DATA 606 Data Project Proposal

Chad Bailley

Data Preparation

```
## load data
    fileLocation1 <- '02 Proficiency Data with Entity Demographics.csv'
    proficiency <-read.csv(fileLocation1, sep = ',')

## check the first few records
    head(proficiency)
```

```
##   AcademicYear ISDCode                                 ISDName
## 1      2014-15    3000 Allegan Area Educational Service Agency
## 2      2014-15    3000 Allegan Area Educational Service Agency
## 3      2014-15    3000 Allegan Area Educational Service Agency
## 4      2014-15    3000 Allegan Area Educational Service Agency
## 5      2014-15    3000 Allegan Area Educational Service Agency
## 6      2014-15    3000 Allegan Area Educational Service Agency
##   DistrictCode                            DistrictName BuildingCode
## 1         3000 Allegan Area Educational Service Agency         6730
## 2         3000 Allegan Area Educational Service Agency         6730
## 3         3000 Allegan Area Educational Service Agency         6730
## 4         3000 Allegan Area Educational Service Agency         6730
## 5         3000 Allegan Area Educational Service Agency         8425
## 6         3000 Allegan Area Educational Service Agency         8425
##                          BuildingName StudentGroup       ContentAreaName
## 1 Hillside Learning & Behavior Center All Students English Language Arts
## 2 Hillside Learning & Behavior Center All Students           Mathematics
## 3 Hillside Learning & Behavior Center All Students               Science
## 4 Hillside Learning & Behavior Center All Students        Social Studies
## 5                STAR Family Literacy All Students English Language Arts
## 6                STAR Family Literacy All Students           Mathematics
##        Grade nValidTested nMetProficient nNotMetProficient
## 1 All Grades           36             19                17
## 2 All Grades           36             15                21
## 3 All Grades           17              4                13
## 4 All Grades         < 10             --                --
## 5 All Grades         < 10             --                --
## 6 All Grades         < 10             --                --
##   PctMetProficient nTotalEnrolled nTestedGrades nAIAN nAsian nBlack
## 1          52.7778             75            38     0      0      3
## 2          41.6667             75            38     0      0      3
## 3          23.5294             75            38     0      0      3
## 4               --             75            38     0      0      3
## 5               --              8             2     0      0      1
## 6               --              8             2     0      0      1
##   nHispanic nNHPI nTMR nWhite nED  nEL nSE nMale nFemale PctTestedGrades
## 1         5     0    2     65  47 < 10  70    51      24           50.67
## 2         5     0    2     65  47 < 10  70    51      24           50.67
## 3         5     0    2     65  47 < 10   70    51      24           50.67
## 4         5     0    2     65  47 < 10   70    51      24           50.67
## 5         0     0    0      7   6 < 10 < 10     1       7           25.00
## 6         0     0    0      7   6 < 10 < 10     1       7           25.00
##   PctAIAN PctAsian PctBlack PctHispanic PctNHPI PctTMR PctWhite PctED
## 1       0        0      4.0        6.67       0   2.67    86.67 62.67
## 2       0        0      4.0        6.67       0   2.67    86.67 62.67
## 3       0        0      4.0        6.67       0   2.67    86.67 62.67
## 4       0        0      4.0        6.67       0   2.67    86.67 62.67
## 5       0        0     12.5        0.00       0   0.00    87.50 75.00
## 6       0        0     12.5        0.00       0   0.00    87.50 75.00
##   PctEL PctSE PctMale PctFemale
## 1    -- 93.33    68.0      32.0
## 2    -- 93.33    68.0      32.0
## 3    -- 93.33    68.0      32.0
## 4    -- 93.33    68.0      32.0
## 5    --    --    12.5      87.5
## 6    --    --    12.5      87.5
```

Research question

To what extent does the rate of economic disadvantage correlate with proficiency on state assessments and by how much does that correlation vary across content areas.

Cases

In this data set each record is a single case representing a school's rate proficiency on the state assessment for a given content area (Mathematics, English Language Arts, Science, and Social Studies) and academic year. The dataset contains multiple years and so has 56372 cases.

Data collection

This dataset was acquired through a request to the Michigan Department of Education.

Type of study

This dataset contains observational data.

Data Source

This data is publically available at https://raw.githubusercontent.com/ChadRyanBailey/606-Statistics-and-Probability/master/606-Final-Project/02%20Proficiency%20Data%20with%20Entity%20Demographics.csv

Dependent Variable

The response variable in this dataset is [PctMetProficient] which gives the percent of students that are proficient on the state assessment for the given year, building, and content area.

