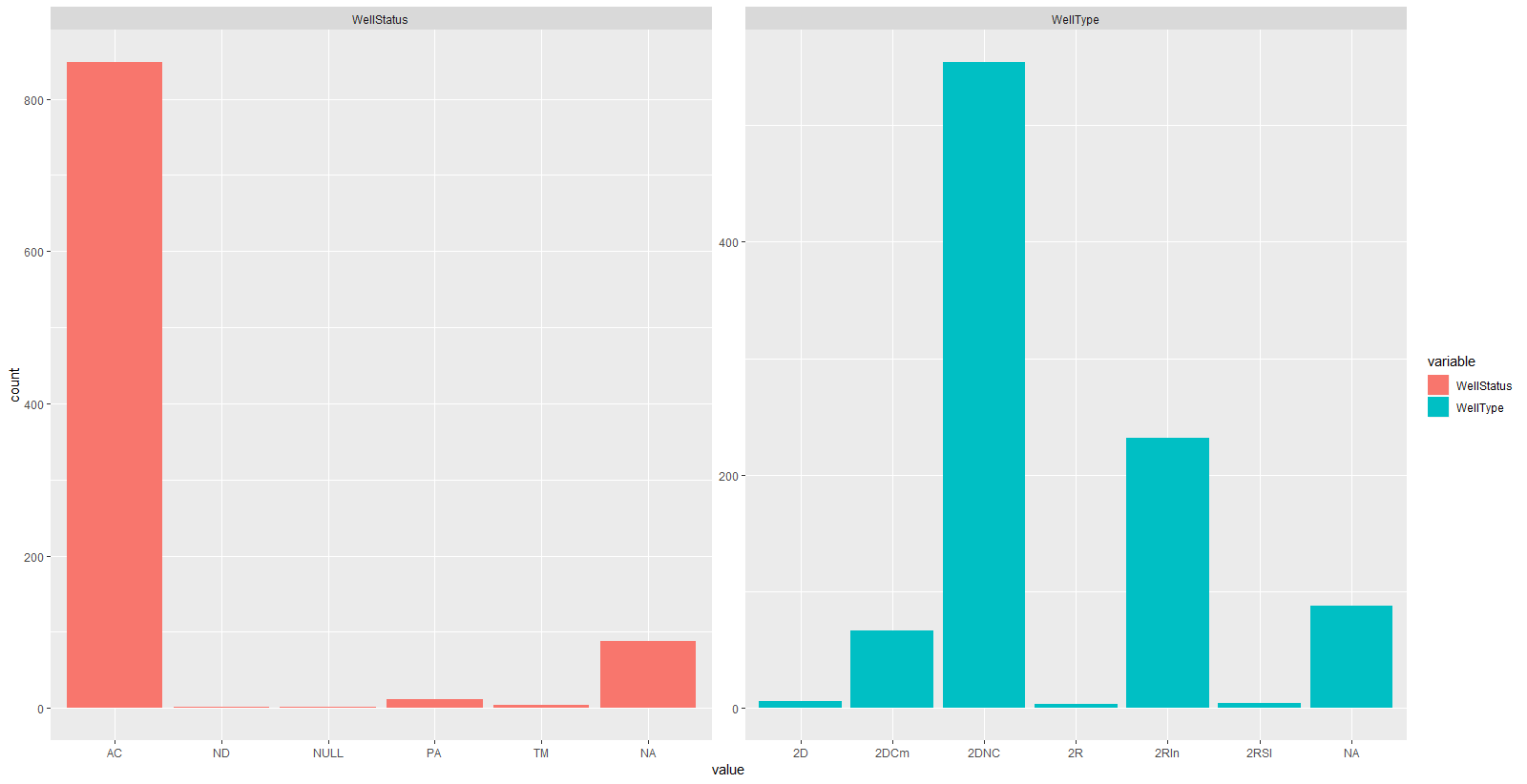
**Data Visualization**

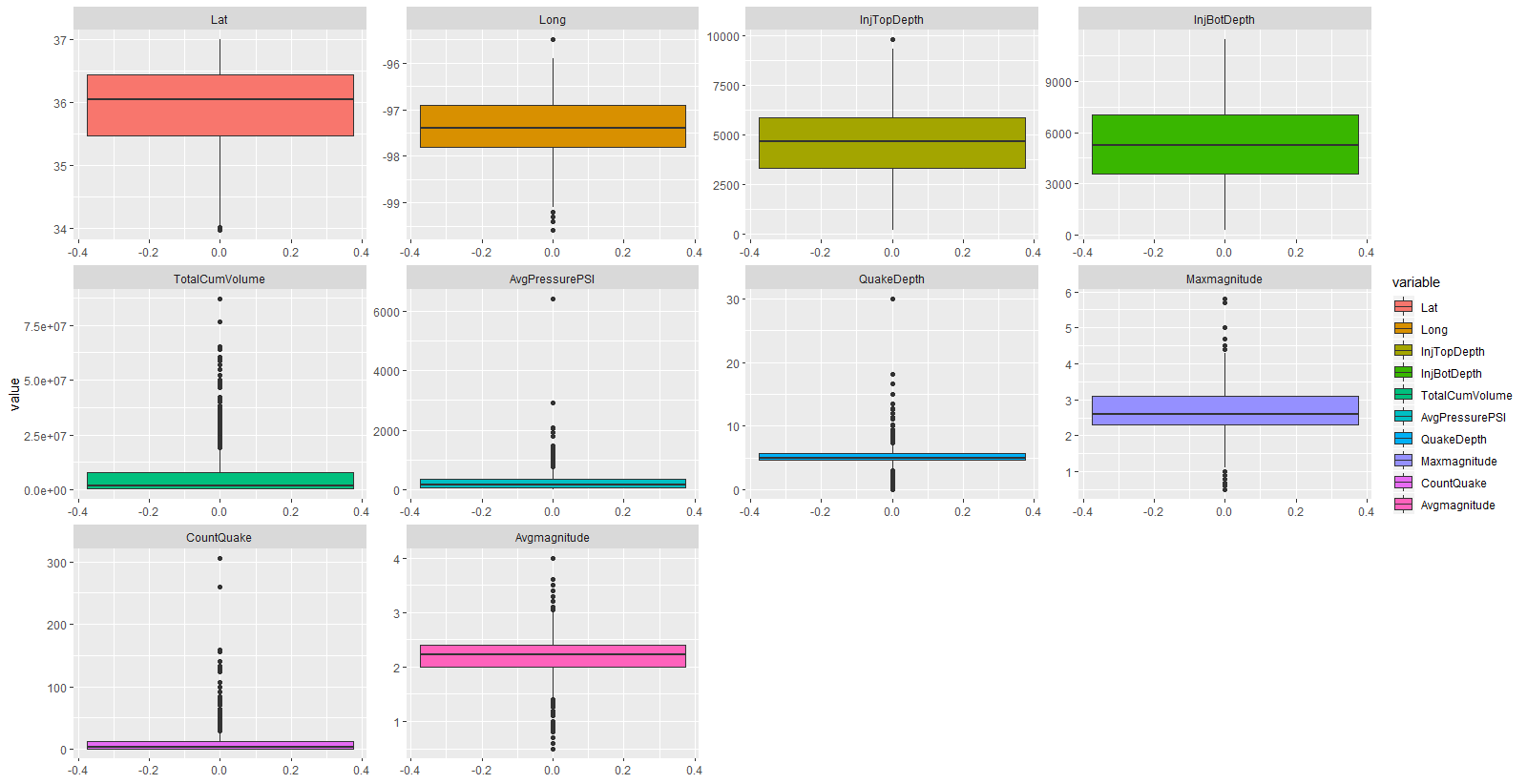
Histogram of numeric variables:

Numeric variable such as latitude and longitude of the wells, Top formation depth, bottom formation depth, total cumulative volume, average injection pressure, quake depth, maximum quake magnitude, quake count, and average magnitude.



Histograms of categorical variables: (well type and well status)

