|  |
| --- |
| Close-up image showing the leaf-sides of two oversized books side-by-side on a bookshelf, with additional books in soft focus background |
| Data 670 Data Analytics University of Maryland University College Professor Dr. Jon McKeeby Assignment 6: Data Analysis Project Final Report |
| |  |  |  | | --- | --- | --- | | Bohannon Smith | 7/29/17 | Data 670 | |

As the lead of a data analytics team contracted to NASA I am assisting in addressing why the United States falls within a list of countries with moderate to high fatalities from landslides. In this project I conducted an analysis on landslide data in order to make an educated guess to why the U.S has such a high fatality rate due to landslides although being one of the most well developed and wealthiest countries in the world. I addressed specific technology needs as well as a broad view of data integration. I demonstrated data integration techniques though the use of IBM Watson Analytics.

Project Scope

I selected the NASA Global Landslide Catalog data set in order to solve the question fatalities in the U.S due to landslides. Starting with the landslide data set I determined recommended resolution insights suggested by the data. I gained insight through analysis to predict outcomes and discover underlying patterns. The new insights I gained allowed me to suggest what actions can be taken to solve this fatality problem.

Data Set Description

The Global Landslide Catalog (GLC) was developed with the goal of identifying rainfall-triggered landslide events around the world, regardless of size, impacts or location. The GLC considers all types of mass movements triggered by rainfall, which have been reported in the media, disaster databases, scientific reports, or other sources. This project marks how the U.S compares to other countries in terms of landslide fatalities. I am now suggesting possible reasons why the U.S makes such a particular list in terms of fatalities caused by landslides. The target variable was fatalities compared with several other variables. With fatalities as the target variable all other variables was used as predictors. There is little information on the historical occurrence of landslides at the global scale with information based on media reports, online databases, and other sources. Landslides were reported most frequently from July to September (As shown in Figure #1) with most events occurring in Asia, North America and Southeast Asia.

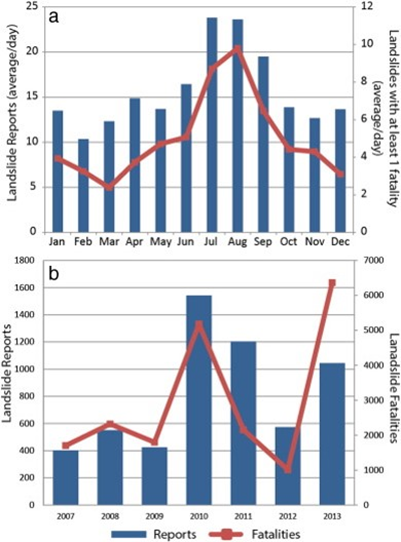


Figure #1: The Global Landslide Catalog (GLC)

a) Reports and events with at least 1 fatality by month (shown as average number of reports per day)

b) Reports and total landslide fatalities by year from 2007–2013. The peak of landslide fatalities in 2013.

Data Analysis Methods and Strategy

I integrated the Systemic Approach which coalesces and distillates on the interaction between elements while studying the effects of interactions. I emphasized the global fatality rate in comparison to the US while modifying groups of variables concurrently (Rosnay, 1997). Using visualization services I discovered patterns and meaning in the data using automated predictive analytics and cognitive capabilities with displays and dashboards to examine and spot trends in landslide fatalities in the US compared to other countries.

Using Watson I interconnected comparisons, relationships, and trends in the landslide data target and predictor variables that will emphasize and clarify the data. We will use Watson to analyze our reliable data by first making sure it’s in synch while we explore, analyze predictors and execute assembling which will thus create a conviction in our understandings into our relief efforts in regards to landslide fatalities. Using Watson Analytics to analyze or landslide data we will show analysis and visualizations in order to discover patterns and meaning in our data. We will utilize this strategy to interact with our data conversationally to get answers we can understand.

We included quality data fields or excluded fields that Watson Analytics has recommended to be excluded. An example of this would be to include location of catastrophic landslides solely to have account of those places in the data but that has more than 25% missing values. We used the refine technique to execute the following data polishing practices:

• Retitle column titles.

• Filter data into a specific subgroup of rows.

• Enhance new columns based on intentions.

• Consolidate our data using groupings and hierarchies.

• Indicate which columns to contain or exclude in our data analysis.

• Journal data values and data marks for each column

Data Refine and Preparation

I viewed the data metrics and scores that displayed in Refine with Watson Analytics automatically producing information about the data. A quality score for each column indicated a column’s potential readiness for use in our prediction as well as the percentage of data absent from each column. We also utilized distribution graphs of the data values in the individual numeric column. I used Watson to communicate the analysis and insights discovered through combining visualizations with text, images, and shapes. I created a dashboard and expert storybook where I and other end users can monitor and share insights.

Describe Descriptive Statistics Used

In the following figures I viewed the statistical details for bar charts and a heat map visualization. These statistical details provided information about the statistical analysis that was performed to produce the visualization. The descriptive statistics are used to perform more advanced analyst and understand the details and accuracy of the analysis. I have a heat map displayed in Figure #2 showing the country’s population for landslide size and type.

