A Statistical Analysis Report of Resident Happiness for Durham County Policymakers

ISDS Group 40

Page count:

Executive Summary: 1

Findings: 3

Methodology: 6

Appendix: 8

Executive Summary

Our team has made use of advanced data analysis techniques to help local authorities in County Durham, UK, improve residents' overall satisfaction with local life. We believe our findings will help local policymakers focus on the most important issues to have the greatest positive impact on community well-being. In this study, we used a mathematical modeling method called multiple linear regression, combined with other statistical means, to analyze the impact of several variables on the overall satisfaction of residents in the North of England. We primarily utilized five variables to examine the satisfaction of local residents:

- Influence_Decisions: The proportion of residents who feel they can influence decisions made in their local area.
- Get_On_Well: The proportion of residents who believe people from different backgrounds get along well in their community.
- Belong: The proportion of residents who feel they belong to their neighborhood.
- Drug_Use_And_Selling: The proportion of residents who consider drug-related issues to be a problem in their area.
- Area: the region an authority is located within.

We have observed that these factors are closely associated with individuals' satisfaction regarding where they live. The biggest impact was on a factor that negatively correlated with happiness: drugs. Areas with fewer issues related to drugs demonstrated markedly higher levels of satisfaction compared to other regions. Get_On_Well and Belong index emerged as the strongest positive determinants, with almost the same coefficient. This indicates that mutual trust and understanding among neighbors as well as harmonious get along well with people from diverse backgrounds, contribute significantly to overall satisfaction. Additionally, regional factors also affect people's happiness index to a certain extent. For instance, in the Northern model, people in Yorkshire and The Humber regions have slightly higher overall satisfaction. However, there is a specific interaction with the Get_On_Well index, which indicates that socio-cultural differences in different regions will affect the relationship between variables. Further research could explore these unique regional patterns in more detail. Based on these findings, our policy recommendations are as follows:

The government and relevant authorities should advocate for comprehensive education and treatment programs aimed at addressing drug addiction, and help residents in need return to normal life. At the same time, law enforcement officials should strengthen neighborhood patrols to curb drug trafficking and abuse. Community leaders should organize appropriate cultural exchanges or volunteer services to bring different groups together to foster trust and friendship, thereby enhancing community cohesion and sense of belonging. Finally, policies should be formulated according to the economic and cultural characteristics of County Durham and information feedback strategies should be collected regularly to adapt to changes in people's actual needs.

By focusing on reducing drug-related issues, enhancing community relationships, and involving residents in local decision-making processes, Durham's local authority can actually improve the overall satisfaction of local residents. It is suggested that the government could conduct further research on the differences between different regions to investigate why certain regions have different patterns of satisfaction. Comprehending these underlying cultural and social determinants can guide the development of more precise and effective policies.

Findings

The final multivariate linear regression model we produced predicts overall satisfaction, defined as the proportion of residents in northern regions (North East, North West, and The Yorkshire and the Humber) who responded positively to their satisfaction with living in their authority area (scaled from 0 to 1). The model includes the following predictors:

- Influence_Decisions: The proportion of respondents answering 'Yes' to the question; "Do you feel able to influence decisions made in your local area?"
- Get_On_Well: The proportion of respondents answering 'Yes' to the question; "Do you believe people from different backgrounds get on well together in your local area?"
- Belong: The proportion of respondents answering 'Yes' to the question; "Do you feel you belong to your neighbourhood?"
- Drug_Use_And_Selling: The proportion of respondents answering 'Yes' to the question; " Do you consider drug use and/or drug selling to be a problem in the local area?"
- Whether or not the authority is located in Yorkshire and The Humber: A dummy variable (1 = in Yorkshire area, 0 = not).
- Area Yorkshire and The Humber*Get_On_Well: An interaction term capturing how the effect of 'Get_On_Well' differs for York and The Humber residents.

The model presented below:

Overall Satisfaction = 0.38943 + 0.19412 * Influence_Decisions + 0.33665 * Get_On_Well + 0.33632 * Belong - 0.36863 * Drug_Use_And_Selling + 0.18831 * Area Yorkshire and The Humber - 0.23456 * (AreaYorkshire*Get_On_Well)

(where Area Yorkshire and The Humber = 1 if the authority is in Yorkshire and The Humber, and 0 if not)

The reason we focus on a smaller dataset is because one of the most important findings of our model is that the UK is not split evenly in terms of their overall satisfaction with living in an area. Our original model, that included the entirety of England, found a statistically significant increase of the overall satisfaction levels for living in a certain area. Of course, we cannot recommend for the Durham County Council to ask people to move to a different part of England if they want to have a higher level of satisfaction, but it does show that there are underlying factors other than the four main variables produced by this model. We believe this to be caused by external factors, such as the economic situation in areas in the South surrounding the wealthier areas of London and Surrey, and the availability of high-paying jobs in the North in comparison to the South.

Considering these factors, we chose to focus our final model (seen above) on just the northern most areas in England. Although there was an area that accounted for just the North East area, there were only three authorities in the data for the region, which is not enough for us to conduct meaningful research. As a response to this, we use three regions, the North East, North West and Yorkshire. These areas should bear the most resemblance to the cultural factors displayed in Durham, hence helps remove the issue of area when it comes to explaining how to improve the overall satisfaction of individuals in Durham.

Using this model, we can explain 89.82% of why changes to the overall index happens, in comparison to only 86% when we do not remove the South from our analysis, indicating the importance of why our model is useful for explanatory purposes relating to satisfaction levels.

To improve the model's stability, variables were standardized by converting the percentages into proportions (scale 0 to 1). Each coefficient represents the change in the overall satisfaction score for a one-unit change in the corresponding variable. For example, a one-unit increase in 'Drug_Use_And_Selling' reduces satisfaction by 0.36863 units.