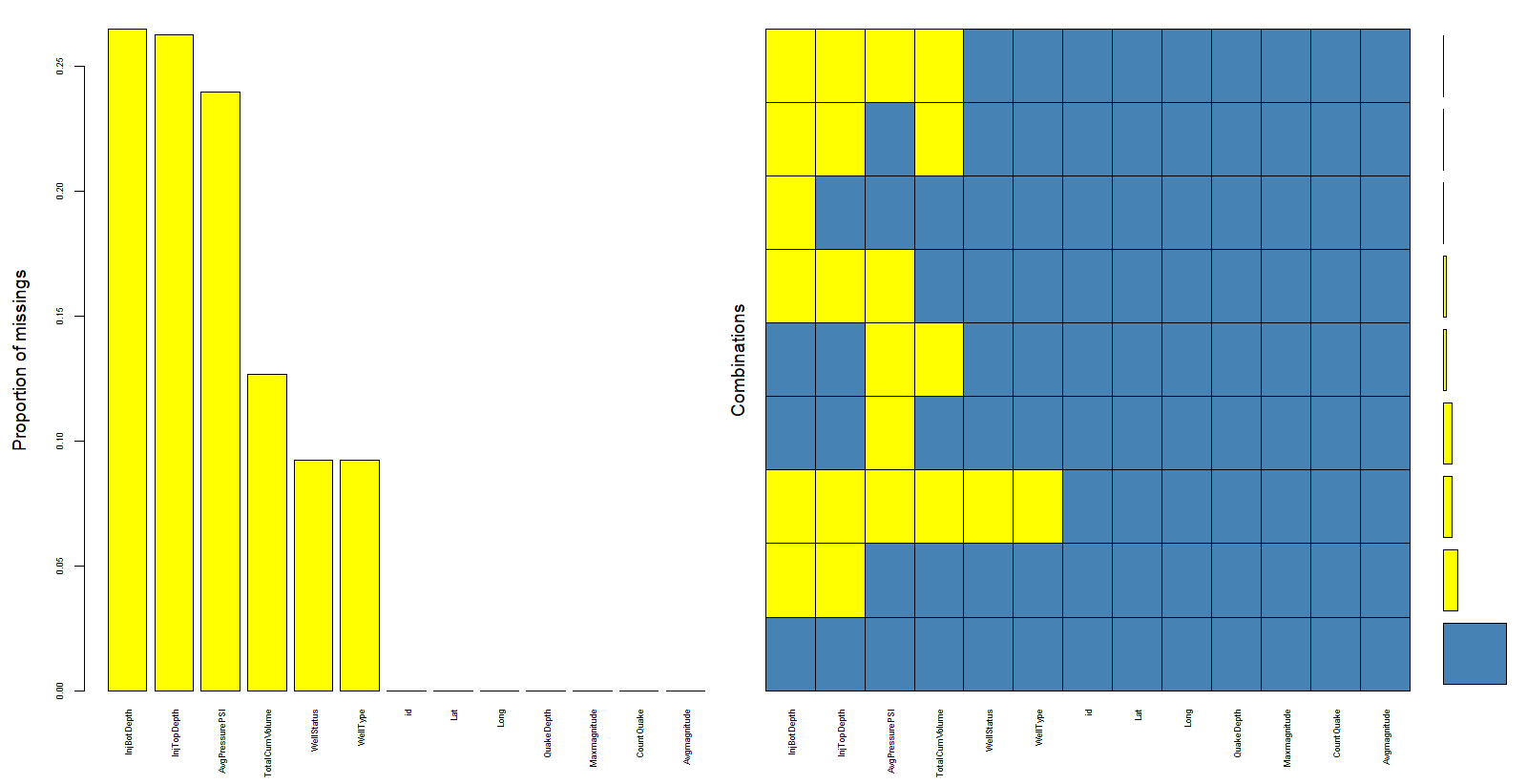




It is clearly observed in the histogram plots that InjTopDepth, InjBotDepth, AvgMagnitude are normally distributed. QuakeDepth also follows normal distribution with some outliers at the high end (~ 30 km). Beside, TotalCumVol and AvgPressurePSI distribution are highly left skewed. In addition, the box-plots provide more information about the number of outliers which lie outside the boxes.

