DSCI 551 project proposal

• Project title

How dangerous is people's life? From car accidents to natural disasters in US

• Project description

Recently, wildfires across west coast ring the alarm bell for everyone about climate change. Not only wildfires, natural disaster happen to US every year, but it becomes more frequent.

Back to people's daily life, car accidents has become the second cause of death in the United States. According to data collected, we want to have a more detailed understanding of how dangerous people's life is, in terms of both natural disasters and car accidents.

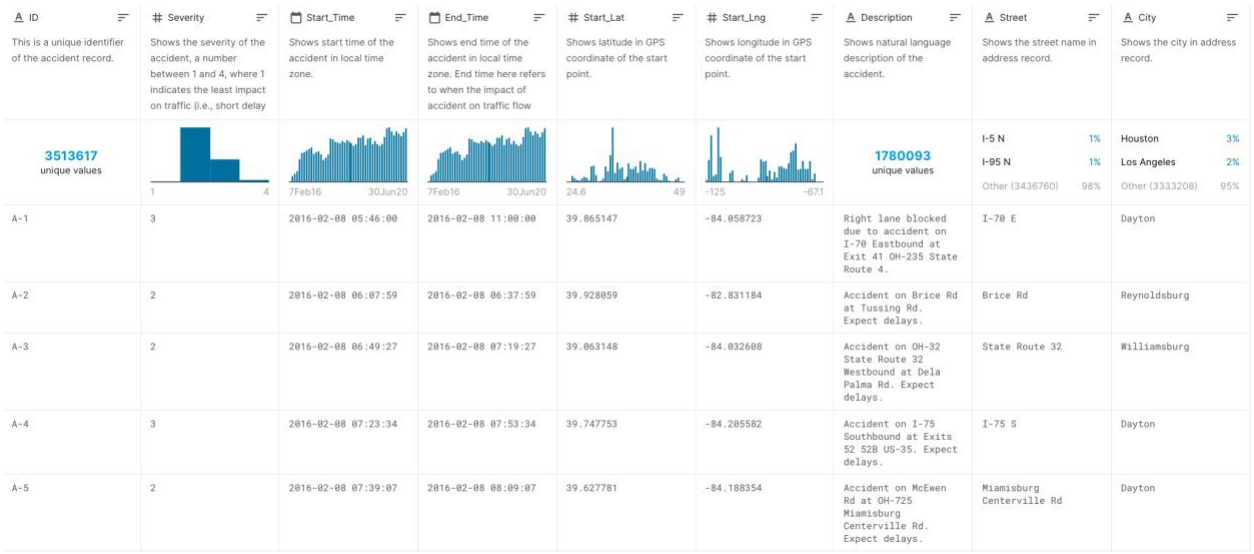
We will use two datasets, US car accidents dataset and US natural disaster dataset to analyze the possible influence of natural disasters had on the weather, furthermore, influence on frequency and/or severity of car accidents. For US car accidents dataset, we will analyze the factor(s) that mostly result in car accidents. Train a prediction model, to predict the severity of car accidents given some description of it. For US natural disaster dataset, we will analyze the frequency, location of disasters by years.

• Data sets

1. US accidents dataset (2016-2020)
   It is a countrywide car accident dataset, which covers 49 states of the USA. The accident data are collected from February 2016 to June 2020, using two APIs that provide streaming traffic incident (or event) data. Currently, there are about 3.5 million accident records in this dataset. The dataset is more than 1GB large.

