Siddhant Rai

Team Members: Caleb Kao , Katia Garcia

PREPARING DATA

1. Education

```
#Create continuous for EDUCD
acs_data = pd.merge(acs_data, crosswalk, left_on='EDUCD', right_on='educd')
acs_data
```

	YEAR	SAMPLE	SERIAL	CBSERIAL	HHWT	CLUSTER	STRATA	GQ	PERNUM	PERWT	NCHILD	NCHLT5	SEX	AGE	MARST	RACE
0	2019	201901	2611	2019000016124	19504.8	2019000026111	80001	1	1	19659.6	1	1	1	29	1	1
1	2019	201901	6016	2019000247422	1702.8	2019000060161	230001	1	1	1702.8	3	1	1	41	1	1
2	2019	201901	6790	2019000301008	8668.8	2019000067901	190001	1	3	20278.8	0	0	1	21	6	1
3	2019	201901	9577	2019000497180	42105.6	2019000095771	10001	1	1	42105.6	0	0	1	20	6	1
4	2019	201901	9731	2019000507929	7430.4	2019000097311	270201	1	1	7275.6	0	0	1	33	6	2
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
9049	2019	201901	1220732	2019000282619	26006.4	2019012207321	230548	1	1	25851.6	2	1	1	47	1	1
9050	2019	201901	1222589	2019000308283	33746.4	2019012225891	680448	1	2	39628.8	0	0	1	43	6	1
9051	2019	201901	1237295	2019000512472	20433.6	2019012372951	670248	1	2	21207.6	2	0	2	59	1	1
9052	2019	201901	1302776	2019001404140	27554.4	2019013027761	461548	1	1	27554.4	0	0	2	49	3	7
9053	2019	201901	1308039	2019000485767	13312.8	2019013080391	5700249	1	1	13312.8	2	0	1	61	1	1

RACED	HISPAN	HISPAND	EDUC	EDUCD	EMPSTAT	EMPSTATD	INCWAGE	VETSTAT	VETSTATD	educd	educdc
100	0	0	8	81	1	10	59000	1	11	81	14.0
100	0	0	8	81	1	10	50000	2	20	81	14.0
100	0	0	8	81	1	10	10000	1	11	81	14.0
100	0	0	8	81	1	10	800	1	11	81	14.0
200	0	0	8	81	1	10	45000	1	11	81	14.0
...	...	...	...	...	...	...	...	...	...	...	...
100	1	100	2	22	1	10	25000	1	11	22	5.0
100	1	100	2	22	1	10	21600	1	11	22	5.0
100	1	100	2	22	1	10	18000	1	11	22	5.0
700	4	416	2	22	1	10	17000	1	11	22	5.0
100	1	100	1	14	1	10	38000	1	11	14	1.0

2. Dummy Variables

```
#Get Dummies
acs_data['hsdip'] = np.where(acs_data['EDUCD'] == 63, 1, 0)
acs_data['coldip'] = np.where(acs_data['EDUCD'] == 101,  1, 0)
acs_data['white'] = np.where(acs_data['RACE'] == 1,  1, 0)
acs_data['black'] = np.where(acs_data['RACE'] == 2,  1, 0)
acs_data['hispanic'] = np.where(acs_data['RACE'] != 0,  1, 0)
acs_data['married'] = np.where(acs_data['MARST'] == 1,  1, 0)
acs_data['female'] = np.where(acs_data['SEX'] == 2,  1, 0)
acs_data['vet'] = np.where(acs_data['VETSTAT'] == 2,  1, 0)
acs_data.head()
```

educd	educdc	hsdip	coldip	white	black	hispanic	married	female	vet
81	14.0	0	0	1	0	1	1	0	0
81	14.0	0	0	1	0	1	1	0	1
81	14.0	0	0	1	0	1	0	0	0
81	14.0	0	0	1	0	1	0	0	0
81	14.0	0	0	0	1	1	0	0	0

