<u>Patterns of strategic absence / presence amongst parliamentarians</u>

A - Clustering

The data shows the absence and presence patterns for each of the 166 TD's in 6 salient votes, using a '1' to denote absence and a '2' to denote presence. As the data is binary and ordinal in nature, hierarchical clustering is likely to be more effective than k-means clustering.

Exploratory data analysis showed strong correlations between some of the six votes. For example, there is a correlation of 0.95 between votes ED1 and ED2, suggesting that TD's who took part in one of these votes also participated in the other. There are low correlations (~ 0.1 – 0.2) between the ED votes and the other votes, suggesting that TD's present for the former ED votes were absent for other votes. Similarly, there is a high correlation (0.95) of TD presence for the two confidence votes. The pattern of absence for the remaining four votes is again replicated. This suggests at least two initial clusters of TD's – those who are present for ED votes, but absent for most others; and those who are present for confidence votes, but absent for most others.

Binary distance was selected for *initial* clustering work, and this was implemented with single, average and complete linkages. This was not very successful as it tended to lead to 'chaining' – namely the (imbalanced) assignment of most observations to a single cluster, with small numbers of observations being sent to the remaining clusters. In many cases, with k=3, 160+ observations were in cluster 1, with 1-3 observations in each of clusters 2 and 3. Later work used euclidean distance instead for the dissimilarity matrix. When implemented with single and average linkage it returned similarly modest results. However, the use of complete linkage and setting of k=3, returned three clusters of 113, 34 and 19.

The issue of the optimum value of k (ie. the most appropriate number of clusters) remained. The NbClust package in R was used to determine this. This uses two measures (Hubert index and the D index) and takes a majority vote from amongst several iterations (**figure 3**). Based on its' conclusions**, a value of k=5 with euclidean distance and complete linkage was eventually chosen**.

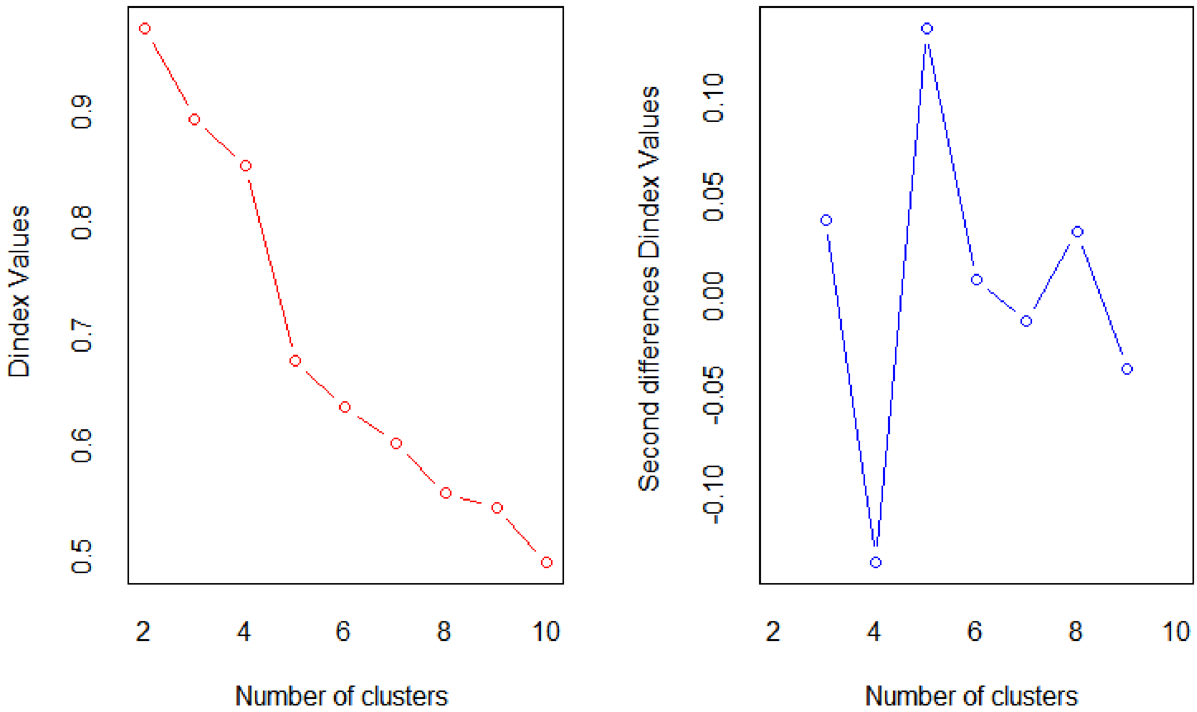

Figure 3 – Selection of value of k

The absence / presence patterns of the 166 TD's were therefore assigned to 5 clusters of size 72, 26, 19, 41 and 8 respectively. Cluster membership is discussed futher in the context of question 2C and 2D.