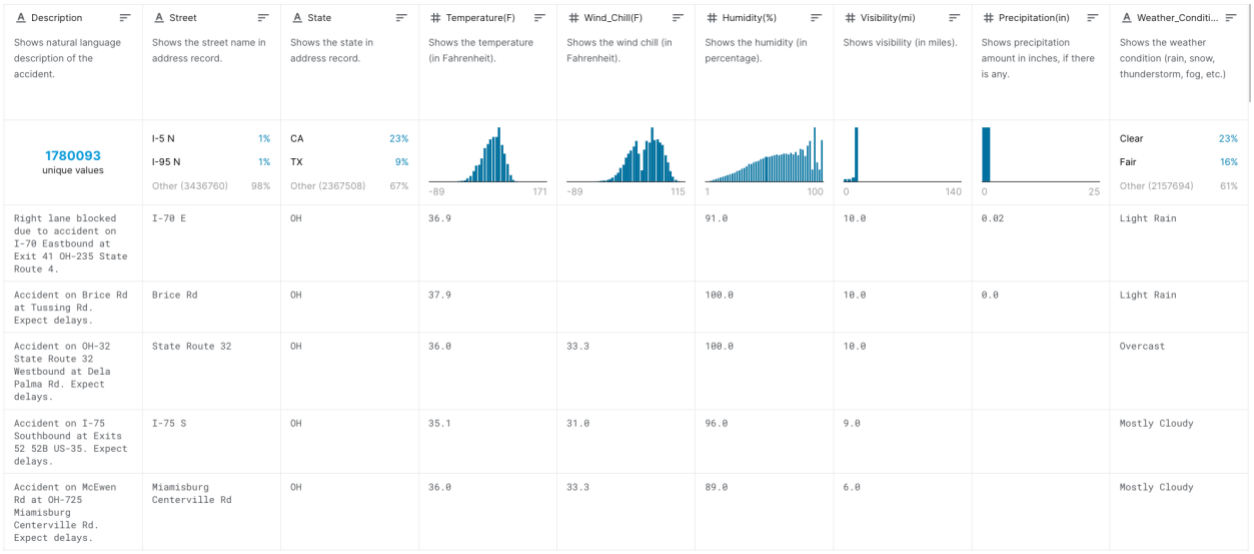
   This dataset can be used to do numerous analysis, such as real-time car accident prediction, studying car accidents hotspot locations, casualty analysis and extracting cause and effect rules to predict car accidents, and studying the impact of precipitation or other environmental stimuli on accident occurrence.

   In this project, we are focusing on the weather condition influence of car accidents. Weather conditions are also related to the natural disaster dataset. We want to find that, whether natural disaster, which severely affect weather condition, can have underlying influence on car accidents, in terms of severity and/or frequency.



ID	Severity	Start_Time	End_Time	Start_Lat	Start_Lng	Description	Street	City
This is a unique identifier of the accident record.	Shows the severity of the accident, a number between 1 and 4, where 1 indicates the least impact on traffic (i.e., short delay	Shows start time of the accident in local time zone.	Shows end time of the accident in local time zone. End time here refers to when the impact of accident on traffic flow	Shows latitude in GPS coordinate of the start point.	Shows longitude in GPS coordinate of the start point.	Shows natural language description of the accident.	Shows the street name in address record.	Shows the city in address record.
3513617 unique values	1 – 4	7Feb16 – 30Jun20	7Feb16 – 30Jun20	24.6 – 49	-125 – -67.1	1780093 unique values	I-5 N 1% I-95 N 1% Other (3436760) 98%	Houston 3% Los Angeles 2% Other (3333208) 95%
A-1	3	2016-02-08 05:46:00	2016-02-08 11:00:00	39.865147	-84.058723	Right lane blocked due to accident on I-70 Eastbound at Exit 41 OH-235 State Route 4.	I-70 E	Dayton
A-2	2	2016-02-08 06:07:59	2016-02-08 06:37:59	39.928059	-82.831184	Accident on Brice Rd at Tussing Rd. Expect delays.	Brice Rd	Reynoldsburg
A-3	2	2016-02-08 06:49:27	2016-02-08 07:19:27	39.063148	-84.032608	Accident on OH-32 State Route 32 Westbound at Dela Palma Rd. Expect delays.	State Route 32	Williamsburg
A-4	3	2016-02-08 07:23:34	2016-02-08 07:53:34	39.747753	-84.205582	Accident on I-75 Southbound at Exits 52 52B US-35. Expect delays.	I-75 S	Dayton
A-5	2	2016-02-08 07:39:07	2016-02-08 08:09:07	39.627781	-84.188354	Accident on McEwen Rd at OH-725 Miamisburg Centerville Rd. Expect delays.	Miamisburg Centerville Rd	Dayton

