Abstract (200 words):

Problem Statement (250 words):

Earthquakes are one of the most dangerous phenomena which strike unexpectedly. It is a result of releasing tension at faults where slowly moving tectonic plates get stuck. As Earth’s lithosphere is made up of multiple plates, earthquakes occur frequently worldwide which can cause severe disruptions to services and more destructive impacts such as tsunamis wreaking havoc amongst cities like the Boxing Day Tsunami (reference), ultimately leading to loss of life.

Hence much geological research is focused here (reference) for example proposals on constructing ‘earthquake-proof’ buildings (reference) and developments of deep-learning models for earthquake detection (nature).

We aim to answer the following:

1. Is there a detectable seasonality in earthquakes or peculiar years in terms of magnitude?
2. Can we identify the most susceptible countries, and identify countries expected to receive increased impact if earthquakes exhibit a temporal increase?
3. Can we develop a reliable time series model to predict the expected number of earthquakes in following years?

This analysis will prove useful in possibly identifying earthquake patterns and suggesting to planners which countries require more protection to minimize future damages.

The dataset available (reference) is suitable as it contains spatial-temporal data on earthquakes spanning from 1965 to 2016 where temporal patterns can be investigated under varying granularities. Multiple datasets will be combined such as plate boundary locations, allowing for spatial assessments of earthquakes along these. Whilst factors such as underlying soil condition (reference) are unavailable, our data should nonetheless suffice and combined with over 20,000 observations, provide a detailed analysis.

250 WORDS!!

‘Planners?’

State of the Art: (500 words)

Earthquakes have always been a major concern due to its impacts and therefore has various subtopics of interest, for example, developments of earthquake monitoring software (reference) and resistant structural engineering (reference). Combined with the rise in visual analytics approaches (reference) (reference), it is now easier to efficiently investigate earthquakes and their consequences.

Azis et al. (reference) worked with identical data and used time series plots assessing how numbers of earthquakes varied annually, aggregated by plate boundaries and also by time zone. To predict the number of earthquakes in following years, 152 models were developed; for each of the 24 time zones, there were two time series, a regular time series and its stationary counterpart, similarly for the 52 plate boundaries. A combination of Linear Regression and LSTM Models were deployed, where both models showed varying performance depending on time zone and plate. Predictive results were primarily considered by assessing R2 values. It was found that stationary models performed better, and hence suggested that further research should look at implementations of stationary time series modelling techniques.

Hence, we plan to implement ARIMA modelling to predict the number of earthquakes in following years and furthermore, use diagnostic visualisations to evaluate model effectiveness which the author did not carry out, and also consider time granularity.

Yang et al. (reference) explored spatial-temporal properties of earthquakes between 1960 and 2014 using a spatial-temporal scanning technique to identify two kinds of clusters, burst and persistent. This was carried out by modelling the whole space as a spatial-temporal cube, detecting the type of clusters using cylinders, expanding their base and height until a threshold was met. The relative risk of each cluster was calculated. Space-Time Plots were used to visualize clusters as well as two-dimensional spatial plots examining how clusters varied along plate boundaries.

Our spatial methodology will differ here slightly, as we will be using instead OPTICS and DBSCAN Clustering due to their abilities to adapt with geo-spatial data (reference) (reference). This will be applied on spatial data to eliminate noise and locate clusters of dangerous at-sea earthquake hotspots since most occur off-shore. This report will aid our research as we will similarly be utilizing two-dimensional spatial plots to help visualize clusters.

Battul et al. (reference) explored spatial-temporal techniques using similar data, for earthquakes between 1912 to 2009 in India. Feature engineering was carried out by binning features such as magnitude, and visualisations such as bar charts and histograms were used to assess distributions of features, as well as spatial visualisations assessing the distribution of earthquakes, coloured by binned groups. A linear regression model was fitted examining how numerical features related to earthquake size, visualised by plotting comparisons of true values against predicted.

Whilst some histograms had yielded some conclusions, the author did not consider temporal changes, this would be more meaningful to assess which is what we’ll implement in combination with binning to improve insight. We will also build on this by examining how earthquakes vary given plate boundary locations and examine seasonality via heatmaps.