Finally, both a cluster plot and a silhouette plot of the results were set out (**figure 4**). The former shows a degree of overlap between the larger clusters. What is noticable from the latter is the low silhouette widths for some of the clusters, combined with a number of observations showing negative widths (indicating that they might be better assinged to a different cluster).

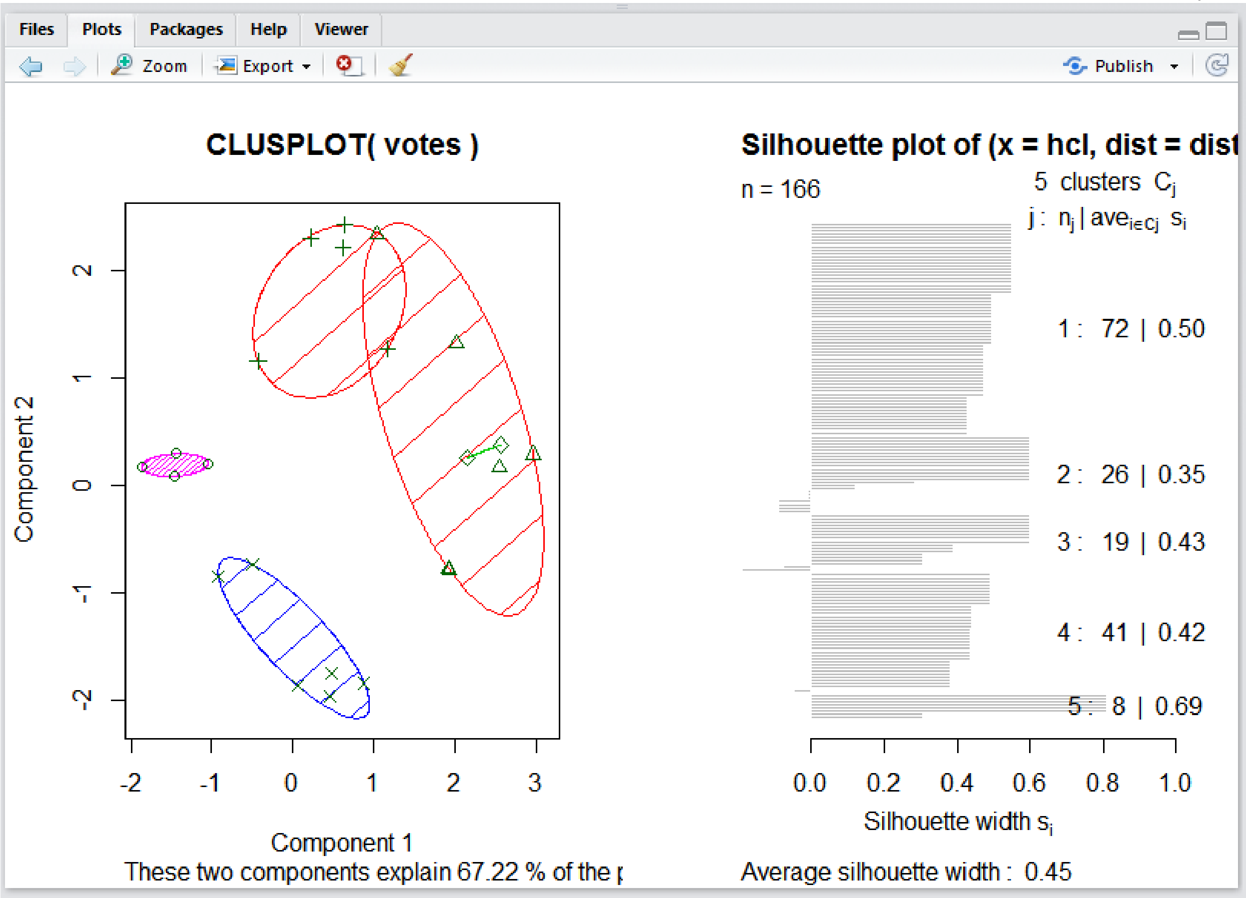

Figure 4 – Cluster and silhouette plots of the 5 clusters.

