Before dealing with the data

We are asked to find out the factors that affect the voters' decision regarding the UK's departure from the EU and predict the percentage of 'leave' votes in wards with missing voting data. The given dataset comprises details about the EU referendum votes from 1070 wards in England. Among all the data, there are 267 missing values needed to predict.

The voting pattern of the EU Referendum is too comprehensive to analyse simply by looking at the data. Thus, we first discuss demographic and socio-economic characteristics and factors which might affect, regardless of the original dataset given.

Age, education level, employment status, housing tenure, and ethnicity are strong demographical indicators of Leave votes: individuals are more likely to support Leave if they are older, less qualified, retired or unemployed and are from a majority ethnic group (white). (Alabrese, E., Becker, S. O., Fetzer, T., & Novy, D., 2019)

The region is also an affecting factor in the voting pattern: Among England, the Leave vote is highest in the East and West Midlands, and lowest in London. What's more, in the most urban category, Remain took over 50% of the vote in a higher number of local authorities than Leave. (Uberoi, E., 2016)

From an economic perspective, according to an article published by Joseph Rowntree Foundation, it was areas where people tended to earn less that voted for Brexit even if these were not always the communities that had been the most badly affected in recent years. (Goodwin & Heath, 2016)

Also, cross effects of factors also account for people's incentive to vote "leave". For example, a more prosperous area may attract highly-educated people, which may further affect the Leave votes.

After gaining these valuable insights, we can prioritise our focus on the most correlated variables for further analysis.

Introduction to the data

The dataset contains over 40 variables. The average proportion of Brexit votes was 51.90%, and the highest percentage of Leave votes in a ward was 78.97%. To streamline the analysis and take all potential variables into account, we classified the original variables into the following: types of areas, including 'unitary authorities', non-metropolitan districts', 'metropolitan districts', and 'London boroughs'; the mean age of adults in a ward; different age groups, including children, young voters, working-age, and retirement; ethnic groups; whether postal votes are counted; current property status; education levels; employment status; population density; proportions of deprived households and social grades. We will further discuss the relationship between these variables and the proportion of Leave votes by box plots, scatterplots, and correlation matrix to discover the correlation between the covariates and "LeavingRate".

Figure 1 suggests a strong negative linear relationship between individuals with level 4 education and the Leave vote. It appears that those with higher levels of education tend to vote for staying in the EU. In contrast, individuals with no qualifications or level 1 qualifications exhibit a strong and similar preference for voting Leave, with correlation coefficients of 0.68 and 0.81, respectively.

From Figure 2, we can see that deprived residents have a greater intention to vote and are less likely to vote Leave. Also, we can find a similar relationship as in the relationship between social grade and the Leave vote. Ward with a higher proportion of people with social grade DE has more Leave votes.

Figure 3 shows a strong positive relationship between adult mean age and the proportion of Leave votes. A ward had a greater proportion of older people had a higher proportion of Leave votes, which we can find a similar trend in the correlation between different age groups and proportions of votes. Young voters between the ages of 18 and 29 tend to vote for Remain, while people at retirement age (beyond 64) have a stronger preference for 'leave'.

Figure 4 shows that individuals in higher-level occupations and students are less likely to vote for leaving the EU. Conversely, those who are short-term or long-term unemployed, or in routine occupations, are more likely to vote for 'leave'. RoutineOccupOrLTU has the biggest proportion under the sector.

Since white is the majority ethnicity in most of the wards. We further categorised the ethnicity into two groups: white-majority with more than 50% of white residents and other-majority. In Figure 5, we can observe a difference between the two box plots. The proportion of Leave votes is significantly higher in areas that have mostly white residents. Thus, we further explored the relationship between ethnicity and the proportion of Leave votes by correlation matrix. Table 1 is the correlation matrix analysing ethnicity and votes. We could see all ethnic minorities prefer to vote for Remain.

Figure 6 represents the relationship between regions and Leave votes. We can observe a significant difference between London and other regions. There are regions with similar box plots, which leads us to introduce hierarchical clustering.