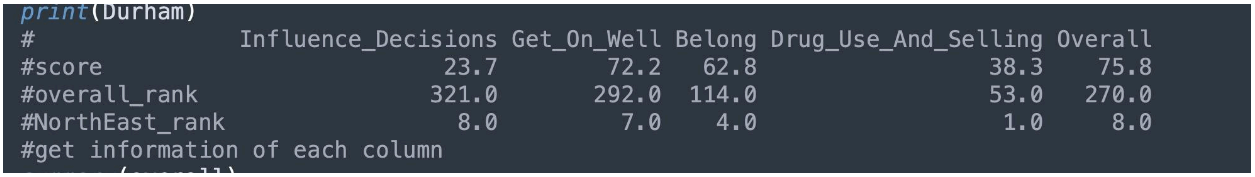
By comparing the coefficients, we identify three most influential factors affecting satisfaction including 'Drug_Use_And_Selling', 'Get_On_Well', and 'Belong'. This suggests that these are three factors any government in the north area should focus on, in order to improve the overall satisfaction of their residents. Specifically, 'Drug_Use_And_Selling' is the largest negative driver of satisfaction with a coefficient of -0.36863, which means addressing drug-related concerns is essential for improving satisfaction. In contrast, both 'Get_On_Well' and 'Belong' have strong positive influences on satisfaction, with coefficients of 0.33665 and 0.3363,2, respectively. This suggests that fostering social cohesion and a sense of belonging should remain key priorities for local government. While the coefficient of 'Influence_Decisions' is small (0.19412), it still yields a statistically significant result, therefore this variable has a moderate but meaningful impact on satisfaction, suggesting that efforts to increase public engagement in decision-making can still yield benefits.

As of the category variable, Area, the model reveals that living in Yorkshire and the Humber has a positive direct effect on satisfaction (+0.18831). However, this effect is moderated by the interaction term with 'Get_On_Well' (- 0.23456). While Yorkshire residents generally report higher satisfaction, those who feel strongly positive about 'Get_On_Well' may experience lower satisfaction levels. This could reflect unique social dynamics in different regions, such as perceived social isolation or unaddressed local challenges. Further research is needed to explore these factors in detail.

The model suggests a linear relationship between predictors and satisfaction, with no evidence of diminishing returns. For example, improvements in 'Get_On_Well' consistently increase satisfaction, regardless of initial levels. However, this assumption may oversimplify real-world dynamics, as diminishing returns could occur when predictors already perform exceptionally well. For instance, further enhancing 'Get_On_Well' in communities with high baseline levels of cohesion may yield smaller incremental gains in satisfaction than the model predicts. This limitation highlights the need for caution when interpreting the results and suggests exploring non-linear models to better capture such nuances in future analysis.

Although the model has a high adjusted R-squared score (0.8982), indicating a good fit to the data, it simplifies complex relationships. Factors such as urban vs. rural differences, economic disparity, or other unmeasured influences may also impact satisfaction but are not captured in this analysis. Further research should consider incorporating additional variables or exploring region-specific nuances to provide a more comprehensive understanding of resident satisfaction.

Overall, we can identify that Drug_Use_And_Selling is the most significant negative influential factor of satisfaction in the northern regions. Durham ranks the highest in the North East (among 8 authorities) in concerns over drug-related issues, indicating that this problem is particularly acute for its residents. In contrast, Get_On_Well is the strongest positive driver of satisfaction for residents in the northern area. However, Durham ranks second to last in the North East on this factor, reflecting weaker social cohesion compared to other authorities in the North East. Therefore, to improve the satisfaction score in Durham, addressing the drug-related challenges and fostering stronger community relationships is key focuses for local government. By addressing these two aspects, the government can significantly enhance the quality of life for its residents while improving their overall satisfaction.

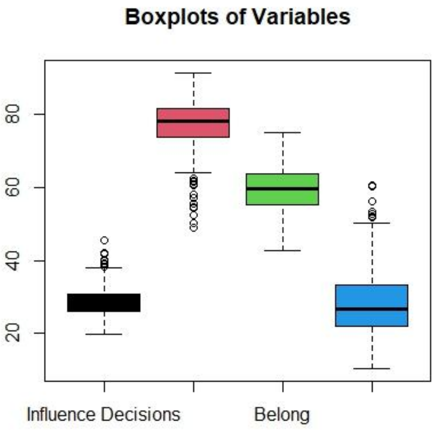


```
print(Durham)
#               Influence_Decisions Get_On_Well Belong Drug_Use_And_Selling Overall
#score                        23.7        72.2   62.8                 38.3    75.8
#overall_rank                321.0       292.0  114.0                 53.0   270.0
#NorthEast_rank                8.0         7.0    4.0                  1.0     8.0
#get information of each column
```

Methodology

The findings and the analysis of the factors listed above are based on the multivariable linear regression model we constructed from the UK social health dataset. This section will give a brief outline of the construction of the model, focusing on the modification of the dataset, variables correlation, the variable combinations and assessing the credibility of the final model.

The modifications to the dataset for facilitating the regression



The outliers were found during our first assessment of the dataset. It is not clear if the outliers in the data represent a statistical error rather than some regional pattern in social culture, not to mention whether the data really shows the preference of local residents. For instance, while most of the authorities have a low common agreement on *the Influence_Decisions* variable, most authorities in the *London Area* have the highest percentage; these data points are the outliers of this variable. So excluding outliers should be assessed carefully for not undermining the model's performance. The following analysis will be based on the outliers' influence in the initial model, which is constructed on the initial dataset without any other modification.

The outliers found using the quantile method exist in each variable except the *Belong* variable. The normality test on each numerical data shows that only the *Belong* data follows a normal distribution, which other variables do not. Hence different standards are used in finding the outliers: for the *Belong* variable, the Z-score method is applied; and for other variables, the quantile method will be used.

Two multivariable linear regression models were built using the initial data and default linear regression, with and without outliers, respectively. The comparison in the R^2 values between the two models shows that the model's performance had declined by two percent if we excluded the outliers (with outliers: 0.8492, without outliers: 0.8267). The predictor coefficient in both models is statistically significant under 0.05 significance level. From the residual analysis perspective, the RSE is slightly reduced (with outliers: 2.764, without outliers: 2.625). Also, the normality of the residuals improved (p increased from 0.08642 to 0.2369), indicating that the residual is basically in line with the normal distribution.