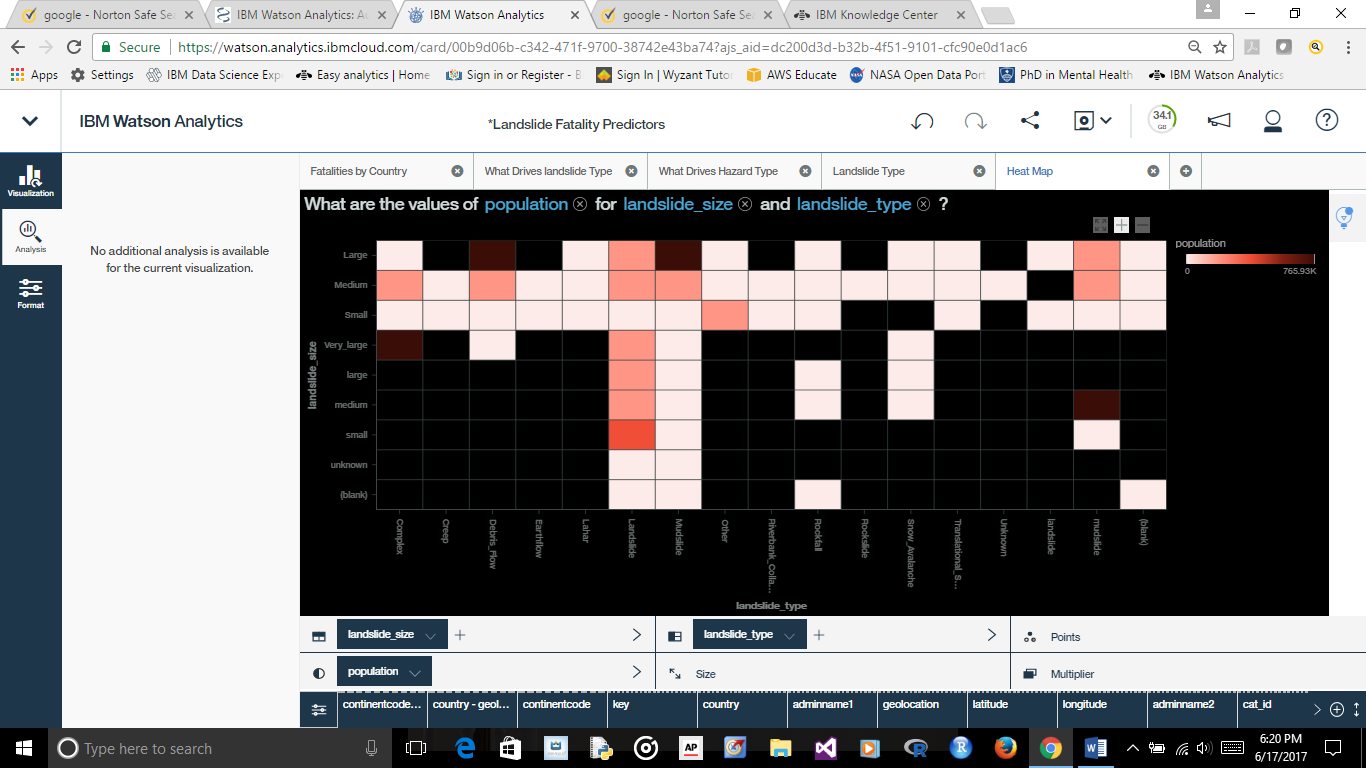


Figure 2: Heat Map population for landslide size and type

### In Figure 3 I ask the question how the values of population compare by landslide size and landslide type. I am then given a bar chart showing mudslides as the mean type. In figure 4 the landslide type median is shown as the teal representation the bar chart. In figure 5 Sri Lanka is show as the country with the median value for population. Watson Analytics enables me to build displays, dashboards and a story book using all of these visualizations. I used these tools in order to create an expert storybook for the landslide data analysis.