500 Words!

Data (500 words)

The dataset was extracted from Kaggle (REFERENCE) containing 23,230 earthquakes between 1965 and 2016, hence a sufficient timescale to identify patterns. It consisted of earthquakes with magnitudes over 5.5 making up 21 features; ‘main’ features included Date, Time, Latitude, Longitude, Depth and Magnitude. Since magnitudes of below 5.5 were not included, we implicitly assumed each earthquake was distinct rather than grouped by foreshocks, mainshocks and aftershocks which would be unrepresentative. Spatial features spanned worldwide; in terms of precision, Latitude and Longitude were given to 3 decimal places, sufficient for earthquakes accounting for random error, with Time given down to the second. Remaining features were geology specific for example Earthquake ID and Magnitude Source which were irrelevant, as well as Azimuthal Gap etc. which contained 30% to 99% missing values. Given these proportions and the fact that imputing these would yield unfair analysis, these were dropped. No duplicates appeared.

Missing value analysis was carried out; only 3 missing values were found for Magnitude Type (the way magnitude was measured), these were imputed using the mode since it made up a miniscule amount. Earthquake magnitudes were measured via different metrics for example Moment Magnitude, Richter Magnitude etc., these measurements are valid for certain frequency and distance ranges; the range of validity for these all lie on the same scale. Where possible, magnitudes were converted to the uniformly applicable Moment Magnitude (reference). Where conversions were not possible, we implicitly assumed their magnitudes were measured using the Moment Magnitude since previously converted magnitudes did not deviate much from their original values.

Outlier analysis was carried out using histograms and boxplots with the latter acclaimed at its ability to detect outliers (reference). Only Depth exhibited these across the 50 years. Earthquakes with negative depths had their depths set to zero, which were close to 0 anyways. 300 earthquakes were identified with larger than usual depths - we decided to not drop these – we interpreted these as rare earthquakes in which we cannot know when or how often they occur, attributed to the Black Swan Principle (reference).

Extensive feature engineering was implemented; geocoding was employed to retrieve country names based on latitude and longitude, and subsequently continent names. Around 50% of earthquakes occurred at sea, hence their Country and Continent label was ‘Undefined’. Magnitude and Depth were grouped to create two new features using MTU’s Magnitude Groupings (reference) and the USGS Depth Groupings (reference) respectively. The Time feature was kept but broken down into Year and Month allowing for analysis over different granularities.

K-Means Clustering was applied using a large number of clusters (40) not for the purpose of clustering, but to create many small groupings of points. By assessing the spatial distributions of each cluster, two new features were created – the Plate Boundary these earthquakes lie on and the Type of Plate Boundary (Convergent, Divergent and Transform). K-Means surprisingly performed well by producing many tight clusters lying precisely on plate boundaries, given that latitude and longitude were used with the algorithm implementing Euclidean distances. (reference)

500 Words!

(Figure of KDE? – if so need to convert to histogram with side-by-side boxplot) – after outlier

(Table of all vars. w/ description of each) at end

Analysis Process (500 words)

We now discuss the analysis approach taken to answer our questions, highlighting how human reasoning with visual displays are repeatedly utilized together to aide our analysis. (Fig X) summarizes our approach.

Our approach to pre-processing was explained earlier, human reasoning was essential in understanding visualisations like histograms and determining how to analyse outliers in an earthquake context. It was used to engineer a further meaningful dataset through determining suitable numbers of clusters for K-Means and assignment of plate boundaries. It was required to rectify errors during geocoding ensuring correct country assignments, identify suitable magnitude transformations, to make assumptions and ensure a suitable final dataset.

We now discuss our analysis in further detail.

To expose possible seasonality and peculiarities, heatmaps will be implemented, human reasoning will be used by examining how magnitudes differ across varying combinations of years and months aiding in seasonality identification. Given the high number of years, time-series plots by both year and month will also be used to help reasoning. Tweaking colour schemes of the heatmap also will assist reasoning in identifying peculiar periods by highlighting these in dark colours. Where peculiar periods occur, deeper investigation will be performed providing insight on responsible plate boundaries for increased magnitudes via bar charts aggregated by magnitude group. Spatial visualizations will be utilized, examining most hit countries during these periods, attempting to hypothesize the reasoning behind these using domain knowledge.

Tree-maps will be utilized providing brief ideas of most susceptible countries. Human judgement will be used identifying whether fair conclusions from this can be derived. Given over 50% of earthquakes occur offshore, density-based clustering methods will be implemented using default and tuned parameters (more detail later) to identify spatial clusters at sea visualised using a map. Human judgement will be used identifying oceanic hotspots, further considering how depth and magnitude vary in these regions and considering proximity to countries, in combination with the tree-map helping identify overall vulnerable countries.

In examining countries expected to receive increased impact, temporal visualisations will be used assessing whether increasing trends appear, aggregated by continent. Where trends appear, analysis showing changes in earthquakes numbers on surrounding plate boundaries will be explored, including changes in magnitude and depth assessing risk. Spatial visualisations of earthquakes surrounding worsening plate boundaries will be examined using judgement to identify at-risk countries in the future.

