ECON 4232 Labor Economics

Analysis of Factors Affecting Earnings Using Current Population Survey Data From The Bureau of Labor Statistics

Karen Loscocco

April 30, 2020

Contents

1 Executive Summary

In this paper, I examine the causal relationship between education and earnings and also perform regression analysis to determine the best predictors of earnings. I perform the analysis on earnings data from the Current Population Survey (CPS) which is conducted each month by the U.S. Census Bureau for the Bureau of Labor Statistics. Regression is a powerful approach used in a wide variety of applications for predictive analysis in both industry and academia. After exploring a multiple regression model using R, I have come to the conclusion that education, sex, and race are the best predictors of earnings. I include my data exploration, mathematical regression models, and implementations in R, as well as my solutions and comments on the applications of my findings.

2 Introduction

Education plays a key role in U.S. labor markets. Literature suggests a positive correlation between a person's level of education and his or her labor market status and earnings. Numerous studies have shown that better-educated individuals not only earn higher wages but experience less unemployment, and work in more prestigious occupations. It is however, difficult to prove that there exists a causal relationship here. The higher earnings for the better-educated could be due to the fact that individuals with greater earning capacity choose to acquire mores school. This suggests that the inherent ability differences between people could be the cause of the earnings differences, rather than just the schooling level itself. There are a large number of economists that believe education serves mainly to signal a worker's qualifications or innate ability, and not actually increase a worker's productivity at all. What else besides a degree might influence an individual's potential earnings? The goal of this paper is to examine the impact of education on earnings and determine which other factors have the most impact on earnings.

3 Problem Description

My analysis proposes to determine which variables have the most effect on the observed differences in earnings. I tested a multiple linear regression models with earnings as the dependent variable in order to find the model which contains the most statistically significant predictor variables. I also made observations on the data itself and included visuals to give a deeper understanding of the data.

4 Data Exploration

4.1 Data Description

The CPS Earnings (LE) data is based off of data from the Current Population Survey (CPS). Below, I include definitions of the CPS survey, description of the CPS earnings data collection, and structure of the earnings data directly from the BLS data source.

The CPS is a sample survey of the population 16 years of age and over conducted each month by the U.S. Census Bureau for the Bureau of Labor Statistics and provides comprehensive data on the labor force, the employed, and the unemployed, classified by such characteristics as age, sex, race, family relationship, marital status, occupation, and industry attachment. The information is collected by trained interviewers from a sample of about 60,000 households located in 754 sample areas. These areas are chosen to represent all counties and independent cities in the United States, with coverage in 50 States and the District of Columbia.

The earnings data consists of quarterly earnings averages for all wage and salary workers broken down by age, race, Hispanic or Latino ethnicity, sex, occupation, usual full-or part-time status, educational attainment, and other characteristics. All self-employed workers are excluded. The earnings reported are before taxes and other educations, and include any overtime pay, commissions, or tips usually received. Earnings reported on a basis other than weekly are converted to weekly. All data expressed in constant dollars are deflated by the Consumer Price Index for All Urban Consumers (CPI-U).

There are twenty-four tables that make up the CPS Earnings data. They include time series table, a data table, and mapping tables.

A time series refers to a set of data observed over an extended period of time over consistent time intervals. The BLS time series data are typically produced at monthly intervals and represent data ranging from a specific consumer item in a specific geographical area whose price is gathered monthly to a category of worker in a specific industry whose employment rate is being recorded monthly, etc. The series file contains a set of codes which, together, compose a series identification code that serves to uniquely identify a single time series. Additionally, the series file also contains the period and year corresponding to the first data observation and the period and year corresponding to the most recent data observation.

The mapping files are definition files that contain explanatory text descriptions that correspond to each of the various codes contained within each series identification code.

The data file contains one line of data for each observation period pertaining to a specific time series. Each line contains a reference to the following: series identification code, year in which data is observed, period for which data is observed, value.

