# Final Project code

April 23, 2021

## 1 Cleanning the data

```
[1]: import pandas as pd
```

### 1.1 Data 1

```
[2]: # Read csv file
     # Original data: https://www.kaggle.com/wsj/college-salaries?
      ↪select=salaries-by-region.csv
     data1 = pd.read_csv('salaries-by-region.csv', error_bad_lines=False)
```

```
[3]: # Overview of the data
     data1.head(20)
```

```
[3]:
                                        School Name      Region  \
0                               Stanford University  California
1          California Institute of Technology (CIT)  California
2                               Harvey Mudd College  California
3                University of California, Berkeley  California
4                                Occidental College  California
5                          Cal Poly San Luis Obispo  California
6    University of California at Los Angeles (UCLA)  California
7        University of California, San Diego (UCSD)  California
8                                    Pomona College  California
9           University of Southern California (USC)  California
10                  University of California, Davis  California
11           University of California, Irvine (UCI)  California
12                 San Jose State University (SJSU)  California
13   University of California, Santa Barbara (UCSB)  California
14         California State University (CSU), Chico  California
15    California State University, Fullerton (CSUF)  California
16            San Francisco State University (SFSU)  California
17                San Diego State University (SDSU)  California
18  California State University, Long Beach (CSULB)  California
19    California State University, East Bay (CSUEB)  California
```

```
   Starting Median Salary Mid-Career Median Salary  \
0              $70,400.00              $129,000.00   
1              $75,500.00              $123,000.00   
2              $71,800.00              $122,000.00   
3              $59,900.00              $112,000.00   
4              $51,900.00              $105,000.00   
5              $57,200.00              $101,000.00   
6              $52,600.00              $101,000.00   
7              $51,100.00              $101,000.00   
8              $48,600.00              $101,000.00   
9              $54,800.00               $99,600.00   
10             $52,300.00               $99,600.00   
11             $48,300.00               $96,700.00   
12             $53,500.00               $95,600.00   
13             $50,500.00               $95,000.00   
14             $47,400.00               $88,100.00   
15             $45,700.00               $87,000.00   
16             $47,300.00               $86,400.00   
17             $46,200.00               $85,200.00   
18             $45,100.00               $84,700.00   
19             $49,200.00               $84,300.00   

   Mid-Career 10th Percentile Salary Mid-Career 25th Percentile Salary  \
0                         $68,400.00                        $93,100.00   
1                                NaN                       $104,000.00   
2                                NaN                        $96,000.00   
3                         $59,500.00                        $81,000.00   
4                                NaN                        $54,800.00   
5                         $55,000.00                        $74,700.00   
6                         $51,300.00                        $72,500.00   
7                         $51,700.00                        $75,400.00   
8                                NaN                        $63,300.00   
9                         $49,700.00                        $73,800.00   
10                        $52,000.00                        $71,600.00   
11                        $47,800.00                        $66,000.00   
12                        $50,700.00                        $70,500.00   
13                        $51,300.00                        $71,200.00   
14                        $46,800.00                        $62,800.00   
15                        $45,400.00                        $62,500.00   
16                        $45,100.00                        $62,700.00   
17                        $45,500.00                        $61,800.00   
18                        $47,400.00                        $62,500.00   
19                        $46,000.00                        $62,400.00   

   Mid-Career 75th Percentile Salary Mid-Career 90th Percentile Salary  
0                        $184,000.00                       $257,000.00  
```

```
1                         $161,000.00                                 NaN
2                         $180,000.00                                 NaN
3                         $149,000.00                         $201,000.00
4                         $157,000.00                                 NaN
5                         $133,000.00                         $178,000.00
6                         $139,000.00                         $193,000.00
7                         $131,000.00                         $177,000.00
8                         $161,000.00                                 NaN
9                         $140,000.00                         $201,000.00
10                        $135,000.00                         $202,000.00
11                        $123,000.00                         $172,000.00
12                        $122,000.00                         $156,000.00
13                        $129,000.00                         $173,000.00
14                        $122,000.00                         $154,000.00
15                        $119,000.00                         $158,000.00
16                        $114,000.00                         $150,000.00
17                        $116,000.00                         $158,000.00
18                        $113,000.00                         $154,000.00
19                        $115,000.00                         $155,000.00
```

[4]:

```
# Rename columns
data1 = data1.rename(
    columns = {
        "School Name": 'name',
        'Starting Median Salary':'starting_median_salary'
    }
)
```

[5]:

```
# Reformatting starting_median_salary from '$AAAA'(str) to AAA.00(float)
for i in range (len(data1)):
    row = data1['starting_median_salary'][i][1:]
    lst = list(row.split(','))
    data1['starting_median_salary'][i] = float(''.join(lst))
```

