**Data Set Table of Contents**

In this document, you will be briefly describing each data set you used / created in this project. These data sets should be in .csv or Excel forms so your clients can use them. You need to include (1) the name of each data set , (2) what each data set contains, and (3) the name of the R script in the R script folder that was used to create the data set, if applicable.

Example:

1. Explanatory1.xlsx: file contains information on the variables acquittal rates, dismissal rates, prosecutorial district, judicial district, volume of cases, and defense counsel types. This dataset was created using excel outside R.
2. Combine3.csv: file contains information of all explanatory variables and y2 using merge function in R.
3. Combine2.csv: file contains information of all explanatory variables and y1 using merge function in R.

Under the Folder: Goal2LinearModels\_Nico

1. Combine\_1\_Updated.xlsx: The file containing all the information of all explanatory variables plus the added political views data that match with each of the data rows.
2. North Carolina Election Data.xlsx: This file includes data on the percent that voted republican in each of the 100 North Carolina for the 2008, 2012, 2016, and 2020 Presidential elections as well as other information such as the name of the Republican candidate and the number of votes for that candidate, as well as the total number of votes in the county.
3. PoliticalViewsNCData2008.xslx: This file includes data on the percent that voted republican in each of the 100 North Carolina for the 2008 Presidential election as well as other information such as the name of the Republican candidate and the number of votes for that candidate, as well as the total number of votes in the county.
4. PoliticalViewsNCData2012.xslx: This file includes data on the percent that voted republican in each of the 100 North Carolina for the 2012 Presidential election as well as other information such as the name of the Republican candidate and the number of votes for that candidate, as well as the total number of votes in the county.
5. PoliticalViewsNCData2016.xslx: This file includes data on the percent that voted republican in each of the 100 North Carolina for the 2016 Presidential election as well as other information such as the name of the Republican candidate and the number of votes for that candidate, as well as the total number of votes in the county.
6. PoliticalViewsNCData2020.xslx: This file includes data on the percent that voted republican in each of the 100 North Carolina for the 2020 Presidential election as well as other information such as the name of the Republican candidate and the number of votes for that candidate, as well as the total number of votes in the county.

Under the folder Goal1VariableCreation, there are two folders.

1. sourceFiles Folder: It has North Carolina felony case activity reports from 2010 to 2022, without the year 2018-2019, as our project decided not to study this year. They are named by the year. They are used in the R script: Goal1VariableCreation.rmd.
2. resultFiles Folder: It has four files. They are created in the Goal1VariableCreation.rmd.
   1. nostate.csv–the combined dataset with all variables selected and explanatory variables calculated based on clients' needs with all years combined. It is the source file of y1.csv and y2.csv.
   2. nostateNoNA.csv–the combined dataset with no missing value that contains the same information as nostate.csv.
   3. y1.csv–the response variable y1 calculated based on the clients' needs. The missing value is not removed. In Yifan's folder for the heatmap, the holder.csv is created based on the y1.csv with the missing value removed.
   4. y2.csv–the response variable y2 calculated based on the clients' needs. The missing value is not removed. In Yifan's folder for the heatmap, the holder2.csv is created based on the y2.csv with the missing value removed.

Under the folder Goal2VariableOrganization, there are three folders.

1. CrimeRatesSourceFiles: There are two subfolders inside this folder
   1. OriginalFiles: Those files are downloaded from the NC State Bureau of Investigation website based on year. They are transformed into documents with Adobe for data extraction. We use Table 6, "Crime Rates Per 100,000 by County."
   2. xlsxFormatted: Those files are extracted from the documents mentioned above. They include information on two explanatory variables "Violent Crime Rate" and the "Property Crime Rate." Each xlsx file is named based on the year. They first are organized and formatted in excel, and more work is done in Goal2VariablesOrganization.Rmd.
2. EthinicMakeupSourceFiles:
   1. 2010Cencus.xlsx: This file is downloaded from [www.census.gov](http://www.census.gov/) and organized in Excel. For each county in 2010, the total county population, the percentage of that total that is Black, and the percentage of the total population that is Hispanic ethnicity are inside. More work is done based on this file in Goal2VariablesOrganization.Rmd.
   2. 2020Cencus.xlsx: This file is downloaded from [www.census.gov](http://www.census.gov/) and organized in Excel. For each county in 2020, the total county population, the percentage of that total that is Black, and the percentage of the total population that is Hispanic ethnicity are inside. More work is done based on this file in Goal2VariablesOrganization.Rmd.
3. FinalDatasetsForModeling:
   1. Combine2.csv: This file contains all the explanatory variables and response variable y1 for our models. It is created in Goal2VariablesOrganization.Rmd.
   2. Combine3.csv: This file contains all the explanatory variables and response variable y2 for our models. It is created in Goal2VariablesOrganization.Rmd.