New Mexico Oil and Gas Field Spill Incidents Inferential Statistics Analysis

The New Mexico oil and gas field spills data collected by Oil Conversation Division (https://wwwapps.emnrd.state.nm.us/ocd/ocdpermitting/Data/Spills/Spills.aspx) is used to study the correlations among causes, impacts, locations and operators of incidents. This document is submitted to my GitHub account together with Jupyter Notebook file "New_Mexico_Oil_and_Gas_Field_Spill_Incidents_Inferential_Statistics_Analysis.ipynb".

Step 1: Data Import and Wrangling

Import the table of spills containing oil and gas field spills data in New Mexico, and the columns are going to be analyzed are:

Incident Number: incident identifier
Facility: facility identifier if the incident happened in a facility
API: well identifier if the incident happened in a well
Operator Name, Severity, Incident Type, Incident Date, Material Spilled, Volume Spilled, Volume Recovered, Spill Cause, Spill Source, District, County, Waterway Affected, Ground Water Impact
Then Forward and backward fill the NaNs in the table with the information of the same incident, drop duplicated rows, and convert the incident date as pandas datetime and set it as the index of the table.

Step 2: Exploratory Data Analysis

From EDA, it's found that the major incidents and minor incidents have strong correlation between each other, from the incident number, types, causes, spilled materials and volumes, ground water and waterway impacts, and incidents number caused by operators.

Since the major incidents tend to spill more volumes of materials and have more impact on groundwater and waterway, I'd like to further analyze whether the facility or well has a higher probability to cause a major incident when an incident occurred.

Step 3: Inferential Statistics

To answer the question above, both bootstrap and z proportion tests are conducted.

The null hypothesis is the probability of an incident is a major incident is equal in facilities and wells; and the alternative hypothesis is the probability of an incident is a major incident in wells is larger than in facilities, assuming the significance level is 5%.

Step 4: Conclusion:

1. The probability that an incident is a major incident in wells is larger than in facilities.

2. There are more major incidents in the history since 1980s. The monthly number of incidents has increased to around 60/month in 2018.
3. The volume spilled in minor incidents (below 50 barrels) is much less than the volume spilled in major incidents (up to 100,000 barrels).
4. The correlation between monthly major and minor incidents indicates a positive linearly relationship between major and minor incidents.
5. The top three incident types are produced water release, oil release and natural gas release which is corresponding to the top three spilled materials; The top three spill causes are equipment failure, corrosion and human error; and the top three spilled sources are tank (any), flow line - production, and pipeline (any). Fire, triethylene and sulphuric acid spill, and generator have 100% probability to cause major incident.
6. The ground water is more likely impacted by major incidents than minor incidents, and the waterway are 100% affected by major incidents.
7. The most of major and minor incidents happens in county Lea in Hobbs, Eddy in Artesia, San Juan in Aztec, and 0 (missing county name probably) in Santa Fe.
8. There is a positive linear relationship between the total number of major and minor incidents from each operator. COG OPERATING LLC and EOG Y RESOURCES,INC. are the first two operators cause most of incidents in New Mexico.