Hierarchical clustering helps us to discover the most similar groups. We utilise this method to classify four different groups, the most similar area groups are within the same group. East Midlands and London are categorised as Group 1 and 3 respectively; Group 2 includes East of England, South East, and South West. West Midlands, North East, North West, and Yorkshire and The Humber are in Group 4.

East Midlands (Group 1) have the highest percentage of industrial output and manufacturing employees, with high levels of employment but a small proportion of higher-skilled occupations. (Beaumont, J., 2009) For regions in Group 2, there are relatively low levels of deprivation and residents are more educated overall. London (Group 3) is different from other regions: the overall education level is highest in England, but the unemployment rate is higher than average.

Regions in Group 4 have common characteristics as they have the highest proportion of the working-age population having no qualifications, and the most manufacturing activities among all the regions. Regions allocated to Group 1 generally have lower population density, while Group 3 has the highest.

To discover the relationship between population density, region, and proportion of Leave votes, we introduced Figure 7. The scatter plot reveals that Group 1 and 4 has the highest proportion of votes for leaving the EU and lowest population densities, followed by Group 2 and Group 3, with Group 3 having the lowest proportion of Leave votes and highest population densities. These findings suggest that the lower the population density of the area, the higher the proportion of votes to leave.

Overall, we conclude that education level, deprivation, age, occupation, ethnicity, region and population densities might be factors affecting leaving rates. While building the model, we should consider both the characteristics of covariates and their interaction effect.

Model building

Decisions of model

From the exploratory analysis we observed a generally linear relationship between all the potential factors and the proportion of Leave votes, hence we used a linear regression model. The variance of the data is not constant, considering that the data is generated from a binomial distribution where variance is equal to $p(1-p)/n$, with p and n varying between samples. The homoscedasticity assumption of linear models cannot be satisfied. Therefore, we carefully started from a generalised linear model with a binomial distribution, considering the binary nature of our response variable. We use the weights function to account for varying sample sizes (number of votes from 1039 to 15148).

Backward Elimination

First, we built a simple generalised linear regression model Model_0 which includes all untransformed covariates with binomial distribution. However, we found that the value of the dispersion parameter is around 53, which is significantly larger than 1, leading us to apply quasi-binomial distribution with all covariates in model Model_1. Although Model_1 seems to have good statistics, some covariates represent the proportions, such as education level or ethnicity, which add up to 1. Their values are perfectly correlated, which may introduce multicollinearity. Therefore Model_2 is built by deleting some redundant covariates based on Model_1, utilising backward elimination at 95% significance level.

We chose our covariates in Model_2 based on the statistical summary of Model_1. The covariates with a p-value larger than 0.05 may be considered uncorrelated.
We didn’t remove all the covariates with a p-value larger than 0.05, since some actually contribute to the model. There are two reasons that account for the large p-value: some covariates are influential after interacting with other covariates; and the p-value might be affected by the quasi-binomial techniques performed in the GLM model. For example, ethnicity covariates are not statistically significant, but we still preserved them because of the interaction effect, referring to one of BBC’s reports, saying ‘Ethnicity was crucial in some places, with ethnic minority areas generally more likely to back Remain.’ (Rosenbaum, M., 2017)

We also made decisions based on our exploratory analysis. We chose L4Quals to represent qualifications because other two covariates have similar linear relationship with response variable and these three covariates can explain the education level of most of the population. Similarly, we selected covariates with largest proportion under that category since they might affect the Leave votes the most, such as 'RoutineOccupOrLTU' under occupation.

To improve our model, we tried lots of combinations of interactions, which will be illustrated in the following.

Interactions

A more urbanised region may have more diverse ethnic groups, which might affect residents' intention to vote "Leave". For example, London is the most diverse area, while the North East has the highest proportion of White British residents. Also, from previous analysis, we can see White, Black, and Asian account for the "LeavingRate" the most. Therefore, we introduce our first three interaction called 'NewGroup:White', 'NewGroup:Black', 'NewGroup:Asian' respectively into a new model called Model_3 to investigate the potential combined effect on Leave votes. We've also added an interaction between regions and percentage of young voters, according to BBC's report that 'Turnout was low in areas with more young people. (BBC News, 2021)

Moreover, we've tried several other interactions. After careful considerations, we added two more interaction terms called 'AdultMeanAge:L4Quals_plus' and 'L4Quals_plus:Deprived' to better explain the interaction effects between: age and qualifications, and qualifications and deprivation. 'Density:NewGroup' is introduced based on our exploratory analysis.

