Investigating Pipeline Spills and Leakages in the United States from 2010-2016

Anne Evered

MUSA620, Fall 2019

Contents

1 Introduction

For the final project, I investigated the number and characteristics of oil pipeline leaks and spills across the United States from 2010 to 2016. In particular, I created two dashboards with several different visualizations or maps each that explore various features of these incidents. This report discusses the specific data and methods involved in the project, including techniques used in the collection, analysis and visualization steps. Finally, I discuss the main questions driving each of the visualizations on the dashboards, as well as key insights that can be seen in each.

2 Data and Methods

2.1 Dataset Overview

The primary dataset used for this analysis contains information on each reported oil pipeline leak or spill in the United States from 2010 to 2016. While the original data is collected and published by the Pipeline and Hazardous Materials Safety Administration, the specific dataset for this analysis is an aggregated version of this data available on Kaggle.[1] The dataset has a row for each incident and includes geolocation information, as well as details such as the company involved, the type of pipeline and liquid, the number and type of injuries involved, and the breakdown of various monetary costs.

In addition to the pipeline incidents dataset, county level shape files were also used, as well as county level demographic and income measures from the American Community Survey (2013-2017). These additional datasets (joined to the incidents dataset) were used for examining demographic trends related to the geographic distribution of the pipeline incidents.

2.2 Variables of Interest

Of the variables in the pipeline incidents dataset discussed above, several were of key interest for this analysis. Accident Year gives the year the incident occurred, while Accident Latitude and Accident Longitude give the location. Two other variables are explored in depth in the analysis: the Cause Category, which provides a general cause of the incident and Operator Name, the operator responsible for the pipeline involved. These variables will be discussed further in the data cleaning and data visualization sections.

2.3 Data Exploration and Cleaning

While the pipeline accidents dataset downloaded from Kaggle was relatively clean to begin with, several additional data preparation steps were taken based on exploratory analysis prior to creating the final visualizations.

For example, grouping the incidents by year showed that the dataset has many rows for the years 2010-2016, but that there are only two incidents in the dataset reported to have occurred in 2017. Given that all 2017 reported incidents do not appear to be in the data, these two incidents were dropped and the focus of the analysis limited to years 2010-2016.

In addition, initial data exploration suggests possible data quality issues with the fields providing fatality and injury information about the pipeline incidents. Compared to similar analyses, the values for these fields appear very low. Therefore, while I was initially planning on including visualizations comparing fatalities and injuries by location, year, and operator, given the potential data quality issues with these fields, I did not include these visualizations in the final report.

Initial data exploration also revealed a high degree of skew in the distribution of number of incidents per state (with many states having no incidents, particularly when looking at a single year, and some states having dozens of incidents). Given this distribution, I used log scaling for visualizing the number of incidents per state (discussed later, as well).

Along with the general data exploration and cleaning, I took additional data manipulation steps for the particular visualizations. For example, for the visualizations that also include census information, I joined the incidents dataset on county ID to the census dataset and used a geospatial join to add in the county shape files.

3 Visualizations

Following the data cleaning and preparation and doing some exploratory analysis, I created two data

[1] https://www.kaggle.com/usdot/pipeline-accidents

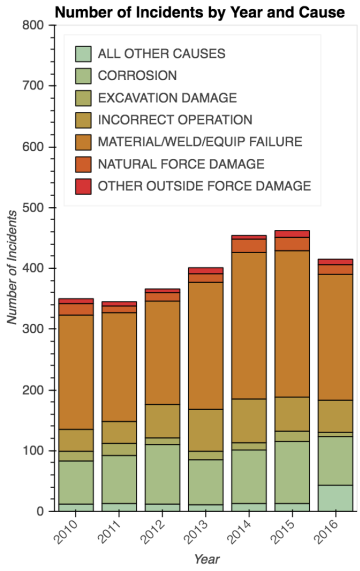
dashboards that address specific questions or components of the data. The first focuses on questions related to the pipeline incidents themselves, such as cause and operator, and the second looks at how the locations of the incidents are related to various demographic features. For both dashboards, I used panel to combine and display the different visualizations. Some of these visualizations are similar to a 2012 ProPublica report[2] on pipeline incidents (but which only included data up to 2012), and others are additional analyses I was interested in including.

