Project idea #1: Exploratory analysis of "Electric Disturbance Events Annual Summary" data

Utility companies report power outages and other electric disturbance events to the Department of Energy (DoE). The DoE then publishes aggregated reports in a table format, dating back the 2000. For each event, the data includes: start and end dates of the incident, time of restoration, affected area, type alert that was issued, type of event, and the number of customers affected. Analyzing this data can provide us insights into the following:

a. How did the response time change over time? Which type of events did become more/less frequent in time?
b. How long it takes to respond an incident based on the type and location of the event? Knowing this information, companies can issue alerts to inform the customers how long they will have to wait before the power is restored.
c. Company resources can be diverted to areas that are frequently affected.
d. Is there periodicity in the events based on the time of the year? If so, utility company can increase workforce to prepare for incidents that will most likely occur.

Project idea #2: YouTube comments detection: spam or ham?

The dataset includes 1956 real messages collected from five popular YouTube videos. The data has 4 attributes (comment id, author, date, and content) and 1 output (class). A sample data shown in table format below:

COMMENT ID	AUTHOR	DATE	CONTENT	CLASS
z13nuzezjmfpcjqzm04cftuaxpb3gjmxevw0k	Charles Baptist	2014-08-24T03:57:52	Help Please!! http://www.gofundme.com/RJanimalcareï»¿	1 (spam)
z121tz2zhzjgercem23yttsqvnuijljql04	Daniel Korp	2014-08-24T17:17:17	katy perry does remind me of a tiger,like as if its her spirit animal :3 <3ï»¿	0 (ham)

The comments section in YouTube videos can easily be populated by spams. People who share videos for constructive feedback and watchers who like to spend time in comments section can get frustrated and discouraged by these spams, which can affect their engagement, yielding a decrease in the profits. Furthermore, scammers can use comments section as a platform to deceive people and spread false information. Therefore, it is important to detect spams and eliminate them to protect viewers and preserve website's credibility. My approach would be analyzing the spam content for commonly occurring words and word combinations, word ordering, use of pronouns, requests, links, etc. Finally, train a classification algorithm to distinguish between spam and ham comments.

Project idea #3: Analysis of bike sharing data

Bike sharing systems are becoming more attractive in big cities for they provide affordable and practical means of transformation. Aside from their actual use as vehicles, they can also collectively act as a sensor network providing real-time information regarding mobility in the city, such as traffic flow and accidents. We can combine bikeshare data with publicly available datasets such as: weather data, news data, and gas prices to evaluate customer behavior and detect major events. The dataset was taken from the Capital Bikeshare, a Washington DC based company. Two sample rows from the data are given below:

Duration	Start date	End date	Start station number	Start station	End station number	End station	Bike number	Member type
221	1/1/2017 0:00	1/1/2017 0:04	31634	3rd & Tingey St SE	31208	M St & New Jersey Ave SE	W00869	Member
1676	1/1/2017 0:06	1/1/2017 0:34	31258	Lincoln Memorial	31270	8th & D St NW	W00894	Casual

a) Exploratory data analysis:

- Where do the riders go, how far they travel, how long they travel?
- Which stations are more popular? What times in the day and what times of the week are most rides taken on?
- Do people prefer casual rides? How does the membership numbers change over the years and throughout the year?

b) Predictions and event detection

- Predict the number of rentals on a given day based on the seasonal changes
- Detect major events from the bikeshare activity (first, find events that occurred in the past by doing a web search)