The following is a breakdown of all the earnings tables and their respective meanings:

Table	Meaning
contacts	Contacts for LE survey
ages	Age codes mapping
cert	Certification and licensing status codes mapping
class	Class of Worker codes mapping
data.0.Current	All current year-to-date data
data.1.AllData	All data
earn	Earnings codes mapping
education	Education codes mapping
fips	Federal Information Processing Standards codes mapping
indy	industry codes mapping
lfst	Labor Force Status codes mapping
occupation	Occupation codes mapping
orig	Ethnic Origin codes mapping
pcts	Percent codes mapping
periodicity	Periodicity codes mapping
race	Race codes mapping
seasonal	Seasonal codes mapping
seq	Sequence codes mapping
series	All series and their beginning and end dates
sexs	Sexes codes mapping
stype	Seasonal Type codes mapping
tdata	Time Data codes mapping
txt	General information
unin	Union codes mapping

After writing a script that pulls all of the data from the BLS website, I joined all the data and corresponding mapping tables to the series table. There are a total of 9,221 series in this dataset.

4.2 Data Exploration

To get a better understanding of the data structure, I decided to examine a single series titled '(Seas)- Employed full time, Wage and salary workers.' This series summaries the number of people, in thousands, every quarter from 1979 to 2020 that meet the following criteria: 16 years and over, Wage and salary workers (excluding incorporated self employed), All educational levels, All states, All Industries, Employed full time, All Occupations, All Races, Both Sexes. The yearly average number of full-time employees, in thousands, is shown in figure (4.1). As seen from the figure, the number employed shows a 65% increase from 1979 to 2020.

Since this dataset breaks down the summarized values into different series, it means that I have to select specific series (from those available) in order to show the data by different breakdowns over time. For example, in order to show this specific series breakdown by gender, it needs to be available broken down by gender. If it is, I can select the three series, making sure to group the data by the series title. Figure (4.2) shows the breakdown of the number of people in the labor force over time by gender. The data show the average number women in the workforce to always be lower than the number of men over time.

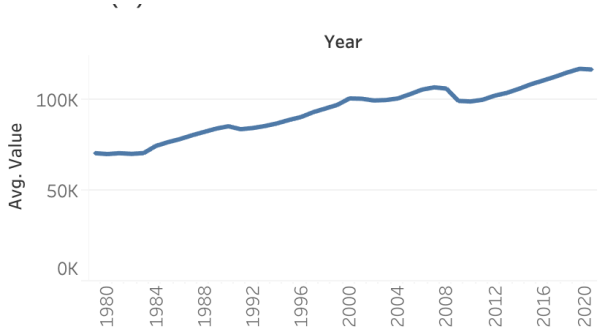


Figure 4.1

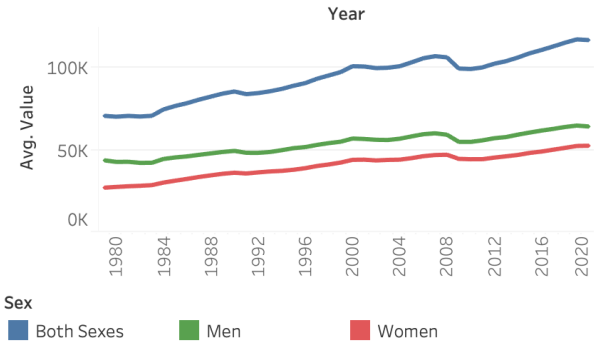


Figure 4.2

Similarly, to look at the progression of median usual weekly earnings of men and women over time, I examine another three series: '(Seas)- Median usual weekly earnings (second quartile), Employed full time, Wage and salary workers', '(Seas)- Median usual weekly earnings (second quartile), Employed full time, Wage and salary workers, Men', and '(Seas)- Median usual weekly earnings (second quartile), Employed full time, Wage and salary workers, Women.' It is important to note, that figure (4.3) shows averages of separately summarized series values. This is why the line for both sexes is not showing the sum of both men and women.