Figure 1. screenshot of US accidents dataset



Description	Street	State	Temperature(F)	Wind_Chill(F)	Humidity(%)	Visibility(mi)	Precipitation(in)	Weather_Conditi...
Shows natural language description of the accident.	Shows the street name in address record.	Shows the state in address record.	Shows the temperature (in Fahrenheit).	Shows the wind chill (in Fahrenheit).	Shows the humidity (in percentage).	Shows visibility (in miles).	Shows precipitation amount in inches, if there is any.	Shows the weather condition (rain, snow, thunderstorm, fog, etc.)
1780093 unique values	I-5 N 1% I-95 N 1% Other (3436760) 98%	CA 23% TX 9% Other (2367508) 67%	-89 – 171	-89 – 115	1 – 100	0 – 140	0 – 25	Clear 23% Fair 16% Other (2157694) 61%
Right lane blocked due to accident on I-70 Eastbound at Exit 41 OH-235 State Route 4.	I-70 E	OH	36.9		91.0	10.0	0.02	Light Rain
Accident on Brice Rd at Tussing Rd. Expect delays.	Brice Rd	OH	37.9		100.0	10.0	0.0	Light Rain
Accident on OH-32 State Route 32 Westbound at Dela Palma Rd. Expect delays.	State Route 32	OH	36.0	33.3	100.0	10.0		Overcast
Accident on I-75 Southbound at Exits 52 52B US-35. Expect delays.	I-75 S	OH	35.1	31.0	96.0	9.0		Mostly Cloudy
Accident on McEwen Rd at OH-725 Miamisburg Centerville Rd. Expect delays.	Miamisburg Centerville Rd	OH	36.0	33.3	89.0	6.0		Mostly Cloudy

Figure 2. screenshot of US accidents dataset

2. US natural disaster declarations dataset (1953-2020)
   The United States experience a large variety of natural disasters each year: devastating hurricanes, seasonal tornadoes, and scorching wild fires are among the events that endanger many lives and cause billions of dollars in damages. This dataset contains all the natural disaster records in US from 1953 to 2020. It contains columns such as location information (latitude and longitude), time of disaster happened, disaster type, etc.

   In this project, we are focusing on the influence of weather condition, which is directly related to natural disaster, on car accidents. Therefore, we will use natural disaster dataset as constraints, to retrieve data in car accidents dataset, and do analysis.

fema_declaratio...	disaster_number	state	declaration_type	declaration_date	fy_declared	incident_type	declaration_title	incident_begin_d...
Agency standard method for uniquely identifying Stafford Act declarations. Concatenation of declaration_type,	Sequentially assigned number used to designate an event or incident declared as a disaster.	US state, district, or territory.	One of "DR" (= major disaster), "EM" (= emergency management), or "FM" (= "fire management")	Date the disaster was declared.	Fiscal year in which the disaster was declared	Type of incident such as "Fire", "Flood", or "Hurricane". The incident type will affect the types of assistance available.	Title for the disaster. This can be a useful identifier such as "Hurricane Katrina" or "Covid-19 Pandemic".	Date the incident itself began.
4442 unique values	1 5350	TX 8% MO 5% Other (52612) 88%	DR 68% EM 29% Other (1734) 3%	1May53 6Sep20	1953 2020	Severe Storm(s) 28% Hurricane 19% Other (31505) 52%	Covid-19 Pandemic 7% Severe Storms And ... 6% Other (52293) 87%	1May53 6Sep20
DR-1-GA	1	GA	DR	1953-05-02T04:00:00Z	1953	Tornado	Tornado	1953-05-02T04:00:00Z
DR-2-TX	2	TX	DR	1953-05-15T04:00:00Z	1953	Tornado	Tornado & Heavy Rainfall	1953-05-15T04:00:00Z
DR-3-LA	3	LA	DR	1953-05-29T04:00:00Z	1953	Flood	Flood	1953-05-29T04:00:00Z
DR-4-MI	4	MI	DR	1953-06-02T04:00:00Z	1953	Tornado	Tornado	1953-06-02T04:00:00Z
DR-5-MT	5	MT	DR	1953-06-06T04:00:00Z	1953	Flood	Floods	1953-06-06T04:00:00Z
DR-6-MI	6	MI	DR	1953-06-09T04:00:00Z	1953	Tornado	Tornado	1953-06-09T04:00:00Z
DR-8-IA	8	IA	DR	1953-06-11T04:00:00Z	1953	Flood	Flood	1953-06-11T04:00:00Z