3. Interaction Terms

```
#Interaction
acs_data['EDUC:educdc'] = acs_data['EDUC'].mul(acs_data['educdc'])
acs_data.head()
```

EDUC:educdc
112.0
112.0
112.0
112.0
112.0

4. Created Variables

```
#Age Squared
acs_data['AGE^2'] = np.power(acs_data['AGE'],2)
acs_data.head()
```

AGE^2
841
1681
441
400
1089

```
#Log of Wage
acs_data = acs_data[acs_data['INCWAGE'] != 0] #Only one row had a 0 so needed to remove it
acs_data['LNINCWAGE'] = np.log(acs_data['INCWAGE'])
acs_data.head()
```

LNINCWAGE
10.985293
10.819778
9.210340
6.684612
10.714418

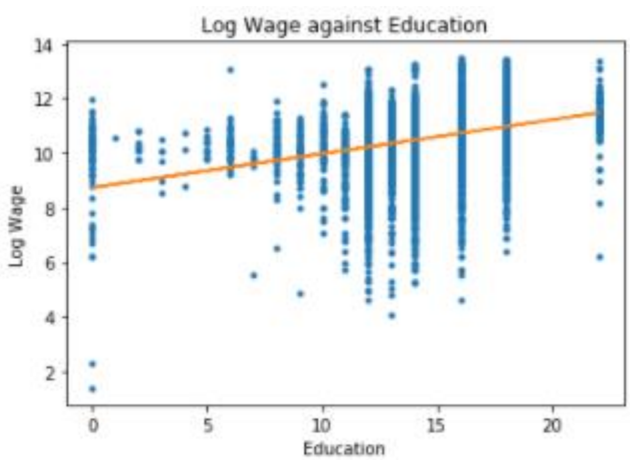
DATA ANALYSIS

1.

	YEAR	INCWAGE	LNINCWAGE	educdc	female	AGE	AGE^2	white	black	hispanic	married
count	8606.0	8606.000000	8606.000000	8606.000000	8606.000000	8606.000000	8606.000000	8606.000000	8606.000000	8606.0	8606.000000
mean	2019.0	58420.063212	10.496826	14.187427	0.486521	41.849059	1931.513479	0.773298	0.090286	1.0	0.525912
std	0.0	68115.268196	1.097734	2.857626	0.499847	13.423513	1126.920775	0.418723	0.286607	0.0	0.499357
min	2019.0	4.000000	1.386294	0.000000	0.000000	18.000000	324.000000	0.000000	0.000000	1.0	0.000000
25%	2019.0	22725.000000	10.031219	12.000000	0.000000	30.000000	900.000000	1.000000	0.000000	1.0	0.000000
50%	2019.0	41300.000000	10.628615	14.000000	0.000000	42.000000	1764.000000	1.000000	0.000000	1.0	1.000000
75%	2019.0	70750.000000	11.166889	16.000000	1.000000	54.000000	2916.000000	1.000000	0.000000	1.0	1.000000
max	2019.0	717000.000000	13.482831	22.000000	1.000000	65.000000	4225.000000	1.000000	1.000000	1.0	1.000000

NCHILD	vet	hsdip	coldip	EDUC:educdc
8606.000000	8606.000000	8606.000000	8606.000000	8606.000000
0.784801	0.049733	0.210783	0.236812	117.493028
1.100708	0.217405	0.407888	0.425150	51.649895
0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	0.000000	0.000000	72.000000
0.000000	0.000000	0.000000	0.000000	98.000000
1.000000	0.000000	0.000000	0.000000	160.000000
9.000000	1.000000	1.000000	1.000000	242.000000

2.



