FEMA Data Analyst Interview – Public Assistance Funded Projects

Sabina Bhuiyan

Dataset and Software Used

Public Assistance Funded Projects Details (V1)

  - https://www.fema.gov/openfema-data-page/public-assistance-funded-projects-details-v1

- Tools used:
  - Analysis: Python (Jupyter Notebooks)
  - Visualizations: Power BI

Part 1 – Analysis with Python

- Target: Project Amount
- Initial observations:
  - 22 columns and 780894 rows
  - There are some integer and float datatypes, while others are object
  - Average project amount is $269,280

```
#Distribution of project amount
df['projectAmount'].describe()
```

```
count    7.808940e+05
mean     2.692804e+05
std      1.668811e+07
min     -3.726871e+08
25%      3.521775e+03
50%      1.079688e+04
75%      3.810900e+04
max      9.553782e+09
Name: projectAmount, dtype: float64
```

Part 1 – Analysis with Python

- One-hot encoding using pandas
  - Transform categorical (i.e. type object) attributes to binary
  - 'incidentType', 'dcc', 'projectSize', 'state' attributes chosen as the variables that would most likely affect project amount

```
dummy_df = pd.get_dummies(df, columns = ['incidentType', 'dcc', 'projectSize', 'state'])
dummy_df.head()
```

	projectAmount	projectSize_Large	projectSize_Small	incidentType_Biological	incidentType_Severe Storm	state_Puerto Rico	dcc_C	dcc_F
0	-1351843.17	1	0	0	0	0	0	0
1	3161467.69	1	0	0	0	0	0	0
2	2497127.50	1	0	1	0	0	0	0
3	3030387.44	1	0	0	0	1	0	0
4	0.00	1	0	0	0	0	0	0

Part 1 – Analysis with Python

- Determining most correlated features to target variable from encoded attributes

```
# Put the target back in the dataframe
dummy_df['projectAmount'] = df['projectAmount']
# Correlations in one-hot encoded dataframe
dummy_df.corr()['projectAmount'].sort_values(ascending=False)
```

  - Calculated using the correlation coefficient, or r-value

```
projectAmount                1.000000
federalShareObligated        0.998575
totalObligated               0.998564
projectSize_Large            0.033624
incidentType_Biological      0.024007
```

  - Unnecessary attributes were then dropped from the dataframe, such as federal share obligated, applicant id, and pw number. Since this analysis is focusing on disaster factors that directly affect the project amount granted to each incident, these factors were dropped

Part 1 – Analysis with Python

- Determining most correlated feature to target variable after combining categorical and numerical attributes
  - Calculated using the r-value

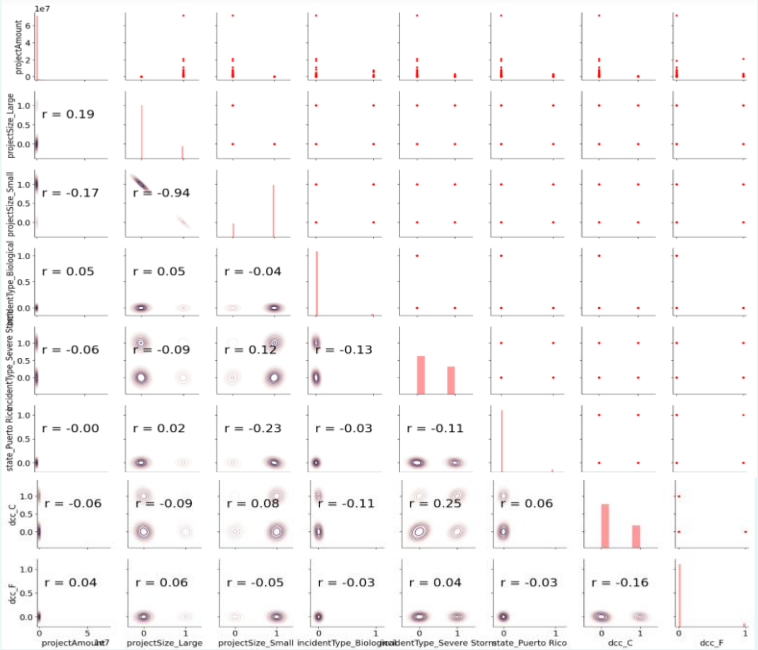
```
Most Correlated Features:

projectAmount                    1.000000
projectSize_Large                0.033624
projectSize_Small                0.031789
incidentType_Biological          0.024007
incidentType_Severe Storm        0.010989
state_Puerto Rico                0.009746
dcc_C                            0.008627
dcc_F                            0.006780
Name: projectAmount, dtype: float64
```