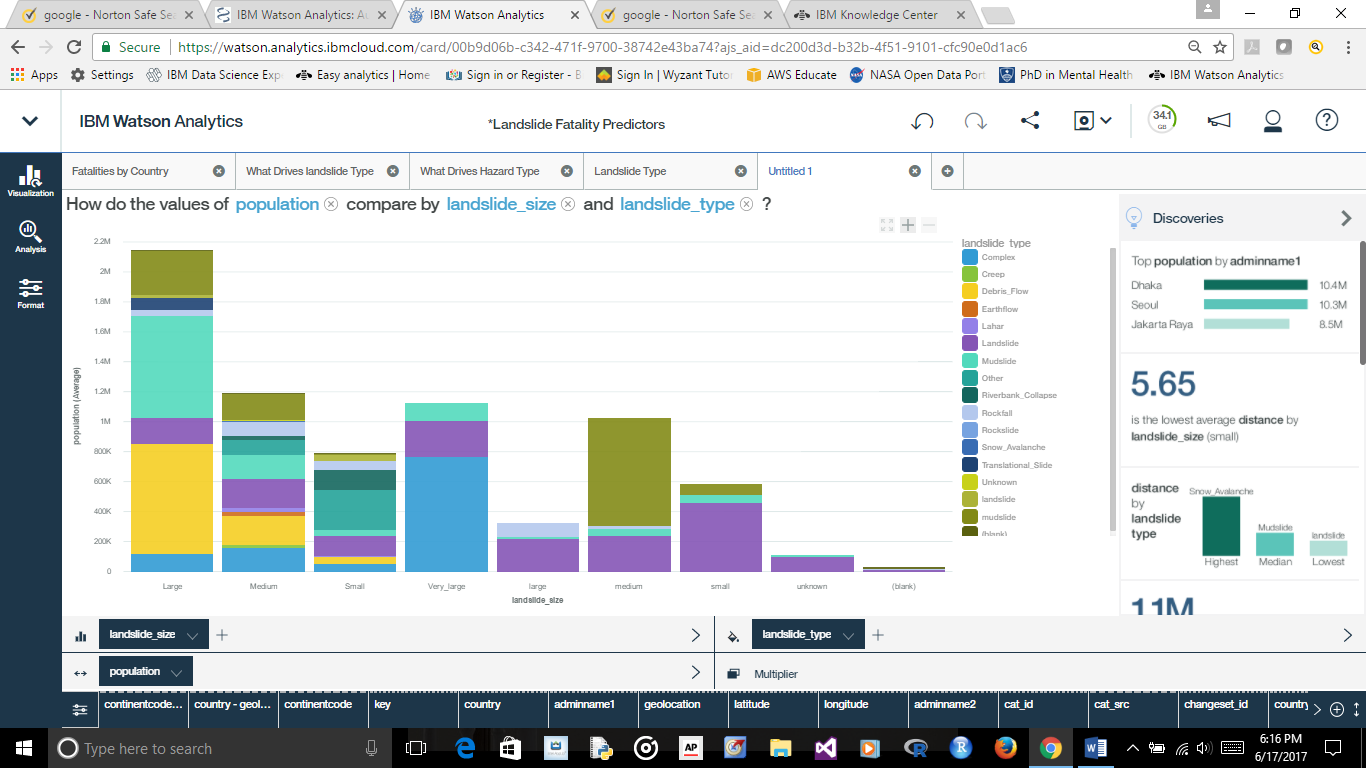


Figure #3 – Mudslide as Mean Type

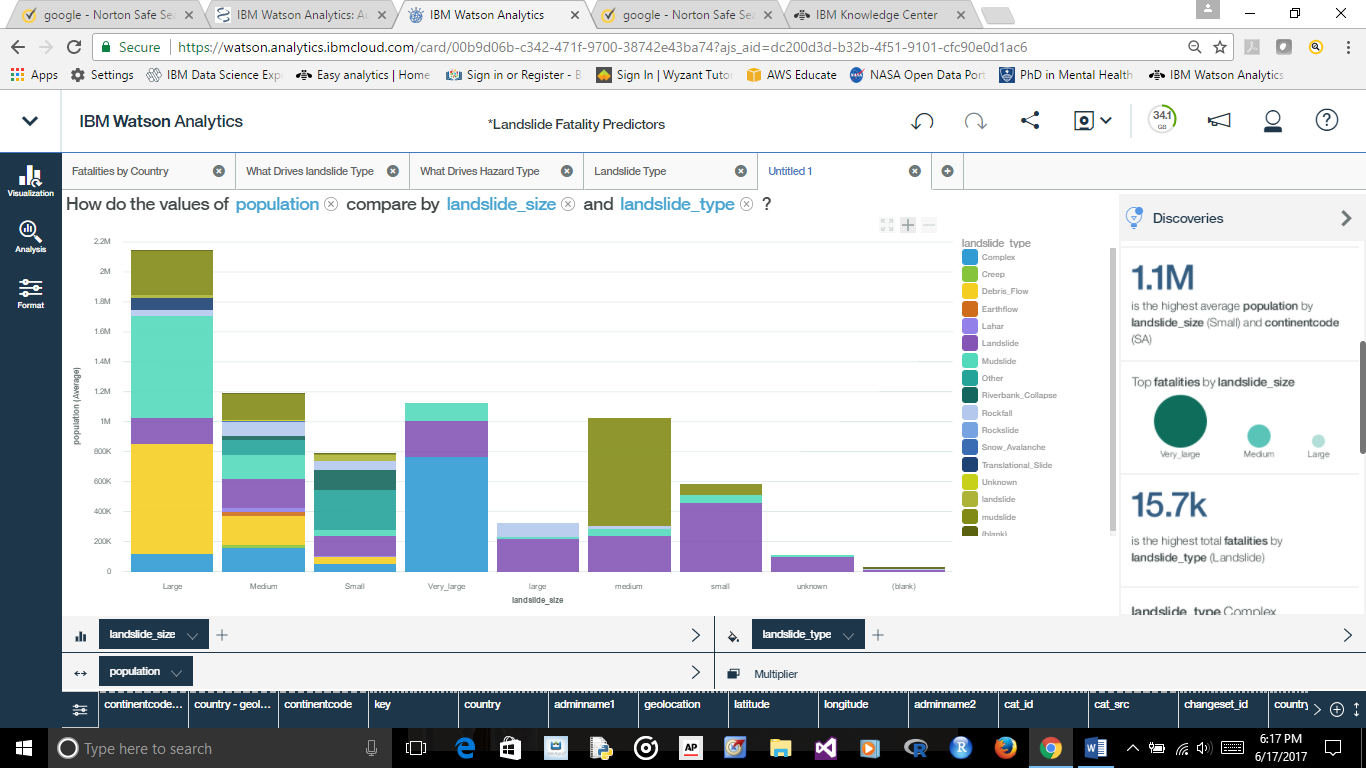


Figure 4 – Landslide Type Median

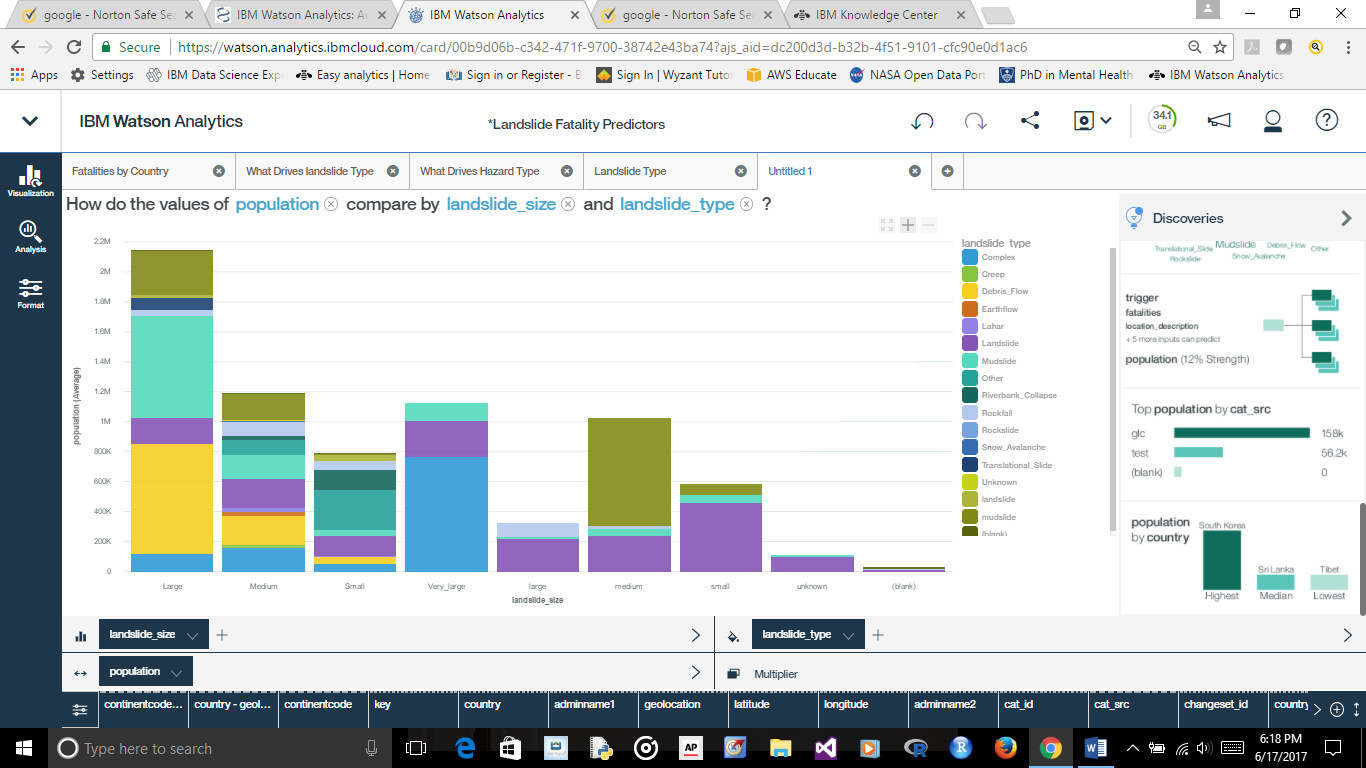


Figure #5 – Sri Lanka as Median Population

Descriptive Statistics 2.6

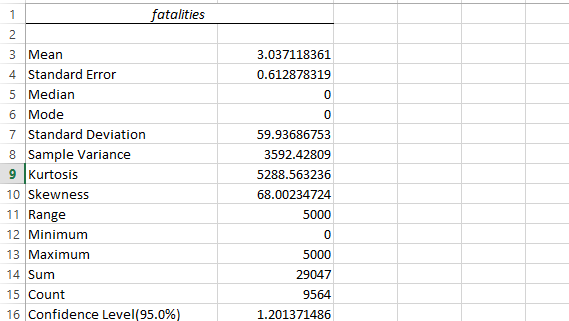


Figure #6 – Descriptive Statistics

Figure 6 shows the descriptive statistics of the fatality data as a whole. As you can see above the mean amount of fatalities is approximately 3. However, the standard deviation is at approximately 60 which is a cause for concern statistically. With a minimum of 0 and a maximum of 5000 we can see how the standard deviation would 60. There are many countries analyzed in this data set with many resulting in 0 fatalities. However, the countries that tend to have fatalities often have many which is another factor of the data. An option would be to exclude all countries with fatalities.

I used Microsoft Excel for calculating descriptive statistics in the Figures 7 and 8. By examining the descriptive statistics for fatalities using the Descriptive Statistics tool I summarized the landslide data set to get an idea of the distribution of the data. All of the statistical data shown here will be represented in the displays, dashboard and storybook.

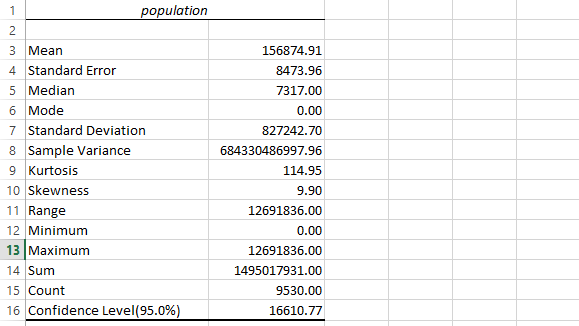


Figure 8: Descriptive Statistics

I based our U.S comparison per capita which means per country in this case. Per capita is a Latin term which represents "by the head." We used this perception of data in this project grounded on the average fatality rate per country in relation to population. Using the per capita method of statistical analysis we are providing an approximation how landslide fatalities affect each country. Thus, using Watson we will be executing a per capita analysis by dividing our statistical measurement for landslide fatalities per country by its population.

