Sriram Sai Sumanth
G-01230629

AIT-580 FINAL PROJECT

1.INTRODUCTION

The undertaken dataset comprises of geological hazards data which is being used to assist in understanding the occurrence of all kinds of natural phenomena having abnormal magnitudes leading to the loss of life, injuries and property damage. This data is analyzed and the valuable insights are presented using visualizations.

2.NATURE OF THE DATA

Who collected the data?

National Centers for Environmental Information(NCEI) is the official caretaker of NWS climate records and reacts to the demands for ensured records for litigation purposes. This information is gathered from varied sources such as NWS damage surveys, media, law enforcement/ government agencies, private companies, individuals, newspaper services. It also does not support accuracy of data but provides the best it can. National center for Environmental Information collect their data from national weather service who come under United States federal government and provide all the necessary weather forecasts and warnings, weather forecasting equipment, public safety and appropriate measures to protect property and economy. (service, n.d.)

Why did they collect the data?

The main reason for collecting this data is to analyze the how various disasters can affect an economy and also providing awareness to the public regarding the safety measures that need to be taken during a particular disaster. (service, n.d.)

What is the nature of the data given the purpose of the data collection?

The data set consists of Discrete, Nominal, Categorical and ordinal data.

Pros

- This data set contains the numerical values like magnitude which can be used to predict other variables like direct injuries.
- As the deaths and injuries are given as numerical values it is easy to show them in a visualization and to identify the states with highest number of deaths and injuries.

Cons

- The dataset contains many unimportant columns like Episode ID, LAST MOD TIME, LAST CERT TIME, ADDCORR_DATE which does not provide any information. All these empty columns are dropped for better visualizations.
- There are many null values in the dataset which are replaced with zero.

3.Questions

What are the questions of your interest that can be answered through the data that you choose?

1. What are the average magnitudes recorded in a particular month in a particular state?
2. The month that recorded highest number of injuries?

3. The month that recorded highest number of deaths?
4. What are the maximum injuries for an average magnitude recorded?
5. Which state has the highest number of injuries?
6. Which state has the highest number of deaths?

Justifications on why questions are important

Questions are very much important because they can be used to tell which states are safe based on the average magnitude recorded, which months recorded more deaths and injuries.

Is there any privacy, quality, or other issues with this data?

The data need to be cleaned properly in order to make proper visualizations as there are many missing values in the dataset, there are few empty columns in the dataset but this information can be used to make visualizations if cleaned properly.

Do you have any hypothesis for your questions? If so, Provide justifications. If not, provide justification on why it is hard to hypothesize your question.

1. Is the magnitude of Winter storm more than the magnitude of floods?
2. Are the deaths caused due to winter storm more than the deaths caused due to flood?
3. Are the deaths caused due to winter storm less than the deaths caused due to lightening?

Who can benefit from your data analysis (i.e., who are stakeholders)? Specify detailed justifications on why answering your questions would be beneficial to particular groups of people, researchers, or organizations.

The stakeholders of this data are the councils of the states and the government, they can analyse which states are having more deaths during which month of year and how they would utilize this to protect the people of their country.

4.Requirements and Resources needed

What software and hardware resources you have used in this project?

Hardware: macOS, Intel Core i5 3.1 GHz, Ram 8GB

Software: R studio is used analysis and hypothesis testing, Jupyter Notebook is used for cleaning, Network visualizations, Correlation matrix and for displaying accuracy, Tableau is used for visualizations.

What kinds of pre-processes were needed to make use of the data, and why?

The data need to be pre-processed properly, checking all the required columns and replacing nulls with zeros and also dropping the columns that are of no use and also simultaneously dropping all duplicates thereby making the dataset ready for visualization

Sriram Sai Sumanth
G-01230629

What are the advantages and limitation of the target dataset in answering your question?

The advantages of the storm dataset are we can tell about the deaths, injuries and the amount of damage caused by all natural calamities and thereby help us identify which states were affected more by them The target dataset is not clearly organised making it difficult to analyse properly and thereby making it hard to visualize the data properly.

