Väinö Kauppila

Research Question: A research on predicting a tsunami's wave height from the magnitude of a preceding earthquake

1. **Introduction**
2. **Background information**
    a. Hypothesis: the relationship between these two variables is linearly correlated
    b. What is linear regression?
    c. Correlation is not necessarily causation
3. **Statistical approach**
    a. The dataset used
        i. Tsunami dataset (NGDC/WDS Global Historical Tsunami Database), includes earthquake information linked to each tsunami
    b. Processing the datasets for use
        i. Making the data usable for the experiment (Python libraries such as pandas can be used to clean up dataset quickly)
            1. Converting the provided compressed (7zip) dataset .KMZ file to a human-readable standard .CSV file
            2. Cleaning of uncertain data points (those recorded before introduction of the WWSSN, those without a recorded water height…), as stated by the NOAA [1]
    c. The experimental procedure; the type of statistical model used
        i. Establishing the independent and dependent variable (earthquake magnitude and tsunami wave height)
        ii. Fitting a model to the dataset
4. **Model results**
    a. Tabular and graphical presentation of results, relationship between earthquake magnitude and tsunami wave height
    b. Analysis of results
5. **Conclusion**
6. **Evaluation of the experimental method**
    a. Strengths and weaknesses of the method used and of the general process (dataset could be lacking)
7. **Further research opportunities**
    a. Mention possible use of machine learning for this type of statistical model [2]
8. **Works Cited**
9. **Appendix**

[1] https://www.ngdc.noaa.gov/hazard/tsunami-db-intro.html#uncertainty
[2] https://github.com/kinnounko/poseidon