The formula is:

Fatalities / Population = Fatalities per Capita.

The storybook I created is a packaged asset with data models, analysis, visualizations and insights focused on solving the fatality problem in the U.S as compared to other countries. I created displays and dashboards which I then applied to the storybook. The end user can see the visualizations in each top finding by tapping a top finding and then explore the details.

### Results.

Fatality Predictors

As we see below in Figure 9 India has the greatest amount of fatalities with unknown almost equal. China is the second highest known country in fatalities with Afghanistan coming close but since being significantly smaller than China would be higher per capita. As you can see there aren't any developed countries surrounding the U.S in the data.

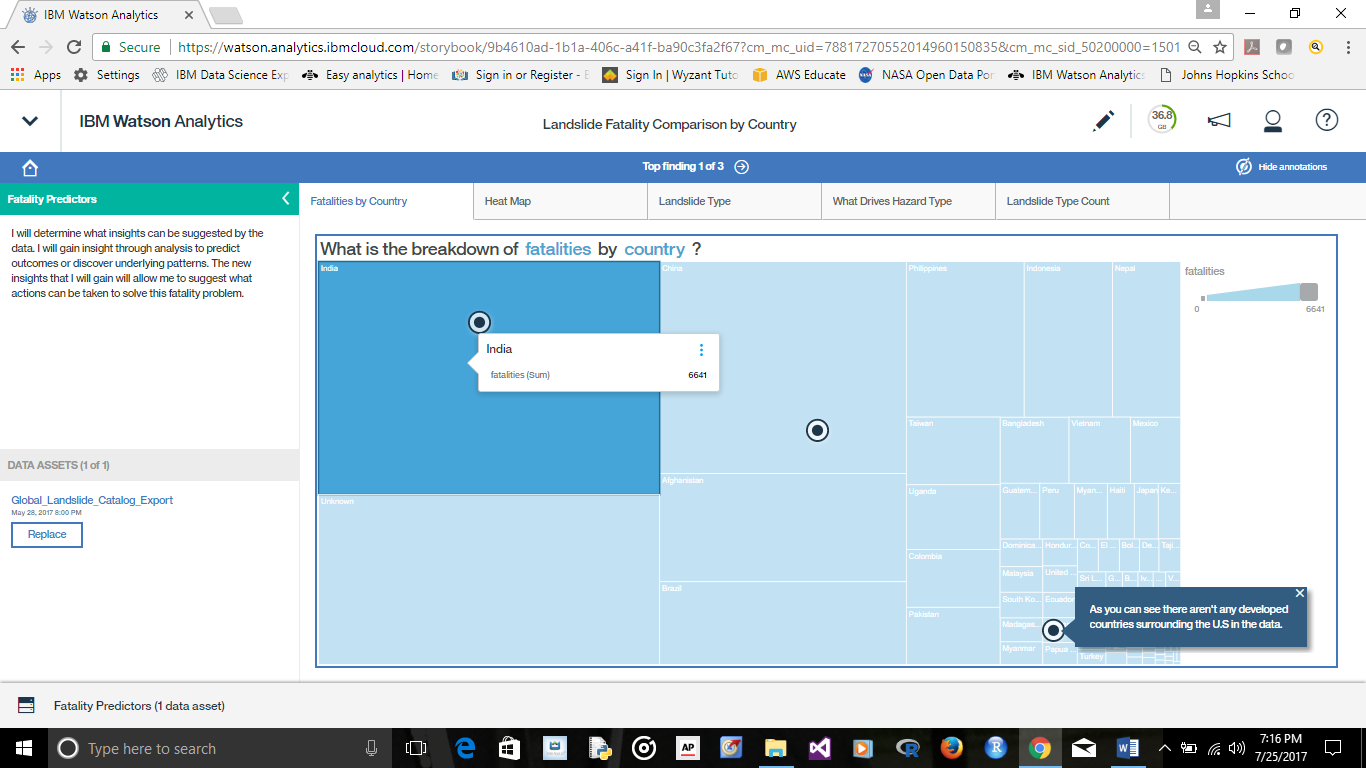


Figure 9 – India top country in fatalities

As seen below in Figure 10 landslides are not the most common type in the most heavily populated areas. The most populated areas are prone to having mudslide, debris flow, and complex. The visualization in Figure 11 paints a colorful picture of the landslide type.

