Introduction

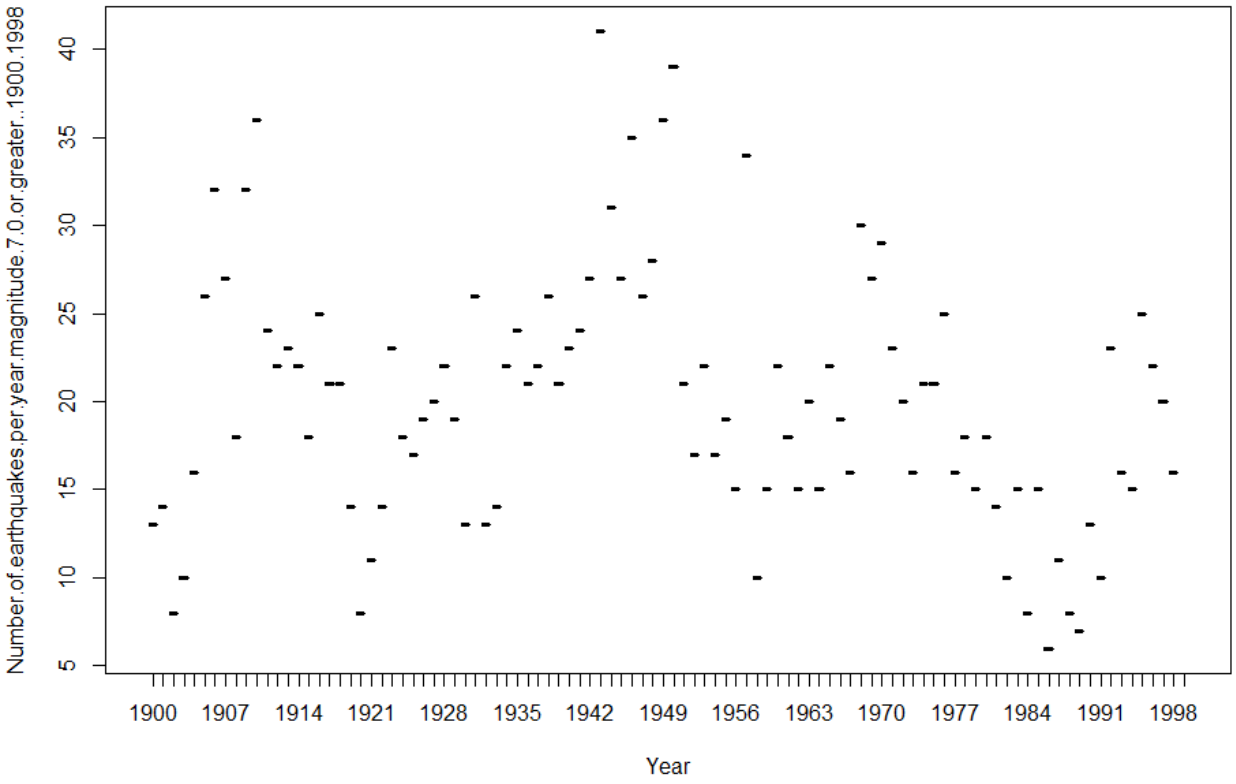
Today I want to use the outlier detection commands and complete an analysis using outlier detection techniques in R. I wanted to specifically use packages and commands to explore the data and then create anomaly detection plots and be able to detect anomalies in data for the analysis about earthquakes, which interested me. I felt it was better for this data to use the packages AnomalyDection, tsoutliers, and h2o to detect anomalies in the data and complete the analysis.

Methods and Results

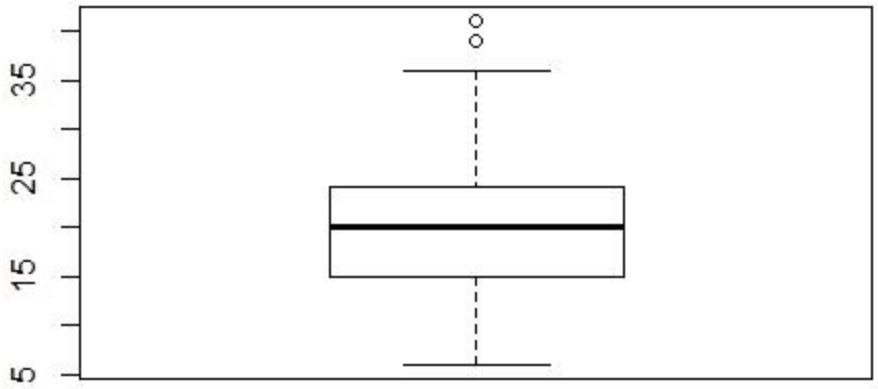
First, we were trying to see the structure of the data and describe what we saw along with some initial inferences. We read in the data first and then must make sure it is in the correct format. Once it is in this format we can complete the analysis and start to look for anomalies. We also want to run a basic anomaly analysis with a boxplot before continuing with the suggested packages for the assignment.

```
> str(nepy)
'data.frame':   100 obs. of  2 variables:
 $ Year                                                : Factor w/ 100 levels "1900","
1901",..: 1 2 3 4 5 6 7 8 9 10 ...
 $ Number.of.earthquakes.per.year.magnitude.7.0.or.greater..1900.1998: int  13 14 8 10 16 26 32 27 1
8 32 ...
```