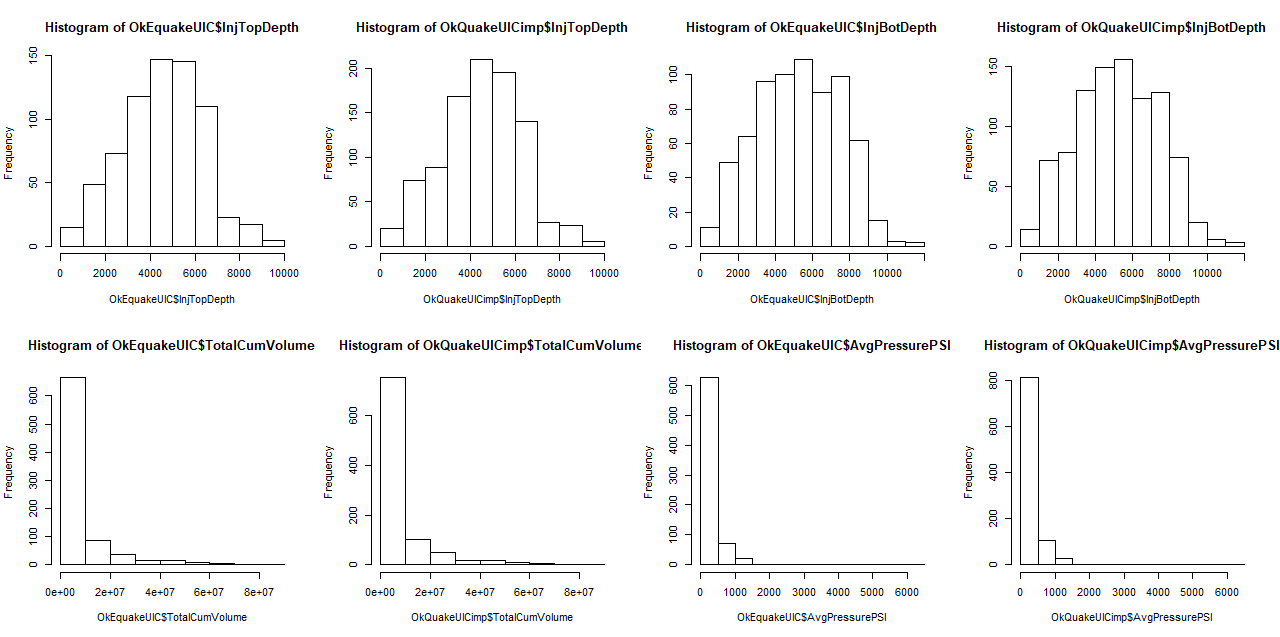
**Missing values**

Looking into the missingness in the data, the following plots indicate that there are some missing values in the InjBotDepth, InjTopDepth, AvgPressurePSI, TotalCumVolume, WellStatus, and Welltype variables. The missingness is of an accepted proportion (less than 27%) that we can deal with by data imputation. We dealt only with the first four mentioned variables’ missingness, because the last two variables are categorical variables that are not important to our study.



**Missing values imputation:**

We tried different imputation approaches and we found that the most reliable approach for our data is the multiple regression imputation. Following are the histograms of each of the four variables that had missing values before the imputation (to the left) and after the imputation (to the right) indicating that the data distributions were not distorted by the imputation. That is, for InjTopDepth and InjBotDepth, the data are still approximately normally distributed after the imputation, and for the other two variables, the data distribution stayed the same as well after imputation.



**Feature engineering**

We added the following predictors to our dataset to help us get more useful insights from our modeling study:

Formation Thickness: by subtracting the Injection top depth from the injection bottom depth, we could investigate the formation thickness impact on the quake’s average magnitude.

Fracture pressure: by adding the hydrostatic pressure at the injection bottom depth to the injection pressure, we could calculate if the total pressure would exceed the formation fracturing pressure and cause it to be fractured or not and if that could have an impact on the quake average magnitude.

Depth Difference: By subtracting injection bottom depth from the depth at which the quake occurred and measured, we could investigate the impact of the relationship between the two depths on the quake average magnitude.

**Analysis plan**

Using several modeling tools, we tried to find the best model to represent the data. We used pls, ridge regression, lasso, and elastic net regression as follows.

1- PLS Model:

The first model we trained using PLS method. Cross-validation is applied to select the optimal component n which yields the smallest error (RMSE).

