PART 1

QUESTION A

The aim of this report is to identify which types of people are most supportive of longer prison sentences, while simultaneously evaluating how much each characteristic matters in explaining support. Before choosing the variables from the logit model, I will be assessing all variables altogether (using logit transformation). As can be seen, the following variables are not statistically significant (with significant threshold at 5%) in regards to the logit model regressions: sex, age, working class and married. To compare whether one model is better than this initial one, I performed training and test data to see the difference in error rates.

Table 1: Logit model for original data (complete variables)

	Estimate	**Standard Error**	**Z value**	**P value**
Intercept	0.812	0.276	2.947	0.003
Sex	-0.159	0.093	-1.702	0.089
Age	0.002	0.003	0.768	0.442
Right-wing scale (1 as the lowest level and 5 as the highest)	0.137	0.062	2.193	0.028
Working Class(=1) or not (=0)	-0.112	0.133	-0.844	0.399
University educated (=1) or otherwise (=0)	-0.929	0.100	-9.310	<2e-16
Urban area	0.280	0.103	2.711	0.007

(=1) or otherwise (=0)				
Married (=1) or otherwise (=0)	0.203	0.104	1.957	0.050
Income decile (1=lowest to 10=highest)	-0.066	0.020	-3.322	0.001

As can be seen from Table 2: The error rate illustrates that our model wrongly classifies our test data in 32.02% of the cases, regardless of the direction of the mistake (false positives and false negatives). The sensitivity is the proportion by which our model correctly classifies an observation as 1 if the true value of the observation is indeed 1 and is at 97.15% which is very high in this case. Lastly, specificity is the times our model correctly classifies true 0's as a 0 and here is very low at 11.5%.

Table 2: Test data for original data

Type	**Percentage, %**
Error rate	32.02
Sensitivity rate	97.15
Specificity rate	11.5

For model 2, I have chosen to remove the following variables: Working Class (or otherwise) because it can be similarly defined by level of income. For example, to define working class is not only in terms of the education background or type of skills needed, but one could define whether someone is in the working class by average yearly income. Moreover, living in urban area, marriage status and sex because I personally do not think the given variables has a strong relevance in determining whether one supports longer prison sentences. From its logit regression summary, there are only two variables that are not statistically significant: age and sex. Next, when using the test data the error rate is slightly bigger at 34.37%, sensitivity increased to 99.74% and specificity at 5% (a relatively big decrease).

However, when focusing on training data, the latter performs better than the former; with model 2 having 34.03% error rate in comparison to 54.9%, 99.62% for sensitivity compared to 96.73% and lastly, (though it did not perform better), it had specificity of 3.19% in comparison to 9.84%. Therefore, the new model performs worse in terms of error rate and specificity using the test data but better in terms of error rates and sensitivity in the training data. For this reason, I will just be using model 2 in the following question - also to focus on each variable more closely.

Table 3: New data (i.e. Model 2)

	Estimate	**Standard Error**	**Z value**	**P value**
Intercept	0.972	0.238	4.093	4.26e-05
Age	0.002	0.003	0.567	0.570
Right-wing scale (1 as the lowest level and 5 as the highest)	0.148	0.061	2.410	0.016
University educated (=1) or otherwise (=0)	-0.922	0.097	-9.465	< 2e-16
Income Decile (1=lowest to 10=highest)	-0.054	0.018	-3.075	0.002

Table 4: Test for Model 2

Type	**Percentage, %**
Error rate	34.37
Sensitivity rate	99.74
Specificity rate	5

Table 5: Training Data for Model 1 and 2

Type	**Percentage, %**	
	Model 1	**Model 2**
Error rate	54.9	34.03
Sensitivity rate	96.73	99.62
Specificity rate	9.84	3.19

To find how much the variables matter in explaining support for longer sentences, I will be conducting two interpreting tools; Average marginal effects (to find a general overview - especially for the binary variables) and change in predictive probabilities (for Right Wing Scale and Income levels).

Using AME, the following can be found: If the significant level is 5%, we cannot reject the null hypothesis that the average marginal effects of age on the probability of supporting longer sentences is zero. Because p value is bigger than 0.05, at 0.57. Whereas the AME of how right wing a person is, whether they went to university and their level of income are statistically significant at the 5% level. The marginal effect of having gone to university suggests that, on average, across our dataset, is associated with decreasing the probability of supporting longer prison sentences by 0.20.

Table 6: Average Marginal Effects of Model 2

Marginal Effects:				
	dF/dx	**Standard Error**	**z**	**P value**
Age	0.000	0.001	0.567	0.571
Right Wing Scale	0.031	0.0128	2.396	0.017
University (=1) or otherwise (=0)	-0.203	0.021	-9.459	< 2e-16
Income Decile (1=lowest to 10=highest)	-0.011	0.004	-3.046	0.002

I particularly want to see if there will be a change in supporting longer prison sentences when someone becomes more right wing. To do this, I estimated the difference in the predicted probability of voting leave for a university-educated person with a political right wing level of 1 to someone who is also university-educated but a political right wing level of 5. As can be seen, the probability of voting leave is 0.142 higher for someone in level 5 than someone in level 1. I did the same thing with income levels - as I estimated the difference in the predicted probability of voting leave for a university-educated person with income decile 1 to someone who is also university-educated but with income decile 10. The probability of voting leave is 0.117 less likely for someone (university-educated) in income decile 10, than for someone (university-educated) in income decile 1. It is also statistically significant with 95% confidence interval -0.820,-0.041.