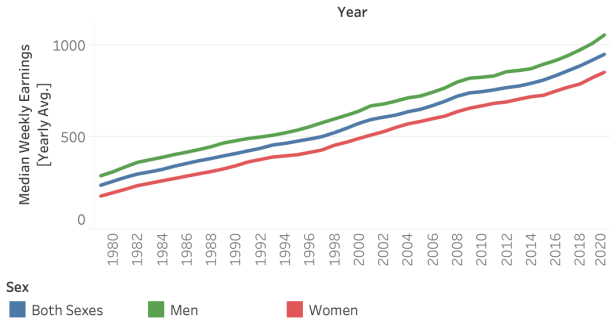
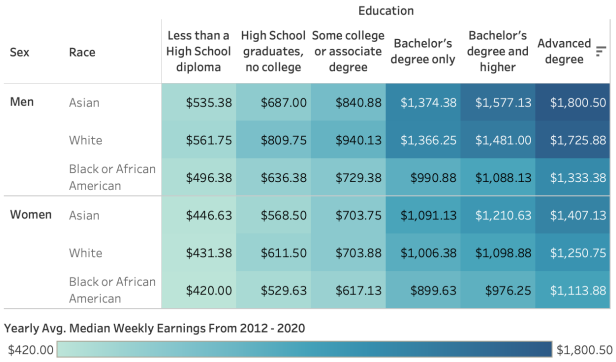


Figure 4.3

The following are some visuals that show the breakdown of data that is available for differing education level. In short, the only series that included a breakdown of education level had a breakdown of sex, race, and birthplace (foreign or native). I was hoping to be able to get additional breakdowns by at least age and region, but this was unavailable. I include some analysis in a later section

Figure (4.4) shows an average of the median earnings from 2012 to 2020 by sex, race, and education level. This data is broken down by year and quarter. The data shown here is an average of all years. Another thing to note is that this series included median earnings for all people 25 years and over, from every state and from all occupations and industries.



		Education					
Sex	Race	Less than a High School diploma	High School graduates, no college	Some college or associate degree	Bachelor's degree only	Bachelor's degree and higher	Advanced degree
Men	Asian	$535.38	$687.00	$840.88	$1,374.38	$1,577.13	$1,800.50
	White	$561.75	$809.75	$940.13	$1,366.25	$1,481.00	$1,725.88
	Black or African American	$496.38	$636.38	$729.38	$990.88	$1,088.13	$1,333.38
Women	Asian	$446.63	$568.50	$703.75	$1,091.13	$1,210.63	$1,407.13
	White	$431.38	$611.50	$703.88	$1,006.38	$1,098.88	$1,250.75
	Black or African American	$420.00	$529.63	$617.13	$899.63	$976.25	$1,113.88

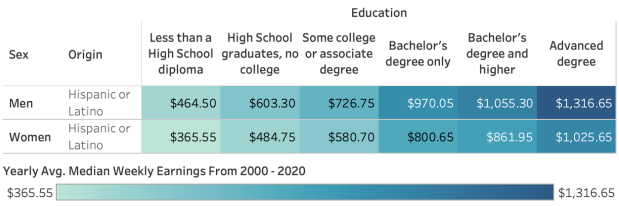
Yearly Avg. Median Weekly Earnings From 2012 - 2020

$420.00 $1,800.50

Figure 4.4

As you can see, the highest yearly average of weekly earnings are from Asian men with advanced degrees. They earn on average $1,800.50 per week, or $93,626 per year. With this breakdown alone, you can see a correlation between attaining higher education levels and higher earnings. The lowest values are those who have less than a high school degree, making around $500 per week, or $26,000 per year. This graphic also confirms the above ones with respect to lower average weekly earnings for women relative to me.