3.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:              LNINCWAGE   R-squared:                       0.309
Model:                            OLS   Adj. R-squared:                  0.308
Method:                 Least Squares   F-statistic:                     426.4
Date:                Sat, 30 Jan 2021   Prob (F-statistic):               0.00
Time:                        19:24:58   Log-Likelihood:                -11425.
No. Observations:                8606   AIC:                         2.287e+04
Df Residuals:                    8596   BIC:                         2.294e+04
Df Model:                           9
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
educdc         0.1112      0.004     31.564      0.000       0.104       0.118
female        -0.4338      0.020    -21.577      0.000      -0.473      -0.394
AGE            0.1567      0.006     27.646      0.000       0.146       0.168
AGE^2         -0.0016    6.7e-05    -24.270      0.000      -0.002      -0.001
white         -0.0297      0.029     -1.024      0.306      -0.087       0.027
black         -0.1875      0.043     -4.410      0.000      -0.271      -0.104
hispanic       5.6555      0.114     49.594      0.000       5.432       5.879
married        0.1955      0.023      8.450      0.000       0.150       0.241
NCHILD        -0.0063      0.010     -0.616      0.538      -0.026       0.014
vet           -0.0396      0.046     -0.856      0.392      -0.130       0.051
==============================================================================
Omnibus:                     2437.163   Durbin-Watson:                   1.867
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            10300.166
Skew:                          -1.335   Prob(JB):                         0.00
Kurtosis:                       7.647   Cond. No.                     2.60e+04
==============================================================================
```

a) The fraction of variation is the R^2 which from our results is 0.309 which indicates that around 30.9% of the log wage as the dependent variable can be explained by the predictor
b) The F-statistics: 426.4 and Prob(F-statistic): 0 indicates that we reject the null hypothesis that ln(wage) is not affected by any of the predictors
c) An additional year of education increased the log wage by (e^(0.112)-1)*100 =11.85129 ~ 11.985% and is statistically significant because we have a p-value of 0 which is less than 0.05.
d) According to the model, the ln(wage) will achieve the highest wage at Age 50.

```
d = {'AGE': [50]}
df = pd.DataFrame(data=d)
predictions = result.get_prediction(df)
predictions.summary_frame(alpha=0.05)
```

	mean	mean_se	mean_ci_lower	mean_ci_upper	obs_ci_lower	obs_ci_upper
0	10.923088	0.06992	10.786029	11.060148	8.99937	12.846806