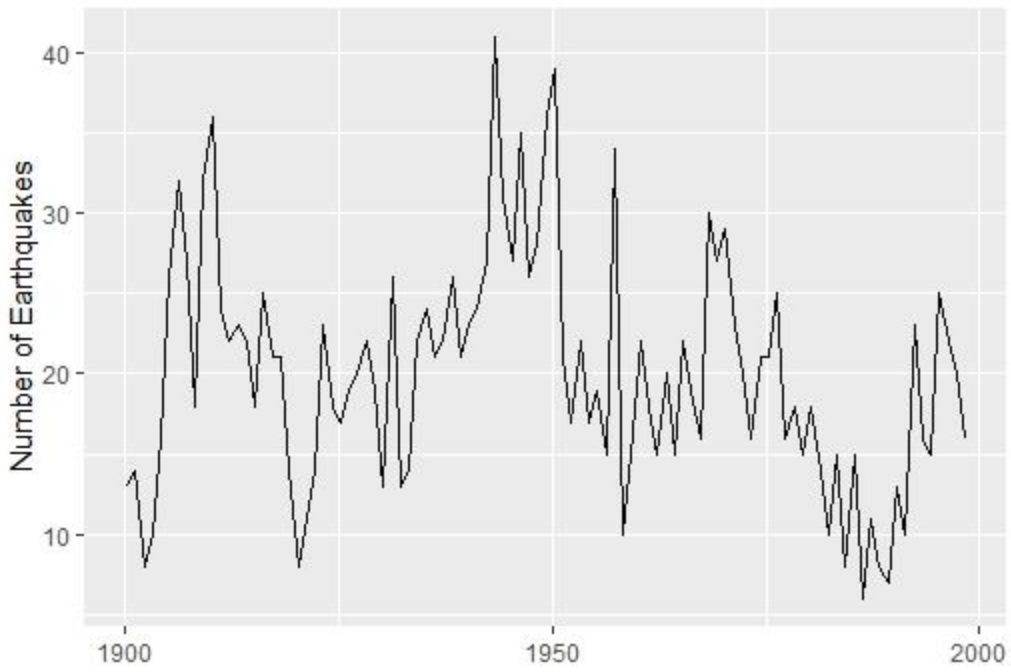
```
> summary(nepy)
      Year    Number.of.earthquakes.per.year.magnitude.7.0.or.greater..1900.1998
 1900   : 1   Min.    : 6.00
 1901   : 1   1st Qu.:15.00
 1902   : 1   Median :20.00
 1903   : 1   Mean   :20.02
 1904   : 1   3rd Qu.:24.00
 1905   : 1   Max.    :41.00
 (Other):94   NA's    :1
```