5.Descriptive Analysis

The storm event dataset contains the records of all kinds of storm data, the size of the original dataset is 39.9Mb, it has 50973 rows and 58 columns. The dataset contains the following attributes Under ordinal datatype we have BEGIN_YEARMONTH, BEGIN_DAY, BEGIN_TIME, END_YEARMONTH, END_DAY, END_TIME, STATE_FIPS, YEAR, CZ_FIPS and BEGIN_DATE_TIME. INJURIES_DIRECT, INJURIES_INDIRECT, DEATHS_DIRECT, DEATHS_INDIRECT, DAMAGE_PROPERTY, DAMAGE_CROPS, MAGNITUDE, TOR_LENGTH and TOR_WIDTH come under discrete data types. The dataset also contains attributes of nominal data types like EPISODE_ID, EVENT_ID, STATE, MONTH_NAME, CZ_NAME, WFO, CZ_TIMEZONE and SOURCE. EVENT_TYPE, CZ_TYPE, MAGNITUDE_TYPE, FLOOD_CAUSE, TOR_F_SCALE fall under categorical data type.

Since there are many unwanted attributes in the data set, I have used the following for my analysis

MONTH_NAME: Name of the month for the event in this record. Ex: January, February, March

CZ_FIPS: The county FIPS number is a unique number assigned to the county by the National Institute for Standards and Technology (NIST) or NWS Forecast Zone Number falls under discrete data type. Ex: 88, 53.

CZ_NAME: This attribute tells about the county/zone name and falls under nominal type Ex: PASCO, SARASOTA

INJURIES_DIRECT: The number of injuries directly related to the weather event and falls under discrete data type. Ex: 0,1,2.

DEATHS_DIRECT: The number of deaths directly related to the weather event and falls under discrete data type. Ex: 0, 1,2.

TOR_LENGTH: Length of the tornado or tornado segment while on the ground (in feet) and falls under discrete data type. Ex: 0.1, 0.2.

TOR_WIDTH: Width of the tornado or tornado segment while on the ground (in feet) and falls discrete data type. Ex: 0.1,0.2,0.25.

STATE: This attribute tells about the state names and fall under categorical type Ex: FLORIDA, ALASKA. (service, n.d.)

Sriram Sai Sumanth
G-01230629

Provide some descriptive statistics so readers can understand the data without actually looking into it

Descriptive statistics of the dataset

```
#displaying the mean,standard deviation,min and max value of the data set

data1=data.describe()
print(data1)
```

```
         EPISODE_ID      EVENT_ID    STATE_FIPS    MONTH_NAME       CZ_FIPS  \
count  5.097300e+04  5.097300e+04  50973.000000  50973.000000  50973.000000
mean   1.014418e+06  5.652699e+06     30.410845      5.886960     79.730190
std    2.784291e+05  1.478213e+04     15.468765      3.023738     83.325578
min    5.507200e+04  5.627083e+06      1.000000      1.000000      1.000000
25%    1.074564e+06  5.639901e+06     19.000000      4.000000     23.000000
50%    1.078946e+06  5.652692e+06     30.000000      6.000000     59.000000
75%    1.084054e+06  5.665503e+06     42.000000      7.000000    110.000000
max    1.151786e+06  5.678291e+06     99.000000     12.000000    840.000000

       INJURIES_DIRECT  DEATHS_DIRECT     MAGNITUDE    TOR_LENGTH  \
count     50973.000000   50973.000000  50973.000000  50973.000000
mean          0.218527       0.013438     10.766697      0.081265
std           7.394486       0.278228     21.849761      1.151790
min           0.000000       0.000000      0.000000      0.000000
25%           0.000000       0.000000      0.000000      0.000000
50%           0.000000       0.000000      0.000000      0.000000
75%           0.000000       0.000000      1.500000      0.000000
max         800.000000      32.000000    131.000000    100.000000

          TOR_WIDTH
count  50973.000000
mean       3.587741
std       48.077317
min        0.000000
25%        0.000000
50%        0.000000
75%        0.000000
max     2200.000000
```

