# Kumar - SDS293 Final Project 0

## Part I: Loading and Cleaning the Data

Load libraries

```
library(tidyverse)
library(fs)
```

load data on various bills in and out of the state legislatures

```
## locate where all of the "bills.csv"s are for all the different years
bills_dir = "/Users/sirohikumar/Library/CloudStorage/GoogleDrive-sirohikumar185@gmail.com/My Drive/UNDE

## list these csv files
bills_files = fs::dir_ls(bills_dir)

## load in all of the bills csvs, adding an id column with the years and creating a column for legislat
bills_csv = bills_files %>%
    map_dfr(read_csv, .id = "source") %>%
    mutate(year = substr(source, 185, 193),
           legislature = case_when(year == "2009-2010" ~ "124",
                            year == "2011-2012" ~ "125",
                            year == "2013-2014" ~ "126",
                            year == "2015-2016" ~ "127",
                            year == "2017-2018" ~ "128",
                            year == "2019-2020" ~ "129",
                            year == "2021-2022" ~ "130",
                            year == "2023-2024" ~ "131",
                            year == "2025-2026" ~ "132",)) %>%
    select(-source) %>%
    mutate(bill_id = as.character(bill_id),
           status = as.character(status),
           committee_id = as.character(committee_id),
           session_id = as.character(session_id))           ## convert ids to chrs
```

```
## Warning: One or more parsing issues, call ‘problems()‘ on your data frame for details,
## e.g.:
##   dat <- vroom(...)
##   problems(dat)
## One or more parsing issues, call ‘problems()‘ on your data frame for details,
## e.g.:
##   dat <- vroom(...)
##   problems(dat)
## One or more parsing issues, call ‘problems()‘ on your data frame for details,
## e.g.:
##   dat <- vroom(...)
##   problems(dat)
```

```
head(bills_csv)
```

```
## # A tibble: 6 x 16
##   bill_id session_id bill_number status status_desc status_date title
##   <chr>   <chr>      <chr>       <chr>  <chr>       <date>      <chr>
## 1 191384  48         LD9         6      Failed      2009-03-25  An Act To Prohi~
## 2 191385  48         LD10        6      Failed      2009-06-08  An Act To Autho~
## 3 191386  48         LD11        4      Passed      2009-04-16  Resolve, To Enc~
## 4 191387  48         LD12        4      Passed      2009-04-14  Resolve, Direct~
## 5 191388  48         LD13        6      Failed      2009-03-17  An Act To Expan~
## 6 191389  48         LD14        6      Failed      2009-03-03  An Act To Prohi~
## # i 9 more variables: description <chr>, committee_id <chr>, committee <chr>,
## #   last_action_date <date>, last_action <chr>, url <chr>, state_link <chr>,
## #   year <chr>, legislature <chr>
```

load data on the actions taken on each bill

```
## locate where all of the "history.csv"s are for all the different years
history_dir = "/Users/sirohikumar/Library/CloudStorage/GoogleDrive-sirohikumar185@gmail.com/My Drive/UN

## list these csv files
history_files = fs::dir_ls(history_dir)

## load in all of the history csvs, adding an id column with the years and creating a column for legisl
history_csv = history_files %>%
    map_dfr(read_csv, .id = "source") %>%
    mutate(year = substr(source, 187, 195),
           legislature = case_when(year == "2009-2010" ~ "124",
                                   year == "2011-2012" ~ "125",
                                   year == "2013-2014" ~ "126",
                                   year == "2015-2016" ~ "127",
                                   year == "2017-2018" ~ "128",
                                   year == "2019-2020" ~ "129",
                                   year == "2021-2022" ~ "130",
                                   year == "2023-2024" ~ "131",
                                   year == "2025-2026" ~ "132",)) %>%
    select(-source) %>%
    mutate(bill_id = as.character(bill_id),
           sequence = as.character(sequence))             ## convert several ids to characters

head(history_csv)
```

```
## # A tibble: 6 x 7
##   bill_id date       chamber sequence action                      year  legislature
##   <chr>   <date>     <chr>   <chr>    <chr>                       <chr> <chr>
## 1 191283  2009-01-13 <NA>    1        (H) READ and ADOPTED.       2009~ 124
## 2 191283  2009-01-13 <NA>    2        (H) Sent for concurrenc~ 2009~ 124
## 3 191283  2009-01-13 <NA>    3        (S) Under suspension of~ 2009~ 124
## 4 191284  2009-02-05 <NA>    1        (H) READ and ADOPTED.       2009~ 124
## 5 191284  2009-02-05 <NA>    2        (H) Sent for concurrenc~ 2009~ 124
## 6 191284  2009-02-10 <NA>    3        (S) READ and ADOPTED, i~ 2009~ 124
```