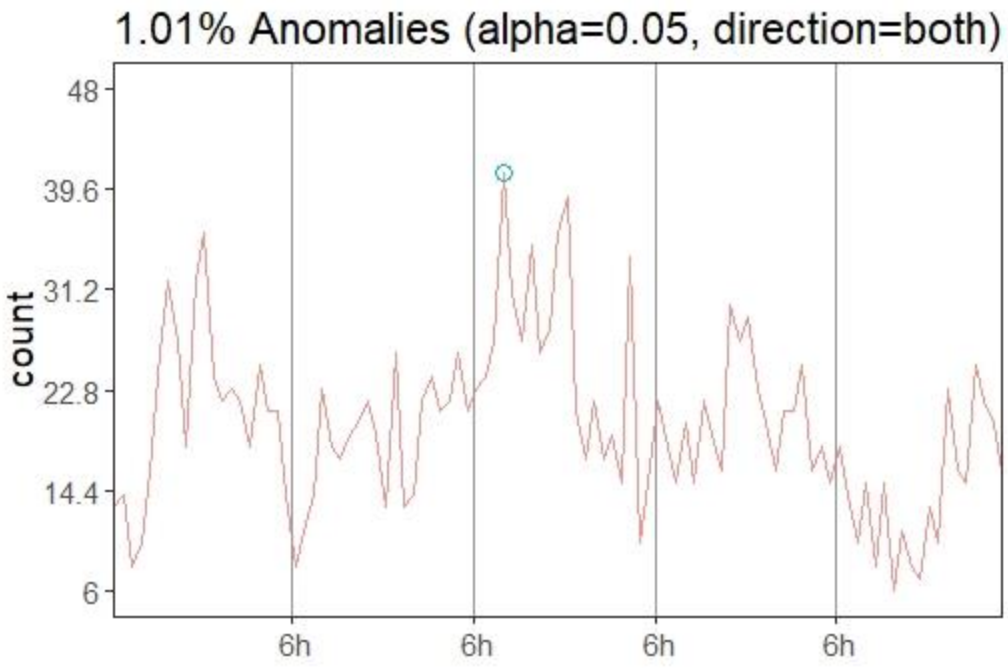
After looking at the data we can see, that the data appears to be very simple data that cumulates over time, one record for each year, such as last weeks data which was a time series. However, this week we are only looking for anomalies, and not trying to predict the future with the data. So, the most basic way to detect anomalies is to use a simple boxplot technique; this will help us to identify outliers. But, since our data is sequential, and outliers are not always an anomaly, we should use caution with our analysis and steps. We can see the boxplot below to give us an idea about the outliers that exist.



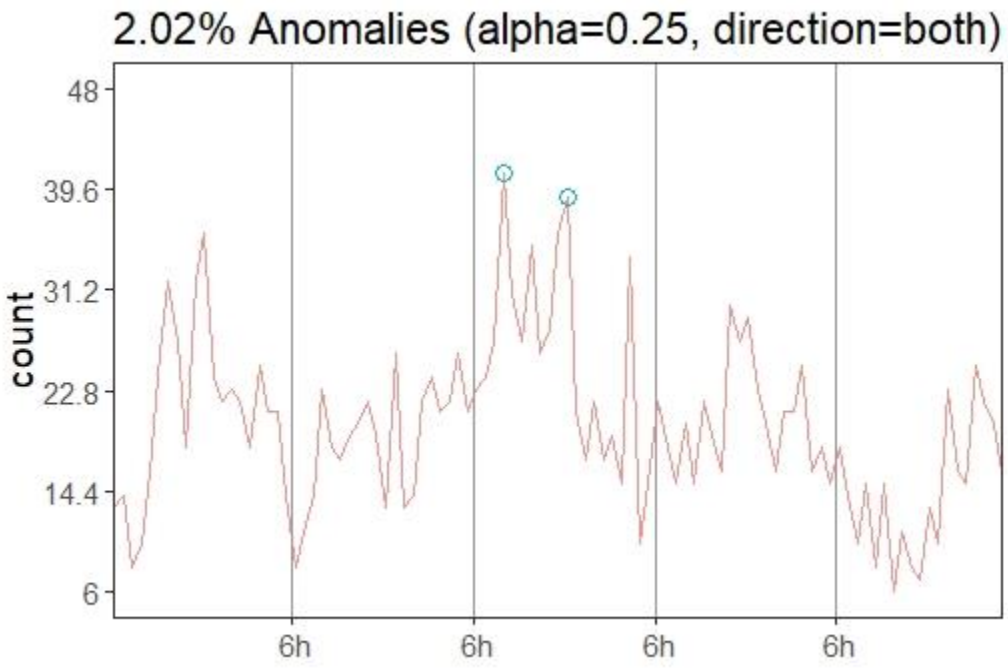
So, above we see that we have two outliers in the data. However, even though our boxplot can show us the true anomalies in the data it does not take into consideration the time series aspect of the data. So, I wanted to start to look at the suggested commands in order to start informing my decisions for the assignment. What we can do next is use a package designed to identify anomalies for time series data. I then loaded the package and transformed the data to meet the needs of the package for analysis. The Year field needed to be converted to a date and I also uploaded the ggplot2 package to show plots of the data.



As we can see above the plot of the new data came out well from date from. We then want to use the AnomalyDetection package that we downloaded to look at the actual anomalies in the data. One of a couple of the commands imbedded in this package are, max.anoms, which is the argument that determines how many maximum anomalies there could be. The alpha argument determines a level of significance for qualifying anomalies. We may have to try a few of these out in different scenarios to see how the command words. The package requires is a period of more than 1 year in order to function. We will look at data in increments of 5 years in order to get results using this method of analysis.