Human judgement will be used in deciding the time granularity to predict the future number of earthquakes on; an initial time series plot complimented with its ACF and PACF plots will be made where judgement is required identifying whether differencing is needed. Once differenced, if required, human judgement is needed to determine ARIMA model coefficients based on the (new) ACF and PACF visualizations whereby a model is fitted. Assessment of the model coefficients significances, results of the Ljung-Box and Jarque-Bera Test and examination of residual diagnostics will be done via human judgement. Using the final model, predictions will be done, to be compared to true values whereby conclusions will be made on quality and predictive power.

500 words!!

(Create flow chart after first paragraph in terms of … via previous paper) + task subheadings

Analysis (1500 words)

Heatmap analysis in figure (INSERT) showed average magnitude did not yield trends by month, year or a combination, supported also by time series examinations by year and month even when accounting for magnitude and depth, suggesting absent seasonality supporting expectations that earthquakes unexpectedly strike.

However, the heatmap suggested peculiarities during August 1965, December 1966 and February 1969 exhibiting unusually high average magnitudes.

We found the Australian-Pacific boundary primarily responsible here. Analysis showed annually, roughly 65-70% of earthquakes magnitudes are Band 0. With 53% of earthquakes in Bands 1 and 2 with remaining in Band 0 during this period, this explained increases in average magnitude. Highest magnitude earthquakes were concentrated on the Australian-Pacific boundary, with depth mostly shallow. Combinations of high magnitude and shallow earthquakes imply high chances of damages to land in close proximity. Spatial analysis showed much of these earthquakes were surrounding Vanuatu which borders the boundary – correlating with literature mentioning Vanuatu being hit with a series of destructive earthquakes then (reference).

Similar conclusions yielded compared to August 1965 with most earthquakes focused on the Australian-Pacific plate, including the highest magnitude ones, although a shallow Band 2 earthquake occurred on the South American – Nazca boundary. Most earthquakes were shallow and spatial analysis showed these surrounded Papua New Guinea and Vanuatu, as well as a destructive earthquake on the Chilean coast causing $400,000 dollars’ worth of damages (1966 rate) (reference\*\*).

February 1969 was interesting because whilst there was dominance of higher band magnitude earthquakes like before, with much of the earthquakes focused on the Philippine-Eurasian plate boundary, what was interesting was that strongest earthquakes still occurred on Australian plate boundaries, e.g., Australian-Eurasian and Australian-Pacific, with Band 0 earthquakes occurring along Philippine-Eurasian plate and elsewhere. In contrast, earthquakes that occurred along Australian plates were deep depth-wise hence not deemed as destructive due to seismic waves losing energy travelling far to the surface (reference). Spatial visualisation showed the Philippines and parts of Indonesia primarily affected here.

In conclusion we observe the source of peculiarity narrowed down to high magnitude earthquakes surrounding the Australian plate in the 60s, where boundaries are convergent, primarily affecting Vanuatu and Papua New Guinea. Since the 60s, earthquake magnitudes surrounding Australian plates have dropped and stabilised to reasonable levels via time series visualisations explaining peculiarities in this period, although no domain literature found explaining why. With insufficient data, we cannot conclude why these boundaries had such high numbers however literature suggested the Australian plate is the fastest moving plate (reference), hence we could hypothesize that the plate was moving at its fastest in the 60s and due its speed, huge build-ups of stress were created where the plate got stuck. Once weaker areas of crust slipped, releases of tension created a series of dangerous earthquakes affecting surrounding countries compared to slower moving plates where stress build-up is naturally weaker. We may further hypothesize that since the 60s, this plate speed has slowed down(485).

Fig X shows a tree-map showing the most susceptible countries; however, this is unrepresentative. Over 50% of earthquakes occurred offshore, with spatial visualisations showing many earthquakes occurring close to countries where geocoding failed at assigning countries. Spatial visualisations were highly congested, so DBSCAN and OPTICS clustering were employed to identify oceanic hotspots due to flexibilities working with varying distance metrics (reference, reference). We will be working with offshore earthquakes here.

Spatial coordinates were given via latitude and longitude, hence the Haversine metric was used in calculating great-circle distances between earthquakes assuming Earth was spherical to develop a distance matrix, first converting spatial data into radians.

A default DBSCAN algorithm was initially implemented – yielding poor results, essentially all points allocated to a single cluster, likely attributed to the epsilon parameter in scikit-learn being unsuitable; DBSCAN results are highly sensitive to epsilon (reference) with large values yielding large clusters with less noise and vice versa. Attempts at tuning both the epsilon and minimum number of samples parameters were implemented using heuristic approaches (reference)(reference). This was our best DBSCAN outcome - clusters resembled plate boundaries, shown in Fig X. Whilst helpful in showing most earthquakes occurred on boundaries, this was still too noisy.