ncomp RMSE Rsquared MAE

#1 0.4232620 0.001471229 0.3066016

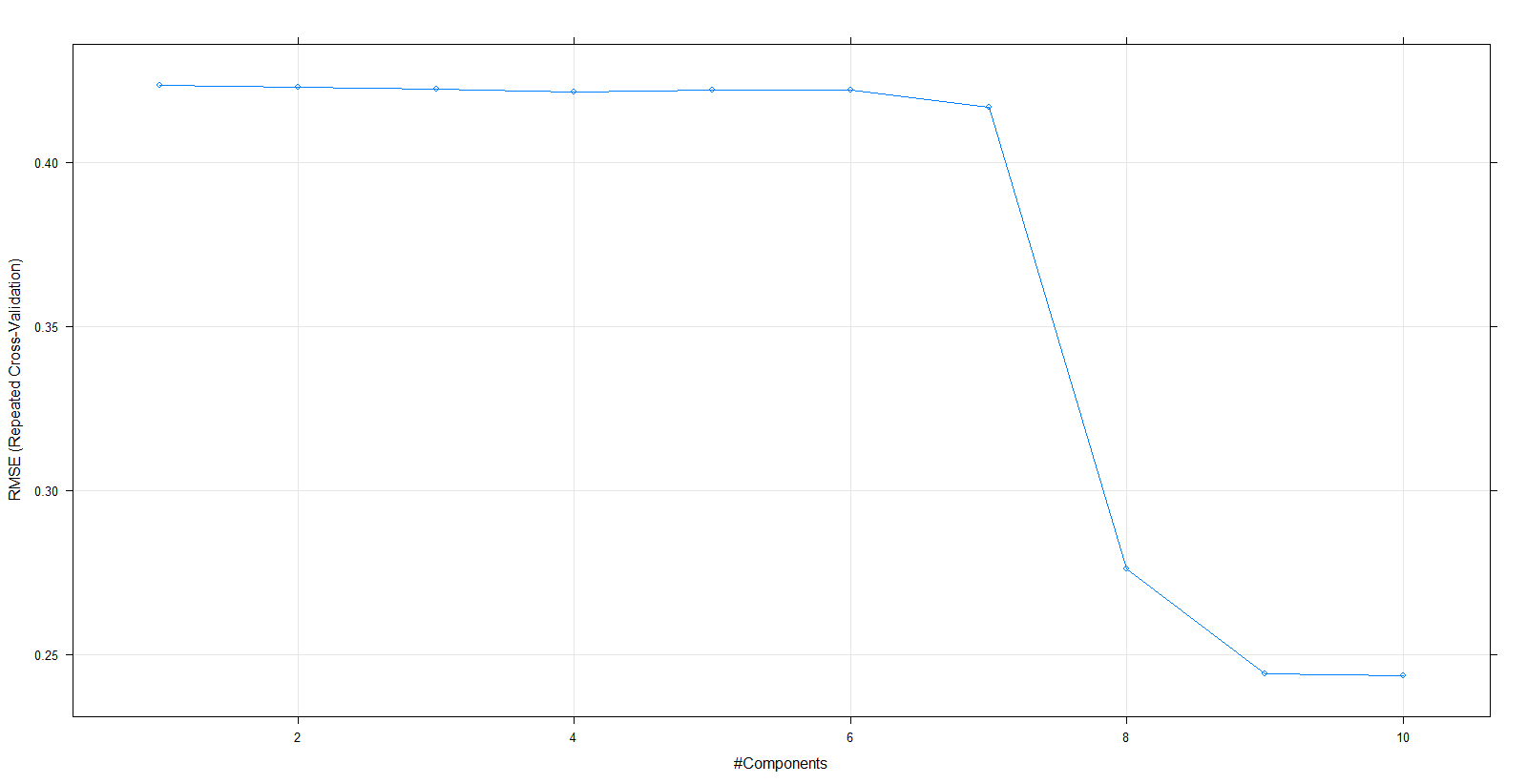
…..

#8 0.2761265 0.573424598 0.2048703

#9 0.2440894 0.667498946 0.1803778

#10 0.2436752 0.668723122 0.1805082

It turned out that the model with ncomp = 10 is considered the best. Cross Validation plot for RMSE against ncomp is shown below.



Then, we extracted the model results with all performance metric as follows:

ncomp RMSE Rsquared MAE RMSESD RsquaredSD MAESD

10 0.2436752 0.668723122 0.1805082 0.01887480 0.056724385 0.01071111

2- Ridge Regression model:

Ridge regression algorithm is trained to evaluate our data. Our cross-validation is focused on finding the optimal lambda value given alpha value is equal to 0.

