Final-ABS

Introduction:

After looking through the questions, we decided that in the time given, we should focus on building a model that can predict q106, which asks if the respondent thinks equal treatment of citizens by their government has gotten better or worse. We believe this question is important in order to compare how different respondents have viewed trends towards a "better" government system in their country. For instance, it's possible that a perception of positive trends is strongly correlated with a specific country or demographic group, indicating targeted improvements in equality. Alternatively, if perceptions of unequal treatment are strong correlated with certain groups, these groups may have been disproportionately "left out" of recent progress.

Overall, this analysis could provide grounds for further research into how different groups perceive recent developments around equality in East and Southeast Asia.

EDA:

Before starting our exploratory data analysis, we subsetted the merged Wave 1 data to only include the most important variables as provided in the documentation.

Afterwards, we explored which variables had the most missing data and investigated whether data was randomly or non-randomly missing (see Table 1 in the appendix). We discovered that there was no data for se004, q121, and q028 from Mainland China and no data on respondent age from Mongolia.

After visualizing the breakdown of responses to our variable of interest by age group (table 2), we determined that age was unlikely to be a significant predictor in our final model. Due to the missing data, we decided to drop se004, q121, q028, and se003a from our dataset.

After visualizing our variable of interest (Graph 1), we determined that there were a good number of observations for each category of our response variable. This allowed us to proceed with putting together a multinomial lasso model.

Proposed Methodology:

Our selected variables: Se002(gender), se005(education), se009(income), country, q027(participation in last election), q105(everyone is free to say what they think)

In building our model, we used various methods to narrow down a range of potentially strong and important predictor variables.

First off, we choose to include a few typical indicators of socio-economic status as predictor variables – gender, education level, and income level. These ___________. Although we considered adding age as a predictor, we found that there was significant missing data on age from Mongolia, and it was therefore compromised as a predictor variable. Our exploratory data analysis confirmed that age was likely not a significant predictor of responses to our question of interest, so we proceeded without age (see appendix).

In addition, we added question q105 as a predictor, which asked whether survey respondents felt that free speech was getting better or worse in their country. We felt that this was potentially relevant to include, because perspectives on equality and free speech are tied to a survey respondent's notions of rights and justice. There were no systematically missing observations from q105 and no obvious barriers to it being a predictor variable.

Finally, we included country because country is a key aspect of the response variable we are assessing. We are attempting to predict a survey respondent's perspective on trends in their government's equal treatment of citizens, so the respondent's country is a key aspect to include.

We decided to transform our response variable q106 into two different categories. "Much worse," "Somewhat worse," "Much the same," were encoded as 0, and "Somewhat better,"Much better than before" were encoded as 1. As a result, our goal is to predict if trends in the equal treatment of citizens has gotten worse/stayed the same, or gotten better. We decided to drop all NA's from the dataframe. We considered using KNN to impute missing values, but we thought the missing data may have a systematic pattern, as seen in the Table 1, and the large amount of missing values were computationally expensive to fit given the number of dimensions.

We then decided to fit a logistic lasso regression in order to decrease the variance of our model and also perform variable selection. Shrinkage of the coefficients also may help with multicollinearity in the model. We used cross validation to find the optimal lambda shrinkage parameter, and used that as the lambda parameter in our final model.

Analysis:

```
## 27 x 1 sparse Matrix of class "dgCMatrix"
##                                                s0
## (Intercept)                         -1.3343002616
## se002female                         -0.0963043889
## se005Incomplete elementary school    0.0143863534
## se005Complete elementary school      .
## se005Incomplete secondary school     .
## se005Complete secondary school      -0.1601890133
## se005Incomplete high school         -0.0437237364
## se005Complete high school           -0.0002552002
## se005Some university_college education -0.0799765318
## se005University_college degree       .
## se005Post graduate degree            0.3761538422
## se0092nd quintile                   -0.0476837074
## se0093rd quintile                   -0.1024841350
## se0094th quintile                   -0.1311929814
## se0095th quintile                   -0.1154351663
## countryHong Kong                    -1.0234966990
## countryKorea                        -0.5341041232
## countryMainland China                0.3221378706
## countryMongolia                     -0.5295745795
## countryPhilippines                  -0.2954362164
## countryTaiwan                        0.0853827923
## countryThailand                      1.1514918441
## q027Yes                              0.0187273032
## q105Somewhat worse                  -0.2408888461
## q105Much the same                    .
## q105Somewhat better                  1.8206298440
## q105Much better than before          2.4197431773
```

```
## [1] 0.0009257176
```

