Capstone Project 1: Investigation the correlation between seismic activities and injection activities in Oklahoma

The seismic activities significantly increase in Oklahoma in the past decade and the rise of seismic activities draws a lot of attention and concern from scientists, government, and the public. During the same period, the oil & gas operation activities increase a lot in Oklahoma. Among all the possible reasons, water injection (water disposal) is considered as the most important reason for the rise of seismic. Water disposal is a process to inject the produced water back to the underground to balance the reduced reservoir pressure in the oil production. From the scientific point of view, the water injected back to the underground might behaves as lubricant, which makes the faults or strakes easier to slip, causing seismic activities. The objective of this project is to evaluate the correlation between seismic activities and injection activities from the point of data analysis. Until now, I have completed three steps of data analysis for this topic: data cleaning, storytelling, and statistical inference.

1. data cleaning

There are two datasets in my capstone project 1: one is the water injection data in Oklahoma from 2006 and 2017 and the other one is seismic data in Oklahoma from 2000 to 2018. The objective is to find correlation between water injection and seismic activities in Oklahoma in the past decade.

Step 1: import data
These two data sets are from the website. I utilized request.get() command to obtain the data in Json type and converted into dataframe. One main issue what I encounter is that there is data limitation for each request from API and I was only able to obtain 2000 data points for each request. To tackle this issue, I utilized while loop and resultoffset command to obtain the full set of data.

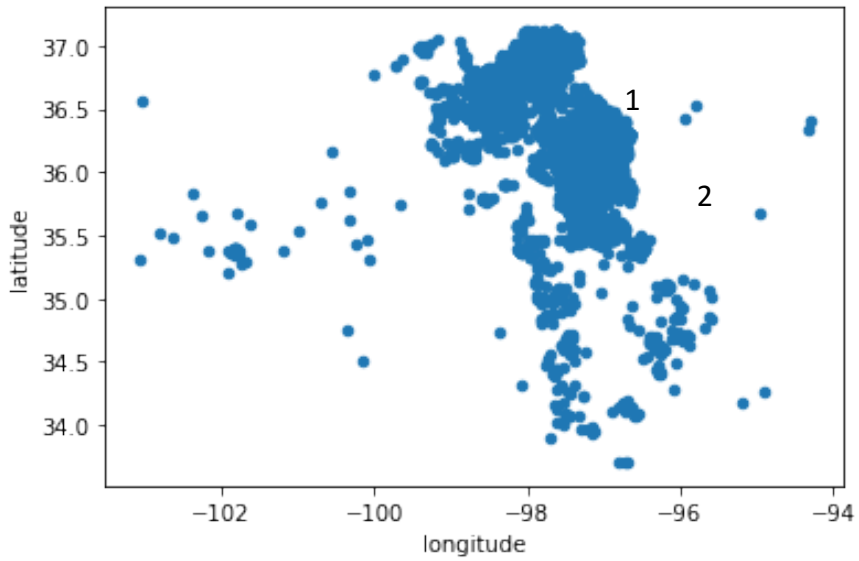
Step 2: data cleaning
I utilized command info() and head(), columns to get a general idea of the data. I dropped unnecessary columns (for example, well names, well numbers) in the injection data as they are not needed in my analysis. I found that there are injection volumes for each year and also for each month from year 2006 and 2017, thus there are more than 100 columns. In my analysis, I would like to analyze the data on a yearly basis. As a result, I created a new injection data by selecting only the yearly volume. Similar issues occur to the seismic data as well. I also dropped some unnecessary columns as well. In addition, the time I initially imported are not in the right format, then I utilized to_datetime command to transform into the right time format. After that, I created a new column to collect the year information of seismic events based on the time.

Step 3: group data
This step is the most important steps in data cleaning. In these two data sets, there is location information for each injection activity and seismic activity by telling the latitude and longitude. As a result, each data points are scattered points in space. In order to find correlations, I have to group these scattered points into some regions. I first looked at the seismic map in Oklahoma and figured out that there are two main seismic activity regions and then measured the latitude and longitude of these two regions (In region one, latitude [36.25, 37], longitude [-99, -97.7]; and in region two, latitude [35.5, 36.6] and longitude [-97.7, -96.7]). Then I grouped the injection and seismic data points into groups by generation a new column named region based on the value of latitude and longitude: if the data is in region one, then the value is 1, if the data is in region two, then the value is 2, otherwise the data is NaN.