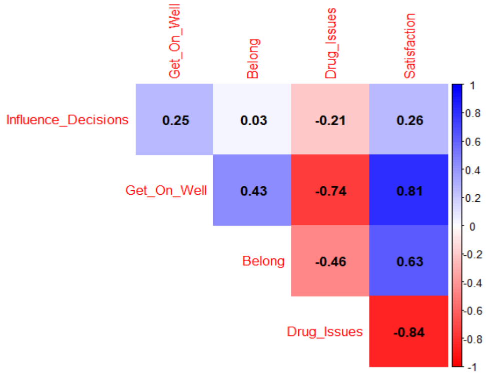
Generally speaking, removing outliers improves the residual normality, but it reduces the explanatory power and overall performance of the model. Based on the interpretation above, we decided not to remove the outliers in order to build a more accurate and explanatory model.

Another modification to the model is that we divide the numerical data by 100 since they represent the percentage of agreement on the questions. This will facilitate the variable combination process and make the combination more reasonable as this roughly shows the overlap in multiple variables.

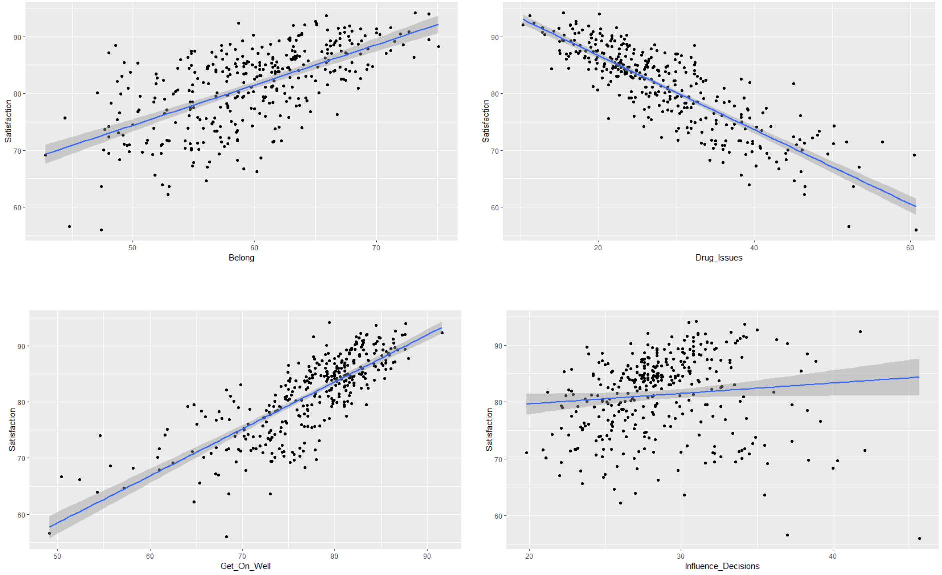
The decisions in predictor variable selection and combination

Checking the linearity of the relationship between predictor and response variables

To decide which variables from our dataset would be appropriate to include in our model, we first made sure that all the continuous numeric variables had a suitable linear relationship with the general satisfaction. A linear relationship is a prerequisite to creating a multivariate linear regression model, and thus any individual variable that failed to display this sort of relationship would have to be excluded from the model. After identifying that all of the numeric variables are not the normal distribution except Belong, we decided to calculate Spearman's correlation coefficients between each of these variables and general satisfaction rather than calculating Pearson's correlation coefficients.



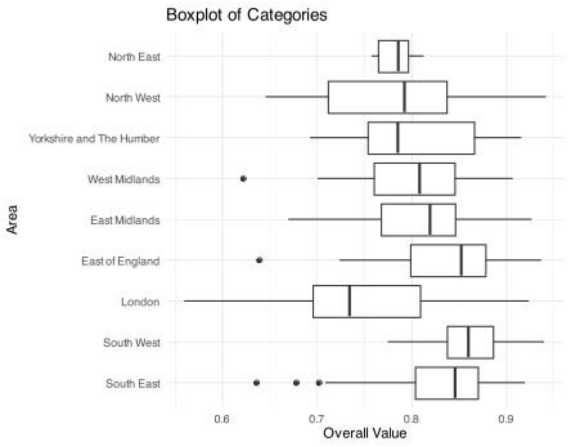
These showed moderately strong correlations with satisfaction, with the notable exception of Influence Decisions. To check whether this was a consequence of a non-linear relationship between Influence Decisions and Satisfaction, and that there was no hidden non-linear relationship between one of the other variables and satisfaction, we created a series of scatter plots. These are included below:

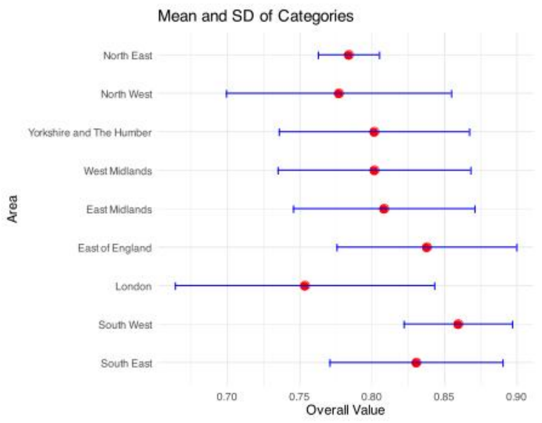


The scatter plot of Influence Decisions and Satisfaction demonstrates a slight and weak positive linear relationship. Except for Drug Issues, which reveal a highly negative linear relationship with satisfaction, the other two variables display a highly positive linear relationship with satisfaction.

Dealing with the categorical data and the selection of the data points

The correlations between the numeric data and the Overall variables have emerged as highly correlated. One problem has been raised for the only categorical data Area, which is divided into nine categories corresponding to nine regions in England. It is hard to directly tell if it is helpful to include these variables for better regression and what things need to be done. We need more descriptive data from each category for a more plausible linear regression model for further implication and modification. Hence, we calculated the descriptive data for each region's Overall variable. Though for the purpose of the suggestions on the government, the factor of the Area may be of no use for the county, the coefficient is still calculated for regression purposes and the analysis of the overall England data.