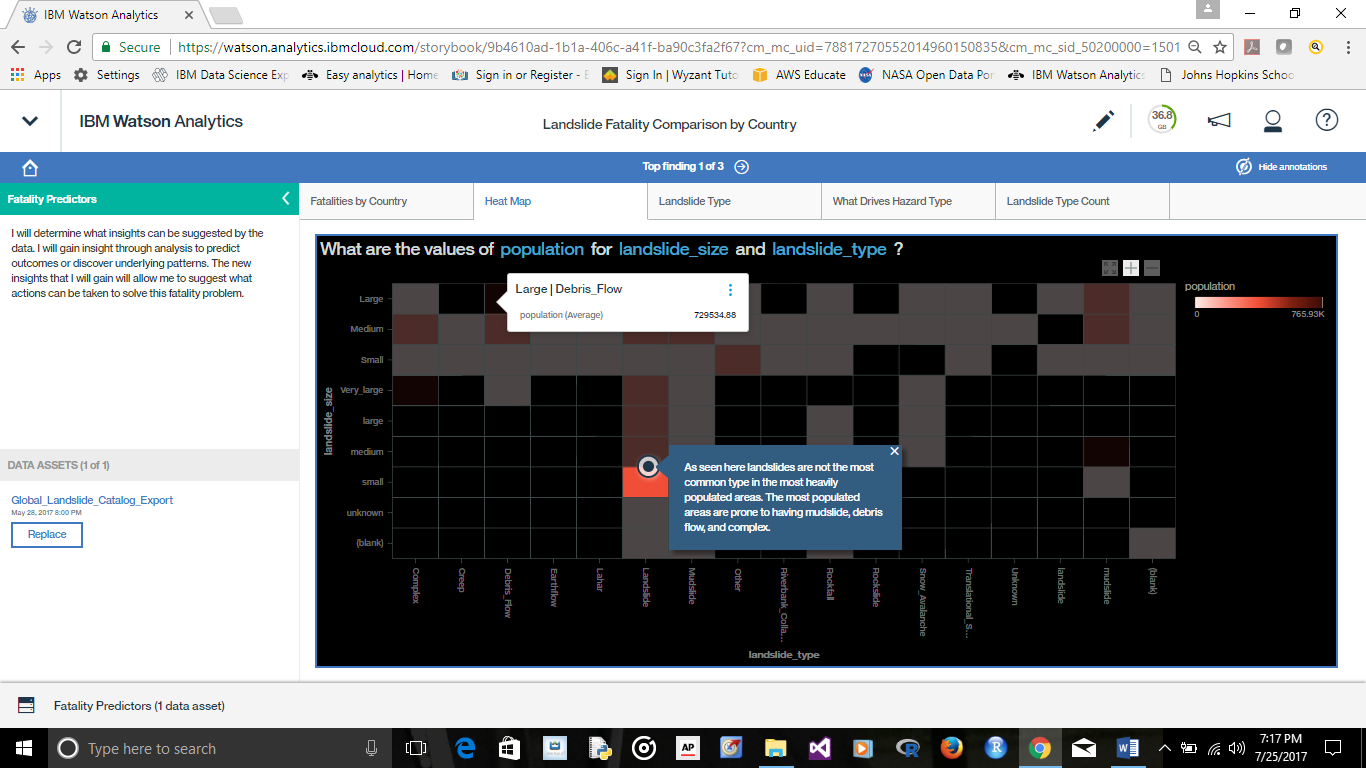


Figure 10 – Mudslides common in heavily populated areas

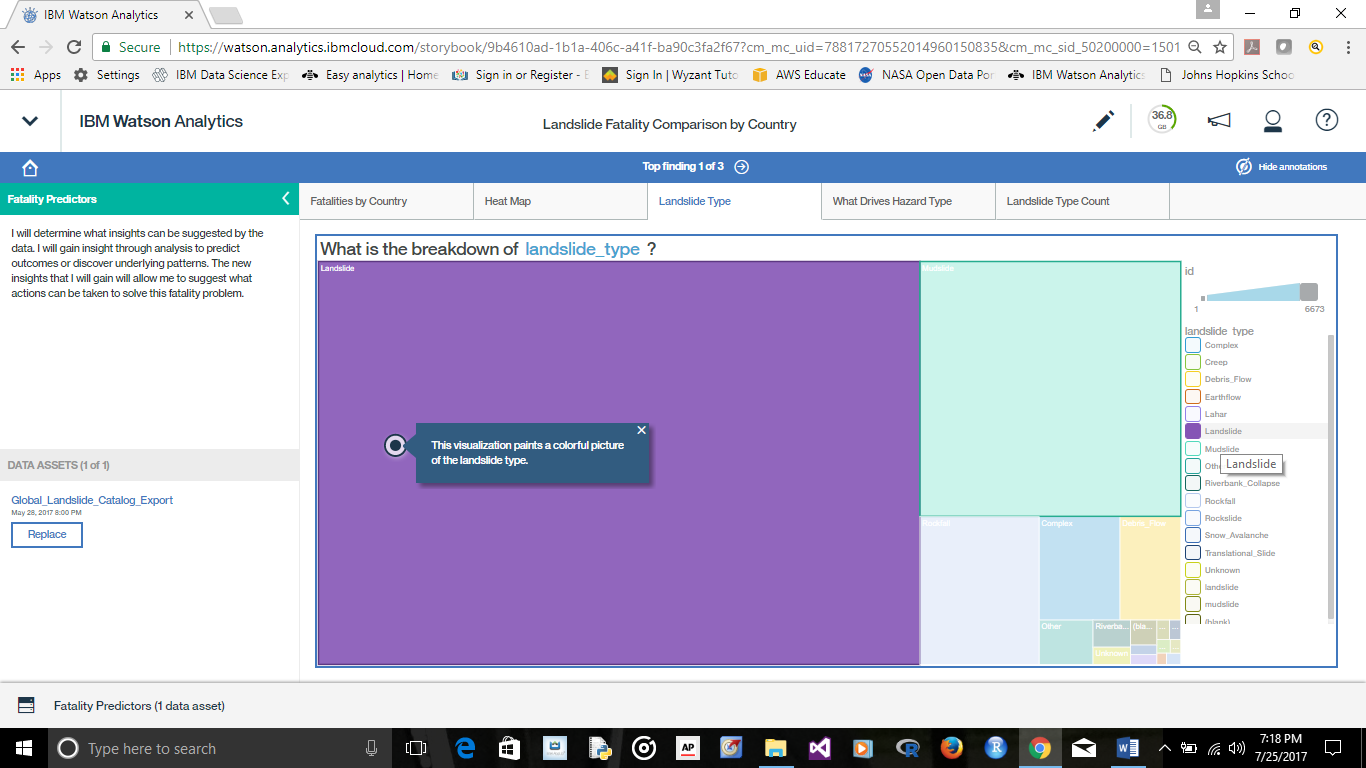


Figure 11 – Landslide Type

As seen here below in Figure 12 location description and trigger are what drives hazard type with a 100% accuracy. The greatest trigger of landslides is downpour with rain being in second.

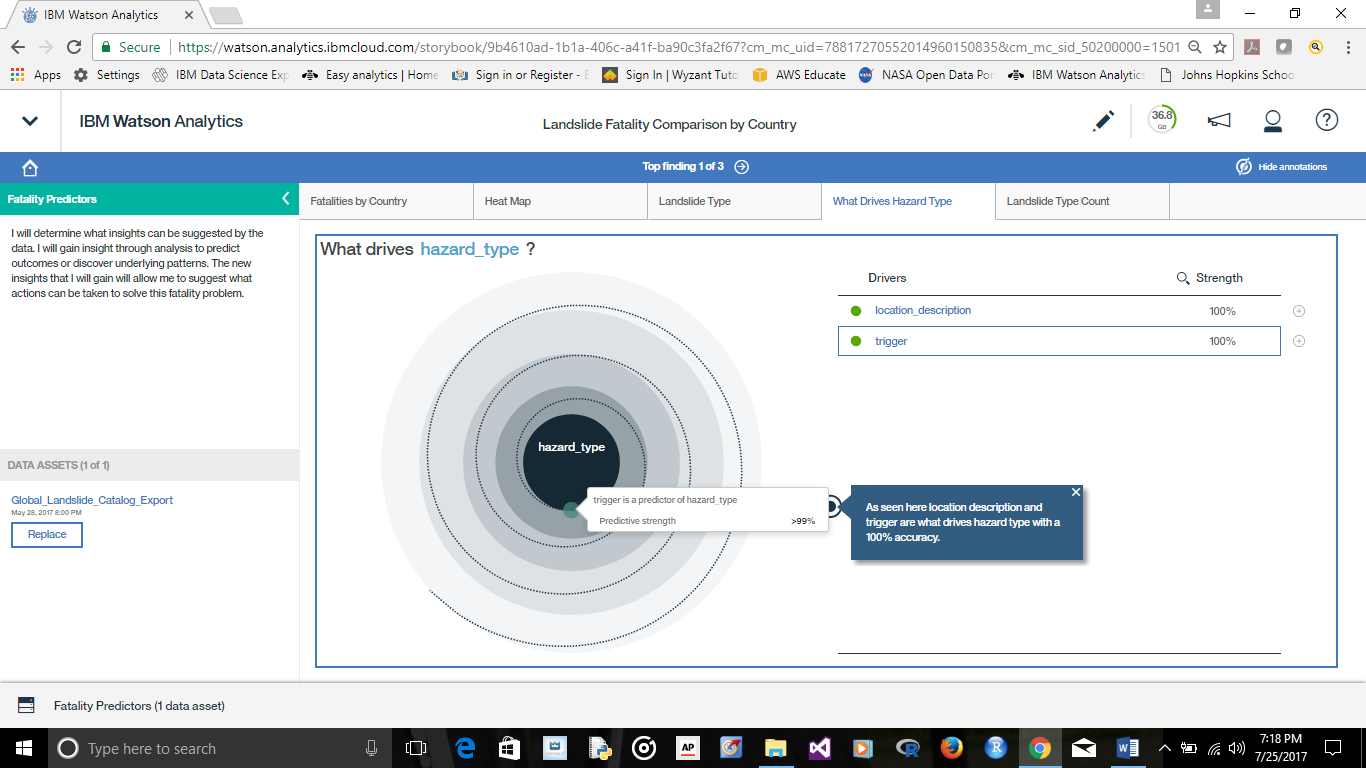


Figure 12: Location description and trigger drive hazard type

As we see here in Figure 13 downpour is by far the greatest trigger of landslides with rain being in second. However, as also seen here in Figure 14 the data on what drives trigger is not statistically significant.