Looking at our model, we see that certain levels of some predictors may have coefficients of 0, but for all predictors most if not all levels are nonzero and significant. We can tell that for example, a female is less likely to think that equal treatment of people by the government has gotten better. Countries Mainland China, Mongolia, and Thailand seem to have be more likely to think that equal treatment of people by the government has gotten better. Meanwhile, countries Hong Kong, Korea, Philippines, and Taiwan are less likely to think so. Interestingly enough, we found that compared to the 1st quintile of wealth (poorest), each other quintile of wealth is less likely to believe that equal treatment of people has gotten better.

Conclusion:

Given more time, we would like to explore more interaction effects. In addition, we would also include Wave2 as a predictor variable and merge in the Wave2 data. Also, another thing we want to consider is the handling of missing data. Part of the reason we didn't use certain predictors, for example age, is because certain countries did not report them at all. We were wondering if there was a better way to fix this problem. Plenty of other observations had missing values as well, so we wanted to find a statistically sound way to impute them. Finally, we wanted to find other models that might fit well. For example, we fit a SVM but scrapped it because we were not sure about performing proper variable selection with SVM's. In addition, we would have liked to implement decision trees since they can handle categorical predictors and responses really well.

Appendix

Table 1 - table of missing values

```
##  [1]    8 3201   46  958 1184    0 1266  730  742  889 1231  904 1033   70 1829
## [16] 6899  730  980 3585  829  869
```

Table 1: Table of NA values

Variable	Number of NA
se002	8
se004	3201
se005	46
se009	958
se003a	1184
country	0
q007	1266
q008	730
q009	742
q010	889
q006	1231
q098	904
q128	1033
q005	70
q027	1829
q028	6899
q105	730
q106	980
q121	3585
q123	829
q127	869