B - Latent Class Analysis

The poLCA function was then applied to the same dataset, and the background to latent class analysis is discussed further for question 2D.

To begin with, each of the six votes were brought together into a single formula, for processing by the function. In this approach, the number of classes required is selected by the user. Initially, five versions of the poLCA function were set up – with the number of classes (clusters) running from one to five. Each of these returned both an AIC (Akaike information criterion) and a BIC (Bayesian information criterion). Both of these fall sharply between models m0 (1 class) and m1 (2 classes) and, again, between models m1 and m2 (3classes). They both increase when we move to m3 (4 classes) , with AIC falling when we move to m4 (5 classes), whilst BIC increases in this scenario. This suggested that m2 (the three class model) was the most parsimonious.

# Classes	1	2	3	4	5
AIC	1333	1178	1028	1038	993
BIC	1351	1218	1091	1122	1099

Figure 5 – AIC and BIC for the five per number of latent classes

The presence and absence patterns of the TD's are therefore set out in terms of three latent varibles as in **figure 6**. The red bars represent the conditional probabilities, by latent class, of being present for the votes in question. Taller bars correspond to conditional probabilities closer to one. The values along the bottom represent the proportion of the TD's belonging to each latent variable. Reading from left to right, we can see that the first latent class is made up of TD's who were strongly present for the two confidence votes, who were present to a smaller extent for the trade and credit votes, and who were largely absent for the emergency department votes. The second class consists of TD's who were strongly present for all votes, whilst the third shows TD's who were effectively absent for the two confidence votes and who registered modest levels of presence for each of the other four votes. The number of TD's in each class (based on the predicted latent variable membership) came to 41,75 and 50 respectively.

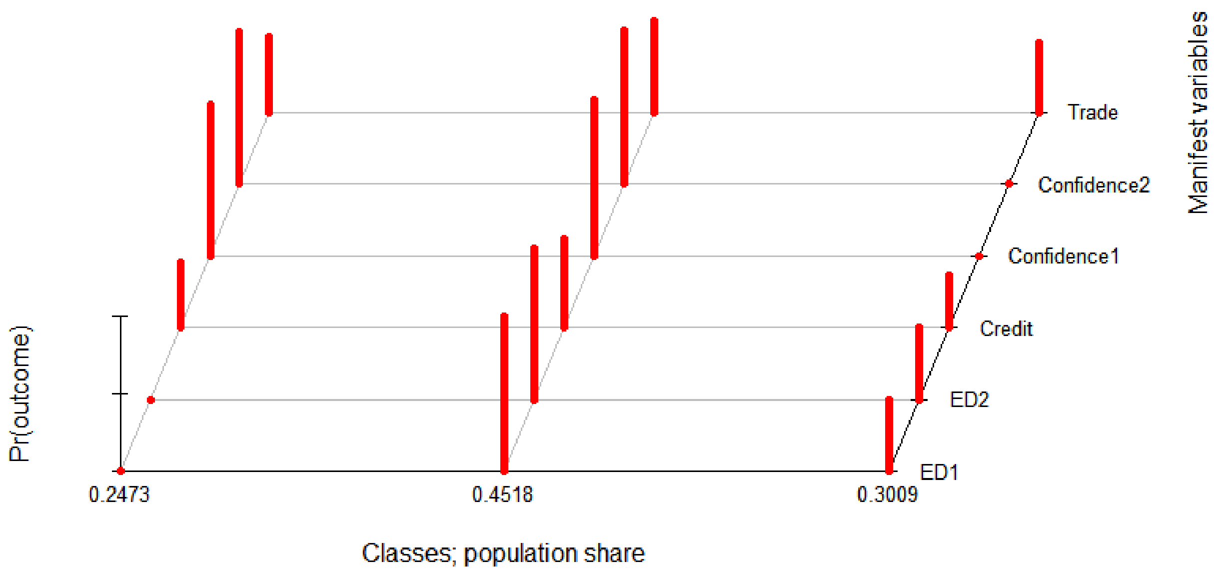

Figure 6 – Latent class analysis for the TD dataset

C - Comparison of results

Therefore, the original clustering solution gave us five clusters of TD's, whilst the LCA approach gave us three classes. The Table in figure 7 shows the extent to which the two approaches concurred and varied. The first latent class (present mainly for confidence votes) aligns closely with cluster 4 in the earlier approach with 39/41 observations matching up. The second latent class (strong presence for all votes) matches primarily with cluster 1 of the earlier approach (72/75 observations match). The third class (significant presence for ed, trade and credit votes – low presence for confidence votes) matches strongly with no fewer than three of the clusters from before – they are effectively integrated into a single class.

```
> # Q2C - Results Comparison
> table(hcl, m2$predclass)

hcl  1  2  3
  1  0 72  0
  2  2  0 24
  3  0  1 18
  4 39  2  0
  5  0  0  8
>
```