Table 7: Difference in Predicted Probability - Right Wing Level

Mean	**Quantile**	
	2.5%	**97.5%**
0.142	0.023	0.263

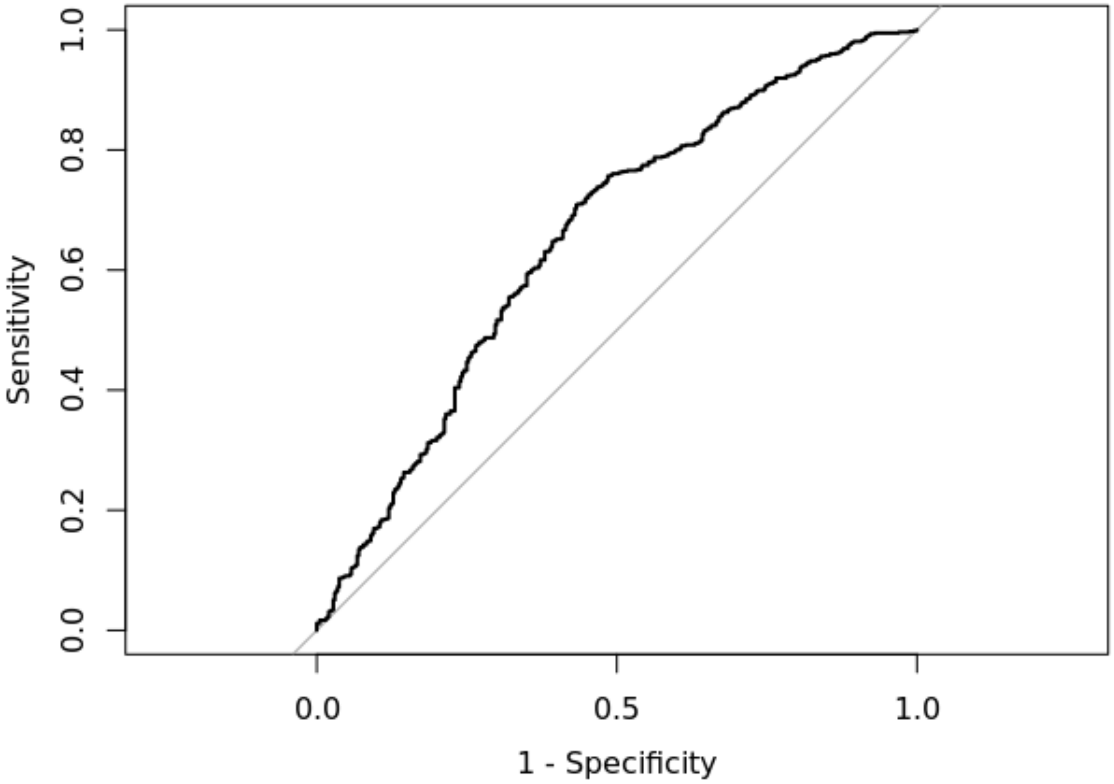
Table 8: Difference in Predicted Probability - Household Income Decile

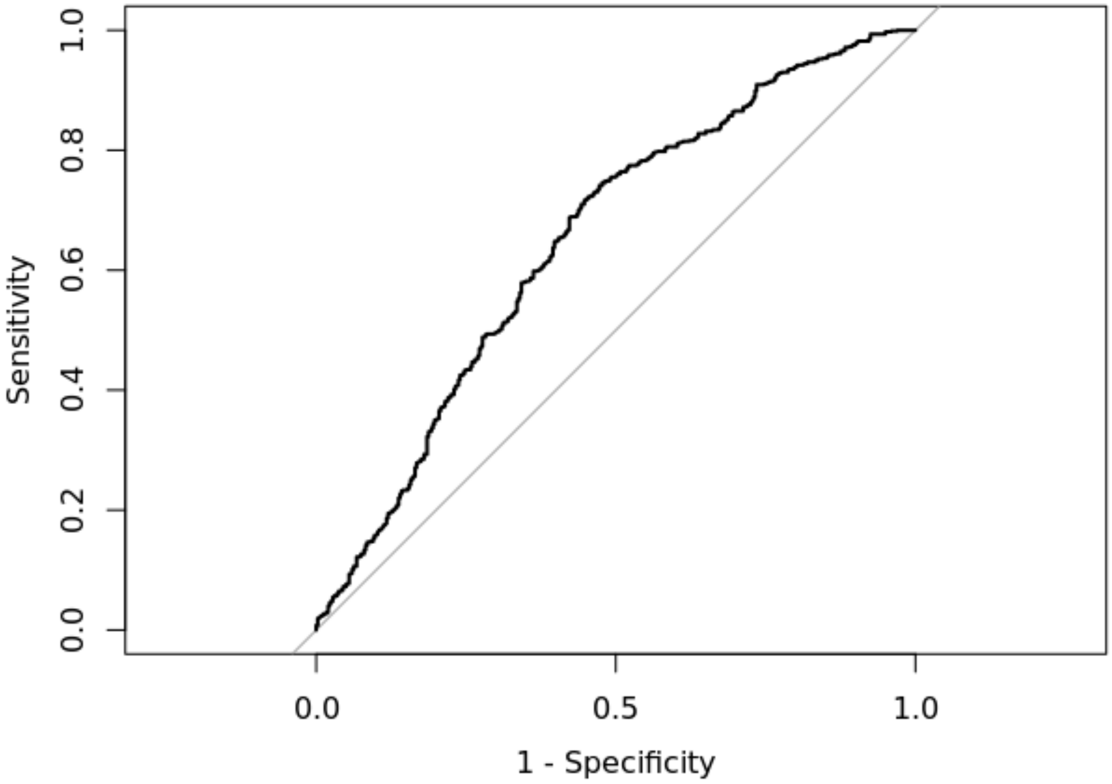
Mean	**Quantile**	
	2.5%	**97.5%**
-0.117	-0.190	-0.041