Independent Variable

There are multiple independent variables in this dataset. The ones used for this project will be * [PctEd], a quantative which gives the percent of students that are economically disadvantaged * [ContentAreaName], a qualitative variable giving the content area of the state assessment

Relevant summary statistics

Initial review of the data

```
## get a general summary of each field
summary(proficiency)
```

```
##   AcademicYear      ISDCode                 ISDName       DistrictCode
##  2014-15:12692   Min.   : 3000   Wayne RESA     : 9051   Min.   : 1010
##  2015-16:12605   1st Qu.:33000   Oakland Schools: 5777   1st Qu.:33220
##  2016-17:12508   Median :52000   Macomb ISD     : 4069   Median :52110
##  2017-18: 9285   Mean   :51749   Kent ISD       : 3649   Mean   :52080
##  2018-19: 9282   3rd Qu.:74000   Genesee ISD    : 2284   3rd Qu.:74040
##                  Max.   :84000   Ottawa Area ISD: 1662   Max.   :84060
##                                  (Other)        :29880
##                                     DistrictName    BuildingCode
##  Detroit Public Schools Community District:  943   Min.   :   1
##  Grand Rapids Public Schools              :  808   1st Qu.:1547
##  Detroit City School District             :  706   Median :3153
##  Utica Community Schools                  :  669   Mean   :4126
##  Dearborn City School District            :  619   3rd Qu.:6553
##  Ann Arbor Public Schools                 :  540   Max.   :9994
##  (Other)                                  :52087
##                        BuildingName         StudentGroup
##  Central Elementary School   :  134   All Students:56372
##  Washington Elementary School:   97
##  North Elementary School     :   88
##  Roosevelt Elementary School :   85
##  Central High School         :   68
##  Lincoln School              :   65
##  (Other)                     :55835
##               ContentAreaName         Grade        nValidTested
##  English Language Arts:16167   All Grades:56372   < 10   : 2847
##  Mathematics          :16168                      54     :  313
##  Science              : 9430                      60     :  302
##  Social Studies       :14607                      14     :  296
##                                                   11     :  291
##                                                   13     :  289
##                                                   (Other):52034
##  nMetProficient  nNotMetProficient PctMetProficient nTotalEnrolled
##  < 3    : 5011   --     : 7858     --     : 7858    Min.   :   1.0
##  --     : 2942   54     :  410     33.3333:  278    1st Qu.: 248.0
##  3      : 1337   56     :  379     25     :  262    Median : 391.0
##  4      : 1236   55     :  375     50     :  248    Mean   : 471.1
##  5      : 1031   57     :  374     20     :  200    3rd Qu.: 568.0
##  6      :  977   58     :  370     16.6667:  173    Max.   :5960.0
##  (Other):43838   (Other):46606     (Other):47353
##  nTestedGrades        nAIAN             nAsian            nBlack
##  Min.   :   0.0   Min.   :  0.000   Min.   :   0.00   Min.   :   0.00
##  1st Qu.: 105.0   1st Qu.:  0.000   1st Qu.:   0.00   1st Qu.:   3.00
##  Median : 199.0   Median :  1.000   Median :   2.00   Median :  15.00
##  Mean   : 251.1   Mean   :  3.149   Mean   :  15.69   Mean   :  85.42
##  3rd Qu.: 331.0   3rd Qu.:  2.000   3rd Qu.:   9.00   3rd Qu.:  82.00
##  Max.   :1831.0   Max.   :512.000   Max.   :1672.00   Max.   :3834.00
##
##    nHispanic           nNHPI             nTMR            nWhite
##  Min.   :   0.00   Min.   : 0.000   Min.   :  0.00   Min.   :   0.0
##  1st Qu.:   5.00   1st Qu.: 0.000   1st Qu.:  3.00   1st Qu.:  80.0
##  Median :  16.00   Median : 0.000   Median : 10.00   Median : 257.0
##  Mean   :  35.22   Mean   : 0.323   Mean   : 16.64   Mean   : 314.6
##  3rd Qu.:  36.00   3rd Qu.: 0.000   3rd Qu.: 23.00   3rd Qu.: 418.0
##  Max.   :2784.00   Max.   :44.000   Max.   :442.00   Max.   :5394.0
##
##       nED              nEL             nSE            nMale
##  Min.   :   0.0   < 10   :34612   < 10   : 5643   Min.   :   0.0
##  1st Qu.:  92.0   10     :  984   43     :  836   1st Qu.: 128.0
##  Median : 175.0   12     :  793   47     :  826   Median : 200.0
##  Mean   : 222.1   11     :  783   34     :  806   Mean   : 241.3
##  3rd Qu.: 282.0   14     :  715   41     :  806   3rd Qu.: 292.0
##  Max.   :4544.0   13     :  690   49     :  797   Max.   :3074.0
##                   (Other):17795   (Other):46658
##     nFemale       PctTestedGrades     PctAIAN          PctAsian
##  Min.   :   0.0   Min.   :  0.00   Min.   : 0.0000   Min.   : 0.000
##  1st Qu.: 119.0   1st Qu.: 33.47   1st Qu.: 0.0000   1st Qu.: 0.000
##  Median : 190.0   Median : 50.98   Median : 0.2200   Median : 0.640
##  Mean   : 229.8   Mean   : 54.64   Mean   : 0.9646   Mean   : 2.464
##  3rd Qu.: 278.0   3rd Qu.: 64.99   3rd Qu.: 0.5900   3rd Qu.: 1.880
##  Max.   :2976.0   Max.   :100.00   Max.   :89.1100   Max.   :78.820
##
##     PctBlack       PctHispanic         PctNHPI            PctTMR
##  Min.   :  0.00   Min.   :  0.000   Min.   : 0.0000   Min.   :  0.000
##  1st Qu.:  0.97   1st Qu.:  2.020   1st Qu.: 0.0000   1st Qu.:  1.080
##  Median :  4.09   Median :  4.190   Median : 0.0000   Median :  2.830
##  Mean   : 18.70   Mean   :  7.501   Mean   : 0.0677   Mean   :  3.705
##  3rd Qu.: 21.46   3rd Qu.:  7.870   3rd Qu.: 0.0000   3rd Qu.:  5.250
##  Max.   :100.00   Max.   :100.000   Max.   :12.1900   Max.   :100.000
##
##     PctWhite          PctED            PctEL           PctSE
##  Min.   :  0.00   Min.   :  0.00   --     :34612   --     : 5643
##  1st Qu.: 49.70   1st Qu.: 33.33   3.27   :   46   100    : 1482
##  Median : 79.92   Median : 53.02   2.04   :   44   14.29  :  157
##  Mean   : 66.57   Mean   : 52.71   2.18   :   42   11.11  :  152
##  3rd Qu.: 90.21   3rd Qu.: 72.00   3.1    :   42   12.5   :  137
##  Max.   :100.00   Max.   :100.00   1.94   :   41   16.67  :  120
##                                    (Other):21545   (Other):48681
##     PctMale         PctFemale
##  Min.   :  0.00   Min.   :  0.00
##  1st Qu.: 49.49   1st Qu.: 46.15
##  Median : 51.50   Median : 48.50
##  Mean   : 52.50   Mean   : 47.50
##  3rd Qu.: 53.85   3rd Qu.: 50.51
##  Max.   :100.00   Max.   :100.00
##
```