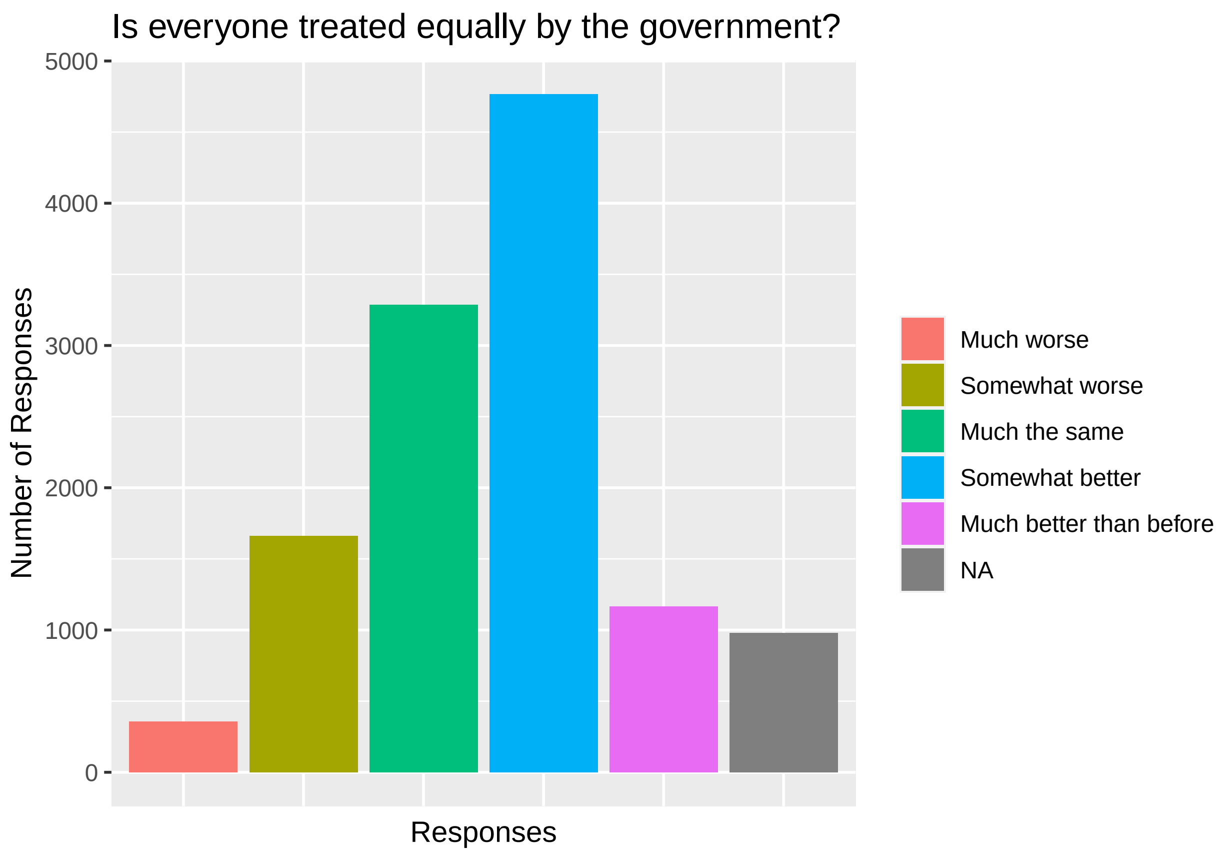
Table 2- Table of country breakdown

Table 2: Survey Respondents: Breakdown by Country

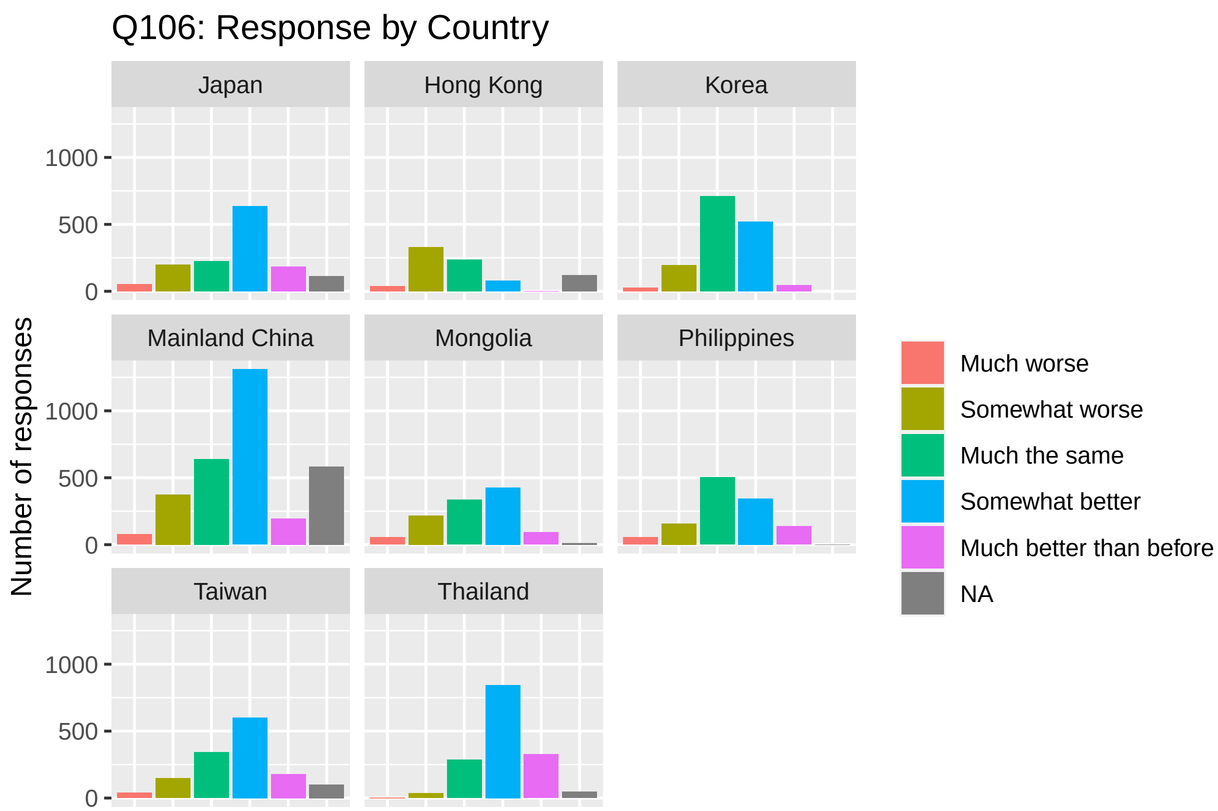
Country	Percent of Respondents
Hong Kong	6.638
Mongolia	9.364

