Data_Bootcamp_Final_Project

Household Structure and Higher Education: An Analysis on The Impact of Childhood Living Arrangements on Attaining Higher Education

Cayley Boyd, Meetali Gupta, Angela Yang
Data Bootcamp, Section 2
Fall 2018

Abstract

Nature vs. Nurture - our environment has long been thought to have an effect on our success in life. We decided to explore the relationship between household structures of children growing up and their likelihood to attain higher education. To clarify, by "children" we are talking about young adults who are around the ages of 18/19 who are dependents finishing their last year of high school. Later, we look at young adults/adults who have received their Bachelor's and Associate Degrees. Looking at panel data across all 50 states, in addition the District of Columbia and Puerto Rico, across the years 2000-2017, we compared varying aspects of household structure with different levels of educational attainment on the state level. We were unable to draw a solid correlation between the two variables and suspect more influential ommitted variables are responsible for between-state differences in educational attainment.

Set Up

Importing Packages

In [14]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
import statsmodels.formula.api as smf
```

Retrieving Data

We used the ACS Household Structure Data pulled from the Kids Count Data Center for the years 2000-2017 (with some data points omitted in this most recent year). We pulled data on both household structures and children's education levels.

In [15]:
```
# Data containing the household structure of various states in the US
over several years
child_pop_household_type = pd.read_excel('C:/Users/meeta/Desktop/Data
Bootcamp/Data/Child population by household type.xlsx')
child_neither_parent = pd.read_excel('C:/Users/meeta/Desktop/Data Boot
camp/Data/Children living with neither parent.xlsx')

# Data containing the education statistics of adults between 25 and 34
in various states in the US over different years
edu_pop_25_to_34 = pd.read_excel('C:/Users/meeta/Desktop/Data Bootcamp
/Data/Educational attainment of population ages 25 to 34.xlsx')
```

Defining the Variables: Household Data

Child Population by Household Type
Definition: Percent of total child population in married-couple, father only, and mother only households.

Children living with Neither Parent
Definition: The share of children under age 18 living in households where neither parent resides.

Defining the Variables: Education Data

Educational Attainment of population ages 25-34
Definition: The share of all adults ages 25 to 34 by educational attainment.

Cleaning the child_pop_household_type and child_neither_parent datasets

In [16]:
```
## Cleaning the child_pop_household_type data

# making location the index
child_pop_household_type_new = child_pop_household_type.set_index(['Lo
cation'])

# dropping the fields with United States with them
child_pop_household_type_new = child_pop_household_type_new.drop(['Uni
ted States'])

# dropping N.A
child_pop_household_type_new = child_pop_household_type_new[child_pop_
household_type_new.Data != 'N.A.']

# dropping the fields with Number in them
child_pop_household_type_new = child_pop_household_type_new[child_pop_
household_type_new.DataFormat != 'Number']

# dropping the DataFormat field
child_pop_household_type_new = child_pop_household_type_new.drop(['Dat
aFormat'], axis = 1)
```

In [17]:
```
## Cleaning the child_neither_parent data using the same process as th
at of child_pop_household_type data

child_neither_parent_new = child_neither_parent.set_index(['Location']
)
child_neither_parent_new = child_neither_parent_new.drop(['United Stat
es'])
child_neither_parent_new = child_neither_parent_new[child_neither_pare
nt_new.Data != 'N.A.']
child_neither_parent_new = child_neither_parent_new[child_neither_pare
nt_new.DataFormat != 'Number']
child_neither_parent_new = child_neither_parent_new.drop(['DataFormat'
], axis = 1)

#Creating a list for the child_neither_parent data
Household_Type = []
for i in range(879):
    Household_Type.append('Neither parent Households')

#Adding the Household_Type list to the child_neither_parent_new data
child_neither_parent_new['Household Type'] = Household_Type

##Changing the order of columns
child_neither_parent_new = child_neither_parent_new[['Household Type',
'TimeFrame', 'Data']]
```

In [18]:
```
## Combining the child_pop_household_type and child_neither_parent_new datasets together
Household_structure = child_pop_household_type_new.append(child_neither_parent_new)

## Reseting the index
Household_structure = Household_structure.reset_index()

## Changing the string type to float type
Household_structure['Data'] = pd.to_numeric(Household_structure['Data'], errors = 'coerce')
```

In [19]:
```
Household_structure
```

Out[19]:

	Location	Household Type	TimeFrame	Data
0	Washington	Mother only Households	2008	0.20
1	Montana	Married-couple Households	2000	0.77
2	Montana	Father only Households	2000	0.08
3	Montana	Mother only Households	2000	0.15
4	South Carolina	Married-couple Households	2011	0.59
5	South Carolina	Father only Households	2011	0.07
6	South Carolina	Mother only Households	2011	0.33
7	South Carolina	Father only Households	2012	0.07
8	South Carolina	Married-couple Households	2012	0.59
9	South Carolina	Mother only Households	2012	0.34
10	South Carolina	Father only Households	2009	0.08
11	South Carolina	Married-couple Households	2009	0.61
12	South Carolina	Mother only Households	2009	0.31
13	South Carolina	Father only Households	2007	0.06
14	South Carolina	Married-couple Households	2007	0.62
15	South Carolina	Mother only Households	2007	0.31
16	South Carolina	Father only Households	2010	0.07
17	South Carolina	Married-couple Households	2010	0.60
18	South Carolina	Mother only Households	2010	0.33
19	South Carolina	Married-couple Households	2000	0.64
20	South Carolina	Father only Households	2000	0.05

21	South Carolina	Mother only Households	2000	0.30
22	South Carolina	Married-couple Households	2001	0.63
23	South Carolina	Married-couple Households	2006	0.61
24	South Carolina	Father only Households	2001	0.05
25	South Carolina	Father only Households	2006	0.06
26	South Carolina	Mother only Households	2001	0.31
27	South Carolina	Mother only Households	2006	0.32
28	Montana	Married-couple Households	2002	0.75
29	Montana	Married-couple Households	2003	0.72
...	...	...	...	...
3642	Wisconsin	Neither parent Households	2016	0.04
3643	Wyoming	Neither parent Households	2016	0.06
3644	Wyoming	Neither parent Households	2014	0.06
3645	Wyoming	Neither parent Households	2012	0.05
3646	Wyoming	Neither parent Households	2011	0.05
3647	Wyoming	Neither parent Households	2015	0.05
3648	Wyoming	Neither parent Households	2010	0.04
3649	Wyoming	Neither parent Households	2009	0.06
3650	Wyoming	Neither parent Households	2013	0.04
3651	Wyoming	Neither parent Households	2008	0.06
3652	Wyoming	Neither parent Households	2007	0.06
3653	Wyoming	Neither parent Households	2006	0.07
3654	Wyoming	Neither parent Households	2004	0.06
3655	Wyoming	Neither parent Households	2005	0.05
3656	Wyoming	Neither parent Households	2003	0.04
3657	Wyoming	Neither parent Households	2001	0.06
3658	Wyoming	Neither parent Households	2002	0.06
3659	Wyoming	Neither parent Households	2000	0.05
3660	Puerto Rico	Neither parent Households	2005	0.09
3661	Puerto Rico	Neither parent Households	2006	0.09
3662	Puerto Rico	Neither parent Households	2007	0.09

3663	Puerto Rico	Neither parent Households	2008	0.06
3664	Puerto Rico	Neither parent Households	2013	0.05
3665	Puerto Rico	Neither parent Households	2009	0.05
3666	Puerto Rico	Neither parent Households	2010	0.05
3667	Puerto Rico	Neither parent Households	2015	0.04
3668	Puerto Rico	Neither parent Households	2011	0.04
3669	Puerto Rico	Neither parent Households	2012	0.04
3670	Puerto Rico	Neither parent Households	2014	0.04
3671	Puerto Rico	Neither parent Households	2016	0.04

3672 rows × 4 columns

Cleaning the edu_pop_25_to_34 dataset

In [20]:

```
## Cleaning the edu_pop_25_to_34_new
edu_pop_25_to_34_new = edu_pop_25_to_34.set_index(['Location'])
edu_pop_25_to_34_new = edu_pop_25_to_34_new.drop(['United States'])
edu_pop_25_to_34_new = edu_pop_25_to_34_new[edu_pop_25_to_34_new.DataFormat != 'Number']
edu_pop_25_to_34_new = edu_pop_25_to_34_new.drop(['DataFormat'], axis = 1)
edu_pop_25_to_34_new = edu_pop_25_to_34_new.reset_index()
```

In [21]:

```
edu_pop_25_to_34_new
```

Out[21]:

	Location	Education	TimeFrame	Data
0	Alabama	Not a high school graduate	2000	0.17
1	Alabama	High school diploma or GED	2000	0.52
2	Alabama	Associate's Degree	2000	0.08
3	Alabama	Bachelor's Degree	2000	0.18
4	Alabama	Graduate degree	2000	0.05
5	Alabama	Not a high school graduate	2001	0.14
6	Alabama	High school diploma or GED	2001	0.56
7	Alabama	Associate's Degree	2001	0.08
8	Alabama	Bachelor's Degree	2001	0.17

9	Alabama	Graduate degree	2001	0.05
10	Alabama	Not a high school graduate	2002	0.14
11	Alabama	High school diploma or GED	2002	0.54
12	Alabama	Associate's Degree	2002	0.08
13	Alabama	Bachelor's Degree	2002	0.17
14	Alabama	Graduate degree	2002	0.06
15	Alabama	Not a high school graduate	2003	0.16
16	Alabama	High school diploma or GED	2003	0.54
17	Alabama	Associate's Degree	2003	0.07
18	Alabama	Bachelor's Degree	2003	0.17
19	Alabama	Graduate degree	2003	0.06
20	Alabama	Not a high school graduate	2004	0.15
21	Alabama	High school diploma or GED	2004	0.51
22	Alabama	Associate's Degree	2004	0.08
23	Alabama	Bachelor's Degree	2004	0.18
24	Alabama	Graduate degree	2004	0.08
25	Alabama	Not a high school graduate	2005	0.16
26	Alabama	High school diploma or GED	2005	0.52
27	Alabama	Associate's Degree	2005	0.08
28	Alabama	Bachelor's Degree	2005	0.18
29	Alabama	Graduate degree	2005	0.06
...	...	...	...	...
4625	Puerto Rico	Not a high school graduate	2012	0.14
4626	Puerto Rico	High school diploma or GED	2012	0.46
4627	Puerto Rico	Associate's Degree	2012	0.11
4628	Puerto Rico	Bachelor's Degree	2012	0.23
4629	Puerto Rico	Graduate degree	2012	0.07
4630	Puerto Rico	Not a high school graduate	2013	0.14
4631	Puerto Rico	High school diploma or GED	2013	0.45
4632	Puerto Rico	Associate's Degree	2013	0.12
4633	Puerto Rico	Bachelor's Degree	2013	0.22
4634	Puerto Rico	Graduate degree	2013	0.07

4635	Puerto Rico	Not a high school graduate	2014	0.12
4636	Puerto Rico	High school diploma or GED	2014	0.45
4637	Puerto Rico	Associate's Degree	2014	0.12
4638	Puerto Rico	Bachelor's Degree	2014	0.23
4639	Puerto Rico	Graduate degree	2014	0.08
4640	Puerto Rico	Not a high school graduate	2015	0.11
4641	Puerto Rico	High school diploma or GED	2015	0.47
4642	Puerto Rico	Associate's Degree	2015	0.13
4643	Puerto Rico	Bachelor's Degree	2015	0.22
4644	Puerto Rico	Graduate degree	2015	0.07
4645	Puerto Rico	Not a high school graduate	2016	0.10
4646	Puerto Rico	High school diploma or GED	2016	0.47
4647	Puerto Rico	Associate's Degree	2016	0.13
4648	Puerto Rico	Bachelor's Degree	2016	0.23
4649	Puerto Rico	Graduate degree	2016	0.07
4650	Puerto Rico	Not a high school graduate	2017	0.10
4651	Puerto Rico	High school diploma or GED	2017	0.45
4652	Puerto Rico	Associate's Degree	2017	0.13
4653	Puerto Rico	Bachelor's Degree	2017	0.25
4654	Puerto Rico	Graduate degree	2017	0.07

4655 rows × 4 columns

Merging the Household_structure and edu_pop_25_to_34_new datasets for years 2000, 2005, 2010 and 2015

In [22]:
```
## Selecting four years - 2000, 2005, 2010 and 2015 from the Household
_structure and edu_pop_25_to_34_new
## datasets and then combining the two datasets for the year and makin
g a pivot table

# Selecting the years for the Household_structure data
House_2000 = Household_structure.loc[Household_structure['TimeFrame']
== 2000].drop(['TimeFrame'], axis = 1)
House_2005 = Household_structure.loc[Household_structure['TimeFrame']
== 2005].drop(['TimeFrame'], axis = 1)
House_2010 = Household_structure.loc[Household_structure['TimeFrame']
== 2010].drop(['TimeFrame'], axis = 1)
House_2015 = Household_structure.loc[Household_structure['TimeFrame']
== 2015].drop(['TimeFrame'], axis = 1)

# Making a pivot table for the Household_structure datatsets
H_Pivot_2000 = House_2000.pivot_table(index = 'Location', columns = 'H
ousehold Type')
H_Pivot_2005 = House_2005.pivot_table(index = 'Location', columns = 'H
ousehold Type')
H_Pivot_2010 = House_2010.pivot_table(index = 'Location', columns = 'H
ousehold Type')
H_Pivot_2015 = House_2015.pivot_table(index = 'Location', columns = 'H
ousehold Type')

# Selecting the years for the edu_pop_25_to_34_new data
Edu_2000 = edu_pop_25_to_34_new.loc[edu_pop_25_to_34_new['TimeFrame']
== 2000].drop(['TimeFrame'], axis = 1)
Edu_2005 = edu_pop_25_to_34_new.loc[edu_pop_25_to_34_new['TimeFrame']
== 2005].drop(['TimeFrame'], axis = 1)
Edu_2010 = edu_pop_25_to_34_new.loc[edu_pop_25_to_34_new['TimeFrame']
== 2010].drop(['TimeFrame'], axis = 1)
Edu_2015 = edu_pop_25_to_34_new.loc[edu_pop_25_to_34_new['TimeFrame']
== 2015].drop(['TimeFrame'], axis = 1)

# Making a pivot table for the edu_pop_25_to_34_new datasets
E_Pivot_2000 = Edu_2000.pivot_table(index = 'Location', columns = 'Edu
cation')
E_Pivot_2005 = Edu_2005.pivot_table(index = 'Location', columns = 'Edu
cation')
E_Pivot_2010 = Edu_2010.pivot_table(index = 'Location', columns = 'Edu
cation')
E_Pivot_2015 = Edu_2015.pivot_table(index = 'Location', columns = 'Edu
cation')
```

In [23]:
```
## Merging the Household_structure and edu_pop_25_to_34_new datasets f
or the four years
Merge_2000 = pd.merge(H_Pivot_2000,E_Pivot_2000, on = 'Location', how
= 'inner')
Merge_2005 = pd.merge(H_Pivot_2005,E_Pivot_2005, on = 'Location', how
= 'inner')
Merge_2010 = pd.merge(H_Pivot_2010,E_Pivot_2010, on = 'Location', how
= 'inner')
Merge_2015 = pd.merge(H_Pivot_2015,E_Pivot_2015, on = 'Location', how
= 'inner')
```

In [24]:
```
## Cleaning the pivot tables

# removing the data field on the top
Merge_2000.columns = Merge_2000.columns.droplevel(0)
Merge_2000 = Merge_2000.reset_index().rename_axis(None, axis = 1)

# renaming the columns
Merge_2000.columns = Merge_2000.columns.str.strip().str.lower().str.re
place(' ', '_').str.replace("'", "").str.replace("-","_")

# Repeating the cleaning process with the datasets from the other year
s
Merge_2005.columns = Merge_2005.columns.droplevel(0)
Merge_2005 = Merge_2005.reset_index().rename_axis(None, axis = 1)
Merge_2005.columns = Merge_2005.columns.str.strip().str.lower().str.re
place(' ', '_').str.replace("'", "").str.replace("-","_")

Merge_2010.columns = Merge_2010.columns.droplevel(0)
Merge_2010 = Merge_2010.reset_index().rename_axis(None, axis = 1)
Merge_2010.columns = Merge_2010.columns.str.strip().str.lower().str.re
place(' ', '_').str.replace("'", "").str.replace("-","_")