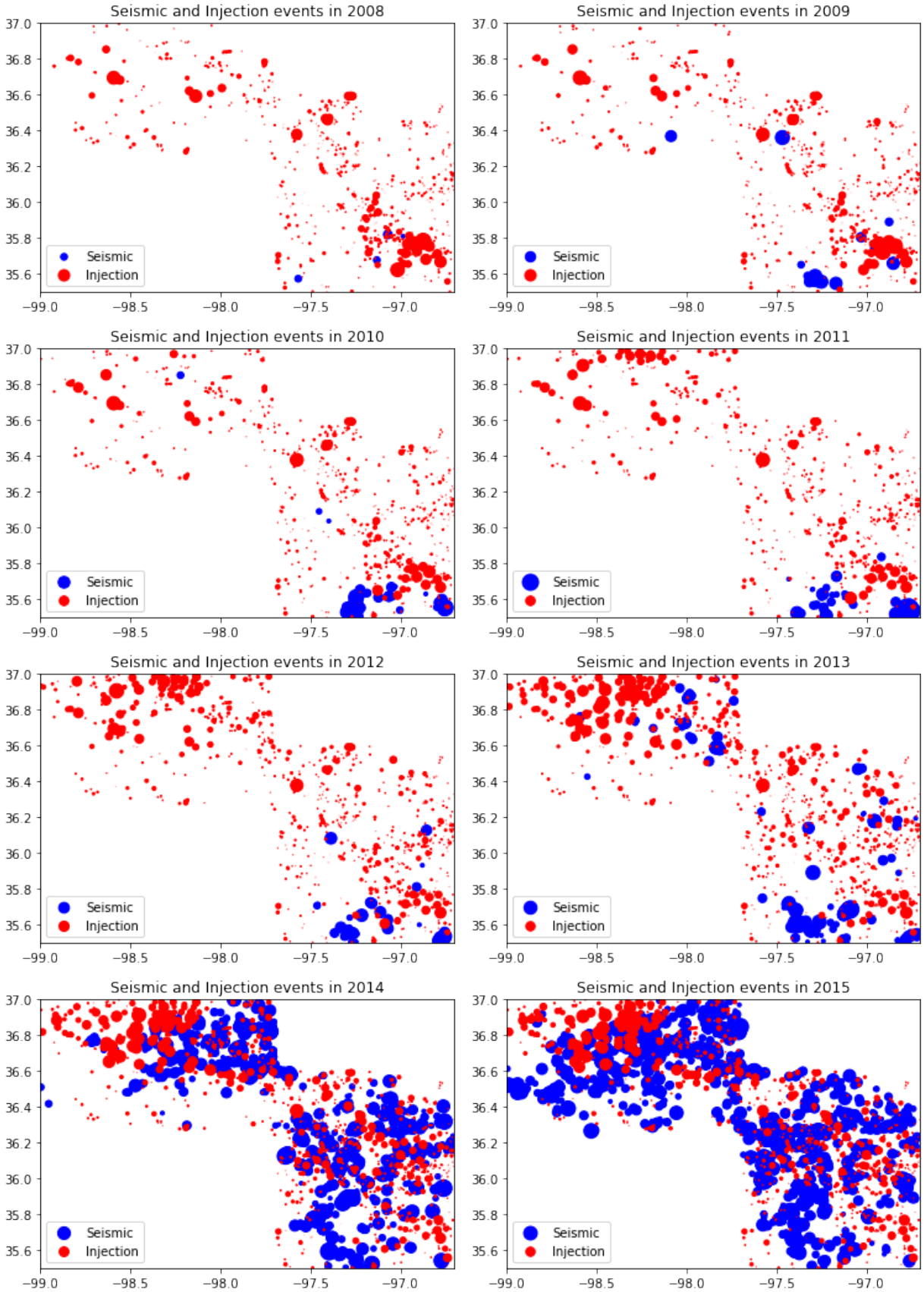
Step 4. Drop NaN values
I utilized info() command to check if there is any NaN values in the dataframe. I found that that the only NaN values occur in the column group. Then I dropped the NaN values by using the dropna function.