We applied the Chi-squared test and F test to test the goodness of fit of two nested models. The p-value of both tests suggests that the improvement in model fit provided by the more complex model (Model_3) is statistically significant. Among all our models, Model_3 is the best model with suitable covariates, hence we embrace Model_3 as our final model.

Transformation

We tried square root and log transformation to some of the covariates that seem to have nonlinear relationships. However, we found out that the transformation can barely improve the model fit. Thus, we decided not to use any transformation.

Model Checking

Our final model Model_3 represents the relationship between the proportion of Leave votes in the EU referendum and several covariates: routine occupation, population density, proportions of residents owning properties, deprivation, social grades DE, different areas in England, adult mean age, proportions of higher than L4 qualifications and ethnicities (white, black and asian in particular).

After calculation, around 90% of the variation in the proportion of Leave votes can be explained by the model. The QQ plot (Figure 8) of the model generally fits well, suggesting the model is a good fit for the data.

We now check for the assumption:

1. Linearity

We can observe there is no obvious pattern in the residuals vs fitted plot, suggesting that the linearity assumption holds.

2. Independent Errors

In the residuals vs fitted plot, the data is generally randomly scattered around 0, indicating the independent error assumption holds.

3. Distribution

The response variable we examined is a binary outcome which follows the definition of binomial distribution.

Conclusion

In conclusion, some of the covariates appear to be useful in the Leave votes model. To be more specific, deprivation seems to be positively correlated with the Leave votes, suggesting that voters who are more deprived are more likely to vote in favour of leaving the EU. On the other hand, students tend to vote for Remain, indicating that education may be a significant factor in shaping voting behaviour.
Property ownership and population density both have a negative association with voting for Leave. It indicates that those who own properties may be more inclined to vote in favour of remaining in the EU, and areas with higher population densities tend to have less support for leaving the EU.

Age and education level seems to have a combined negative effect on the proportion of votes for Leave. With another negative interaction term between deprivation and qualifications, we could confirm that education is a significant factor. Additionally, different ethnicities and regions in England may also exhibit varying voting behaviours, suggesting that cultural and regional factors may play a role in the voting behaviour of the EU referendum.

Limitations

However, there are some limitations in Model_3. First, we used GLM to build our model, but GAM may still produce good results as it could detect more complex interactions and it is more flexible with non-linear relationships. Moreover, there are several outliers in the data. Although we choose to not exclude them after considerations, they might still affect our model predictions and reliability. Collinearity exists in the research data, which might influence Model_3's reliability. Also, in the tail region, the QQ plot of Model_3 deviates noticeably below the expected diagonal line, which might indicate a lack of fit.

Reference

Uberoi, E. (2016). *Analysis of the EU Referendum results 2016. House of Commons Library Briefing Paper.* Available at: https://commonslibrary.parliament.uk/research-briefings/cbp-7639/ (Accessed: 23 April 2023).
Goodwin, M. & Heath, O. (2016). *Brexit Vote Explained: Poverty, Low Skills and Lack of Opportunities. Joseph Rowntree Foundation.* Available at: https://www.jrf.org.uk/report/brexit-vote-explained-poverty-low-skills-and-lack-opportunities (Accessed: 23 April 2023).
Beaumont, J. (2009). *Portrait of the East Midlands. Office for National Statistics. Available at:* https://doi.org/10.1057/rt.2009.5 (Accessed: 23 April 2023)
BBC News. (2021). *EU referendum: The result in maps and charts. BBC News.* Available at: https://www.bbc.co.uk/news/uk-politics-36616028 (Accessed: 23 April 2023).
Rosenbaum, M. (2017). *Brexit: All you need to know. BBC News.* Available at: https://www.bbc.co.uk/news/uk-politics-38762034 (Accessed: 23 April 2023).
Alabrese, E., Becker, S. O., Fetzer, T., & Novy, D. (2019). *Who voted for Brexit? Individual and regional data combined. European Journal of Political Economy*, *56*, 132-150. https://doi.org/10.1016/j.ejpoleco.2018.08.002 (Accessed: 23 April 2023).