The plots showed descriptive data about each category, also known as the Area. We visualized the Overall variable of each category in the order of the regions, roughly from north to south. From the plots, we can tell that the distribution of the Overall variable follows a particular pattern, except for London. So, it is reasonable to conclude that the northern and southern regions can have a difference in the regression pattern on the overall variable. Also, the plot shows that the median value of the three northern regions is close.

Therefore, we defined two different models. One model represents the performance of England's nine regions, while the other concentrates on three areas—the North East, North West, Yorkshire, and Humber—to provide more local suggestions for County Durham, which is located in the North East region. This model will be named the North model. It will contribute mainly to our analysis, while the whole England model offers a reference.

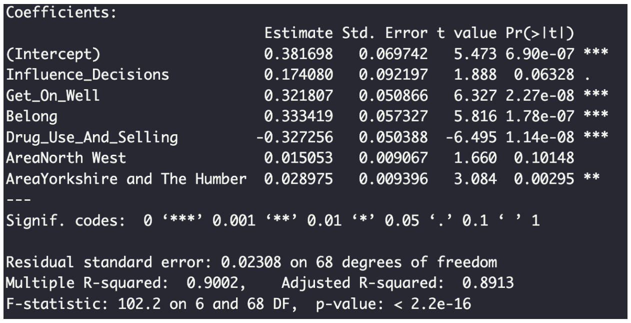
We selected these three areas for the North model rather than focusing on a specific region because the North East, where County Durham is located—only contains eight authorities. The lack of observations will make the model less reliable for regression. Therefore, two regions besides the North East are included in the regression. With the addition of the other two areas, the overall rank will be 81. This shows enough credibility for regression with five predictor variables.

Another issue to be concerned with is whether the categorical data area should be included as the predictor variable. As the variables show the difference of the regression model in different places, the County of Durham isn't able to change where it is located, so the utility of including this factor may not be helpful for the county to make constructive decisions. However, It may provide some implicit information about the region and connect the Overall satisfaction mark with the region itself, meaning that the region's specific culture and society contribute a particular part to the satisfaction. Though further quantitative and qualitative research is needed, this will also provide supportive information on decision-making. For regression and implication purposes, we need the categorical data included.

Accordingly, we decide to include ‘Area’ as a dummy variable in our modeling. We opt to cluster England into the north part as the above and a south part. We combine North West and North East as None-Yorkshire for the north part since these two regions have close mean overall satisfaction. Ultimately, the initial performance showed that the North model performs better than the whole England model (the adjusted R-squared: England: 0.8617, North: 0.8913). We will use the north model for further modification.

Variable selection and combination options

As mentioned above, the Influence_Desicion variable has a weak correlation with the Overall satisfaction data, and Area data is included as dummy variables. The Initial models for the North area show significance in each variable coefficient except the Influence_Decisions variable and most Area dummy variables under 0.05 significance level, which means they are not able to reject the null hypothesis that their coefficient is equal to zero. Also, the model has a high performance(Whole England adjusted R-squared: 0.8617, North Model Adjusted R-squared: 0.8913). Moreover the high correlation between the Drug variable and other numerical variables needs assessment, this has raised the concerns for the colinearity. The contradiction between high performance and low significance needs to be addressed by the variable selection and combinations, as well as the colinearity issues.



```
Coefficients:
                            Estimate Std. Error t value Pr(>|t|)    
(Intercept)                 0.381698   0.069742   5.473 6.90e-07 ***
Influence_Decisions         0.174080   0.092197   1.888  0.06328 .  
Get_On_Well                 0.321807   0.050866   6.327 2.27e-08 ***
Belong                      0.333419   0.057327   5.816 1.78e-07 ***
Drug_Use_And_Selling       -0.327256   0.050388  -6.495 1.14e-08 ***
AreaNorth West              0.015053   0.009067   1.660  0.10148    
AreaYorkshire and The Humber  0.028975   0.009396   3.084  0.00295 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02308 on 68 degrees of freedom
Multiple R-squared:  0.9002,	Adjusted R-squared:  0.8913 
F-statistic: 102.2 on 6 and 68 DF,  p-value: < 2.2e-16
```

We used the best subgroup selection method to show the significance of each variable. This also suggests that the Influence_Decision variable is not as significant as other numerical variables. The five best variable selection orders for the North model are Drug_Use_And_selling, Get_On_Well, Belong, AreaYorkshire and The Humber, and Influence_Decisions. This indicates that the Influence_Desicions variable is less significant than other numerical variables, and the AreaNorthWest dummy variable(which represents the North West Area category) needs modification.

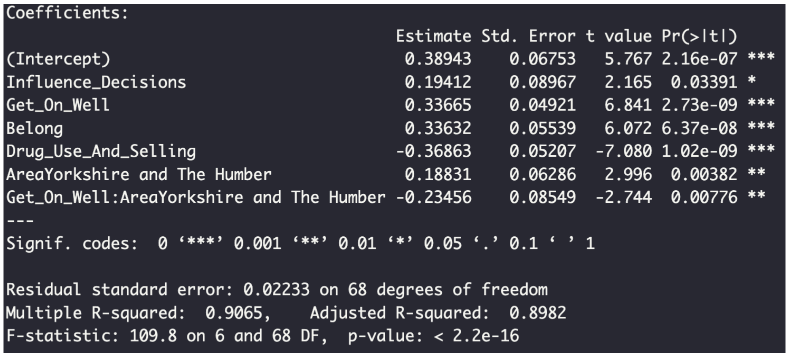
First, we modified the area data and combined the North West and North East into the non-Yorkshire category, as in the initial model, the North East is the reference category, and the significance of the AreaNorth West coefficient is relatively low. A new dataset is therefore constructed for the next steps. For the Influence_Decision, we built two models with and without the Influence_Decisions variable, respectively, to compare the performance. The result shows there is only a slight difference between models and totally not affect the performance (adjusted R-square: with Inf: 0.8886, without inf: 0.885), so we decided not to exclude the Influence_Decision variable. This is entirely plausible as in a democratic society, residents' impact on decisions will give them more satisfaction in society.

