EDAOutput

Summary Notes:

1. Country has the least amount of missingness (at most 0.9%), which tells us that we can use country as our individuals.

2. Gender can be successfully recoded as binary for 2014-2018. However, missingness is as bad as 35% for 2017 and 2018. We would need to do some sort of fancy imputation method (Inverse Probability weighting or logistic regression). **More research required**.

3. When subsetting the data by FT employed developers, missingness in Gender is a bit better (as bad as 32%). However, we would need to eliminate 2014 dataset because they don't have a question that separates by Employment Status.

4. Missingness in Salary is the terrible for original dataset (as bad as 80%), so we could not use it as a covariate. However, if we adjust for Employment Status, there are good decreases in missingness. But, it's still a terrible situation (as bad as 66%). If our imputation methods that we research work for missingness as bad as 66%, then we can use Salary.

Type of Responses

Year	Options	Type
2018	Male, Female Transgender, gender non-conforming, genderqueer, Non-binary NA	Mark all that apply
2017	Male, Female Transgender, Gender non-conforming Other, NA	Mark all that apply
2016	Male, Female Other Prefer not to disclose, NA	Choose one
2015	Male, Female Other Prefer not to disclose, NA	Choose one
2014	Male, Female Prefer not to disclose, NA	Choose one

Data Reduction / Derived Variables

Recoding Algorithm

1. Any answer indicating 'Male' exclusively was coded as '1'. 2. 'Prefer not to disclose' was changed to 'NA'. 3. All other responses were coded as '0'.

For Country variables, it appears that we can continue with complete cases. There is only missingness in 2016 (n=502, 0.9%) and 2018 (n=412, 0.4%). The missingness is clearly a very small fraction of the entire respective datasets.

\begin{table}[!h]

\caption{Table of Proportions (%) of Gender for 219 Potential Countries Across Time}

	Other	Males	NA
2018	5.09	60.40	34.51
2017	6.72	61.55	31.73
2016	6.21	92.04	1.75
2015	6.11	91.81	2.08
2014	4.61	89.80	5.59

\end{table}

Solutions to Missingness

Missingness in Gender is very problematic for 2018 and 2017. Otherwise, we can continue with complete case for 2014-2016. Or, if we come up with an imputation method for 2017-18, then might as well apply it to 2014-2016.

(Imputation on Binary Variables)[https://niasra.uow.edu.au/content/groups/public/@web/@inf/@math/documents/mm/uow228467.pdf] See slides 15 specifically; alludes to using logistic regression to impute for missing values in gender. Jaylen also recommend Inverse Probability Weighting; there's an R package called ipw.