Appendix

Figure1: Leaving Rate vs. Qualification Level

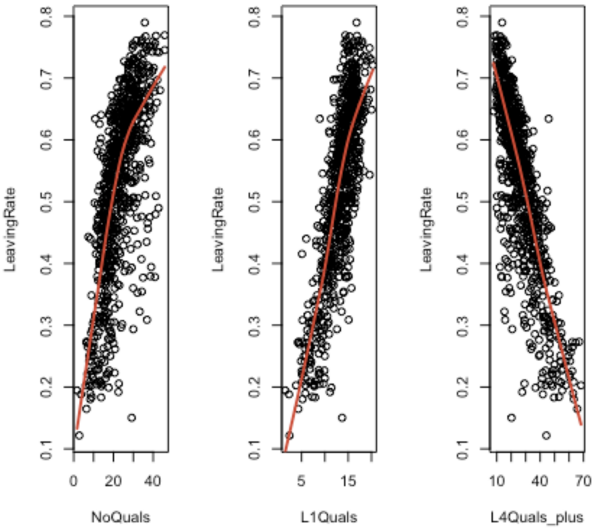


Figure2: Leaving Rate vs. Deprivation / MultiDeprivation

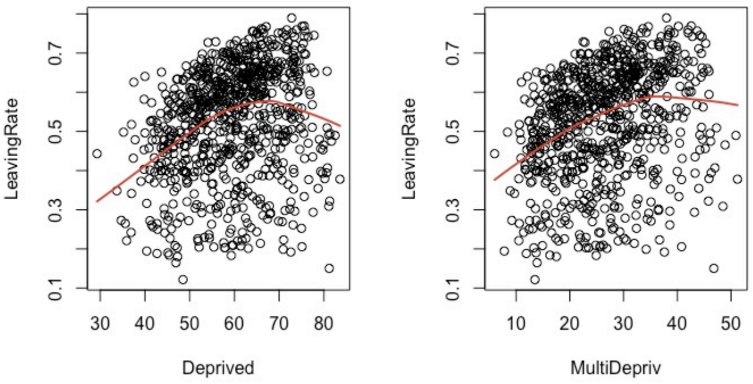


Figure3: Leaving Rate vs. Adult Mean Age

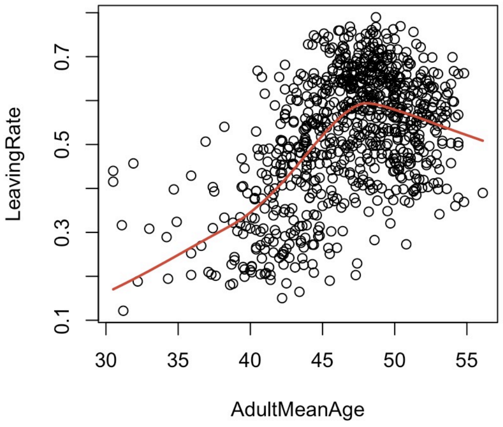


Figure4: Leaving Rate vs. Occupation

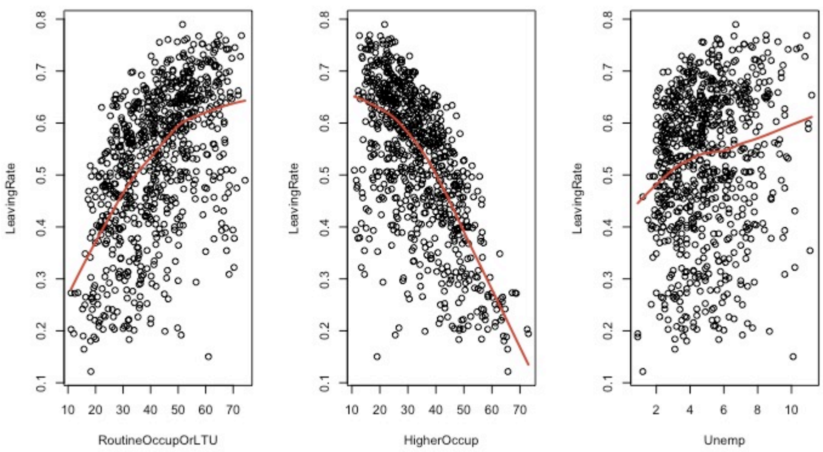


Figure5: Boxplots of LeavingRates on different ethnic groups

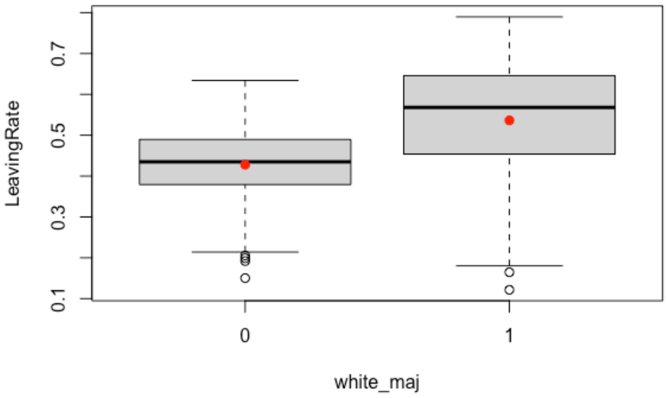


