**Meeting Notes: 4 July 2024**

*Post EDA video plan – Some of these to be asked*

**Further Data Cleaning**

* Instead of outlier removal process: take a natural logarithm of the income and annualised premiums.
  + Reasoning: we might be getting rid of high income and thus high premium paying customers due to face value, for example, for products like income protection.
* Further visualisations to get a better idea
  + NEOS vs. Non NEOS -> this was feedback from our EDA video.
  + Correlation matrix with the NEOS vs. Non NEOS ideology.
* Zero Income people -> imputation? Removal? [ASK]
  + We were thinking that it doesn’t make sense for so many professionals to have zero income. Proof below:
  + 
* Occupation: There are 807 different occupation levels which is too high to use in a modelling technique. We were thinking to either only take the top 30 occupations (or any other reasonable number) OR group occupations manually (in excel? Although this will be very tedious) [ASK]

**Modelling Process [ASK]**

* NEOS vs Non NEOS indicator to be used in the modelling.
  + Models being considered (3-4 should be enough):
    - Binomial GLMs (logistic regression) -> Can run forward / backwards selection?
    - Classification Trees
    - Random Forests -> can look at variable importance charts
    - Clustering Techniques (Unsupervised Learning: we learn these in week 7)
      * K-Means etc. (different to KNN)
* Should we also split the dataset into the four focus products?
  + Just because NEOS may be performing stronger in a specific one but weaker in another due to different circumstances of customers and the products they may need.