Figure 7 – Crosstab of the two approaches

D – Latent Class Analysis

Introduction

The objective is to assess the extent to which TD presence and absence reflects a wider strategic motivation. To this end, the LCA approach is briefly reviewed. The initial model from question 2B will be slightly modified, and then a number of analyses, in the wider context of party affiliation, will be conducted.

Latent Class Analysis

A latent class model uses the response patterns in multivariate data to propose similar, though presently unobserved groups. The aim is to identify sub-groups of individuals based on the latent (unobserved) characteristics.

Initial poLCA implementation and refinement

When running LCA, it is possible for the computer to emphasise a local maximum value, rather than a global one. To avoid this, we can run the function several times over, to ensure that there is agreement on the final outcome. The original m2 model was therefore run for 200 iterations, as in **figure 8**. This left us with the same three class model as before, with a majority group heavily present for most votes (45% of the TD's) along with two smaller groups, one with low levels of presence at confidence votes and modest levels of presence at other votes (30% of TD's), and a final group with high levels of presence at confidence votes, modest presence at trade / credit votes, and almost total absence at emergency department votes (~25% of TD's).

```
# refinement - nrep to avoid local maxima
m2 <- poLCA(formula1,votes,nclass=3,maxiter=8000, nrep=200,graphs=TRUE)
```

Figure 8 – Elimination of local maxima

Party Membership

Party membership data was then merged with the original presence and class data, as in **figure 9**. A table showing counts of party membership (**figure 10**) was then set up. For later analysis, it was decided to focus on the five largest groupings in terms of numbers of TD's (Fine Gael (44% of TD's), Labour (22%), Fianna Fail (12%), Independent (9%), Sinn Fein (8%)).

	ED1	ED2	Credit	Confidence1	Confidence2	Trade
O Caolain, Caoimhghin	2	2	2	2	2	2
O Cuiv, Eamon	1	1	2	1	1	1
O Fearghail, Sean	2	2	2	1	1	2
O Riordain Aodhan	1	1	1	2	2	2
O Snodaigh, Aengus	2	2	1	2	2	1
Adams, Gerry	2	2	1	2	2	1
Aylward, Bobby	2	2	2	2	1	1
Bannon, James	1	1	1	1	1	1
Barrett, Sean	1	1	1	1	1	1

Party	Name	M2
Sinn Fein	O Caolain, Caoimhghin	1
Fianna Fail	O Cuiv, Eamon	2
Fianna Fail	O Fearghail, Sean	2
Labour	O Riordain Aodhan	3
Sinn Fein	O Snodaigh, Aengus	1
Sinn Fein	Adams, Gerry	1
Fianna Fail	Aylward, Bobby	1
Fine Gael	Bannon, James	2
Ceann Comhairle	Barrett, Sean	2

Figure 9 – Latent class and party affiliation

```
> count(x, 'Party') # number of tds in each party
                      Party freq
1       Anti-Austerity Alliance    1
2             Ceann Comhairle    1
3                Fianna Fail   20
4                  Fine Gael   73
5                Independent   15
6                     Labour   37
7 People Before Profit Alliance    2
8                  Sinn Fein   14
9             Socialist Party    3
>
```

Figure 10 – Party membership counts

Analysis 1 – Class membership -> Political party

The majority class (class 2 in figure 6) consists of 75 TD's (45% of the total). The class is characterised by heavy presence for all votes, running at ~98% for **both** emergency department votes and confidence votes, and around ~55% for trade and credit votes. It is noticable that Fianna Fail are under-represented in this group (only 4 of their 20 TD's are listed), whilst Sinn Fein is strongly over-represented (with 11 of their 14 TD's) being assigned to this group.

The next biggest class (class 3 in figure 6) consists of 50 TD's (or 30% of the total). The class is marked by the almost complete absence of TD's at confidence votes, combined with modest presence for all other votes (~ 30 – 50% depending on the vote). In terms of party affiliation, this class is almost the converse of the previous one, with Fianna Fail being strongly over-represented (13 of their 20 TD's are here) and Sinn Fein being strongly under-represented (2 of their TD's are here – less than half of what one might expect).