Part 1 – Analysis with Python

- Visualizing correlations via pairplot using seaborn

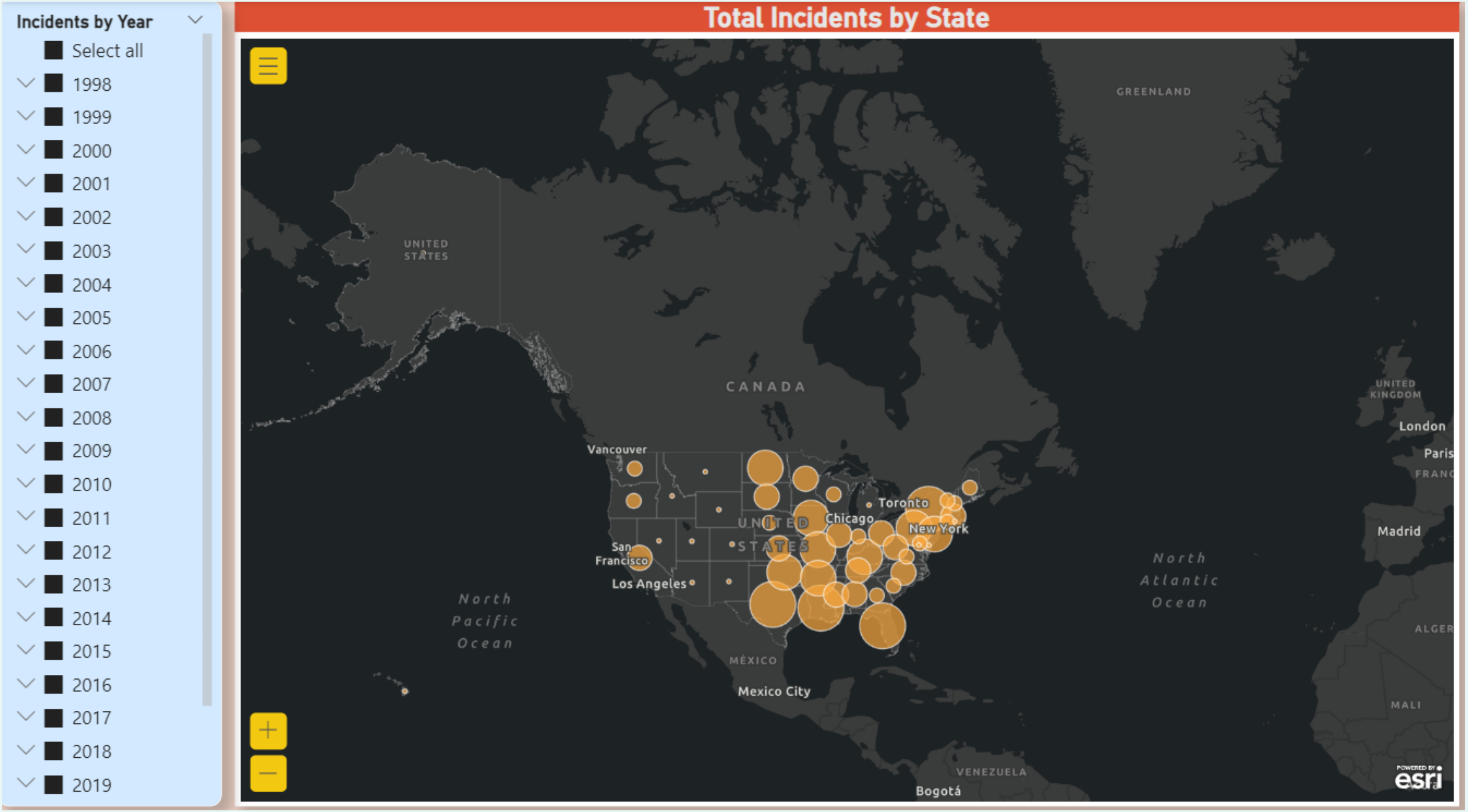
Most correlated features:
- Large project size
- Small project size
- Biological incident
- Severe storm incident
- State of Puerto Rico
- DCC C – Roads and Bridges
- DCC F – Public Utilities



Part 2 – Creating Visualizations with Power BI

Incident Map Dashboard:

- ArcGIS map with slicer – visualize total incidents by state per year
  - Location: state
  - Size: count of incident type
  - Tooltips: count of incident type
- Slicer – filter by year, damage category code, and incident type