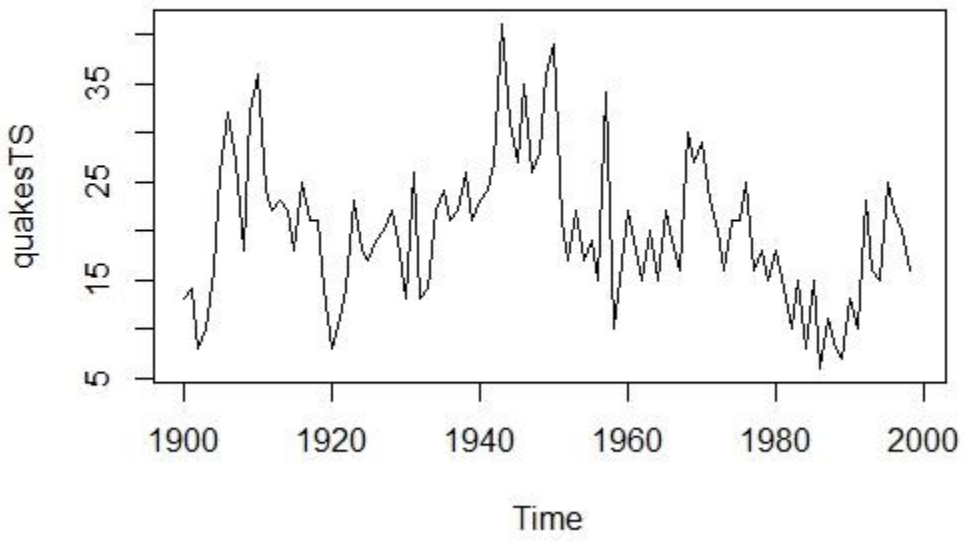


Here we interestingly only find one anomaly. This is less than the two we found when we looked at the box plot. So, what we can do is lower the alpha in order to see if we can detect the anomalies better with a lower threshold.

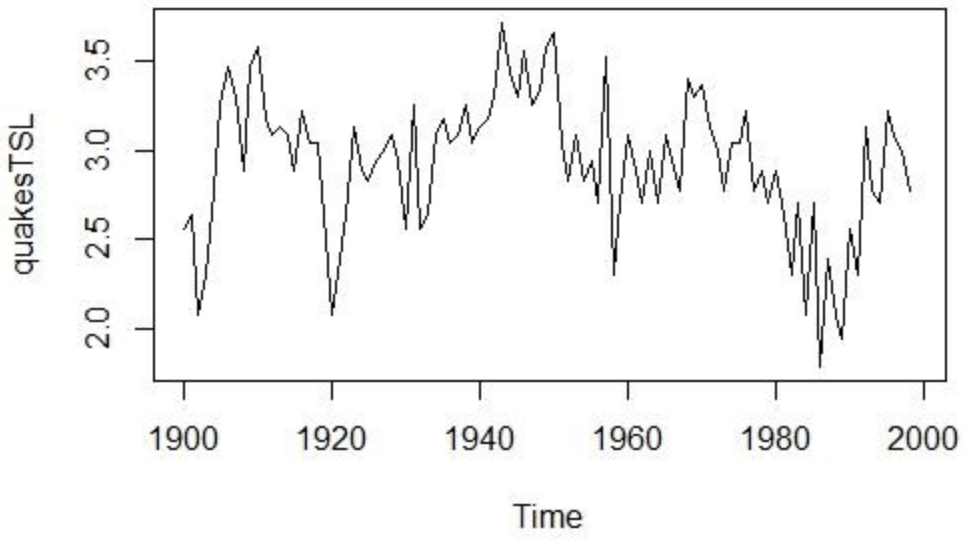


At .25 we get a second anomaly that matches what we found in our box plot. However, its probably not significant at that level either. The problem that we start to see with using this package is that we cannot see the dates at the bottom which would help our analysis. This is not something that hurt the validity or functioning of the package, but, is not very helpful during a quick glance in the analysis phase. It also, feels like the argument are very sensitive in the package when changing the values in them. For example, if we specified everything that matched our data set, we would probably get errors from the system.

However, to this point, we know that the year with approximately 40 earthquakes or so is probably an anomaly. Using the boxplot from earlier, and when we lower the significance, the year with slightly under 40 earthquakes also seems like it may be an anomaly. These two points in time are fairly close neighbors so they may not be anomalies but could be outliers. To investigate more, we can try a few other methods that were suggested to us in the assignment. We can look at the package "tsoutliers" to see if it helps with our analysis any. However, this package requires time series objects, so we transformed our data again and plotted the data as such.