Country	Percent of Respondents
Philippines	9.822
Taiwan	11.582
Japan	11.607
Korea	12.278
Thailand	12.654
Mainland China	26.054

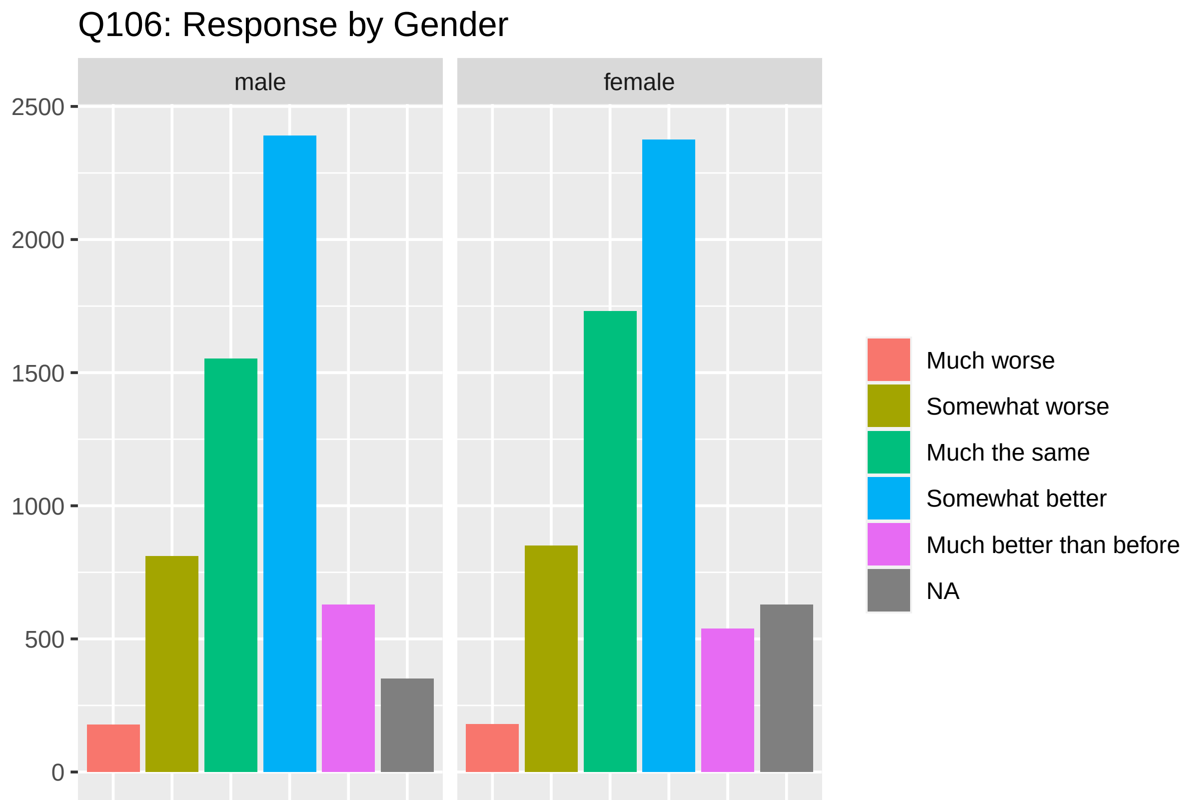
Graph 1- Visualize Q106



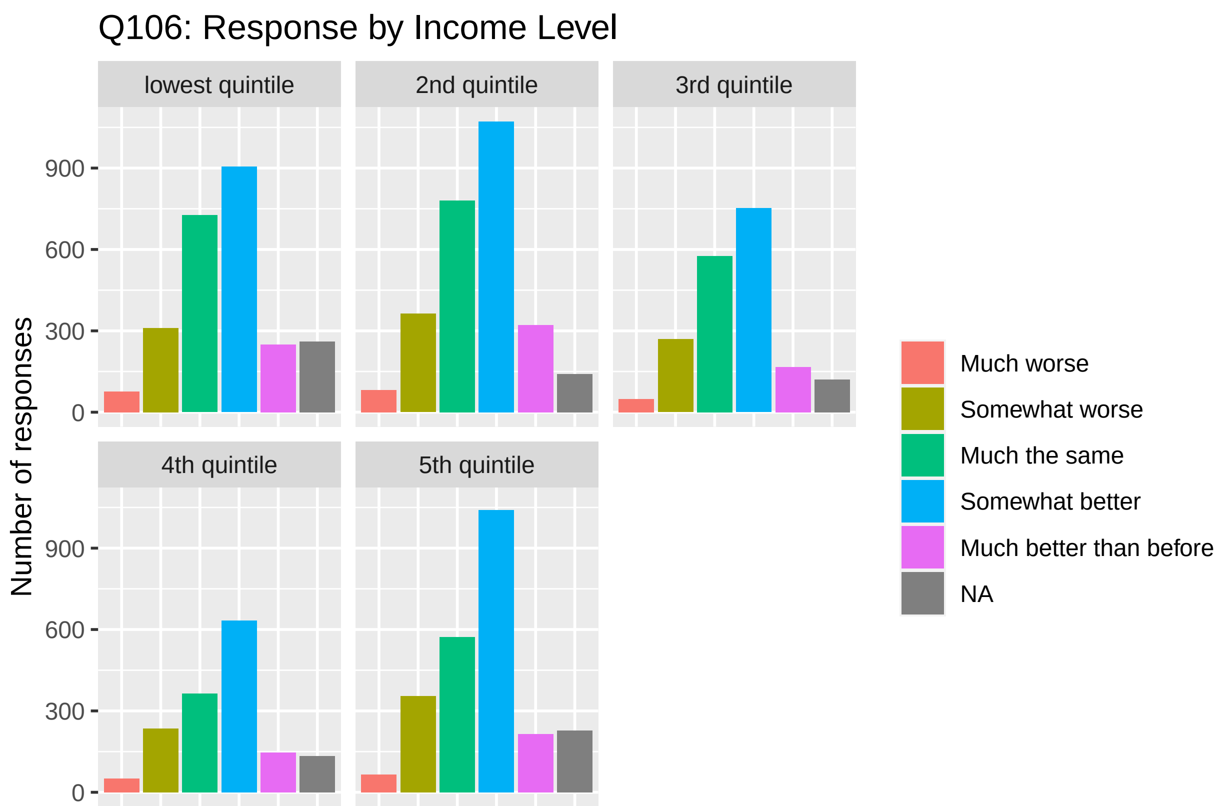
Graph 2- Visualize Q106 by country



Graph 3- Visualize Q106 by Gender



Graph 4- Vizualize Q106 by Income