```
Intercept          8.5613      0.088     96.999      0.000       8.388       8.734
C(AGE)[T.19]       0.1403      0.118      1.192      0.233      -0.090       0.371
C(AGE)[T.20]       0.5995      0.119      5.020      0.000       0.365       0.834
C(AGE)[T.21]       0.8163      0.120      6.803      0.000       0.581       1.052
C(AGE)[T.22]       0.8900      0.113      7.889      0.000       0.669       1.111
C(AGE)[T.23]       1.2769      0.117     10.936      0.000       1.048       1.506
C(AGE)[T.24]       1.5024      0.116     12.901      0.000       1.274       1.731
C(AGE)[T.25]       1.5190      0.115     13.157      0.000       1.293       1.745
C(AGE)[T.26]       1.7094      0.115     14.859      0.000       1.484       1.935
C(AGE)[T.27]       1.8766      0.118     15.876      0.000       1.645       2.108
C(AGE)[T.28]       1.8940      0.114     16.649      0.000       1.671       2.117
C(AGE)[T.29]       1.7146      0.113     15.213      0.000       1.494       1.936
C(AGE)[T.30]       2.0496      0.111     18.489      0.000       1.832       2.267
C(AGE)[T.31]       1.9967      0.114     17.552      0.000       1.774       2.220
C(AGE)[T.32]       2.0790      0.115     18.134      0.000       1.854       2.304
C(AGE)[T.33]       1.9255      0.116     16.658      0.000       1.699       2.152
C(AGE)[T.34]       2.1251      0.112     18.928      0.000       1.905       2.345
C(AGE)[T.35]       1.9479      0.115     16.892      0.000       1.722       2.174
C(AGE)[T.36]       2.0875      0.112     18.593      0.000       1.867       2.308
C(AGE)[T.37]       2.2044      0.117     18.855      0.000       1.975       2.434
C(AGE)[T.38]       2.1583      0.114     18.868      0.000       1.934       2.382
C(AGE)[T.39]       2.3228      0.112     20.689      0.000       2.103       2.543
C(AGE)[T.40]       2.1019      0.115     18.313      0.000       1.877       2.327
C(AGE)[T.41]       2.1998      0.114     19.295      0.000       1.976       2.423
C(AGE)[T.42]       2.1307      0.114     18.770      0.000       1.908       2.353
C(AGE)[T.43]       2.1793      0.115     18.988      0.000       1.954       2.404
C(AGE)[T.44]       2.0735      0.115     17.960      0.000       1.847       2.300
C(AGE)[T.45]       2.2876      0.115     19.931      0.000       2.063       2.513
C(AGE)[T.46]       2.2064      0.115     19.134      0.000       1.980       2.432
C(AGE)[T.47]       2.1709      0.115     18.870      0.000       1.945       2.396
C(AGE)[T.48]       2.2629      0.115     19.738      0.000       2.038       2.488
C(AGE)[T.49]       2.3234      0.115     20.196      0.000       2.098       2.549
C(AGE)[T.50]       2.3618      0.113     20.975      0.000       2.141       2.583
C(AGE)[T.51]       2.2763      0.112     20.389      0.000       2.057       2.495
C(AGE)[T.52]       2.2442      0.112     19.989      0.000       2.024       2.464
C(AGE)[T.53]       2.1435      0.112     19.217      0.000       1.925       2.362
C(AGE)[T.54]       2.2303      0.111     20.085      0.000       2.013       2.448
C(AGE)[T.55]       2.1614      0.114     19.040      0.000       1.939       2.384
C(AGE)[T.56]       2.1321      0.108     19.776      0.000       1.921       2.343
C(AGE)[T.57]       2.3128      0.114     20.373      0.000       2.090       2.535
C(AGE)[T.58]       2.1994      0.111     19.736      0.000       1.981       2.418
C(AGE)[T.59]       2.1471      0.114     18.792      0.000       1.923       2.371
C(AGE)[T.60]       2.1243      0.116     18.355      0.000       1.897       2.351
C(AGE)[T.61]       2.0855      0.113     18.448      0.000       1.864       2.307
C(AGE)[T.62]       2.1106      0.116     18.192      0.000       1.883       2.338
C(AGE)[T.63]       1.9634      0.120     16.336      0.000       1.728       2.199
C(AGE)[T.64]       1.9695      0.120     16.439      0.000       1.735       2.204
C(AGE)[T.65]       2.0175      0.126     16.063      0.000       1.771       2.264
==============================================================================
```

e) The model predicts that men will have a slightly higher wage than females and according to our results the coefficient for females is -0.4338. We might investigate this pattern to see when males start to have a higher wage gap than females at a certain age, married status, or birth to a child

```
d = {'SEX': [1]}
df = pd.DataFrame(data=d)
predictions = result.get_prediction(df)
predictions.summary_frame(alpha=0.05)
```

	mean	mean_se	mean_ci_lower	mean_ci_upper	obs_ci_lower	obs_ci_upper
0	10.69294	0.016234	10.661118	10.724762	8.577313	12.808567