In summary, I have presented a logit multilevel regression model to predict support for longer prison sentences. After removing certain variables: working class, living in urban area, marriage status and sex to create our second model, and using test and training data to compare my second model with its initial one, I found that the second model performs better in terms of error rate and specificity when using training data. I proceeded to use the second data to present the model's findings that explain how much the chosen variables matter in explaining support for longer sentences. I first conducted Average Marginal Effects to analyse the relationship, and found that I cannot reject the null hypothesis that the AME of age on the probability of supporting longer sentences is zero. Using change in predicted probabilities, I found how the probability of voting leave changes when right wing level is at its minimum (=1) to its maximum(=5) - the same way with household income. The relationship is positive for the former and negative for the latter.

Areas of improvement would be in improving model performance as illustrated by the ROC curve. When illustrating each model to the ROC Curve, it shows that the initial model has an area under the curve of 0.653 while the second one has an area of 0.650 (as shown in the diagrams below). Given that the closer to 1 it is, the better, the second model has worse performance than the first one.

Diagram: ROC Plot of Model 1 and Model 2





Question B

This report aims to do the following: i) to produce estimates of the percentage of voters that voted ‘leave’ in every constituency and ii) to use the results to explain why people voted to leave. To estimate an appropriate logistic multilevel model, I firstly find the Intraclass Correlation Coefficient. This measures how much variance there is between groups compared to between individuals, from values 0 to 1. Our ICC value, 0.054 indicates that group level predictors account for a small proportion of the variance in our model.

Given that this deals with many binary variables such as: Voting Conservative and UKIP in the 2015 election (or otherwise), being female (or otherwise), having education with a degree level or higher (or otherwise), etc., multilevel logistic regression is needed to assess the relationship between the coefficients and the dependent variable. Table 9 describes the justification to the coefficients used in the regression - I have used a YouGov poll (2016) that measured how voting to Leave differs across party votes, gender, age, education, and constituency variables. I have used other polls (Goodwin, M and Heath, O., 2022 and Martin, N. and Sabolewska, M., 2017) and surveys to justify the remaining variables.

Table 9: Chosen variables for Multilevel Logistic Regression

Variables	Derivation	Justification
Voted Conservative in 2015 election	Voted for Conservative =1, otherwise =0	Conservatives voted to leave 61% to 39%, indicating that there is a relationship between having voted Conservative and voting leave EU
Voted UKIP in 2015 election	Voted for UKIP=1, otherwise =0	The same with UKIP when 95% of those who voted for the party, voted to leave the EU

Sex	Female = 1,
	Otherwise = 0
		Though not as prevalent as the rest in terms of its justification, I believe

		gender wealth gap can play a key determinant for voting to Leave the EU. As females on average receive lower wages than their male counterparts, and seeing how Brexit is rooted in economic and employment ideas - gender would have a relationship with the vote to Leave the EU.
Age	Continuous	Under 25s were more than twice as likely to vote Remain (71%) than Leave (29%). Whereas over 65s, were almost twice as likely to vote Leave (64%) than Remain (36%)
Educated to a degree level or higher	Educated to a degree level or higher = 1, Otherwise = 0	Those with a degree, 68% voted to Remain whereas the rest voted to Leave indicating a negative relationship (Higher degree more likely to vote Remain)
No educational qualifications	No educational qualifications = 1, Otherwise = 0	Similarly, those with no education (GCSEs or lower), 70% voted to leave and the remaining to remain - illustrating a strong relationship between having no educational qualifications and voting to leave.
Percent vote for Conservative party in the constituency, 2015 election	Continuous	The same justification as Voted Conservative in 2015 election
Percent vote for UKIP party in the constituency, 2015 election	Continuous	The same justification as Voted UKIP in 2015 election
Constituency UE rate	Continuous	Among the most popular arguments defended by

		Brexit supporters were its anti-immigration policies - to reduce the number of immigrants which would “secure” more jobs for British workers. It would be to no surprise that this sentiment would be supported by those who are unemployed.
(Percent of constituency population who are) **White British**	Continuous	A paper by Martin and Sobolewska (2017) explored the racial determinants behind Brexit and found that ethnic minorities are strongly in favour of Remaining the EU. This may give weight in the constituency-level variables of percent of White British with it being higher, the more likely they would vote Leave.
(Percent of constituency population living in) **poverty**	Continuous	Same argument with constituency UE rate.

