main

September 18, 2025

[3]:
```
import pandas as pd
import numpy as np

# Load data from CSV file in the folder
df = pd.read_csv('downsample.csv')
# down sample to half of size to be less than 10mb
# df_half = pd.read_csv('serious-incidents-expensive.csv')
# df_half = df_half.sample(frac=0.5, random_state=42).reset_index(drop=True)
# df_half.to_csv('downsample.csv', index=False)
# Drop columns where most values are missing (e.g., more than 80% are NaN)
threshold = 0.8
missing_ratio = df.isnull().mean()
cols_to_drop = missing_ratio[missing_ratio > threshold].index
df = df.drop(columns=cols_to_drop)
X = df.drop(columns=['Total Amount Of Damages'])
y = df['Total Amount Of Damages']
```

```
/var/folders/8f/7hlpwq3n64b0x53crzr0w9pc0000gn/T/ipykernel_79286/4090986999.py:5
: DtypeWarning: Columns (11,100) have mixed types. Specify dtype option on
import or set low_memory=False.
  df = pd.read_csv('downsample.csv')
```

[4]:
```
X
```

[4]:
```
     Report Submission Source Multiple Rows Per Incident  \
0                       Paper                         No
1                       Paper                         No
2                       Paper                        Yes
3                       Paper                         No
4                         Web                         No
…                         …                          …
5113                      Web                        Yes
5114                    Paper                         No
5115                    Paper                         No
5116                    Paper                         No
5117                      Web                         No

                                          Report Number  \
```

```
0     <a href = https://portal.phmsa.dot.gov/PDFGene…
1                                          I-1991020567
2     <a href = https://portal.phmsa.dot.gov/PDFGene…
3     <a href = https://portal.phmsa.dot.gov/PDFGene…
4     <a href = https://portal.phmsa.dot.gov/PDFGene…
…                                                   …
5113  <a href = https://portal.phmsa.dot.gov/PDFGene…
5114  <a href = https://portal.phmsa.dot.gov/PDFGene…
5115  <a href = https://portal.phmsa.dot.gov/PDFGene…
5116  <a href = https://portal.phmsa.dot.gov/PDFGene…
5117  <a href = https://portal.phmsa.dot.gov/PDFGene…

                      Report Type Date Of Incident  Time Of Incident  \
0     A hazardous material incident       2004-12-06            1330.0
1     A hazardous material incident       1991-02-05            1020.0
2     A hazardous material incident       2000-11-04            2230.0
3     A hazardous material incident       1990-03-10            1215.0
4     A hazardous material incident       2011-06-22            1715.0
…                               …              …                …
5113  A hazardous material incident       2005-07-19            1830.0
5114  A hazardous material incident       2019-03-12             500.0
5115  A hazardous material incident       1993-11-15            1800.0
5116  A hazardous material incident       1995-08-04             330.0
5117  A hazardous material incident       2009-10-06             810.0

      NRC Report Number Incident City Incident County Incident State  …  \
0                   NaN      LA PORTE          HARRIS             TX  …
1                   NaN    SEELYVILLE            VIGO             IN  …
2              547322.0   SCOTTSBLUFF    SCOTTS BLUFF             NE  …
3                   NaN    CLOVERDALE          SONOMA             CA  …
4              980481.0        SONORA          HARDIN             KY  …
…                   …             …               …             …  …
5113                NaN      EASTOVER        RICHLAND             SC  …
5114                NaN   E LEWISBURG  NORTHUMBERLAND             PA  …
5115                NaN   ORLAND PARK            COOK             IL  …
5116                NaN       EMPORIA            LYON             KS  …
5117           919848.0   GREEN RIVER      SWEETWATER             WY  …

     Hmis Serious Fatality Hmis Serious Injury Hmis Serious Flight Plan  \
0                       No                  No                       No
1                       No                  No                       No
2                       No                  No                       No
3                       No                  No                       No
4                       No                  No                       No
…                      …                   …                        …
5113                    No                  No                       No
5114                    No                  No                       No
```

```
5115                    No                  No                         No
5116                    No                  No                         No
5117                    No                  No                         No

     Hmis Serious Evacuations Hmis Serious Major Artery  \
0                          No                        No
1                         Yes                        No
2                         Yes                        No
3                          No                        No
4                          No                       Yes
…                         …                          …
5113                       No                        No
5114                       No                       Yes
5115                       No                        No
5116                       No                        No
5117                      Yes                       Yes

     Hmis Serious Bulk Release Hmis Serious Marine Pollutant  \
0                          Yes                           No
1                          Yes                           No
2                          Yes                           No
3                          Yes                           No
4                          Yes                           No
…                         …                             …
5113                       Yes                           No
5114                        No                           No
5115                       Yes                           No
5116                       Yes                           No
5117                       Yes                           No

     Hmis Serious Radioactive Contact Business Name Contact Country
0                          No                   NaN              US
1                          No                   NaN              US
2                          No                   NaN              US
3                          No                   NaN              US
4                          No                  ERTS              US
…                         …                     …               …
5113                       No    Finnchem USA, Inc.              US
5114                       No                   NaN              US
5115                       No                   NaN              US
5116                       No                   NaN              US
5117                       No      PC TRANSPORT INC              US

[5118 rows x 162 columns]
```

[5]:
```
y
```

```
[5]: 0         18500
     1        121013
     2       3074000
     3         10030
     4         10000
              …
     5113      61500
     5114      53000
     5115      40225
     5116      50000
     5117      28500
     Name: Total Amount Of Damages, Length: 5118, dtype: int64
```