The summary shows two issues:

- There are many unneeded columns and
- Some of the columns of interest have supressed values. These records will need to have values imputed or be removed.

Reducing the width of the data

A new dataset is created to onlyl contain the columns of interest and to rename some columns to have shorter names

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)

proficiencySlim <- proficiency %>%
  ## limit to only the fields of current interest
  select (AcademicYear
          ,BuildingCode
          ,ContentAreaName
          ,nValidTested
          ,nMetProficient
          ,nNotMetProficient
          ,PctMetProficient
          ,nTotalEnrolled
          ,nED
          ,PctED) %>%
  ## rename to shorter field names
  rename(nTested = nValidTested
         ,nProf = nMetProficient
         ,nNonProf = nNotMetProficient
         ,PctProf = PctMetProficient
         ,nEnrolled = nTotalEnrolled)
```

Dealing with suppression

As can be seen in the columns {nTested, nProf, nNonProf, and PctProf}, the file has records that have been suppressed. This is typical for public education data. The suppression is done to protect the privacy of small groups of students.

Flag records that have suppression applied and count records with each type of suppression case

```
# add flags to review each suppression condition
    proficiencySlim <- proficiencySlim %>%
      mutate( HasTestedLT10 = ifelse(nTested == '< 10', 1, 0)
              ,HasProfLT3 = ifelse(nProf == '< 3', 1, 0)
              ,HasNonProfLT3 = ifelse(nNonProf == '< 3', 1, 0)
              ,HasEitherProfOrNonProfLT3 = ifelse(HasProfLT3 == 1 | HasNonProfLT3 == 1, 1, 0)
              ,HasRecord = 1)