Figure (4.5) shows a similar breakdown, but includes origins of Hispanic or Latino. According to the U.S. Census, people who identify their origin as Hispanic, Latino, or Spanish may be of any race. I include this here to make the comparison because the series was available with the same education breakdown.



		Education					
Sex	Origin	Less than a High School diploma	High School graduates, no college	Some college or associate degree	Bachelor's degree only	Bachelor's degree and higher	Advanced degree
Men	Hispanic or Latino	$464.50	$603.30	$726.75	$970.05	$1,055.30	$1,316.65
Women	Hispanic or Latino	$365.55	$484.75	$580.70	$800.65	$861.95	$1,025.65

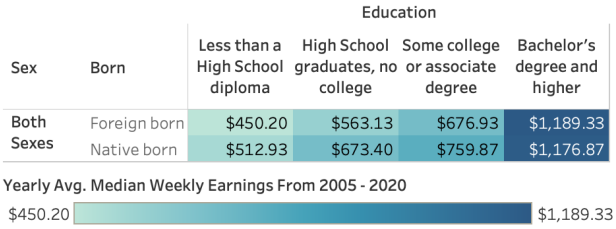
Yearly Avg. Median Weekly Earnings From 2000 - 2020

$365.55 $1,316.65

Figure 4.5

The third breakdown available by education was one broken down by if the person was foreign or native born. Figure (4.6) shows this breakdown. It is interesting to note that Foreign born persons with a Bachelor's degree and higher earn the most per week, at $1,189.33. Unfortunately, this data was not broken out by gender, but it is still interesting to note that foreign born persons with a Bachelor's degree and higher earn more than native born persons on average. It is also interesting to see that foreign born persons with anything less than a bachelors degree earn less than their native born counterparts. The theories learned in class suggest that immigrants are not randomly selected from the population but that only foreign persons who have exceptional ability are likely to immigrate to the U.S.. This could explain the higher earnings for foreign born persons with a Bachelor's degree and higher. I am not sure if this data includes illegal immigrants. This could

explain that for all foreign born persons with less than a Bachelor's degree earn less than their native born counterparts.



		Education			
Sex	Born	Less than a High School diploma	High School graduates, no college	Some college or associate degree	Bachelor's degree and higher
Both Sexes	Foreign born	$450.20	$563.13	$676.93	$1,189.33
	Native born	$512.93	$673.40	$759.87	$1,176.87

Yearly Avg. Median Weekly Earnings From 2005 - 2020

$450.20 $1,189.33

Figure 4.6

To look further into factors that impact earnings other than education, I include the following graphs that show the available series data broken out by state. Figure (4.7) shows this breakdown below. it is interesting to note the top 10 highest earnings states, which are District of Columbia ($1,228.50) , Massachusetts ($1,025.63), Connecticut ($1,012.88), Maryland ($991.63), New Jersey ($976.75), Washington ($939.25), Alaska ($931.63), Minnesota ($929.13), Virginia ($925.63), and New Hampshire ($915.63). This could be due to the cost of living be higher. This could also be due to tax laws and regulations. For example, minimum wage laws, income tax rates, and sales tax rates.