Figure6: Boxplots of LeavingRates on different regions

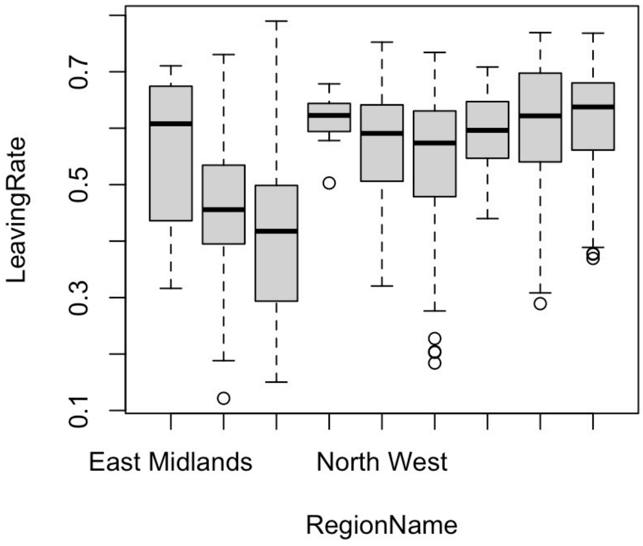


Figure7: Scatterplot between LeavingRate and population density

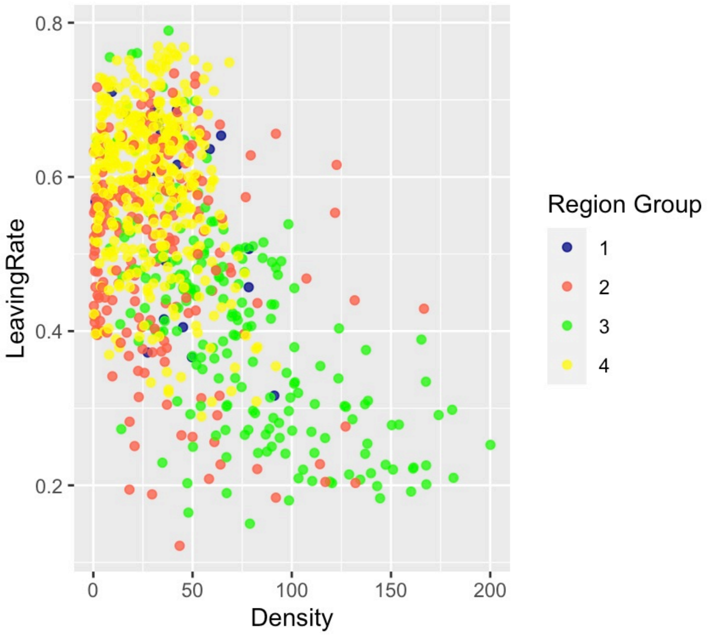


Figure8: various plots of Model_3

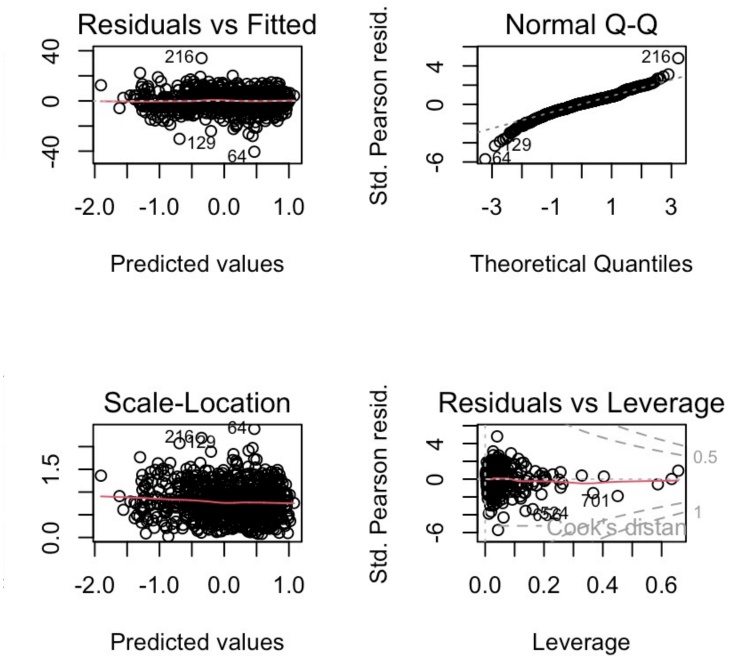
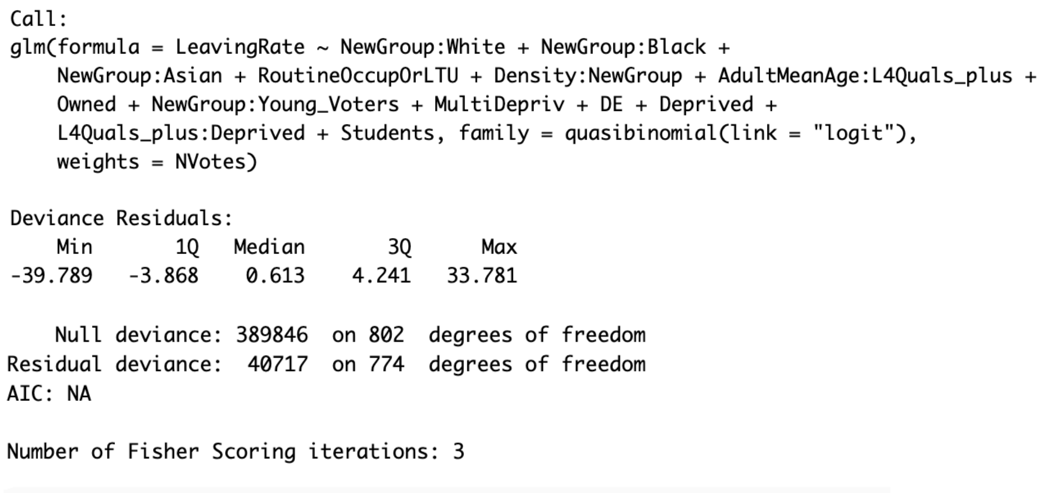


Table 1:



```
Call:
glm(formula = LeavingRate ~ NewGroup:White + NewGroup:Black +
    NewGroup:Asian + RoutineOccupOrLTU + Density:NewGroup + AdultMeanAge:L4Quals_plus +
    Owned + NewGroup:Young_Voters + MultiDepriv + DE + Deprived +
    L4Quals_plus:Deprived + Students, family = quasibinomial(link = "logit"),
    weights = NVotes)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-39.789   -3.868    0.613    4.241   33.781

    Null deviance: 389846  on 802  degrees of freedom
Residual deviance:  40717  on 774  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 3
```

Two of us contributed equally.