**Problem Statement**

In the USA, drivers are required to have at least third party insurance. Comprehensive insurance is also available and covers damage to the insured car & injury in addition to third party insurance.

Insurance Companies are concerned with paying out fraudulent claims. The purpose of this project is to quantify the impact of fraudulent claims and devise a decision making model to detect validity of claims and whether more investigations is necessary to prevent fraud.

1. Clean data
2. Quantify or analyse impact and frequency of fraudulent claims.
3. Design and deploy a model that can detect claims that are likely to be fraudulent and therefore require further validations before paying out.

**Clean data:**

Missing Data & summaries in excel and Power BI

*I used Pivot tables and PowerBI dashboards to impute missing data. See comments in excel for the way data was cleaned.*

There are 22 missing values in the overall data.

21 rows, each missing only 1 value and 1 row missing 2 inputs (policy\_annual\_premium & incident\_state)

10 columns are missing data in the range of 1 to 4 values.

Conclusion: Given a 1000 rows and 39 columns, having only 22 missing values spread across variables and row labels, we can impute this data rather than removing it altogether.

178 collision types were fill in as “?”, these were replaced by “Unknown”.

360 property claim ?. Replace with Yes if there is property claim value, with No if property claim value is zero.

343 records have ? on policy report. Made these “Unknown”