```
d = {'SEX': [2]}
df = pd.DataFrame(data=d)
predictions = result.get_prediction(df)
predictions.summary_frame(alpha=0.05)
```

	mean	mean_se	mean_ci_lower	mean_ci_upper	obs_ci_lower	obs_ci_upper
0	10.289845	0.016677	10.257153	10.322536	8.174204	12.405485

f) Interpreting the coefficients White: -0.0297, Black:-0.1875, Hispanic: 5.655 indicates that Hispanics have the highest percentage increase in wages controlling for all variables, whites had a small decrease, but is NOT statistically significant, and blacks have the highest decrease in wages in 2019.

g) Null Hypothesis: $H_0 : B_{race} = 0$

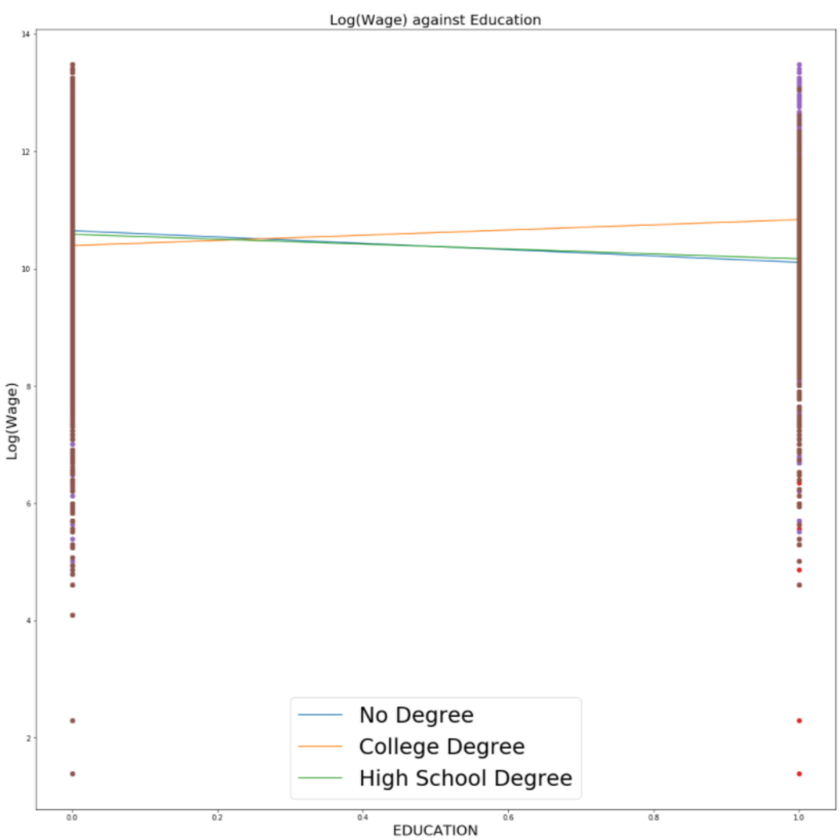
Alternative Hypothesis: $H_A : B_{race} \neq 0$

```
                            OLS Regression Results
==============================================================================
Dep. Variable:              LNINCWAGE   R-squared:                       0.001
Model:                            OLS   Adj. R-squared:                  0.001
Method:                 Least Squares   F-statistic:                     12.56
Date:                Sun, 31 Jan 2021   Prob (F-statistic):           0.000397
Time:                        17:10:32   Log-Likelihood:                -13007.
No. Observations:                8606   AIC:                         2.602e+04
Df Residuals:                    8604   BIC:                         2.603e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     10.5374      0.016    639.840      0.000      10.505      10.570
RACE          -0.0225      0.006     -3.543      0.000      -0.035      -0.010
==============================================================================
Omnibus:                     1902.558   Durbin-Watson:                   1.627
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             6042.310
Skew:                          -1.122   Prob(JB):                         0.00
Kurtosis:                       6.437   Cond. No.                         3.89
==============================================================================

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

Running an OLS of LNWAGE on RACE. We reject the null hypothesis that race has nothing to do with wages because it is not statistically significant , but the variation is very small of R^2 at 0.001.

4. All three of the scatter plots in the graph below tend to overlap each other. Also, College is shown to increase, while High School Diploma and No degree show a decrease in wages.