Table 10: Multilevel Logistic Regression

Fixed Effects				
	Estimate	**Standard Error**	**T value**	**P value**
(Intercept)	-3.419	0.694	-4.925	8.41e-07
Voted Conservative in 2015 election (=1, otherwise =0)	1.004	0.109	9.182	<2e-16
Voted UKIP in 2015 election (=1, otherwise =0)	4.673	0.719	6.499	8.11e-11
Gender (female = 1, male = 0)	-0.037	0.099	-0.368	0.713
Age	0.016	0.003	5.448	5.11e-08
Educated to degree level or higher (=1, otherwise =0)	-0.641	0.114	-5.615	1.96e-08
No educational qualifications (=1, otherwise =0)	-0.308	0.137	-2.245	0.025
Percent vote for Conservative party in the constituency, 2015 election	0.010	0.006	1.727	0.084
Percent vote for UKIP party in the constituency, 2015 election	0.006	0.011	0.526	0.599
Constituency	0.040	0.091	0.443	0.658

UE rate				
(Percent of constituency population who are) **White British**	0.008	0.004	2.011	0.044
(Percent of constituency population living in) **poverty**	0.043	0.032	1.351	0.177

Table 11: Average Fixed Effects

Variables	**Average Fixed Effects**
(Intercept)	-0.184
Voted Conservative in 2015 election (=1, otherwise =0)	0.233
Voted UKIP in 2015 election (=1, otherwise =0)	0.630
Gender (female = 1, male = 0)	-0.006
Age	0.003
Educated to degree level or higher (=1, otherwise =0)	-0.128
No educational qualifications (=1, otherwise =0)	-0.056
Percent vote for Conservative party in the constituency, 2015 election	0.002
Percent vote for UKIP party in the constituency, 2015 election	0.001
Constituency UE rate	0.008
(Percent of constituency population who are) **White British**	0.002
(Percent of constituency population	0.008

living in) **poverty**	

Table 11 illustrates the average effect of having voted Conservative in the last 2015 election is 0.233 and for having voted UKIP is 0.630. These two are both statistically significant, and have a relatively higher coefficient than other variables (For example, the average effect of being educated to a degree level of higher is -0.128 and having no educational qualifications is -0.056). White British is also statistically significant.

Because this is a logit multilevel regression model, odds ratios were used to conduct a deeper level of interpretation. The following points can be deduced from Table 12: 1)The odds are 2.730 higher to vote to leave the EU for someone who voted Conservative in the last election. 2) It is drastically higher for someone who voted UKIP with the odds being 107.029 higher. 3) The odds are 0.964 higher for someone who gets one year older to vote to leave. 4) The odd ratio is smallest when it comes to whether or not someone had an education of degree-level or higher - with the odds of voting leave being 0.527 higher for someone with university education or higher. 5) Lastly, constituency-level coefficients have similar units of odd-ratios (being in the range of 1.006 to 1.044).

Table 12: Odd Ratios

Variables	**Odd Ratios**
Voted Conservative in 2015 election (=1, otherwise =0)	2.730
Voted UKIP in 2015 election (=1, otherwise =0)	107.029
Gender (female = 1, male = 0)	0.964
Age	1.016
Educated to degree level or higher (=1, otherwise =0)	0.527
No educational qualifications (=1, otherwise =0)	0.735
Percent vote for Conservative party in the constituency, 2015 election	1.010

Percent vote for UKIP party in the constituency, 2015 election	1.006
Constituency UE rate	1.041
(Percent of constituency population who are) **White British**	1.008
(Percent of constituency population living in) **poverty**	1.044

To produce post-stratified estimates of our dependent variable, I have incorporated the post-stratification data for the same 631 constituencies. Our goal of post-stratification is to estimate outcome in small geographic areas, which in this case is the percentage of voting leave by constituency. Post-stratification uses two sources of data to get there: 1) Predicted outcome for demographic groups in every area and 2) census data on the demographic composition of every area.

Firstly, I added a new variable called “prediction” to the “post” dataset for the predicted support to leave the EU for each demographic subgroup in the data. Secondly, I added a weight variable, which weights each prediction group by the group’s percentage of the constituency's population. Lastly, I produce estimates of the constituency's opinion from the regression shown in the previous question by adding up weighted predictions for each constituency.

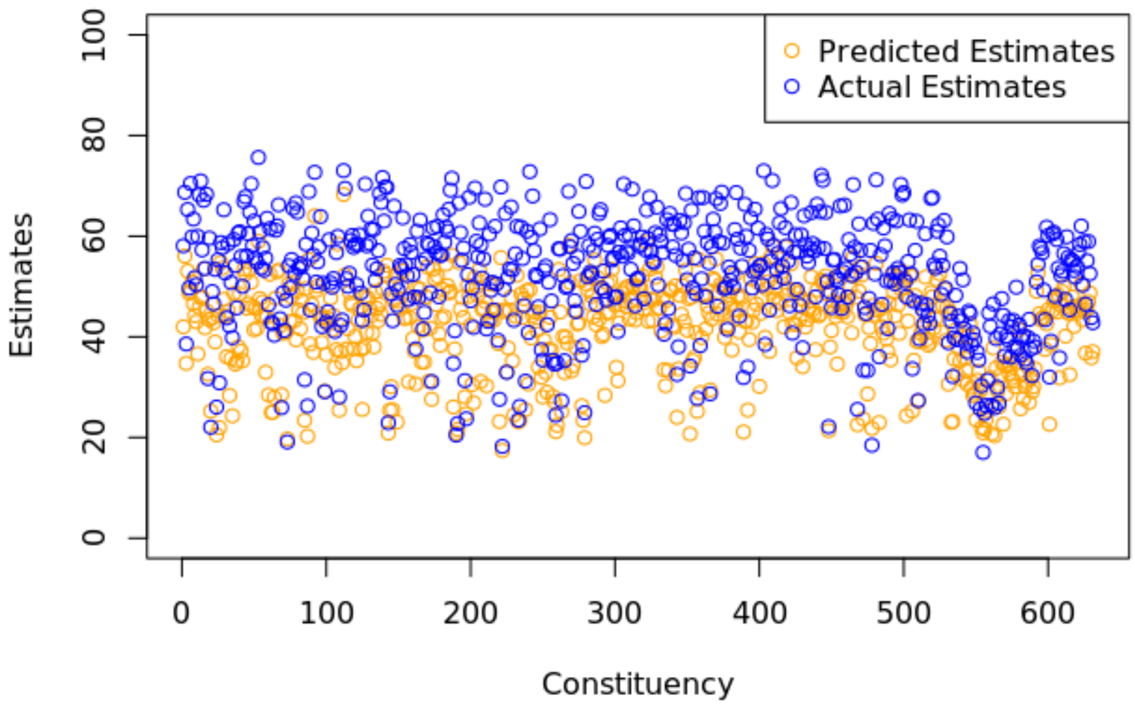
Table 13 illustrates 10 constituencies with the highest post-stratified estimates including: Clacton (68.287), Castle Point (64.044) and Boston and Skegness (58.983). Also, ten constituencies with the lowest post-stratified estimates: Homsey and Wood Green (17.407), Bristol West and Manchester(19.704), Withington (19.961).