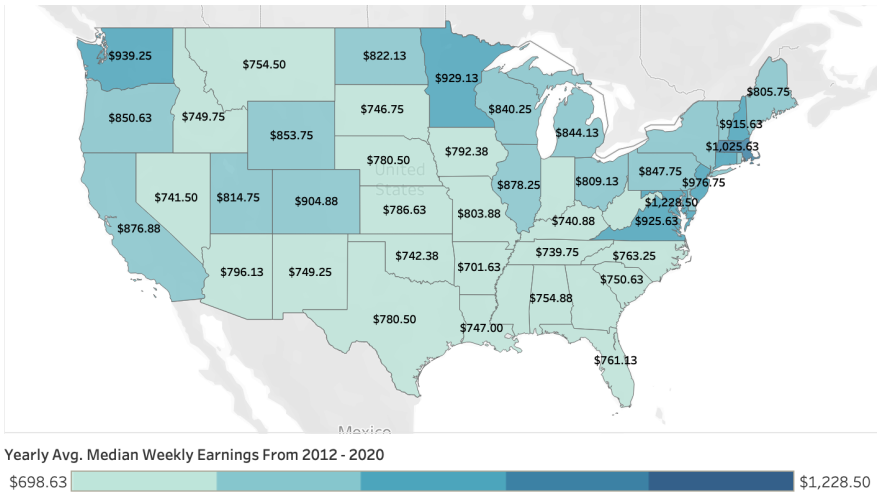


Figure 4.7

Figure (4.8) shows a breakdown of the highest paying occupations with the highest paying being the largest box. The top five are the following: Architecture and Engineering occupations ($1,529.40), Computer and Mathematical occupations ($1,509.20), Management Occupations (%1,457.80), Legal Occupations (%1,430.80), Management, Business, and Financial Operations Occupations ($1,391). This could be due to a correlation in level of education attainment and certain occupations.

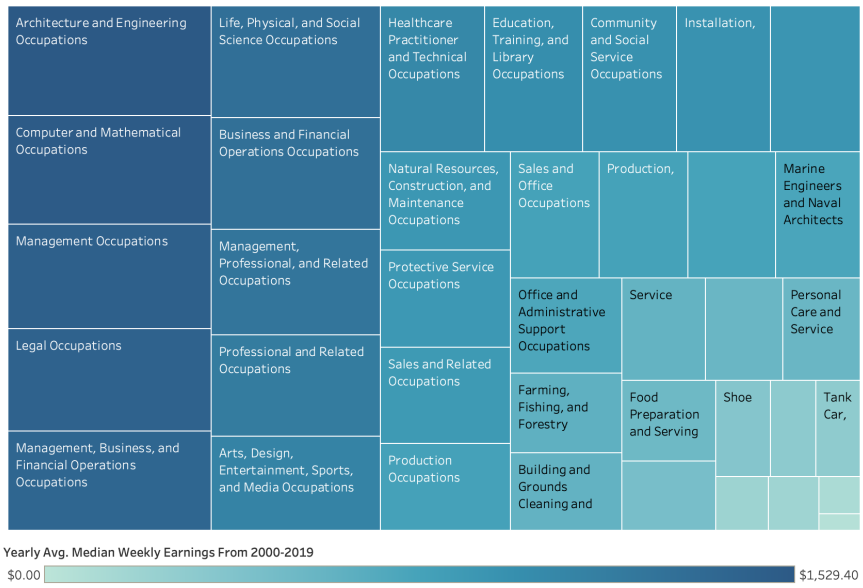


Figure 4.8

Figure (4.9) and figure (4.10) shows the top industries with the highest number of people (in thousands) who are at the prevailing federal minimum wage and below the prevailing federal minimum wage, respectively. There are a very large number of people who are earning below the minimum wage in the industries indicated below. This shows that certain industries might earn less than others.