Figure 3. screenshot of US natural disaster dataset

• Data problems to be addressed (data cleaning, transformation, integration, aggregation, etc)

1. Processing large dataset problem
   For US accidents dataset, it contains all the accidents' record from 2016-2020, which is more than 3.5 million data points. The dataset is more than 1GB large. It can be processed on Amazon EC2 instance, but may be more efficient to use cloud database such as MongoDB or Firebase to do storing and processing.

2. Data cleaning problem
   US accidents dataset contains features such that location of accident (latitude, longitude, state, county), description of accident, weather condition of accident, severity level of accident, etc. Many features (columns) are useless, also many data points have missing values. For those, I need to do data cleaning process.

3. Data integration
   Since I want to use natural disaster's records to analyze the underlying weather condition influence on car accidents. I will need to integrate the two datasets according to location information. More specifically, in a certain time and location, one of natural disaster happened, I will extract the records during this time and location on car accidents dataset. And analyze the severity and/or frequency of car accidents.
   In this process, Spark would be very helpful for extracting the records under some constraints.

4. Data analysis
   After I extract some meaningful records from the dataset, Spark can be used to apply some machine learning technique for parallel processing. For example, I want to predict the severity of car accidents given a set of features. This needs to train machine learning model, not only to extract meaningful features, but also to process data in a parallel way. In this case, Spark is very efficient for transforming the data to some smaller spaces, training the model, and give prediction.

• Databases to be used and how to use them

1. Firebase
   Firebase cloud database is going to be used for storing and processing large dataset, i.e. US accidents dataset on cloud.
   In terms of how to use it, I will first load the whole dataset on it. Do data cleaning to clear some missing value records, and get rid of some columns that are not related to this project. Then retrieve data records that is under some certain constraints, such as states and time. Those data records are the car accidents happened when some natural disaster happened at that time and states. This is also how I integrate the two datasets.

2. Relational database
   Store US natural disaster dataset. Since natural disaster dataset is only 14MB, I can only use relational database to do the store and processing. Similarly, I will first clean the dataset, clear some columns that are mostly flags, place code, etc. Then I will retrieve data and do the analysis, such as analyzing the natural disaster frequency by year, location etc.

• Team members, background and skills.

Yao Zhu, yaozhu@usc.edu

I'm currently a 2nd year PhD student in ECE department under supervision of Professor C.-C. Jay Kuo. My research area is image forensics. Currently I'm working on image stenagalysis using explainable machine learning techniques. I'm fluent in python and have been using python for 3 years. I have solid knowledge of machine learning, and have had several image-related projects using machine learning algorithms. But I'm new to big data, database, data management, data mining etc. I'm eager to learn more about parallel computing, high performance computing, etc. What's more, I don't have any front-end development experience, nor have I taken web technologies course. I will try my best to learn and create a web browser based interface for users to search, extract, and explore the data.

• Milestones and timelines

September 28th: finish creating database, link of two database, basic analysis of individual dataset, learn how to create web browser interface

October12th : Project midterm report

October 26th : finish all analysis of two dataset using spark, learn how to create web browser interface

November 9th : finish establishing web browser interface

November 23rd : Final project demo