[6]:

```
# Necessary columns for analyze
kept_cols = [
    'name',
    'starting_median_salary',
]

# Drop unnecessary columns
for col in list(data1.columns):
    if col not in kept_cols:
        data1.drop(col, axis='columns', inplace=True)
```

[7]:

```
data1
```

[7]:
```
                                                name starting_median_salary
0                                Stanford University                  70400
1           California Institute of Technology (CIT)                  75500
2                                Harvey Mudd College                  71800
3                 University of California, Berkeley                  59900
4                                 Occidental College                  51900
..                                               ...                    ...
315   State University of New York (SUNY) at Potsdam                  38000
316                               Niagara University                  36900
317  State University of New York (SUNY) at Fredonia                  37800
318                     University of Southern Maine                  39400
319                                    Mercy College                  43700

[320 rows x 2 columns]
```

## 1.2 Data 2: College characteristics data

[8]:
```
# Load data 2 with college demographics
# Orginal data link: https://opportunityinsights.org/data/
# Name data: College Level Characteristics from the IPEDS Database and the␣
 ↪College Scorecard
data2 = pd.read_csv('mrc_table10.csv')
```

[9]:
```
data2.head(10)
```

[9]:
```
   super_opeid                                                name  region  \
0        30955     ASA Institute Of Business & Computer Technology       1
1         3537                        Abilene Christian University       3
2         1541                Abraham Baldwin Agricultural College       3
3         7531                           Academy Of Art University       4
4         1345                              Adams State University       4
5         2666                                  Adelphi University       1
6         2860  Adirondack Community College - SUNY Office Of ...       1
7         2234                                      Adrian College       2
8        11484                   Advanced Institute Of Hair Design       2
9        31275                       Advanced Technology Institute       3

  state  fips     cz          czname  cfips           county    zip  ...  \
0    NY    36  19400        New York  36047            Kings  11201  ...
1    TX    48  32501         Abilene  48441           Taylor  79699  ...
2    GA    13   8503        Valdosta  13277             Tift  31793  ...
3    CA     6  37800   San Francisco   6075    San Francisco  94105  ...
4    CO     8  34805         Alamosa   8003          Alamosa  81101  ...
5    NY    36  19400        New York  36059           Nassau  11530  ...
6    NY    36  18600          Albany  36113           Warren  12804  ...
```

```
7    MI    26  11500         Jackson  26091             Lenawee  49221  ...
8    WI    55  24100       Milwaukee  55079           Milwaukee  53228  ...
9    VA    51   2000  Virginia Beach  51810  Virginia Beach City  23462  ...

   hisp_share_fall_2000 alien_share_fall_2000  pct_arthuman_2000  \
0              0.073324              0.071229           0.000000   
1              0.056724              0.039943          10.785619   
2              0.016730              0.013688           0.000000   
3              0.205893              0.271817          96.091957   
4              0.275481              0.002404          10.850440   
5              0.067118              0.040981           9.009009   
6              0.009203              0.002856           0.000000   
7              0.017576              0.019426          19.631903   
8              0.036201              0.000000           0.000000   
9              0.021978              0.000000           0.000000   

   pct_business_2000  pct_health_2000  pct_multidisci_2000  \
0           6.603774        11.425576             0.000000   
1          22.503330         5.059920             9.720373   
2           4.100228        12.072893            47.152618   
3           0.000000         0.000000             0.000000   
4          24.046921         0.000000            33.431084   
5          28.078077        13.963964             3.603604   
6          19.017094        12.606837            50.000000   
7          28.834356         0.613497            12.269938   
8           0.000000         0.000000             0.000000   
9           0.000000         0.000000             0.000000   

   pct_publicsocial_2000  pct_stem_2000  pct_socialscience_2000  \
0               0.000000      81.970650                0.000000   
1               8.788282      11.318242               31.691078   
2               6.378132      29.612757                0.000000   
3               0.000000       0.000000                3.908046   
4               0.000000      13.489737               18.181818   
5               8.558559       6.006006               30.780781   
6               7.478632       8.547009                2.136752   
7               5.521472      16.564417               16.564417   
8               0.000000       0.000000                0.000000   
9               0.000000      15.163935                0.000000   

   pct_tradepersonal_2000  
0                0.000000  
1                0.133156  
2                0.683371  
3                0.000000  
4                0.000000  
5                0.000000  
```

```
6            0.213675
7            0.000000
8          100.000000
9           84.836067

[10 rows x 49 columns]
```

