Disaster at St. Himark!  
Project I: Crowdsourcing for Situational Awareness

Isaac McPadden

Data Science Capstone Final Project

# About St. Himark

St. Himark is a large city with a population of 246,839 located on a peninsula. Residents can access the mainland via one of 5 bridges and 1 highway. The city is divided into 19 neighborhoods (interchangeably referred to as locations) each with their own unique characteristics. A map of the city can be seen in Figure 1. The city has 8 hospitals, 34 fire stations with fully trained and equipped emergency response teams, and is powered by renewable energy sources and the Always Safe Nuclear Power Plant. The only downside to St. Himark is it is prone to earthquakes.

## The Rumble App

In anticipation of future earthquakes, city management has recently released the Rumble app to crowdsource damage reporting in the event of an earthquake. Damage reports are recorded by time and which neighborhood the damage is observed in. Citizens can use the app to report shaking intensity of an ongoing or recent earthquake and damage observations in up to 5 categories, each on a scale of 0 to 10. The damage categories are sewer and water systems, power systems, roads and bridges, medical facilities, and buildings. Reports are collected and recorded in 5 minute intervals and in the event of a service outage, all data reports recorded by users during the outage are included in the first 5 minute interval after the outage is over.

Map

Description automatically generated

Figure 1. Map of St. Himark and its neighborhoods

## The Earthquakes

On April 6th, St. Himark experienced a minor earthquake followed by a major earthquake on April 8th. A map of the measured shake intensities for the April 6th and 8th earthquakes can be seen in Figure 2 and Figure 3. Data from the Rumble app was collected from 12:00 AM on April 6th to 12:00 AM on April 11th.

Map

Description automatically generated

Figure 2. Map of Measured Shake Intensity of Minor Earthquake April 6th

Map

Description automatically generated

Figure 3. Map of Measured Shake Intensity of Major Earthquake April 8th

# Goals

The primary goal of the data from the Rumble app is to prioritize which neighborhoods need emergency responders the most as of April 11th. Secondary goals are to determine if any areas are providing unreliable information, how emergency response needs change with time, and if reliability of data changes with time. With 34 fire departments with emergency response teams, the primary goal will be accomplished by determining the 34 neighborhoods and damage types that require the most urgent response.

# Exploratory Data Analysis

The initial exploratory data analysis was performed using Pandas Profiling. The full report can be found in “report.html” included with this document. General summary statistics were unremarkable, largely because they did not account for location. Interactions between damage types were weakly linear which is not surprising as you would expect to see something like a heavily damaged building cause damage to a nearby road. Interactions between reported shake intensity and reported damage were weakly logarithmic. An example chart of interactions between damage types can be seen in Figure 4. That also makes sense as a small increase in shake intensity can have major effects on structural damage. An example chart of interactions between shake intensity and reported damage can be seen in Figure 5.

A picture containing bubble chart

Description automatically generated

Figure 4. Typical graph of interactions between damage types. The data show a weakly linear trend.

A picture containing chart

Description automatically generated

Figure 5. Typical graph of interactions between shake intensity and damage types. The data show a weakly logarithmic trend.

Missing values (NaNs) mostly occur in two columns of data: the medical facilities damage column, and the shake intensity column. These can be seen in Figure 6. The medical facilities column is expected to have NaNs because there are only 8 hospitals in the city, but 19 neighborhoods and the data are collected by neighborhood and time. NaNs in shake intensity require further analysis. There are some NaNs in sewer and water systems and buildings. That occurs in location 7 which is mostly undeveloped forest.

Chart, bar chart

Description automatically generated

Figure 6. Distribution of NaNs in the data.

Correlation between variables can be seen in the heatmap in Figure 7. Damage reports are positively correlated with each other as expected. Location has a negative correlation with damage reports but that is random chance as the numbering of the neighborhoods is not logical or in a particular order. Shake intensity is somewhat uncorrelated with damage reports but is correlated with medical facilities damage reports. This is likely due to the NaNs in both columns.