Frequency Count of the column event type

```
In [66]: data['EVENT_TYPE'].value_counts()
```

```
Out[66]: Thunderstorm Wind       13603
         Hail                    12561
         Flash Flood              3582
         Flood                    2318
         Heavy Snow               2197
         High Wind                2165
         Winter Storm             2025
         Drought                  1774
         Tornado                  1529
         Heat                     1404
         Ice Storm                1158
         Heavy Rain                981
         Strong Wind               920
         Lightning                 901
         Winter Weather            833
         Cold/Wind Chill           505
         Funnel Cloud              442
         Dense Fog                 407
         Blizzard                  389
         High Surf                 299
         Hurricane (Typhoon)       191
         Waterspout                176
         Tropical Storm            144
         Coastal Flood             123
         Wildfire                  113
         Storm Surge/Tide           55
         Sleet                      50
         Rip Current                38
         Lake-Effect Snow           36
         Avalanche                  17
         Frost/Freeze               12
         Dust Storm                 10
         Seiche                      8
         Dust Devil                  4
         Debris Flow                 3
         Name: EVENT_TYPE, dtype: int64
```

6.Results/Finding

1.Total number of injuries and deaths due to these hazards every month

Number of injuries recorded in every month in the year 1998

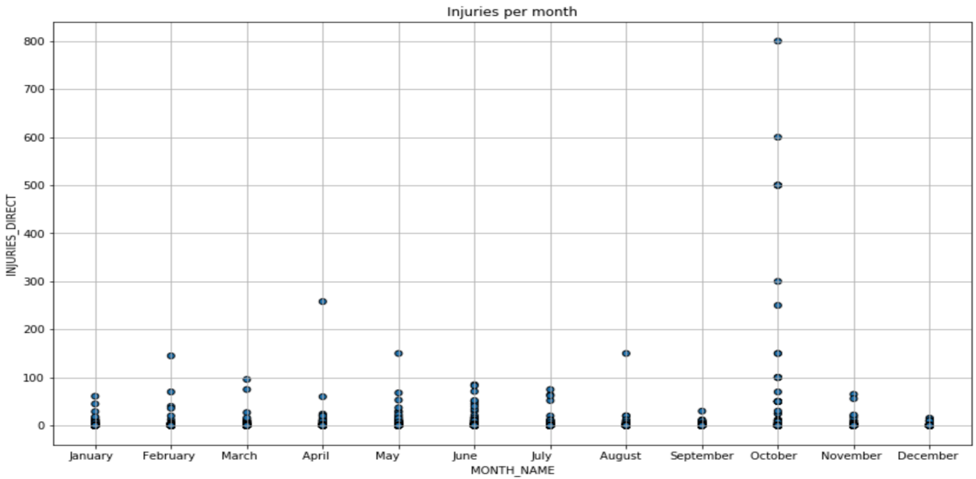

Fig.1 Scatterplot for Injuries Direct and Month

Justification: The above visualization states that more number of injuries are recorded in the month of October and the least number of injuries occurred in the month of December. This helps us in understanding the injuries across United States in the year of 1998.