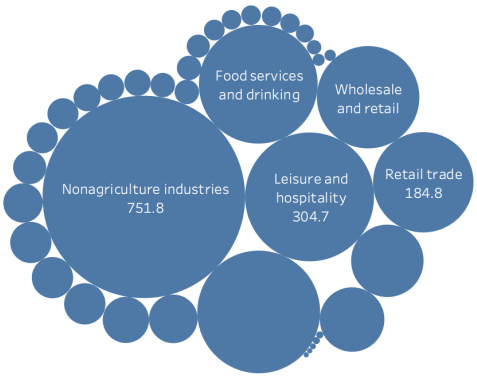


Figure 4.9

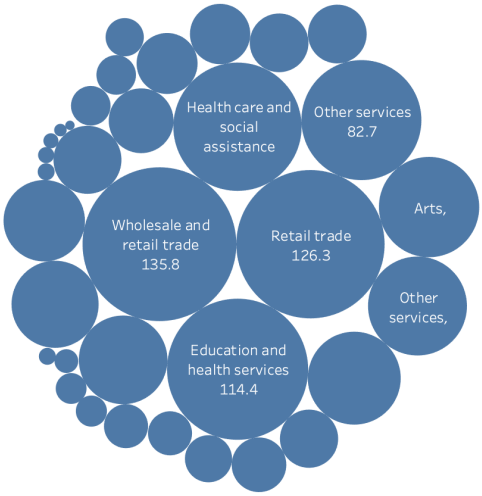


Figure 4.10

These data visualizations are helpful because they prove that there may be other factors that influencing earnings than just education.

5 Regression Analysis

After looking at the time series data, I decided that in order to perform a regression analysis, I would need to examine either one year, or an average over a set of years or data. I would also need to combine data from multiple series. I wanted to break down the data by education level, gender, age, region, and race. I decided to exclude looking at all industries and all occupations for this analysis because I knew it would require a lot of data processing and creation of a lot of dummy variables in my model.

I found that a complete set of all the series which breaks down the summaries by all of these criteria did not exists. This meant that I could not perform regression analysis with all of these criteria. I decided to take the average of series data from 2000-2019 that was broken down by sex, race, and education level.

The summary of this model in figure (5.1) reveals that the level of education, sex, and race are significant. All variables have p-values significantly less than 0.05. The analysis can also be found on page 26 of the appendix (**??**).

```
model = lm(avg_val ~ + high_nocol + bachandhigh
           + bachonly + adv + lessthanhigh
           + white + black + men)

summary(model)
```

```
##
## Call:
## lm(formula = avg_val ~ +high_nocol + bachandhigh + bachonly +
##     adv + lessthanhigh + white + black + men)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -162.57  -72.79   15.68   43.17  222.10
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    742.45      55.40  13.402 3.48e-13 ***
## high_nocol    -113.45      63.45  -1.788 0.085428 .
## bachandhigh    444.45      63.45   7.005 1.94e-07 ***
## bachonly       338.00      63.45   5.327 1.42e-05 ***
## adv            629.59      63.45   9.923 2.49e-10 ***
## lessthanhigh  -250.87      66.88  -3.751 0.000892 ***
## white         -116.42      46.09  -2.526 0.017973 *
## black         -278.13      44.86  -6.199 1.47e-06 ***
## men            206.37      37.30   5.532 8.30e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 109.9 on 26 degrees of freedom
## Multiple R-squared:  0.9328, Adjusted R-squared:  0.9122
## F-statistic: 45.13 on 8 and 26 DF,  p-value: 2.616e-13
```

```
anova(model)
```

```
## Analysis of Variance Table
##
## Response: avg_val
##              Df  Sum Sq Mean Sq F value    Pr(>F)
## high_nocol    1  638081  638081  52.837 1.016e-07 ***
## bachandhigh   1  301498  301498  24.966 3.391e-05 ***
## bachonly      1  171287  171287  14.184 0.0008578 ***
## adv           1 2209374 2209374 182.951 2.825e-13 ***
## lessthanhigh  1  205534  205534  17.020 0.0003368 ***
## white         1    1002    1002   0.083 0.7755584
## black         1  464135  464135  38.433 1.475e-06 ***
## men           1  369591  369591  30.605 8.299e-06 ***
## Residuals    26  313985   12076
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 5.1

6 Conclusion

My findings show a relationship between education and earnings and also a relationship between other factors such as race, sex, and location. I would have liked to perform further regression analysis for the other break downs. I would also like to see the raw data from the BLS.

7 References

https://download.bls.gov/pub/time.series/le/

Import Satements

In [1]:
```
import sys
import urllib
import datetime
import pandas as pd
from urllib.request import urlopen, urlretrieve
from bs4 import BeautifulSoup
import pandasql
import mysql.connector
from sqlalchemy import create_engine
import pymysql
```

Functions

In [2]:
```
def get_data(href):
    url_open = urllib.request.urlopen('https://download.bls.gov' + hre
f)
    html_doc = url_open.read().decode('utf-8')
    soup = BeautifulSoup(html_doc,'html.parser')
    soupstr = str(soup)
    souplist = soupstr.split('\r\n')
    column_names = souplist[0].split('\t')
    data = [row.split('\t') for row in souplist[1:]]
    df = pd.DataFrame(data, columns = column_names)
    return df
```

Get Earnings Data From BUREAU OF LABOR STATISTICS

In [3]:
```
bls = urllib.request.urlopen('https://download.bls.gov/pub/time.series
/le/')
html_doc = bls.read().decode('utf-8')
soup = BeautifulSoup(html_doc,'html.parser')