For the Drug variable’s high correlation with the Get_On_Well Variable, we performed Spearsman’s correlation between this variable and other variables showing that this variable has a high correlation with the Get_On_Well variable, as it is reasonable that higher drug issues mean more violence, crime and unsafety in local area. However, the VIF test shows that the Drug variable has a low significance with other numerical variables(VIF = 2.56, which is below 5), and the best subgroup selection method also suggests this variable as the prior variable. Hence this will not be excluded.

Due to these outcomes, we combined the North West and North East categories to non-Yorkshire and chose not to exclude the Influence_Decision variable and the Drug Variable. The up-to-date model still has a low significance in the Influence_Decisions coefficient, so the final step we made is to combine variables with each other variable, including Area data, to show the cumulative effect and the difference of each variable coefficient in different areas. Based on the new coefficient significance, we decided to add the Area*Get_On_Well variable as only this coefficient shows considerable significance and improved the performance (adjusted R-squared: 0.8982). This gives us an outcome where all of our variables pass the 0.05 significance test, that is to say we can reject the null hypothesis that there is no impact of any of the variables on overall satisfaction levels.

The final model

This final model uses 81 authorities data points from three regions: the North East, the North West, and Yorkshire and the Humber. It includes all numerical variables and combines the North East and North West categories as the non-Yorkshire category for higher performance. All the numerical variables are on the same scale. An additional Area*Get_On_Well variable is added for better regression and interpretation purposes. This model shows high and credible performance in analyzing the satisfaction factors in the North Area.

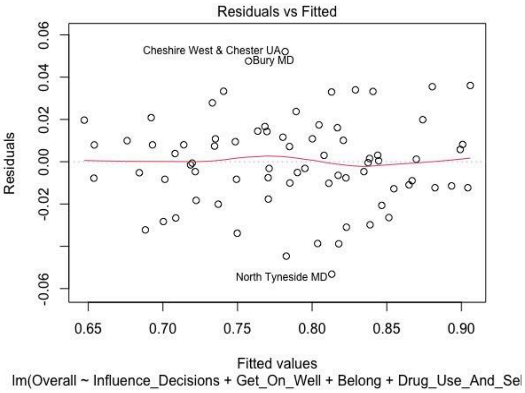


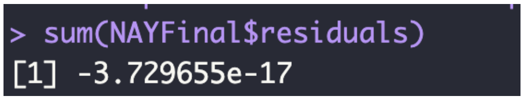
```
Coefficients:
                                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)                            0.38943    0.06753   5.767 2.16e-07 ***
Influence_Decisions                    0.19412    0.08967   2.165  0.03391 *  
Get_On_Well                            0.33665    0.04921   6.841 2.73e-09 ***
Belong                                 0.33632    0.05539   6.072 6.37e-08 ***
Drug_Use_And_Selling                  -0.36863    0.05207  -7.080 1.02e-09 ***
AreaYorkshire and The Humber           0.18831    0.06286   2.996  0.00382 ** 
Get_On_Well:AreaYorkshire and The Humber -0.23456    0.08549  -2.744  0.00776 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02233 on 68 degrees of freedom
Multiple R-squared:  0.9065,	Adjusted R-squared:  0.8982 
F-statistic: 109.8 on 6 and 68 DF,  p-value: < 2.2e-16
```

The final model we constructed has shown good properties as a multiple linear regression model. The F-test on the model is hugely significant (p-value < 2.2e-16). The R-squared is 0.8982, which means it can explain 89% of the North dataset. Also, this model performs well in the assumption tests.

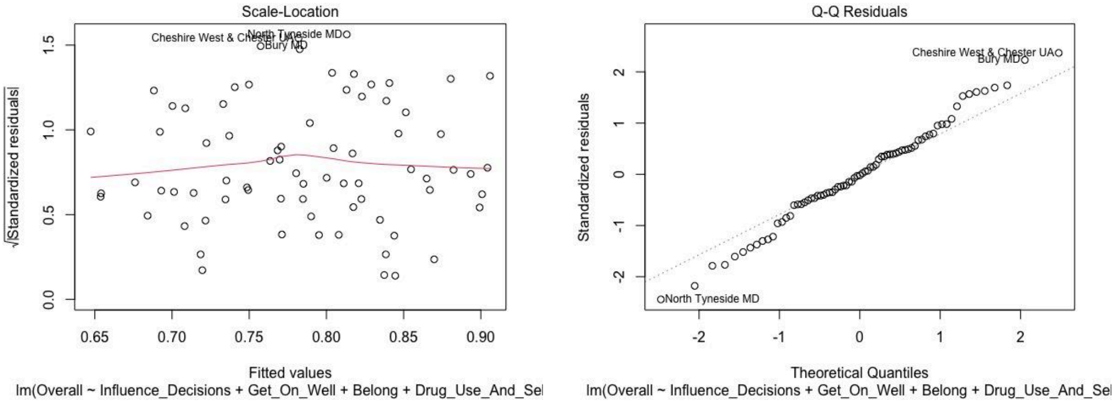
This model meets the assumption of linearity, as the residuals are similar above and below zero. From the residuals vs. Fitted plot, no apparent patterns could be found, as the residuals are even on different fit values. The residual sum is close to 0, which also highly supports the assumption.



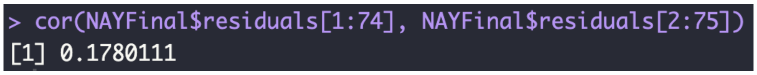


```
> sum(NAYFinal$residuals)
[1] -3.729655e-17
```

This model also meets the assumption of homoscedasticity. The fitted-value VS standardized residuals plot shows a constant, random, and even distribution of the residuals. There is no pattern in the spread of residuals, as the standard errors are not biased by the plot.