A picture containing chart

Description automatically generated

Figure 7. Correlation Heatmap for variables in the collected data

# Continued Analysis

## Cubic Splines

A quick explanation of cubic splines is needed for further analysis. Cubic splines are piecewise polynomial fits to data points that are smooth and continuous and are smooth and continuous in the first and second derivatives of the piecewise function. They are very useful in data visualization for interpolating a smooth curve along a time series scatterplot without overfitting the curve (source: <https://pythonnumericalmethods.berkeley.edu/notebooks/chapter17.03-Cubic-Spline-Interpolation.html>). In order to reduce noise in the data and create spline curves that make sense, the damage reports were aggregated into 1-hour intervals from the original 5-minute intervals.

## Shake Intensity NaNs Analysis

The first use of cubic spline fitting is analyzing the shake intensity NaNs. The NaNs in the shake intensity column were counted and aggregated by hour and location. The full set of graphs can be found in Appendix I: NaN Counts in Shake Intensity Column. In Figure 8, two representative graphs can be seen. The gaps are intentionally produced in the spline setup to occur in places where multiple consecutive hours of data are not observed. In the case of these graphs, that means no NaNs were observed between hours 58 and 60 at location 1 and between hours 58 and 73 and 86 and 108 at location 3. At location 3 there is a spike in NaN count at around hour 15. Otherwise, NaNs appear to fluctuate in count at a low level. We can use the collected shake intensity reports to determine what is happening.Graphical user interface, application, table

Description automatically generated

Figure 8. Representative count distributions of NaNs by hour and location

In Figure 9, there is a spike in total counts at the same time as the spike in NaNs, suggesting it is just a proportional increase in NaNs. In Figure 10, the shake intensity reports get frequent and consistent. Clearly, this is the time on April 6th that the minor quake occurred.

Graphical user interface, chart, application, table

Description automatically generated

Figure 9. Total count of shake intensity reports by hour and location

Graphical user interface

Description automatically generated

Figure 10. Reported shake intensities by hour and location.

## Damage Map Analysis

Using some image processing tricks, the neighborhood labels were extracted from Figure 1 and overlaid with the damage map from the major earthquake on April 8th to produce Figure 11. From this, you can see that location 1 was almost completely spared, locations 3 and 4 were among the strongest hit, and location 2 was in the middle levels of shake intensity. Rather than use full 8.5 x 11 pages of graphs for analyzing damage reports, I will use the first four locations for relevant analysis of damage reports and damage report counts. The actual analysis covered the full set of graphs which are included in Appendix II: Hourly Damage Reports by Location and Appendix III: Hourly Damage Report Counts by Location.

Map

Description automatically generated

Figure 11. Measured shake intensity map overlaid onto the neighborhood map of St. Himark

## Damage Report Analysis

Damage is reported by Rumble users on a scale of 0 to 10. The median value for each hour at each location was chosen for analysis. The median was chosen over the mean because some hours have very few reports and the median prevents any extreme values from affecting the average. Additionally, the mean and median shake intensity graphs were compared and as can be seen in Figure 12 and Figure 13, there is little qualitative difference between the two.

Graphical user interface, chart

Description automatically generated

Figure 12. Mean reported shake intensities by hour and location

Graphical user interface, table

Description automatically generated

Figure 13. Median reported shake intensities by hour and location

Looking at the shake intensity graphs, location 1 has low shake intensity the whole time, with consistent reports of intensity level 1 around hour 55 with a possible power/internet outage at hour 84. Location 3 has low shake intensity around hour 15, then middle intensity shaking at around hour 55. There is an outage from hours 59 to 72 and from around hours 86 to 108. Late reports arrive all at once at hour 73, 86, and 108. Location 2 and 4 report low and mid-level shaking around hour 55. There isn’t an outage covering the entirety of each neighborhood, but areas seem to be reporting shaking data late between hours 86 and 96. Likely this is smaller areas within the neighborhoods getting their power/internet restored.

