Roberto Reyes
CS591 T1
Homework 6

Data Mining Tools and Techniques Homework 6:

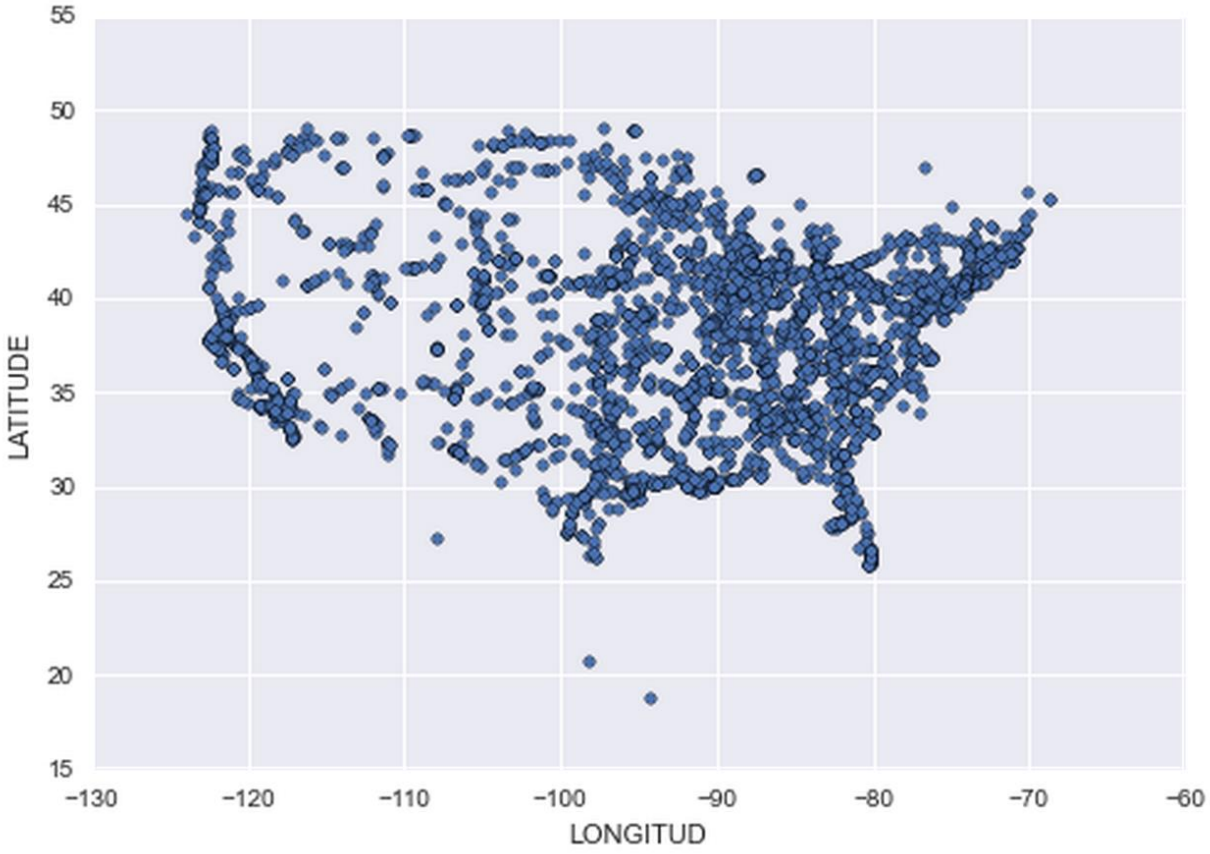
I am using a dataset obtained from the Federal Railroad Administration Office of Safety Analysis. The dataset gives information on railroad-related casualties in the United States. I am focusing on data gathered from January 1, 2014 to December 31, 2014.

The link for the dataset is:
https://safetydata.fra.dot.gov/OfficeofSafety/publicsite/Query/casrpt.aspx.

The fields in the dataset and their respective fields are:

- Railroad company (256 values)
- Date of incident
- Condition of casualty (406 values)
  - Various injuries
  - Trauma
  - Fatalities
- Cause of casualty (80 values), including:
  - Equipment malfunction
  - Person slipped
  - Stabbing
  - Highway-rail collision
- Type of victim (7 unique types)
  - On-duty employee
  - Off-duty employee
  - Tresspasser
  - Non-tresspasser
  - Passenger on train
  - Contractor
  - Volunteer
- Job, if employee
- Age of Victim
- State / County
- Latitude /Longitude - I have included here a map of all the casualties to illustrate the geographic distribution of the data points.

Roberto Reyes
CS591 T1
Homework 6



Roberto Reyes
CS591 T1
Homework 6
The total number of records is 9215.

The fields in the dataset are:

Field	Number of Missing Values
Railroad Company	0
Date	0
Condition	0
Cause / Event	0
Type of Victim	0
Employee Job	0
Age	616
State	1
County	0
Long / Lat	4730

The geospatial data might not necessariy be required, depending on which railroad company. For example, one such event that is missing latitude/ longitude information occurred on the PATH train. The PATH train is 13.8 mile long system operating between NJ and NYC. Depending on the hypotheses to be tested, such specific geospatial data might not prove to be useful.

The only other data missing in this dataset is the age of the victim, and less than 10% of the population is missing. The worst case scenario, I would discard these records if age is an important variable.

Hypotheses

I am interested in exploring the link between railroad companies and casualties such as injuries and fatalities. Several railroads I am familiar with are in the dataset, such as the MBTA, Path, MetroNorth. One hypothesis to explore is if there is a link between more populated states and a certain type of casualty.

Another interesting point to explore are trans-state railroad services. I would like to explore whether certain railroads are prone to casualties in certain areas or uniformly across their areas of service.

There is also the option of exploring seasonal variation on casualties. It is possible that casualties are more common in the colder and hotter months when mechanical failures can be expected.

Several more hypotheses can be explored with this dataset.