Merge_2015.columns = Merge_2015.columns.droplevel(0)
Merge_2015 = Merge_2015.reset_index().rename_axis(None, axis = 1)
Merge_2015.columns = Merge_2015.columns.str.strip().str.lower().str.re
place(' ', '_').str.replace("'", "").str.replace("-","_")
```

In [25]:
```
Merge_2000
```

Out[25]:

	location	father_only_households	married_couple_households	mother_only_households
0	Alabama	0.07	0.61	0.3
1	Alaska	0.07	0.69	0.2
2	Arizona	0.08	0.68	0.2
3	Arkansas	0.05	0.64	0.3

4	California	0.07	0.70	0.2
5	Colorado	0.05	0.75	0.1
6	Connecticut	0.05	0.74	0.2
7	Delaware	0.06	0.66	0.2
8	District of Columbia	0.06	0.34	0.6
9	Florida	0.07	0.63	0.2
10	Georgia	0.06	0.63	0.3
11	Hawaii	0.05	0.71	0.2
12	Idaho	0.06	0.80	0.1
13	Illinois	0.06	0.71	0.2
14	Indiana	0.06	0.70	0.2
15	Iowa	0.06	0.76	0.1
16	Kansas	0.06	0.74	0.1
17	Kentucky	0.06	0.71	0.2
18	Louisiana	0.06	0.60	0.3
19	Maine	0.08	0.72	0.1
20	Maryland	0.05	0.64	0.3
21	Massachusetts	0.04	0.72	0.2
22	Michigan	0.07	0.68	0.2
23	Minnesota	0.04	0.78	0.1
24	Mississippi	0.06	0.59	0.3
25	Missouri	0.08	0.67	0.2
26	Montana	0.08	0.77	0.1
27	Nebraska	0.04	0.74	0.2
28	Nevada	0.09	0.67	0.2
29	New Hampshire	0.09	0.72	0.1
30	New Jersey	0.05	0.75	0.1
31	New Mexico	0.06	0.68	0.2
32	New York	0.07	0.66	0.2
33	North Carolina	0.07	0.64	0.2
34	North Dakota	0.06	0.73	0.2

35	Ohio	0.06	0.71	0.22
36	Oklahoma	0.08	0.68	0.23
37	Oregon	0.09	0.67	0.23
38	Pennsylvania	0.05	0.70	0.23
39	Rhode Island	0.04	0.65	0.31
40	South Carolina	0.05	0.64	0.30
41	South Dakota	0.06	0.77	0.16
42	Tennessee	0.05	0.66	0.27
43	Texas	0.06	0.71	0.23
44	Utah	0.06	0.76	0.18
45	Vermont	0.05	0.71	0.23
46	Virginia	0.06	0.69	0.24
47	Washington	0.06	0.75	0.19
48	West Virginia	0.04	0.72	0.23
49	Wisconsin	0.06	0.73	0.19
50	Wyoming	0.10	0.75	0.14

In [26]:
```
Merge_2005
```

Out[26]:

	location	father_only_households	married_couple_households	mother_only_households
0	Alabama	0.06	0.64	0.30
1	Alaska	0.09	0.70	0.21
2	Arizona	0.08	0.68	0.24
3	Arkansas	0.07	0.66	0.27
4	California	0.08	0.71	0.22
5	Colorado	0.07	0.73	0.19
6	Connecticut	0.05	0.71	0.24
7	Delaware	0.07	0.66	0.26
8	District of Columbia	0.08	0.36	0.56
9	Florida	0.07	0.65	0.28
10	Georgia	0.06	0.65	0.28

11	Hawaii	0.06	0.74	0.20
12	Idaho	0.06	0.77	0.17
13	Illinois	0.06	0.70	0.24
14	Indiana	0.07	0.70	0.23
15	Iowa	0.06	0.74	0.20
16	Kansas	0.06	0.73	0.21
17	Kentucky	0.06	0.70	0.24
18	Louisiana	0.06	0.59	0.36
19	Maine	0.10	0.69	0.21
20	Maryland	0.06	0.68	0.25
21	Massachusetts	0.06	0.71	0.23
22	Michigan	0.07	0.69	0.24
23	Minnesota	0.07	0.75	0.19
24	Mississippi	0.07	0.55	0.38
25	Missouri	0.07	0.68	0.25
26	Montana	0.07	0.72	0.21
27	Nebraska	0.05	0.75	0.20
28	Nevada	0.08	0.69	0.23
29	New Hampshire	0.06	0.76	0.18
30	New Jersey	0.06	0.72	0.22
31	New Mexico	0.08	0.64	0.28
32	New York	0.07	0.66	0.28
33	North Carolina	0.07	0.66	0.27
34	North Dakota	0.06	0.77	0.17
35	Ohio	0.07	0.68	0.25
36	Oklahoma	0.07	0.68	0.25
37	Oregon	0.07	0.71	0.22
38	Pennsylvania	0.06	0.69	0.25
39	Puerto Rico	0.06	0.57	0.37
40	Rhode Island	0.06	0.67	0.27
41	South Carolina	0.06	0.63	0.31

42	South Dakota	0.07	0.72	0.21
43	Tennessee	0.07	0.66	0.27
44	Texas	0.06	0.69	0.24
45	Utah	0.05	0.83	0.12
46	Vermont	0.09	0.69	0.22
47	Virginia	0.06	0.71	0.23
48	Washington	0.07	0.72	0.21
49	West Virginia	0.07	0.71	0.22
50	Wisconsin	0.07	0.72	0.21
51	Wyoming	0.09	0.75	0.16

In [27]:
```
Merge_2010
```

Out[27]:

	location	father_only_households	married_couple_households	mother_only_households
0	Alabama	0.07	0.61	0.32
1	Alaska	0.09	0.69	0.21
2	Arizona	0.09	0.64	0.26
3	Arkansas	0.07	0.62	0.30
4	California	0.08	0.68	0.23
5	Colorado	0.08	0.70	0.22
6	Connecticut	0.06	0.68	0.25
7	Delaware	0.08	0.64	0.27
8	District of Columbia	0.08	0.41	0.50
9	Florida	0.08	0.61	0.30
10	Georgia	0.07	0.63	0.30
11	Hawaii	0.07	0.71	0.22
12	Idaho	0.06	0.75	0.19
13	Illinois	0.07	0.68	0.25
14	Indiana	0.07	0.66	0.26
15	Iowa	0.07	0.71	0.21
16	Kansas	0.07	0.70	0.22
17	Kentucky	0.07	0.65	0.26

18	Louisiana	0.07	0.56	0.36
19	Maine	0.08	0.66	0.25
20	Maryland	0.07	0.65	0.27
21	Massachusetts	0.06	0.69	0.25
22	Michigan	0.07	0.66	0.26
23	Minnesota	0.07	0.72	0.20
24	Mississippi	0.07	0.56	0.37
25	Missouri	0.07	0.66	0.25
26	Montana	0.07	0.71	0.21
27	Nebraska	0.07	0.71	0.20
28	Nevada	0.11	0.64	0.24
29	New Hampshire	0.08	0.73	0.18
30	New Jersey	0.06	0.71	0.22
31	New Mexico	0.11	0.59	0.29
32	New York	0.07	0.65	0.27
33	North Carolina	0.07	0.64	0.28
34	North Dakota	0.06	0.74	0.19
35	Ohio	0.07	0.65	0.27
36	Oklahoma	0.07	0.66	0.26
37	Oregon	0.08	0.68	0.23
38	Pennsylvania	0.07	0.66	0.26
39	Puerto Rico	0.09	0.48	0.43
40	Rhode Island	0.06	0.64	0.29
41	South Carolina	0.07	0.60	0.33
42	South Dakota	0.07	0.71	0.22
43	Tennessee	0.07	0.63	0.29
44	Texas	0.07	0.66	0.27
45	Utah	0.05	0.82	0.13
46	Vermont	0.08	0.69	0.21
47	Virginia	0.07	0.69	0.24
48	Washington	0.07	0.71	0.21

49	West Virginia	0.08	0.68	0.23
50	Wisconsin	0.08	0.68	0.23
51	Wyoming	0.06	0.75	0.18

In [28]:
```
Merge_2015
```

Out[28]:

	location	father_only_households	married_couple_households	mother_only_households
0	Alabama	0.06	0.61	0.32
1	Alaska	0.11	0.67	0.21
2	Arizona	0.10	0.63	0.27
3	Arkansas	0.08	0.64	0.27
4	California	0.09	0.68	0.23
5	Colorado	0.08	0.73	0.19
6	Connecticut	0.06	0.68	0.25
7	Delaware	0.07	0.61	0.31
8	District of Columbia	0.07	0.47	0.45
9	Florida	0.09	0.60	0.30
10	Georgia	0.07	0.62	0.30
11	Hawaii	0.08	0.71	0.20
12	Idaho	0.07	0.74	0.18
13	Illinois	0.07	0.67	0.25
14	Indiana	0.09	0.65	0.25
15	Iowa	0.09	0.70	0.21
16	Kansas	0.08	0.70	0.21
17	Kentucky	0.08	0.65	0.26
18	Louisiana	0.08	0.56	0.35
19	Maine	0.11	0.65	0.24
20	Maryland	0.08	0.66	0.26
21	Massachusetts	0.06	0.68	0.25
22	Michigan	0.08	0.65	0.26
23	Minnesota	0.08	0.71	0.20
24	Mississippi	0.08	0.54	0.38

25	Missouri	0.08	0.66	0.25
26	Montana	0.08	0.72	0.19
27	Nebraska	0.07	0.71	0.21
28	Nevada	0.10	0.61	0.28
29	New Hampshire	0.07	0.70	0.22
30	New Jersey	0.06	0.71	0.22
31	New Mexico	0.11	0.60	0.29
32	New York	0.08	0.64	0.28
33	North Carolina	0.08	0.64	0.28
34	North Dakota	0.08	0.73	0.18
35	Ohio	0.08	0.64	0.27
36	Oklahoma	0.09	0.65	0.25
37	Oregon	0.08	0.69	0.22
38	Pennsylvania	0.08	0.65	0.26
39	Puerto Rico	0.10	0.44	0.46
40	Rhode Island	0.07	0.61	0.31
41	South Carolina	0.07	0.60	0.32
42	South Dakota	0.09	0.68	0.22
43	Tennessee	0.08	0.63	0.28
44	Texas	0.07	0.66	0.26
45	Utah	0.05	0.81	0.14
46	Vermont	0.08	0.72	0.19
47	Virginia	0.07	0.69	0.23
48	Washington	0.08	0.70	0.20
49	West Virginia	0.10	0.63	0.26
50	Wisconsin	0.09	0.68	0.23
51	Wyoming	0.11	0.72	0.16

Limitations with the Data

Given that our investigating question sought to explore effects over a sustained time period, our data is not a perfect fit for our topic. However, given the sensitivity of child-related data on an individual level, we decided that exploring our question on a more aggregate level (state-specific) would be an effective way to test our hypothesis.

We also face challenges regarding inconsistencies in the data. Namely, that the data over the 17 year time-span captures the 50 most populous cities (and Puerto Rico and District of Columbia) of any given year, meaning that there is the possibility that different cities are accounted for in different years. Because the state-level data is an average, the percentages within a state do not always add up to 100%.

Regression Analysis

In [29]:

```
## Running a regression for Merge_2000
print(smf.ols("high_school_diploma_or_ged~father_only_households+marri
ed_couple_households+mother_only_households+neither_parent_households"
, data = Merge_2000).fit().summary())
```