The smallest class (class 1 in figure 6) has 41 TD's (or 24% of the total). This class is characterised by strong presence (97% +) for confidence votes, modest presence for trade and credit votes, and 0% presence for emergency department votes. Labour are over-represented in this group (35% of their TD's), whilst Sinn Fein is strongly under-represented (1 of their 14 TD's is here) and Fianna Fail is significantly under-represented (15% of TD's against an expected 25%).

Analysis 2 – Political part -> Class membership

Based on this, a picture for the behaviour of the each of the parties emerges. The pattern for **Fine Gael** TD's is closest to that of the expected latent classes. The majority (49%) show strong presence for each vote, nonetheless almost a quarter absent themselves from confidence votes, whilst a further quarter do not participate in emergency department votes. **Labour** TD's show similar adherence to the majority class as Fine Gael TD's – but they are more likely absent themselves from emergency department votes, whilst turning up for confidence votes**. Fianna Fail** TD's are under-represented in the majority class, with only four of their twenty members included. Almost two thirds of their TD's absent themselves from confidence votes – more than double the rate one might expect. **Independents** broadly follow the overall latent class distribution, though they have a slightly higher than expected tendancy to absent themselves from confidence votes. Finally, **Sinn Fein** TD's belong predominantly to the majority class (80%), showing high presence for all votes, particularly for confidence votes.

Analysis 3 - Smaller parties

The 3 TD's of the **Socialist Party** are split between the three classes (one each), with Clare Daly being absent for confidence votes and Ruth Coppinger being absent for emergency department votes. The **Anti Austerity Alliance** TD (Paul Murphy) is in the class that is absent for emergency department votes but present for confidence votes. The two members of the **People Before Profit Alliance** are similarly split with Joan Collins being absent for emergency department votes (though not confidence votes) and Richard Boyd Barrett represented in the majority class. Finally, the **Ceann Comhairle** is a special case, being absent for all votes – impartial, and strategic in a different sense.

Conclusion

Patterns of strategic presence and absence at TD's votes can be summarised as follows:

1 – The majority of TD's (45%) are almost uniformly present for all votes. This is a significant characteristic of Fine Gael, Labour and especially Sinn Fein TD's.

2 – A large group (30%) of TD's strategically avoid confidence votes. This group has significant numbers from all the main parties, but contains 65% of all Fianna Fail TD's at one extreme and 14% of Sinn Fein TD's at the other.

3 – A final significant group is strategically absent for votes on emergency departments, whilst being present for those relating to confidence. Again, this has a significant component of each of the main parties. However, it is most strongly represented by Labour TD's (35% of that party), whilst Fianna Fail and Sinn Fein are under-represented (with 15% and 7% of their respective TD's in this group).

Appendix – R Code for all questions

```
#Q1 - Hotellings t test

# Main Code

# EDA
summary(prices)
par(mfrow = c(3, 1))
boxplot(AskingPrice ~ Area, prices, main = "Asking Price", horizontal = T)
boxplot(NoBedrooms ~ Area, prices, main = "No of bedrooms",horizontal = T)
boxplot(SqFoot ~ Area, prices, main = "Square Footage",horizontal = T)

# Segment dataset
PA = subset(prices, prices$Area=="PA")
MP = subset(prices, prices$Area=="MP")
summary(PA)
summary(MP)

# Remove non numerical variables
PA$Area<-NULL
MP$Area<-NULL


# Main Function
hotelling = function(y1, y2) {

  # calculate sample size and observed means
  k = ncol(y1)
  n1 = nrow(y1)
  n2 = nrow(y2)
  ybar1= apply(y1, 2, mean); ybar2= apply(y2, 2, mean)
  bardiff = ybar1-ybar2

  # calculate the variance of the difference in means
  v = ((n1-1)*var(y1)+ (n2-1)*var(y2)) /(n1+n2-2)

  # calculate the test statistic and associated quantities
  t2 = n1*n2*bardiff%*%solve(v)%*%bardiff/(n1+n2)
  f = (n1+n2-k-1)*t2/((n1+n2-2)*k)
  pvalue = 1-pf(f, k, n1+n2-k-1)

  # return the list of results
  return(list(pvalue=pvalue, f=f, t2=t2, diff=bardiff))
}


# Run function
res1 = hotelling(PA, MP)
res1

# Compare with the inbuilt R function

library(Hotelling)

res2<-hotelling.test(.~Area, data = prices)
res2

###############################################################################

#Q2A - Absent / present vote hierarchical clustering

library(cluster)
library(factoextra)
library(NbClust)
library(HSAUR)
library(ltm)
options(max.print = 10000)