# get the count of records by suppression conditions
    proficiencySlim %>%
      summarise(nTotal = sum(HasRecord)
                ,nTestedLT10 = sum(HasTestedLT10)
                ,nProfLT3 = sum(HasProfLT3)
                ,nNonProfLT3 = sum(HasNonProfLT3)
                ,nEitherProfOrNonProfLT3 = sum(HasEitherProfOrNonProfLT3))
```

```
##   nTotal nTestedLT10 nProfLT3 nNonProfLT3 nEitherProfOrNonProfLT3
## 1  56372        2847     5011          95                    5106
```

Remove records where values cannot be imputed

Remove records with less than 10 valid tested students as all data for those records is supressed and a value for imputation cannot be applied

```
## remove records with < 10 valid tested; all data for these records are suppressed
  proficiencySlim <- proficiencySlim %>%
    filter(HasTestedLT10 == 0)

  nrow(proficiencySlim)
```

```
## [1] 53525
```

Impute supressed values where possible

Since <3 is equal to the set {0, 1, 2}, the middle value "1" will be used as the imputed value. Also, percentages will be calculated for suppressed records using the imputed value.

```
## deal with cases where suppression is applied because nearly all nor nearly
## none of the students were proficient
proficiencySlim <- proficiencySlim %>%
    #convert factors to characters
    mutate(nTested = as.character(nTested)
           ,nProf = as.character(nProf)
           ,nNonProf = as.character(nNonProf)
           ,PctProf = as.character(PctProf)
           ) %>%

    #convert the characters to numerics
    mutate(nTested = as.numeric(nTested)
           ,nProf = as.numeric(nProf)
           ,nNonProf = as.numeric(nNonProf)
           ,PctProf = as.numeric(PctProf)
           ) %>%

    # for count variables (nProf and nNonProf) replace the suppression flag with imputed count
    mutate(nProf = ifelse(HasProfLT3 == 1, 1, nProf)
           ,nProf = ifelse(HasNonProfLT3 == 1, nTested - 1, nProf)

           ,nNonProf = ifelse(HasNonProfLT3 == 1, 1, nNonProf)
           ,nNonProf = ifelse(HasProfLT3 == 1, nTested - 1, nNonProf)
           ) %>%

    # for percentage variables (PctProf and PctNonProf) replace the suppression flag
    # with a calucuated percentage using the imputed counts
    mutate(PctProf = ifelse(HasProfLT3 == 1, round(nProf*1.0/nTested*100.0, 2), PctProf)
           , PctProf = ifelse(HasNonProfLT3 == 1, round(nProf*1.0/nTested*100.0, 2), PctProf))
```

```
## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion
```

Summarize the “cleaned” dataset

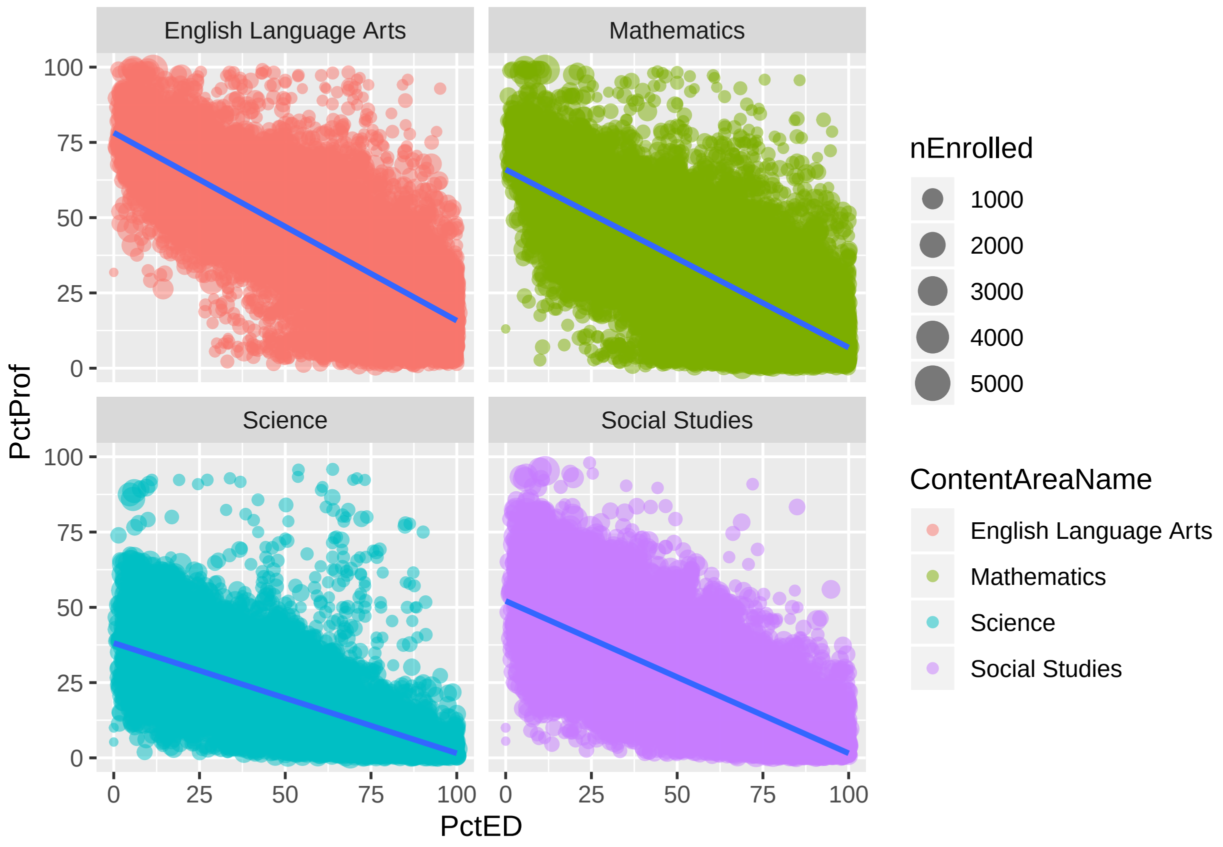
Summary statistics for each variable