Total number of deaths recorded in every month in the year 1998

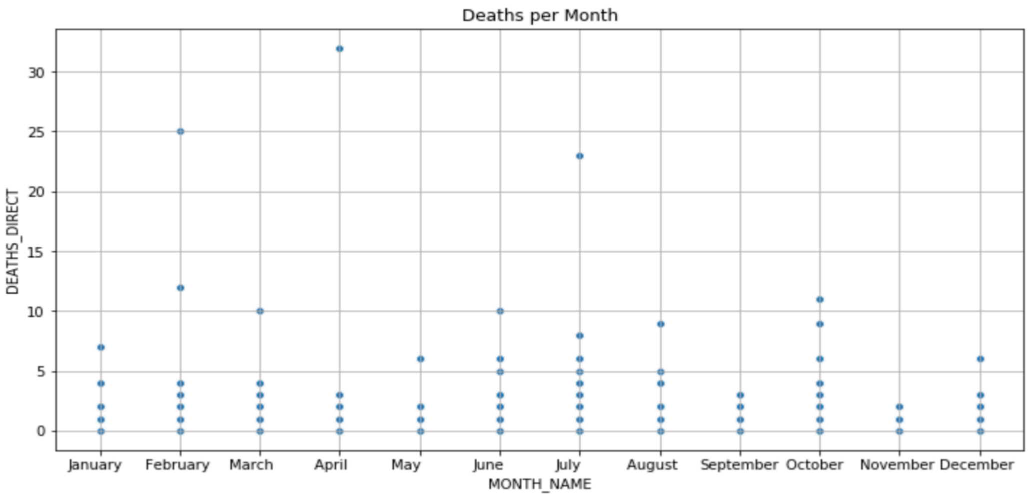

Fig.2 Scatterplot for Deaths Direct and Month

Justification: The above visualization states that more number of deaths are recorded in the month of July and the least number of deaths occurred in the month of November. This helps us in understanding the deaths across United States in the year of 1998.