```
                                OLS Regression Results
=====================================================================
==================
Dep. Variable:     high_school_diploma_or_ged   R-squared:
0.209
Model:                                    OLS   Adj. R-squared:
0.140
Method:                         Least Squares   F-statistic:
3.031
Date:                        Fri, 21 Dec 2018   Prob (F-statistic):
0.0267
Time:                                01:08:53   Log-Likelihood:
80.140
No. Observations:                          51   AIC:
-150.3
Df Residuals:                              46   BIC:
-140.6
Df Model:                                   4
Covariance Type:                    nonrobust
=====================================================================
=========================
                                coef    std err          t      P>|t
|      [0.025      0.975]
---------------------------------------------------------------------
-------------------------
```

```
Intercept                       3.2269      1.133      2.849      0.00
7        0.947       5.507
father_only_households         -2.5302      1.303     -1.942      0.05
8       -5.153       0.093
married_couple_households      -2.7067      1.137     -2.382      0.02
1       -4.994      -0.419
mother_only_households         -2.9386      1.109     -2.651      0.01
1       -5.170      -0.707
neither_parent_households       0.1689      0.740      0.228      0.82
1       -1.322       1.659
====================================================================
==========
Omnibus:                        0.609   Durbin-Watson:
2.178
Prob(Omnibus):                  0.738   Jarque-Bera (JB):
0.143
Skew:                           0.079   Prob(JB):
0.931
Kurtosis:                       3.206   Cond. No.
385.
====================================================================
==========

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors
is correctly specified.
```

From the Merge_2000 dataset, we see that the R-squared is very low, at only 20.9%. It is interesting too how the only statistically significant relationships seems to be with married and mother-only households, and those relationships are negative. Strictly looking at coefficients, the only positive coefficient belongs to the neither-parent households. This contradicts our hypothesis the most, as it indicates the least stable family structure.

In [30]:

```
## Running a regression for Merge_2005
print(smf.ols("high_school_diploma_or_ged~father_only_households+marri
ed_couple_households+mother_only_households+neither_parent_households"
, data = Merge_2005).fit().summary())
```

```
                                OLS Regression Results
====================================================================
==================
Dep. Variable:     high_school_diploma_or_ged   R-squared:
0.442
Model:                                    OLS   Adj. R-squared:
0.394
Method:                         Least Squares   F-statistic:
9.303
Date:                        Fri, 21 Dec 2018   Prob (F-statistic):
```

```
1.27e-05
Time:                            01:08:56   Log-Likelihood:
90.143
No. Observations:                      52   AIC:
-170.3
Df Residuals:                          47   BIC:
-160.5
Df Model:                               4
Covariance Type:                nonrobust
=====================================================================
=========================
                              coef    std err          t      P>|t
|      [0.025      0.975]
---------------------------------------------------------------------
-------------------------
Intercept                   0.2241      1.514      0.148      0.88
3      -2.821       3.270
father_only_households      1.0240      1.563      0.655      0.51
6      -2.121       4.169
married_couple_households   0.2226      1.519      0.146      0.88
4      -2.834       3.279
mother_only_households     -0.4674      1.513     -0.309      0.75
9      -3.511       2.576
neither_parent_households   2.8098      0.612      4.595      0.00
0       1.580       4.040
=====================================================================
==========
Omnibus:                         0.439   Durbin-Watson:
1.883
Prob(Omnibus):                   0.803   Jarque-Bera (JB):
0.583
Skew:                           -0.047   Prob(JB):
0.747
Kurtosis:                        2.490   Cond. No.
598.
=====================================================================
==========

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors
is correctly specified.
```

The Merge_2005 data has vastly more explaining power, with an R-squared of 44.2%; however none of the variables appear statistically significant beyond neither-parent households (which is again a positive coefficient). It appears that in the span of 5 years, this coefficient has notably strengthened.

In [31]:
```
## Running a regression for Merge_2010
print(smf.ols("high_school_diploma_or_ged~father_only_households+marri
ed_couple_households+mother_only_households+neither_parent_households"
, data = Merge_2010).fit().summary())
```

```
                            OLS Regression Results
======================================================================
==================
Dep. Variable:     high_school_diploma_or_ged   R-squared:
0.475
Model:                                    OLS   Adj. R-squared:
0.431
Method:                         Least Squares   F-statistic:
10.65
Date:                        Fri, 21 Dec 2018   Prob (F-statistic):
3.17e-06
Time:                                01:08:58   Log-Likelihood:
87.535
No. Observations:                          52   AIC:
-165.1
Df Residuals:                              47   BIC:
-155.3
Df Model:                                   4
Covariance Type:                    nonrobust
======================================================================
=========================
                                coef    std err          t      P>|t
|      [0.025      0.975]
----------------------------------------------------------------------
-------------------------
Intercept                    -0.7389      1.219     -0.606      0.54
7      -3.191       1.713
father_only_households        2.2580      1.485      1.521      0.13
5      -0.729       5.245
married_couple_households     1.1865      1.223      0.971      0.33
7      -1.273       3.646
mother_only_households        0.4154      1.202      0.346      0.73
1      -2.002       2.833
neither_parent_households     3.5935      0.710      5.062      0.00
0       2.165       5.022
======================================================================
==========
Omnibus:                        5.546   Durbin-Watson:
2.048
Prob(Omnibus):                  0.062   Jarque-Bera (JB):
4.455
Skew:                          -0.655   Prob(JB):
0.108
Kurtosis:                       3.585   Cond. No.
```

```
473.
==============================================================================
==========

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors
is correctly specified.
```

In the Merge_2010 data, the R-squared is similarly high at 47.5%, and once again neither-parents is the only significant coefficient. It has grown since 2005 and is still positive.

In [32]:
```
## Running a regression for Merge_2015
print(smf.ols("high_school_diploma_or_ged~father_only_households+marri
ed_couple_households+mother_only_households+neither_parent_households"
, data = Merge_2015).fit().summary())
```

```
                                OLS Regression Results
==============================================================================
==================
Dep. Variable:     high_school_diploma_or_ged   R-squared:
0.306
Model:                                    OLS   Adj. R-squared:
0.247
Method:                         Least Squares   F-statistic:
5.190
Date:                        Fri, 21 Dec 2018   Prob (F-statistic):
0.00152
Time:                                01:09:01   Log-Likelihood:
81.355
No. Observations:                          52   AIC:
-152.7
Df Residuals:                              47   BIC:
-143.0
Df Model:                                   4
Covariance Type:                    nonrobust
==============================================================================
=========================
                                coef    std err          t      P>|t
|      [0.025      0.975]
------------------------------------------------------------------------------
-------------------------
Intercept                     0.8109      1.715      0.473      0.63
9      -2.640       4.261
father_only_households        0.8440      1.634      0.517      0.60
8      -2.443       4.131
married_couple_households    -0.4665      1.742     -0.268      0.79
0      -3.971       3.038
mother_only_households       -0.7891      1.714     -0.460      0.64
```

```
7      -4.236       2.658
neither_parent_households     2.0845      0.755      2.761      0.00
8       0.566       3.603
==============================================================================
==========
Omnibus:                       34.670   Durbin-Watson:
2.163
Prob(Omnibus):                  0.000   Jarque-Bera (JB):
114.100
Skew:                          -1.735   Prob(JB):
1.67e-25
Kurtosis:                       9.374   Cond. No.
560.
==============================================================================
==========

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors
is correctly specified.
```

The Merge_2015 dataset shows a decrease in the R-squared to 30.6%. Neither-parent household is still a positive, significant coefficient.

Determining the Most and Least Educated States

In [33]:
```
## grouping the edu_pop_25_to_34_new data by the education type for al
l the years from 2000 to 2017 and
## then taking the average percentage of education level in different
states over the years.

# grouping the edu_pop_25_to_34_new data by not a high school graduate
education level and taking the average
no_highschool_avg = edu_pop_25_to_34_new.loc[(edu_pop_25_to_34_new['Ed
ucation'].str.contains('Not a high school graduate'))]
no_highschool_avg = no_highschool_avg.groupby('Location').agg({'TimeFr
ame':np.mean, 'Data':np.mean})

# repeating the grouping part for other education levels
high_school_avg = edu_pop_25_to_34_new.loc[(edu_pop_25_to_34_new['Educ
ation'].str.contains('High school diploma or GED'))]
high_school_avg = high_school_avg.groupby('Location').agg({'TimeFrame'
:np.mean, 'Data':np.mean})

associate_avg = edu_pop_25_to_34_new.loc[(edu_pop_25_to_34_new['Educat
ion'].str.contains("Associate's Degree"))]
associate_avg = associate_avg.groupby('Location').agg({'TimeFrame':np.
mean, 'Data':np.mean})

bachelor_avg = edu_pop_25_to_34_new.loc[(edu_pop_25_to_34_new['Educati
on'].str.contains("Bachelor's Degree"))]
bachelor_avg = bachelor_avg.groupby('Location').agg({'TimeFrame':np.me
an, 'Data':np.mean})

grad_avg = edu_pop_25_to_34_new.loc[(edu_pop_25_to_34_new['Education']
.str.contains('Graduate degree'))]
grad_avg = grad_avg.groupby('Location').agg({'TimeFrame':np.mean, 'Dat
a':np.mean})
```

In [34]:
```
## sorting the data from highest to lowest and finding the state with
the highest and the lowest percentage of
## adults who are not high school graduates
no_highschool_avg = no_highschool_avg.sort_values('Data', ascending =
False)
print(no_highschool_avg.head(1)) ## gives us Texas
print(no_highschool_avg.tail(1)) ## gives us North Dakota