a = []
for link in soup.find_all('a'):
    a.append(link.get('href'))

data = {}
for i in range(1,len(a)):
    if 'txt' in a[i]:
        continue
    data[a[i].split('/le/le.')[1].replace('.','_')] = get_data(a[i])
```

In [4]:
```
list(data.keys())
```

Out[4]:
```
['ages',
 'born',
 'cert',
 'class',
 'contacts',
 'data_0_Current',
 'data_1_AllData',
 'earn',
 'education',
 'fips',
 'footnote',
 'indy',
 'lfst',
 'occupation',
 'orig',
 'pcts',
 'race',
 'seasonal',
 'series',
 'sexs',
 'tdata',
 'unin']
```

In [5]:
```
data['data_1_AllData'] = data['data_1_AllData'].rename(
    columns={'series_id        ': 'series_id', '       value':'value'}
)

data['data_0_Current'] = data['data_0_Current'].rename(
    columns={'series_id        ': 'series_id', '       value':'value'}
)

data['series'] = data['series'].rename(
    columns={'series_id        ': 'series_id'})
```

Database Connection

In [6]:
```
engine = create_engine("mysql+pymysql://root:********@localhost/bls_ea
rnings")
con = engine.connect()
```

Insert All Tables Into Database

In [7]:
```
for tablename in data.keys():
    data[tablename].to_sql(name = tablename, con = con)

con.close()
```

Appendix

REGRESSION ANALYSIS

Loading the original data:

```
project_data = read.csv("reg_educ_race_sex.csv")
```

Omitting null values:

```
nrow(project_data)
```

```
## [1] 35
```

```
project_data = na.omit(project_data)
nrow(project_data)
```

```
## [1] 35
```

```
project_data
```

```
##                        education_text                 race_text sexs_text
## 1  High School graduates, no college                     White       Men
## 2       Bachelor's degree and higher                     White       Men
## 3             Bachelor's degree only                     White       Men
## 4                    Advanced degree                     White       Men
## 5    Less than a High School diploma Black or African American       Men
## 6    Less than a High School diploma                     White     Women
## 7  High School graduates, no college                     White     Women
## 8       Bachelor's degree and higher                     White     Women
## 9             Bachelor's degree only                     White     Women
## 10                   Advanced degree                     White     Women
## 11 High School graduates, no college Black or African American       Men
## 12      Bachelor's degree and higher Black or African American       Men
## 13            Bachelor's degree only Black or African American       Men
## 14                   Advanced degree Black or African American       Men
## 15   Less than a High School diploma Black or African American     Women
## 16 High School graduates, no college Black or African American     Women
## 17      Bachelor's degree and higher Black or African American     Women
## 18            Bachelor's degree only Black or African American     Women
## 19                   Advanced degree Black or African American     Women
## 20   Less than a High School diploma                     Asian       Men
## 21 High School graduates, no college                     Asian       Men
## 22  Some college or associate degree                     Asian       Men
## 23      Bachelor's degree and higher                     Asian       Men
## 24            Bachelor's degree only                     Asian       Men
## 25                   Advanced degree                     Asian       Men
## 26   Less than a High School diploma                     Asian     Women
## 27 High School graduates, no college                     Asian     Women
## 28  Some college or associate degree                     Asian     Women
## 29      Bachelor's degree and higher                     Asian     Women
## 30            Bachelor's degree only                     Asian     Women
## 31                   Advanced degree                     Asian     Women
```