2.The average magnitude recorded in a particular state in a particular month

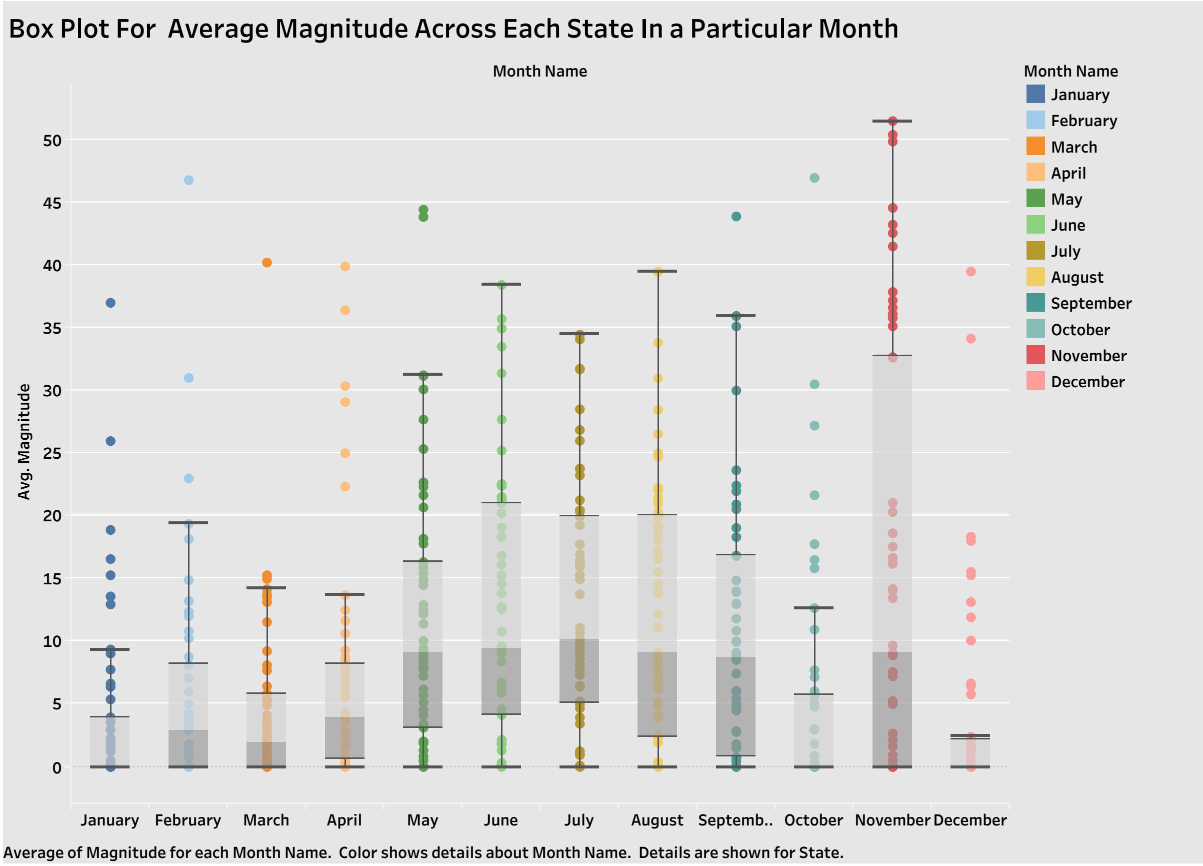

Fig.3 Boxplot for Average magnitude, states and Months

Justification: The above visualization tells us the average magnitudes recorded in a particular state in a particular month in the year of 1998. By this we can say which state is getting more magnitude in which month of the year. Thereby taking necessary precautions in the future.

3.Correlation across the columns of the dataset

Correlation matrix: Correlation matrix is a table showing correlation coefficients between variables, each cell in the matrix shows the correlation value between the two variables.

	EPISODE_ID	EVENT_ID	STATE_FIPS	CZ_FIPS	INJURIES_DIRECT	DEATHS_DIRECT	MAGNITUDE	TOR_LENGTH	TOR_WIDTH
EPISODE_ID	1	0.084	-0.011	-0.032	0.003	0.0042	0.0063	0.0071	0.0076
EVENT_ID	0.084	1	0.072	-0.066	0.022	0.0062	0.049	-5.3e-05	-0.01
STATE_FIPS	-0.011	0.072	1	0.14	0.016	-0.0029	-0.084	-0.0059	0.0082
CZ_FIPS	-0.032	-0.066	0.14	1	0.018	0.0023	0.032	0.0098	0.012
INJURIES_DIRECT	0.003	0.022	0.016	0.018	1	0.29	-0.011	0.054	0.062
DEATHS_DIRECT	0.0042	0.0062	-0.0029	0.0023	0.29	1	-0.018	0.11	0.12
MAGNITUDE	0.0063	0.049	-0.084	0.032	-0.011	-0.018	1	-0.034	-0.035
TOR_LENGTH	0.0071	-5.3e-05	-0.0059	0.0098	0.054	0.11	-0.034	1	0.43
TOR_WIDTH	0.0076	-0.01	0.0082	0.012	0.062	0.12	-0.035	0.43	1

Fig.4 Correlation Matrix

Justification: From the correlation matrix, I thought of using the correlation between Deaths and magnitude, to see if the level of magnitude increases the deaths or not but as the correlation is very less so I thought of using injuries direct and tor width. This correlation matrix can be used to visualize various variables across the dataset in order to determine the relationship between them. The columns in the storm dataset are less correlated making it difficult to relate various columns and make necessary visualization.

4.In order to find out the maximum number of injuries for an average magnitude recorded.

Regression analysis is a prominently used statistical tool to establish a relationship between two variables.