```
## get
summary(proficiencySlim)
```

```
##   AcademicYear    BuildingCode               ContentAreaName
##  2014-15:11932   Min.   :   1   English Language Arts:15540
##  2015-16:11940   1st Qu.:1519   Mathematics          :15540
##  2016-17:11899   Median :3142   Science              : 8810
##  2017-18: 8877   Mean   :4073   Social Studies       :13635
##  2018-19: 8877   3rd Qu.:6396
##                  Max.   :9994
##     nTested           nProf           nNonProf         PctProf
##  Min.   :  10.0   Min.   :  1.00   Min.   :   1.0   Min.   : 0.32
##  1st Qu.:  67.0   1st Qu.: 11.00   1st Qu.:  46.0   1st Qu.:14.97
##  Median : 143.0   Median : 41.00   Median :  83.0   Median :31.25
##  Mean   : 192.7   Mean   : 75.98   Mean   : 116.8   Mean   :33.01
##  3rd Qu.: 251.0   3rd Qu.:104.00   3rd Qu.: 150.0   3rd Qu.:48.53
##  Max.   :1865.0   Max.   :868.00   Max.   :1575.0   Max.   :99.71
##    nEnrolled           nED              PctED        HasTestedLT10
##  Min.   :   1.0   Min.   :   0.0   Min.   : 0.00   Min.   :0
##  1st Qu.: 274.0   1st Qu.: 104.0   1st Qu.: 32.68   1st Qu.:0
##  Median : 404.0   Median : 184.0   Median : 52.48   Median :0
##  Mean   : 493.2   Mean   : 232.2   Mean   : 52.30   Mean   :0
##  3rd Qu.: 582.0   3rd Qu.: 291.0   3rd Qu.: 71.54   3rd Qu.:0
##  Max.   :5960.0   Max.   :4544.0   Max.   :100.00   Max.   :0
##    HasProfLT3       HasNonProfLT3      HasEitherProfOrNonProfLT3
##  Min.   :0.00000   Min.   :0.000000   Min.   :0.00000
##  1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:0.00000
##  Median :0.00000   Median :0.000000   Median :0.00000
##  Mean   :0.09362   Mean   :0.001775   Mean   :0.09539
##  3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:0.00000
##  Max.   :1.00000   Max.   :1.000000   Max.   :1.00000
##    HasRecord
##  Min.   :1
##  1st Qu.:1
##  Median :1
##  Mean   :1
##  3rd Qu.:1
##  Max.   :1
```

Scatter plot of the relationship of interest

```
ggplot(proficiencySlim
       , aes(x = PctED, y = PctProf)) +
  geom_point(aes(size = nEnrolled, color = ContentAreaName), alpha = 0.5) +
  facet_wrap(~ContentAreaName)+
  geom_smooth(method=lm)
```


English Language Arts
Mathematics
Science
Social Studies
PctProf
PctED
nEnrolled
ContentAreaName
English Language Arts
Mathematics
Science
Social Studies