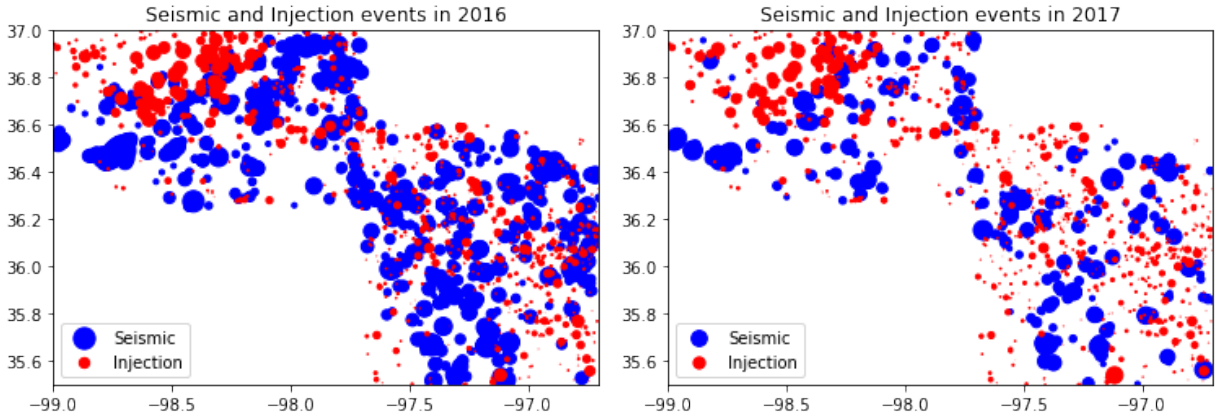
2. story telling

After importing data and data cleaning, I first visualized the seismic activities from 2008 to 2017 in OK by plotting all the events in the space, which is shown below. As illustrated, the main seismic activities occur in two regions, namely region 1 and region 2. Each seismic activity is a single data point; we have to group these scattered data points into groups for the sake of data analysis. As a result, I only take the data in region 1 and 2 in the data analysis.



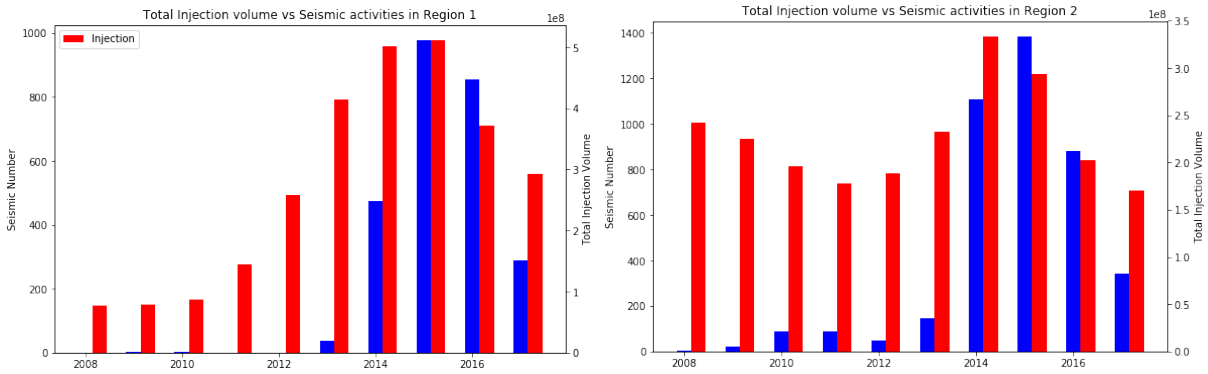
After grouping into two regions, I plotted injection wells and seismic activities in the same figure from 2008 to 2017. In the figure, the red dots indicate injection wells and the larger the red dot, the more injection volume in that well. The blue dots indicate seismic activities and the larger the size, the larger magnitude it is. As indicated in the figures, there is few seismic activities before 2013 and the seismic activities rise significantly in 2014 and peaks in 2015. In the same time, the injection activities also increase from 2008 to 2015. After 2015, the seismic activities decrease.