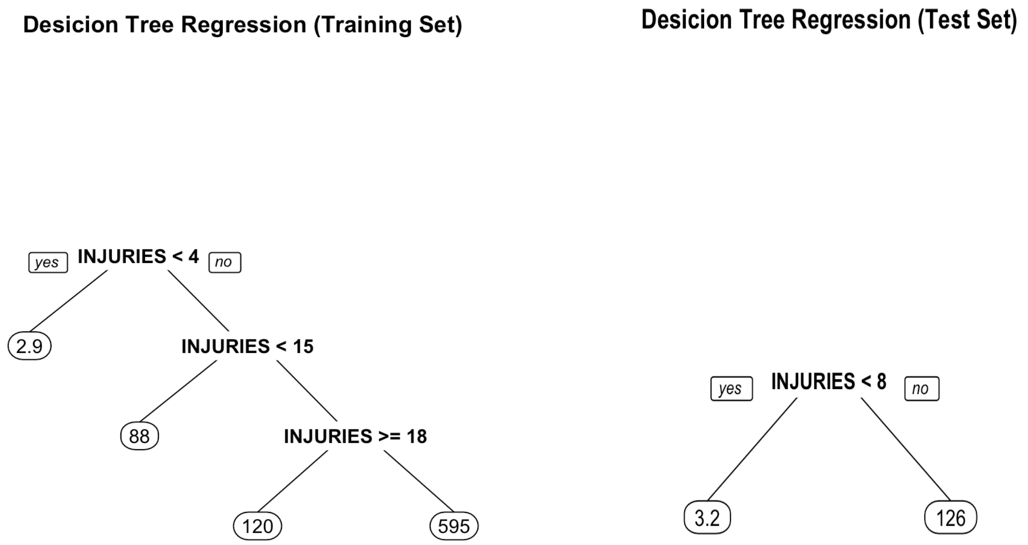

Fig.5 Regression tree for Injuries Direct and Magnitude

Justification: From the above grpah we can say the maximum injuries for an average magnitude recorded. Out of the entire dataset 75 percent of data is used for training purpose and 25 percent of data is used for testing purpose and then we tell the number of injuries. The accuracy is found out using random forest in python and the visualization is done in R.

```
Accuracy on training set: 0.998
Accuracy on test set: 0.998
```

5.To check whether the deaths occurred due to winter storm are more than the deaths occurred due to flood.

Hypothesis testing is used to check whether the given hypothesis is accepted or rejected.

5.1 Is the magnitude of Winter storm more than the magnitude of floods?

```
> t.test(data$MAGNITUDE[winter_storm],data$MAGNITUDE[flood], alternative="greater", conf.level = 0.95)

	Welch Two Sample t-test

data:  data$MAGNITUDE[winter_storm] and data$MAGNITUDE[flood]
t = 1, df = 2024, p-value = 0.1587
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -0.01881027         Inf
sample estimates:
mean of x mean of y
0.0291358 0.0000000
```

Justification: As the P value is 0.1587, as it is greater than the significance level (0.05) we consider null hypothesis is true, that is the magnitude of winter storm is not more than the magnitude of flood.

5.2 Are the deaths caused due to winter storm more than the deaths caused due to flood?

```
	Welch Two Sample t-test

data:  data$DEATHS_DIRECT[winter_storm] and data$DEATHS_DIRECT[flood]
t = 0.62804, df = 3447.9, p-value = 0.265
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -0.004219472          Inf
sample estimates:
  mean of x   mean of y
0.010370370 0.007765315
```

Justification: As the P value is 0.265, as it is greater than the significance level(0.05) we consider null hypothesis is true, that is deaths occurred due to winter storm are not more than the deaths occurred due to flood.

5.3 Are the deaths caused due to winter storm more than the deaths caused due to lightning?

```
        Welch Two Sample t-test

data:  data$DEATHS_DIRECT[winter_storm] and data$DEATHS_DIRECT[lightning]
t = -4.6328, df = 1309.7, p-value = 1.984e-06
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
        -Inf -0.02479796
sample estimates:
 mean of x   mean of y
0.01037037 0.04883463

>
```

Justification: As the P value is 1.984e-06, as it is less than the significance level(0.05) we consider alternates hypothesis is true, that is deaths occurred due to winter storm are less than the deaths occurred due to lightening.