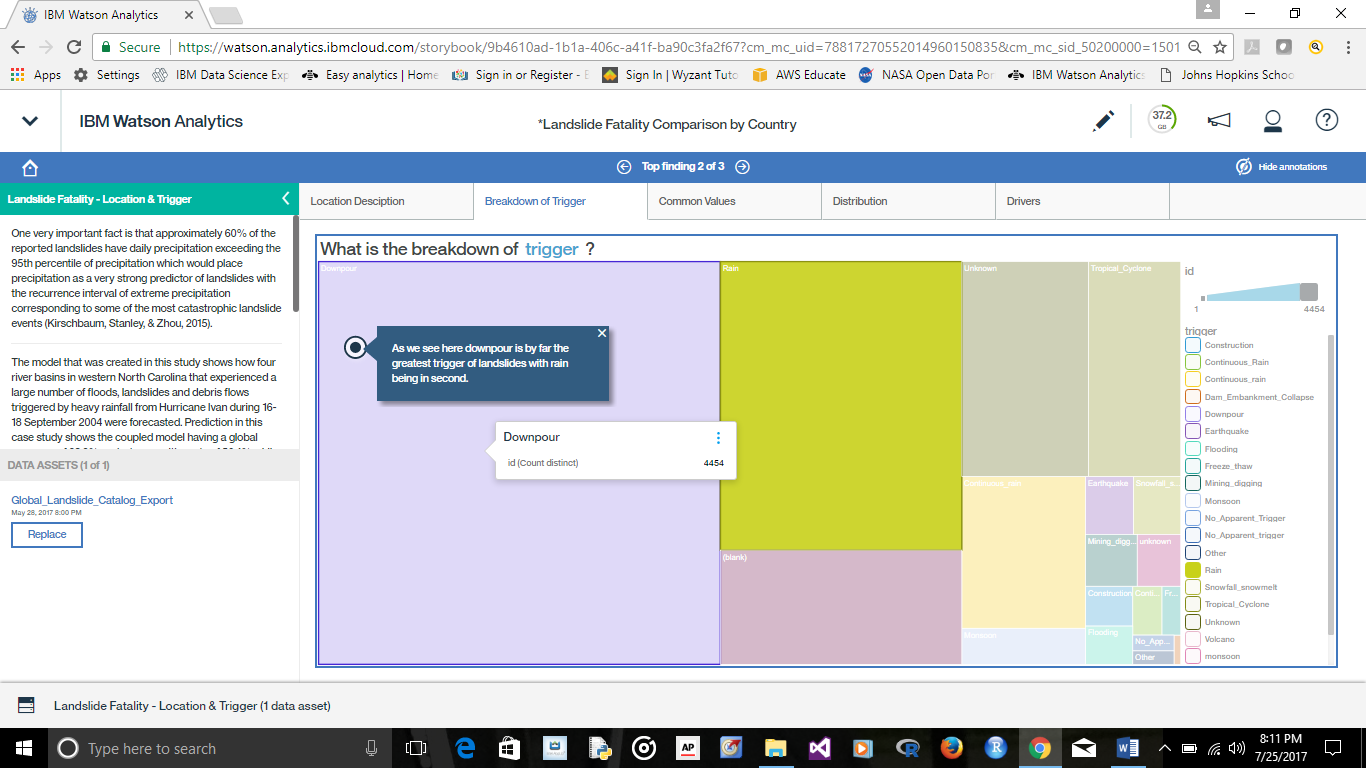


Figure 13 – Downpour greatest trigger of landslides

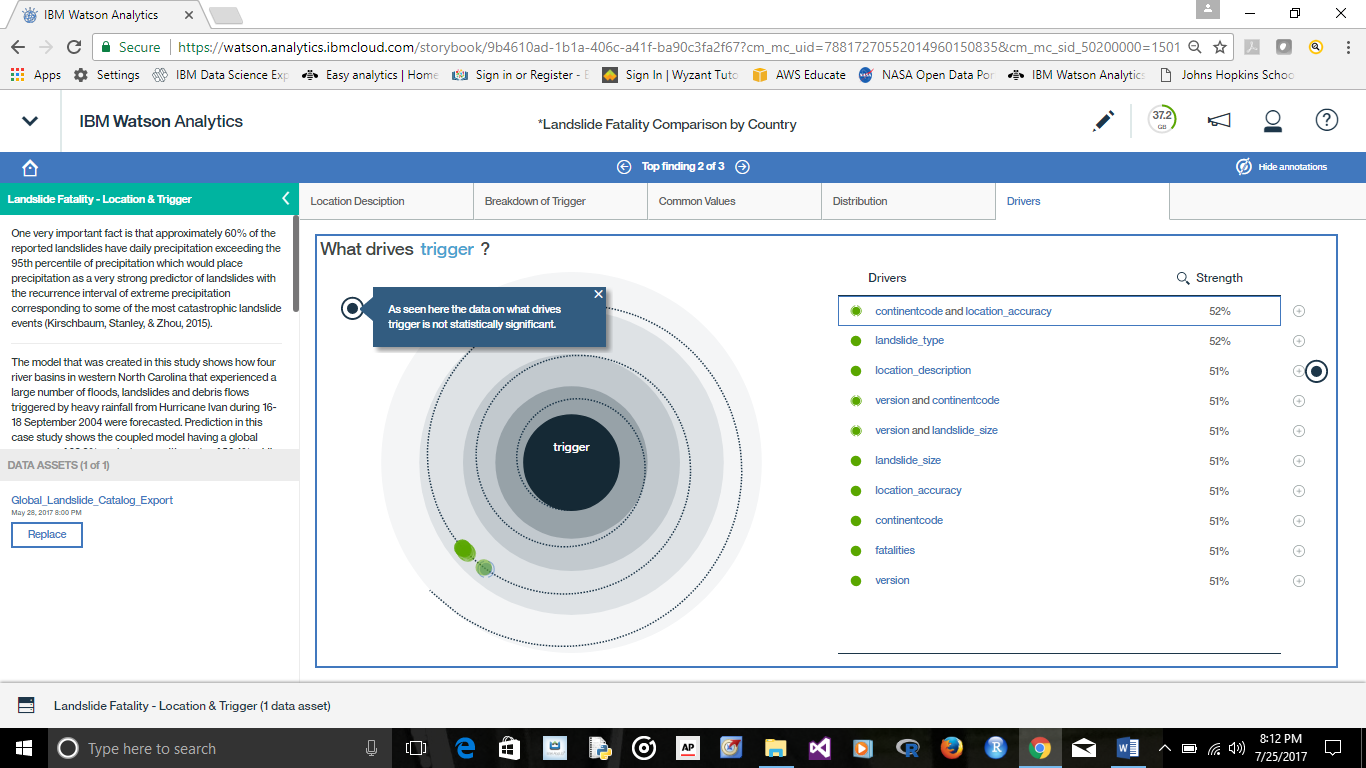


Figure 14 - Data on what drives trigger not statistically significant

Population & Type of Landslide

As seen here in Figure 15 mudslides are most common in densely populated areas. As seen here below in Figure 16 Uttarakhand has an extremely high number of fatalities as it is in the form of an outlier.