In this section, I will discuss each of these visualizations in detail, including the tools used, the question focused on and some key insights that can be seen from the visualizations.

3.1 First Dashboard

3.1.1 Incidents by Year and Cause

The first visualization is a stacked bar plot that shows the number of incidents by year broken down by cause. To create this visualization, I first grouped the dataset by Accident Year and Cause and took a count of the Report Numbers. A static view of this visualization can be seen below.

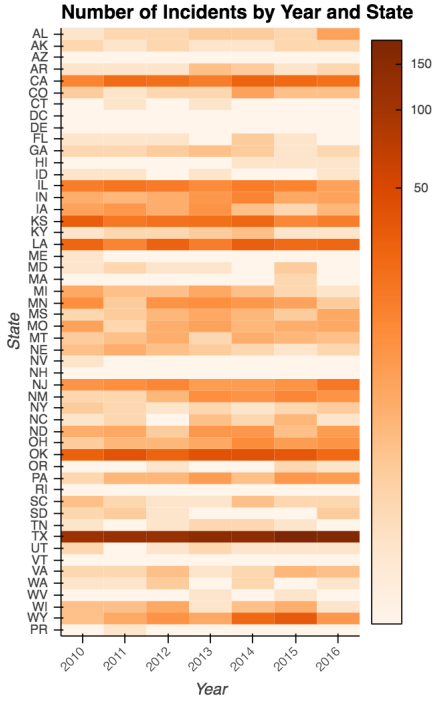


Like the other visualizations in the dashboard, I used Holoviews as the data visualization package for this chart, specifically using the hv.plot function and customization options. The color map for the stacked bar graph was chosen using an online color picker tool. This chart also uses the built in Holoviews hovertool functionality to allow the user to see the number within each breakdown by scrolling over the bar plot.

This plot suggests that the main cause category for pipeline spills and leakages in the United States is Material or Equipment Failure (material/weld/equip failure), followed by Corrosion and Incorrect Operation. This is true across all of the years included in the analysis. The graph also suggest a general increase in the total number of incidents in the years from 2011 to 2015 with a slight decrease in 2016.

3.1.2 Incidents by Year and State

The second visualization I created is a heatmap that shows the distribution of number of incidents by year and state. Again, a static version of this chart is provided below.



Similar to the stacked barplot, to create this visualization, I first grouped the dataset by Accident Year and State and used count of the Report Numbers to get the Number of Incidents. I again used Holoviews as the data visualization package for this chart, specifically the hv.HeatMap function. As mentioned in the data exploration sec-

[2] https://projects.propublica.org/pipelines/

tion, I used a log scale for the color scale/legend given that the distribution of number of incidents by state and by year has a high degree of skew. I used a single hue gradient colormap in order to allow the viewer to visually see the scale differences in number of incidents across year and state.

In creating the first version of this visualization, I also realized that for years and states with no incidents, since there are no rows in the data set for those categories, the aggregate tables after the group by did not include them either. This meant that when I originally plotted the heatmap, some of the states (those with zero incidents) were showing as missing (i.e. no color). To handle this, I added a row in the aggregated year and state dataset for each year-state combination not already in the dataset and set the value for count of number of reports to zero.

As discussed briefly already, and as might be expected, this map shows that for all of the years considered, Texas has a significantly higher number of pipeline leakages and spills than other states, with some states such as New Hampshire and Massachusetts having almost no incidents. This visualization and analysis might be improved by also considering the number of pipelines or operators or some other measure of frequency within each state and standardizing the number of spills/leakages by that number. For instance, Texas has more oil industry activity in general than most other states, so it is difficult to tell if this is the main reason why there is such a large discrepancy.

3.1.3 Incidents by Operator

I also added a table to the dashboard, using the Holoviews hv.Table function, that shows the number of incidents over the time range considered (2010-2016) by operator. I grouped the dataset by Operator Name and again took a count of the Report Number to get the Number of Incidents. This table allows the user to sort either by the Operator Name alphabetically or by the Number of Incidents to explore this variable. An image of this table can be seen below.