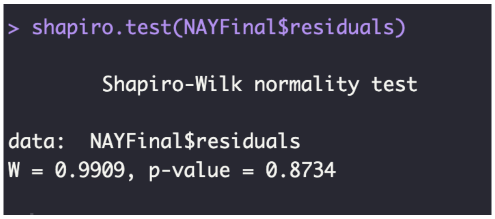


The final model also fulfilled the assumption of no autocorrelation of residuals. The correlation between each consecutive residuals is 0.17. This is not high enough for considering having auto correlation. Therefore this shows the residuals of the model are independent from each other.



```
> cor(NAYFinal$residuals[1:74], NAYFinal$residuals[2:75])
[1] 0.1780111
```

This final model also fulfilled the normality test on residuals. The Shapiro test on residuals has a very low significance therefore the residuals are normally distributed(p-value = 0.8734). Also, the QQ plot shows that the residual is close to a straight line in the plot. According to the central limit theorem, the size of the residual is greater than 30 and hence can be considered following a normal distribution.

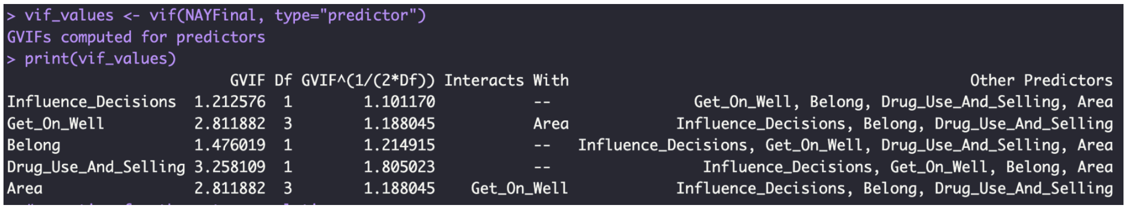


```
> shapiro.test(NAYFinal$residuals)

	Shapiro-Wilk normality test

data:  NAYFinal$residuals
W = 0.9909, p-value = 0.8734
```

The final model met the assumption of no multicollinearity. The GVIF is tested in the model with all values less than 5, which shows that the multicollinearity is not significant in the final model. The model is reliable.



```
> vif_values <- vif(NAYFinal, type="predictor")
GVIFs computed for predictors
> print(vif_values)
                         GVIF Df GVIF^(1/(2*Df)) Interacts With                                             Other Predictors
Influence_Decisions  1.212576  1        1.101170           --                Get_On_Well, Belong, Drug_Use_And_Selling, Area
Get_On_Well          2.811882  3        1.188045           Area            Influence_Decisions, Belong, Drug_Use_And_Selling
Belong               1.476019  1        1.214915           --   Influence_Decisions, Get_On_Well, Drug_Use_And_Selling, Area
Drug_Use_And_Selling 3.258109  1        1.805023           --                 Influence_Decisions, Get_On_Well, Belong, Area
Area                 2.811882  3        1.188045    Get_On_Well            Influence_Decisions, Belong, Drug_Use_And_Selling
```