lambda RMSE Rsquared MAE

0.00 0.2465873 0.6664280 0.1818106

0.01 0.2465873 0.6664280 0.1818106

0.02 0.2465873 0.6664280 0.1818106

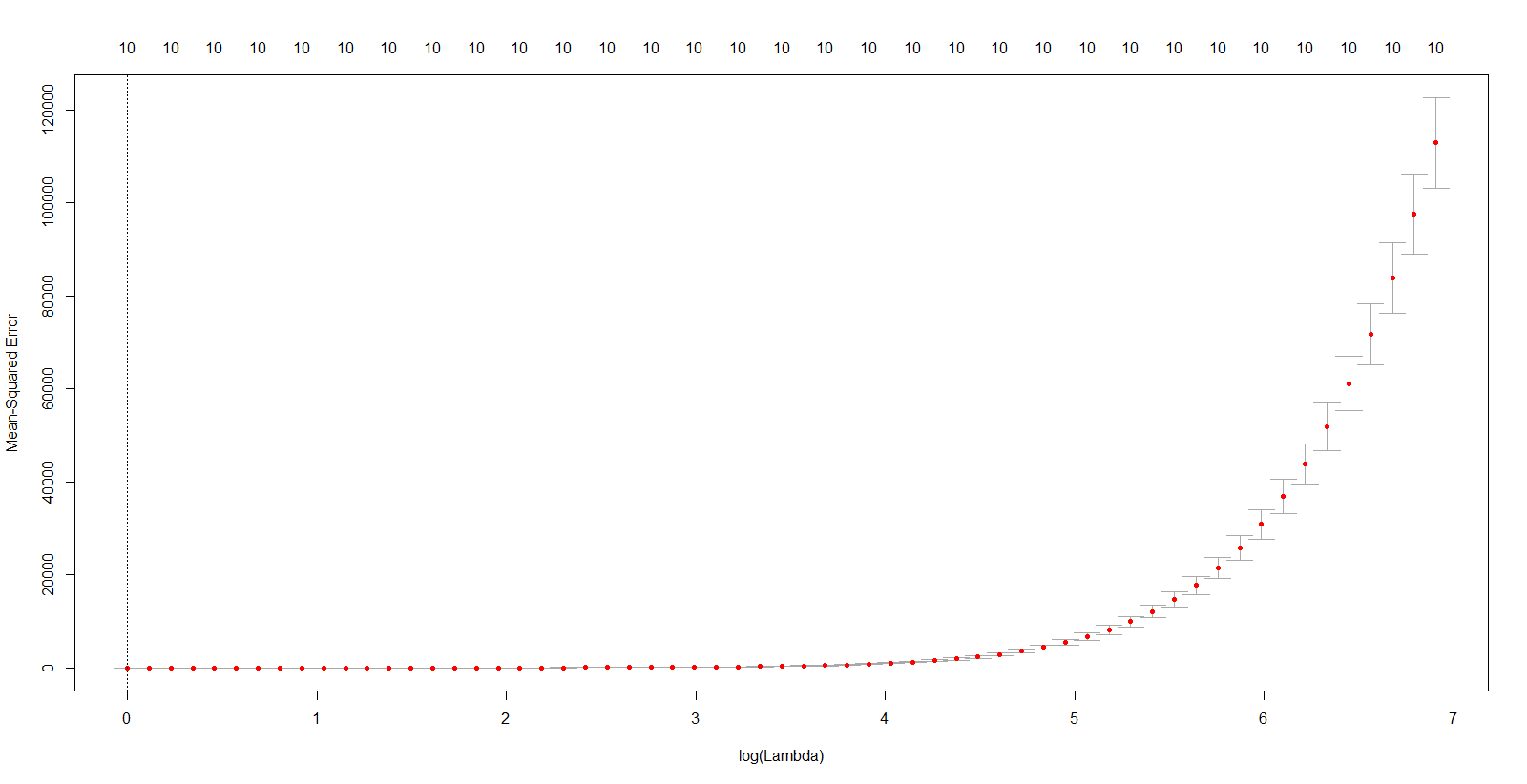
0.03 0.2466412 0.6664086 0.1818324

0.04 0.2481078 0.6655862 0.1825870

0.05 0.2498597 0.6644709 0.1835550

And to be consistent, RMSE was used to select the optimal model using the smallest value.

The final values used for the model were alpha = 0 and lambda = 0.02.



Ridge model result associated with lambda = 0.02 is presented in the following:

alpha lambda RMSE Rsquared MAE RMSESD RsquaredSD MAESD

#3 0 0.02 0.2465873 0.6664280 0.1818106 0.01825642 0.05318186 0.01059382

3- Lasso Model:

Next, we tried Lasso model. As the same with Ridge, we found the optimal lambda value which gives the best model. However, the alpha value is given at 1. Cross-validation results for different values of lambda:

lambda RMSE Rsquared MAE

0.000 0.2186788 0.6778726 0.1640854

0.001 0.2186431 0.6779265 0.1638589

0.002 0.2185300 0.6782614 0.1637503

0.003 0.2184747 0.6784839 0.1637179

0.004 0.2185140 0.6784751 0.1637517

0.005 0.2186097 0.6783477 0.1638215

0.006 0.2187194 0.6782249 0.1639165

0.007 0.2188470 0.6780913 0.1640191

0.008 0.2190069 0.6779071 0.1641384

0.009 0.2192025 0.6776581 0.1642839

0.010 0.2194335 0.6773396 0.1644524

Tuning parameter 'alpha' was held constant at a value of 1

RMSE was used to select the optimal model using the smallest value.

The final values used for the model were alpha = 1 and lambda = 0.006.

alpha lambda RMSE Rsquared MAE RMSESD RsquaredSD MAESD

1 0.006 0.2187194 0.6782249 0.1639165 0.009049659 0.02597753 0.005774218

4- Elastic net regression model:

The final model we train the data is elastic net regression. In this algorithm, we need to find both alpha and lambda in our cross-validation step.