## repeating the sorting part for other education levels
high_school_avg = high_school_avg.sort_values('Data', ascending = Fals
e)
print(high_school_avg.head(1)) ## gives us Alaska
print(high_school_avg.tail(1)) ## gives us District of Columbia

associate_avg = associate_avg.sort_values('Data', ascending = False)
print(associate_avg.head(1)) ## gives us North Dakota
print(associate_avg.tail(1)) ## gives us District of Columbia

bachelor_avg = bachelor_avg.sort_values('Data', ascending = False)
print(bachelor_avg.head(1)) ## gives us District of Columbia
print(bachelor_avg.tail(1)) ## gives us New Mexico

grad_avg = grad_avg.sort_values('Data', ascending = False)
print(grad_avg.head(1)) ## gives us District of Columbia
print(grad_avg.tail(1)) ## gives us Nevada
```

```
          TimeFrame      Data
Location
Texas        2008.5  0.178333
              TimeFrame      Data
Location
North Dakota     2008.5  0.043889
          TimeFrame      Data
Location
Alaska       2008.5  0.603889
                      TimeFrame      Data
Location
District of Columbia     2008.5  0.258889
              TimeFrame      Data
Location
North Dakota     2008.5  0.148889
                      TimeFrame      Data
Location
District of Columbia     2008.5  0.024444
                      TimeFrame      Data
Location
District of Columbia     2008.5  0.328889
            TimeFrame      Data
Location
New Mexico     2008.5  0.151111
                      TimeFrame      Data
Location
District of Columbia     2008.5  0.304444
          TimeFrame      Data
Location
Nevada       2008.5  0.049444
```

State-Specific Analyses

We sought to look more closely at states that seemed to be the top/bottom of each educational attainment standard (Not high school graduate, high school diploma, Associate's, Bachelor's, Graduate) To do this, we averaged accross the entire time period (2000-2017) and averaged the data points related to each educational attainment level on a state level.

In [35]:
```
## Showing the states with the highest and the lowest percentage of di
fferent education levels

select_state_degree = pd.DataFrame({ "Education Level": ["No_highschoo
l", "High_school_diploma_or_GED","Associates", "Bachelors", "Graduate"
],
                            "Highest_percentage_state": ['Texas', 'Ala
ska', 'North Dakota', 'District of Columbia', 'District of Columbia'],
                            "Lowest_percentage_state": ['North Dakota'
, "District of Columbia", "District of Columbia", "New Mexico", "Nevad
a"]})
select_state_degree
```

Out[35]:

	Education Level	Highest_percentage_state	Lowest_percentage_state
0	No_highschool	Texas	North Dakota
1	High_school_diploma_or_GED	Alaska	District of Columbia
2	Associates	North Dakota	District of Columbia
3	Bachelors	District of Columbia	New Mexico
4	Graduate	District of Columbia	Nevada

As we can see with the table above, our hypothesis is essentially disproven, as the District of Columbia has the highest numbers of Bachelors/Graduate degrees and one of the lowest percentages of married-family households.
We believed that the stability of married-family househoulds would have a significant effect on influencing children to attain higher education. However, it appears that state-related factors (potentially region-specific occupations, religion, and other ommitted variables) must be responsibly for the distribution of educational attainment levels.

One important note is that our initial investigation of using High-School Diploma as a proxy for educational attainment was not effective for several reasons. Firstly, High-school Diploma did not include the percentage of people who attained education beyond a high-school diploma. A better proxy would have been the No High-school Diploma group, as this was an inclusive population of indviduals who did not finish high-school or anything beyond it.

In [41]:
```
## Selecting the states of Alaska, District of Columbia, Nevada, New M
exico, North Dakota and Texas
## and finding out the average household structure of those states ove
r the years

Alaska_edu = Household_structure.loc[(Household_structure['Location'].
str.contains('Alaska'))]
Alaska_edu = Alaska_edu.groupby('Household Type').agg({'TimeFrame':np.
mean, 'Data':np.mean})

Columbia_edu = Household_structure.loc[(Household_structure['Location'
].str.contains('District of Columbia'))]
Columbia_edu = Columbia_edu.groupby('Household Type').agg({'TimeFrame'
:np.mean, 'Data':np.mean})

New_Mexico_edu = Household_structure.loc[(Household_structure['Locatio
n'].str.contains('New Mexico'))]
New_Mexico_edu = New_Mexico_edu.groupby('Household Type').agg({'TimeFr
ame':np.mean, 'Data':np.mean})

Nevada_edu = Household_structure.loc[(Household_structure['Location'].
str.contains('Nevada'))]
Nevada_edu = Nevada_edu.groupby('Household Type').agg({'TimeFrame':np.
mean, 'Data':np.mean})

North_Dakota_edu = Household_structure.loc[(Household_structure['Locat
ion'].str.contains('North Dakota'))]
North_Dakota_edu = North_Dakota_edu.groupby('Household Type').agg({'Ti
meFrame':np.mean, 'Data':np.mean})

Texas_edu = Household_structure.loc[(Household_structure['Location'].s
tr.contains('Texas'))]
Texas_edu = Texas_edu.groupby('Household Type').agg({'TimeFrame':np.me
an, 'Data':np.mean})
```

In [42]:
```
## Showing the household structure of the different states

# Merging the state_edu datasets for different states to be combined
merge1 = pd.merge(Alaska_edu, Columbia_edu, on ='Household Type', how
= 'inner')
merge2 = pd.merge(New_Mexico_edu, Nevada_edu,on ='Household Type', how
= 'inner' )
merge3 = pd.merge(North_Dakota_edu, Texas_edu,on ='Household Type', ho
w = 'inner')
merge4 = pd.merge(merge1,merge2, on ='Household Type', how = 'inner')
select_state_household = pd.merge(merge4,merge3, on ='Household Type',
how = 'inner')

# Cleaning the select_state_household dataset
select_state_household = select_state_household.drop(columns = ['TimeF
rame_x_x', 'TimeFrame_y_x', 'TimeFrame_x_y', 'TimeFrame_y_y', 'TimeFra
me_x', 'TimeFrame_y'], axis = 1)
select_state_household = select_state_household.rename(index = str, co
lumns = {"Data_x_x": "Alaska", "Data_y_x": "Columbia", "Data_x_y": "Ne
w Mexico", "Data_y_y": "Nevada", "Data_x": "North Dakota", "Data_y": "
Texas" })
select_state_household = select_state_household.transpose()
select_state_household
```

Out[42]:

Household Type	Father only Households	Married-couple Households	Mother only Households	Neither parent Households
Alaska	0.091667	0.692222	0.206111	0.058235
Columbia	0.077222	0.401111	0.515000	0.085882
New Mexico	0.093889	0.615556	0.285556	0.062941
Nevada	0.091111	0.655000	0.244444	0.058235
North Dakota	0.069444	0.742778	0.178333	0.040588
Texas	0.062778	0.678333	0.251111	0.055882

In [43]:
```
## Making a dataset for Alaska and District of Columbia to understand the household structure trends in the two states
## for years from 2000 to 2017. Alaska and District of Columbia have been specifically chosen because the states
## have the highest and the lowest percentage of high school graduates respectively

# Dataset for Alaska
Alaska_household = Household_structure.loc[Household_structure['Location'].str.contains('Alaska')]

# dropping location and making a pivot table
Alaska_household = Alaska_household.drop(['Location'], axis = 1)
Alaska_household = Alaska_household.pivot_table(index = 'TimeFrame', columns = 'Household Type')

# removing the data field on the top
Alaska_household.columns = Alaska_household.columns.droplevel(0)
Alaska_household = Alaska_household.reset_index().rename_axis(None, axis = 1)

# renaming the columns
Alaska_household.columns = Alaska_household.columns.str.strip().str.lower().str.replace(' ', '_').str.replace("-","_")

#assigning the year as index
Alaska_household = Alaska_household.set_index(['timeframe'])

# Repeating the process for District of Columbia
Columbia_household = Household_structure.loc[Household_structure['Location'].str.contains('District of Columbia')]
Columbia_household = Columbia_household.drop(['Location'], axis = 1)
Columbia_household = Columbia_household.pivot_table(index = 'TimeFrame', columns = 'Household Type')
Columbia_household.columns = Columbia_household.columns.droplevel(0)
Columbia_household = Columbia_household.reset_index().rename_axis(None, axis = 1)
Columbia_household.columns = Columbia_household.columns.str.strip().str.lower().str.replace(' ', '_').str.replace("-","_")
Columbia_household = Columbia_household.set_index(['timeframe'])
```

In [44]:
```
## Making a dataset for Alaska and District of Columbia to understand the education level trends in the two states
## for years from 2000 to 2017. Alaska and District of Columbia have been specifically chosen because the states
## have the highest and the lowest percentage of high school graduates respectively

## Making a dataset for Alaska and cleaning it
Alaska_education = edu_pop_25_to_34_new.loc[edu_pop_25_to_34_new['Location'].str.contains('Alaska')]

# dropping location and making a pivot table
Alaska_education = Alaska_education.drop(['Location'], axis = 1)
Alaska_education = Alaska_education.pivot_table(index = 'TimeFrame', columns = 'Education')

# reseting the index
Alaska_education.columns = Alaska_education.columns.droplevel(0)
Alaska_education = Alaska_education.reset_index().rename_axis(None, axis = 1)

# renaming the columns
Alaska_education.columns = Alaska_education.columns.str.strip().str.lower().str.replace(' ', '_').str.replace("-","_").str.replace("'", "")