Table 13: Highest and Lowest results of Post-Stratified Estimates

Highest results of Post-Stratified Estimates	
Constituency Name	**Post-Stratified Estimates**
Clacton	68.287
Castle Point	64.044
Boston and Skegness	58.983
North Thanet	58.874
Christchurch	58.470
Louth and Horncastle	58.200
South Holland and The Deepings	58.068
South Thanet	57.853
Great Yarmouth	56.216
Aldridge-Brownhills	56.049
Lowest results of Post-Stratified Estimates	
Homsey and Wood Green	17.407
Bristol West	19.704
Manchester, Withington	19.961
Cambridge	20.252
Glasgow North	20.351
Bermondsey and Old Southwark	20.518
Poplar and Limehouse	20.710
Glasgow Central	20.718
Dulwich and West Norwood	20.874
Hackney North and Stoke Newington	20.883

To assess the model's performance across all 631 constituencies, we can calculate the Mean Absolute Error (MAE). The formula of MAE is $\sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{n}$ where y_i is the actual outcome for actual constituency vote share and $\hat{y}_i$ is the predicted outcome for constituencies from our MRP analysis. In words, we take the average absolute difference between the actual and predicted outcomes of all geographic areas. Here, the MAE is 10.33, a relatively inaccurate model. From plotting the estimates for both the predicted and actual data, we can see that my MRP analysis underestimates. There are many reasons behind this, with one being omitted variables. As we only have 10 variables to predict our estimates, we might have left out important variables that could have made our model more accurate.

Scatter Plot 1: Predicted vs Actual Estimates



In summary, I have estimated an appropriate logistic multilevel model explaining voting for leave using all the variables provided in the dataset. I first used ICC to see the extent group level predictors account for the variance in the model. To analyse the regression model, I used odds ratios and found that the odds are higher to vote to leave the EU for someone who voted Conservative and UKIP in the 2015 election, and through age (one year older). Then, I produced post-stratified estimates and illustrated the highest and lowest 10 constituencies to vote leave. Lastly, to assess our model's performance across all 631 constituencies, I conducted Mean Absolute Error, which I found was 10.33 - reflecting a relatively high inaccuracy. When plotting the predicted estimates and the actual estimates in a scatter plot, I found that my estimates underpredicts. This might be due to omitted variables - causing biases to my estimates.