load data on the members of each legislature

```
## locate where all of the "people.csv"s are for all the different years
people_dir = "/Users/sirohikumar/Library/CloudStorage/GoogleDrive-sirohikumar185@gmail.com/My Drive/UND

## list these csv files
people_files = fs::dir_ls(people_dir)

## load in all of the people csvs, adding an id column with the years and creating a column for legisla
people_csv = people_files %>%
    map_dfr(read_csv, .id = "source") %>%
    mutate(year = substr(source, 186, 194),
           legislature = case_when(year == "2009-2010" ~ "124",
                        year == "2011-2012" ~ "125",
                        year == "2013-2014" ~ "126",
                        year == "2015-2016" ~ "127",
                        year == "2017-2018" ~ "128",
                        year == "2019-2020" ~ "129",
                        year == "2021-2022" ~ "130",
                        year == "2023-2024" ~ "131",
                        year == "2025-2026" ~ "132",)) %>%
    select(-source) %>%
    mutate(people_id = as.character(people_id),
      party_id = as.character(party_id),
      role_id = as.character(role_id),
      committee_id = as.character(committee_id),
      knowwho_pid = as.character(knowwho_pid)) %>%      ## convert ids to chrs
    select(-followthemoney_eid, -votesmart_id,          ## remove unnecessary columns
           -opensecrets_id, -knowwho_pid,
           -ballotpedia, -committee_id)

head(people_csv)
```

```
## # A tibble: 6 x 14
##   people_id name       first_name middle_name last_name suffix nickname party_id
##   <chr>     <chr>      <chr>      <chr>       <chr>     <chr>  <chr>    <chr>
## 1 8737      Hannah Pi~ Hannah     M.          Pingree   <NA>   <NA>     1
## 2 8738      Elizabeth~ Elizabeth  H.          Mitchell  <NA>   <NA>     1
## 3 8739      Lisa Marr~ Lisa       Tessier     Marrache  <NA>   <NA>     1
## 4 8740      Kevin Raye Kevin      L.          Raye      <NA>   <NA>     2
## 5 8741      Joshua Ta~ Joshua     A.          Tardy     <NA>   <NA>     2
## 6 8742      Wendy Pieh Wendy      <NA>        Pieh      <NA>   <NA>     1
## # i 6 more variables: party <chr>, role_id <chr>, role <chr>, district <chr>,
## #   year <chr>, legislature <chr>
```

##Part II: Describe the Data

This data is on bills in Maine State Legislature. Details of all individual bills introduced are contained in `bills.csv` for each legislature, and the actions taken on each bill are in `history.csv` for each legislature. Further, the members of each state legislature are in `people.csv`.

For all three of these csvs, I've compiled long versions of each one across all legislatures. I've also converted a number of these columns which have ID numbers from numerical (which they were read in as) to categorical variables.

**bills_csv**

- `bill_id` - LegiScan bill identifier
  - categorical data, ranges from 191384 to 2022283
- `session_id` - LegiScan session identifier
  - categorical data, ranges from 48 to 2181
- `bill_number` - Bill number
  - categorical data, contains LD (legislative document), HP (), SP (senate document)
- `status` - Status value for bill
  - categorical, values from 0-6
- `status_desc` - Description of bill status value
  - categorical, written complement to `status`: (0 - NA, 1 - Introduced, 2 - Engrossed, 3 - Enrolled, 4 - Passed, 5 - Vetoed, 6 - Failed)
- `status_date` - Date of status change
  - numeric, dates ranging from January 13, 2009 to April 11, 2025, when this data was most recently pulled from the database
- `title` - Short title of bill
  - categorical, incredibly wide range of values
- `description` - Long title of bill
  - categorical, also wide range of values
- `committee_id` - LegiScan committee identifier
  - categorical, ranges from 0 to 4984
- `committee` - Pending committee name
  - categorical, also wide range of values
- `last_action_date` - Date of last action
  - numeric, dates ranging from January 13, 2009 to April 11, 2025, when this data was most recently pulled from the database
- `last_action` - Description of last action
  - categorical, also wide range of values
- `url` - LegiScan URL for bill detail
  - categorical, also wide range of values
- `state_link` - State URL for bill detail
  - categorical, also wide range of values
- `year` - Year range of the legislature
  - categorical, generated by me, ranges from 2009-2010 to 2025-2026
- `legislature` - Maine State Congress
  - categorical, generated by me, ranges from 124 to 132