Seismic and Injection events in 2008
Seismic and Injection events in 2009
Seismic and Injection events in 2010
Seismic and Injection events in 2011
Seismic and Injection events in 2012
Seismic and Injection events in 2013
Seismic and Injection events in 2014
Seismic and Injection events in 2015
Seismic
Injection
37.0
36.8
36.6
36.4
36.2
36.0
35.8
35.6
−99.0
−98.5
−98.0
−97.5
−97.0

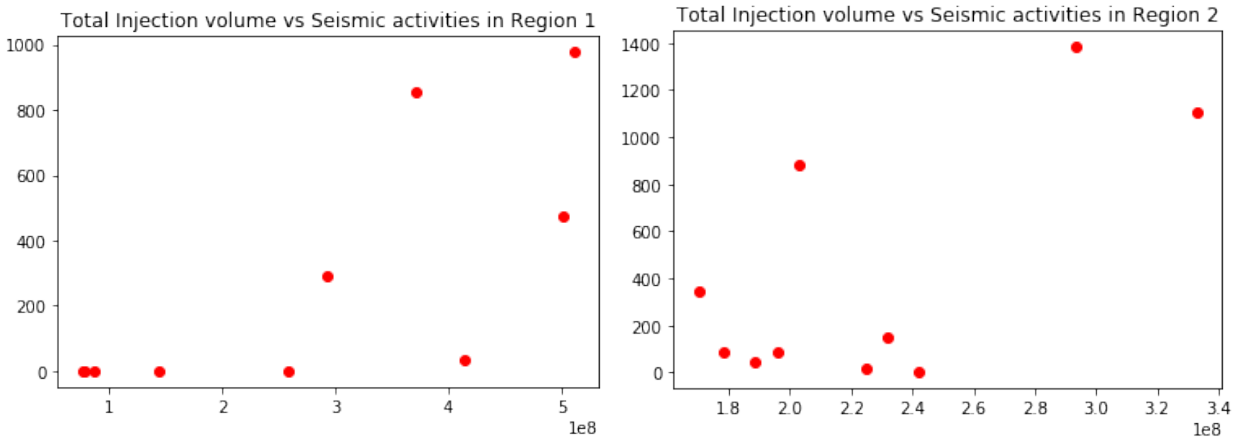


The relationship between total injection volume and seismic activities

In this section, I investigated the relationship between the total injection volume and seismic activities in regions 1 and 2. In the following bar plots, the red bar is the total injection volume and the blue bar is the total seismic activity numbers. As indicated in the bar plots, the total injection volume increases from 2008 to 2015 and decreases from 2015 to 2017 in region 1 and in the meantime, the seismic activities increase before 2015 and peak in 2015 and decreases afterwards. Similar observation was observed in region 2. However, there is very few seismic activities before 2012 in region 1 while there is relatively larger number of seismic activities in region 2. One possible reason might be larger injection volumes in region 2 compared to region 1 before 2012. Compared to region 2, region 1 is a relatively new developed field.