6.In order to find out which state has more number of direct injuries

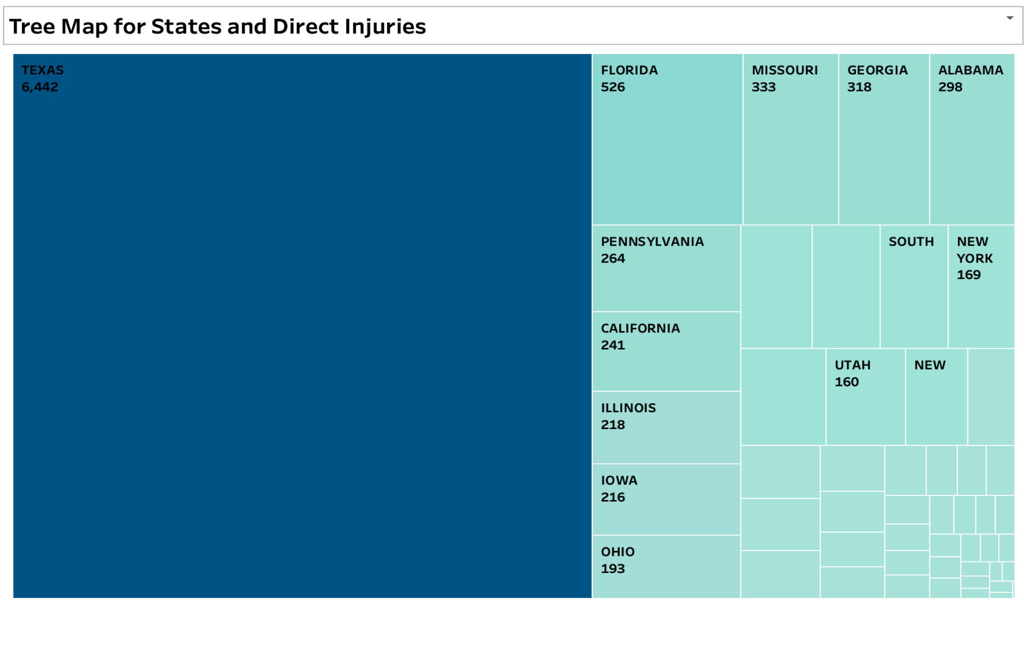

Fig.6 Tree map for Injuries Direct and State

Justification: From the above tree map we can say that Texas is having the highest number of injuries and Utah is having the lowest number of injuries, so using the tree map we can easily say which state is having highest number of injuries and which state is having the lowest number of injuries recorded.

Sriram Sai Sumanth
G-01230629

7.In order to find out which state has more number of direct deaths

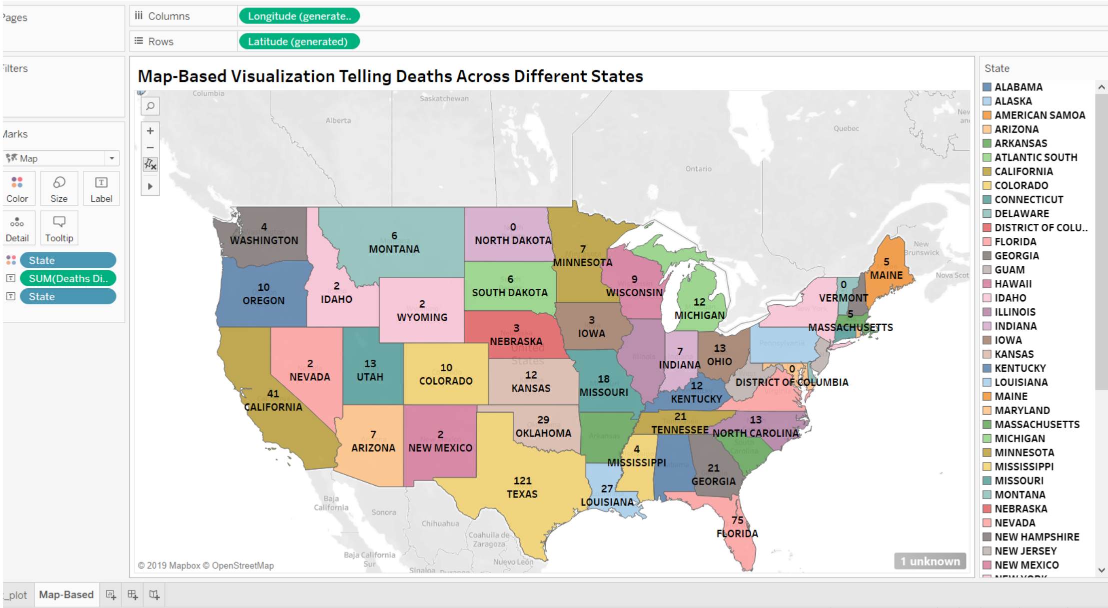

Fig.7 Map Based Visualization for Deaths Direct and State

Justification: From the above map based visualization we can easily say which states have the highest number of deaths and which states have the lowest number of deaths there by bringing awareness among public telling them which state is safe to live.

