# Summary

Conclusions

* Of the incident types, vehicle collisions (multi or single vehicle collisions) seem to be more likely to be fraud cases with 27% and 29% of cases being fraud, respectively, compared to parked car and vehicle theft, which have 9.5% and 8.5% cases being fraud respectively.
* In general, incidents with major severity are more likely to be fraud (60.5%) than legitimate. They made up 27.6% of the cases in total, but 67.6% of all fraudulent cases.
* The location of major incidents can tell us even more about the likelihood of the case being fraud. Some locations where major incidences occurred have 100% fraud cases, which include Northbend, NC (5/5), Arlington, OH (4/4), Columbus, OH (1/1), Northbrook, OH (2/2), Riverwood, OH (1/1), Columbus, PA (1/1). The locations with the highest number of fraud cases are Arlington, SC (11/12, 92%), Columbus, SC (11/16, 69%), Springfield, NY (9/11, 82%), Springfield, SC (8/13, 61.5%), Arlington, NY (7/8), 87.5%).

Techniques used

* Excel
* Opened up csv file on Excel
* Identified potential duplicates using conditional formatting to highlight any duplicate cells in the policy number column (as each policy number should be unique). No duplicates were identified.
* Selected the whole dataset and clicked remove duplicates. Still no duplicates removed.
* Standardised data – kept all data title cased, and changed abbreviations to full words (e.g. Y and N to Yes and No) – used find and replace function to do this by selecting column and control h.
* Removed the \_c39 column.
* Used pivot tables to identify possible patterns and relationships in the data.
* MySQL
* Converted excel dataset to csv file and imported it to MySQL
* Investigated missing values in the collision\_type column – turns out all the missing values were there for incident types “vehicle theft” and “parked car” which weren’t applicable to the collision type column, so changed the missing values “?” to “n/a”.
* Created a new column which combined incident state with incident city.
* Modified the date columns to the correct format
* Python (Jupyter notebook)
* Inspected the data
* Calculated the percentage of fraud insurance claims
* In the collision\_type column, “n/a” was being treated as a missing column, so changed it to “no collision”.
* Carried out a chi-square test on incident severity and fraud reported, and found a significant association between the two, where major damage incidents were significantly more associated with fraud reports than any other incident severity type.
* Tableau Public
* Used tableau public to create a dashboard to visualise the data.