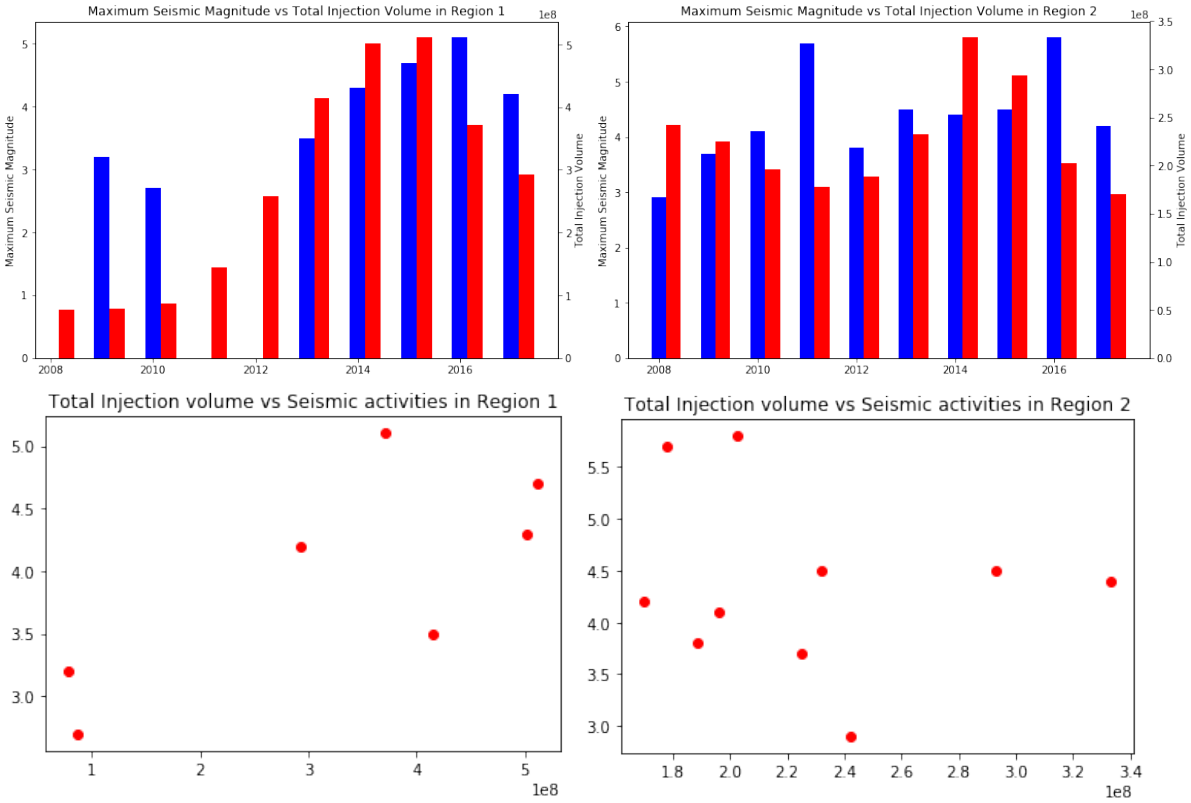


The dot plots are also shown to illustrate the relationship between total injection volume and seismic activities. The plot shows that for both regions 1 and 2, the seismic activities increase as the total injection volume increases.



The relationship between maximum seismic magnitude and total injection volume

I also investigated the relationship between the injection volume and the maximum seismic magnitude. As shown in the plots, the seismic magnitude increases as the total injection volume increases in region 1. However, there is no correlation between seismic magnitude and the total injection volume in region 2. A possible reason might be that region 2 is a more matured field. If enough water is already injected, many faults are already activated and energy is already released. In that case, injecting more water might cause fault slips, but may not cause large slips in the fault as the fault slipped before and energy was released. However, for a relative new field like region 1, region 1 is a relative new field and many faults are not activated. Water acts like lubricant and Injection more water might cause more slips in the fault, causing larger magnitude seismic.



3. statistical inference

The main objective of my capstone project 1 is to evaluate the correlation between seismic activities and the injection activities. To evaluate whether seismic activities and injection activities are strongly correlations, I apply two methods: confidence interval and T-test. The data we are focusing on is the total seismic numbers and total injection volumes between 2008 and 2017 in Oklahoma states. For the sake of illustration, I will only choose the data from zone 2 (there two zones of data sets in my analysis based on the location). The hypothesis we are making is that seismic distribution and injection volume distribution are the same.