# reseting the index
Alaska_education = Alaska_education.set_index(['timeframe'])

# Repeating the process for District of Columbia
Columbia_education = edu_pop_25_to_34_new.loc[edu_pop_25_to_34_new['Location'].str.contains('District of Columbia')]
Columbia_education = Columbia_education.drop(['Location'], axis = 1)
Columbia_education = Columbia_education.pivot_table(index = 'TimeFrame', columns = 'Education')
Columbia_education.columns = Columbia_education.columns.droplevel(0)
Columbia_education = Columbia_education.reset_index().rename_axis(None, axis = 1)
Columbia_education.columns = Columbia_education.columns.str.strip().str.lower().str.replace(' ', '_').str.replace("-","_").str.replace("'", "")
Columbia_education = Columbia_education.set_index(['timeframe'])
```

In [45]:
```
## Plotting graphs to understand the household structure and education level trends in the states of
## Alaska and District of Columbia

plt.style.use('seaborn')
fig, (ax, ax2) = plt.subplots(ncols=2, sharey=True)

Alaska_household.plot(ax = ax, figsize = (20,7))
ax.set_xlim(2000,2017)
ax.set_title('Alaska Household Structure', fontsize = 20)
ax.set_xlabel('Year', fontsize = 15)
ax.set_ylabel('Percentage of Households', fontsize = 15)

Columbia_household.plot(ax = ax2, figsize = (20,7))
ax2.set_xlim(2000,2017)
ax2.set_title('Columbia Household Structure', fontsize = 20)
ax2.set_xlabel('Year', fontsize = 15)
ax2.set_ylabel('Percentage of Households', fontsize = 15)

plt.style.use('seaborn')
fig, (ax, ax2) = plt.subplots(ncols=2, sharey=True)

Alaska_education.plot(ax = ax, figsize = (20,6))
ax.set_xlim(2000,2017)
ax.set_title('Alaska Education Levels', fontsize = 20)
ax.set_xlabel('Year', fontsize = 15)
ax.set_ylabel('Percentage of Population', fontsize = 15)

Columbia_education.plot(ax = ax2, figsize = (20,6))
ax2.set_xlim(2000,2017)
ax2.set_title('Columbia Education Levels', fontsize = 20)
ax2.set_xlabel('Year', fontsize = 15)
ax2.set_ylabel('Percentage of Population', fontsize = 15)
```