[10]:

```
# Show all columns of data2
data2.columns
```

[10]:

```
Index(['super_opeid', 'name', 'region', 'state', 'fips', 'cz', 'czname',
       'cfips', 'county', 'zip', 'tier', 'tier_name', 'type', 'iclevel',
       'public', 'barrons', 'exp_instr_pc_2000', 'exp_instr_pc_2013', 'multi',
       'hbcu', 'flagship', 'ipeds_enrollment_2013', 'ipeds_enrollment_2000',
       'sticker_price_2013', 'sticker_price_2000', 'grad_rate_150_p_2013',
       'grad_rate_150_p_2002', 'avgfacsal_2013', 'avgfacsal_2001',
       'sat_avg_2013', 'sat_avg_2001', 'scorecard_netprice_2013',
       'scorecard_rej_rate_2013', 'scorecard_median_earnings_2011',
       'endowment_pc_2000', 'exp_instr_2012', 'exp_instr_2000',
       'asian_or_pacific_share_fall_2000', 'black_share_fall_2000',
       'hisp_share_fall_2000', 'alien_share_fall_2000', 'pct_arthuman_2000',
       'pct_business_2000', 'pct_health_2000', 'pct_multidisci_2000',
       'pct_publicsocial_2000', 'pct_stem_2000', 'pct_socialscience_2000',
       'pct_tradepersonal_2000'],
      dtype='object')
```

[11]:

```
# Necessary columns for analyze
kept_cols = [
    'name',
    'starting_median_salary',
    'zip',
    'tier',
    'sat_avg_2013',
    'asian_or_pacific_share_fall_2000',
    'black_share_fall_2000',
    'hisp_share_fall_2000',
    'alien_share_fall_2000'
]

# Drop unnecessary columns
for col in list(data2.columns):
    if col not in kept_cols:
        data2.drop(col, axis='columns', inplace=True)
```

[12]:

```
# Rename columns
data2 = data2.rename(
    columns = {
        'asian_or_pacific_share_fall_2000': 'asian_or_pacific_share',
        'black_share_fall_2000':'black_share',
        'hisp_share_fall_2000': 'hisp_share',
        'alien_share_fall_2000': 'alien_share'
    }
)
```

### 1.3 Data 3: Collge by city and population

[13]:
```
# Import data
url = 'https://docs.google.com/spreadsheets/d/e/
 ↪2PACX-1vTcLn5ahGpzQAAK4MZqlZ-vcsR2OjjxECPN4hOMnegZphhmKP6VGG4GhXVBV9qnjU4azuRC_OaURp8X/
 ↪pub?gid=1955362680&single=true&output=csv'
data3 = pd.read_csv(url)
```

[14]:
```
data3.head(10)
```

[14]:
```
                                      name           city population_in_2010
0                      Stanford University       Stanford             13,809
1                      Harvey Mudd College      Claremont             34,926
2                       Occidental College    Los Angeles          9,818,605
3                           Pomona College      Claremont             34,926
4                San Jose State University       San Jose            945,942
5       California State University, Chico          Chico             86,187
6   California State University, Fullerton      Fullerton            135,161
7           San Francisco State University  San Francisco            805,235
8               San Diego State University      San Diego          1,307,402
9  California State University, Long Beach     Long Beach            462,257
```

[15]:
```
# Formatting column population_in_2010
# Changing the string to float
for i in range (len(data3)):
    row = data3['population_in_2010'][i]
    lst = list(row.split(','))
    data3['population_in_2010'][i] = int(''.join(lst))
```

[16]:
```
data3.head(10)
```

[16]:
```
                                      name           city population_in_2010
0                      Stanford University       Stanford              13809
1                      Harvey Mudd College      Claremont              34926
2                       Occidental College    Los Angeles            9818605
3                           Pomona College      Claremont              34926
4                San Jose State University       San Jose             945942
5       California State University, Chico          Chico              86187
```

```
6   California State University, Fullerton      Fullerton            135161
7           San Francisco State University  San Francisco            805235
8               San Diego State University      San Diego           1307402
9  California State University, Long Beach     Long Beach            462257
```

### 1.4 Data 4: Parent income by college

[17]:

```
# Load data 2 with parent income
# Orginal data link: https://opportunityinsights.org/data/
# Name of dataset: Baseline Cross-Sectional Estimates of Child and Parent Income␣
 ↪Distributions by College
data4 = pd.read_csv('mrc_table2.csv')
```