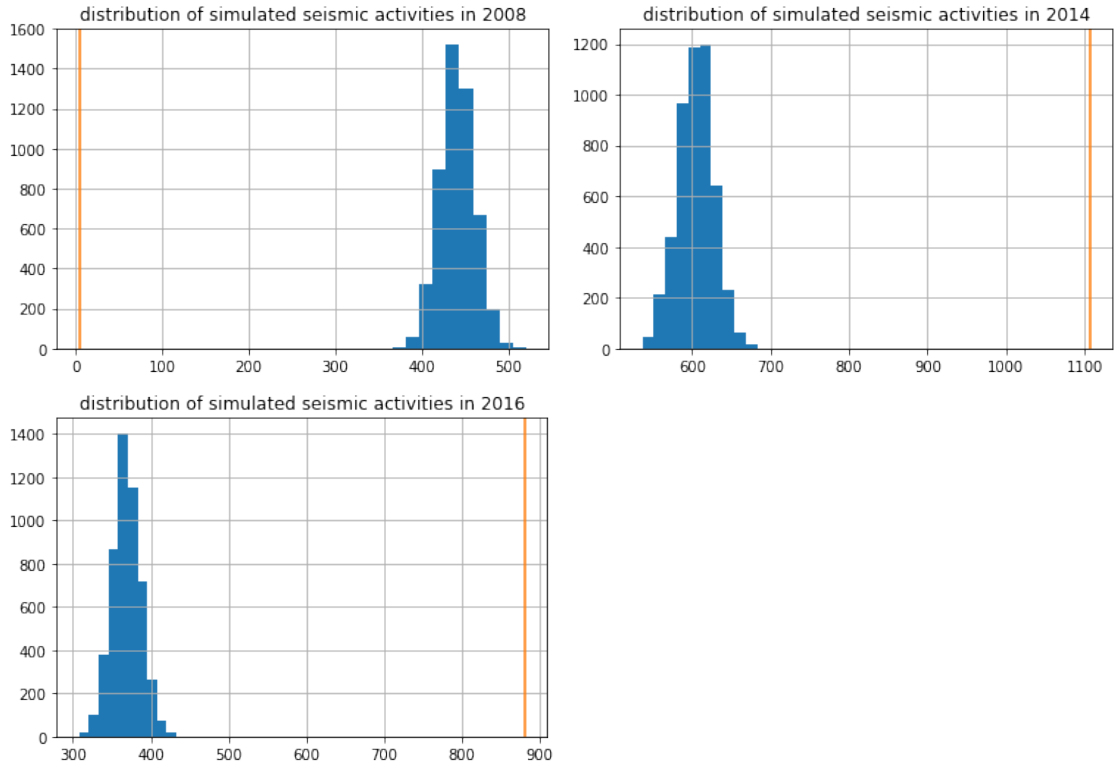
1. Confidence interval

This method is what I learned while taking the course. The basic idea is that assuming the seismic follows the distribution of injection volumes, then we take random simulations for thousands of times and utilize the distribution of injection volumes to get a distribution of estimated seismic activities. If our assumption is true, the real seismic data should be within the confidence interval of the estimated seismic distribution. Otherwise, our assumption is not correct, which indicates the distribution of seismic activities and injection volumes are statistically different.

The following the functions to perform random simulations, n is the total seismic activities between 2008 and 2017, Inj_sum_2_normalized is the percentages of injection volumes between 2008 and 2017.

```
def simulate(n):
  return pd.DataFrame({'year': np.random.choice([2008,2009,2010,2011,2012,2013,2014,2015,2016,2017],
size=n, p=Inj_sum_2_normlized)})
```

For each random test, I can obtain a group of seismic data between 2008 and 2017. I did 5000 tests in total and obtained 5000 data sets of seismic numbers for each year between 2008 and 2017. Then I plotted the distribution of simulated seismic activities and the real seismic activities for a year. If our assumption is correct, the real seismic number should be within the distribution of the simulated seismic numbers. Here I show the results for three years: 2008, 2014, 2016 for the sake of illustration though I tested for all the years and the conclusion are the same. In these plots, the yellow line indicates the real seismic data and the blue blocks shows the histogram plots of the simulated seismic activities. The results indicate that the real seismic number is far away from the simulated seismic numbers, which indicates that the distribution of seismic number and injection volume is statistically different.



2. T-test

T-test compares two means of the distribution and tells the difference between each other. I utilized scipy built-in function scipy.stats.ttest_ind to compare the distribution of seismic activities and total injection volume by evaluating t-values and p-values. T-values tells how large the difference is between two groups: the larger the t score, the more difference there is. P-values is the probability that the results occurred by chance. To test my hypothesis that the seismic distribution and injection volume distribution are the same, I compared two groups of distributions: one is the simulated distribution of seismic number using 5000 random simulations by utilizing the percentages of injection volumes between 2008 and 2017, which I already obtained in the 1st method. The other data sets I selected is the simulated distribution of seismic number using 5000 random simulations by utilizing the percentages of seismic numbers between 2008 and 2017. Thus, these two data sets have the same sample size.

I tested the t-values and p-values for three years: 2014, 2016, and 2017. In 2014, the t-value is 1131 and p-value is 0. In 2016, the t-value is 984 and p-value is 0. In 2017, the t-value is 98 and the p-value is 0. The extreme large value of t-values and 0 p-values indicate that we are very confident that seismic numbers and injection volumes are statistically different.

4. In-depth Data Analysis

In Capstone project 1, our problem is a regression problem. That is to say we would like to predict the total seismic events based on the total injection volume. In addition, I also predicted the maximum seismic magnitude based on the total injection volume in this analysis.

The main difficulty in this project is that there is limited data for regression analysis. Even though there are totally 8000+ seismic data and 10000+ injection volume data between 2008 and 2017 in Oklahoma, we grouped them into two groups based on the locations. As a result, we have totally 2 groups of data, with 10 data points in each group, illustrating the seismic and injection data in each year. In order to tackle this issue, we utilized the cross-validation method. Cross-validation is a resampling procedure used to evaluate machine learning models based on limited number of data. Among various models of cross validation, leave one out cross validation (LOOCV) is utilized in this study. For the progression model, I tried linear regression and decision tree model.