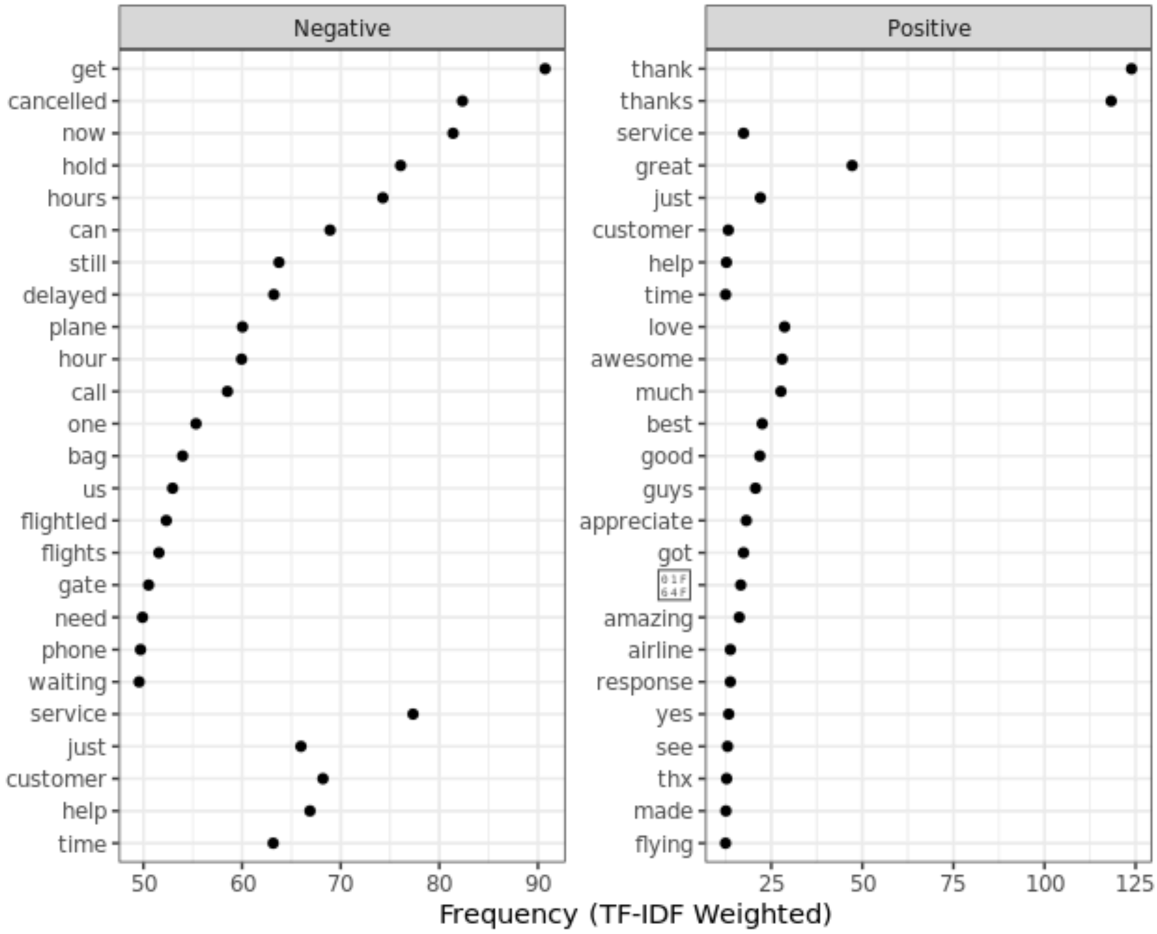
PART 2

QUESTION C

The aim of this report is to find out how many people talk about airlines on Twitter, and then build a predictive tool that can classify tweets in future into ‘negative’ or ‘positive’ sentiment toward airlines. Prior to analysing the 11,541 tweets about the following American airlines: United, JetBlue, American Airlines, US Airways, Virgin America, and Southwest, I have turned the corpus into a document term matrix. Moreover, I have cleaned up the corpus by removing numbers, making everything lowercase, removing stopwords, applying a proportional weighting, removing words that appears less than 3 times, and removing the following words to avoid repetition: "@americanair","@united", "@USAirways", "@SouthwestAir", "@JetBlue", "@VirginAmerica",and "flight". Lastly, I replaced proportional weighting Term Frequency - Inverse Document Frequency (TF-IDF) in order to lessen the amount of uninformative words in the corpus. TF-IDF weighting is calculated by multiplying the frequency of term t_j in document i (TFi) by the log of the number of documents divided by the number of documents containing tj (IDFj). In sum, this removed unique words from 13,066 to 1,339 words.

In analysing the positive or negative sentiment received by airlines, American Airlines received the most negative sentiment with one user saying “@AmericanAir YOU FUCKING SUCK” whereas United received the most positive sentiment “@united Thanks. Hopefully this is easily resolved”. To analyse in more detail, Table 14 illustrates the top words associated with positive and negative sentiments for the whole corpus.

Table 14: Top 25 most negative and most positive words



To see how word usage differs across different airlines, one can refer to Table 15 which illustrates the most frequent words for each airline. This does not illustrate much - thus, to find more information, Table 16 illustrates the proportion of positive and negative tweets by airline. All airlines receive more negative reviews than positive. It can be seen that for American Airline, United and US Airline, a big proportion of tweets have negative sentiments: For American Airline negative reviews make up 85.366%, for United makes 84.256% and lastly for US Airways it makes up 89.376%. For Southwest and JetBlue, the proportion of negative to positive is not that big - with JetBlue having 63.709% of tweets being negative. Virgin America received the least reviews overall (333 tweets) with 54.000% being negative.