```

```
# EDA
mode(votes)
head(votes)
descript(votes)
cor(votes)
names(votes)

# Implement - Distance based on euclidean / binary
dist.eucl<-dist(votes, method = "euclidean")

# Then - hierarchical clustering of this based on hclust (average, single, complete)
cl.complete=hclust(dist.eucl, method="complete")
plot(cl.complete)

# How many clusters? - 5
clustercount<- NbClust(votes, distance = "euclidean", min.nc = 2,
                       max.nc = 10, method = "complete", index ="all")

# Implement - cut dendrogram into 5 clusters
hcl=cutree(cl.complete, k=5)
table(hcl)

# Visualize
clusplot(votes, hcl, color=TRUE, shade=TRUE, plotchar=TRUE, labels=0, lines=0)
plot(silhouette(hcl, dist.eucl))

# Cluster Membership
x<-cbind(votes,hcl)
x1<- subset(x, hcl==1)
x2<- subset(x, hcl==2)
x3<- subset(x, hcl==3)
x4<- subset(x, hcl==4)
x5<- subset(x, hcl==5)

# Tidying up
x["hcl"] <- hcl
x[order(x$hcl),]

####################################################################

#Q2B - poLCA applied to voting dataset

library(poLCA)

# Need to create the formula to pass to the poLCA function
formula1 <- cbind(ED1,ED2,Credit,Confidence1,Confidence2,Trade) ~ 1

# m0: Loglinear independence model
# m1: Two-class latent class model
# m2: Three-class latent class model
# m3: Four-class latent class model
# m4: Five-Class latent class model

# m2 is our preference in all of these
m0 <- poLCA(formula1,votes,nclass=1) # cannot produce a graph for single latent var.  log-likelihood:
m1 <- poLCA(formula1,votes,nclass=2, graphs=TRUE) # log-likelihood:
m2 <- poLCA(formula1,votes,nclass=3,maxiter=8000, graphs=TRUE) # log-likelihood:
m3 <- poLCA(formula1,votes,nclass=4,maxiter=8000, graphs=TRUE) # log-likelihood:
m4 <- poLCA(formula1,votes,nclass=5,maxiter=8000, graphs=TRUE) # log-likelihood:

# observed and estimated frequencies for each of the response patterns
m0$predcell
m1$predcell
m2$predcell

# Predicted latent variable membership for m2
m2$predclass
table(m2$predclass)
```

```
x<-cbind(votes,m2$predclass)
x1<- subset(x, m2$predclass==1)
x2<- subset(x, m2$predclass==2)
x3<- subset(x, m2$predclass==3)

x["hcl"] <- hcl
x["M2"]<- m2$predclass
x[order(x$hcl),]

#####################################################################

# Q2C - Results Comparison
table(hcl, m2$predclass)

######################################################################
#Q2D - Strategic absence / presence

library(plyr)

# refinement - nrep to avoid local maxima
m2 <- poLCA(formula1,votes,nclass=3,maxiter=8000, nrep=200,graphs=TRUE)

# Preliminary - Integrate Party Membership Data / Tidy up

x<-votes
members<-members.party
x["Party"]<-members$Party
x["Name"]<- members$TD

count(x, 'Party') # number of tds in each party

# Class membership by party
x1<- subset(x, m2$predclass==1)
x1 <- x1[order(x1$Party), ]
x2<- subset(x, m2$predclass==2)
x2 <- x2[order(x2$Party), ]
x3<- subset(x, m2$predclass==3)
x3 <- x3[order(x3$Party), ]

# Party membership by class - top 5 parties

xfg<-subset(x, x$Party=="Fine Gael")
xfg <- xfg[order(xfg$M2), ]

xla<-subset(x, x$Party=="Labour")
xla <- xla[order(xla$M2), ]

xff<-subset(x, x$Party=="Fianna Fail")
xff <- xff[order(xff$M2), ]

xin<-subset(x, x$Party=="Independent")
xin <- xin[order(xin$M2), ]

xsf<-subset(x, x$Party=="Sinn Fein")
xsf <- xsf[order(xsf$M2), ]
```