# Analysis of the 1992 US Presidential Elections

## Dataset Selection

We have selected 1992 US presidential vote by country. The link to the data source is:
https://www.statcrunch.com/app/index.php?dataid=493100

## Overview of Dataset

1992 US presidential vote county by county dataset has 3141 observations with 19 variables. The following is a list and brief description of the variables:

1. County - County
2. State - State
3. MSA - Metropolitan Statistical Area
4. PMSA - Primary Metropolitan Statistical Area
5. Pop.density - 1992 population per 1990 miles^2
6. Pop - 1990 population
7. Pop.change - Percent population change 1980-1992
8. Age 65-74 - Percent population age 65-74, 1990
9. Age75 - Percent population age >= 75, 1990
10. Crime - Serious crimes per 100,000, 1991
11. College - Percent with bachelor's degree or higher of those age>=25
12. Income - Median family income, 1989 dollars
13. Farm - Farm population, % of the total, 1990
14. Democrat - Percent votes cast for Democratic president
15. Republican - Percent votes cast for Republican president
16. Perot - Percent votes cast for Ross Perot
17. White - Percent white, 1990
18. Black - Percent black, 1990
19. Turnout - 1992 votes for president / 1990 pop x 100

## Cleansing and Preparing the dataset

- In order to prepare and cleanse the dataset, we have deleted unnecessary data that we are not going to use during our analysis such as “msa”, “pmsa” (we are using county-level data, not metropolitan statistical area nor primary metropolitan statistical area) and ”pop.change” ( we did not feel we needed to include this variable as we have 13 variables that are more important).
- We combined the ‘Age 65-74’ and ‘Age75’ into a single column to have data for percentage of the population whose age is above 65.
- We have transformed the following columns from percent to a headcount for ease of interpretation when we build our model
    - Population above the age 65 (column name: AgeAbove65)
    - College graduates with bachelor’s degree or higher of those age =>25 (column name: College)

    - Farm population (column name: Farm)
- We have inspected the data and noticed missing data from all Alaska counties that voted Democrat, Republican, and Perot that we will exclude Alaska as we have enough data points from the other State/Counties. We shall exclude missing data while importing the data into R by using na.omit().
- We will transform State into individual categorical variables in R using factors.
- We have removed the rows when there is a tie between two parties. Ideally, we would have to go for the particular electoral college to see what decision was made to be considered as a tiebreaker. We have considered this out of our scope for now.

## Data Analysis Plan

**Objective: The main objective of our analysis is to identify the variable(s) most influential in determining the Democrat win in the 1992 presidential election.**

**Tools:** R- Studio, MS Excel.

**Variable Selection:** In order to find the answer to the question "What variables most influenced the Democrat to win in 1992?", we will use following explanatory variables in our model:

1. State - Each of the 49 States will be transformed into a categorical variable i.e. Arkansas (AR), Alabama (AL).
   Note: The State of Alaska will be eliminated due to missing data.
1. Population - 1990 population count
2. Population Density - 1992 population count per 1990 miles^2
3. AgeAbove65 - Population Age 65 years and over
4. Crime - Count of serious crimes per 100,000 in 1991
5. College - Headcount with bachelor's degree or higher of those age>=25
6. Income - Median family income, 1989 dollars
7. Farm Count - Headcount of Farm Population
8. White - Headcount of White people in 1990
9. Black - Headcount of Black people in 1990
10. Turnout - Percent of Voter turnout (based on 1990 population)

The dependent variables in our model are as below:

Logit: This column has either zero or one where zero represents democrats not winning the majority and one represent democrats winning the majority.

**Statistical Method:** In our model, we have three dependent variables. We will code the three level of dependent variables (Democrat, Republican, and Perot) in Excel and will use multinomial logistic regression during our analysis. The four levels of coding will use the nested if formula in excel.

=IF(AND(Democrats>Republicans, Democrats>Perot),1, 0)

1: If Democrats win

0: If Democrats lose

**Steps for Analysis:**

1. Create a factor variable for the column “logit”.
2. Estimate the regression with logit as the y variable and State, Population, AgeAbove65, Crime, College, Income, Farm Count, White, Black, Turnout as our explanatory variables.
3. Check for multicollinearity using vif() function.
4. Transform the data if necessary.
5. Using glm() function predict the logistic regression model
6. Create residual plots.
7. Identify which variables are significant.
8. Conduct partial f-test.
9. List the most likely parsimonious model.

Multinomial Logistic Regression is the linear regression analysis to conduct when the dependent variable is nominal with more than two levels. This regression process uses maximum likelihood estimation to evaluate the probability of categorical membership.