```
Out[45]: Text(0,0.5,'Percentage of Population')
```

One interesting insight from these plots are how in the District of Columbia, married couple households and mother only households have a strong inverse relationship, suggesting that shifts in household structure are largely related to these two structures-- when people aren't married, the children live with the mothers.

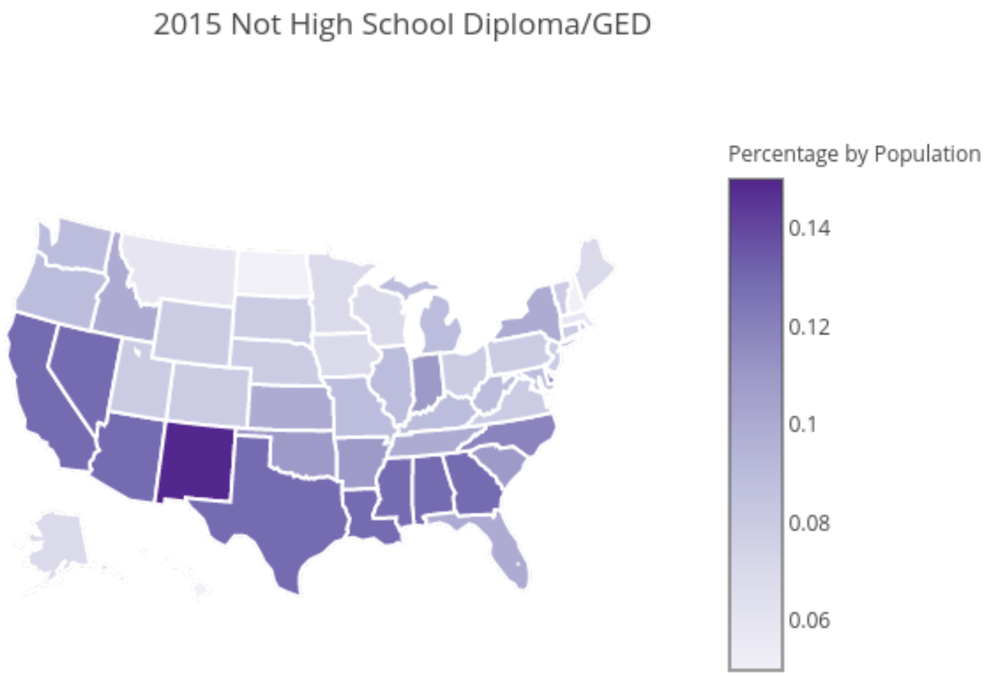
It is also interesting how Alaska has such high levels of Associate's degrees. We beleive this suggests that there is a region-specific need or educational opportunities are targetted at attaining this level of education. Similarly, the very low number of Associate's degrees in the District of Columbia suggest that this educational level is not very valuable for the area or possibly that opportunities to achieve this educational level are limited there.

Mapping Merge_2015 Data

As an additional tool to understanding the analyses, we also mapped the 2015 pivot table. The manipulated data allows us to have a general overview of the percentage of population within each category we looked at. It does not include Puerto Rico or District of Columbia since the map package Basemap and Plotly were limited to the 50 states. However, because our purpose for mapping is only to see the densities of our manipulated data, it has served its purpose.

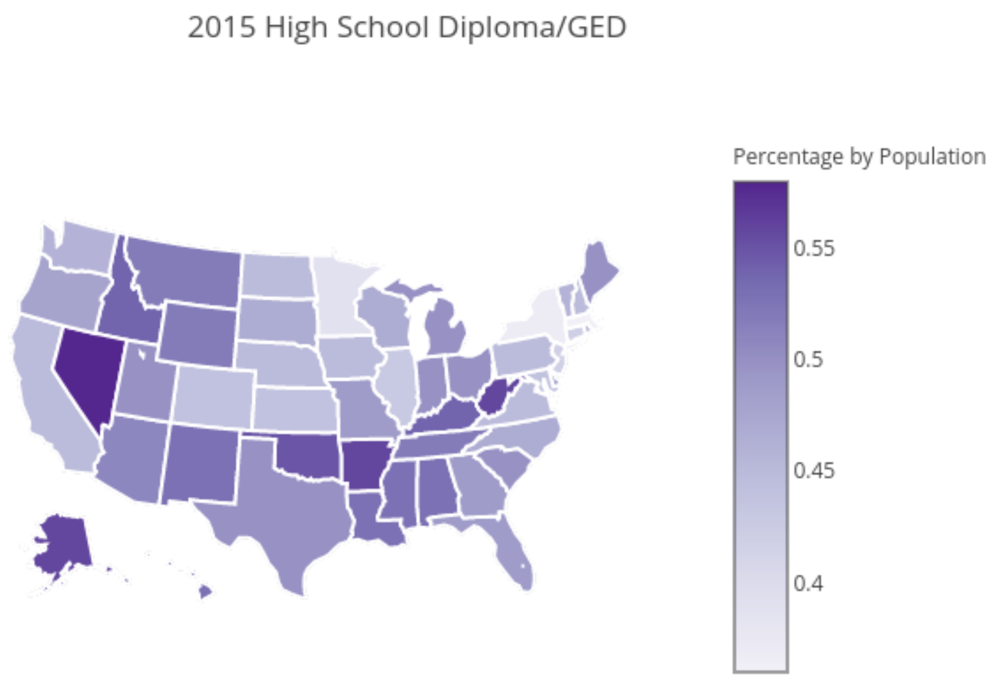
In [12]:
```
from IPython.display import Image
url = 'https://angelayxng.files.wordpress.com/2018/12/nothighschoolorged-1.png'
Image(url)
```

Out[12]:



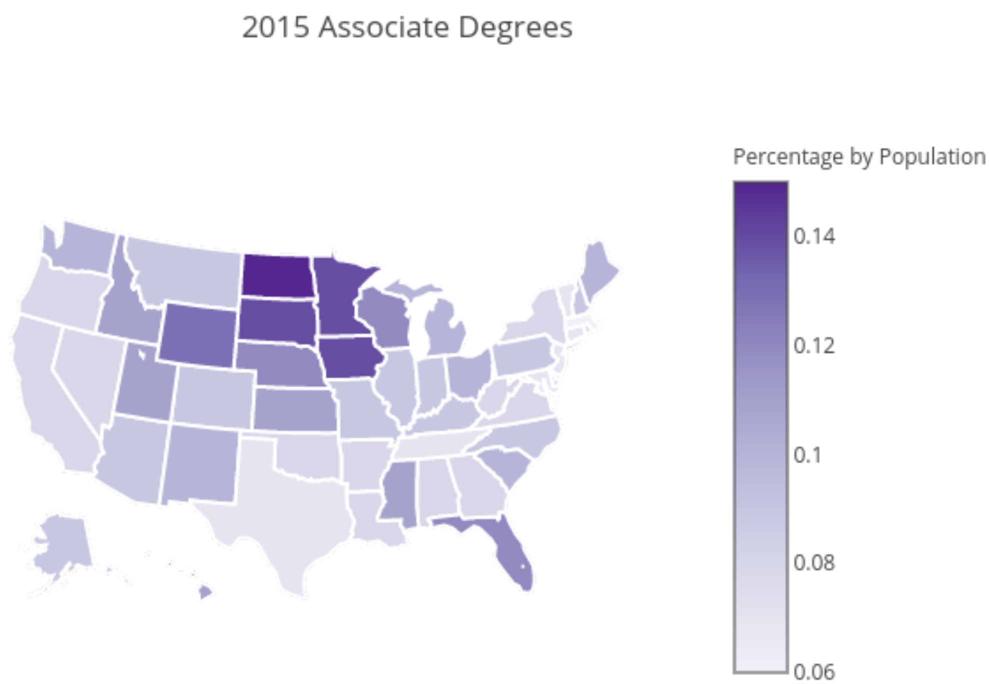
In [6]:
```
from IPython.display import Image
url = 'https://angelayxng.files.wordpress.com/2018/12/highschoolged.pn
g'
Image(url)
```

Out[6]:



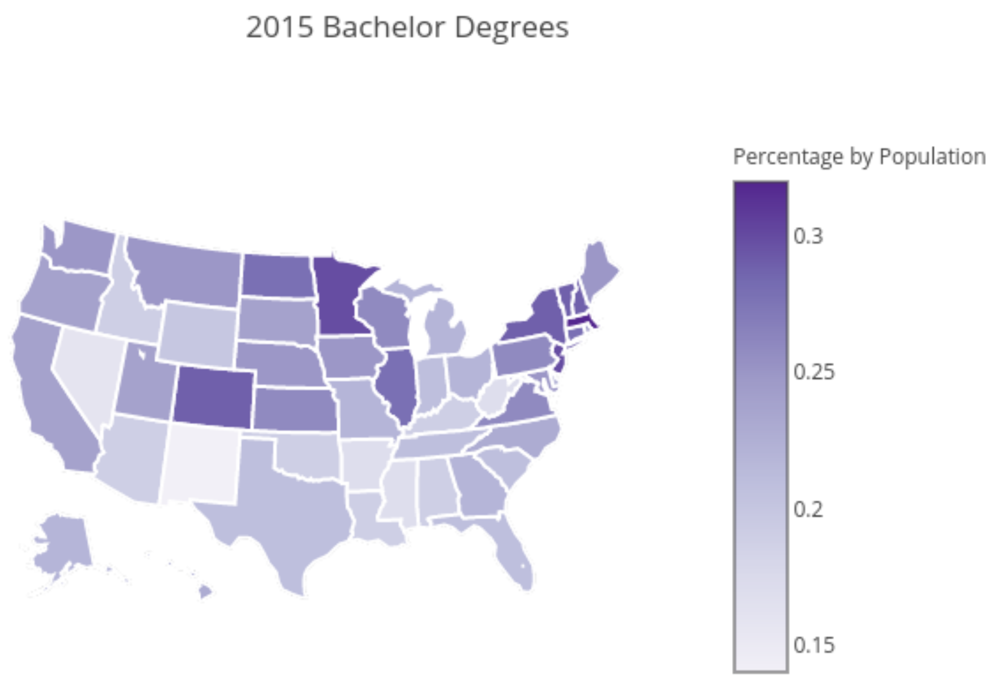
In [1]:
```
from IPython.display import Image
url = 'https://angelayxng.files.wordpress.com/2018/12/assoc-degree.png
'
Image(url)
```

Out[1]:



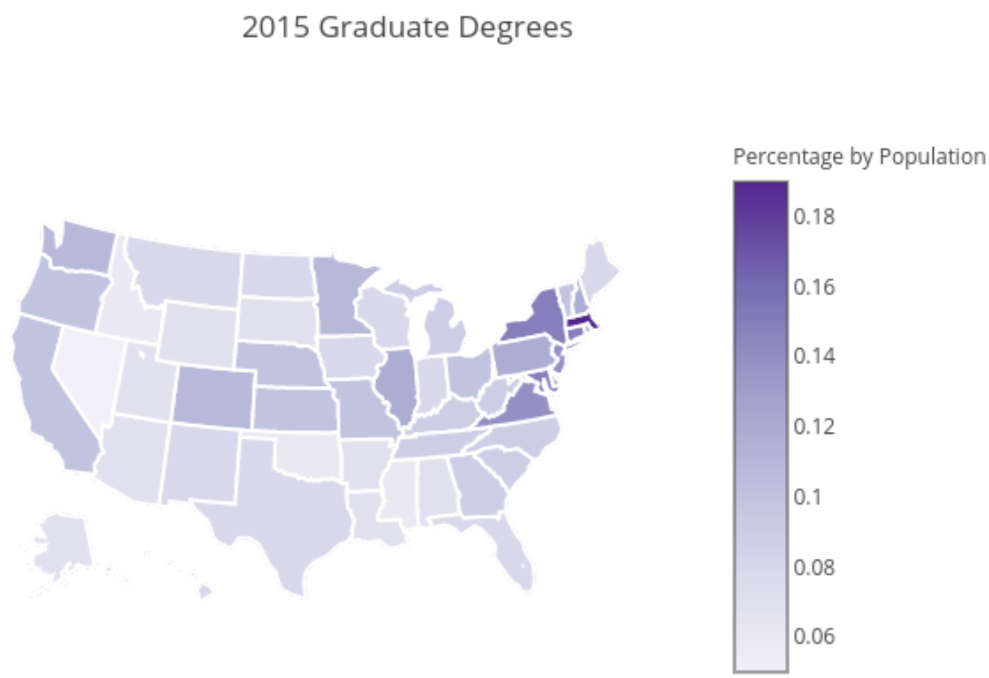
In [2]:
```
from IPython.display import Image
url = 'https://angelayxng.files.wordpress.com/2018/12/bachelor-degree.png'
Image(url)
```

Out[2]:



In [5]:
```
from IPython.display import Image
url = 'https://angelayxng.files.wordpress.com/2018/12/graduate-degree.
png'
Image(url)
```

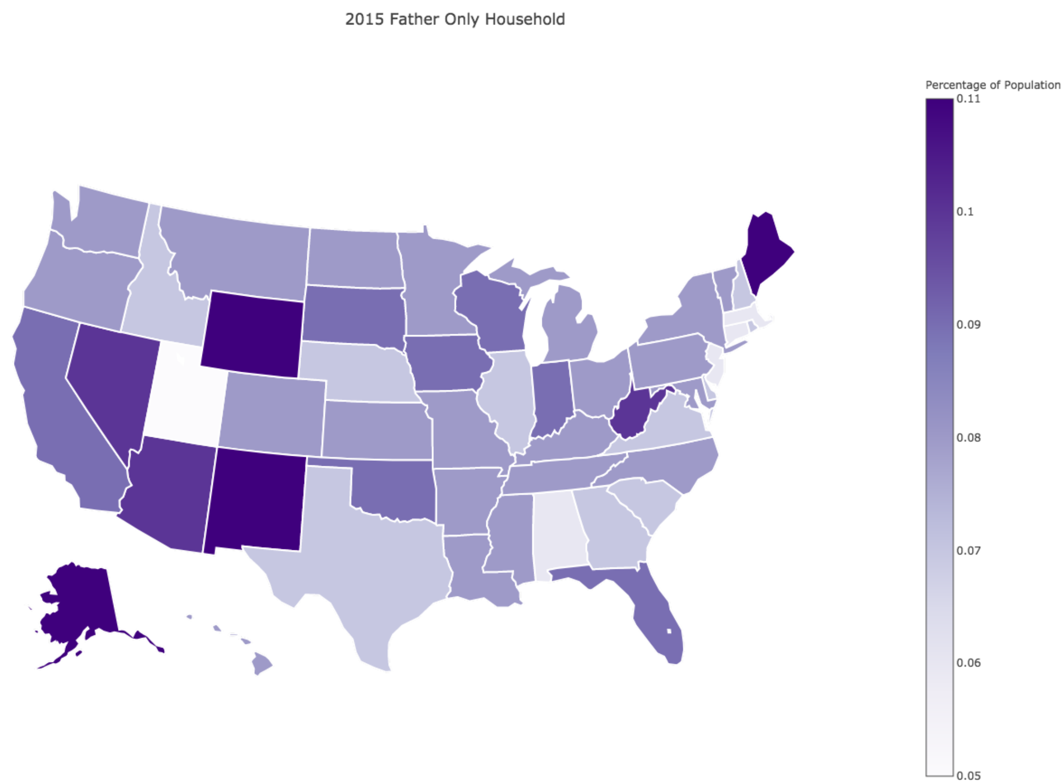
Out[5]:



In [ ]:

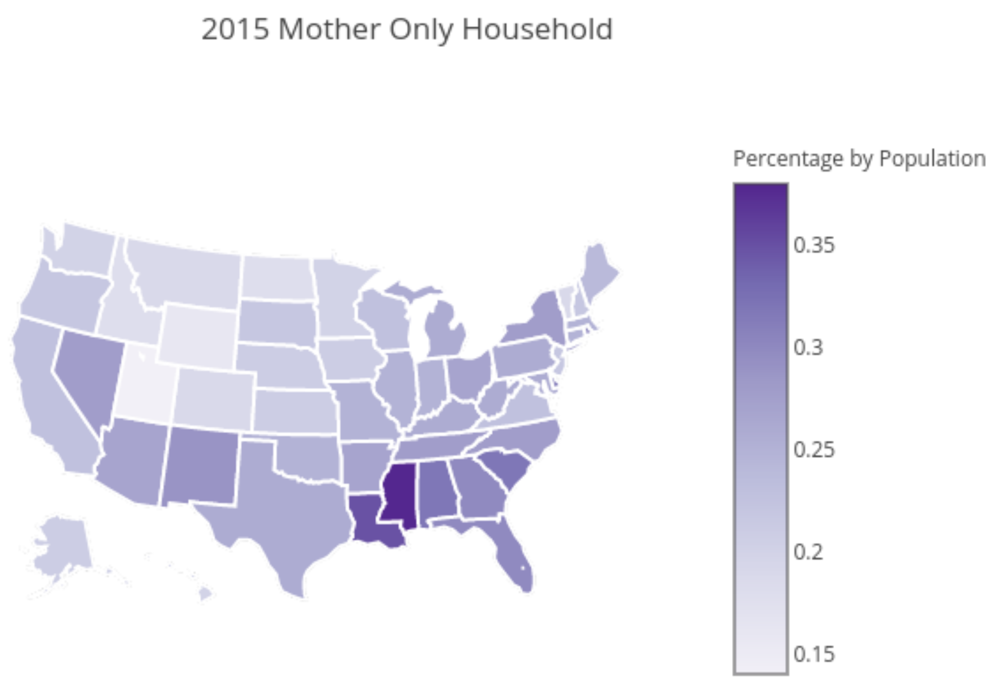
In [4]:
```
from IPython.display import Image
url = 'https://angelayxng.files.wordpress.com/2018/12/father-only.png'
Image(url)
```

Out[4]:



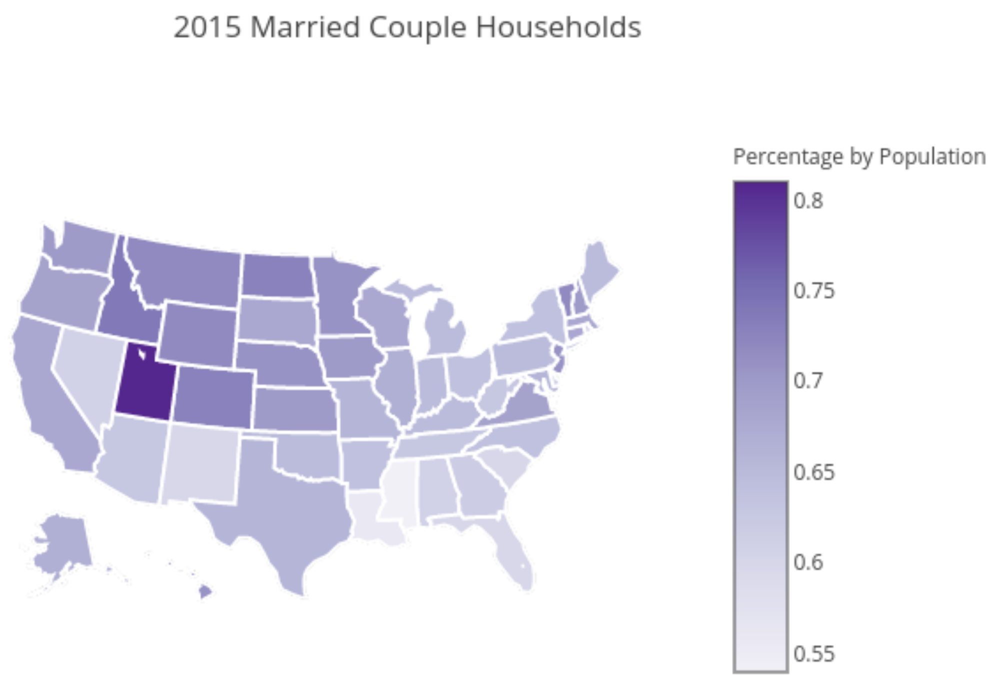
In [8]:
```
from IPython.display import Image
url = 'https://angelayxng.files.wordpress.com/2018/12/mother-only.png'
Image(url)
```

Out[8]:



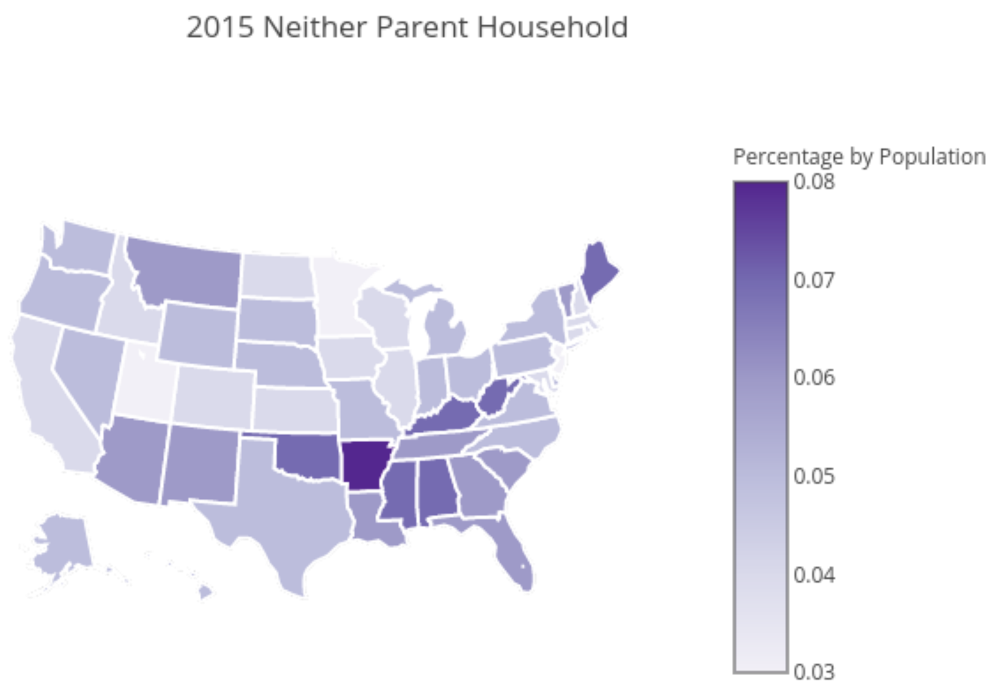
In [7]:
```
from IPython.display import Image
url = 'https://angelayxng.files.wordpress.com/2018/12/married-couple.p
ng'
Image(url)
```

Out[7]:



In [9]:
```
from IPython.display import Image
url = 'https://angelayxng.files.wordpress.com/2018/12/neither-parent.p
ng'
Image(url)
```

Out[9]:



Conclusion

The goal of this project was to understand whether growing up in a certain kind of household (father only household, mother only household, married couple household and neither parent household) has an effect on the children (who are finishing high school) and young adults' educational attainment. Specifically, we were interested in finding out whether education attainment levels of children/young adults who grew up in married couple households were higher than of those children/young adults who grew up in single parent or neither parent households because of greater emotional involvement.