Diagram, table

Description automatically generated

Figure 14. Median reported sewer and water system damage by hour and location

Graphical user interface, table, schematic, Excel

Description automatically generated

Figure 15. Median reported power system damage by hour and location

Table

Description automatically generated

Figure 16. Median reported damage to roads and bridges by hour and location

Similar patterns occur in the reports of the other damage types. One important pattern that stands out is that people are reporting damage before either of the earthquakes happened. In Figure 17, Neighborhoods 1 and 3 have a lot more medical facilities damage reports than Neighborhoods 2 and 4. This is because Neighborhoods 2 and 4 don’t have hospitals so they are likely reports from people near bordering neighborhoods with medical facilities that have damage.

Graphical user interface, diagram, schematic

Description automatically generated

Figure 17. Median reported damage to medical facilities by hour and location

Diagram

Description automatically generated with medium confidence

Figure 18. Median reported damage to buildings by hour and location

# Insights and Next Steps

Based on the damage reports, people are reporting damage before the earthquakes happened so the urgency assessment needs to consider a change in before and after. Additionally, the earthquakes happened days ago and data from during the quake took over a day to arrive. Assessing the current urgency of damage will need to consider only the most recent damage reports. For future quakes, this system can be deployed shortly after the quake for more useful emergency response direction. Some areas of the city had roadwork, drainage work, powerline work, and other types of construction before the earthquake. If we consider the change in damage reports over the period, the urgency should add more significance to new damage and account for people making damage reports before the earthquakes. Last, the truthfulness of reports can be evaluated by comparing reported shake intensity in a neighborhood to the measured shake intensity from the shake map overlaid on the neighborhood map.

# Urgency Evaluation

Urgency of each damage type at each location will be determined by taking the mean of the last 12 hours of median damage reports (right side of Figure 19) and subtracting the mean of the first 48 hours of median damage reports (left side of Figure 19) to determine the change in damage.

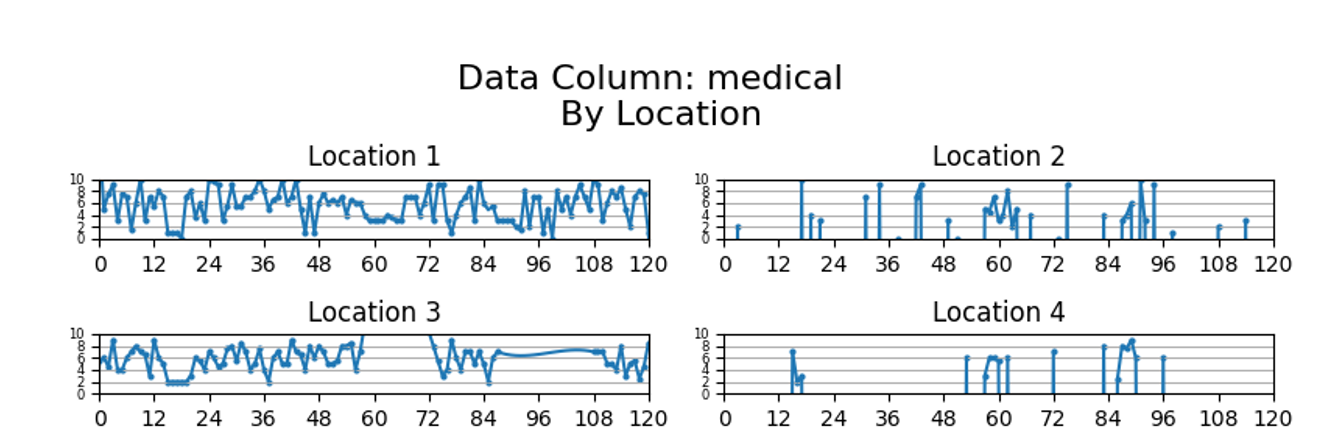


Figure 19. Graph showing the time periods used to calculate current mean damage (black box on the right) and previously present damage (black box on the left)

Adjusted damage is the mean of the last 12 hours of damage reports multiplied by 1 plus the change in damage. This adds urgency to new damage without discounting existing damage entirely. The max measured shake at a location is added to that to get a base urgency value. For the roads and bridges damage reports, if the location has access to the mainland, 2 is added to the base urgency to add weight to the importance of those roads.