7.Conclusion and Future work

So, we can say that the project describes about the possible ways to analyze the data and we can generate some useful results from those visualizations. These visualizations tell us about the statistics and conclusions of different attributes of the dataset. The analysis method used is random forest which shows that the model generated is the best fit for the dataset.

Advantages

The visualizations done provided good results so the deaths can be predicted with the magnitude value and provide better results.

Disadvantages

Due to more number of empty values first the dataset need to be cleaned in order to make the analysis, there are many models that can be used for analysis such as gradient boost, adaboost etc.

Future work

A dataset can be made which contains only numerical data for predicting data with the help of categorical value, so that the visualizations and analysis can be done efficiently and effectively.

8.References

service, N. w. (n.d.). *NAIONAL CENTERS FOR ENVIRONMENTAL INFORMATION*. Retrieved from ftp://ftp.ncdc.noaa.gov/pub/data/swdi/stormevents/csvfiles/legacy/

Stephanie. (2016, May 11). *Statitics how to* . Retrieved December 15, 2019, from https://www.statisticshowto.datasciencecentral.com/correlation-matrix/

Business Jargons. (n.d.). Retrieved December 15, 2019, from https://businessjargons.com/hypothesis-testing.html

Tutorials Point. (n.d.). Retrieved December 15, 2019, from https://www.tutorialspoint.com/r/r_linear_regression.htm

Techopedia. (n.d.). Retrieved December 15, 2019, from https://www.techopedia.com/definition/26136/statistical-mean

Investopedia. (n.d.). Retrieved December 15, 2019, from https://www.investopedia.com/terms/s/standarddeviation.asp

Encyclopedia. (2019, December 6). Retrieved Decemeber 15, 2019, from https://www.encyclopedia.com/humanities/encyclopedias-almanacs-transcripts-and-maps/frequency-count

DeepAI. (n.d.). Retrieved December 15, 2019, from https://deepai.org/machine-learning-glossary-and-terms/random-forest

9.Explain/Define terms

Correlation Matrix: A correlation matrix is a table showing correlation coefficients between sets of variables. Each random variable in a table is correlated with each of the other values in the table. This allows you to see which pairs have the highest correlation. (Stephanie, 2016)

Hypothesis Testing: The Hypothesis Testing is a statistical test used to determine whether the hypothesis assumed for the sample of data stands true for the entire population or not. Simply, the hypothesis is an assumption which is tested to determine the relationship between two data sets.

Hypothesis testing is used to check whether the given hypothesis is accepted or rejected. If the p value is less than significance level (0.05) we accept alternate hypothesis and if the p value is greater than significance level (0.05) we accept null hypothesis. (Business Jargons, n.d.)

Regression Analysis: Regression analysis is a very widely used statistical tool to establish a relationship model between two variables. One of these variable is called predictor variable whose value is gathered through experiments. The other variable is called response variable whose value is derived from the predictor variable. (Tutorials Point, n.d.)

Random Forest: The random forest is a supervised learning (DeepAI, n.d.) algorithm that randomly creates and merges multiple decision tree into one "forest." The goal is not to rely on a single learning model, but rather a collection of decision models to improve accuracy. The primary difference between this approach and the standard decision tree algorithms is that the root nodes feature splitting nodes are generated randomly. (DeepAI, n.d.)

Mean: The statistical mean refers to the mean or average that is used to derive the central tendency of the data in question. It is determined by adding all the data points in a population and then dividing the total by the number of points. The resulting number is known as the mean or the average. (Techopedia, n.d.)

Standard Deviation: The standard deviation is a statistic that measures the dispersion of a dataset relative to its mean and is calculated as the square root of the variance. It is calculated as the square root of variance by determining the variation between each data point relative to the mean. If the data points are further from the mean, there is a higher deviation within the data set; thus, the more spread out the data, the higher the standard deviation. (Investopedia, n.d.)

Frequency Count: An attempt to discover the number of occurrences of particular unit in particular contexts of language use. (Encyclopedia, 2019)