lambda fraction RMSE Rsquared MAE

0.0 0.5000000 0.2662171 0.6321153 0.1942136

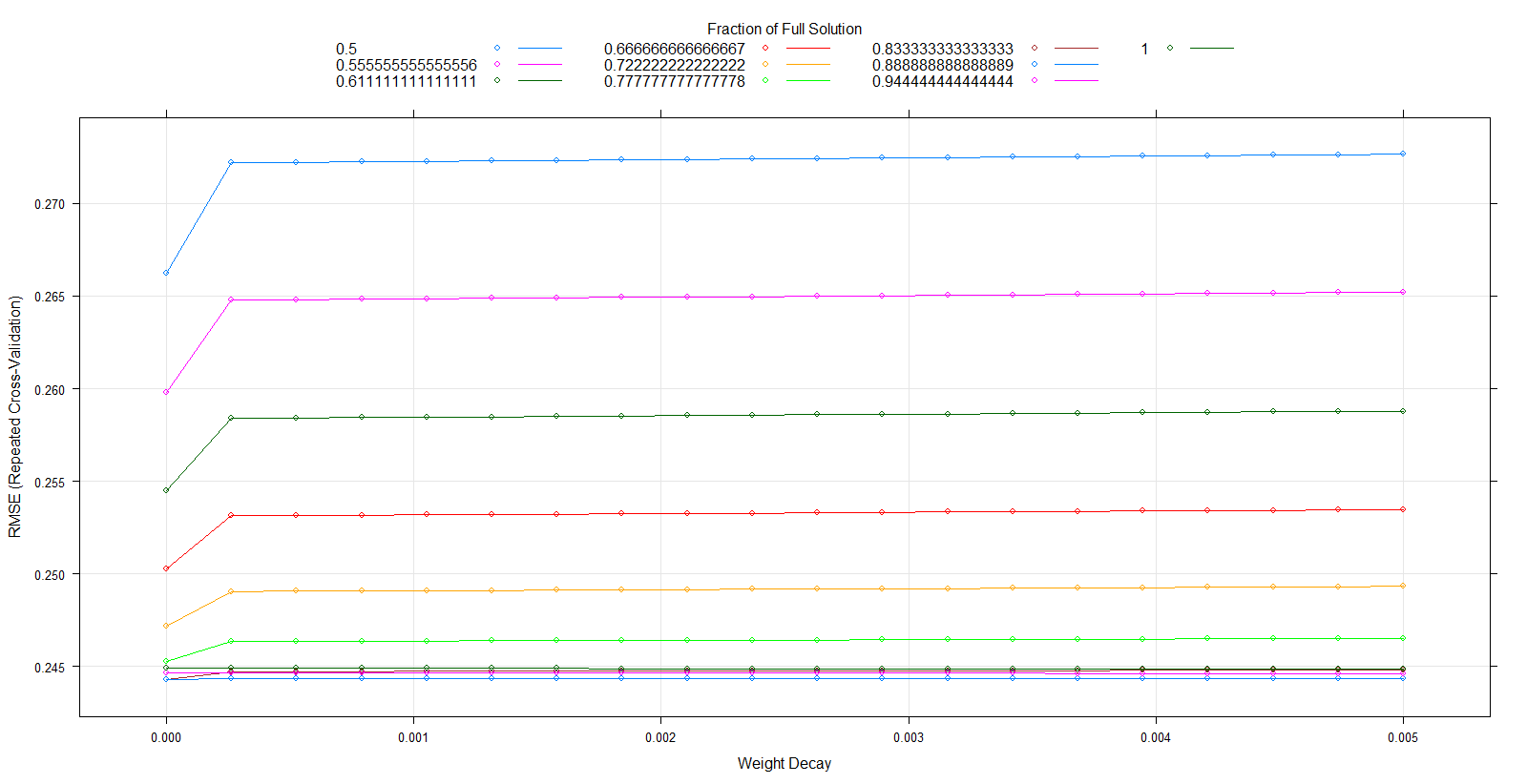
0.0 0.5555556 0.2598033 0.6455985 0.1899760

0.0 0.6111111 0.2544875 0.6552262 0.1865865

0.0 0.6666667 0.2502843 0.6617120 0.1839023

0.0 0.7222222 0.2471671 0.6657729 0.1819234

0.0 0.8888889 0.2442909 0.6678670 0.1800885



RMSE was used to select the optimal model using the smallest value.