```
## 32  Some college or associate degree                     White       Men
## 33  Some college or associate degree Black or African American       Men
## 34  Some college or associate degree                     White     Women
## 35  Some college or associate degree Black or African American     Women
##     avg_val
## 1   736.550
## 2  1326.450
## 3  1223.500
## 4  1542.600
## 5   454.750
## 6   382.950
## 7   549.300
## 8   978.250
## 9   901.000
## 10 1120.800
## 11  580.750
## 12  978.100
## 13  904.550
## 14 1179.850
## 15  378.250
## 16  481.900
## 17  880.800
## 18  818.100
## 19 1011.350
## 20  535.375
## 21  687.000
## 22  840.875
## 23 1577.125
## 24 1374.375
## 25 1800.500
## 26  446.625
## 27  568.500
## 28  703.750
## 29 1210.625
## 30 1091.125
## 31 1407.125
## 32  861.300
## 33  669.300
## 34  641.700
## 35  567.750
```

```
attach(project_data)
```

```
high_nocol = as.numeric(education_text == 'High School graduates, no college')
bachandhigh = as.numeric(education_text == "Bachelor's degree and higher")
bachonly = as.numeric(education_text == "Bachelor's degree only")
adv = as.numeric(education_text == 'Advanced degree')
lessthanhigh = as.numeric(education_text == 'Less than a High School diploma')
white =  as.numeric(race_text == 'White')
black = as.numeric(race_text == 'Black or African American')
asian = as.numeric(race_text == 'Asian')
men = as.numeric(sexs_text == 'Men')
```

Creation of Model:

```
model = lm(avg_val ~ + high_nocol + bachandhigh
           + bachonly + adv + lessthanhigh
           + white + black + men)

summary(model)
```

```
##
## Call:
## lm(formula = avg_val ~ +high_nocol + bachandhigh + bachonly +
##     adv + lessthanhigh + white + black + men)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -162.57  -72.79   15.68   43.17  222.10
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    742.45      55.40  13.402 3.48e-13 ***
## high_nocol    -113.45      63.45  -1.788 0.085428 .
## bachandhigh    444.45      63.45   7.005 1.94e-07 ***
## bachonly       338.00      63.45   5.327 1.42e-05 ***
## adv            629.59      63.45   9.923 2.49e-10 ***
## lessthanhigh  -250.87      66.88  -3.751 0.000892 ***
## white         -116.42      46.09  -2.526 0.017973 *
## black         -278.13      44.86  -6.199 1.47e-06 ***
## men            206.37      37.30   5.532 8.30e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 109.9 on 26 degrees of freedom
## Multiple R-squared:  0.9328, Adjusted R-squared:  0.9122
## F-statistic: 45.13 on 8 and 26 DF,  p-value: 2.616e-13
```

```
anova(model)
```

```
## Analysis of Variance Table
##
## Response: avg_val
##              Df  Sum Sq Mean Sq F value    Pr(>F)
## high_nocol    1  638081  638081  52.837 1.016e-07 ***
## bachandhigh   1  301498  301498  24.966 3.391e-05 ***
## bachonly      1  171287  171287  14.184 0.0008578 ***
## adv           1 2209374 2209374 182.951 2.825e-13 ***
## lessthanhigh  1  205534  205534  17.020 0.0003368 ***
## white         1    1002    1002   0.083 0.7755584
## black         1  464135  464135  38.433 1.475e-06 ***
## men           1  369591  369591  30.605 8.299e-06 ***
## Residuals    26  313985   12076
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```