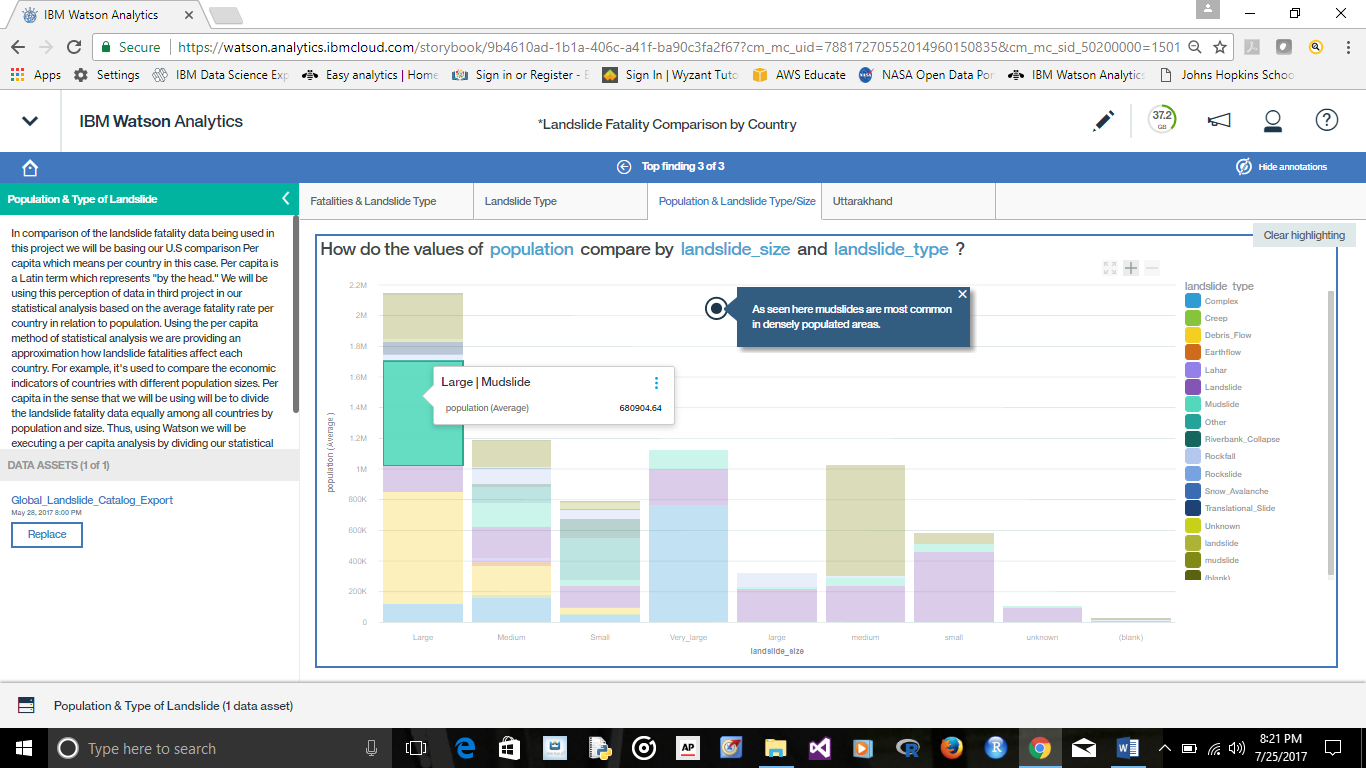


Figure 15 - Mudslides are most common in densely populated areas

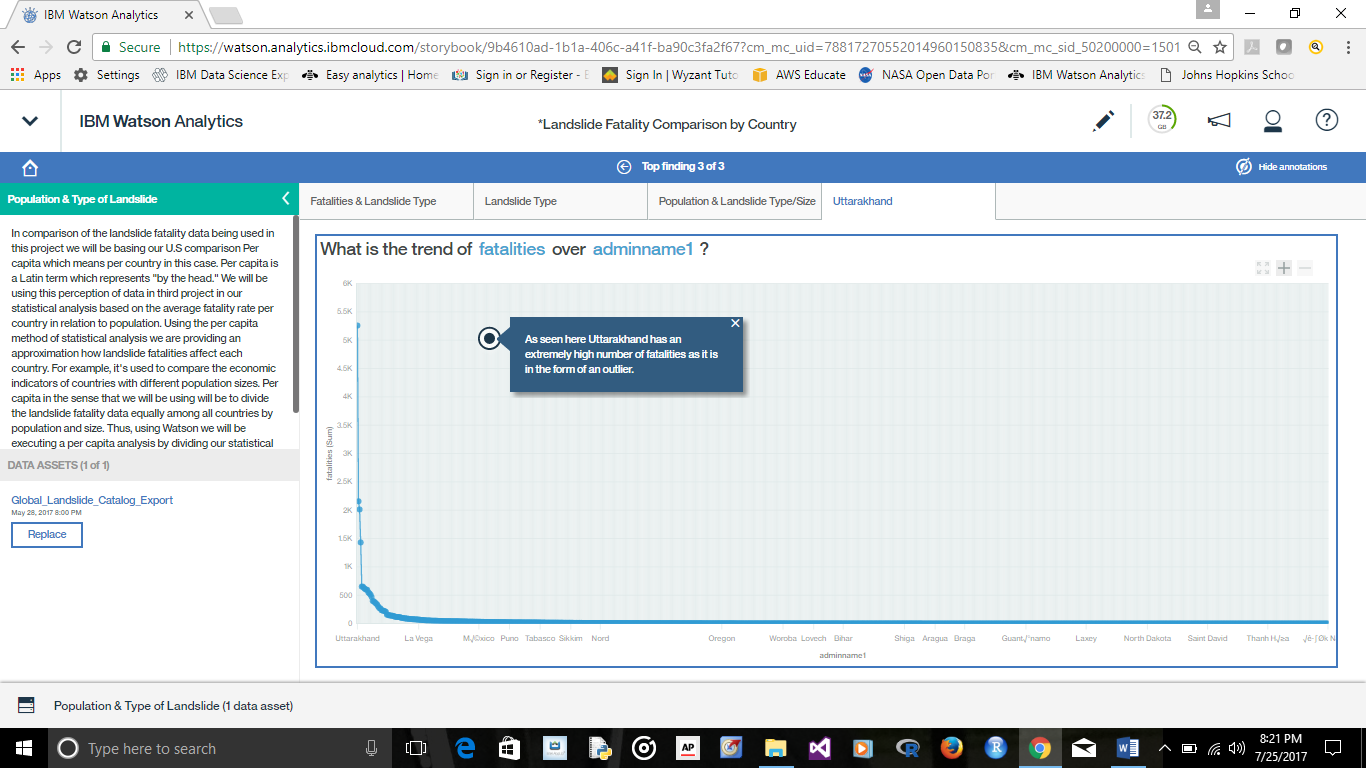


Figure 16 - Uttarakhand has an extremely high number of fatalities

Forecast Outcomes

IBM Watson does not have the advantage of forecasts in order to analyze and discover hidden patterns. However, predictions can be improved through data driven decision making. We will not be building predictive models in order to forecast landslide fatalities in real time; however, in regards to forecasting outcomes we will refer to Ke, Xianwu, Yang, Gourley, Ning, Zhanming, & Wooten (2016) displays (in the below figures) how severe storm-triggered floods and landslides are two major natural hazards in the US that cause between 110 to 160 fatalities per year nationwide. Since floods and landslides often occur in a cascading manner, thus posing significant risk, losses that are significantly greater than the sum of the losses from the hazards when acting separately are acquired (Ke, Xianwu, Yang, Gourley, Ning, Zhanming, & Wooten, 2016).

In a study by Rawat, Uniyal, Dobhal, Joshi, Rawat, Bartwal, & Aswal (2015). Study of landslide hazard zonation in Mandakini Valley, Rudraprayag district, Uttarakhand using remote sensing and GIS. Uttarakhand observed extraordinary damage to life, property, infrastructure and landscape in June 2013 due to torrential rains. This study shows that the outlier I our data with the most fatalities from landslides also happened to experience the greatest trigger. Rainfall in this area surpassed the limit and the excess of rivers led to landslide in the region and flash floods in the downstream areas which is a consequence of complex interaction amid several factors, chiefly involving geological, geomorphological and meteorological factors (Rawat, Uniyal, Dobhal, Joshi, Rawat, Bartwal, & Aswal, 2015 – pg. 158-170).

Conclusion

My initial thoughts are that since the U.S has so many mountain ranges and such vast amounts of land the sheer level of opportunity may have something to do with the fatality rate. However, there are many other factors that may negate and affect the U.S landslide fatality rate compared to other countries. Consequently, since the U.S also has various types of geological locations including mountains, rivers, and complex environments the fatality rate is able to climb very high. The U.S experiences multiple episodes of flooding and swelling of waterways which causes mudslides thus the highest combination factor for fatalities.

Fewer than 5% of the fatalities were reported in North America, suggesting a significant amount of under-reporting in other regions as well as potential discrepancies between developing and developed regions (Kirschbaum, Stanley, & Zhou, 2015). The U.S is shown having a relatively high fatality rate due to landslides and if only 5% of the fatalities are reported this would put the U.S at a very high rate depending on the report rate of other countries. Approximately 60% of the reported landslides have daily precipitation exceeding the 95th percentile of precipitation which would place precipitation as a very strong predictor of landslides with the recurrence interval of extreme precipitation corresponding to some of the most catastrophic landslide events (Kirschbaum, Stanley, & Zhou, 2015).

Recommendations