Appendix: code

```r
install.packages("corrplot")
install.packages("ggcorrplot")
library(ggplot2)
if (!requireNamespace("car", quietly = TRUE)) {
  install.packages("car")
}
library(car)

########################
#initialize the dataset#
########################
UK <- read.csv('UKContentment.csv')
str(UK)
#There are a total of 352 rows of observations and 7 variables in the data, five of
which are numeric variables and two of which are character variables.
summary(UK)
anyNA(UK)
#The results show that there are no missing values in the data set.

#assign the authority name as a row name #
overallRAW <- data.frame(UK[, -1], row.names = UK[, 1])
overall <- data.frame(UK[, -1], row.names = UK[, 1])
overall$Influence_Decisions <- overall$Influence_Decisions/100.0
overall$Get_On_Well <- overall$Get_On_Well/100.0
overall$Belong <- overall$Belong/100.0
overall$Drug_Use_And_Selling <- overall$Drug_Use_And_Selling/100.0
overall$Overall <- overall$Overall/100.0

# Extract the relevant variable
Influence_Decisions <- UK$Influence_Decisions
Get_On_Well <- UK$Get_On_Well
Belong <- UK$Belong
Drug_Issues <- UK$Drug_Use_And_Selling
Satisfaction <- UK$Overall

# Extract the area identifiers
Authority <- UK$Authority
Area <- UK$Area

data <- data.frame(
  Authority = Authority,
  Area = Area,
  Influence_Decisions = Influence_Decisions,
  Get_On_Well = Get_On_Well,
  Belong = Belong,
  Drug_Issues = Drug_Issues,
  Satisfaction = Satisfaction
)

####################
#tests on variables#
####################

#normality test
shapiro.test(Influence_Decisions)
shapiro.test(Get_On_Well)
shapiro.test(Belong)
shapiro.test(Drug_Issues)
shapiro.test(Satisfaction)
#Showing that every variable except belong is not normally distributed

# Correlation data
num_UK<-UK[sapply(UK,is.numeric)]
library(corrplot)
cor_matrix <- cor(num_UK, method = "spearman")
corrplot(cor_matrix, type = "upper", method = "color", col =
colorRampPalette(c("red", "white", "blue"))(200),
         addCoef.col = "black", tl.col = "red", tl.srt = 90, diag = FALSE,)

#The outliers
sapply(num_UK,function(x) c(Mean = mean(x),Median = median(x),SD = sd(x)))
lapply(names(num_UK), function(x){
  hist(num_UK[[x]],
       main = paste("Histogram of",x),
       xlab = x,
       col = "lightblue",
       border = "black")
})

#Use boxplot to find outliers.
boxplot(num_UK[,
c("Influence_Decisions","Get_On_Well","Belong","Drug_Use_And_Selling")],
        main = "Boxplots of Variables",
        col = c(1,2,3,4,5),
        vlab = "Values",
        names = c("Influence Decisions","Get on well","Belong","Drug use and
Selling"))

outliers_list <- lapply(num_UK,function(x){
  Q1 <- quantile(x,0.25)
  Q3 <- quantile(x,0.75)
  IQR <- Q3-Q1
  lower_bound <- Q1 - 1.5 * IQR
  upper_bound <- Q3 + 1.5 * IQR
  outliers <-x[x < lower_bound | x > upper_bound]
})
print(outliers_list)

#Use Z-scores to detect outliers.
z_score_outliers <- lapply(num_UK, function(x){
  z_scores <- scale(x)
  outliers <- x[abs(z_scores) > 3]
  return(outliers)
})
print(z_score_outliers)

##########################################################
#assign the authority name as row names#
#standardise the percentage data for further combinations#
##########################################################
dim(overall)
overallrows = nrow(overall) # numbers of values
overallcols = ncol(overall)-1 # numbers of predictor variables

#####################################################
#give the summary of County Durham and each category#
#####################################################
NoArea <- subset(overall, select = -c(Area))
Durham<- NoArea[0,]
Durham["score",] <- NoArea["County Durham UA",]
print(Durham)

#get the rank
Durham["overall_rank",] <- list(1,1,1,1,1)
for(cols in colnames(Durham)){
  for(rows in rownames(NoArea)){
    if(Durham["score", cols] < NoArea[rows, cols]){Durham["overall_rank",cols] =
Durham["overall_rank",cols]+1}
  }
}
Durham["NorthEast_rank",] <- list(1,1,1,1,1)
for(cols in colnames(Durham)){
  for(rows in rownames(overall)){
    if(overall[rows, "Area"] == "North East" && Durham["score", cols] <
NoArea[rows, cols]){Durham["NorthEast_rank",cols] =
Durham["NorthEast_rank",cols]+1}
  }
}
print(Durham)
```

```
#               Influence_Decisions Get_On_Well Belong Drug_Use_And_Selling Overall
#score                         23.7        72.2   62.8                 38.3    75.8
#overall_rank                 321.0       292.0  114.0                 53.0   270.0
#NorthEast_rank                 8.0         7.0    4.0                  1.0     8.0
#get information of each column
summary(overall)


#descriptive data for regions
London <-subset(overall, overall$Area == "London")
East_Midlands <-subset(overall, overall$Area == "East Midlands")
East_Of_Englands <- subset(overall, overall$Area == "East Of England")
North_East <- subset(overall, overall$Area == "North East")
North_West <- subset(overall, overall$Area == "North West")
South_East <- subset(overall, overall$Area == "South East")
South_West <- subset(overall, overall$Area == "South West")
West_Midlands <- subset(overall, overall$Area == "West Midlands")
Yorkshire_and_The_Humber <- subset(overall, overall$Area == "Yorkshire and The
Humber")

overall_stats <- data.frame("mean" = numeric(),
                            "sd" = numeric(),
                            "median" = numeric(),
                            "max" = numeric(),
                            "min" = numeric())

regions <- unique(overall$Area)

for(areas in regions){
  sub <- subset(overall, overall$Area == areas)
  quantile_value <- quantile(sub$Overall, probs = c(0.25, 0.75))

  overall_stats[areas, "mean"] = mean(sub$Overall)
  overall_stats[areas, "sd"] = sd(sub$Overall)
  overall_stats[areas, "max"] = max(sub$Overall)
  overall_stats[areas, "Q3"] = quantile_value[2]
  overall_stats[areas, "median"] = median(sub$Overall)
  overall_stats[areas, "Q1"] = quantile_value[1]
  overall_stats[areas, "min"] = min(sub$Overall)
  overall_stats[areas, "Area"] = areas
}

desired_order <- c("South East", "South West", "London", "East of England", "East
Midlands", "West Midlands", "Yorkshire and The Humber", "North West", "North East")
overall$Area <- factor(overall$Area, levels = desired_order)
overall_stats$Area <- factor(overall_stats$Area, levels = desired_order)
#visualise the descriptive data
```

```r
library(ggplot2)

# Create the boxplot
ggplot(data = overall, aes(x = Overall, y = Area)) +
  geom_boxplot() +
  theme_minimal() +