Base urgency is then multiplied by location modifiers. Location 7 has a modifier of 0 because it is undeveloped. For medical facilities damage, locations without hospitals get a multiplier of 0. The truthfulness modifier is calculated by taking the difference between maximum reported shake intensity and maximum measured shake intensity (reported – measured), dividing that by 10, and subtracting it from 1. This gives a truthfulness modifier of 0.90 for locations that report a shake intensity of 1 over the measured shake intensity, 0.80 for locations that report a shake intensity of 2 over the measured shake intensity, 1.10 for locations that report a shake intensity of 1 under the measured shake intensity, etc. The Base Urgency is then multiplied by the Truthfulness Modifier for the locations and damage types that did not get a multiplier of 0. This proportionally penalizes overvaluing and proportionally rewards undervaluing shake damage in the urgency calculation. This assumes the neighborhoods making those assessments are off in their values by similar amounts in the other types of reports.

# Results

The Nuclear Power Plant in Neighborhood 4 was in one of the highest shake intensity areas and gets its own Fire Department Emergency Response Team (FD-ERT). That leaves 33 FD-ERTs to assign to various locations and damage types. The tabulated results of the urgency calculations can be found in Appendix IV: Tabulated Urgency Results. A more valuable tool is a map of the most urgent damage reports as seen in Figure 20. The color and size of the circles represent the same thing: urgency of the damage report. Despite Neighborhoods 3 and 4 having some of the highest intensity shaking, there are several damage reports more urgent in other Neighborhoods. This could be due to better construction or a flaw in the urgency evaluation formula. It would likely be more useful to use data just after the earthquake when more people are making damage reports but with this earthquake occurring days ago it doesn’t make sense to do that.

Diagram

Description automatically generated

Figure 20. Map of top 33 most urgent damage reports. Darker reds and larger circles correspond to increased urgency values.

Using the maximum reported median damage over the entire five day period, the urgency map changes to what is seen in Figure 21. The new urgency map suggests even less urgency in Neighborhoods 3 and 4. The description of Neighborhood 3 (Old Town) is it is historic brick buildings suggesting a possibility of sturdy construction that has outlasted multiple earthquakes. The description of Neighborhood 4 (Safe Town) is older single-family homes that may also have survived past earthquakes. If this is not the case, the urgency calculation requires tuning to give more weight to shake intensity.

Chart

Description automatically generated

Figure 21. Map of top 33 most urgent damage reports when using the maximum median reported damage instead of the average reported damage in the last 12 hours. Darker reds and larger circles correspond to increased urgency values.

Other observations from the urgency data are that most of the most urgent damage happened to power systems as seen in Table 1. Areas 7, 12, and 15 have no damage in the top 33 urgent damage reports. Most areas had a median reported shaking intensity close to the truth or completely truthful compared to the measured shaking intensity. Location 2 under-valued shaking intensity by 1, locations 3, 4, 14, 15, 18, and 19 over-valued shaking intensity by 1, and location 12 over-valued shaking intensity by 2.5 making location 12 the least truthful.

Table 1. Counts of damage types in the 33 most urgent damage reports

Graphical user interface, application

Description automatically generated

# Appendix I: NaN Counts in Shake Intensity Column

Graphical user interface, application, table

Description automatically generated

# Appendix II: Hourly Damage Reports by Location

Word

Description automatically generated with low confidence

A picture containing text

Description automatically generated

A picture containing text

Description automatically generated

A picture containing text

Description automatically generated

Graphical user interface, application

Description automatically generated

A picture containing text

Description automatically generated

# Appendix III: Hourly Damage Report Counts by Location

Table

Description automatically generated

Table

Description automatically generated

Graphical user interface, application, table

Description automatically generated

Table

Description automatically generated

Graphical user interface, application, table

Description automatically generated

Graphical user interface, application, table

Description automatically generated

# Appendix IV: Tabulated Urgency Results

Table 2. The 33 most urgent damage reports by location and damage type

Table

Description automatically generated with medium confidence

Table 3. Location and count of damage reports in the top 33 most urgent damage reports

A picture containing chart

Description automatically generated