Promotion and advocacy for data collection. We need to communicate with other countries that have high fatality rates that it is important to report fatalities and practice efficient data collection. I will get more involved in the actual account of the techniques used in the collection of this data and how it can be improved. I will also do further research on the evaluation of data in regards to the statistical significance tests and training set results.

References

Alcántara-Ayala I, López-García J, Garnica R. On the landslide event in 2010 in the Monarch Butterfly Biosphere Reserve, Angangueo, Michoacán, Mexico. Landslides [serial online]. June 2012;9(2):263-273. Available from: Academic Search Complete, Ipswich, MA. Accessed June 14, 2017.

Dowling, C., & Santi, P. (2014). Debris flows and their toll on human life: a global analysis of debris-flow fatalities from 1950 to 2011. Natural Hazards, 71(1), 203-227. doi:10.1007/s11069-013-0907-4

Google - https://www.google.com. Cover Page Photo

Hoyt, R. E., Snider, D., Thompson, C., & Mantravadi, S. (2016). IBM Watson Analytics: Automating Visualization, Descriptive, and Predictive Statistics. JMIR Public Health and Surveillance, 2(2), e157. http://doi.org/10.2196/publichealth.5810

IBM.com - https://www.ibm.com/support/knowledgecenter/en/SS4QC9/com.ibm.solutions.wa\_an\_overview.2.0.0.doc/c\_wa\_an\_ovrvw\_stat\_tests.html#c\_wa\_an\_ovrvw\_stat\_tests

IBM Watson Analytics Quality In, Quality Out, McKinley Stacker IV

https://www.ibm.com/communities/analytics/watson-analytics-blog/quality-in-quality-out/

IBM.com https://www.ibm.com/support/knowledgecenter/en/SS4QC9/com.ibm.solutions.wa\_an\_overview.2.0.0.doc/wa\_an\_hlp\_adjust\_intro.html

<https://www.ibm.com/us-en/marketplace/expert-storybooks>

Ke, Z., Xianwu, X., Yang, H., Gourley, J. J., Ning, L., Zhanming, W., & ... Wooten, R. (2016). iCRESTRIGRS: a coupled modeling system for cascading flood-landslide disaster forecasting. Hydrology & Earth System Sciences, 20(12), 5035-5048. doi:10.5194/hess-20-5035-2016

Kirschbaum, D., Stanley, T., & Zhou, Y. (2015). Spatial and temporal analysis of a global landslide catalog. Geomorphology, 249 (Geohazard Databases: Concepts, Development, Applications), 4-15. doi:10.1016/j.geomorph.2015.03.016

Kirschbaum, D. B., Adler, R., Hong, Y., Hill, S., & Lerner-Lam, A. (2010). A global landslide catalog for hazard applications: method, results, and limitations. Natural Hazards, 52(3), 561–575. doi:10.1007/s11069-009-9401-4. [1]

Microsoft.com - https://docs.microsoft.com/en-us/azure/data-catalog/data-catalog-how-to-data-profile

NASA.com

“This dataset on data.nasa.gov was a one-time export from the Global Landslide Catalog maintained separately. It is current as of March 7, 2016. The original catalog is available here: http://ojo-streamer.herokuapp.com/”

Rawat, M. S., Uniyal, D. P., Dobhal, R., Joshi, V., Rawat, B. S., Bartwal, A., & ... Aswal, A. (2015). Study of landslide hazard zonation in Mandakini Valley, Rudraprayag district, Uttarakhand using remote sensing and GIS. Current Science (00113891), 109(1), 158-170.

Retrieved from: <http://eds.a.ebscohost.com.ezproxy.umuc.edu/eds/detail/detail?vid=1&sid=f46bf554-3b5f-47c5-a258-0360ccd19601%40sessionmgr4009&bdata=JnNpdGU9ZWRzLWxpdmUmc2NvcGU9c2l0ZQ%3d%3d#AN=108333106&db=a9h>

Rosnay J de. (1997 &1979) Analytic vs. Systemic Approaches