### history_csv

- `bill_id` - LegiScan bill identifier
  - categorical data, ranges from 191384 to 2022283
- `date` - Date of history step
  - numeric, dates ranging from January 07, 2009 to April 11, 2025, when this data was most recently pulled from the database
- `chamber` - Chamber the action occurred in
  - categorical, includes values NA, Joint, House, and Senate
- `sequence` - Ordered sequence of history step
  - categorical, ranges from 1-99
- `action` - Description of history step
  - categorical, wide range of values, including vote counts, referrals to committees, and committee actions
- `year` - Year range of the legislature
  - categorical, generated by me, ranges from 2009-2010 to 2025-2026
- `legislature` - Maine State Congress
  - categorical, generated by me, ranges from 124 to 132

### people_csv

```
head(people_csv)
```

```
## # A tibble: 6 x 14
##   people_id name       first_name middle_name last_name suffix nickname party_id
##   <chr>     <chr>      <chr>      <chr>       <chr>     <chr>  <chr>    <chr>
## 1 8737      Hannah Pi~ Hannah     M.          Pingree   <NA>   <NA>     1
## 2 8738      Elizabeth~ Elizabeth  H.          Mitchell  <NA>   <NA>     1
## 3 8739      Lisa Marr~ Lisa       Tessier     Marrache  <NA>   <NA>     1
## 4 8740      Kevin Raye Kevin      L.          Raye      <NA>   <NA>     2
## 5 8741      Joshua Ta~ Joshua     A.          Tardy     <NA>   <NA>     2
## 6 8742      Wendy Pieh Wendy      <NA>        Pieh      <NA>   <NA>     1
## # i 6 more variables: party <chr>, role_id <chr>, role <chr>, district <chr>,
## #   year <chr>, legislature <chr>
```

- `people_id` - LegiScan person identifier
  - categorical, ranges from 191283 to 2023520
- `name` - Full name
  - categorical, wide range of values
- `first_name` - First name
  - categorical, wide range of values
- `middle_name` - Middle name

  - categorical, wide range of values

- `last_name` - Last name
  - categorical, wide range of values
- `suffix` - Suffix
  - categorical, wide range of values, many NAs
- `nickname` - Nickname
  - categorical, wide range of values, many NAs
- `party_id` - LegiScan political party identifier
  - categorical, ranges from 0-6
- `party` - Political party description
  - categorical, 1 corresponds to D (Democratic), 2 corresponds to R (Republican), 3 corresponds to I (Independent). 0 are NAs, 5 are L (Libertarian), 6 are N (Native, non-voting members).
- `role_id` - LegiScan legislative role identifier
  - categorical, ranges from 1-2
- `role` - Legislative role description
  - categorical, 1 corresponds with State Representative, 2 corresponds with State Senator
- `district` - Legislative district
  - categorical, ranges from HD-001 (house district 1) to HD- (house district 2) and SD-001 to SD-035 (senate district 35). Additionally, small group of HD-TRIBE for tribal representatives.
- `year` - Year range of the legislature
  - categorical, generated by me, ranges from 2009-2010 to 2025-2026
- `legislature` - Maine State Congress
  - categorical, generated by me, ranges from 124 to 132

## Part III: Pressing Issues

1. This data is incomplete, as the 2025-2026 session is not complete yet, which means it isn't complete and maybe shouldn't be used as training data. This could make good test data, depending on the questions being asked.
2. This data contains easy join values (`people_id` and `bill_id`), but this might make the data too large to be manageable.