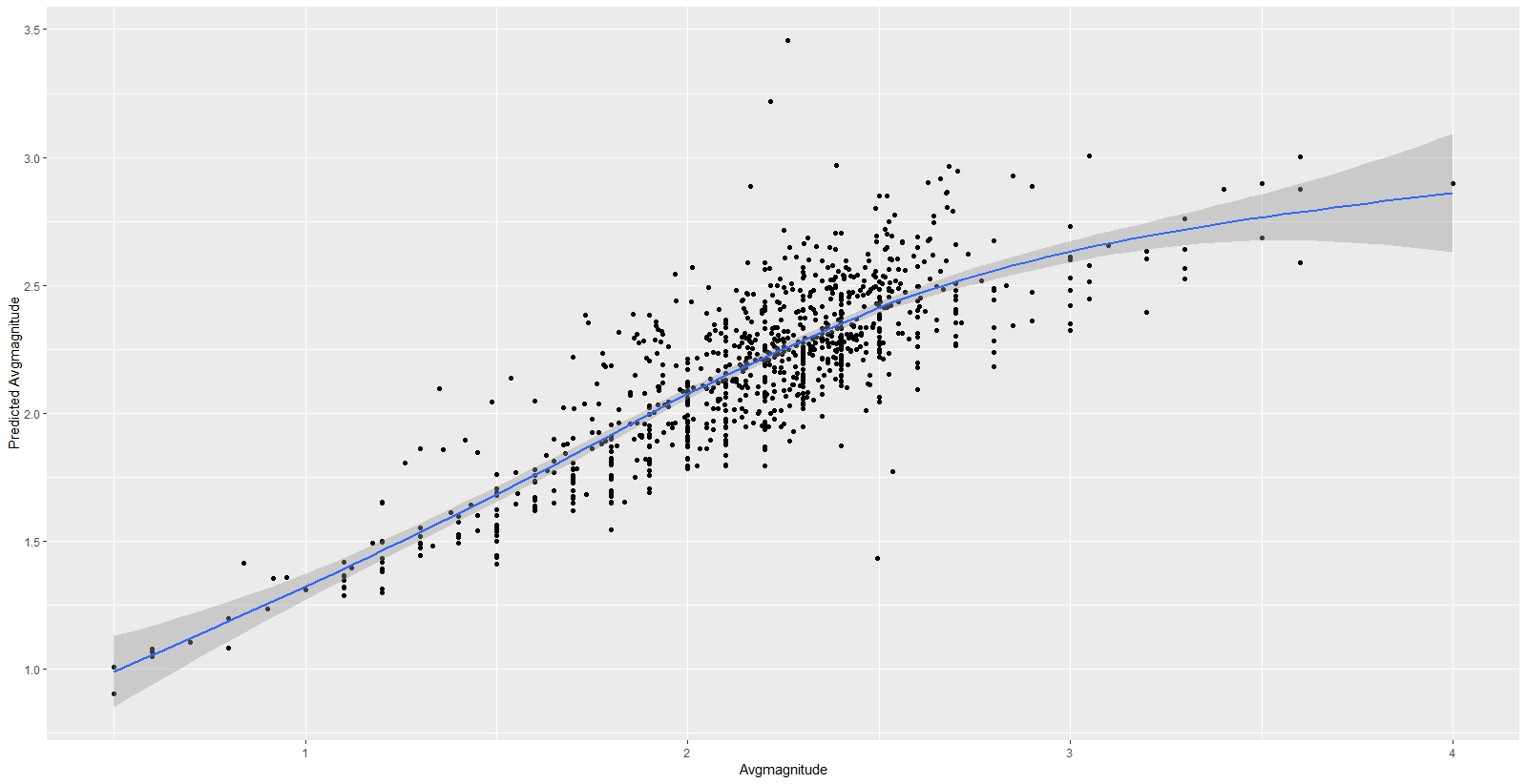
The final values used for the model were alpha = 0.889 and lambda = 0.

lambda fraction RMSE Rsquared MAE RMSESD RsquaredSD MAESD

0.0 0.8888889 0.2442909 0.6678670 0.1800885 0.01532513 0.04675756 0.008034092

A summary of all models associating with their peformance metrics are presented below:

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Tuned hyper-parameters** | **R-squared** | **RMSE** |
| PLS | Ncomp = 10, alpha = 0.889, lambda = 0 | 0.668723 | 0.243675 |
| Ridge | Alpha = 0, lambda = 0.02 | 0.666428 | 0.24659 |
| Lasso | Alpha = 1, lambda = 0.006 | 0.678225 | 0.21872 |
| Elastic-Net | Alpha = 0.889, lambda = 0 | 0.66787 | 0.24429 |



**Conclusion**

We trained the regression model to predict the average earthquake magnitude using different supervised algorithms encompassing PLS, Ridge, LASSO and Elastic Net. Cross-validations were also evaluated using different metrics such as R2, RMSE and MAE. It turned out that LASSO model outperforms the remaining models based on its smallest RSME and highest R2 metrics.

In addition, data processing, exploratory and feature engineering are the critical parts of the entire modeling process. Starting from two separated data sets including injection volume file history and earthquake catalog, we proposed a reasonable and concreted approach to merge them into a consistent data frame which can be used in the model training section.

In summary, regression models can be used to build the relationship between the earth quake and injection volume data. Regression model results provide a considerable reliability with R2 = 68% in Lasso technique. Moreover, the approach used in this study can be incorporated with more geological features such as fault, seismicity to provide more insights into earth quake events.