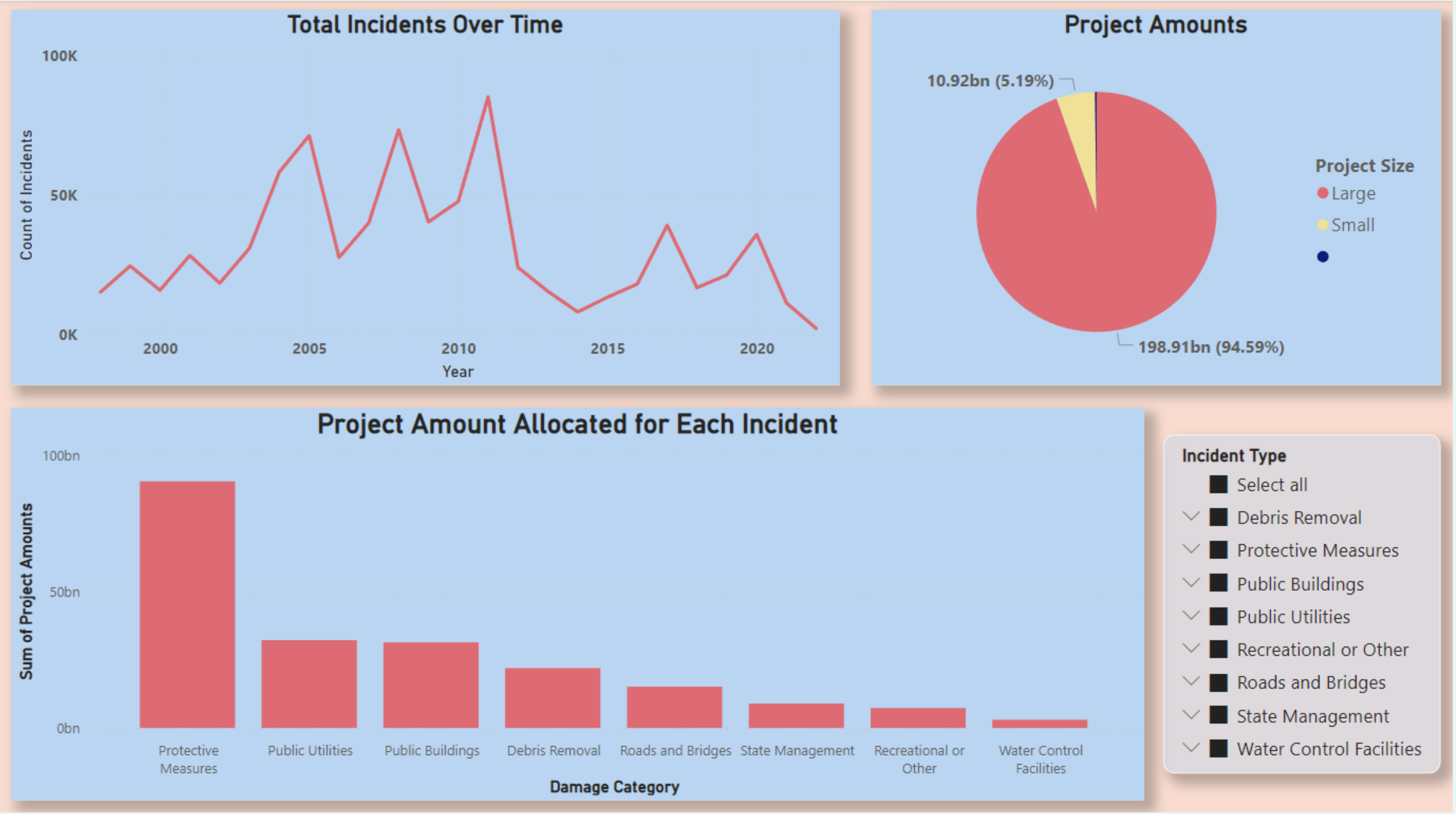


Map of total incidents by state

Part 2 – Creating Visualizations with Power BI

Statistics Dashboard:

- Line chart – visualize total incidents over time
  - X- axis: year
  - Y- axis: count of incidents
- Pie chart – visualize project amount spent for projects of various sizes
  - Values: sum of project amount
  - Legend: project size
- Stacked column chart – visualize project amount spent on each incident
  - X-axis: damage category
  - Y-axis: sum of project amount
- Slicer – filter by damage category code and incident type



Visualizations of public assistance per year and incident type

Conclusions

- There were more incidents in the eastern part of the country
  - Highest in 2011
- Most incidents fell under protective measures
- Factors that have a direct impact on project amount allocated are project size, biological incident, severe storm incident, and damage control codes C – Roads and Bridges and F – Public Utilities
- Future work could be done to mitigate these issues in respective states to prevent further incidents