Table 15: Different Words Usage Across Airlines

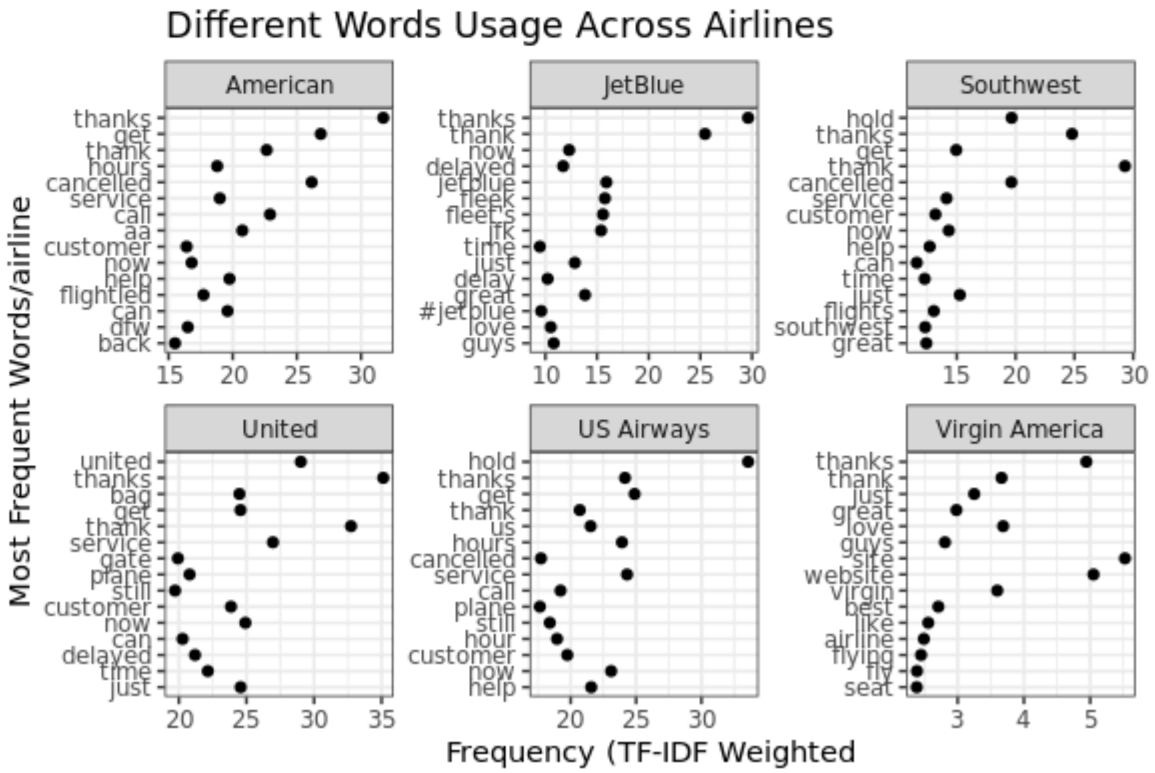
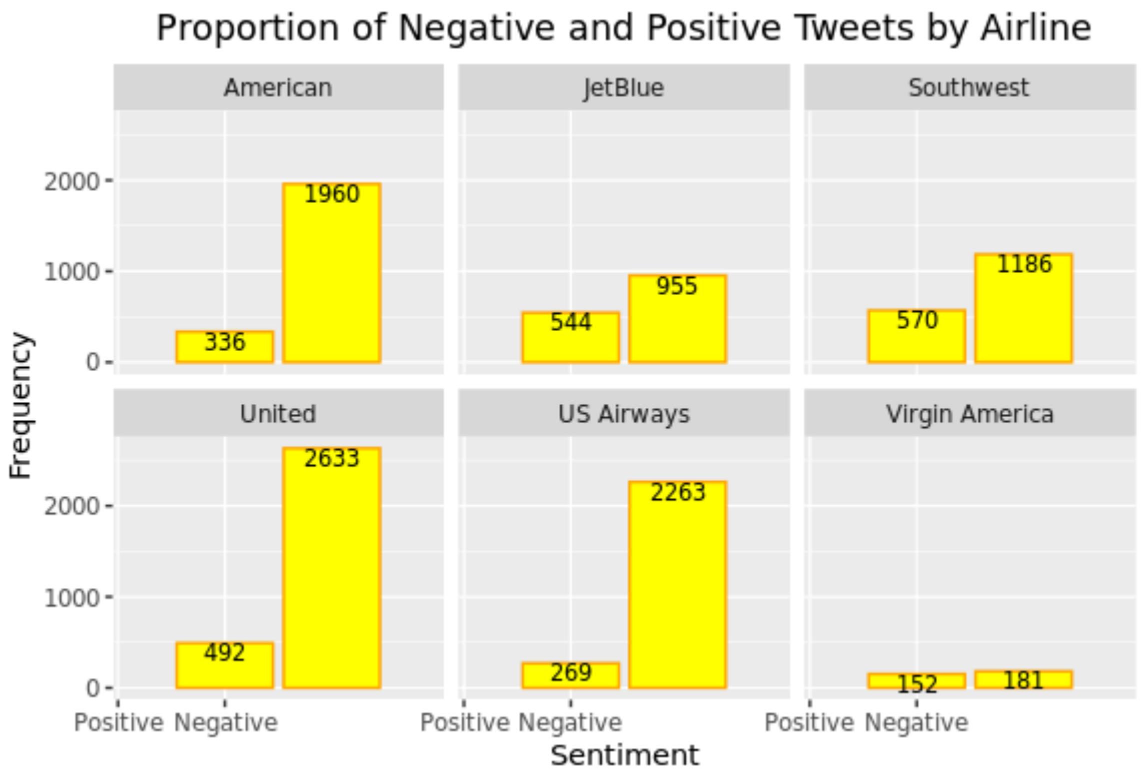