Number of Incidents by Operator

#	Operator Name	Number of Incidents
0	ALYESKA PIPELINE SERVICE CO	8
1	AMOCO OIL CO	3
2	ASIG - HONOLULU	1
3	BELLE FOURCHE PIPELINE CO	13
4	BHP BILLITON PETROLEUM (EAGLE FORD GATHERING) LLC	1
5	BKEP CRUDE, LLC	2
6	BKEP PIPELINE, LLC	7
7	BLUE RACER MIDSTREAM, LLC	1
8	BP OIL PIPELINE CO	1
9	BP PIPELINE (NORTH AMERICA) INC.	10
10	BP PIPELINES (ALASKA), INC	2

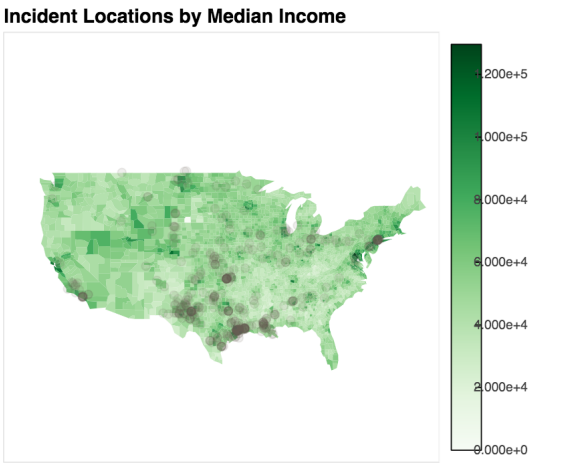
3.1.4 Incidents by Location

Lastly, the first dashboard also includes a map that shows the location distribution of the pipeline spills and leakages that can be filtered by the year of the incident.

Additional analysis of the location distribution of the pipeline spills, particularly as it is related to various demographic features, can be further explored in the second dashboard. This dashboard is described in more detail below.

3.2 Second Dashboard

The visualizations in the second dashboard show pipeline leakage and spill locations against county level census data from the 2013-2017 American Community Survey. Each visualization focuses on a particular census variable, including median income of the county and percent of the population identifying as various race categories. These maps only include 2016 pipeline spills and leakages. Initially, I included all years in the dataset in the maps, but later filtered to just 2016 to more easily view trends. As discussed briefly in the data preparation section, the data for these maps comes from joining the pipeline incidents dataset to county level census information and doing a geospatial join to get the county shape files. An example of one of these graphs (for Median Income) can be seen below.



The purpose of these visualizations is to allow for exploring trends of how the pipeline locations relate to demographic features. While it is somewhat difficult to see trends on these maps when viewing the United States as a whole, they can be zoomed in on (all of the maps zoom in simultaneously) to explore specific regions.

4 Summary and Next Steps

4.1 Key Insights

Together, these two dashboards allow the viewer to explore various questions related to the prevalence of oil leaks and spills in the United States over the given time period, such as comparing number of incidents geographically and over time. For instance, as discussed in this report, we can see an increase in number of incidents from 2011 to 2015 and that material and equipment failures are the most common cause category of oil spills and leakages. We also saw that there is a wide variation in where most of these incidents are located within the United States. The exploratory analysis stage of this project also raised potential data quality concerns with certain fields in the dataset, such as the number of injuries and fatalities.

4.2 Further Analyses

While the dashboards provide many different cuts of the data, the dataset contains many other variables that could also be explored. Additional visualizations and analyses could be conducted, such as exploring the more detailed cause data, doing more analysis on operators involved, and looking at cost data. For example, additional visualizations might include:

- A visualization showing a breakdown of incident costs by state, year, or operator with a drop-down allowing for selection of different types of costs, as well as total cost.
- Allow the user to select a particular incident on the location map and filter a table based on the selection that gives additional details (operator, cause, type of accident, etc.) about that incident.
- An investigation into how the geographical distribution of the incidents are related to operator, monetary costs, or other details. For instance, are the particular operators involved clustered in certain locations?
- Other visualizations exploring variables not considered in this analysis, such as liquid type, barrels released, evacuations and different types of monetary costs.

In addition to these other visualizations and analyses (of which there are also many others, as well!), another possible next step might be to further validate the incidents dataset used. For example, this dataset could be compared against other pipeline spill and leakage datasets or analyses to better understand the reporting accuracy and completeness of the dataset.