We analyzed data from 50 states and District of Columbia and Puerto Rico over a period from 2000-2017. After cleaning our data and regressing percentages of children/young adults with high school diplomas or GEDs on the household structure for the years 2000, 2005, 2010 and 2015, we found that the household structure and child's education attainment weren't correlated. To be specific, the coefficients of the dependent variables changed signs. For instance, while the coefficient for father-only household was negative for 2000, it came out to be positive for 2005, 2010 and 2015. Moreover, the coefficients of married couple households fluctuated as it went from negative (2000) to positive (2005 and 2010) and then negative (2015).

The change in the signs of the coefficients along with the values of R^2 and p values suggest that there are other factors at play influencing levels of higher education beyond household structure. It is possible that factors such as occupations/job potential, incomes, and adverse life events play more significant roles in a child's potential to graduate high school and pursue college. While household structure is a good predictor of such factors, it is not the whole story.

Finding data on children was extremely challenging due to ethical concerns. We opted instead to use the ACS data, as it was the most comprehensive dataset we could find for the variables we wanted to explore. This data takes the 50 largest cities along with their states (and Puerto Rico and District of Columbia) in a given year based on survey results. The data we pulled covers the years 2000-2017. Based on the longer time frame, we believed the data would accurately portray how children made education choices as they became young adults-- given that they moved from being within households as children on to higher education in this time span. Although the data did not track children individually, we believe that the aggregate data on household structures and the longitudinal horizon for education levels of young adults would be indicative of the effects of household structure on education levels of young adults. One of the obvious drawbacks to this is that we are are not able to track children over time, regarding changes in household structure that occur or direct education levels received by specific children given their household structure.

Since we were not able to find much correlation when we regressed the education level on household structure for all the states, we decided to take another approach. We decided to take 2 states - District of Columbia and Alaska - based on the criteria that these states have the highest and the lowest aggregate percentage of high school graduate young adults (respectively) over 2000-2017 to understand if there were any trends. We thought that when we used a more specific type of regression, rather than an aggregate regression, we may have better and more accurate results. Upon doing so, we found that this was not the best metric, given that it excluded individuals who received educations beyond high school diplomas. By comparing the highest/lowest education level states by degree, the insights were much more interesting. We were able to essentially see which states had the highest percentage of people attain a given degree. While the District of Columbia had the highest percentage of Bachelor's and Graduate degrees, it had the lowest percentage of High School Diplomas or Associate's Degrees. We interpreted this as possibly meaning that those individuals who are receiving higher educations are choosing to pursue these higher degrees, likely as a result of the employment opportunities made available for them. While North Dakota had the lowest number of non-high school graduates and the highest number of Associate's Degrees, we argue that it has the highest base-line for education.

Sources

https://datacenter.kidscount.org/locations (https://datacenter.kidscount.org/locations)

Github Link

https://github.com/clb536/Data_Bootcamp_Final_Project (https://github.com/clb536/Data_Bootcamp_Final_Project)