We expected OPTICS would improve results given independence of epsilon and consideration of clusters having local densities (opticspaper). Using the same minimum number of samples as the ‘tuned’ DBSCAN model, results are shown in Fig X, showing improved results with noisier observations such as those far from countries removed, hence easier to identify susceptible countries. Many countries in (south) eastern Asia such as Japan, Indonesia, Papua New Guinea and Vanuatu, and most lying on the west of South America such as Chile and Mexico were found to be highly susceptible alongside New Zealand. Most countries within the Greater Antilles are susceptible and also those around the Mediterranean and Arabian Sea, all of which noticeably lie close to plate boundaries; highlighting why utilizing tree-maps alone were unrepresentative; for example, Japan, one of the most susceptible countries(reference), was not reflected in Fig X as nearly all earthquakes happened offshore. By combining the tree-map with above clusters, it was easily identifiable what the most susceptible countries are, namely Japan, Indonesia, Papua New Guinea, Chile, Vanuatu and Philippines in which the latter 5 dominated the tree-map whilst exhibiting oceanic hotspots.

Time series plots were used assessing changes in numbers of earthquakes annually by continent. All continents showed stability except ‘Undefined’, exhibiting significant increases i.e., rises in offshore earthquakes, prompting temporal examinations of changes in earthquakes aggregated by plate boundary only applied to ‘Undefined’, identifying responsible boundaries. Most boundaries remained stable, except 5 (Australian-Pacific, Caribbean-Cocos, Nazca-Pacific, Antarctic-Pacific and South America-Nazca) exhibiting slow increases which collectively gave significant increases overall. An investigation into risk was done. In all mentioned plates, increases in earthquakes were attributed to rises in Band 0 earthquakes, which were all shallow hence low risk consequences for example light ground shaking. Higher bands of earthquakes remained stable on most plates hence risk levels only marginally increasing for surrounding countries, although the Australian-Pacific and Caribbean-Cocos boundaries are experiencing gradual increases in Band 1 earthquakes alongside, though the depths of these are intermediate so risk levels are also marginal increased.

Whilst risk levels to countries surrounding these boundaries seem unchanged, these countries should expect further earthquakes in the upcoming years primarily the west coast of South America, countries in the Greater Antilles, and those bordering the Australian-Pacific plate such as New Zealand, Vanuatu and Fiji(580).

Building upon (references) paper, we now develop ARIMA models to predict the number of earthquakes. The granularity of time was selected to be months, the reasoning being because this struck a good balance between years and days. Prediction based on year means we only have 50 data points which is insufficient, and the use of days is far to granular with the time series being far to erratic or random.

Fig X shows the series we are dealing with as well as its respective ACF plot. It is clear an increasing trend is visible meaning the series is not stationary combined with the fact that slow decay occurs with points not staying within bounds on the ACF plot at high lags. The trend looks roughly linear or rough a shallow quadratic trend suggesting we should try both first and second differences and assess the difference between results.

Both the first and second differenced series look much more stationary by consideration of the raw differenced series albeit some large spikes in the latter, as well in ACF plots although this is clearly not perfect with multiple points on this plot reaching outside of bounds as higher lags. No pattern was determined with these out of bounds spikes, indicating the absence of seasonality.

Considering the first differenced series, one significant spike is shown at lag 1 on the ACF plot, indicating an MA(1) model supported by the exponential decay in the PACF plot. Fitting this model returned coefficients all of which were statistically significant through assessment of p-values or by the general rule of thumb: |coefficient/standard error| > 2 (reference). Whilst the model passed the Ljung-Box Test indicating the model does not show of fit i.e., the autocorrelations of the residuals are not statistically different from zero which is good we expect model residuals to resemble white noise which has this property, the model fails the Jarque-Bera Test indicating model residuals do not follow a Gaussian Distribution. Examination of the Kernel Density estimate showed a slight skew supporting this.

The final model, predictions will be done, to be compared to true values whereby conclusions will be made on quality and predictive power.

435 Words

FIGURES – Ljung Box? KDES?

\*\* Vanuatu by magnitude

+ task subheadings, I’ve made subheading for heatmap!

Subheadings for clustering techniques

Double clustering plot – mention without noise

\*\*\* both clustering outcomes

10 FIGURE LIMIT OVERALL WITH LIMITS ON EACH SECTION – SECTION LIMITS ADD UP TO 12

References found via wiki

Consideration of double plots in a single figure