  labs(title = "Boxplot of Categories",
       x = "Overall Value",
       y = "Area")

# Create the mean and standard deviation
ggplot(overall_stats, aes(x = mean, y = Area))+
  geom_point(data = overall_stats, aes(x = mean, y = Area),
             color = "red", size = 3) + # Add mean points
  geom_errorbarh(data = overall_stats,
                 aes(xmin = mean - sd, xmax = mean + sd, y = Area),
                 height = 0.2, color = "blue") + # Add SD error bars
  theme_minimal() +
  labs(title = "Mean and SD of Categories",
       x = "Overall Value",
       y = "Area")

North <- rbind(North_West, North_East, Yorkshire_and_The_Humber)
MidAndSouth <-rbind(London, East_Midlands, East_Of_Englands, South_East,
South_West, West_Midlands)

#########################
#investigating the model#
#########################
#Initial model using the R's default linear modelling function
modeldraft = lm(Overall~., data = overall)
summary(modeldraft)
plot(overall)
#it has shown that we are able to reject the null hypothesis that the coefficients
for the numeric variables are not zero, at 99.9% confidence
#the adjusted R-squared also reached 0.8617, showing that about the model is 86.17%
useful for predicting the overall score
#However in some areas(Mostly in the North of England and South West), we cannot
reject the null hypothesis that the coefficient is equal to 0 at the same
confidence level

#testing the outlier exclusion
num_UK_clean <- num_UK
belong_outliers <- z_score_outliers[["Belong"]]
num_UK_clean <- num_UK_clean[!(num_UK_clean$Belong %in% belong_outliers),]
print(num_UK_clean)
```

```
other_vars <- setdiff(names(outliers_list),"Belong")
for (var in other_vars){
  outliers <- outliers_list[[var]]
  num_UK_clean <- num_UK_clean[!(num_UK_clean[[var]] %in% outliers),]
}
print(num_UK_clean)

#Now model the data with and without outliers
model_with_outliers <- lm(Overall~ Belong+Influence_Decisions+ Get_On_Well +
Drug_Use_And_Selling,data = num_UK)
summary(model_with_outliers)
#R-squared: 0.84
model_without_outliers <- lm(Overall~ Belong +Influence_Decisions+ Get_On_Well +
Drug_Use_And_Selling,data = num_UK_clean)
summary(model_without_outliers)
#R-squared: 0.82
#decide to include the outliers

#Initial northern model using the R's default linear modelling function#
northdraft = lm(Overall~., data = North)
summary(northdraft)
#Adjusted R-squared: 0.8913
#high enough though, but we have two coefficient with low significance
#choosing to use north model for further investigation

#selecting variables
library(ElemStatLearn)
library(leaps)

#find the best subset
bss<-regsubsets(Overall~.,data=North,method="exhaustive",nvmax=overallcols)
summary(bss)

#combine the North west with north east
North_and_yorkshire <- North
North_and_yorkshire$Area <- ifelse(North_and_yorkshire$Area == "North East", "Not
Yorkshire", North_and_yorkshire$Area)
North_and_yorkshire$Area <- ifelse(North_and_yorkshire$Area == "North West", "Not
Yorkshire", North_and_yorkshire$Area)

#test the best subset variables
bss<-regsubsets(Overall~.,data=North_and_yorkshire,method="exhaustive",nvmax=overal
lcols)
summary(bss)
#suggesting that the
par(mfrow = c(1, 3))
adjr <- summary(bss)$adjr2
```

```r
cp <- summary(bss)$cp
bic <- summary(bss)$bic

plot(adjr)
plot(cp)
plot(bic)

NAY1 <- lm(Overall~., data= North_and_yorkshire)
summary(NAY1)
# adjusted R-squared: 0.8886
NAY2 <- lm(Overall~.-Influence_Decisions, data= North_and_yorkshire)
summary(NAY2)
# adjusted R-squared: 0.885
#slightly different, choose not to exclude the Influence_decision

#test the combination of the variables
NAY3 <-
lm(Overall~Influence_Decisions+Get_On_Well+Belong+Drug_Use_And_Selling+Area*(Get_On
_Well)+Area, data= North_and_yorkshire)
summary(NAY3)
#adjusted R-squared: 0.8982, therefore the area*getonwell is included, each
coefficient has a P-value less than 0.05
par(mfrow = c(1, 1))
plot(NAY3)

#standardise each variable into a same scale
#min to max linear scale
scale_to_01 <- function(x) {
  (x - min(x, na.rm = TRUE)) / (max(x, na.rm = TRUE) - min(x, na.rm = TRUE))
}
NAY_01 <- North_and_yorkshire
NAY_01$Overall <- scale_to_01(NAY_01$Overall)
NAY_01$Influence_Decisions <- scale_to_01(NAY_01$Influence_Decisions)
NAY_01$Get_On_Well <- scale_to_01(NAY_01$Get_On_Well)
NAY_01$Belong <- scale_to_01(NAY_01$Belong)
NAY_01$Drug_Use_And_Selling <- scale_to_01(NAY_01$Drug_Use_And_Selling)

NAY01 <-
lm(Overall~Influence_Decisions+Get_On_Well+Belong+Drug_Use_And_Selling+Area*(Get_On
_Well)+Area, data= NAY_01)
summary(NAY01)
#R-squared: 0.8982, not a big change compared to previous NAY3

#Z-score
scale_to_z <- function(x){
  (x-mean(x, na.rm = TRUE))/sd(x,na.rm = TRUE)
}
```

```
NAY_Z <- North_and_yorkshire
NAY_Z$Overall <- scale_to_z(NAY_Z$Overall)
NAY_Z$Influence_Decisions <- scale_to_z(NAY_Z$Influence_Decisions)
NAY_Z$Get_On_Well <- scale_to_z(NAY_Z$Get_On_Well)
NAY_Z$Belong <- scale_to_z(NAY_Z$Belong)
NAY_Z$Drug_Use_And_Selling <- scale_to_z(NAY_Z$Drug_Use_And_Selling)
NAYZ <-
lm(Overall~Influence_Decisions+Get_On_Well+Belong+Drug_Use_And_Selling+Area*(Get_On
_Well)+Area, data= NAY_Z)
summary(NAYZ)
#R-squared: 0.8982, not a big change compared to previous NAY3

#only belong is normally distributed
#set belong to z-score, others to linear scale
NAY_0Z<- North_and_yorkshire
NAY_0Z$Overall <- scale_to_01(NAY_0Z$Overall)
NAY_0Z$Influence_Decisions <- scale_to_01(NAY_0Z$Influence_Decisions)
NAY_0Z$Get_On_Well <- scale_to_01(NAY_0Z$Get_On_Well)
NAY_0Z$Belong <- scale_to_z(NAY_0Z$Belong)
NAY_0Z$Drug_Use_And_Selling <- scale_to_01(NAY_0Z$Drug_Use_And_Selling)
NAY0Z <-
lm(Overall~Influence_Decisions+Get_On_Well+Belong+Drug_Use_And_Selling+Area*(Get_On
_Well)+Area, data= NAY_0Z)
summary(NAY0Z)
#it shows that different scale will not affect the model performance
#we will use NAY3 as the finalmodel as it has the lowest residual standard
deviation
NAYFinal <- NAY3

#################
#assumption test#
#################
#assumptions for residual expectation
sum(NAYFinal$residuals)
shapiro.test(NAYFinal$residuals)
#highly convinced that the residuals are normally distributed
plot(NAYFinal)

#assumption for multicolinearity
vif_values <- vif(NAYFinal, type="predictor")
print(vif_values)
#the VIF for every variable is less than 5 and thus we can conclude that the model
meets the assumption for non-multicollinearity
#the adjusted R-squared for the north model is 0.8913

#asumption for the auto correlation
cor(NAYFinal$residuals[1:74], NAYFinal$residuals[2:75])
```