Table 16: Proportion of Negative and Positive Tweets by Airline



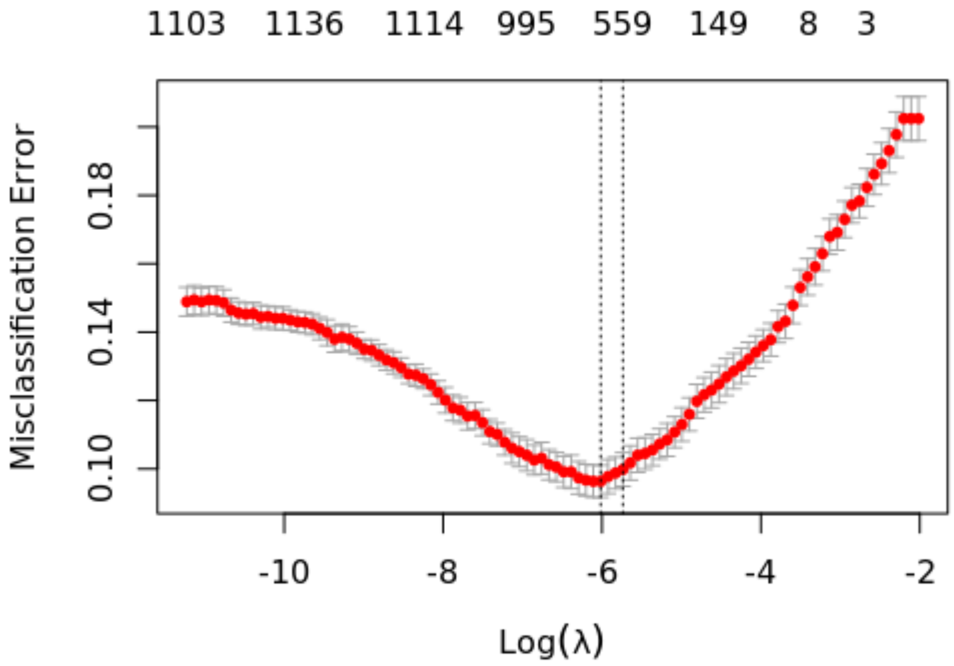
Now that the tweet corpus has been analysed, a dictionary can be manually created using the dictionary() function. This is considered a good way to classify tweets as it concerns terms that are specific to the corpus - which measures sentiments in regards to airlines. This would mean that they would have used many airline jargons that some existing dictionaries in R do not have. I have used the top 10 positive terms to define positive sentiments and the top 10 negative words to define negative sentiments (based on the analysis above) in my dictionary - as illustrated in Table 17.

Table 17: Positive and Negative words in My Dictionary

Positive	**Negative**
1. Thank 2. Thanks 3. Great 4. Love 5. Awesome 6. Much 7. Best 8. Good 9. Just 10. Guys	1. Get 2. Cancelled 3. Now 4. Hold 5. Hours 6. Service 7. Can 8. Just 9. Still 10. Delayed

However, there are downsides of using the dictionary method, among them being that it is difficult to build, as the computer will have to firstly learn the dictionary prior to analysing the corpus. Moreover, there is the problem of having a vast size of many corpora, lack of automation and its reliability. To solve this issue, we can use Lasso (Least Absolute Selection and Shrinkage Operator). Lasso does this by aiming to 1) fit the best-fitting model by minimising squared errors and also, 2) to penalise overly-complex models by reducing the number and size of coefficients. Out of the 11,541 tweets, the process only used 560 for classification. The scatter plot below illustrates the y axis being the average test error (misclassification) rate across folds and the left most dashed line shows the optimal lambda (i.e. lowest average error rate) with 560 words being fairly close to it (dashed line on the right).

Scatter plot 2: Misclassification Error of Lasso Model



To see which of the two is better in analysing the tweets, a comparison of the performance of the two classifiers is made by analysing the error, sensitivity and specificity rate. It can be seen that My Dictionary has a higher error rate of 15.007% whereas the Lasso Classifier's is 9.496%. They both have the same sensitivity rate and specificity rate. Thus, it can be said that in terms of error rate, the Lasso Classifier performs better but aside from that, My Dictionary performs relatively well as it was manually created. This may be because both classifiers have very similar most important words. This can be seen when analysing the most important words for the Lasso Classifier and comparing it with the words used in My Dictionary.

Table 18: Performance Comparison of My Dictionary and Lasso Classifier

	My Dictionary	**Lasso Classifier**
Error rate	15.007%	9.496%
Sensitivity rate	91.392%	91.392%
Specificity rate	60.135%	60.135%

Table 19: Most Important Words in Lasso Model

Positive		**Negative**	
Words	**Lasso coefficient**	**Words**	**Lasso coefficient**
thank	-19.454	hold	9.933
thanks	-18.925	1k	9.763
awesome	-14.453	hours	9.053
kudos	-13.245	cancelled	8.912
love	-12.569	worst	8.717
great	-12.190	nothing	7.359
abest	-12.116	ruining	7.249
amazing	-10.854	late	7.028
rock	-10.760	delayed	6.790
excellent	-10.401	fare	6.322

In summary, this report aims to find how many people talk about 7 different airlines on Twitter. Moreover, I created my personal dictionary manually and compared its performance using the lasso classifier. The comparison shows that though in terms of error rate, the Lasso classifier is better, they both nevertheless perform similarly in terms of sensitivity and specificity rate. This may be because both classifiers use similar words to classify tweets (which were top 10 most positive and negative terms). The Lasso Classifier can be said to have performed well given from the Scatter plot 2 that it is very close to the optimal lambda. Areas of improvement in the report could be identifying the negative and positive words for each airline so individual airlines could have specific targets on areas of improvements. Moreover, we could explore different dictionary and classification methods to compare and evaluate the corpus. The same could be said in terms of the diversification of the tweets - where one could use word scraping to acquire more diverse and up-to-date opinions(/tweets).

References

Goodwin, M. and Heath, O., 2022. *Brexit vote explained: poverty, low skills and lack of opportunities*. [online] JRF. Available at: <https://www.jrf.org.uk/report/brexit-vote-explained-poverty-low-skills-and-lack-opportunities> [Accessed 17 May 2022].

Martin, N. and Sobolewska, M., 2017, June. Ethnic minority support for leaving the European Union. In *EU Referendum Project Symposium*.

Moore, P., 2022. *How Britain voted at the EU referendum | YouGov*. [online] Yougov.co.uk. Available at: <https://yougov.co.uk/topics/politics/articles-reports/2016/06/27/how-britain-voted> [Accessed 17 May 2022].