Data quality report FOD

# Executive summary

# Approaches taken for data understanding

# Found data quality violations and applied fixes

# Approaches taken to deal with missing data

**Primary\_Agency\_ID**: When this column data is missing/9999 we are going to replace it with the Agency\_ID. We can not find any way to associate a primary agency with the agency in question so will replace it with its ID.

**Prod\_abbr:** Mapping the values to numbers, we threat the 12 versions of products as different products.

**Prod\_Line**: We will be splitting the dataset into a separate PL and CL dataset, this is because for whichever line you use, there is different data available.

**State\_ABBR:** Mapping the values to numbers.

**Prev\_Poly\_Inforce\_QTY:** Contains 9999 values. These mean there are no previous years or this year did not have data (CL)

**(NB) WRTN\_PREM:** Calculate average based on Year, Line, State and product. Add this to a new column to save old data. This calculated might not be accurate but adds a lot to the dataset.

**PREV\_WRTN\_PREM:** Adjust to the amount of months to be 12, when 0/9999 generate based on previous year. If there is no previous year, take 0.

**Months:** All values that are not 12 will be calculated to be 12, this column becomes unused afterwards.

**Retention\_Ratio:** More than 60% of the data is missing and this column is a calculation based on 2 other columns. We will remove this column as it has no use.

**Loss\_ratio (3YR) and Growth\_Ratio 3YR:** Recalculate the values by using our new columns, keep old values for reference. These columns will be used as labels.

**Agency\_appointment\_year:** 5000 missing data, irrelevant column, will be replaced with average.

**Active\_Producers, Max\_age and min\_age:** If the amount of producers is not set (9999) this can be set to 0, this also goes for max and min age.

**Vendor\_IND:** Convert to 0 for no and 1 for yes.

**Vendor:** Map the values to numbers, change unknown to 0.

**(VENDOR)\_START\_YEAR:** Values are varied between 1994 and 2015. When the value is missing (9999) the vendor was not used. All 9999 values can be replaced with 0.

**(VENDOR)\_END \_YEAR:** The year the agency stopped using the vendor. This can either mean the vendor is never used or the vendor is still in use. If the vendor has never been used we can replace the value with 0, if the vendor is still in use we can replace it with *Hight number or 1*.

**PL\_BOUNDS/QUO\_(UNIT):** Contains a lot of missing data, if the product is within the CL, all the PL columns should be 0, other way around is vice versa. All the 9999 values can be replaced with 0.

# Any other data preparation that you performed

# Conclusion