Graph 5- age vs answer to Q106

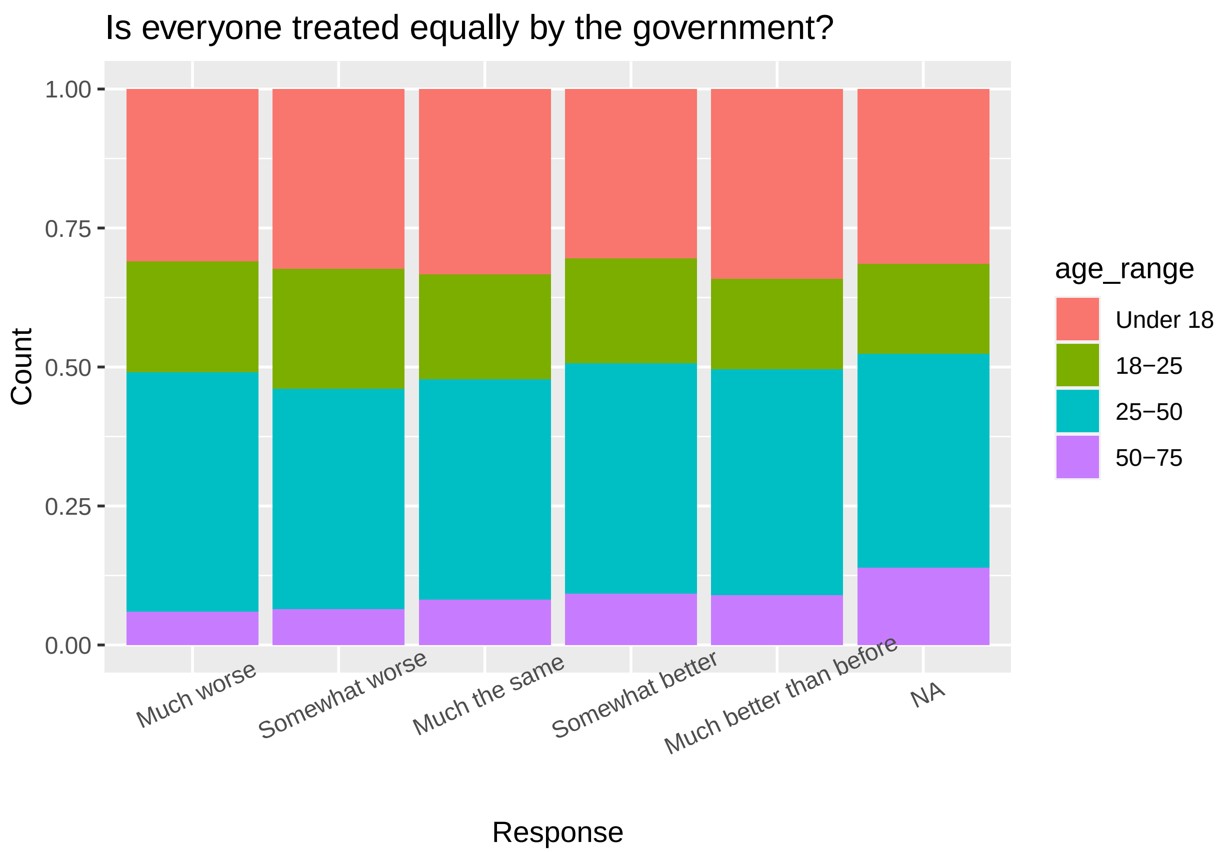


Table 3- Response of Q106 by Country

Table 3: Country Breakdown of Response to Q106

Q106 Response	Japan	Hong Kong	Korea	Mainland China	Mongolia	Philippines	Taiwan	Thailand
Much worse	3.949	4.932	1.800	2.451	4.895	4.583	2.827	0.259
Somewhat worse	14.104	40.691	13.000	11.781	19.056	13.083	10.530	2.393
Much the same	16.008	29.470	47.467	20.138	29.371	41.917	24.240	18.435
Somewhat better	44.852	9.864	34.667	41.156	37.413	28.833	42.544	54.657
Much better than before	13.047	0.247	3.067	6.158	8.129	11.417	12.721	21.216
NA	8.039	14.797	NA	18.316	1.136	0.167	7.138	3.040