5. **LNINCWAGE = B_0 + B_1 (educdc) + B_2 (coldip) + B_4 (hsdip) + B_3 (female) + B_4 (AGE) + B_5 (AGE^2) + B_6 (white) + B_7 (black) + B_8 (Hispanic) + B_9 (married) + B_{10} (NCHILD) + B_{11} (educdc*hsdip)+ B_{11} (educdc*coldip)**

   Probability = e^(LNINCWAGE REGRESSION OUTPUT)

   I believe this is the best model because we can control for all variables and include these three categorical variables because we are trying to find correlations between the education levels and INCWAGE not the causation. Also, the parameters B_0, B_1 etc. best explain the observed data and with this probability model we can explain more than 2 categorical variables

6.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:              LNINCWAGE   R-squared:                       0.266
Model:                            OLS   Adj. R-squared:                  0.265
Method:                 Least Squares   F-statistic:                     282.7
Date:                Mon, 01 Feb 2021   Prob (F-statistic):               0.00
Time:                        21:25:45   Log-Likelihood:                -11684.
No. Observations:                8606   AIC:                         2.339e+04
Df Residuals:                    8594   BIC:                         2.348e+04
Df Model:                          11
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
Intercept         4.0434      0.035    114.563      0.000       3.974       4.113
female           -0.4459      0.021    -21.508      0.000      -0.487      -0.405
AGE               0.0168      0.005      3.226      0.001       0.007       0.027
AGE ^ 2           0.0040      0.005      0.782      0.434      -0.006       0.014
white            -0.0294      0.030     -0.981      0.327      -0.088       0.029
black            -0.1537      0.044     -3.505      0.000      -0.240      -0.068
hispanic          4.0434      0.035    114.563      0.000       3.974       4.113
married           0.2366      0.024      9.945      0.000       0.190       0.283
NCHILD            0.0803      0.010      8.090      0.000       0.061       0.100
vet              -0.0246      0.048     -0.516      0.606      -0.118       0.069
educdc            0.1110      0.004     27.367      0.000       0.103       0.119
hsdip            -0.0003      0.000     -1.434      0.152      -0.001       0.000
coldip            0.0007      0.000      6.720      0.000       0.000       0.001
educdc:hsdip     -0.0033      0.002     -1.434      0.152      -0.008       0.001
educdc:coldip     0.0109      0.002      6.720      0.000       0.008       0.014
==============================================================================
Omnibus:                     2434.650   Durbin-Watson:                   1.858
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             9490.073
Skew:                          -1.362   Prob(JB):                         0.00
Kurtosis:                       7.364   Cond. No.                     8.92e+17
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 4.41e-29. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
```

a) The wages which predict from our model which we get is a mean of 7.1465 which equals e^(7.1465) = $1269.65

			High School	College
Intercept	4.034			
female	-0.4459		1	1
age	0.0168		22	22
age^2	0.004		484	484
white	-0.0294		0	0
black	-0.1537		0	0
hispanic	4.0434		0	0
married	0.2366		0	0
NCHILD	0.0803		0	0
vet	-0.0246		0	0
educdc	0.111		12	16
hsdip	-0.0396		1	1
coldip	0.1747		0	0
educdc:hsdip	-0.0033		12	16
educdc:coldip	0.0109		0	0
			7.1465	7.5773
			1269.654379	1953.348

b) Yes, the individuals do receive a higher wage, and according to our model for the female student who got a college education, of about 7.5773 , e^(7.5773)= $1953.35 which is a $683 dollar increase with a college degree

c) From my point of view, it is crucial to expand college education to increase wages for male and female workers. However, allowing subsidies will increase the cost of attending college which can cause the students to fall into debt knowing college administrators can raise the prices because banks are providing easy money loans without any application.

7. To improve this model, I would probably include more interaction between education and variables such as Age, Married, and NCHILD to see where the increase or drop off is in wages. The reason why is when women give birth to a child wages tends to slip as they focus on more part-time jobs.