The code is submitted and the procedure of analysis in coding will not be discussed in detail. For instance, the following is the code to analyze seismic and injection data in region 2 with decision tree and LOOCV.

```
# region 2, injection volume - total seismic number, decision tree
# import library
from sklearn.model_selection import LeaveOneOut
from sklearn.model_selection import cross_val_score
from sklearn.tree import DecisionTreeRegressor

lm = DecisionTreeRegressor()
cv = LeaveOneOut()

# transform ot numpy array
Inj_sum_2_array = np.array(Inj_sum_2)
Seis_count_2_array = np.array(Seis_count_2)

# normalize
norm = np.linalg.norm(Inj_sum_2_array)
Inj_sum_2_array_s = Inj_sum_2_array/norm
norm = np.linalg.norm(Seis_count_2_array)
Seis_count_2_array_s = Seis_count_2_array/norm

# model, cross validation
model = lm.fit(Inj_sum_2_array_s.reshape(-1,1), Seis_count_2_array_s)
scores = cross_val_score(model, Inj_sum_2_array_s.reshape(-1,1), Seis_count_2_array_s,
scoring='neg_mean_absolute_error', cv=cv, n_jobs = -1)

print("Folds" + str(len(scores)) + ", MSE:" + str(np.mean(np.abs(scores))) + ", STD:" + str(np.std(scores)))
```

I collected cross validation scores and calculated the means and standard variations of the score values of the cross validation to evaluate the accuracy of linear regression and decision tree method. The following table illustrates the summary of means and standard variations of scores. The results indicate that the accuracy is

quite low and standard variations is relative high. Unfortunately, linear regression and decision tree method does not provide good accuracy of the data analysis in this case.

			MSE	STD
Region 1	Seis Number vs. Injection volume	Linear Regression	0.162	0.141
		Decision Tree	0.228	0.221
	Seis magnitude vs. Injection volume	Linear Regression	0.142	0.084
		Decision Tree	0.206	0.134
Region 2	Seis Number vs. Injection volume	Linear Regression	0.173	0.126
		Decision Tree	0.153	0.126
	Seis magnitude vs. Injection volume	Linear Regression	0.060	0.036
		Decision Tree	0.072	0.045

5. Summary of Capstone Project 1

Capstone Project 1 is my first project in data science. It is a good learning process for me and step by step, I learned and applied various data analysis tools to analyze real engineering problems. I feel really excited about it.

The main objective is to investigate the correlations between seismic activities and injection activities in Oklahoma states. In this project, I first imported seismic data and injection data in Oklahoma between 2008 and 2017 from the website. Each data set contains about 10000 data points. However, the injection data and seismic data cannot be directly correlated because the injection location is not the same location of seismic events. As a result, the key step of data cleaning in this project is to group the seismic data and injection data based on the location. As a result, I cleaning the data and created two groups of injection and seismic events. After data cleaning, I did story telling by visualizing the data using scatter and histogram plots to get a direct idea of the distribution of data. At this step, I did see some rough correlations between injection and seismic as seismic activities decreases as injection volume decreases especially between 2014 and 2017. However, in order to obtain whether injection and seismic activities are statistically correlated, statistical interference is needed. In the statistical interference analysis, I utilized two methods: confidence interval and T-test. The final conclusion of the interference analysis is that seismic numbers and injection volumes are statistically different. In the final step, I applied two regression models (linear regression and decision tree) to find the correlation between seismic data and injection volume. In addition, cross-validation method is also utilized to tackle the issue of very limited data numbers. The final conclusion of regression analysis is that both linear regression and decision tree model perform poor accuracy. As a conclusion of this project, the seismic activity and injection activity are statistically different and we are not able to accurately predict the seismic activities (seismic number and seismic magnitude) based on the injection volumes.

6. Future Plans

Generally speaking, there are only 20 data points that are utilized for the final machine learning analysis in capstone project 1. As a result, only simple regression models can be utilized. In the capstone project 2, I want to tackle more complicated and challenging problems with more data points, more parameters, and etc.. And I want to explore other machine learning techniques (for example random forest, SVM) to solve different problems (classification or clustering).