**With existing data**

Correlation 🡪 heatmap (pearson) or Association 🡪 chi-sq

these vars have a relationship: each of age, gend, dis, loc vs each of 7gphs

do we really care tho about introspective analysis, for the scope this time?

also, as scores are basically 3-5 (agree), all we can hope for is something like young

strongly agree but old slightly agree

*Tokenisation (tag, lemmatise, noun chunking)*

*tag will identify nouns, verbs, etc.*

*lemmatisation will identify roots, e.g., do/doing/done*

*noun chunking helps show context*

*unsure if we need this level of granularity*

*Confidence scores*

*didn’t show much, just that we’re 95% sure in those means*

**With more data**

5 Topic models

was v valuable this time

*Clustering, respondent profiling (km, gmm, db/hdb/opt, agg, msc, birch, aff, spc, llm)*

*may be hard as scoring mostly 5s, how do you differentiate?*

*6 predictive, Pr(score | these demographics) (nbc, svm, knn, dtree, rf, bst), iterative*

*understand which factors influence the score*

*hard tho as labels v imbalanced, e.g., hardly any 1s + 2s even if pooled*