Objective: based on all data tables available, forecast categorisation and give reasons (depends on model).

# Stage 1: Data preprocessing

## Featrues description

### File 1: Master data set (the main file we work on)

**Contains following fields:**

CNID: identifier.

ParentCN: we can at the moment not consider parent contract misclassification and focus on other aspects. – Discard;

Current\_Agency\_Abbr: possibly, as a shorter version, can be replaced through idagency.

SupplierName: important indicator. Need stemming?

PDate; ADate:don’t think matter. Discard

Status: don’t matter. Discard

SContract, EContract: merge to work out the period. Possibly construct new variable like daily rate to indicate labour hire?

ValueContract: value. Can determien SI service (if above 1 MM. can add rule?)

Oldstart:

Oldvalue:

Title: or the description of the contract. Main analysis tool. The majority can be determined. Maybe focus on this field before moving on to the rest fields.

AgencyRef: ?

austender\_cat: spscs, can be indexed. – maybe together with agency can yield certain indications? – if someone were to consistently wrongly assign categories.

Procmethod: unsure.

AtmID: linked to procurement table. – another module to import

SonId: linked to panel table, can incorporate in python directly.

AmendReason: unsure.

ATSupplierName: ?

SupplierPostcode: don’t know. Maybe indicates IT like in north sydney? Or use cluster to determine?

SupplierABN: maybe link to ABN search? Not important at the moment.

Category: **Results**

SubCategory: **Results**

Comment: **Results**

InternalComment: to standardise in the future, include.

GAPS ID: unsure;

AgencyTier:

Original\_Agency\_Abbr: figure out diff. get rid of one either.

OriginalSupplierName: figure out diff. get rid of one either.

FinYear; FinQuarter; Month; Month Name: can discard.

ValueBand: could replace value? Correlated.

## File 2: Supplier table.

## File 3: Panel Table. Can give strict rule and change some obvious ones encountered.