[18]:

```
data4.columns
```

[18]:

```
Index(['super_opeid', 'name', 'type', 'tier', 'tier_name', 'iclevel', 'region',
       'state', 'cz', 'czname', 'cfips', 'county', 'multi', 'count', 'female',
       'k_married', 'mr_kq5_pq1', 'mr_ktop1_pq1', 'par_mean', 'par_median',
       'par_rank', 'par_q1', 'par_q2', 'par_q3', 'par_q4', 'par_q5',
       'par_top10pc', 'par_top5pc', 'par_top1pc', 'par_toppt1pc', 'k_rank',
       'k_mean', 'k_median', 'k_median_nozero', 'k_0inc', 'k_q1', 'k_q2',
       'k_q3', 'k_q4', 'k_q5', 'k_top10pc', 'k_top5pc', 'k_top1pc',
       'k_rank_cond_parq1', 'k_rank_cond_parq2', 'k_rank_cond_parq3',
       'k_rank_cond_parq4', 'k_rank_cond_parq5', 'kq1_cond_parq1',
       'kq2_cond_parq1', 'kq3_cond_parq1', 'kq4_cond_parq1', 'kq5_cond_parq1',
       'kq1_cond_parq2', 'kq2_cond_parq2', 'kq3_cond_parq2', 'kq4_cond_parq2',
       'kq5_cond_parq2', 'kq1_cond_parq3', 'kq2_cond_parq3', 'kq3_cond_parq3',
       'kq4_cond_parq3', 'kq5_cond_parq3', 'kq1_cond_parq4', 'kq2_cond_parq4',
       'kq3_cond_parq4', 'kq4_cond_parq4', 'kq5_cond_parq4', 'kq1_cond_parq5',
       'kq2_cond_parq5', 'kq3_cond_parq5', 'kq4_cond_parq5', 'kq5_cond_parq5',
       'ktop1pc_cond_parq1', 'ktop1pc_cond_parq2', 'ktop1pc_cond_parq3',
       'ktop1pc_cond_parq4', 'ktop1pc_cond_parq5', 'k_married_cond_parq1',
       'k_married_cond_parq2', 'k_married_cond_parq3', 'k_married_cond_parq4',
       'k_married_cond_parq5', 'shareimputed', 'imputed'],
      dtype='object')
```

[19]:

```
# Necessary columns for analyze
kept_cols = [
    'name',
    'par_median'
]

# Drop unnecessary columns
for col in list(data4.columns):
    if col not in kept_cols:
        data4.drop(col, axis='columns', inplace=True)
```

```
[20]: data4
```

```
[20]:                                          name  par_median
      0     ASA Institute Of Business & Computer Technology       29000
      1                   Abilene Christian University      101000
      2            Abraham Baldwin Agricultural College       66000
      3                       Academy Of Art University       92300
      4                          Adams State University       67200
      ...                                            ...         ...
      2197               Yuba Community College District       48700
      2198                            Zane State College       53800
      2199                            Late College Goers       43300
      2200          Never Attended College (up to year 2013)       35200
      2201               Colleges with insufficient data       50500

      [2202 rows x 2 columns]
```

## 1.5 Merge datasets

```
[21]: ## Merge data1 and data2 by college name, only keeps colleges in both dataset
      data5 = pd.merge(data1, data2, on='name', how='inner')
```

```
[22]: ## Merge data3 and data4 by college name
      data6 = pd.merge(data3, data4, on='name', how='inner')
```

```
[23]: ## Merge data5 and data6 by college name
      data = pd.merge(data5, data6, on='name', how='inner')
```

```
[24]: ## Drop NA row
      data = data.dropna(axis = 0, how ='any')
```

```
[25]: data
```

```
[25]:                           name starting_median_salary    zip  tier  \
      0          Stanford University                  70400  94305     1
      1          Harvey Mudd College                  71800  91711     2
      2            Occidental College                  51900  90041     2
      3               Pomona College                  48600  91711     2
      4     Humboldt State University                  42600  95521     5
      ..                         ...                    ...    ...   ...
      112       Quinnipiac University                  43000   6518     4
      114           Skidmore College                  41600  12866     4
      115           Moravian College                  42500  18018     6
```

```
116         Suffolk University                  42100   2108     6
118         Niagara University                  36900  14109     6

     sat_avg_2013  asian_or_pacific_share  black_share  hisp_share  \
0          1475.0                0.209992     0.073294    0.086482
1          1494.0                0.210600     0.004184    0.040446
2          1300.0                0.180857     0.061656    0.140341
3          1460.0                0.137230     0.037484    0.076239
4           985.0                0.031071     0.024115    0.082393
..            ...                     ...          ...         ...
112        1090.0                0.019203     0.023333    0.029940
114        1240.0                0.035904     0.023664    0.049776
115        1020.0                0.011608     0.016251    0.024956
116        1040.0                0.054392     0.038976    0.043339
118        1035.0                0.007683     0.040841    0.019410

     alien_share              city population_in_2010  par_median
0       0.040451          Stanford              13809      172600
1       0.034868         Claremont              34926      139800
2       0.041691       Los Angeles            9818605      122400
3       0.019695         Claremont              34926      161600
4       0.004483            Arcata              17231       96000
..           ...               ...                ...         ...
112     0.003304            Hamden                879      127000
114     0.010200  Saratoga Springs              26586      175400
115     0.010447         Bethlehem              71133       97100
116     0.157068            Boston             617594       88100
118     0.048120    Niagara County             216479       92300

[94 rows x 12 columns]
```

```
[26]: data.to_csv('data.csv')
```

```
[ ]:
```