We can then see above that we plotted the time series object we created from the data to plot with the auto ARIMA function in R.



Above we then plot the log of the points of the original set to help with the distances between the points in the data. After creating the plots above we see another possible point of interest and decide to check it out with the ARIMA analysis of time series. We can then see that we have a third point in the data right after 1960 that in fact does skew the data and the plots. This, by definition, means that this is in fact an outlier that's classified as an innovative outlier because of the skew it causes.

```
Series: quakesTS
Regression with ARIMA(1,0,0) errors

Coefficients:
         ar1  intercept     IO44     AO58
      0.5849    19.2748  17.1340  20.6306
s.e.  0.0827     1.2722   5.4121   4.5964

sigma^2 estimated as 29.27:  log likelihood=-305.79
AIC=621.57   AICc=622.22   BIC=634.55

Outliers:
  type ind time coefhat tstat
1   IO  44 1943   17.13 3.166
2   AO  58 1957   20.63 4.488
```

We then look at a few more points and analysis below however, we only find the one other outlier after 1960 in the data according to the below outputs.

```
Series: quakesTS
Regression with ARIMA(1,0,0) errors

Coefficients:
         ar1  intercept     IO44     AO58
      0.5849    19.2748  17.1340  20.6306
s.e.  0.0827     1.2722   5.4121   4.5964

sigma^2 estimated as 29.27:  log likelihood=-305.79
AIC=621.57   AICc=622.22   BIC=634.55

Outliers:
  type ind time coefhat tstat
1   IO  44 1943   17.13 3.166
2   AO  58 1957   20.63 4.488
```

```
Series: quakesTS
Regression with ARIMA(1,0,0) errors

Coefficients:
         ar1  intercept     IO44     AO58
      0.5853    19.2772  17.1176  20.6321
s.e.  0.0826     1.2733   5.4111   4.5958

sigma^2 estimated as 29.28:  log likelihood=-305.79
AIC=621.58   AICc=622.23   BIC=634.56

Outliers:
  type ind time coefhat tstat
1   IO  44 1943   17.12 3.163
2   AO  58 1957   20.63 4.489
```

```
Series: quakesTSL
Regression with ARIMA(2,0,0) errors

Coefficients:
         ar1     ar2  intercept    AO58
      0.4666  0.2367     2.9050  0.8743
s.e.  0.0985  0.0985     0.0968  0.2654

sigma^2 estimated as 0.09017:  log likelihood=-19.62
AIC=49.25   AICc=49.89   BIC=62.23

Outliers:
  type ind time coefhat tstat
1   AO  58 1957  0.8743 3.294
```

We then look to do the h2o package as suggested in order to see if any changes happen to the data or if we find any other anomalies. In order to use this package, we need to convert the data to factors due to it being a deep learning algorithm. So, we make the years factors. We reduce the dataset to the data that functions as well as we set the seed for reproducibility. We then create the training data sets needed and use the autoencoder function in order to help us with this piece of the model. All of this was done after starting h2o with the command localH2O = h2o.init(). We then used the h2o.anomaly function to reconstruct the original data set using the reduced set of features and calculate a means squared error between both. Here we set per feature parameter to FALSE in the h2o.anomaly function as we want a reconstruction mean error based on observations, not individual features. After looking at the plot we saw that there were only the two outliers identified, and not the 1960-year anomaly like in the last package used.

	Year <fctr>	Earthquakes <int>
44	1943	41
51	1950	39

2 rows

Analysis and Interpretations

From the charts and packages that were used, it is probable that 1943 was a year that is likely an anomaly in terms of the number of earthquakes that were seen. Also, 1950 is probably an anomaly year for earthquakes as well according to the data and the plots. 1957 could also be considered an anomaly, but has little impact on the data, as it spikes rather than misleads a trend or skews the data in any way.

Finally, 1960 may also be an anomaly year however, there was not enough evidence to truly prove this in the plots of more than one package.

Conclusion

This analysis helped facilitate the transformation of the data, cleaning, analysis, and understanding and interpretations of the different packages and commands in R for deep learning and outliers. This would be a great way to look at data from multiple angles and decide what data should and should not be in the dataset. It seemed like a great way to have proof and back up that the data was truly anomalous or if it was a legitimate increase or decrease in what you are seeing.

References:

1. Han, J. Kamber, M. and Pei, J (2012). Data Mining: Concepts and Techniques. UIUC. Retrieved from: http://hanj.cs.illinois.edu/bk3/bk3_slidesindex.htm
2. R help at get hub. Available at: http://amunategui.github.io/anomaly-detection-h2o/